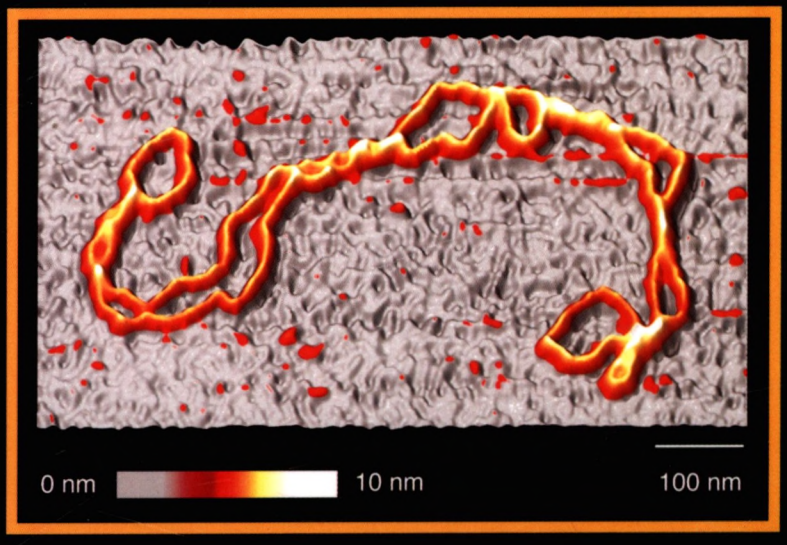
AN INTRODUCTION TO BIOPHYSICS

WITH MEDICAL ORIENTATION

EDITED BY

G. RONTÓ AND I. TARJÁN



1828

KIADÓ, BUDAPEST

AN INTRODUCTION TO BIOPHYSICS

WITH MEDICAL ORIENTATION

AN INTRODUCTION TO BIOPHYSICS

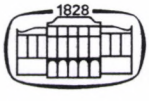
WITH MEDICAL ORIENTATION

EDITED BY

G. RONTÓ AND † I. TARJÁN

CONTRIBUTORS

L. BERKES, S. GYÖRGYI, G. RONTÓ, † I. TARJÁN



AKADÉMIAI KIADÓ, BUDAPEST

This book is the translated version of the revised and enlarged Hungarian original
A BIOFIZIKA ALAPJAI
published by Semmelweis Kiadó, Budapest 1997

Translated by

† Z. Morlin, † M. Mátrai and J. Kerpel-Fronius

Translation revised by

D. Durham, †M. Mátrai and I. Voszka

Figures constructed by

K. Módos

Mathematical Appendix done by

L. Herényi

Cover page

Scanning force microscopic image of a 3.7 kb superhelical plasmid DNA measured in aqueous solution (courtesy of Dr. Karsten Rippe, DKFZ Heidelberg)

ISBN 963 05 8088 8

First English edition: 1987
Second revised and enlarged English edition: 1991
Third revised and enlarged English edition: 1999
Fourth English edition: 2003

Printed in Hungary

CONTENTS

PREFACE

Just like in the past, the new edition is at the same time a revised one. In some chapters the revision means only reedition, while in others it consists of additions or updating. In the 3rd edition almost all of the chapters have been more or less rewritten; it is worthwhile to mention as examples especially the part of Chapter 5 dealing with thermodynamics and the complete Chapter 6 which changed also its title and became the host of medical electronics.

Moreover, Chapter 4 became updated by the treatment of the new types of microscopes. These instruments present a great challenge and new possibilities for research and practice alike. While the earlier "high-resolution" diffraction and spectroscopic structure-examining methods make possible the joint study of a population of molecules, by means of the more recent types of microscopes single molecules may be "dealt with". Thus the former methods allow of the statistical treatment of the molecular parameters, the latter render possible the examination of individual molecules.

Further modernizations are the inclusion of the environmental aspects of ultraviolet radiation based on the definitions of WHO/CIE, and the revision of the section on radiation protection with consideration to the most recent ICRP recommendations. By emphasizing these parts we want to serve the preventive aspects of medicine. Namely at the present stage of industrialization it is necessary to get acquainted with the role, importance and potency of environmental factors damaging our health and the whole biosphere, in order to establish preventive medicine.

The Appendix has been extended, too. In the present edition, beside the mathematical topics, the treatment of a few physical problems with medical relevance is also included. Among them there are some which previously belonged to the "stock" and have now been transferred to the Appendix, and there are others which had been treated in the early Hungarian editions, were later omitted in order to reduce the size of the book but are now reincorporated in agreement with the opinion of some of our colleagues. The change in the preliminary knowledge of our students plays also a role in all these.

The major part of the physical, chemical, biological and mathematical rudiments and relations is not included in this book; these are taken for granted to an extent determined by our experience, taking into consideration the high- school curricula.

All these did not change the nature of the book: its purpose is to serve the training of medical students also in the future, in the hope, however, that it shall be able to satisfy the occasional minor demands of physicians (biologists) and physicists even in the postgradual period.

The illustrations of the volume are partly new, partly were prepared on the basis of those of the previous editions, making use of computer graphics.

Our thanks are due to our colleagues, Drs L. Fedina, J. Fidy, I. Préda, K. Blaskó, N. Rozlosnik, M. Szőgyi, K. Tóth and I. Voszka, for their remarks and advise concerning certain details. We are especially grateful to K. Módos and Dr. I. Voszka for their careful participation in the editing. The cooperative administrative and technical help of Mrs. Gy. Bányay and Mrs. É. Matus-Stein is appreciated also with regard to this edition.

Budapest, 1999.

The Authors

1. THE STRUCTURE OF MATTER. MOLECULAR BACKGROUND OF STRUCTURE AND FUNCTION

The relation between the structure and function of matter is a fundamental problem of science. Research toward this relation became of particular importance in biology as soon as up-to-date and sophisticated methods of structure analysis were elaborated, leading to a better understanding of the functioning of living organisms; knowledge in this field has been considerably extended by the use of optical and electron microscopes. The development of the most recent scanning microscopes shall certainly further improve our knowledge. It is no exaggeration to say that the profound biological studies may be extended to molecular and even atomic levels. Figure 1.1 not only summarizes the progress in the structural investigations, but also demonstrates the historical fact that new methods and a deeper insight usually call for new disciplines; these are not restricted to the emergence of a new branch of science, but are usually connected with a new approach to reasoning, and with a new scientific attitude.

The considerable development of physics in recent decades has led to a great deal of new knowledge about subatomic structures. However, since these are stable from a biological aspect, we shall discuss subatomic processes only occasionally in this chapter.

Every body, living or inanimate, is built up from elements (atoms). Molecules and complex systems of molecules, and the organization within these, are determined by

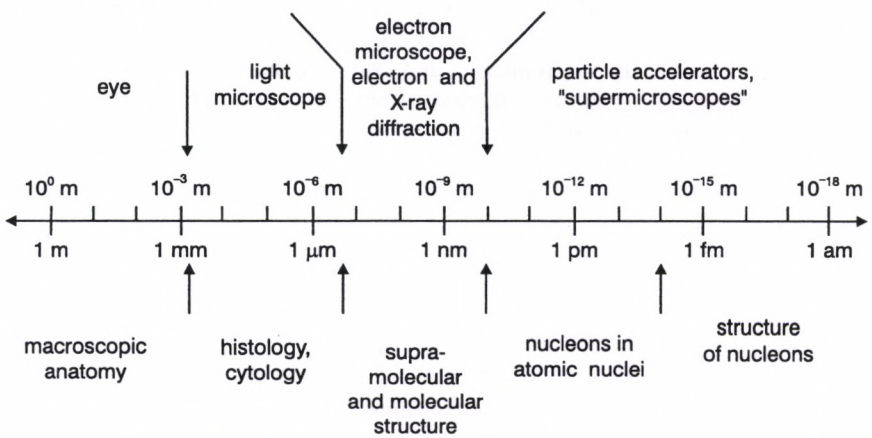


Fig. 1.1. Some of the more important stages of structure analysis
Above the scale the device or method is shown; arrows indicate the smallest detectable size.
Below, examples are given for the characterization of these details

atomic and molecular interactions, respectively. The various systems display the general properties of matter, but they also have special properties due to their composition and organization. On Earth life developed from inanimate nature over millions of years, mainly through particular combinations and interactions of light atoms. Consequently, living matter shows both *general* and *special* properties. The special properties account for the phenomenon called life. In this chapter we shall deal mainly with the general properties of matter, though some of the biological aspects will also be mentioned. The biologically important methods of structure analysis will be discussed in Chapter 4.

1.1. The basic forms of matter

The knowledge related to the basic properties of matter has increased considerably in the first decades of this century. Much earlier, the most important general characteristic of matter was considered to be its *corpuscular structure,* and the only type of matter was thought to be the chemical (mechanical) substance built up from atoms. This conformed well with the later realization that even the atom has a complex structure: it is comprised of positively charged protons and neutral neutrons forming the atomic nucleus, and negatively charged electrons surrounding the nucleus in shells. According to our present knowledge the *force fields* (gravitational, electromagnetic fields, force fields between nucleons, etc.) acting between microparticles, and also between the macroscopic bodies built up from these particles, are special forms of matter, too. These two forms of matter, fields and corpuscular particles, are of equal importance and complement each other. Without any aim at completeness, a few empirical facts may be mentioned to demonstrate this.

(a) Under certain conditions, physical fields display corpuscular properties. As an example, the electromagnetic field, or more exactly light, may be mentioned. In some phenomena light behaves as an electromagnetic wave, but in its interactions with molecules, atoms or electrons light behaves as a corpuscle. Thus, light is both a wave and a corpuscle. Light corpuscles of given energy, mass and momentum, possessing simultaneously a well-defined frequency and wavelength, are called *light quanta* or *photons.* The energy E, mass m and momentum I of a photon of frequency ν and wavelength $\lambda = c/\nu$ are expressed by the relations

$$E = h\nu, \qquad [1.1a]$$

$$m = \frac{h\nu}{c^2} \qquad [1.1b]$$

$$I = \frac{h\nu}{c} = \frac{h}{\lambda} \qquad [1.1c]$$

where h ($= 6.63 \times 10^{-34}$ Js) is Planck's constant, also called the action quantum, and c ($= 3 \times 10^8$ ms^{-1}) is the velocity of light.

Relation [1.1a] stated by Planck is one of the greatest scientific achievements giving rise in our century to the quantum conception in natural sciences. — [1.1b] expresses the

relation of Einstein between mass and energy according to which a mass m corresponds (is equivalent) to an energy mc^2. — [1.1c] refers to the fact that momentum is given by the product of mass and velocity.

The interaction between nucleons inside the nuclei is a result of the nuclear force field, which is ineffective for other particles (e.g. electrons or photons). The quanta of the nuclear force field are the *π-mesons*, also called *pions*. While the photon exists only at the velocity of light, with zero rest mass, the mass of the pion, similarly to those of electrons or nucleons, is not zero. The rest mass of the pion is about 270 times larger than that of the electron. According to the present knowledge, even the nucleons are not the "most elementary" particles: they are built up from the *quarks*, the interaction between which being mediated by *gluons*.

(b) Under certain conditions, corpuscular particles display wave properties, which are otherwise features of physical fields. If, for instance, a radiation of high-velocity electrons passes through a thin metal film, a photographic plate placed behind this film will show interference patterns similar to those generated when X-rays are transmitted through the same object. The so-called *matter wave* belonging to a particle of momentum I has the wavelength (λ)

$$\lambda = \frac{h}{I} \qquad [1.2]$$

which corresponds to [1.1c].

(c) According to classical physics, particles and force fields may transform into each other. If, for instance, a positively charged electron (positron) collides with a negatively charged electron, the electron pair transforms into electromagnetic radiation, or more exactly into γ-photons (in most cases two γ-photons are produced in this process). The reverse process is also known: γ-radiation may produce electron pairs (pair formation).

Positrons are sometimes referred to as antielectrons, and the *positron–electron pair* is then called an *antiparticle pair*. Antipairs are pairs of particles whose mass, intrinsic angular momentum or spin, magnetic moment, electric charge and other characteristic data are identical in absolute values, but opposite in sign. The antiparticles of the proton, the neutron and the neutrino are known: the antiproton, the antineutron and the antineutrino. (The photon is identical with its antiparticle.) It is generally true that when a particle collides with its antiparticle both are annihilated.

If only antiparticles are present they behave as common particles. Thus, atoms with a nucleus consisting of antiprotons and antineutrons and a shell containing positive electrons (antiatoms) are conceivable. Similarly to antiatoms, antimolecules and anticrystals may exist, in exactly the same way as common molecules or bodies are built up by common atoms. All of the macroscopic properties of antibodies would be the same as the properties of common bodies. Under the conditions on the Earth, there is no possibility for antibodies to exist. If an antiparticle is somehow created, within a very short time it encounters its corresponding particle and both are annihilated. In principle it is not impossible that antimatter does exist somewhere in the Universe, but there is no definite evidence of it.

(d) In all of its forms matter has mass and energy, and may have also momentum and angular momentum. These quantities are general characteristics of matter. In the different processes of matter (interactions, transformations) exchanges of mass, energy and momentum may occur, though the *laws of conservation* of mass, energy, momentum,

angular momentum and electric charge remain valid. A generally accepted, consistent picture of matter has been developed by accepting that *matter is a corpuscular particle and a field. The particles are the quanta of the fields.*

Electrons, nuclei, photons and pions belong to the family of *elementary particles.* At present approximately 200 elementary particles are known, but the above particles play the primary roles in atomic processes.

1.2. Atoms

1.2.1. The principal characteristics of quantum theory (quantum mechanics)

The basic property that matter is both a particle and a field is frequently expressed by the term *wave particle.* Wave particles are not comparable with any physical body of everyday life. The wave and the particle natures of matter can be described separately by classical models, but attempts to reflect the reality of unified wave particles by some similar descriptive model have remained unsuccessful. In spite of this, a theory has been developed which describes the behaviour of the smallest particles of matter, especially the events on an atomic scale. This theory, the *quantum theory,* also permits an understanding of numerous macroscopic properties of matter.

The earliest developed and basic part of quantum theory is *quantum mechanics,* which deals with the laws of motion of atomic particles. Quantum mechanics allowed an understanding of the structures of atoms, molecules and solids, and its importance is continually increasing as concerns the interpretation of chemical and biological processes *(quantum chemistry, quantum biology).* For this reason, mainly quantum mechanics will be referred to in the following discussions.

Another, similarly important branch of quantum theory embracing electromagnetic phenomena is *quantum electrodynamics.* The most up-to-date interpretation of the laws of motion of matter, unifying both quantum mechanics and quantum electrodynamics, is provided by the *quantum-field theory.*

Quantum mechanics sets out from the dualistic behaviour of matter, and draws its conclusions by means of mathematical reasoning, an understanding of which requires deep mathematical knowledge. Discussion of the higher mathematics involved would exceed the scope of this book; only a few results will be summarized.

It is a consequence of the principles of quantum mechanics that the electrons in the atom cannot be in arbitrary states: their energy and angular momentum can assume only well-defined values, which means that these quantities can change only by well-defined *quantized* values. (The name quantum mechanics refers to this property.) With the aid of quantum mechanics the frequencies of emitted and absorbed radiation can be obtained in accordance with experience. Similarly, the intensity and polarization of radiation are calculable. The theory explains the rotational and vibrational properties of molecules, the atomic, electronic and molecular interactions, and the chemical bonds; together with the interpretation of the periodic system and the most important optical, electric and

magnetic properties of solids, this constitutes an especially impressive result of quantum mechanics.

Beside the insufficiency of the simple visualization, another main property of quantum mechanics is, in many cases, its probability character. It is not possible to define the exact localization of an electron within the atom; only the probability distribution of its position can be given. However, this is not a deficiency of the theory, but a specific property of the microworld. It follows that the electrons do not actually revolve around their nucleus, as the planets in the solar system do, for instance. No atomic orbits exist in this sense, though the expression *atomic orbital* is used for a mathematical function, the *wave function* derived by quantum mechanical methods. Various physical quantities describing the state of the electrons can be calculated by means of this function.

Quantum mechanics should be regarded as a theory which does not simply extend our knowledge, but leads to more universal natural laws than the laws of classical mechanics. It is generally accepted that classical mechanics can be regarded as a special case of quantum mechanics. This generalization can be advanced even further by stating that it is valid for the relation of quantum physics and classical physics as well.

Finally one more remark. The importance of quantum mechanics (quantum chemistry) is still being increased by new informations and applications. A prospective direction of research is, for example, the quantum-chemical characterization of the active atomic groups of the drugs, thereby the prediction of effective compounds, or the exploration of certain pathological mechanisms, in general, *the investigation of the relationship between the structure and the effect of compounds.*

1.2.2. Quantum numbers

1. Electron shells, principal quantum number. According to quantum mechanics, and in accordance with experience, several characteristics of the atomic electrons (in general the bound electrons), such as energy and angular momentum, do not vary continuously, but in a stepwise way; these properties are said to be quantized, and the electron states and their changes are described by discrete numbers. The states of the atomic electrons are determined by four numerical data, called *quantum numbers:* the *principal quantum number* (n), the *angular orbital momentum* (l), the *spin* (s) and the *magnetic quantum number* (m). Any change of state is connected with a change in at least one quantum number.

The possible states of the electrons, the atomic orbitals, show a well-defined arrangement: they form *shells* around the nucleus. The shells are at different distances from the nucleus, and are denoted (moving outwards from the nucleus) by the capital letters *K, L, M, ... The principal quantum numbers of electrons in the same shell are identical.* The principal quantum numbers of the electrons in the *K, L, M, ...* shells are successively the integers 1, 2, 3, ...

2. Angular orbital momentum. The angular orbital momentum quantum number associated with a given principal quantum number specifies the possible values of the *angular momentum* of the electron. The name refers to the outdated picture of electrons rotating around the nucleus on a circular or elliptical orbit, as had been conceived by

Bohr and Sommerfeld. According to this older theory the angular momentum is equal to the product of the momentum of the rotating electron and its distance from the point-like nucleus. Quantum mechanics has proved that this picture is incorrect, though it is still considered that the electron within the atom does possess an angular momentum; however, this does not originate from rotation. For a given n the possible values of the orbital angular momentum quantum number are

$$l = 0, 1, 2, 3, \ldots, (n-1) \qquad [1.3]$$

Accordingly, for any given principal quantum number the *s, p, d, f,* ... states exist. For instance, electrons characterized by the quantum number $n = 2$ and $l = 0$ are $2s$ electrons, and the $3p$ electrons have the quantum numbers $n = 3$ and $l = 1$. The magnitude of the orbital momentum for orbital quantum number l is given by

$$\hbar\sqrt{l(l+1)}$$

where $\hbar = h/2\pi$; h is the Planck constant.

3. Spin. For an understanding of the empirical facts it has to be assumed that, besides the orbital momentum, the electron has an additional angular momentum, due to its spinning around its symmetry axis. This momentum is the *spin,* and the electron is thought of as a small gyroscope. Though this model is incorrect, the existence of spin has been proved. (Incidentally, not only electrons, but also other elementary particles, such as photons and nucleons, possess spin.) The quantum number characterizing the electron spin is

$$s = \frac{1}{2} \qquad [1.4]$$

and the magnitude of the spin is

$$\hbar\sqrt{s(s+1)}$$

4. Magnetic quantum number. The angular momentum represents a vector with a corresponding direction, which must be considered whenever a preferred orientation in space exists, induced for instance by an external magnetic or electric field.

The only directions of the angular momentum of the atomic electrons that are permissible are those where the projection of the orbital momentum on the preferred direction is $m_l\hbar$, and the projection of the spin $m_s\hbar$ (Fig. 1.2). The quantities m_l and m_s are the *orientation or magnetic quantum numbers.* m_l may assume every integer value from $-l$ to $+l$, including zero:

$$m_l = -l, -(l-1), \ldots, -1, 0, +1, \ldots, +(l-1), +l \qquad [1.5]$$

The total number of possible values is consequently $2l + 1$. The quantum number m_s has only two values

$$m_s = -\frac{1}{2} \text{ or } +\frac{1}{2} \qquad [1.6]$$

The positive sign indicates the projection in the preferred direction, and the negative sign that in the opposite direction.

5. Internal quantum number. The electrons have a *magnetic moment* connected with their angular momentum. The atomic magnetic moment may be conceived as the magnetic moment of revolving or spinning electric charges. However, the direction of the magnetic moment in a magnetic field cannot be arbitrary; the possible directions are determined by directional quantization.

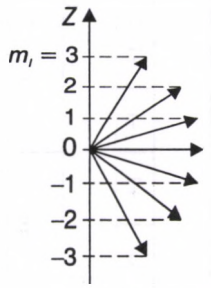


Fig. 1.2. Possible orientations of the orbital angular momentum with respect to a preferred direction (Z) for $l = 3$

The momentum due to the orbital momentum of an electron moving around the nucleus and to the spin can be added vectorially. The resultant total angular momentum is described by the *internal quantum number* (j), which can easily be obtained from [1.3] and [1.6]:

$$j = l \pm \frac{1}{2} \qquad [1.7]$$

In [1.7] the positive and negative signs (in agreement with experience) refer simply to the fact that the spin is only bidirectional with respect to the orbital angular momentum. Consequently, the spin either increases or decreases the orbital angular momentum quantum number by 1/2. Moreover, [1.7] states that the projection of the orbital angular momentum on any direction (in the present case on the direction of the resultant angular momentum) is an integral multiple of $\hbar$, and the projection of the spin may have only the value $1/2\ \hbar$.

Similarly as in the previous expressions, the magnitude of the resultant angular momentum is given by the quantity

$$\hbar \sqrt{j(j+1)}$$

6. Total magnetic quantum number. By introducing the quantity j, the orientation or magnetic quantum number is generalized in the sense that its permissible values are

$$m = -j, -(j-1), \ldots, +(j-1), +j \qquad [1.8]$$

Therefore, there are altogether $2(2l + 1)$ values (with the exception of $l = 0$, since in this case j has only one value). The generalized magnetic quantum number is also called *the total magnetic quantum number* (m).

For clarity, the quantum numbers defining the permitted electron states and the exact notations of these states for the values $n = 1$, 2 and 3 are listed in Table 1.1.

Table 1.1. Quantum states of a one-electron system

Principal quantum number (n)	1	2				3								
Orbital angular momentum quantum number (l)	0	0	1			0	1			2				
Spin quantum number (s)	1/2													
Magnetic quantum number (m_l)	0	0	0	−1	+1	0	0	−1	+1	−1	+1	0	−2	+2
(m_s)	−1/2 +1/2													
Internal quantum number (j)	1/2	1/2	1/2	3/2		1/2	1/2	3/2		3/2		5/2		
Total magnetic quantum number (m)	−1/2 +1/2	−1/2 +1/2	−1/2 +1/2	−3/2 −1/2	+1/2 +3/2	−1/2 +1/2	−1/2 +1/2	−3/2 −1/2	+1/2 +3/2	−3/2 −1/2	+1/2 +3/2	−1/2 +1/2	−5/2 −3/2	+3/2 +5/2
Shell notation	K	L				M								
State notation	1s	2s	2p			3s	3p			3d				
Number of states	2	2	6			2	6			10				
	$2 = 2 \times 1^2$	$8 = 2 \times 2^2$				$18 = 2 \times 3^2$								

1.2.3. Hydrogen atom

1. Simplified structure. Ground and excited states. As an example, *the energy level system of the hydrogen atom,* which has only one electron, will now be discussed. Figure 1.3 depicts the simplified structure, which is obtained when the spin of the electron is not taken into consideration in the quantum mechanical calculations. Even with this simplification, the emission and absorption spectra of hydrogen obtained with a spectroscope of medium resolution can be suitably interpreted. The horizontal lines denote the permitted energy states of the electron. In this case these states depend only on the principal quantum number. The lowest energy belongs to the principal quantum number $n = 1$; this is the *ground state.* The atom attains *excited states* of higher energy by energy uptake, e.g. by the absorption of photons or by collision with an electron or some other particle of appropriate energy. These processes are indicated in the diagram by arrows pointing upwards. Any transition from a higher to a lower energy state takes place by energy release, for instance by light emission or by some interaction with another particle. These transitions are shown by arrows pointing downwards. The energy values connected with the permitted transitions are presented on the vertical axes in eV units. With increasing n values the energy levels become more dense, and for $n = \infty$ an energy is obtained which corresponds to the removal of the electron from the influence of the nucleus; this process is called ionization. The amount of energy needed to ionize a hydrogen atom is 13.53 eV. If the energy uptake is higher, the removed electron has a kinetic energy equal to the excess energy. The energy of the free electron may have any arbitrary value. This is demonstrated in the diagram by the range denoted continuum. It is generally true for any particle in a bound state that its energy can have only certain definite values, but the energy of free particles may change continuously.

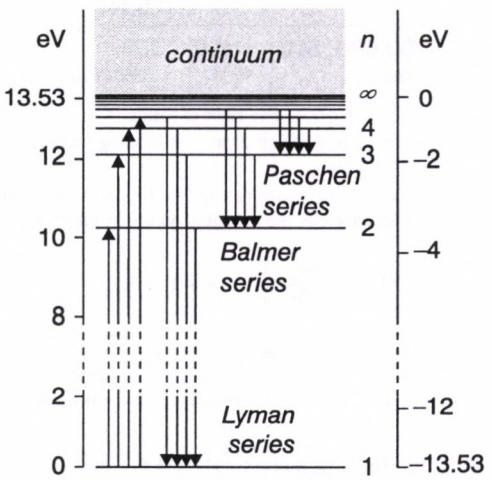


Fig. 1.3. The energy level system of the hydrogen atom
The energy of the ground state is indicated by zero on the left-hand ordinate, while the right-hand ordinate shows the energy necessary for ionization

The *binding energy* of an electron within the atom is defined as the energy required to remove the electron from the atom. It follows from the previous section that the binding energy is higher if the electron is closer to the nucleus, i.e. if it is located in an orbital of lower energy. Thus, the electrons in the K shell are the most strongly bound. For the hydrogen atom the binding energy of the electron in the K and L shells is 13.53 eV and 3.38 eV, respectively.

2. Series of spectral lines. The frequencies of the spectral lines are described by simple equations. The system of spectral lines given by the same mathematical relation is called a *series,* and the equation describing this series is the *series formula.* The better-known series for the hydrogen atom are as follows:

$$\nu_{1,n} = R\left(\frac{1}{1^2} - \frac{1}{n^2}\right), \quad n = 2, 3, 4, \ldots \textit{ Lyman series} \qquad [1.9a]$$

$$\nu_{2,n} = R\left(\frac{1}{2^2} - \frac{1}{n^2}\right), \quad n = 3, 4, 5, \ldots \textit{ Balmer series} \qquad [1.9b]$$

$$\nu_{3,n} = R\left(\frac{1}{3^2} - \frac{1}{n^2}\right), \quad n = 4, 5, 6, \ldots \textit{ Paschen series} \qquad [1.9c]$$

R is the *Rydberg constant;* its value is $3.29 \times 10^{15}\ s^{-1}$. If the integers n are substituted into the above formulae, the frequencies of the series are obtained. The frequency of the first line of the best-known Balmer series (Fig. 1.4) is $4.6 \times 10^{14}\ s^{-1}$, its wavelength is 656 nm, and the emitted energy is 1.9 eV. The lines of any series are initially at a greater distance from one another; they subsequently lie closer and closer to one another, and the Balmer series, for example, ends with $n = \infty$ at the frequency $\nu_{2,n} = R/4$ (*frequency limit*).

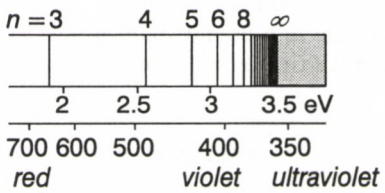


Fig. 1.4. The Balmer series
The domain of grey tone indicate the limiting continuum

The experimental facts can be interpreted in the following way. According to quantum mechanical calculations, the energy of an electron in the state characterized by the principal quantum number n is

$$E_n = -\frac{e^4 m_e}{8\varepsilon_0^2 h^2}\frac{1}{n^2}, \quad n = 1, 2, 3, \ldots, \qquad [1.10]$$

where e is the elementary electric charge, m_e is the mass of the electron and ε_0 is the permittivity of vacuum. The ordinate on the right side of Fig. 1.3 corresponds to the energy values calculated from [1.10]. In order to raise the electron from the $n = 1$ state into a state with $n > 1$, for instance, the energy needed is

$$E_n - E_1 = -\frac{e^4 m_e}{8\varepsilon_0^2 h^2}\left(\frac{1}{1^2} - \frac{1}{n^2}\right), \quad n = 2, 3, 4, \ldots \qquad [1.11]$$

and the same energy is liberated when the electron returns from the $n > 1$ state to the ground state $n = 1$. In the case of photon absorption the energy difference $E_n - E_1$ is of the form $h\nu_{1,n}$, which leads to the equation

$$\nu_{1,n} = -\frac{e^4 m_e}{8\varepsilon_0^2 h^3}\left[\frac{1}{1^2} - \frac{1}{n^2}\right], \quad n = 2, 3, 4, \ldots \qquad [1.12]$$

Naturally, [1.12] also gives the frequency of the emitted light. The similarity between the empirically obtained [1.9a] and [1.12] calculated by means of quantum mechanics, is obvious, and the insertion of the values of the constants yields the experimentally measured R value with satisfactory accuracy. Though [1.11] and [1.12] relate to the Lyman series, the spectral lines of the other series can be interpreted in a similar way.

3. Fine and hyperfine structures. If also the spin is introduced into the calculations, the energy no longer depends only upon the principal quantum number, but also upon the internal quantum number. Since j may assume various values for a given value of n, the single energy level associated with n splits into several sublevels. In this way *fine structure* is obtained. Figure 1.5 illustrates the splitting of the energy levels (also called terms) of the principal quantum numbers $n = 2$ and $n = 3$, respectively. (Some of the lines denoting the energy levels are actually split into two, but are drawn as single lines because the levels are too close to each other.)

The magnetic quantum number becomes important in the energy terms of the electrons only if an electron localized in the electric field of a nucleus is perturbed by some other force field. This occurs when the atom is placed in a magnetic or electric field (Zeeman effect, Stark effect). Under the influence of external force fields the energy levels are split further. However, even the atomic nucleus itself may perturb the electron state, since not only the electron, but also the nucleus may have a magnetic moment, for the nucleus may also possess mechanical angular momentum. The effect of the magnetic field of the nucleus is that *hyperfine structure* is obtained. The expressions fine structure and hyperfine structure, the latter displaying even finer details, stem from spectroscopic studies, the results of which have been interpreted with high precision by the quantum mechanical method, as mentioned above.

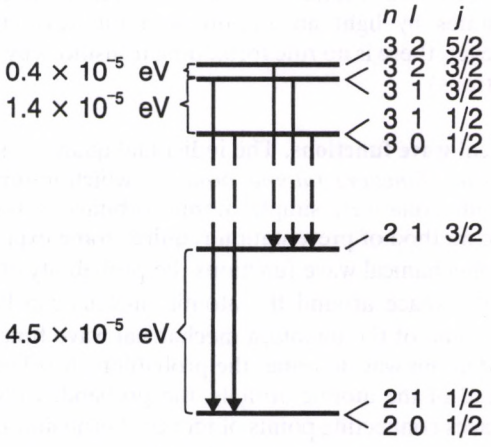


Fig. 1.5. The fine structure of the energy levels of the hydrogen atom for $n = 2$ and $n = 3$

4. Selection rules. Metastable states. It is found in practice that the emission and absorption spectra consist of fewer lines than expected from the calculated energy levels. This fact too can be accounted for by quantum mechanics. In an interaction with photons any change in the electron states is restricted by *selection rules.* No restriction exists for the principal quantum number (Fig. 1.3), but the orbital momentum quantum number can change only by one, and the total internal and magnetic quantum numbers must either be unchanged, or change by one. Consequently

$$\Delta n = \text{any arbitrary integer}$$

$$\Delta l = \pm 1; \quad \Delta j = 0 \pm 1; \quad \Delta m = 0 \text{ or } \pm 1 \qquad [1.13]$$

The arrows in Fig. 1.5 indicate the permitted transitions of light emission. The selection rules account for physical properties connected with interactions between the atomic system and the angular momentum of the photon. The angular momentum of atoms has already been discussed. The photon also has spin: $s = 1$. Only those interactions (emission and absorption) exist which are permitted by the law of conservation of angular momentum. The possible transitions are given by the selection rules. There is one essential condition which deserves consideration. Intense emission and absorption of radiation are possible only if the dipole moment of the system is changed in the transition. It can be demonstrated that the transitions permitted by the selection rules conform to this condition. — Those excited states from which transition to a lower energy level is forbidden by any of the selection rules are *metastable states.* The restrictions must be interpreted statistically, which means that the probability of the transition is small. The time for spontaneous decay from a normal excited state to a lower-energy one is approximately 10^{-8} s, whereas the life-time of the metastable states is longer by several orders of magnitude, it may be seconds or even more. The atom cannot be brought into metastable excited states by light absorption, and the restriction is thus valid for absorption, too. However, there is no rule forbidding transitions by collision (e.g. collision with accelerated electrons).

5. Quantum mechanical wave functions. The individual quantum states are described by *quantum mechanical wave functions (atomic orbitals),* which in turn are functions of the space coordinates. Some relatively simple atomic orbitals of the hydrogen atom are shown in Fig. 1.6. The method of presentation requires some explanation.

With the quantum mechanical wave functions the probability of localizing an electron at some position in the space around the atomic nucleus can be calculated. For this purpose the absolute value of the quantum mechanical wave function must be squared. The function obtained in this way describes the probability distribution of the localisation of the electron. Instead of the atomic orbitals, the probability distributions are usually demonstrated by surfaces connecting points of identical probabilities around the nucleus. The various probabilities are represented by the different surfaces. Of course, surfaces may be selected within which an electron in a given state is localized with some fixed probability (e.g. 0.5 or 0.9). For simplicity, not only the wave function is called an atomic

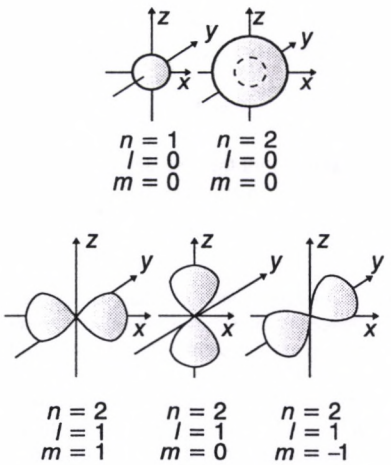


Fig. 1.6. Some atomic orbitals of the hydrogen atom

orbital (especially in connection with chemical or biological applications), but also the surface within which an electron in a given quantum state may be found with a probability of 0.9. These types of surfaces for the hydrogen atom are presented in Fig. 1.6. The *s* orbitals have spherical symmetry, the radius of the sphere containing the 1 *s* electron with a probability 0.9 is 140 pm. The probability of finding the electron outside this sphere is only 0.1, but in principle becomes zero only at an infinite distance from the nucleus. The probability of finding the electron in the 2*s* state vanishes at a distance of 106 pm from the nucleus (this spherical nodal surface is indicated by a dotted circle). In the 2*s* state the radius of the sphere within which the electron is to be found with a probability of 0.9 is 320 pm. The 2*p* orbitals show the symmetry of a rotational body resembling a figure 8. The axes of rotation for $m_l = +1, -1$ and 0 are the *X*, *Y* and Z axes of the coordinate system. The regions of zero probability are on the coordinate planes normal to the actual axis of rotation.

If the probability of localization is multiplied by the charge of the electron, the spatial distribution of the electron charge is obtained. Accordingly, the charge of an electron may be regarded as a charge cloud distributed in the space around the nucleus, and the actual charge density within this cloud is proportional to the probability of finding the electron at the given place.

6. Many-electron systems. So far, only the motion of a single electron moving in the force field of the atomic nucleus has been discussed (one-electron system). In a *many-electron system,* however, it is not sufficient to consider only the nucleus to describe the motion of an electron, since the effect of the other electrons, the electron coupling, must also be taken into account. From the structure of optical spectra it may be inferred that for most atoms the so-called *LS-coupling (Russel–Saunders coupling)* is valid. This involves the calculation of the resultant of the orbital angular momentum and the spin momentum (**L** and **S**, respectively) for every electron separately; subsequently, the total angular

momentum (**J**) of the atomic electron shell is derived from these quantities. The atomic properties are determined by the vectors **L, S** and **J,** where

$$\mathbf{J} = \mathbf{L} + \mathbf{S} \qquad [1.14]$$

The quantum numbers L, S and J associated with the resultant vector are usually denoted by capital letters. A relation similar to that for the one-electron system also holds for many-electron systems:

$$J = L + S \qquad [1.15]$$

L is always an integer, and quantum states corresponding to the quantum numbers $L = 0, 1, 2, 3, \ldots$ are denoted by the capital letters $S, P, D, F, \ldots$ (Generally, the states of one-electron systems are described by small letters, whereas capital letters are used for many-electron systems. The S notation of the resultant spin quantum number should not be confused with the S notation of the $L = 0$ state.) Naturally, in the case of many-electron systems the resultant spin quantum number is not confined only to 1/2; it may be zero or a multiple of 1/2. As an example, consider the spins of two electrons which may be parallel or antiparallel. In the first case $S = 1$, and in the second $S = 0$. For three electrons the permissible values of S are 3/2 and 1/2, respectively. Since S and L may assume various values in relation to each other, J may assume various values for given L and S. If $L \geq S$:

$$J = L + S, L + (S - 1), \ldots, L - (S - 1), L - S \qquad [1.16a]$$

and for $L \leq S$:

$$J = S + L, S + (L - 1), \ldots, S - (L - 1), S - L \qquad [1.16b]$$

Thus, J can assume $2S + 1$ and $2L + 1$ values, respectively. This number, called the *multiplicity,* refers to the multiple splitting of the energy levels for given L and S values. With $S = 1/2$ and $L = 0$ the multiplicity is 1, but for every other L value it is 2. These are the doublet levels (including the case $L = 0$) observed in the hydrogen atom. With $S = 0$ the levels are singlets, with $S = 1$ triplets, and so on. The selection rules of the optical transitions are

$$\Delta L = 0 \text{ or } \pm 1; \quad \Delta S = 0; \quad \Delta J = 0 \text{ or } \pm 1 \qquad [1.17]$$

Calculation of the resultant momentums is facilitated by the fact that the electrons in the fully occupied shells (see section 1.2.4) can be neglected, since both the resultant orbital momentum and spin momentum are zero for these. This also holds for the electrons in the s and p states within a given shell, if these electrons constitute a closed system (saturated subshell). Consequently, only the electrons outside the saturated shells (subshells) need to be taken into account.

1.2.4. The periodic system

In the periodic system (Table 1.2; cf. pp. 30–31) the chemical elements are arranged in a sequence of increasing electric charge of the nucleus or the electron shells. The *atomic number* of an element gives the number of elementary charges in the nucleus, or in the electron shells of the neutral atom which agrees with the number of protons in the nucleus and with that of the electrons in the shells. The atomic electron shell of an element in the periodic system is obtained by adding a further electron to the previous element in the system. In an atom in the ground state, every electron occupies the lowest possible energy state. It might be expected that in the ground state every electron of the atom would be

situated in the K shell. However, though some of them actually are in this shell, their number is limited by the *Pauli exclusion principle,* which states that no two electrons in an atom can simultaneously have all four quantum numbers identical. The maximum number of electrons with the same principal quantum number n is $2n^2$ (cf. Table 1.1). It follows that the K shell contains at most 2 electrons, the L shell 8 electrons, the M shell 18 electrons, and so on. Table 1.3 (cf. p. 32) lists the electron shell structures of the first 54 elements, giving the numbers of *s, p, d,* etc. electrons in the individual shells.

The expression periodic system of elements refers to the fact that several essential properties of the elements vary periodically with the atomic number. For instance, there is a striking similarity in chemical behaviour and in the nature of the optical spectra of the elements in a given column (multiplicity, cf. section 1.2.3), since the number and arrangement of the electrons in the outermost shell are identical. Thus, the alkali metals have one electron, the alkaline earth metals two electrons, and the halogens seven electrons in the outermost shells of the neutral atoms. These are the *valence electrons* (also called photoelectrons), which determine the type of the optical spectrum. The electrons occupying lower electron shells are more strongly bound to the nucleus; these are the *core electrons,* which together with the nucleus form the atomic core. The differentiation of electrons into valence and core electrons is justified by the fact that the atomic core contains full (closed) shells, whereas the valence electrons occupy partly filled shells. As pointed out in section 1.2.3, the full shells have minimum energy states and the resultants of the angular momentum and the spin, and consequently of the magnetic moment are zero. For this reason the full shells are stable and virtually indifferent. This is also true within a given shell for electrons in the s and p states, for the saturated s and p subshells are closed systems. It clearly follows that the "activity" of an atom is predominantly due to electrons outside the saturated shells (subshells), and the resistance of the inert gases against chemical effects is understandable.

1.3. Molecules

1.3.1. Chemical bonds. Bond energies

Interactions between atoms are transmitted mainly by the valence electrons. Though several bond types exist, in every case the outermost electron shell is transformed so that the system attains a *stable state characterized by an energy minimum.* This holds for the formation of molecules and for the interactions between them.

1. Bond types. The binding in *heteropolar* or *ionic compounds* (e.g. NaCl, CaO) results when one electron or more are transferred from one atom to another. The atom losing the electrons is transformed into a positive ion (cation), while the atom to which they are transferred is converted into a negative ion (anion). For instance, in the formation of the NaCl molecule one electron of the Na atom is transferred to the Cl atom. In this process energy is released, and Na^+ and Cl^- ions are formed. Cations and anions attract each other by electrostatic (Coulomb) forces.

Table 1.2. Periodic system of elements*

Ia	IIa	IIIb	IVb	Vb	VIb	VIIb		VIIIb
1 H Hydrogen 1.0								
3 Li Lithium 6.9	4 Be Beryllium 9.0							
11 Na Sodium 23.0	12 Mg Magnesium 24.3							
19 K Potassium 39.1	20 Ca Calcium 40.1	21 Sc Scandium 45.0	22 Ti Titanium 47.9	23 V Vanadium 50.9	24 Cr Chromium 52.0	25 Mn Manganese 54.9	26 Fe Iron 55.8	27 Co Cobalt 58.9
37 Rb Rubidium 85.5	38 Sr Strontium 87.6	39 Y Yttrium 88.9	40 Zr Zirconium 91.2	41 Nb Niobium 92.9	42 Mo Molybdenum 95.9	43 Tc Technetium 98.9	44 Ru Ruthenium 101.1	45 Rh Rhodium 102.9
55 Cs Cesium 132.9	56 Ba Barium 137.3	57 La Lanthanum 138.9	72 Hf Hafnium 178.5	73 Ta Tantalum 180.9	74 W Tungsten 183.9	75 Re Rhenium 186.2	76 Os Osmium 190.2	77 Ir Iridium 192.2
Lanthanides:			58 Ce Cerium 140.1	59 Pr Praseodymium 140.9	60 Nd Neodymium 144.2	61 Pm Promethium 146.9	62 Sm Samarium 150.4	63 Eu Europium 152.0
87 Fr Francium 223	88 Ra Radium 226.0	89 Ac Actinium 227.0	104 257	105 257	106 263	107 262	108 264	109 266
Actinides:			90 Th Thorium 232.0	91 Pa Protactinium 231.0	92 U Uranium 238.0	93 Np Neptunium 237	94 Pu Plutonium 239	95 Am Americium 241

* The number beside the chemical symbol is the atomic number and the number below the name is the atomic mass. In the case of transuranium elements the mass number of a produced isotope is given instead of the atomic mass. At present there is still no international convention for the name of elements with atomic number above 103. Their mass number values are also uncertain.

							VIIIa
							2 He Helium 4.0
			IIIa	IVa	Va	VIa	VIIa	
			5 B Boron 10.8	6 C Carbon 12.0	7 N Nitrogen 14.0	8 O Oxygen 16.0	9 F Fluorine 19.0	10 Ne Neon 20.2
			13 Al Aluminum 27.0	14 Si Silicon 28.1	15 P Phosphorus 31.0	16 S Sulfur 32.1	17 Cl Chlorine 35.5	18 Ar Argon 39.9
	Ia	IIb						
28 Ni Nickel 58.7	29 Cu Copper 63.5	30 Zn Zinc 65.4	31 Ga Gallium 69.7	32 Ge Germanium 72.6	33 As Arsenic 74.9	34 Se Selenium 79.0	35 Br Bromine 79.9	36 Kr Krypton 83.8
46 Pd Palladium 106.4	47 Ag Silver 107.9	48 Cd Cadmium 112.4	49 In Indium 114.8	50 Sn Tin 118.7	51 Sb Antimony 121.8	52 Te Tellurium 127.6	53 I Iodine 126.9	54 Xe Xenon 131.3
78 Pt Platinum 195.1	79 Au Gold 197.0	80 Hg Mercury 200.6	81 Tl Thallium 204.4	82 Pb Lead 207.2	83 Bi Bismuth 209.0	84 Po Polonium 210	85 At Astatine 210	86 Rn Radon 222

64 Gd Gadolinium 157.3	65 Tb Terbium 158.9	66 Dy Dysprosium 162.5	67 Ho Holmium 164.9	68 Er Erbium 167.3	69 Tm Thulium 168.9	70 Yb Ytterbium 173.0	71 Lu Lutetium 175.0

96 Cm Curium 247	97 Bk Berkelium 247	98 Cf Californium 251	99 Es Einsteinium 254	100 Fm Fermium 257	101 Md Mendelevium 257	102 No Nobelium 255	103 Lr Lawrencium 257

Table 1.3. Electronic structures of the first 54 elements of the periodic system

Z		K	L		M			N			O	
		s	s	p	s	p	d	s	p	d	s	p
1	H	1										
2	He	2										
3	Li	2	1									
4	Be	2	2									
5	B	2	2	1								
6	C	2	2	2								
7	N	2	2	3								
8	O	2	2	4								
9	F	2	2	5								
10	Ne	2	2	6								
11	Na	2	2	6	1							
12	Mg	2	2	6	2							
13	Al	2	2	6	2	1						
14	Si	2	2	6	2	2						
15	P	2	2	6	2	3						
16	S	2	2	6	2	4						
17	Cl	2	2	6	2	5						
18	Ar	2	2	6	2	6						
19	K	2	2	6	2	6		1				
20	Ca	2	2	6	2	6		2				
21	Sc	2	2	6	2	6	1	2				
22	Ti	2	2	6	2	6	2	2				
23	V	2	2	6	2	6	3	2				
24	Cr	2	2	6	2	6	5	1				
25	Mn	2	2	6	2	6	5	2				
26	Fe	2	2	6	2	6	6	2				
27	Co	2	2	6	2	6	7	2				
28	Ni	2	2	6	2	6	8	2				
29	Cu	2	2	6	2	6	10	1				
30	Zn	2	2	6	2	6	10	2				
31	Ga	2	2	6	2	6	10	2	1			
32	Ge	2	2	6	2	6	10	2	2			
33	As	2	2	6	2	6	10	2	3			
34	Se	2	2	6	2	6	10	2	4			
35	Br	2	2	6	2	6	10	2	5			
36	Kr	2	2	6	2	6	10	2	6			
37	Rb	2	2	6	2	6	10	2	6		1	
38	Sr	2	2	6	2	6	10	2	6		2	
39	Y	2	2	6	2	6	10	2	6	1	2	
40	Zr	2	2	6	2	6	10	2	6	2	2	
41	Nb	2	2	6	2	6	10	2	6	4	1	
42	Mo	2	2	6	2	6	10	2	6	5	1	
43	Tc	2	2	6	2	6	10	2	6	5	2	
44	Ru	2	2	6	2	6	10	2	6	7	1	
45	Rh	2	2	6	2	6	10	2	6	8	1	
46	Pd	2	2	6	2	6	10	2	6	10		
47	Ag	2	2	6	2	6	10	2	6	10	1	
48	Cd	2	2	6	2	6	10	2	6	10	2	
49	In	2	2	6	2	6	10	2	6	10	2	1
50	Sn	2	2	6	2	6	10	2	6	10	2	2
51	Sb	2	2	6	2	6	10	2	6	10	2	3
52	Te	2	2	6	2	6	10	2	6	10	2	4
53	I	2	2	6	2	6	10	2	6	10	2	5
54	Xe	2	2	6	2	6	10	2	6	10	2	6

The binding in *covalent* or *homopolar* compounds (e.g. most organic compounds) has been established by quantum mechanical methods. Without a rigorous discussion, a few special properties of these bonds may be mentioned. In covalent compounds electrons are not removed from the atoms; the outermost electrons of two atoms are shared between both of them and the atoms are united into one system, called a homopolar compound. For instance, two hydrogen atoms combine to yield a single H_2 molecule by sharing their electrons. The shared electrons are with high probability to be found between the two protons, holding them together in this way.

Metals are held together by *metallic bonds.* The valence electrons of the metal atoms are removed and shared by all the other atoms, with the result that the lattice points of metal crystals are occupied by positive metal ions. The valence electrons move more or less freely in the lattice, binding together the positively charged ions. These "free" electrons are not associated with any single atom, but belong collectively to the atomic assembly. The metallic bond resembles the covalent bond in that the atoms are held together by shared electrons in both cases, but with metals the collectivization of the electrons is more pronounced.

Covalent bonds must be considered in somewhat more detail, since they are important in the structure of biological molecules. In some covalent compounds the centres of positive and negative charges coincide; these involve *pure* covalent bonds. However, a considerable number of covalent compounds are known in which the shared electrons are found with a higher probability in one part of the molecule than in another part. The centres of positive and negative charges in this case are separated from each other, and the molecules behave as electric dipoles. Covalent *dipole molecules* represent a transition between ionic molecules, which are always dipoles, and pure covalent molecules. The atoms in H_2, Cl_2 and O_2 molecules, for instance, are linked to each other by pure covalent bonds, whereas the molecules of H_2O and HCl are dipoles. Similarly, amino acid molecules, lipids and proteins involve dipole properties.

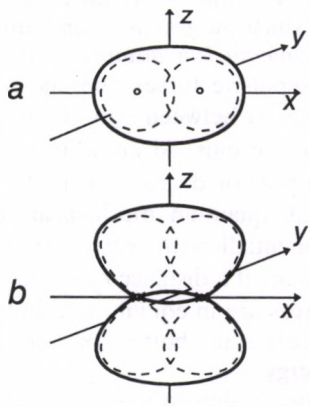


Fig. 1.7. Molecular orbitals

a: σ orbital created by coalescence of the atomic *s* orbitals; *b*: π orbital formed from the *p* orbitals. The dashed curves indicate the *s* and *p* orbitals, and the small circles denote the nuclei of the coupled atoms

Two types of covalent bonds are distinguished, σ and π bonds. The first type is formed in the case of single bonds between electrons, be they in an *s, p,* or any other state. π bonds, on the other hand, are formed at multiple bonds, but one of the bonds is a σ bond. *s* electrons do not participate in the formation of π bonds.

The individual electron states and the characteristic properties (e.g. the localization, or the probability distribution) are described and calculated from quantum mechanical functions called (analogously to atomic orbitals) *molecular orbitals.* These are conceived of in the same way as atomic orbitals.

With respect to the straight line between the nuclei of the two atoms, the bond axis, the σ orbital has the rotational symmetry depicted in Fig. 1.7*a* where *X* is the axis of rotation. In σ bonds the linked atoms can rotate around the bond axis with respect to each other. The π orbital has only mirror symmetry. The atomic nuclei are situated in the plane of symmetry, and the probability of finding an electron in this plane is zero. Figure 1.7*b* depicts the π orbital (with the *XY* symmetry plane). Atoms connected by π bonds cannot rotate around the bond axis with respect to each other. Whereas σ bonds are localized to two atoms, in multiatomic systems π electrons may be found with high probability in the vicinity of more than two atoms, i.e. they are delocalized and have *delocalized molecular orbitals.* This type of orbital can quite frequently be found in organic molecules, e.g. in the rings of aromatic compounds (see point 3 below), in proteins, nucleic acids, etc. (cf. sections 1.5.2 and 1.5.3).

2. Bond energy. One of the most important data characterizing bonds is the bond energy. This is most easily demonstrated for ionic bonds. Two ions with opposite charges are attracted by an electrostatic force, which is balanced by a repulsive force. (The repulsion will be explained later.) Both forces increase with the decrease of the distance between the two ions, though the repulsive force increases more rapidly than the attractive one. As long as the ions are not too close to each other, the attractive force is predominant; within a certain distance, however, due to its faster increase the repulsion becomes predominant. The distance at which the repulsive and attractive forces just cancel each other is called the *equilibrium internuclear distance,* and the ions oscillate around this. Instead of the attractive and repulsive forces, the energies are given in Fig. 1.8. The abscissa represents the distance (r) between two atomic nuclei, and the ordinate the interaction energy (E) between the ions. At an infinite distance E becomes zero. The potential due to the Coulomb attraction decreases hyperbolically with increasing distance (curve *a*). On the other hand, quantum mechanical calculations indicate that the repulsive energy increases exponentially with decreasing distance (curve *b*). The resulting potential (curve *c*) has a minimum at a distance r_0. r_0 corresponds to the distance where the attractive and repulsive forces are in equilibrium. In this state the ion is found in a potential valley, similarly to a ball at the bottom of a crater, and a displacement in any direction would increase its energy.

Repulsion is a tendency of two bodies to move away from each other. The repulsive potential can be deduced from Pauli's principle, since the electrons in the outermost shells of the linked atoms occupy the lowest energy state, and the shells are saturated.

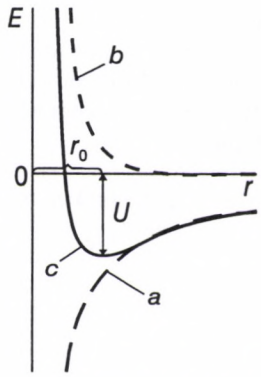


Fig. 1.8. Variation of the interaction energy (E) with the inter-ionic distance r

Any additional electrons have to be placed in a new shell, but this involves an energy uptake. If the ions came too close to each other, their electron clouds would merge. The Pauli principle would hold for the resulting system, and consequently the electrons of one of the ions should be brought to a higher energy level. The requirement of energy uptake, however, is equivalent to the development of a repulsive force, i.e. a repulsive potential.

The distance dependence of the interaction energy in covalent bonds is described by a similar curve. The *bond energy* is defined as the energy required to remove the partner particles from their potential valley and to separate them by an infinite distance. This energy is denoted by U in the diagram. As an example U = 5.2 eV for KCl, with r_0 = 280 pm. The same amount of energy is released when the two ions come close enough to form the KCl molecule.

The bond energy is usually given for 1 mole, and not for a single bond. In ionic, covalent and metallic bonds the bond energies are in the range of 100–400 kJ/mol, which is equivalent to 1–5 eV per bond.

For crystals, the bonds are characterized by the *lattice energy,* which is a measure of the work required to separate the lattice elements of 1 mole crystal by an infinite distance from each other.

In numerous covalent molecules the bond energy between two given types of atoms is practically independent of any other bonds which may exist in the molecule. The total bond energy of the molecule is simply the sum of all of the bond energies present. Table 1.4 lists the energies of some covalent bonds frequent in organic compounds.

Table 1.4. Energies of various covalent bonds

Bond	Bond energy (kJ/mol)	Bond	Bond energy (kJ/mol)
H – H	430	O – H	461
C – H	358	N – H	391
C – C	263	N – C	292
C = C	424	N – O	255
C – O	352	N = O	452

When the various bond energies are additive, the bond distances between two atoms can easily be calculated as the sum of the atomic radii associated with the covalent bonds. The value of the radius depends only upon the multiplicity of the bond.

Table 1.5 presents a few data; the bond length between oxygen and hydrogen atoms, for instance, is 66 pm + 30 pm = 96 pm.

Table 1.5. Atomic radii (pm) in covalent bonds

	C	O	N	H
Single bond	77	66	70	30
Double bond	67	55	62	—

3. Conformation. Molecular conformations and crystal structures are similarly determined by the electronic structures of the associating atoms, again on the principle of *minimum energy*. We shall not go into details; only a few examples of covalent compounds will be presented.

The outermost shell of the carbon atom contains two electrons in the *s* state and two electrons in the *p* state. However, its bonds are not formed by isolated *s* and *p* electrons, but by hybridized electrons, which are equivalent from the aspect of chemical binding. The experimental fact that the four hydrogen atoms of the methane molecule, for instance, are in exactly identical positions relative to the carbon atom can be explained by the existence of equivalent bonds. The four hydrogen atoms occupy the apices of a tetrahedron, with the carbon atom at the centre (Fig. 1.9). The H–C–H bond or valence angle is 109.5°. If one of the hydrogen atoms is replaced by an OH radical, the bond angles at the carbon atom remain practically unaltered. There are only slight changes even in methyl iodide or glycine. Table 1.6 presents some bond angles, which help to illustrate the conformations of some simple molecules.

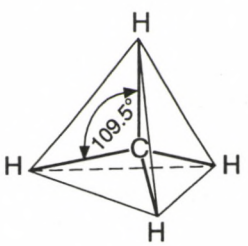


Fig. 1.9. The tetrahedral model of methane

The binding in the rings or ring systems of aromatic compounds should be discussed briefly. As a simple example, benzene may be considered. According to Kekulé's formula, the benzene molecule contains six C–H, three C–C and three C=C bonds (Fig. 1.10*a*). It might be concluded that the three C=C bonds should be shorter than the three C–C bonds (cf. Table 1.5). In fact, however, all of the bonds between the carbon atoms are equivalent. This can be explained in the following way. The C–H bonds are σ bonds, and the carbon atoms are linked by six σ bonds and three π bonds. The π bonds are not

localized to two adjacent carbon atoms, but are found with equal probability between any two adjacent carbon atoms. This makes the bonds between the carbon atoms equivalent. One of the usual notations is presented in Fig. 1.10*b*. This conception correctly explains the experimental results obtained for the bond energies.

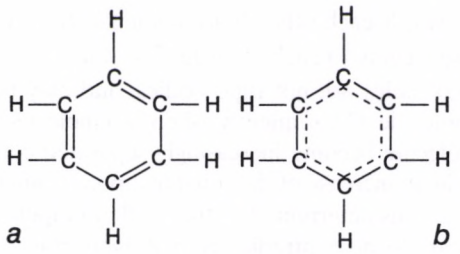


Fig. 1.10. Benzene bonds according to Kekulé (*a*) and according to experience (*b*)
The dashed lines denote delocalized π electrons

Table 1.6. Bond angles

Compound	Bond	Bond angle (degree)
Methane (CH_4)	H – C – H	109.5
Methanol (CH_3OH)	H – C – H	109.3
Methyl iodide (CH_3I)	H – C – H	111.4
Glycine ($H_2N - CH_2 - COOH$)	C – C – N	111.8
Water (H_2O)	H – O – H	104.5
Dimethyl ether ($H_3C - O - CH_3$)	C – O – C	111
Ammonia (NH_3)	H – N – H	107.3
Trimethyl amine ($(CH_3)_3N$)	C – N – C	108

1.3.2. Van der Waals bonds. Hydrogen bonds

Atoms, ions and molecules are frequently linked by van der Waals bonds. This type of bond is weaker by at least one order of magnitude (with values of 0.001–1 eV, or 0.08–80 kJ/mol) than ionic, covalent or metallic bonds. For this reason van der Waals bonds are usually neglected when the chemical bonds in molecules or crystals are discussed. However, in the interaction of molecules with each other or with ions and atoms, the van der Waals forces become important. The attraction between neutral gas molecules, the condensation of gases, the cohesion of the molecules in liquids, and hydrate and solvate envelopes are all due to van der Waals forces. From biological aspects these forces are of considerable importance in the formation of the secondary, tertiary, ... structures of macromolecules, and in the interactions between the structural elements of the cells. The variety, changeability and sensitive responsiveness of biological processes are connected with the low energies of van der Waals bonds. The weakness of this bond type results in the easy disruption and reformation of the individual linkages.

Van der Waals bonds. This type of bond is also called a dipole-dipole or ion-dipole bond. This nomenclature reflects the nature of the bond. As already mentioned, certain molecules have a dipole moment as a result of their structure. If the dipoles are sufficiently close to each other, repulsive forces act between poles of the same sign, and attractive forces between poles of opposite sign. As a result of these forces the molecules are oriented so that opposite charges turn towards each other *(orientation effect)*. The attractive forces due to opposite charges in close vicinity to each other lead to van der Waals binding between the molecules. However, molecules are not rigid bodies, and may become deformed in the force field of another molecule. Consequently, when two molecules approach, the opposite charges within each molecule become increasingly separated compared to their original positions, which results in an increase of the original dipole moment (*induction effect*). The strength of the bonding is thus determined by the resultant dipole moment.

At first sight it appears to be contradictory that even molecules in which the dipole moment is initially zero can be bound by dipole forces. However, it should be remembered that within an atom the positively charged nucleus and the negative electron cloud are not at rest with respect to each other. The same applies for molecules, too. The centres of charges of opposite sign alternately move away from and approach each other. The overall effect is that one charged centre appears to vibrate with respect to the other centre or to rotate around it. The charges appear in positions changing with respect to each other. With molecules of zero dipole moment this means that only the average value of the moment is zero, since the centres are not coincident but are in a vibrational or rotational state. As a result of the continuous charge displacements, the molecules are actually dipoles which are associated with each other by the continually changing moment.

All this is also true for the binding between molecules and ions, molecules and atoms, or ions and atoms.

Hydrogen bonds. This type of bond is also formed by interacting dipoles. The hydrogen atom has only one electron and can form a covalent bond. In some compounds, however, the hydrogen atom can be bound to two other atoms instead of one. The additional bond is found in compounds containing fluorine, oxygen and nitrogen atoms, or FH, OH and NH radicals. The example of H_2O is presented in Fig. 1.11. The hydrogen atoms are bound to two oxygen atoms. One O–H distance is approximately 100 pm, while the other

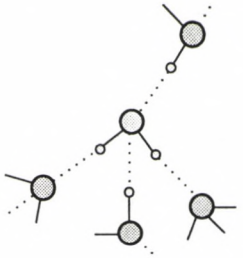


Fig. 1.11. Coupling of H_2O molecules by hydrogen bonds
The larger circles denote oxygen atoms and the smaller ones hydrogen atoms.
The continuous lines indicate covalent bonds and the dotted ones hydrogen bonds

is 180 pm. The former is a stronger (covalent), and the latter is a weaker bond: a hydrogen bond. The dipole character of the hydrogen bond is quite obvious, since the above-mentioned radicals have a relatively large dipole moment, the positive charge of the dipole being the hydrogen. This type of radical can readily bind electronegative atoms through the positive end. Hydrogen bonds play an important role in the structures of many compounds, e.g. alcohols, carboxylic acids, amines and the biologically important fats, carbohydrates, nucleic acids and proteins. The binding energy of a hydrogen bond is generally several times higher than that of a van der Waals bond.

1.3.3. The energy states of simple molecules

The energy of any molecule consists of the following three components (possible translational motion is not considered): the *electronic energy* (E_{el}); the energy resulting from the vibration of the molecular atomic nuclei along their connecting lines (axes), the *vibrational energy* (E_v); and the *rotational energy* (E_r), due to the rotation of the molecule itself. All these energies can have only well-defined, discrete values determined by the quantum conditions.

Let us consider a simple diatomic molecule with a structure similar to that of a dumb-bell. The molecule rotates around an axis normal to the straight line connecting the nuclei. The energy of the rotation is described by the equation

$$E_r = \frac{\hbar^2}{2K} m\,(m+1), \quad m = 0, 1, 2, \ldots \qquad [1.18]$$

K is the moment of inertia. The atoms oscillate along their connecting axis with an energy

$$E_v = \left(n + \frac{1}{2}\right) h\nu_0, \quad n = 0, 1, 2, \ldots \qquad [1.19]$$

where ν_0 is the eigenfrequency of the oscillating system. The value of E_v is never zero, even in the lowest energy state; this is the zero-point energy, existing at 0 K, too.

It follows that the total energy of a molecule increases or decreases only by definite values. The electron state and the vibrational and rotational states of the molecule usually change simultaneously, and consequently the total energy gained or lost is the sum of three terms:

$$\Delta E = \Delta E_{el} + \Delta E_v + \Delta E_r \qquad [1.20]$$

The largest change in energy is due to the change in the electronic configuration. The order of magnitude of the term ΔE_{el} is eV. The energy change due to the change in the vibrational energy is smaller by one order of magnitude, and the rotational energy change is smaller by a further order of magnitude.

Figure 1.12 shows the essential features of the energy level system of a diatomic molecule. Level system A may be considered to belong to an electron in the ground state, system B to one in the first excited state, followed by more such systems. Of the vibrational and rotational levels also just a few are shown, with the lowest energies. To this

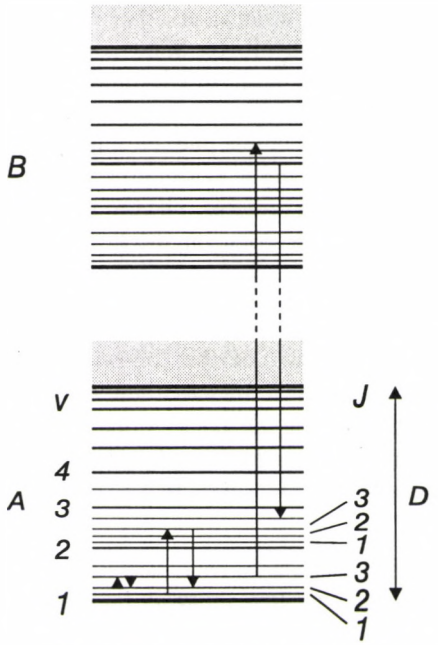


Fig. 1.12. Part of the energy level system of diatomic molecules
The thick horizontal lines denote electron levels, medium ones vibrational and thin ones rotational levels. Arrows pointing upwards: mean energy uptake (e.g light absorption), downwards: energy release (e.g. light emission). On the left pure rotational, in the middle vibrational-rotational, on the right electron transition are shown

fact refer the vibrational and rotational quantum numbers (v and J, respectively). If the vibrational energy is high enough ($>D$, dissociation energy) the molecule dissociates and from this point the atomic energies vary already continuously. The shading indicates just this region.

Basic information about the energy levels of molecules may be obtained from the *optical spectra,* though not every conceivable transition between the energy levels appears as a spectral line. For molecules, too, the actual transitions are restricted by selection rules. Even so molecules present complex spectra consisting of a considerable number of lines; several vibrational states are associated with each electron state, and several rotational states with each vibrational state. If only the electron states were to change, relatively simple line spectra, similar to those for atoms, would be obtained, and the spectral lines would indicate frequencies due to changes in the state of the molecular electrons. However, the changes in the vibrational state modify every frequency, leading to multiplication of the lines. Moreover, the changes in the rotational energy yield additional modifications, resulting in further splitting of the already multiple lines. In low resolution spectroscopes some spectral lines merge into bands, that is why molecular spectra are frequently mentioned as band spectra.

1.4. Condensed systems. Order and disorder

1.4.1. The universal gas law and its interpretation

According to the *universal gas law (Clapeyron–Mendeleev equation),* the product of the pressure (p) and the gas volume (V) is proportional to the product of the quantity of the gas (mole number v) and the absolute thermodynamic temperature (T):

$$pV = RvT \qquad [1.21]$$

The proportionality factor R is independent of the nature of the gas, and for this reason R is called the universal gas constant. R is sometimes also referred to as the molar gas constant

$$R = 8.314 \frac{\text{J}}{\text{mol K}} \qquad [1.22]$$

In practice, deviations from [1.21] are observed; these are the smaller, the more point-like the molecules can be considered and the more negligible the molecular interactions. For simplicity, in the following sections we shall deal only with *ideal gases,* for which no deviations are observed, and restrict the discussion to the limiting case presented by [1.21]. The most important properties of the gaseous state can easily be understood if it is assumed that the molecules are constantly in motion *(thermal motion),* and collide elastically with each other and with the vessel walls. Further, the molecules are assumed to move in straight lines between two collisions. The motion is totally random, in every direction and with widely varying velocities (Maxwellian velocity distribution; cf. Fig. 1.13).

a) Interpretation of the gas pressure. The pressure exerted by the gas on the walls of the vessel is assumed to be a result of the elastic collision of the molecules with the walls of the vessel. If the collision of the randomly distributed molecules also occurs at random and the collision frequency is sufficiently large, the effect is manifested macroscopically by a uniform pressure distribution, produced by compressive forces acting on the walls of the vessel. The number of molecules in a comparatively small volume of gas is very high. For instance, 1 m^3 air in the normal state contains 2.69×10^{25} molecules, and even at a low pressure of approximately 10 mPa there are still 10^{18} molecules in the same volume. Calculations show that the gas pressure (p) is proportional to the concentration (n) and to the average kinetic energy of the molecules, as expressed by the mass of one molecule (μ) times the average of its square velocity ($\bar{v}^2$):

$$p = \frac{1}{3} n\mu\bar{v}^2 \qquad [1.23]$$

The concentration of the molecules is defined as

$$n = \frac{N}{V} \qquad [1.24]$$

where N denotes the number of molecules in a volume V, and $\bar{v}^2$ is the arithmetic mean of the square velocities. For a given pressure and volume, n is the same for every gas *(Avogadro's law);* its value for the normal state is 2.69×10^{25} m^{-3} *(Avogadro's constant).*

b) Interpretation of temperature. From [1.24] and [1.23]

$$pV = \frac{1}{3} N\mu\bar{v}^2 \qquad [1.25]$$

If [1.25] is compared with the universal gas law [1.21]:

$$\frac{1}{3} N\mu\bar{v}^2 = vRT \qquad [1.26]$$

The numerical value of N/v is equal to the number of molecules per mole, thus this quotient is the same for every gas; this is the *Loschmidt constant and is denoted by L:*[1]

$$\mathrm{L} = 6.02 \times 10^{23}/\mathrm{mol} \qquad [1.27]$$

If both sides of [1.26] are divided by N:

$$\frac{1}{3}\mu\bar{v}^2 = \frac{R}{L}T \qquad [1.28]$$

The quotient R/L is the *Boltzmann constant,* usually denoted by k:

$$k = \frac{R}{L} = \frac{8.31\ \mathrm{J/mol\ K}}{6.02 \times 10^{23}/\mathrm{mol}} = 1.38 \times 10^{-23}\ \mathrm{J/K} \qquad [1.29]$$

Minor rearrangement of [1.28] leads to

$$\frac{1}{2}\mu\bar{v}^2 = \frac{3}{2} kT \qquad [1.30]$$

which permits a more profound interpretation of temperature as well as the quantity of heat. The temperature is in a clearly-defined relation with *the mean kinetic energy of the molecules,* which is *linearly proportional to the absolute temperature.* An increase or decrease in temperature means an increase or decrease in the mean kinetic energy of the molecules. A temperature increase clearly requires additional energy, whereas a temperature decrease leads to a decrease in the mean kinetic energy of the molecules.

Volume, pressure and temperature are characteristic quantities of the *macroscopic* state of the gas, whereas the space coordinates, the velocity and the energy of the individual molecules relate to the *molecular* state. The macroscopic properties are determined by the "average behaviour" of the molecules.

[1] This nomenclature is not universally accepted; some authors interchange the *Loschmidt constant* with *Avogadro's* constant. The dimensionless numerical values of these constants are called the *Avogadro* and *Loschmidt* numbers.

If the definitions of density (ρ) and molar mass (M) are introduced, [1.21] can be written in the following form:

$$p = \rho \frac{RT}{M}, \text{ where } \rho = \frac{m}{V}, \; M = \frac{m}{\nu} \qquad [1.31]$$

where m is the mass of a gas of volume V.

Another form sometimes used is

$$p = \rho \frac{kT}{\mu} = nkT \qquad [1.32]$$

obtained from [1.31] using the relations

$$\rho = n\mu, \; M = L\mu \text{ and } k = \frac{R}{L}$$

Thus according to [1.31] and [1.32] at a constant temperature and for a given quantity of gas, the pressure, density and molecular concentration are proportional to one another.

Some informative data on gas molecules can be summarized as follows. Any individual gas molecule undergoes collision several times within 1 s, and the number of collisions is proportional to the square of the molecular radius, the concentration of the molecules and their mean velocity. The *number of collisions per second* of a gas at atmospheric pressure and at 273 K is of the order of magnitude of 10^9.

The *mean free path length* between two consecutive collisions is inversely proportional to the square of the molecular radius and the molecular concentration. Its value at atmospheric pressure is of the order of magnitude of 10^{-8} m, but at 1 mPa it is approximately 10 m.

c) Maxwellian velocity distribution. Figure 1.13 depicts the velocity distribution of oxygen molecules according to their velocity (*Maxwellian velocity distribution)* at temperatures of 273 K and 373 K. The abscissa represents the velocity v, while the ordinate gives the percentage of molecules whose velocity lies between v and $v + \Delta v$. (One has to consider a Δv finite interval, because the probability that a molecule has exactly v velocity is zero. Δv is a small but arbitrarily chosen value.) In Fig. 1.13 Δv is 10 m/s, which is small compared to the values in the figure, thus the curve may be considered practically continuous. The maximum in the curve relates to the velocity of a relatively large number of molecules. If the velocities of the molecules were studied individually, this velocity value would be found most frequently. The most probable velocity is around 350 m/s at 273 K and 450 m/s at 373 K. The asymmetry of the curves means that the mean velocity differs from the most probable velocity, the former being somewhat larger. With

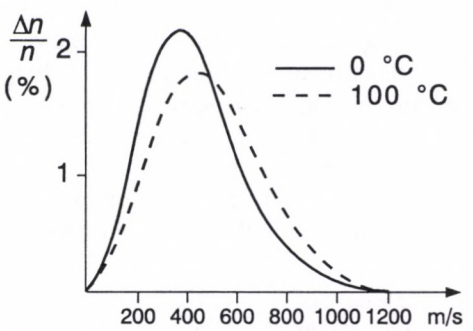


Fig. 1.13. The velocity distribution of oxygen molecules

increasing temperature the curve shifts towards higher velocities, because the number of relatively slow molecules decreases while the number of faster ones increases.

The kinetic gas theory permitted a more profound interpretation of several well-known phenomena, such as friction, heat conduction and diffusion. The experimental results are the most important evidence in support of the theory.

1.4.2. Kinetic heat theory

The molecules of the perfect gases discussed above were regarded as points, and consequently only their translational motion was considered. However, besides undergoing translation, real molecules also rotate, and moreover the atoms forming the molecules oscillate with respect to each other. For a correct interpretation of the empirical results, not only the kinetic energy of translation, but also the kinetic energies due to rotation and vibration must be considered.

The molecules of fluids and solids (crystals)[2] too are in continuous, but irregular motion *(thermal motion),* and it is generally true that the heating of a body is equivalent to an increase in the *mean velocity* of its molecules. In solids the thermal motion of the atomic constituents is mainly restricted to vibration and rotation around their equilibrium positions. The attractive forces keep the molecules together in liquids too, and consequently the thermal motion consists mainly of vibrations and rotations, though the migration of the molecules can no longer be neglected. With increasing temperature the structure of a liquid becomes more and more similar to that of a gas, and besides vibrations and rotations translational motion gradually comes into prominence.

Any change in phase of matter is accompanied by structural changes. The melting point is the temperature at which the concentration of the lattice defects in the crystal increases to such an extent that the lattice becomes unstable, and the degree of order decreases (cf. also sections 1.4.3 and 1.4.4). Freezing involves the opposite processes. In the case of evaporation (sublimation) molecules with relatively high kinetic energy overcome the attractive forces and break away from the liquid (solid). The average distance between the molecules in vapour is so large that their interconnection (apart from random collisions) is virtually non-existent. Condensation involves the transformations in the opposite direction.

Melting and evaporation are energy-requiring structural changes. The melting heat and the heat of evaporation supply the energy needed for the variation in the positions of the molecules that leads to the structural changes at the melting point and evaporation temperature, respectively. In freezing or condensation the opposite transformations are accompanied by heat liberation. Thus, the transformation heat is a measure of the changes in the mutual *potential energies* of the molecules.

In chemical reactions the changes in the mutual positions of the atoms (atom groups) are also accompanied by the liberation or absorption of heat (exothermic or endothermic

[2] We differentiate between "solid bodies" and "solids". The former concept refers to bodies which keep their shape and volume. (The adjective solid can frequently be substituted by hard.) The second expression, on the other hand, is equivalent to the term crystals.

processes). The reaction heat is a measure of the overall change in the potential energies of the combining atoms (cf. also sections 5.3.1 and 5.3.2).

1.4.3. Defects in the order of atoms

Solids or crystals are discussed below, but many of the findings also apply to macromolecules (cf. sections 1.5.2–1.5.5).

1. Missing atoms, vacancies. As already pointed out, the order in crystals is determined by the electronic structures of the associating atoms, and is characterized by an energy minimum. In fact, perfect crystals do not exist. One reason for defective states is the thermal motion of the atoms, ions or molecules (i.e. the lattice elements). As a result of energy exchange between the lattice elements, some of these elements may obtain such a high energy that they overcome the attractive forces and break away from their neighbours. A small number of lattice elements of such high energy always exists at low temperatures, and this number is larger at high temperatures. For instance, evaporation or sublimation may be explained in that lattice elements with sufficiently high energy escape from the surface. This escape of lattice elements from the surface is responsible for empty lattice sites, called *vacancies,* within the crystal. When a particle has left the surface its place may be taken by a neighbouring particle from some deeper position, whose empty site in turn may be occupied by another particle from a still deeper region. In this way the vacancy migrates within the crystal (Fig. 1.14). The vacancies created by this means are called *Schottky defects.* Vacancies are continuously created and annihilated, for instance by migrating to the surface. Crystals always contain vacancies, the number of which increases with increasing temperature. The vacancy concentration depends upon the activation energy necessary for the formation of the vacancies, too. At a given temperature the vacancy concentration is lower in a material in which the activation energy is higher. At thermal equilibrium the concentration of Schottky defects (n_s) is given by

$$n_s = ne^{\frac{-\varepsilon_s}{kT}} \qquad [1.33]$$

where n is the concentration of lattice points in the material, T is the absolute temperature, ε_s, is the energy required for the formation of a vacancy, and k is the Boltzmann constant. The value of ε_s lies in the range of the binding energies of the lattice elements.

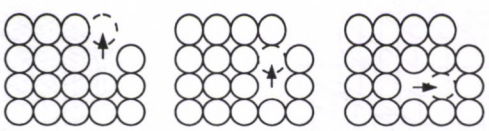


Fig. 1.14. Vacancy migration
Arrows indicate displacement of the crystal element to the vacancy and dashed circles its position after displacement

Near the melting points in metals, the number of vacancies is a few thousandths of the number of lattice points.

In some substances the vacancies are created not only at the surface, but also within the crystals. In this case particles leaving their lattice sites become wedged between other regular lattice sites and occupy *interstitial positions.* A defect pair consisting of a vacancy and an interstitial particle is a *Frenkel defect.* The concentration of this type of defect pair (n_F) is given by a relation similar to [1.33].

The defect concentration can be increased above its temperature-dependent equilibrium concentration by various treatments, e.g. by deformation, by quenching from some higher temperature, or by irradiating the crystal with some high-energy electromagnetic or corpuscular radiation. The excess of defects thus created can be conserved for some time (possibly years), though their distribution may change, the defects accumulating locally, for instance. The equilibrium concentration is restored by heating the crystal to a temperature close to the melting point, and subsequently cooling it slowly to room temperature (this cooling may last for several days).

An important group of lattice defects are the *chemical defects* created by the incorporation of foreign atoms (ions, molecules) in the crystal. These defects may be localized at regular lattice sites, or in interstitial positions, either individually or in groups (as complexes). It frequently occurs with compound crystals that some of their components are not built into the crystals in proper stoichiometric ratio.

2. Dislocations. Surface defects. Another type of lattice defects are *dislocations,* which are created by mechanical stresses, for instance by deformation, uneven cooling or heating, or as a result of the accumulation of vacancies or interstitial particles. Two types of dislocation exist: *edge dislocation* and *screw dislocation.* For example, if an initially regular crystal is pressed in the direction of the arrow shown in Fig. 1.15, the lattice planes are displaced with respect to one another and the result is that one plane appears to be wedged in as an extra plane (denoted by *A* in the figure) not in correspondence with the adjacent planes. The lattice structure in essence becomes distorted along one line, the dislocation line, close to the edge of the extra plane. This line represents the edge dislocation. In the example in the diagram the dislocation line lies normal to the plane of the drawing, through the extension of the dashed line. Figure 1.16 shows a screw dislocation generated by a deformation acting in the vicinity of the vertical dashed line in the crystal. This line represents the screw dislocation, because a path proceeding round it (starting at point *P*) describes a spiral.

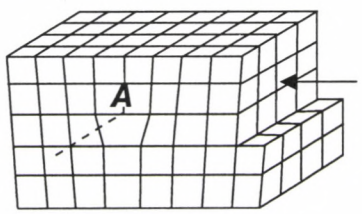


Fig. 1.15. Edge dislocation

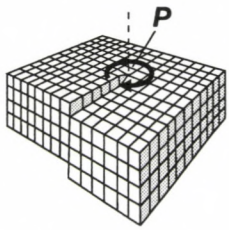


Fig. 1.16. Screw dislocation

Vacancies and interstitial particles are frequently referred to as point defects, and dislocations are also called line defects. *Surface defects* are a third type of lattice defects. The grain boundaries in polycrystalline materials are surface defects. Single crystals too may have a mosaic-like structure, consisting of crystalline blocks whose orientations differ from one another only slightly. The domains connecting adjacent blocks (block boundaries, grain boundaries) of necessity contain a large number of point and line defects, thereby allowing the smoothing-out of the orientational differences between neighbouring grains.

Some properties of crystals are especially sensitive to structural irregularities, e.g. diffusion, thermal conduction, plasticity, luminescence, ionic conduction, photoelectric properties, etc., i.e. all those properties connected with the migration of neutral or charged particles or energy transport.

3. Biological aspects. Structural defects also exist in biological macromolecules and macromolecular systems, where the likelihood of defects is enhanced, for instance, by thread- or chain-like molecular configurations with rotational and bending possibilities. The probability of defects is especially high for the relatively weak intra- and intermolecular bonds (van der Waals bonds, hydrogen bonds). In this case too the formation of defects is a result of various factors: a temperature increase, the presence of various substances, a change in the hydrogen ion concentration, etc.

In many cases the presence of structural irregularities is unfavourable, and defect formation must be avoided or at least decreased. However, it would be a mistake to think that the irregularities result in unfavourable macroscopic or harmful functional consequences in every case. From among the examples for crystals, it may suffice to mention the basic role played by the doping additives in doped semiconductors or activated luminophores. Other well-known examples include channel formation in membranes, and the role of local denaturation in the DNA chain in connection with its biological functions (cf. also sections 1.5.2–1.5.5).

1.4.4. Liquids and amorphous solids. Mesomorphous state

1. Liquids. As concerns their structure, liquids comprise a transition between the solid and gaseous states. Near to the melting point liquids are still ordered to some extent, and the disorder characteristic of gases develops only gradually with increasing temperature. At a temperature close to the melting point liquids are referred to as "molten crystals", and close to the critical temperature they may be thought of as condensed (liquefied) gases. The atomic order present in liquids is restricted to small volume units. In some cases the order extends over groups consisting of a few hundred molecules, but these elements with ordered structure are disordered with respect to one another. With increasing temperature the number and extent of the ordered volume units decrease. Long-range order is typical only of crystals, and liquids display only short-range order. The ordered groups in the liquids are not stable, but are continually breaking and reforming due to the thermal motion. The presence of the small ordered groups and their disordered orientation explains the directional independence of the physical properties

of the liquids: they are said to be isotropic. Anisotropy, on the other hand, is a characteristically crystalline property.

As already mentioned, the thermal motion in liquids consists mainly of oscillations, though the equilibrium positions are less well defined than in solids. The number of vacancies in liquids is considerably larger than in solids (on melting the number is multiplied several-fold). This allows frequent displacements of particles, i.e. translational motion, which explains the fluidity of liquids.

2. Amorphous solids (glasses) are structurally liquids, and consequently do not have sharp, well-defined melting points, in contrast with crystals. On cooling, no qualitative change occurs in their short-range order, that characteristic of liquids simply being "frozen in". Amorphous solids are supercooled and strongly viscous liquids without fluidity.

3. The mesomorphous or liquid crystalline state is intermediate between liquids and crystals. The common liquids are isotropic in every respect, whereas crystals may exhibit anisotropy in many respects; for example the elastic constant, the dielectric constant and the optical refractive index may be direction-dependent in crystals. The *mesomorphous state* is defined as an *anisotropic liquid phase;* it is frequently referred to as an *anisotropic liquid* or *liquid crystal.*

The mesomorphous state can be observed in substances consisting of molecules with non-spherical symmetry, but of anisodimensional, e.g. rod-like or thread-like molecules. In a consideration of the structure in the mesomorphous state not only the ordering of the molecular mass centres, but also the directional arrangement of the molecules *(translational and orientational order)* must be taken into account. The order of the intermediate phases is lower than that found in the solid phase, though higher than in real liquids.

Two types are distinguished, the classes of *thermotropic* and of *lyotropic* liquid crystals. In the former class the mesophase is induced by a temperature change. Lyotropic systems, on the other hand, are aqueous solutions of amphiphilic molecules, the phase properties depend on the concentration too. Figure 1.17 depicts the most frequently occurring arrangements. In the *smectic* state the centres of mass of chain-like molecules are situated on planes equidistant from one another (Fig. 1.17*a*). The molecular axes are generally normal to these planes, though oblique directions are also observed sometimes. In the *nematic* state the ordered structure is reduced to a nearly parallel arrangement of elongated molecules, no layers are formed and the molecules can move with respect to one another along their longitudinal axes (Fig. 1.17*b*). In the *cholesteric* (twisted nematic) state the longitudinal molecular axes are in parallel planes (shown in Fig. 1.17*c*), but axes in neighbouring parallel planes are rotated with respect to one another. Consequently, the axes of the molecules in planes situated one above the other display a helical arrangement.

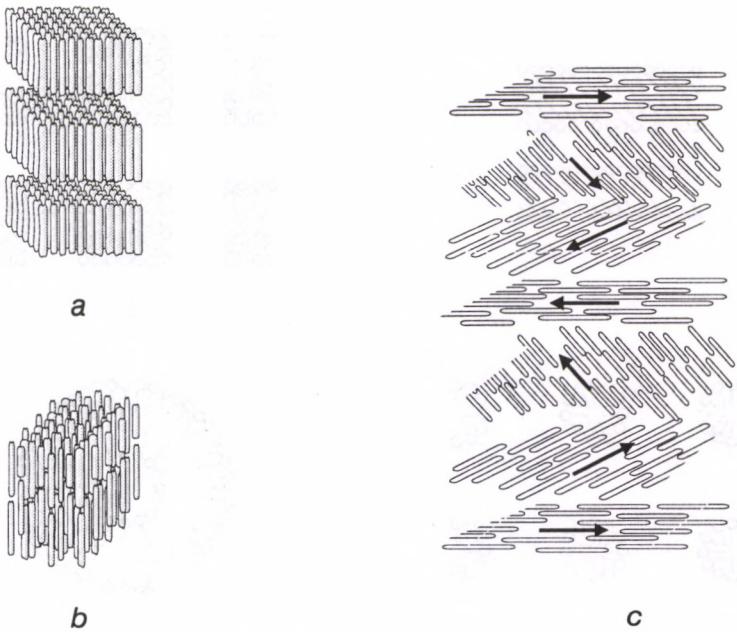


Fig. 1.17. Mesomorphous states, *a,* smectic state, *b,* nematic state, *c,* cholesteric state. The arrows represent the direction of molecules

A sequence of decreasing order is shown in Fig. 1.18.

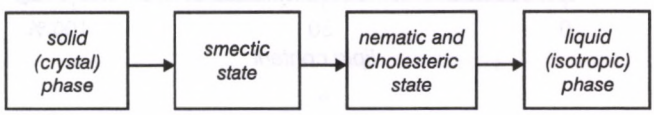


Fig. 1.18. Phase transitions in the sequence of decreasing order

4. Biological and other practical aspects. Figure 1.19 depicts lyotropic systems. Soap solutions, for instance, are lyotropic systems, and nucleic acids, and many polypeptides become lyotropic when present in appropriate concentration in some solvent (mainly water). This condition is usually given in the intracellular structure. Similarly, aqueous solutions of biologically important lipids exhibit lyotropic liquid crystal structure. Figure 1.19 may also be regarded as depicting some possible forms of lipid–water systems. The small circles represent the hydrophilic polar head groups of the lipid molecules, while the tails are the lipophilic (i.e. hydrophobic) hydrocarbon chains. The polar groups always turn towards the aqueous phase, interacting with the polar water molecules. The hydrophobic parts, on the other hand, interact with one another by van der Waals forces. Figure 1.19*a* shows bimolecular layers of lipid molecules in section. The molecule pairs are either normal to the plane of the layer or tilted to it at an oblique angle. The diagram

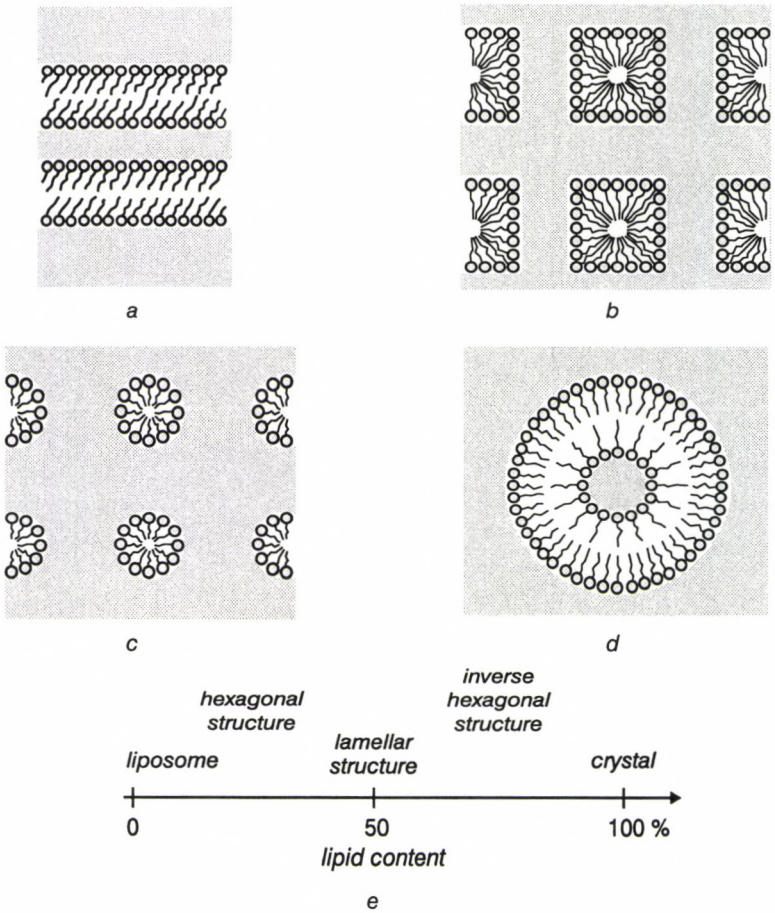


Fig. 1.19. Lyotropic systems
a: layered; *b*: prismatic; *c*: cylindrical; *d*: tubular or vesicular arrangement; *e*: connection between the lipid content and the structure

illustrates this latter case. The layers are separated by water, which is indicated by the shading. Figures 1.19*b* and 1.19*c* depict square and circular cross-sections of lipid rods. Many other arrangements are possible, for instance tube-like structures containing water or aqueous solution in the inside (Fig. 1.19*d*). Amphiphilic molecules may also form spheres or spherical vesicles consisting of bilayers (liposomes). These may display crystallographic structures, for instance face-centred or body-centred lattices. Figures 1.19*c* and 1.19*d* show structures of spherical symmetry.

Similar structures are sometimes also formed in the reverse sense, since the lipid and aqueous phases may exchange roles: in a lipid medium, layers and threads containing water molecules are formed. The basic condition for the development of the various

structures in these cases too is the interaction between the ionic groups of the lipid and polar water molecules. The formation of a given structure is influenced by several factors, e.g. the quality and type of its amphiphilic component, water content (lipid concentration), ionic environment, pH value, temperature, (hydrostatic) pressure, and the presence of "pollutants". Since the stability of the structure is based on low-energy bonds, these structures may be easily (upon a slight change in the environment) transformed. Figure 1.19*e* shows the relationship between the structure and lipid content of the solution at a given temperature and other given conditions, in case of a certain lipid. Small lipid concentration is advantageous for the formation of liposomes, while large concentration for inverted structures.

In addition to the typical lamellar structure of the cell membrane several other biological structures display a characteristic, more or less ordered liquid crystalline structure. An example is the arrangement of the myosin and actin molecules in the A and I bands of the striated muscles. Both proteins are thread-like, the first being 4.5 μm, the second 3.6 μm long, respectively. In the A band the myosin has a smectic, while in the I band the actin has a nematic liquid crystalline structure. The structural changes of the macromolecular systems brought about by environmental effects may also be considered to be connected to the liquid crystalline characteristic: e.g. the change in the proportion of the alpha/beta conformations of some proteins caused by a change in the ionic strength, or the change of the B formation of DNA to A formation upon the reduction of the water content (cf. also section 1.5.5).

Concerning the practical application of thermotropic systems, two phenomena should be mentioned: the *electro-optical* and the *thermo-optical* one.

Due to the dipole moment of a liquid crystal, its molecular structure can be rearranged with an electric field. This rearrangement may change the optical transparency, etc. of these substances. This property can be used to transform electric signals into optical ones (cf. section 6.6.1).

A change in temperature changes the colour of a cholesteric film. The helical arrangement (the helical pitch) and consequently the reflected light can likewise be changed by a temperature change. When such films are placed on the human body, for example the thermal inhomogeneities of the skin can be observed by means of colour changes. Even a difference of one-tenth of a degree can be discerned.

1.4.5. The electronic structure of solids (macromolecules). Energy band model

In both individual atoms and simple molecules, the *bound* electrons are localized in separated, well-defined energy levels, followed by the continuum, the energy region occupied by electrons removed from the atoms (*free* electrons) (cf. the energy level system of the hydrogen atom in Fig. 1.3). The situation is different for systems consisting of many atoms. Crystals, with their periodic structures, are the most easily accessible of many-atomic systems for various, detailed investigations, and a considerable amount of knowledge has therefore accumulated on the physical properties of crystals. The methods developed for crystals can be successfully applied to establish the electronic structure of biologically important macromolecules, since periodic crystals are to a first approxi-

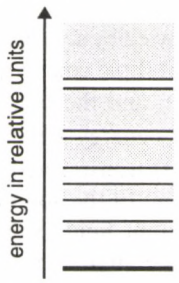


Fig. 1.20. The electron energy band system in a crystal

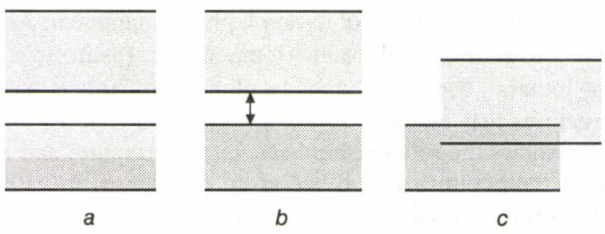


Fig. 1.21. Positions of empty, partially occupied and occupied energy bands. The occupied parts are indicated by shading

mation good models of macromolecules, as far as the structure of the latter is also periodical in a certain sense (cf. sections 1.5.2–1.5.5).

Energy bands. As a result of the interactions between the components, the electrons of a complex system are not strictly confined to individual levels, though they cannot move completely freely either. The reason for this is that, because of the interactions between the components, the electrons, mainly the valence electrons, may be exchanged. However, the electrons are not free either, since they move in the electric field of the other components, and consequently, in connection with the electronic structure of solids and macromolecules, one cannot speak of the energy levels of a single component, but of a system of energy levels. In this case, *energy bands* are formed instead of sharp energy levels (Fig. 1.20). If the system consists of N atoms forming a crystal lattice, the individual atomic electron states are split into N levels, which appears as the broadening of N discrete levels. The electrons of the system can exist only in states permitted by the broadened levels. The energy bands are the broader, the stronger the interactions between the components of the system. The lower bands are usually narrower than the higher ones, which may become so broad that the bands in this region overlap.

In the ground state the electrons occupy the lowest energy states and are excited to a higher band only by some energy uptake (e.g. light absorption, or collision with high-energy particles). However, the Pauli exclusion principle is valid for these systems, too and limits the number of electrons occupying one band. Consequently, only a fraction of the electrons can be found in the lowest band, and some of the electrons are in a higher band. If this band is also filled, further electrons will occupy an even higher band, and so on. Figure 1.21 demonstrates the various possibilities. Figure 1.21*a* depicts the case where the uppermost band is only partially filled. Figure 1.21*b* shows the case when this band is also completely filled, and no partially filled band exists. In Fig. 1.21*c* a completely empty band overlaps a completely filled band. This case is similar to that shown in diagram *a*, since only a partially filled band results from the overlapping.

Cases *a* and *c* are essentially different from case *b;* in cases *a* and *c* the electrons may take up any arbitrarily small energy, whereas in case *b* the uptake of energies smaller than a certain energy limit is forbidden. The energy limit is determined by the energy required

to promote an electron from the top of the highest filled band into the bottom level of the lowest empty band. This is indicated by the double arrow in the diagram. Many optical and electric properties can be easily explained on the basis of these diagrams. Only a few examples are given below.

Examples of the application of the band model

1. Quantum mechanical calculations indicate that the energy bands of metals which are good electronic conductors can be represented by Figs 1.21*a* or 1.21*c*, whereas the band systems of insulators and semiconductors follow the scheme given in Fig. 1.21*b*. Case *a* is observed for the alkali metals, and case *c* for other metals; the band system of e.g. the insulator NaCl and the semiconductor Ge is similar to case *b*.

Electronic conduction occurs only if the electrons can take up energy from the surrounding electric field. (The result will be accelerated electron motion within the metal, whereby the electrons lose their energy by collision with the atoms, though subsequently they are again accelerated, and so on.) This type of process can occur only in cases *a* and *c*, when the electrons in the partly empty band are gradually promoted to higher levels, followed by their dropping to a lower level. This process is then repeated. It is clear that no such possibility exists in case *b*, which demonstrates the situation in insulators and semiconductors. In case *b* the electrons can absorb energy from the surrounding electric field only if they are previously somehow transferred from the filled band into an empty band. It has been observed that this transition may occur by light absorption (photoconductivity), for instance, or on bombardment of the crystal with high-energy particles. Collisions due to the thermal motion of the atoms may also promote electrons into an empty band. According to theoretical considerations but in good agreement with experience too, for the number of these electrons (n) the expression

$$n \sim e^{\frac{-\Delta\varepsilon}{2kT}} \qquad [1.34]$$

is valid, where $\Delta\varepsilon$ is the energy difference, the "gap" between the two bands, i.e. the width of the *forbidden band,* T is the temperature and k the Boltzmann constant. At electronic conduction [1.34] plays a decisive role in the dependence of the electric conductivity on the width of the gap and the temperature. Thus the larger is $\Delta\varepsilon$ the better insulator is the substance and vice versa, and with rising temperature the insulating power decreases rapidly, the conductivity increases. The value of $\Delta\varepsilon$ for insulators is of the order of magnitude of eV, for semiconductors a tenths of eV or even smaller. (Semiconductors will be discussed in detail in point 3.)

2. A well-known *optical* property of insulators is their transparency in the visible range, whereas they absorb strongly in the UV and IR ranges. *Ultraviolet absorption* is related to electronic transitions, whereas *infrared absorption* is due to the vibrational excitation of the atoms, and the rotational excitation of atom groups. In the present case we consider only the possibilities of electronic transitions, i.e. we should like to explain the trans-

parency of insulators in the visible range, and their absorption in the ultraviolet optical range. Consider Fig. 1.21*b*. Biologically important macromolecules behave in very much the same way (cf. sections 1.5.2–1.5.5). The figure shows that electrons can be excited only by high-energy photons which can raise the electrons from the uppermost filled band into the lowest empty band. It has already been mentioned that insulators require an energy of a few eV, which corresponds to the energy of the ultraviolet photons.

Metals are non-transparent in the whole optical range. This experimental fact can easily be explained; since the free electrons in the metal can take up any energy, they absorb throughout the total spectral range. The high reflectivity of the metals is connected with the same fact.

3. The electrical conduction of *semiconductors* playing an essential role in electronics can also be discussed quite easily in terms of the band model as it was already mentioned (Fig. 1.21*b*). First let us consider *intrinsic semiconductors.* These are substances where the width of the forbidden band is small enough for the electric conductivity to be sufficiently high even at room temperature. The current consists of two parts. One is created by the motion of electrons promoted into the band above the forbidden band and moved by the electric field. This originally empty band is therefore called the *conduction band.* The other part of the current is the result of a new situation in the band below the forbidden band. This is the *valence band,* since it accommodates the electrons responsible for the chemical binding. The electrons removed into the conduction band leave *holes* (defect electrons) in the valence band, which means that this band is no longer filled, and the electrons left in the valence band can take up energy from the surrounding electric field. This part of the current is the *hole current,* and the conduction is called *hole conduction.* This may be explained in the following way. The holes are created at the interatomic bonds, and they move randomly among the atoms in the same way as the electrons split off from the bonds. The displacement of a hole can be conceived most simply in that a valence electron from its surroundings occupies the position of the hole, which in turn is shifted to the original position of the electron. This step may be repeated, with the consequence that the holes migrate in every direction within the substance. The electric field induces ordered motion, and this is superimposed on the random shifts and produces the hole current. The holes are set in motion by the electric field and drift in the opposite direction to the electrons; the holes thus behave as positive charge carriers.

If a crystal contains foreign atoms, the force field and consequently the energy level system will be changed in the vicinity of these atoms. The situation is particularly interesting if the foreign atoms incorporated into the crystal produce free charge carriers. Such crystals are called *doped* or *impurity semiconductors.* The foreign doping atoms either provide electrons and are called *donors,* or capture electrons in which case they are called *acceptors.* The donors increase the concentration of the electrons in the conduction band, whereas the acceptors increase the concentration of the holes in the valence band. If the donors predominate, the current is mainly due to electrons, i.e. negative charge is carried; this is *n-type conduction,* and the crystal is called an *n-type semiconductor.* With acceptors, the current is mainly due to holes, i.e. positive charge is carried. In this case *p-type conduction* is involved, and the crystal is a *p-type semiconductor.* For instance, a

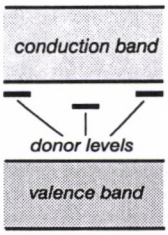


Fig. 1.22. The energy band system of a *n*-type semiconductor

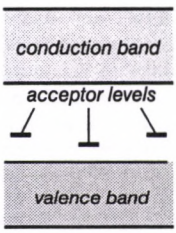


Fig. 1.23. The energy band system of a *p*-type semiconductor

germanium crystal doped with pentavalent arsenic is an *n*-type, and one doped with trivalent indium is a *p*-type semiconductor.

Figures 1.22 and 1.23 depict not only the band systems characteristic of the basic material, but also the energy levels produced by the incorporated foreign atoms. These specific levels, however, are localized in the vicinity of the dopants and are indicated in the diagrams by short bars. The fact that the donors release electrons more easily than the basic material implies according to the band model that the highest filled donor level lies considerably nearer to the conduction band than the filled valence band of the basic material. The electrons promoted from the donor levels into the conduction band leave behind holes but, since the donor levels are localized, the holes also become localized and do not participate in the conduction. The acceptors, on the other hand, accept electrons readily, since the empty energy level of the acceptors lies close to the valence band. The electrons promoted from the valence band into the acceptor level do not participate in the conduction, and the conducting charge carriers are the positive holes left behind in the valence band.

Everything mentioned above with regard to the electronic conduction of the semiconductors is valid or may be applied appropriately also for biological macromolecules.

1.4.6. Energy propagation in crystals (macromolecules)

1. Lattice vibrations (phonons). The components of an atomic system, for instance a crystal, are never at rest; they vibrate (and rotate) around their equilibrium positions. Since forces act between the components, they do not vibrate independently of one another. This is similar to a system of balls situated at equal distances and connected by elastic springs. This chain produces standing waves. The situation is much the same in crystals, though more complicated, because the atomic arrangement is three-dimensional and the vibrations propagate in every direction. The wavelength of a standing wave in a crystal may assume numerous, but only discrete values, since the only waves leading to a stationary state are those which have nodal surfaces on the boundary of the crystal. The longest wavelength is determined by the dimension of the crystal, and the shortest by its lattice constant.

Even in a given crystal, two types of lattice vibrations may exist. One is a *simple elastic acoustic vibration,* and the other type is the so-called *optical vibration.* The latter is a vibration accompanied by a change in the electric dipole moment; in contrast, the acoustic wave is due to a simple elastic oscillation. The dipole vibrations are involved in the phenomenon that the crystal can emit electromagnetic radiation (infrared light) due to lattice vibrations.

In the course of their propagation the lattice vibrations amplify each other at certain places in the crystal, and attenuate each other elsewhere; further, the vibrations are scattered by lattice defects and reflected from internal boundaries. All these effects result in a complex, continually changing situation, even if the system is otherwise

stationary. It follows that a crystal or a macromolecule is a coherent entity and that even in the most stable systems continuous processes, changes and transformations are going on.

Experimental evidence indicates that, similarly as for electromagnetic waves, in many cases corpuscular properties can be attributed to the propagation of lattice vibrations. In just the same way as light quanta lattice vibration quanta, called *acoustic quanta* or *phonons*, can be conceived. This name relates to the fact that the propagation of lattice vibrations is similar to the propagation of acoustic vibrations. The motion of the lattice elements forming the crystal can be described by phonons with various wavelengths, energies, polarizations, etc. It may also be said that there is a phonon field within the crystal, and on heating, for instance, the total phonon energy increases, whereas on cooling it decreases. Every external effect which perturbs the atomic motion produces new phonons which propagate in the system: the perturbation changes the phonon spectrum of the system.

2. Electrons, defect electrons, excitons. If a crystal takes up energy, not only the lattice vibrations (the phonon field), but also the electron states of the crystal may change. Such changes were discussed in the previous section in connection with the energy band model. Let us recall, for instance, those processes in which valence electrons of insulators or semiconductors were promoted to the conduction band by light absorption, whereby defect electrons were produced in the valence band. The electrons and holes migrate separately in the crystal for some time, until their recombination. The energy liberated by the recombination may be transformed into photon emission (luminescence), a rather rare occurrence in complex systems at room temperature. More frequently, the recombination energy is transformed into lattice vibration, i.e. phonons are produced.

From experience it may be assumed that in a number of cases the energy transfer is achieved by a coupled migration of the energy-carrying electrons and defect electrons, rather than by their separate motion. Such a bound electron–defect electron pair is called an *exciton.* It resembles a hydrogen atom, the defect electron taking over the role of the proton. The exciton may be in various discrete energy states.

The exciton states are created by the cloud-like expansion of the electron and hole in the crystal (cf. section 1.2.2), therefore they form energy bands extending over the whole crystal. These bands are very narrow, and are localized in the forbidden gap just below the conduction band. With increasing energy the separate bands become increasingly dense, and finally coalesce with the conduction band. The exciton dissociates if it attains a sufficiently high energy, and similarly to the case discussed earlier the electrons promoted in this way to the conduction band migrate independently from the defect electrons left in the valence band. As long as the exciton does not dissociate, it does not participate in the electric conduction. Charge can be carried only by excitons dissociated into electrons and defect electrons.

Exciton states are characteristic of ideal periodic systems. If the lattice contains defects, the expansion of the loosely coupled electron–defect electron pair is stopped, and the exciton becomes localized at the defect. This occurs in biological macromolecules, for instance in nucleic acids, which may be thought of as crystals enriched with lattice defects. As a result, the excitons in them are strongly localized. In practice the effect of excitation is found to extend only to a sequence of 3–4 bases of the macromolecule. Consequently, the fate of absorbed energy, e.g. in the form of a photochemical reaction or phonon creation, will be governed by the immediate environment of the absorption.

1.5. Structure and function

1.5.1. The properties and structure of water

The properties of water. Water is of fundamental importance not only in the development of life, but also in its maintenance. The density of water is highest at 4 °C; freezing is accompanied by a density decrease. It follows that freezing starts on the surface of lakes and seas, while below the ice cover the conditions of life in water at 4 °C are still maintained. Living organisms generally exist within fairly narrow temperature intervals. The high heat capacity of water means that this condition is ensured for organisms living in rivers, lakes or seas. In terrestrial life water is an important thermal regulator in two ways. As a result of the high water content of the tissues and the high heat capacity of

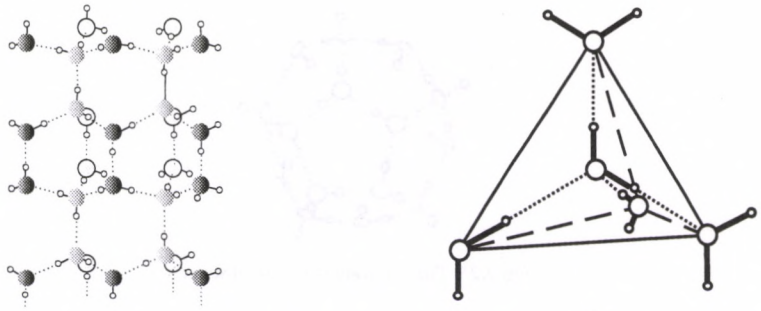


Fig. 1.24. The structure of ice
The larger circle denote oxygen atoms and the smaller ones hydrogen atoms.
The covalent bonds are indicated by continuous lines and the hydrogen bonds by dotted lines

water, the heat produced in the metabolic processes would increase the temperature of the body only slowly, even without thermal regulation. However, due to the high evaporation heat of water the superfluous heat is easily lost. The considerable surface tension of water plays an important role in the formation of the lipid and protein layers of the cell membranes. Further, since water is a good solvent for many inorganic and organic compounds, it is an excellent medium for biochemical reactions.

The structure of water. The properties of water are natural consequences of its structure. The water molecules are interconnected by hydrogen bonds in both the liquid and the solid state. Figure 1.24*a, b* shows the three-dimensional structure of ice. Every O atom occupies the centre of a nearly regular tetrahedron, and the adjacent O atoms occupy the vertices of the tetrahedron. For clarity, diagram *b* shows an arbitrary O atom and its environment. The H atoms between the O atoms are situated so that four H atoms are linked to one O atom by two covalent and two hydrogen bonds.

The melting of solids (including ice) involves a gradual splitting of the molecular bonds and an increase in the number of lattice defects (vacancies). In liquids the molecular order is restricted to much smaller volumes than in the solid state. With increasing temperature the number of defects increases, i.e. the order diminishes further. Simultaneously with the melting processes, ice undergoes structural transformation. The new structure consists of pentagonal dodecahedron units, as demonstrated in Fig. 1.25. These units are connected by further water molecules in such a way that one water molecule occupies the centre *(clathrate structure).* It is important that in this new structure the water molecules are more closely packed than in ice. All these processes together lead to a characteristic density change. The decrease of the long-range order and the increase in the vacancy concentration result in a density decrease. On the other hand, the change in the tetrahedral structure increases the density. This latter process is already apparent on melting, and is most marked at 4 °C. The factors causing a decrease in the density of water predominate only at higher temperatures.

It follows that an increase in temperature is accompanied by an increase in the mean kinetic energy of the molecules, together with the progressive splitting of the hydrogen

Fig. 1.25. The clathrate structure of water

bonds. This latter process requires a considerable amount of energy, which explains the high specific heat of water. The energy of one hydrogen bond in water is about 0.2 eV, which means that a single water molecule is bound with an energy of about 0.4 eV (~ 40 kJ/mol). The large values of the evaporation heat and surface tension are also explained by these relatively strong bonds.

The good solvent power of water is due to the relatively large dipole moment of water molecules. These weaken the bonds between the ions (atoms) in crystals by forcing apart the opposite charges, and this enables the water molecules to become wedged in between the lattice elements. The individual ions are then no longer surrounded by their original partners, but by water molecules (*hydrate sheath).*

Bound water requires special mention. This is a result of an interaction between water and the surface of some hydrophilic substance. This interaction may either be a hydrogen bond or it may be due to van der Waals forces. As a result, the structure of the water near the surface of the substance differs from the structure at a distance of a few molecular layers, which means that the structure of the bound water is different from that of the bulk water. Bound water freezes at a temperature below 0 °C, and its density is larger than that of bulk water. The difference is the larger, the closer the water molecules come to the surface of the hydrophilic substance.

The functional aspects of bound water are of considerable importance in biological systems. For instance, the water bound to membranes plays an essential role in the formation of the dynamic (liquid-crystalline) structure of the membranes, which means that it is an indispensable condition of transport processes (cf. section 1.5.5). Under normal conditions, a considerable proportion of the water content of cells is bound water. As an example, one lipid molecule in a phospholipid membrane or one base pair of a native nucleic acid ordered in a higher order structure (e.g. chromosome) bind ca. 10 water molecules.

1.5.2. Common features in the structure of macromolecules

The order of magnitude of the mass of biological macromolecules is in general from ten thousand to million dalton. All life processes are related to these molecules, and take place by their aid, are mediated by them. *Nucleic acids,* representing the genetic material of cells, enzyme and structural *proteins* (hemoglobin, myosin, collagen, etc.), storing and structural *carbohydrates* (starch, glycogen, chitin, cellulose, etc.) all belong to these molecules.

Though the chemical structure of these molecules is different, common structural features can be found in them, which play a role in the formation of their special properties and characteristic functions. In the following we deal exactly with these common features. Two such structural features are emphasized:

– their building up from subunits,

– the role of low-energy bonds in the maintenance of higher order structures.

The *subunits as building elements* are joined to form chain-like molecules (biopolymers) by covalent bonds. The bonds are usually formed by water release. In the case of proteins the building elements are *amino acids,* and the covalent bonds forming between them, the peptide bonds, produce the so-called polypeptide chains. In the case of nucleic acids the building elements are *nucleotides,* and the polynucleotide chain is formed by the phosphate-ester bonds between them. The subunits of structural and reserve carbohydrates are various *monosaccharides,* they are bound into a chain by the glycoside bond between the carbon atoms 1 and 4, or 1 and 6.

Further on only proteins and nucleic acids will be dealt with in detail. They are characterized by *the strictly,* genetically *determined number of their subunits* and by the fact that the monomers are usually linked into a single chain without branching.

An important property of the building elements is that the size and structure of the subunit parts forming the "backbone" of the macromolecule chain are identical, i.e. the *structure of the molecular backbone is periodical,* and the *aperiodical* feature of the molecule is related to the presence of *side groups* branching off the main chain. Figure 1.26*a* shows this periodicity and aperiodicity in the case of proteins, while Fig. 1.26*b* in the case of nucleic acids. The periodicity in the case of proteins is represented by the carboxyl, alpha carbon atom and amino groups and the aperiodicity by the various side chains of the 20 different amino acids occurring in the living organisms. The former groups are shown in the figure in detail; the latter ones are denoted by the symbols R_1, R_2, R_3, ... The periodic part of nucleic acids is formed by the sugar (pentose) phosphate backbone, while the bases (denoted by B_1, B_2, B_3, ... in Fig. 1.26*b*), oriented nearly perpendicularly to the backbone, produce the aperiodicity within the molecule. In the nucleic acids four different bases are present.

The consequence of the system of building element subunits is an important property of macromolecules, namely *the possibility of a great diversity* in their structure. This means

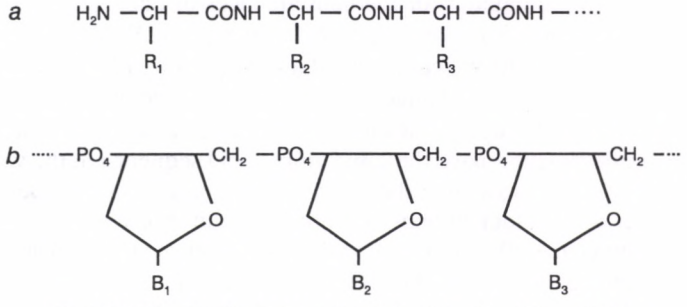


Fig. 1.26. The subunit structure of proteins (*a*) and nucleic acids (*b*)

that in the case of a chain consisting of a determined finite number of elements a high number of various polypeptide and polynucleotide chains can be arranged from the 20 different amino acids and the four types of nucleotides present in biological systems. In other words: since the primary structure of proteins and nucleic acids is given by the sequence of amino acids and nucleotide bases, in the case of a fixed chain length a great number of primary molecular structures can be constructed from comparatively few building elements. As an example let us consider a polynucleotide chain containing 10^6 nucleotides. The question is the following: in how many different ways can this chain be constructed from the four nucleotide bases? The calculation can be easily carried out with the following conditions:

– one of the four nucleotides can be selected to each site of the polynucleotide chain in *a random way;*

– the selection of each base is *independent* from the other bases.

It can be proved directly that in this case $4^{(10^6)}$ different base sequences are possible since for a DNA molecule constructed from k number of chain loops the number of possible base sequences is given by the number of arrangements with repetitions of k-th order and this is 4^k. – In the case of proteins the number of elements is 20 and the order of the arrangement is determined by the number of peptide groups, i.e. also by the chain length.

The great number of possible arrangements forms the basis of the *great diversity* found in nature with respect to both genetic material and proteins. The *great quantity of information* stored in these macromolecules is also related to the great variety of the primary structure of nucleic acids and proteins. This information can be estimated in the following way. In the case of a DNA molecule containing k number of nucleotides the realization of *one selected sequence* from the 4^k different possibilities has a probability of $1/4^k$. From this it follows that the uncertainty concerning the base sequence of the chain is $\log 4^k$, which is equivalent to the statement that the average information content of the molecule is 2^k bit (cf. section 8.1). If according to our previous example $k = 10^6$, the average information stored by the molecule is 2×10^6 bit. (In reality a smaller number is involved, as it is certain that the second condition does not hold.)

The replaceable or *changeable character* of the building elements is also the consequence of the construction of macromolecules from subunit systems, since the incidentally defective subunits can be removed any time and replaced by a perfect one. This feature has manifold *functional significance;* for example it ensures the possibility of internal control in the elimination of the consequences of damages caused by external effects (e.g. radiation, chemicals) as well. – With respect to DNA we refer to the functioning of repair mechanisms which ensure that the genetic information stored in the nucleic acid is constant, and the change of a single nucleotide may lead to the variation of a biological function through point mutation. – In the case of proteins: the proteins performing the same function have the same amino acid sequence in the various species, but there is a different amino acid in the chain which is related just to species specificity. As an example let us consider human and horse insulin molecules. In the so-called A chain of the two molecules 20 amino acids and their sequence are identical, only the ninth member of the chain is different: in human insulin it is serine, in horse insulin glycine.

The primary structure formed by means of covalent bonds determines the secondary

and higher-order structure of the macromolecule as well. In a given environment (pH, ion concentration, temperature, etc.) the chains assume a definite spatial arrangement, while within and between the chains *low-energy bonds* are forming which stabilize the structure. In this stabilization hydrogen bonds, ionic and van der Waals bonds play a role (cf. section 1.3.2).

The energy of a weaker bond is only between one tenth and one hundredth of an electronvolt, thus the bond can be broken thermally as well. However, due to the vast number of bonds the molecule may be stabilized.

The presence of these bonds ensures not only the stabilization but also the *dynamics of the molecules.* Namely, based on the results gained by the most recent structure-analysing methods (e.g. by the fine analysis of X-ray diffraction) the following picture may be formed concerning the macromolecules. Within the energy minima (potential valleys) determining the spatial configuration there are also smaller, some tenth to hundredth eV potential valleys, between which continuous transitions (fluctuations, fine structural rearrangements) may be formed as a consequence of the thermal motion. This phenomenon plays a very important role in the function of the macromolecules: based on it such intermolecular interactions can be explained which are significant steps in the regulation of a given function (cf. section 8.2.1). For example, on the grounds of the known "static" structure of hemoglobin, the oxygen molecule could not reach its binding site (the heme group) due to its size, the binding takes place nevertheless, since at the right moment a path is opened for the oxygen molecule by means of fluctuation.

1.5.3. Structure and some properties of proteins

Proteins are essential constituents of living cells, accounting for more than half (e.g. about 60% for bacteria) of the dry weight of the cells. Proteins display catalytic and contractile properties, and they also act as supports, provide protecting functions, participate in transport processes, and are basic substances of antibodies and certain hormones. In one bacterial cell e.g. any of approximately 1000 different types of protein molecules may be present, performing a large variety of functions. The multiple functions of proteins are connected with the nature, the number and the sequence of their components (*primary structure*) and their spatial arrangement (*secondary* and *tertiary structure*), which in turn depends on the primary structure.

1. Briefly on the structure. (a) *The primary structure.* Proteins are built up from *amino acids.* The smallest proteins, peptides, contain only a few amino acid residues, whereas the larger ones contain several thousand. In the various proteins 20 different amino acids may be present in practice. Their frequencies of occurrence and their sequences differ, but are characteristic of each protein type. In neutral aqueous solution the amino acid residues behave predominantly as zwitter ions. The dipole moment of an amino acid residue is approximately eight times larger than that of water. In an acidic medium amino acids behave as cations, and in alkaline solution as anions. Both their dipole moments and charges may be influenced by the R groups. Naturally, not only the individual amino acids, but also the polypeptide chains built up from the amino acid residues may have

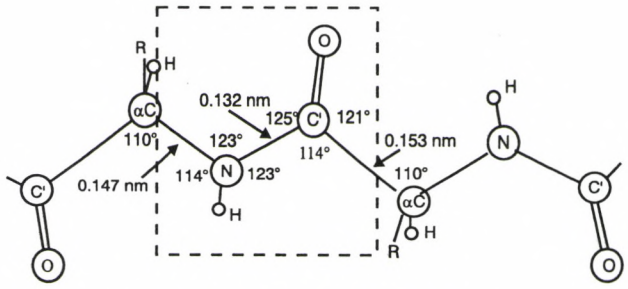


Fig. 1.27. Extended polypeptide chain
C′ denotes a carbon atom in a carboxyl group, and αC a carbon atom adjacent to the carboxyl group

dipole moments and electric charges, which are determined by the terminal amino acid units and the *R* groups.

(b) *The secondary structure.* A polypeptide chain may have a given specific form which, though determined by the primary structure, may also be influenced by other circumstances, such as the hydrogen ion concentration or the temperature. In their native environment, polypeptide chains are not extended. They always assume the energetically most favourable spatial arrangement (energy minimum) associated with the primary structure and the given environment.

Figure 1.27 demonstrates some specific features of peptide bonding and an extended part of a peptide chain, including the bond lengths and the bond angles. The chain is of a zigzag shape. The carbon (C′) and nitrogen atoms linked by peptide bonds are in nearly the same plane with the α-carbon and oxygen, further with the hydrogen and α-carbon atoms attached to them. This is indicated by the dashed frame in the figure. The planar arrangement of the peptide group can be accounted for by delocalized π-electrons which are responsible for the bond formation of the N–C′–O atom group in the same way as in the benzene rings of aromatic compounds. For this reason the C′–N linkage is a partial double bond, while the C′=O linkage cannot be regarded as a pure double bond; consequently, rotation around the bond axis is restricted. Free rotations are possible only around the α C–N and the C′–αC bond axes. The *secondary structures* formed by the peptide chains differ in the positions of the bond planes with respect to one another. These structures are in all cases stabilized by infra- and intermolecular hydrogen bonds.

There are several types of secondary structure. One of the most frequent is the *beta-form* (also called pleated sheet). As may be seen in Fig. 1.28*a*, the beta-type peptide chains are arranged in slightly folded layers. Figure 1.28*b* depicts part of a single layer chain, whose axis consists of the αC–C′O–NH–αC groups. The *R* groups are perpen-

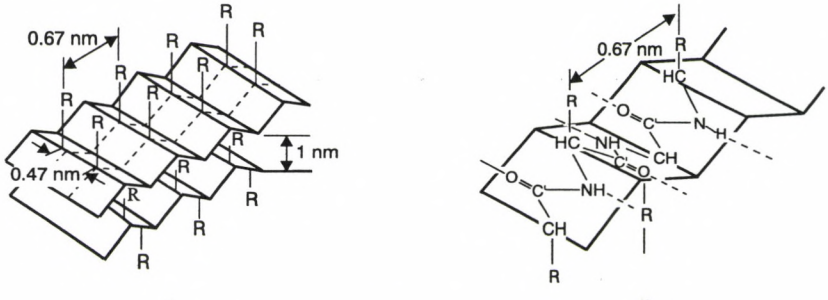


Fig. 1.28. Outline of the beta-form

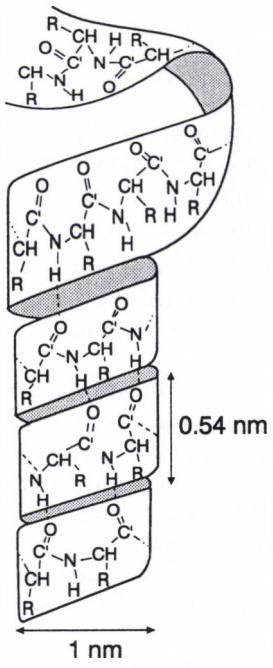


Fig. 1.29. Outline of the alpha-form

dicular to the axis and are situated alternately on opposite sides of the chain. One layer consists of several parallel chains, whose axes are shown in diagram *a* by dotted lines. The stability of the chain array is ensured by hydrogen bonds (denoted by dashed lines in diagram *b*) linking the C′ = O and N–H groups with the N–H and C′ = O groups of the adjacent chain. The *R* side groups are situated between the adjacent sheets. Because of the relatively large distance between the neighbouring sheets, they can easily slide on each other.

Another frequently occurring form is the right-hand *alpha-helix*, whose more important features are depicted in Fig. 1.29. The peptide chains are situated along a helix on the lateral surface of an imaginary cylinder. The diameter of the cylinder is ca. 1.0 nm and the pitch ca. 0.54 nm and an average of 3.6 amino acids are found per turn. It follows from these data that the adjacent amino acid moieties are rotated with respect to each other by 100° around the cylinder axis, and displaced by approximately 150 pm relative to each other along the axis. The individual peptide groups are situated in tangential planes of the cylindrical shell. The *R* groups protrude from the axis. The turns of the helix are linked by hydrogen bonds between the hydrogen atom of the N–H group and the oxygen of the C′ = O group of the adjacent turn. Consequently, the hydrogen bonds are nearly parallel to the cylinder axis.

The importance of the environment of the molecule for the formation of the secondary structure is demonstrated by two examples. Both examples are associated with the synapses which play a significant role in the conduction of stimuli (cf. section 7.3.3).

The transmission of the stimulus is carried out by the acetylcholine molecule, which is in interaction with the acetylcholine receptors of the synaptic membrane; the latter are of protein nature. In the active state, for example, in the cell membrane of the torpedo ray (torpedo marmorata) 47% of the chain of the receptor protein contain α helix, 25% have a β lamellar structure, while the rest of the chain does not have an organized structure. If

the receptor gets into interaction with tubocurarine, a molecule competitively inhibiting the effect of acetylcholine, the stimulus conduction becomes impossible. All this means the following changes in the structure of the receptor protein: the proportion of the α structure in the chain decreases by about 10%, and the proportion of the unorganized parts increases accordingly. – Another example. The irreversible inhibition of a cholinesterase enzyme catabolizing acetylcholine is brought about by the interaction between the enzyme and the inhibitor. Under the influence of the inhibitor the proportion of the α helical structure in the enzyme protein is decreased by about 10%, that of the β structure is increased by about the same extent, while the proportion of the unorganized structures remains unchanged. Both examples may be also considered as structural rearrangements of proteins as liotropic liquid crystalline systems (cf. section 1.4.4) under environmental influences.

(c) *The higher-order structure* is a comprehensive name including our data on the shape of the protein molecule, on the position of its subunits related to each other, and on the chains in the subunits ordered in regular secondary structure or in "disordered" state.

Classified by shape, proteins are fibrillar or globular. To the first group belong e.g. keratin and collagen, to the latter one hemoglobin and myoglobin. It has to be noted that in a given molecule – though in a way not yet widely known – always the primary structure (amino acid sequence) of the chain determines the spatial arrangement, the ratio of the chains ordered into the secondary structure to the elements of the "disordered" structure, etc.

Recently considerable success has been achieved in the sequencing of numerous proteins, but for the elucidation of the higher-order structure it is not essential to know fully the primary structure, moreover information obtained from the determination of the higher-order structure may promote amino acid sequencing. There exist already highly developed physical methods of structure analysis rendering possible – naturally if possessing some fundamental data on the molecule – to reveal the structure down to *atomic level* with a resolution of 0.2 nm. The determination of the spatial structure of a protein to such depth is an enormous task, since depending on the molecule size (disregarding hydrogen atoms) the exact localization of one thousand to one hundred thousand atoms has to be dealt with. Among physical methods X-ray and electron diffraction play a considerable role (cf. section 4.5).

To illustrate the main phases of structure determination by X-ray diffraction the hydrogen peroxide decomposing catalase enzyme from *Penicillium vitale* was chosen. It is known that catalase consists of four protein subunits of identical size (tetramer), its molecular mass being about 300,000. Each subunit is a polypeptide chain and their characteristic constituent – similar to the hemoglobin or cytochrome *c* molecules – is a heme group containing iron.

For the measurement, a *crystalline* protein is necessary, which can be produced in the case of globular proteins in appropriate conditions. In Picture 1.1*a* (in the Supplement) the catalase crystals of 0.3–0.5 mm size, produced for X-ray analysis, can be seen. The lattice points of the crystals are occupied by two subunits, forming one half of the catalase molecule.

The extremely complex composition of the molecules means that the X-ray diffraction pattern is rather complicated. However, evaluation may be facilitated by special methods, such as *heavy atom substitution.* This method is based on the observation that complexes containing heavy metals or other heavy elements are bound to certain specific, well-defined sites in the molecules, which in this way become labelled. Since the heavy atoms are

surrounded by a cloud of many electrons which strongly diffract X-rays, the reflexions due to this effect are well visible in the diffraction pattern. However, this method yields valuable information only if labelling does not change the molecular configuration, and the crystal containing the labelled molecules is isomorphous with the original one.

In case of the fungal catalase, e.g. platinum, lead, mercury and uranium heavy metal ions were applied to produce isomorphous proteins.

Picture 1.1*b* (in the Supplement) shows the X-ray diffraction pattern of a native catalase crystal. The regularly placed interference spots prove that the macromolecules forming the crystal are structurally identical. One diffraction pattern contains – depending on the size of the molecule – several thousand, even several ten thousand interference spots of different position and intensity. In case of catalase the reconstruction of the density of scattering electrons belonging to individual atoms of the protein molecule was carried out via the analysis of nearly 70 thousand such spots. Producing and evaluating the data of such enormous number are naturally possible only by the aid of automated measuring technique and computerized data processing.

Figure 1.30*a* shows the electron density map of one single amino acid, phenylalanine, and Fig. 1.30*b* that of a peptide chain fragment. The curves returning to themselves connect the locations of identical electron density, similar to level lines in geographical maps. From the figures it is conceivable that e.g. the electron density map of a catalase subunit containing about 660 amino acids consists of quite numerous elements, therefore if the chain is long, it is divided into several different segments and the analysis is carried out segment by segment. From the electron density map of the peptide chain one may infer e.g. the secondary structure of the different chain segments (alpha-helix, beta-form, disordered), and by "putting together" these segments the model of the molecule can be constructed.

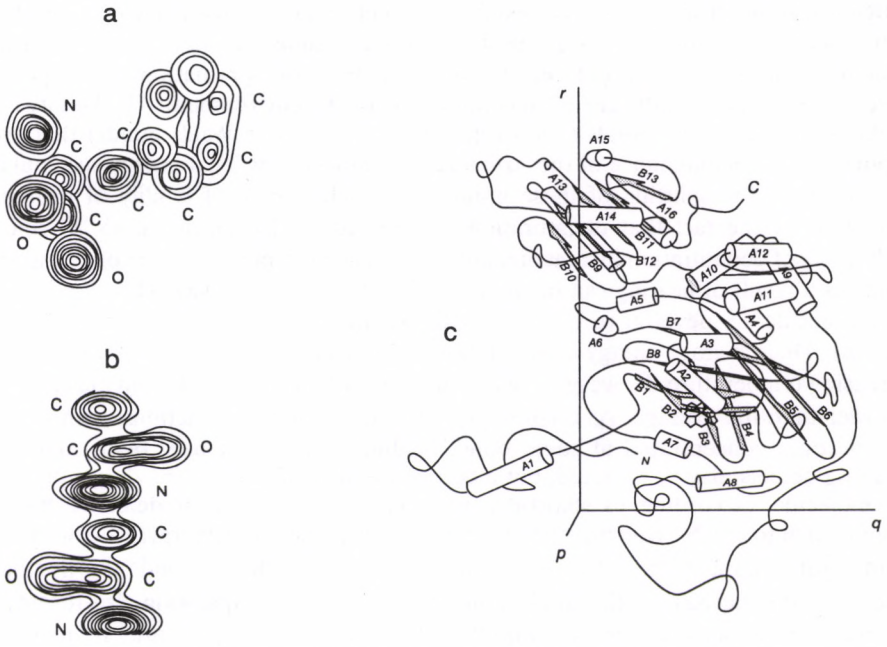


Fig. 1.30. Diagram relating to the X-ray diffraction analysis of proteins (data of B. K. Vainshtein and W. R. Melik-Adamyan)

a: electron density maps of phenylalanine and *b:* polypeptide chain segment. The positions of oxygen, nitrogen and carbon atoms are indicated by O, N, and C, respectively; *c*: reconstruction of the three-dimensional structure of a catalase enzyme subunit based on the electron density. The heme group is in the interior of the molecule; details denoted by A1–A6 represent alpha-helical and arrows denoted by B1–B13 beta-sheeted structures. Single line refer to disordered chain parts

Figure 1.30*c* shows the spatial arrangement of a catalase subunit and the localisation of the regular secondary structures. It can be seen that the segments of the 660 amino acids are arranged in a structure of 16 alpha-helices (A1–A16) and 13 beta-sheets (B1–B13). About one sixth of the amino acids forms a "disordered" chain or links the various ordered domains of the molecule. Inside all four subunits are situated the most important parts, the four so-called *active centres.* An iron-containing heme group is situated in each centre which is oxidized by hydrogen peroxide. According to measurements, the active centres are located at 1.7 nm from the molecular surface and 2.3 nm from the molecular centre. A chosen iron atom is to be found at distances 3.1, 3.4 and 4.5 nm, respectively, from the other three iron atoms in the molecule.

Comparison of the primary and higher-order structures of proteins with the same function but of different origin presents possibilities for interesting conclusions. According to experience, in case of proteins (e.g. catalases) of identical function but originating from cells of different living organisms, the three-dimensional structure is much more conservative than the amino acid sequence itself. This statement is especially true for secondary structures forming the environment of the active centre, which shows that in respect to evolution the stereochemical mechanism of enzymatic effect is more stable than the genetic information on the structure of proteins.

2. Denaturation. Proteins are very unstable molecules; their structures may easily change so that the molecules can no longer perform their functions, i.e. they become denatured. Structural changes may be induced, for instance, by various radiations, a temperature increase or a change in the chemical composition of the environment. The lability is due to the weakness of the bonds (hydrogen bonds and van der Waals bonds) that are of importance for maintenance of the structure. A relatively low local energy accumulation may result in the splitting of these bonds. The high degree of lability of proteins is illustrated by the fact that denaturation can be observed even at ca. 45 °C, and the probability of denaturation increases rapidly with rising temperature. The breaking of the bonds as a result of heating can be interpreted in the following way. The atoms making up a molecule are never at rest, but oscillate around the equilibrium positions (thermal motion). Their average energy is well-defined at a given temperature and increases with increasing temperature. However, the actual energy of an individual atom may be higher or lower than the average. As a consequence of the atomic interactions, higher energy may become concentrated at some bond, leading to its weakening or breaking. The resulting defects may be reverted, but a certain number of defects are always present in the molecule. According to statistical mechanics, the number of defective bonds is proportional to $e^{-\Delta\varepsilon/kT}$ (Boltzmann factor), where $\Delta\varepsilon$ is the bond energy, T is the absolute temperature and k is the Boltzmann constant. With low-energy bonds the number of defects is large because of the small value of $\Delta\varepsilon$. The above expression also reveals why the number of defects increases rapidly with increasing temperature. In reality the situation is even worse, since in the environment of defective bonds another defect may develop easier: the environment behaves as if $\Delta\varepsilon$ would be smaller. This rapidly progressing process leads to denaturation which may be observed already around 45 °C in case of proteins.

The thermal denaturation of proteins plays a decisive role in the thermal inactivation of living cells. Experience shows that the thermal denaturation kinetics of a given cell (e.g.

yeast, bacterium, *Drosophila melanogaster)* coincides with the denaturation kinetics of the critical protein[3] of the cell, which means that the denaturation of this protein leads to the killing of the cell.

1.5.4. Structure and some properties of nucleic acids

Nucleic acids play an extremely important role in various cell functions, since these substances are responsible for the storage, transformation and transmission of genetic information (cf. section 8.1). The storage of genetic information is achieved by the *identical replication* of the nucleic acids, which means the formation of molecules completely identical to the original one. Transmission and implementation of the information are accomplished by *transcription* and *translation* of the signals carrying the information. Thus, since the structure of proteins is determined by the nucleic acids, and since the function of proteins depends upon their structure, it may be said that the structure and function of a living cell are determined ultimately by the nucleic acids.

The great variety of living organisms is thus connected with the great variety of genetic information, which in turn depends upon the numerous variations possible in the primary structure of nucleic acid molecules.

1. Briefly on the structure

a) *The chemical structure.* Depending on their chemical structures, nucleic acids can be classified into two groups: *deoxyribonucleic acids* (DNA) and *ribonucleic acids* (RNA).

Nucleic acids, as it was mentioned, are macromolecules formed by the linkage of *nucleotides.* Polynucleotide chains contain several thousand, several hundred thousand, or even several million nucleotides. Nucleotides consist of nitrogen-containing bases, phosphate groups and sugars with five carbon atoms (DNA contains 2-deoxy-D-ribose, and RNA contains D-ribose).

Four kinds of *nucleotide bases* are found in DNA: adenine (A) and guanine (G) with a purine skeleton; thymine (T) and cytosine (C) with a pyrimidine ring. In RNA the same two purine bases and cytosine occur as in DNA, but the second pyrimidine is uracil (U) instead of thymine. (U differs from T only in not containing a methyl group on carbon atom 5; cf. Fig. 1.31). Depending upon the environment, the bases may exist in different tautomeric forms. Figure 1.31 shows the structures most frequently occurring in nature.

The different bases are not present in equal numbers in a given nucleic acid, but the double-stranded DNA and RNA contain the same number of adenine as of thymine (uracil) and as many guanine as cytosine.

b) *Three-dimensional structure.* The first conformation successfully elucidated was that of the most frequently occurring double-stranded DNA (Watson and Crick 1953, Wilkins et al. 1953).

The most important method of structure analysis for studying the three-dimensional structure of DNA is X-ray diffraction (cf. section 4.5.1). The investigations are carried out

[3] The critical protein is the protein molecule which undergoes denaturation at the lowest temperature.

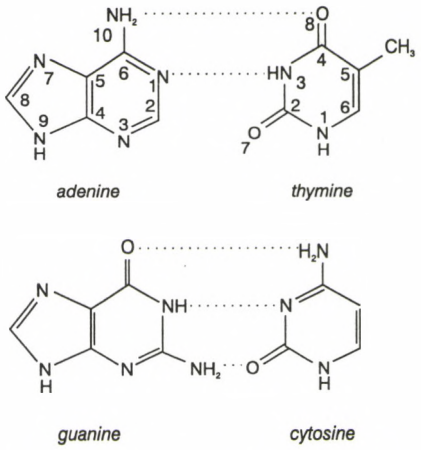


Fig. 1.31. Structural formulae of nucleotide bases
The bases on the left contain purine rings, and those on the right contain pyrimidine rings. The numbering sequences for A and T also apply to G and C, respectively. The dotted lines indicate the sites of hydrogen bond formation between the A–T and G–C base pairs in DNA

either on single DNA fibres or on polycrystalline samples (Laue and Debye–Scherrer method, respectively). As examples two photographs are shown in Pictures 4.8 and 4.9 (in the Supplement). Essential information on DNA structure can be obtained by almost every variation of optical spectrometry (cf. section 4.4). A common advantage of optical methods is that they can be applied for nucleic acids in solution too, therefore it is not necessary to produce molecular crystals for the studies.

Summarizing all experimental results the following statements are generally true for the three-dimensional structure of DNA. The DNA molecule consists of two helical polynucleotide chains (Fig. 1.32). The helices are generally right-handed. The outer parts of the chain are occupied by the sugar and the hydrophilic phosphate

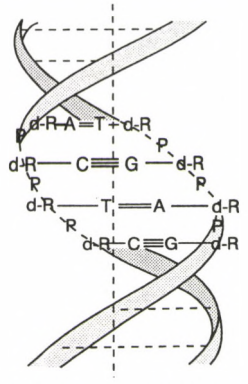


Fig. 1.32. Outline of the DNA structure
d–R: deoxy-Ribose, *P:* phosphate, *A:* adenine, *T:* thymine *C:* cytosine, *G:* guanine. The thin lines drawn within the helices, and the double and triple lines between the bases in the centre of the figure denote hydrogen bonds

groups, while the hydrophobic bases are inside the double helix. The atoms of one base are nearly coplanar, and the planes of the bases are parallel to each other. The two helices are connected by the bases opposite each other in the individual chains. Further, opposite the positions occupied by adenine in one of the DNA chains the other chain contains thymine. Guanine is similarly always paired with cytosine. The two members of a base pair are situated in one plane, and the members of the complementary pairs are connected by hydrogen bonds: for the A–T pairs by 2 hydrogen bonds, and for the G–C pairs by 3 hydrogen bonds per pair. (Figure 1.31 shows the atom pairs forming these hydrogen bonds.) These base pairings explain the analytical chemical findings relating to the frequencies of occurrence of the bases, and also account for their spatial arrangement, since the smaller pyrimidine is always coupled with the bulkier purine.

The special three-dimensional structure (conformation) of DNA molecules depends considerably on their environment (water content, presence and concentration of mono- and multivalent ions, interaction with proteins and other DNA segments, etc.).

The most frequent conformation is the so-called *B-form* stable in an ion environment corresponding to the composition present in the cells, but due to local changes in the environment it may be transformed into other forms, too. The model of the B-form and its cross-section constructed from the X-ray diffraction results are shown in Figs 1.33*a* and *b*. In part *a* the regular progress of the sugar–phosphate backbone can be clearly seen, producing on the structure alternately major or minor grooves. In this form a 3.4 nm high turn contains 10 nucleotide bases, the base planes are nearly perpendicular to

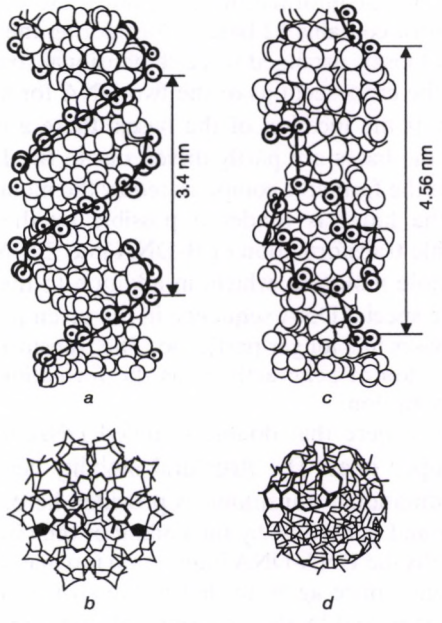


Fig. 1.33. Structural characteristics of B- and Z-DNA. In the ball model the backbone is marked by a thick line
a: ball model of B-conformation; *b:* cross-section of B-conformation: one base pair inside the structure and the sugar part belonging to it are marked; outside, the phosphate groups are located; *c:* ball model of Z-conformation; *d:* cross-section of Z-conformation, the marked base pair is in the outer part of the structure

the axis of the helix. It can be seen on the cross-section of 2.0 nm diameter that the hydrophobic bases (a G–C pair in the figure) are inside the structure, while the phosphate groups are located on the surface. In a medium with less water content a conformation slightly different from the B-form (though also right-handed) is found. In a dried bacterium spore, for instance, the so-called *A-form* occurs which, considering geometric parameters, is more compact than the B-form: one turn of the right-handed helix is 2.8 nm high and the diameter of the helix is also greater than at the B-form: 2.2 nm.

The base planes are also parallel here but they form an angle of about 20° with the plane normal to the helical axis. According to experience the change of the environment may transform easily the A- and B-forms into each other.

In physiological environment the so-called *Z-DNA* (Z-form) is less stable than B-DNA, but the B→Z transformation may occur under *in vivo* conditions, too. The Z-form is produced by multivalent metal ions (e.g. Mg^{2+}), alkaline polyamines (spermine, spermidine), special proteins that stabilize the structure. The transformation is reversible, i.e. another change (restoration of the former state) brings about again the B-form. Figures 1.33*c* and *d* show the characteristics of the Z-form. In this conformation – differently from A- and B-DNA – the two polynucleotide chains are wound in a left-handed way. In Fig. 1.33*c* the behaviour of the sugar-phosphate backbone can be well seen. While in the B-form it is running regularly, in the Z-form it displays characteristic "zig-zag" shapes, hence comes the name Z-conformation. Comparison of Figs 1.33*a* and *c* shows further that in the Z-conformation the grooves are irregular and of about the same depth. Here one turn contains 12 bases. Considering the helical pitch of 4.56 nm too, Z-DNA is the most loosely arranged three-dimensional structure known up to now. From Figs 1.33*b* and *d* the cross-sections of the two DNA forms can be compared. The most striking difference is the location of the bases: there are phosphate groups also inside the Z-DNA and the bases are partly driven out to the DNA fibre surface. (The electrostatic repulsion of the former is compensated by the metal or positive organic ions and the presence of the latter is rendered possible by the interaction with other molecules.) The reversible transformation of B-DNA into Z-DNA takes place usually in some sections of the whole molecule, which, in addition to the mentioned interactions may be promoted by the special DNA sequence in the given place (e.g. the variation of purine and pyrimidine bases in a longer part), too. The transformation and the reversion play a role in the regulation of gene activity, as the formation of the Z-conformation makes possible the transcription.

It should be mentioned here that double-stranded DNA forms are closed circular molecules rather than open ones. The structural mobility necessary for the biological functions, i.e. the conformational transitions, is provided here, too, partly by the interaction with the proteins and partly just by the coiling of the circular chains, since at the formation of the superhelix the DNA-DNA interaction too may produce Z-conformation. Here the attention is called once again to that in the structural changes of the nucleic acids an important role is played by the environment, more exactly by their interaction with the molecules in the environment. This indicates that the nucleic acids, similarly to the proteins, have a liquid crystalline character.

The double-stranded RNA may form helical conformations more or less similar to

DNA, but the complementary base of adenine in this case is uracil. It often occurs that certain sections of a single RNA chain are complementary to one another, and an intrachain helical structure stabilized by hydrogen bonds is formed affecting only one detail of the molecule (e.g. transfer RNA, virus-RNA).

From the discussion so far, partial answers can now be given to the questions concerning the relation of the structure and function of the nucleic acids. Such problems are e.g. the storage, conservation, transfer and translation of the genetic information.

c) *A few words on electronic structure.* Figure 1.34 depicts the absorption of a DNA solution. The spectrum shows that the molecules absorb in the ultraviolet range; the first absorption maximum is at 260 nm, which means that the smallest excitation energy of the DNA electron structure is about 4.8 eV. Quantum chemical methods exist for the determination of the electronic states (e.g. charge density characteristics for individual atoms, bond strengths and ionization energy) for the individual nucleotides. However, as yet no theoretical method is known for interpreting the total electronic configuration of the aperiodic DNA system. At any rate, the optical absorption suggests that the width of the forbidden band is about 4.8 eV.

Though the optical absorption spectrum of DNA is similar to the spectrum of the individual nucleotides making up the macromolecule, it still cannot be constructed as the sum of these component spectra. In fact, the π electrons distributed around the atoms of the bases above one another interact by van der Waals type forces (stacking interaction) both in the ground state and in the excited states generated by the absorbed UV light. These interactions are manifested in a weaker absorption than the sum of the absorptions of the individual components. For a double helix, for instance, the attenuation may amount to 30–40%. This is the *hypochromic effect.* Study of the absorption spectrum yields information not only on the primary, but also on the secondary structure, since the height of the absorption band maximum is characteristic of the strength of the stacking interaction.

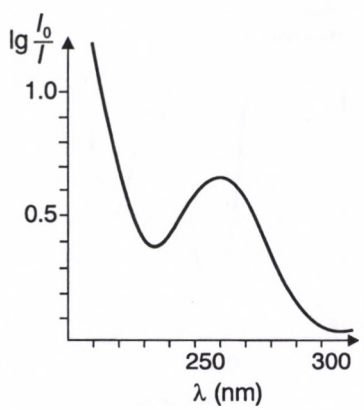


Fig. 1.34. Ultraviolet absorption of DNA obtained from a bacteriophage. On the ordinate the extinction is given

2. Some properties. A comprehensive study of the relations between the conformation and electronic structure and the biological function of DNA molecules is in progress. Among the essential questions are the elucidation of the structural transformations leading to mutations, and the understanding of the connections between the DNA structure and the mechanism of duplication. Detailed research work is continuing to investigate the structural and functional changes due to spontaneous effects and various agents.

In this section some interesting results will be discussed.

a) What are the reasons for the defects in the base sequence, and how do these defects eventually lead, for example, to mutations? This question has several answers; actually the *quantum mechanical tunneling effect* may be taken as a possible departure for assessing the formation of mutations. According to classical mechanics, a particle in a potential valley can surmount its energy barrier only if it has a sufficiently high energy to reach the top of the potential barrier. Quantum mechanics, however, teaches us that particles of lower energy may also cross the barrier, as if some tunnel existed. Thus tunnel effect plays an important role in many cases, for instance in radioactive α-decay, or the migration of electrons through the potential barriers in connected atomic systems. The tunneling process also accounts for the crossing of electrons through the connections of electric power lines. In this context let us investigate more closely the motion of hydrogen atoms or protons in the hydrogen bonds connecting bases. In a hydrogen bond the hydrogen atom moves in a force field where the potential energy exhibits minima at two positions (Fig. 1.35). In the following discussion we shall distinguish between a DNA molecule in the ground state, and one in an excited state due to UV irradiation or some other excitation. The potential curve in the first case is strongly asymmetric, as shown in Fig. 1.35, but in the second case the curve is nearly symmetrical, and the depths of the two energy valleys are approximately the same. In the ground state of the molecule the hydrogen atom is localized with high probability in the deeper valley, which also means that the hydrogen atom can usually be found in the vicinity of the atom with which it was associated before formation of the hydrogen bond. The hydrogen atom can pass into the shallower valley by means of tunneling only if it takes up energy (e.g. infrared light) to attain the level corresponding to the bottom of the shallow valley, as indicated by the dashed line in the figure. With excited molecules the situation is different, since in this case the hydrogen atom may occupy either of the two potential valleys with approximately the same probability, i.e. it may be found in the vicinity of either pillar atom, i.e. in the vicinity of either member of the base pair. The transfer of the hydrogen atom from one base to its pair implies that *tautomeric bases* different from the original ones are created. If this occurs in DNA duplication, the complementary bases corresponding to the unusual tautomeric forms appear. This process may lead to defective base sequences, which possibly results in mutation. This picture explains the experimental fact that mutations arise very rarely under ordinary circumstances, whereas they are produced with high probability by ultraviolet light or high-energy photons or particles.

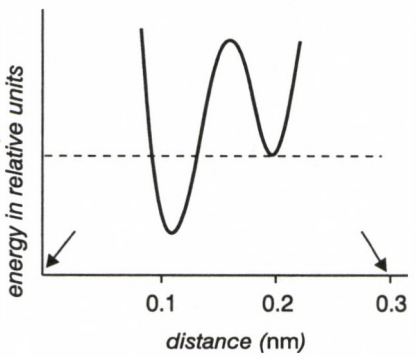


Fig. 1.35. Variation in potential energy along the distance between the pillar atoms in a hydrogen bond
The arrows indicate the positions of the pillar atoms

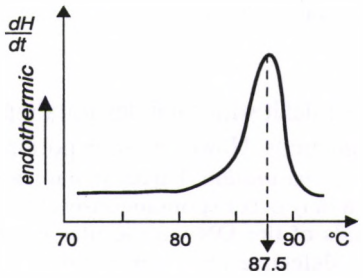


Fig. 1.36. Melting of DNA obtained from phage T7
The abscissa shows the temperature of the solution and the ordinate the change of its enthalpy per unit time. The melting point is also indicated

b) The fundamental and extensively studied effect of *denaturation* in biological macromolecules means the loss of the characteristic biological properties. As mentioned already with regard to proteins (cf. also section 1.5.3), denaturation is connected with structural changes and is reminiscent of the melting of solids, since in both cases ordered structures become less ordered. As an example, mention may now be made of solutions containing DNA molecules. From experience it may be concluded that the double helices of the DNA molecules become separated first at some places, and subsequently along the whole chain, and the resulting single chains assume a less ordered coil-like form. This process may be compared to the melting of solids in that it is accompanied by an energy uptake, the amount of which can be measured by microcalorimetry (Fig. 1.36). The diagram demonstrates the melting of T7-DNA as an example. The transition takes place in the temperature interval between 80 and 92 °C. The melting point is to be found roughly at the centre of this interval, at 87.5 °C. It should be stressed, however, that the interval of transition and the melting point as defined depend upon the circumstances, for instance upon the hydrogen ion and other ion concentrations of the solution containing the DNA molecules. For DNA molecules of various origins the melting point also depends on the relative quantities of the G–C and A–T base pairs. The higher the G–C content, the higher the melting point under otherwise identical conditions. For instance, at neutral pH and 0.2 mol/l KCl concentration the DNA molecule of *Diplococcus pneumoniae* with a 40% G–C content melts at ca. 85 °C, whereas the melting point of the DNA molecule of *Mycobacterium phlei* with a G–C content of ca. 75% under identical circumstances is approximately 97 °C.

If a heated DNA solution is cooled at a sufficiently slow rate, the vast majority of the separated chains can recombine, and double helices are formed again. This process is called *renaturation.* While denaturation is reminiscent of the melting of crystals, renaturation rather resembles freezing; the interpretation of these processes is similar too. With increasing temperature, i.e. with increasing thermal motion of the atoms, the low energy (hydrogen, van der Waals) bonds break first, and the two helices become uncoiled. New bonds are formed between the atoms of the isolated chains, which finally results in coil-shaped molecules. The melting point is the temperature above which this

coil-like structure becomes relatively frequent, whereas below it the helix form is energetically more stable.

c) In the previous point we dealt with total denaturation, which finally results in the cessation of the biological functions. However, an important role in biological functions is attributed to *partial (local) denaturation.* Partial denaturation means the local splitting of the hydrogen bonds, which may occur spontaneously even at 37 °C. If the mean binding energy of the hydrogen bonds of the DNA molecules is 30 kJ/mol, 10^5 base pairs in a DNA molecule contain one defective hydrogen bond; this can be calculated from the Boltzmann distribution, considering thermal motion alone. It has been experimentally demonstrated that during molecular mechanisms connected with DNA (e.g. transcription, duplication, genetic recombination) the presence of the appropriate enzymes or enzyme substrates leads to a considerable increase locally in the number of hydrogen bonds broken in the environment of the interaction. In the case of transcription this means a 50-fold increase in the defect concentration. This local denaturation permits mRNA synthesis corresponding to complementarity.

Nucleic acid–protein complexes. Nucleic acids usually exist in nature in interaction with proteins, in the form of nucleic acid–protein complexes. Both DNA and RNA may form complexes; examples of the former case are chromatin, i.e. the material of the cell nucleus, and DNA-containing viruses, and examples of the latter are the cell ribosomes and the RNA viruses.

Some of the interacting proteins exhibit enzyme activity; these take part, for example, in the transformation and implementation of the genetic information stored in the nucleic acids. Other proteins are essential in the higher-order structure of nucleic acids, thereby ensuring the information-storing function of the genetic material (mainly DNA). Since the storage and transformation are accomplished by the interaction of one and the same nucleic acid with different proteins, this is possible only if the complexes possess a high structural mobility ensured just by the low-energy bonds (cf. section 1.5.2), i.e. even small local changes (e.g. the appearance or disappearance of some ions or molecules) may result in the dissociation of the complex and produce a new one, or modify the three-dimensional structure of the nucleic acid (cf. also Picture 4.4*c*, in the Supplement). Deeper insight into these processes is given by an understanding of the molecular structure.

To establish the molecular structures of the complexes, almost all up-to-date methods of physical structure analysis are applied, e.g. small-angle X-ray and neutron scattering, Raman spectroscopy and UV absorption and luminescence spectroscopy (cf. sections 4.3–4.6) which yield information from various aspects on the nature of the nucleic acid–protein interaction and its consequences. Some characteristic, common structural features of DNA-containing complexes will be discussed in this section.

The inner core of the complex is a protein; the double-stranded DNA is coiled around this, forming a superhelix. Further protein molecules are attached to the outer side of the complex. The DNA conformation in the superhelix is nearly of the B-form (see Figs 1.33*a*, *b*), but due to the interaction with the proteins it differs somewhat from it, the regular arrangement of nucleic acid bases being distorted: the distance between the base planes, the positions of the planes relative to each other and to the helix axis, etc. are changed. This indicates that the π electron interaction between the base pairs is smaller in the complexes than in B-form DNA (e.g. in solution). On the other hand, if the nucleic acid–protein interaction ceases partially or totally (e.g. due to slow heating or

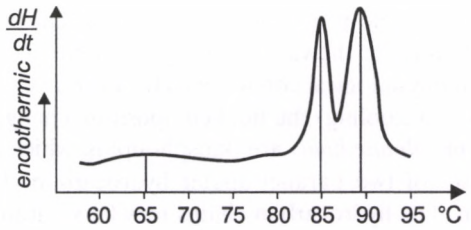


Fig. 1.37. Phase transitions of phage T7

removal of the protein), a hypochromic effect can be detected in the UV absorption band of the nucleic acid, which shows that the π electron interaction between the base pairs increases and a DNA with regular B-conformation appears in the solution.

Another example of the behaviour of the nucleic acid–protein complexes is that heating causes a structural rearrangement of chromatin and virus nucleoproteins. As an example of this, Fig. 1.37 shows the result of a microcalorimetric measurement on the phase transitions of a bacterial DNA virus, phage T7. The diagram shows transitions at three different temperatures, characterized by different heat absorptions. The difference from Fig. 1.36 is striking; this is related to the fact that we were dealing there with pure DNA and here with a nucleoprotein. From a biological aspect the transition at 65 °C is of special interest. This can be attributed to the fact that due to heating the nucleic acid–protein complex partly dissociates and is partly transformed, and the superhelical structure of the DNA is destroyed. The process may take place under *in vivo* conditions too, though in that case it results not from a temperature increase, but from a change in the physico-chemical environment. This was mentioned above in connection with the significance of local changes in the breaking of weak bonds and the formation of new ones. This fact produces a situation favourable for the transmission of information stored in the DNA (replication, translation).

Finally, the double maximum (at 85 °C and 89 °C) should be mentioned, though it has no direct biological significance. The double maximum is also characteristic of the melting of DNA (cf. Fig. 1.36), but the process is now modified by the residual protein–DNA interaction: the DNA stabilized by the interaction undergoes a phase transition at 89 °C, while in the maximum at 85 °C the conformational change of the proteins also plays a role.

1.5.5. Structure and some properties of biological membranes

The membranes of cells and of their individual cell constituents (e.g. the mitochondria) are composite systems consisting of smaller molecules, whose main components are lipids and proteins. The structure of the membranes – considering mainly their lipid constituents – reminds of that of the macromolecules, since they are also composed of subunits, and the weak energy bonds characteristic of macromolecules are important in the preservation of the molecular structure. However, they differ from the macromolecules, since the subunits – in the present case the lipid molecules – are not connected into chains (cf. section 1.5.2), instead they form layers, whose monotony is disrupted by larger protein molecules.

The membranes play a determining role in maintaining and regulating the functions of the cells and organisms. Membranes with various structures and functions constitute approximately 80% of the total dry-matter content of animal cells; they ensure the constant shape and the mechanical stability of the cells, and the concentration difference (essential in maintaining the life functions) between the intracellular and extracellular space. Further, the membranes are responsible for the transport of ions and molecules participating in metabolic processes.

1. Briefly on the structure. The proportion of the lipids and proteins constituting the membranes, though different for the various membrane types, may be regarded constant within given genetic and physiological conditions. However, in case of a lasting change of the external conditions (e.g. cooling) the lipid composition changes (cold adaptation).

The majority of the *membrane lipids* are phospholipids, which consist of a polar head group and in most cases of two parallel apolar hydrocarbon chains containing 14–18 carbon atoms per chain. The hydrocarbon chains may be saturated, or may contain one or more double bonds. Figure 1.38 depicts the phosphatidylcholine (lecithin) molecule, which is a component of every biological membrane. The molecule consists of a glycerine backbone (*a*), a choline phosphate group (*b*), and two palmitic acids (16 carbon atoms) linked (*c*) with the hydroxyl groups of glycerine. Besides the polar phospholipids (e.g. phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine), the various membranes contain among others considerable but different amounts of apolar cholesterol, which is important in the formation of the membrane structure. The chemical composition allows the lipid molecules to form van der Waals interactions with one another and also with other molecules.

The quantity and amino acid composition of the *membrane proteins* also vary considerably in the different cell types. The protein/lipid ratio ranges between 0.3 (myelin) and 3.0 (bacterial membranes), but in most cases it may be taken as 1.0 (erythrocytes, the outer membranes of mitochondria, etc.). Similarly, membrane proteins with various functions contain polar, weakly polar and apolar amino acids in approximately the same proportions. Accordingly, the proteins may develop stronger (electrostatic) or weaker (van der Waals type) interactions with one another and with their environment.

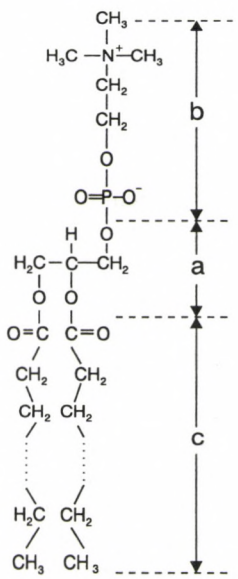


Fig. 1.38. Structural formula of dipalmitoyl phosphatidylcholine

The structure of membranes is mainly determined by the dual character (polar and apolar) of their lipid and protein content, i.e. their amphiphilic nature. In aqueous medium the molecules are ordered so that the polar groups turn towards the aqueous phase and get into electrostatic interaction with one another and with the dipolar water molecules. The hydrophobic parts are linked by van der Waals forces inside the membrane. As pointed out in connection with liquid crystals (cf. section 1.4.4), plane and bent lipid bilayers are to be found with high probability among the thermodynamically possible configurations of the lipid–water lyotropic systems. The structural basis of the membranes is the lipid bilayer, in which (as mentioned above) the polar head groups interact with water on the two sides of the membrane, while the hydrocarbon chains of the two layers are turned towards each other, resulting in an easily changing *liquid crystalline* structure of the membrane. This hydrophobic part is responsible, for instance, for the high electric (and diffusional) resistance and the high electric capacity of the membranes. The membrane proteins are intercalated into this "fluid" lipid layer to an extent depending upon their geometric and charge configurations, determined by their amino acid content and sequence. Some proteins may reach completely across the lipid bilayer. This arrangement called *fluid mosaic* model is shown in Fig. 1.39. The small spheres on both sides of the membrane represent the polar head groups of the lipids, the tails represent the hydrocarbon chains, and the shaded regions represent the *integrant* proteins penetrating into, or even through the lipid layer. Recent investigations indicate that a small proportion of the membrane proteins are not incorporated into lipid layer, and are only loosely bound to the surface of the membrane. These are the *peripheral* proteins which, together with the branching glycoproteins (hydrocarbon–protein complexes) emerging perpendicularly from the lipid layer, play an important role in the interactions between the membrane and its environment (for instance in the immune processes).

The *water layer bound* by the polar groups on the surface is also part of the membrane structure, though not much is known about its molecular configuration. Experimental results suggest that some of the surface water is strongly bound and consequently more

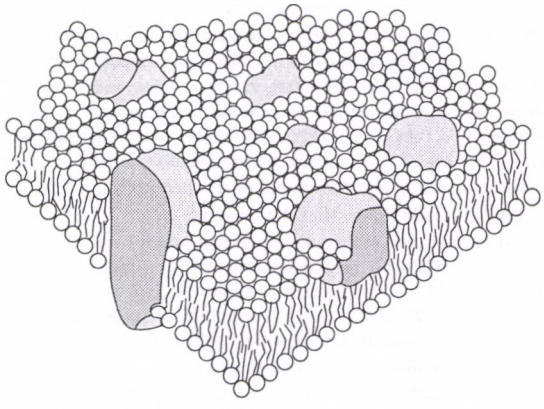


Fig. 1.39. Fluid mosaic model of the structure of biological membranes

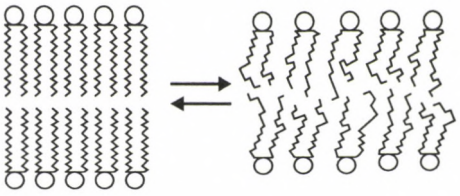


Fig. 1.40. Outline of the structural change in the hydrocarbon chains of the lipid bilayer at the phase transition temperature

ordered. A loosely-bound intermediate layer with dynamic configuration covers the strongly-bound water, followed by the intra- or extracellular solution. (This ordering of the water bound to the surface can be found in all biological systems.)

2. Some properties. Most properties of biological membranes can be attributed to the liquid-crystalline behaviour of the lipid layers. The structural transformation of the apolar phase is of fundamental importance in the function of the membrane. In straight, ordered hydrocarbon chains with *all trans* configurations, one or more breaks (gauches) appear above the *temperature of the phase transition,* resulting in a less ordered configuration (Fig. 1.40). In this process the otherwise weak bonding between the individual chains becomes even weaker, with the result that the molecules are more widely separated from one another and the number of structural defects suddenly increases.

Various types of structural defects may appear, depending upon the external effects (e.g. temperature change, pressure, electric field, drugs) modifying the membrane structure. The number and type of the structural defects are decisive in the development of the membrane functions. The defects determine mainly the permeability of the lipid layer, though they may be important as concerns the mechanical properties (elasticity) of the membrane and the lateral motion of the membrane components (e.g. proteins). Two basic types of defects will be mentioned:

(a) First *intramolecular defects,* which are produced by the rotational isomerization of the hydrocarbon chains of the lipid molecules. Rotational isomers are formed if the chain segment is rotated by $\pm 120°$ around a C–C bond of the hydrocarbon chain. In Fig. 1.41*a* two isomers are presented in which this rotation takes place around one and two C–C groups, respectively. This type of defect may be created by a temperature increase. Because of the loosening of the membrane structure, the permeability of the membrane increases in these cases.

(b) *Of the intermolecular defects,* the *domain wall* is a disordered region between the differently oriented ordered domains (Fig. 1.41*b*). The membrane permeability is larger at the domain wall than in the ordered domains.

Intermolecular defects are also formed in the transitional phase present at the temperature of the phase transition, when the membrane structure is characterized by large lateral density fluctuations. The temporarily opening holes (*hydrophobic pores;* Fig. 1.41*c*) increase the permeability of the membrane considerably. For this reason the permeability becomes larger in the transitional phase than in the original crystalline phase prior to the phase transition or in the liquid-crystalline state after the transition.

The *hydrophilic pore* is also an important intermolecular defect, which causes a topological change in the bilayer structure of the membrane (Fig. 1.41*d*). In the phenomenon of *electroporation* the pore-forming effect of the electric field (short electric pulses of high voltage) is used also in practice (e.g. in gene technology). Namely through the pores molecules of such a big size (e.g. DNA segments corresponding to a complete gene) may enter into the cell for which the membrane is usually not permeable. After the closure of the pore the nucleic acid which entered the cell may carry out the function coded by it. The fusion of several cells upon the effect of electric field, *electrofusion,* is also due to pore formation. In the latter case the whole substance of the cells participating in the fusion is united into a single formation, possibly a giant cell, visible even by naked eye.

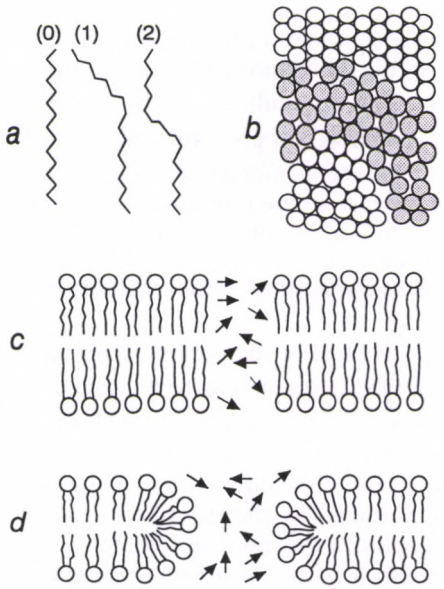


Fig. 1.41. Structural defects in lipid membranes
a: hydrocarbon chain with one (1) and two (2) rotational defects (isomers) (a perfect chain is shown on the left); *b*: lipid membrane (viewed from above) with differently oriented ordered domains and the domain wall (shaded circles); *c*: hydrophobic pore; *d*: hydrophilic pore

The structural changes occurring at the temperature of the phase transition can be followed well with physical structure analysis methods which are sensitive to changes in the molecular order. Primarily electron and X-ray diffraction methods, laser-Raman-spectroscopy, magnetic resonance spectroscopy (ESR, NMR) and the microcalorimetric methods (e.g. DSC) are suitable for determination of the transition temperature, and the extent of the structural changes occurring at this temperature (cf. sections 4.4 and 4.6). Because of the structural weakening, the permeability of the membranes increases at and above the phase transition temperature (cf. section 5.5.4), and consequently the ion and molecular transport through the membrane will also increase. It is interesting to note that the phase transition temperature of the lipid layers of membranes maintaining an intensive transport (e.g. the mitochondrium) are found close to or somewhat below the physiological temperature. At the same time, however, membranes displaying a slow metabolism (for instance the myelin sheath of nerve cells) exist below the phase transition temperature, i.e. in a state of higher order.

In the regulating function of the membrane it is essential that, in particular, close to the phase transition temperature, the structure of the lipid layer may be changed considerably by slight environmental effects (e.g. changes in the concentrations of hydrogen and other ions, or the membrane potential). This allows sensitive regulation of the transport processes. It is an experimentally observed fact that cholesterol increases the molecular order in the lipid layer above the temperature of the phase transition, and

decreases it below this temperature; consequently, the permeability of the membrane can be regulated by changing the cholesterol content of the membrane.

The dynamic structure of the lipid phase promotes the finer regulation of the function of the membrane proteins. The lipid bilayer in part ensures a mobile medium for the conformational change of the integrant proteins, and in part permits the lateral motion of the proteins in the plane of the membrane. Both processes are indispensable for the functions connected with the membrane proteins, e.g. the active transport based on the enzymatic function (cf. section 5.5.5) or the immune processes. Through the lipid–protein interactions the structural changes in the lipid layers modify the function of the proteins. The reverse process too takes place, which allows the transmission of information by the membranes.

The hydrophilic channels formed by the proteins reaching across the membrane play an important role in the regulation of ion transport through membranes, i.e. the ion permeability. These channels promote ion transport through an otherwise apolar membrane. In this case too the permeability is regulated by the structure of the proteins and the surrounding lipids. In this way external effects (e.g. an altered membrane potential) may induce a considerable ion permeability change (cf. also section 7.2). Similarly, the water layer bound at the surface plays a determining role, for by interacting with the hydrate shells of the ions and polar molecules (e.g. amino acids) it participates in the regulation of the transport of these substances.

REFERENCES

Books

Bittar, E., Membrane Structure and Function. John Wiley and Sons, New York 1980

Brown, G. H., Wolken, J. J., Liquid Crystals and Biological Structures. Academic Press, New York 1979

Cantor, Ch. R., Schimmel, P. R., Biophysical Chemistry. I. The Conformation of Biological Macromolecules. W. H. Freeman and Company, San Francisco 1980

Davies, D. B., Saenger, W., Structural Molecular Biology. Plenum Press, New York 1982

De Gennes, P. G., The Physics of Liquid Crystals. Clarendon Press, Oxford 1974

Duchesne, J., Physico-Chemical Properties of Nucleic Acids, I–II–III. Academic Press, New York 1973

Jain, M. K., Wagner, R. C., Introduction to Biological Membranes. John Wiley and Sons, New York 1980

Kittel, Ch., Introduction to Solid State Physics, 5th edition. John Wiley and Sons, New York 1982

Kreher, K., Festkörperphysik. Akademie-Verlag, Berlin 1973

Landau, L. D., Lifshitz, E. M., Quantum Mechanics. Pergamon Press, Oxford 1974

Tien, H. T., Bilayer Lipid Membranes. Marcel Dekker, New York 1974

Tinoco, I., Sauer, K., Wang J. C., Physical Chemistry (Principles and Application in Biological Sciences). Prentice-Hall International Inc., London 1978

Villars, H., Benedek, G., Physics, Vol. 2: Statistical Physics. Addison-Wesley Publishing Company, Reading, Ma 1974

Wang, S. J., Photochemistry and Photobiology of Nucleic Acids, I–II. Academic Press, New York 1967

Adams, M. D. et al., Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature, 377,* 3 (1995)

Aslanian, D., Gróf, P., Renault, F., Masson, P., Raman spectroscopic study of conjugates of butyrylcholinesterase with organophosphate. *Biochim. Biophys. Acta 1249,* 37 (1995)

Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V., Kopka, M. L., The Anatomy of A-, B-, and Z-DNA. *Science, 216*, 475 (1982)

Fekete, A., Rontó, Gy., Feigin, L. A, Tikhonychev, V. V., Módos, K., Temperature dependent structural changes of intraphage T7 DNA. *Biophys. Struct. Mech., 9,* 1 (1982)

Li, J., Carroll, J., Ellar, D. J., Crystal structure of insecticidal δ-endotoxin from *Bacillus thuringiensis* at 2.5 Å resolution. *Nature, 353,* 815 (1991)

Preisler, R. S., The B-DNA and Z-DNA transition in alkali and tetraalkylammonium salts correlated with cation effects on solvent structure. *Biochem. Biophys. Res. Commun., 148,* 609 (1987)

Rontó, Gy., Agamalyan, M. M., Drabkin, G. M., Feigin, L. A., Lvov, Yu. M., Structure of bacteriophage T7. Small angle X-ray and neutron scattering study. *Biophys. J., 43,* 309 (1983)

Sugár, I. P., The effects of external fields on the structure of lipid bilayers. *J. Physiol, 77,* 1035 (1981)

Vainshtein, B. K., Melik-Adamyan, W. R., Barynin, V. V., Vagin, A. A., Grebenko, A. I., Borisov, V. V., Bartels, K. S., Fita, I., Rossmann, M. G., Three-dimensional structure of catalase from *Penicillium vitale* at 2.0. Å resolution. *J. Mol. Biol., 188,* 49 (1986)

Wang, A. H-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., Rich, A., Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature, 282,* 680 (1979)

Watson, J. D., Crick, F. H. C., Molecular structure of nucleic acids. *Nature, 171,* 737 (1953)

Wilkins, M. H. F., Stokes, A. R., Wilson, R. H., Molecular structure of deoxypentose nucleic acids. *Nature, 171,* 738 (1953)

2. LIGHT AND X-RADIATION

We would like to point out already here that the next chapter deals with the nuclear radiations. These two chapters contain several informations which are useful for both groups of phenomena. Such are some processes of the interaction between radiations and the material, or the measurement and dosimetry of the X- and γ-radiations. Common features may be found also in the field of their therapeutic application. Nevertheless, there are several substantial differences between the two groups: the light and X-radiations are the consequences of processes taking place outside the atomic nucleus, while the nuclear radiation – as indicated by its name – is produced during nuclear processes.

2.1. The complete electromagnetic spectrum

The electromagnetic spectrum with its ranges is shown in Fig. 2.1. Only visible light (VIS) stimulates the visual receptors of the human eye. Its wavelength range is a fairly narrow one, from 400 to 800 nm. However, the concept of light includes not only visible, but also infrared (IR) and ultraviolet (UV) radiation. These three forms are collectively known as the optical interval of electromagnetic radiation. X-radiation, with an extremely short wavelength (less than 100 nm), is also called light, which expresses the fact that any electromagnetic radiation resulting from changes in state of electrons, atoms or molecules (multi-atomic systems), i.e. from processes outside atomic nucleus is called

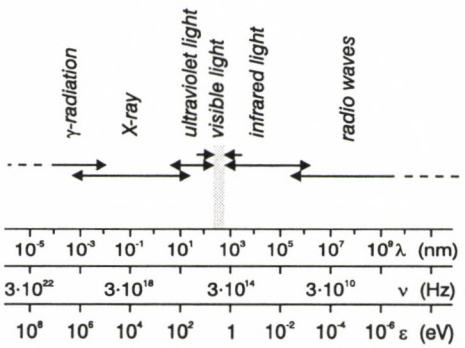


Fig. 2.1. The complete electromagnetic radiation spectrum
λ: wavelength; *ν*: frequency; *ε*: photon energy

light. On the short wavelength side of the light range follows (with a considerable overlap) γ-radiation produced by nuclear processes, and on the long wavelength side the electromagnetic waves produced by the telecommunication techniques.

Nearly all properties of the telecommunication range can be interpreted by the physics of waves, whereas almost all properties of γ-radiation are rather of a corpuscular character. In the intermediate interval of the electromagnetic spectrum neither property predominates. Numerous processes may be interpreted electromagnetically, and others by means of the corpuscular character.

In the next section only the *optical range* will be discussed. X-radiation is dealt with separately, and γ-radiation is treated together with nuclear radiations.

2.2. Interaction with atomic systems

The emission and absorption of light are produced by changes in state of atomic systems or more exactly by changes in the *electric dipole moment of these systems.* Just these radiative transitions are allowed by the selection rules discussed earlier (sections 1.2.2 and 1.3.3). *The electric dipole moment may change during electronic transitions,* but also as a result of changes in the rotational or vibrational states of atoms or molecules. Light emission is connected with a change in state of optical electrons, whereas X-radiation is a consequence of a change in state of the inner shell electrons.

1. Emission. An atomic system (atoms, molecules, multi-atomic systems) may emit light only if the system possesses excess energy in relation to its ground state, i.e. if it is in an excited state. Emission ensues when the system gets from the excited state into a state of lower energy. An atomic system may be excited in different ways: thermally, electrically or optically. An example of the first case (thermal excitation) is the coloration of the Bunsen flame by metal salts, the second case (electrical excitation) may be exemplified by discharge tubes, while the photosynthesis of green plants is a result of optical excitation.

An excited atomic system may emit photons in two ways: either *spontaneously* or by *induced* emission. In the first case the emission is produced without any external effect (e.g. external force field). The induced emission is triggered by another photon of the same frequency as the emitted one. In photon-induced emission the electron is displaced from a higher energy level to some lower state due to the effect of inducing photon. The spontaneous emission shows an accidental, statistical distribution concerning the time and direction of the emission, the phase of the wave train, etc. With other words: spontaneous emission occurs statistically in every direction and produces incoherent light. Induced emission, on the other hand, consists of highly coherent photons, which propagate in the same direction as the light-inducing photons. The spontaneous emission may be illustrated by relatively short, some mm, at most some cm wave trains, the induced emission on the other hand by those of 10^2 to 10^8 m, respectively. The usual light sources produce only spontaneous emission, and the induced emission is practically negligible. However, lasers are light sources based on induced emission (cf. section 2.6).

2. Absorption. In optical excitation a photon may interact with an atomic system in three different ways.

(a) The photon only *perturbs* the electronic state. The perturbation time is approximately the same as the oscillation period of the photon (in the visible range this is approximately 10^{-15} s). As a result of the perturbation, almost exclusively such photons are emitted, in rather irregular directions, whose frequency is the same as that of the incident photons. The probability of production of photons with different frequencies is rather low. Light scattering at the same frequency is usually called *coherent* (classical or Rayleigh) scattering, whereas photons with frequencies different from the exciting frequency constitute *Raman scattering* (section 4.4.3). Light of any frequency may produce light scattering, which is also responsible for the phenomenon of light reflection and refraction.

(b) The absorbed photon *excites* the atomic system by raising it to a higher energy level. The lifetime of the excited state is generally several orders of magnitude longer than the perturbation time: in the case of allowed optical transitions it is at most 10^{-8} s, but for metastable levels 10^{-3} s or even longer (cf. sections 1.2.3 and 2.5).

The excitation energy may be emitted as a photon, this phenomenon being called *luminescence* (cf. section 2.5). However, the excited atoms or molecules may trigger chemical transformations without any light emission; these processes are *photochemical reactions.* In most cases, however, the excitation energy is transformed into heat (phonon emission), i.e. in a strict sense *light absorption* occurs only if absorbed light *energy* is converted into heat.

For any given system, excitation is observed only at some well-defined frequency or in a definite frequency range. In the excitation processes, mainly because of scattering and heat conversion, the intensity of the exciting light decreases as it passes through the system (cf. section 2.3.1). The spectral distribution of the light attenuation is characteristic of the absorbing system.

(c) The photon behaves as a classical particle and produces a *photoelectric effect.* With high-energy electromagnetic radiation (X-radiation and γ-radiation) the Compton effect and pair production may also be considerable (cf. section 2.10.2).

2.3. Radiometry–photometry

In the following sections optical radiation will be dealt with as an energy transporting process and the fact that it may produce a light sensation as well will be considered only in section 2.3.2. Thus first the *measurement of radiation,* i.e. *radiometry,* will be discussed and only subsequently will follow the measurement of visible radiation weighted according to light sensation, i.e. the *measurement of light* or *photometry.* The concepts and relations to be discussed in radiometry can be applied not only in the optical range but all over the complete electromagnetic spectrum.

2.3.1. Radiometry

1. Basic concepts

a) Radiant flux. Let dQ denote the *radiant energy* emitted by a radiation source in time dt. *Radiant flux* is defined as

$$\Phi = \frac{dQ}{dt} \qquad [2.1]$$

i.e. as the energy emitted in unit time. Its unit is W.

b) Radiant intensity. The radiant flux emitted by a radiation source in various directions is usually different. The dependence on direction is characterized by the *radiant intensity* defined by the quantity

$$I = \frac{d\Phi}{d\omega} \qquad [2.2]$$

where $d\Phi$ denotes the radiant flux propagated in an element of solid angle $d\omega$ containing the given direction. Thus the radiant intensity gives the flux emitted in unit solid angle[1] in the given direction. Its unit is W sr^{-1}.

c) Irradiance. In practice it is often necessary to know the radiant flux density at a given place of an irradiated surface, i.e. the energy incident per unit time and surface. This is given by the quantity

$$E = \frac{d\Phi}{dA} \qquad [2.3]$$

where $d\Phi$ denotes the radiant flux incident on the surface element with area dA at the given place. It is called *irradiance.* Its unit is W m^{-2}.

d) Radiant intensity. In everyday practice the expression radiant intensity and its symbol I is often used differently from [2.2]. The quantity giving the radiant flux incident on a unit surface perpendicular to the propagation of radiation is also called *radiant intensity:*

$$I = \frac{d\Phi}{dA} \qquad [2.4]$$

where in this case dA is the area of the surface element perpendicular to the radiation. Obviously [2.4] is more closely related to [2.3] than to [2.2]. Its unit too is the same as that of irradiance: W m^{-2}.

The concept of intensity denoted by I will be used according to [2.4] also further on.

[1] "sr" means *steradian,* i.e. the symbol of the unit of solid angle. – In a plane an angle is measured as the quotient of an arc and the radius belonging to it. A *solid angle* is defined by the quotient of a spherical cap's surface (S) and the radius belonging to it, i.e. the total solid angle is given by the dimensionless quantity $\omega = S/r^2$. For the total solid angle (the total space) S is equal to the surface $4r^2\pi$ of a sphere with radius r, therefore the total space is described by the quantity 4π sr.

The energy radiated by the Sun is 4×10^{26} J s^{-1}, and the intensity at the atmospheric boundary (the Sun–Earth distance is ca. 1.5×10^{11} m) is approximately 1.35×10^{3} W m^{-2} (this is the solar constant). At most only half of this energy reaches the Earth, the rest being absorbed and scattered by the atmosphere. The human eye is most sensitive to yellowish-green light, and an intensity of approximately 2×10^{-12} W m^{-2} can already be perceived.

2. The law of radiation attenuation. On passing through some medium, radiation loses intensity due to scattering and transformation into some other energy, mainly heat (absorption). As concerns the degree of attenuation, we shall consider only the case of a parallel beam striking some medium normally.

Let the intensity of the radiation striking a layer of the medium of thickness x be denoted by I_0, and the intensity of radiation having passed through this layer by I (Fig. 2.2). The intensity decreases exponentially with the increase of x:

$$I = I_0 e^{-\frac{x}{\delta}} \qquad [2.5]$$

where δ is the layer thickness which decreases the intensity by a factor e ($e = 2.71...$, the base of natural logarithms).

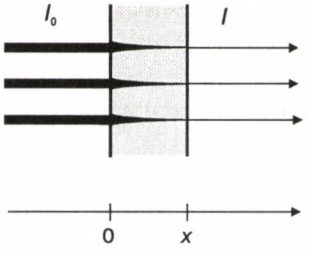


Fig. 2.2. Diagram relating to the Lambert–Bouguer law

The intensity decrease is frequently characterized by the half-value thickness, i.e. the layer thickness D which decreases the intensity by half. Of course, δ is always larger than D:

$$D = 0.693\ \delta \qquad [2.6]$$

The meaning of δ is obvious, since with $x = \delta$, $I/I_0 = e^{-1}$. If $x = D$, from the definition of D we have $I = I_0/2$, and consequently $e^{D/\delta} = 2$. If the logarithms of both sides are taken, [2.6] is obtained.

The quantity $\mu = 1/\delta$ is the attenuation or *extinction coefficient.* μ is the reciprocal of that layer thickness which decreases the intensity to $1/e$ of its original value: its unit is consequently m^{-1}, cm^{-1}, etc. The substitution of μ into [2.5] leads to

$$I = I_0 e^{-\mu x} \qquad [2.7]$$

while from [2.6] we have

$$\mu = \frac{0.693}{D} \qquad [2.8]$$

From simple mathematical considerations, [2.7] is equivalent to

$$-dI = \mu I dx \qquad [2.9]$$

This means that for a sufficiently small layer thickness (dx) the intensity change (dI) is proportional to the layer thickness and the intensity I measured at the point of incidence of the beam on the layer. The negative sign refers to the fact that dI is negative for an intensity decrease.

The law expressed by [2.5] or the equivalent relations [2.7] and [2.9] are known as the *Lambert–Bouguer law.*

The common logarithm (to the base 10) of the quotient I_0/I is the decadic extinction or simply the extinction (often called optical density; *OD*), while the natural logarithm (to the base e) of I_0/I is called the natural extinction. For solutions, if μ is proportional to the molar concentration c, the following relation holds: $\mu = \varepsilon c$ *(Beer law).* The proportionality factor is the *molar extinction coefficient.*

D and consequently δ and μ depend upon the nature of the scattering and/or absorbing medium and also upon the wavelength of radiation.

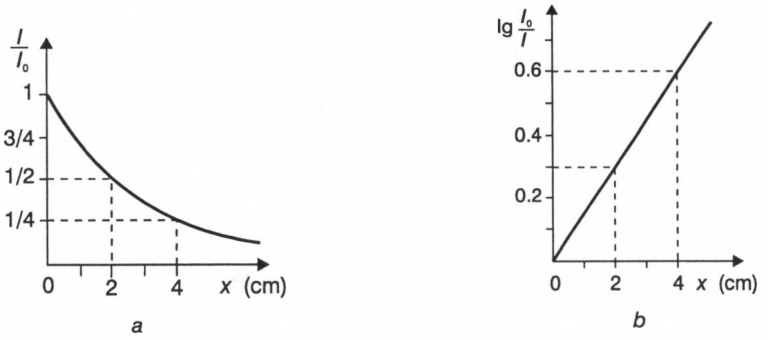


Fig. 2.3. Variation in the relative intensity (I/I_0) (*a*) and the extinction (lg (I_0/I)) (*b*) with layer thickness (x) for $D = 2$ cm

Figure 2.3 depicts the effect of attenuation for $D = 2$ cm. The decrease in intensity is frequently due to absorption in the narrower sense, if the light scattering is negligible. This explains why the term absorption coefficient is used instead of the extinction coefficient.

3. Further concepts and quantities

(a) The *reflectivity* (reflectance) of a body is the fraction of the total incident radiation reflected from some site on the surface. The reflectivity of high-quality mirrors (silver surfaces) in the visible range is 0.9–0.96, i.e. 90–96%.

(b) The *transmittivity* (transmittance) is defined by the ratio of the transmitted and the incident energy.

(c) The *absorbance* is the ratio of the absorbed and incident radiation. This quantity is always a fraction. The absorbance of soot for visible light is close to 1, i.e. nearly 100%.

The quantities defined in points (a)–(c) depend upon the angle of incidence. If this is not given, incidence normal to the surface must be assumed. It follows from the definitions that in a given case the sum of the quantities defined in (a)–(c) is 1. If the penetrating energy is considered instead of the incident energy, one obtains the pure transmittance for case (b) and the pure absorbance for case (c). The value of the penetrating energy is obtained by subtracting the reflected energy from the incident energy.

2.3.2. Photometry

1. Spectral luminous efficiency. The sensitivity of the human eye varies with the frequency (or wavelength) of light. Figure 2.4 shows the *average spectral sensitivity*, i.e. the *spectral luminous efficiency* $[V(\lambda)]$ of the human eye. The abscissa represents the wavelength and the ordinate the sensitivity (luminous efficiency). The maximum value of the sensitivity at 555 nm is taken as 1. If a k times higher intensity is required to produce the same sensation of light at some other wavelength, then the sensitivity of the eye for this wavelength is only $1/k$.

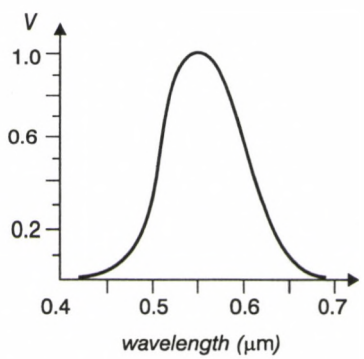


Fig. 2.4. Sensitivity curve of the eye in daylight

2. Photometric quantities. Only the quantities corresponding to the previously discussed radiometric quantities will be discussed here.

a) Luminous flux is a quantity derived from radiant flux by weighting it with the luminous efficiency. It is denoted by Φ_v. The subscript v here and further on refers to visual, i.e. to visibility.

The relation of Φ_v, to Φ can be obtained by the following consideration. Let $d\Phi$ denote the radiant flux between the wavelength λ and $\lambda + d\lambda$ and V the luminous efficiency at λ. The luminous flux in the above wavelength range is given *by* $V d\Phi$. The total luminous flux in the visible range can be obtained by integration:

$$\Phi_v \sim \int V d\Phi, \quad \Phi_v = K \int V d\Phi \qquad [2.10a]$$

The proportionality factor K depends on the units chosen. The unit of radiant flux is watt, that of luminous flux is lumen (lm) which will be defined later on. Since according to the definition and the measurements at the maximum of the spectral luminous efficiency ($V = 1$) 1 W corresponds to 683 lm, therefore

$$K = 683 \text{ lm W}^{-1} \qquad [2.10b]$$

b) Luminous intensity is related to radiant intensity. In the case of a point-like source in a given direction it is expressed by the quantity

$$I_\upsilon = \frac{d\Phi_\upsilon}{d\omega} \qquad [2.11]$$

where $d\Phi_\upsilon$ denotes the luminous flux propagated in an element of solid angle $d\omega$ containing the given direction. Thus the luminous intensity is the luminous flux propagating in unit solid angle.

If the emission of the light source is uniform in every direction, i.e. I_υ is the same in every direction, the total luminous flux Φ_0 of the light source is obtained according to [2.11] by multiplying I_υ by the total solid angle 4π:

$$\Phi_0 = 4\pi I_\upsilon \qquad [2.12]$$

c) Illuminance corresponds to irradiance. If a surface element of area dA receives a luminous flux of $d\Phi_\upsilon$ (Fig. 2.5) illuminance is given by the quantity

$$E_\upsilon = \frac{d\Phi_\upsilon}{dA} \qquad [2.13]$$

A valuable relation is obtained by comparing [2.11] and [2.13]: $E_\upsilon = d\omega I_\upsilon / dA$. If the surface is perpendicular to the direction of propagation of the beam, from the definition of the solid angle ($d\omega = dA/r^2$) it follows that

$$E_\upsilon = \frac{I_\upsilon}{r^2} \qquad [2.14]$$

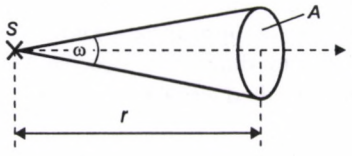


Fig. 2.5. Diagram relating to the definition of luminous intensity and illuminance

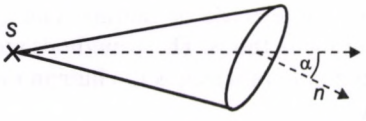


Fig. 2.6. Diagram relating to [2.15]
n: the normal to the surface

Consequently the illuminance produced by a point-like source is proportional to the luminous intensity and inversely proportional to the square of the distance. If the sur-

face is oblique with respect to the incident beam (Fig. 2.6), the following relation holds instead of [2.14]

$$E_v = \frac{I_v}{r^2} \cos \alpha \qquad [2.15]$$

which means that the more obliquely a beam strikes a surface the lower the illuminance.

3. Units. In photometry the basic unit is the unit of luminous intensity and the units of other quantities are derived from this (cf. Table 8.1).

The unit of luminous intensity is the candela (denoted by cd). The candela is the luminous intensity, in a given direction of a source that emits monochromatic radiation of frequency 540×10^{12} Hz (i.e. of wavelength 555 nm in vacuum) and has a radiant intensity in that direction of (1/683) W sr^{-1}. This is approximately equivalent to the luminous intensity produced by a simple candle in the horizontal direction. For the production of a given luminous intensity, standard lamps made and calibrated according to internationally accepted standards can be obtained from the various Bureaux of Standards. These lamps play the same role in photometry as the standard metre bar in longimetry or the standard resistances in electrical measurements.

The unit of luminous flux is the lumen (denoted by lm) which is the luminous flux emitted by a light source of 1 cd in unit solid angle.

The unit of illuminance is the lux (denoted by lx). One lm luminous flux produces 1 lux illuminance on a surface of 1 m^2.

Reading requires at least 200 lx, while fine work needs about 1000 lx. At noon in summer the sunshine produces approximately 10^5 lx, compared with approximately a few tenths lx by the full moon.

As concerns health regulations, mainly the measurement of the illuminance is important; this is most easily carried out with photoelements. The sensitivity curve of a selenium or silicon photoelement provided with an appropriate filter is approximately the same as that of the human eye. The instrument connected to the element in most cases permits reading of the intensity of illumination directly in lux units.

2.3.3. Measuring methods

Radiation (light) energy is measured by transforming it into some easily measurable energy, e.g. electric or thermal energy, then applying the appropriate known physical measuring methods. These methods are called objective methods, as opposed to those the "instrument" of which is the human eye. The latter are called visual or subjective photometry.

In the following some objective methods will be dealt with shortly. Radiant (light) intensity is the most frequently measured quantity; this is the case e.g. at the measurement of the extinction coefficient or the spectral distribution of the intensity in the emission spectrum of a radiation source.

Intensity measurement based on the *photoelectric effect* is a readily available, very sensitive and subsequently frequently used method, since the photocurrent is directly

proportional to the intensity incident on the photoelectric cell. In the visible and ultraviolet range photocells with Cs, Na, K, etc. cathodes are used together with Si, GaAs, etc. photodiodes while those in the infrared range have lead sulphide, lead selenide, etc. photoconductors. One drawback of these methods is the dependence of the photoelectric effect on the radiation (light) frequency. As a consequence, the photoelectric methods are used only to measure the intensities of radiation (light) of identical spectral distributions. If different frequencies are compared, the frequency-dependence of the photocell or photoelement, etc. must be taken into consideration.

If a *thermocouple* or thermopile is used, the incident radiation (light) strikes the soldering seam, which is coated with some appropriate absorbing material (e.g. for visible light soot). The seam is heated up by the absorbed radiation, and the thermocurrent produced is proportional to the intensity of the incident radiation. This method is readily applicable throughout the whole spectral range. Its great advantage is that it is independent of the frequency. – Detectors based on the *pyroelectric effect* are similarly advantageous; they are more sensitive than the above-mentioned ones but respond only to the change of radiation intensity.

For the measurement of radiation the *photochemical effect* is often used. The chemical transformation (e.g. in case of film darkening) of the light-sensitive layer is nearly proportional to the incident energy. Accordingly, not the intensities, but the products of the intensities and exposure times are compared in this case. With equal exposure times the darkening is proportional to the incident intensity. Because of the frequency dependence of the darkening, the considerations discussed in connection with the photoelectric effect hold here too. A further disadvantage is the different photosensitivities of different plates, and even of the various sites on a given plate.

2.4. Thermal radiation

Radiation produced at the expense of thermal energy (phonon space) of a body, and depending only on the temperature of this body, is thermal radiation. Every body emits thermal radiation at any temperature. As long as the temperature of the body is lower than approximately 750 K, it emits practically only infrared light, and consequently its radiation cannot be observed by the human eye. The visible components of thermal radiation, with wavelengths shorter than those in the infrared range, appear with intensities perceptible to the eye only at temperatures higher than approximately 750 K. As the temperature rises above this value, the body glows first dark red, then bright red, yellowish-red, and finally at approximately 1800 K white. Ultraviolet radiation, with wavelengths shorter than those of visible light, begins only at a temperature higher than about 2000 K with such intensity that its biological effects should be taken into account.

More exact data are given in Fig. 2.7, which illustrates the wavelength (λ) dependence of the emittance for various temperatures (T) of an absolute black body.[2] An incandescent lamp, a heating radiator, the human body and also the Sun radiate almost as black bodies.

In the case of the absolute black body the emission in the wavelength range between λ and $\lambda + d\lambda$ is characterized by the quantity $E(\lambda, T)\, d\lambda$, where $E(\lambda, T)$ denotes the radiant flux emerging perpendicularly to the emitting surface in unit wavelength range, surface area and solid angle. $E(\lambda, T)$ is the emittance of the absolute black body. With the aid of $E(\lambda, T)$ the emittance of any other body, $e(\lambda, T)$ can be determined knowing its absorbance $a(\lambda, T)$, since according to *Kirchhoff's law* the following equation holds:

$$e(\lambda, T) = E(\lambda, T)\, a(\lambda, T), \quad \text{i.e.} \quad E(\lambda, T) = \frac{e(\lambda, T)}{a(\lambda, T)} \qquad [2.16]$$

[2.16] expresses the experimental fact that at a given T and λ the quotient of $e(\lambda, T)$ and $a(\lambda, T)$ is the same for every body. If a body emits some radiation more strongly, it also absorbs the same radiation more strongly, and conversely. In general, $a(\lambda, T) < 1$, and therefore $e(\lambda, T) < E(\lambda, T)$; this means that the emissivity of any body at a given wavelength is smaller than that of an absolute black body at the same temperature.

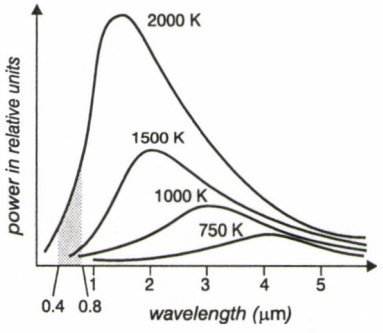


Fig. 2.7. Spectral energy distribution in black body radiation
The visible region is shaded

Figure 2.7 demonstrates that with increasing temperature the maxima in the curves are shifted towards shorter wavelengths. The energy distribution in the spectrum of the Sun is approximately the same as that of a black body at 6000 K. At this temperature the maximum in the distribution curve (not shown in the Figure) is already in the visible range. Thus, the human eye is sensitive in the range of the maximum of solar radiation. Besides ultraviolet radiation, solar emission contains a considerable amount of X-radiation. However, X-radiation and the very short-wavelength ultraviolet radiation are

[2] The term *absolute black body* refers to a body whose absorbance is 1, and transmittance is zero at all temperatures and all wavelengths. The absolute black body is an ideal limiting case, but it can be attained to a good approximation. Let us take, for example, a closed metal box with a sooted inner surface and a small hole in one of its faces. The light incident through this hole becomes continually weakened by multiple reflection, and in the case of a properly made box only a very small amount of the incident radiation will emerge through the hole. Thus, the hole behaves as a black spot on the surface of the hollow body. If this body is heated to different temperatures, the hole radiates virtually as an ideal black body at the respective temperatures.

absorbed by the atmosphere (especially by the ozone in the upper atmosphere) so that radiation with a wavelength shorter than 290 nm does not reach the Earth's surface. Of course, infrared radiation of very long wavelength is also found in the spectrum of the Sun; this is detectable as a group of ultrashort radiowaves.

Figure 2.7 also shows that with increasing temperature the emitted power increases at every wavelength. The *emitted total power at a given temperature* is demonstrated by the area under the curve. According to the *Stefan–Boltzmann law* the emitted total power per unit area of an absolute black body is proportional to the fourth power of the absolute temperature

$$E = \sigma T^4 \qquad [2.17a]$$

where

$$\sigma = 5.7 \times 10^{-8}\ \mathrm{W\ m^{-2}\ K^{-4}} \qquad [2.17b]$$

[2.17] allows the estimation of the heat loss due to the radiation of a body at temperature T_1. However, it should be taken into consideration that the body also absorbs heat radiated from the surrounding bodies, e.g. the walls, at a temperature T_2 ($T_1 > T_2$). Consequently, the heat loss of a body is calculated from

$$E = E_1 - E_2 = \sigma (T_1^4 - T_2^4) \qquad [2.18]$$

At rest, the human organism releases somewhat more than 50% of its heat in the form of radiation from the surface of its body into its environment at room temperature. From the aspect of radiation, at 303 K the human body behaves as a body with an absorbance of 95%. According to [2.18] the emitted power of a living organism in an environmental temperature of 293 K is approximately 60 W m^{-2}. The emitted energy falls in the infrared range of the spectrum, with a maximum at 9.5 μm wavelength.

[2.18] permits the estimation of T_1 (if T_2 is known) by measuring E, or determination of the difference $T_1 - T_2$. As an example, this possibility is used for diagnostic purposes when the temperature of the skin surface is measured at different points by means of thermal radiation (infradiagnostics; cf. section 6.7.2).

2.5. Luminescence

1. Production of luminescence. Light emission which cannot be attributed to the phonon space, but is a result of some other excitation, is called *luminescence.* Luminescent light may be excited by electromagnetic radiation, corpuscular radiation, the effect of an electric field, or chemical processes. In the following we shall mainly deal with the luminescence produced by light, though the conclusions can be extended to other types of excitation.

By the above definition, light scattering produced by light might be regarded as luminescence, but it is a very fast process, occurring in less than 10^{-10} s. However, light emission is regarded as luminescence only if it follows

excitation on the average after longer than 10^{-10} s. Thus, the definition given should be complemented by excluding light scattering and other fast light-emitting processes, such as Cherenkow radiation from the notion of luminescence.

In luminescence the simultaneously excited atoms and molecules *(the luminescence centres)* do not emit simultaneously. When the excitation ends, the excited centres do not stop emitting, but decay during some shorter or longer time. These processes proceed in time similar to the radioactive decay or to the discharging of the RC circuit (cf. sections 3.1 and 6.2.1) that is, the luminescence decays exponentially as well. For the characterization of time dependence a quantity, very similar to the mean lifetime (radioactive decay) or to the time constant (discharge of the RC circuit) can be used. This characteristic is the *lifetime of the excited sate* (or *luminescence lifetime)* which is equal to the time required for the number of luminescence centres to decay to the e-th part of its initial value.

The intensity of the luminescent light (in a given wavelength range) is always higher than the intensity of the thermal radiation of the emitting body at the given (usually room) temperature; luminescent light is therefore called *cold light.*

2. Fluorescence. Phosphorescence. Luminescence may take place both in case of isolated molecules or atomic groups (e.g. in low-pressure gases) and in case of those which are *in interaction with their environment* (e.g. those in macromolecules or components of solid bodies). The "origin" of luminescence – as already mentioned – is always a transition between the (outer) electron orbitals of the molecule or atomic group, in the course of which the electron gets from a higher energy state into a lower energy state and the energy difference appears in the form of optical photons (in case of transitions between inner orbitals in that of X-ray photons). The higher energy state is always the result of the uptake of external energy, i.e. *excitation.* According to the complexity of the processes two types of luminescence can be distinguished: one is *fluorescence* and the other is *phosphorescence.* There is much uncertainty in the definitions one can find in scientific literature, in our book, however, we give preference to the terms generally used for organic substances.

We speak about fluorescence when at light emission an electron returns to the lower energy state from the *same* higher energy state into which it got by excitation. – On the other hand, we speak about phosphorescence when the excited or separated electron is transferred to *another* so-called metastable state (cf. section 1.2.2), from which emission of light is starting out.

3. Luminescence of organic molecules. In the following the *luminescence of organic molecules* will be dealt with in more detail because of its biological importance. In the case of organic molecules in the ground state there are mostly two electrons in the outermost electron shell which is relevant for the excitation, thus the resulting spin quantum number is zero (cf. section 1.2.2). This is called *singlet state* (S_0). Transition to a higher level can be reached if during the transition the spin quantum number does not change, i.e. the resulting spin quantum number of the excited electron and of that, staying on the original orbital remains zero. This is the *singlet excited state:* S_1. However, the lifetime (about 10^{-9} s) of the excited state S_1 makes it possible that, with some probability, as a result of

the interaction with the environment, the spin state of the electron may change to its opposite. It was revealed that in this case the energy of the excited level also changes, it decreases. In this excited state the resulting spin quantum number is 1, this is called *triplet state* (the level splits namely into three levels in magnetic field). Its sign is T_1. The return from the triplet state in the ground state is possible only if the electron regains its original spin state, for which a further interaction is necessary. Therefore the lifetime of the T_1 state is longer than that of the S_1 state, it may have values of μs to s, depending on the system. These levels of long life-time from which the return to the ground state is forbidden (has a small probability) for some reason, are called *metastable levels.*

In accordance with the definition mentioned in the previous point, luminescence is called *fluorescence*, if the light emission takes place during the $S_1 \rightarrow S_0$ transition, and *phosphorescence* during the $T_1 \rightarrow S_0$ transition. One and the same system may possess both fluorescence and phosphorescence, the occurrence of which may be characterized with a certain probability.

The above discussed changes in the states are demonstated in Fig. 2.8. The thick horizontal lines designate electron levels, the thin lines represent vibrational levels. The upward pointing arrow shows the energy uptake during excitation. The downward pointing thin straight arrows refer to energy changes taking place by radiationless heat release, the wavy lines to energy changes accompanied by light emission.

By the way, the radiationless and radiating transitions "compete" with each other, therefore the probability of energy release by light emission is low even in the case of fluorescence. If the excited molecule is not isolated, the emission is always preceded by a radiationless transition. Thus it may be conceived from the figure that

$$(h\nu)_{exc} > (h\nu)_{fluo} > (h\nu)_{phosph}$$

i.e., phosphorescence may be differentiated from fluorescence both by its long lifetime and its spectrum. Moreover, phosphorescence has a substantially lower, temperature-dependent intensity.

A good example for the behaviour corresponding to Fig. 2.8 is given by the tryptophan-contatining proteins. In their case, depending on the structure of the macromolecule, upon UV excitation of about 290 nm, around 340 nm fluorescence, while under special circumstances around 450 nm phosphorescence may be observed. In the first case the lifetime of the excited state is 2 to 5 ns, while in the second case it may be about 100 million times more, that is 0.5 to 1 s. The influence of the temperature (in the absence of other interacting molecules) may be also explained on the basis of the figure. The electron in the metastable state may get again to the S_1 level by means of thermal energy, and from here it returns to the ground level. Thus with increasing temperature the number of the $T_1 \rightarrow S_0$ transitions decreases , while that of $S_1 \rightarrow S_0$ transitions increases. The cessation time of the latter is determined by the lifetime of T_1 level, nevertheless, the process may be considered fluorescence on the basis of the emitted energy. In the literature this is sometimes called *delayed fluorescence*, which refers also to the long lifetime.

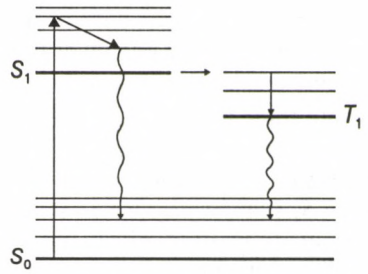


Fig. 2.8. Energy scheme of luminescence transitions

4. Types of luminescence. As mentioned previously, the introductory act of luminescence is the excitation or ionization of atoms or molecules. Various types of luminescence exist, depending upon the process of producing the required exciting and ionizing energy. *Photoluminescence* is luminescence produced by light, while *radioluminescence* is due to ionizing radiation. The luminescence produced by fast electrons is *cathodoluminescence.* The energy released by certain chemical processes may also result in luminescence; this is *chemoluminescence,* which produces the luminescence of some animals or the light of phosphor due to slow oxidation in the air. The name phosphorescence is derived from this process. The *luminescence immunoassay (LIA)* methods, which are taking the place of the radioimmunoassay (RIA) more and more, are also based on chemoluminescence. The fracture or friction of certain materials may also result in luminescence *(triboluminescence).* For instance, when a lump of sugar is broken into two parts, a weak gleam may be observed in the dark. Luminescence may similarly result from an electric field *(electroluminescence).*

5. Practical application. The majority of luminescent substances applied in practice are crystals. Crystalline phosphors are used in X-ray and television screens, cathode-ray tubes and fluorescent lamps, and also for the detection of radioactive radiation. Luminescent crystals are never pure, but are always doped with foreign atoms, called *activators* or *photocentres.* Pure crystals exhibit practically no luminescence. In pure crystals the excitation energy is transformed into phonons. The colour of the emitted light depends both on the basic material and on the incorporated activator. By a suitable choice of the activator, practically any colour can be selected even at the same basic material and luminescent phosphors emitting any desired wavelength can be made. For instance, the colours of fluorescent lamps can be designed to fit the requirements of industrial health regulations.

6. Application in the scientific research. The living tissues are only slightly damaged by light absorption and luminescence, therefore in the past decade there has been an increased attempt at their application in diagnostics. For example, *optical tomography* seems to be such a promising method. – A new therapeutic possibility is involved in the phenomenon that the excitations caused by light absorption lead to highly reactive states. Thus with selective excitation special chemical reactions may be induced which compete with

the luminescence. Great expectations are attached for example to the therapeutic results of the photochemical degradation processes following the excitation (T_1 state; cf. point 3 of this section) of porphyrin derivatives concentrating in tumours (photodynamic therapy).

The study of luminescence also plays a significant role in structural research, especially in the case of biological macromolecules (cf. section 4.4). As an example aromatic amino acids, such as tryptophan, tyrosine and phenylalanine should be mentioned, which are natural luminophores present in proteins. In proteins their interaction with the environment, its rigidity or flexibility may influence to a great extent the characteristics (e.g. lifetime, intensity, polarization) of the emitted luminescence light. From these data it is possible to get information on the environment or its changes. In the more rigid beta-form (cf. section 1.5.3) for example the phosphorescence lifetime of tryptophan can reach even 1 s. The sensitivity of the characteristics of the luminescence light to the molecular environment is utilized in the widely used methods of *luminescence labelling,* e.g. for tracing the changes in calcium content, pH, membrane fluidity and membrane potentials (cf. also section 7.2.1).

2.6. Light sources

1. Light sources based on thermal radiation. In some light sources the photon emission of substances at high temperature is produced at the expense of thermal energy (cf. section 2.4). The most important natural light source, the Sun, and a large number of artificial light sources belong in this category. Common incandescent lamps with tungsten filaments glow at about 2700 K: their spectrum is approximately the same as the emission spectrum of black body radiation at this temperature (cf. section 2.4). The glass bulb, however, transmits only the range of the spectrum between 350 nm and 2.8 μm. In special cases, if just the emission of infrared light is desired, tungsten spiral lamps, called infralamps, glowing at a lower temperature of approximately 1300 K, are used. The *sollux lamp* is a high-power tungsten spiral lamp glowing at roughly 3300 K. These lamps are provided with filters to eliminate the long-wavelength infrared radiation from the spectrum. The near infrared radiation of these lamps is used in therapy. The *evolite* or *bioptron* lamps are lamps of various power supplemented with polarizing filters.

2. Metal vapour lamps belong to luminescent light sources. In these lamps, metal (e.g. mercury, sodium) vapour is contained in a glass or silica tube. Electrodes protrude into the sealed tube, and the vapour is excited by an electric discharge produced by the voltage across the electrodes. These types of lamps yield line spectra. The 589 nm spectral line of the *sodium lamp* (the sodium D line) is used in laboratory practice as monochromatic light. Of the metal vapour lamps, *mercury lamps* are of special medical interest. The tube walls of these lamps are made of silica, which transmits ultraviolet light (down to 200 nm). For this reason they are frequently referred to as quartz lamps. Two types are distinguished: low (1–100 Pa) and high (0.01–10 MPa) pressure mercury vapour lamps. *Germicidal* lamps are of the first type. Approximately 75% of the emitted energy of these lamps is observed as the 254 nm spectral line. High-pressure lamps give a spectrum of many broad lines in the ultraviolet range. (When high-pressure lamps are switched on,

the intensity of the emitted light increases for some time and becomes stable only after a few minutes.)

Because of their economicalness, *fluorescent* lamps (or *F-tubes*) are gaining ground in illumination technics. In most cases, these tubes are low-pressure mercury vapour lamps, whose inner walls are coated with some luminescent material. 254 nm light excites the luminescent substance, which emits visible radiation (optical excitation). The tube wall is made of glass which transmits only the visible light. By variation of the luminescent material, the composition of the emitted radiation, i.e. the colour of the light, can be changed. *Erythemal lamps* as well as *solarium lamps* operate on a similar principle. The bulk of light emitted by erythemal lamps lies in the range of 280–320 nm and that by solarium lamps at 320–340 nm. The F tube of the blue light therapy used in the treatment of the icterus of premature babies is one of them; its spectrum is especially rich at 400–500 nm. The lamp walls are made of a special glass which, while absorbing light at 254 nm, transmits UV radiation in the required range. More recently *xenon lamps* made of silica glass have been applied in phototherapy. These lamps supply a spectrum very close to that of the Sun, and are thus suitable for producing nearly the same effects.

3. Lasers. In the following special light sources will be dealt with, the light of which has some particular features. Their function is based on the *induced (stimulated) emission* (cf. section 2.2, point 1) which is also responsible for the characteristics of the laser light.

The essence of induced emission is the following (Fig. 2.9): Under the effect of an "appropriate" photon (electromagnetic field) an electron being in the excited state (on the E_2 level) gets to a lower energy state (E_1 level) immediately, before it would do so spontaneously, while a photon with an energy corresponding to the difference between the energy levels ($h\nu$) is produced. The photon is "appropriate" for eliciting the phenomenon if it has the same energy (frequency) as the arising photon. It has to be emphasized that the stimulating photon is not used up: one photon arrives, two proceed further, in fact in the same – original – direction, having the same energy and vibrational phase. Due to the induced emission more and more identical photons may be produced and proceed in a parallel beam, if there are a sufficient number of electrons in the excited state along the path of the photons.

Thus the peculiarity may occur that the intensity of a light beam increases rather than decreases while passing through a material. (The name laser comes just from this: *light amplification by stimulated emission of radiation*.) The condition of light amplification is usually that more electrons be on the higher than on the lower energy level. This distri-

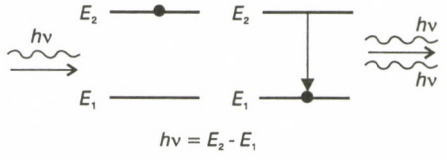


Fig. 2.9. Induced emission: one photon arrives, two leaves; their energy, direction and vibration phase is the same

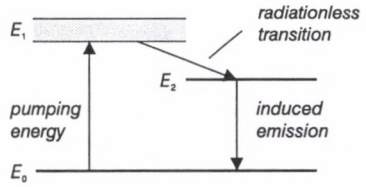


Fig. 2.10. Three-level laser system: E_0 denotes the ground level, E_1 and E_2 are the excited levels, E_2 is metastable level

bution of the electrons in the two energy states differing from the thermic equilibrium (from the Boltzmann distribution, cf. Appendix A1), the so-called *population inversion*, may be attained only if the *excited state* is a *metastable state*, the lifetime of which is long enough ($\tau > 10^{-8}$ s).

Therefore the material of the laser should contain such atoms or atomic groups, ions, which have metastable states among their energy levels. Let us consider as an example *ruby*, which – as an artificial ruby crystal a few cm long and with a cross-section of 1–2 cm^2 – was the material of the first laser. Ruby is an Al_2O_3 crystal containing Cr^{3+} in a concentration of some ‰ as dope. The basic material – the Al_2O_3 crystal – is an insulator having a broad forbidden band, thus in its pure state it is transparent for the visible light. Both the colour of ruby and the function of the ruby laser are related to the dope Cr^{3+}: the basic material has only a carrier function, the substantial processes take place at the energy levels of the chromium ions; by the way, the latter reside in the forbidden band of the basic crystal as shown simplified in Fig. 2.10. The excitable electrons are on the ground level E_0, from where they move to state E_1 as a result of illumination by flashlight (i.e. optical excitation). A part of the electrons get from level E_1 to the metastable level E_2 with a radiationless transition. The lifetime of the metastable level E_2 is 10^{-3} s which is long enough for the realization of population inversion and short enough for the spontaneous production of emitted photons following the excitation. Each spontaneously produced photon elicits induced emission, and the process continues in such a way that the number of the photons proceeding together in the same direction and in the same vibrational phase increases exponentially with the length of distance covered.

The laser material is enclosed by two mirrors normal to the axis of the laser; one of them is completely reflecting, the other one is semitransparent (Fig. 2.11). This way that part of the light – developed mainly by induced emission – which proceeds parallel to the axis of the laser (i.e. normal to the mirrors), covers the distance several times within a very short time, while its intensity increases up to a saturation value. The distance of the mirrors (l) is the multiple (m) of the half-wavelength of the laser light ($\lambda/2$); thus on this wavelength the mirrors act as *optical resonators* between which a *standing wave* is formed. This resonator, in which the light covers the distance several times, makes very strict conditions for the wavelength, therefore the light leaving through the semitransparent mirror is much more monochromatic – i.e. its photons are much more identical – than the light of the spontaneously emitting conventional monochromatic light sources (e.g. a Na-lamp).

Let us compare the light of the conventional (spontaneously emitting) light source and that of the laser (Fig. 2.12): from the conventional light source photons of various energy

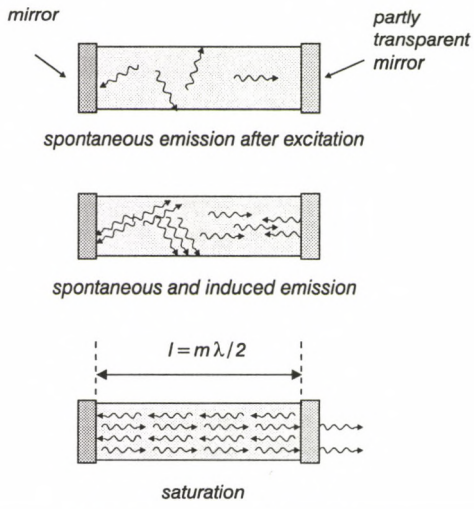


Fig. 2.11. Production of laser light

leave independently of each other in every direction; on the other hand, the photons of the light leaving the laser are identical, proceed in the same direction and join each other in the same vibrational phase. The laser light may be conceived as an almost infinitely long sine wave train, while in the case of conventional light sources the length of a wave train (the so-called coherence length) is not more than a few cm.

According to the above, laser light may be considered special for three reasons:

a) *The laser light is monochromatic.* The induced emission itself is favourable for the formation of identical photons which corresponds to the 1 GHz bandwidth of the con-

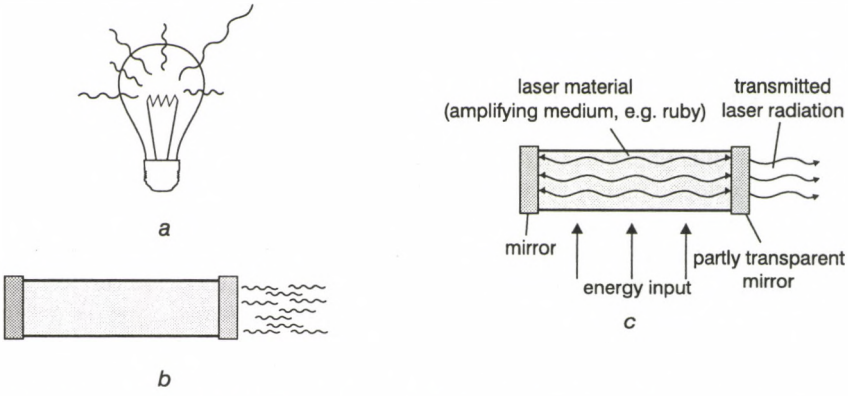


Fig. 2.12. a) Incoherent emission of a traditional light source: the energy, direction and phase of photons have stochastic distribution; *b)* Coherent emission of laser: photons have the same energy, direction and phase; *c)* Outline of a laser setup. A standing wave is formed in the laser material between the mirrors, the transmitted laser light represents travelling wave

ventional monochromatic spectrum line (i.e. to a relative frequency variation of about 10^{-6}). The optical resonator allows a much narrower frequency range, the monochromaticity of the laser light is better by a few orders of magnitude than that of the conventional light. Such a high degree of monochromaticity is of no importance from a medical view.

b) *The laser light is highly coherent* (its coherence length is large). This characteristic is not important either for its medical application; holography, on the other hand, could be developed only after the appearance of lasers (cf. section A5 of the Appendix).

c) *Laser light arises as a parallel beam.* Of course, this is not quite true, it has a slight divergence, but the so-called angle of divergence is only about a few millirad (in case of a very short distance between the mirrors it is more). The diameter of the laser-light beam is between 1–2 mm and 1–2 cm, depending on the construction of the laser, and since the beam is almost completely parallel, the energy of the laser light stays in a concentrated beam.

The power density (intensity) of the laser light is considerable already originally: for example, a ruby laser may radiate an energy of 10 J in a pulse of 1 ms (at a wavelength of 694.3 nm), and since the cross-section of the beam is about 1 cm^2, the intensity of the laser light is about 10^8 Wm^{-2} during the duration of the pulse. If this light is focused on an area having a diameter of 10 μm, a power density of 10^{14} Wm^{-2} may be reached, which is approximately 10^{12} times larger than the power density of the Sun on the surface of the Earth. With such a power density *every material may be "cut" and even evaporated.* For the same reason laser light may serve as a scalpel in the hands of the physician.

In medicine many types of lasers are used. The great variety manifests itself partly in different operating wavelengths, partly in different powers. The range of the applied powers is between a few mW (so-called soft lasers) and about 100 W (surgical lasers).

The wavelength of the laser light is determined by the material of the laser. Many kinds of materials are suitable for the generation of laser light, *lasers may be constructed practically at every wavelength*. The laser material enclosed by the resonator mirrors (Fig. 2.12) must be excited by an appropriate energy input. The excitation (energy pumping) is carried out with light (optical pumping) in the case of crystalline, so-called solid lasers and dye lasers (organic dye solutions) and by an electric discharge in gas-loaded lasers. For the sake of interest it should be mentioned that laser function may be produced not only by electron transition. For example, in the CO_2 lasers the vibration energy states of the CO_2 molecules play a role, while in the diode lasers the photons of the laser light arise from the recombination of the electrons and defect electrons. In the latter case the pumping is done by electric current, which provides for a sufficient number of electrons and holes in the *p–n* transition layer of the diode.

Finally, some laser types used in medicine and their wavelengths are listed below, without the intention to be exhaustive.

Gas lasers: He–Ne (633 nm); CO_2 (10.6 μm), argon ion (488 nm and 514 nm); excimer (excited dimer) lasers, their material is a mixture of an inert gas and a halogen, e.g. Kr–F, they radiate in the UV range.

Solid lasers: ruby (694 nm); the carrier material of the YAG lasers is $Y_3Al_5O_{12}$ (yttrium aluminium garnet) which is doped with Nd^{3+} (or with another lanthanide), their opera-

tional wavelength is in the range of 1–3 μm, depending on the doping. For the sake of interest it should be mentioned that amorph glas material was also doped with neodymium (Nd-phosphate glass laser, 1.05 μm).

Some laser operate in the pulse mode, i.e. they emit laser-light pulses of short duration either as individual pulses or pulse series. Other lasers radiate continuously, and there are lasers which may be operated in all of the three ways.

In medicine lasers are used for many purposes, beginning from the laboratory and clinical diagnostics through dermatology, cosmetology and ophthalmology to the various fields of surgery. The variability of the possibilities offered by the lasers will be demonstrated by just a few selected examples. Their application in surgery was already mentioned; here mainly CO_2 lasers and Nd–YAG lasers are used (with a power of about 1 W to 100 W). The light of the latter may be led also by an optical fibre which makes possible for example intracavitary radiations too. Due to coagulation, the incision made by the laser bleeds only slightly or not at all, the remaining tissues are spared, the healing process is much more rapid than after the conventional surgical interventions. These circumstances render possible in many cases the so-called one-day surgery. In dermatology and cosmetology, among others such laser lights (argon or Nd–YAG) are successfully used for the elimination of haemangiomas and birthmarks which are absorbed only slightly by the cutaneous tissue and to a great extent by the blood in the vessels present in the lesions; thus these vessels become coagulated, the blood supply of the lesions is terminated, and the lesions will be absorbed.

The laser methods took the place of many conventional methods in ophthalmological surgery. Some examples: interventions on the fundus (obliteration of capillaries, correction of the retinal injuries or the welding of the retina in case of its detachment) with laser lights (argon, krypton) which are hardly absorbed by the intermediate parts of the eye but are well absorbed in the target area; burning of holes on the iris for the alleviation of the symptoms of glaucoma (Nd-YAG); the correction of the ocular refraction by the appropriate "planing" of the superficial layer of the cornea (excimer laser).

In certain cases light sources are equipped by *light conducting optical fiber.* It is based on total reflection of light (Fig. 2.13). The material of long, thin, cylindrical fiber is transparent in the wavelength range of the applied light (e.g. proper glass). It has a core of relatively high refractive index (n_c) surrounded by jacket of lower refractive index (n_j). Light getting into the fiber with not too high angle of incidence is totally reflected on the boundary of core and jacket. Until reaching the other end of the fiber the number of total reflections can be several thousand. The diameter of fiber is around 10 μm and a flexible

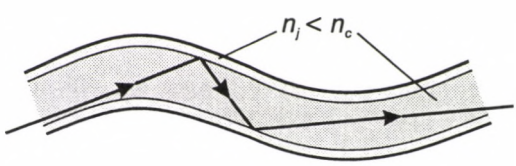


Fig. 2.13. Spreading of light in optical fiber

light conducting fiber bundle of some mm^2 cross-sectional area can consist of some ten thousand elementary fibers. (Optical fibers of 1–2 mm diameter are also used.) Ordered fiber bundle can be used for picture conduction (cf. section 6.7.1).

2.7. The biological effects of light

1. Division of the optical range. With regard to the effects of light on the living world the following further division is used:

100– 280	nm:	*UV C*	(far UV)			
280– 315	nm:	*UV-B*	(Dorno range)	400–420	nm:	violet
315– 400	nm:	*UV-A*	(near UV)	420–490	nm:	blue
400– 760	nm:	*VIS*	———	490–540	nm:	green
760–1400	nm:	*IR-A*	(near IR)	540–600	nm:	yellow
1.4– 3	μm:	*IR-B*	(medium IR)	600–760	nm:	red
3–1000	μm:	*IR-C*	(far IR)			

There are no sharp boundaries; transitions between the ranges are gradual.

The part of the *UV-C* range below about 180 nm is also called vacuum *UV* range, since it is absorbed in air and propagates only in vacuum.

2. Mechanism of action. The main phases of the development of biological effects are shown in Fig. 2.14.

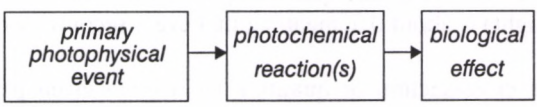


Fig. 2.14. Diagram of the development of the biological effect of light

The first step is always the absorption of the photon, the *primary photophysical event.* As a consequence the molecules of the biological system *get excited,* or in case of sufficiently high photon energy are *ionized.* Ultraviolet and visible light cause changes mainly in the electron energies, while infrared light in the vibrational and rotational energies.

If a molecule of essential importance in the biological effect is excited, a *direct photochemical reaction* takes place. As an example the ultraviolet (*UV-B, UV-C* ranges) damage of DNA should be mentioned. The absorbing molecules are the nucleotide bases and in the photochemical reaction the C(5)–C(6) double bonds of two pyrimidine bases (cf. section 1.5.4) stacked over each other are split, and a cyclobutane ring is formed with covalent bonds between the two bases. The photoproduct is the pyrimidine dimer. The biological consequence of this point-like injury, i.e. the *photobiological effect,* may be e.g. mutation.

Excitation may lead to an *indirect photochemical reaction* as well. In this case the absorbent, the so-called *sensitizer* molecule either transfers light energy to a molecule in the biological system on which the photochemical reaction takes place or gets itself into reaction with the molecule essential in the biological effect. An example for the first case is given by pigment molecules, playing a role in photosynthesis, which absorb visible light and transfer energy to the so-called reaction centres where the reaction itself is effectuated. For the second case an example may be a whole series of *photo-chemotherapeutical* agents (psoralens, hematoporphyrins, etc.). The reaction of psoralens is induced by near UV photons. The photoproduct inhibiting the nucleic acid replication is obtained via the formation of covalent bonds between the psoralen and the DNA molecule. Thus the photobiological effect is manifested in the inhibition of cell proliferation.

The effectiveness of light is usually described by several characteristics. One of them is the *cross-section* showing the surface of the given object (cell, molecule, etc.) pointing towards *one incident photon* with respect to the investigated effect (damage). This surface is usually smaller, but may be also greater than the geometric cross-section. It is obvious that the greater the cross-section, the more sensitive is the object. From among the nucleotide bases the pyrimidine bases in the double helical structure show a surface of about 8×10^{-20} cm^2 towards a photon of *UV-C* range, the pyrimidine bases in a single-stranded DNA (e.g. in certain bacterial viruses) show about one order of magnitude greater, and uracil in an uracil crystal shows 10^{-17} cm^2. (The geometric cross-section is about 10^{-16} cm^2.) This means that the bases of double helical DNA are less sensitive to UV photons.

For the characterization of the effectiveness of light also the *quantum yield* is applied. This quantity gives the ratio of either molecules transformed in a photochemical reaction or objects showing biological response to the effect of light related to one absorbed photon. In the uracil crystal e.g., the quantum yield of the formation of an uracil dimer (in case of 254 nm light) is about 0.5, meaning that every second absorbed photon causes dimerization.

The value of the cross-section or quantum yield for a given photobiological effect (photochemical reaction) depends usually on the wavelength of the inducing light. The function expressing this dependence is called *action spectrum;* it reveals the most effective wavelength(s) producing this effect. Figure 2.15 shows two remarkable action spectra. One of them is related to the damage of DNA, the other one to photosynthesis. In the damage of DNA the UV range, while in the photosynthesis the visible spectrum is effective, respectively. From Fig. 2.15*b* it becomes also apparent that in case of DNA damage the effectivity decreases by several orders of magnitude toward the longer wavelengths as compared to the maximal effect (at 260 nm). It can be easily seen that the action spectrum in a direct photochemical reaction is similar to the absorption spectrum of the molecule responsible for the biological effect, and in an indirect case to that of the sensitizer molecule (cf. Fig. 2.15*a*).

3. Photosynthesis. This chapter deals with the effects of light mainly in relation to the human organism, however light, more exactly visible light, affects through photosynthesis the whole living world, since chemical energy transformed from light energy by green plants ensures the energy supply of the living world to a decisive extent. The primary

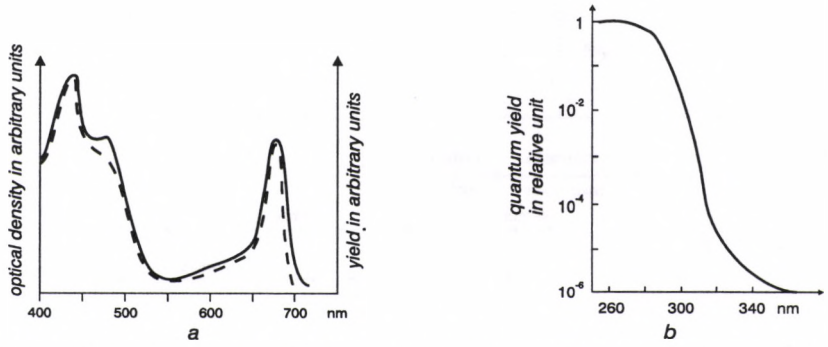


Fig. 2.15. a: Absorption spectrum of *Chlorella vulgaris* (full line) and action spectrum of its oxygen production proportional to photosynthesis (dashed line); *b:* action spectrum of DNA damage (the scale on the ordinate is logarithmic)

process takes place in the plant chloroplast, more exactly in its chlorophyll–protein complex photosynthetic units: on effect of light chlorophyll molecules are excited. Energy gets into the reaction centres specialized for photochemical energy conversion and causes charge separation, leading via a reaction chain partly to hydrolysis and partly to the incorporation of carbon dioxide into high-energy organic compounds. The process can be characterized by the following global reaction equation:

$$6H_2O + 6CO_2 + \text{light energy} \rightarrow (HCOH)_6 + 6O_2$$

This means: from carbon dioxide and water sugar and oxygen gas are produced. The absorbed light energy on the left side of the equation appears on the right side in the *chemical bond energies* of the products. In other words: the absorbed light energy accumulates in photosynthetic organisms in the form of chemical energy.

In Fig. 2.15*a* the continuous curve shows the absorption spectrum of a green alga, *Chlorella vulgaris.* The spectrum corresponds to the characteristic absorption of chlorophylls in the blue and red wavelength range (with maxima at 437, 475 and 676 nm, respectively). The dashed line is the action spectrum of oxygen formation and consequently of photosynthesis as well. The development of plants rendered possible the emergence of animal and other organisms which are unable to photosynthesize, moreover in their existence an opposite process, respiration and oxidation, is playing the main part. Oxygen necessary for these processes had been and is still developing in the Earth's atmosphere and animal and other organisms obtain the energy required for the sustenance of life ultimately from the plant organic substances produced by photosynthesis. They build their own organic substances by transformation of these compounds.

4. Effect on the eye. Light in its various ranges penetrates into the eye to different depths, thus it is absorbed and exerts its effect in different tissues. In the following the different photochemical reactions will not be discussed in detail, only a brief summary of the effects of light on the eye will be given.

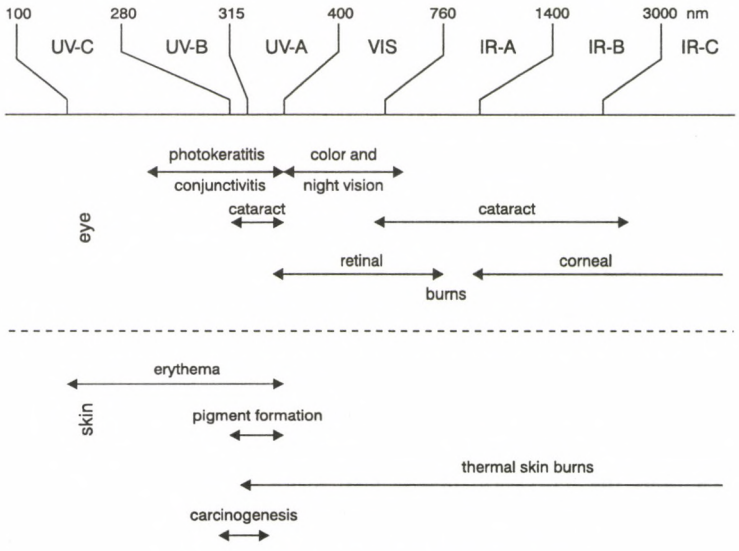


Fig. 2.16. Essential effects of the optical range
Wavelength scale is logarithmic

The *UV-A* range and the longer wavelength part of the *UV-B* range are absorbed in the cornea and the eye lens. According to the place of absorption their damaging effect is manifested partly as cataract, partly as photokeratitis and conjunctivitis. *UV-C, IR-B* and *IR-C* light does not penetrate behind the outer part of the eye, but is mainly absorbed in the cornea and causes mainly keratitis and, in case of the *IR* ranges, corneal burns. The various effects of the whole optical range on the eye are summarized in Fig. 2.16. The extent of the damage depends on the intensity and duration of the affecting light. This holds also for the visible light. In the *visual process* power densities occurring in the everyday life (some tenth to hundredth Wm^{-2}, cf. section 2.3.2) play a role, however, high-power lasers operating in the visible range (cf. section 2.6) may cause burns in various parts of the eye. The visible light passes through the cornea, the eye lens and the vitreous body. Through the refracting media light becomes converging and the photons are absorbed in the pigments of the retina, more exactly in those of the rods and cones.

The pigment in the rods is contained by a chromoprotein, *rhodopsin,* with an absorption maximum at about 500 nm. In the cones three light sensitive pigments are present with absorption maxima at about 570, 535 and 445 nm, respectively. The initial photophysical event in vision takes place in the pigment molecules and it triggers a reaction chain. In the first step, e.g. in the case of rhodopsin, a stereoisomerization of the pigment takes place with a rather high quantum yield of about 0.7. The further steps of the reaction chain affect the protein part and this results in the excitation process of the optic nerve (cf. section 7.5). Rods are essential in night vision and cones in daylight (colour) vision. The sensitivity curve of the eye in daylight has been shown in Fig. 2.4, which may be also considered the

action spectrum of colour vision corresponding to the averaged sensitivity of the cones containing three different pigments for light.

5. The effect on the skin depends partly on the penetration depth of the given light, and partly on where and by what kind of pigment molecules it is absorbed. The *UV-C* range and the shorter wavelength part of the *UV-B* range are absorbed mainly in the top skin layer, the stratum corneum, and thus this layer protects the deeper layers against these wavelengths. – However, one part of the *UV-B* range, especially the longer wavelengths and the *UV-A* light, penetrate into the epithelial layer of about 0.2 mm thickness, too, furthermore 30–40% of 400 nm light gets even into the corium. For *UV-B* light the absorbing molecules are the constituents of nucleic acids and proteins, i.e. nucleotide bases and aromatic amino acids (e.g. tryptophan, tyrosine) and their derivatives (e.g. DOPA, melanin). – As *UV-A* and visible light (penetrating in 60–70%, even into the corium) absorbing molecules we mention hemoglobin, carotenes, bilirubin, melanin.

The *IR-A* and *IR-B* ranges penetrate deepest into the skin. About 20% of the first one is absorbed at a depth of 3–3.5 mm from the skin surface, already in the subcutis. The more important light effects on the skin and the wavelength ranges inducing them are also summarized in Fig. 2.16.

The figure shows the effects of the *UV-C* range as well, though these components of sunlight are either absent or are present only very slightly on the Earth's surface. The ozone content of the atmosphere plays an important role in the absorption of the *UV-B* range. The depletion of the ozone layer encountered in the past two decades affects the whole living world due to the biological effects of the *UV-B* range (cf. also Fig. 2.15*b*). Ozone destruction is caused by air pollution, by the increase of the concentration of extraneous gases (freon, oxidation products of nitrogen and carbon, etc.). – Obviously, the biological effect of the *UV-C* range may be of interest at the application of various artificial light sources (e.g. quartz lamp, bactericidal lamps) as well.

A characteristic effect of *UV-A* and *UV-B* ranges is *pigmentation.* Light causes polymerization of the pigments in skin melanocytes and this provides protection for the deeper lying cells of the epithelium and corium. For example, light transmission of strongly tanned skin in the whole *UV* range is about half of the transmission of white skin.

The most striking and quickly appearing effect of *UV-A, UV-B* and *UV-C* ranges is *erythema* (skin reddening). At present it can only be assumed that this reaction chain leading to photobiological effect is related to the damage of the epithelial cells (mainly of their nucleic acid content) and to the toxic effect of substances released from the cells due to the damage.

The appearance of erythema is a sensitive indicator of light exposure. The *minimal erythema dose* (MED) is frequently used to characterize the effect of different wavelength ranges. MED is the incident threshold energy per unit surface inducing erythema. According to experience this value is about 3.7 mJ/cm^2 at 254 nm and 7 mJ/cm^2 and 13 mJ/cm^2 at 297 and 300 nm, respectively.

The effect of *UV-A* and *UV-B* exposure may be manifested in the formation of malignant melanomas and nonmelanoma skin tumours; this process is the *photocarcinogenesis.* According to statistics on a big population living at identical latitudes (i.e. exposed to the

same average sunlight intensity), among 100,000 white-skinned people nonmelanoma skin cancer occurs in 35 cases in age groups 35–44 years and in about 800 cases at 75–84 years. This is due to the fact that the light energy incident on the skin and its biological effect are summed up, i.e. with increased doses the probability of photocarcinogenesis increases. Progressing towards the equator the frequency of skin cancer increases gradually and at the equator it reaches about 3% at 75–84 years.

For the prevention of the above outlined damaging effects an international recommendation compiled from numerous experimental facts gives the *exposure limit* (EL), below which the health risk of persons exposed to sunlight or *UV*-light during work is still acceptable. The *UV* exposure limit similarly to the action spectrum of DNA depends on wavelength: the function takes into account the damages of the skin and the eye. To demonstrate this, Table 2.1 shows a few EL data for information. It can be seen that the EL value is the lowest at 270 nm, consequently this wavelength is the most dangerous. With both increasing and decreasing the wavelength the danger decreases.

Table 2.1. Exposure limit (EL) values for ultraviolet light*

Wavelength (nm)	EL (mJ/cm^2)
200	100
254	6.0
270	30
290	4.7
300	10
305	50
310	200
315	1.0×10^3
320	2.9×10^3
340	1.1×10^4
360	2.3×10^4
400	1.0×10^5

* Data according to the 1988 Recommendations of IRPA/INIRC (*Health Physics,* 56, 97, 1989)

2.8. On X-rays in general

X-radiation is produced whenever electrons become stopped[3] after striking some target with a sufficiently high velocity.

The practically important effects may be summarized as follows:

luminescence: certain materials (e.g. barium platinocyanide, calcium tungstate, zinc silicate, zinc sulphide doped with silver or copper) luminesce in response to X-ray irradiation;

photographic effect: a photographic plate is darkened similarly as in the case of light;

ionizing effect: the electrical conductivity of some materials is increased (this phenomenon is especially well observable with gases);

[3] X-radiation is also induced by other charged particles, but for practical purposes only electrons are used.

chemical effect: in water, for instance, hydrogen peroxide is produced;

biological effect: e.g. the production of morphological and functional changes in cells.

The primary effect produced by X-rays in atoms is in every case *excitation* or *ionization.* All the other effects are only consequences (secondary, tertiary effects and so on), i.e. indirect effects. Particularly complex processes precede the biological effects. The primary phenomena induce chemical processes in the molecules of the cell; the biological effects are results of these processes.

The common character of all these effects is the transformation of the X-radiation energy into some other energy. Beside these effects, the *formation of secondary X-radiation (X-ray scattering)* is of importance. This is a fundamental effect too, and accompanies *always* the propagation of X-rays in some medium.

The *detection* and *measurement* of X-rays are generally carried out via one of the effects listed. For example, in diagnostic X-ray irradiation the physician obtains a shadow picture on a luminescent screen or a photographic plate, and generally measures the radiation incident on the body, or absorbed by the body, by means of the ionization of the air or the darkening of the photographic plate due to the radiation (cf. section 3.3).

The wavelength of the X-radiation used in medical practice lies in the range 5–120 pm. This corresponds to a photon energy of 0.2–0.01 MeV. However, in recent decades X-rays of shorter wavelength (less than 1 pm), and consequently of higher photon energy (up to several MeV), have also become of increasing importance.

Though generally not valid, it is accepted in medical practice that the penetrating power of X-rays of shorter wavelength is greater than that of X-rays of longer wavelength (cf. section 2.10). The shorter wavelength radiation with its higher penetrating power is said to *be hard,* whereas the radiation of lower penetrating power is *soft.*

X-rays are used for both diagnostic and therapeutic purposes. The *diagnostic* application is based upon the fact that the various tissues absorb X-radiation to various degrees. The soft tissues are more transparent to X-rays than, for instance, the bones. Consequently, if the organism is transilluminated, brighter and darker domains are observed, depending upon the absorbance of the tissues.

In radiation effects, only the quantity of absorbed radiation is important in fact. The different cells show various sensitivities, young and actively dividing cells being especially sensitive. For this reason, every tissue and organ in which intensive cell regeneration is taking place, such as the bone marrow, the lower skin layers, the gonads, etc., should be protected with special care. For the same reason, certain pathologically reproducing cells are destroyed more quickly. This destructive effect is in part the basis of the *therapeutic application* of X-radiation (e.g. in the treatment of malignant tumours).

2.9. X-ray sources and their spectra

1. X-ray tube. In medical practice highly evacuated sealed tubes made of glass are generally used to produce X-radiation. The construction is outlined in Fig. 2.17. The electrons are supplied by the hot cathode, opposite which the anode (anticathode) is placed. A voltage (usually 10-400 kV) between the cathode and anode accelerates the

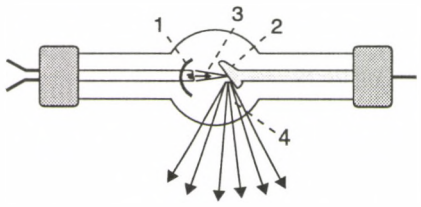


Fig. 2.17. Outline of an X-ray tube
1: hot cathode; *2:* anode (anticathode); *3:* cathode rays; *4:* X-rays

electron current, and the X-rays are produced by the electron impact on the anode. Only a few tenths of a percent of the energy of the electrons is transformed into X-radiation, the rest being converted into heat and raising the temperature of the anode.

The spot on the anode where the X-rays are produced is the *focus* of the tube. The smaller (the more point-like) the focus, the more perfect are the shadow images obtained on the luminescent screen or the photographic plate. In therapeutic tubes the area of the focus is less important whereas the performance of diagnostic tubes is much better if the focus is more point-like. The focus will be small if the electron beam striking the anode is concentrated; this can be achieved by applying a suitable electric field (electric lens). The X-radiation leaves the focus with maximum intensity nearly normal to the direction of the electron beam. In practice these rays are used; the other parts of the tube are covered with a lead coating, which absorbs the radiation propagating in other directions.

2. Particle accelerators. With X-ray tubes, depending upon the voltage across the tube, the maximum energy of the electron beam is only a few tenths of a MeV. In order to produce higher-energy electrons, and consequently harder X-rays, particle-accelerating equipments, recently mainly linear accelerators, developed in nuclear physics are applied (cf. section 3.2).

In medical practice both electrons with high energy (up to ca. 50 MeV) and X-rays induced by these electrons are used.

3. Bremsstrahlung and characteristic radiation. An X-ray tube simultaneously emits waves of various wavelengths and moreover the emitted power is wavelength-dependent. The curves in Fig. 2.18 show the power emitted in the various wavelength ranges. The kV values relating to the curves are the voltages on the tube (the accelerating voltages of the electrons).[4] Consider first Fig. 2.18*a*. The following characteristic properties are observed:

(a) The tube radiates in a broad wavelength range. Every wavelength is represented within this range, i.e. *the spectrum is continuous.*

(b) The left side of the spectrum has a sharp cut-off, which shifts towards shorter wavelengths as the voltage is increased. There is no limit on the right side, the emitted power gradually decreases with increasing wavelength.

[4] The ordinates of Figs 2.18*a* and *b* cannot be compared, since *b* has been reduced. For quantitative relations cf. paragraph 4 of this section.

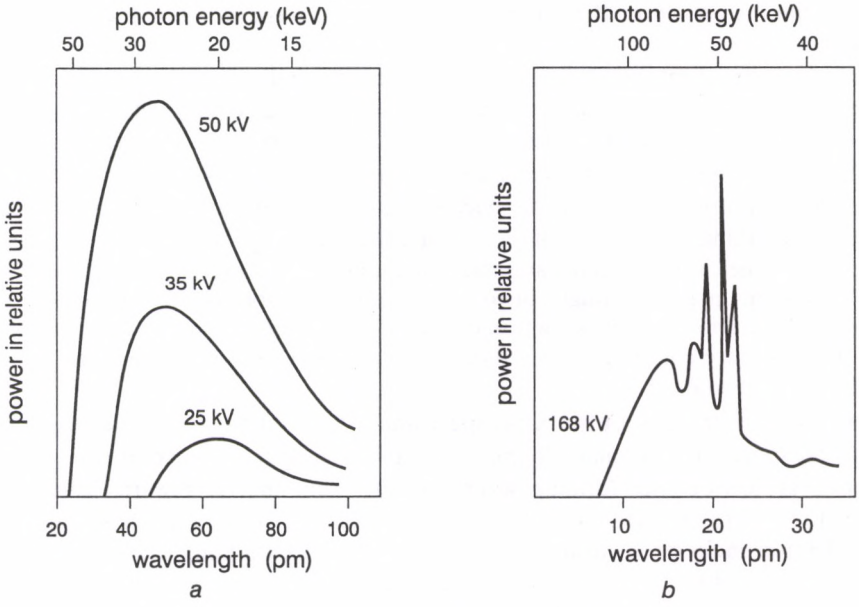


Fig. 2.18. Wavelength distribution of the emitted power for a tungsten anode
a: at lower; *b:* at higher voltages

(c) Only one well-developed maximum is observed; this points to the existence of a distinct wavelength represented by the maximum power in the continuous spectrum. With increasing voltage, this maximum shifts towards shorter wavelengths. Neither the short wavelength cut-off nor the position of the maximum depends on the material of the anode. The area under the curves is proportional to the total emitted power; the latter increases rapidly with increasing voltage (cf. also point 4).

(d) Consider Fig. 2.18*b*. If the voltage across the electrodes in the tube is sufficiently large, sharp peaks appear at certain sites in the curve indicating that some wavelengths are present in the spectrum with high power. This leads to the conclusion that the spectrum of an X-ray tube is actually composed of two parts. At lower voltage only the above continuous spectrum appears (Fig. 2.18*a*), but with increasing voltage a *line spectrum* is superimposed on the continuous one (Fig. 2.18*b*).

The two types of X-ray spectrum are due to two kinds of origin of radiation. *Bremsstrahlung* gives the continuous spectrum, whereas the other type, the *characteristic radiation,* is responsible for the line spectrum. The Bremsstrahlung is produced by electrons colliding with the atoms and subsequently being stopped by the atomic force field (or more precisely by the atomic nucleus rather than the whole atom). The characteristic radiation, on the other hand, is produced by the atoms of the anode, and is characteristic of the anodic target material. In this case too radiation is initiated by electron impact: however, X-radiation is due to the atoms excited by the electrons and not to the electrons themselves; when getting from an excited state to a lower energy

state, the atoms get rid of the released energy in the form of X-ray photons (cf. also section 2.11).

Optical spectra display considerable differences, depending on whether they relate to single atoms, or molecules, or larger continuous systems (e.g. solids). The difference is so large that in the spectrum of a molecule, for instance, the spectra of the individual atoms comprising the molecule are no longer discernible. However, in X-ray spectra there is no fundamental difference between the spectra of atoms and of continuous systems (in this respect we are thinking primarily of the characteristic X-ray spectrum). The characteristic properties of the atomic spectra are conserved, and the characteristic spectra of more complex systems may be thought of to a first approximation as the summation of the spectra of the atoms. The line structure of the atomic spectra is maintained when several atoms form a more complex system, though the lines may become broadened and their positions shifted somewhat.

Obviously the structure of the X-ray spectrum is more simple and clearer than that of the optical spectrum. The main characteristic of X-ray spectra is the presence of well-separated groups of lines forming several series. From shorter towards longer wavelengths, the groups are denoted by the capital letters *K, L, M, N,* etc. These letters refer to the electron shells or electron states producing the spectral lines (cf. section 2.11).

The positions of some characteristic X-ray emission spectrum series on the wavelength scale are presented in Fig. 2.19. For elements of lower atomic number, only the *K* series is emitted. On increase in the atomic number of the target material, the series is shifted towards shorter wavelengths, and the *L* series, and later the *M* and *N* series, too, appear.

Optical spectra are produced either by light refraction (more exactly dispersion) or by diffraction, X-ray spectra are obtained only by diffraction. The phenomenon of refraction is rather inextensive for X-rays, since at a wavelength less than 10 nm the refractive index of every material is close to 1.

4. Emitted power. The greater part of the power of the emitted radiation due to the Bremsstrahlung. In practice, therefore, it is sufficient to discuss only this type of radiation. The following relation holds to a good approximation: the power P of the radiation is

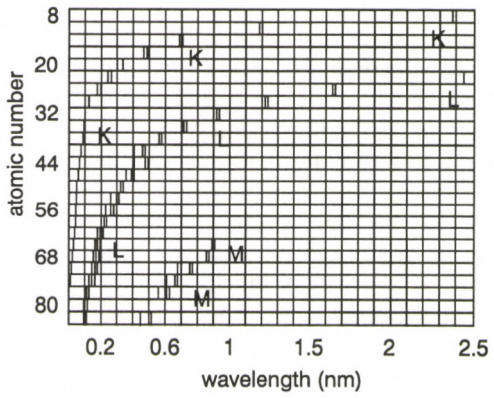


Fig. 2.19. K, L and *M* spectral series of the elements

proportional to the square of the voltage U on the tube, the intensity I of electron current, and the atomic number Z of the target (anode) material:

$$P = cU^2IZ \qquad [2.19]$$

If the voltage is given in V, the current in A, and the power in W units, the value of the proportionality factor c is approximately 10^{-9}. (Its unit is 1/V.)

From the results obtained in the previous section, the following practical consequences can be drawn:

(a) An increase in the voltage results in shifts in the short wavelength cut-off and the maximum power so that the emitted radiation becomes richer in short wavelength components, which allows regulation of its hardness.

(b) The power of the radiation increases in proportion to the intensity of the electron current, and to the square of the voltage, and consequently the intensity of the radiation, too increases. If it is desired to change the intensity without varying the hardness of the radiation, only the electron current should be changed. This is attained by variation of the cathode filament heating.

Taking into account [2.19], the quantitative relation expressing the *efficiency* of the X-ray tube as a radiation source can be given. If the voltage is denoted by U, and the electron-current intensity by I, the invested electric power will be $P' = IU$. The degree of efficiency (η) is then the quotient of P given in [2.19] and P':

$$\eta = cUZ \qquad [2.20]$$

2.10. The attenuation of X-radiation

2.10.1. The law of attenuation

If a monochromatic, parallel X-ray beam propagates in some medium, its *intensity* decreases according to a law similar to that for some other photon radiation (cf. section 2.3.1), i.e.

$$I = I_0 e^{-\mu x}, \quad \text{or} \quad I = I_0 e^{-\frac{0.693}{D} x} \qquad [2.21]$$

where I_0 is the intensity of the radiation penetrating the medium, I is the intensity of the radiation which passed through a layer of thickness x of the medium, μ is the attenuation coefficient, and D is the half-value thickness.

The beam may also be characterized by the *photon flux*, instead of intensity. The photon flux is defined as the number of photons flowing per unit time through unit cross-section. Let N_0 denote the photon flux penetrating into a layer of thickness x, and let N denote the photon flux passing through the layer. The following equation then holds:

$$N = N_0 e^{-\mu x}, \quad \text{or} \quad N = N_0 e^{-\frac{0.693}{D} x} \qquad [2.22]$$

The value of μ depends upon the energy of radiation and the material of the medium. This latter dependence does not refer only to the fact that μ is different, for instance, for water and air; it also depends upon the density of a given substance. Accordingly, if the density of some substance changes, μ will also change. As an example, the value of μ is higher in the solid phase than in the gaseous state of the same material. With X-rays μ changes in proportion to the density (ρ), from which it follows that the quantity

$$\mu_m = \frac{\mu}{\rho} \qquad [2.23a]$$

called the *mass-attenuation coefficient,* is independent of the density and depends for a given substance only on the energy of the radiation. If for instance μ is measured in cm^{-1} and ρ in g cm^{-3}, μ_m is obtained in cm^2 g^{-1}. In order to distinguish μ and μ_m, the former is frequently called the *linear attenuation coefficient.*

For a clear interpretation, let us transform the exponent in [2.22]. It is obvious that $\mu x = \mu_m x_m$, where

$$x_m = \rho x \qquad [2.23b]$$

Consequently

$$N = N_0 e^{-\mu_m x_m} \qquad [2.23c]$$

Since ρ is the mass of material in a volume of 1 cm^3 (the shaded volume in Fig. 2.20) $x_m = \rho x$ gives the mass of material in a prism with a length of x and a cross-section of 1 cm^2. Let x_m be the reciprocal of μ_m. In this case the incident flux (or the intensity) decreases by a factor e. Hence, μ_m is the reciprocal of the mass of a prism, whose cross-section is 1 cm^2, and whose length decreases the incident flux (intensity) to $1/e$ of its original value. This mass is clearly the same in the gaseous and in the solid states of the substance, the only difference being that it fills the volume of a shorter or longer prism (i.e. a thinner or thicker layer).

Analogously to the half-value thickness one may define the half-value mass D_m. This is the mass (for a cross-section of 1 cm^2) which decreases the incident flux (intensity) by half

$$D_m = \rho D \qquad [2.23d]$$

The usual units for the half-value mass are g cm^{-2}.

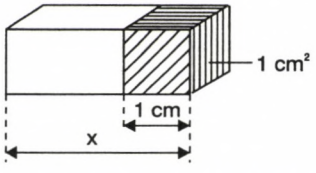


Fig. 2.20. Diagram relating to the interpretation of μ_m

2.10.2. Processes leading to attenuation

a) The photoelectric effect (photoeffect, photoelectric absorption) consists of an interaction between a photon of energy $h\nu$ with one of the electrons (usually an inner shell electron) of an atom; by transferring *all of its energy* the photon is annihilated (Fig. 2.21). The energy received causes the electron (photoelectron) to rise to the surface of the atom, and with the remaining energy as kinetic energy escapes from the atom. The following energy balance applies:

$$h\nu = A + \frac{1}{2}mv^2 \qquad [2.24]$$

where $mv^2/2$ denotes the kinetic energy of the moving electron and A is the work of release necessary to raise the electron from some inner level to the atomic surface (its value for the K shell is 5–100 keV). After leaving the atom, the electron induces excitation and ionization until its excess energy is lost. The production of an ion pair in the air (or in tissues) consumes on average an energy of 34 eV, and thus a single photoelectron with an energy of 340 keV can produce 10,000 ion pairs. It follows that X-radiation does not excite and ionize directly; this is done rather by the high-energy electrons produced by the X-rays (for further details cf. sections 3.2.3 and 3.3).

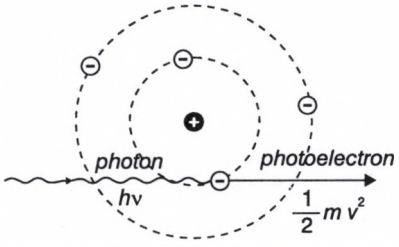


Fig. 2.21. The photoelectric effect

The exciting and ionizing energy may be the source of numerous processes, such as heat formation, luminescence, formation of active chemical radicals, etc. However, in many cases the atoms return to their ground state by the *emission of characteristic X-radiation.* It should be noted in this connection that in some rare cases one photoelectron (every hundredth or thousandth) may suddenly be stopped in the field of an atom and, similarly to the processes in the X-ray tube, produces *Bremsstrahlung.* The series of processes continues, since the resulting characteristic radiation and Bremsstrahlung may be the source of further processes.

b) In the **Compton effect** a photon of energy $h\nu$ again interacts with an electron, but here it transfers only *part of its energy* to the electron, and continues moving with a smaller energy $h\nu'$ in a changed direction (Fig. 2.22). In contrast with the photoeffect, the Compton effect occurs with great probability with loosely bound (or free) electrons. The process may be described by the following energy balance

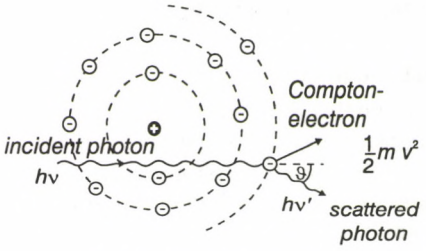


Fig. 2.22. The Compton effect

$$h\nu = A + \frac{1}{2}mv^2 + h\nu' \qquad [2.24]$$

Since $h\nu'$ is smaller than $h\nu$, it follows that λ' is larger than λ, which means that this process leads to the softening of the radiation; this is the Compton scattering, the scattered electrons are Compton electrons. The change in the wavelength is independent of the wavelength of the incident photon and depends only on the direction (ϑ) of the scattering.

The fate of the Compton electrons is subsequently similar to that of photoelectrons, and the scattered photons behave in the same way as other photons.

c) **Pair production.** In this process an electron–positron pair is produced by an X-ray photon in the vicinity of an atomic nucleus (Fig. 2.23). The process is governed by Einstein's energy–mass equivalence: $E = mc^2$. The rest mass of an electron (or a positron) is equivalent to 0.51 MeV, and consequently that of one pair to roughly 1.0 MeV. This means that pair production occurs only if the energy of an X-ray photon is at least 1.0 MeV. If its energy is higher, the excess appears in the kinetic energies of the electron and the positron. The components of the pair subsequently produce ionization and excitation in the same way as photoelectrons or Compton electrons with the corresponding energy. On slowing down, the positron component unites with an electron, their encounter leading to their annihilation, generally with the emission of two γ-photons. The γ-photons induce the same effects as X-ray photons of equivalent frequency.

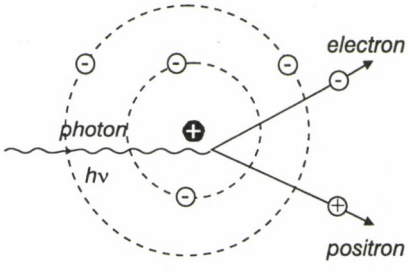


Fig. 2.23. Pair production

d) Coherent or classical scattering. The above three processes result in the total or partial transformation of the photon energy. However, the photon sometimes changes only its direction, without energy loss. This type of coherent (classical) scattering occurs mainly on electrons; the scattering on atomic nuclei is negligible compared with that on electrons.

e) Nuclear reactions. High-energy photons may interact with atomic nuclei. The total energy is transferred to the nucleus, more exactly to a nucleon in it. This nucleon may then have sufficient energy to escape from the nucleus. In most cases a neutron is released; proton release is rare. The binding energy per nucleon in the different nuclei is about 7–8 MeV. Nuclear reactions can be produced only with X-rays having a higher photon energy. With two of the lightest elements in the periodic system (heavy hydrogen and beryllium), however, a photon energy of even a few MeV is sufficient to induce nuclear transformations: the nucleus absorbs a photon and releases neutron. In this way a new nucleus is produced, and the released neutrons readily give rise to further nuclear reactions.

2.10.3. Attenuation (absorption) spectra

The processes discussed above do not occur with equal probability. At low photon energies (especially with elements of high atomic numbers) the photoelectric effect prevails; with increasing energy the probability of the photoeffect decreases, and (especially with elements of low atomic numbers) the Compton effect comes into prominence. The coherent scattering is appreciable only in the case of low photon energies (below 50 keV), but even here it is merely 6–10% of the Compton scattering. Above 1 MeV, pair production too assumes increasing importance. Besides these effects, the probability of nuclear reactions is generally small.

For the different processes the attenuation coefficients μ and μ_m are considered to consist of several terms:

$$\mu = \tau + \sigma + \kappa \text{ and } \mu_m = \tau_m + \sigma_m + \kappa_m \qquad [2.26]$$

where τ, σ, and κ denote the linear attenuation coefficients due to the photoeffect, the Compton scattering and the pair production, respectively, while τ_m, σ_m and κ_m refer to the respective mass attenuation coefficients. In [2.26] the nuclear reactions are usually neglected, and the classical scattering is considered together with the Compton effect. However, if the radiation is sufficiently hard, compared with the Compton scattering the classical scattering is disregarded.

The values of the attenuation coefficients depend on the material of the medium and the energy of the photons. As an example, the values for lead, playing an important role in radiation protection, are presented. Figure 2.24 refers to lower, and Fig. 2.25 to higher photon energy. The former presents the absorption spectrum produced only by the predominant photoelectric effect (absorption spectrum in the narrower sense), while the latter depicts not only the resulting attenuation spectrum, but also the component spectra. Figure 2.25 illustrates the above observations relating to the courses of the component spectra too.

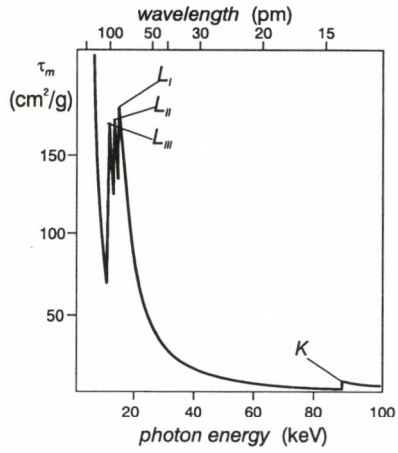


Fig. 2.24. Absorption spectrum of lead

From Figs 2.24, 2.26 and 2.27, further conclusions may be drawn with regard to the attenuation (absorption) of X-radiation.

a) Not only the emission, but also the absorption spectra are more simple in the X-ray than in the optical range. Thus, for instance, it is striking in Fig. 2.24 that at certain wavelengths the curve displays peaks, called *absorption edges.* Similarly as for the emission lines, these edges form groups, and are appearing in about the same energy range as the emission lines. The edges are also denoted by capital letters *K, L, M,* etc. on proceeding from higher to lower energy values (i.e. from shorter to longer wavelengths). The spectrum shows the single *K* edge and the *L* triplet of lead (lower energy groups, including the *M* edges, are not depicted).

b) Figure 2.26 shows the absorption spectra of some elements with low atomic numbers. Neither edges nor peaks are to be seen in the curves; *K edge with highest energy* appears below 0.01 MeV (i.e. at a wavelength longer than 120 pm), and consequently

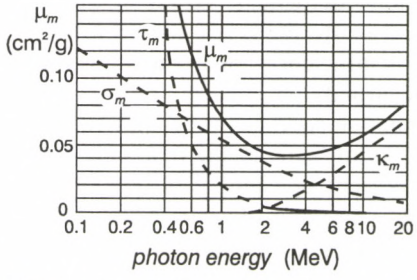


Fig. 2.25. Total and partial attenuation coefficients of X-rays as functions of photon energy, for lead
Logarithmic scale on the abscissa

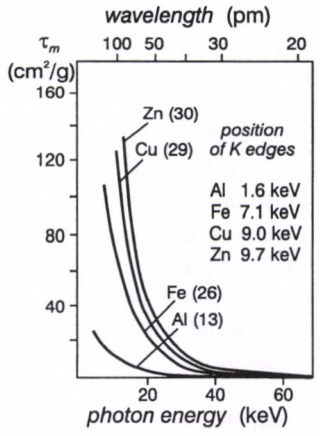


Fig. 2.26. Absorption spectra of some elements of low atomic number

outside the range shown in the diagram. For elements with even smaller atomic numbers, which are especially important in building up living organisms, the K edges lie at even lower photon energies (i.e. at longer wavelengths).

c) In the range between two neighbouring edges, τ_m increases roughly in proportion to the third power of the wavelength (λ). This is the mathematical reflection of the empirical fact that softer X-rays are generally more absorbed.

d) Apart from the peaks, at a given wavelength τ_m increases with the atomic number Z of the element; the increase is roughly proportional to the third power of the atomic number. This expresses the important empirical fact that a given radiation is absorbed more strongly by elements of higher than of lower atomic number. This may be generalized: absorbents which are richer in elements of higher atomic number absorb the radiation more strongly than absorbents composed of substances of lower atomic numbers.

The results of points (c) and (d) can be given in a single relation:

$$\tau_m = C\lambda^3 Z^3 \qquad [2.27]$$

The value of the proportionality factor C on the short-wave side of the K edge is 5.5–6.5 (if λ is measured in nm and τ_m in cm^2g^{-1}). If not elements but complex substances are investigated, Z is replaced by the mean atomic number, also called the *effective atomic number,* which is calculated from the relative amounts and the atomic numbers of the components. The effective atomic number of air is approximately 7.3, and that of water somewhat higher. Obviously, both are smaller than the atomic number of oxygen (which is 8), since both nitrogen and hydrogen precede oxygen in the periodic system.

e) [2.27] is applied in practice in several cases. The reason why soft tissues do not absorb X-rays very strongly is that they consist mainly of elements of low atomic numbers, whereas bones absorb more strongly, since they also contain relatively large proportions of elements of higher atomic numbers.

If there is no essential difference between the tissue or organ to be investigated and its surroundings, an artificial contrast may be made by applying some *contrast substance.* The stomach and intestines, for instance, are well outlined if a barium sulphate suspension is introduced into the system; this absorbs X-radiation more strongly than the surroundings do (the atomic number of barium is 56). *Liquid* contrast substances, for instance iodine-containing solutions, can also be used to obtain contrast images of the kidneys, the gall-bladder, the blood vessels, etc., which thereby become well outlined (the atomic number of iodine is 53). In some cases *gaseous* contrast materials are used. Their low density means that they absorb less strongly than the tissues, making thus possible the observation of details. Materials which absorb more strongly than their surroundings are called *positive* contrast materials, and the weaker absorbents *negative contrast materials.*

Because of its large atomic number (82), lead is a rather good absorbent. For this reason it is used in radiation protection. It is also understandable why surface therapy makes use of longer wavelengths, which are more strongly absorbed, whereas radiation of shorter wavelengths, which penetrates deeper into the tissues, is used in deep therapy.

As already outlined, the radiation emitted by an X-ray tube consists of components of different wavelengths which are absorbed to different extents. Of the radiation encountering the surface of the body, mainly the longer wavelength radiation is absorbed by the surface layers. Hence, the propagating radiation gradually becomes poorer in softer components and richer in harder components. In therapeutic treatment the softer radiation absorbed in large quantities by the surface layers (e.g. the skin) may cause undesired damage; in order to avoid this, the softer radiation is filtered out with intermediate metal plates (copper, aluminium, etc.; Table 2.2). The *filters* also attenuate the stronger components, but to a lesser extent than the softer ones. As a consequence, the radiation falling on the body will be more homogeneous, and at the same time the damage to the surface layer will be decreased.

Table 2.2. Half-value thicknesses of some substances

Wave-length (pm)	Photon energy (keV)	Air (standard state)	Half-value thickness (cm)			
			Water	Al	Cu	Pb
10	124	3800	4.55	1.72	0.25	0.017
30	41	2150	2.9	0.55	0.017	0.0044
50	25	1120	1.45	0.122	0.0042	0.0010
100	12	205	0.25	0.017	0.00061	0.000077
200	6	25.5	0.033	0.0024	0.000071	0.000012

f) Figure 2.27 depicts the absorption curves of air and water. The soft tissues of the body absorb in practically the same way as water, and consequently the data relating to the water curve are also valid for the tissues.

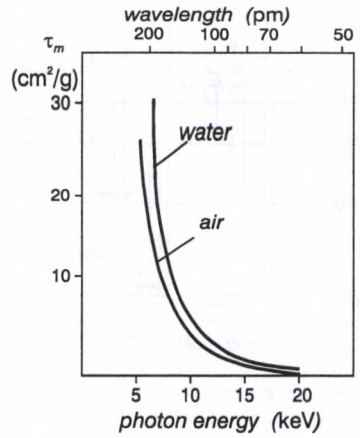


Fig. 2.27. Absorption spectra of air and water

It should be noted that the curves for air and water are nearly parallel and the ratio of their absorption constants is practically the same throughout the whole range. This fact plays an important role in *X-ray dosimetry.* The X-radiation absorbed in the air is relatively easily measured, in contrast with the radiation absorbed by water or by tissues (cf. section 3.3.2). From the similarity of the curves it follows that the energy absorbed by the tissues at some site may be considered proportional to the energy absorbed by the air at the same place, and what is even more essential, the proportionality factor is the same in the wavelength range of practical importance. Consequently, it is sufficient to measure the energy absorbed by the air, for if this is found to be higher by a factor of k, then the tissues at the same site will also absorb k times more energy, regardless of whether the X-ray spectrum is richer in harder or in softer components.

The attenuation of X-radiation has been seen to result from *scattering* too (classical or Compton scattering). The mass scattering coefficient for light elements in the range generally used in medical practice is nearly independent of the wavelength; its value is ca. 0.2 $cm^2\ g^{-1}$, except for hydrogen, which has a value of ca. 0.4 $cm^2\ g^{-1}$. The mass scattering coefficient of the body tissues is close to 0.3 $cm^2\ g^{-1}$. In practice the scattering is never negligible. It must be taken into account in the determination of the dose when radiation is used for therapeutic purposes, and it must never be neglected when the health protection of the medical staff is concerned.

2.11. Interpretation of X-ray spectra

1. Origin of characteristic radiation. Whereas the optical spectrum provides information about the changes in the state of the outer electrons, also called optical electrons, and thus about processes occurring in the outer energy levels, the characteristic X-ray spectra shed light on the processes within the inner electron shells. This situation is presented in Fig. 2.28, which shows as an example the energy levels of the copper atom. The diagram

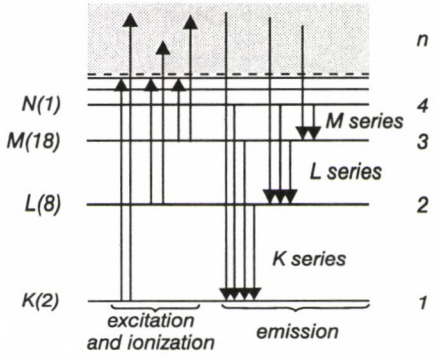


Fig. 2.28. Simplified energy level system of the copper atom
The numbers in brackets refer to the number of electrons in the individual shells

is only an outline, since every principal quantum number $n = 1, 2, 3, \ldots$ is represented by only one level, though they actually consist of several levels close to one another (see section 1.2.2). Consequently, every line represents a group of energy levels. Optical spectra are generated by exciting the *outer* electrons from the state with principal quantum number $n = 4$ into some higher unoccupied state (possibly in the continuum), after which the electron returns in one or more steps to the level of the principal quantum number $n = 4$. X-radiation, however, is generated if some of the electrons of the *K*, *L* or *M* shells are excited to higher levels or into the continuum (this process is represented in the diagram by arrows pointing upward), and electrons in higher energy state return to the holes generated by the excitation in these shells (arrows pointing downward). The *K* series is produced by the transitions whose final state is the *K* state, while the *L* series and the *M* series are the transitions to the *L* and *M* states, respectively. The whole picture is reminiscent of the generation of the hydrogen atom spectrum and the various series can be described with relations similar to the Lyman, Balmer, etc. series. The similarity also exists insofar as every series is joined by a continuum.

In order for X-radiation to be produced, the creation of free sites or holes in the inner shells is necessary. Until this occurs, the inner shells cannot accommodate electrons, since according to Pauli's principle they are completely filled. Consequently, it is impossible to excite only optical electrons so that they are moved, say, into the *K* shell instead of some outer (partly or fully) empty level. In the same way it follows that an excited *K* shell electron cannot move directly into the filled *L* shell above the *K* level. This transition can occur only if there is an empty site (hole) in the *L* shell, but this is highly improbable. This explains why the arrows indicative of excitation without exception point upwards to the upper empty levels, i.e. to the continuum in the diagram.

The differences between the inner energy levels are much larger than those in the optical range, and hence the energy of X-ray photons is much larger than the energy of optical photons. This train of thought makes it obvious that elements whose electrons occupy more and more levels with increasing atomic number yield an increasing number of series, which in turn are increasingly richer in lines, and the spectra become shifted

towards higher energy values. In the case of hydrogen and helium, which have no inner shells, there is no X-ray spectrum, but only an optical one. From lithium to neon, only the *K* series is present; from sodium to argon, the *L* series also appears; and so on.

Excitation and ionization, i.e. creation of free sites in the inner shells, occur in X-ray tubes if high-energy electrons impinge on the anode, but excited states can be created by any charged particle (e.g. an alpha-particle) striking some target with sufficiently high energy. Excitation or ionization is always accompanied by the release of characteristic X-radiation (cf. section 3.2).

2. Interpretation of absorption edges. Absorption of X-radiation also produces excitation and ionization. The transitions on the left side of Fig. 2.28 demonstrate the processes occurring when X-radiation is absorbed. Based on it the appearance of the absorption edges is easily followed. Let us explain, for instance, the *L* edge (or edges). As long as the energy of the incident photons is not large enough to raise an *L* electron to the lowest empty level (or partly empty level), no absorption occurs in the *L* shell. If the photon energy is large enough to produce this transition, however, there is some probability that the atom will absorb the photon, and the absorption suddenly increases. This increase in absorption produces the *L* absorption edges. Photons whose energy is larger than the limiting energy may be absorbed too, though the probability of absorption decreases with increasing energy. For this reason, on passing from the *L* edges to the *K* edge the absorption constant decreases; however, it increases again if the photon energy is sufficiently high to raise the electron from the *K* shell. The occurrence of multiple edges is connected with the fact that several energy levels belong to the same shell, and excitations may occur in any of them. Each of them gives rise to individual absorption edges, but these are situated close to one another (cf. section 2.10.3).

3. Interpretation of Bremsstrahlung. So far, the only electron transitions considered have been those in which the initial or final state (possibly both) belong to some *discrete* energy level. However, transitions also exist in which both the initial and final states belong to the *continuous* energy range (the continuum). In this case free electrons pass over from one state into another. Two possibilities exist: the free electron returns from a higher to a lower state, or conversely the transition is from a lower to a higher energy state. In the first case the electron becomes decelerated in the atomic or ionic force field, while in the second it is accelerated. The first case may be accompanied by the emission of radiation, whereas in the second case radiation may be absorbed. Since both the initial and the final states now lie in the continuous energy range, the spectrum will be continuous.

From the practical viewpoint, especially emission spectra induced by the deceleration of high-velocity electrons in the force field of an atom or ion are important. The radiation produced is the Bremsstrahlung and its continuous spectrum is called decelerating continuum. This process occurs, for instance, in X-ray tubes when accelerated electrons are decelerated in the force field of the atoms of the anticathode.

The decelerated electrons lose various amounts of energy. The energy may be lost step by step in small doses, or it may be lost in a single act. The various energy losses result in

the emission of photons of various energies. Maximum-energy photons, i.e. photons of the smallest wavelength, are emitted if the electrons lose their total energy in a single act; the short-wavelength limit of the spectrum in Fig. 2.18 can be interpreted by this. In any given case the maximum photon energy corresponding to the minimum wavelength is easy to calculate from the law of conservation of energy. If the maximum photon energy is denoted by hv_l, and the kinetic energy of the impinging electron is $m_e v^2/2$:

$$\frac{1}{2} m_e v^2 = h v_l \qquad [2.28]$$

where v_l is the frequency limit. [2.28] can be written in the form used in practice by expressing the kinetic energy in terms of the accelerating voltage U ($m_e v^2/2 = eU$), and introducing the limiting wavelength $\lambda_l (v_l = c/\lambda_l)$ instead of the limiting frequency. Substituting into [2.28] leads to

$$\lambda_l = \frac{hc}{eU} \qquad [2.29]$$

[2.29] is known as the *Duane–Hunt law* in radiology, for the relation between U and λ_l was carefully measured by these authors.

REFERENCES

Books

Gilford, G. A., Handbook of Physics for Radiologists and Radiographers. John Wiley & Sons, New York (1984)

Holick, M. F., Jung, E. G., Biologic Effect of Light. Walter de Gruyter, Berlin–New York (1996)

L'Ésperance, F. A. Jr., Ophthalmic Lasers, Photocoagulation, Photoradiation and Surgery. C. V. Mosby Company, St. Louis (1983)

Müller, G., Chance, B., Alfano, R. et al., Medical Optical Tomography: Functional Imaging and Monitoring. SPIE Optical Engineering Press, Washington (1993)

Silney, D., Wolbarsht, M., Safety with Lasers and Other Optical Sources. Plenum Press, New York–London (1980)

Ultraviolet Radiation. An Authoritative Scientific Review of Environmental and Health Effects. WHO, Geneve (1994)

Young, A. R., Björn, L. O., et al., Environmental UV Photobiology. Plenum Press, New York–London (1993)

Papers

de Gruijl, F. R., Van der Leun, J. C., Systemic influence of pre-irradation of a limited skin area on UV-tumorogenesis. *Photochem. Photobiol., 35,* 379 (1982)

Diffey, B. L., McKinlay, A. F., The UVB content of 'UVA fluorescent lamps' and its erythemal effectiveness in human skin. *Phys. Med. Biol., 28,* 351 (1983)

Elson, E. L., Magde, D., Fluorescence correlation spectroscopy. *Biopolymers, 13,* 1 (1974)

Kripke, M. L., Immunologic mechanisms in UV-radiation carcinogenesis. *Adv. Cancer Res., 34,* 69 (1981)

Rontó, Gy., Gróf, P., Gáspár, S., Bérces, A., Biologic Dosimeters in Photomedicine. In: *Biologic Effects Light,* Eds., Holick, M. F., Jung, E. G., Walter de Gruyter, Berlin–New York (1996)

3. NUCLEAR RADIATIONS AND THEIR APPLICATIONS

In the last decades the application of nuclear radiations in the medical science, together with other radiations, has significantly broadened. At the beginning of the century nuclear radiations were used mainly for therapeutic purposes. In the 40s (in Hungary only in the 50s), on the other hand, the appearance of the artificial radioactive isotopes opened a new era both in scientific research and in the field of diagnostic and therapeutic methods. As for the last 10 to 20 years, nuclear radiations gained a further important role in the so-called iconographic methods: by their means namely many-sided and direct information can be obtained concerning the condition and function of the organism. In the following the nuclear physical bases of these possibilities will be discussed. The theoretical bases of the function of the apparatuses will be presented in Chapter 6, together with other instruments related to electronics, automation and computer technology.

3.1 Radioactive isotopes. The decay law. Biological half-life

1. Radioactive isotopes. Certain atoms (isotopes) are unstable and their nucleus disintegrates with the emission of some particle. Such atoms are *radioactive atoms (radioactive isotopes).* Numerous radioactive isotopes occur naturally, mainly among the heavy atoms at the end of the periodic table (e.g. uranium, thorium, actinium).

The *natural* radioactive isotopes of the heavy elements belong to the uranium, thorium and actinium families. Each family is headed by a long-living primary isotope: at the head of the thorium series is $^{232}_{90}$Th, at that of the actinium series is $^{235}_{92}$U, and at that of the uranium–radium series is $^{238}_{92}$U. The decay of the primary isotope leads to another radioactive isotope which decays further, the extensive decay chain leads finally to a stable state. The three radioactive families contain about 44 members; the final product in each case is a stable lead isotope.

Some lighter elements too are known to occur naturally as radioactive isotopes. In natural potassium, for instance, besides the stable $^{39}_{19}$K and $^{41}_{19}$K a very small quantity (0.012%) of radioactive $^{40}_{19}$K can be found; rubidium too has a natural radioactive isotope. $^{147}_{62}$Sm is similarly radioactive by nature. From a biological aspect the presence of $^{40}_{19}$K in natural potassium is of special interest, since potassium is an important component of human food and the human organism.

Not only natural radioactive elements, but also more than a thousand *artificial* radioactive isotopes are known. They are produced by changing the original proton–neutron ratio in a stable nucleus by some suitable method (e.g. initiating nuclear processes by bombarding the nucleus with high-energy particles). Numerous radioactive isotopes of various elements of the periodic system are used in medical science. For instance, besides

the single stable isotope of iodine ($^{127}_{53}$I) more than ten radioactive iodine isotopes exist; the radioactive variants of sodium ($^{22}_{11}$Na, $^{24}_{11}$Na) are also frequently used.

Since the radioactive decay and connected processes of natural and artificial radioactive elements are governed by the same laws, no differences will be made between them in the following sections.

2. The decay law. Consider a radioactive preparation containing N radioactive atoms of the same kind. Each of them possesses an identical excess energy, but nevertheless their decay does not occur simultaneously. It cannot be stated with certainty at what time a given atom will decay. However, if sufficient atoms are involved, the question of how many of them will decay in a given time can be answered. The *decay number per unit time,* or more exactly the decay rate *dN/dt, is proportional to the number of undecayed atoms at a given time:*

$$\frac{dN}{dt} = -\lambda N \qquad [3.1]$$

The proportionality factor λ is different for the different isotopes, but it is constant (the *decay constant*) for the same kind of atoms. According to [3.1] λ defines the fraction of the total number of atoms which decay in unit time. If, for instance, $\lambda = 1\ h^{-1}$, 1/3600 of the total number of atoms disintegrate per second. (The negative sign indicates that the number of radioactive atoms decreases in time.) Integration of [3.1] gives

$$N = N_0\, e^{-\lambda t} \qquad [3.2]$$

where N_0 denotes the number of undecayed atoms at $t = 0$ (at the beginning of the observation), and N is the number of undecayed atoms at time t. [3.2] indicates that the decay of radioactive substances is governed by an exponential law; $N = 0$ at $t = \infty$, which means that in principle a given radioactive substance disintegrates totally only in an infinite time.

Instead of the decay constant generally the *half-life* (T) is used; this is the period during which the number of disintegrating atoms decreases by half. [3.2] can be rewritten:

$$\frac{N_0}{2} = N_0\, e^{-\lambda t} \quad \text{from which} \quad T = \frac{0.693}{\lambda} \quad \text{or} \quad \lambda = \frac{0.693}{T} \qquad [3.3]$$

An unambiguous relation exists between λ and T, expressed by [3.3]. The half-life of $^{131}_{53}$I is 8 days, that of $^{32}_{15}$P 14.3 days, that of $^{226}_{88}$Ra 1600 years, and that of $^{238}_{92}$U 4.5×10^9 years. The half-life is not influenced by the usual physical and chemical effects. The half-life of an isotope incorporated in some compound is identical with that of the elementary isotope. Pressure, a magnetic field and heating do not change the decay rate of an isotope. (Only at very high temperatures of 10^4–10^5 K is some detectable deviation observed.)

With the use of the half-life, the decay law can be reformulated to give

$$N = N_0\, e^{-\frac{0.693}{T}t} \qquad [3.4]$$

The *mean lifetime* (τ) is the time in which the number of undecayed atoms decreases by a factor e. Straightforward calculation yields

$$\tau = \frac{1}{\lambda} = 1.443T \qquad [3.5]$$

3. Biological half-life. The quantity of a radioactive isotope introduced into the organism and distributed either uniformly or enriched in some organ or tissue decreases not only because of the physical disintegration, but also as a result of the *biological* elimination characteristic of the isotope as a chemical element. Consequently, if the isotope content of the organism or an organ is measured continuously, a half-life shorter than that expected from the physical decay is obtained. The resultant of the physical decay and the biological decrease is the *effective half-life.* The physical half-life (T_{phys}), the biological half-life (T_{biol}) and the effective half-life (T_{eff}) are connected by the relation

$$\frac{1}{T_{eff}} = \frac{1}{T_{phys}} + \frac{1}{T_{biol}} \qquad [3.6]$$

Since T_{phys} and T_{eff} are directly measurable, T_{biol} can be calculated from [3.6]. The *biological half-life* is defined as the time in which half the quantity of the element or compound in the organism, organ or tissue is eliminated by biological processes.

The biological half-life is independent of whether the isotope is stable or radioactive. Stable and radioactive isotopes behave similarly in the life processes. As an example $^{131}_{53}\mathrm{I}$ has a physical half-life of 8 days, but when present in the thyroid gland it has an effective half-life of 7.5 days. From [3.6] the biological half-life is therefore 120 days. This is the time normally necessary to remove half the iodine present in the thyroid gland. Of course, the eliminated quantity is continuously replaced.

Once in the body, radioactive isotopes may become enriched in certain organs, the *critical organs* (e.g. the liver). Consequently, careful attention must be given to protecting these organs from hazardous radiation effects. The thyroid gland is the critical organ for iodine. The critical organ for $^{51}_{24}\mathrm{Cr}$ is the kidney. $^{24}_{11}\mathrm{Na}$ is fairly evenly distributed in the extracellular fluid, and consequently the whole organism is considered to be critical for this isotope. These or other organs (e.g. the liver) are also considered to be critical, if their functions are especially important for the normal functioning of the whole organism.

4. The activity of a radioactive substance is characterized by the decay rate, which is the decay number per second. The unit of activity is the becquerel (denoted by Bq):[1]

$$1\ \mathrm{Bq} = 1\mathrm{decay/s}, \quad \text{or} \quad 1\ \mathrm{Bq} = 1\ \mathrm{s}^{-1}$$

Since the decay rate is proportional to the number of undecayed atoms present [3.1], the decay law [3.4] may be used for the time course of the activity:

$$A = A_0 e^{-\frac{0.693}{T}t} \qquad [3.7]$$

where A_0 is the initial activity of the preparation, and A is the activity still present after time t.

[1] Previously the curie (denoted by Ci) was used: $1\ \mathrm{Ci} = 3.7 \times 10^{10}$ Bq.

The concept of *specific activity* is frequently used. This denotes the activity as related to unit mass. Its unit is Bq/kg.

In a broader sense the notion of specific activity is associated not only with a pure radioactive substance but also with its mixture with e.g. the same inactive substance (the *carrier substance*[2]); moreover, the specific activity of liquids or tissues is sometimes also used. The activity related to unit volume is the radioactive concentration; its unit is Bq/l.

3.2. Nuclear radiations

Radioactive nuclei may release their excess energy in various ways. In a considerable number of cases this energy or part of it is carried off by some particle. In these cases the nuclear transformation is accompanied by corpuscular radiation. However, the release of excess energy may also produce electromagnetic radiation, and it is an even more frequent occurrence that part of the excess energy is released in the form of corpuscular radiation (of course, the energy released also includes the energy proportional to the mass of escaping particle), and the remaining excess energy is then emitted as electromagnetic radiation. Though not belonging strictly to radioactive phenomena, the processes of *nuclear fission* and *spallation* may be mentioned here. The essence of the former effect is the fission of certain (mainly heavy) nuclei into two parts of comparable mass *(fission products).* Nuclear fission was first observed with $^{235}_{92}U$; besides being radioactive this isotope can capture a slow neutron and then undergoes fission into two medium heavy nuclei, at the same time emitting 2 or 3 neutrons. *Nuclear spallation* is induced by extremely high-energy alpha, deuteron or similar radiation. A nucleus bombarded in this way readily disintegrates into smaller nuclei, nuclear fragments, protons and neutrons.

3.2.1. Alpha-radiation

1. α-decay, α-particles. In the course of α-decay an atomic nucleus releases a high-energy *helium nucleus,* the *α-particle.* As a result, the atomic number of the nucleus decreases by 2, and its mass by 4. For instance:

$$^{226}_{88}Ra \rightarrow {}^{222}_{86}Rn + {}^{4}_{2}\alpha \qquad [3.8]$$

2. The interaction of α-radiation with matter. The initial velocity of α-particles is several thousand km/s, which is equivalent to a kinetic energy of several MeV. The energy of these particles is lost through ionization or excitation of the molecules and atoms of the surrounding medium, for instance the air. The ionizing power of an α-particle is characterized by the *linear ion density (specific ionization)* produced along the particle

[2] In most samples to be administered the mass of the radioactive isotope is extremely small, e.g. the mass of 37 MBq (1 mCi) $^{131}_{53}I$ is 8.1×10^{-9} g, and the mass of $^{24}_{11}Na$ is only 0.113×10^{-9} g. For administration, these small quantities of radioactive compounds are generally mixed with some other substance (usually the same, inactive compound); this is the carrier substance.

path. If the particle produces n ion pairs along a track of length l, the linear ion density is given by the quotient n/l; this is the number of ion pairs produced per unit track length. The linear ion density of α-particles is 20,000–80,000 ion pairs/cm in air at normal atmospheric pressure. The production of one ion pair in air requires an energy of 34 eV. If the velocity of an α-particle decreases to the thermal value it is transformed into a helium atom by the capture of 2 electrons.

The path of an α-particle is straight except for the rare cases when the α-particle interacts not with an electron shell, but with a nucleus, which is small even compared to the electron shell. In this case the α-particle is scattered on the nucleus. The mass of the α-particle is considerably (approximately 7000 times) larger than that of the electron, which explains why a collision with electrons does not influence the direction of the motion of α-particles. The *effective range* R is the distance covered by an α-particle in a medium of density ρ until its energy E decreases to its thermal value. For substances consisting of elements of low atomic numbers (e.g. air, water, the body tissues) the following relation holds to a good approximation:

$$R = k\frac{E^{3/2}}{\rho} \qquad [3.9]$$

If E is measured in MeV, ρ in g cm^{-3} and R in cm, the numerical value of k is 4.15×10^{-4}. For instance, in atmospheric air for ^{226}Ra (E = 4.8 MeV) R = 3.4 cm. In liquids or soft tissues the effective range of an α-particle is 10–100 μm.

The ionizing power of an α-particle is almost constant at the beginning of its path, but towards the end increases about fourfold, and attains the above value of approximately 80,000 ion pairs/cm, from which it falls abruptly to zero as shown in Fig. 3.1. Thus, the linear ion density is smaller with high-energy particles, and increases with decreasing energy.

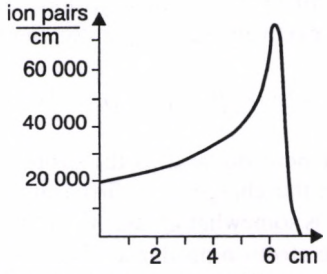


Fig. 3.1. Specific ionization of the ^{214}Po α-particle as a function of its track length (in air)

The behaviour of α-particles in different media is usually characterized by the *stopping power* of the medium, which is defined by the energy loss of the α-particle per unit track length. Instead of the stopping power, the expression *linear energy transfer (LET)* is frequently used. By definition, the stopping power is equal to the product of the linear ion density and the energy required to produce one ion pair. It follows that the stopping power depends upon the energy of the α-particle in the same way as the linear ion density.

If the stopping power is divided by the density of the medium, the *mass stopping power* is obtained; this is the energy loss of the particle after its passage through a layer in which behind a 1 cm^2 surface the unit mass of the medium is situated. In practice mainly the *relative stopping power* is used; this is generally defined as the stopping power related to air at 101 kPa and 15 °C. The advantage of the relative stopping power is that it is practically independent of the particle energy.

α-radiation has a line spectrum, which indicates that a given radionuclide emits only α-particles of a given energy: the particles carry with them from the nucleus only discrete, possible energies. For instance, 93% of the α-particles of radium escape from the nucleus with an energy of 4.8 MeV, while 7% escape with an energy of 4.6 MeV. Every α-particle released from radon has an initial energy of 5.5 MeV. In accordance with the discrete energy values, the effective range too of these particles attains only definite values.

The ionization and excitation produced in a medium may induce various processes. Thus, atoms returning to their ground state emit *characteristic X-radiation.* Luminescent material (e.g. barium platinocyanide, or silver-activated zinc sulphide) produce visible light pulses (scintillations) on collision with α-particles. From a biological aspect the most important fact is that ionization and excitation may induce *chemical processes* resulting in *functional and morphological changes* of the tissues. The majority of the energy absorbed is finally transformed into *heat* in several steps.

There is also a small probability that an α-particle may interact with an atomic nucleus. If the energy of the α-particle is high enough, this interaction may lead to a *nuclear transformation.*

3.2.2. Beta-radiation

1. β-decay. In the course of β-decay a *negative* electron or a *positive* electron *(positron)* is emitted from the nucleus. Electrons are not present in the nucleus, and accordingly β-decay requires some explanation. In processes within the nucleus one neutron may be transformed into one proton and one electron, or one proton may be transformed into one neutron and one positron. The nuclear symbols for the two kinds of β-particles are ${}^{0}_{-1}\beta$, β^-, β and ${}^{0}_{+1}\beta$, β^+ respectively. Consequently, the transformations within a nucleus are

$$ {}^{1}_{0}n \rightarrow {}^{1}_{1}p + {}^{0}_{-1}\beta \quad \text{and} \quad {}^{1}_{1}p \rightarrow {}^{1}_{0}n + {}^{0}_{+1}\beta \qquad [3.10] $$

The symbol ${}^{1}_{0}n$ denotes the neutron, and ${}^{1}_{1}p$ the proton. The upper index is the mass number, and the lower index the charge. The first transformation is easy to interpret, since the mass of the neutron is somewhat greater than that of the proton (cf. also [3.15]) and can cover the mass of the electron produced. The transformation of a proton into a neutron may be explained in that part of the excess energy of the nucleus is devoted to ensuring the higher mass of the neutron (cf. [3.15]).

Negative β-decay is produced whenever more neutrons are in the nucleus than the number required to maintain stability; in positive β-decay the number of neutrons is smaller than required. In the case of negative β-decay the atomic number of the nucleus clearly increases by 1, whereas in the case of positive β-decay it decreases by 1. The mass number remains the same in both cases. Negative β-decay is observed in both natural and artificial radioactive isotopes, but positron radiation is found only with artificial radioactive isotopes. As an example, the two radioactive isotopes of phosphorus may be mentioned:

$$^{32}_{15}P \rightarrow {}^{32}_{16}S + {}^{0}_{-1}\beta \quad \text{and} \quad {}^{30}_{15}P \rightarrow {}^{30}_{14}Si + {}^{0}_{+1}\beta \qquad [3.11]$$

2. Inverse β-decay. A nucleus with excess protons may decrease its positive charge not only by positron emission, but also by the capture of an electron from an inner shell, mainly from the *K* shell. As a result of the capture one proton is transformed into one neutron. This process results in the decrease of the atomic number of the nucleus by 1, while the mass number remains unchanged (similarly as in positron emission). This process is also called *shell electron capture,* or *K capture. K* capture occurs within the atom, and cannot be observed directly. However, when an electron jumps from one of the outer shells into the hole resulting from *K* capture, characteristic X-radiation is produced, which thus gives information about the *K* capture. The symbol of *K* capture is *K*.

3. β-radiation. Neutrino. The energy distribution of the β-radiation of an isotope is continuous: the energy of β-particles varies from zero up to the maximum value. However, it is surprising that the energy loss of a disintegrating nucleus is the same in all cases. This phenomenon is explained in that the β-particle is not emitted alone, but in the company of a neutral particle whose mass is a thousand times smaller than that of the electron; this particle is the *neutrino.* In the decay the two particles carry with them an identical overall energy, but this energy can be distributed between them in an infinite number of ways. If the electron acquires the total energy, a β-particle of maximum energy is ejected. The various isotope tables contain the maximum energy (E_{max}) of β-radiation. For instance, E_{max} is 1.7 MeV for the β-radiation of $^{32}_{15}P$. The energy spectrum of this isotope is depicted in Fig. 3.2. The lower-energy (softer) components of the β-radiation of an isotope obviously produce a smaller effect in any substance, or in the living organism, than higher-energy components. For an estimate of the expected effect, calculations are made only with an *average energy* (for $^{32}_{15}P$ this is 0.68 MeV), as if every electron had this same energy value.

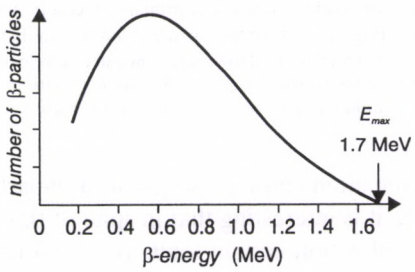


Fig 3.2. The β-spectrum of $^{32}_{15}P$

The abscissa gives the initial energy of the β-particles, and the ordinate the number of β-particles per unit energy interval. The maximum of the curve is at about 0.51 MeV. In the decay of $^{32}_{15}P$, β-particles of this energy are produced with the highest probability. In the event of a symmetric curve, 0.51 MeV would be the average energy of the β-particles. Since the number or higher-energy particles is greater, the average energy is higher: 0.68 MeV

The initial velocity of the β-particle approaches the velocity of light, and for this reason the relativistic mass increase must also be taken into account.

4. Interaction of β-radiation with matter. Because of the small mass of the electron and as a result of collisions and scattering, the track of a β-particle is rather zigzagged (in contrast to α-particles, β-particles are scattered by electrons). The degree of scattering may be even larger than 90° *(back-scattering)*. This circumstance must be considered in measuring technique as well as in radiation protection. Because of the continuous energy distribution, no uniform effective range exists; the effective range extends from 10 cm to several m in air, but is only a few mm in water and living tissues. The *specific ionizing power* of the β-particle is approximately 1000 times smaller than that of the α-particle. Naturally, the *stopping powers* of the various substances are smaller by the same order of magnitude for β-particles than for α-particles. Instead of the *stopping power,* the expression *linear energy transfer* (*LET*) is also used in the present case. For a velocity v the specific ionization in air is given by

$$s = \kappa \left[\frac{c}{v}\right]^2 \qquad [3.12]$$

where $c = 3.0 \times 10^{10}$ cm s^{-1} and $\kappa = 46$ ion pairs/cm. In the case of high velocities ($v \approx c$), the specific ionization approaches the value of κ, but for lower velocities s is larger than κ.

Similarly to α-radiation, when β-radiation passes through various media it produces not only excitation and ionization, but also (as a consequence) chemical, photochemical, biological, etc. effects. *Characteristic X-radiation* too is generated, and (rather rarely) β-particles may suddenly be stopped in the field of an atom. When this occurs, *Bremsstrahlung* is produced.

An interesting phenomenon, *Cherenkov radiation,* should be mentioned here. Cherenkov radiation is observed if high-energy β-particles or other charged particles (accelerated electrons, protons, pions, etc.) interact with some medium. Cherenkov radiation is generated whenever a charged particle moves in some medium (e.g. water) with a velocity larger than that of light in the same medium. The primary process in this case too is the excitation of the atoms or molecules of the medium. If the velocity of the particle is lower than that of light, the light waves induced by excitation extinguish each other by interference. However, if the particle velocity is higher than that of light, the extinction is not total. The remaining bluish-white light is Cherenkov radiation. Similarly to other effects of the charged particle radiations this radiation too may be used to detect the (high-velocity) particles, to count them, to measure their velocity, etc.

The process leading to the attenuation (absorption) of β-particles is thus an extremely complex one. Nevertheless, it is interesting that in spite of this variety the absorption of β-radiation can be described within certain limits (cf. sections 2.3.1 and 2.10.1) by the equation

$$I = I_0 e^{-\mu x} \qquad [3.13]$$

For media consisting of elements of low atomic number, μ is linearly proportional to the density of the medium for a given radiation. Consequently, similarly as for X-radiation it is convenient to express the layer thickness not in cm or in mm, but in g cm^{-2} or mg cm^{-2}. In this way μ becomes independent of the nature of the substance, and depends only on the energy of radiation. Besides of μ, the *half-value thickness* D is frequently used in practice; this is given in units of g cm^{-2} or mg cm^{-2} instead of cm or mm. As an example, the half-value thickness of $^{32}_{15}$P β-radiation is approximately 110 mg cm^{-2}, which means

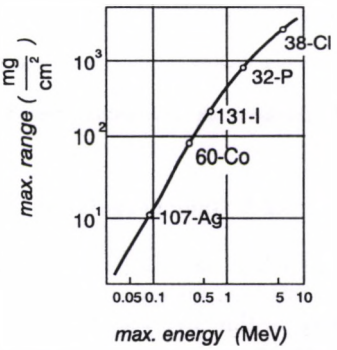


Fig. 3.3. Maximum range of β-radiation as a function of the maximum energy
Logarithmic scale on both axes

that a 0.4 mm thick layer of aluminum with a density of 2.7 g cm^{-3}, or a 1.1 mm thick water layer, is required to reduce the intensity of radiation by half.

The above exponential law holds only for 3–4 half-value thicknesses; after this the intensity decreases more rapidly, and at 7–8 half-value thicknesses the intensity is suddenly extinguished (maximum range). The maximum ranges of the β-radiation of some isotopes are presented in Fig. 3.3. To a first approximation the following equation holds between the maximum energy (E_{max}) and the maximum range (R_{max}) in the range 0.8–3.0 MeV:

$$R_{max} = aE_{max} - b \qquad [3.14]$$

If E_{max} is expressed in MeV and R_{max} in mg cm^{-2}, the numerical value of a is 542, while b is 133.

There is no essential difference between negative and positive β-particles with regard to ionization, excitation, scattering and absorption. However, the lifetime of a positron is extremely short. It interacts with a negative electron (with great probability towards the end of its track, when most of its kinetic energy has been lost), and on encounter their charges neutralize each other, and the two particles are annihilated as photons (γ-photons). The mass of the two interacting electrons corresponds to an energy of 1.02 MeV and this results in two annihilating γ-photons with 0.51 MeV energy propagating in opposite directions. In the case of positron radiation, particle annihilation always occurs, and consequently the presence of γ-radiation must always be taken into account in measurement technique and radiation protection.

It has already been mentioned that the continuous β-spectrum can be interpreted by assuming the existence of the *neutrino*. Since every β-decay is accompanied by the simultaneous appearance of a neutrino, all β-radiating isotopes are also strong neutrino-emitters. The neutrino may interact with a neutron or a proton, but the probability of this is extremely small. For instance, if 10^{12} neutrinos pass through the Earth, on average only one of them interacts with either a neutron or a proton. Thus the neutrino may cover a very large distance without transferring any energy to the medium through which it passes. For this reason their presence is neglected in medical isotope diagnostics and measurement techniques.

If the neutrino too is taken into consideration, the processes in [3.10] may be written in a more correct form:

$$n \rightarrow p + \beta^- + \bar{\upsilon};\ p \rightarrow n + \beta^+ + \upsilon \qquad [3.15]$$

It should be remembered that negative β-decay produces an antineutrino ($\bar{\upsilon}$), whereas positive β-decay a common neutrino (υ).

3.2.3. Gamma-radiation

1. Prompt γ-radiation. After a nucleus has released a particle, it frequently still has some excess energy. In this case the resulting nucleus remains in an excited state. The excited nucleus releases its excess energy within a very short time (10^{-13}–10^{-18} s) in one or several steps by emitting γ-radiation. By this means the excited state is converted into a stable one. γ-radiation generally does not occur spontaneously: it is rather an effect accompanying some sort of corpuscular radiation. Independent γ-radiation is found only in rare isomeric transitions (see below).

γ-radiation is not accompanied by any change in atomic number or mass number. Let us consider as an example the decay of $^{198}_{79}$Au, which emits negative β-radiation with a half-life of 2.89 days, and a maximum β-energy of 0.96 MeV. The $^{198}_{80}$Hg nucleus formed immediately after the β-decay still possesses an excess energy of 0.41 MeV, which is released by the ejection of a single γ-photon. After this the nuclear derivative is stable. The γ-radiation is emitted by the excited nuclear derivative and not by the original isotope, yet in practical terms the prompt γ-radiation is ascribed to the primary nucleus, and $^{198}_{79}$Au is said to be an isotope emitting β- and γ-radiation. Isotopes which emit γ-radiation besides some particle are *mixed-radiating* isotopes, in contrast to the *purely* α- or β-*radiating* ones.

2. Isomeric transition. Some radioactive nuclei do not radiate their excess energy as γ-photons immediately after the particle emission, and the nuclear derivative remains for a relatively long time ($> 10^{-10}$ s) in an excited state, returning to the ground state only with a definite half-life. Such a transformation takes place e.g. in the following process:

$$^{99}_{42}\mathrm{Mo} \xrightarrow[\beta^-]{66\ \text{hours}} {}^{99\mathrm{m}}_{43}\mathrm{Tc} \xrightarrow[\gamma]{6\ \text{hours}} {}^{99}_{43}\mathrm{Tc}$$

Above the arrows the half-lives, below them the type of decay are indicated. The $^{99\mathrm{m}}_{43}$Tc isotope, an intermediate product of the reaction, is transformed into a stable technetium isotope with a half-life of 6 hours by γ-emission. Thus technetium nuclei may exist simultaneously in excited and stable state for a well measurable time. The excited nucleus is called an *isomer* of the nucleus in the ground state and the phenomenon is *nuclear isomerism*. The isomeric (metastable) nucleus is denoted by the letter m beside the mass number.

The $^{99\mathrm{m}}_{43}$Tc isotope is used in medical practice, since its properties are advantageous in several respects (cf. section 3.5.2). Here it only has to be noted that the lifetime of $^{99\mathrm{m}}_{43}$Tc nuclei is long enough to ensure the practically usable quantity of active technetium nuclei in a technetium preparation separated chemically from molybdenum. Thus, one has a

method to produce a substance which emits only γ-radiation. With the above outlined *technetium generator* newly produced isotope may be obtained at the site of applications daily several times.

Cases are known when the radioactive decay of an excited nucleus follows the primary emission of γ-photons.

3. Interaction of γ-radiation with matter. The nature of γ-radiation is similar to that of X-radiation, and thus their effects are essentially the same (cf. section 2.10).

Mention may be made of the *internal photoeffect.* With certain atom types the γ-photon emitted by an excited nucleus may produce photoeffect in the *same atom,* i.e. the excitation energy is captured by a shell electron which can then escape from the atom. In these cases the escaping electron, called a *conversion electron* (and the X-radiation produced) is observed instead of γ-radiation. The conversion electron is generally ejected from the *K* shell; its symbol in nuclear reactions is e^-.

The data discussed in connection with the attenuation (absorption) of X-radiation in section 2.10 also relate to γ-radiation. Some additional findings may also be dealt with. Figure 3.4 depicts the dependence of the mass attenuation (absorption) constant for water on the photon energy (or the wavelength). Conclusions can be drawn from the curves about the absorption by the soft tissues. From a practical aspect the following data are of interest: the half-value thickness of lead upon γ-radiation is ~7 mm for $^{226}_{88}$Ra, ~14 mm for $^{24}_{11}$Na, ~3 mm for $^{131}_{53}$I and ~6 mm for $^{137}_{55}$Cs. Depending upon the wavelength, the half-value thickness of air is several hundred m, and that of the living organism several dm.

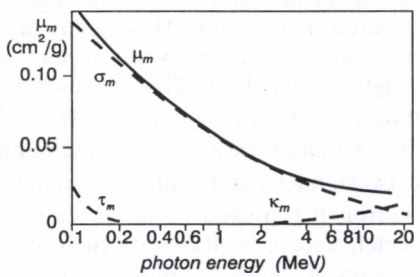


Fig. 3.4. Total and partial attenuation coefficients of water for X-radiation and γ-radiation

3.2.4. Neutron and proton radiation

1. Neutron radiation is produced by certain nuclear processes, mainly by bombardment of the nucleus with appropriate particles (including photons). The resulting highly-excited nucleus emits a neutron. The capture of the particle and ejection of a neutron occur within a very short time (10^{-15}–10^{-18} s). (Only a small number of nuclear processes are known in which the neutron emission occurs with a measurable half-life.) As an example, the bombardment of $^{9}_{4}$Be with γ-photons may be mentioned. Following γ-energy absorption, the nucleus emits one neutron and is transformed into the $^{8}_{4}$Be nucleus:

$$^{9}_{4}\mathrm{Be} + h\nu \rightarrow {}^{8}_{4}\mathrm{Be} + {}^{1}_{0}n$$

or in brief:

$$^{9}_{4}\mathrm{Be}(\gamma, n)\,{}^{8}_{4}\mathrm{Be}$$

The first symbol in the bracket indicates the bombarding species entering into the nucleus, while the second refers to the ejected particle.

The *free neutron* is an *unstable product,* with a half-life of approximately 13 minutes; it disintegrates into a proton and an electron.

Since the neutron has no charge, it *does not* cause *direct ionization.* On passing through some medium it interacts only with the nuclei, and not with the electrons. In this respect, two processes are of interest: *neutron scattering,* and the *production of nuclear reactions.*

Neutrons are generally ejected from the site of their generation with high energy (a few MeV). During their progress, however, they transfer part of their energy by scattering to the atomic nuclei, and continue travelling in a changed direction. In the case of *elastic scattering,* the nucleus takes up in the form of kinetic energy the kinetic energy lost by the neutron in the scattering process. The energy transfer is larger if the masses of the colliding particles are comparable. Consequently, the energy loss of fast neutrons affects mainly light nuclei, e.g. $^{1}_{1}\mathrm{H}$, $^{2}_{1}\mathrm{H}$. The biological effect of neutrons is due essentially to the struck and ionizing (exciting) protons. In the event of a central collision with a proton (similarly to the collision of a moving billiard ball with another ball at rest), the neutron may lose its total kinetic energy in one step, after which it maintains only thermal motion with an energy of 0.1–0.01 eV *(thermal neutron).*

On colliding with nuclei, neutrons (and mainly the high-energy neutrons, the fast neutrons) may also be scattered inelastically. This means the transfer not only of the kinetic energy, but also of additional energy to the nucleus, which results in the promotion of the nucleus to a higher energy level. The nucleus becomes excited, and releases its excess energy by emitting one or more γ-photons. Thus, the inelastic scattering of neutrons is accompanied by γ-radiation emitted by the excited nuclei.

The other interaction between nuclei and neutrons is manifested in the production of nuclear reactions. The neutron has proved to be an extremely suitable bombarding particle in transforming nuclei, since it is electrically neutral, and easily penetrates into the nuclear force field. The neutron is captured by the nucleus, which generally ejects one proton or one γ-photon. More rarely, after capturing a fast neutron, the nucleus may eject another neutron, or possibly even two. Neutron capture is sometimes followed by the ejection of an α-particle.

Both scattering and nuclear reactions result in the attenuation of neutron radiation. Considering the low probability with which a nucleus captures a fast neutron as compared to a slow (thermal) neutron, in practice fast neutrons become decelerated in a series of elastic and inelastic scattering steps, and the resulting slow neutron finally produces the nuclear reaction. The attenuation of neutron radiation is described by the well-known equation

$$I = I_0 e^{-\mu x} \qquad [3.16]$$

Here too the attenuation constant depends on the material of the absorbing medium and on the energy of the neutrons.

2. Proton radiation is produced either by the acceleration of hydrogen ions or (similarly to neutron radiation) by bombardment of the nucleus with some particle (including photons). In the latter case the capture of the missile and the ejection of the proton generally follow each other within a very short time. Like all other particles bearing an electric charge, the proton causes ionization and excitation as it passes through a medium. The linear ion density is smaller than for α-particles, but larger than for electrons.

3.2.5. Cosmic radiation

The *cosmic radiation* arriving from the Universe can be divided into two groups. One group consists of *primary radiation* at altitudes above 25–30 km, and comprises mainly (approximately 91%) protons and to a smaller extent (ca. 8%) α-particles; more complex nuclei (up to nickel) also occur in traces. The mean energy of these particles is very high (of the order of 10 GeV), but particles with energies of even 10^{14}–10^{19} eV are found too. (The largest accelerator constructed to date produces protons of approximately 500 GeV.)

The primary particles arriving in the atmosphere of the Earth interact with the atmospheric atoms, and in doing so induce various processes (secondary radiation). The atmospheric atomic nuclei may undergo fission, some of them are transformed into radioactive nuclei, neutrons are emitted, and γ-radiation, electron pairs and various neutral and charged particles of short lifetimes (10^{-16}–10^{-6} s) are produced, which in turn initiate further processes. The masses of some of the short-lifetime particles are greater than the mass of the electron, but smaller than that of the proton (mesons), whereas the masses of others lie between the mass of the proton and that of the deuteron (these are the hyperons).

The secondary cosmic radiation consists of soft (easily absorbed) components and hard components with a considerable penetrating power. The soft components are mostly electrons and photons, and the harder ones are mainly mesons. For instance, more than 10 mesons per second pass through the human body. The soft components are absorbed by a lead layer a few cm thick, whereas the hard components easily pass through a 1.0 m thick lead wall.

It is currently assumed that *cosmic radiation originates in the supernovae,* stars of varying brightness. These suddenly increase enormously in brightness, while their internal temperature attains a value of 10^9 K The high-energy protons and other particles emitted race throughout the Universe for millions of years before reaching some celestial body, e.g. the Earth.

3.2.6. Particle accelerators in medicine

Charged particles (electrons, protons, deuterons, helium ions and other ions) are accelerated in an electric field. The high-energy particles obtained are utilized in various ways in both medical practice and research. In recent years different types of accelerators have been developed, mainly for use in nuclear physics. Of these, however, only the basic types which are of interest in medicine will be described.

1. In the cyclotron the particles (mainly ions) are accelerated by a voltage of several hundred thousand volts, but they can traverse the accelerating space approximately 200 times and accumulate excess energy with every turn.

An essential part of the equipment is a flat, cylindrical metal box (with a diameter of several m), which is cut in two halves along its diameter (Fig. 3.5). The accelerating voltage is applied to the two parts of the box, called dees (*D*-electrodes). An electric field is produced practically only in the slit between the electrodes, the field intensity within the boxes being zero. The ions to be accelerated are produced by the ion source protruding into the centre of the box. Within this box the ions move with increasing velocity along a spiral path. Acceleration is provided whenever the ions pass through the slit. This occurs twice per rotation, in opposite directions. In order to induce

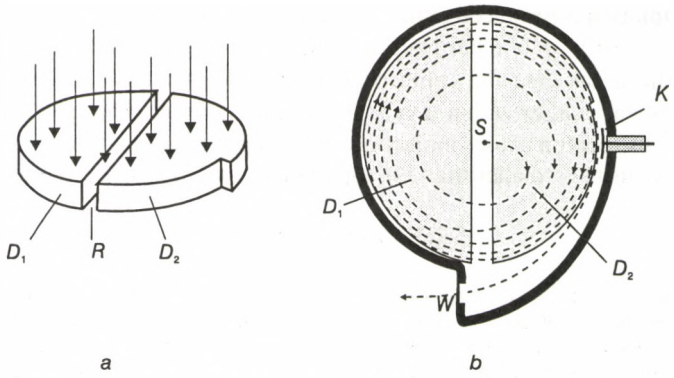


Fig. 3.5. Principle of operation of the cyclotron
a: the dees (D_1 and D_2) and the magnetic lines of force; *R* is the slit between the dees;
b: the dees viewed from above with the deflecting electrode *K*; the helical line is the path of the ions starting from the source *S*; *W* is the exit window

an acceleration in every period, an alternating field is required. Consequently, an alternating voltage is applied to the dees. The curvature of the path is ensured by a magnetic field, whose lines of force are perpendicular to the plane of the box. With increasing velocity the radius of the orbital path increases too; this is the reason for the spiral particle orbit. The accelerated ions are deviated from their orbit after passing the electrode *K*, escape through the window *W* and hit the target.

With the first cyclotrons, protons could be accelerated to an energy of 20 MeV. In the variants developed later, such as the synchrocyclotron, the synchrophasotron, etc., protons with an energy several orders of magnitude higher (up to a few GeV) may be obtained. For medical purposes, mainly small and medium cyclotrons (up to approximately 50 MeV) are used.

The importance of these cyclotrons lies in the fact that the accelerated particles can be employed to initiate a large number of nuclear reactions providing radioisotopes with nuclear parameters favourable for tracing purposes (for use in both research and diagnostics, cf. section 3.5). The use of *isotopes with shorter half-life* (^{123}I, ^{11}C, ^{81}Rb, etc.) is advantageous in medical practice, since with these the dose exposure is smaller (cf. section 3.5, point 2), several examinations can be carried out in a short time, the medical check-up is simplified, and so on.

On-site isotope production is especially favourable with elements possessing only isotopes with very short half-lives, for instance oxygen and nitrogen, which are important in some functional examinations (the isotope parameters are listed in Table 9.11).

These ideas are obviously also fundamental in scientific research.

In connection with the medical applications of cyclotrons, it should be emphasized that some of the short half-life isotopes produced by these accelerators emit only γ-rays, which are obviously more favourable from the viewpoint of dose exposure than those isotopes of the same element which emit mixed radiation. The possibility of producing short half-life positron-emitting isotopes should also be borne in mind. These isotopes allow very exact localization in certain diagnostic examinations.

The cyclotrons (and their developed variants) may become an important tool in therapy. As a result of recent investigations it is expected that besides X-radiation, γ-radiation and accelerated electrons (not to mention α- and β-radiation), high-energy protons and heavier charged particles (atomic nuclei, fission products), neutrons and pions (cf. section 1.1) will also be of therapeutic use.

Appropriately accelerated *protons* and *heavier particles* are obtained directly from the accelerator. *Neutrons,* on the other hand, are usually produced by the impact of accelerated deuterons on a beryllium target according to the following nuclear reaction:

$$^{9}_{4}\mathrm{Be}\ (d, n)\ ^{10}_{5}\mathrm{B} \qquad [3.17]$$

In this way, neutrons of practically any desired energy may be obtained. Therapeutically, fast neutrons (>30 MeV) are of considerable importance because of their great biological effectiveness (cf. section 3.6).

Pions are produced by the collision with atomic nuclei of protons accelerated to an energy of several hundred MeV. Positive, negative and neutral pions exist, though therapeutically only negative pions are of interest, since they are captured with high probability by positively charged atomic nuclei. As a result of the large energy uptake, these nuclei explode and the therapeutic effect is induced by the fission products.

The so-called *synchrotron radiation* should be mentioned here which is becoming more and more important in various areas of the scientific research. The primary phenomenon: every charged particle having a curvilinear path emits *electromagnetic radiation.* This phenomenon is especially marked in the cyclotron-type accelerators (as their name indicates), in which the particles move on a circular path. The particles accelerated to a velocity close to that of the light are directed into a ring-shaped vacuum chamber, the so-called storage ring, then the arising electromagnetic radiation (Fig. 3.6) is made to leave the ring for further use. The radiation appears at the cost of the energy of the particle, tangentionally to its path, in the form of a narrow beam.

For the production of high-intensity radiation accelerated *electrons* are used. The power of the radiation is in direct proportion to the intensity of the electron current, to the fourth power of the kinetic energy of the electron, and in inverse proportion to the radius of the curvature of the path. The spectrum of the radiation is continuous, and with increasing electron energy the emission maximum is shifted towards the shorter wavelengths. Thus in case of electrons with sufficiently high velocity every wavelength may be found in the spectrum of the synchrotron radiation, from the radio waves through the visible light to the hard X-rays. The intensity of the synchrotron radiation in the X-ray region can be larger by several orders of magnitude than that of the radiation produced by a conventional X-ray tube. This may be very favourable in various cases, for instance a diffraction picture made to explore the molecular structure may be obtained within seconds instead of hours. Aside from its high intensity, the natural alignment of the synchrotron radiation presents favourable conditions for the production of nearly monochromatic radiation with an appropriate intensity, if necessary. In a given case the high polarization of the radiation may be advantageous.

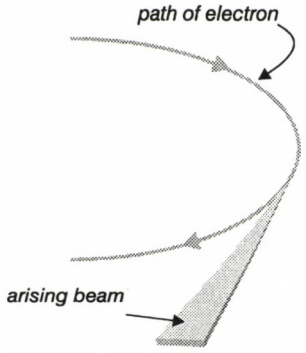


Fig. 3.6. Synchrotron radiation

2. Linear accelerators accelerate ions as well as electrons; with these equipments it became possible to produce electrons of many GeV, and protons of a few GeV. (The accelerated electrons may be used subsequently to produce hard X-rays.) Differently from cyclotrons, here the particles are accelerated on a linear path, hence the name: linear accelerator. Their advantage, as opposed to accelerators of other type, is the production of beams of relatively large intensity, and homogeneous irradiation fields which is favourable concerning their therapeutic application.

Figure 3.7 demonstrates an arrangement in which the path leads within a sequence of coaxially placed, tube-like electrodes working with an alternating current of properly selected high frequency. In these equipments – similarly to cyclotrons – electric field is produced only in the gaps between the electrodes, within the cylinders the electric field is zero. The particles are accelerated only when passing through the gaps. The motion of the particles as well as the length of the cylinders and the frequency of the alternating current must be co-ordinated so that the most favourable part of the period of the alternating field (concerning direction and field strength) should affect the particles passing through the gaps (resonance method). By increasing the velocity the path length of the particles will be increasingly longer, which must be taken into account by increasing the length of the electrodes, which is shown in the figure.

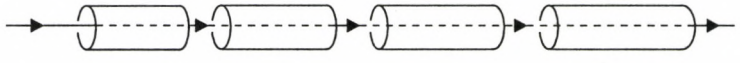


Fig. 3.7. Diagram of a linear resonance accelerator with cylindric electrodes
The dashed line with arrows denotes the path of the particles

Every accelerator belongs to the so-called large equipments, whose operation and maintenance require a specially trained technical personnel and specially constructed laboratories.

3.3. Measurement of nuclear radiations. Dosimetry

Various methods have been developed to detect and measure nuclear radiations. These allow selection of the most suitable procedure in every case. The methods to be outlined are also applicable to the measurement of X-radiation.

The measurement of radiation is always based on the interaction between the radiation and some substance, the medium (detector material). Radiation transfers some of its energy to the medium. Radiation passing through without energy loss does not leave a trace in the medium. In the case of radiation of electrically charged particles (e.g. α- or β-radiation) the primary phenomena consist of ionization and excitation, followed by secondary, tertiary, etc. processes in the various substances. Such processes are thermal effects, luminescence, photochemical processes, and so on. In some cases the ionization, and in others the secondary, tertiary, etc. effects are used to measure the radiation. Electrically neutral radiation (γ- and neutron radiation) first produces electrically charged particles in the detector, and these subsequently display the effects already discussed. Thus, in the measurement of γ-radiation the photoelectrons, Compton electrons and electron–positron pairs are utilized, while in the case of neutron radiation the neutron-struck protons or the charged particles (e.g. α-particles) produced by neutron-induced nuclear reactions are used. Radiation consisting of charged particles is frequently referred to as *directly ionizing,* whereas radiation of neutral particles is *indirectly ionizing* radiation.

Of the large variety of types of nuclear radiation, γ- and β-radiation are used most frequently in medical practice and in biological investigations. Accordingly, stress will be laid on some possibilities of their measurements.

3.3.1 Possibilities of measurement

1. Measuring devices based on gas ionization. One large group of measuring instruments is that of *ionization chambers, proportional counters,* and *Geiger–Müller tubes.* In this family of instruments charged particles move in an electric field and collide with gas molecules, which become ionized. The ions or ion pairs produced are accelerated by an electric field in the tube or chamber. Depending on the sign of their charge, the ionized particles move to the positive or negative electrode of the apparatus (Fig. 3.8). The

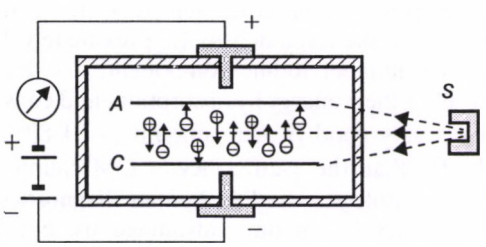


Fig. 3.8. Outline of operation of an ionization chamber
The small circles between the electrodes (*A* and *C*) denote the charge carriers produced by radiation. *S* is the source (e.g. γ-source)

measurement of the radiation is based on measurement of the ionization current. In ionization chambers the total ionization produced by a large number of ionizing particles is generally measured, whereas the Geiger–Müller counter (abbreviated to GM counter) is used to count the number of current or voltage pulses (particle counting). The proportional counter allows not only counting of the particles, but also their distinction according to type and energy. (The ionization current is proportional to the energy of the particle stopped completely in the gas-filled chamber.)

Ionization chambers are used especially frequently at present in individual radiation protection. They are easily manageable devices (e.g. pocket dosimeters of the size and shape of a fountain pen); they can be fastened on the clothing, which permits easy control of the dose obtained during work by personnel handling radiating substances. In principle these chambers are charged electrometers, which progressively become discharged in response to the radiation. The dosimeters for calibration purposes and thimble chambers frequently used in radiological practice are also ionization chambers (cf. section 3.3.4).

GM tubes are mainly used to detect β-particles. Tubes made of some suitable material and filled with some appropriate gas are also applied to count neutrons. For γ- and X-ray photons, especially high-energy ones, the efficiency of GM tubes is only 0.1%, and for this reason they are used only exceptionally in these cases.

Recently the multifibre (multianodic) panel variation of proportional counters, the so-called *proportional chamber* has been developed which is becoming widely used in medicine, in connection with both γ- and X-radiation.

2. Detectors based on luminescence. These detectors are usually made of inorganic or organic crystalline materials, but fluid luminophores are also used. Every charged particle striking the detector produces a scintillation (light pulse). The total light emitted is usually measured, and from the intensity of light the intensity of the nuclear radiation is inferred. Another method of application is to count the individual light pulses and hence to determine the activity of the radioactive sample. In either light intensity measurements or the scintillation counting method, light is first converted into an electric signal by a photomultiplier; this signal is then amplified and processed (cf. section 6.6.2).

The measurement of γ-radiation is being frequently applied in medical practice. For this purpose mainly scintillation counters are used; thallium-doped (activated) NaI crystals a few cm in size are generally built in as detectors. By means of well-known effects (cf. section 2.10.2), on striking the large density (approximately 2.5 g cm^{-3}) NaI crystals containing the high atomic number iodine component, it is highly probable that the γ-photons lose some or all of their energy by absorption, and consequently the efficiency of the photon counting is very good (it may amount to 60–80%). In fact, it is several orders of magnitude better than the γ-efficiency of GM counters. In clinical practice, where the quantity of radioisotopes used in human diagnostics should be as low as possible, the large γ-efficiency is not only advantageous, but is considered to be a requirement for protection. The unit consisting of the scintillation detector, the photomultiplier and the first electronic amplifier is the *scintillation head.* Figure 3.9 outlines the construction of the detector and the photomultiplier.

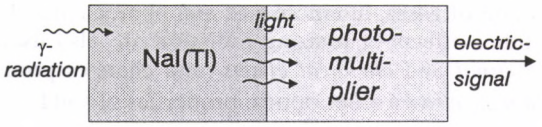


Fig. 3.9. Outline of scintillation head for γ-counting

A particle stopped in the scintillation detector (for instance a photoelectron produced by a γ-photon) interacts several times with the detector material until its total energy is lost. As a result, one particle creates several luminescent photons. The average number of these photons is proportional to the energy of the interacting particle; this greatly enhances the applicability of luminescence detectors, since from the magnitude of the light pulse produced and after conversion from the magnitude of the electric pulse, the energy of the particle can be determined. If the particle is, for instance, a photoelectron produced by a γ-photon, the energy of this γ-photon can be determined from the generated electric pulses. The scintillation technique allows the determination of the γ-spectrum of the radioactive preparation, and hence the identification, control and determination of its composition.

β-particles are detected with anthracene crystals or detectors made of plastic (plastic phosphors). Fast neutrons are usually detected with some plastic substance rich in hydrogen; slow neutrons are measured with substances containing boron, and α-particles are most suitably detected with thin layers of zinc sulphide doped with silver.

The measurement of β-particles of very low energies (e.g. the maximum of the β-particle energy of tritium is only 0.0183 MeV) is most conveniently carried out with *liquid scintillators.* In this method the sample to be measured is mixed with the scintillator fluid, and the mixture is placed on the window of a photomultiplier.

High-energy charged particles are also measured by means of *Cherenkov radiation* (cf. section 3.2.2). As a medium, generally distilled water is used and the light signals are processed in the same way as in the former methods.

3. Photochemical measuring devices. The effects produced on photographic plates are frequently used to detect and measure radioactive radiation. This method is suitable for study of the total emission, i.e. the *total amount of radiation,* and also the *individual particles.* In the former case the darkening of the emulsion due to radiation is determined by means of a photometer (densitometer), and hence the energy falling on unit surface of the emulsion is deduced. In the latter case the tracks of the individual ionizing particles in the emulsion are studied microscopically, and as a result the type, energy, etc. of the particle are determined.

The *photoemulsion method* in many instances allows a relatively exact determination of the location and distribution of the radioactive substances. This may provide valuable information in biological and histologic studies. While measurements with the GM counter or the scintillation technique yield information of the total isotope content of some larger size tissue or organ, the latter method, *autoradiography,* permits the localization of the isotope within a given organ, cell cluster, or even individual cells (cf. Picture 3.1. in the Supplement). In a frequently used procedure 5–20 μm thick sections are made from the biological object to be examined and are then contacted with a photosensitive emulsion (e.g. by smearing a 1–2 μm thick emulsion on the section). After a suitable exposure time and subsequent development, depending upon the concentration of the isotope, dark spots of different sizes are observed. The autoradiogram provides the more information, the better it reflects the distribution of the isotope in the histologic section. With the most advanced methods, even spots only 1–2 μm apart from each other can be distinguished.

Besides the ionization of gases, luminescence and photochemical effects produced in certain substances, other effects of ionizing radiation are also used for detection and measurement, e.g. *thermal* and *chemical effects,* and changes produced in the *electric conductivity, absorption spectra* or other optical properties of solids (cf. section 3.3.4).

3.3.2 Dosimetry

This section will deal with those problems of radiation measurement which are important because of biological effects.

Ionizing radiation initiates in the human body destructive processes. However, the damage is either manifested at once, or after some time, possibly years or even decades later. Consequently, quantities must be found which permit the characterization of various types of radiation from the aspect of their expected biological effects, so that these may be inferred in advance. Such problems are dealt with in *dosimetry.*

The concept of dose has been adopted from pharmacology, where it means the drug quantity administered into the organism in various ways. More exactly, the term dose denotes the administered drug quantity per weight or mass unit. The expression *radiation dose* refers quite generally to biologically effective quantities taken by the organism and related to mass (or volume) units. There is a considerable difference between the two concepts of the dose. In pharmacology the total drug quantity administered is considered, regardless of the effective quantity taken by the organism, and also of the quantity taken by the organism but secreted without exerting any effect. With radiation, however, the situation is different, since in radiodosimetry only the effective quantity is considered, i.e. only the energy absorbed by the body or organism. Radiation passing through the organism without any interaction is not taken into consideration. After these preliminaries the basic objective of dosimetry, more exactly of *physical radiodosimetry* may be formulated as the *determination of the energy absorbed by the tissues in a certain region.* Frequently also the opposite task arises: the unknown dose of exposure must be inferred from well-measurable, statistically evaluable biological changes. This is the domain of *biological dosimetry.*

It is important to emphasize that the absorbing region of the organism, which may be located in various depths, may contain different tissues. Because of the scattered radiation, surroundings of these regions are also considered. In several cases (for instance in case of the therapeutic application of the radiations) the dose must be known from area to area and almost from point to point, and the measurements are carried out only where this is actually necessary and justified by the circumstances. (In other cases, e.g. in radiation protection, it is sufficient to know the average dose for a given tissue or organ.)

It should also be taken into consideration that, though the knowledge of the absorption of biologically effective radiation is of basic importance, it is in itself not sufficient, because other physical, chemical and biological factors are also important in the development of the biological effects.

It clearly follows that the measurement of *radiodoses* requires *special* efforts, and the solution of the problems involves new concepts and measuring techniques. Many problems have still not been satisfactorily solved, and a more exact knowledge of the physical, chemical and biological effects will probably lead to further development in this field.

3.3.3. Dose concepts

1. Absorbed dose. Denote the energy absorbed by some mass dm ($= \rho\, dV$) of the body by dE. The absorbed dose in this region is given by

$$D = \frac{dE}{dm} = \frac{1}{\rho}\,\frac{dE}{dV} \qquad [3.18]$$

The numerical value of D is equal to the energy absorbed by unit mass; its unit is J kg^{-1}, i.e. the *gray* (denoted by Gy). This is the most fundamental dose concept. In the radiation therapy it is used almost exclusively, and the radiation dose eliciting the so-called deterministic effect (cf. section 3.4.1) may be also characterized by it.[3]

2. Exposure. While the absorbed dose is valid for any kind of absorbed radiation, the exposure refers only to *X*- and *γ-radiation.* An even more essential difference between the two types of doses is that, whereas the former provides direct information on the energy absorbed, the latter characterizes only the *ionizing capability* of radiation in the air, which consequently gives only indirect information about radiation actually absorbed by the tissue. If the exposure in some region of the tissues is mentioned, this actually refers to the ionization produced by X- or γ-rays in the air rather than in the tissues.

Before the exposure is defined, some considerations are required, since ionization measurements in the air may be carried out under different conditions. The definitions should also contain the conditions of measurement.

Figures 3.10 and 3.11 depict two extreme cases. A common feature of the two cases is the air-filled cavity separated by a wall from the surrounding tissues. In both cases the

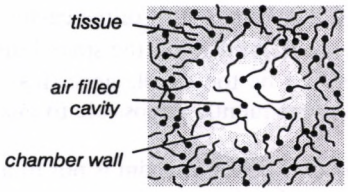


Fig. 3.10. Diagram relating to the measuring chamber operating on the principle of electron equilibrium

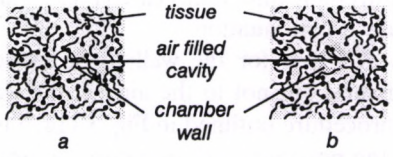


Fig. 3.11. Diagram relating to the measuring chamber operating on the Bragg–Gray principle

[3] The former unit was the rad (radiation absorbed dose): 1 rad = 0.01 Gy.

free positive and negative charges produced by the ionization of the air in the cavity are measured, though the number of charges produced in the air is influenced in different ways by the environment. The role of the surroundings may be understood if the processes induced by the radiation are considered in detail.

a) Not only in the air, but also in the tissues and in the wall surrounding the cavity does radiation induce photoelectrons, Compton electrons, and possibly electron–positron pairs (referred to below as *secondary electrons*), which cover various distances before losing their energy by ionization (and excitation). In the diagrams the points indicate the sites where the electrons are produced, and the irregular lines denote the tracks of the secondary electrons. It should be emphasized once more that only charges produced by secondary electrons in the air are measured.

These secondary electrons may possibly travel only a short distance in the air, and cross the wall before being stopped. Obviously, the charges produced by the secondary electrons in the remaining section of their track will not be measured. However, some secondary electrons produced in the wall or (if the wall is thin) in the surrounding tissue may complete their track in the air, and the charges they produce in the air will be measured. The material (and the thickness of the wall) can be chosen so that the number of charges lost due to the secondary electrons leaving the cavity will be balanced by the charges due to electrons emitted from the surroundings of the cavity into the air. In this case the charge density within the cavity will be the same as if the air in the cavity were surrounded by air. This leads to *electron equilibrium,* depicted in Fig. 3.10. Every substance which absorbs and scatters similarly to the air is said to be air equivalent.

In order to ensure electron equilibrium the cavity walls are made of some *air-equivalent material,* which must be thick enough to block the access of the electrons ejected from the tissues into the interior of the cavity. The dimensions of the cavity are not fixed, but with a smaller cavity better information is obtained about the spatial distribution of the dose. On the other hand, the cavity must not be too small, since this results in the released charge quantity being very low and consequently impossible to measure.

b) Figure 3.11 shows the extreme case, when the aim is not to attain electron equilibrium characteristic of the air in the cavity. Just the opposite is achieved: the *secondary electron density in the cavity is the same as in the tissues.* This happens whenever the dimensions of the cavity are small compared to the effective range of the electrons in the tissues. The cavity wall is extremely thin in this case (Fig. 3.11*a*), so that the electrons can pass through with practically no attenuation.

Another solution (Fig. 3.11*b*) is for the wall to be made of some material which behaves similarly to the tissues, but not to the air, i.e. the wall is made of some *tissue-equivalent substance.* The procedure outlined in Fig. 3.11*a*, *b* is the *Bragg–Gray method.*

Both solutions provide the possibility of dosimetry, i.e. one can measure on the basis of electron equilibrium, or according to the Bragg–Gray method. The former one was actually the first to be achieved, and whenever exposure is mentioned this method is thought of (cf. also section 3.3.4).

After these preliminaries the exposure can be defined. Let dq denote the positive or negative charge produced at *electron equilibrium* by ionization in air of mass dm and volume dV ($dm = \rho dV$). The exposure in this region is defined by

$$X = \frac{dq}{dm} = \frac{1}{\rho}\frac{dq}{dV} \qquad [3.19]$$

Thus exposure is measured by the charge produced by ionization at electron equilibrium in air of unit mass. Its *unit* is C kg^{-1}.[4]

It might well be asked why the ionizing effect produced *in air* is used for dosimetry. There are both theoretical and practical reasons for this. It should again be mentioned that ionization produced in air can in practice be measured fairly exactly, with good reproducibility and relatively simply. Therefore it is used in measuring technique mainly for calibrating purposes. In principle it is also important to know that ionization produced in air runs parallel to the biological effects and is *independent of the wavelength* of the radiation. If the ionizing effects produced by radiation of two different wavelengths are the same in air, the biological effects in a given tissue under given conditions will also be practically the same. This holds only approximately (or not at all) for other effects, such as produced in a photographic emulsion or luminescence, since their wavelength dependence is different from the wavelength dependence of the biological effects. The degrees of darkening of photographic film, for instance, may be the same for radiation of different hardnesses, whereas the expected biological effects will be quite different. The correspondence between the effects of ionization in air and the biological effects is attributed to two circumstances. One has already been discussed (section 2.10.3; Fig. 2.27): the ratio of the absorption (and scattering) coefficients in the air and in the tissues is practically the same at various wavelengths. The other circumstance is the fact that the energy required to produce one ion pair is also independent of the wavelength; its value for electrons is approximately 34 eV in air (and on average is the same in the tissues).[5]

3. Derived dose concepts. A given dose of some radiation may be obtained by the body during different times. In this case *dose rate* is important; this quantity is defined as the quotient of the dose absorbed by the body and the duration of irradiation. Depending upon the dose type the units may be Gy h^{-1}, mGy h^{-1}, C kg^{-1} h^{-1}, etc. For instance, the absorbed dose rate at some region of the body is 1 unit if the absorbed dose per unit time is 1 Gy. Another derived concept is the *integral* or *volume* dose. If the dose is the same at every site in some part of the body of mass m (homogeneous dose distribution), the integral or volume dose is given by the product of the mass and the dose. Thus, the integral absorbed dose gives the energy absorbed by a tissue of mass m, for instance. The unit of the integral dose depends upon the dose actually used: J, C. Consider an inhomogeneous distribution in a sufficiently small domain within which the dose is

[4] The earlier unit was the Roentgen (denoted by R). 1 R is the exposure that in 0.00129 g air (1 cm^3 of normal state) produces positive or negative ions carrying 1 electrostatic unit (esu) of electricity (= 3.34×10^{-10} C) of either sign. Under the same circumstances 2.6×10^{-7} C positive and negative charges are produced in 1 g air. 1 R = 2.6×10^{-4} C kg^{-1}

[5] Approximately 35 eV for protons, α-particles, etc.

constant; the integral dose is then calculated separately for every small domain, and the results are summed.

3.3.4. Dosimetry in practice

1. Small ionization chambers. Dosimetry in medical practice is mainly restricted to the measurement of photon and electron radiation. In this section, only the measurement of photon radiation is dealt with, though the results are also applicable to electron radiation. As concerns interaction with the medium, there is no basic difference between photon and electron radiation, since the effects actually measured with photons are produced by electrons.

It clearly follows from the foregoing that the dosimetry of photon radiation is particularly conveniently carried out by methods based on ionization of air. The use of the smallest possible ionization chambers is advisable, since they virtually do not perturb the radiation field, and permit a relatively exact mapping of the spatial distribution of the radiation. A frequently used group of small ionization chambers are the thimble chambers (Fig. 3.12). One electrode is the chamber wall itself, while the other is a rod protruding into the chamber through an insulator. The measuring space is the air space of the chamber. Ionization chambers can be used to measure either dose or dose rate. Their volume can be reduced below 1 cm^3. With this small size the device can be placed into the cavities of the body, thereby allowing direct measurement of the dose at certain sites within the body in treatment with X-radiation or γ-radiation.[6]

By appropriate selection of the material and the chamber wall thickness, the secondary electron density characteristic of either the air or the surrounding tissues is produced in the air space of the chamber. The former method is applied if photon energies up to 0.6 MeV are used, and the latter with photons of higher energy. For the lower energy range the chamber wall is made of air-equivalent material with a thickness large enough to prevent the secondary electrons produced in the surrounding tissues from entering the measuring volume. This condition can be satisfied with some suitable material with a wall thickness of 2 mm or less. (The chamber wall behaves as a compressed air layer within the

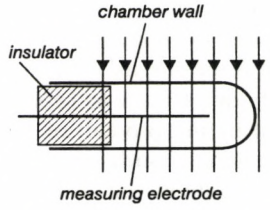


Fig. 3.12. Outline of the thimble chamber
The arrowed lines denote the radiation to be measured

[6] For a more careful measurement of the dose distribution, phantoms are made of some appropriate material (water, wax, etc.); in shape and radiation absorption, these are similar to the part of the body to be irradiated. If the chamber is placed at various sites of the phantom, the dose distribution can be mapped.

chamber.) The higher the photon energy, the thicker the wall required for electron equilibrium characteristic of the air in the chamber to be established. Thick-walled and large chambers, however, are unsuitable. Accordingly, a different measuring principle is employed if photon energies higher than 0.6 MeV are to be measured. In this case the chamber wall is made so thin that secondary electrons produced in the surrounding tissues pass through it almost without attenuation. Consequently, a secondary electron density characteristic of the surrounding medium is produced in the small air space of the chamber.

The chambers developed for the lower energy range are *photon detectors,* whereas the high-energy chambers are *electron detectors.* The nomenclature is an indication that in the former case the measurement is based on the absorption of *photons* in the chamber (mainly the chamber wall), whereas in the second case the photons absorbed by the chamber are negligible, and it is rather the *secondary electrons* entering the chamber (from its surroundings) that are important.

The photon detectors measure directly the exposure, however, the dose absorbed both by the air and the neighbouring tissues may be determined from it. In the case of electron detectors the energy absorbed by the tissues is directly determined by the measurement of the charge released in the chamber or of the energy absorbed by the air, on the basis of the Bragg–Gray principle.

The results with these two methods are discussed separately below.

A) The photon detector

a) *Absorbed dose in air.* Since the charge of an electron is 1.6×10^{-19} C, 1 C kg^{-1} exposure is produced by the release of $10^{19}/1.6$ electrons, or by the same number of ion pairs. The production of a single ion pair requires on average 34 eV, or $34 \times 1.6 \times 10^{-19}$ J; consequently, 34 J is needed to produce $10^{19}/1.6$ ion pairs. Thus, 1 C kg^{-1} exposure in air is equivalent to 34 J kg^{-1}, i.e. 34 Gy absorbed dose. From this, however, it follows that for an exposure X the absorbed dose D_{air} is calculated from the equation

$$D_{air} = f_0 X, \quad \text{where} \quad f_0 = 34 \frac{\text{Gy}}{\text{C kg}^{-1}} \left[= \frac{\text{J}}{\text{C}} \right] \qquad [3.20]$$

Since from the exposure the dose absorbed in air can be easily calculated, recently instead of exposure rather the dose absorbed in air is used and accordingly the instruments too are calibrated in Gy units.

b) *Absorbed dose in tissues.* Since the absorptive power of tissues is greater than that of air, the photon radiation losing 34 J kg^{-1} energy in air, loses more in tissues. The absorbed energies in the case of photon detector are in the same proportion to each other as the respective *mass attenuation coefficients,* i.e.

$$\frac{D_{tissue}}{D_{air}} = \frac{\mu_{m,tissue}}{\mu_{m,air}} \quad \text{or} \quad D_{tissue} = \frac{\mu_{m,tissue}}{\mu_{m,air}} D_{air} \qquad [3.21]$$

from which, with regard to [3.20]:

$$D_{tissue} = \frac{\mu_{m,tissue}}{\mu_{m,air}} f_0 X \qquad [3.22]$$

B) With an **electron detector** – as it was mentioned earlier – one obtains directly the energy absorbed by the tissues by the Bragg–Gray principle. Though the secondary electron density is the same in the chamber and in its surroundings, the absorbed doses are still different, because the stopping powers of air and tissues for electrons are different. The ratio of the absorbed doses is equal to the ratio of the *mass stopping powers* (S_m):

$$\frac{D_{\text{tissue}}}{D_{\text{air}}} = \frac{S_{m,\text{tissue}}}{S_{m,\text{air}}} \quad \text{or} \quad D_{\text{tissue}} = \frac{S_{m,\text{tissue}}}{S_{m,\text{air}}} D_{\text{air}} \qquad [3.23]$$

Table 3.1 gives some mass attenuation coefficient ratios. If 1.08 is accepted as an average value, then according to [3.22] at an exposure of 1 C kg^{-1} the soft tissues absorb approximately 37 J kg^{-1}.

Table 3.1. Some data on the mass attenuation coefficient of tissues as related to air for photon radiation

Photon energy (MeV)	$\mu_{m,\text{tissue}}/\mu_{m,\text{air}}$	
	soft tissues	bones
0.1	1.07	3.54
0.2	1.08	2.4
0.4	1.1	1.25

The Table also shows the radiation hardness dependence of the energy absorbed by the tissues at a given exposure, i.e. the wavelength dependence of the proportionality of the exposure and the absorbed dose. Disregarding the extremely soft and the ultra-hard radiation (not presented in the Table) for *soft tissues,* the proportionality is practically independent of the wavelength. This is to be expected from what was said above, as otherwise the two dose types could not be used together to characterise the biological effects. However, parallelism can exist between the biological effect and the absorbed dose, and between the biological effect and the exposure only if it also exists between the absorbed dose and the exposure. A substantial wavelength dependence is observed only with the bones.

Table 3.2 contains data referring to the relative mass stopping power of carbon, but these yield information on the soft tissues as well.

Table 3.2. Some data on the mass stopping power of carbon as related to air for electrons

Electron energy (MeV)	$S_{m,\text{carbon}}/S_{m,\text{air}}$
0.1	1.016
0.3	1.007
1.0	0.985
3.0	0.946

2. Other methods of dose determination. Besides the methods based on the ionization of air, other radiation effects are also used in practice for dosimetry, e.g. photographic, luminescent, and more rarely thermal or chemical effects. In any actual case, however, the method most suitable for measuring the radiation should be selected, with particular regard to the spectral distribution (in either the wavelength or the energy spectrum), the dose range, and so on. In the following section some remarks are made concerning the procedures used most frequently.

A) Film dosimeters (Film badges). The absorption spectrum of a photoemulsion differs from that of the tissues, film badges displaying a considerable energy dependence, which is especially strong with photon energies between 0.04 and 0.4 MeV. Consequently, photoemulsions are mainly used with *radiation sources exhibiting identical spectra.* For instance, film badges may readily be used to compare different radiation sources containing radium, but to compare the radiation emitted e.g. by a radium preparation and an X-ray tube they can be employed only in conjunction with special evaluation methods. Persons who work with radiation are usually provided with film badges, which are evaluated centrally.

B) Luminescence dosimeters. In these dosimeters the detector is a small plate made of some luminescent material enclosed in a light-proof case. The light produced by radiation is transmitted through an optical system and measured with a photocell or a photomultiplier. The photocurrent yields information about the dose rate.

The disadvantages mentioned in connection with film badges also hold for this method, except for those increasingly frequent cases when the dosimeters are made of air- or tissue-equivalent materials.

C) Calorimetric method. The essence of this method is the transformation of the absorbed radiation energy into heat, the absorbed dose rate being inferred from the temperature change. This method is mainly suitable for the measurement of X-radiation and γ-radiation. Its advantage is that it is independent of the *spectral distribution* of the radiation. A disadvantage, however, is the long time required to obtain the result, and the caution that must be practised, which makes the use of this method rather difficult. The temperature changes due to radiation are extremely small, and a smaller dose rate than 0.5 Gy/min is not measurable with this method. The temperature increase is of the order of magnitude of 10^{-4} °C.

D) Other methods

a) The *electric conductivity* of some *semiconductors* increases on irradiation. The measuring instrument in the circuit is directly calibrated in dose units. Due to its small dimensions the semiconductor crystal can be fixed on the end of a thin, flexible probe. This type of detectors can readily be introduced into the interior of the organism or into the cavities of the human body.

b) Certain *solutions* and *gases* are also used, since radiation induces *chemical changes* in them. Fe^{2+} oxidation (ferrous sulphate dosimeter) is frequently made use of, but the precipitation of iodine from alkyl iodides, the decomposition of formic acid, the bleaching of methylene blue, and other reactions are also applicable.

c) Some crystals (e.g. alkali halides), *glasses* (e.g. phosphate glasses containing silver, glasses containing cobalt or lead), *organic solids* (e.g. plexi, PVC) and so on become coloured on irradiation. (This is the effect which can be observed on the glass coating of X-ray tubes after prolonged use.) The intensity of the colouration is measured by means of colorimetry. The colouration effects on consecutive irradiations are additive. The colorimetric measuring light has practically no effect on the irradiation-induced colour, and hence the evaluation can be repeated several times. Since the colouration can be eliminated by heating, the same crystal can be repeatedly used.

d) Appropriately activated glasses can be used in a different way, too. It is well known that certain substances which have previously been exposed to X-radiation or γ-radiation become luminescent in the visible range if excited with ultraviolet light. The emitted luminescent light, measured for instance with a photomultiplier, is proportional to the X-ray or γ-ray dose *(photoluminescence dosimeter).*

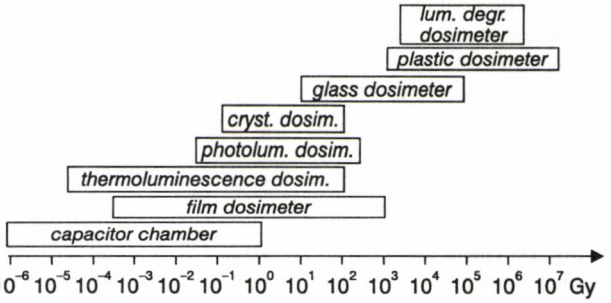


Fig. 3.13. Application ranges of various dosimeter types. Capacitor chambers are ionization chambers used in individual dose measurement

e) Certain substances (e.g. anthracene) which are luminescent if irradiated with ultraviolet light gradually lose their luminescence if exposed to X-radiation or γ-radiation. *Luminescence degradation dosimeters* are based on this phenomenon.

f) A number of materials (e.g. manganese-doped CaF_2, LiF crystals) emit light on heating, after previous X-irradiation or γ-irradiation. This effect is utilized in *thermoluminescence dosimeters.*

The above effects are used in practice in different dose ranges (Fig. 3.13). Individual pocket dosimeters are made as small cubes, needles or foils, and are carried suitably encased either in the pocket or round the neck.

3.4. Radiation protection

3.4.1. Classification of radiation effects

Ionizing radiation damages the living organism. This effect is primarily produced by the resulting *ionization* (and *excitation*) in the organism. The charged particles cause ionization directly, while X-ray and γ-photons and neutrons do so indirectly. The ionization may occur in the molecule playing the key role in the development of the hazard, or in the water molecules comprising the bulk of the mass of the living organism. The former case is usually referred to as a *direct,* and the latter as an *indirect* radiation effect. In the indirect case, monovalent free hydroxyl radicals may be formed from water molecules in contact with the air; these free radicals undergo intermediate processes and are finally transformed to hydrogen peroxide. Free radicals migrate by diffusion; the result of both direct and indirect radiation effect may therefore appear in one or another of the biologically important molecules in the form of some lesion which changes the biochemical reactions or produces mutations resulting in damage, and possibly in the killing of the cells or even the whole organism. These processes are depicted in Fig. 3.14.

A basic role in radiation damage of cells can be attributed to the effect on DNA. The DNA molecule is several orders of magnitude larger than any other molecular component of a cell, so that molecular damage by ionization occurs with the greatest probability in DNA or its surroundings. The functional consequences of the molecular damage manifested at a cell level are of decisive importance in the case of DNA. As regards diploid cells, some chromosomal fragment in a single cell occurs in at most two

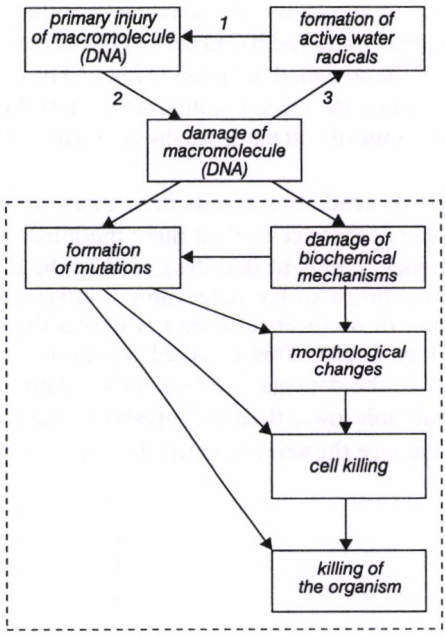


Fig. 3.14. Scheme of processes connected with the development of biological radiation effects
Arrows 1 and 3 refer to reactions between active water radicals and macromolecules, arrow 2 to the intramolecular energy transfer, and unmarked arrows to metabolic processes. The series of processes following the primary radiation effect can be seen in the frame

copies, and the damage of either may result in the loss of some given cellular function. On the other hand, damage to any molecule which exists in several thousand copies in the same cell is less harmful.

Based on the ever increasing experience, but also merely from the reasoning outlined above it seems to be justified to divide radiation effects into two groups.

The division is based on the experience that the radiations may cause two types of damages at the level of the cells: the cell either dies or it survives the damage but its function (metabolism) will be modified. These *two types of cell damage* carry different consequences also for the living organism as a whole.

a) Deterministic effect. Most organs and tissues of any organism remain functional until the amount of the destroyed cells is not too large, i.e. does not reach a certain limit. At this point the signs of the loss of the given function become perceptible. Thus at low doses enough the appearance of such damages is improbable, in other words, has a probability of zero, while above a certain dose, the so-called *threshold dose,* the probability suddenly reaches certainty, i.e. the unit value (100%). The above is shown in Fig. 3.15*a*. The *severity* of the damage increases together with increasing doses. This type of damage is called the *deterministic effect.* Such damages may be caused by the high-dose irradiation of the haemopoietic cells (red bone marrow) with the following consequences: haemorrhagic

diathesis, lack of resistance to infections, and anaemia because of the lack of replacement of the thrombocytes, leukocytes and red blood cells in the circulating blood. These are the characteristic symptoms of the so-called *radiation sickness*. The deterministic effect of high radiation doses is utilized when the cells of a tumour are killed for therapeutic purposes. The dose eliciting the deterministic effect is usually characterized by the absorbed dose.

b) Stochastic effect. If the irradiation results in *modified somatic cells,* but it does not influence the reproductivity of these cells, then these modified cells may form cell groups which, after a certain latency (years to decades), lead to the development of *malignant tumours* (carcinomas). In contrast to the deterministic effect, the probability of tumour development (its frequency of occurrence) does not have a threshold dose, but it is proportional to the dose (Fig. 3.15*b*). This is called *stochastic effect,* which refers to the random, statistical nature of the damage. We emphasize again that in this case the doses on the abscissa are considerably lower than the threshold dose of the deterministic effect. In the case of stochastic damage the severity of the disease does not depend on the magnitude of the dose.

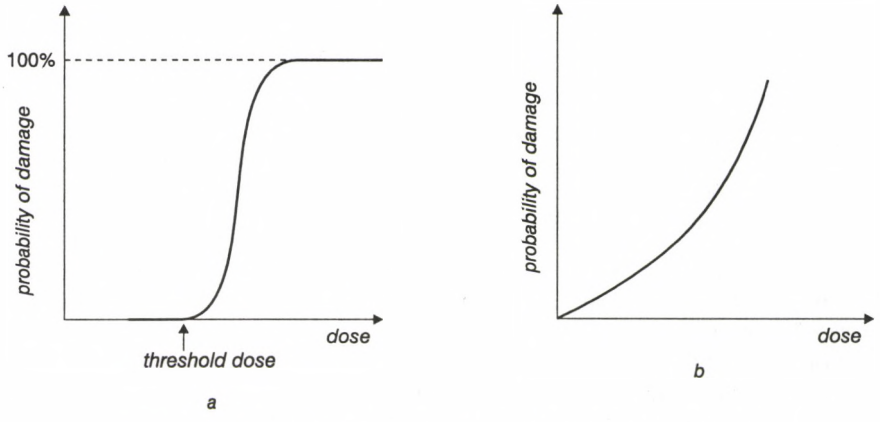


Fig. 3.15. Relationship between the probability of occurrence of the stochastic *(a)* and the deterministic *(b)* effect and the dose of exposure

3.4.2. Dose concepts used in radiation protection

The task of radiation protection is partly to avoid the deterministic effect, partly to prevent or minimalize the stochastic effect. In the following the latter will be discussed. This task involves the estimation and prediction of the expected damage of a given dose (e.g. in connection with the occupation or a diagnostic procedure). For the sake of its solution special dose concepts were introduced which correspond to the purposes of radiation protection. These concepts are renewed and actualized from time to time, as our knowledge concerning the effects of radiation becomes more profound and exact. If we want to infer the effect from the dose, we have to consider the type and energy of the

radiation on one hand, and its distribution in the individual organs/tissues within the organism on the other hand.

a) Equivalent dose. If for radiation R under study the average value of the absorbed dose $D_{T,R}$ in a given T tissue or organ is known, this should be multiplied by a weighting factor w_R characteristic for the radiation to get the equivalent dose H_T:

$$H_T = w_R D_{T,R} \qquad [3.24a]$$

The *radiation weighting factor* shows the effectivity of a given radiation in eliciting the stochastic effect as compared to that of the γ- and X-radiations. First the doses of X-radiation and of the radiation under study which cause an effect of a given quality and magnitude are determined, then their quotient is produced. As many times smaller the dose of the radiation required for the same effect than that of X-radiation, as many times higher its effectivity. This quotient gives the value of w_R (Table 3.3).

Table 3.3. Weighting factors of different radiations and energies

Radiation and energy range		Radiation weighting factor
Photons, at every energy		1
Electrons, muons, at every energy		1
Neutrons, if the energy is	<10 keV	5
	10 keV–100 keV	10
	100 keV–2 MeV	20
	2 MeV–20 MeV	10
	> 20 MeV	5
Protons, if the energy is	> 2 MeV	5
α-particles, nuclear fission products, heavy nuclei		20

If the radiation consists of *more components* (radiations of different type and energy), the different radiations are considered separately, their average doses ($D_{T,R}$) are multiplied by the corresponding radiation weighting factors, and the sum of these weighted doses is the equivalent dose:

$$H_T = \sum_R w_R D_{T,R} \qquad [3.24b]$$

The *unit* of equivalent dose is joule/kg, since the weighting factor is a dimensionless quantity; its name is *sievert* (*Sv*).

b) Effective dose. The relationship between the equivalent dose and the probability of the stochastic effect depends on the irradiated organ. The effective dose (E), which is derived from the equivalent dose, characterizes the total stochastic effect. It is calculated by multiplying the equivalent dose of the given tissue or organ (H_T) by the *tissue weighting factor* (w_T). The tissue weighting factor shows the proportion of the given tissue or organ in the whole damage caused by the homogeneous irradiation of the total body. The tissue weighting factors are presented in Table 3.4.

Table 3.4. Weighting factors of various tissues and organs

Tissue/Organ	Weighting factor, w_T
Gonads	0.20
Red bone marrow	0.12
Large intestine	0.12
Lungs	0.12
Stomach	0.12
Bladder	0.05
Breast	0.05
Liver	0.05
Oesophagus	0.05
Thyroid gland	0.05
Skin	0.01
Surface of the bones	0.01
Other	0.05

The effective dose for the whole body is obtained by the summation of weighted doses for all tissues/organs:

$$E = \sum_T w_T H_T \qquad [3.25]$$

The *unit* of the effective dose is joule/kg, its name is *sievert* (*Sv*) like that of the equivalent dose. Since the sum of the tissue weighting factors is one unit, it is easy to see that in case of homogeneous whole body irradiation the numerical value of the effective dose is the same as the numerical value of the equivalent dose.

c) **Further dose concepts.** The dose concepts discussed above characterize the radiation load of the *individual*, or the risk connected to it. The radiation load of a given *population* is characterized by the *collective effective dose (S)*. If in the ith group of a population N_i persons suffered an average effective dose E_i, the collective effective dose is given as

$$S = \sum_i E_i N_i \qquad [3.26]$$

This formula considers every group of the population, the average load in every group (E_i) and the number of persons in each group (N_i) alike. Its unit: persons × sievert (persons × Sv).

The radiation load from an *incorporated radioactive material*, i.e. from a so-called internal radiation source, is characterized by the *committed equivalent dose*, its sign is $H_T(\tau)$. This corresponds to the already mentioned equivalent dose, the expression *committed* refers to that the radiation source is, so to say, committed in the organ or tissue. The committed equivalent dose considers the dose rate changing in time $[dH_T(t)/dt]$ as well as the time (τ) spent by the isotope in the given organ or tissue. Their product gives the committed equivalent dose. If the time cannot be specified (e.g. in case of isotopes with long effective half-life), for adults 50 years, for children the period from the incorporation up to the age of 70 years should be considered for the calculation. Its unit: sievert (Sv).

The *committed effective dose, $E(\tau)$,* a concept corresponding to the above-mentioned effective dose, gives the dose from the incorporated isotope for the whole body. Its unit too is sievert (Sv).

The *committed collective dose* resulting from the incorporation of radioactive isotopes is calculated from the average committed effective doses, according to [3.26].

3.4.3 Exposure. Dose levels

Every human being is exposed to radiations from natural and artificial sources alike, thus to radiation load. The radiation load may come from the *environment*, from a *medical*

(diagnostical or therapeutical) *activity* using ionizing radiation, or may be connected to *an occupation with ionizing radiations.* The radiation loads following *nuclear accidents and disasters* deserve special attention.

1. Environmental radiation. The environmental or background radiation comes mainly from natural sources, its *yearly effective dose is about 2 mSv.* This yearly dose is tolerated by our organism without any apparent damage. However, if the load is higher than this, we have to reckon with the risk of the stochastic effect, for instance with a certain probability of tumour development. On the basis of the experience on nuclear disasters, it can be estimated that a dose increase of 1 Sv results in tumour development with a probability of 0.04–0.07.

The various components of the background radiation and their equivalent doses are presented in Table 3.5. According to the experience, the doses coming from both the natural and artificial isotopes may reveal considerable differences depending on the geographical location, on the applied building material, etc. For example, the dose of ^{222}Rn produced in the course of the decay of ^{226}Ra may vary between 1 μSv/year and 10 μSv/year at various places of the Earth. If we compare the up-to-date and traditional coal-heated power-plants, in the environment the collective dose related to 1 GW electric power may be 0.5 and 6 persons × Sv, respectively, as a consequence of the emission of the radioactive contamination of the coal. The doses coming from the cosmic radiation vary according to the height above sea-level. Their value is about 2 mSv/year in Tibet, at 3600 m above sea level, and about 0.3 mSv/year at sea-level, respectively.

Table 3.5. Distribution of equivalent doses obtained yearly from background radiation

Origin of radiation		Critical organ	Quantity in whole body (kBq)	Equivalent dose per year (mSv)
Cosmic radiation		Whole body	—	≈ 0.4
Environmental radiation		Whole body	—	0.5–1*
				1.0–4**
Incorporated	^{40}K	Muscles, whole body	≈ 4	≈ 0.2
	^{14}C	Fatty tissue, whole body	≈ 4	≈ 0.02
	^{226}Ra	Bones, haematopoietic organs	0.2–0.02	0.1–0.5
	^{222}Rn	Lung	–	0.3–2.5

* in open air

** referring to radiation in houses and depending on building materials

2. Exposures related to medical activity. Both diagnostic and therapeutical radiations are accompanied by radiation load. According to our present knowledge, every dose increase above the background radiation increases the risk of the stochastic damage (see previous point). The average value of the medical exposures is about 0.4–1.0 mSv/year per person, i.e. it is about the same as that of the background radiation. Most part, about 80–90%, of this dose comes from the radiological diagnostical examinations. Nevertheless, it has to be taken into consideration when estimating the risk that the given

average value is the average for the whole world, and there may be very substantial differences depending on the state of the health care in the different countries. An example: on the average about 1.5×10^9 radiological examinations are performed yearly in the world, which means a collective dose of about 1.6×10^6 persons $\times$ Sv. However, only a quarter of the world population are living in countries with highly developed health care, and three-quarters of the examinations are performed on this quarter; thus the collective dose of the population with advanced health care is much more than the quarter of the total collective dose: about 1.3×10^6 persons $\times$ Sv.

Some informative data: the dose measured on the skin surface facing the X-ray tube is a few tenths mGy in case of a roentgenogram, several hundred mGy during X-ray transillumination; CT examinations mean usually a dose 10–20-times higher than a simple roentgenogram. The effective doses caused by the most frequently performed examinations are summarized in Table 3.6.

Table 3.6. Approximate effective doses of the most frequent radiological examinations

Examination	E (mSv)
Chest X-ray	0.04
Chest CT	7.8
Cranial X-ray	0.1
Cranial CT	1.8
Abdominal X-ray	1.2
Abdominal CT	7.6
X-ray of the dorsal vertebrae	1.0
X-ray of the lumbar vertebrae	2.1
Barium enema with fluoroscopy	8.7

Thus, the dose connected with medical activity (diagnostics) may reach the effective dose of the background activity and may be even 3–4 times higher than that in the course of only one examination of an individual. Considering the risk, it is the responsibility of the physician to carry out indeed only the important examinations or therapy in a given patient. Nevertheless, the medical (both diagnostic and therapeutical) dose is considered in a way different from the other doses, since, although the irradiation of a part of the body increases the risk of the stochastic effect for the whole body, it is associated with a direct advantage for the patient.

3. Occupational exposures affect only a relatively small part of the population: the medical staff using ionizing radiation, the personnel of nuclear power plants and of institutions producing or using radioactive isotopes. In this population the collective dose is about 7000 persons $\times$ sievert yearly. Concerning the amount of the individual occupational dose, the International Safety Standards for Radiation Protection should be observed. According to these, the design, use and operation of radiation sources and any activity accompanied by radiation should be carried out in such a way, which ensures a level of exposure below the reasonable lowest limit taking into consideration the economical and social factors.

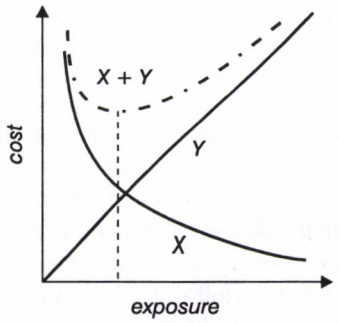


Fig. 3.16. Optimalization of radiation protection

The concept of reasonableness can be demonstrated with examples. Human life is full of activities in which for the sake of the expectable advantages or benefits one has to take risks. If e.g. we drive somewhere in a car we take the risk of the travel to reach our destination. Particularly the extent of intervention is often decided by a risk versus benefit analysis carried out instinctively or consciously. This situation arises frequently in surgical activity but also in drug treatment and in the diagnostic and therapeutic applications of radiations.

The reasonableness or more exactly the *optimizing requirements* are shown in Fig. 3.16. For easier understanding let us think of a laboratory where the staff works with radiating substances. The total radiation exposure of the working team is indicated on the horizontal axis and the cost involving stochastic health detriment of the workers and radiation protection is given on the vertical axis. The lower is the exposure permitted the higher will be the cost of protection, but the risk will be smaller, consequently the cost of health detriment will be also smaller. The first relationship can be considered nearly hyperbolical, the latter one linear and they are denoted in the figure by X and Y, respectively. The total cost is the sum of the two partial costs. The minimum of the resulting $X + Y$ curve determines the reasonable total exposure.

The outlined optimizing analysis cannot be carried out in an exact way in every respect, thus one has to be satisfied with qualitative or semiquantitative approximations.

According to the present knowledge, based on the optimizing principle and on the increasing number of statistical data an effective dose, the *occupational limit* may be established; exposures below this limit present a still *acceptable risk.* Its value is 20 mSv/year in the average of 5 years, with the restriction that it must not exceed 50 mSv in any year, furthermore that only persons above the age of 18 years may be employed at working places using ionizing radiations. With respect to the medical activity (e.g. interventional radiology), the dose limits for some organs may be of interest: eye lens 150 mSv/year, hands (feet) 500 mSv/year.

4. The nuclear disasters, from which our knowledge concerning their consequences has been gained, occurred in connection with the employment of the atomic energy for the purposes of both war (atomic bomb) and peace (nuclear reactors). An example for

the first is the nuclear attack on Japan in 1945, while for the second, the nuclear accident in Chernobyl in 1986 should be mentioned as the closest in time and space.

With respect to the injuries to health, there is a considerable difference between the exposures in the immediate vicinity of the event (explosion) and those present at a distance. In the immediate vicinity the exposure (usually of γ- and neutron radiation) is much higher than the occupational limit, it may even exceed the limit for the deterministic effect; depending on the exposure, the symptoms of *radiation sickness* develop within a few minutes or hours. Mild radiation sickness ensues upon an exposure of 1–2 Gy: 2–3 hours after the exposure nausea and vomiting set in, and leukocytopenia develops. In case of radiation sickness of medium severity (following 2–4 Gy) there is a considerable decrease in the number of thrombocytes, too; this leads to increased haemorrhagic diathesis and a decreased resistance against infections. Upon 4–6 Gy these symptoms become more severe; without treatment 50% of the persons exposed to 4–5 Gy die within 30 days: this is the so-called *median lethal dose* (LD_{50}). Following higher exposures (exceeding 6 Gy) gastrointestinal and neurological symptoms also develop, and the outcome is usually fatal.

In areas at a distance from the disaster the exposure is the consequence of the emission of radioactive materials (medium size nuclei, e.g. isotopes of iodine and cesium). The severity of nuclear accidents is characterized also by the extent of the *emitted activity.* The emission after the disaster of Chernobyl, the most serious so far, was in the order of TBq. Radioisotopes getting into the air may reach places far away from the explosion and by falling to the ground may increase the background radiation on one hand, and on the other hand may be incorporated by eating, drinking and breathing. Naturally the extent of the radioactive contamination decreases with increasing distances from the accident. For instance, while in the area of the disaster the emission caused a contamination of 0.5–1 MBq m^{-2}, in Hungary the measured values were smaller by several orders of magnitude (1–10 kBq m^{-2}).

3.4.4. Radiation hazards and chemical hazards

In everyday life we come into contact with numerous chemicals which may be biologically hazardous. The injury starts with molecular interactions, but the consequences may be manifested at cell, tissue or even organ level. The processes induced by radiation and by chemicals are in many instances similar, differing only in the initial steps of the process. Even here, however, similar phenomena are known, for instance the strand breaks induced by ionizing effects or chemical agents in the nucleic acids, which are of fundamental importance from the aspect of hazardous effects. The differences displayed at a molecular level gradually disappear at higher levels. Hence, any of the physical or chemical agents may produce mutations, malignant transformations, cell death, or cell aging.

A scheme of the processes induced by chemicals is presented in Fig. 3.17. This demonstrates that only some of the hazardous chemicals are active, i.e. only a proportion of them develop direct interactions, the alkylating agents. Another group of chemicals become hazardous only if they are previously activated by metabolic processes, e.g. benzopyrene or ethylene. On the other hand, active chemicals may also be inactivated by metabolic processes.

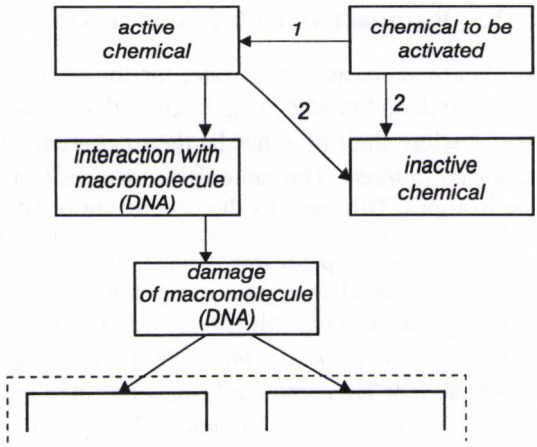


Fig. 3.17. Scheme of the development of chemical damage
Arrows 1 and 2 indicate the activation and inactivation possibilities during metabolic processes. Processes following primary damage are only referred to (cf. Fig 3.14)

On the basis of the similar molecular and cell damage, the effects of radiation and chemicals may be compared. Table 3.7 presents some data on chemicals which are present as the combustion products of various substances in urban atmospheric air. The last column contains the radiation doses, which produce the same biological effects as the chemicals in the given concentration during *an exposure of one week.* The Table demonstrates that the dose relating to even a single combustion product is higher than that due to the background radiation; moreover, some of the values roughly attain the dose equivalent limit at occupational exposure (cf. section 3.4.3).

Table 3.7. Comparison of the hazardous effects of some combustion products and ionizing radiation

Compound	Concentration	Biological effect studied	Dose equivalent per week (mSv)
Ethylene	0.05 mm^3/l	Mutation in mice	≈ 0.1
Ethylene	1 packet of cigarettes daily	Mutation in mice	≈ 0.7
Formaldehyde	2×10^{-2} $\mu g/l$	Cell killing in *in vitro* culture	≈ 0.5
Benzopyrene	2×10^{-3} $\mu g/l$	Cell mutation in *in vitro* culture	≈ 0.06

Chemicals are usually tested first *in vitro* to reveal possible lethal mutations in some virus strain. In order to obtain a simple quantitative characterization of the effect, the probability of injury development in the initial stage of the incubation ($t = 0$) is determined, and related to unit time and unit concentration. This quantity shows differences of several orders of magnitude for various compounds.

3.5. Radioactive isotopes as tracers

The following statements are valid, as appropriate, for the whole body and molecular levels, in the scientific research and practice (e.g. medical diagnostics) alike.

Radioactive isotopes[7] disclose their presence by their radiation, and thus their movement and fate can be traced *(tracers)*. The radioactive isotopes form compounds in the same way as the stable isotopes. This permits the production of labelled molecules and compounds, which behave virtually identically to the unlabelled ones in the various chemical, biochemical and biological processes. Besides the simpler inorganic labelled compounds, a very wide range of labelled organic compounds exist (labelled amino acids, hormones, pharmaceutical products, etc.). Moreover, it is also possible to exchange the $^{12}_{6}C$ atom by the $^{14}_{6}C$ isotope at *a fixed site in a molecule* containing several carbon atoms. In many instances the simple labelled compound is incorporated into a living organism, where it is transformed into a more complex organic molecule by the normal function of the organism *(biosynthesis)*.

3.5.1 The importance of the method

1. The "new" atom or molecule may be distinguished from the "old" one. Radioactive isotopes introduced into an organism (a chemical or other system) are distinguishable by their radiation from the atoms already present. This permits the relatively simple acquisition of *information* about the *dynamics of processes* of uptake, incorporation, exchange, secretion, etc. Many problems can only be solved by this means.

2. High sensitivity. The tracer method is extremely *sensitive*. The lowest limit of the most sensitive microanalytical technique is of the order of 10^{-11} g, which requires the simultaneous presence of 10^{10}–10^{12} atoms. With up-to-date radioactive methods, on the other hand, in principle even the presence of only one atom can be detected; at any event, 10^{5}–10^{6} atoms are sufficient for quantitative measurement. It follows that the sensitivity of the radioactive method is 6–8 orders of magnitude higher than that of microanalytical methods.

3. The physiological conditions do not change. The high sensitivity allows the study of various processes with amounts of substances so small that they have practically no influence on the life processes. For instance, the mass of 0.4 MBq radioactive $^{131}_{53}I$ necessary for examination of the thyroid function is approximately 8×10^{-11} g. This small quantity disturbs neither the normal functions of the organism, nor the state of the thyroid gland. The presence of the radioactive decay products is in most cases negligible. For instance, $^{131}_{54}Xe$ produced by the radioactive decay of $^{131}_{53}I$ has no disturbing effect at all. With appropriate doses, the biological effects of the radiation of the isotope are similarly of no consequence.

[7] The radiophysical data on isotopes dealt with in this chapter are given in Table 11, Chapter 9.

3.5.2 Minimalization of the diagnostic exposure

1. The role of the half-life. Certain examinations may be carried out with several elements or with several isotopes of the same element. A basic aspect in the selection of an isotope suitable for the solution of a given problem is the requirement that the organism be exposed to the *minimum possible dose.* This is especially important if repeated examinations are to be made in the same individual. The dose is proportional to the number of atoms disintegrating in the organism, and also to the energy transferred to the tissues by radiation produced in the disintegration of a single atom. It is therefore advantageous to use isotopes of *short half-life* (of the order of minutes or hours) which interact with the tissues mainly through γ-photons, since they are absorbed to a smaller extent.

The importance of the half-life is obvious if the decay law (cf. [3.1]) is written in the form

$$A \sim \frac{N}{T} \qquad [3.27]$$

The activity A can be measured with sufficient accuracy with a given apparatus only if it is above a certain limiting value. It follows from [3.27] that a smaller number (N) of radioactive atoms of short half-life (T) are required to attain a certain activity. Consequently, the introduction of a smaller quantity of a shorter half-life isotope into the organism is sufficient. One examination may be carried out with variously labelled compounds. With the selected isotope it is preferable to label a compound with a short biological half-life (e.g. a few hours).

Isotopes of very short half-life (some hours or minutes) can be used for tracing only if they are obtained at the place of application. This may be achieved by means of *isotope generators.* The most widespread among these is the *technetium generator* which is produced by putting the $^{99}_{42}$Mo isotope on an adsorbent and this is transported to the user (cf. section 3.2.3, point 2). $^{99}_{42}$Mo decays (by negative beta-decay) with a half-life of 66 hours to $^{99m}_{43}$Tc which can be easily washed out from the generator by simple physiological NaCl solution and can be used promptly. Its half-life is 6 hours.

Some health institutions equipped with *cyclotrons* produce very short half-life isotopes. This has opened up further possibilities of medical examinations (cf. section 3.2.6). Some of the isotopes produced by the cyclotron are $^{11}_{6}$C (20.4 min), $^{15}_{8}$O (2.1 min) and $^{13}_{7}$N (10 min). These are elements of great importance in both medical research and medical practice since they are integral constituents of the living organism. Oxygen and nitrogen isotopes with longer half-lives are not known, and consequently these radioactive isotopes can be used only in the way described. The examples given are all positive β-decaying isotopes. Other positron-radiating isotopes of short half-life are also used as tracing radioisotopes, for measurement of the annihilation photons flying in opposite directions permits localization of the radiation source with high accuracy (cf. positron scanner and PET, section 6.7.4).

2. Selection of the emitted radiation. Another aspect in the selection of a suitable isotope concerns the *manner of decay* and the *type of radiation emitted.* In the case of tracer isotopes, radiation is of importance as regards detection. γ-radiation emerging

from the organism with sufficient energy is generally used for detection. In the tracing technique, therefore, corpuscular radiation, even if stopped by the tissues, is unnecessary and indeed to be avoided. For this reason, if possible, only those isotopes are used whose dose load is practically due to their γ-radiation alone. (Isotopes which decay by shell electron capture are also suitable, since these are detectable by their X-radiation.)

In order to protect the organism from radiation, *harder* γ-radiation is more favourable. However, for technical reasons isotopes emitting relatively *soft* γ-radiation have to be used, since in the measurement it is important that as much as possible of the radiation be absorbed by the detector [e.g. a NaI(Tl) crystal a few cm thick]. As a compromise, 0.1 MeV photons are frequently used.

A favourable isotope as concerns the above conditions is $^{99m}_{43}Tc$, which emits 140 keV γ-photons with a half-life of 6 hours. Though some of the photons undergo internal conversion, the energy of the conversion electrons is also predominantly transformed into photon energy (characteristic X-radiation). A further advantage of technetium is that the TcO_4^- anion may be used to label a large number of compounds. Nowadays about 70% of the *radiopharmacons*[8] used in the *in vivo* isotope diagnostics is labelled by $^{99m}_{43}Tc$.

3.5.3 Examples for the possibilities provided by the method

1. Application in medical diagnostics

a) In vivo methods. The advantages of the tracing with radioactive isotopes (cf. section 3.5.1) are widely used in medical diagnostics, first of all for the examination of the function and malfunction of the different organs.

For historical reasons, the use of radioiodine in the diagnostics of the thyroid gland should be mentioned as the first method in which the metabolic processes of an organ were measured by means of the appropriate radioactive isotope. The essence of the method was the following: about 0.2–0.4 MBq ^{131}I isotope was administered orally into the organism in the form of a NaI solution; the amount of the active iodine taken up by the thyroid gland was measured in consecutive intervals by a scintillation counter placed close to the thyroid gland. The so-called iodine accumulation curve gained this way served as basis for the diagnosis. Nowadays the changes in the thyroid function are still examined by radioisotopic methods, but instead of the determination of the iodine accumulation curves – for reasons of measurement technology and radiation protection – the in vitro methods and the application of the ^{99m}Tc isotope are preferred.

While the function of the thyroid gland may also be followed by the in vitro determination of the hormones produced by it and present in the blood, in the case of several other organs only in vivo methods can be applied. An example is the similarly widely used *renography.* In order to study the renal function, a radioactive substance accumulating in the kidneys is labelled with a γ-radiating isotope, e.g. ^{99m}Tc, and the velocity of its accumulation and excretion is registered continuously by a scintillation detector placed above the kidneys. The resulting renograms yield information on the condition of the kidneys.

[8] Radiopharmacons are radioisotope-labelled compounds, which are used on the basis of their biological features to reveal the functional disorders of the organs or tissues.

Nowadays the renal function is examined mainly by the γ-camera method, too (cf. section 6.7.4). The advantage of this method is that by suitable processing of the information collected by the apparatus during the measurement, in addition to the renograms indicating the function of the kidneys, also the isotope distribution belonging to any time point of the examination may be presented graphically, thus dynamic and static examinations may be performed simultaneously and with great accuracy.

The radioisotope examination of the function of other organs is similar to the foregoing. The applied radiopharmacon is in every case a compound which is excreted with good efficiency by the organ to be studied (e.g. human serum albumin aggregate with a diameter of 10–70 μm for the examination of the lungs, or pyrophosphate for the study of the skeletal system).

Nuclear cardiology deals with relatively fast processes. By the appropriate detection of an isotope (e.g. human serum albumin labelled with ^{99m}Tc or ^{131}I) introduced into the circulation data characteristic of the cardiac function and circulation may be obtained. In case of γ-camera or SPECT examinations (cf. section 6.7.4) the computer processing of the isotope distribution maps recorded in about each 0.1 sec makes possible, for example, the presentation of the time course of the isotope content of the left ventricle and, based on this – also by means of a built-in computer program – the computation of the so-called ejection fraction (EF) characterising the condition of the left ventricle.

Without going into details, we just mention that by means of the above method the blood volume flowing into the aorta from the left ventricle (the so-called *cardiac output)* and, in the knowledge of the pulse rate, naturally the *volume per pulse* and the *volumes of the ventricles* can be determined, too. Other circulation characteristics (*pulmonary circulation*, *limb circulation*) may be determined according to similar principles.

The decomposition and regeneration of red blood cells may be followed by their labelling. (Similar examinations may be carried out also for other cells and tissues.) The activity of the labelled red blood cells introduced into the circulation decreases namely with time. The decrease is due partly to the physical decay of the labelling isotope (e.g. ^{51}Cr), partly to the spontaneous decomposition or destruction of the red blood cells. If blood sampling is carried out over a prolonged period (e.g. 30–40 days), the time during which a certain amount (e.g. 50%) of the red blood cells have been destroyed may be determined. According to the measurements, under physiological circumstances the *biological half-life of the red blood cells* in the circulation is about 130 days. Thus, from the decreased activity of the labelled red blood cells information may be obtained concerning their decomposition. On the other hand, if radioactive iron (e.g. ^{59}Fe) is administered intravenously, the continuous blood sampling reveals the increasing ^{59}Fe activity of the red blood cells (the iron introduced into the circulation is transported to the bone marrow and is continuously incorporated into the newly forming red blood cells). Thus, from the ^{59}Fe content of the blood samples information may be obtained with respect to the regeneration and formation of the red blood cells. Frequently, both isotopes (^{51}Cr and ^{59}Fe) are used simultaneously for the study of this haematological problem. In such cases the blood samples contain two different isotopes (*double labelling*).

b) In vitro methods. Recently in vitro methods have been developed in increasing numbers which are more reliable than the in vivo methods and do not involve radiation exposure either. The in vitro methods were elaborated first of all for the determination of various hormones (e.g. insulin, growth hormone, sexual hormones, thyroid hormones), but they play an increasingly important role in the determination of the quantitites of other, non-hormonal substances (e.g. immunoglobulins, bile acids, vitamin B_{12}) and drugs (e.g. barbiturates, morphine).

One group of the in vitro methods is the *radioimmunoanalytical method*, which combines the high specificity of the immune reactions and the sensitivity of the isotope technique. This is indicated also by its name. The method is suitable for the independent determination of small quantities of substances with similar chemical structures (e.g. serum hormones).

Of the radioimmunoanalytical methods the *radioimmunoassay (RIA)* is mentioned. Its principle is the following: a known amount of the antigen (hormone) to be determined is labelled with radioisotope, then mixed with a solution containing its specific binding protein, the antibody. During a sufficiently long incubation period the antibody binds the labelled antigen and – in case of a proper stoichiometric ratio – becomes saturated with it. If the unlabelled antigen to be determined is added to this labelled antigen–antibody complex, then it "displaces" the labelled antigen from the complex by competing with it. The extent of the displacement depends on the amount of the unlabelled antigen (hormone). If the concentration of the antigen increases, the radioactivity of the complex decreases and the amount of the displaced, i.e. free, labelled antigen increases. The labelled antigen bound to the complex (B) and the free labelled antigen (F) are separated by a suitable method (e.g. chromatography, filtration) and their amounts are independently determined. With the help of the B/F ratio a calibration curve may be obtained from which the amount of the antigen in the sample may be determined.

2. Application at a molecular level. The mechanisms of many of the complex *chemical reactions* (synthesis, decomposition of various substances) occurring in living systems may be elucidated with the use of radioactive isotopes. In investigations carried out to reveal the genetic code, for instance, the meaning of the UUU triplet was solved by adding poly-U as synthetic mRNA to as many cell-free protein-synthesizing systems as the number of the existing amino acids. Each of these cell-free systems contained not only the mRNA, but also all the amino acids required for protein synthesis, but in every system a different amino acid was labelled with $^{14}_{6}C$. Based on the information carried by the poly-U the same polypeptide was synthesized in all the systems, but the product was radioactive only in the system in which phenylalanine was labelled. It could be proved by this series of experiments that the UUU triplet corresponds to the genetic code of phenylalanine. In even more complex experimental systems the codes of the other amino acids were elucidated by means of similarly applied radioactive amino acids.

Labelling with radioactive isotopes is frequently used in *pharmacokinetics.* The uptake, distribution and elimination processes of drugs are easily followed via the tracer method. From a knowledge of the site and rate of decomposition and binding, information is obtained about the mechanism of action.

3.6. Therapeutic applications

The biological effects produced by radiation may also be used for *therapeutic purposes,* especially in tumour therapy. The basis of this therapy was the recognition that rapidly dividing cells with simple (undifferentiated) structure are more sensitive to radiation than

strongly differentiated cells undergoing slow division. From this aspect healthy cells too are of different types, but tumour cells usually belong to the former type and can be destroyed without any substantial radiation injury of the healthy cells.

For therapeutic purposes, X-radiation, γ-radiation, accelerated electrons and β-radiation are used. Extensive and promising research work is being carried out on the applications of *neutrons, protons, heavy ions* and *pions.*

Teletherapy. Earlier, only the γ-radiation of the decay products of radium was utilized. Since the appearance of the artificial radioisotopes, mainly $^{60}_{27}$Co and more rarely $^{137}_{55}$Cs, which are relatively simple and cheap to produce, *cobalt* and *caesium units* have become widespread (Picture 3.2 in the Supplement).

The most important radiophysical data of the used radiation sources are presented in Table 3.8. The radiosources emit γ-photons of various energies, but the Table gives only the effective photon energies, which are usually a sufficient guide with respect to their use. Each radiosource also emits β-radiation (the radium series contains α-radiating members too), but this is stopped by the wall of the irradiation head containing the radiosource, which results in the emission of X-rays and secondary electron radiation. In the preparation of a careful irradiation plan these types of radiation too must be taken into consideration.

Table 3.8. Radiation data of gamma-unit charges

Radiation source	Half-life (year)	Effective photon energy (MeV)
Radium series	1600	0.75
Cobalt-60	5.27	1.25
Caesium-137	30.1	0.662

The advantage of γ-radiation sources over X-ray tubes is that they emit harder photons of definite energy (line spectrum), which permits an increase of the relative depth dose[9] and simultaneous decrease of the lateral scattering. The operation of γ-sources is simple and independent of the electric power supply. The load of γ-units ranges between TBq and PBq.

The tissues to be treated are situated at a distance of 20–80 cm or more from the radiosource. This is the reason for the high load since the source must deposit a sufficient dose at this distance.

If the distance *RS* between the radiosource and the skin is increased, the absolute value of the dose decreases on the body surface and within the body, but the percentage depth dose increases. This is understandable by reference to Fig. 3.18. The distance *RS* in one case is 10 cm, and in the other 20 cm, while in both cases the centre *P* under treatment lies 10 cm below the skin. If, for the sake of simplicity, the absorption is disregarded, and it is considered that the intensity of radiation for a point-like source decreases with the square of the distance, it is easy to see that in the first case the depth dose is 25% of the skin dose, whereas in the second case it is 44.4%.

[9] The relative depth dose is defined as the ratio of the dose taken by the focus at a certain depth below the skin to the dose measured on the surface of the body (skin dose), usually expressed as a percentage.

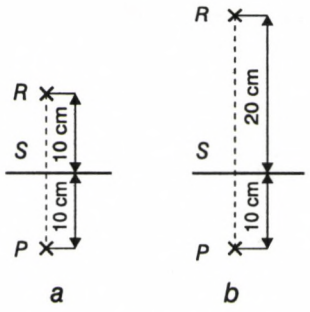


Fig. 3.18. Diagram relating to the increase of the relative depth dose

The spreading use of high-energy photons is justified by the very favourable dose distribution produced in many cases. The effective range of photoelectrons, Compton electrons and electron–positron pairs (together called secondary electrons) produced by sufficiently hard X- or γ-radiation is several mm, or even a few cm in the tissues. In such cases the dose distribution may be influenced substantially not only by the photon scattering, but also by the *scattering of the secondary electrons.* Figure 3.19 illustrates the situation arising close to the air–skin boundary. The dashed lines indicate the boundaries of the photon beam, the curved lines are the secondary electron tracks along which they cause ionization, and the points denote the sites of emission of secondary electrons. The density of these sites is higher in the tissues than in the air, and as a result more electrons pass from the tissues into the air than in the opposite direction. (A similar phenomenon may naturally occur wherever media absorbing X-radiation or γ-radiation to different extents are in contact with each other, for instance on the boundary surfaces between the bones and the soft tissues.)

The dose distribution in the tissues is demonstrated in Fig. 3.20 on actual examples. The abscissa gives the depth of penetration measured from the surface of the body, and the ordinate the percentage depth dose. Curve *a* relates to 0.2 MeV photons, and curve *b* to 22 MeV photons. In the first case the dose decreases monotonously with the distance from the surface, which is a consequence of the divergence and extinction of the radiation beam. These factors are also effective in case *b* but here an additional effect is observed. As a consequence of the secondary electron scattering, on progressing from the surface towards the tissues deeper below the skin the dose first increases, and the decrease

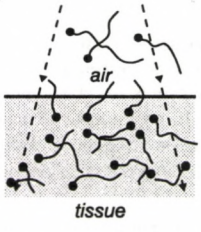


Fig. 3.19. Scattering of secondary electrons in the vicinity of a boundary surface

becomes predominant only later. Instead of a monotonously decreasing function, the curve obtained exhibits a *maximum.* It may readily be seen that the peak in the curve appears the later, the longer the stopping path length of the secondary electrons, i.e. the harder the radiation. With harder radiation the shift of the maximum towards greater depths is further enhanced by the fact that the secondary electrons are increasingly scattered in the forward direction, i.e. in the direction of the incident beam. (This asymmetry is not shown in Fig. 3.20.) For instance, with 22 MeV photons the maximum lies at a depth of 4–5 cm, and with 35 MeV photons at a depth of 6–7 cm below the surface. In principle the formation of a maximum is also to be expected with softer radiation, but it is so close to the surface that it practically coincides with it. By a proper selection of the hardness of the radiation, the dose maximum can be obtained at the focus to be destroyed, and this receives a sufficiently high dose without injury to the healthy tissues close to the surface.

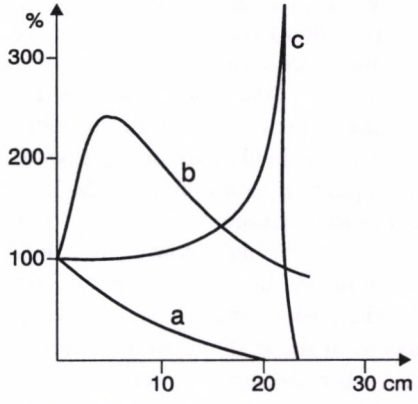


Fig. 3.20. Relative depth dose as a function of depth

Similar phenomena also occur with *high-energy electrons.* In this case too the dose maximum shifts towards the deeper tissues, the higher the energy of the radiation applied. With fast electrons, however, after the maximum is attained the decrease is faster than in the case of photon irradiation. In some therapeutic applications, therefore, accelerated electrons are used instead of photons, in order to protect the deeper tissues.

This latter advantage is even more marked with *heavier charged particles,* e.g. with monochromatic proton radiation (curve *c* in Fig. 3.20). This is connected with the rapid increase in the ionizing power towards the end of the path of the protons, followed by a sharp decrease (cf. Fig. 3.1).

The above dose distribution types may also be observed if fast neutrons are applied. The outlined characteristics of the dose distribution are due to the scattering of the protons. The position of the dose maximum in the tissues lies the deeper, the higher the energy of the neutrons used for irradiation, i.e. the higher the energy of the protons inducing the ionization.

Contact methods. For irradiation, the radiosource is either fastened to the surface of the body, or introduced into the pathologic tissue or into the natural or pathologic cavities (close to the pathologic tissues) of the organism. The great variety of artificial radioisotopes allows in every case the selection of the isotope with the most favourable radiation parameters. Because of its short effective range, β-radiation is used only when the thickness of the tissues to be treated is of the order of at most a few mm. It is a basic requirement that radiation should strike the pathologic tissues in a uniform *distribution,* so that the dose is therapeutically optimum and at the same time an *unnecessary radiation* load to the surrounding healthy tissues is avoided.

For contact irradiation both unsealed and sealed radiosources are used. The *sealed* sources are preparations from which the radiating substance, if applied properly, cannot escape by evaporation, powdering, dissolution or abrasion. If merely one of these conditions is not satisfied, the radiosource is unsealed. Examples of sealed sources are the needle, rod or pearl-shaped solid $^{60}_{27}$Co or $^{137}_{55}$Cs sources, which are positioned according to a carefully prepared therapeutic plan on the surface or in the cavities of the body. *Radioisotopes in solution,* on the other hand, are unsealed sources: for instance, a colloidal solution of $^{198}_{79}$Au, which is administered by infiltration to the tumour (in the event of an appropriate colloidal grain size, the $^{198}_{79}$Au remains at the site of infiltration).

As an example for the therapeutic application of the solution of radioactive isotopes the radioiodine treatment of the hyperactivity of the thyroid gland (hyperthyreosis) is mentioned. For therapeutical purposes a few hundred MBq ^{131}I is administered orally to the patient, which corresponds to a therapeutical dose of 70–100 Gy. The iodine becomes accumulated in the thyroid gland within a few hours, thus in 24 hours 70% of the administered amount are found in the thyroid gland. First of all the energy of the β-particles originating from the radioiodine is absorbed and has a therapeutic effect. As long as the radioiodine dwells in the body of the patient (about 5 days), the patient – mainly his/her thyroid gland – is radiating and excretes radioiodine. From this it follows that the patient has to be considered partly as a radiation source (e.g. at a distance of 1 m from a ^{131}I preparate of 370 MBq the dose rate is 22 μSv/h!), partly as a radioactive contaminating source. Thus the rules of radiation protection apply for both the patient and his/her environment. It is generally valid that the application of therapeutic radioactive preparations is regulated by *the prescriptions of radiation protection* which must be strictly observed by everybody.

REFERENCES

Books

Greening, J. R., Fundamentals of Radiation Dosimetry. Adam Hilger Ltd, Bristol (1981)

Greening, J. R. (ed.), Medical Physics (Proceedings of the International School of Physics Enrico Fermi 1979). North-Holland Publ. Comp., Amsterdam (1981)

Krieger, H., Petzold, W., Strahlenphysik, Dosimetrie und Strahlenschutz. B. G. Teubner, Stuttgart (1992)

Parker, R. O., Smith, P. H. S., Taylor, D. M., Basic Science of Nuclear Medicine. Churchill Livingstone, Edinburgh (1978)

Pohl, R. W., Optik and Atomphysik, 12. Auflage. Springer-Verlag, Berlin (1976)
Rollo, F. David (ed.), Nuclear Medicine, Physics, Instrumentation and Agents. C. V. Mosby Company. St. Louis (1977)
Schlungbaum, W., Flesch, I., Stabell, U., Medizinische Strahlenkunde. Walter de Gruyter, Berlin, New York (1994)
Suess, M. J., Benwell-Morrison, D. A. (eds.), Nonionizing Radiation Protection, 2nd edition. Bureau of Radiation and Medical Devices, Environmental Health Directorate, Health and Welfare Canada, Ottawa (1980)

Paper

1990 Recommendations of the ICRP, *Annals of ICRP, 21,* 1 (1980)

4. MICROSCOPIC AND SUBMICROSCOPIC METHODS IN BIOLOGICAL STRUCTURE ANALYSIS

A common feature of the methods of structure analysis is that the specimen is subjected to some effect and the consequences are subsequently investigated. The analysis can be extended to changes occurring in the agent, the sample, or both. The possible agents include thermal effects, magnetic, electric or other fields, and the sample is often exposed to a radiation or particle beam.

The latter technique of structure analysis is a particularly wide-ranging one, since various radiation and particle types can be used, and numerous interactions with the sample may be produced. Examples are illumination with light or irradiation with X-rays, electrons, ions or atoms, where the interactions may result in scattering, diffraction or absorption, or the emission of radiation or a particle different from the incident one. These methods include studies with the light microscope based on light scattering, which have a long tradition in biology and medicine, while a more recent application is the study of the energy spectrum of electrons produced in the sample by X-ray irradiation.

The methods of structure analysis are undergoing continuous development and new techniques are constantly emerging. A variety of possibilities is of great advantage, for the different methods reveal different properties of the system under investigation, and provide mutually complementary information.

The development is continuous also in the processing of experimental data and imaging. In this chapter we shall not deal with this, only with the theoretical bases of the different methods. The general questions of electronic signal processing and imaging will be discussed in Chapter 6.

4.1. Traditional light microscopes

4.1.1. Construction of the generally used light microscope

The observed objects (and their details) appear to be the larger, the greater the visual angle associated with them (Fig. 4.1). The *visual angle* (φ) is defined as the angle enclosed by the rays arriving from the outermost points of the object and traversing the central nodal point (C) of the eye. The smallest visual angle at which two object points can still be distinguished with the naked eye (without any special aid) has been found to be 1 angular minute. If the visual angle of two points is smaller than this, they cannot be resolved by the eye, which in this case sees only one point instead of two. If the object is

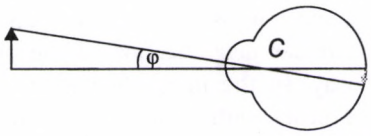


Fig. 4.1. Visual angle

at the distance of clear sight (25 cm), two points can be distinguished only if they are at least 70 μm apart, since they then subtend an angle of 1 angular minute from 25 cm. With simple magnifying glasses the dimensions that can be observed may be reduced by one order of magnitude, while with combined magnifiers, in other words with the light microscope the reduction is of several orders.

The essential parts of the common light microscope are:

(a) the illuminating system, which projects the light through a convex lens system, the *condenser,* onto the object;

(b) the *objective,* a convex lens system directed onto the object;

(c) the *eyepiece,* another convex lens system, through which the produced image is observed. The image formation is demonstrated in Fig. 4.2. For simplicity the exact construction of the image is omitted, and the lens systems are depicted as thin lenses. F_1 and F'_1 are the focal points of the objective, while F_2 and F'_2 are those of the eyepiece. The distance between the points F'_1 and F_2 is usually called the optical tube length and is denoted by d. The objective is brought nearer to the object *(AA')* so that it is situated outside but close to the focal point F_1. In this way the objective produces from the object a magnified, reversed image *(BB')* on the opposite side and at a large distance relative to the focal length. This image is viewed through the eyepiece, which acts as a simple magnifying glass. The image produced by the objective lies within the focal length of the eyepiece, close to the focal point F_2. From the first image of the object the eyepiece produces a virtual, magnified and erect image *(CC')*. Thus, in *the microscope a virtual, magnified and reversed image of the object is observed.*

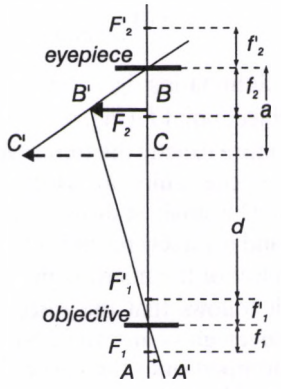


Fig. 4.2. Image formation in the microscope

The magnification of a single lens is equal to the ratio of the image and object distances. Consequently, the linear magnification of the outlined microscope can be estimated in the following way. In the image formation of the objective the object distance is nearly equal to the focal length f_1 and the image distance is nearly equal to the optical tube length d of the microscope, which gives the magnification N_1 of the objective as d/f_1. In the image formation of the eyepiece the object distance is approximately equal to the focal length f_2 of the eyepiece, and the image distance is $-a$, where a (= 25 cm) is the distance of clear sight. (The negative sign indicates that the virtual image CC' is on the same side of the eyepiece *as* BB'.) It follows that the magnification N_2 of the eyepiece is equal to $-a/f_2$. Thus, the total magnification is

$$N = N_1 N_2 = -\frac{da}{f_1 f_2} \qquad [4.1]$$

The magnification of the microscope (as an absolute value) will be the higher, the smaller the focal lengths of the objective and the eyepiece.

4.1.2. Resolving power of the light microscope

The microscope reveals many fine details of the object, which could not be observed with the naked eye. However, even with properly corrected lens systems of small focal lengths, the resolution of small details is limited by the wave nature of light. The problem cannot be solved with the use of more lens systems, since the details of the object are revealed by the objective lens. The eyepiece does not add new detail, only magnifies the image produced by the objective to ensure good observation of the already revealed details. This task is satisfactorily achieved by the eyepiece without too large magnification. The use of further lens systems adds no new information and is consequently superfluous.

We next deal with the smallest distance δ between two object details (also called object points) still distinguishable in the microscope. The considerations originating from *E. Abbe* (1870) lead (in agreement with practical experience) to the equation

$$\delta = 0.61 \frac{\lambda}{n \sin\omega} \qquad [4.2]$$

Here λ is the wavelength of the illuminating light, ω is *half the aperture angle,* in which the objective lens is seen from the object point P (Fig. 4.3), and n is the refractive index of the medium between the object (or more exactly the thin cover glass covering the object) and the objective. δ is obtained in the units in which λ is measured. The quantity $n \sin \omega$ is the *numerical aperture.* The smallest distance still resolved by the microscope is proportional to the wavelength and inversely proportional to the numerical aperture. The reciprocal of δ is the *resolving power* of the microscope.

From the above discussion it follows that the resolving power can be increased by filling the space between the cover glass and the objective lens with a liquid of high refractive index. The liquid is dropped onto the cover slide, and the frontal lens of the objective is immersed in this drop. The objective in this case is the *immersion objective.* Of the liquids used, the refractive index of cedar oil is 1.51, and that of monobromonaph-

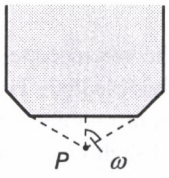


Fig. 4.3. Aperture angle of the objective (2ω)

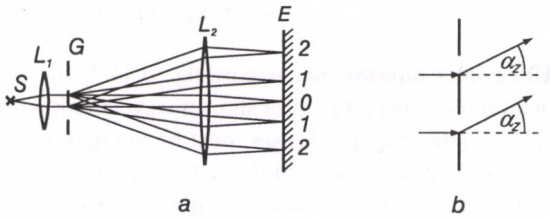


Fig. 4.4. Diagrams relating to the resolving power of the microscope

thalene is 1.66. Since the maximum value of ω in practice is about 70° (sin 70° ≈ 0.95), the smallest distance still resolved for visible light (λ = 450 nm) is approximately 200 nm.

Equation [4.2] can be obtained from the diffraction of light. Consider Fig. 4.4*a*, which demonstrates the diffraction of light by a diffraction grating. The slit-shaped light source S is perpendicular to the plane of the drawing. The rays from the light source pass as a parallel beam through the lens L_1 and arrive at the lens L_2, which projects the image of the slit onto the screen E situated in its focal plane. If a diffraction grating (G) is placed between the two lenses (the slits of the grating are parallel to the slit-shaped light source), the image will be multiplied as a result of diffraction: on both sides of the original central image (also called *zeroth order image,* or *main maximum*) first, second, etc. order diffraction images *(subsidiary maxima)* will appear. For light of wavelength λ these latter images are formed only at those angles α_z (z = 1, 2, ...) which satisfy the equation

$$\delta \sin \alpha_z = z\lambda, \quad \sin \alpha_z = z\frac{\lambda}{\delta} \quad (z = 1, 2, ...) \qquad [4.3]$$

For illustration of [4.3], Fig. 4.4 shows only two adjacent slits and only one pair of diffracted rays propagating in the same direction. δ is the grating constant. The first subsidiary maximum is at the angle α_1, the second at α_2, etc., these angles being defined by the equations $\sin \alpha_1 = \lambda/\delta$, $\sin \alpha_2 = 2\lambda/\delta$, etc.

Figure 4.4*a* may be regarded as a model of microscopic image formation. L_1 is the condenser and L_2 the objective. The object is a diffraction grating. In the focal plane of L_2 a diffraction pattern is produced as described above (even without the screen). The real image of the object (in the present case the grating) appears at some further distance not shown in the diagram. The image of the object is produced by the rays responsible for the diffraction image of the light source in the focal plane of L_2. An interesting phenomenon is obtained if the higher order images are covered by a diaphragm in the focal plane. The more the higher order images are covered, the less perfect the image of the grating. If every subsidiary maximum is covered, and the image is produced only by the primary beam passing through the main maximum, the image of the grating becomes unrecognizable: the lines of the grating cannot be distinguished. The illumination of the field of sight is uniform, and no details of the grating can be observed. In order for the structure of the grating to be seen, it is necessary that not only the rays passing through the main maximum but also the rays propagating through the first order diffraction image participate in the image formation. However, this condition is satisfied only if the grating constant (δ) is not too small. This can easily be understood, because in the case of a too small δ the rays arriving from the grating and propagating towards the first order (z = 1) diffraction image would enclose such a large angle with the lens axis that they would not fall on the lens. It follows that angle α_1 associated with the first order diffraction image must not be larger than the half aperture angle ω of the objective; its maximum value is $\alpha_1 = \omega$, and hence $\sin \alpha_1 = \sin \omega$. For a given ω and λ, the lines of the grating become visible only for the grating constant δ which satisfies the equation $\delta = \lambda/\sin \omega$. With more exact reasoning, δ is given by $\delta = 0.61\ \lambda/\sin \omega$. It has so far been assumed that the space between the object and objective is filled with air, and λ indicates the wavelength in air. In a medium with refractive index n the wavelength will be λ/n. Consequently, if there is a medium of refractive index n between the object and the objective, λ/n should be used instead of λ, which leads to [4.2].

4.1.3. Special light microscopes

1. Ultraviolet microscope. According to [4.2] the resolution of the microscope can be improved if ultraviolet light of shorter wavelength is used instead of visible light. However, with this method quartz lenses must be applied, since glass absorbs ultraviolet radiation. The image is not visible with the naked eye; if the eyepiece is raised slightly, a real image is produced and can be displayed on a luminescent screen or a photographic plate.

2. Ultramicroscope. With the usual illumination the object details appear as more or less dark domains in the otherwise illuminated visual field. However, if the object is illuminated so that only the rays diffracted by the details of the object can enter the objective (which is equivalent with the covering of the main maximum), the boundary lines of the object details become visible as bright spots on a dark background. Picture 4.1 (in the Supplement) shows photographs of the same biological object taken with a normal light microscope (*a*) and with an ultramicroscope (*b*). In the ultramicroscope, particles become visible whose size is below the resolving power. These small particles shine against the dark background like stars in the nocturnal sky. *Details are not visible; only their presence, position and motion can be observed.* The conditions discussed above for the resolving power of the microscope hold here too for the distinction of two such particles. The particle is the better discerned, the more its refractive index differs from that of the environment. For instance, the conditions for the observation of protein particles are less favourable than those for metal colloids. In this latter case even particles 10 nm in size can be well observed; the name ultramicroscope refers to these favourable conditions.

3. 3D condenser. The quality of the microscopic image (within the limits of resolution) is strongly influenced by the illumination of the object. The observations mentioned in the previous point relate to this fact, and the principles of the *3D* condenser too are based on it. With the aid of this condenser the axial illumination is combined with lateral illumination. This system results in increases in the plasticity and contrast of the image. The notation *3D* refers to the three dimensional character of the image obtained.

4. Phase contrast microscope. The observation of certain properties of various objects is made possible by phase contrast microscopy. The eye can distinguish only details which differ from each other either in illumination or in colour. This is clearly due to the details of the object absorbing the illuminating light in different ways. However, the object may have properties which do not give any contrast of illumination or colour; for instance, the various parts of the object may differ in thickness or refractive index and otherwise transmit the light with almost no absorption. The phase contrast microscope allows observation of these properties by converting them into differences of illumination (cf. Picture 4.1*c* in the Supplement). Without going into details, we shall mention only that for this purpose a disc or ring-like small transparent plate of suitable thickness and a diameter of a few mm, the *phase plate,* is placed in the focal plane of the objective, and the illumination is modified accordingly. Every microscope can also be used as a phase contrast microscope if its usual objective is replaced by an objective provided with a phase plate.

5. Polarization microscope. The microscope constructed for the study of birefringence contains not only the lens systems of the ordinary microscope but also two *polarization filters.* One of the filters operates as a polarizer and the other as an analyzer. The polarizer is situated below the condenser lens, and the analyzer above the objective. The analyzer can be rotated around the light beam as axis. In this way its plane of polarization can be changed from the parallel to the crossed position relative to the polarizer. The object stage, provided with an angular scale, can also be rotated. The examination is carried out by rotation of the filters into the crossed position prior to insertion of the object, the field of view thereby becoming dark. On insertion of the object and rotation of the stage, the birefringent details of the object become bright in certain positions and then darken on further rotation (cf. Picture 4.2 in the Supplement).

6. Luminescence microscope. Luminescence can be used in microscopic examinations, because most organic compounds, including those found in the living organism, emit visible luminescent light characteristic of their chemical structure when illuminated with ultraviolet light. The different compounds generally display different colours when luminescing, and thus the various cells or cellular components with different chemical compositions can be well distinguished in the section. Besides the intrinsic luminescence, another phenomenon can also be used: cells and tissues adsorb the luminescent dye from extremely dilute (1:1,000–1:5,000,000) aqueous solutions of these compounds *(fluorochromes),* and the adsorbed dyes emit luminescent light of various colours, depending upon the cell or tissue type. The dilute solution does not affect the structure or functions of living cells.

The great advantage of both luminescent methods over the classical microstaining procedures is the avoidance of the rough chemical effects of fixing and staining and of the possible change of the living tissues. A further advantage is the possibility to examine biopsy samples within a few hours. The method can also be used to reveal the presence of acid-resistant bacteria.

The construction of the luminescence microscope is very similar to that of the ordinary microscope, the only difference being that the condenser and the slides are made of special glass transmitting ultraviolet light; further, usually in the frontal piece of the objective lens system, a filter is used which filters out the ultraviolet light transmitted by the object.

The object is normally illuminated by a mercury vapour lamp, or an arc lamp with metal electrodes. The visible light of these lamps is filtered out to prevent interaction with the luminescent light.

We have so far spoken merely of ultraviolet illumination, without regard to its spectral composition. However, it is a well-known fact that ultraviolet (and sometimes also visible) light of various frequencies may excite different domains of the object. Consequently, with the application of suitable filters, transmitting the exciting light only in a narrow frequency band, additional fine details may be distinguished. The possibilities are even wider, because the luminescence can be excited not only by ultraviolet light, but also by short wavelength visible light.

7. Binocular microscopes. Microscopic observation is more convenient and less tiring for the eye if both eyes can be used. This may be achieved with the binocular microscope, which has only one objective but two eyepieces. The rays passing through the objective are separated into two beams by a semitransparent and semireflecting prism before the formation of the real image. Each of the separated beams produces an image which can be observed with the two eyepieces. The distance between the eyepieces can be adjusted to accord to the individual interocular distance.

8. Stereomicroscope. Three-dimensional images are obtained with the stereomicroscope, which contains two objectives and two eyepieces. This, microscope consists essentially of two microscopes built together. One of the microscopes forms the image of the object from a little to its left, and the other one from its right side, in this way two slightly different images being obtained. If one of these images is viewed with one eye and the other image with the other eye, a realistic three-dimensional image is seen. Stereomicroscopes can be used only at low magnification (at most 100×), for at higher magnification the focal depth is small and stereoscopic observation becomes impossible.

4.2. Traditional electron microscopes

In case of electron microscopes – as indicated by their name – the image is produced by electrons instead of light. Two main types can be distinguished: the transmission electron microscope (TEM) and the scanning electron microscope (SEM). They were developed in the 30s and 60s, respectively.

4.2.1. Transmission electron microscope

1. Electron lenses. In case of optical lenses light can be focused by light refraction, while with electric and magnetic lenses – jointly referred to as electron lenses – the path of electrons is influenced by electric or magnetic fields. Electron lenses are electric or magnetic fields focusing at one point the electrons arriving from one point.

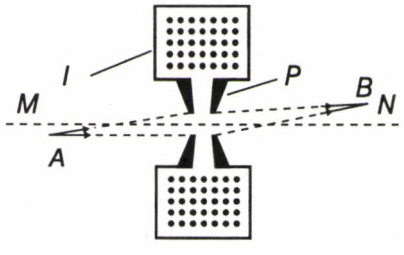


Fig. 4.5. Schematic drawing of a magnetic lens
The black points indicate the intercept of the coil windings with the plane of the drawing. The straight line *MN* indicates the lens axis; the electron beam arriving from the object point *A* is focused at *B*

Nowadays mainly *magnetic lenses* are used in electron microscopes. They are produced by coils in which electric current flows (Fig. 4.5). The electrons travel through a circular slit surrounded by the coil. The coil is covered by an iron coating (*I*) with a gap in its inner side. This arrangement allows the production of a strong magnetic field of cylindrical symmetry within a small space. The focal length can be decreased by inserting iron pole pieces (*P*) into the gap, since this increases the field strength.

Besides these factors, the focal length also depends upon the electron velocity. A smaller velocity results in a smaller focal length, and vice versa. The accelerating voltage is usually several ten thousand volts, and in this case the focal length is a few mm.

The relations of image formation used in light optics hold for electron lenses too. The aberrations of image formation are similar in the two cases. Even chromatic aberration has its electron optical counterpart, and becomes observable if the velocity of the electrons is not homogeneous. Using magnetic lenses a further aberration is due to the spiral path of the electrons in the magnetic field. Thus the image is usually rotated with respect to the object around the lens axis. The aberrations for electron lenses can be decreased only by applying a monochromatic electron beam enclosing a small angle with the lens axis.

2. The construction of the electron microscope is similar to that of the light microscope (Fig. 4.6). The light source is substituted by an *electron source,* which is an electrode system producing electrons with identical high energy. The electron beam leaving the source and accelerated is focused by the electron lens (condenser) onto the object. The electrons are scattered to various extents by the details of the object (Fig. 4.7). Some of the scattered electrons are arrested by a circular aperture. Fewer electrons are transmitted by the aperture from the details scattering more strongly, and a greater number from those scattering more weakly. After the aperture the electron beam passes the objective lens,

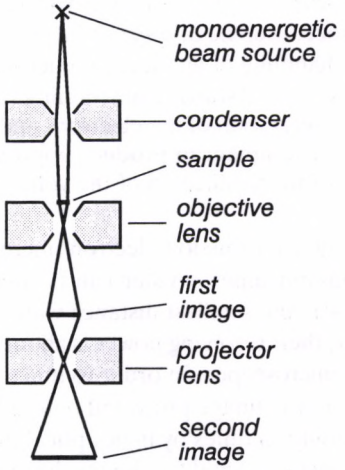


Fig. 4.6. Outline of an electron microscope constructed from magnetic lenses

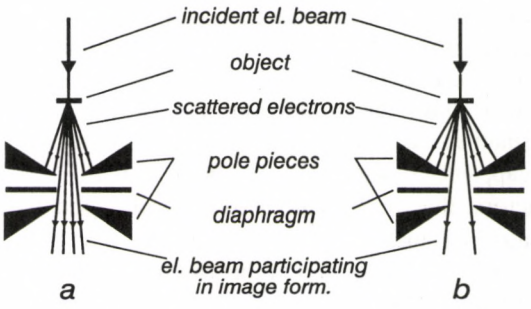


Fig. 4.7. The role of scattering in the resolution of details
Electron beam for a weakly scattering detail (*a*) and a strongly scattering detail (*b*)

which focuses in one point the divergent beam arriving from the object and, depending on the adjusted object distance, produces a real and magnified image of the object. The image can be made visible on a luminescent screen or a photographic plate. Lighter or darker spots are produced, depending upon the stronger or weaker scattering of the electrons on the object points. In this way well-contrasted images are obtained, the details of the object being observable with the human eye. However, a screen is placed in the image plane of the objective only rarely, for control purposes, and the image is magnified further by the projector lens, which from the first image produces a magnified, real image on a luminescent screen or photographic plate. With recently developed electron microscopes it is possible to display the image on a screen by means of television techniques. The position of this final image with respect to the object is practically immaterial; if magnetic lenses are used, the image is rotated. In contrast to the light microscope the electron microscope is a closed system. It requires a vacuum suitable for eliminating or at least minimizing collisions of the electrons with air or other molecules. The object and the photographic plate are placed in the microscope via an "air-lock". Although the vacuum deteriorates by a few millibars during this procedure, the original situation may be restored with a vacuum pump in a few minutes. The handling of an electron microscope requires much care, and for its operation the continuous, undisturbed functioning of several subsidiary devices is necessary. For instance, the acceleration of the electrons and the operation of the electron lenses require special electric equipment to produce appropriately high voltages. Special care must be taken with regard to stabilization of the voltage.

3. Resolving power. With the transmission electron microscopes generally used, object details approximately a thousand times smaller can be resolved than with light microscopes dealt with in section 4.1 (the smallest distance which can be resolved at present is about 0.2 nm). Consequently, their resolving power is about a thousand times higher than that of the well-known light microscopes. In order to make the resulting resolved details observable with the naked eye, the image projected onto a luminescent screen or photoplate is magnified further about ten times by light optical means.

Otherwise a similar relationship is valid for the resolving power of the TEM as for that of the light microscope. However, in case of the TEM the place of the wavelength of light

is taken by the *wavelength of the so-called matter wave* of the electrons (cf. section 1.1). This wavelength is inversely proportional to the velocity of the electrons, and e.g. at an accelerating voltage of 50 kV it is about five orders of magnitude smaller than the wavelength of the visible light. Therefore it should be expected that the resolving power may also be increased by five orders of magnitude. In principle it is so. Nevertheless, in practice we have to be satisfied with the already mentioned thousandfold increase. The reason for this is that in order to decrease the lens aberrations we have to use beams of very small aperture, i.e. very small numeric aperture for the electron microscopes. The increase of the resolving power is impeded also by the fact that it is not easy to prepare sections which are appropriately thin and at the same time distortionless, furthermore it is difficult to prevent the damage of the fine object during preparation.

In the field of particle and nuclear physics the name *supermicroscope* is also used. These are particle accelerators by which electrons of such high velocity can be produced that the wavelength of the matter wave belonging to them is shorter than the size of nuclei and even of protons and neutrons. This high resolution led, e.g. to the exploration of the structure of protons and neutrons and to the discovery of their constituents, the so-called *quarks.*

4. X-ray microanalysis. Most of the electron microscopes are also suitable for X-ray microanalysis. Its main point is that, by bombardment of the sample with focused electron beam, a characteristic X-ray is produced. From the X-ray spectrum the chemical composition and the amount of the components of the sample can be determined. Fixed and scanning modes of operation are available in this case, too. By using the scanning mode the local changes in the composition of the sample can also be revealed.

4.2.2. Scanning electron microscope

It is a surface-exploring instrument; its essence is the following: by means of an electron lens a focused, needle-like electron beam is scanned on the surface of the object just as the electron beam moves in a cathode-ray tube (e.g. the video tube). The electrons impacting with a high velocity release secondary electrons from the surface of the sample the number of which is usually different at the various points of the surface. The electric signals thus produced and amplified modulate the intensity of the electron beam of a television monitor. The beam of the monitor is run synchronously with the beam scanning the object, thus an image of the object is formed on the display which is determined by the distribution of the secondary electrons released from the surface atoms. The image is amazingly beautiful and plastic.

The magnification and resolving power of the scanning microscope are smaller than those of the TEM, but the plasticity of the image frequently makes up for this disadvantage. A great advantage of the SEM that *lenses are not required for the imaging.*

Picture 4.3*a* in the Supplement shows a photo made by a TEM, while photos in Pictures 4.3*b-c* made with SEM are seen.

There are also combinations of the transmission and scanning electron microscopes; these are rather sophisticated instruments, but they have both a high resolving power and the appearance of three-dimensionality.

Another remark. The SEM is suitable also for *X-ray microanalysis.* The essence of the latter is that by means of the focused electron beam characteristic X-ray is elicited from the sample and the spectrum of this is analysed. From the spectrum conclusions may be drawn concerning both the quantitative and qualitative relations and the local changes of the chemical elements.

4.3. Novelties in the field of microscopes

a) At the beginning of the 80s a new kind of the scanning microscope has been developed which differs from the previous ones even in its principle and which is remarkable for its efficiency, theoretical and practical simplicity and elegance alike. This is the *scanning tunneling microscope (STM) based on the quantum mechanical tunneling effect* (cf. section 1.5.4, point 2).

According to the quantum mechanics, the "shared" electrons (cf. section 1.3.1) of a metallic object are found with some probability also on the external side of the electric potential barrier covering the surface of the metal, although their energy is not enough for "jumping over" the barrier. The barrier behaves as if it would be crossed by tunnels. Consequently, the surface of the metal is covered with an electron cloud consisting of "tunnel electrons". The thickness of the cloud is in the order of tenth nm, its density decreases rapidly with the distance from the surface. The relief of samples with electrically conducting surface can be examined by tunneling microscope. Thus *there is no need for an external electron source* in the imaging, the tunnel electrons are used for this purpose. In the case of non-metallic, e.g. biological samples it is sufficient to form by evaporation a fine metal cover with a thickness of some nm and even macromolecules on a metallic surface may be directly examined.

The construction and functioning of the instrument is in principle simple. The surface of the sample is scanned by an especially sharp needle along consecutive lines parallel to each other (Fig. 4.8). (The scanning can be naturally performed so that the sample moves instead of the needle.) The tip of the needle reaches into the electron cloud found on the surface of the sample. If a voltage of some mV is applied between the needle and the

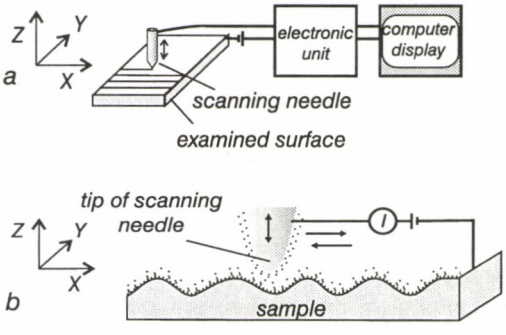


Fig. 4.8. Working principle of STM
a: main functional units; *b:* site of the essential process.
Arrows indicate the direction of needle movement. The points refer to the electron cloud

sample as electrodes, an electron current of nA intensity, the so-called *tunnel current* is produced. Since the density of the electron cloud decreases rapidly with the distance from the surface, the intensity of the tunnel current changes also sensitively according to the gap between the tip and the surface. As the tip scans the surface of the sample (in the X–Y plane in the case outlined in Fig. 4.8), an electronic unit senses the tunnel current and if this changes, the needle is moved perpendicularly to the surface by means of a feed-back so that the intensity is restored to its original value. Thus the tip follows the relief of the sample by moving up and down along the Z axis. The needle is placed at the end of a piezo crystal which makes possible this up-and-down movement. The movement of the tip is processed by a computer and presented on its screen. In the course of the scanning a system of parallel lines is produced which forms a stereoscopical relief with atomic details (cf. Picture 4.4 in the Supplement).

Concerning the *resolving power* of the STM, general rules cannot be given. It depends e.g. on the sharpness of the tip, the local density of the tunnel electrons, and on the geometry of the relief. In the most recent types the smallest distance which may still be resolved is in the order of several hundredth of nm both along the surface and normal to it. We talk about a real case when we mention that about 1 cm on our picture corresponds to a distance of 0.1 nm. Thus the total magnification is about hundred-millionfold.

b) The success of the STM has initiated recently the development of a series of microscopes. Each type belongs to the family of the *scanning microscopes* and – just as the STM – gives information about the atomic details of the surface of a sample. However, the different types reveal the surface from various aspect. From among them the *atomic force microscope (AFM)* will be discussed in the following as the most promising one from a biological point of view. It is based on the perception of the attractive and repelling forces between the surface atoms of the sample and the tip of a needle-like probe. The needle is placed at the end of a fine plate spring, thus it is able to follow the "force relief" of the surface. The movement of the tip is processed by appropriate electronics and on the screen already the image of the relief is seen.

Nevertheless, the sources of information – as already mentioned – are different in the two cases: the STM presents the surface from the aspect of tunnel electrons, the AFM from the aspect of interatomic forces, respectively. In the first case the tip reacts to the changes occurring in the intensity of the tunnel current, which is determined partly by the topography, partly by the electron density of the surface. In the latter case the information concerning the topography of the surface is obtained by following the changes of the interatomic forces.

The method can be applied for samples in air (and other gases or vacuum) and in liquids (water, oil, liquid nitrogen, etc.) alike. From a biological point of view it is especially favourable that the sample may be studied in a *native environment.* The AFM also has the advantage that, while the STM is feasible only with electrically conducting samples (or with those covered in this sense), this restriction does not exist here. The needle may be made also of e.g. quartz or silicon nitride instead of a metal. The method makes possible the individual examination of selected details of biologically important macromolecules (DNA, proteins, etc.) or the following of some processes, such as

transcription, thrombus formation or the penetration of viruses into the cell membrane, at the molecular level (Picture 4.4*a-c*).

With respect to the extreme sensitivity of the instrument it should be mentioned that it makes possible the perception of forces which are smaller by several orders of magnitude than the forces of 10^{-12} N. For comparison: for example the acting forces are in the order of 10^{-7} N in the case of ionic or chemical bonds and in the order of 10^{-12} N in van der Waals bonds, respectively.

c) The family of scanning microscopes also comprises optic microscopes. One of them is the recently developed *near-field optical scanning microscope (NFOS).*

The essential features of the NFOS microscope can be perceived from Fig. 4.9. The thin film-like sample is illuminated in an extreme case only through an aperture of 5–10 nm radius produced in a metal (e.g. Al) film. A laser (e.g. 490 nm argon laser) is applied as light source. Scanning is carried out by moving the sample while the object passes before the aperture at a distance of its radius or even nearer. The constant distance at scanning is ensured by the tunnel current between the sample and the metal film, similarly to the STM described in point (a).

The near proximity of the aperture and the object, expressed also by the name of the method (near-field), is advantageous, even necessary in several respects. Without proof, or detailed explanation of this, only a few essential facts will be given below. The aperture, being small even compared to the wavelength of light, produces extraordinary conditions for the propagation of light. Consequently the more or less absorbing parts of the object influence most the light incident from the aperture in the proximity zone, and this enhances the contrast of the image. The resolution is best in case of objects placed in this zone, corresponding approximately to the diameter of the aperture, i.e. for visible light it is one or two orders of magnitude better than the resolution of traditional optical

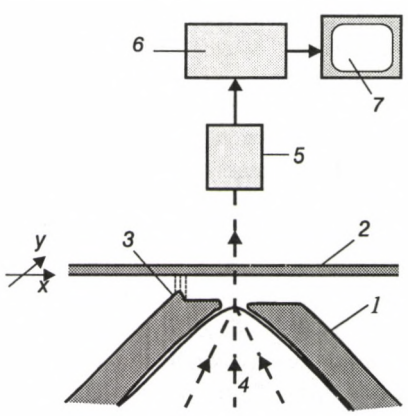


Fig. 4.9. Schematic diagram of the NFOS microscope
1: tapered quartz tip covered by a thin metal (Al) film with a circular aperture in the middle; *2*: sample; *3*: metal microtip and tunneling electrons regulating the distance between *1* and *2*; *4*: illuminating laser light; *5*: photomultiplier; *6*: electronic unit; *7*: display; *x* and *y*: direction of the scanning motion

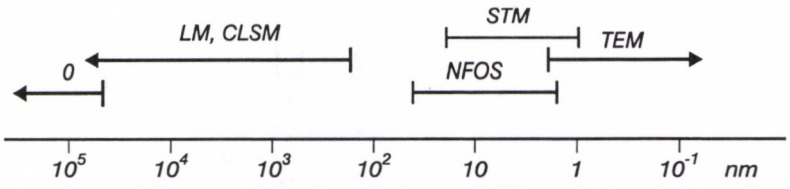


Fig. 4.10. Approximate resolved distances
0: eye; *LM, CLSM:* light microscope, confocal laser scanning microscope;
NFOS: near field optical scanning microscope; *STM:* scanning tunneling microscope;
TEM: transmission electron microscope

microscopes. The proximity of the aperture and the object is indispensable for the production of the tunnel current controlling the distance of the aperture and the object, too. Moreover, a metal tip is also applied in order to decrease this distance locally from 5–10 nm to the required 1–2 nm.

The advantage of the method is that it renders possible the study of such micro-region which up to now has belonged to the "no man's land" between light and electron microscopes (Fig. 4.10). From the biological point of view it is advantageous, too, that the object can be studied in a native environment.

The method, as mentioned already, is still under development. Its extension to the infrared and UV ranges is expected, promising further possibilities.

d) The confocal laser scanning microscope (CLSM) does not scan the surface of an object but rather a selected plane within the object (e.g. a cell). The light illuminating the object – the light of a laser – enters the microscope via a hole diaphragm and reaches a point of the selected plane through the objective (i.e. focused) (Fig. 4.11). The light scattered from this point – or the luminescence light produced at the illuminated point (cf. section 4.1.3, point 6) – is collected by the objective by means of a semitransparent mirror onto another hole diaphragm, i.e. the object-point is projected on the latter. The two diaphragms are at an optically identical distance from the objective (confocal diaphragms). This setup ensures that the object-point to be scanned gets a focused illumination on one hand, and on the other hand, that on the second diaphragm only the image of the object-point is sharp, i.e. mainly the light coming from the object-point gets through the diaphragm to the detector. The detector gives an appropriate signal to the computer even if there are further light-scattering and -absorbing details of the object between the object-point and the objective.

The image of the selected object-plane (in fact, of a thin segment of the object) is produced in course of the scanning from point to point. Both the controlling of the scanning and the appropriate collection and storing of the coordinates and luminosity data of the scanned points are performed by a computer. The scanning is done usually in a plane perpendicular to the axis of the microscope (X–Y plane), and may involve several hundreds of layers parallel to each other (optical slicing).

From the stored data not only the images of the layers in the X–Y plane, but also those of the X–Z layers parallel to the optical axis may be reconstructed. In this way a three-

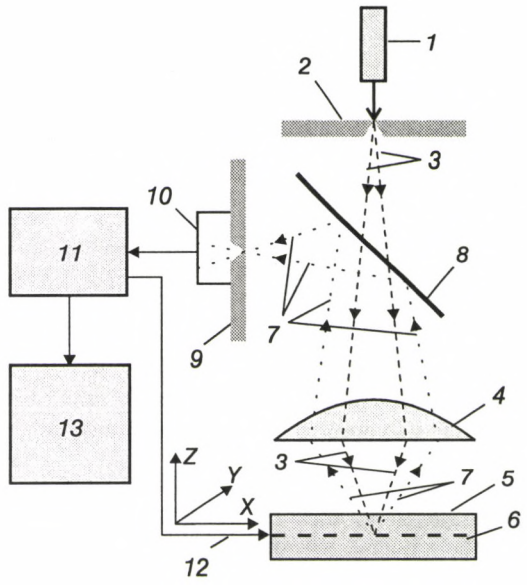


Fig. 4.11. Functioning of the CLSM
1: laser; *2*: illuminating hole diaphragm; *3*: illuminating (exciting) light; *4*: objective; *5*: object; *6*: selected object plane; *7*: reflected light or luminescent light; *8*: semitransparent mirror; *9*: hole diaphragm of the detector; *10*: detector; *11*: computer; *12*: scanning; *13*: screen

dimensional survey of even moderately transparent microscopical objects may be obtained. This method presents an especially significant perspective for the medical and biological research, since for example in case of vital staining by fluorochromes the confocal scanning makes possible the observation of the dynamics of processes taking place in the living cells (cf. Picture 4.5 in the Supplement).

The resolving power of the CLSM is about the same as that of the conventional light microscopes (cf. section 4.1.2).

e) Optical trapping. Optical tweezers. They represent a new possibility for micromanipulations rather than a new type of light microscopes.

In microscopy the light may be used not only as a means for imaging but also as a manipulator. In this respect the biological applications are especially important. Namely the light produces heat, splits chemical bonds, exerts pressure or force on the objects getting in its way. By means of lasers all this may be done at a scale of nm, thus microscopic biological objects, e.g. cells, cell components or biologically important macromolecules may be cut, drilled, welded, moved, etc.

In the following the operation mentioned at last will be dealt with in detail, i.e. for example with the confinement and transportation of suspended particles, with other words with the mechanism of *optical trapping* and *optical tweezers.* The discovery of this possibility in microscopy is only a few years old and is very promising both for the scientific research

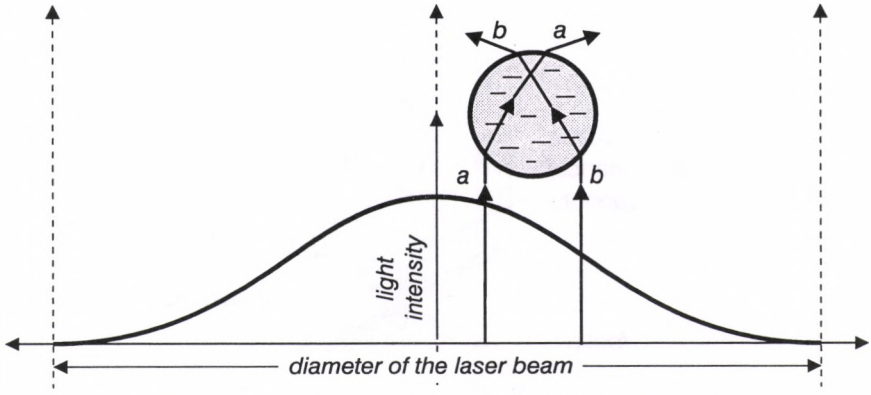


Fig. 4.12. Illustration of the optical trapping. For further details see the text

and practice. The possibilities of the examinations become namely more extensive, and it should be particularly emphasized that the operations are performed in a native environment, without damaging the studied object or influencing the examined process in an undesired way or extent. The laser light is an accurate, gentle and sterile instrument.

The mechanical effects of light, the trapping and tweezer effects under discussion, are not surprising if we consider that the photons have a mechanical momentum. Figure 4.12 is instructive for the understanding the function of the trapping and the tweezers. The lateral dashed lines indicate the edge of a parallel laser beam, while the broken line in the middle stands for the middle, i.e. the axis of the beam. The intensity is the greatest in the middle of the beam, it decreases towards the edges. This is illustrated by the bell-shaped curve. The circle in the figure indicates a transparent globular particle, e.g. a cell, the refractive index of which is greater than that of its environment. Let us consider the direction in which the cell moves under the effect of light. For this, we follow the paths of rays *a* and *b*. Due to the double refraction ray *a* deviates somewhat to the right, ray *b* to the left, respectively. According to the theorem of the mechanical momentum, the deviation of the photons of ray *a* to the right produces an impulse on the cell directed to the left, while the deviation of the photons of ray *b* to the left produces an impulse directed to the right. However, since the intensity is higher in the environment of *a* than in that of *b*, the momentum directed to the left is effective on the cell, the cell moves towards the axis of the light beam, i.e. towards the range of higher intensity. It may be easily seen that in the described case the cell so to say becomes captured in the axis of the beam. Namely if the cell would get to the left side of the axis, forces directed to the right would appear which would divert the cell once again towards the axis. It is true in other cases too that at a light intensity of inhomogeneous distribution a nearly globular object which has a greater refractive index than its environment is diverted towards the *higher intensity* range of the beam.

Figure 4.13 shows a case in which high-intensity light produced in the focus of a lens acts as a trap. The cell which got to the vicinity of the focus and became captured moves thereupon together with the focus, the focus functions as a tweezer.

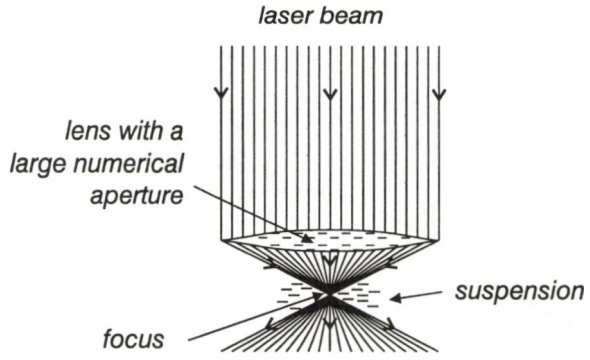


Fig. 4.13. Outline of an optical tweezer. For further details see the text

As examples a few studies will be mentioned, based on the literature, in which the discussed "instrument" was applied:

- the development of a contact between a malignant cell and a killer cell, the penetration of the killer cell through the membrane of the malignant cell;
- microinjection into an intracellular structure and the examination of its effect;
- preparation and isolation of chromosomal segments, e.g. for the analysis of the human genom.

In the case of optical trapping the light absorption (heat production) must be reduced to the minimum by the selection of the wavelength of the light, since this is an exchange of momenta rather than energy transmission. In cytological, microbiological, molecular biological, etc. studies a continuously emitting infrared laser, e.g. Nd-YAG laser emitting at 1064 nm, serves well this purpose. The smallest "light spot" which may be produced by focusing in this case is the size of about 2.5 μm. The depth of penetration is several mm. If we want to damage the studied object deliberately, another laser emitting a light with a proper wavelength should be used, e.g. an argon ion laser emitting green light, or a source emitting UV light, i.e. a laser the light of which becomes absorbed.

4.4. Optical spectrometry

The transition permitted by selection rules between two possible states of atomic systems (atoms, molecules, solids, liquids) may be achieved either by photon uptake (absorption) or by photon release (emission). The energy of the absorbed or emitted photons is equal to the difference between the energies for the two states. These energies may be determined from spectral studies. Information can be obtained on the electronic transitions, and the vibrational and rotational states. Moreover, in an indirect way spectrometry yields still further information, e.g. the following data can be determined:

(a) the geometric positions (distance and directions) of the atoms or atom groups within the molecules;

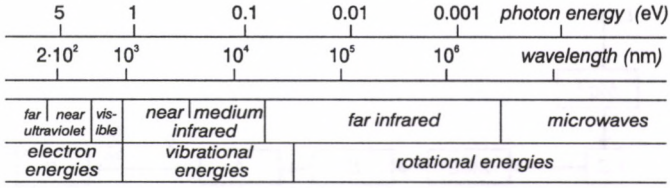


Fig. 4.14. Quantized energy types of molecules, wavelength and photon energy ranges of the spectra corresponding to the energy transitions

(b) the bond strengths between the atoms and atom groups, and the bond types (ionic, covalent bonds, etc.);

(c) the conditions of molecular dissociation, and the energies of dissociation in the ground and the excited states;

(d) conclusions can be drawn about the changes occurring in the molecular configurations in response to environmental changes.

Changes in the electron energies are of the order of magnitude of an electronvolt; the vibrational energy differences are lower by one order of magnitude and the rotational energies are a further order smaller. The electron energies are in the visible and ultraviolet, and the vibrational energies in the infrared range (Fig. 4.14). The rotational energies are comparable with the energy quanta of the far infrared, the microwaves and the radio waves.

4.4.1. Emission spectrometry

The *emission spectrum* is the distribution of the emitted light intensity as a function of wavelength (frequency, photon energy). The spectrum is obtained after excitation of the sample, for instance by heat or electric energy (flame colouring, arc or spark discharge). Optical excitation is frequently applied; this is the *photoluminescent method.* In the event of thermal or electric excitation the substances dissociate into their atoms or ions, such excitation therefore resulting in atomic or ionic spectra. Molecules can be excited only optically by irradiating them in their optical absorption band (cf. section 4.4.2). The energy absorbed on excitation is not always released in the form of an emitted photon, for the molecule may return to its ground state in a radiationless transition. Nucleotide bases at room temperature behave in the latter way, whereas aromatic amino acids emit a greater part of the absorbed energy (cf. section 2.5).

In photoluminescence two kinds of spectra may be obtained. One is the *emission spectrum,* obtained when exciting light of a given wavelength is used and the spectrum of the resulting luminescent light is determined. The other type is the *excitation spectrum,* obtained when the light intensity emitted at a given wavelength is recorded as a function of the wavelength of the exciting light.

For the production and study of both types of spectra the apparatus outlined in Fig. 4.15 is used. It is described as used for measurement of the emission spectrum, but it can equally be applied to produce excitation spectra. Excitation is usually performed with a

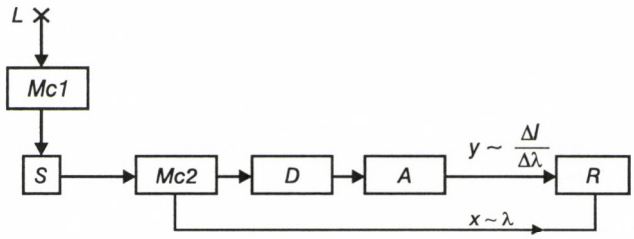


Fig. 4.15. Block diagram relating to measurement of luminescence spectra

light source (*L*) emitting in the UV and visible range, provided with a colour selective filter or a monochromator *(Mc1)*. The monochromator resolves the incident light into its components by means of an optical prism or grating. On rotation of the resolving elements, the wavelength of the light falling on the exit slit of the monochromator can be varied and hence the wavelength of the exciting light can be changed. The light emitted by the sample (*S*) in response to the excitation is analysed by a second monochromator *(Mc2)*. Rotation of the resolving element of *Mc2* causes different wavelength ranges of the studied luminescence spectrum to fall on the exit slit and thus the whole emission spectrum can be obtained. The spectral bandwidth ($\Delta\lambda$) of the light emerging through the slit can be varied to some degree, since it depends on the optical parameters of the resolving element (e.g. the dispersion of the prism, the grating constant of the grating) and on the width of the slit. The emerging light intensity (*I*) is converted by the detector (*D*; e.g. a photomultiplier) into an electric signal, which is amplified by an electronic amplifier unit (*A*). The amplified signal is recorded by an *X–Y* recorder (*R*). *Mc2* and *R* are coupled in such a form that the wavelength is recorded on the *X* axis.

As an example, the fluorescence emission spectra of aromatic amino acids are shown in Fig. 4.16. In practice the identification and simultaneous detection of amino acids run into difficulties because of the broad, overlapping emission bands. However, with respect to the quantum efficiency of the emission (the number of emitted photons related to the number of absorbed photons) the tryptophane signal dominates, therefore the appli-

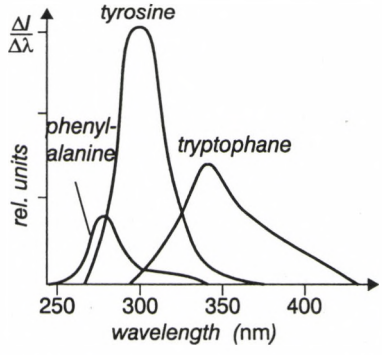


Fig. 4.16. Luminescence spectra of aromatic amino acids

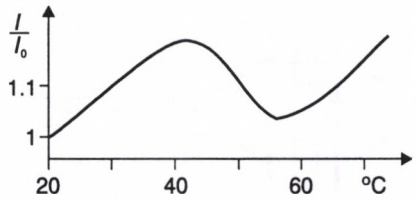


Fig. 4.17. Detection of the conformational rearrangement of the T7 phage-nucleoprotein with the aid of light emitted (at 510 nm) by proflavine molecules bound to nucleic acid. The abscissa gives the temperature, and the ordinate the light intensity emitted at the studied temperature (I) relative to that emitted at room temperature (I_0)

cations are usually based on this. The light emission of the tryptophane molecule is very effectively used both for analytical purposes and protein-structural examinations.

The basis of the applications in structural examinations is that the parameters of the spectrum of the luminescent molecule defined as monitor may yield information about the environment of the emitting molecule (if not only the intensity of the emitted light but also the degree of polarization is known) (cf. section 2.5). In addition to the emission of the aromatic amino acids occurring also under native circumstances, specifically bound fluorescent markers are widely used in the examination of the structures of proteins, nucleic acids and biological membranes. As examples those molecules should be mentioned which can bind to the nucleic acid, or more exactly which may be intercalated between the base planes, having an emitted light intensity depending on the number of intact H-bonds within the DNA chain. Thus, with the aid of intercalated luminescent molecules information may be obtained on the higher order structure of nucleic acids. An example is given in Fig. 4.17. This relates to different structural transitions induced in a nucleoprotein (phage T7) by a temperature increase. From other studies it is known that in the course of phase transition between 45 and 55 °C the ordering of the DNA molecule gradually increases and approaches the ordering in solution (regular *B* conformation: cf. section 1.5.4). This change is indicated by a decrease in the emitted light intensity. The increase of the light intensity at higher temperatures indicates the decrease of the number of hydrogen bonds. This is also in correlation with the conception about the denaturation of nucleic acids.

4.4.2. Absorption spectrometry

The absorption spectrum is the distribution of the extinction (cf. section 2.3.1) as a function of wavelength. It is measured with the apparatus shown in Fig. 4.18. The continuous spectrum of a light source (L) is wavelength-resolved by a monochromator *(Mc)*. The light selected by the exit slit and considered to be monochromatic is split into two beams. One beam passes through the investigated sample (S), and the other through the reference sample (S_0) used for comparison. The rotating sector (Rs) lets the beams pass the samples alternately and thus the detector (D) converts the intensities I and I_0 into electric signals alternately. $\mathrm{Log}(I_0/I)$ is produced by an electronic system (*El*). The *X–Y* recorder (R) plots the wavelength-dependent extinction. The use of the reference sample

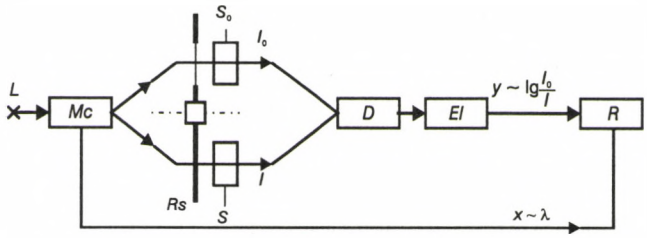


Fig. 4.18. Block diagram relating to measurement of absorption spectrum

eliminates in a simple way the intensity losses due to reflexion and (in the case of solutions) to solvent and cuvette absorption.

As concerns biological structure analysis, the visible and ultraviolet absorption spectra relating to electron transitions are of the same importance as the *infrared* spectra relating to *vibrational* transitions. Certain atomic groups of the molecules give rise to characteristic light absorption. The existence of these *chromophores* can be detected by the absorption spectra. From the shift and intensity changes in their absorption bands (cf. section 1.5.4) conclusions can be drawn regarding their environment. In this respect the hypochromic effect already mentioned in the case of DNA may be referred to: the intensity of the characteristic absorption band (about 260 nm) decreases when the nucleic acid molecule is formed from its constituents, as the intramolecular interaction of the nucleotide bases is stronger than when they are "isolated" in solution.

Infrared absorption measurement is a particularly important method in biological and organic chemical structure analysis. By this means, many functional groups (carbonyl, hydroxyl, amino, etc.) can be detected simply, even in relatively complex molecules, since the vibrational frequencies of these groups (determined by the strengths of the chemical bonds and the atomic weights of the vibrating atoms) depend only slightly upon the carbon chains coupled to the groups. From the small wavelength shifts in the absorption bands characteristic of the functional groups, information can be obtained about the structure of the surroundings of these groups.

4.4.3. Light scattering. Raman spectrometry

Studies involving scattering with electromagnetic and corpuscular radiation began several decades ago in structure research. These methods are based on the experience that the properties of the scattering matter (as revealed by the interactions of its particles with the radiation) exert a considerable influence on the characteristics of the scattered radiation, e.g. its wavelength, frequency, intensity, the spatial intensity distribution, the polarization, etc. These are the properties which are measured experimentally, but the results also allow conclusions on the dimensions, shape, internal density changes or even structure of the scattering particles. The advantage of methods based on scattering in biological research is that they can be applied for objects in aqueous solution too.

The methods based on *light scattering* give information about the morphology of particles of submicroscopic size (macromolecules, viruses, chromosomes, etc.). Further ad-

vantage of this method is that neither the preparation nor the examining agent (light) cause changes in the biological structure. By means of light scattering further information can be obtained on the mutual interactions of the particles, and their interactions with the molecules of the solvent. For solutions containing different kinds of particles, the proportions of the particles of various dimensions and densities can be determined.

The principles of interaction in light scattering can be summarized as follows. The electromagnetic field of the light passing through the solution induces dipole oscillations of the particles. The oscillating dipoles emit light in every direction (cf. section 2.2). In the simplest case, when the interactions between the particles can be neglected, the light waves scattered by the particles in a given direction meet in random phases; consequently, they cannot interfere, and the intensities of the light scattered by the particles are simply added. In every case when the phase difference of the light rays scattered by the individual particles is constant in time, however, they interfere and light diffraction occurs. This latter phenomenon and its practical applications are discussed separately (cf. section 4.5). Electromagnetic waves scattered from various points of the same particle may also produce interference; this internal interference will be dealt with later in this section.

Two types of light scattering are distinguished: Rayleigh and Raman scattering.

1. Rayleigh scattering is also known as elastic or coherent scattering. In this case the wavelengths of the exciting and scattered waves are the same.

(a) For dilute solutions containing (dielectric) *particles of much smaller size than the wavelength,* the scattered light intensity (I_s) is inversely proportional to the fourth power of the wavelength λ, independently of the direction of scattering:

$$I_s \sim \frac{1}{\lambda^4}$$

In the case of natural illumination, for example, the blue component is scattered 16 times more strongly than the red light with twice the wavelength. This explains the bluish colour of colloid solutions, the blue of the sky, or the red colour of the sunset. Light is scattered by the submicroscopic density changes in the air and by the colloid particles always present in the atmosphere. The intensity of the scattered light also depends on the size *(a)* of the scattering particles. If $a << \lambda$, a simple power relation holds:

$$I_s \sim a^6$$

On the polymerization of macromolecules or the production of aggregates, the light scattering (i.e. the turbidity of the solution) increases rapidly with increasing particle size. In this case the dimensions of the oscillating dipoles can be determined directly from the intensity of the scattered light. This dimension is approximately equal to one of the geometrical dimensions (a) of the particles, e.g. the length of a DNA fibre or the radius of a protein coil.

From the intensity of the scattered light, conclusions can be drawn on the refractive index of the particles, which is closely connected with the particle density. If the dimensions and the density of the particles are known, the molar mass also can be determined.

(b) *For particles whose size is comparable with or even larger than wavelength,* the dependence of the scattered intensity on the wavelength and the particle dimensions can no longer be described by a simple power equation. The intensity is then highly dependent on the orientation too. However, this spatial distribution permits inferences as to the shape and size of the scattering particle. As an example, Fig. 4.19 presents a light scattering curve for a suspension of *E. coli* B bacteria of greater size (about 1×2 μm) than the light wavelength. The orientational inhomogeneity can be seen clearly: measured from the direction of the direct light beam in the angle range 20–140°, the light intensity displays maxima in certain directions ($\approx 40°$, $\approx 60°$ and $\approx 80°$ in the diagram). The inhomogeneity of the distribution (i.e. the positions of the maxima and their intensities) is characteristic of the shape and size of the scattering particle. For comparison, the theoretical scattering of ellipsoidal model particles (shaded "ribbon") is also illustrated. The ribbon represents a set of curves, the individual curves relating to model particles of different sizes: the shorter diameter of the ellipsoids is 0.86–0.92 μm, while the longer one is twice this. A similar comparison of theoretical and experimental curves in other cases too may give acceptable information concerning the geometrical data of the particles.

The intensity distribution of the scattered light depends not only on the geometrical parameters, but also on the distribution of the refractive index and the density within the particles. Thus, the method can give information on the distribution of these quantities as well. For example, the wall thickness of a bacterium cell can be determined in this way.

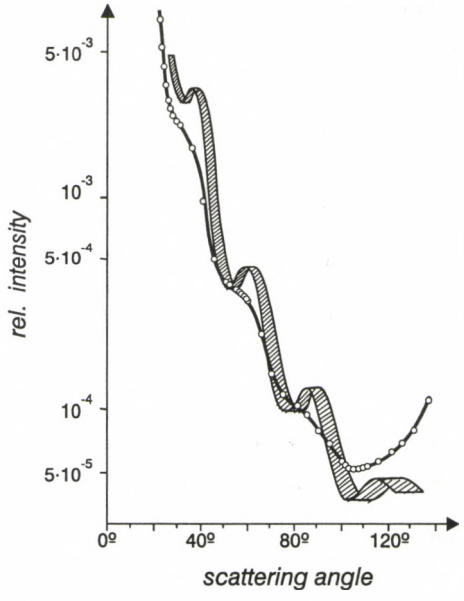


Fig. 4.19. Comparison of experimentally measured light scattering of *E. coli* B bacteria (full curve) with the theoretical scattering of ellipsoid-shaped model particles (shaded ribbon)
The abscissa gives the scattering angle (measured from the direct light beam), and the ordinate the relative intensity of the scattered light (compared to the direct light)

The application of lasers has led to a considerable development in light scattering measurements. The higher intensity of the laser beam improves the sensitivity, and lasers also permit the study of dynamic light scattering. With this technique the hydrodynamic parameters (e.g. the diffusion coefficient) associated with the motion of the particles can be established, and hence the geometrical data and molecular weights can be determined more exactly.

The scattering of other electromagnetic waves can be described similarly as light scattering. From this aspect the ratio a/λ is of interest, since identical values of this are associated with identical scattering laws. With the aid of microwaves for instance, artificial satellites can be investigated, while with X-ray diffraction molecular configurations may be studied. X-ray diffraction has been used to determine the superhelical structure of DNA inside bacteriophages (an example is given in section 4.5.1).

2. Raman scattering (Raman spectrometry) is the light scattering observed if the wavelength of the scattered light is not restricted to the wavelength of the incident monochromatic light, but contains spectral lines of shorter and longer wavelengths too. The name spectrometry refers to the fact that a wavelength-dependent scattered light intensity can be measured at wavelengths different from that of the exciting light.

This phenomenon can be explained in the following way. The $h\nu_0$ photons of the incident light interact with the scattering molecules, whereby changes take place in their vibrational or rotational energy. Two types of energy change exist. Either the vibrational or rotational energy of the molecules increases compared with the original state (in this case the energy of the scattered photon will be smaller than the incident energy $h\nu_0$), or the excited molecule passes into a vibrational or rotational state of lower energy (when the energy of the scattered light will be larger than $h\nu_0$). Both energy changes are unambiguously defined by the possible vibrational or rotational energy changes of the molecule. As a result of the described processes, new lines appear on both sides of the ν_0 line in the spectrum of the scattered light. The distance of these new lines from ν_0 is equal to the vibrational or rotational frequency of the spectral lines of the molecule.

The Raman spectrum can be studied with equipment similar to the luminescence spectrometer in Fig. 4.15. The probability of Raman scattering is extremely small; consequently, the intensity of the spectral lines is very weak, and they cannot be detected with the naked eye, but only on photographic plates after prolonged exposure (several hours). Laser radiation may now be used to increase the light intensity, and photomultipliers to detect the radiation.

The information obtained from Raman spectra does not depend upon the wavelength of the illuminating light. This important circumstance allows the use of Raman spectroscopy in structural research, since the vibrational spectra can be transferred from the experimentally difficult infrared region into the convenient visible or ultraviolet range with a suitable choice of ν_0. A further advantage of this method is that it permits the observation of vibrational and rotational transitions which are forbidden in infrared absorption.

4.4.4. Optical activity

Most biologically important molecules possess characteristic *structural asymmetry,* as they generally have no mirror planes about which the molecular symmetry is invariant *(chiral molecules).* The polarized absorption spectroscopic methods are especially sensitive techniques for studying these types of structures. However, before a discussion of these methods the basic types of polarized light will be surveyed.

In an isotropic medium the electric field vector (the light vector) of light oscillates in a plane perpendicular to the propagation of light. In the case of *natural light,* the direction of this vector may be arbitrary within this plane (Fig. 4.20*a*). However, if the direction of the oscillation is limited and the vector oscillates only along a straight line, i.e. during propagation the light vector always lies in a plane defined by the direction of the propagation and the direction of the electric field vector, then the light is *linearly polarized (plane polarized).* (In Fig. 4.20*b* the plane of oscillation is denoted by σ.[1])

Circularly or elliptically polarized light can also be produced. For circular polarization the magnitude of the light vector is constant, but its direction changes along a circle in a plane perpendicular to the direction of propagation. The velocity of this circulation depends upon the frequency of the light (Fig. 4.20*c*). If the tip of the light vector moves along an elliptical path, the light is elliptically polarized (Fig. 4.20*d*). The circular and the elliptic rotation can be either left or right handed. Both linearly and circularly polarized light may be regarded as special cases of elliptically polarized light. In the former case one axis of the ellipse is zero and in the latter the axes are equal.

Plane polarized light may always be regarded as the resultant of two circularly polarized light beams of the same velocity, frequency and amplitude, but rotating in

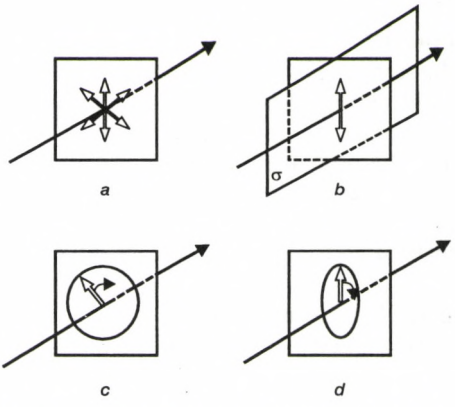


Fig. 4.20. Diagram relating to the polarization of light
a: natural light; *b*: linearly; *c*: circularly; *d*: elliptically polarized light. Solid arrows indicate the direction of the propagation of light; double arrows indicate the wave vector of light

[1] For historical reasons the plane perpendicular to this plane is called *the plane of polarization,* i.e. the plane in which the magnetic field vector oscillates normal to the electric field vector.

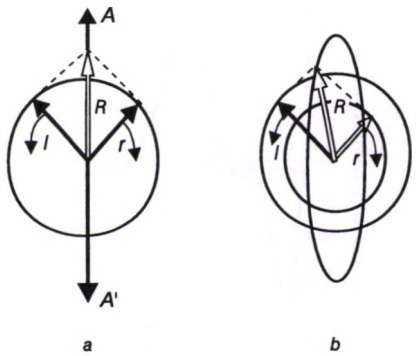


Fig. 4.21. Resolution of linearly (*a*) and elliptically (*b*) polarized light into circular components with opposite directions

opposite senses. This is demonstrated in Fig. 4.21*a*. The circularly polarized light rays rotating to the right (*r*) and to the left (*l*) propagate in a direction perpendicular to the plane of the drawing and moving away from the reader. The tip of their resultant *R* moves in the plane of the diagram along the double arrow *AA'*. The oscillation plane of the resultant plane polarized light is perpendicular to the plane of the drawing and contains the double arrow *AA'* .

A solution of asymmetric molecules interacts with the two circularly polarized components of the linearly polarized light in different ways, with the result that the velocities of propagation (refractive indices) will be different. Examination of Fig. 4.21*a* clearly shows that if the two circular components propagate with different velocities, the plane of oscillation of the resultant plane polarized light will be rotated. The extent of rotation is proportional to the difference between the refractive indices of the two components. Substances with chiral structure thus rotate the plane of linearly polarized light passing through them; substances which display this optical property are called *optically active.* If the rotation of the oscillation plane viewed in the opposite direction to

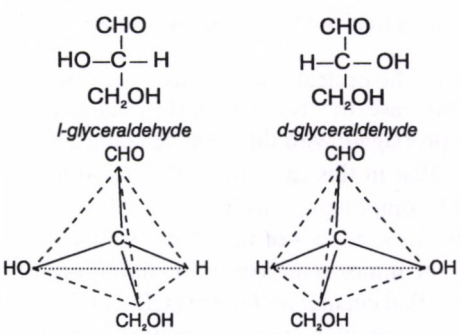


Fig. 4.22. Enantiomorphous modifications of glyceraldehyde
The upper diagrams give the structural formulae, the lower ones the conformations

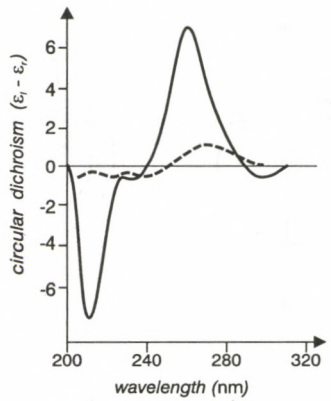


Fig. 4.23. CD spectra
ε_l and ε_r are the molar extinction of the circular components. The full line relates to the RNA of a plant virus, the dashed curve to the nucleotide bases forming the RNA (*Samejima* et al., J. Mol. Biol., *34*, 39, 1968: *Cantor* et al., Biopolymers, *9*, 1059, 1970)

that of propagation is clockwise, the rotation is right handed, and in the alternate case it is left handed, denoted by the + and – signs, respectively.

With many substances both left and right rotating modifications can be produced synthetically. The two geometric structures are mirror images of each other (enantiomorphous molecule pairs, Fig. 4.22); they can easily be distinguished by the direction of rotation. However, the biological systems usually favour one modification; in the enzyme-catalysed synthesis of organic substances, usually only one modification is produced, e.g. *l*-amino acid, *d*-glucose (the letters *l* and *d* refer to the spatial configuration of the group responsible for the asymmetry).

In the above examples the asymmetry is due to the primary structure of the molecule. However, there are cases when slightly active molecules (e.g. nucleotides) can form macromolecules which have considerable optical activity due to their secondary structure (e.g. DNA, RNA). The helical biological structures (DNA, α-helical proteins) are typically chiral systems, whose optical activity sensitively follows the changes in the secondary structure.

It sometimes occurs that the optically active substance absorbs at the wavelength of the illuminating light. In this case the two circularly polarized components of the plane polarized light not only propagate with different velocities, but are absorbed to different extents. It is easy to see that in this case, when the amplitudes (radii) of the oppositely circularly polarized light components differ, the resultant will be elliptically polarized light (Fig. 4.21*b*). The major semi-axis of the ellipse will be equal to the sum of the radii of the components, and the minor semi-axis to the difference of the radii. This change in the polarization state is called *circular dichroism* (CD), and is measured as the difference of the extinctions of the two circular components. If this difference is plotted against the wavelength, the CD spectrum is obtained. Figure 4.23 shows the CD spectrum of an RNA solution (full line), compared with the CD spectrum of a solution of the nucleotide

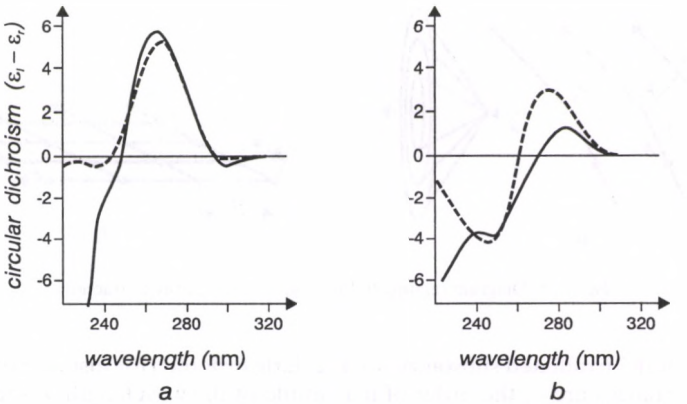


Fig. 4.24. CD spectra of nucleoproteins and isolated nucleic acids (dashed line) in the same solvent ε_l and ε_r are molar extinctions of the circular components. *a*: RNA-containing nucleoprotein (phage MS2) and its nucleic acid; *b*: DNA-containing nucleoprotein (phage T7) and its nucleic acid

monomers forming the macromolecule (dashed line). The spectra clearly demonstrate that a considerable part of the optical activity of RNA is due to the secondary structure.

The structures (conformations) of nucleic acids and hence their optical activities may also be influenced by the interaction with proteins, to various degrees depending upon the interaction. Let us compare parts *a* and *b* of Fig. 4.24. The former shows the CD spectra of an RNA-containing bacteriophage (MS2) and its isolated nucleic acid, and the latter those of the DNA-containing phage T7 and its nucleic acid. The spectra of the isolated nucleic acids (dashed lines) display a maximum at about 270 nm (cf. the RNA spectrum in Fig 4.23), which is characteristic of the chirality due to the structure of the nucleic acid in solution. This chirality is essentially the same in the phage protein-RNA complex (full line in Fig. 4.24*a*), but a considerable difference appears between the CD spectra of the DNA-containing phage T7 and T7-DNA This shows that the conformation of DNA within the phage head is of lower chirality than that of the isolated DNA.

Not only the degree of the CD, but also the rotation of the plane of the linearly polarized light (optical rotation) depends upon the wavelength. The technique by which the wavelength dependence of the rotational angle is determined is *optical rotatory dispersion* (ORD). If the ORD and CD methods are applied, the same phenomenon may be studied from different approaches.

4.5. Diffraction

4.5.1. X-ray diffraction

The X-rays incident on atoms are scattered by the atomic electrons. For structure analysis only the coherent scattering is used (cf. section 2.10.2). The diffraction is the stronger the higher the number of electrons associated with the atom, i.e. the larger the atomic number.

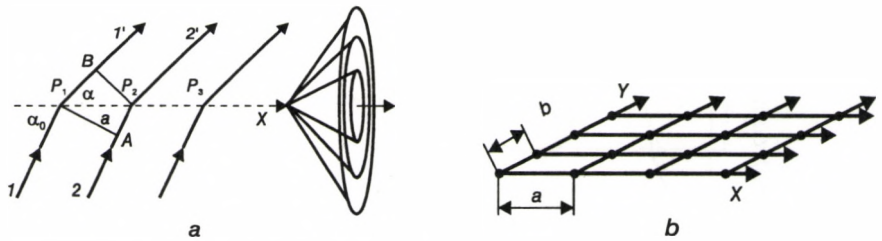


Fig. 4.25. Diagram relating to the production of X-ray diffraction

If the atoms of the irradiated substance are regularly ordered (for instance in crystals) and the atomic distances are of the order of magnitude of the wavelength of the X-radiation, then the rays diffracted by the individual atoms amplify each other in some directions and attenuate each other by interference in other directions. If a single crystal is inserted in the path of the incident radiation and a photographic plate is placed behind the crystal, the parts of the plate where the diffracted rays reinforce each other will be darkened *(Laue method*; Picture 4.6 in the Supplement). The image obtained on a screen is called the *X-ray diffraction pattern.* If the wavelength of the X-radiation is known, the atomic positions and the interatomic distances can be determined from the arrangement of interference spots.[2]

In order to understand the conditions of diffraction, consider Fig. 4.25*a*, which for simplicity depicts a single atomic row (one-dimensional lattice), in which the atoms denoted by the points P_1, P_2, P_3, ... are at the same distance a from each other. The wavelength of the parallel beam is denoted by λ, and the angle of incidence is α_0. Let us consider rays 1 and 2. The angle α denotes the direction in which the X-rays (1' and 2') diffracted from the atoms P_1 and P_2 amplify each other by interference. However, amplification is possible only if the difference between paths P_1B and AP_2 is an integral multiple of the wavelength λ, i.e. if

$$a\,(\cos\alpha - \cos\alpha_0) = e\lambda,\ e = 0, 1, 2, \ldots \qquad [4.4a]$$

Condition [4.4a] is obviously satisfied by all diffracted rays which lie on the surface of the cones with half aperture angle α (the *Laue cones*). The axis of the cones is determined by the lattice line X.

For a plane lattice another similar condition must be satisfied besides [4.4a] (Fig. 4.25*b*):

$$b\,(\cos\beta - \cos\beta_0) = f\lambda,\ f = 0, 1, 2, \ldots \qquad [4.4b]$$

where b denotes the atomic distance along the Y axis, and β_0 and β are the angles between the Y axis and the directions of the incident and diffracted rays, respectively.

In the case of a space lattice a third equation is also necessary:

$$c\,(\cos\gamma - \cos\gamma_0) = g\lambda,\ g = 0, 1, 2, \ldots \qquad [4.4c]$$

In this equation c is the distance between atomic planes situated above each other along the Z axis, while γ_0 and γ are the angles with Z. Condition [4.4b] is associated with Laue cones on the Y axis, and [4.4c] with Laue cones on the Z axis.

[2] For the sake of simple illustration the darkening of the photographic plate will be mentioned also later on as the method for detection, however, nowadays already other methods are used in practice (cf. Chapter 2).

With space lattices, diffracted beams can be observed only in directions whose angles α, β and γ simultaneously satisfy the three *Laue equations*. Expressed in a different way, the diffracted beams amplify each other only in the direction which is simultaneously the generatrix of the three Laue cones.

Further important information can be obtained from the intensities of the diffraction spots (i.e. from the darkening on the photographic plate), which depends on the amplitude of the interfering waves (scattering amplitude). The amplitude is determined by the interaction of the scattering electron and the X-rays. Thus, by a careful study of the diffraction image the dimensions and shape of the electron shell of the atoms (ions) and the electron density distribution can be determined, thereby providing information on the bonds.

While the *Laue* method can be applied only for single crystals, the method developed by *Debye* and *Scherrer* gives evaluable diffraction patterns on powdered crystals. With the Laue method more or less point-like diffraction spots are observed, whereas the latter method yields *diffraction rings* (Picture 4.7, in the Supplement).

The more complex the crystal lattices, the more complex the diffraction patterns. For instance, crystals obtained from protein or DNA molecules display several thousand diffraction spots in the Laue diagrams (Picture 1.1*b*, in the Supplement). The determination of such complex crystal structures is possible only with proper ordering and classifying techniques and computerized evaluation. In order to facilitate the evaluation of the diffraction patterns, special techniques are used, in many cases, for instance the method of heavy atom substitution (cf. section 1.5.3).

Structured X-ray patterns may also be obtained from substances exhibiting only short-range order, for instance water, fibrillar biological substances or liquid crystals. However, the greater the disorder, the more blurred the diffraction pattern, the more difficult its evaluation, and the less the information to be obtained from it.

The applicability of the X-ray diffraction method is widely extended by the fact that a diffraction pattern of spatially disordered atoms or molecules (e.g. in a gaseous state) can also be produced, which is the result of interference of radiation scattered on the different parts of the atomic or molecular electron shell. These patterns give information on the electron distribution *within* the atoms or molecules.

In connection with the study of the structure of biological macromolecules and supramolecular systems (e.g. ribosomes), the *small angle diffraction method* should be mentioned. In the case of the substances discussed above for instance, if the distance between the diffracting lattice points is one or two orders of magnitude larger than the wavelength of the diffracted X-rays, the angle of diffraction satisfying the Laue equations deviates from the angle of incidence only slightly. Consequently, the interference patterns obtained from X-rays scattered in a small angle supply information on the arrangement of larger structural subunits (e.g. the ribosome subunits, mRNA), and the effect of smaller structural elements (e.g. water) is not visible in the pattern. This method can be successfully applied in the study of samples containing water.

Finally, just for the sake of demonstration, two pictures are shown in the Supplement which refer to the molecular-biological importance of the X-ray diffraction method. Picture 4.8 shows a Laue-type picture, while in Picture 4.9 a Debye–Scherrer-type picture may be seen.

4.5.2. Electron and neutron diffraction

Electron diffraction. Due to their wave properties, electrons are also diffracted by regularly arranged atoms (ions). The diffraction pattern is similar to that obtained for X-rays and can be evaluated similarly (Picture 4.10 in the Supplement). The two methods complement each other. Because of their large penetrating power, X-rays are more suitable for the study of thick layers, whereas electrons do not penetrate deeply, and from their diffraction the structure of the surface layers can be determined. As mentioned above, X-rays are diffracted by the electron shells of the atoms, whereas electrons are diffracted mainly by the atomic nuclei. Electrons are especially suitable for the study of surface phenomena (catalysis, adsorption).

Neutron diffraction. By utilization of the wave properties of neutrons (cf. section 1.1), diffraction structural analysis can also be carried out with neutron radiation. For this purpose mainly *thermal neutrons* are used (cf. section 3.2.4), whose wavelength corresponds to the wavelength of the X-radiation used in X-ray diffraction (0.1–0.2 nm). The positions of the interference spots are determined by the distance between the scattering centres (Laue equations), which in this case are the atomic nuclei. While electrons are scattered mainly by heavy nuclei, neutrons are also scattered by protons, and thus neutron diffraction can be used with good results to determine the structures of substances containing hydrogen atoms. Many applications are based on the considerable difference between the scattering amplitudes of hydrogen and deuterium. Thus substitution of hydrogen atoms of special interest by deuterium may lead to the determination of their position. The small angle technique in this case can equally be used for the study of large structural elements (cf. section 4.5.1).

4.6. Other methods

4.6.1. Magnetic resonance spectrometry

This section deals with methods of structural analysis based on the magnetic properties of the atoms.

These methods give information on the static and dynamic conditions in the environment of atoms and atomic nuclei.

It has already been mentioned in section 1.2.2 that a magnetic moment is associated with each non-zero spin momentum and with the non-zero angular momentum of charged particles. Thus, both the nucleus and the electron shell may possess a non-zero resultant magnetic moment. Such atoms undergo ordering in a magnetic field. Only those directions of the magnetic moment are possible for which the projection of the resultant angular momentum along the external magnetic field is Mh, where M is the *orientation* or *magnetic quantum number*. The turning of the magnetic moment from one possible direction into another is a process associated with energy change. This means the splitting of the energy levels of the nucleus or the electron shell in the magnetic field: one level is

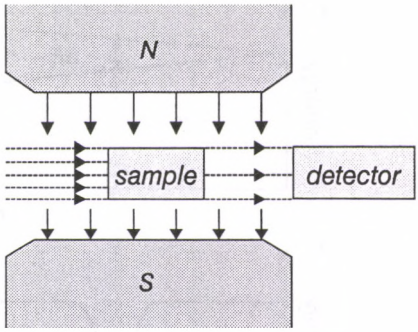


Fig. 4.26. Schematic diagram relating to magnetic resonance spectroscopy
The solid arrows indicate the lines of force of the magnetic field; the dashed arrows refer to the electromagnetic radiation passing through the sample

replaced by a number M of *Zeeman levels.* The energy difference between the levels is proportional to the magnetic field strength causing the splitting of the levels. The nucleus or the shell can be excited from a lower to a higher Zeeman level by electromagnetic radiation of frequency ν, which satisfies the relation

$$\Delta E = h\nu \qquad [4.5]$$

where ΔE is the energy difference between the two Zeeman levels.

Resonance can be achieved (i.e. the above equation is satisfied) in two ways. One possibility is to apply a constant magnetic field, in which case ΔE assumes a given value, and the exciting frequency is varied until [4.5] is satisfied. In the second case, irradiation of constant ν is applied and the magnetic field is varied. In practice, mainly the latter is the method of choice. The sample is placed between magnet poles and excited by electromagnetic waves propagating perpendicularly to the magnetic field vector (Fig. 4.26). The radiation coming from the sample is measured with a detector. By gradual increase of the magnetic field strength, [4.5] is satisfied at a given value, observed via the detector as a minimum in the transmitted radiation, since the sample absorbs the incident radiation to a higher degree. It is not a simple absorption measurement for besides the transmitted radiation, that emitted by the sample returning to the ground state is also measured, thus influencing the line-shape of the resulting curve.

Figure 4.27 shows the simple case when M can have only two values: 1/2 and – 1/2, i.e. only two Zeeman levels are present. Diagram *a* illustrates the above statement that with increasing magnetic field strength the splitting of the energy levels also increases. Diagram *b* depicts the transmittance of the sample as a function of the magnetic field strength, from which the absorption can be determined.

For a deeper understanding of the application of magnetic resonance spectroscopy, one important aspect of this method must be discussed. The return of the system from the excited state to the ground state takes place via *relaxation processes.* Similarly to radioactive decay, a relaxation process follows an exponential law. The time constant of

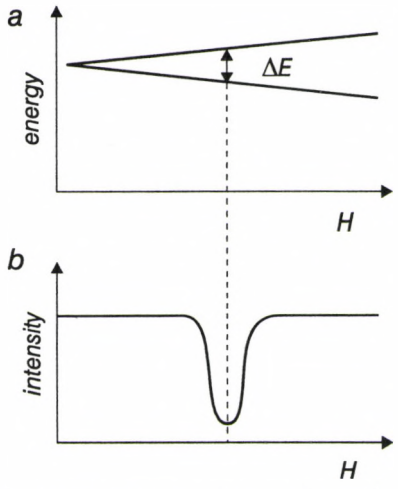


Fig. 4.27. Diagram relating to measurement of magnetic resonance
a: dependence of energy level splitting on magnetic field (*H*); *b*: dependence of detected intensity on magnetic field (*H*) for exciting radiation of given frequency *v*.
The dashed line indicates the magnetic field at which the resonance condition $\Delta E = h\nu$ is satisfied

the process is the *relaxation time,* defined as the time at which the deviation from the equilibrium value of the parameter characterizing the state decreases by a factor *e*. The line-shape of the absorption curve (the height and width of the signal) is determined by the relaxation processes. (The relaxation time is analogous to the luminescence lifetime; cf. section 2.5.)

In the present case a system of particles with magnetic moments is investigated. Let us regard it as a system of spins. The spins occupy the Zeeman levels produced by the strong magnetic field in accordance with Boltzmann distribution (cf. Appendix A1) until the exciting electromagnetic radiation no longer perturbs the stationary state. Naturally, electromagnetic radiation of a given frequency excites only that part of the spin system which satisfies the resonance condition [4.5]. These are spins of identical magnitude and identical environment. The excited spins dispose of their excess energy by interaction with the environment. This is *spin–lattice relaxation.* A different relaxation process can occur if the spins excited with the same energy are situated close to one another. They can then interact and exchange energy in the excited spin system. This is *spin–spin relaxation.* The two relaxation times can be determined from analysis of the line-shape of the spectrum. These values give information on the dynamic properties (molecular structure, motion) of the investigated system. This method is therefore suitable for study of the conformational changes of macromolecules and macromolecular systems.

Magnetic resonance spectroscopy is a very convenient tool for the investigation of just these dynamic properties. In the following we discuss separately the resonance methods based on the magnetic properties of the nucleus and of the electron shell and mention concrete examples of biological applications.

NMR method. The method based on the magnetic properties of atomic nuclei is known as *nuclear magnetic resonance* (*NMR*). This method can be used in cases when the resultant nuclear spin, and consequently the resultant magnetic moment, is not zero. This property is characteristic of nuclei whose proton or neutron number, or possibly both are odd.

The magnetic inductions applied in NMR examinations are 0.3–0.8 T high, the exciting frequencies fall into the range of radiowaves (some 100 MHz).

Some examples of application are listed below.

(a) The energy necessary for the change in direction of the magnetic moment of a nucleus depends upon its *environment.* It follows that NMR yields information about the nature of the bonding of the studied nucleus (or atom); moreover, finer differences can be revealed between bonds of very similar type. In case of the CH_3–CH_2–OH molecule, for instance, three absorption peaks (resonance peaks) appear which can be attributed to protons. This is a result of the fact that the environments of the individual protons (though in each case the bonding is covalent) differ somewhat from each other in the CH_3, CH_2 and OH groups. The areas below the absorption peaks (resonance peaks) are proportional to the numbers of atoms in identical environments, so that the relative numbers of atoms in different environments can be determined.

(b) By means of the NMR signal due to protons, the *bound water* content of any biological object can be determined. Let us take the eye lens as an example, which is cooled gradually, meanwhile the NMR absorption signal produced by the protons of the water is measured. Since free water freezes at higher temperature than bound water, protons which have lost their mobility as a result of freezing do not give an NMR signal, and the signals observed at lower temperature are therefore due to bound water. The measurements indicate that approximately 25% of the water content of the human eye is present in bound form. (This ratio is lower in some pathological cases.)

(c) Finally, mention may be made of the recently rather frequently used method of incorporating atomic nuclei (atoms) with non-zero magnetic moment into the molecule to be examined. Some biological processes can be studied by this method *(nuclear spin labelling).*

(d) **NMR-tomography** *(magnetic resonance imaging, MRI).* The new diagnostic method of MR imaging has recently been developed. This permits the study of larger samples too, such as the liver, kidney, head or the whole body. With an inhomogeneously varying magnetic field it can be achieved that the resonance condition [4.5] will be fulfilled only in a small volume element of the whole sample, whereas the other parts of the body do not give an NMR signal. On alteration of the inhomogeneous magnetic field, the resonance signal always comes from different sites of the sample, so that the whole body can be scanned (cf. section 6.7.5).

The signals arising from different sites of a three-dimensional body are analysed with a computer. The computer determines the absorption and the different relaxation times. The intensity is proportional to the proton concentration in the volume element of the sample from which the signal comes, and the relaxation times are related to its molecular environment. MR imaging is more suitable for the examination of soft tissues, while X-ray imaging is used for tissues containing heavy elements too because it is more sensitive for heavier elements. Typical application fields of MR imaging are the diagnosis of tumours, blood circulation anomalies and the metabolism of soft tissues.

ESR method. *Paramagnetic resonance* or *electron spin resonance* (ESR) is based on magnetic investigation of the electron shells. This method permits the study of atoms (atomic groups, molecules) which contain uncompensated electron spins. The ESR method requires a larger excitation energy than NMR, since in the case of the usual magnetic field induction (0.3–0.8 T) the exciting frequencies fall in the microwave (some GHz) range.

(a) With the ESR method, *free radicals* can easily be detected even in a concentration of roughly 10^{-11} mol/mol. For instance, certain free radicals participating in or produced by the life processes were detected by ESR. This method also furnishes favourable conditions for indication of the presence of active radicals produced by ionizing radiation and also for measurement of their lifetime. The ESR spectra of healthy and pathological tissues differ too.

(b) Further possibilities are given by the *spin labelling* method. Atoms or atom groups with uncompensated electron spins are incorporated into larger molecules and used to follow the changes in their motion and conformation in various biological processes.

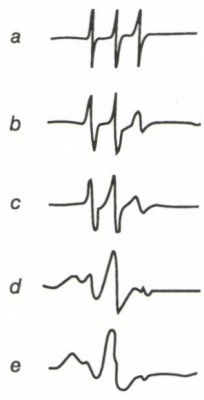


Fig. 4.28. An example of spin labelling. ESR spectra of spin labelling (nitroxide) molecules
a: in aqueous solution: *b*: in a pure phospholipid membrane bound to the lipid;
c–e: in phospholipid membranes containing proteins in various concentrations
(0.49, 0.24, 0.10 mg lipid/mg protein)

With the aid of spin labelling it can be established how and to what degree the motion of a molecule is restricted by its surroundings. The influence of the surrounding lipids on the functions of membrane proteins, for instance, is an essential question, as is the effect of the presence of the proteins on the lipid structures. Among others, nitroxide can be used as a spin label. Figure 4.28*a* depicts the ESR spectrum of an aqueous solution of this compound. If this molecule is bound to the lipid molecules of the membranes for the purpose of spin labelling, its motion is modified by its surroundings. Examinations show that the motion is slower in a protein than in a lipid environment. These results are demonstrated in Fig. 4.28*b–e*. In a pure lipid environment the ESR spectrum is very similar to the spectrum of the freely moving spin labels, i.e. curve *b* is similar to curve *a*. However, an essentially different result is obtained in an environment rich in protein. This can readily be observed by comparing curves *a* and *e*.

ENDOR method. The electron nuclear double resonance (ENDOR) method has been developed from the simultaneous application of ESR and NMR. This method yields more concrete information on the electron structure and configuration of the sample than the individual methods separately.

4.6.2. Mass spectrometry

If a beam of ions with various positive charges, masses and velocities propagates through electric and/or magnetic fields, the ions with different parameters will deviate to different degrees. With appropriately selected electric and/or magnetic fields, the ions of different velocities but of the same mass/charge ratio can be focused at one point (Fig. 4.29). In mass spectrometry an ion beam is generally produced from the sample investigated (e.g. by electron bombardment in vacuum) in such a way that the bulk of the ions have only one positive charge. Thus, the separation of the ions due to the effect of the electric and/or magnetic fields will characterize their different masses. This separation yields the *mass spectrum.* The ions focused on the individual parts of the spectrum are usually detected by means of a secondary electron multiplier provided with a slit. In practice the detector is not moved along the spectrum, but the electric and magnetic fields are changed so that the ions with various masses reach the slit at different times. The masses of the ions are determined from the degrees of the magnetic and/or

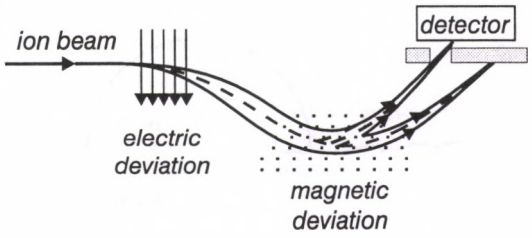


Fig. 4.29. Outline of operation of a mass spectrometer
The electric field vectors are parallel to the plane of the drawing, while the magnetic field vectors are perpendicular to it

electric deviation, and the strengths of the detected signals allow determination of the concentrations.

Extremely *small concentrations* of the various atoms and their isotopes can be measured with mass spectrometry; thus, the method permits the detection of trace elements in biological substances. From the majority of stable small molecules, molecular ions can be produced in the mass spectrometer by electron bombardment, and in this way an extremely exact *molecular weight determination* is possible. The mass spectrometric method also plays an important role as a very sensitive evaluating procedure in the isotope tracer technique (cf. section 3.6). Mass spectrometry is important in *pharmaceutical structure analysis.* In the course of mass spectrometric investigation, not only molecules are ionized by electron bombardment, but electrically charged fragments are also produced. From these easily identifiable mosaic particles, the macromolecular structure of the original substance can be deduced. One great advantage of the method is that the analysis can be carried out with even a few tenths of a mg of substances.

4.6.3. Electron spectrometry for chemical analysis

Several types of electron spectrometry exist: one is electron spectroscopy for chemical analysis (ESCA). In the case of chemical bonds changes taking place in the outer shell electrons may alter the energy states of the electrons in the full core electron shells by 0.1–0.01%. The measurement of these variations thus permits the investigation of the *atomic bonds* and their *changes.* Electrons are ejected from the inner orbitals of the atoms to be investigated by X-ray photons (photoeffect) and their energy is measured with high-precision electron spectrometers. The energy E of the emerging electrons is smaller than the energy $h\nu$ of the X-ray photon.The energy difference (W) between $h\nu$ and E is equal to the bonding energy within the atom, i.e.

$$E = h\nu - W \qquad [4.6]$$

The electron spectrometer measures E and since $h\nu$ is constant and known, the value of E supplies W directly.

Usually a magnetic field is applied for the measurement of E. Figure 4.30 outlines a magnetic spectrometer. The characteristic X-rays emerging from the anode (e.g. Mg) of the X-ray tube fall on the sample, resulting in electron emission from its surface layer. These electrons are brought into a circular path by a magnetic field (which in the diagram is perpendicular to the plane of the drawing). The electrons with different energies are focused at different points (P, P', ...) producing the energy spectrum. The detector is usually a GM counter or an ionization chamber.

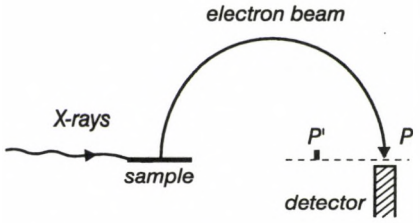


Fig. 4.30. Diagram relating to the ESCA method

As an example, the use of the ESCA method is illustrated with measurements on insulin. The carbon, nitrogen, oxygen and sulphur atoms constituting insulin have characteristic, well-defined electron spectra. The sulphur atoms can be found in the cystine components of the molecule, two of them being situated between the chains and one in the *A*-chain. Figure 4.31 depicts the energy spectrum of the 2*p* electrons of sulphur in normal bovine insulin (diagram *a*), and in the insulin molecule oxidized with periodate (diagram *b*). Curve *a* has only one peak, which shows that the environments of the sulphur atoms in the molecule are identical insofar as every atom participates in a disulphide bond. Curve *b* shows that the periodate oxidation is selective. The double peak refers to the presence of sulphur atoms in different environments, which means that the treatment does not influence the disulphide bond in the *A*-chain, only the sulphur atoms between the *A* and *B*-chains being oxidized. The ratio of the intensity maxima in the spectrum is 2:1, which corresponds to the ratio of the sulphur atoms between the *A* and *B*-chains of the insulin and the sulphur atom within the *A*-chain.

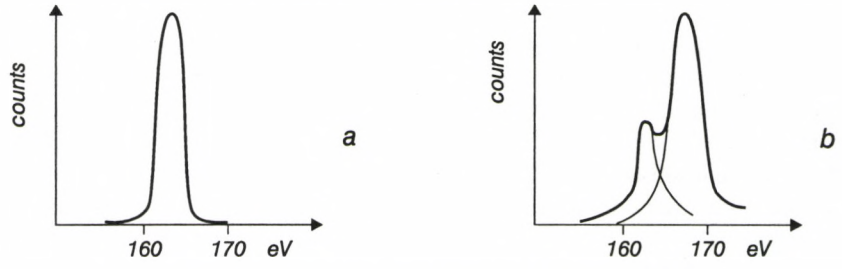


Fig 4.31. Energy spectrum of the 2*p* electrons of sulphur for normal insulin (*a*) and insulin selectively oxidized with periodate (*b*)
The abscissa gives the bonding energy, the ordinate the number of measured electrons (*K. Siegbahn* et al., Ann. Phys. *3*, 281, 1968)

4.6.4. Microcalorimetry

The higher order structure of the biologically important molecules (e.g. proteins, nucleic acids, phospholipids) and the systems built up from them (e.g. membranes, chromatin) may be altered considerably by a small change in the external conditions (temperature, pressure, ion concentration); this results in changes in their functions. The high sensitivity is connected with the relatively weak interactions (van der Waals forces, hydrogen bonds) which stabilize the higher order structures. In these cases, and also in more complex systems such as viruses and cells, the structural changes (phase transitions) resulting from small temperature changes can be sensitively studied by microcalorimetric methods. The transition temperatures and heats (more precisely the transition enthalpies) can be

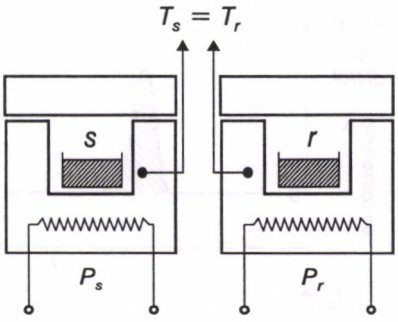


Fig. 4.32. Schematic diagram of the DSC method
s: sample; *r*: reference material; P_s and P_r: power supplied to the heating filament of sample and reference side; T_s and T_r: temperature of sample and reference material

determined by microcalorimetry. The main type of this method is described in the following.

Differential scanning calorimetry (*DSC*). The sample (*s*) is placed into a metal vessel in a metal case, while a reference material (*r*) into a similar one (Fig. 4.32) and both are heated in a controlled way. The reference material should differ from the sample as slightly as possible, but must not have any heat-induced transition in the investigated temperature range. Thus, e.g. for a protein solved in some buffer solution, a protein-free buffer solution may serve as reference material.

In course of the measurement the temperature of the sample and the reference material is increased uniformly in time by appropriate selection of the heating rates so that the two temperatures should be always equal: $T_s = T_r$. For this, certainly different amounts of heat should be given to the two systems because of their different heat capacity. Being no transition in the reference material, its heating power (P_r) remains nearly constant during the whole measurement. On the other hand, the power taken up by the sample (P_s) is constant only until the beginning of the (endothermic or exothermic) transition in the sample. In this case more (or less) heat has to be given to the sample for a uniform increase of the temperatures. After the transition P_s returns to a constant value not necessarily the same as the previous one. The power difference ($P_s - P_r$) is recorded during the measurement as the function of time and the transition of the sample will be indicated in the record by a peak (Fig. 4.33*a*). If the transition is endothermic, a positive peak is obtained since in this case the sample needs more heat ($P_s > P_r$). In case of exothermic transition the situation is reversed.

Let us consider the meaning of P_s and P_r from the thermodynamic point of view. Since the pressure of the system (sample and reference) is practically constant during the measurement, the amount of heat given to the system in unit time (P_s and P_r) gives the enthalpy change of the system within unit time: dH_s/dt and dH_r/dt, respectively (cf. section 5.3.3).

In most cases our aim is to study the transition in the sample, more exactly the determination of enthalpy change accompanying the transition. Concentrating on the main points one can make some simplifications. Before and after the transition (in the

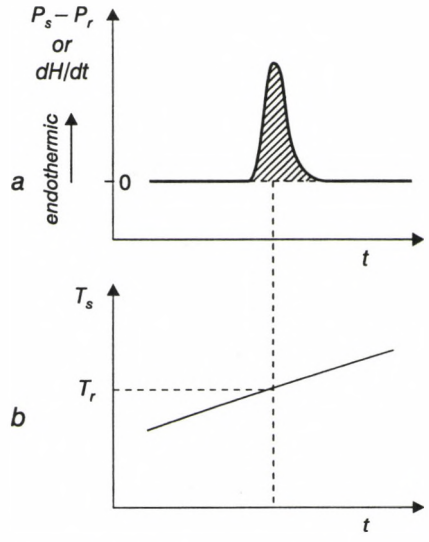


Fig. 4.33. DSC curve of endothermic transition (*a*) and its heating curve (*b*)

figure before and after the peak), where $P_s - P_r$ is constant, this constant is taken as zero. Thus the difference $(P_s - P_r)$ during the transition (in the figure the vicinity of the peak) gives simply the enthalpy change in unit time related to the transition in the sample denoted by *dH/dt,* marked on the vertical axis of Fig. 4.33*a*. Integrating the function *dH/dt* for the time of transition, the enthalpy change of the transition, i.e. the transition heat, is received. The numerical value of the integral is given by the area under the curve between the given boundaries (cf. section B4 in the Appendix). According to these the shaded area in Fig. 4.33*a* represents the transition heat.

The temperature of the sample (T_s) is also recorded in the function of time during the measurement (Fig. 4.33*b*). Comparing the two curves the transition temperature (T_r) can also be determined, as can be seen in the figure. Since the temperature of the sample increases uniformly in time, this may also be indicated – using appropriate scale – on the horizontal axis in Fig. 4.33, instead of the time. This simplified mode of representation has been widely referred to in the literature.

In addition to the transition enthalpy and temperature, the temperature dependence of the specific heat of the sample, the kinetic order, rate constant and activation energy of chemical reactions can also be determined by means of the DSC method.

One application of this method, the determination of the bound water content of biological systems by DSC, is mentioned below. The left side of Fig. 4.34 relates to this method. The peak at 0 °C is connected with the free water–ice transformation. The peak height increases in proportion to the water content. With an 80% lipid content this peak disappears, which means that the system contains only bound water, which freezes at a lower temperature. The peak at 60–65 °C on the right side of the figure relates to the crystalline–liquid-crystalline phase transition of the DSPC membrane.

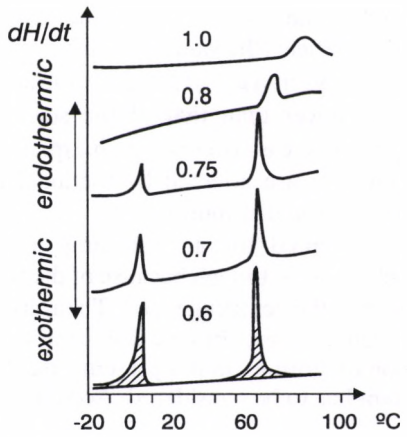


Fig. 4.34. DSC curves of mixtures of a membrane constituent lipid, distearoylphosphatidylcholine (DSPC), and water in various ratios
The numerical data indicate the lipid/water ratio.
The shaded area at the lowermost curve corresponds to the enthalpy of transition

4.6.5. Sedimentation

The measurement of the sedimentation rate of suspended particles, and after the termination of this process, the investigation of the particle distribution in the solution allow conclusions about the dimensions, shape, density, molecular weight, etc. of the particles. Further, it is possible to separate particles with various parameters. The sedimentation method is used in biology to study cells, cell components, viruses, molecules, etc. In most cases the gravitational field strength is not sufficient. Instead, the centrifugal forces of rotating systems are used; in suitable equipment *(centrifuges)* these exceed the gravitational force by several orders of magnitude. For instance, in a centrifuge (ultracentrifuge) performing approximately 1000 revolutions per second, the centrifugal force may be 10^5 times larger than the gravitational force.

1. Investigation of particle distribution. At the beginning of centrifugation the particles migrate in the solution in accordance with the generalized Archimedes law, in the direction which is the resultant (F) of the centrifugal force (F_{cf}) and the lifting force (F_l) acting in the direction opposite to F_{cf}.[3] The magnitude of the resultant force is given by the relation

$$F = (\rho - \rho')\, Vr\omega^2 \quad \text{or} \quad F = mr\omega^2 \qquad [4.7]$$

where ρ and ρ' are the densities of the particles and of the solvent, respectively, V is the volume of the particle, r is the distance of the particle from the rotational axis of the centrifuge, and ω is the angular velocity of the rotation (see also the right side of

[3] The lifting force is the compressive force of the solvent on the particles.

Fig. 4.35). Here $m = (\rho - \rho')V$ is the effective mass; due to the lifting force, this is not equal to the actual mass of the particle (the value of which is ρV). Particles whose density is higher than the density of the solvent ($\rho > \rho'$) migrate towards the bottom of the tube, while particles with a density lower than that of the solvent move in the opposite direction. In practice mainly the first case occurs, which explains the name *sedimentation.* Besides these forces, the particles are affected by frictional forces, but these can be neglected in investigations of particle distribution.

If the centrifugation is continued for a sufficiently long period, the migration (sedimentation) of the particles ceases, though because of the thermal motion they do not all cluster tightly at the bottom of the centrifuge tube. Their distribution can be described by the *Boltzmann relation,* which also describes the distribution of dust and other particles suspended in air as a function of altitude. In the present case the potential energy of the particles in the rotating system has to be considered. According to Boltzmann:

$$\frac{N_1}{N_2} = \exp\left[-\frac{\frac{1}{2}m\omega^2 (r_2^2 - r_1^2)}{kT}\right] \qquad [4.8]$$

where k is the Boltzmann constant, T is the absolute temperature, and N_1 and N_2 are the equilibrium concentrations of the particles at distances r_1 and r_2, respectively, from the axis of rotation (left side of Fig. 4.35). The numerator of the exponent contains the difference between the potential energies of the particles of effective mass m due to the change of their distance. It follows from [4.8] that larger particles (for instance cells, or larger cell components) are deposited at the bottom of the centrifuge tube even at not too high rotation rates. With smaller particles (e.g. macromolecules), even if an ultracentrifuge is used, the particle distribution according to [4.8] must be taken into consideration.

It has been assumed that the density distribution of the suspending medium does not change during the centrifugation, i.e. that ρ' is the same at every site in the centrifuge tube. In practice this is usually the case. In the interest of separating particles with different densities, however, it is advantageous to use a medium in which a density gradient described by the Boltzmann relation develops due to the centrifugation *(centrifugation in a density gradient).* Such a medium, for instance, is a solution of some heavy metal salt (most frequently CsCl). Thus, ρ' in [4.7] is not constant, but increases with the distance from the axis of rotation. It is clear from [4.7] that the particles

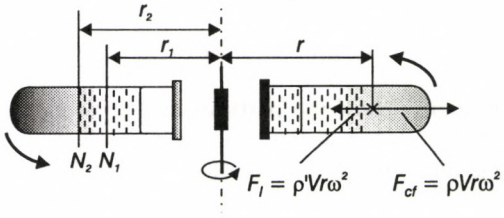


Fig. 4.35. Outline of centrifugation

of density ρ collect in a given region of the centrifuge tube, where $\rho = \rho'$. The particles of higher density collect further from, and those of lower density closer to the axis of rotation.

2. Measurement of sedimentation rate. As already mentioned, in the sedimentation process the migrating particles are affected by frictional forces besides the centrifugal and lifting forces. If the resultant of these three forces is nearly zero, the constant rate of sedimentation (v) may be regarded as proportional to the centrifugal acceleration ($r\omega^2$), i.e.

$$v = sr\omega^2 \qquad [4.9]$$

where the proportionality factor s (the sedimentation constant) has dimensions of time. [4.9] is usually satisfied in practice. Some centrifuge types permit the measurement of v and, if r and ω are known, s can be determined. From the value of the sedimentation constant, conclusions can be drawn as to the size of the particles (or of their hydrate sheath), their shape and their molecular weight. Since the viscosity of the solvent also influences the sedimentation constant, s depends very sensitively upon temperature. For this reason the value of s is usually given related to water at 20 °C. In this case, instead of s the proportionality factor is denoted by $s_{20,w}$. Its unit is the *svedberg* which corresponds to 10^{-13} s. For informative purposes, Table 4.1 lists sedimentation constants and molecular weights for some biological macromolecules and viruses.

Table 4.1. Some data on macromolecules and viruses

Particle	Relative molecular weight	$s_{20,\omega}$ (svedberg unit)
Myoglobin	1.7×10^4	2.0
Haemoglobin	6.8×10^4	4.0
Botulinus toxin	9.5×10^5	17.0
E. coli ribosome	2.8×10^6	69.1
Poliomyelitis virus	6.7×10^6	120.0
Phage T2	2.0×10^8	900.0

REFERENCES

Books

Barltrop, J. A., Coyler, J. D., Principles of Photochemistry. John Wiley, New York (1978)

Davies, D. B., Saenger, W., Danylak, S. S. (eds), Structural Molecular Biology. Plenum Press, New York (1981)

Fairing, J. D., Fairing, E. A, Image and Signal Processing in Electron Microscopy, Scanning Microscopy, International, Chicago (1988)

Hedvig, P., Experimental Quantum Chemistry. Akadémiai Kiadó, Budapest (1975)

Ibach, H. (ed.), Electron Spectroscopy for Surface Analysis. Springer-Verlag, Berlin (1977)

Pethig, R., Dielectric and Electronic Properties of Biological Materials. John Wiley, New York (1979)

Rochow, T. G., Rochow, E. G., An Introduction to Microscopy by means of Light, Electrón, X-rays or Ultrasound. Plenum, New York–London (1979)

Schuster, T. M., Laue, T. M. (eds), Modern Analytical Ultracentrifugation. Birkhäuser, Boston (1994)
Theophanides, Theo M., Infrared and Raman Spectroscopy of Biological Molecules. D. Reidel, Dordrecht, Holland (1979)

Papers

Binnig, G., Rohrer, H., The scanning tunneling microscope. *Scientific American, 252,* No. 8, 40–46 (1985)
Binnig, G., Rohrer, H., Scanning tunneling microscopy. *Trends in Physics, 1,* 38–46 (1984)
Binnig, G., Rohrer, H., Gerber, Ch., Stoll, E., Real-space observation of the reconstruction of Au(100). *Surface Sci., 144,* 321 (1984)
Dunlap, D. D., Bustamante, C., Images of single-stranded nucleic acids by tunneling microscopy. *Nature, 342,* 204 (1989)
Dürig, U., Pohl, D. W., Rohner, F., Near-field optical-scanning microscopy. *J. Appl. Phys., 59,* 3318–3327 (1986)
Fekete, A., Rontó, Gy., Feigin, L. A., Tikhonychev, V. V., Módos, K., Temperature dependent structural changes of intraphage T7 DNA. *Biophys Struct Mech., 9,* 1–9 (1982)
Hansma, P. K., Tersoff, J., Scanning tunneling microscopy. *J. Appl. Phys., 61,* R1–R23 (1987)
Kurz, A., Lampel, S., Nickolenko, J. E., Bradl, J., Brenner, A., Zirbel, R. M., Cremer, T., Lichter, P., Active and inactive genes localize preferentially in the periphery of chromosome (territories). (Confocal fluorescence scanning microscopy). *J. Cell. Biol 135,* 1195 (1996)
McNaughton, J. L., Mortimer, C. T., Differential scanning calorimetry, IRS; Physical Chemistry Series 2, Vol. 10. Butterworths, London (1975)
Rippe, K., Mücke, N., Langowski, J., Superhelix dimensions of a 1868 base pair plasmid determined by scanning force microscopy in air and in aqueous solution. *Nucleic Acid Research, 25,* 1736 (1997)
Rontó, Gy., Tóth, K., Feigin, L. A., Svergun, D. I., Dembo, A T., Symmetry and structure of bacteriophage T7. *Comput. Math. Appl., 16,* 617–628 (1988)
Tóth, K., Études des bacteriophages T7, MS-2 et ΦX–174 par dichroisme circulaire et l'absorption UV. Thèse de 3e cycle, Univ. P. et M. Curie, Paris (1981)

5. TRANSPORT PROCESSES THERMODYNAMIC BASIS OF LIFE PROCESSES

This chapter is divided into three main parts. We first treat the transport processes which are of outstanding biological importance, while in the second part the basis of thermodynamics is dealt with. Finally, the material in the first two parts is used to discuss the biophysical aspects of membrane transport phenomena.

5.1. Flow of fluids and gases

In the life processes, especially in those of more highly developed organisms, the circulation of various fluids and gases is of prominent importance. Consider for instance blood circulation or respiration. These processes can be modelled from a physical viewpoint by the flow of fluids and gases in tubes.

5.1.1. Basic concepts

As long as the flow velocity is below a value of approximately 50 m/s, the compressibility of gases plays practically no role in flow phenomena. For this reason, both gases and fluids are regarded as incompressible. Since the velocities do not exceed the above critical value, the flow of gases can be discussed together with the flow of fluids. Though only fluids are mentioned in the following treatment, the results also hold for gases.

The discussion is restricted to flow in rigid-walled tubes; tubes with elastic walls will be dealt with in section 5.1.6. The flow of fluids is characterized by the fluid volume flowing through the tube cross-section in unit time, or more exactly by the *volume current strength* (I), which can also be defined as the intensity of the current. By definition, we have

$$I = \frac{\Delta V}{\Delta t} \qquad [5.1a]$$

where ΔV denotes the volume of fluid flowing through the tube cross-section in time Δt.

With *ideal* fluids, i.e. frictionless (and incompressible) fluids, the flow velocity is the same at every point of the tube, and a simple equation describes the relation between this common velocity v and the current intensity I. If a fluid volume element $\Delta V = q\Delta s$ flows through a tube of length Δs and cross-section q during time Δt, from the definition of current intensity:

$$I = q\frac{\Delta s}{\Delta t} = qv \qquad [5.1b]$$

In real fluids the velocity differs at the individual points of a given cross-section of the tube, being maximum in the tube axis and decreasing from the axis towards the tube wall. In these cases it is customary to work with the *mean velocity* $\bar{v}$ belonging to the given cross-section, as defined by [5.1b]. Consequently, [5.1b] is considered as valid for real fluids too, and the *mean velocity* is given by

$$\bar{v} = \frac{I}{q} \qquad [5.1c]$$

i.e. the mean velocity is the quotient of the current intensity and the cross-section.

If the characteristic quantities of the current (velocity, current intensity, pressure) are time-independent, and at most vary from place to place, the current is said to be *stationary.* It is clear that for stationary currents the velocity is larger with a small than with a large tube cross-section: the velocity is inversely proportional to the cross-section.

5.1.2. Bernoulli's law

In this section we study the pressure distribution of ideal fluids undergoing stationary flow through tubes of different cross-sections (Fig. 5.1). For simplicity we consider a horizontal tube. The pressure (or more exactly the hydrostatic pressure) is given by the height of the fluid column in a vertical side-tube. According to Bernoulli's law, at any point in the tube

$$p + \frac{1}{2}\rho v^2 = \text{constant} \qquad [5.2a]$$

where ρ denotes the fluid density, while p and v are its pressure and velocity, respectively. p in [5.2a] is often referred to as the static pressure, and the quantity $1/2\ \rho v^2$ as the *dynamic pressure,* the sum of these two quantities is the total pressure. Thus, from Bernoulli's law the *sum of the static and the dynamic pressures is constant and equal to the total pressure.*

If the tube is not horizontal, the change in the potential energy of the fluid particles due to gravitation must also be considered and instead of [5.2a] we have

$$p + \frac{1}{2}\rho v^2 + \rho g h = \text{constant} \qquad [5.2b]$$

Bernoulli's law essentially expresses the *law of conservation of energy* for fluids.

5.1.3. Internal friction. Stokes' law

Here again fluids will be discussed, though the results are also valid for gases. If a body moves in a liquid medium, or the medium flows with respect to the body, a force due to friction is exerted on the body. The friction between the medium and the surface of the body is referred to as *external friction,* while the displacements of the layers of the medium with respect to each other give rise to *internal friction.* In practice the fluid usually wets

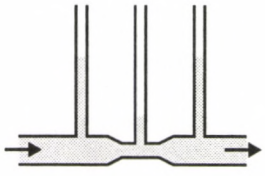


Fig. 5.1. Diagram relating to Bernoulli's law

the surface of the body, which means that a liquid layer of given thickness (sometimes only a monomolecular layer) adhering to the surface moves together with the body. In the relative motion of the body and its liquid environment, fluid generally moves on fluid and only internal friction occurs.

The internal friction is characterized by the *internal friction coefficient,* i.e. the *viscosity.* The introduction of this concept is associated with a phenomenon important in practice. The viscosity can be studied by letting spherical bodies fall in some liquid. As the body falls, it is first accelerated by gravitation, but later acceleration gradually decreases to zero and the body attains a constant velocity. This phenomenon is explained in that the frictional force acting on the body increases with increasing velocity. Finally, when the frictional and driving forces become equal, the velocity of the falling body becomes constant. Experimental evidence demonstrates that in most cases the internal frictional force (F_f) is proportional to the relative velocity (v) of the body:

$$F_f \sim v \qquad [5.3a]$$

The frictional force depends upon the shape of the body and the nature of the medium. In the case of spheres, the form factor is the radius of the sphere (r), and the internal friction is proportional to r. The nature of the medium is included in the proportionality factor; for spheres (without going into detail) this can be expressed by the quantity $6\pi\eta$. Consequently,

$$F_f = 6\,\pi\eta r v \qquad [5.3b]$$

The factor η is the *internal friction coefficient* or simply the *viscosity,* its unit is Pa s. For instance, the viscosity of water at 18 °C is 1.1 mPa s, and that of the air at atmospheric pressure is 0.018 mPa s. The viscosity of an "easily flowing" fluid is small, whereas viscous fluids have high viscosities.

As mentioned above, the liquid layer adhering to the surface moves together with the body, though its effect decreases with increasing distance from the body. The fluid layer in which this effect is still observable is called the *boundary layer.* For water the boundary layer has a thickness of a few mm, whereas for fluids of higher viscosity the boundary layer is thicker, e.g. for blood a few cm.

The reciprocal of the viscosity is the fluidity. The quantity η is frequently referred to as *dynamic* viscosity and the quotient η/ρ as *kinetic* viscosity (ρ denotes the density of the liquid).

The viscosity depends very sensitively upon the temperature. For gases, the viscosity grows proportionally to the square root of the absolute temperature T. The qualitative explanation is that the interaction of the gas layers sliding on each other is the more intensive, the higher the mean thermal velocity of the molecules. For fluids, however, the viscosity decreases with increasing temperature. The mechanism of internal friction is different for liquids since the molecular gaps, vacancies, play an essential role in the displacement of the liquid layers on one another. With more vacancies the molecules can jump more easily into the adjacent vacancies, with the result that with an increasing vacancy concentration the mutual displacement of the layers becomes easier and the viscosity decreases. This also means that the viscosity is nearly inversely proportional to the corresponding Boltzmann factor (cf. Appendix A1), more exactly

$$\frac{1}{\eta} \sim T e^{-\frac{\varepsilon}{kT}} \qquad [5.3c]$$

The letter ε here denotes the *activation energy of molecular migration,* whose value in the case of liquids is a few tenths, or frequently only a few hundredths of an eV. (The factor T before the exponential term generally plays only a minor role as compared to that of the rapid exponential change.)

[5.3b] has been used to introduce the concept of viscosity, but it is also frequently used as a special relation called Stokes' law. By its aid the constant velocity v attained by a sphere of radius r and density ρ' falling in air of viscosity η and density ρ can be calculated:

$$v = \frac{2g}{9\eta}(\rho' - \rho)\, r^2 \qquad [5.3d]$$

This relation explains the relatively low falling velocity of small fog droplets or dust particles in the air. The equation can also be used for viscosity measurements or to determine the radius of spherical particles (such as, for instance, colloidal particles or macromolecules).

Finally, a further remark is made in connection with [5.3a]. In the case of motion at constant velocity, the driving and the frictional forces differ from each other only in sign, and consequently the driving force F_d is proportional to the velocity. This relation can be written in the form

$$v = uF_d \qquad [5.4a]$$

where the coefficient u is called the *mobility.* The value of u gives the mean velocity of a colloid particle or macromolecule moved in some medium by unit driving force. In the case of spherical particles we have

$$u = \frac{1}{6\pi\eta r} \qquad [5.4b]$$

5.1.4. The Hagen–Poiseuille law

Consider a fluid flowing in a horizontal tube of constant circular cross-section (Fig. 5.2). Let the volume of fluid flowing across the cross-section $r^2\pi$ in time Δt be ΔV. The pressure on the tube wall (static pressure) is measured by the height of the fluid column in a vertical tube connected to the horizontal system. Let the pressures at the two ends of the horizontal tube length l be p_1 and p_2, respectively. The pressure decrease per unit length, as expressed by the quantity $(p_1 - p_2)/l$, is called the *presssure drop.* According to the Hagen–Poiseuille law, ΔV is proportional to the flow time, the pressure drop and the fourth power of the tube radius, and inversely proportional to the viscosity (η) of the fluid. Without going into a detailed discussion, the proportionality factor is $\pi/8$. Consequently

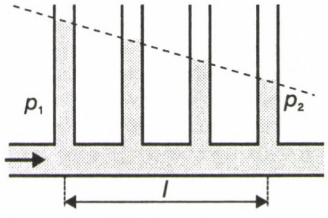


Fig. 5.2. Diagram relating to the Hagen–Poiseuille law

and

$$V = \frac{\pi}{8} \frac{r^4}{\eta} \frac{p_1 - p_2}{l} \Delta t \qquad [5.5a]$$

$$I \equiv \frac{\Delta V}{\Delta t} = \frac{\pi}{8} \frac{r^4}{\eta} \frac{p_1 - p_2}{l}$$

The pressure distribution along the tube length is characterized by the quantity dp/dl, called the pressure gradient. For identical cross-sections the pressure gradient is constant in the various regions, and therefore the pressure gradient can also be written in the form $(p_2 - p_1)/l$. In [5.5a] we have $-(p_2 - p_1)/l$, since $p_1 > p_2$. This means that the intensity of the volume flow is proportional to the negative pressure gradient (cf. section 5.5.1).

To measure the tube resistance (frictional resistance), various quantities are used:

(a) [5.5a] can be written in the following form

$$p_1 - p_2 = RI \qquad [5.5b]$$

where

$$R = 8\pi\eta \frac{1}{\pi^2 r^4} \qquad [5.5c]$$

and the resistance is characterized by the quantity R. [5.5b] is similar to Ohm's law for electric current, which describes the relation between the potential difference, the electric current intensity and the resistance of the conductor. The Hagen–Poiseuille law expresses a similar relation between the pressure difference, the liquid flow intensity and the frictional resistance.

The electric resistance is proportional to the length of the conductor, and inversely proportional to its cross-section. The frictional resistance R is proportional to the length of the tube and inversely proportional to the square of the tube cross-section.

(b) In the case of a circular cross-section, from [5.1c] $I = \bar{v} r^2 \pi$, and thus [5.5a] can be rewritten in the form

$$(p_1 - p_2) r^2 \pi = 8 \pi \eta l \bar{v} \qquad [5.5d]$$

The left-hand side of [5.5d] (pressure difference multiplied by cross-section) is equal to the compressive force required to maintain the flow. At constant $\bar{v}$, however, the frictional force acting against the current is equal to this force. Consequently, a tube of length l exerts a frictional force $8\pi\eta l\bar{v}$ against the flow of a liquid of velocity $\bar{v}$ and viscosity η. In the present case the frictional resistance is characterized by this frictional force.

Further remarks

(a) A tube with varying cross-section may be regarded as consisting of parts with different cross-section connected in series (Fig. 5.3). The question arises as to what law governs the flow in this type of tube system. The cross-sections of the individual tube segments are constant, which means that for each tube segment of the system [5.5a] and [5.5b] can be applied. Without going into details, calculations relating to the overall tube system yield the equation

$$p_i + \frac{1}{2}\rho \bar{v}_i^2 - \left(p_0 + \frac{1}{2}\rho \bar{v}_0^2\right) = (R_1 + R_2 + \ldots + R_n) I \qquad [5.6]$$

where I is the constant current intensity, R_1, R_2 ... R_n are the resistances calculated from [5.5b], and p_i and $\bar{v}_i$ and p_0 and $\bar{v}_0$ are the static pressure and the velocity at the points of inflow and outflow, respectively. The left-hand side of [5.6] gives the *total pressure difference* between the ends of the tube system. The sum of the partial resistances

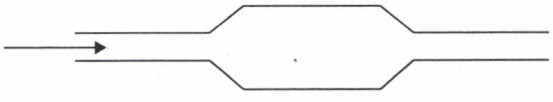


Fig. 5.3. Resistances connected in series

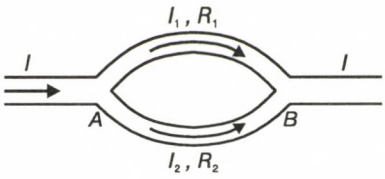


Fig. 5.4. Resistances connected in parallel

is called the total resistance, and thus [5.6] expresses the fact that the total pressure difference between the tube ends is equal to the product of the current intensity and the total resistance. [5.6] may be regarded as the generalized Hagen–Poiseuille law, which includes [5.5b] as a special case. [5.6] reduces to [5.5b] if the dynamic pressures at the ends of the tube are negligible as compared to the static pressures. Similarly, [5.5b] results if the tube cross-sections are the same, at least at the ends, i.e. when $\bar{v}_i = \bar{v}_0$.

(b) In Fig. 5.4 the sum of the current intensities in the branches is equal to the current intensity in the main branch, i.e.

$$I = I_1 + I_2 \qquad [5.7a]$$

From [5.5b] the pressure difference between points A and B is equal to the product of the current intensity and the resistance, i.e. to $I_1 R_1$ and $I_2 R_2$, respectively; consequently

$$I_1 R_1 = I_2 R_2 \quad \text{and} \quad \frac{I_1}{I_2} = \frac{R_2}{R_1} \qquad [5.7b]$$

This means that the current intensities are inversely proportional to the resistances. The current intensity I and the pressure difference associated with the section AB remain unchanged if the branches are substituted by a single tube of resistance R, such that $IR = I_1 R_1$ and $IR = I_2 R_2$. From these relations and [5.7a], the following equation is obtained:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \qquad [5.7c]$$

The reciprocal of the resistance is called the conductivity, and thus (5.7c] expresses the fact that the resultant conductivity is equal to the sum of the conductivities of the individual branches. This also holds for systems of more than two branches.

So far, the viscosity has been regarded as having a constant value characteristic of the fluid and changing only with temperature. However, there are some fluids whose viscosity also depends upon the pressure inducing the flow. The former group of fluids is called the *normal* or *Newtonian fluids,* while the latter *anomalous* or *non-Newtonian.* Pure liquids and real solutions belong in the first group, whereas the second group includes colloid solutions, emulsions and suspensions. In the latter case the dispersed particles are lamellar or fibrillar ones, and only a loose connection exists between them and the dispersing medium. The pressure difference inducing the flow orders the elongated particles and disrupts the loose structure between them with the consequence that in both cases the viscosity of this type of fluid decreases with increasing pressure drop. Blood behaves anomalously; at body temperature its viscosity is roughly 4.5 mPa s in the greater arteries and 2 mPa s in the smaller arteries.

The Hagen–Poiseuille law may have various applications, depending upon the known quantities, which in turn allow the unknown ones to be found. Depending upon the nature of the problem, this law may be used, for example, to determine the viscosity,

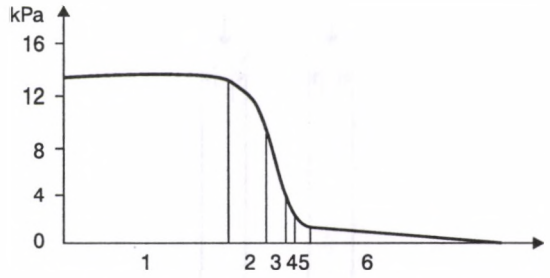


Fig. 5.5. Pressure drop in the greater circulation
1. great arteries, *2*: small arteries; *3*: arterioles; *4*: capillaries; *5*: venules; *6*: veins.
The ordinate gives pressure values above 101 kPa in kPa units

pressure distribution, and so on. It must be stressed, however, that the Hagen–Poiseuille law holds for the stationary flow of Newtonian fluids only if their velocity is below a certain critical value (cf. section 5.1.5). A small cross-section means that the tube radius is smaller than the boundary layer; in the case of water, for instance, this critical value is at most a few mm, and for blood at most 1–2 cm.

The above conditions are only partly satisfied in biological experiments and for this reason the Hagen–Poiseuille law can be used in these cases only to provide an approximate description of the actual situation, though this may be useful as a starting point.

Some informative data will now be given on the greater circulation. The flow velocity of blood in the aorta, with an inner diameter of approximately 2 cm, is at most 30–40 cm/s. The great arteries branch off into arteries of smaller diameter; these in turn divide into the still smaller arterioles, and finally into the capillaries. The branching can be regarded as a tube system connected in parallel, whose total cross-section increases with increasing branching. The flow velocity in the capillaries is only about 0.05–0.08 cm/s, from which it follows that the total cross-section of the capillaries is approximately 600–800 times larger than the cross-section of the great arteries. The capillaries unite into venules, and these into veins, which results in the decrease of the total cross-section and consequently in an increase of the flow velocity. For example, the flow velocity in the vein joining the right atrium is 6–14 cm/s. On contraction, the pressure in the left ventricle increases to about 13–16 kPa above atmospheric pressure, whereas in the vein flowing into the right atrium the pressure is atmospheric. Figure 5.5 depicts the distribution of the pressure above 101 kPa along the greater circulation. The pressure is seen to decrease considerably in the arterioles and the capillaries, due to the large frictional resistance of these systems. At first sight it might appear surprising that the frictional resistance is large in the arterioles and the capillaries, whose total cross-section is high. In order to explain this fact, the capillary system may be modelled by n tubes of identical radius r, connected in parallel. From [5.5c], the resistance of a single tube is inversely proportional to r^4. It follows from [5.7c] that the resultant resistance of n tubes connected in parallel is inversely proportional to nr^4. Since the total cross-section q is proportional to nr^2, the resultant resistance of the tube system is inversely proportional to qr^2. Thus, it is quite conceivable that, even with a large q, qr^2 will be small if r is sufficiently small, and consequently the resultant resistance will be high. This is the situation with the arterioles and capillaries.

5.1.5. Laminar and turbulent flow

Let water flow downwards in a vertical glass tube (Fig. 5.6). The water source consists of two vessels; from one, coloured water flows into the centre of the tube through an opening, while colourless water flows from the other into the part surrounding the opening. The velocity of flow can be regulated by a tap at the lower end of the tube. As

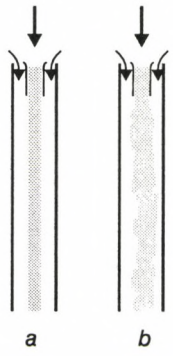


Fig. 5.6. Laminar flow (*a*); turbulent flow (*b*)

long as the flow velocity is small, the coloured water does not mix with the colourless one (Fig. 5.6*a*), and a coloured fluid thread well separated from its surroundings can be observed in the tube axis. This type of flow is called *layered* or *laminar* flow. If a given velocity is exceeded, however, rotations (eddies) are added to the unidirectional motion of the fluid particles and a confused flow results (Fig. 5.6*b*). The flow pattern changes continually. This type of flow is called *turbulent* flow. Everything discussed in the previous sections refers only to laminar flow. Turbulence increases the frictional resistance, and the statement that the internal frictional force is proportional to the velocity as expressed in [5.3] and [5.4] and in the Hagen–Poiseuille law [5.5d] no longer holds. In the case of turbulence the frictional force is approximately proportional to the square of the flow velocity; the compressive forces maintaining the flow perform work against the frictional forces. Since the frictional resistance increases with the occurrence of turbulence, more work is required to maintain the same current intensity.

The velocity above which laminar flow passes over into turbulent flow is the *critical velocity.* From the investigations of Reynolds, the critical velocity v_c depends upon the viscosity η and the density ρ of the fluid, and upon the radius r of the tube:

$$v_c = Re \frac{\eta}{\rho r} \qquad [5.8]$$

The dimensionless factor *Re* is called the *Reynolds number,* whose value is 1160 for smooth-walled tubes. *Re* is smaller for tubes with rough walls. Thus from [5.8] the flow of water (at 18 °C) in a glass tube of 1 mm radius becomes turbulent only above a velocity of 127 cm/s; in the case of a tube with a radius of 1.0 cm, the critical velocity is one tenth of this value: v_c = 12.7 cm/s. The critical velocity of blood in a smooth-walled tube of 1 cm radius would be 50 cm/s; the actual value in the blood vessels is generally smaller than this.

Under healthy conditions, the vascular flow of the blood is laminar. Turbulence occurs only at some places, e.g. in the aorta behind the semilunar cardiac valves. In certain pathological cases, however, v_c decreases because of the decrease in η, and the turbulence may extend over larger sections. The larger the sections of turbulent flow, the more work the heart has to do to ensure the blood supply.

Turbulent motion is accompanied by a low humming sound, which can be heard in the artery of the arm, for instance, when blood pressure is taken. This sound is due to the fact that as the cross-section of the artery decreases the flow velocity increases and exceeds the critical value.

The air flow in the nasal canals is laminar under healthy conditions. Under pathological conditions, however, the nasal canals may become so narrow that the air flow becomes turbulent in some sections. In such cases breathing becomes difficult, resulting in increased work by the breathing muscles.

5.1.6. Flow in tubes with elastic walls

In the previous section, no difference was made between the flow processes in tubes with rigid or with elastic walls, for under conditions of stationary flow the discussed relations can be applied to both cases. Though an elastic tube yielding to pressure expands at the onset of flow, the geometrical dimensions (radius, length, etc.) do not change in the course of flow after it has become stationary. Of course, the new geometrical values of the tube formed during stationary flow must be inserted into the relations describing the flow.

The situation is considerably more complicated in the case of fluctuating pressure values. Instead of a quantitative discussion, we shall be satisfied with a description of an experiment which reveals the essential differences between tubes with rigid or with elastic walls. In Fig. 5.7 the glass tube protruding from a water vessel branches into two parts. One branch is connected to a rubber tube A, and the other via a short rubber connection to a glass tube B. The cross-sections are selected so that stationary flow of identical intensity is produced in both tubes. If the rubber tubes are compressed by the bar C at short time intervals (1–2 times per second) during the flow, the flow becomes intermittent; the water flows will differ in the two branches, and in a given time considerably more water will emerge from the end of tube A than from tube B.

This phenomenon can be explained in the following way. Because of its elasticity, the rubber tube expands and contracts periodically, in accordance with the pressure changes. On expansion the kinetic energy of the water is partly transformed into elastic energy, and on contraction part of the elastic energy of the wall is retransformed into kinetic energy. No such energy transformations are produced in the rigid-walled tube, and the lost mechanical energy is transformed into heat as a result of friction.

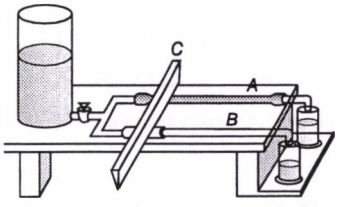


Fig. 5.7. Flow in tubes with elastic and with rigid walls

This experiment is instructive from the viewpoint of blood circulation and supports the physiological observation that the elasticity of the blood vessels is an essential and deciding factor of normal circulation.

5.2. Diffusion and osmosis

5.2.1. Fick's laws

If density or concentration differences exist in some medium, material flow ensues from the higher to the lower densities or concentrations. The spontaneous equalization of the density and concentration differences is called diffusion, which can be observed in gases, fluids and solid states. The phenomenon can be interpreted in terms of molecular thermal motion.

The quantitative discussion of this phenomenon is carried out with reference to Fig. 5.8. For simplicity we assume that the concentration c of the solution[1] in the vessel changes only in one direction (Z), upwards, and that the diffusion takes place in this direction. Let us denote the concentration change along the length dz in the direction of the diffusion at the investigated height A by dc. The quantity dc/dz is called the *concentration gradient.* By *Fick's first law,* the amount dv of substance which migrates by diffusion in time dt across the cross-section q is proportional to the concentration gradient, the cross-section and the time, i.e.

$$dv = -Dq\frac{dc}{dz}dt, \quad \text{or} \quad \frac{dv}{dt} = -Dq\frac{dc}{dz} \qquad [5.9]$$

The quotient dv/dt is the *rate of diffusion.* (The negative sign is necessary since dc/dz is a negative quantity.) The proportionality factor D (the *diffusion coefficient*) gives the amount of substance migrating in unit time across unit cross-section in the case of a unit concentration gradient.

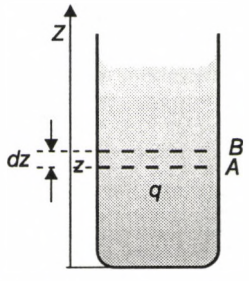


Fig. 5.8. A scheme relating to diffusion

[1] In this case the *concentration of substance* is used, which is the number of moles of the substance in question related to unit volume of solution. Its dimensions are mol/m^3; mol/l is also used, the latter being the *molarity.*

Fick's first law holds lastingly only if the concentration distribution does not change in time *(stationary diffusion)*. With *non-stationary diffusion* the concentration of the material studied is a function not only of place, but also of time. *Fick's second law* holds for this case. Assuming flow only in the Z direction in this case too, the relation

$$\frac{\delta c}{\delta t} = D \frac{\delta^2 c}{\delta z^2} \qquad [5.10a]$$

is obtained. Thus, the change in the concentration in time at a given place is proportional to the change in the concentration gradient in space, with the proportionality factor D again denoting the diffusion coefficient. (The partial differentiation is a consequence of the fact that in this case c is a function not only of the z coordinate but also of the time t; cf. Appendix, section B.5.)

[5.10a] may be obtained in the following way. Consider the layer of thickness dz (Fig. 5.8). In the foregoing the concentration gradients were taken as equal at the bottom (A) and at the top (B) of the layer, but now we assume different gradients at the two boundaries of the layer. Consequently, in the first case the concentration in the layer remains constant in the course of diffusion (stationary diffusion), in the second one it changes in time (non-stationary diffusion). In other words: in the first case the number of molecules diffusing into the layer through surface A is equal to that of molecules leaving through surface B but in the latter case these values are different. Thus, from [5.9] for surface A

$$(dv)_A = -Dq \left(\frac{\delta c}{\delta z}\right)_A dt \qquad [5.10b]$$

and for surface B

$$(dv)_B = -Dq \left(\frac{\delta c}{\delta z}\right)_B dt \qquad [5.10c]$$

can be written.

The subscripts indicate the relevant surfaces. The partial differentiation gives the derivative of c with respect to z at a given time t.

It can be easily seen that

$$\left(\frac{\delta c}{\delta z}\right)_B - \left(\frac{\delta c}{\delta z}\right)_A = \frac{\delta^2 c}{\delta z^2} dz \qquad [5.10d]$$

According to mathematical analysis the difference between the function values taken at z and $z + dz$ is equal to the product of the derivative taken at z and dz. The function in this case is the concentration gradient. Since the concentration gradient is the first derivative of concentration with respect to z, on the left side the difference of these derivatives, and on the right side the second derivative of concentration multiplied by dz can be seen.

Subtracting [5.10c] from [5.10b] and taking into account [5.10d]

$$(dv)_A - (dv)_B = Dq \frac{\delta^2 c}{\delta z^2} dz\, dt \qquad [5.10e]$$

is obtained. The left side gives the change of the diffusing amount of substance in volume $q\, dz$ during time dt. Dividing [5.10c] by the expression $q\, dz\, dt$, one obtains on the left side the variation of concentration in unit time which may be denoted *by* $\delta c/\delta t$. Thus, after division the equation can be written in the form

$$\frac{\delta c}{\delta t} = D \frac{\delta^2 c}{\delta z^2}$$

which is identical to [5.10a].

In practice one deals mostly with non-stationary diffusion and to maintain the stationary process presents a particular task. Accordingly, the second law is more important than it seems at first sight. The starting point for the determination of D is generally also the equation [5.10a].

For gases, the diffusion coefficient is approximately proportional to the square root of the absolute temperature, quite similarly to the average thermal velocity of the molecules. In fluids or solids, which are much more close-packed, diffusion is possible because intermolecular gaps (vacancies) always exist which mediate the molecular migration. Consequently, it is understandable that in a given case the diffusion coefficient is proportional to the Boltzmann factor:

$$D \sim e^{-\frac{\varepsilon}{kT}} \qquad [5.11]$$

where ε is the *activation energy of the migration of the investigated molecular species.* According to [5.11], D varies exponentially with the temperature in condensed systems. If the molecules in question are those of the system itself, the diffusion is called self-diffusion and the energies of activation characteristic of the viscosity and diffusion, respectively, are identical.

It follows from [5.3c] and [5.11] that in a given case viscosity η and diffusion coefficient D are related. For spherical macromolecules or colloidal particles, for instance, regardless of whether the diffusion occurs in gases or liquids, the following equation derived by *Einstein* describes the process to a good approximation:

$$D = \frac{kT}{6\pi\eta r} \qquad [5.12]$$

where r is the radius of the diffusing particle (cf. section 5.1.1). [5.12] is used in several ways, mainly to determine the dimensions and masses (molar mass) of macromolecules.

Table 5.1 lists the values of the diffusion coefficient for a few molecules. Because of the molecular interactions, the value of D depends upon the concentration (the effect may be considerable even at small concentrations in the case of macromolecules), and hence these values are extrapolated to infinite dilution.

Table 5.1. Diffusion coefficients of some compounds in aqueous solution at 20 °C

Compound	D (m^2/s)
Glycine	9.5×10^{-10}
Leucyl-glycyl-glycine	4.6×10^{-10}
Ribonuclease	10.2×10^{-11}
Human serum albumin	6.1×10^{-11}
Human haemoglobin	6.8×10^{-11}
Tobacco mosaic virus	3.0×10^{-12}

The tabulated data show that for proteins the value of the diffusion coefficient is in the range 10^{-12}–10^{-10} m^2/s, and is more than one order of magnitude smaller than for more simple molecules (e.g. glycine). The differences are due to the differences between the molecular dimensions.

5.2.2. Van't Hoff's law

In this section we deal separately with the case of diffusion across a wall, when the wall can be permeated only by some components (*semipermeable wall* or membrane), but is impenetrable for others. For simplicity, we shall discuss only solutions consisting of two components, and the wall is permeable only for the *solvent.* Immerse a cellophane bag filled with sugar solution into pure water. The bag will swell in a few hours. The compartment outside the bag still contains only water, but the solution within the bag becomes more dilute. Though both the solute and the solvent strive towards uniform distribution, the possibility of this is given for only one component. In our experiment the pressure within the bag steadily increases due to the influx of water, but after some time an equilibrium (dynamic equilibrium) is attained, when the same solvent quantity diffuses into the bag in unit time as is forced out from it by the pressure difference. This

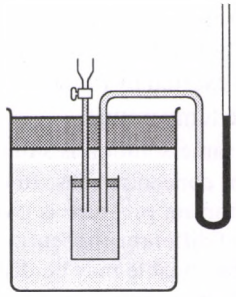


Fig. 5.9. Equipment for osmotic pressure measurement

phenomenon is called *osmosis,* and the pressure difference which can compensate the influx of the solvent into the bag is the *osmotic pressure.*

Figure 5.9 depicts a simple set-up for measuring, e.g., the osmotic pressure of sugar solutions. The suitably prepared wall of the inner clay vessel lets through the water, but not the sugar. The sugar solution is poured into the inner, and the pure water into the outer vessel. After equilibrium has been attained, the osmotic pressure can be read off the mercury manometer.

The osmosis can be interpreted in terms of the tension decrease of the solutions. The solvent, with the higher tension, is situated on one side of the semipermeable wall, and the solution, with a lower tension, on the other side. The molecules of the solvent can pass through the wall; as a result of the higher tension, more molecules will diffuse through the wall from the pure solvent than in the reverse direction, and hence more solvent will pass into the solution. However, as the pressure within the vessel containing the solution increases due to the solvent migration, the number of molecules diffusing from the solution towards the solvent will also increase. If the pressure becomes large enough, the same number of molecules diffuses towards the pure solvent as in the opposite direction, and dynamic equilibrium is attained between the solution and the solvent. The osmotic pressure (p_{osm}) is the pressure required for the development of dynamic equilibrium at a given temperature. For fairly dilute solution of a non-volatile solute, the experimental results lead to a simple relation, called *van't Hoff's law:*

$$p_{osm} = RTc \qquad [5.13a]$$

where c is the concentration of the solution, and R is the universal gas constant. If c is given in mol/m^3, and R in J/(mol K), p_{osm} is obtained in Pa. Thus, for a dilute solution the osmotic pressure at a given temperature is proportional to the concentration of the solution, and is independent of the material of the solvent and solute. According to [5.13a] the osmotic pressure of a solution at 20 °C with a concentration of 10 mol/m^3 is 24.3 kPa.

If a solution of volume V contains v mol solute, $c = v/V$. Substitution into [5.13a] yields the equation

$$pV = vRT \qquad [5.13b]$$

[5.13b] formally corresponds to the universal gas law and expresses the fact that the osmotic pressure of a dilute solution (independently of the nature of the solvent and solute) is the same as the pressure exerted by the solute in gaseous form if it occupies the volume of the solution at the same temperature.

Osmosis also takes place if the same solution is situated on both sides of the semipermeable membrane, but at different concentrations. In every case the more concentrated solution will be diluted. The equilibrium pressure is given by the difference between the osmotic pressures of the solutions of different concentrations. Naturally, different substances for which the membrane is impermeable may be dissolved on the two sides. From the viewpoint of osmosis the nature of the solute is immaterial. At a given temperature, only the (molar) concentration is important, and the process of osmosis is induced if the concentrations are different on the two sides. Solutions of equal osmotic pressure are called *isotonic* solutions.

5.3. Basic concepts of thermodynamics

How can it be explained that in the living organism the same amount of heat is produced, e.g. during the oxidation of lipids and carbohydrates as – under much simpler circumstances – in a bomb calorimeter? In other words: why is *Hess' law* (1840) true, according to which during chemical transformations the sum of the transformational heats depends only on the initial and final states and is independent of the intermediate phases? – The principle of the conservation of energy "allows" a natural process to go in both directions, "there" and "back". In reality this is not that simple. Namely the processes proceed spontaneously in one direction, in the opposite direction only with some "external" help. For example the equalization of differences in temperature, pressure and concentration is a spontaneous process, but for their development external help is required. How can the spontaneous direction be characterized? – How can the chemical equilibrium be characterized? – How much of the energy released during the chemical reactions may be used for mechanical or electric, etc. work and how much of it appears unavoidably as heat? What are the relationships in the vital processes?

One of the fundamental problems of physiology are the excitation processes. Where are we in their understanding?

These and similar questions will be answered in what follows. All these questions pertain to the domain of thermodynamics the object of which is in general the investigation of processes and equilibria occurring during the different (mechanical, electric, chemical, etc.) interactions from the aspect of energetics. Thus thermodynamics is the basic science of every basic and applied natural science, such as chemistry, biology, physiology, etc. Our examples will also be taken from here and there, but the essence of the statements have a general validity.

5.3.1. Formulation of the first law. Internal energy

It follows from the law of energy conservation that, if the energy supplied to a so-called thermodynamic system containing chemical substances or released from the system to its

environment is known, the change in the energy content of the system is the algebraic sum of the energies taken and released by it.[2]

In thermodynamics we always deal with cases in which the changes in the energy supply of the system arise from the changes of the *internal energy of the system.* The internal energy includes all kinds of energies belonging to the atoms and molecules: thus it comprises also the kinetic energy of the constant, random motion of the atoms and molecules and the potential energy of the interacting atoms and molecules (the bond energy).

The energies taken and released can be divided into two groups. One group involves only thermal energy and the second all the various other energy types (mechanical, electromagnetic, chemical), and these are taken into account as work done. (The separation is motivated by the special role of thermal energy in Nature, which will be expressed in the second law of thermodynamics.)

Let us denote the energy content of a system at the beginning of some process by U_1 and at the end of the process by U_2, while for the sake of brevity the energy change $U_2 - U_1$ is denoted *by* ΔU. Let us further denote the heat transferred during the process by Q, and the work done by W. According to the general energy theorem:

$$\Delta U = Q + W \qquad [5.14a]$$

which means that *the change of the energy content in the process is equal to the algebraic sum of the heat transferred and the work done during the process.* Q is a positive quantity if it denotes the heat taken by the system, and is negative in the case of a heat loss; further, W is positive if work is done on the system and negative if it is done by the system.

For infinitesimally small changes the following form is used

$$dU = dQ + dW \qquad [5.14b]$$

The changes refer to positive or negative quantities, respectively, if they increase or decrease the energy of the system.

Both forms of [5.14] are the *basic equations of the first law of thermodynamics.* If it is added that the *internal energy is a state function,* the formulation of the first law of thermodynamics is complete.

The behaviour of the internal energy as state function means that it, as the function of state markers or thermodynamical parameters (temperature, volume, mass, pressure, concentration, etc.), depends only on the initial and final state, but is independent of the ways and phases by which the system gets from the initial state to the final state. The immediate consequence of this feature is Hess' law mentioned in the introduction of section 5.3 and within this the possibility that the physiologically useful value of nutrients measured in units of energy may be determined also relatively simply, in a bomb calorimeter.[3]

[2] The taken and released energies are considered to have opposite signs.

[3] The transferred heat and the work done themselves generally depend not only on the initial and final states, but also on the means of change. The symbol d usually denotes only the change in the state functions, while the changes in the other functions are denoted by the letters D or δ. In this book we disregard this notation and the infinitesimally small changes are uniformly denoted by the symbol d.

The first law in its forms given above should be considered a "frame law", its practical applications will be discussed below.

5.3.2. Addenda to the first law. Enthalpy

1. Simple cases of the thermal and mechanical interactions. For the sake of illustration consider the heating or cooling of a gas closed in a cylinder with piston at a constant volume (isochoric) or under a constant pressure (isobaric).

The substantial difference between the isochoric and isobaric conditions is that in the first case the change of the internal energy (ΔU) is provided only by the exchanged heat (Q_V), while in the latter case in addition to the heat (Q_p) also the volumetric work connected to the expansion or contraction ($-p\Delta V$) should be taken into consideration. Thus in isochoric cases

$$\Delta U = Q_V, \qquad [5.15]$$

in isobaric cases

$$\Delta U = Q_p - p\Delta V. \qquad [5.16]$$

The index V refers to the constancy of the volume, index p to that of the pressure, respectively. By writing the work on the left-hand side of the equation the content of equation [5.16] may be formulated so that only a part of the supplied heat increases the internal energy in case of isobaric heating, the other part is converted into work.

The role of the negative sign in the expression of the volumetric work may be easily seen if we consider the following. During its expansion, i.e. at positive ΔV, the work is done by the system (body) and this is given by the expression with a negative sign, in accordance with the sign convention (cf. section 5.3.1). During contraction, i.e. at negative ΔV, on the other hand, the work is done on the system, and the expression gives this work as a positive quantity by the negative sign, in accordance with the convention.

2. Enthalpy. In practice the different physical and chemical processes take place at constant pressure, usually under atmospheric conditions. This statement is also valid for living processes.

Processes at constant pressure can be described in a simple way, if a new state function, called *enthalpy* (H), is introduced:

$$H = U + pV \qquad [5.17a]$$

where U is the internal energy of the system, V is the volume and p the pressure of the system in *equilibrium* at a given volume and temperature. It follows from the definition that the equilibrium pressure is always equal to the external pressure (for instance the atmospheric pressure) acting on the system. For processes at constant pressure the change in the enthalpy due to the changes in U and V will be

$$\Delta H = \Delta U + p\Delta V \quad \text{and} \quad dH = dU + pdV \qquad [5.17b]$$

respectively. Thus, the enthalpy change is given by the sum of the internal energy change and the volumetric work. This allows the change in the internal energy and the volumetric work in *isobaric* processes to be expressed as the change in a single quantity. For instance, after appropriate rearrangement [5.16b] can be written in the form $\Delta H = Q_p$. It is generally true that in *isobaric* processes, where there is only volumetric work, the transferred heat is equal to the change in the enthalpy:

$$\Delta H = Q_p \qquad [5.18]$$

and in those *isosteric* processes where no work other than the naturally missing volumetric work is involved, the transferred heat is equal to the change in the internal energy

$$\Delta U = Q_V \qquad [5.19]$$

3. Transition heat and enthalpy (internal energy). If a solid is melted at constant temperature and the small volume change on melting is disregarded, the total heat input will be used to increase the internal energy. Consequently, we may write a relation similar to [5.19], where the heat input (and consequently the change in the internal energy) can be expressed with the melting heat. On evaporation, however, the volume increase resulting from the transformation into vapour is accompanied by considerable work, which must be taken into account. In this case the energy relations are described by [5.18]. The heat uptake which can be expressed by the heat of evaporation is equal to the enthalpy change. If the volumetric work is subtracted from this, the internal energy change due to evaporation is obtained. The first law can be applied in a similar way to processes in the opposite direction, i.e. to freezing and condensation, which are accompanied by a decrease in the internal energy, i.e. decrease in the enthalpy. Our findings accordingly hold for all types of phase transitions.

Measurements indicate that a work of 40.7 kJ is required for the evaporation of 1 mol (18 g) water at 100 °C and a pressure of 101 kPa. This is the value of the enthalpy increase in this case. Part of this work is used to perform the volumetric work against the atmospheric pressure; its value is

$$W = -p(V_{\text{vapour}} - V_{\text{water}}) \approx -pV_{\text{vapour}}$$

Namely V_{water} may be neglected in comparison to V_{vapour}.

However, according to the universal gas law

$$pV_{\text{vapour}} = vRT$$

Since $R = 8.31$ J/(mol K) and in this case $v = 1$ mol and $T = 373.16$ K, in our example we have

$$W \approx -3.1 \text{ kJ}$$

and the increase in the internal energy is only 37.6 kJ. Consequently, approximately 0.4 eV per molecule is required to vaporize water at 100 °C.

4. Reaction heats and enthalpy (internal energy). An understanding of chemical reactions requires a knowledge of the heat released or taken during the reaction at constant temperature (isothermal process). The amounts of substances participating in a reaction are usually given in mol units, and consequently the released or absorbed heat is related to 1 mol of substances; this is called the *reaction heat.*

In isothermal–isosteric reactions the reaction heat is measured at constant volume, and in isothermal–isobaric reactions at constant pressure. In the first case the change in the internal energy (also related to one mol) is equal to the reaction heat Q_V, i.e.

$$\Delta U = Q_V \qquad [5.20]$$

while in the latter case the heat of reaction is partly required for work associated with volume changes, and thus the reaction heat Q_V is equal to the change in the enthalpy and not in the internal energy, i.e.

$$\Delta H = Q_p \qquad [5.21]$$

In reactions in liquid or solid phase the volumetric work can be neglected and [5.20] accepted as valid. In reactions in the gaseous phase, however, when the volumetric work in general cannot be neglected, calculations should be carried out by [5.21].

Measurements on the conversion of 1 mol dextrose to 2 mol ethyl alcohol and 2 mol carbon dioxide at 25 °C and 101 kPa pressure show that 71.2 kJ heat is released. The enthalpy change for 1 mol dextrose is thus $\Delta H = -71.2$ kJ. The change in the internal energy can be determined by considering the volumetric work done during the process, which is obtained mainly from the formation of 2 mol gaseous carbon dioxide. This work can be calculated in a similar way as for the evaporation of water in the previous example. The result, i.e. the work done by the system, is $-p\Delta V = -5$ kJ. Thus, the change in the internal energy is $\Delta U = \Delta H - p\Delta V = -76.2$ kJ. This means that rearrangement of the atoms of a dextrose molecule into two ethyl alcohol and two carbon dioxide molecules results in a more stable configuration than the original one, as characterized by the release of approximately 0.8 eV.

5. Determination of the enthalpy (internal energy). In practice usually the changes of the enthalpy (internal energy) during different processes and transformations are of interest. For this only the absorbed or released heat and the volumetric work must be known, which does not present a special problem. For the sake of uniformity, however, international conventions regulate the initial values and states to be considered when determining the changes. Accordingly, the enthalpy (internal energy) is fixed so that the enthalpies (internal energies) of the *chemical elements* at 25 °C and 101 kPa (in the state in which they are stable) are considered to be zero. From this it follows that the enthalpies (internal energies) of the *chemical compounds* at 25 °C and 101 kPa are equal to their heats of formation at constant pressure (constant volume). This heat of formation (or more exactly its value related to one mol) is called the *standard heat of formation* or *standard enthalpy,* denoted *by* H° or ΔH°. Table 5.2 lists the standard enthalpies of some substances.

Table 5.2. Standard enthalpies (H°) of some substances

Element or compound	State	H° (kJ/mol)
H_2	g	0.0
O_2	g	0.0
C (graphite)	s	0.0
H_2O	l	−286.0
H_2O	g	−242.0
CO_2	g	−394.0
Acetic acid	l	−487.4
Lactic acid	l	−677.0
Ethyl alcohol	l	−278.0
Glycerine	l	−666.6
Glucose	s	−1280.1

g = gas or vapour; l = liquid; s = solid

5.3.3. Formulation of the second law. A statistical interpretation of entropy

1. A certain formulation of the law. According to the experience, as already mentioned (cf. the introduction of section 5.3), in reality not every process takes place which is possible on the basis of the first law. The direction of the spontaneous processes is given by the second law.

The factors determining the direction of spontaneous processes are related to *heat,* and in every case can be ascribed to the universal experience that *heat always flows spontaneously from a warmer to a cooler body.* Various, equivalent definitions of the second law are known, although perhaps the above is the simplest of all; its content can readily be followed as concerns atomic or molecular aspects. Heat is a form of energy, which has its origin in the random motion of atoms or molecules. When two systems come into contact, for instance, that in which the mean kinetic energy of the molecules is higher can transfer heat to the other. This seems to be quite natural. If billiard balls collide, it is more probable that the ball with higher energy will transfer some of its energy to the ball with lower energy, and not vice versa; the latter process may also occur, but with much lower probability. The situation is quite similar to the case of molecular collisions. In principle it may be possible that the molecules of the warmer body gain energy from the molecules of the colder one, i.e. the warmer body obtains energy from the colder one, but this is so improbable that it does not occur spontaneously in practice. This type of process can take place only with some external aid. (Consider e.g. the refrigerator.) According to this concept, the *second law of thermodynamics is a statistical one in character and simply expresses the fact that the thermodynamic processes progress spontaneously towards the more probable state.* In principle the opposite processes may also occur but this is not probable spontaneously in practice. Processes in the opposite direction can occur in reality only if they are accompanied by some other changes. In an isolated system, processes of any kind can be observed only until the system reaches its most probable state; when this has been

attained, the system has reached thermodynamic equilibrium, from which it can be displaced only by some external effect.

2. Entropy. The quantitative formulation of the first law was made possible by the use of a state function, the internal energy. The quantitative definition of the second law requires a new state function, *entropy.* Entropy can be expressed by probabilities, more exactly as will be shown below, *by thermodynamic probabilities* associated with the individual states. According to Boltzmann, the entropy S of a system in a given state is proportional to the logarithm of the thermodynamic probability (w) associated with this state. If natural logarithms are used the proportionality factor k is the Boltzmann constant; thus, we have

$$S = k \ln w \qquad [5.22a]$$

Let us denote the thermodynamic probability at the beginning of the process by w_A and the entropy by S_A, and in the final state let the probability be w_B and the entropy S_B. With these notations the entropy change will be expressed by the relation

$$S_B - S_A = k \ln \frac{w_B}{w_A} \qquad [5.22b]$$

It is generally true that in any isolated system the processes progress in the direction of an increase in the thermodynamic probability and hence in the entropy. The thermodynamic equilibrium is characterized by the maximum of the thermodynamic probability, and by that of the entropy.

The determination of the thermodynamic probability is illustrated by a simple example. Let us consider an isolated vessel which contains a gas of 4 identical point-like molecules denoted by the letters *a, b, c* and *d.* We now examine the possible distributions of these molecules in the two halves (cells I and II) of the vessel. The possibilities are summarized in column 3 of Table 5.3. Altogether 16 random distributions or *microstates* are conceivable. However, measurement of the various physical properties (e.g. the density) reveals no difference between the various microstates, i.e. from a macroscopic point of view e.g. the 4 various microstates in the 2nd row are identical. The 6 microstates in the 3rd row are different from the states in the 2nd row, but identical with each other. Similarly, the microstates in the 4th row, although differing from the states in the other rows, are again found to be identical macroscopically. In reality only five different states can be observed; these *macrostates* are denoted by capital letters in column 1. The number of microstates related to the individual macrostates are given in column 4. If it is assumed that the individual microstates are equally probable, it is quite clear that the most probable macrostate will be the state associated with the highest number of microstates. In the above case, therefore, the most probable state is state C (i.e. uniform distribution), and states A and E are the least probable, the distributions there being the least uniform. *The thermodynamic probability of a macrostate is characterized by the number of microstates associated with the respective macrostate.*

Table 5.3. Example to illustrate thermodynamic probability

Macro-states	Molecules in cell I		Number of microstates
	number	notation	
A	4	abcd	1
B	3	abc, abd, acd, bcd	4
C	2	ab, ac, ad, bc, bd, cd	6
D	1	a, b, c, d	4
E	0	—	1

The thermodynamic (statistical) probability is not identical with the mathematical probability. For calculation of mathematical probability associated with some macrostate, the quotient of the numbers of favourable and possible cases is needed. The number of favourable cases of a certain macrostate is given by the number of the *associated* microstates, and the number of possible cases (for any macrostate) is equal to the number of *all* possible microstates. For instance, the thermodynamic probability associated with macrostate *B* in the above example is 4, while the mathematical probability is 4/16; the corresponding quantities in case *C* are 6 and 6/16.

In practice one usually deals with the *changes of state* of a system. In this context the *quotient* of the probabilities is considered as in [5.22b]. It follows that, because of the quotient formation in the case of changes of states, the thermodynamic probabilities yield the same results as the mathematical probabilities.

In the previous example the possibilities of the spatial distribution have been investigated. Similar calculations could be carried out with respect to the distribution of the kinetic energy. The result would show in this case too that the number of microstates is highest when the kinetic energy, i.e. the temperature, is identical in the cells.

If the system consists of multiatomic molecules, the motion of the molecules includes their rotations and the vibrations of the atoms in the molecule. All these motions must be taken into consideration in a calculation of the total entropy of the system. It follows that the entropy is the sum of several terms. All processes increasing the molecular mobility also increase the entropy of the respective system. Examples of such processes are melting, evaporation, dissolution, diffusion, the expansion of gases, etc. The loosening or cleavage of the bonds between the atoms within the molecules, e.g. molecular dissociation, results in an increase of entropy. In contrast, all processes leading to a strengthening of the atomic bonds or limiting the molecular motion decrease the entropy. Such processes are freezing, condensation and the formation of molecules from atoms or atomic groups.

An increase in the thermodynamic probability can be regarded as associated with a decreased ordering, and vice versa. For instance, the uniform distribution of gases in a given volume is viewed as a lower degree of ordering than their distribution in only part of the volume. For this reason it may be stated that *thermodynamic processes in a system without external influences proceed in the direction of lower ordering.*

It follows from the statistical meaning of the second law that the equilibrium state of a thermodynamic system is its most probable state; however, this may allow some local and transient fluctuations. For instance, temperature, pressure, density and concentration fluctuations may be expected. Such phenomena can be observed, and are experimental proof of the molecular heat theory. As an example, the blue colour of the sky may be explained by the irregular, slight density fluctuations of the air. Brownian motion is also a fluctuation phenomenon.

5.3.4. The phenomenological formulation of entropy

1. Definition. In practice entropy is usually determined by means of directly measurable quantities, according to the development of the concept, rather than by calculating the thermodynamic probability. The development of the concept will not be detailed, only its definition will be dealt with.

Let dQ stand for the heat exchanged during the infinitesimal change of state between a system and its surroundings at T temperature. The infinitesimal change of entropy (S) is given by

$$dS = \frac{dQ}{T} \qquad [5.23a]$$

In case of finite changes, when the temperature also changes, the total change of entropy is given by the summation of the elementary changes (integration, cf. Appendix B4).

Let a thermodynamic system get from state A in state B. Let the entropy of the system be S_A in state A and S_B in state B. The change in entropy during the process $A \rightarrow B$ is given by the following quantity expressed in definite integral:

$$S_A - S_B = \int_A^B \frac{dQ}{T} \qquad [5.23b]$$

The above must be completed by an important statement. Namely the exchanged heat depends on the way leading from state A to state B. For example, the amount of heat developed during the transformation of glucose to lactic acid may be larger but also smaller and thus less but also more work may be obtained, depending on the circumstances of the process (e.g. whether it is more rapid or slower). The definition of entropy refers to the case in which the process in question takes place almost infinitely slowly, *infinitesimally closely to the prevailing state of equilibrium, in other words quasi-statically.* In this case the process passes through the same states in both directions, thus it returns to the initial state of the system without any changes in the environment. Such processes are *reversible* in the thermodynamical sense. *In the definition of entropy the heat exchanged during a reversible process is given.*

In Nature there are no reversible processes in a strict sense, the real processes are more or less irreversible. This means that a system can return to its initial state only with some external help and therefore there is always some change remaining in the surroundings. The reversible change is an ideal borderline case which may be conceived and followed by calculations. Its study is important for the practice, because the obtained results give information concerning also the real processes. This is proven by the following examples which show the determination of entropy according to its definition in concrete cases and within this also the possible reversibility of the given processes.

The statistical and phenomenological definitions of entropy are very different from each other formally, but their equivalence has been proven by *Boltzmann* already in the second half of the last century. In one of our examples this equivalence will be demonstrated.

The entropy is a state function, its value depends only on the initial and final stages of the system, it is independent of the "way" by which the change of state occurs. Its unit is J/K.

2. Examples

1. Let us calculate the change of entropy of a gas in a cylinder closed by a piston moving "without friction" at *isothermal expansion* or *compression.* This example is a useful preliminary study for several problems, e.g. for the examination of isothermal diffusion and mixing or for the construction of the definition of the chemical potential.

Let us consider the isothermal expansion. For the calculation of entropy a reversible process must be examined, i.e. a state of change during which the external and internal pressures are in equilibrium in every moment. The pressure p of moles v of a gas for different V volumes at temperature T is given by the universal gas law:

$$p = vRT/V$$

where R is the universal gas constant. The relation between V and p is illustrated by Fig. 5.10. On isothermal expansion the gas performs work against the external pressure and takes an according amount of heat from its surroundings. The work done (W_{AB}) may be calculated by integration:

$$W_{AB} = -\int_{V_A}^{V_B} p\Delta V \qquad [5.24a]$$

Considering the foregoing and by integration:

$$W_{AB} = -vRT \ln \frac{V_B}{V_A} \qquad [5.24b]$$

In Fig. the work is shown by the darkened area under the curve.

Since the heat taken (Q_{AB}) is equal to the work done:

$$Q_{AB} = -vRT \ln \frac{V_B}{V_A} \qquad [5.24c]$$

The change of entropy:

$$S_B - S_A = -vR \ln \frac{V_B}{V_A} \qquad [5.24d]$$

Since the heat is taken up, the entropy increases which is in accordance also with the statistical interpretation of entropy.

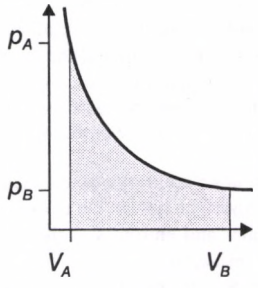


Fig. 5.10. Diagram relating to the isothermal change of state of gases

In reality the expansion takes always place against a pressure lower than the equilibrium pressure, consequently the work done by the gas is also smaller than during a reversible process. It is generally true that during an isothermal change of state the work done by a system (and also the heat taken by it) is maximal if the system gets from one state in the other in a reversible way.

The work done and the heat exchanged during isothermal compression may be calculated in exactly the same manner. During expansion the work is done by the gas which takes up heat accordingly, now the work is done on the gas which releases heat. The results differ from the previous ones only in their signs. The first process is accompanied by the increase of entropy, the latter one by its decrease, respectively. It is generally true that the work done on a system (and the heat released by it) is minimal if the process is reversible.

2. In what follows the isothermal expansion of gases will be given as an example for demonstrating that *the statistical and phenomenological definitions of entropy are equivalent, in spite of their considerable formal differences.*

According to [5.23], i.e. the phenomenological definition and considering also [5.24d]:

$$S_B - S_A = -\nu R \ln \frac{V_B}{V_A}$$

and by further transformation

$$S_B - S_A = Nk \ln \frac{V_B}{V_A} = k \ln \left(\frac{V_B}{V_A}\right)^N$$

where $N = \nu L = \nu R/k$ is the number of gas molecules. On the other hand, from the statistical interpretation and [5.22b] we have

$$S_B - S_A = -k \ln \frac{w_B}{w_A}$$

It must be proved that the relations derived from the two types of interpretation can be transformed into each other. This task is equivalent to proving the equality

$$\frac{w_B}{w_A} = \left(\frac{V_B}{V_A}\right)^N$$

For this purpose let us investigate more closely the meaning of the quotient w_B/w_A; this tells us how many times the probability of uniform occupation of the volume V_B is greater than the probability that only the smaller volume V_A is filled. It is readily seen that this quotient is equal to the power expression on the right-hand side. Since the equation contains quotients, in the calculations mathematical probabilities may be used instead of thermodynamic probabilities. The mathematical probability of finding N molecules in the total volume V_B is equal to 1, i.e. certainty. On the other hand, the probability of finding N molecules in the volume V_A smaller than V_B is V_A/V_B for the case $N = 1$, $(V_A/V_B)^2$ for $N = 2$, and $(V_A/V_B)^N$ for the case $N = N$. Thus, the above equivalence is proved.

Our example also shows that, in the statistical interpretation of entropy, calculations cannot be made simply with the thermodynamic probability; instead, its logarithm is required. In other words, in place of [5.22a] one cannot simply write a proportionality between the entropy and the thermodynamic probability, though this would also express the basic requirement that the two quantities should change in the same sense. The statistical interpretation leads to a result identical with the phenomenological definition only if instead of the thermodynamic probability its logarithm is taken into account. The essence of the logarithmic relation becomes even clearer in the following consideration.

Let us take any thermodynamic system in an arbitrary macrostate. Let the thermodynamic probability of the macrostate be w, and let S denote the entropy of the system. Let us divide in imagination our system into two parts, and denote the thermodynamic probabilities and entropies of the subsystems by w_1 and w_2 and by S_1 and S_2, respectively. The following facts must be considered:

(a) Entropy is an extensive quantity, i.e. the entropy of the total system is equal to the sum of the entropies of the subsystems:

$$S = S_1 + S_2$$

(b) From probability theory the microstates associated with all the macrostates of the total system can be produced by combining every microstate in one subsystem in every possible way with the corresponding microstates in the other subsystem. If the number of microstates in one subsystem is w_1 and that in the other subsystem is w_2, the number of microstates in the total system is given by the equation

$$w = w_1 w_2$$

Consequently, the relation between the entropy and the thermodynamic probabilities must be of a form which satisfies (a) with the use of (b). It is immediately seen that a logarithmic relation corresponding to [5.22a] satisfies this requirement.

3. In the next example the changes of entropy occurring in the case of *isothermal diffusion (isothermal mixing)* will be dealt with.

Consider a case in which, at the beginning of the process, the lower half of a vessel is filled with an ideal solution with a pure solvent above it (Fig. 5.11). In the initial state the solute has the volume V_A, while in the final state it is evenly distributed in the total volume V_B. Let us denote the initial concentration by c_A, the final concentration by c_B. The system can get from its initial state to its final state by several ways. How can the above dilution process be conceived quasi-statically? It will be shown that the diffusion may also proceed reversibly, namely by a *quasi-static osmotic process.*

For this purpose imagine the pure solvent to be separated from the solution by means of a piston which is permeable only for the pure solvent. The osmotic pressure may be compensated if an appropriate weight is placed on the balance. Subsequently, we begin to decrease the force acting on the balance in infinitesimally small steps. During this process the piston gradually rises while the solution becomes continuously more dilute. The pressure acting on the balance at every moment is in quasi-equilibrium with the osmotic pressure. The work done by the system during the dilution (W_{AB}) can be calculated similarly as in the isothermal expansion of gases. From the analogy between the universal gas law and van't Hoff's law, the final result is similar to [5.24*b*], but in this case v denotes the amount of solute. Consequently

$$W_{AB} = -vRT \ln \frac{V_B}{V_A}. \qquad [5.25a]$$

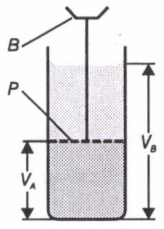


Fig. 5.11. Diagram relating to reversible diffusion
P: piston made of semipermeable material; *B*: balance

and since $c_A = \nu/V_A$ and $c_B = \nu/V_B$, instead of [5.24a] we may write

$$W_{AB} = -\nu RT \ln \frac{c_A}{c_B}. \qquad [5.25b]$$

The dilution accompanied by work is associated with heat exchange:

$$Q_{AB} = -\nu RT \ln \frac{V_B}{V_A} = -\nu RT \ln \frac{c_A}{c_B}. \qquad [5.25c]$$

Hence the increase of entropy:

$$S_B - S_A = -\nu R \ln \frac{V_B}{V_A} = -\nu R \ln \frac{c_A}{c_B} \qquad [5.25d]$$

In the same way, also the reversibility of a process opposite to diffusion or dilution, i.e. concentration may be conceived and the accompanying decrease of entropy may be determined.

4. In the following a chemical reaction will be examined (Fig. 5.12). What is the extent of the change of entropy when water is formed from oxyhydrogen? Let both the initial and the final temperatures be room temperatures. Considering that entropy is a state function, with respect to its change it is indifferent whether the process takes place as an explosion or by means of a catalyst quasi-statically. For the determination of the change of entropy the latter case has to be examined.

Let us start from equilibrium in the reaction space. Equilibrium pressures are denoted by p_{H_2}, p_{O_2} and p_{H_2O}. During the reaction these must remain constant, only infinitesimally small changes can be allowed. Thus hydrogen and oxygen can be introduced into the reaction space only at a pressure infinitesimally higher than p_{H_2} and p_{O_2}, while exhaustion of water vapour from the reaction space takes place at a pressure slightly lower than p_{H_2O}. This requirement can be fulfilled by a rather slow driving of the piston.

The presence of an appropriate *catalyst* is indispensable for the reversible conduction of the process (in our case, e.g., a platinum sponge). This ensures that the reaction takes place immediately and this way produces the conditions in the reaction space for a constant equilibrium. The dashed lines at the mouth of the cylinders denote walls permeable only to the component or product in the cylinder. The heat conducting wall of the reaction vessel ensures quick heat exchange with the environment and consequently the isothermal course of the process.

If the initial pressure of hydrogen and oxygen in the pistons is not equal to the corresponding equilibrium pressures, first the pistons have to be manipulated by closing the mouth of the cylinders to produce this pressure in both cylinders. The reversible conduction of this process (see the previous example 1), as well as bringing the water vapour reversibly to a final pressure differing from the equilibrium pressure are easily conceivable.

The reversible process in the opposite direction can be conducted similarly. Considering all this, obviously every state can be restored not leaving any changes either in the system or in its environment.

The example illustrates well the volumetric works exerted and also the heat exchanged during the process. In real processes instead of the whole volumetric work also heat appears.

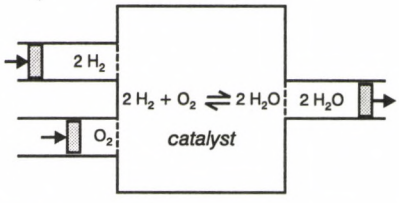


Fig. 5.12. Diagram relating to reversible formation of water from hydrogen and oxygen

3. Standard entropy. In practice, mainly the change of entropy is involved, though determination of *the absolute value of the entropy* is also possible. (This possibility does not generally exist with the internal energy or enthalpy; cf. section 5.3.3.) From considerations not discussed here, *the entropy of every chemically uniform substance is zero at absolute zero temperature.* Starting from this, with the aid of the specific heats and the heats of phase transitions the absolute values of the entropy of individual substances can be determined for any temperature. The entropy values associated with 25 °C and 101 kPa pressure are usually given for 1 mol amount of substance. This value of the entropy is called the *standard entropy* (Table 5.4).

Table 5.4. Standard entropies (S°) of some substances

Element or compound	State	$S°$ (J/mol × K)
H_2	g	130.6
O_2	g	205.2
C (graphite)	s	5.9
H_2O	l	69.9
H_2O	g	188.8
CO_2	g	214.0
Acetic acid	l	159.9
Lactic acid	l	192.2
Ethyl alcohol	l	160.8
Glycerine	l	208.1
Glucose	s	212.3

g = gas or vapour, l = liquid; s = solid

5.3.5. Direction and equilibrium of isolated and adiabatic processes. Life processes and the second law of thermodynamics

In connection with the statistical interpretation of entropy, it is easy to see (cf. section 5.3.3) that in an isolated system real processes proceed in the direction of increase in the entropy until it reaches the maximum possible value for the given system. However, experience reveals that this also holds under conditions where the system is not strictly isolated. The finding is also true for adiabatic processes, since the system must be isolated from its surroundings as concerns heat exchange, but not necessarily as concerns the work done. The processes in adiabatic systems can be characterized briefly by the formula

$$dS \geq 0 \qquad [5.26]$$

where the inequality sign relates to adiabatic processes, and the equality sign to thermodynamic equilibrium.

The validity of [5.26] can be demonstrated by a simple example. Consider an adiabatically closed vessel whose volume is divided into two parts by a wall made of some heat-conducting material. One part of the volume is filled with a gas at temperature T_1 and the other part with the same gas at temperature T_2. Let us now follow the initial process of temperature equalization. It will be shown that this process is characterized by [5.26].

Since the system is adiabatic, the algebraic sum of the heat exchange between the two volume parts is zero. i.e.

$$dQ_1 + dQ_2 = 0, \quad \text{or} \quad dQ_2 = -dQ_1$$

The entropy change (dS) in the course of this process is the sum of the entropy changes (dS_1 and dS_2) of the component systems, i.e.

$$dS = dS_1 + dS_2 = \frac{dQ_1}{T_1} + \frac{dQ_2}{T_2} = dQ_1 \left(\frac{1}{T_1} - \frac{1}{T_2} \right)$$

The following conclusions can be drawn from the equation:

(a) if $T_1 > T_2$, $dQ_1 < 0$, and consequently $dS > 0$;
(b) if $T_1 < T_2$, $dQ_1 > 0$, and consequently $dS > 0$;
(c) if $T_1 = T_2$, equilibrium exists and $dS = 0$.

The conclusions thus prove [5.26].

Of course, in non-adiabatic systems processes may also take place in which the entropy decreases. For instance, the entropy of liquids freezing at constant temperature decreases. [5.26] holds only if, besides the solidifying liquid, the bodies taking the heat released on freezing are also taken into consideration. In the isolated (or only adiabatic) systems formed in this way, processes involving both entropy increases and entropy decreases occur, and only the entropy of the overall system increases.

From the viewpoint of the validity of the above statements it is worthwhile to study the life processes more closely.

The living organism, such as the human being, is an open system exchanging substances and energy with its environment. Therefore in the study our environment (i.e. the Earth) also has to be taken into account since the condition of a closed system can be satisfied only in this way. In some cases even this is not enough and the environment of the Earth too has to be included in our considerations. In the course of life, processes accompanied by both entropy increase and entropy decrease take place. In the formation of macromolecular systems e.g. entropy decreases, since in the components the atoms are dispersed in a less ordered way compared to the more concentrated and ordered structure of macromolecules. An example of this is the synthesis of proteins from amino acids by the organism. – Contrary to this, solution formation, evaporation, etc. are accompanied by entropy increase.

The living organism is in a low-entropy state just because of its ordered structure and it maintains this state. The problem is how to make this consistent with the second law of thermodynamics.

In the course of metabolism the organism takes and releases substances. This means both entropy increase and decrease. The ingested substances – see e.g. the carbohydrate or protein content of foodstuffs – are of low entropy, but much entropy leaves with the released heat, the secreted and discharged waste materials, e.g. urine, the evaporated water, the expired carbon dioxide, etc. Summing up all this, even by purely qualitative consideration it is acceptable that the living organism produces entropy in the course of metabolism and releases the "excess" to its environment, thus ensuring the low entropy state inherent in life. The second law is obviously valid since meanwhile the entropy of the whole system consisting of the organism and its environment increases.

This last statement involves a further problem: the entropy on Earth should increase constantly; what happens with this accumulating entropy? The essence of the answer: the Earth releases the "excess" entropy to the cosmos. The process can be understood by the study of water circulation.

Water evaporates from the Earth, vapour gradually rises, cools, and gets to the upper atmosphere, where via radiation it cools further, condenses, and the precipitation returns to Earth. Evaporation on the Earth takes place at $T_1 \sim 300$ K maintained by solar radiation, while radiative heat loss in the atmosphere at $T_2 \sim 250$ K. Let Q denote the thermal energy taken at evaporation of a given quantity of water and later on released by radiative heat loss during a given time. According to the phenomenological interpretation of entropy the taken entropy is Q/T_1, the released one Q/T_2. Considering that $T_2 < T_1$, entropy leaves the Earth, and this is the answer to the above question.

This points out from a further aspect the fundamental importance of solar radiation (cf. section 2.7) and water (cf. section 1.5.1) in the formation and maintenance of life. Let us also notice that these factors can ensure life only in the present astronomical, geophysical, physical-chemical, etc. conditions (allowing only slight deviations). For example:

– if the size of Earth and consequently its gravitation were smaller, it could retain less water vapours; on the other hand, with greater mass, only little vapour would reach the air layers of sufficiently low temperature;

– solar physical and astronomical data, e.g. the size and surface temperature of the Sun, the Sun–Earth distance are also essential, since their change would produce basically different conditions for water on the surface of Earth;

– the role of water's properties (cf. section 1.5.1): high specific and evaporation heat, good solvent, etc., is again coming to the fore.

5.3.6. Direction and equilibrium of isothermal processes. Helmholtz and Gibbs free energy

In practice, situations are fairly frequently encountered where the temperature of the system is the same at the beginning as at the end of a process. Such processes may be regarded as isothermal from the viewpoint of their direction as well as their equilibrium. The chemical (biochemical) reactions in laboratory experiments or in life processes are of this type. Isothermal processes can also be studied on the basis of the entropy theorem as discussed in the previous section, if every part of the environment is regarded as belonging to the system, which therefore becomes isolated or at least adiabatic. This type of investigation is sometimes rather cumbersome and even unnecessary. In order to study isothermal processes, the use of the internal energy (enthalpy) and the entropy allows the introduction of state functions which give the direction and equilibrium of the processes directly. The state function which can be used in processes at constant volume (isochoric processes) is called the free energy, and the concept of Gibbs free energy is used to describe isobaric processes at constant pressure. The free energy function is also termed the *Helmholtz* function and the *Gibbs* free energy is sometimes called free enthalpy.

1. Free energy. Direction and equilibrium of isothermal and isochoric processes. The free energy is defined by the state function

$$F = U - TS \qquad [5.27a]$$

In isothermal processes, where T is constant

$$dF = dU - TdS \qquad [5.27b]$$

From experience it may be said that *isothermal processes proceed spontaneously in the direction of a decrease in the free energy. The end of the change of state, i.e. the equilibrium state, is characterized by the minimum free energy attainable in the given situation.* The decrease means a negative change, and the minimum attained refers to a zero change. These statements can be briefly expressed by the relation

$$dF \leq 0 \qquad [5.28a]$$

dF has a simple physical meaning. To illustrate this, let us write the first law for a reversible process

$$dU = dQ_{rev} + dW_{rev} \qquad [5.28b]$$

Since the volume is constant in our case no volumetric work is done and dW_{rev} may denote some other, e.g. electric work. However, by definition $dQ_{rev} = TdS$, and thus [5.28b] can be rewritten to yield

$$dW_{rev} = dU - TdS \qquad [5.28c]$$

From a comparison of [5.27b] and [5.28c] we have

$$dF = dW_{rev} \qquad [5.28d]$$

which means that the change in the free energy in the case of isothermal–isochoric processes is equal to the work done in a reversible process, so that condition [5.28a] can be put into the form

$$dW_{rev} \leq 0 \qquad [5.28e]$$

The work is less than zero if it is done by the system. *According to* [5.28e] *the isothermal–isochoric processes proceed spontaneously in the direction in which work is done by the system, or more exactly in the direction in which the system can do work.* This correction is necessary, since in reality the process may proceed without doing any work, only heat exchange occurs. However, in the determination of the direction of the spontaneous process information must be obtained about the possibilities. The maximum work which can be done by a system is that performed reversibly as expressed in [5.28e]. The system attains thermodynamic equilibrium, even if in principle no more work can be gained.

The correctness of [5.28a] can be proved by a simple example. Let us consider a perfect gas at temperature T enclosed in a vessel of volume $V = V_1 + V_2$, where the pressure in V_1 is p_1 and that in V_2 is p_2. The two volumes in the closed vessel are separated by a piston moving without friction, which allows a spontaneous pressure

equalization. The walls of the vessel are made of some good heat-conducting material, so that the pressure is equalized at a constant temperature. We shall prove that in an isothermal–isochoric process the free energy of the total gas quantity decreases in accordance with [5.28], and the equilibrium is characterized by the minimum free energy. Let us apply the first law of thermodynamics to an elementary pressure equalization for both volume parts. In the case of a reversible process

$$dU_1 = TdS_1 - p_1\,dV_1 \quad \text{and} \quad dU_2 = TdS_2 - p_2\,dV_2$$

The change in free energy of the whole system, i.e. the quantity

$$dF = dU - TdS$$

is obtained by adding the two equations. This leads to the relation

$$dF = (p_2 - p_1)\,dV_1 \qquad [5.28f]$$

The addition is carried out by using the fact that the total gas volume is constant, so that

$$dV = dV_1 + dV_2 = 0 \quad \text{and} \quad dV_2 = -dV_1$$

Further, it is taken into account that the internal energy and the entropy of the total system are obtained as the sums of the internal energies or entropies of the parts of the system (extensive quantities), i.e. $dU = dU_1 + dU_2$ and $dS = dS_1 + dS_2$, respectively. From equation [5.28f] the following conclusions can be drawn:

(a) if $p_2 > p_1$, $dV_1 < 0$ and consequently $dF < 0$;
(b) if $p_2 < p_1$, $dV_1 > 0$ and consequently $dF < 0$;
(c) *if* $p_2 = p_1$, equilibrium exists, and $dF = 0$.

Thus, the conclusions fully prove [5.28a].

2. Gibbs free energy. Direction and equilibrium of isothermal–isobaric processes. The Gibbs free energy is defined by the state function

$$G = H - TS \qquad [5.29a]$$

which differs from the free energy only by substituting the internal energy by the enthalpy. The change in the Gibbs free energy in the case of isothermal processes can be described by

$$dG = dH - TdS \qquad [5.29b]$$

Isothermal and isobaric processes proceed spontaneously in the direction of a decrease in the Gibbs free energy. The equilibrium is characterized by the minimum value that can be reached in the given case. This can be written briefly as

$$dG \leq 0 \qquad [5.30a]$$

In isothermal–isobaric processes, the change in the Gibbs free energy is equal to the reversibly performed work:

$$dG = dW_{\text{rev}} \qquad [5.30b]$$

Thus, instead of [5.30a] we may write

$$dW_{rev} \leq 0 \qquad [5.30c]$$

which means that the isothermal–isobaric processes proceed spontaneously in a direction in which the system can do work along a reversible path.

The entropy, free energy, Gibbs free energy and chemical potential (to be introduced later; cf. section 5.4.1) in certain respects play the same role in Nature as the potential energy in pure mechanics or the electric potential in the field of electrical phenomena. The equilibrium state is characterized by a potential energy minimum, while the characteristics of the thermodynamic equilibrium are

– the entropy maximum in the adiabatic systems;
– the free energy minimum for isothermal–isochoric processes;
– the Gibbs free energy minimum for isothermal–isobaric processes.

We are thus justified in calling the above state functions *thermodynamic potentials.*

3. Determination of thermodynamic potentials. In this section only the Gibbs free energy will be dealt with, for in the applications mainly processes proceeding at constant pressure are encountered. However, the results also apply to the free energy. In the most frequent applications, in chemical reactions, the difference between the free energy and the Gibbs free energy cannot be neglected, especially for processes accompanied by gas formation or consumption. In these cases, it is advisable always to perform calculations with the Gibbs free energy, which otherwise practically agrees with the free energy.

The change in the Gibbs free energy can be *measured* directly only in cases, when a reversible change of state can be achieved to a good approximation and the work done in this process is measurable. It generally holds that, in connection with chemical reactions, the change in the Gibbs free energy can be measured only in reactions functioning as reversibly working galvanic cells. *Calculation* of the Gibbs free energy, however, is always possible if the enthalpy and entropy are known.

Table 5.5. Standard enthalpies ($H°$), standard entropies ($S°$) and standard Gibbs free energies ($G°$) of some substances

Element or compound	State	$H°$ (kJ/mol)	$S°$ (J/mol × K)	$G°$ (kJ/mol)
H_2	g	0.0	130.6	0.0
O_2	g	0.0	205.2	0.0
C (graphite)	s	0.0	5.9	0.0
H_2O	l	– 286.0	69.9	– 237.4
H_2O	g	– 242.0	188.8	– 228.6
CO_2	g	– 394.0	214.0	– 394.8
Acetic acid	l	– 487.4	159.9	– 392.7
Lactic acid	l	– 677.0	192.2	– 520.4
Ethyl alcohol	l	– 278.0	160.8	– 175.0
Glycerine	l	– 666.6	208.1	– 475.6
Glucose	s	– 1280.1	212.4	– 915.7

g = gas or vapour; l = liquid; s = solid

Let us calculate, e.g., by using the data in columns 3 and 4 of Table 5.5 and equation [5.29b], the change in the Gibbs free energy in the production of 1 mol water from 1 mol hydrogen and 0.5 mol oxygen at atmospheric pressure and 25 °C. In this case ΔH = 286.0 kJ, ΔS = 69.9 – 130.6 – 102.6 = – 163.3 J/K, and $T\Delta S$ = – 48.6 kJ. Consequently, ΔG = – 237.4 kJ.[4] This value is given in column 5 of the Table, which also contains the *Gibbs free energy of formation* of other substances *related to 1 mol at 25 °C and 101 kPa.* These data are called *standard Gibbs free energies.* The Table shows that the Gibbs free energy is standardized similarly to the enthalpy, due to the fact that the absolute value of the Gibbs free energy is generally not known either, and only its changes can be calculated.

4. Bound energy. [5.29b] may also be written as

$$dH = dG + TdS$$

i.e. the change in the enthalpy is composed of two terms. The first gives the maximum work to be obtained from the enthalpy, and the second term provides information on the remainder, which cannot be used as work, but represents the *inevitable heat.* This latter term is called the *bound energy.* According to Table 5.5, of the enthalpy released on the formation of 1 mol water from its elements a maximum of 237.4 kJ can be used as work, and at least 48.6 kJ is dissipated as heat.

The statements concerning the direction of isothermal processes can be illustrated with the aid of the free energy or Gibbs free energy in the following way. Let us consider again the relation

$$dG = dH - TdS$$

and investigate separately the roles of the terms on the right-hand side. Two extreme situations may be observed. The first involves processes in which the entropy does not change, while in the second case the enthalpy (internal energy) remains constant. The processes proceed spontaneously in the direction of enthalpy decrease in the first case, and in the direction of entropy increase in the second case. A decrease in enthalpy (internal energy) means energy released, which occurs if the attractive forces between the atoms or molecules become stronger in the process, which finally results in a more compact arrangement. An increase in entropy, on the other hand, is related to a decrease in the bonding between the particles, i.e. to structural loosening. Hence, two opposite effects exist, which together determine the direction of the process. For instance, the attractive forces predominate in condensation or in the synthesis of molecules, and the direction of the process is therefore determined by the decrease in enthalpy (internal energy). Conversely, scattering tendencies become dominant in evaporation, mixing, dissociation of molecules, etc., and in these cases the entropy increase will be the most important factor. It is generally true that the direction of the processes at sufficiently low temperature and high pressure is determined by the decrease in enthalpy (internal energy), whereas if the temperature is raised the entropy increase becomes increasingly more dominant.

[4] The maximum work to be obtained in the formation of 1 mol water from 1 mol hydrogen and 0.5 mol oxygen can also be obtained from electric data. Platinum electrodes surrounded by gaseous hydrogen and oxygen, respectively, are immersed in weakly acidified water. The pressure of the gas surrounding the electrodes is kept constant (e.g. at 101 kPa). The resulting system is a galvanic cell, whose electromotive force at 25 °C and 101 kPa is 1.23 V. The electric work is gained by the formation of water in the cell. By Faraday's law, an electric charge of 193,000 C passes through the system on the formation of 1 mol water. The maximum obtainable work is given by the product of the charge and the e.m.f. The result, as before, is 237.4 kJ.

5.4. Additions and applications

5.4.1. Gibbs free energy of mixtures. Chemical potential

The problems arising in practice are usually not associated with systems consisting of a single pure substance, but with mixtures (gas mixtures, solutions). The extra- and intracellular spaces are filled with mixtures, and metabolic processes also proceed in mixtures. Though the direction of the processes is unambiguously defined by the results derived in the previous section, their application still requires some consideration.

1. Ideal liquid mixtures. For simplicity, let us assume that the mixing occurs at 25 °C ($T \approx 298$ K) and a pressure of 101 kPa. The number of components is denoted by n and the amount of the i-th component (in mol) by ν_i. *The Gibbs free energy of the mixture is the sum of the Gibbs free energies of the components:*

$$G = \sum_1^n \nu_i \mu_i \qquad [5.31]$$

where μ_i denotes the Gibbs free energy of the i-th component of the system examined relating to 1 mol of the component. Thus, μ_i is *partial molar Gibbs free energy,* also called the *chemical potential* of that component.

The task is to define μ_i. If the components were not in mixtures, but in their pure form (or in saturated solution), their Gibbs free energies relating to 1 mol at 25 °C and 101 kPa could be obtained from the tables. The same should be valid, if the components were present in a concentration of 1 mol/l, for these data could also be found in tables. For instance the chemical potential of a glucose solution of unit molarity is –904 kJ/mol, that of lactic acid –530.1 kJ/mol. The value of μ_i in our case differs from the data of tables, because the concentrations of the components in the mixture are generally different from those which the tables refer to. Let μ_i^0 denote, in the case of the i-th component, the chemical potential of the solution of unit molarity. For an ideal solution by using [5.24b] in the proper sense:

$$\mu_i - \mu_i^0 = RT \ln \frac{c_i}{c_i^0} \qquad [5.32a]$$

or

$$\mu_i = \mu_i^0 + RT \ln \frac{c_i}{c_i^0} \qquad [5.32b]$$

where $c_i^0 = 1$ mol/l, while c_i is the concentration of the i-th component in the solution.

μ_i^0 is referred to as *chemical normal potential* and its value at 25 °C is the *chemical standard potential.* The expression in [5.32b] with the concentrations is called *mixing term.* When formulating [5.32b] the concentration of the components was expressed in molarity (mol/l), but mole fractions could also have been used. The value of μ_i is independent of the concentrations used to characterize the composition of the mixture, but this does not hold separately for either the mixing term or for the chemical normal and standard potentials. The difference in the mixing term is a reasonable consequence of the fact that at a given composition the numerical values of the different concentrations are not equal.

Further, the normalization and standardization refer in one case to a solution of unit molarity – as in [5.32b] too – and in the other to the pure state of the component, from which it follows that the normal and standard values are also different.

An additional remark: instead of [5.32b] one encounters very often the following form:

$$\mu_i = \mu_i^0 + RT \ln c_i \qquad [5.32c]$$

i.e. c_i^0 is omitted from the formula considering that its value is 1. From a dimensional point of view, of course, this form is not correct.

2. Real mixtures. Activity. Van't Hoff's law is only more or less valid for real mixtures, consequently these formulas of chemical potential hold only with better or worse approximation. However, this can be improved if the concentration is replaced by a new quantity which depends on the concentration so that the respective relations of ideal mixtures remain valid in their original form for real mixtures too. This quantity is the *activity* (a_i) in the case of liquid mixtures. The activity values can be determined and their relations with the concentrations are tabulated. Thus, the chemical potential of the i-th component for real mixtures, on the basis of [5.32c], can be written in the following form:

$$\mu_i = \mu_i^0 + RT \ln a_i \qquad [5.32d]$$

The chemical standard potentials are related to unit activity.

The findings for the Gibbs free energy also hold for the chemical potentials and in the case of mixtures they give information on the direction and the equilibrium of the changes in the individual components. If the chemical potential of one component is different at different points of the mixture, the respective component will migrate from a site of higher concentration to one of lower concentration. The overall mixture will be in equilibrium only if the chemical potential of each component is identical throughout the mixture.

From a study of the Gibbs free energy of mixtures, several practical results are obtained. For instance, the equilibrium constant of a chemical reaction or the equilibrium change due to a change in temperature or pressure can be calculated, and the chemical affinity too can be quantitatively characterized. Thermodynamic considerations lead to the derivation of a relation which permits a comparison of solution concentrations via measurement of the e.m.f. of a concentration cell. Similar relations can be derived between the potentials of redox systems and the concentrations of reduced and oxidized compounds. Some of the results will be dealt with in more detail below.

As concerns the metabolic reactions of the life processes, the change in the Gibbs free energy is itself of interest, because it gives information about the maximum mechanical, electric or other work which can be done at the cost of the energy released in the reactions, or which may be used to initiate other energy-requiring chemical processes. This is an important point, since the energy of the metabolic processes in the cells is usually expended initially for producing compounds (e.g. ATP) which have relatively large Gibbs free energies. The decomposition of these compounds at appropriate sites releases the energy necessary for the work required.

Let us calculate the Gibbs free energy change due to the decomposition of 1 mol glucose to 2 mol lactic acid at 25 °C and 101 kPa. The calculations are carried out for the case when the concentration of the mixture is 0.01 mol/l for glucose and 0.002 mol/l for lactic acid. These are the average concentrations in the living organism. The tabulated value of the chemical potential (standard potential) of a glucose solution of unit molarity is –904 kJ/mol, and that of the lactic acid solution –530.1 kJ/mol. These values are changed by the mixing terms, which in the given case, according to [5.32b] are –11.3 kJ/mol for glucose and –15.5 kJ/mol for lactic acid. Thus, the chemical potentials are –915.3 kJ/mol and –545.6 kJ/mol, respectively. Since 2 mol lactic acid is produced from 1 mol glucose, the Gibbs free energy change is –175.9 kJ. The maximum work that can be obtained by the conversion of 1 mol glucose into lactic acid will therefore be 175.9 kJ. Lactic acid dissociates at 25.1 nmol/l H^+ concentration (pH = 7.6) of the organism into hydrogen and lactate ions. When this process too is considered, the maximum work that can be obtained theoretically will be more: approximately 209.4 kJ. The organism uses this conversion in muscular activity, for instance. Measurements show that approximately 117.2 kJ work is gained, which corresponds to an efficiency of more than 50%. In reality the situation is even better, since the reaction takes place on the surface of the enzymes, where the concentrations differ from the data used above, i.e. from the mean concentration values. Near the surface of the enzymes the glucose concentration is obviously lower and the lactic acid concentration is higher than the mean value, since the glucose consumed cannot be replaced immediately and the lactic acid produced cannot move away at once from the site of its production. A further circumstance may be considered in connection with the efficiency. The work is actually measured on a group of several fibres and the individual fibres do not contract simultaneously. As a result, the measurements definitely yield values smaller than the actual ones.

It is worthwhile emphasizing in this context that the human organism is a system which covers from the Gibbs free energy all the energy necessary for its functioning and activity not required as heat. With the aid of the enzymatic system, the processes occur nearly reversibly in the living organism. Consequently, from the aspect of the work performed, the organism is a nearly ideal thermodynamic system.

5.4.2. The quantitative description of chemical affinity

Let us consider a reversible chemical reaction at constant temperature and pressure. Reversible reactions are generally described by the stoichiometric equation

$$r_A A + r_B B + \ldots \Leftrightarrow r_K K + r_L L + \ldots \qquad [5.33a]$$

A, B, ... refer to the *initial substances,* and *K, L,* ... to the *products.* The quantities r_A, r_B, ... r_K, r_L ... denote the *stoichiometric coefficients* in the studied processes. Instead of the above notation frequently the more concise symbolism

$$\Sigma r_A A \Leftrightarrow \Sigma r_Z Z \qquad [5.33b]$$

is used, where A stands for the initial substances and Z for the products.

In the course of the reaction the quantity of some substances decreases whereas that of others increases. These changes take place in ratios determined by the stoichiometric coefficients, which enables to describe the degree of the progress in any chemical reaction by a single quantity briefly called *reaction coordinate* (ξ). In order to determine the value of ξ one starts from the fact that while at the beginning of the process the product quantity is zero, it will be ν_K, ν_L, ... after some time.

Since

$$\nu_K : \nu_L : \ldots = r_K : r_L : \ldots \qquad [5.34a]$$

one may write

$$\nu_K = r_K \xi, \quad \nu_L = r_L \xi \qquad [5.34b]$$

and for further changes one has

$$d\nu_K = r_K d\xi, \quad d\nu_L = r_L d\xi, \ldots \qquad [5.34c]$$

ξ is a positive quantity which increases as the reaction advances. By similar reasoning the changes in the quantity of the initial substances are described by the equations

$$-dv_A = r_A d\xi \quad \text{and} \quad -dv_B = r_B d\xi \ldots \qquad [5.34d]$$

The negative sign indicates the decrease of the quantity of the initial substances.

It is suitable to characterize the affinity among the substances participating in isothermal–isobaric reactions with a quantity which is positive in the direction of the spontaneous process and becomes the larger the further is the system from equilibrium, while in equilibrium this quantity is zero. Such quantity may be deduced from the change of the Gibbs free energy.

Recalling the reaction described by [5.33b] let us assume that it proceeds from left to right. The Gibbs free energy change of the total system will be

$$dG = \mu_A dv_A + \mu_B dv_B + \ldots + \mu_K dv_K + \mu_L dv_L + \ldots \qquad [5.35a]$$

or in a more concise notation

$$dG = \Sigma \mu_A dv_A + \Sigma \mu_Z dv_Z \qquad [5.35b]$$

where μ_A, μ_B ... and μ_K, μ_L ... denote the chemical potentials of the initial substances, as well as of the products in a given state of the system, while dv_A, dv_B... and dv_K, dv_L ... indicate the changes in the quantity of the initial substances and the final products, respectively.

Equation [5.35b] may be rewritten in a better arranged form expressing the changes in the amount of substances with the reaction coordinates. By the use of relations [5.34c] and [5.34d] one has

$$dv_A = -r_A d\xi, \ \ldots \quad \text{and} \quad dv_Z = r_Z d\xi, \ \ldots$$

consequently

$$dG = (\Sigma r_Z \mu_Z + \Sigma r_A \mu_A)\, d\xi \qquad [5.35c]$$

Since for isothermal–isobaric processes

$$dG \leq 0 \qquad [5.36a]$$

(cf. section 5.3.6), taking into consideration [5.35c] we may write

$$(\Sigma r_Z \mu_Z - \Sigma r_A \mu_A)\, d\xi \leq 0$$

and

$$\Sigma r_Z \mu_Z - \Sigma r_A \mu_A \leq 0 \qquad [5.36b\text{–}c]$$

respectively, where the inequality refers to spontaneous processes, and the equality indicates the equilibrium state.

Equations [5.36*b–c*] express the fact that the chemical reactions proceed in the direction in which the sum of the chemical potentials of the products weighted by their stoichiometric coefficients is smaller than that of the reacting substance. Equilibrium will be at equality.

The *chemical affinity* is characterized by the quantity

$$A = -(\Sigma r_Z \mu_Z - \Sigma r_A \mu_A) \qquad [5.37a]$$

It clearly follows from the previous reasoning that A is positive and its value is the greater the further is the system from the equilibrium. In equilibrium $A = 0$, consequently the above definition of affinity satisfies the previously formulated conditions. On the basis of the above it may be also said that A is equal to the work which could be done by the system during reversible processes while from r_A moles of the initial substance r_Z moles of the product are formed.

In order to compare the affinity of the various reactions the affinity is normalized and/or standardized. The *normal affinity* ($\mathring{A}$) related to the temperature T is obtained if instead of the chemical potentials μ_Z and μ_A of [5.37] the chemical normal potentials μ_Z^0 and μ_A^0 are used. Thus:

$$\mathring{A} = -(\Sigma r_Z \mu_Z^0 - \Sigma r_A \mu_A^0) \qquad [5.37b]$$

The normal affinity at temperature 25 °C ($T \approx 298$ K) is called *standard affinity.*

5.4.3. The law of mass action. Equilibrium constant

The determination of thermodynamic equilibrium for reversible chemical reactions is one of the fundamental and most important questions from the theoretical viewpoint as well as from practical aspects. The thermodynamic equilibrium is properly defined by the law of mass action, which is the result of the general thermodynamic equilibrium conditions (cf. section 5.3.6).

As has been already briefly pointed out, the chemical reactions are described by the stoichiometric equation

$$\Sigma r_A A = \Sigma r_Z Z$$

The r_A factors indicate the stoichiometric coefficients of the substances denoted by A, and the factors r_Z refer to the stoichiometric coefficients of the substances Z.

The direction of the reaction is determined by the activities (concentrations) which the components possess in the mixture. The activities are denoted by the symbols a_A, a_B, ..., a_K, a_L, ... The activities associated with the equilibrium state are denoted by a bar ($\bar{a}_A$, etc.). The relation existing between the equilibrium activities is called the law of mass action:

$$\frac{(\bar{a}_K)^{r_K}(\bar{a}_L)^{r_L}\ldots}{(\bar{a}_A)^{r_A}(\bar{a}_B)^{r_B}\ldots} = \mathrm{K} \qquad [5.38a]$$

The quantity K, the *equilibrium constant,* is constant at given temperature and pressure and is characteristic of the given reaction. If the activity of one of the components of the mixture is changed, the others also change so that the value of K remains constant. The equilibrium constant can also be expressed in terms of chemical normal potentials of the individual components. With the usual notations, we have

$$\ln K = -\frac{\Sigma r_Z \mu_Z^0 - \Sigma r_A \mu_A^0}{RT} \qquad [5.38b]$$

Considering also [5.37b], let us recognize that between two main characteristics of a reaction, namely its equilibrium constant and affinity, a simple relationship exists which makes possible the calculation of one of them from the other, since

$$\mathring{A} = RT \ln K \qquad [5.38c]$$

The general derivation of the mass action law is omitted here. We shall restrict ourselves to a special case to illustrate the train of thought which leads to the law of mass action starting from the general conditions of thermodynamic equilibrium. Let us discuss, for example, the simple dissociation process:

$$C_{r_C}A_{r_A}\ (\text{crystalline salt}) \Leftrightarrow r_C C + r_A A\ (\text{dissolved salt})$$

In the present case r_C denotes the stoichiometric number of cations C, and r_A that of anions A. Equilibrium exists if the Gibbs molar free energy of the salt in the solution is equal to the Gibbs molar free energy of the crystalline salt:

$$r_C(\mu_C^0 + RT \ln \bar{a}_C) + r_A(\mu_A^0 + RT \ln \bar{a}_A) = \mu_{CA}^0$$

The left-hand side refers to the salt in the solution, and the right-hand side to the crystalline salt. The first term of the left-hand side is associated with the cations, and the second with the anions. μ_C^0 and μ_A^0 are the chemical normal potentials of the cations and anions in the solution while μ_{CA}^0 refers to the chemical normal potential of the crystalline salt. $\bar{a}_C$ and $\bar{a}_A$ are the *equilibrium* activities (concentrations) of the cations and the anions in the solution. After rearrangement, we have

$$r_C\mu_C^0 + r_A\mu_A^0 - \mu_{CA}^0 = -RT \ln (\bar{a}_C)^{r_C}(\bar{a}_A)^{r_A}$$

Since the left-hand side is constant, the right-hand side must of necessity be constant too, from which it follows that

$$(\bar{a}_C)^{r_C}(\bar{a}_A)^{r_A} = K_{\text{diss}}$$

where the constant is now called the *solubility product.*

It can immediately be seen that the result is the special case of the law of mass action for unit denominator. In the present case the denominator is the activity characteristic of the crystalline salt, which can indeed be taken as unity.

In connection with the above example, the relation between the equilibrium constant and the chemical normal potential is obtained directly:

$$\ln K_{\text{diss}} = \frac{r_C\mu_C^0 + r_A\mu_A^0 - \mu_{CA}^0}{RT}$$

Let us apply the above relation to the dissociation of water, i.e. to the process

$$H_2O \Leftrightarrow H^+ + OH^- \quad \text{or} \quad 2H_2O \Leftrightarrow H_3O^+ + OH^-$$

Calculation with the molar concentrations c_{H+} and c_{OH-} instead of the activities involves the relation

$$c_{H^+} \times c_{OH^-} = K_{\text{water}}$$

The value of K_{water} at 25 °C is 10^{-14} (mol²/l²). (For the temperature dependence of the equilibrium constant cf. Appendix A1.)

5.4.4. Electrode potentials. Nernst's equation

The body fluids of the living organism contain various ions. The thermodynamic phenomena connected with these ions are of vital importance in the life processes. In the following section (cf. also sections 5.5.2 and 5.5.3) we shall deal with some basic relations necessary to obtain a deeper insight into the properties of electrolytes and other solutions containing charged particles.

1. Electrode potentials. If a metal (the electrode metal) is immersed into a solution of its ions, an electric potential difference results between the metal and the electrolyte. Its production can be interpreted in the following way. When the metal is immersed in the electrolyte, depending upon the concentration (activity) positive metal ions either pass from the metal into the solution, or vice versa. In the former case the metal will become negatively charged, while in the latter case it will be positively charged with respect to the solution. The potential difference between the electrode and the solution will in both cases increase to a certain extent and within a fairly short time *dynamic equilibrium* will be established at the boundary surface between the metal and the solution. Here too ions move from one phase into the other, but in a given time the voltage produced will cause the same number of ions to move in the opposite direction. The potential difference associated with the equilibrium state is the *electrode potential.*

Similar processes take place at both electrodes of galvanic cells, and the electromotive force (e.m.f.) of the cell results from the *algebraic difference* of the two electrode potentials.

The electrode potential ε can be calculated from thermodynamic considerations (see below). The relation

$$\varepsilon = \varepsilon^0 + \frac{RT}{zF} \ln a \qquad [5.39a]$$

is obtained, where R is the universal gas constant, T is the temperature, F is Faraday's constant,[5] z is the valence of the electrode metal ions, and a denotes the activity of the metal ions in the solution. ε^0 is the electrode potential of a solution of unit activity; it is called the *normal electrode potential* at a pressure of 101 kPa. The normal potential at 25 °C is the *standard electrode potential* of the respective electrode.[6] By convention, the values ε and ε^0 are positive if the electrode metal has a positive potential with respect to the electrolyte, while in the opposite case ε and ε^0 are negative.

[5] F = 96,500 coulombs/gram-equivalent; it gives the charge of 1 gram-equivalent ions.

[6] Many authors use the expression electrochemical normal potential and electrochemical standard potential instead of normal electrode potential and standard electrode potential, respectively. The expression electrochemical potential is used in this book, too, but in another sense (cf. section 5.5.2).

Electrode potentials are measured in the following way. A galvanic cell is made, one electrode of which is the investigated metal together with the electrolyte in contact with the metal,[7] while the other electrode is the reference electrode, usually a standard hydrogen electrode (see below). The electrode potential of the studied electrode can be characterized by the e.m.f. of this cell. The normal potential of the standard hydrogen electrode used for comparison is taken arbitrarily as zero at the temperature of measurement. The tabulated data are usually the potentials of electrodes relative to the standard hydrogen electrode at 25 °C.

The thermodynamic considerations concern the case when the electrode metal dissolves on being immersed in the solution. Thermodynamic equilibrium between the metal and the solution is reached when the work associated with the transition of the ions is zero. In the present case this work consists of two parts: the Gibbs free energy change (W_1) accompanying the dissolution of the metal, and the electric work (W_2) due to the potential difference between the metal and the solution. Both components are related to 1 mol ions. With the usual notations, W_1 is given by

$$W_1 = \mu_{ion} - G_{metal} \quad \text{or} \quad W_1 = \mu^0_{ion} + RT \ln a_{ion} - G_{metal}$$

The electric work is

$$W_2 = zF\,(\varphi_{solution} - \varphi_{metal})$$

$(\varphi_{solution} - \varphi_{metal})$ is the potential difference between the solution and the metal. At equilibrium, $W_1 + W_2 = 0$; i.e.

$$\mu^0_{ion} + RT \ln a_{ion} - G_{metal} + zF\,(\varphi_{solution} - \varphi_{metal}) = 0$$

from which the electrode potential of interest is

$$\varphi_{metal} - \varphi_{solution} = \frac{\mu^0_{ion} - G_{metal}}{zF} + \frac{RT}{zF} \ln a_{ion}$$

or with a simplified notation

$$\varepsilon = \varepsilon^0 + \frac{RT}{zF} \ln a$$

Thus, [5.39a] has been proved.

2. Nernst's equation. Galvanic cells whose electrodes are identical, and in which only the concentrations (activities) of the electrolyte solutions around the two electrodes are different, are called *concentration cells.* The production of e.m.f. can be interpreted according to the scheme outlined in point 1, since [5.39a] shows that the electrode potential depends upon the concentration of the solution. Consequently, different electrode potentials are created on the two electrodes of the cell. The e.m.f. (disregarding the liquid potential) is given by their algebraic difference, which means that the e.m.f. of the concentration cell is

$$E = \varepsilon_1 - \varepsilon_2 = \frac{RT}{zF} \ln \frac{a_1}{a_2} \qquad [5.39b]$$

where the subscripts 1 and 2 are used to distinguish the two electrodes. [5.39a] or [5.39b] is called the *Nernst equation.*

[5.39b] provides an easy method of determining the ion concentration: if the e.m.f. is measured, the concentration of one solution can be calculated if that of the other is known. A particularly frequent task in practice is the determination of the hydrogen ion concentration of a solution. For this purpose hydrogen electrodes are used, which belong in the group of gas electrodes. A gas electrode too contains a metal, but it is always surrounded by the corresponding gas, and the metal is only the gas carrier. For the hydrogen electrode the carrier electrode is platinum and the whole system is immersed in a solution containing hydrogen ions. Two electrodes are always required to carry out the measurements: one is immersed in the solution to be studied, and the other in the solution used for comparison. In the case of hydrogen ions a molar solution (strictly speaking a solution of unit activity) is

[7] In the present case the name electrode may mean not only the metal conductor immersed in the electrolyte, but (as frequently used in electrochemistry) the *system* consisting of the respective element and the solution containing its ions.

used, and this system is then the *standard hydrogen electrode.* Since the hydrogen electrode is sensitive only to the hydrogen ion concentration, this procedure can be applied to solutions (e.g. blood) which contain many other components, so that any other method of determination would be complicated and less exact. With appropriate electrodes, the concentrations of other ions can be determined in a similar way.

5.4.5. Some remarks

As has been mentioned previously, the first law of thermodynamics given in the form of equation [5.14] may be thought of as a general law. This will be written in a case when *thermal, mechanical* and *chemical* interactions exist in isothermal and isobaric conditions of a chemical system. The changes of the internal energy resulting from these interactions may be written for the quasi-static case in the form

$$dQ = TdS; \quad dW_{\text{mech}} = -pdV; \quad dW_{\text{chem}} = \sum_1^n \mu_i dv_i \qquad [5.40a]$$

consequently

$$dU = TdS - pdV + \sum_1^n \mu_i dv_i \qquad [5.40b]$$

Note that each energy or work term on the right side may be constructed as the product of the intensive and extensive quantity (more exactly of the change of the extensive quantity) corresponding to the relevant interaction. *Intensive* quantities are the parameters the value of which is the same for the parts of the system as for the whole system in equilibrium, *extensive* are the parameters the value of which is in correlation with the extension of the system if the latter is homogeneous. Examples for the first are the temperature, pressure, electric potential, chemical potential, etc. Extensive quantities are e.g. the volume, mass, electric charge, internal energy, enthalpy, entropy, free energy. Accordingly the quantities related to these interactions, including also the electrostatic interactions, are collected in Table 5.6.

Table 5.6. Quantities characterizing energetic interactions

Interaction	Characteristic quantity		Work or energy
	extensive	intensive	
Mechanical	Volume (V)	Pressure ($-p$)	Volumetric work ($-pdV$)
Electrostatic	Electric charge (q)	Electric potential (φ)	Electric work (φdq)
Chemical	Amount of component (v_i)	Chemical potential of component (μ_i)	Work required for transport of molecules ($\mu_i\, dv_i$)
Thermal	Entropy (S)	Temperature (T)	Heat (TdS)

Thereupon the first law takes the simple form

$$dU = \sum_1^s y_i dx_i \qquad [5.40c]$$

where x_i and y_i are the extensive and the intensive quantities characterizing the i-th interaction, and s denotes the number of interactions.

5.5. Return to transport processes

In our discussion of thermodynamic processes and the determination of their direction, we have been engaged mainly in the study of equilibrium states. In the following treatment, applying the results of equilibrium thermodynamics, we return to non-equilibrium processes, such as diffusion, the flow of heat and electric charge, etc. All these phenomena will be reviewed from a new aspect.

5.5.1. Onsager's linear relations

Any isolated system is in equilibrium if the intensive quantities are the same at every point of the system. This statement is frequently referred to as the *zero-th law of thermodynamics.* If this condition is not fulfilled, *transport processes* (flows) take place leading to the equalization of the differences in the various intensive quantities. In all of these processes some extensive quantity flows, as for instance, the volume flow and the charge flow (electric current) leading to the equilibrium of pressure differences and electric potential differences, respectively, as well as the material flow resulting from the differences of the chemical potentials in chemical reactions as well as in dissolution and precipitation processes, etc.

The transport is characterized by the *flux (current density)* of the flowing quantity. The flux (J) is given by the extensive quantity passing through unit cross-section in unit time. The transport is due to the inhomogeneity in the spatial distribution of the intensive quantities, which is characterized by the *gradient* of these quantities. For simplicity we discuss only cases where a single quantity participates in the process, and its value changes only in one dimension, the x coordinate, *the positive X axis pointing in the direction of the transport.* Let us denote the change in the intensive quantity in question over the length dx by dy. The gradient is then defined as dy/dx. The gradient of the intensive quantities plays the same role in thermodynamics as the force in mechanics. For this reason *the gradient of the intensive quantities, or more exactly its negative value, is called the generalized or thermodynamic force (X).* So:

$$X = -\frac{dy}{dx}$$

A comparison of the equations describing the various transport processes (Ohm's law, Fick's first law, the Hagen–Poiseuille law, etc.) results in the following general statement: *the flux (J) is proportional to the corresponding thermodynamic force (X),* i.e.

$$J \sim X, \quad \text{or} \quad J = LX \qquad [5.41]$$

where the coefficient L is called the *phenomenological coefficient.*[8] The above statement will be referred to in the forthcoming discussions as *Onsager's linear law.* Its validity may

[8] [5.41] accounts for the use of the negative sign in the definition of thermodynamic force, since transport processes always proceed in the direction of the decrease of intensive quantities. This requires the use of the negative gradient as thermodynamic force.

Fig. 5.13. Diagram for the interpretation of diffusion
Notations associated with thermodynamic force (*a*) and with flux (*b*).
The X axis denotes the direction of both the thermodynamic force and the flux

be especially easily seen in the case of the Ohm's law: the current density is proportional to the negative electric potential gradient, the proportionality factor is the specific conductivity.

Let us study the *diffusion* more closely. In this case the thermodynamic force is the chemical potential gradient (or rather the negative value) of the solute, and we have

$$X = -\frac{d\mu}{dx} \qquad [5.42a]$$

where $d\mu$ is the chemical potential change along the length dx (Fig. 5.13*a*). In the present case the flux is

$$J = \frac{dv}{dt}\,\frac{1}{q} \qquad [5.42b]$$

where dv denotes the amount of substance transported across the surface q in time dt (Fig. 5.13*b*). From [5.41]

$$\frac{dv}{dt}\,\frac{1}{q} \sim -\frac{d\mu}{dx} \qquad [5.42c]$$

In the following we shall prove that [5.42c] corresponds to Fick's first law, and additional information will be obtained on the diffusion coefficient.

Let c denote the concentration of the solute, and dV the volume of solution containing the amount of solute dv. It follows from the concept of the concentration of substance that $dv = cdV$.

It is also taken into consideration that $dV = qds$, where ds denotes the distance moved by the diffusing molecules in time dt in the direction of diffusion. ds/dt gives the mean diffusion velocity of the molecules, which will be denoted by v. The flux can then be expressed as

$$J = cv \qquad [5.42d]$$

This relation can be transformed if the diffusion of the molecules is considered as frictional motion, and use is made of the empirical relation (cf. section 5.1.3) that the velocity of a moving particle is proportional to the driving force. The driving force in the present case is the thermodynamic force given in [5.42a], and hence

$$v = -u\,\frac{1}{N_A}\,\frac{d\mu}{dx} \qquad [5.42e]$$

where u denotes the mobility of the solute molecules. The division by the *Loschmidt constant* (Avogadro's constant, N_A) is explained by the requirement that in [5.42e] we calculate with the force acting on one single molecule, whereas $d\mu/dx$ represents the force acting on one mol. With the use of [5.42e], [5.42d] can be rewritten as

$$J = -uc \frac{1}{N_A} \frac{d\mu}{dx} \qquad [5.42f]$$

which is a more developed form of [5.42c], since it is an equality instead of a proportionality. In the next step we transform the gradient $d\mu/dx$ by making use of relation [5.32c] for the chemical potential:

$$\mu = \mu^0 + RT \ln c \qquad [5.42g]$$

where c is the concentration of the solute. From [5.42g] it follows by derivation (cf. Appendix, section B3) that (if T is constant)

$$\frac{d\mu}{dx} = \frac{RT}{c} \frac{dc}{dx} \qquad [5.42h]$$

Consequently

$$J = -ukT \frac{dc}{dx} \qquad [5.42i]$$

where $k = R/N_A$ is the Boltzmann constant.

It is clear that [5.42i] is identical with Fick's first law (cf. section 5.2.1), and the diffusion coefficient is given by the equation

$$D = ukT \qquad [5.43a]$$

For spherical particles we have from [5.4b] $u = 1/6\pi\eta r$, and thus

$$D = \frac{kT}{6\pi\eta r} \qquad [5.43b]$$

In this way, Einstein's relation defining the diffusion coefficient [5.12] has been proved.

A relation similar to [5.41] can be found for chemical processes. Instead of the flux, in these cases the calculations involve the rate of formation or depletion of one of the participating components, and the role of the thermodynamic force inducing the process is taken over by the affinity. Consequently, the following statement corresponds to [5.41]: *the rate of formation (or depletion) is proportional to the chemical affinity.*

5.5.2. Diffusion of electrolytes. Diffusion potential

1. Development of the potential. The merging of electrolytes with different concentrations (activities) may lead to the development of an electric potential gradient, the *diffusion potential* (or more exactly the diffusion voltage). This process may be visualized in the following way. Diffusion is always directed from higher to lower concentrations. If the

anion and cation mobilities are the same, they diffuse jointly (quite randomly) and no voltage develops. However, if their mobilities are different, the more mobile ions lead the way and the less mobile ions lag behind. As a result, some order develops in the distribution of the opposite charge carriers and for this reason a potential difference develops between the regions of different concentration. If the anion is the more mobile, the more dilute side will be at a negative potential relative to the more concentrated one, while with more mobile cations the reverse situation arises. The diffusion potential persists until the concentration difference becomes zero. (The diffusion potentials are in practice between 0.01 and 0.1 V.)

Calculations for binary solutions containing monovalent cations and anions in a concentration gradient dc/dx show, in agreement with experience, that the following relation holds for the potential gradient

$$\frac{d\varphi}{dx} = -\frac{RT}{F}\,\frac{u_C - u_A}{u_C + u_A}\,\frac{1}{c}\,\frac{dc}{dx} \qquad [5.44a]$$

where c is the concentration of the electrolyte in the volume in question, u is the mobility of ions (c.f. section 5.1.3), T is the temperature, R is the universal gas constant and F is Faraday's constant. The subscript C means the cation, and A the anion.

Integration of [5.44a] yields

$$\varphi_2 - \varphi_1 = -\frac{RT}{F}\,\frac{u_C - u_A}{u_C + u_A}\,\ln\frac{c_1}{c_2} \qquad [5.44b]$$

[5.44b] defines the potential difference that develops in the contact layer of solutions of concentrations c_1 and c_2.

It is clear from the above equations and also from the qualitative description that the larger the mobility difference between the anions and cations and the concentration ratio of the solutions in contact, the larger is the diffusion potential. If no concentration gradient exists or the mobilities are the same, the diffusion potential is zero.

Table 5.7. Mobilities of some ions (in relative units) in aqueous solution at infinite dilution at 25 °C

Ion	Mobility	Ion	Mobility	Ion	Mobility
H^+	349.8	Mg^{2+}	53.0	OH^-	198.6
Li^+	38.7	Ca^{2+}	59.5	F^-	55.4
Na^+	50.1	Sr^{2+}	59.4	Cl^-	76.4
K^+	73.5	Ba^{2+}	63.6	Br^-	78.1
Rb^+	77.8			I^-	76.8
Cs^+	77.2			$CH_3CO_2^-$	40.9
				SO_4^{2-}	80.0

The mobilities of the individual ions can be determined via their electric conductivities, since a linear relation exists between the two quantities. Table 5.7 lists a few data. The diameter of the alkali metal ions increases on proceeding from Li^+ to Cs^+; their mobilities change in the same sense. At first sight it would appear that this observation contradicts the idea that the diffusion may be compared to the motion of spherical particles in a viscous medium, since the mobility and the particle diameter are inversely proportional to each other in the case of fric-

tional motion. However, the apparent contradiction is explained by taking into consideration that the ions are enveloped in hydrate shells and, as concerns the mobility, not the diameter of the ion but that of the hydrate shell moving together with the ion is decisive. The diameter of the hydrate shell, however, decreases from Li^+ to Cs^+. The mobilities of the H^+ or H_3O^+ ion and the OH^- ion are considerably larger than would be expected from the diameters of their respective hydrate shells. This irregular behaviour is explained by proton exchange. In the case of H_3O^+ ion this means that the H_3O^+ transfers a H^+ ion to a neighbouring water molecule, which is equivalent to the propagation of the H_3O^+ ion. The OH^- ion, on the other hand, migrates by abstracting a proton from a neighbouring water molecule. It should be observed that the mobilities of the K^+ and Cl^- ions are practically equal. This means that for KCl solutions the diffusion potential can be taken as zero. For this reason, a concentrated KCl solution is used as liquid junction when a diffusion potential is to be avoided (cf. the measurement of the resting potential in section 7.1).

2. The way leading to the relation. Electrochemical potential. The relation describing the diffusion potential can be obtained in the following way. For the flux of any ion in an electrolyte a relation similar to [5.42f] holds

$$J = -uc \frac{1}{N_A} \frac{d\mu^e}{dx} \qquad [5.45a]$$

However, in the present case the driving force is not only the chemical potential gradient; the force due to the electric potential gradient must also be considered. Their resultant is called the *electrochemical potential gradient,* which is denoted in [5.45b] by $d\mu^e/dx$. Consequently

$$\frac{d\mu^e}{dx} = \frac{d\mu}{dx} + zF \frac{d\varphi}{dx} \qquad [5.45b]$$

Only the product zF in the second term on the right-hand side requires explanation. In [5.45b] the thermodynamic forces refer to 1 mol amount of substance. However, $d\varphi/dx$ by definition gives only the force acting on unit charge. The electric force on 1 mol of ions will be obtained if $d\varphi/dx$ is multiplied by the charge of 1 mol ion, i.e. by the quantity zF (F is the Faraday constant, z is the valence number of the ion). The *electrochemical potential* is clearly given by the expression

$$\mu^e = \mu + zF\varphi \qquad [5.45c]$$

From [5.42h], [5.45b] can be rewritten as

$$\frac{d\mu^e}{dx} = \frac{RT}{c} \frac{dc}{dx} + zF \frac{d\varphi}{dx} \qquad [5.45d]$$

and we have

$$J = -ukT \left(\frac{dc}{dx} + \frac{zcF}{RT} \frac{d\varphi}{dx} \right) \qquad [5.45e]$$

For simplicity, we calculate the diffusion potential only for binary solutions containing monovalent cations and anions, that is we shall prove only [5.44a]. For our purpose we write the flux of the monovalent cations and anions with the aid of [5.45e]:

$$J_C = -u_C kT \left(\frac{dc}{dx} + \frac{cF}{RT} \frac{d\varphi}{dx} \right) \quad \text{and} \quad J_A = -u_A kT \left(\frac{dc}{dx} - \frac{cF}{RT} \frac{d\varphi}{dx} \right) \qquad [5.46a]$$

Since the diffusion of the ions in itself (i.e. without an external electric field) does not produce an electric current,

$$J_C - J_A = 0 \qquad [5.46b]$$

and

$$u_C \frac{dc}{dx} + \frac{u_C cF}{RT} \frac{d\varphi}{dx} - u_A \frac{dc}{dx} + \frac{u_A cF}{RT} \frac{d\varphi}{dx} = 0 \qquad [5.46c]$$

If this is rearranged to express the gradient $d\varphi/dx$, [5.44a] is obtained directly.

5.5.3. Membrane equilibrium and membrane potentials

In the life processes, the nutrient molecules, the intermediate and final metabolic products, etc. in most cases do not diffuse in a continuous medium. The membranes covering the cells and certain cell components are permeable to various extents for different substances and participate actively in the processes; they therefore exert a profound influence on the transport processes, and hence on the state of the intra- and extracellular space. In the following section we shall discuss the characteristics of transport processes across the membranes.

If a membrane is perfectly permeable for both the solvent and the solutes, at equilibrium the concentrations (activities) on the two sides of the membrane will be the same. From a practical point of view, however, the processes of interest are those in which the membrane is permeable for only some components, and is impermeable for the others. The osmotic processes already dealt with are of this type, as are the phenomena in solutions containing ions and other charged particles.

1. A simple case of membrane equilibrium. Let us consider the simple case when the same electrolyte is situated on both sides (I and II) of the membrane, but in different concentrations, and the membrane is permeable only for the cation of valence z_C. Thus, only the cations can migrate from side I (higher concentration) to side II, the anion migration being stopped by the membrane. As a result an *electric double layer* is formed on the membrane: one layer consists of the anions stopped on side I, while on side II there is a layer of cations attracted by the anions. The concentration difference "drives" the cations, whereas the electric field of the double layer "pulls them back". At equilibrium the potential difference between the two sides of the membrane ($\varphi^{II} - \varphi^{I}$) and the cation concentrations on the two sides (c_C^I and c_C^{II}) at a given temperature T are related in the following way:

$$\varphi^{II} - \varphi^{I} = -\frac{RT}{z_C F} \ln \frac{c_C^{I}}{c_C^{II}} \qquad [5.47a]$$

where R is the universal gas constant and F is the Faraday constant.

[5.47a] may be derived if the electrochemical potentials of the cations are written for both sides. At equilibrium the two electric potentials are equal. [5.47a] follows directly from this equality. – However, [5.47a] may also be obtained by considering the diffusion potentials, more exactly [5.44b]. In our example the membrane impedes the diffusion of the anion, therefore $u_A = 0$. In this special case [5.44b] turns into [5.47a].

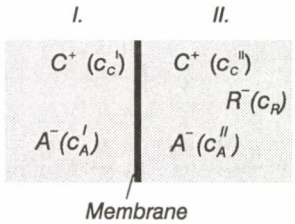


Fig 5.14. Diagram relating to development of Donnan equilibrium

2. Donnan equilibrium. Donnan voltage. Figure 5.14 depicts a somewhat more complicated case. The membrane is fully permeable for the monovalent ions C^+ and A^- of the electrolyte CA but it is impermeable for the similarly monovalent molecular ions R^- on side II. The freely diffusing C^+ and A^- ions are said to be mobile, while the R^- ions stopped on side II of the membrane are immobile. Because of the presence of these latter ions, neither of the mobile ions is evenly distributed on the two sides of the membrane, and a special equilibrium, the *Donnan equilibrium,* is formed. (The expressions in brackets in Fig. 5.14 denote the equilibrium concentrations.) The development of the equilibrium along the membrane is accompanied by the formation of an electric double layer, the equilibrium voltage of which is called the *Donnan voltage.*

The relations for Donnan equilibrium follow from the general conditions of thermodynamic equilibrium. The calculations are carried out in connection with the example outlined in Fig. 5.14. The electrochemical potentials at equilibrium are equal on the two sides of the membrane. The equation must hold separately for the cations C^+ and the anions A^-, and consequently, with the usual notations, we may write

$$\mu_C^0 + RT \ln c_C^{\mathrm{I}} + F\varphi^{\mathrm{I}} = \mu_C^0 + RT \ln c_C^{\mathrm{II}} + F\varphi^{\mathrm{II}}$$

$$\mu_A^0 + RT \ln c_A^{\mathrm{I}} - F\varphi^{\mathrm{I}} = \mu_A^0 + RT \ln c_A^{\mathrm{II}} - F\varphi^{\mathrm{II}}$$

In the terms containing the electric potential, the valence (the number of charges on the ion) has not been denoted separately; it is taken as +1 for monovalent cations and as –1 for monovalent anions. Summation of the above equations gives the following relations for the equilibrium concentrations:

$$c_C^{\mathrm{I}} c_A^{\mathrm{I}} = c_C^{\mathrm{II}} c_A^{\mathrm{II}} \text{ and } \frac{c_C^{\mathrm{I}}}{c_C^{II}} = \frac{c_A^{\mathrm{I}}}{c_A^{II}} = r \qquad [5.47b]$$

According to the left-hand equation, the products of the mobile monovalent ion concentrations are equal on the two sides of the membrane. The equation on the right expresses that the ratio of the different cation or anion concentrations on the two sides of the membrane gives the same number r. Due to the presence of the immobile ions, this number is not unity; it is called the *Donnan ratio.* The Donnan voltage derives immediately from any of the above equations:

$$\varphi^{II} - \varphi^{I} = \frac{RT}{F} \ln \frac{c_C^I}{c_C^{II}} = \frac{RT}{F} \ln \frac{c_A^I}{c_A^{II}} = \frac{RT}{F} \ln r \qquad [5.47c]$$

If the system contains several types of mobile ions, including multivalent ones, power relations hold instead of the above simple concentration ratios, and in this case r denotes their common value. The exponents are different for the individual ions, and are the reciprocals of their valences. Let us denote the valence of the i-th ion by z_i, and the equilibrium concentrations by c_i^I and c_i^{II}, respectively; in this case

$$r = \left(\frac{c_i^I}{c_i^{II}} \right)^{1/z_i} \qquad [5.47d]$$

The expression for the Donnan potential is formally the same in the general case as in the simple example discussed here, except that r must be substituted by [5.47d]. z_i is a positive integer for the cations and a negative integer for the anions.

5.5.4. Transport equations for membranes

1. Permeability constant. The diffusion of neutral solute molecules across membranes can be described by a relation similar to the Fick–Onsager equation [5.42i]:

$$J = -ukT \frac{dc}{dx}$$

However, instead of the concentration gradient dc/dx we employ the quotient $\Delta c/\Delta x$, where the absolute value of Δc is the difference between the concentrations on the two sides of a membrane of thickness Δx. Instead of the mobility of the molecules diffusing in the solution, a mobility-like quantity (u^m) is used, which characterizes the interaction between the molecules and the membrane. Thus, the flux of the i-th component is given by the expression

$$J_i^m = -u_i^m kT \frac{\Delta c_i}{\Delta x} \qquad [5.48a]$$

(The superscript m refers to the presence of the membrane.) Quite frequently, neither u_i^m nor Δx is known exactly, and for this reason they are combined into one quantity which also includes Boltzmann's constant (k) and the temperature (T). The resulting quantity is called the *permeability constant* for the i-th component and is denoted by p_i. Thus

$$J_i^m = -p_i \Delta c_i \qquad [5.48b]$$

p_i is the flux measured in the case of unit concentration difference.

2. Membrane potential. The transport of ions (and other electrically charged particles) in solution may be induced not only by a concentration (chemical potential) gradient, but also by an electric potential gradient. In the case of biological membranes, both factors act simultaneously, for it is found by experience that the concentrations are different on the two sides of the membranes and an electric potential difference too is measured

between the two sides. The electric potential difference (otherwise the electric double layer) is a result of the fact that the permeability of the membrane is different for the different ions in its environment. Consequently, both electric and chemical potential gradients exist inside the membrane. In this case the flux of the k-th ion in the membrane is described by a relation differing from [5.45e] only by the constant $z_k e$ (the charge of the k-th ion):

$$J_k = -u_k z_k \, ekT \left[\frac{dc_k}{dx} + \frac{Fz_k c_k}{RT} \frac{d\varphi}{dx}\right] \qquad [5.49]$$

The difference is due to the fact that [5.45e] holds for the material flow of the ions, whereas [5.49] refers to the electric charge flow of the ions.

From [5.49] it is possible to derive a relation for the *membrane potential.* We do not carry out the detailed calculations, but discuss only the conditions applied and the results obtained.

For biological membranes it may be assumed that the electric potential gradient is constant across the membrane, so that $d\varphi/dx$ may be replaced by the expression $(\varphi_i - \varphi_e)/\delta$ where φ_i and φ_e are electric potentials on the internal and external sides of the membrane (in the intra- and extracellular space), and δ denotes the membrane thickness. Further, it is a well-founded assumption that the ion transport in the membrane is a *stationary* process, i.e. J_k is independent of time. This condition means that J_k is independent of position too: neither charge accumulation, nor charge depletion occur between two arbitrarily selected points of the membrane. Moreover, in full agreement with experience, it may be stated that a charge transport across the membrane produces an electric current only if the system is placed in an external electric field. Under normal circumstances the processes in biological systems proceed without an external electric field, and hence the electric flux flowing through the membranes is zero.

Under such circumstances the potential difference $(\varphi_e - \varphi_i)$ may also be considered as the resting potential of a cell. To obtain it [5.49] has to be written for the electric charge flows of each of the ions, then these have to be summed. Considering that in the summation the "resultant" electric flux is zero, for $(\varphi_e - \varphi_i)$ the following expression is given:

$$\varphi_e - \varphi_i = \frac{RT}{F} \ln \frac{\sum_{k=1}^{m} p_k^+ c_{ke}^+ + \sum_{k=1}^{n} p_k^- c_{ki}^-}{\sum_{k=1}^{m} p_k^+ c_{ki}^+ + \sum_{k=1}^{n} p_k^- c_{ke}^-} \qquad [5.50]$$

where only the n species of monovalent anions and the m species of similarly monovalent cations have been taken into account, since mainly these participate in the ion transport in biological membranes. The plus and minus signs refer to the positive and negative ions, c_{ki} and c_{ke} denote the ion concentrations of the solution on the internal and external sides of the membrane, and p_k are the permeability constants of the k-th ion of the membrane. By the way, [5.50] is the solution of [5.49] for the resting potential, called also Hodgkin–Huxley–Katz equation.

3. The transport of water. Water is a basic component of the living cell, and consequently a discussion of some of the characteristic features of its transport is justified. Most biological membranes are permeable to water, but they display differences in their tolerance of hydrostatic pressure differences. The glomerular membranes are

relatively rigid and are thus able to maintain relatively large pressure differences. Red blood cells are less rigid, and the membranes of amoebae, for example, are even less rigid. In plant cells the rigid cellulose matrices enable the thin membranes to endure relatively large pressure differences.

The flux J^m_{water} across a membrane is determined by two types of pressure differences: the hydrostatic pressure difference (Δp_{hst}) between the two sides of the membrane, and the osmotic pressure difference (Δp_{osm}) due to the concentration differences. The flux of water is proportional to the algebraic sum of these quantities:

$$J^m_{water} = -p_{water}\,(\Delta p_{hst} - \Delta p_{osm}) \qquad [5.51]$$

The proportionality factor p_{water}, the *hydrodynamic permeability coefficient,* depends upon the mobility of water in the membrane and the membrane thickness. In order to explain the signs, it is enough to remember that the direction of water transport is determined by the direction of the hydrostatic pressure decrease and by that of the osmotic pressure increase.[9] It depends upon the relative magnitudes and signs of Δp_{hst}, and Δp_{osm} whether in a given case the water flows from the more dilute to the more concentrated solution or in the opposite direction. With a sufficiently large hydrostatic pressure it can be attained that the water is forced from the solution of higher concentration into the solution of lower concentration. This phenomenon is called *ultrafiltration,* which plays an important role in biological material transport and is also utilized in practice. As an example, [5.51] explains the fact that liquid efflux occurs at the arterial end of capillaries, whereas influx takes place at the venous end.

According to van't Hoff's law, the quantity Δp_{osm} can be substituted by the product $RT\Delta c$, where Δc denotes the difference in concentration of solute on the two sides of the membrane. Thus, instead of [5.51] we may write

$$J^m_{water} = -p_{water}\,(\Delta p_{hst} - RT\Delta c) \qquad [5.52]$$

The van't Hoff relation in this form is a good approximation if the membrane is perfectly impermeable for the solutes. In this case Δc denotes the total molar concentration difference, taking into consideration every solute. For biological membranes this requirement is generally not satisfied, because most of the ions or molecules are passing through the membrane to various degrees. This process is accounted for by introducing the *reflection constant* (δ), which is the quotient of two water fluxes. The numerator contains the flux produced by the semipermeable membrane for a given solute at a given concentration difference, and the denominator contains the flux measured if the membrane is perfectly impermeable for the solute at the same concentration difference. In the case of impermeable material $\delta = 1$, and for perfectly permeable substances $\delta = 0$ (these latter do not produce any osmotic pressure). With biological membranes, for most solutes $0 < \delta < 1$. On introduction of the reflection constant, [5.52] takes the following form:

$$J^m_{water} = -p_{water}\left(\Delta p_{hst} - RT\sum_{i=1}^{n}\delta_i\Delta c_i\right) \qquad [5.53]$$

where the summation must be carried out for every solute species in the system. The above reasoning holds exactly only for simple, homogeneous membranes. With more composite membranes the flux is a non-linear function of the osmotic pressure.

5.5.5. Apparent anomalies in the transport processes

In the previous section such processes were discussed in which the transport took place in the direction of the concentration gradient (thermodynamic force). However, in the biological systems, containing different membranes, there are also such cases in which some molecules or ions move towards the higher concentration. Such an "uphill" process cannot take place by itself, another, so-called "downhill" transport directed towards the lower concentration (or an energy producing chemical reaction) also has to take place – moreover the two processes have to be coupled. The first process takes place with the help of the second.

[9] [5.51] is a consequence of Onsager's law; its derivation is omitted, since the result is simple and illustrative.

Onsager's law may be applied in these cases too and it helps their understanding. Namely, according to the experience, the current of a so-called extensive quantity is determined not only by the inhomogeneity of the "corresponding" intensive quantity, but more or less also by every thermodynamic force present in the system. In the Onsager equations this may be expressed, too.

In the simple case in which only two processes are coupled, i.e. two thermodynamic forces (X_1 and X_2) and two fluxes (J_1 and J_2) are present, the following Onsager equations may be written:

$$J_1 = L_{11}X_1 + L_{12}X_2$$

$$J_2 = L_{21}X_1 + L_{22}X_2.$$

X_1 and X_2 are the thermodynamic forces *corresponding* to J_1 and J_2, respectively, therefore the factors L_{11} and L_{22} are called "direct" phenomenological coefficients. On the other hand, factors L_{12} and L_{21} characterize the relations between fluxes and thermodynamic forces *not associated* with each other, the so-called *cross effects*, therefore they are called cross coefficients. The condition of the presence of *cross effects* is that the factors L_{12} and L_{21} characterizing the relation should be non-zero. The above will be illustrated by examples which are important also in practice.

1. Coupled transport. According to the experience of the last years, the direction of the transport of several neutral molecules and ions (e.g. sugars, amino acids, Ca^{++}, PO_4^{3-}, organic anions) is opposite to the direction which could be expected on the basis of their chemical or electrochemical potential gradient and corresponds – by coupling – to the electrochemical potential gradient of the hydrogen or sodium ions. Molecules or ions having such "anomalous" transport are connected to hydrogen or sodium ions. For example, such coupled transport ensures the uptake of phosphate and other, organic, anions by the mitochondria (by means of thermodynamic forces related to the H^+ ion and directed inwards) or the glucose uptake of the epithelial cells (here the force comes from the Na^+ ions).

Consider the glucose uptake of the tubular cells of the kidney under the effect of the Na^+ gradient and apply the previous system of equations:

$$J_{gl} = -L_{11}\frac{d\mu_{gl}}{dx} - L_{12}\frac{d\mu^e_{Na^+}}{dx}$$

$$J_{Na^+} = -L_{21}\frac{d\mu_{gl}}{dx} - L_{22}\frac{d\mu^e_{Na^+}}{dx}$$

On the left side there are the inwardly directed fluxes of the glucose and Na^+ ions, on the right side the thermodynamic forces creating the fluxes as well as the direct and cross coefficients, respectively.

Consider the first equation. In case of a functioning cell the concentration of glucose is higher in the intracellular space, while that of the Na^+ ion is higher in the extracellular space. The same holds for the chemical potential of glucose and the electrochemical po-

tential of the Na^+ ion. The thermodynamic force acting on glucose is directed towards the extracellular space, while the force acting on the Na^+ ion towards the intracellular space, respectively. The right side of the equation was written by taking this into consideration. – The transport of glucose towards the intracellular space, necessary for the metabolism of the cell, takes place if the resultant of the forces on the right side is directed towards the interior of the cell. This happens in reality, since by means of coupling the *Na^+ ions help* the transport of the glucose molecules towards the inside of the cell, for which it is naturally also necessary that the absolute value of L_{12} be sufficiently large, and that of L_{11} sufficiently small. – The lower equation does not have to be dealt with, since the main driving force of the flux of the Na^+ ions comes from their own electrochemical potential gradient, thus $L_{21} \approx 0$, therefore the right side of the lower equation is reduced to a single term.

Coupled transport may be present also in cases of fluxes with opposite directions. For example, the transport of the Na^+ ions towards the inside of the myocardial cells, more exactly the inwardly directed potential gradient helps the outflow of the Ca^{++} ions.

2. The so-called active transport is also a transport against the concentration gradient, but in this case the energy necessary for the transport of ions and molecules to a site of higher potential is provided by metabolic processes (e.g. the catabolism of ATP). The best-known active transport is the so-called Na^+, K^+ pump.

For the K^+ and Na^+ ions there is a manyfold concentration ratio between the extra- and intracellular spaces (in the opposite directions for the two ions) and these concentration differences make possible the functioning of the cells. The concentration differences can be kept naturally only if, in addition to the so-called passive Na^+ and K^+ transport in the direction of the concentration gradient, there is also a transport against the concentration gradient. This is the transport of K^+ ions towards the interior of the cell and that of the Na^+ ions towards the extracellular space, the so-called active Na^+, K^+ transport the driving force of which comes from the energy-producing ATP $\rightarrow$ ADP + phosphate reaction occurring in the intracellular space. The available energy is taken into consideration with the affinity of the reaction (A_{ATP}, cf. section 5.2.2).

Again, the previous equations are taken as starting points, but for the sake of simplicity only the equation concerning the active transport of sodium is written down

$$J = -L_{11}\frac{d\mu^e_{Na^+}}{dx} + L_{12}A_{ATP}$$

The equation expresses that the flux of the sodium ions is influenced not only by the thermodynamic force arising from the electrochemical potential of the Na^+ ions, but also by the affinity of the energy-producing chemical reaction (ATP catabolism) – in the latter case in the opposite sense. Between the sodium transport and the chemical reaction a cross effect develops which is expressed by that the cross coefficient L_{12} is not zero.

On the basis of the equation a further conclusion may be drawn. Since the flux has a definite direction (it is a vector), the expressions on the right side of the equation must also behave as vectors. The electrochemical potential gradient in the first term is by definition a vector which is multiplied by the scalar L_{11}. Thus the first term is indeed a vector. On the other hand, in the second term the affinity is a scalar, therefore the cross coefficient L_{12} must be vectorial, which means that the chemical reaction proceeds in the form of a di-

rected ion flux. For the synchronous occurrence (coupling) of the above scalar and vectorial processes structural anisotropy must be present. In other words: the active transport may be realized only in an anisotropic medium. The biological membranes are in fact such media: they are built up from oriented molecules (cf. section 1.5.5), and the active transport is made possible by the contribution of an enzyme protein extending through the membrane and facilitating also the catabolism of intracellular ATP. At the molecular level this is the sense of the statement that the cross coefficient L_{12} is not zero and is a vector.

The coupling between the chemical reaction and the ion transport may be established also in the opposite sense, namely: an electrochemical potential gradient may facilitate the synthesis of compounds with higher internal energy. A well-known example for this is the ATP synthesis of the mitochondria (from ADP and phosphate), the driving force of which comes from the electrochemical potential gradient of the hydrogen ions between the two sides of the internal membranes of the mitochondria. The presence of the membrane is necessary for the coupling also in this case.

Let us now compare passive and active transport from an *energetic* aspect.

Passive transport is a spontaneous, isothermal process during which the Gibbs free energy of a system consisting of a membrane and the solutions on the two sides of the membrane *decreases,* i.e. $\Delta G < 0$. From the relation $\Delta G = \Delta H - T\Delta S$, a decrease in the Gibbs free energy may be due to either a decrease in enthalpy ($\Delta H < 0$) or an increase in entropy ($\Delta S > 0$). In the course of passive transport the internal energy of the system does not change considerably; there is usually no appreciable volumetric work, though because of the concentration equalization the entropy of the system increases (the ordering decreases). The decrease in the Gibbs free energy is mainly due to this latter circumstance.

In the course of *active transport* the ordering of the ions on the two sides of the membrane increases, which results in a decrease in the entropy of the transport system. Since the enthalpy does not change considerably in the process, the Gibbs free energy of the system *increases* due to the decrease in the entropy. Active transport can take place only if the strict transport system is associated with a process whose Gibbs free energy decrease can cover the above-mentioned increase. The process with decreasing Gibbs free energy associated with the transport is always some chemical reaction, in most cases the decomposition of adenosine triphosphate (ATP). In this latter case the Gibbs free energy decrease is a consequence partly of the decrease in the enthalpy and within this the internal energy, and partly of the increase in the entropy. The internal energy decreases in this case as a result of the rearrangement of the atoms in the course of the reaction, and the entropy increase follows from the increase in the disorder resulting from the decomposition. In the decomposition of ATP, the enthalpy decrease and the entropy increase depend sensitively upon the concentrations of other substances present in the medium (e.g. hydrogen ion, magnesium ion). In the living organism the two factors participate nearly equally in the change of the Gibbs free energy.

REFERENCES

Books

Ayres, R. U., Informatin, Entropy and Progress. A New Evolutionary Paradigm, AIP Press, New York (1994)

Guggenheim, E. A., Thermodynamics (6th edition). North-Holland Publ. Comp., Amsterdam (1977)

Katchalsky, A., Curran P. F., Non-equilibrium Thermodynamics in Biophysics. Harvard University Press, Cambridge, Mass. (1965)

Kreuzer, H. J., Nonequilibrium Thermodynamics and its Statistical Foundations. Clarendon Press, Oxford (1981)

Lamprecht, T., Zotin, A. I., Thermodynamics of Biological Processes. W. de Gruyter, Berlin (1978)

Sybesma, Ch., Biophysics. An Introduction. Kluwer Academic Publishers, Dordrecht–Boston–London (1989)

Tosteson, D. C., Ovchinnikov, Yu. V., Latore, R., Membrane Transport Processes. Raven Press, New York (1978)

6. BIOMEDICAL ELECTRONICS

A large variety of electronic devices are used in medical practice. These are devices employed mainly in the diagnostic but also in therapeutic practice. Technical development – especially the rapid development in microelectronics and computer technics – brought the qualitative innovation of some methods used for a long time in medicine, but it made possible also the birth of some new possibilities.

This chapter deals mainly with the general physical and technical bases necessary for an understanding of the function of the devices used in medical electronics; however, their general character means that they can serve as the basis of more wide-ranging information.

In the discussion of the subject technical details are avoided, only the basic principles, and the larger functional units will be dealt with, as far as necessary for an understanding of their function.

As it was mentioned, most of the devices in medical technics, that we are interested, serve for diagnostic purposes. That is why we emphasize first, as a guideline, that the basis of diagnosis is the collection of necessary information. The way for it is: *measurement and procession of signals.*

6.1. Signals as information carriers

1. Signals in general, and their role in medicine. The organism, and within it the individual cells, tissues and organs (in brief any biological system), continually interacts with the surroundings and with other systems. In one direction of this interaction the organism perceives effects arriving from the external world and processes the information content of these effects. This is the basis of the adaptation of the organism to the surroundings and its changes. The other direction of the interaction is connected with the fact that the life functions are accompanied by various phenomena which can be observed and recorded and which supply information that may be processed by the environment. The information is always carried by some physical quantity, e.g. light, sound reaching the organism or heat emitted by it. *A quantity (or its change) carrying or storing information is called a signal.* However, the definition means that the concept of a signal is more general than indicated by the above examples. For a presentation of the more general concept of signals, it is worthwhile to mention some additional examples. For instance, the electric potential generated on cardiac action yields valuable

information on the activity of the myocardium; similarly, the blood glucose concentration is an important indicator of the metabolism; the DNA base sequence carries genetic information; and so on.

The signal, or information, may refer to the state of the system, to some process, some phenomenon, etc. One such state parameter is the body temperature. The signal associated with this is continuous, and its magnitude is approximately constant: it is a *static signal.* As an example of periodic processes, the ECG signal associated with the heart function may be mentioned, it is changing in time, a *nearly periodic signal.* An instance of a single process is the stimulation process induced by an electric pulse, accompanied by an electric signal of characteristic form, by a *pulse signal.* A random sequence of individual events (stochastic process) is the γ-radiation emitted by some radioactive preparation. The voltage pulse signals of the scintillation counter correspond to individual γ-photons, this is a *series of stochastic pulse signals.*

Of the quantities used as signals, electric signals can relatively easily be processed by electronic devices. For this reason, as a first step the originally non-electric signals are transformed to electric signals. During processing further transformations of the electric signal may become necessary. However, it is an absolute requirement that the information content of the signal must not change during its transformation.

Two types of signal transformation are known: *analogue* and *digital* transformation. In the former case the time course of the transformed signal is similar (analogous) to that of the original signal. Thus, the electrocardiogram represents a graphically recorded analogue signal of the action potential changes of the myocardium. In the case of a digital transformation, elements of a symbol system, in most cases numbers, are unambiguously assigned to the instantaneous values of the signal. This happens, for instance, when a change in the myocardial action voltage is stored as a series of numerical values in the memory of a computer (cf. converters, section 6.6.1).

Without the intention to be exhaustive, the methods and devices applied in biomedical electronics may be classified as follows:

(a) the processing of the signals associated with the state or function of the system under investigation (cell, organ, organism), with the purpose of obtaining diagnostic information (e.g. electroencephalography, thermography);

(b) a signal produced by an electronic device is introduced into the system to be investigated to obtain information on its structure, state or function (e.g. echocardiography);

(c) a signal produced by an electronic device is introduced into the system to influence its state or function (e.g. high-frequency heat therapy).

Figures 6.1*a, b* symbolize diagnostic applications, and Fig. 6.1*c* therapeutic ones. The arrows in the diagram indicate the direction of information flow, which is always associated with some energy transport. Signal power is mentioned in this sense throughout this chapter.

2. The concept of decibel. Signals can be compared by means of the ratio of their powers. Frequently, the logarithm of this ratio to the base ten is used; or this value multiplied by 10. The scales thus obtained are called the bel and decibel scales (denoted

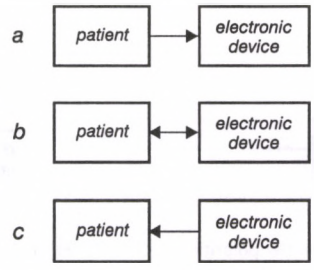


Fig. 6.1. Typical cases of signal energy flow

by B or dB). Let us denote the reference power by P_1 and the power in question by P_2; the relations used for their comparison are then

$$n\,(\mathrm{B}) = \log \frac{P_2}{P_1}\,\mathrm{B} \quad \text{or} \quad n\,(\mathrm{dB}) = 10 \log \frac{P_2}{P_1}\,\mathrm{dB} \qquad [6.1]$$

The bel scale is too large for practice, therefore the decibel scale is frequently used. If the signal is an electric one, the calculations may involve voltage instead of power ratios. In this case

$$n = 20 \log \frac{U_2}{U_1}\,\mathrm{dB} \qquad [6.2]$$

where U_2/U_1 denotes the voltage ratio.

[6.2] can easily be obtained from [6.1] if the relation $P = U^2/R$ is considered and if it is assumed that the two signal powers or signal voltages appear on the same resistance R.

3. Signal-to-noise ratio. The signal to be processed is frequently accompanied by some identical or similar signal produced by the signal source or the surroundings. For instance, the observation of cardiac sounds may be disturbed by various other sounds produced either by the patient or by the environment. This accompanying signal disturbs the information content of the signal to be processed (disturbing signal, noise). The signal-to-noise ratio is usually expressed by the ratio of the signal voltage (U_{sig}) to the noise voltage (U_{noise}), or on the decibel scale.

6.2. Electronic units and basic circuits

Electronic systems are built up from combinations of basic elements for certain purpose. The more important elements are voltage or current sources, constant or variable resistors and capacitors, induction coils, rectifiers (diodes), amplifier units (e.g. transistors), integrated circuits, displays (e.g. cathode-ray tubes, liquid crystals, LEDs, etc.), sensors and transducers (photocells, thermoelements and so on). Any desired function can be obtained by various combinations of the elements; for this reason the user need not study the working of an electronic system in detail; it is sufficient to know its *functional units* and their interrelation. The functional units are frequently called blocks, which are usually represented by rectangles with their function written inside them.

Certain combinations of elements play a fundamental role in the individual blocks; these are basic electronic circuits. In the following sections some units and basic electronic circuits will be considered in more detail.

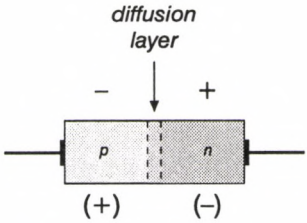


Fig. 6.2. Semiconductor diode; above: the polarity of the diffusion voltage, below: the polarity of the forward voltage

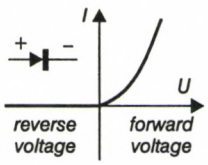


Fig. 6.3. Symbol and characteristics of the diode

1. The semiconductor diode is a semiconductor crystal one half of which has a *p*-type doping, the other half an *n*-type doping, respectively (cf. section 1.4.5). Both are supplied with a conducting contact (Fig. 6.2). In the thin border layer between the two parts the charge carriers diffusing there – electrons from the *n* layer, defect electrons from the *p* layer – *recombine* with each other. All this has two consequences. The first one is that in this so-called *diffusion layer* only a very limited number of free charge carriers are left, therefore its resistance is high. The other consequence is that – since the electrons left the *n* layer and diffused to the *p* side – the *n* layer turns positive, the *p* layer negative. The *diffusion voltage* thus developed (cf. section 5.5.2) brings the diffusion to a stop. Let us now apply voltage to the diode. If its polarity is the same as that of the diffusion voltage, the diode does not conduct *(reverse voltage)*. If, on the other hand, the voltage is opposite to that of the diode, it destroys its effect and guides charge carriers into the diffusion layer: the diode conducts *(forward voltage)*. Therefore the diode may be used as a *rectifier.* Figure 6.3 shows the dependence of the diode current on the voltage applied to the diode *(diode characteristics)*.

2. The transistor consists of three semiconductor layers with three electrodes: *emitter* (*E*), *base* (*B*) and *collector* (*C*). The order of the layers may be *p-n-p* or *n-p-n,* according to their impurity. The middle layer is very thin and joins its neighbours on both sides by a diffusion layer. (The following considerations concern the *p-n-p* transistors, but are valid also for the *n-p-n* transistors as appropriate.) It operates with two circuits (Fig. 6.4): the *base circuit* with the base voltage source U_B, and the *collector circuit* with the collector voltage source U_C. These two circuits are not independent of each other, and this is the basis of the transistor application.

If only the collector circuit voltage would be applied, current would not flow through the transistor in either of its polarity, since one of the two diffusion layers (their order: *p-n–n-p*) would stop the conduction, just like in the case of diodes. Let us, however, apply both voltage sources. U_B is effective as forward voltage at the *p-n* transition of the emitter–base, under its influence charge carriers (defect electrons) overflow the base. This layer, as already mentioned, is very thin, a large part of the charge carriers diffuses into the collector layer; of course, they are also helped in this by the U_C, since it is about ten times higher than the U_B: the defect electrons are practically "sucked" into the collector layer by the higher collector voltage.

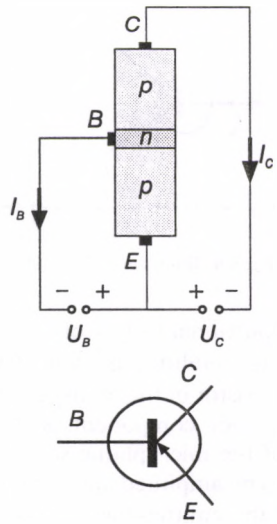


Fig. 6.4. Circuits and graphic symbol of the *p-n-p* transistor

The intensity of the collector current I_C is proportional with a good approximation to the intensity of the base current I_B (Fig. 6.5), their ratio, I_C/I_B, the *current amplification factor* is usually between 10 and 100.

On the basis of the above it is understandable that the transistors may be used both as amplifiers and switching units.

a) Their application as *amplifiers* is shown by the example of a simplified sound amplifier (Fig. 6.6). The sound vibrations are analogously transformed by the microphone into audio-frequency voltage, the latter is the signal voltage to be amplified, superposed on voltage U_B. The current intensity of the base circuit I_B changes in proportion to the signal voltage (according to the sound vibrations), while the current intensity of the

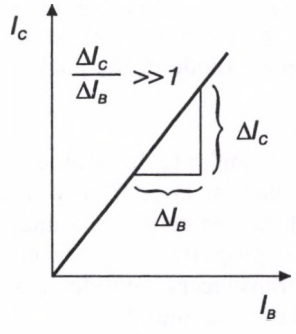


Fig. 6.5. I_C/I_B characteristics of the transistor

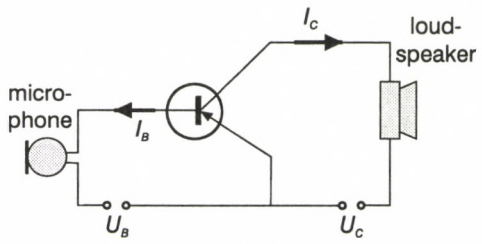


Fig. 6.6. Transistorized amplifier

collector circuit I_C is always proportional to I_B. Consequently I_C will also have an audio-frequency component which is the multiple of that of the base: this sounds the loudspeaker. Naturally, through the resistor of the loudspeaker voltage drop develops; its part proportional to the audio-frequency component is the amplified signal voltage; the relation of this to the voltage of the microphone serves the voltage amplification. If in addition to the signal voltages (to be amplified and amplified) the resistors through which they appear (i.e. the resistor of the emitter–base of the transistor and that of the loudspeaker) are also taken into consideration, the signal powers and their ratio: the power gain may be also calculated.

b) Their use as *switching units* is made possible by the fact, that collector current will flow in a transistor only with a forward base voltage; with a reverse base voltage, the intensity of the collector current is zero. Consequently, the transistor may be used as an electronic switch which, depending upon the base voltage, disconnects or connects the emitter and collector electrodes. If suitable auxiliary circuits are applied, the switching occurs at a preset base voltage.

3. One of the layers of the **photodiode** (either layer *p* or layer *n*) is so thin that the light falling on it reaches the diffusion layer *p-n* and releases charge carriers in it, in other words it forwards electrons into the conducting band, while in the valence band defect electrons are left over. These free charge carriers are moved by the diffusion voltage in the appropriate direction: the electrons get into layer *n*, the defect electrons into layer *p*. Consequently under the effect of illumination the photodiode becomes a voltage source, which makes possible two kinds of its application. On one hand it may be used as light-signal transducer (light detector). On the other hand it may serve as an electronic source of energy. For example, the solar cells are in essence large-surfaced photodiodes made of silicon. The surface of the photodiode in the pocket calculators is some cm^2, in the spaceships several m^2.

4. Phototransistors are devices similar to the photodiode in that also the light produces free charge carriers, namely in the base layer. Thus without base voltage (and even without base circuit), only by applying the collector voltage, current flows in the collector circuit, the intensity of which is proportional to the illuminating light. The phototransistors are used as light-signal transducers (light detectors). Their advantage is that they are very sensitive, due to the transistor function.

5. Integrated circuits *(IC).* Here, a large number of transistors, diodes, resistors, etc. are formed together with their connections in a silicon plate.[1] Integrated circuits, being important building elements in modern electronics are completed with further discrete circuit elements (resistors, capacitors, etc.) to form functional units, such as clocks and pacemakers (cf. section 6.5.2). The solid-state memories of the computers and the *microprocessors* are also integrated circuits (cf. section 8.3.1).

As a result of the development of microelectronics, not only the dimensions of the devices but also their power consumption has been decreased by several orders of magnitude. From the medical viewpoint beside the decrease in size the further advantage is that increasing number of devices are portable and can be operated independently of the electric mains network, which is important from the aspect of safety, too.

6. The CCD (charge-coupled device)-plate (image recorder, image transducer) is an integrated circuit containing about hundred thousand phototransistors (or photodiodes), isolated from each other, arranged in several hundred rows in a matrix on a silicon plate. The real image of the object is formed on the plate by an objective; the individual detectors provide a signal voltage proportional to the intensity of the light illuminating them. The reading of the signals of the image spots takes place by means of the auxiliary circuits integrated into the plate. The CCD-plate works very quickly, e.g. in case of the transmission of TV images with a frequency of 25 images/s. It may be prepared in such a small size (e.g. on a single platelet with an edge of some mm length: plate chip) that it may be built even into an endoscope (cf. section 6.7.1). Its name refers to the charge carriers released by the light which ensure the circuit between the emitter and the collector of the phototransistor.

7. Voltage division. Potentiometer. If some voltage is applied to two or more resistors connected in series (Fig. 6.7*a*), the ratio of the voltages on the resistors corresponds to the ratio of their resistances (e.g. $U_1/U_2 = R_1/R_2$). This is valid for both direct and alternating current, and the resistances may be not only ohmic, but also inductive and capaci-

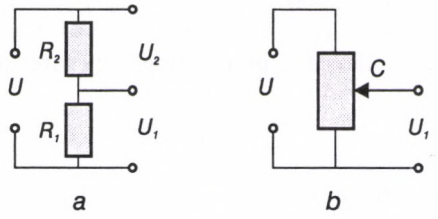


Fig. 6.7. Voltage division

[1] The degree of integration may be characterized by the number of transistors on the semiconductor plate (chip). In the SSI (Small Scale Integration) technology a single silicon plate with a surface of a few tenths of a cm^2 contains not more than 50 transistors, in the VLSI (Very Large Scale Integration) technology the number of transistors is already 10^4–10^5. As a synonym of IC also the term chip is often used which refers to the size of the semiconductor platelet.

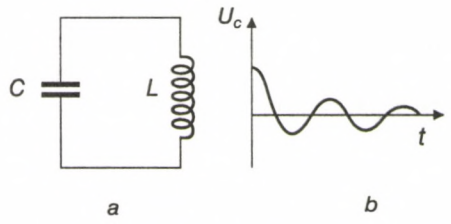


Fig. 6.8. LC circuit *(a)*; voltage changes of the capacitor during damped oscillation *(b)*

tive (cf. the *RC* circuit below). One frequently used solution is depicted in Fig. 6.7*b*. The device, called a potentiometer, allows a continuously variable voltage division by movement of the sliding contact *C*. The knobs regulating amplification, light intensity, sound intensity, etc., on electronic devices are usually the rotatable knobs of potentiometers.

8. *LC* circuit. If a capacitor of capacitance *C* is discharged through a coil with a self-induction coefficient *L*, its energy will be partly radiated as electromagnetic field energy by damped electromagnetic oscillation, partly transformed to heat (Fig. 6.8). The frequency of the oscillation (ν) depends on the characteristics of the circuit:

$$\nu = \frac{1}{2\pi\sqrt{LC}} \qquad [6.3]$$

This is called the *eigenfrequency* of the *LC-circuit*, the circuit itself is called *oscillating circuit.*

If an alternating voltage is applied to the oscillating circuit and its frequency is the same as the eigenfrequency, *resonance* takes place. The eigenfrequency may have any values, e.g. if a capacitor with a variable capacitance is applied; this is the basis e.g. of the tuning of the radio to a desired radio sender. (Its applications will be dealt with in section 6.4.)

9. *RC* circuit. If a capacitor with a capacitance *C* is discharged through a resistor *R* (together with it forms an *RC* circuit), its voltage (U_C) decreases exponentially in time from the initial value (U_0) (Fig. 6.9). The time course of the process is similar to the decrease of the activity of a radioactive preparation, the expression describing the process is also similar:

$$U_C = U_0 e^{-t/RC} \qquad [6.4]$$

It may be easily seen that the product in the denominator of the exponent has the dimension of time:

$$\tau = RC \qquad [6.5]$$

and it is equal to the time required for the voltage of the capacitor to decrease to the *e*-th (approx 37%) of its initial value. This time is characteristic of the *RC* circuit, and analogous with the average lifetime of the radioactive atoms, it is called the *time constant of the RC circuit*.

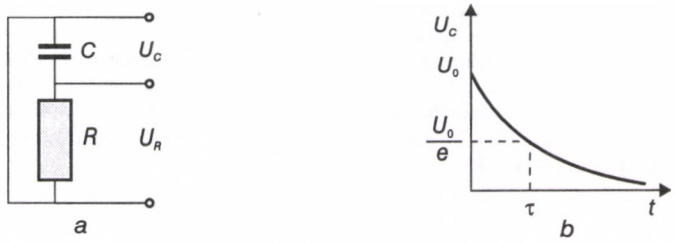


Fig. 6.9. RC circuit *(a)* and its discharge *(b)*

It is obvious that in the *RC* circuit in Fig. 6.9a the sum of the voltages $(U_R + U_C)$ is always zero, thus the prevailing value of U_R is always $-U_C$.

In points 10–13 some applications of the *RC* circuits will be demonstrated; however, it is worthwile to mention also that the discharge of the *RC* circuit may serve as a model of the exponential processes, such as the equalization of temperature or concentration. Similarly, processes resulting in saturation, e.g. the accumulation of a radiopharmacon in the target organ, may be modelled by charging the capacitor.

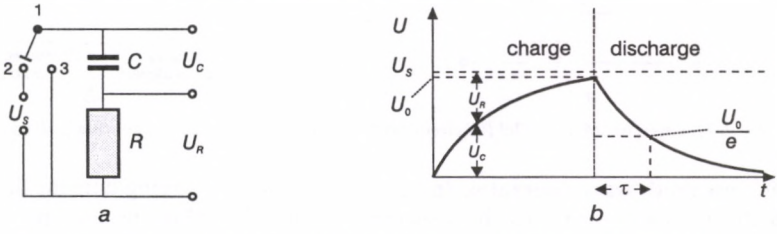


Fig. 6.10. Diagram of *RC* circuit in series *(a)*; the process of charging and discharging *(b)*

a) If a constant direct voltage (U_S) is applied to a series-connected capacitor and resistor (series *RC* circuit), the capacitor is charged through the resistor (Fig. 6.10*a*). The voltage of the capacitor U_C approaches U_S monotonously, but more and more slowly, its polarity is opposite to that of U_S. Therefore during charging the intensity of the current flowing in the circuit – which is proportional in every moment to the difference of U_S and U_C – decreases exponentially with the increase of U_C; U_R, which is proportional to it, behaves similarly:

$$U_R = U_S e^{-t/\tau} \qquad [6.6]$$

Since during charging it is always true that

$$U_R + U_C = U_S \qquad [6.7]$$

from [6.6] and [6.7] it follows that

$$U_C = U_S (1 - e^{-t/\tau}) \qquad [6.8]$$

Figure 6.10*b* illustrates the relations between the voltages expressed in the above three expressions.

b) If such a current source is connected to a parallel *RC* circuit which provides the *RC* circuit with a constant current intensity while the switch is closed, a voltage change may be observed similar to that observed by the investigators who attempted to influence the resting potential of the cell membrane by short-term current pulses of constant intensity (cf. section 7.2.4). Namely: a (decreasing) part of the current of a constant intensity I_0 flowing during the closure of the switch charges the capacitor C, its other (increasing) part flows through the resistor R. It may be easily seen that on one hand U_R is always equal to U_C, on the other hand they increase together (with a decreasing velocity) up to a limit $I_0 R$:

$$U_R = U_C = I_0 R\,(1 - e^{-t/\tau}) \qquad [6.9]$$

When the charging current is stopped by the opening of the switch, the self-discharge of the *RC* circuit immediately begins (Fig. 6.11*b*) as it was seen already in Fig. 6.9 (and in the expression [6.4]).

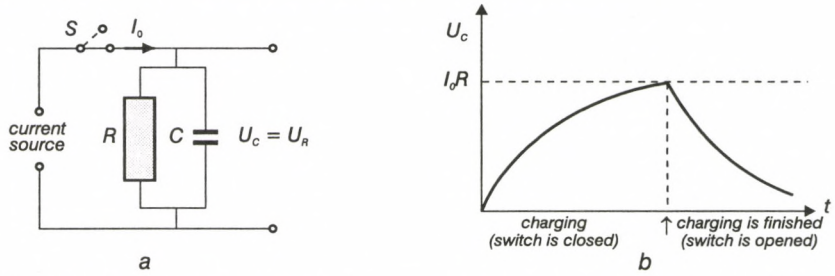


Fig. 6.11. Charging of a parallel *RC* circuit with constant current (*a*) and the voltage diagram of it (*b*)

10. Saw-tooth wave generator. In case of a constant charging current the voltage of a capacitor increases evenly (with a constant velocity). Roughly this takes place in the initial stage of charging of the series *RC* circuit, while $U_C < U_T$. This linear charging process is used to produce the so-called saw-tooth voltage. The saw-tooth voltage generator is outlined in Fig. 6.12*a*. The capacitor voltage increases uniformly up to a certain value; when this has been attained, the switching circuit S (cf. section 6.2) discharges the capacitor, after which the process is repeated (Fig. 6.12*b*). Saw-tooth wave generators are used, for instance, in cases when various periodic phenomena are displayed by cathode-ray oscilloscopes. In these cases the horizontal (time) axis of the oscilloscope is given by the linearly increasing range of the saw-tooth voltage (cf. section 6.3.2).

11. Capacitive current. In conductors electric current is generated by charge flow. Though in the case of ideal insulating dielectrics there is no charge movement, one may e.g. refer to the current flowing through a capacitor. The variation of the electric field strength is also equivalent to current since it creates a magnetic field in the same way as the so-called conduction current in conductors. The current due to field strength variation is called capacitive current.

The intensity of capacitive current (I_C) is proportional to the change of the electric field strength in unit time (dE/dt) and to the area (A) of the capacitor plate

$$I_C = \varepsilon\,\varepsilon_0 A\,\frac{dE}{dt} \qquad [6.10a]$$

where ε is the relative dielectric constant of the dielectric between the capacitor plates and ε_0 is the absolute dielectric constant of vacuum ($\varepsilon_0 = 8.86 \times 10^{-12}$ As/Vm). Let us take into account that the electric field strength between the capacitor plates is

$$E = \frac{U}{d}$$

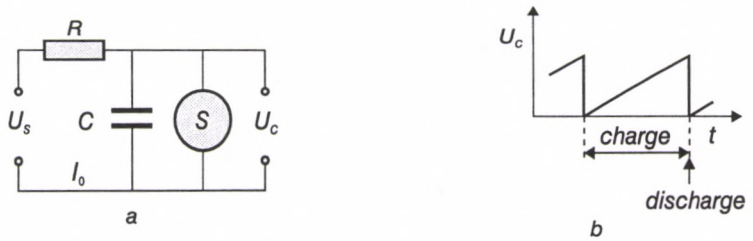


Fig. 6.12. Saw-tooth voltage generator (*a*); saw-tooth oscillation (*b*)

and the capacitance (C) of the capacitor is

$$C = \varepsilon \varepsilon_0 \frac{A}{d}$$

U denotes the voltage and d the distance between the plates. Thus after conversion of [6.10a] the expression

$$I_C = C \frac{dU}{dt} \qquad [6.10b]$$

is obtained for the capacitive current (cf. section 7.2.4: capacitive current).

Thus in the case of a series *RC* circuit, too we are dealing with a closed circuit: in the wires and the resistor conduction current, in the dielectric of the capacitor (or even in vacuum) capacitive current flows.

If the dielectric is not an ideal insulator but it has a finite resistance and conductivity (dissipative dielectric), the capacitor itself shows the behaviour of a parallel *RC* circuit as part of the electric circuit in question (Fig. 6.13). The above-mentioned are valid in both direct and alternating current circuits.

This is the case e.g. in cell membranes which, as nonideal insulators, separate media which conduct better than themselves: thus the membrane, together with its surroundings, behaves as a capacitor with a dissipative dielectric, the relatively highly conducting extra- and intracellular electrolyte spaces being its "capacitor plates", and the membrane its dissipative dielectric. Similar phenomena occur during the warming of the body tissues in high-frequency electromagnetic fields (cf. high-frequency heat therapy, section 6.4.3).

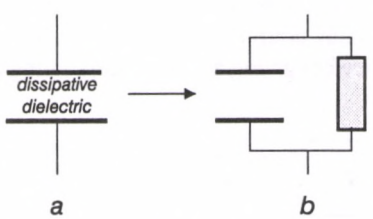


Fig. 6.13. Capacitor with dissipative dielectric *(a)* equivalent to a parallel *RC* circuit *(b)*

12. Ratemeters. A usual way for the processing of pulse signals is the determination of *pulse frequencies,* e.g. the particle flux in nuclear radiation. Such a task is e.g. in isotope diagnostics the determination of the quantity of an isotope accumulation in the target organ and its changes in time. These types of measurements may be carried out with a ratemeter, consisting essentially of *RC* circuits (Fig. 6.14). It is a basic condition of the measurement that every pulse input should produce equal charges on the capacitor C across the resistor R_1. This condition can be achieved by forming signals of identical amplitude and shape prior to the input on one hand, on the other hand by keeping the voltage of the capacitor U_C much smaller than the amplitude (voltage) of the incoming pulses with the selection of the appropriate time constants R_1C and R_2C. Under such circumstances the voltage U_{out} appearing on the resistor R_2 is proportional to the pulse frequency.

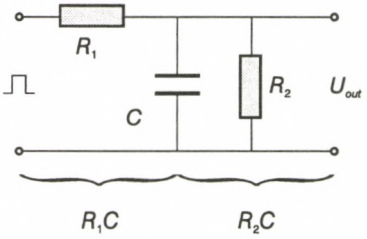


Fig. 6.14. Diagram of the ratemeter

13. Coupling elements. Coupling circuits. These elements connect the patient (or more generally the system to be investigated or influenced) with the electronic device; i.e. they ensure the flow of the signal energy in the required direction.

It has to be mentioned that the powers employed in influencing (therapy) are larger by several orders of magnitude than those reaching the electronic device from the signal source through the coupling element in the course of an examination (diagnostics). Figure 6.15 is similar to Fig. 6.1, the only difference being that the coupling elements are shown as blocks separate from the electronic device.

Various signals may occur between the coupling element and the examined system, but always electric energy is transmitted between the coupling element and the electronic device. In the following the role of the coupling elements will be demonstrated by a few examples, without the intention to be exhaustive.

Let us consider first the case depicted in Fig. 6.15*a*, where the processing of signals arriving from the patient is demonstrated. This signal may be some electric voltage, associated for instance with the functioning of some organ. The purpose of the coupling element in this case is the production of an electric contact between the patient and the signal-processing system. Such coupling elements are called *electrodes.* However, all electroanalytical auxiliary devices giving a voltage depending on the ionic milieu (e.g. ion-

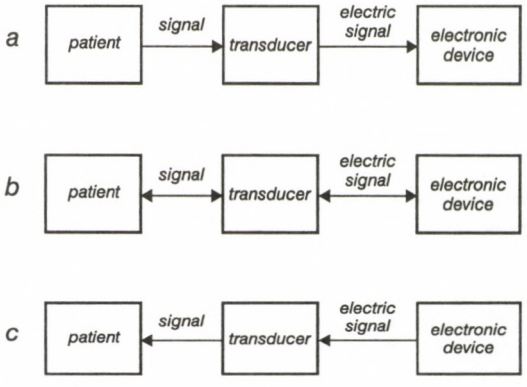


Fig. 6.15. Functions of the coupling elements

selective electrodes) are also called electrodes (e.g. pH electrode). Their application allows the fast, exact and easy determination of the concentrations of inorganic and organic substances in the body fluids. Such are also the electrodes, used in the monitoring of premature babies, which sense the O_2 and CO_2 content in the blood through the skin.

If the signal arriving from the examined system is not an electric one (e.g. the heart sound), the purpose of the coupling element is also to transform the signal into an electric one. (The coupling element of our example is the heart sound microphone.) The coupling elements performing this transformation are *signal transformers, detectors* or *transducers.*

There are several diagnostic procedures in which the patient's body cannot be considered the original, natural signal source: the body participates in the formation of the signal in an active or passive way. Such are all the in vivo radioisotope diagnostic methods the essential of which is the determination of the space distribution and time course of a radiopharmacon introduced in the body orally or intravenously. For these measurements *scintillation detectors* are used most frequently; their scintillating material is usually NaI(Tl), but other crystals (BaF_2, CsF, bismuth germanate [$Bi_4Ge_2O_{12}$]) are also applied. In the sense of the above definition, a detector should be conceived as a unit consisting of the scintillator and the photomultiplier, the output signal of which is a short voltage pulse for each γ-photon. It is worthwhile to mention here also the *semiconductor detectors* (e.g. Ge); in their material the γ-photons set free charge carriers in an amount proportional to their energy; these detectors are very useful energy selective devices for the measurement of the radioactive background irradiation (cf. Fig. 6.16).

Similar is the case in the X-ray diagnostic methods: the body is not "really" a signal source; the intensity of the X-ray passing through the body may be considered as signal which was influenced naturally by the tissue through which it passed. The detector of the first computer tomograph was still a NaI(Tl)+ photomultiplier, in the most modern CT equipments (cf. section 6.7.3) either luminescent materials combined with a photodiode or ionization detectors filled with Xe gas are applied.

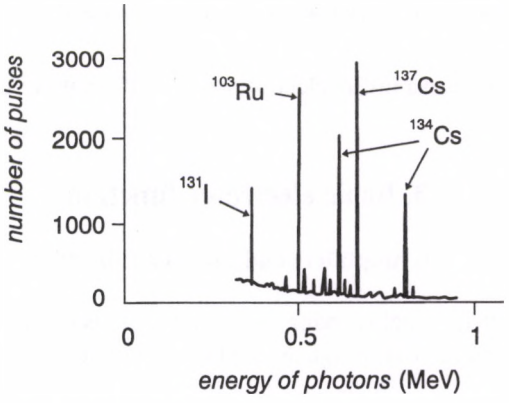


Fig. 6.16. Spectrum of the radioactive contamination recorded with a Ge detector (part of a measurement carried out in the open air in Germany in 1986, 30 days after the reactor accident in Chernobyl)

The detector used in the magnetic resonance imaging (MRI, cf. section 6.7.5) is a coil formed according to the body part to be projected in which radiofrequency voltage is induced as a sign of the relaxation.

As *light detectors* e.g. photodiodes or phototransistors may be applied. On the other hand, for infradiagnostics (thermography, cf. section 6.7.2) the photoconduction of such a semiconductor should be considered which has a very narrow forbidden band, since at body temperature the spectrum of the thermal radiation has its peak at about 10 μm, thus the energy of the photons is only about 0.12 eV. For the same reason, however, the semiconductor used as detector (e.g. indium antimonide or mercury-cadmium-telluride) should be kept at a low temperature during the measurement.

Figure 6.15*b* depicts a coupling element functioning in both directions. Such two-way transducers are used in ultrasound diagnostics (cf. section 6.7.6). By means of this transducer, electric energy of ultrasound frequency is transformed into ultrasound, which is radiated into the system under examination, and additionally the reflected ultrasound is retransformed into an electric signal and is transmitted into the processing electric device.

In the case shown in Fig. 6.15*c* the coupling element transmits energy to the patient. This energy transmission may also be achieved without energy transformation (e.g. in the case of electric excitation). The coupling element here too is called an electrode. An example of energy coupling with energy transformation is the earphone, which in audiometry transmits to the patient electric power of audiofrequency transduced into sound.

Coupling circuits are always electric networks consisting of connecting wires, ohmic resistances, capacitors, induction coils, etc., which ensure the connections between the functional units of an electronic system.

Similarly to coupling elements coupling circuits transmit signals, but these are always electric signals. Quite frequently *RC* circuits are used as coupling circuits, in which case we have *RC* or capacitive coupling. An advantage of *RC* coupling is the direct voltage separation between the connected blocks. Coupling circuits used for direct voltage signals may consist only of connecting wires and ohmic resistances, when the coupling is said to be direct or galvanic. The expression coupling circuit is often used with reference to the connection of the patient and the electronic devices if this connection is achieved with a network consisting of electrodes, electric wires and other connecting units.

The coupling circuits transmitting electric power to the patient are frequently termed *patient circuits.*

6.3. Basic electronic functions

6.3.1. Amplifiers and their amplification

The power of the electric signal carrying the information is usually not high enough for further processing. For this reason signals must be amplified by electronic devices called amplifiers.

1. The amplifier is a functional unit consisting of amplifying and other elements. The signal connected to its input appears on the output in an increased (amplified) form (Fig.

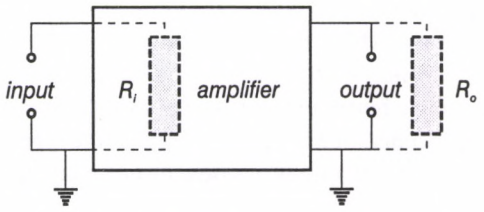


Fig. 6.17. Diagram relating to the amplifier

6.17). (One pole of both the input and the output is usually earthed to decrease the noise voltages originating from the surroundings of the amplifier.) The input of the amplifier is a load for the signal source; this is represented in the diagram by the input resistance (R_i). At the same time the amplifier is also a signal source for the next functional unit, which in turn is a load on the output of the amplifier (R_o in the diagram).

2. Gain. In a given case the gain of the amplifier may be characterized in various ways.

(a) The *power gain* (K_P) is the quotient of the output (i.e. amplified) and the input powers (P_o and P_i);

$$K_P = \frac{P_o}{P_i} \qquad [6.11a]$$

The power gain is frequently given in decibels:

$$K_P\,(\mathrm{dB}) = 10 \log K_P\,\mathrm{dB} \qquad [6.11b]$$

(b) Instead of power, the signal can be characterized by its voltage: therefore, the *voltage gain* (K_U) too is used, which is the quotient of the output and input signal voltages (U_o and U_i):

$$K_U = \frac{U_o}{U_i} \qquad [6.12]$$

(c) The relation between the voltage and the power gain is given by the simple relation

$$K_P = K_U^2 \frac{R_i}{R_o} \qquad [6.13a]$$

or, expressed in decibels:

$$K_P\,(\mathrm{dB}) = \left(20 \log K_U + 10 \log \frac{R_i}{R_o}\right) \mathrm{dB} \qquad [6.13b]$$

To express only the variation of the gain (in the regulation of gain) we may use the equation

$$K_P\,(\mathrm{dB}) = 20 \log K_U\,\mathrm{dB} \qquad [6.13c]$$

As an example, it may be mentioned that the gain required with an ECG apparatus amounts to 60 dB, which corresponds to a power gain of 10^6 or a voltage gain of 10^3.

3. Transfer characteristics. It is a basic condition in amplification that it must be frequency-independent in the whole frequency range. The meaning and importance of this condition become obvious if it is considered that the signal usually cannot be described by means of a single harmonic oscillation (a single sine wave). Aside from static cases, the time-dependent function $U(t)$ corresponds to the signal voltage: this contains the information. Each such signal, either periodic or not, may be obtained as the sum of harmonic components:

$$U(t) = \sum_{k=1}^{n} U_k \sin(\omega_k t + \varphi_k) \qquad [6.14]$$

where k represents integers. The first member of the sum ($k = 1$) is the so-called fundamental harmonic having the amplitude U_1 and the frequency ν_1 ($\omega_1 = 2\pi\nu_1$) which is the frequency of the original – periodic or approximately periodic – signal. The frequencies of all the other components (the so-called overtones: $k > 1$) are the integral multiples (k-folds) of the fundamental harmonic. The above sum is the *Fourier spectrum* of $U(t)$. (The phase angle φ_k of the individual components is either zero or $\pi/2$ (90°), the individual amplitudes U_k are either negative or positive, therefore the Fourier spectrum of the signal $U(t)$ consists in fact of positive and negative sinusoidal as well as positive and negative cosinoidal members.) Hence, it is not sufficient to characterize a signal by only one frequency; instead, a series of frequencies or a frequency interval (frequency band) is required. This series of frequencies, or band, is called the total frequency range of a signal (cf. Table 7.4). If the condition mentioned at the beginning of this point were not satisfied, the various frequency components of the signal would be amplified to various degrees and, as a consequence, the shape of the signal would become distorted. The frequency-dependent description of the amplification of the amplifier gives the *transfer characteristics* (Fig. 6.18). The frequency range within which the gain is independent of the frequency (or more exactly where the frequency dependence remains below 3 dB) is the *transfer band.* Figure 6.18*a* shows the characteristics of an amplifier which can be used in the *range* between the *low-* and *high-frequency limits* ν_l and ν_h, while Fig. 6.18*b* relates to *a direct voltage amplifier,* whose low-frequency limit is zero Hz. The condition mentioned in the introduction can also be stated in that the transfer band of the amplifier must cover the frequency range of the signal. For instance, the action voltage of the heart is a nearly periodic signal with a base frequency of 1.1–1.3 Hz, whose undistorted processing requires a transfer band between 0.1 Hz and 100 Hz.

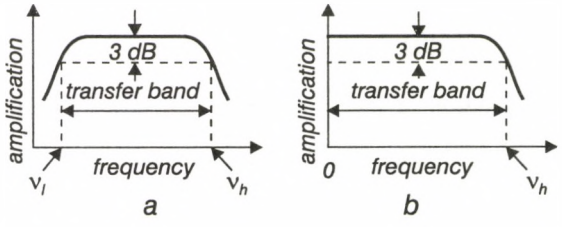


Fig. 6.18. Transfer characteristics of the amplifier

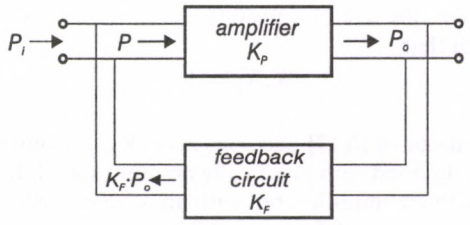


Fig. 6.19. Amplifier with feedback

4. Feedback amplifier. A part of the amplified signal power is frequently fed back by a feedback circuit to the input of the amplifier (Fig. 6.19). The feedback factor K_F is defined as the quotient of the back-fed and output powers. Its numerical value is considerably smaller than 1, so that the back-fed signal is comparable with, or may be even smaller than the input signal. The input signal P_i to be amplified and the back-fed signal $K_F P_o$ are added, which means that the amplifier actually amplifies the resultant power

$$P = P_i + K_F P_o \qquad [6.15]$$

Thus, the power of the output signal can be rewritten as

$$P_o = K_P P \qquad [6.16]$$

where K_P denotes the power amplification of the amplifier without feedback.

The result of the signal summation (interference) on the input depends upon the phase relations of the interfering signals. Of the many possible cases, we shall discuss here only two practically important extreme cases, i.e. when signals of identical or opposite phases are added. In the case of identical phases, the feedback factor is considered to be a positive quantity ($K_F > 0$); this is a *positive feedback.* In the opposite case, the feedback is regarded as negative ($K_F < 0$), and this is a *negative feedback.* Accordingly the power gain (K_{PF}) of the feedback amplifier, defined by the quotient P_o/P_i, is

$$K_{PF} = \frac{K_P}{1 - K_F K_P} \qquad [6.17]$$

In positive and negative feedbacks, the denominator of [6.17] is smaller or larger, respectively, than 1. Thus, in the first case $K_{PF} > K_P$, and in the second one $K_{PF} < K_P$.

A positive feedback increases the power amplification, though this is not necessarily advantageous. Among its drawbacks is the increase of the electronic noises generated in the amplifier (due e.g. to the statistical fluctuations of the flow of the charge carriers), and the distortion of the amplifier usually increases.

With a negative feedback the situation is reversed: the distortion of the amplifier decreases, the transfer characteristics is smoother, etc. A negative feedback is generally used to improve or modify the properties of the amplifier.

However, in some cases the use of a positive feedback may also be of advantage. Let us select a K_F value such that

$$K_F K_P = 1 \qquad [6.18]$$

In this case the denominator of [6.17] becomes zero, K_{PF} becomes infinite, which means that an output signal is obtained with practically no input signal. In this case, however, we are no longer dealing with an amplifier, but with an *oscillator,* which in practice is used as a special electronic energy source (cf. section 6.3.3).

5. The differential amplifier has two inputs and one output (Fig. 6.20). The output signal voltage is proportional to the difference between the two input voltages:

$$U_o = K_U (U_{i1} - U_{i2}) \qquad [6.19]$$

where K_U is the voltage gain. The gain can be expressed in decibels as well. The use of differential amplifiers renders possible the enhancement of the signal-to-noise ratio in cases when noise signals of common mode (identical shape, frequency, amplitude, phase) are superimposed on the signals to be amplified.

If the noise signals on the two inputs of the ideal differential amplifier were to be completely of common mode, their difference would be always zero, and the amplified output signal would not contain at all a component from this noise voltage. In practice this can never be completely fulfilled, the differential amplifier damps the noise signals of common mode to a great extent only, thus the output voltage (power) of the noise signal is much lower than the voltage (power) of input noise signals on the two inputs. The extent of damping is usually given in dB and the method is called *common mode noise suppression* or *common mode rejection.*

The differential amplifier can be used with good effect to measure voltages of biological origin, and accordingly these amplifiers are also called *bioamplifiers.* Consider for instance an ECG diagram. Since this is usually recorded in the disturbing electric field of the electric mains network, the electrodes on the patient transmit not only the useful signal, but also disturbing voltages from the mains. This noise appears as a common mode signal for every electrode on the patient. Thus, if U_{i1} is the voltage obtained on the left arm, and U_{i2} is that on the right arm, the signal measured at the output of the amplifier contains practically no noise from the mains network. A voltage of biological origin may also be a noise signal: for instance, on the electrodes placed on the skull to measure the action potentials associated with the functions of the brain one can observe not only the

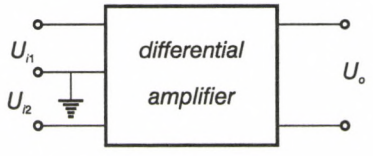


Fig. 6.20. Diagram of the differential amplifier

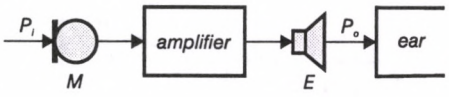


Fig. 6.21. Diagram relating to the hearing-aid device

useful signal, but also the action voltage of the heart, which appears on the EEG electrodes as a common mode signal.

6. Examples of the application of amplifiers. In the previous section we have frequently referred to medical applications. In this section two applications are described.

(a) Figure 6.21 depicts the basic construction of a *hearing-aid device.* A miniaturized microphone (M), as a transducer, detects the input sound power P_i. The amplified audiofrequency electric power is transformed by an equally small earphone (E) into acoustic power (P_o), which is irradiated into the ear of the patient. The hearing-aid operates properly only when the value of the acoustic amplification as defined by the ratio P_o/P_i is the same as the hearing loss. Of course, the electric amplification must be larger than the acoustic amplification, since the transformation losses in the microphone and in the earphone must be accounted for.

(b) Another example is shown from the field of *radiation dosimetry* (Fig. 6.22; cf. also section 3.3). Depending upon whether the dose exposure or the dose exposure rate is to be measured, a capacitor C or a resistor R is connected into the measuring circuit. The charge released in the measuring chamber either charges the capacitor to a degree proportional to the exposure, or a current with intensity proportional to the exposure rate will flow through the resistor. Consequently, the voltage over the capacitor or the resistor will be an analogue signal of the exposure or exposure rate; after amplification, this will be shown by the voltmeter or recorded by the recorder.

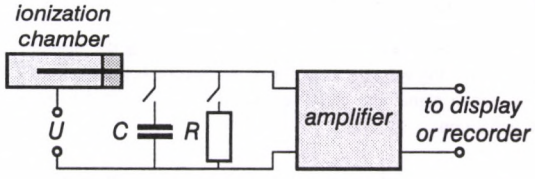


Fig. 6.22. Diagram relating to the measurement of dose and dose rate with an ionization chamber

6.3.2. Displays and recorders

The function of the last unit of the signal-processing systems is usually to display the signal for immediate use or to record it. This may mean the display of the instantaneous or steady signal value, or further characteristics and parameters of diagnostic importance derived from signal processing, but it may also result in a two-dimensional image or a diagram demonstrating the course of a process in time. In the following section some of the more frequent methods of the display and the recording of two-dimensional images and time processes will be discussed.

1. The cathode-ray tube is an electronic device used for displaying time processes and alphanumerical signs (letters, numbers, etc.) graphically and for producing two-dimensional images (Fig. 6.23). Its well-known applications are the displays of computers and

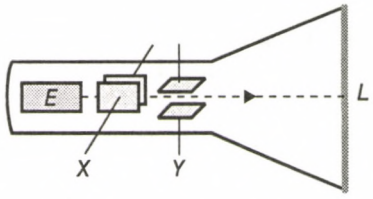


Fig. 6.23. Diagram of the cathode-ray tube

the television tubes. Similarly to TV image tubes, the cathode-ray tube consists of an electron source (hot cathode) and an electrode system (E) allowing the production of a narrow electron beam (cathode-ray) whose intensity can be regulated. After passing two deflecting electrode pairs (X and Y), the beam finally arrives at a luminescent screen (L). The screen luminesces in response to the arriving electrons, thereby indicating the spot of the incident electron beam. With the aid of the deflecting voltage applied to the plate pairs, the cathode-ray can be deflected to any point of the screen. (This deflection can be accomplished not only with electric but also with magnetic fields.)

a) *Time processes* can be displayed by applying a saw-tooth voltage to the deflecting plate pair X. As a result (during the rise of the voltage) the cathode-ray travels across the screen, which represents the time axis. The voltage signal of the studied time process is superimposed on this motion. This latter signal is applied to the deflecting plate pair Y.

A characteristic date of the cathode-ray tube is the *sensitivity,* which is the displacement of the electron beam on the screen relative to unit deflecting voltage; its order of magnitude is usually mm/V. If the sensitivity is known, or after comparison with a known signal the cathode-ray tube may be used to *measure* signal voltages, signal amplitudes, time intervals, and so on.

b) On the screen of the cathode-ray tube the two-dimensional image is built up from spot rows in the following way (Fig. 6.24). For imaging saw-tooth voltages, with frequencies shown in the figure, are connected to both deflecting systems X and Y. In fact, the deflecting voltage in the Y direction does not change continuously: its size remains

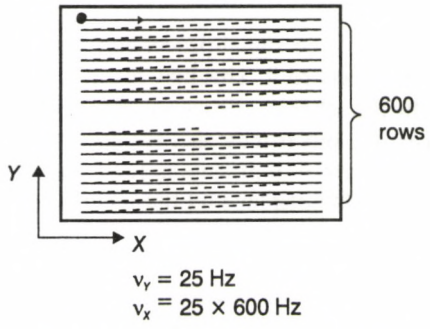


Fig. 6.24. Scanning motion of a cathode-ray

constant during a single run of the cathode-ray in the X direction, after which it changes abruptly to a value corresponding to the height of the next row. From the frequency conditions of the saw-tooth voltages it appears that the cathode-ray would cover the screen 25 times in every second with 600 horizontal lines if its intensity would be constant. However, the intensity of the cathode ray changes while it moves from spot to spot on the screen: the elements of the image drawn during 0.04 s differ from each other in brightness. These two-dimensional images are also called *B images* (*B* for brightness).

The data of Fig. 6.24 correspond to images consisting of 600 spot rows and a frequency of 25 images/s; the latter ensures an image without flashing and is also enough for the displayed movements to appear as continuous.

The intensity of the cathode ray may be changed in two or more steps. In the first case the spots may have two kinds of brightness: dark and bright, this is the *bistable B image*; if, on the other hand, the intensity of the cathode ray allows for more brightness grades, *tonal, grey-scale B images* are obtained.

Of the numerous applications of the two-dimensional display tube, we wish to mention here only the scanning electron microscope (cf. section 4.2.4), in which the deflection of the cathode-ray scanning the object and that of the cathode-ray of the image tube are controlled by the same saw-tooth wave generators.

Applying computers or microprocessors (cf. section 8.3) the image elements can be stored in the memory. In such cases the storage time can be arbitrarily long and on the display besides the line diagrams and two-dimensional images, alphanumeric characters can be displayed too. The two-dimensional images are always tonal, while the graphs, alphanumeric signs and texts consist of bistable spots. (The above refer to the black-and-white television tube; the description of the coloured variety is beyond the scopes of our discussion.)

2. Liquid crystal displays (LCD) are also frequently used for displaying alphanumeric signs and graphs. Here the liquid crystal is placed between two glass plates and the electro-optical phenomenon takes place in the electric field between the transparent electrodes put on the glass plates (cf. section 1.4.4). In digital displays (e.g. in digital clocks) the system of the electrodes is relatively simple (Fig. 6.25*a*). For more demanding displays the electrodes are thin parallel lines placed perpendicularly to each other on the two glass plates (Fig. 6.25*b*); with their help the electro-optical phenomenon may be elicited in each point of the display, and alphanumeric signs and graphs may be produced from bistable spots. The simple LCDs use only the reflected light, while the more sophisticated ones have their own adjustable bias illumination.

3. Recorders. Computer technology plays an ever increasing role in the functioning of the diagnostic signal-processing systems. The processed signals are stored in the memory of the computer from where they can be recalled onto the screen *at any time,* and, in the presence of a picture archiving and communication system (= PACS, cf. section 6.6.4) even *at any point of the system.* However, it may be required that the signal be available also in a recorded form (hard copy).

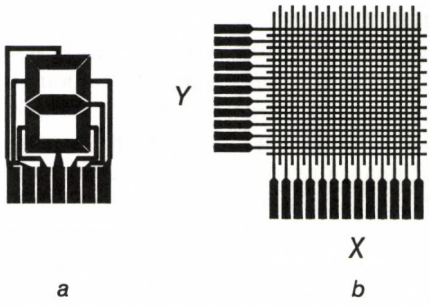


Fig. 6.25. Arrangement of the electrodes in liquid crystal displays in a so-called seven-segment numeric display *(a)* and in a matrix display *(b)*. (The *X* electrodes of the matrix display are on one of the glass plates, while the *Y* electrodes on the other; for technical reasons every second electrode exits on the opposite edge of the glass plate)

a) For the graphic recording of *processes in time* various devices present themselves, which are usually similar to the printer of the computer. The form of the recording paper is traditional (e.g. for ECG recorders a tape with a width 60, 90 or 110 mm), but as opposed to the traditional recordings the seemingly continuous curves usually consist of dots (e.g. 8 dots/mm) and in addition to the curves other informations, e.g. – sticking to the example of ECG, as important data for the diagnosis – amplitudes and times also appear on the tape.

b) The recording of *two-dimensional images* is made more and more also by printers, including the possibility of colour printing too. The colours are usually not real colours. In such cases the recording contains information concerning the meaning of the colours, e.g. on the ultrasound-diagnostic recordings the various colours represent various flow velocities (cf. section 6.7.6).

The recordings of the digital (computerized) X-ray diagnostics fulfil so to say traditional demands: a laser printer exposes the film (from dot to dot), which is then available for the physician in a form similar to the traditional X-ray pictures; the CT pictures are also made like this.

6.3.3. Electronic energy sources

Functional units transforming part of the input electric power into electric power with parameters required for special purpose are frequently found in various electronic devices, when for instance the transformation of direct current into alternating current is desired, or vice versa. Similarly, it may also be necessary to transform a lower voltage into a higher one, etc.

1. The sine-wave oscillator produces a voltage varying sinusoidally ($U = U_0 \sin 2\pi\nu t$). This may be achieved simply with an amplifier having positive feedback, whose output is loaded with an *LC* circuit, as depicted in Fig. 6.26. (The feedback is ensured by an

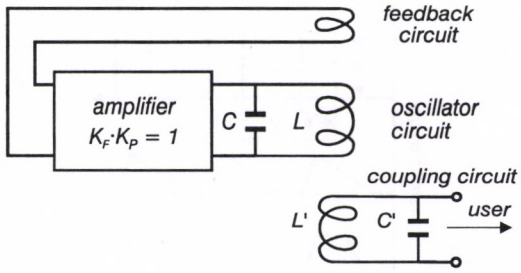


Fig. 6.26 Diagram relating to the sine oscillator

induction coil.) The oscillation begins when the condition $K_F\ K_P = 1$ is satisfied (cf. section 6.3.1). The frequency of the electromagnetic oscillation produced is equal to the resonance frequency of the LC circuit (ν_0; cf. section 6.2), i.e.

$$\nu_0 = \frac{1}{2\pi\sqrt{LC}} \qquad [6.20]$$

ν_0 may be changed according to purpose by a suitable change of C and/or L. The electric energy produced is transmitted to its site of application by an $L'C'$ coupling circuit. The optimum energy coupling can be ensured by the resonance condition

$$LC = L'C'$$

The frequencies used in audiometry are in the range of audible sound (20 Hz to 20 kHz), while those in ultrasonic are higher than 20 kHz. A common feature of all these methods is the transformation of electric power of appropriate frequency into mechanical vibration power. On the other hand, high-frequency heat therapy and surgery utilize radiofrequency range electric power ($\nu > 10^5$ Hz) without any transformation.

2. Pulse generators produce voltage or current pulses and pulse series of a given polarity. The shapes of the pulses (their time course) may differ. In some cases (for instance in determination of the stimulus characteristics) the shape of the pulse is important, but in other cases (e.g. in ventricular defibrillation) it is of minor importance. Figure 6.27 depicts some characteristic pulse shapes; of these, the simplest and most frequently used is the *square-wave pulse.* Its characteristics are the pulse duration time (τ), the period (T), the frequency ($\nu = 1/T$) and the amplitude (a). Pulse generators involve special electronic circuits. In the following, only square-wave generators are dealt with.

Individual square pulses can be generated by a *monostable multivibrator, monoflop* (Fig. 6.28*a*). Besides the stable state, this generator also has an activated (quasi-stable) state, which is produced by a suitable voltage (voltage pulse) applied to the input. The lifetime of this activated state depends upon an RC circuit, and it can therefore be expressed by the time constant $\tau = RC$. After this time the stable state is restored. The output voltage has two values: U_1 in the stable state, and U_2 in the activated state (Fig. 6.28*b*). The

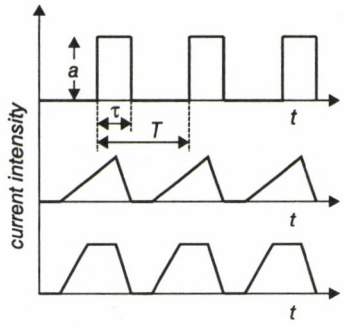


Fig. 6.27. Current pulses of various shapes

monostable multivibrator responds to every activating pulse with a square-wave pulse, the duration of which is determined by the time constant RC, and whose amplitude is equal to the voltage difference $U_2 - U_1$. This means that this functional unit may also operate as a pulse-shaping device. This property is made use of in ratemeters (cf. section 6.2).

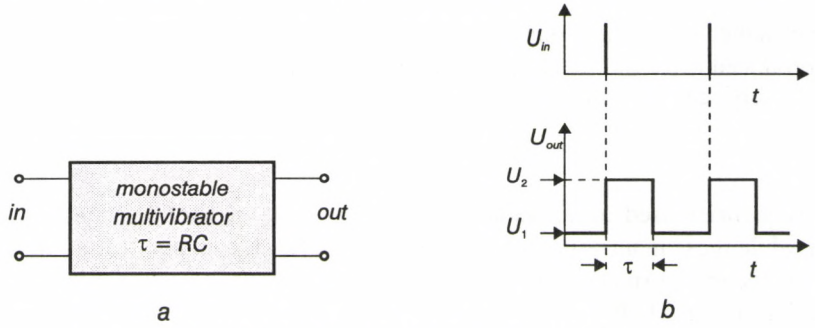


Fig. 6.28. Diagram relating to a monostable multivibrator

If a periodic signal source is added to the monostable multivibrator, a device is obtained whose output releases a series of square-wave pulses. The period is identical with the period of the generating signals. This new functional unit is called an *astable multivibrator,* which naturally operates as a square-wave generator (Fig. 6.29). As the simplest procedure, two monostable multivibrators are connected with each other so that the return of the first multivibrator into the stable state activates the other one, and vice versa. The amplitude of the square waves of the astable multivibrator in this case is again $U_2 - U_1$, the two periodically changing pulse durations are τ_1 and τ_2, respectively, and the period is $\tau_1 + \tau_2$.

The direct applications of pulse generators will be dealt with in section 6.5. Here an indirect application in connection with the use of sine-wave generators (for instance in ultrasound diagnostics, or high-frequency surgery) will be discussed. In these cases the power produced in sine-wave generators is used intermittently; this is pulse mode operation. The special switching circuit (gate circuit) regulating the energy output is controlled by a pulse generator.

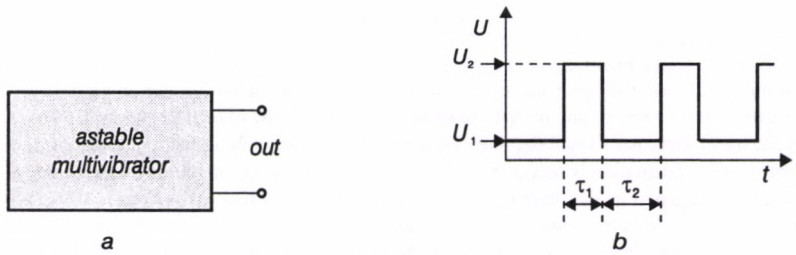


Fig. 6.29. Diagram relating to an astable multivibrator

3. The energy supply of the electronic devices can be obtained in two ways: from the electric mains network or from a built-in disposable or rechargeable battery. Each solution has advantages over the other. For instance, the mains network supply has practically neither time limits expressed in operating time, nor power limits of operation, whereas the battery supply has no restrictions as to location. In the case of a mains network supply the alternating voltage can be transformed by transformators into practically any voltage required, whereas the usually small voltage means that the battery supply is advantageous from the point of view of safety. In some (e.g. hearing-aids, electromechanical limb prosthesis), only battery supplies can be considered.

If the power requirement of an electronic device is known (voltage, current, power, time of operation without battery change or recharging), the planning of a built-in power supply is essentially reduced to the choice between a disposable or rechargeable battery, and often means a compromise between the operating time and the portability. In the case of a pacemaker introduced under the skin of the patient for instance, a considerable part of the volume and mass of the device is due to its battery, but this ensures operation over several years.

The electronic devices usually require a direct current power supply. Consequently, if an alternating current supplies the power, the required direct power is obtained with a unit consisting of a transformer *T*, a rectifier *R* and a filter circuit *F* (Fig. 6.30).

In the simplest case the filter circuit is a capacitor which is charged to peak voltage through the rectifier; this voltage feeds the consumer. The resistance of the comsumer and the capacitor form an *RC* circuit; if the time constant of the latter is large enough (in relation to the period time of the charging voltage), the feeding voltage is satisfactorily constant.

In some cases the operation of a unit requires a voltage of several hundred or several thousand volts (e.g. the acceleration of cathode-rays). Both in mains and battery operation, this high direct voltage is usually obtained by producing an alternating power of a few kHz with a sine oscillator; this is then transformed to the required high voltage, rectified and finally filtered. This method is similar in essence to the operation of the system depicted in Fig. 6.30, with the single modification that the primary coil of the transformer is supplied by a sine oscillator.

The *acceleration voltage supply of the X-ray tubes* deserves special attention. In the X-ray diagnostics – and therapy, too – *the application of homogeneous X-rays* (photons having the same energy) *is desirable.* However, in the spectrum of the tungsten anode X-ray tubes used almost exclusively in medicine the *Bremsstrahlung dominates with a continuous spectrum*, the shortwave cutoff of which depends on the acceleration voltage (cf. section 2.9. and 2.11).

In the simplest, so-called half-wave X-ray apparatuses the X-ray tube itself rectifies the alternating high voltage, thus it works only in the positive half-periods for the anode. During the working half-period the acceleration voltage changes between zero and maximum, and meanwhile, in the most part of the half-period, the critical wavelength is also longer (i.e. the radiation is softer) than desirable.

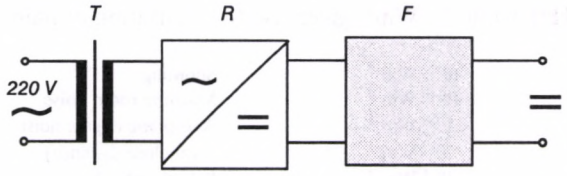


Fig. 6.30. Mains power supply

The best from among the solutions better than the above (corresponding also to the international recommendations) is the *high-frequency X-ray generator.* This produces, as a first step, direct voltage from the mains alternating voltage of 50 Hz to feed a high-frequency (40–80 kHz) sine oscillator. The high-frequency alternating voltage can be transformed to the wanted high-voltage by means of a relatively small transformer; this is followed by a rectifier and filter capacitor (in a way similar to that shown in Fig. 6.30). The use of high frequency has several advantages. Not only the transformer has a small size, but also the capacitor, since only a small capacitance is required: in case of 1 nF the voltage fluctuation is only about 1% even at a tube current of 100 mA. Due to the small sizes the complete high-voltage part may be built in the same bulb as the X-ray tube.

A simpler solution (corresponding to Fig. 6.30) uses the mains voltage with a relatively large filter capacitor: for example, at a capacitance of 1 μF the voltage of the capacitor decreases by 10% (e.g. from 100 kV to 90 kV) during an exposition of 10 mAs.

6.4. Applications of sine-wave generators

6.4.1. The physical basis of audiometry

According to its mechanism of propagation, the sound is a longitudinal pressure wave. This statement means (e.g. in the case of propagation in the air) that the gas molecules perform a back and forth oscillatory motion in the direction of the propagation, consequently ranges with large and small molecular concentrations (with high and low pressures) follow each other at a half-wavelength distance. Both the molecular concentration and the pressure proportional to it fluctuate around the actual atmospheric value. The pressure fluctuation is called alternating sound pressure, in short sound pressure, its unit is the pascal (Pa). In case of its propagation in media which may be considered as incompressible (e.g. the propagation of sound in the endolymph or the propagation of ultrasound in body tissues) the fluctuation of the concentration may be disregarded, only the sound pressure has to be considered.

Naturally, in the sound energy is propagated, therefore, as one possibility, the *strength of a sound stimulus* is characterized by the intensity of the sound waves, called the sound intensity (objective sound intensity). This must not be confused with the loudness (the subjective sound intensity) characterizing the intensity of *sound sensation.* Audiometry studies the relation between the objective and subjective sound intensities.

1. The characterization of sound intensities with the decibel scale. The human ear is receptive to sound stimuli over an extremely wide intensity range. The table below summarizes the stimuli required in the case of a sinusoidal pure sound of 1000 Hz frequency producing sound sensations of various intensities. The data refer to average values for healthy individuals. The auditory threshold is the lowest audible intensity (at a frequency of 1000 Hz), while 10 Wm^{-2} gives rise to a sensation of pain:

Auditory threshold	10^{-12} Wm^{-2}	Shouting	10^{-4} Wm^{-2}
Whisper	10^{-10} Wm^{-2}	Machine room noise	10^{-3} Wm^{-2}
Low-tone conversation	10^{-8} Wm^{-2}	Aeroplane engine noise	
Normal conversation	10^{-7} Wm^{-2}	(at close distance)	1 Wm^{-2}
Urban street noise	10^{-5} Wm^{-2}	Pain threshold	10 Wm^{-2}

In practice usually the concept of relative stimulus intensities is used; this is defined as the ratio of the stimulus intensity to the stimulus threshold intensity. With regard to the very wide range it is convenient to use the logarithms of the relative values, or more exactly to characterize the stimulus intensities by the corresponding decibel values (n). Thus, we have

$$n = 10 \log \frac{I}{I_0} \text{ dB} \qquad [6.21]$$

where I is the intensity level of the sound studied and I_0 is that of the stimulus threshold. In the present case $I_0 = 10^{-12}$ Wm^{-2}. According to the table the intensity level at 1000 Hz corresponding to a low-tone conversation is 40 dB higher, loud shouting is 80 dB higher, and the intensity level corresponding to the pain threshold is 130 dB higher than the stimulus threshold level. The overall intensity range which may be referred to as sound stimulus at 1000 Hz is characterized by a scale ranging from zero dB to 130 dB. For information: 34 dB correspond to 1 mPa sound pressure, 94 dB to 1 Pa, respectively.

2. The phon scale. The human ear is sensitive to various degrees to pure sounds of various frequencies. Figure 6.31 provides some data. The lowest curve depicts the frequency dependence of the auditory threshold. The other curves demonstrate for various frequencies the intensity values which produce sound sensations corresponding to a low-toned conversation, shouting or an unpleasant engine noise. A decibel scale can also be fixed to the intensity scale of Fig. 6.31 in such a way that the intensity belonging to the auditory threshold at 1000 Hz is chosen as reference. So the points of the curves – equal loudness curves – belonging to 1000 Hz are 0, 40, 80 and 120 dB, respectively. Accept these numbers along the whole length of each curve as characteristic data for equal sensation: they express the so-called *phon loudness.*

If, *judged by hearing,* a sound produces in us the same sound sensation as the 1000 Hz pure sound of intensity I, then (by agreement) the phon loudness (H_{ph}) of this sound will be:

$$H_{ph} = 10 \log \frac{I}{I_0} \text{ phon} \qquad [6.22]$$

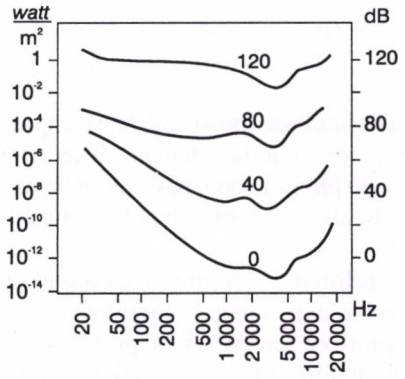


Fig. 6.31. Equal loudness curves of the ear

The table below gives information about the approximative phon loudnesses of sound sensations of various intensities:

Auditory threshold	0 phon	Shouting	80 phon
Whisper	20 phon	Machine room noise	90 phon
Low-tone conversation	40 phon	Aeroplane engine noise	
Normal conversation	50 phon	(at close distance)	120 phon
Urban street noise	70 phon	Pain threshold	130 phon

3. The sone scale. The previously discussed phon scale is based on the *Weber–Fechner psychophysical law* that was considered valid for more than a century. According to this law the intensity of sensation is proportional to the *logarithm* of the relative intensity of the stimulus; this is expressed by [6.22] too. The investigations carried out on the relation of stimuli and sensation disproved the validity of the Weber–Fechner law and it was replaced by the *Stevens psychophysical law* corresponding better to experience. According to this law the intensity of sensation is proportional to the *fractional power* of the relative intensity of the stimulus. More exactly, for sound of 1000 Hz frequency at the intensities above 10^{-8} Wm^{-2} (i.e. above 40 dB or 40 phon) the *loudness level in sones* (H_s) can be calculated with sufficient accuracy by the relation

$$H_s = \left[\frac{I}{I_0}\right]^{0.3} \qquad [6.23]$$

Here $I_0 = 10^{-8}$ Wm^{-2} and I is the intensity of the sound of 1000 Hz frequency, which causes the same sound sensation as the investigated sound.

Naturally the sone loudness of any sound can be determined, not only that of the sound of 1000 Hz and of sinusoidal sounds: let us select the intensity the sound of 1000 Hz frequency so that the *intensity of the sensation* evoked by it will be the same as caused by the sound to be determined. In this case the value calculated by [6.23] gives also the sone loudness of the sound in question.

The loudness values in phons and sones correspond to each other with a good approximation in the following way:

phon	30	40	50	60	70	80	90	100
sone	0.5	1	2	4	8	16	32	64

Thus it holds for a broad loudness range (at least approximately) that if the phon number increases by 10, the sone number doubles. This relation applies for the whole *sound frequency range* above 40 phon–1 sone, but cannot be used below 30 phon–0.5 sone. However, this range of low loudness has only slight practical significance.

4. Harmful effects of noise. Intensive or prolonged sound causes not only indisposition and disturbs human activity, but depending on conditions – on top of various nervous and somatic symptoms – may produce temporary or prolonged loss of hearing and in consequence of the irreversible injuries of the internal ear even permanent impairment of hearing. The sources of unpleasant or even harmful noises are mainly the vehicles in

traffic and certain industrial and agricultural machines, but more and more noise source can be found among the machines used in housekeeping and the kitchen, and even the electroacoustic devices serving originally for entertainment are well-known noise sources. The purpose of various labour-safety and environment-protection regulations aim to prevent or moderate health impairment and to protect our surrounding against noise.

The various recommendations and regulations allow e.g. for intellectual work a noise level of 35–50 dB, in noisy factories the allowed upper limit is 90 dB, but at about 100 dB already a fast temporary loss of hearing may occur.

Obviously, the harmful effects of noise and the protection against it are not restricted to audible sounds but extend to the non-audible low-frequency infrasound range and to the ultrasound range as well. People working with certain tools (e.g. pneumatic hammer, chain saw) suffer beside the considerable audible noise exposure grave infrasound (vibrational) damage as well.

In the various frequency bands different parameters are used in the formulation of limits, e.g. the amplitude of vibrations or the ensuing acceleration, in the ultrasound range the power density. Finally it should be emphasized that the *duration* of the stimulus plays an essential role in each range.

5. Audiometry. The above data relating to healthy persons are averages, from which significant individual deviations can be found. These may have of course also pathological reasons. One method and task of audiometry is the cognition of the actual auditory threshold curve and the establishment of the difference between the actual and the average (normal) auditory threshold curve. The curve of the decrease in hearing as a function of frequency is an *audiogram,* and the device taking the audiogram is the *audiometer.* As concerns its construction, the audiometer is a sine oscillator whose output is connected with an earphone, which transforms the electric signals into mechanical vibrations. The oscillator frequencies can be varied in the frequency range 20–20,000 Hz, and the intensity can be varied at each frequency. The intensity at the measuring frequency must be increased until the individual indicates sound perception. The deviation from the normal auditory threshold can be read directly in decibel (dB) units on the intensity scale. This method, *threshold audiometry*, requires the patient's cooperation, therefore it is naturally burdened by a subjective source of error.

There are methods which are free from the mentioned sources of error, they do not even require the cooperativity of the patients. In the following two objective audiometric methods will be described which yield information concerning also the *causes of auditory disturbances.*

a) The *acoustic-impedance audiometry* examines the acoustic, i.e. mechanic reactivity of the middle ear (eardrum – auditory ossicles – oval window) to sound pressure. The acoustic impedance mentioned in the name of the method, characterized by the ratio of the sound pressure and the velocity of the oscillatory motion evoked by it, expresses in fact the resistance of the system against the oscillatory motion forced upon it; its reciprocal, the so-called admittance reflects the mobility of the middle ear.

During the examination the external ear is closed by a plug which has three holes:

– through one of them e.g. a 200-Hz sound of a small loudspeaker is transmitted into the closed auditory canal where thus a sound level of about 80 dB is established;

– through the second hole a small microphone senses the pressure developing in the middle ear and keeps it at a constant level by means of feed-back; the loudspeaker voltage necessary for it is proportional to the admittance; this can be measured;

– through the third hole the pressure of the middle ear can be adjusted to a level higher or lower than the atmospheric pressure; by this actually the eardrum is tightened outward or inward from its state at rest: with its help the admittance can be measured also in its relation to the pressure (this is the so-called tympanometry).

The possibilities of impedance audiometry will be illustrated with only two examples. The first example: if the eardrum is perforated, the admittance does not depend on the pressure. The second one: the increased value of impedance refers e.g. to a more rigid eardrum.

b) The *examination of the evoked potentials* (ERA: evoked response analysis) is aimed at the converter function of the inner ear (cf. section 7.5.2) and nerve conduction. The action potentials of the auditory nerve are evoked by short-term (200–300 ms) sound pulses of several kHz, the measuring electrodes are placed on the top of the head and on the petrous bone. By the computerized averaging of the reactions to several hundred stimulations a curve with a duration of about 10 ms (corresponding to the propagation from the inner ear to the auditory centre) is produced on which the maxima represent the generation of the auditory pulses and their regenerations at the sites of transmission.

6.4.2. Ultrasound

1. The ultrasound generator. This device basically consists of a sine oscillator and a transducer. The sine oscillator generates high-frequency (more than 20 kHz) electric power, while the transducer unit converts the electric oscillations into mechanical ones.

The electroacoustic transducers used in the ultrasound range operate on the basis of various phenomena.

a) *Piezoelectric ultrasound generation.* If pressure is applied to the surface of appropriately cut plates or discs of certain monocrystals (e.g. quartz, ethylene diamine tartrate, Rochelle salt), electric charges are generated. This is the *piezoelectric effect.* The effect is reversible: if electrodes are placed on the crystal plate and a potential difference is applied to the electrodes, the plate will be deformed by the electric field (inverse piezoelectric effect). In an alternating electric field the size (thickness) of the crystalline plate follows the variation of the electric field, i.e. the crystalline plate vibrates. Resonance occurs whenever the frequency of the alternating voltage agrees with the eigenfrequency of the plate. In order to obtain an intensive oscillation, the plate is cut to dimensions according to the frequency of the ultrasound, and the plate is excited by electric oscillations corresponding to the eigenfrequency.

b) In response to an electric field, a phenomenon similar to the direct or inverse piezoelectric effect occurs in some polycrystalline ceramic insulators. This phenomenon is called *electrostriction,* which can similarly be used to generate ultrasound. Piezoelectric

and electrostriction transducers may be used to produce ultrasound or to retransform ultrasound into electric signals.

c) The *magnetostrictive effect* is similar to the previous one concerning its result, but here the dimensional change takes place in a ferromagnetic material (e.g. iron or nickel) under the influence of a magnetic field. Thus the magnetostrictive ultrasound source is a rod made of a ferromagnetic material enclosed by a coil; in the coil alternating current of ultrasound frequency flows, in the magnetic field of which the length of the rod changes periodically according to the frequency.

2. The propagation of ultrasound. The velocity of propagation depends upon the frequency to only a small degree, and consequently ultrasound propagates with the same velocity as audible sound in the various substances. Sound is reflected at the boundary of media of different acoustic impedance (i.e. the product of the density and velocity of propagation). For normal incidence the *reflectivity* (R; cf. section 2.3.1, point 3) can be calculated via the relation

$$R = \left(\frac{\rho_1 v_1 - \rho_2 v_2}{\rho_1 v_1 + \rho_2 v_2}\right)^2 \qquad [6.24]$$

where ρ_1 and ρ_2 are the densities of the two media, and v_1 and v_2 the velocities of propagation in these media. For liquids and solid media the acoustic impedance is generally considerably larger than for gases, and for this reason $R \approx 1$ at liquid–gas and solid–gas boundaries, i.e. most of the sound energy is reflected. Sound energy can be transmitted between bodies of higher density in air by placing some medium of nearly identical density, a coupling medium, between them. If, for instance, ultrasound is to be transmitted into the tissues, the air layer between the irradiating head and the body should be filled with water or contact gel. As a result of absorption and scattering, sound intensity decreases. For a parallel beam this decrease can be described by an exponential function. The attenuation is generally larger at shorter wavelengths, and hence ultrasound dies away faster than audible sound. At 10 kHz, for instance, the half-value thickness for air is approximately 100 m, while that for water is approximately 100 km. At 1 MHz the corresponding values are approximately 1 cm for air, a few meters for water, 2 cm for muscle and only a few mm for bone.

3. Some medical applications of ultrasound

a) The *therapeutical applications* are based naturally on the physical characteristics of the ultrasound, concretely: the ultrasound is a *longitudinal pressure wave* with a frequency higher than 20 kHz in which *energy is propagated.* For therapeutic purposes usually ultrasounds with a frequency of 0.8–1.2 MHz and with a power of several watts are used. The radiating surface of the transducer (Fig. 6.32) is a few cm^2, the applied intensity ranges between 0.1 W/cm^2 and maximum 3 W/cm^2. In the mechanism of effect an important role is played by the alternating sound pressure the extent of which is much larger than that mentioned concerning audible sounds.

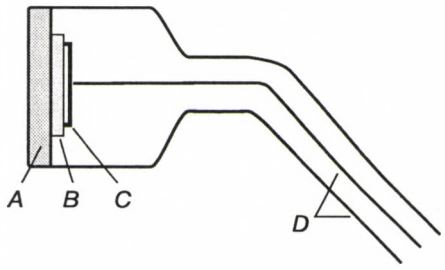


Fig. 6.32. Section of the ultrasound transducer
A: cover plate which is one of the electrodes; *B*: oscillator crystal; *C*: the other electrode; *D*: coaxial cable

Let us consider as an example a power of 10 W at a frequency of 870 kHz and a radiating surface of 4 cm^2: the intensity is now 2.5 W/cm^2, the extent of the sound pressure would be 3 kPa in the air which easily changes its volume (has a small impedance), however, in the muscle tissue the volume of which is practically unchanged (it has a larger acoustic impedance) it reaches 200 kPa, which is roughly the double of the atmospheric pressure. An alternating sound pressure is considered, which means that points with pressures that are much larger and smaller than the atmospheric pressure alternate at a distance of $\lambda/2$ (in our case less than 1 mm).

Thus a considerable tensile and compressive effect is exerted in the irradiated tissue (micromassage); this is one of the causes of the therapeutic effect. The other cause is that the energy of the ultrasound is absorbed in the tissues (the half value thickness in muscles is approximately 2 cm), it is transformed into heat and leads to warming up.

b) The dentists use ultrasound oscillations with a frequency of 20–40 kHz for scaling. The magnetostrictive transducer ferromagnetic rod used for this purpose is fitted with an appropriately formed piece which makes possible a gentle and thorough scaling. In fact, ultrasound irradiation does not take place in this case, the oscillating metal tip transmits energy directly to the odontolith.

c) We mention here the *shock wave therapy* used for the gentle, noninvasive lithotripsy of renal (and other) calculi. A common misunderstanding must be dispelled: *this is not an ultrasound method.*

The essence of the method is the following: a capacitor charged to a voltage of 15–25 kV is discharged through an underwater electrode. As a consequence of this water evaporates explosively and a short pressure pulse develops which then disappears during the condensation of the vapour. The spark-gap which discharges the capacitor is in one of the focuses of an elliptic metal reflector; the semielliptical reflector itself is closed by a plastic bag; the water filling the bag mediates the pressure wave into the patient's body, to the calculus which is placed exactly in the other focus of the reflector under two-directional X-ray or ultrasound control (Fig. 6.33). The pressure attainable in the focus is 10^7–10^8 Pa, depending on the capacitor voltage, under its effect cracks develop in the calculus. Depending on the size and material of the calculus, 500–1000 pressure waves have to be applied for the disintegration of the calculi to a sand-like fineness.

As opposed to the periodic pressure wave of the ultrasound, the shock wave consists of

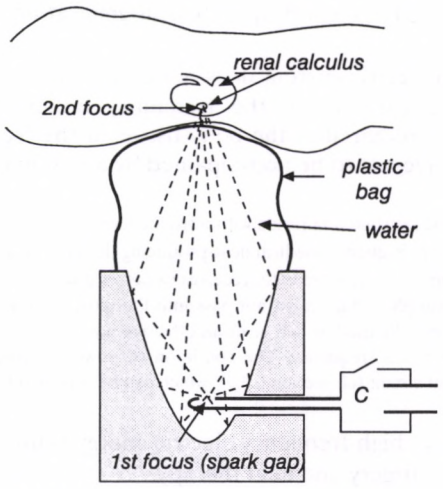


Fig. 6.33. Shock wave generator; propagation of the pulse wave from the first focus of the generator into its second focus where the calculus is

a single positive pressure phase which develops during a ns interval and decays during a time in the order of μs (Fig. 6.34).

d) The *reflection* serving as the basis for the *diagnostic application of the ultrasound* makes possible the noninvasive examination of the different structures of the tissues. Many kinds of ultrasound methods have been developed (particularly following the appearance of computers in medicine), several of them involve imaging, therefore these diagnostic methods – also those which are not imaging – will be dealt with in section 6.7.

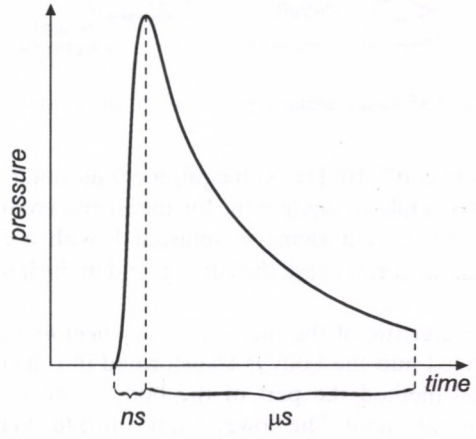


Fig. 6.34. Development and decay of the shock wave

6.4.3. High-frequency heat generation

The heating effect of an electric current does not depend on the current direction or, in the case of alternating current, on the frequency either. Electric power can be transformed into heat as required in the body tissues if the frequency of the current is high enough for its passage not to be accompanied by an excitation effect.

This condition is found to be satisfied if currents with a frequency higher than 10^5 Hz are applied. The energy produced in the oscillator of high-frequency medical heat-producing devices is transferred to the site of treatment by the patient circuit (cf. section 6.2.1). The oscillator receives its power usually from a mains transformer. To avoid electric shocks, the oscillator supply voltage must not pass into the patient circuit. For this reason, an inductive coupling with air insulation is usually used, which gives an effective power coupling in the radiofrequency range, and a total separation at the mains frequency. The condition of power transmission is the resonance of the oscillator circuit and the patient circuit (cf. section 6.3.3), which can be achieved with automatic tuning.

The heat obtained from high-frequency electric energy is used in two fields in medical practice: high-frequency surgery and heat therapy.

(a) In the surgical application of high-frequency power the body tissues are connected to the patient circuit via a large-surface neutral electrode and a cutting electrode with small surface (Fig. 6.35). With this arrangement the current relating to unit cross-section, i.e. the current density, will be high close to the cutting electrode. Since the heat produced in unit volume of tissue is proportional to the square of the current density, the warming-up will be stronger at the cutting electrode. The cutting (tissue-separating) effect results from the intense heating under the cutting electrode causing the cells as they were exploding. An important feature of this method is the coagulation due to the heat production, and there is thus relatively little bleeding.

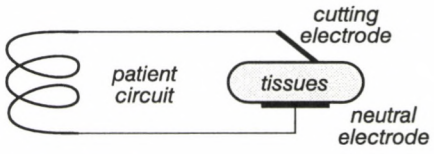


Fig. 6.35. Patient circuit applied in high-frequency surgery

The frequency applied is 10^5–10^6 Hz; with equipment used in major surgery the power is several hundred watts, while in equipment for minor interventions it is lower by one order of magnitude. The current shape is sinusoidal, with a constant or modulated amplitude (Fig. 6.36). In the former case the cutting, and in the latter case the coagulating effect is dominant.

(b) A common characteristic of the high-frequency *heat-therapy* methods is that the electric energy introduced into the body is transformed into heat within the tissues. In contrast to the surgical method the part of the body to be treated is not connected galvanically into the patient circuit. The power can be introduced into the body in various ways, associated with characteristic heating conditions.

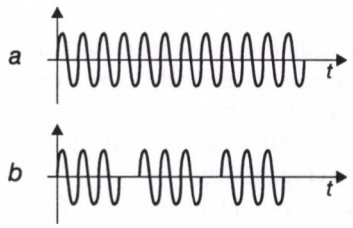


Fig. 6.36. Examples of the high-frequency current forms used in surgery
a: continuous; *b*: pulse modulated shape

With the *capacitor field method,* the part of the body to be treated must be placed between the insulator-coated plates of the capacitor of the patient circuit; in the case of the *coil field method,* the part to be treated must be brought beside or inside the coil. In both methods the frequency is several times 10 MHz, and the power a few hundred watts. The body tissues treated in the capacitor field are warmed up as dissipative dielectrics (cf. capacitive current in section 6.2.1). The electric field produced in the individual tissue layers is determined in a rather complex way by all the electric properties (electric conductivity and dielectric coefficient at the applied frequency) of all the layer together. From the aspect of treatment an important result is that the electric field is lowest in the well-conducting muscle tissues, whereas the heating of the fat tissues is almost ten times that of the muscles. In the treatment with an inductive coil, the intensity of the induced current is proportional to the conductivity of the medium, and accordingly this latter method is favourable from the viewpoint of muscle heating.

One of the *radiation field methods* applies a frequency of approximately 0.43 GHz,

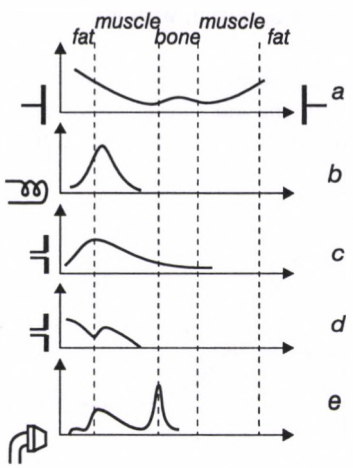


Fig. 6.37. Temperature distributions in the individual tissue layers
for various high-frequency heat treatments
a: capacitor field; *b*: coil field; *c* and *d*: radiation fields; *e*: ultrasound treatment.
The horizontal axis shows the position of the tissue between the capacitor plates and their distance from the coil, i.e. the radiation sources

whereas the other operates at 2.5 GHz; the respective wavelengths are approximately 70 and 12 cm. Electromagnetic radiation is directed with a radiator consisting of a dipole antenna and a reflector to the part of the body to be treated. At the frequencies applied, the attenuation coefficient of the fat tissues is smaller than that of the muscle tissues. This is valid especially for irradiation with longer wavelengths resulting in a considerable heat effect in the inner tissues, too (Fig. 6.37*c*).

Figure 6.37 compares the various high-frequency (and ultrasound) treatments. The diagram demonstrates the warming-up of various tissues treated together.

6.5. Applications of electric pulses

Several diagnostic methods are based on the characteristic electric phenomena associated with the functions of the cells and organs (cf. section 7.2), which permits the study of these functions. Further, these processes can also be influenced by electric stimulation; this may be used not only for research, but also for diagnostic and therapeutic purposes.

The effect of the electric current on a given tissue or cell depends upon the intensity of the current (or more exactly upon the current density), its direction, type, etc. In the following sections this will be illustrated by several applications; only one example is mentioned here. Direct current flowing through the human body remains below the stimulus threshold if the current density is slowly increased up to a current density of approximately 50–200 $\mu A/cm^2$. A rapid increase or a higher current density produces stimulation. The circumstance gives the possibility for the stimulation-free introduction of medicaments electrolytically into the tissues below the skin (iontophoresis).

6.5.1. Stimulation with electric pulses

Experiences can be obtained if a surviving nerve–muscle preparation is stimulated by square-wave current pulses with the aid of electrodes applied to the nerve. The amplitudes and the durations of the individual pulses may differ. It is found that some pulses

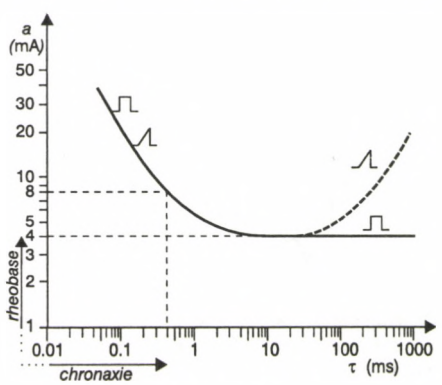


Fig. 6.38. Stimulus characteristics for square-wave and triangle-wave pulses

are ineffective (they are below the stimulus threshold), while others result in muscle contraction. The experimental results are reproduced in Fig. 6.38 (full curve). The horizontal axis gives the duration (τ) of the pulses, and the vertical axis their amplitude (a). The points of the curve denote square-wave pulses which just produce contraction. In other words: The curve depicts the lowest limit of stimulation; this is called the *stimulus threshold* or *stimulus characteristics.* Every point above the curve represents a vs. τ pairs for which stimulation is produced; the points below the curve, on the other hand, relate to a vs. τ pairs for which there is no stimulation. One characteristic datum of the excitability of a nerve-muscle preparation is the lowest pulse amplitude which just produces muscle contraction in a sufficiently long time. In the case of the diagram under discussion, the magnitude of this threshold, the *rheobase* (r), is about 4 mA. In the case of a larger pulse amplitude, a shorter pulse duration is associated with the stimulus threshold. The threshold pulse width associated with the double value of the rheobase is the *chronaxie* (c); its value in our example is approximately 0.4 ms.

The curve can be described by the relation

$$a = \frac{q}{\tau} + r \qquad [6.25]$$

where q, whose dimension is electric charge, is a constant quantity. Its physical meaning can easily be obtained if it is taken into consideration that in the case of sufficiently short pulses (the steep part of the curve) the term r can be neglected relative to q/τ, so that

$$q = a\tau \qquad [6.26]$$

Thus q is the minimum charge (threshold charge) required for stimulation response. The threshold charge may pass through the nerve cell membrane in pulses of varying amplitude and duration. It cannot be applied in very short pulses, since strong warming-up would then be caused. Longer pulses are of course associated with a smaller amplitude, but (as already mentioned) pulses with current intensities lower than the rheobase are ineffective. In the horizontal part of the curve, q/τ can be neglected relative to r, which means that the stimulus threshold is now given by the current intensity and not by the charge. The explanation of this phenomenon is that the passage of the current initiates accommodating processes in the membrane, and these can compensate the effect of the intervention as long as the current intensity is small enough, or more exactly until the value of the rheobase is reached. The compensation is also effective at higher current intensities if this intensity is increased gradually. This is the case if triangular pulses are applied, for instance. Figure 6.38 illustrates two characteristics of the same experimental object. One is valid for square-wave, and the other for triangle-wave pulses. In accordance with what has been said above, the two curves diverge only at higher pulse durations. The adaptability is characterized by the ratio of the amplitude of triangle and square shaped threshold pulses both having a duration of 1 s.

The experimental results obtained by stimulation with square-wave and triangle-wave pulses can also be used in connection with the effects of sinusoidal alternating currents, for the half-periods of the alternating current correspond to individual pulses.

The frequency dependence must be discussed separately, for in its interpretation one should take into consideration that, due to the effect of the electric field, the charge carriers in the membrane and in the intra- and extracellular space are displaced. The motion may be of various types: the translation of ions, the rotation of dipole molecules (atomic groups) and the migration of charge within atoms and molecules. In generating excitation, however, most probably only the translations of the ions (and possibly the rotation of the dipoles) are of importance. However, as a consequence of the relatively large masses of the ions and dipole molecules, these motions are appreciable only if the frequency is not too high. The rapid field-intensity changes accompanying very high frequencies cannot be followed by the "inert" ions (and molecules). The charge motions within atoms and molecules, however, are associated with the oscillations of electrons (of low mass), and consequently these oscillations also occur at high frequencies. These factors do not play a role in the stimulation effects, though they do feature in the production of the heating effect. In the case of higher-frequency alternating current (above 10^5 Hz)

a current with an intensity of several amperes may pass through the human body without any stimulation, practically only the heating effect being observed.

With decreasing frequency, the duration of the half-periods (the unidirectional charge motion) increases; the stimulus threshold appears and gradually becomes lower as the frequency decreases. Below a frequency of several 10 Hz the threshold current intensity increases again, as a result of the compensation mechanism already mentioned. It is unfavourable that the stimulus threshold is lowest near 50–60 Hz, a region of importance due to technical progress (electric hazards, cf. section 6.5.3).

6.5.2. Medical applications of electric pulses

a) The *skeletal muscles* are usually stimulated (e.g. with square-wave pulses) for therapeutic purposes if, for instance, the innervation has been impaired, but there is hope of regeneration. The denervated muscles would begin to undergo irreversible atrophic changes, which would prevent the regeneration of the muscle functions if the innervation were restored. This atrophy can be prevented if the muscles are kept functioning with systematical and steady stimulation. Subsequently, on regeneration of the innervation the muscle may again be able to function.

b) The stimulus characteristics of the skeletal muscles can also be recorded under *in vivo* conditions. The parameters of the resulting curves may supply important diagnostic data: in the event of muscle degeneration, for instance, the rheobase and the chronaxie increase, and the adaptability decreases. This latter circumstance allows a selective stimulation of the degenerated fibres by applying e.g. triangle-wave pulses.

c) As a result of electric or other accidents, and also in the case of surgical interventions, the heart may stop beating or the contraction of the cardiac muscles may become uncoordinated (fibrillation). In such cases the application of a short electric shock of high energy to the heart might be life-saving, since the heart muscles contract simultaneously and subsequently relax. This contraction and relaxation is similar to a natural heart cycle and generally creates conditions favourable for the regeneration of the cardiac functions. The pulse generator applied for this purpose is the *defibrillator.* The defibrillating pulse may be supplied by a capacitor with a capacity of a few 10 μF, charged by a high voltage of several kV. The discharge through the chest takes place within a few ms in the form of a pulse with the following energy:

$$E = -\frac{1}{2} CU^2 \qquad [6.27]$$

The energy may be several 10 J to several 100 J.

d) The *pacemaker* is essentially a square-pulse generator which supplies 70 to 90 pulses per minute. Implanted pacemakers work quite reliably for several years, supplying ms pulses of a few volts, with an energy of ~20 μJ, by which the cardiac function may be controlled if necessary.

6.5.3. Electric hazards and electric safety measures

The hazards of electric current are associated with the stimulating and heat effects of the current. Among the electric power supplies used in everyday life, the electric mains network is of outstanding importance due to its frequent use and inherent dangers. As concerns the sources of danger, we repeat that the stimulus threshold is the lowest around the widely-used frequency of 50 Hz. The danger is increased by the fact that one wire of the mains network is at the earth potential, and thus touching only the other wire may be dangerous if there is no satisfactory insulation between the body and the ground or the other hand touches an earthed object, e.g. a water conduit.

The effect produced by an electric current depends mainly upon the intensity, duration and the *path of the current.* This latter expression refers to the organs through which the current flows and the current density developing in them. An especially critical organ from this aspect is the heart, and within it the sinoatrial node.

Table 6.1 presents data giving the approximate lowest limits of the various effects if the current flows between the two arms for longer than 0.3 s. The current intensity follows Ohm's law, and consequently its actual value is determined by the resistance of the circuit at a given voltage. Under the most unfavourable conditions the only appreciable resistance in the circuit is the human body resistance, in which the skin surfaces play a decisive role. The resistance of dry skin is greater than of moist skin, and thus the resistance of the body depends primarily on the degree of wetness of the skin. Figure 6.39 gives information on the resistance of the human body between two hands of intact skin at various voltages. It may be calculated from the data in the diagram that even 50 V may have serious consequences under unfavourable conditions.

Table 6.1. The effects of electric current

Alternating current (50 Hz)	Direct current	Effect on humans
Current intensity (mA)		
1–1.5	5–6	Weak shock sensation (sensation threshold)
15	70–80	Beginning of danger (painful spasm in limb muscles); "let go" current intensity
25	90–100	Respiratory spasm, heavy pain
80	300	Ventricular fibrillation, danger of death
100	500	Cardiac paralysis, clinical death

* Data of the International Labour Organization (1961)

At lower voltages (a few hundred volts), or at moderate current intensities, the contractive symptoms are dominant. The higher current intensity due to a higher voltage produces a larger current density on the heart, similarly to the effect of a defibrillating pulse. In a "lucky" case the dangerous consequences due to contraction do not appear, and thus at higher voltages the burning symptoms predominate.

The data of Table 6.1 relate to cases when the current enters and leaves the body surface (macroshock). However, if the entrance and exit occur on the cardiac surface, e.g.

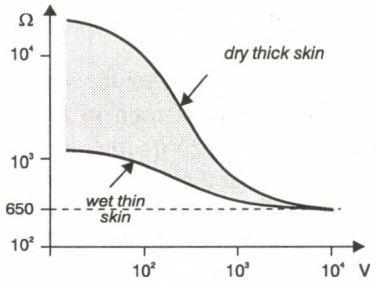


Fig. 6.39. Resistance of the human body

when a catheter is applied (microshock), it is found that a current intensity of even a few 10 μA may cause fibrillation. The generally accepted safety threshold for microshocks is thus 10 μA.

It follows from the above that the use of electric power, i.e. the application of electric/electronic devices, may be dangerous and represents some risk. The possibility of electric accidents can be decreased by means of various shock-proof technical devices and by observing the relevant safety regulations. These were developed on the basis of experience and the actual technical solutions and regulations usually differ depending on the type of the device and the circumstances of its application. It is generally true that the regulations are more rigorous for medical devices than for equipment used in everyday life (for instance in the household), for with certain medical devices *good electric contact* must be made between the patient and the instrument via the electrodes placed on the patient's body. It is extremely important to know and observe the electric safety regulations in every field of application of medical devices.

6.6. Signal processing

Electronic diagnostic devices are generally signal-processing systems. As already mentioned (section 6.1), the signal is related to some process or event (sequence or groups of events) occurring in the system. Accordingly, with some simplification, we may refer to the processing of continuous signals and of pulse signals.

6.6.1. Processing of continuous signals

1. The measuring system is shown schematically in Fig. 6.40; the signal source is the patient. The coupling element (cf. section 6.2.1) ensures a selective connection, specific for the signal to be processed, between the signal source and the processing system. In the case of an originally electric signal (e.g. an action potential associated with the muscle function) the coupling elements are simply contacts (electrodes), but in other cases the coupling element transforms the signal to an electric one (transducer). The amplifier amplifies the signal to the higher level required for further processing. The transfer characteristics of the amplifiers must be adjusted to the signal to be processed. The

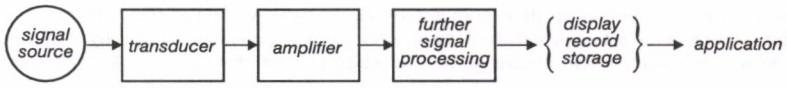


Fig. 6.40. Block diagram relating to the processing of continuous signals

frequencies of the biologically or medically important signals are in the range 0–10 kHz; however, in practice a narrower frequency band within this range is generally used. Direct voltage (0 Hz) is required for measurement of the body temperature or in concentration measurements using electrochemical transducers. In the recording of cardiac sounds the amplifier should amplify signals in the frequency range 10–800 Hz.

The further processing of signal may mean e.g. the elimination of noise signals or the frequency analysis of the signal; however, signal processing with computers (microprocessors) is getting more and more general. In respiratory function analysis e.g. the detector produces a voltage signal proportional to the intensity of expiration and after processing the recorder gives not only the intensity versus time curve, but numerous other diagnostic data too, as e.g. the vital capacity or the highest expiratory intensity even expressed in the percentage of normal (healthy) data. In some cases the transformation of electric signals into electric signals with different parameters may be necessary to simplify or even render possible the further processing.

2. Signal conversion. The transformation of an electric signal into an electric signal with different parameters is called *signal conversion,* and the functional unit converting the signal is the *converter.* The computerized processing of signals requires an *analogue→digital* conversion of the amplified signal which assigns digital data to the numerical values indicating the instantaneous magnitude of the signal voltage. This sampling should be done with a frequency high enough to ensure an unchanged information content of the complex biological signals (cf. section 7.4.2). This is usually attainable if the frequency of sampling is at least double the highest frequency component of the signal (cf. Table 7.4). In the case of a computed X-ray tomograph (cf. section 6.7.3) the signals of the transducer detecting the X-ray intensity transmitted by the body section are fed into the computer after an analogue→digital conversion. The computer stores the calculated elementary density data in digital form and consequently the display of the densitogram wilt be preceded *by digital→analogue* conversion.

Digital computers and calculators operate in a binary system, though the data are fed in and the results are displayed in the decimal system. Both the input and the output involve *digital→digital* conversion. The converters may be accessories or interfaces of the computer.

3. Multichannel measuring systems and other constructions. The system depicted in Fig. 6.40 is suitable for the processing of only one signal at a time. Several such or similar systems *(channels)* are frequently built together into one piece of equipment. With multichannel equipment several signals can be studied or recorded simultaneously (synchronously). Their best-known applications are the simultaneous multielectrode detecting and processing of various action potentials (cf. section 7.4).

For the recording of action potentials, the measuring equipment is often completed with various stimulators; thus, the action potentials of the central nervous system can be stimulated repeatedly with light and sound pulses.

The phono- and photostimulators for these purposes are electronically-controlled pulse-operated sound or light sources. Examination of the action potentials of the muscles, or measurement of the velocity of nerve conduction, requires similar equipment, where electric pulses are employed for the reproducible production of action potentials.

In special cases (during operations, deliveries involving complications, etc.) or in critical conditions (e.g. after an operation), measuring systems may be required which are suitable for the display or possibly the recording of the most important body functions and parameters. Thus, bedside monitors most often contain channels to measure or record the electrocardiogram, the pulse and respiratory rate, the temperature and the blood pressure. It is a general requirement that intensive care systems give alarm signals whenever any parameter goes outside the preset range.

6.6.2. Processing of pulse signals

Determination of the *frequency of some event* is a widely used diagnostic method. Examples of such parameters to be determined are the concentrations of the blood particles, and the distribution of radioactive isotopes in the body. The technical solution of the measurement of such signals requires the processing of pulse signals.

A block diagram of the processing system is depicted in Fig. 6.41. The signal source may be the patient if the determination relates to the distribution of a radioactive isotope introduced into the organism, or it may be a blood sample if the number of red blood cells is to be determined. The role of the detector is to assign voltage pulses to the individual events. After appropriate amplification, these pulses are classified (discriminated) according to amplitude. The further processing involves either *counting* or *frequency determination*, with subsequent display or recording and computer processing. In the following section a few remarks are made in connection with some functional units.

Fig. 6.41. Block diagram relating to the processing of pulse signals

1. Detectors have already been dealt with, mainly in connection with the measurement of nuclear radiation (cf. section 3.3.1). As an example, we describe a detector which is used to count the blood particles (Fig. 6.42). The blood sample, diluted with physiological salt solution, is pumped from the outer vessel to the inner one through a capillary at the bottom of the inner vessel. Meanwhile, by means of the two electrodes a current of low intensity flows through the system. If a blood element passes the capillary, the resistance of the electrolyte in the capillary increases, which produces a voltage pulse.

2. Discriminators. The amplitude of the pulses of the above detector is proportional to the volume of the blood particles. A similar statement can be made in connection with the scintillation detector: its output pulses are proportional to the energy transferred to the scintillator by the gamma-photon. Often the classification of pulses according to amplitude *(pulse-height analysis)* may yield valuable information. For this purpose electronic units, discriminators, are used. These have two modes of operation.

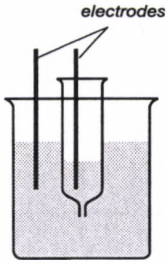


Fig. 6.42. Diagram of the counting of blood particles (measuring capillary)

a) The *integral discriminator* (Fig. 6.43*a*) gives a pulse on its output only when the amplitude of the pulse on the input is larger than the preset discriminating *threshold voltage* (U_{ID}). The integral discriminator responds to every pulse above the threshold with a uniform output pulse. It has an important role in increasing the ratio of signal and noise pulses. The amplitude of the latter is usually smaller than the signal amplitudes, so the optimal value of the discrimination threshold can be found, where the ratio of signal pulse number/noise pulse number is maximum.

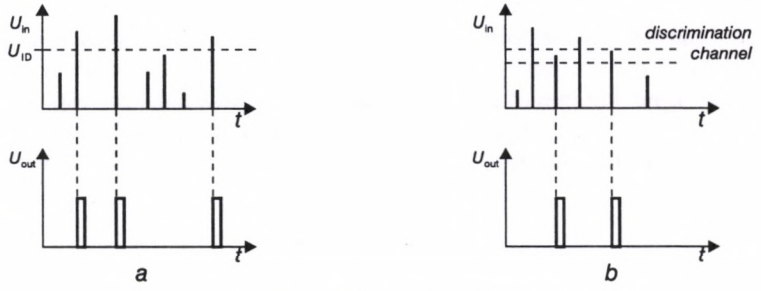


Fig. 6.43. Diagram of the functioning of discriminators
a: integral discriminator (ID); *b*: differential discriminator (DD)

b) Pulse-height analysis is carried out in the *differential discriminator* (Fig. 6.43*b*) operation mode. The difference between the differential and the integral discriminator is that in this mode of operation an upper threshold voltage can be set above the discrimination threshold. A pulse appears on the output whenever the amplitude of the input pulse falls between the two preset limits, i.e. it enters the *discrimination channel.* With the differential discriminator the frequency distribution of the pulse amplitudes, i.e. the pulse amplitude spectrum, can be obtained. A similar purpose is served by the multi-channel pulse amplitude analysers which classify the pulses of a temporal set of pulses according to their amplitudes. This is the way the histograms of the haematological automatic equipments are made which present e.g. the frequency of the leukocytes plotted against the cell volume.

3. Pulse counters. The individual pulses can be counted relatively easily by means of *bistable multivibrators.* This is a functional unit with two stable states (Fig. 6.44) and two

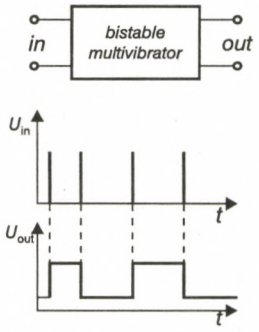


Fig. 6.44. Pulse-counting with a bistable multivibrator

definite values of the output voltage. The change of state (and the change of the output voltage) is triggered by an input voltage of suitable amplitude. The next pulse resets the previous state (together with the corresponding output voltage). Thus the bistable multivibrator responds with one square pulse to the two input pulses. Its pulse width is the time passing between the appearance of the two input pulses. However, this means that the bistable multivibrator halves the pulse number. A pulse-dividing (pulse-counting) chain is obtained if these multivibrators are connected so that the output pulses of one unit are passed to the input of the next one. Counting is carried out in a binary system, the two values of the output voltages of the individual multivibrator units correspond to the two digits of the binary system. Their place values are given by the individual multivibrator units. The results of the counting are converted into the decimal system with a digital–digital converter.

6.6.3. Telemetry

It may sometimes be necessary to process the signals at some distance from the signal source. This situation occurs for instance if the effects of special circumstances or stresses on the physiological parameters are to be examined. A well-known example is the control on the Earth of the vital functions of astronauts in space, but sport physiology and labour hygiene may also acquire valuable information through telemetry. The problem is solved by connecting a telemetric channel between the signal source and the processing system. In more simple cases this can be done with conducting wires; for instance, the patients in the intensive care units of hospitals are connected to the central observing system by means of wires.

In other cases only wireless communication is possible, because the wires would hinder the activities to be observed.

6.6.4. Medical electronics and computers

Complex and exact evaluation of certain signals is possible only with the use of computers. As an example, the analysis of EEG or ECG signals might involve the evaluation of complex signals obtained in 3–250 channels (cf. section 7.4).

Many devices have their "own", *special computer*. For instance, in the laboratory diagnostical automatic equipments the moving of the sample holders containing urine probes, the adding of reagents necessary for the colour reactions, the determination of the concentrations, then the systematic storage and printing of the data are controlled or carried out by computers.

The development in this field was initiated and made possible by the application of microprocessors. Microprocessors can take over the functions of the central unit of digital computers (cf. section 8.3). They are not computers themselves, but can be made suitable for various calculating, counting, comparing or regulating functions by appropriate additions (memories, input and output units). Their application possibilities are practically unlimited. For instance, a microprocessor signal-processing system can automatically carry out its own standardization, zero-point correction, sensitivity regulation, error display, measuring channel change, etc. In this way e.g. after putting on the electrodes and starting the ECG apparatus by pushing a single button it records successively the usual 12 ECG curves (meanwhile stabilizes the zero level, checks the skin-electrode resistances and the sensitivity of recording) and finally it prints the data obtained from the computer analysis of the ECG signals on the recorder chart. It is even possible in intensive care systems for parameters of ECG signals (time intervals, amplitude values) to be compared continuously with preprogrammed data or with the data of the previous cardiac cycle and the device signals any unfavourable tendencies in due time.

The majority of the diagnostic equipments to be described in the following (section 6.7) operate under the control of computers (microprocessors). The first of these, the computer tomograph (CT) still has in its name the reference to the computerized function, nowadays, on the other hand, this is considered obvious.

In the large medical institutes several diagnostic units (CT, PET, ultrasound equipment, etc.) are operated in different rooms or buildings. The connection of every imaging and other diagnostic departments of the institute to a computer network is technically feasible and, in fact, has been realized in many cases. By means of this Picture Archiving and Communication System (PACS) every picture and laboratory finding of a patient is available at any time, on any of the terminals of the institute.

6.7. Iconographic methods in medical diagnostics

Iconography is the collective designation of the various methods employed in medical diagnostics to produce two-dimensional images.

These are generally electronic methods, with a few exceptions, such as conventional X-ray techniques or the traditional and partly fibre optical endoscopy. Computer-aided methods are of ever growing importance: medical iconography is usually based on digital image construction.

Digital processing and treatment of electric signals offer many advantages, e.g. video signal storage in the computer memory, selection and arbitrary modification of any range of image contrast, reduction and enlargement of image parts, various measurements in the recorded picture (distances, directions, areas, etc.). The processed image appears on the screen of a cathode-ray tube and is treated by controlling and performing operations on the screen. For documentation pictures may be made by a polaroid camera or the contents of the screen may be printed out. The product of the recent printers are coloured pictures of photographic quality.

It has to be pointed out that the colours of the pictures (either on the screen or in the print-out) usually refer to colour-coded information, thus are not real colours. An exception is naturally endoscopy where the colour fidelity of the tissues may be important for the physician.

6.7.1. Endoscopy

Medical documentation conventionally uses the imaging of inner and outer surfaces of the body (by drawing or photographic methods). Rigid metal tubes containing lenses and other optical elements used for imaging body cavities have recently been replaced by flexible bundles of light-conducting optical (glass or plastic) fibres. These systems not only introduce light to illuminate the cavity but simultaneously pass the picture to be seen and recorded outside (Fig. 6.45). The illuminated cavity is projected by the objective on the endplate of the fibre bundle. At the outer end of the endoscope the image may be observed, or may be stored or visualized on a screen by means of a CCD camera. In other, flexible endoscopes only the introduction of the light is carried out by the fibres, the image is projected on the CCD image converter chip of a few mm^2 surface by a lens at the inner end of the endoscope, the video signals are conveyed by an electric wire. There are rigid endoscopes with a maximal length of several ten cm, while the flexible ones may be even 1.5 to 2 m long.

Endoscopes are not only diagnostic equipments: with their help also biopsy or certain surgical interventions may be performed. For these purposes they have further channels, in addition to those for the illumination and image transport. Some of these make possible the injection of air or water, while via the so-called working channels grasping, cutting, pinching and burning instruments or laser conducting fibres may be introduced into the cavities of the body which may be operated from outside.

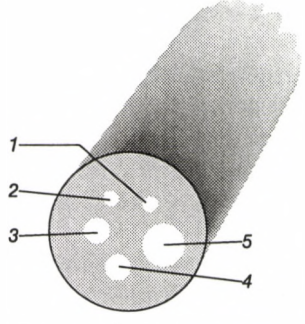


Fig. 6.45. Endplate (or front plate) of an endoscope introduced into a cavity
1 and *2*: end of the tube for air or water; *3*: objective; *4*: endplate of the optical fibre bundle illuminating the cavity; *5*: operating channel

6.7.2. Thermography

Two methods have been elaborated for mapping the body surface according to its temperature. One of them is the so-called *contact thermography* which produces a colour coded temperature map of the body surface on a cholesteric liquid crystalline film placed onto the surface (cf. section 1.4.4). The other method is the so-called *telethermography* or *thermovision.* This is based on the fact that the intensity and spectral distribution of the thermal radiation from the body surface depends on the surface temperature (cf. section 2.4).

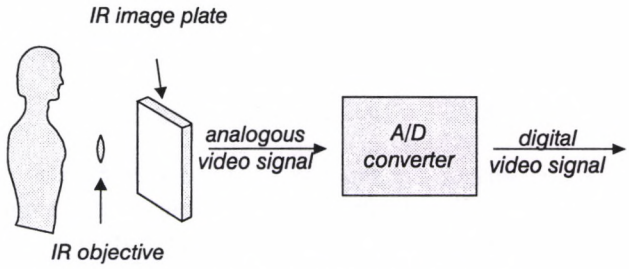


Fig. 6.46. Schematic representation of thermography

Considering that at body temperature the maximum of the emission spectrum of the heat radiation is approximately at 9.5 μm, the task may be carried out by means of an *infracamera*, resembling a video camera, in which both the objective and the image plate are adjusted to this wavelength range. Therefore the material of the objective has to be transparent in this far IR range; such are e.g. zinc selenide (ZnSe) and germanium. The adequate physical phenomenon for detection is photoconduction; a material having the necessary energy band structure is e.g. mercury cadmium telluride (HgCdTe). The number of the detectors on the image plate is more than 60 000 (256 rows with 256 detectors in each row), the video signal is scanned similarly to the case of the CCD image plate (Fig. 6.46).

If processed digitally, the data can be stored, evaluated and recorded any time. In both of the above methods, polaroid photography is used for documentation. Signal processing makes also possible the use of colour printers.

6.7.3. X-ray techniques

X-radiation was introduced in the medical diagnostics already weeks after its discovery (1895) and since that time its importance has constantly increased. By means of it the noninvasive examination of the interior of the human body has become possible for the first time. Up to now, many methods and equipments have been developed for the diagnostic application of the X-radiation, while the development of electronics and measuring techniques, especially the appearance of the computers has presented new possibilities and opened a new era in this field, too.

These methods utilize the fact that different tissues have different attenuation [or density, $\log (I_0/I)$] due to the differences between their effective atomic number (cf. section 2.10.3).

1. Summation images. The conventional X-ray images appear on the luminescent screen (X-ray screen) or X-ray films placed at the side of the patient opposite to the X-ray tube. The X-ray passes through successive tissue layers of different densities, which all participate in decreasing its intensity:

$$I = I_0 \, e^{-(\mu_1 x_1 + \mu_2 x_2 + \ldots)}$$

The intensity of the luminescent light of the X-ray screen and the darkening of the film depend on the *resulting density* which is given by the following expression:

$$\log \frac{I_0}{I} = (\mu_1 x_1 + \mu_2 x_2 + ...) \log e$$

In other words, on the images the shadows of the details behind one another are projected upon one another. Thus the dimension of the three-dimensional body which is in the direction of the X-ray does not appear in this summation or superposed image (dimension reduction).

2. Conventional tomography. On the summation images tissues with similar or higher densities may conceal the image of details the detection of which might be important for the diagnosis. Therefore special methods have been developed for the imaging of individual layers of the body. Here the simplest of them will be described which makes possible that a selected part of the body, which is parallel to its longitudinal axis, is outlined relatively sharply on the film, while the images of details above and below this layer are blurred, so thus e.g. a lung X-ray is not disturbed even by the shadows of the ribs.

Consider Fig. 6.47. During illumination the X-ray tube of the tomograph moves along a circular arc above the body, at the same time the cassette containing the film moves in the opposite direction along a horizontal straight line below the patient. The centre of their common movement is at the height of the layer to be imaged. It can be seen that the shadows of the points in this layer are cast on the same sites of the shifting film: thus the image of the layer (especially around ***O***) will be sharp. The shadows of details above this layer move faster, those below it slower, than the film, therefore they all become blurred.

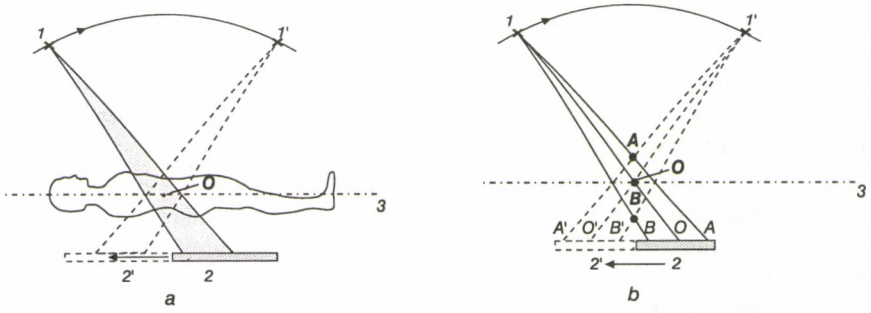


Fig. 6.47. Schematic representation of the conventional tomography. In both pictures:
1–1' displacement of the focus of the X-ray tube; *2–2'* simultaneous displacement of the cassette;
O centre of the displacement; *3* examined layer
b shows separately the point above the examined layer (*A*)
moving faster and the point below it lagging behind (*B*)

3. Electronic X-ray image amplifier. Its schematic structure is seen in Fig. 6.48: in a glass vacuum bulb there are two luminescent screens, a photocathode and a cylindrically symmetric system of electrodes. The image of the illuminated body appears first on the

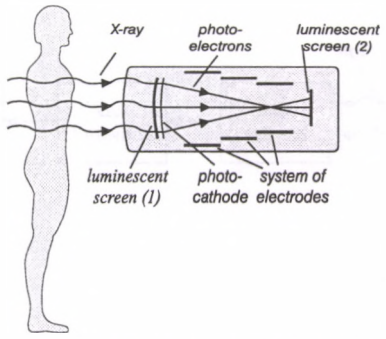


Fig. 6.48. Schematic representation of the electronic image amplifier

luminescent screen 1. The luminescent light *releases* photoelectrons from the photocathode. The number of these electrons at each place is proportional to the intensity of the luminescent light. The voltage connected to the cathode and the system of electrodes (25–30 kV) produces an electric field which not only accelerates the electrons leaving the photocathode but also functions as an imaging electron lens for them. The accelerated electrons reach the luminescent screen 2 on which the image produced on screen 1 is thus repeated through the electron lens. This latter image is real, reversed, reduced and, what is especially important, has a high light intensity due to the high energy (25–30 keV) of the electrons producing it. This makes possible the further processing and use of the image.

The application of the X-ray image amplifier has several advantages:

- by its use the X-ray exposure of the patient and the physician may be decreased;
- by the simultaneous application of the amplifier, video camera and cathode-ray tube the X-ray picture is well visible even at daylight which makes possible various medical interventions under X-ray control (Fig. 6.49);
- the video signal may be recorded on a video tape and may be played back just as a motion picture;
- digital X-ray imaging and processing may be performed.

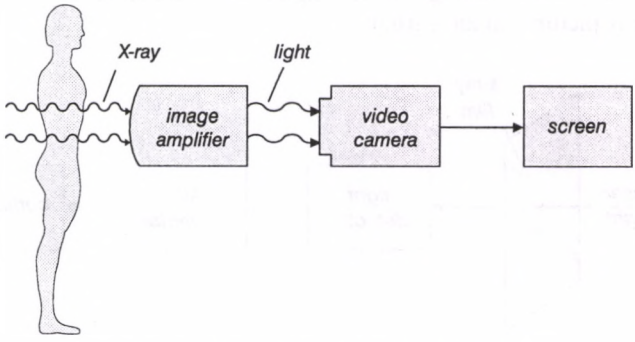


Fig. 6.49. Application of the image amplifier

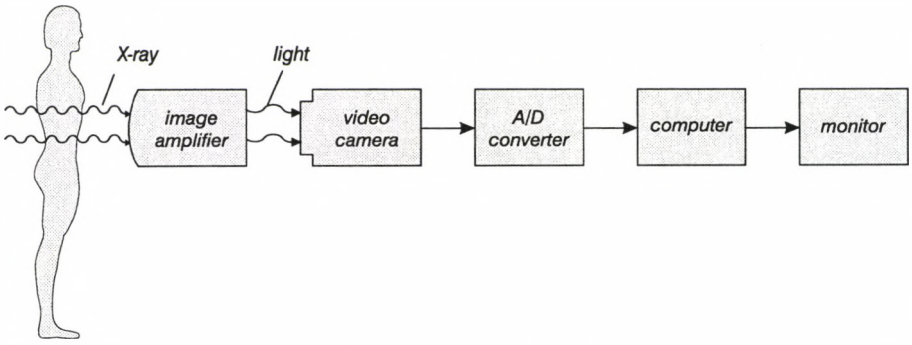


Fig. 6.50. Digital X-ray image processing

Some remarks:

(a) The image amplifier may be well used instead of the X-ray screen or cassette of the traditional X-ray equipments.

(b) The image amplifier is frequently used in the so-called C-armed arrangement in which at one end of a 1-m-diameter semicircular support (resembling the letter C) the X-ray tube, while at its other end the amplifier is located, facing each other. The C arm can be adjusted to the X-ray examination of any part of the patient's body from any direction.

4. Digital X-ray imaging. Figure 6.50 shows the scheme of a possible construction for this task: image amplifier–video camera–A/D converter–computer–monitor. From among the advantages of digital X-ray imaging only the selection of any part of the contrast range is mentioned here. A high-resolution digital X-ray picture may contain as much as 1024×1024 pixels, the stored contrast dynamics is 8 bits. The 8 bits mean $2^8 = 256$ contrast grades from which any brightness range may be selected and displayed on the screen.

In the scanner developed for the digitalization of X-ray films (Fig. 6.51) laser light illuminates every spot in every row of the film. The signal voltage, which is proportional to the intensity of the light passed through each point, is produced by a photosensitive detector (photodiode or phototransistor) and after A/D conversion is forwarded to the computer. This way the advantages of the digital X-ray imaging may be enjoyed even in the case of X-ray pictures made earlier.

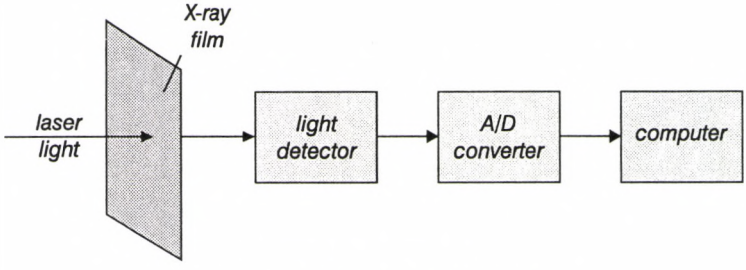


Fig. 6.51. Digital processing of X-ray films

It may be expected that the large *X-ray image storing plates* appearing in the near future will drive the X-ray films out of practice. Their construction and function are similar to that of the dental image storing plate to be described later (cf. point 7), their reading is similar to that shown in Fig. 6.51. The image storing plate may be used many times, its use involves much less exposure than that of X-ray films. The image is digitally processed, however, if necessary, the picture stored in the memory of the computer can be transmitted onto a film by means of a laser light which exposes each point of the film. (The films of the CT images are also made like this.)

5. Digital subtraction angiography (DSA) or digital angiography (DA), takes a series of records of the investigated blood vessels and their environment from a steady recording position. A so-called base record is taken prior to the injection of contrast material and the rest after the injection. The computer subtracts the tone values of the first exposure (as a background) from those of each subsequent member of the series pixel by pixel and the difference pattern appears on the screen with a contrast greater than at any previous angiogram.

6. Computed tomography (CT) (X-ray densitography, computer-aided tomography) reveals the third dimension of the body which remained hidden at conventional summation X-ray imaging. The method yields two-dimensional transverse sections of a few mm thick layers perpendicular to the body axis. The grades of the grey tone correspond to different values of X-ray attenuation (densitogram). Although the first equipments were followed by more and more advanced generations, their principle of operation remained the same. This will be illustrated in the following in the simplest, first-generation variation.

Consider a few mm thick cranial segment perpendicular to the axis of the body (Fig. 6.52). Let the layer be covered by a network of 1.0 mm^2 squares. If a 20 × 20 cm^2 surface is assumed, the network contains 40,000 elements. Valuable information on the layer is obtained if the attenuation coefficient of every element of the net, the *density matrix,* is known. The layer positioned for image formation is transirradiated by a narrow X-ray beam approximately 1.0 mm in diameter from the source (S). The source moves past the layer. On the opposite side the detector (D, e.g. a scintillator crystal), which measures the intensity of the emerging radiation, moves together with the source in approximately 1 mm steps (or continuously). Both the detector and the X-ray tube are supplied with a narrow (approximately 1 mm) lead collimator, the collimator belonging to the detector serves the resolving power, while the one before the X-ray source decreases first of all the exposure of the patient. It is obvious that a row of the density matrix participates in the intensity decrease relating to the momentary position of the radiating source and detector. Clearly:

$$I = I_0 \, e^{-(\mu_{i1}\Delta x + \ldots + \mu_{ij}\Delta x + \ldots + \mu_{in}\Delta x)}$$

where I_0 is the incident, and I the emerging intensity, Δx represents the thickness of the matrix element, and μ_{i1}, ... μ_{ij}, ..., μ_{in} are attenuation coefficients characteristic of the individual elements (in our example the number of elements in a row of the matrix

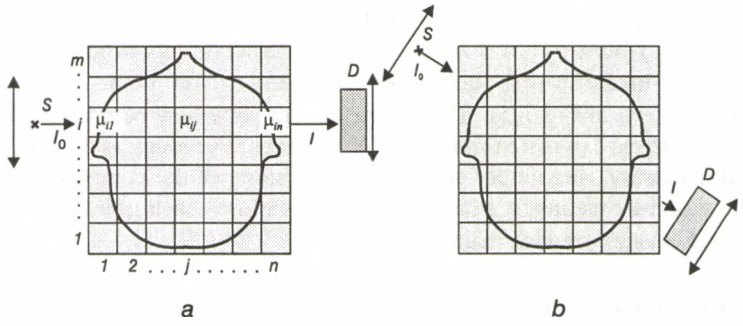


Fig. 6.52. Schematic representation of computed tomography

$n = 200$). Similar relations apply to every matrix row (the number of rows $m = 200$). Overall, a knowledge of $m \times n$, i.e. 40,000 elements, is required for the example discussed. A one-directional displacement of the radiation source and the detector along the layer is clearly insufficient, since this would furnish only 200 equations in our example. However, the problem is solved by the equipment scanning the same sectional layer in several directions. (The first CT repeated the scanning 180 times after rotations of one degree.) In our figure the arrows beside the X-ray source and the detector show the direction of the scanning, in *a* in the first position, in *b* in the position corresponding approximately to the 30th scanning. After a sufficiently large number of scannings the computer solves the system of the equations with 40,000 unknowns of our example on the basis of the available data, i.e. it calculates the density of every matrix element, carries out the coding according to the grey-tone scale and presents the densitogram of the examined layer on the screen (Picture 6/1 in the Supplement).

The first CT with its single detector – having a NaI(Tl) crystal and photomultiplier – could scan one cranial layer in about fifteen minutes, the data processing took a further half hour before the image appeared on the screen. A later, third-generation variation uses a fan-shaped beam of X-rays and several hundred detectors, the motion of the tube and the detector system is simplified to a rotation around the body, thus the measuring and data collecting speed increased considerably. In the fourth generation CT the detectors are placed along a ring form around the body and only the X-ray tube is moving. The X-ray tube of the so-called spiral CT carries out several circular movements successively, while the patient is steadily shifted in the direction of his/her body axis.

CT images are usually made successively from a large number of neighbouring layers. The data of the layers (coordinates, densities) are stored in the memory of the computer, from these a segmental image of any position can be displayed on the screen. Such a segmental image usually consists of 512×512 pixels with a resolution of about 0.5 mm and a contrast dynamics of 10 to 12 bits. (12 bits: $2^{12} = 4096$ densities slightly differing from each other.) From this broad contrast spectrum a narrower range can be selected for processing, according to the object of the examination. There are image processing programs which make possible a quasi three-dimensional display. With their help e.g. in case of a cranial injury the virtual image of the skull may be constructed from optional

layers, leaving out the density range of the soft tissues; and since the image may be rotated in any direction by these programs, the bony substance of the skull may be observed inside and out.

7. Special dental methods. Although in dental radiology the conventional method – the production of an X-ray image on a film placed behind the tooth (in the oral cavity) – is used the most frequently, some more recent methods should be mentioned.

The *panorama tomogram* is somewhat similar to the conventional tomogram in that the X-ray tube and the film cassette pass around the head of the patient in the opposite directions. However, in this case a sufficiently sharp image of a curved object (dentitions, mandible) should be produced on the film. The movement takes place along a horizontal path, that of the X-ray tube behind the head, that of the cassette in front of it, at the height of the dentition around a centre of rotation which passes along the curved object to be imaged during the exposure of 10 to 15 sec. In front of the X-ray tube there is a narrow slit collimator in vertical position, thus the illumination is made with a vertical, fan-shaped, narrow X-ray beam (this decreases first of all the exposure of the patient); another, similarly vertical narrow slit collimator moves in front of the film in a way that in every moment of the exposure only a narrow vertical strip of the mandible or the dentition is projected onto the film. The focus of the X-ray tube, the centre of rotation and the two slits are in the same line during the whole time, meanwhile the cassette rotates around a vertical axis in a way that the just exposed detail is perpendicular to the axis of the X-ray beam.

The methods producing *digital dental X-ray pictures* do not use films and do not require film processing.

a) One of these methods uses instead of the film a CCD image plate built closely together with a luminescent layer. The CCD plate is connected to the digital image processor. The image appears on the screen. The computer offers the advantages of the digital image processing (cf. section 8.3.2) also in this case.

b) Another method applies a special image storing plate instead of the film. In this plate microscopic crystals are evenly distributed the material of which is a doped insulator with a broad forbidden band (cf. section 1.4.5). The doping ions have a metastable level in the forbidden band of the basic material, deep below the conduction band. The X-ray image is stored on the plate by the electrons captured at the metastable level. The exposed plate may be scanned by laser light (laser scanner): the laser light excites the electrons from the trap level to the conduction band from where they get to their original state by emitting luminescent light. (The light also "erases" the plate which can be therefore re-used.) The detector of the luminescent light provides the video signal the digital processing of which is computerized.

The advantage of both methods is that their application involves smaller exposures, namely 10–20% of that necessary for the conventional X-ray films.

6.7.4. Methods employing radioactive isotopes

1. Gamma-camera. It records the two-dimensional projection and time course of the spatial distribution of radioactive isotopes (more correctly: compounds marked with radioactive atoms, radiopharmacons, cf. section 3.5) introduced into the human body for diagnostical purposes.

Its construction and function are shown in Fig. 6.53. The γ-photons leaving the organism reach the scintillation detector crystal through a multi-channel lead collimator; from there the light of the scintillation arrives to the photomultipliers through a few cm thick light guide layer. The material of the large-surface (e.g. with a diameter of 40 cm and a thickness of 1 cm) scintillation detector is most frequently NaI(Tl). The lead collimator below it lets the γ-photons through only in a direction approximately perpendicular to the detector; due to this circumstance the spatial distribution of the scintillation events corresponds to the projection of the isotope distribution in the body. The light of the scintillation spreads in the light guide, thus the arrival of a single γ-photon is detected by several photomultipliers, but they signal it with pulses of different amplitudes depending on the distance from the site of the scintillation. The site (coordinates) of the scintillation is determined by the comparison of the pulse amplitudes of the individual photomultipliers which is performed naturally by the equipment itself.

Though the employed large-surface scintillation crystal plate is covered by not more than a few times ten photomultipliers the resolution is fairly good. The usual pixel number (expressed in a quadratic pixel matrix) is 64×64, but there are gamma-cameras with resolutions of 128×128 or even 256×256 pixels.

The gamma-camera operates quickly, by using an appropriate amount of isotopes as much as 10 isotope distribution maps may be produced per s (in case of a smaller area even more); therefore it is also suitable for the examination of relatively rapid processes (e.g. cerebral circulation, ventilation, cardiac function).

The spatial and temporal coordinates of the scintillations are stored in the memory of the computer from where the distribution maps, the so-called *scintigrams* may be displayed on the screen with colour coding. The different colours represent the different local concentrations of the isotope.

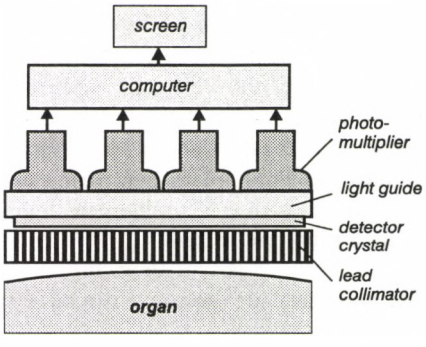


Fig. 6.53. Delineation of the functioning of the gamma-camera

The scintigrams may be *static images,* such as a simple scintigram of the thyroid gland. The individual images of the *process-shots* may be displayed one by one but also as a motion picture *(dynamic examination)*. The region of interest (ROI) with respect to the examination may be marked on any scintigram of the process-shot (on the display with the mouse), then the temporal changes in the activity of any ROI may be displayed graphically. The isotope passage curve of the kidneys, the renogram is made in such a way with the gamma-camera (cf. section 3.5.3).

Similarly to the conventional X-ray pictures, the scintigrams are also summational (superposed) two-dimensional images, i.e. the gamma-camera does not respect the depth distribution of the isotope either.

2. Single Photon Emission Computed Tomography (SPECT). If a gamma-camera is rotated round the patient with a constant slow speed or in steps of a few degrees, in a plane perpendicular to the body axis, then the computer receives summation data of the isotope distribution measured at different directions. From this set of data any distribution patterns for any body section can be displayed. That is, SPECT reveals – similarly to CT – the third dimension being hidden in the summation images produced by the gamma-camera. While, however, the CT examines the thin layers perpendicular to the body axis one after the other, the SPECT surveys the spatial distribution of the activity content of a 30 to 50 cm thick layer at the same time with its large-surface direction-sensitive detector.

3. Positron Emission Tomography (PET). The equipment uses a positron-radiating isotope for marking, its detectors sense the γ-photon pairs radiating in the *opposite* directions after positron–electron recombinations (cf. section 3.2.2).

The predecessor of the PET of our days, the *positron scanner*, is not used any more. Nevertheless, it is worthwhile to get acquainted with it briefly for the understanding of the PET. The essence of its function is outlined in Fig. 6.54 showing the determination of the site of a brain tumour as an example. In the figure each arrow-pair pointing in opposite

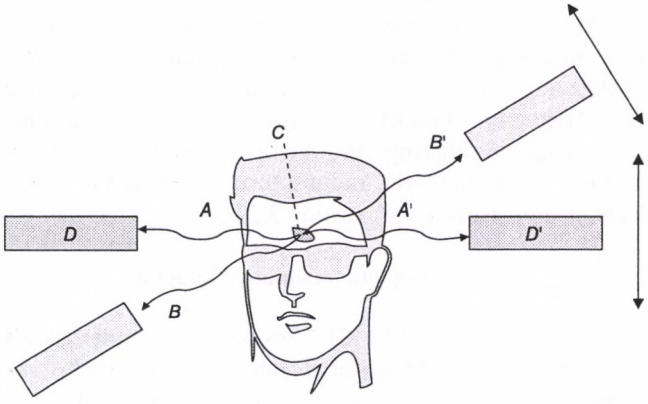


Fig. 6.54. Scheme of function of the positron scanner

directions (*A, A'* and *B, B'*) denote photon pairs produced by the annihilation of a positron. The γ-detectors (*D* and *D'*) placed opposite each other on the two sides of the skull operate in coincidence, i.e. the apparatus counts only photons passing through both detectors simultaneously. Consequently disregarding chance events of low probability, the apparatus counts only the annihilating positrons. There is a notable counting frequency if the radiating centre (*C*) lies in the same line as both detectors and is situated between them. Thus if both detectors are moved in the direction of the double arrows, *AA'* may be determined with respect to the tumour. The same can be done also in other positions of the detectors, e.g. in position *BB'*. The intersection of *AA'* and *BB'* marks the site of the tumour with a sufficient exactness.

As opposed to the positron scanner, the PET is naturally already a computerized equipment, but the basis for its function are still the γ-photon pairs originating from the positron–electron pair, which, as mentioned, moving in the opposite directions make possible the determination of the site of the positron emission. The development was focused on increasing the speed of the scanning. This was accomplished by applying a large number of detectors instead of the detector pair which are placed in a ring form around the body in a plane perpendicular to the body axis. The signal reception and processing speed of the computer makes it possible that not only the opposite detectors of the ring will be coincidence pairs: with the help of the computer each detector has a separate coincidence contact with the group of the opposite detectors. Thus if an emission occurs between any two detectors in the body segment within the detector ring, the signal gets to the computer (if the two γ-photons started towards the detectors at all). If the emission occurred exactly in the middle between the two opposite detectors, the two detections are exactly in the same time; if, however, it is nearer to one of the detectors, the signal from the other one will be delayed: 1 ns time difference corresponds to a path length difference of 30 cm. In practice the computer accepts a pulse pair within 1 ns as coincidence and from the time difference of the two pulses it determines the exact site of the emission. Thus we not only know between which detectors the emission occurred but also its location on the line connecting the detectors.

The scanning of a body segment is performed electronically, without mechanical movement. With the help of an equipment consisting of one ring images from several neighbouring horizontal segments are made consecutively. PETs which are quicker and suitable also for dynamic examinations use several detector rings parallel to each other (e.g. 8 rings with 512 detectors in each), and in them the coincidence contact involves also the detectors of the neighbouring rings. For the operation of a PET a nearby cyclotron is essential, since the positron-radiating radioisotopes used in the PET diagnostics are cyclotron products of short half-life (cf. section 3.2.6).

6.7.5. Magnetic resonance imaging

In addition to the above, abbreviated by MRI, several other names are also used: nuclear spin tomography, NMR (Nuclear Magnetic Resonance) tomography.

The non-zero resultant magnetic moment of nuclei with uncompensated spin can take only certain directions in relation to an external magnetic field to which slightly different

energy values are assigned (cf. 4.6.1). In the tissues of the body first of all the nucleus of the hydrogen atom is taken into consideration (from which there are many in our tissues) and since its spin is 1/2, its energy splits into two levels. The difference of the two energy levels (ΔE) is proportional to the magnetic field strength, the ratio of nuclei being on these levels corresponds to the Boltzmann distribution (cf. Appendix A1). The excitation to the higher level is carried out by radio frequency electromagnetic radiation, but absorption takes place only at a frequency satisfying the following condition (resonance absorption):

$$h\nu = \Delta E$$

After the excitation the spins gradually dispose of the absorbed energy (relaxation) and the population relation corresponding to the Boltzmann distribution is restored.

The amplitude and time course (relaxation time) of the radio frequency signal emitted following the excitation provide information about the concentration and molecular environment of the protons. The excitation frequency being constant, the absorption – and the consecutive relaxation – takes place at a given value of the magnetic field strength. Since we want to obtain detailed information about a body segment, the strength of the magnetic field should meet the resonance condition at a given instant for a small volume element of the body section to be examined. This is attained by simultaneously applying several magnetic fields. The main magnetic field is mostly produced by a superconductor coil. Let this field be parallel to the body axis of the patient and let this direction be that of the Z axis of a coordinate system. Let us now switch on a second coil the field of which is also parallel to the Z axis but uniformly varying along this direction. (The latter is the so-called Z gradient coil.) The resultant of these two fields furnishes the resonance condition in a perpendicular XY plane. A third coil, whose magnetic field varies along the X direction (X gradient coil) limits the validity of the resonance condition to one axis in Y direction within the XY plane and a fourth, the Y gradient coil, to one single point on this axis. By varying the gradient fields, either the mentioned XY plane or any other plane of arbitrary position can be scanned. The protons are excited by a radiofrequency field which, in turn, is produced by a transmitter coil in the form of a short pulse. The signal of the proton relaxation is detected by a receiver coil. The whole equipment operates under computer control, and the signals obtained are also processed by the computer.

In magnetic resonance tomographs the applicable magnetic induction range lies between 0.05 and 2 T (tesla), whereas the frequency of the exciting electromagnetic field corresponds to that of short radio waves: it ranges from a few MHz to several 10 MHz. The method provides a two-dimensional image of an arbitrarily situated section of the examined organ and its vicinity with respect to proton concentration or relaxation times. The main advantages of the method are that – unlike the techniques discussed above – it employs neither X-rays nor radioactive isotopes, and no harmful effect of the used magnetic and electromagnetic fields could be observed up to now. A further advantage is that the image of soft tissues contains fine details and sharp contrasts which is not attainable by CT, since the latter is only sensitive to the elementary composition of the tissues (effective atomic number, cf. section 2.10.3), whereas MRI is sensitive also to the molecular environment of protons, in addition to their concentration.

From among the disadvantages of the MRI method it should be mentioned that the equipment still belongs to the most expensive ones, and the examination may cause unpleasant claustrophobic symptoms, because the patient must stay in a long, narrow tube during the rather lengthy process of scanning.

6.7.6. Ultrasonic imaging

The different body tissues have different acoustic impedances. The background of the *diagnostic application of ultrasound* is its reflection from the boundary media of different acoustic impedance, and the fact that the time interval between the emission of the ultrasound and the return of its echo is proportional to the distance of the reflecting surface. Thus the sites of the body tissues can be determined by the measurement of the time intervals.[3]

The essence of this method is demonstrated in Fig. 6.55. Only one point is stressed. The echo time is represented by a uniform displacement of the cathode-ray in the X-direction. From this it follows that the distance (l) to be determined can be measured by the displacement (l') of the ray as observed on the oscilloscope screen.

1. *A*-image. The echo signals arrive in sequence with a time delay from the acoustically different tissues lying behind one another in the direction of the propagating ultrasound pulse. The distance between the pulses on the oscilloscope screen are proportional to the measured distances of the reflecting surfaces from each other. Further, the amplitudes of the pulses appearing on the screen are determined by the amplitude of the echo pulse. This type of image is an amplitude-modulated or *A* image. The schematic diagram of an

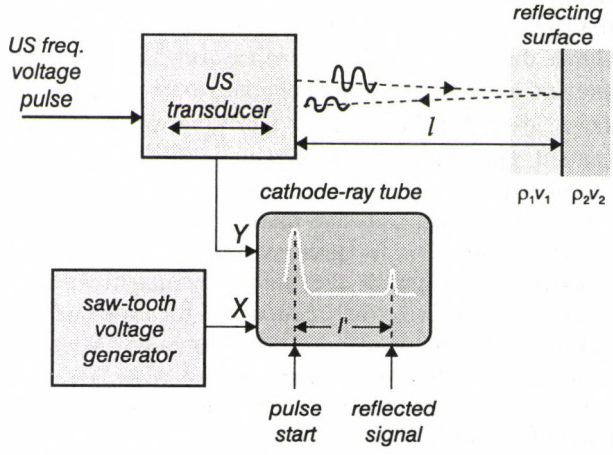


Fig. 6.55. Diagram relating to distance measurement with ultrasound echography

[3] The *radar principle* is frequently mentioned as the basis of the method, because radio waves were used already earlier in a similar way for the determination of the location of airplanes. Nevertheless, it would be more correct to call it "bat" principle, since the reflection of the ultrasound pulses emitted by the bats plays an important role in their orientation. Thus the priority is not due to the radar.

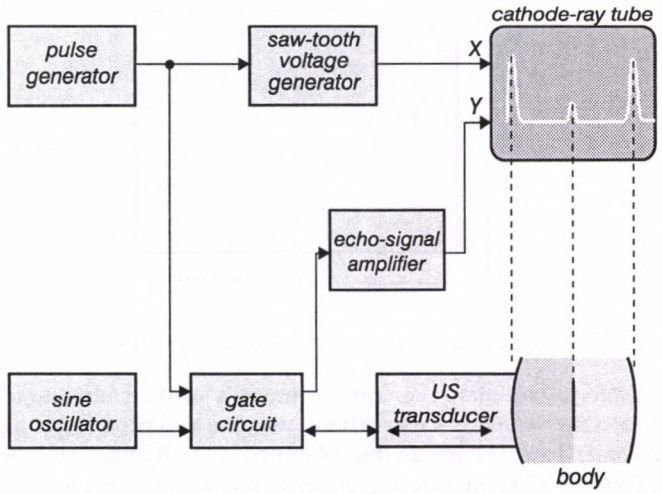


Fig. 6.56. Diagram of ultrasound diagnostic equipment working in the *A*-mode

A-mode ultrasound diagnostic apparatus is presented in Fig. 6.56. With its square-wave pulses of, for example, 1 kHz frequency and 1 μs time period, the pulse generator starts the operation of the saw-tooth wave generator and at the same time opens the gate circuit. From the sine oscillator signals of several MHz frequency in each ms the gate circuit passes an ultrasound frequency signal lasting for 1 μs (several periods) to the ultrasound transducer, which irradiates the ultrasound pulse into the body. The echo signals arrive back in the pauses between the individual input signals and are retransformed into electric signals by the transducer. These signals are then forwarded by the gate circuit (which operates as a receiver in the pulse pauses) into the amplifier. The amplified signals are led to the deflecting system *Y* of the cathode-ray tube and an *A* image is produced.

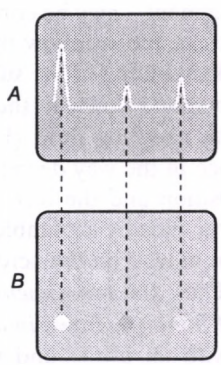


Fig. 6.57. One-dimensional *A* and *B* images
The spots on the *B* image denote light flashes proportional with the echo amplitude

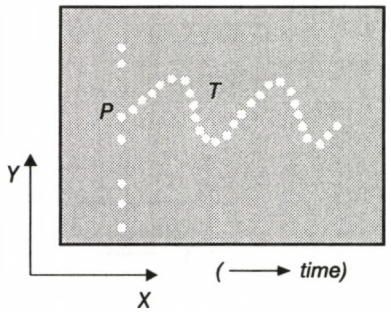


Fig. 6.58. Diagram relating to the formation of an *M* image

It is also possible to modulate the current intensity of the cathode-ray with the echo signals. In this case the screen is dark without an echo, and produces a light flash whenever an echo pulse arrives. The image thus obtained is the *B* image (*B*-type operation, cf. section 6.3.2). Figure 6.57 compares one-dimensional *A* and *B* images.

2. *M*-image. The time course of any periodic motion (e.g. the motion of the cardiac wall) can be observed as an echo image by producing a one-dimensional *B* image of the examined organ with the aid of a fixed transducer. This type of image is demonstrated by the point series in the *Y* direction on the left-hand side of Fig. 6.58. Consider a single point *P* of the image, which should belong to the outer surface of the heart. If the cathode-ray is deflected in the *X* direction with a rather slow saw-tooth voltage, the change in the distance of the selected point in time will be drawn on the screen. In our example the distance of the selected heart surface point from the transducer (skin surface) will be displayed on the screen. This type of image (which visualizes a motion in time) is called *TM* (time motion), or briefly *M* image.

3. *B* image – 2D image. The first equipment producing two-dimensional ultrasound segmental images – the transducer of which had to be moved by hand on the patient's body according to the examined segment – may be considered today only as a curiosity in the history of medicine. Nevertheless, let us follow the early production of two-dimensional *B* images for the better understanding of the subsequent procedures.

For this purpose the transducer slided slowly in the plane to be imaged over the skin surface which was pretreated with coupling fluid (Fig. 6.59), while the emission and detection of echo signals took place in the way described previously. The motion of the transducer, i.e. its momentary position and the direction of its axis, transferred to the deflecting system of the cathode-ray tube by a suitable mechanoelectric transducer. The trace of the cathode-ray spot then moved on the screeen in the same way as the ultrasound pulse moves in the body. Thus, the *two-dimensional B image* was formed, which *consists of a set of one-dimensional B images drawn in consecutive steps.*

Mechanical scanners operate with an ultrasound receiver-transmitter crystal (more exactly piezoelectric ceramic plate) which rotates or performs a fan-like movement along a circular arc in the detector emitting ultrasound pulses at about every thousandth of

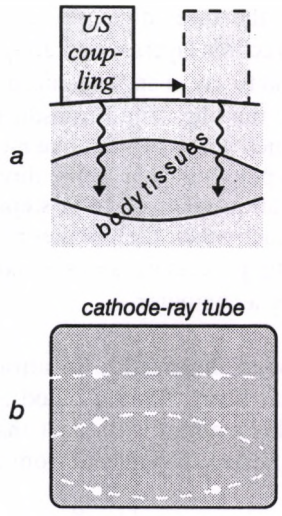


Fig. 6.59. Diagram relating to the production of a two-dimensional echogram
a: motion of the ultrasound transducer on the body surface;
b: appearance of the image elements on the screen in two positions of the ultrasound transducer

a second to scan an about 60° circular sector-shaped body section. The ultrasound echoes received in the pauses between the pulses are processed and they supply a typically sector- (or fan-) shaped image consisting of one-dimensional *B* image lines.

Electronic scanners are built of several hundred narrow piezoelectric plates. During scanning, groups of several plates simultaneously transmit and receive signals for producing a single *B* image line. (The resulting ultrasound beam is directed and focused by appropriately delaying the excitation of the members of the group.) The individual groups receive the signals one after the other: scanning is performed electronically and quickly. The transducer system is about 10 cm long, the form of the image is rectangular or trapezoid and is made up of about 150 lines.

Ultrasonic imaging nowadays is a computer-aided digital process, and this fact was the basis of the dynamic development over the past few years. The obtained image is generally stored as a set of 512 × 512 pixels, the dynamics of the echo amplitude is 6 bit from which a contrast range of 8–16 grey grades can arbitrarily be extracted.

Both mechanical and electronic scanners make possible the production of several ten images per s, which is sufficient even for the observation of the motion of cardiac valves. The images may be observed simultaneously with the recording (real time display), and may be also retrieved from the memory either as motion picture or one by one, for a more detailed observation.

4. Three-dimensional (3D) imaging and plastic display is a relatively new method in the ultrasound diagnostics. Its essence is that several (some ten) tomograms are made of the examined region and the quasi three-dimensional image is reconstructed from the processed and stored data of these.

This seems to be similar to the three-dimensional CT image. The difference (and difficulty) is that while the CT records layers which are parallel to each other in advance and the tomograms are stored so to say in an "organized" way, a series of exactly parallel tomograms cannot be made by the ultrasound transducer. As a solution, e.g. multiunit electronic scanners are used which perform a fan-like movement around an axis which is in the plane of the units and is perpendicular to the direction of the ultrasound. Rather than being continuous, the rotation takes place in steps, from layer to layer. Thus the tomograms are not parallel to each other, their planes pass through the rotational axis of the scanner head. The collecting, processing, storing and naturally the reconstruction of the echo signals is performed by a computer.

5. Doppler effect in ultrasound diagnostics. In ultrasound examinations of moving structures (the heart, the foetal heart, flowing blood), use is frequently made of the Doppler effect. The basis of this method is the fact that the frequency ν of the ultrasound reflected from a moving surface is different from the original frequency ν_0:

$$\nu = \nu_0 \left(1 \pm \frac{v'}{v}\right)$$

where v denotes the velocity of the ultrasound in the medium and v' is the component of the velocity of the reflecting surface in the direction of ultrasound propagation. The positive sign corresponds to the approach of the examined surface towards the transducer, and the negative sign to its moving away. The modified frequency ν or the frequency difference $(\nu - \nu_0)$ depends upon v'. The examination is carried out with continuous ultrasound irradiation. The transducer contains two crystals (or two ceramic plates), one of them is the ultrasound transmitter, the other one is the echo receiver. The signal voltage of the reflected ultrasound (having a changed frequency) is made to interfere with the original one. The difference frequency can easily be observed if the ultrasound oscillations of the original and the changed frequency are made to interfere. The difference frequency appears among the interference products; under the usual examination conditions, this falls in the audible range and can actually be made audible after amplification. This solution is used in a device, for instance, which transforms the beating of the foetal heart into audible sounds.

The blood flow can be studied in a similar way. This is possible, because ultrasound is scattered on the blood particles (also backwards), so the blood acts as a reflecting medium. The Doppler-sound of blood flow is well recognizable and the arterial or venous flow or a stenotic flow can be differentiated. The difference frequency is displayed on a cathode-ray tube in function of time; this is not a velocity/time graph but displays a given velocity belonging to a given frequency value.

In cardiovascular diagnostics the Doppler effect is frequently employed simultaneously with grey-scale 2D imaging. In one version the point of the examined vessel is marked in the two-dimensional real-time image with the cursor from which the above-mentioned frequency difference–time graph may be displayed with the help of a continuous Doppler procedure. Thus here the screen displays simultaneuosly a 2D image and a graph which gives information about the blood flow velocities. The 2D image is made by ultrasound

pulses, the graph by continuous radiation. Another version makes use of the fact that through the frequency difference the reflected pulses may carry information not only about the spatial coordinates of the reflecting surface but also about the velocity of the moving surface (or the circulating blood). The velocity of the motion appears in the grey-scale *B* image in a colour-coded form: the motion approaching the transducer appears in warm colours, the one moving away from it in cold colours, respectively; the colouring may even show the possible turbulences of the flow. The Doppler methods are important noninvasive tools of the cardiovascular diagnostics.

The average ultrasound intensities applied in ultrasound diagnostics are in the order of 10 mW/cm^2. It has to be emphasized that this is an average value, since it is obvious from the relation of the pulse duration (μs) and pulse spacing time (ms) that the average 10 mW/cm^2 means an intensity of 10 W/cm^2 during the 1 μs of the pulse. In spite of this, according to the experience, the ultrasound diagnostics does not seem to have damaging effects.

REFERENCES

Books

Berlien, H. P., Müller, G., Angewandte Lasermedizin. Ecomed Verlagsges., Landsberg–München–Zürich (1989)

DeMarre, M., Bioelectric Measurements. Prentice-Hall International, Englewood Cliffs, N.Y. (1983)

Geddes, L. A., Baker, L E., Principles of Applied Biomedical Instrumentation (2nd edition). John Wiley, New York (1975)

Kaufmann, L., Crooks, L. E., Margulis, A. R., NMR-Tomographie in der Medizin. Schattauer Verl., Stuttgart–New York (1984)

Kresse, H., Kompendium Elektromedizin. Siemens Aktiengesellschaft, Berlin–München (1982)

Krestel, E., Bildgebende Systeme für die medizinische Diagnostik. Siemens Aktiengesellschaft, Berlin–München (1988)

McMullan, J. T. (ed.), Physical Techniques in Medicine. Vol. I. John Wiley, New York (1977)

Millman, J., Halkias, C. C., Electronic Fundamentals and Applications for Engineers and Scientists. McGraw-Hill Book Company, New York (1975)

Serway, R. A., Physics for Scientists and Engineers with Modern Physics. (3rd edition). Saunders College Publ., Philadelphia (1990)

Sohn, C., Bastert, G., Die dreidimensionale Ultraschalldiagnostik. Springer Verl., Berlin–Heidelberg–New York (1994)

7. EXAMPLES OF PHYSICAL MODELLING: THE BIOPHYSICS OF EXCITATION PROCESSES

7.1. On modelling in general

The most important phases of the natural scientific investigations, including also modelling, are shown in Fig. 7.1. First the phenomenon or process is *perceived* and *observed*. The thorough cognition of the phenomenon usually requires a great number of systematically performed *experiments (measurements)* on the basis of which the phenomenon is described and if possible, quantitatively characterized. With the help of the experimental results *relationships* and *laws* are established, then the phenomena are *interpreted*: *hypotheses, theories* are created. By comparing the models and the experience usually *conclusions* pointing ahead may be drawn the control of which may be the starting point of further studies and experiments. Thus a new investigational cycle is initiated. Modelling may be considered as part or even synonym of the interpretational and theorizing activity.

Models have played a significant role in the development of the natural sciences, first of all in that of physics. As examples taken from physics just consider the models of the centre of mass, rigid body, perfect liquid, ideal gas or the models of light, atoms and molecules. The quantum mechanical model of the atoms and atomic systems is one of the most fruitful discoveries of this century. Modelling plays an important role in biology and medicine as a useful tool for the exploration, comprehensive recognition and interpretation of the phenomena.

The role, significance and limits of modelling will be shown in the following through concrete examples but some general statements should be already formulated:

- a good model creates connections between various groups of phenomena,
- the conclusions drawn from the model provoke new thoughts and may point out new directions of investigation,
- the models which can be formulated also mathematically are especially valuable because they can be compared quantitatively with the real system,
- computers (cf. section 8.3) have an outstanding role in modelling, since by their means several mathematical models or different solutions of one model may be examined without the necessity of exposing the examined real system e.g. to extreme experimental conditions which would be intolerable for the biological systems. This is the *computer simulation* of the examined system,
- each model is only a certain approximation of reality which requires the strict control of the conclusions,
- only some aspects of the examined phenomenon or process are highlighted by modelling, the investigation of other aspects is limited or is not possible at all.

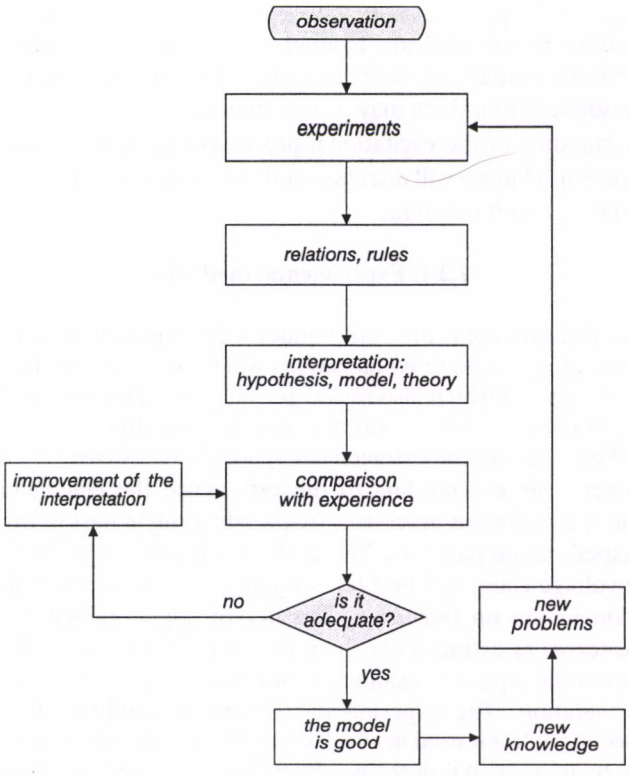


Fig. 7.1. Scheme of the aquisition of scientific knowledge

In this chapter a group of biological phenomena will be discussed concerning which a wast factual knowledge is available, therefore it is very suitable for the demonstration of the application of various models. The models built on physical knowledge have proven very productive. Excitation processes, more exactly the electric phenomena associated with them, will be dealt with; through them also the interaction between experience and theory may be demonstrated which has been an important driving force of scientific development.

7.2. Resting cells

Excitability is a characteristic of living cells at any degree of organization. It is an important condition for the adaptation of living organisms to their environment. In higher organisms excitability is primarily a striking feature of certain specialized cells or cell groups. Outstanding examples of these are muscle and nerve cells. In examinations concerning excitability the characteristics of certain groups of these cells may be studied

in vivo but *special cell groups* or a *single isolated cell* may be investigated also in vitro (by providing the conditions for survival). Examples of the in vivo studies are those of the cardiac and cerebral functions, while concerning in vitro investigations muscle or nerve bundles or cells isolated from them may be mentioned.

For the understanding of the excitational processes the characteristics of the resting cells should be known. Thus we will discuss first these, mainly with respect to their electric features, in compliance with our aims.

7.2.1. Experimental methods

In case of in vitro measurements the object under study is placed in such an environment (usually a solution of appropriate composition) which provides for the life functions of the cells or cell groups and which also makes possible the alteration of the solutions, i.e. the experimental conditions of the environment within certain limits. In fact, the solution corresponds in these cases to the extracellular space. In many cases a single cell is still a too complex system. Since according to the experience the cell membrane plays an important role in the excitation processes, frequently a single part of the cell membrane is selected for experimental purposes. This is the *patch clamp* method which is a special variation of the voltage clamp method (cf. section 7.3.2). Its sketch is shown in Fig. 7.2. Here the solution in the tip (having a diameter of approximately 1 μm) of a finely polished pipette serves as extracellular fluid. In this case the part of the membrane delimited by the tip of the pipette is examined, moreover, it is possible to change the composition of the solution. The experimental system is simplified further if a *model membrane* (cf. section 1.4.4) is used in which case not only the composition of the solution surrounding the membrane, but also the composition of the membrane can be changed according to the requirements of the experiment.

The majority of the measuring methods are "traditional" electric measurements, namely measurements of voltage and current intensity. For this purpose usually non-polarizing electrodes are used. Both electrodes may be placed on the surface of the cell or cell bundle or else, one of them may be introduced through the cellular membrane into the intracellular space. In the latter case the measurement presents the voltage between

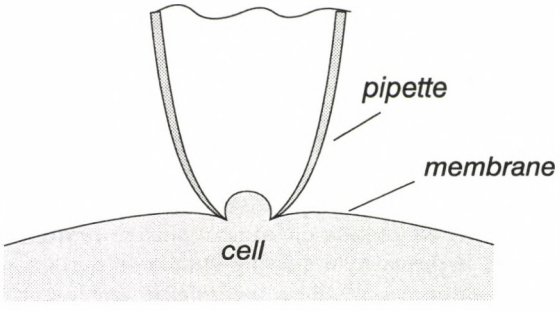


Fig. 7.2. Design of the patch clamp method

the extracellular and intracellular spaces or the intensity of the current passing through the cell membrane. This method, even in its fine form in which a glass capillary with a diameter of only some hundred nm is pierced through the cell membrane, involves the damage of the cell membrane thus influencing also the result of the measurement.

For the measurement of the electric potentials *luminescence labeling* (cf. sections 2.5 and 4.4.1) may also be used; it has the advantage that it does not cause injuries in the cell membrane. It is namely characteristic of several fluorochrome families (e.g. mesocyanines, rhodamine derivatives) that they may enter into the cell through the cell membrane and bind to the components of the cell. The point is that the electric potential in the cell determines the amount of a given luminescent substance getting into the cell and binding. The parameters of the luminescent light (e.g. emitted intensity, emission spectrum, polarization conditions) are changed by the binding, and from the changes the intracellular potential may be inferred.

7.2.2. The resting potential

If, for instance, nonpolarizing electrodes are positioned within a resting muscle cell (in the intracellular space) and on some point of the cell surface (in the extracellular space) (Fig. 7.3), a potential difference, the *resting potential,* can be measured between the electrodes. The intracellular electrode is always found to be at a negative potential with respect to the extracellular electrode. The values of the resting potential differ depending upon the cell type and the animal from which the cell originates. Even with identical cells, this value also depends upon the composition and concentration of the ionic constituents of the solution surrounding the cell. The resting potential corresponding to the normal ion composition in the intra- and extracellular spaces is generally 80–100 mV.[1] In some cases the resting potential is given as a negative quantity, for the potential of the extracellular space is usually taken as zero, when the potential of the intracellular space will be negative. In the

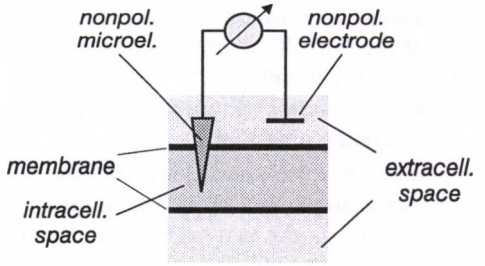


Fig. 7.3. Diagram relating to measurement of the resting potential
The convenient extracellular medium is ensured by a salt solution

[1] A potential difference of 100 mV between the two sides of a cell membrane approximately 10 nm thick corresponds to a field strength of about 10^7 V/m. At a somewhat higher field strength already breakdown ensues which is even utilized in some experimental techniques. It should be mentioned for comparison that the electric field strength of the atmosphere on the surface of the Earth is approximately 130 V/m.

following treatment the resting potential is regarded as negative only in the diagrams; the discussion relates to its absolute value.

7.2.3. Interpretation of the resting potential

Several models have been developed to interpret the resting potential (and generally the stimulatory processes). The older though still most frequently used models describe the processes *phenomenologically by thermodynamical reasoning,* which is still effective and by far not weakened by the more recent developments based on a deeper knowledge of the details of the molecular mechanism. Since the thermodynamic models connect the development of the potential with the diffusion of the ions within the cells and in the intercellular space across the membranes, they are called *electrodiffusion models.* Such are, for example, the Donnan model discussed below and the transport model. In another group of these models an attempt is made to explain the movement of the ions across the membrane as well as the blocking of the movement by the characteristic properties of the structural elements of the membrane. Since the *molecular interpretation* considers the lipid double layers and the properties of the proteins connected looser or tighter to the lipids (structural defects, pores, channels; cf. section 1.5.5), and the lyotropic liquid crystal is considered a quasi-ordered system, these models can be referred to as *solid-state physical* or *liquid crystalline* models.

The behaviour of the resting (and excited) cells – in accordance with the concept of electrodiffusion and solid-state physical models – can be characterized by *equivalent circuit models.*

In the following only the best-known electrodiffusion models providing a quantitative picture and the electric models associated with them will be dealt with.

1. Donnan (equilibrium) model. The simplest electrodiffusion model treats the living cell and its intra- and extracellular spaces as a *Donnan system* (cf. section 5.5.3) and disregards the other factors which determine the distribution of the mobile ions. (From among the latter the role of the Na^+–K^+ pump is considered important at present, cf. section 5.5.5). In Donnan model the membrane potential produced by the electric double layer developed by the presence of immobile ions and the semipermeable membrane is the basic phenomenon of the resting potential.

The living cell (as a Donnan system) is characterized by the presence of immobile ions on both the intra- and the extracellular side of the membrane; these together determine the distribution of the mobile ions on the two sides of the system (Fig. 7.4). The protein and phosphate anions of the intracellular space are immobile, and to a first approximation the cell membrane is also impermeable to the Na^+ ions which are found mainly in the extracellular space. From the aspect of the Donnan model, the mobile K^+ and Cl^- ions are the most important ions in the cells and in the intracellular space.

In this model, because of the immobile ions, the concentration of K^+ ions ($[K^+]$) is higher in the intracellular space than in the extracellular fluid, whereas for the Cl^- ion concentration ($[Cl^-]$) the situation is the reverse (cf. section 5.5.3):

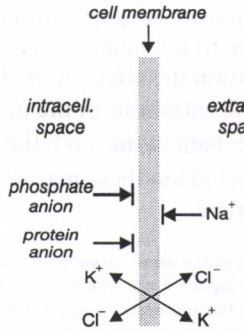


Fig. 7.4. Distribution of mobile and immobile ions in the intra- and extracellular spaces

$$\frac{[K^+]_i}{[K^+]_e} = \frac{[Cl^-]_e}{[Cl^-]_i} \qquad [7.1]$$

where the subscripts i and e refer to the intracellular and the extracellular space. The potential difference $(\varphi_e - \varphi_i)$ between the two sides of the membrane may be described by the equation

$$\varphi_e - \varphi_i = \frac{RT}{F} \ln \frac{[K^+]_i}{[K^+]_e} = \frac{RT}{F} \ln \frac{[Cl^-]_e}{[Cl^-]_i} \qquad [7.2]$$

Some results of measurements for various tissues are given in Table 7.1. Naturally, these values pertain to the living cell in which the function of the above-mentioned ion pump is also effective. Using the data of the table [7.1] and [7.2] give the closest agreement for the rat muscle, inasmuch they yield 95 mV for the K^+ ions and 86 mV for the Cl^- ions. These data approximate satisfactorily the directly measured value of 92 mV. The result is less satisfactory for the squid giant axon and the frog muscle. For these tissues the calculations lead to 91 mV and 103 mV for the K^+ potential, and to 56 mV and 89 mV for the Cl^- potential.

Table 7.1. Measured values of ion concentrations and resting potentials for a few types of tissue

Tissue	Intracellular conc. (mmol/l)			Extracellular conc. (mmol/l)			Resting potential
	$[Na^+]_i$	$[K^+]_i$	$[Cl^-]_i$	$[Na^+]_e$	$[K^+]_e$	$[Cl^-]_e$	(mV)
Squid giant axon	72	345	61	455	10	540	62
Frog muscle	20	139	3.8	120	2.5	120	92
Rat muscle	12	180	3.8	150	4.5	110	92

The Donnan model differs from the real situation in several respects:

- in contrast with the actual situation, the model regards the cell and its environment as a thermodynamically closed system and studies the equilibrium conditions accordingly (this is reflected also by the term "equilibrium model");

- the immobile ions are assumed to be perfectly immobile, and the membrane is assumed to present no obstacle to the mobile ions;
- it does not take into consideration the effect of the ion pump in the development of the extra- and intracellular concentrations of the mobile ions;
- the interactions between the membrane and the ions are not taken into consideration, though for any selected ion these may vary considerably depending upon the composition of the membrane.

It has to be mentioned here that, in spite of the above drawbacks, the applicability of the Donnan model has been recently supported from the molecular aspect. For example, in lobster muscle a good agreement was found between the voltage values calculated from the measured concentrations of the negative charges fixed on the intracellular myosin and actin filaments and those measured directly in the *A* and *I* filaments.

2. Transport model. This is also an electrodiffusion model but in several respects it contains fewer simplifications than the Donnan model (cf. section 5.5.4, point 2). The main characteristics of this model can be summarized as follows. Constant concentration differences exist between the outer and inner sides of the membrane, which results in a constant material transport across the membrane. (The model is not concerned with the processes maintaining the concentration differences.) According to transport model the migration of ions across the membrane being hindered to various extents, and hence an electric double layer is produced on the two sides of the membrane. *The resting potential is equal to the potential difference characterizing the double layer.* One of the advantages of this model is the possibility that all of the ion species on the two sides of the membrane can be considered simultaneously. Further, it also takes into account the empirical fact that the membrane is neither perfectly permeable nor totally impermeable for any type of ions. The permeability of the membrane is different for the different ions.

Consequently, it follows that the model is based on an equation describing the ion transport across the membrane (cf. section 5.5.4, relation [5.49]). For the membrane potential, i.e. the resting potential $(\varphi_e - \varphi_i)$, this equation yields the following relation

$$\varphi_e - \varphi_i = \frac{RT}{F} \ln \frac{\sum_{k=1}^{m} p_k^+ c_{ke}^+ + \sum_{k=1}^{n} p_k^- c_{ki}^-}{\sum_{k=1}^{m} p_k^+ c_{ki}^+ + \sum_{k=1}^{n} p_k^- c_{ke}^-} \qquad [7.3]$$

[7.3] is a solution of [5.49] the conditions of which were also given in section 5.5.4. The experimental basis of these conditions consists of the results of Hodgkin, Huxley and Katz. [7.3] is generally called *Goldman–Hodgkin–Katz equation.*

Since generally monovalent ions are considered in the development of the resting potential [7.3] refers only to such ions. c_{ki} and c_{ke} denote the concentrations measured in the intracellular and extracellular spaces, while the superscripts "+" and "–" refer to the cations and the anions. p_k is the permeability constant of the membrane for the k-th ion. On the basis of equation [7.3] the resting potential can be calculated using either absolute or relative permeability constants. In a given case p_k can be substituted by these (Table 7.2).

Table 7.2. Relative permeability constants of some resting cells*

Tissue	p_{Na}	p_K	p_{Cl}
Squid giant axon	0.04	1	0.45
Frog muscle	0.01	1	2

* Related to the permeability constant for potassium

With the tabulated data, [7.3] yields 61 mV for the squid giant axon and 90 mV for the frog muscle at 25 °C. The agreement between the measured and calculated values is satisfactory. Attempts to obtain even better agreement appear superfluous, for the differences are within the error of measurement.

Further *electrodiffusion* models developed to explain the resting potential with the migration of ions across the cell membrane emphasize certain characteristic properties of the molecular mechanism. Special attention is usually given to the interpretation of the membrane permeability for Na^+ and K^+ ions. With appropriate selection of the conditions, the solution of these models generally leads directly or indirectly to [7.3] or some similar relation.

7.2.4. Electrotonic potential change

1. Basic phenomena. Besides the electrodes detecting the resting potential, let us place another pair of electrodes inside a fibre and on its surface (Fig. 7.5). With the aid of these latter electrodes, electric current pulse (usually square pulses) are passed through the membrane. In this way a transient change of the membrane potential can be produced. The former electrode pair is the measuring, and the latter the exciting electrode pair. If the current direction (the direction of the shift of the positive charges) is from the surface electrode towards the intracellular space, the numerical value of the membrane potential increases; for the opposite direction, the potential decreases, or may even change in sign. The first case is called *hyperpolarization* (Figs 7.5*d* and 7.6*a*), and the second *depolari-*

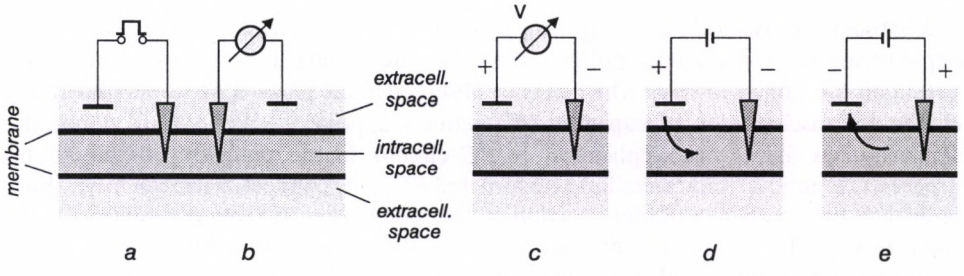


Fig. 7.5. Diagram relating to the change of the resting potential
On the left *(a)* the exciting electrode pair; *(b)* the recording electrode pair.
The distance between the two electrode pairs is of the order of a tenth of a mm.
On the right: potential conditions: *(c)* in the case of resting potential; *(d)* during hyperpolarization; *(e)* during depolarization. The arrows show the directions of the currents

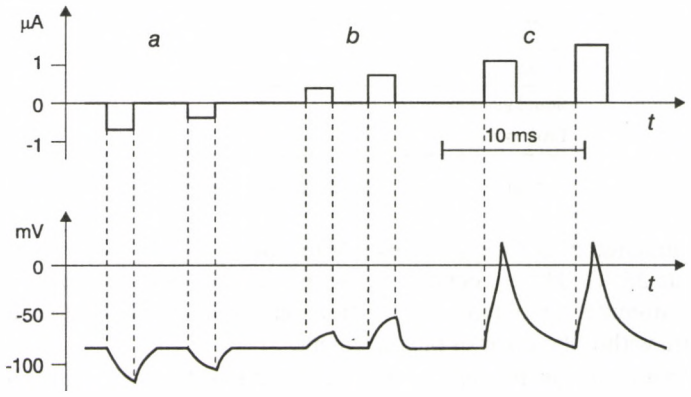


Fig. 7.6. The effect of square-wave current pulses (upper diagram) on the membrane potential (lower diagram)
The upper ordinate shows the amplitude of the current pulses, and the lower one the membrane potential. The abscissae give the time

zation (Figs 7.5*e* and 7.6*b*). The comprehensive name of these two phenomena is *electrotonic potential change.* Hyper- and depolarization as depicted in Figs 7.6*a* and *b* can be characterized in a relatively simple way: the maximum change of the voltage observed on the measuring electrode develops later relative to the exciting pulse. The difference between the excitation and response is more marked when the depolarization attains a certain value (Fig. 7.6*c*), since in this case an essentially new phenomenon is produced. A stimulus inducing only a local depolarization is a stimulus *below the threshold,* while a stimulus which by producing a depolarization induces the excitation process of the cell is *a stimulus above the threshold.* (In the example presented in the diagram the stimulus below the threshold amounts to a few tenths of a μA, and it is a square pulse with a duration of a few ms.) In this section only the electrotonic potential change is dealt with; the excitation processes of the cell will be discussed in section 7.3.

2. Modelling. According to Fig. 7.6, in case of hyperpolarization and (subliminal) depolarization – using square pulses – the curves describing the changes of membrane potentials are similar to those which may be observed in the parallel *RC* circuit during the charge and discharge of the capacitor (cf. section 6.2, point 9). Without any doubt, this similarity has led to the application of *RC* circuits in the modelling of membrane processes. Figure 7.7 shows similarly constructed *RC* units connected to each other. Each individual unit models a definite section of the membrane, of cross-resistance R_m, and capacity C_m. The individual units are connected with each other by the longitudinal resistances on the intra- and extracellular sides, R_i and R_e. Every quantity is related to unit membrane length. U is the membrane potential, U_0 is its resting value, and E is the voltage due to the concentration difference of the mobile ions on the two sides of the membrane, as obtained from the Donnan model by the Nernst equation (cf. sections 5.4.4 and 5.5.3). Consequently, using the notations of [7.2], we have

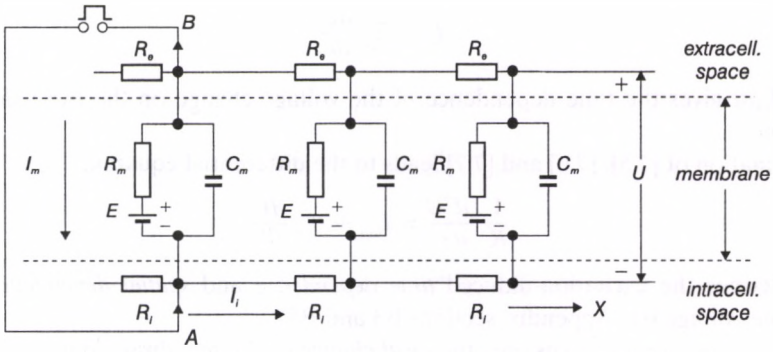


Fig. 7.7. Schematic circuit of the electric model of the cell for the interpretation of the effect of a pulse on the membrane
A and *B*: exciting electrodes

$$U = \varphi_e - \varphi_i; \qquad E = \frac{RT}{F} \ln \frac{[K^+]_i}{[K^+]_e} = \frac{RT}{F} \ln \frac{[Cl^-]_e}{[Cl^-]_i} \qquad [7.4]$$

U and E act in opposite directions. At rest (in equilibrium): $U_0 = E$.

For the electric quantities present in the model relations may be written with the help of well-known laws (e.g. Ohm's law, Kirchhoff's laws). In a given case the values of the quantities may be determined experimentally and substituted in the relations obtained from the model. A model is correct if the "theoretical" results gained this way are in a good agreement with the experienced ones. This can be said in the case of the cable model.

In the following also some specific aspects of the models will be mentioned. A current pulse induces a current both along and across the membrane. From Fig. 7.7, applying the laws of Ohm and Kirchhoff, the following relation is obtained for the current intensity across the membrane (referred to a unit membrane length, I_m):

$$I_m = \frac{1}{R_i} \frac{d^2U}{dx^2} \qquad [7.5]$$

where d^2U/dx^2 is due to the change of the modified membrane potential along the membrane (X-axis). In the derivation of [7.5] it was taken into consideration that the change of the current intensity along the membrane I_i equals the current intensity across the membrane. Since $R_i >> R_e$, it is sufficient to consider only R_i in the derivation. A parallel RC circuit being the case, I_m consists of two parts. One is the ion migration (I_{ion}) through the resistance R_m, and the other is given by the capacitive current (I_C) on the capacitor. Thus:

$$I_m = I_{ion} + I_C \qquad [7.6]$$

Every symbol denotes the current intensity referred to unit membrane length. From the definition of the capacity, I_C is given by

$$I_c = C_m \frac{dU}{dt} \qquad [7.7]$$

where dU/dt gives the time dependence of the voltage change on the two sides of the membrane.

Combination of [7.5], [7.6] and [7.7] leads to the differential equation

$$\frac{1}{R_i} \frac{d^2U}{dx^2} = I_{\text{ion}} + C_m \frac{dU}{dt} \qquad [7.8]$$

[7.8] describes the excitation-induced *time dependence* and *spatial dependence* of the membrane voltage (cf. Appendix, sections B3 and B5).

If the electric pulse induces only the *local change* of the membrane potential (cases *a* and *b* in Fig. 7.6), the value of the cross-resistance of the membrane, R_m, is constant: this is the situation until the membrane reaches the depolarization threshold level. [7.8] has a solution which describes the *time dependence* of the local change of the membrane voltage (the deviation from U_0; ΔU) at a given distance from the exciting electrode. Further, a solution is also obtained which characterizes the voltage change at a given time as *a function of the distance x from the exciting electrode.* Figure 7.8 presents some solutions of the differential equation [7.8]. Figure 7.8*a* depicts the time dependent change of the resting potential from the beginning of the exciting pulse ($t = 0$) at the position of the exciting electrode ($x = 0$), and at a distance λ from it (see below the definition of the so-called length constant, λ). It should be noted that the characteristics of the curves are similar to those of the voltage change recorded in the time period of the pulse in cases *a* and *b* in Fig. 7.6. For the voltage change after the pulse is stopped, [7.8] yields a theoretical solution similar to the declining branches of the curves in Fig. 7.6. The λ value in Fig. 7.8a is the *length constant,* which is the distance at which the pulse-induced voltage change has decreased by a factor *e* from its initial value. The τ_m value on the abscissa is the *time constant.* The value of the membrane time constant is given by the product $C_m R_m$, while the expression $\sqrt{R_m/R_i}$ gives approximately the value of the *length constant.* Figure 7.8*b* depicts the solution of [7.8] which yields the expected voltage change on moving away from the point of excitation ($x = 0$). The curve $t = \infty$ demonstrates the change in

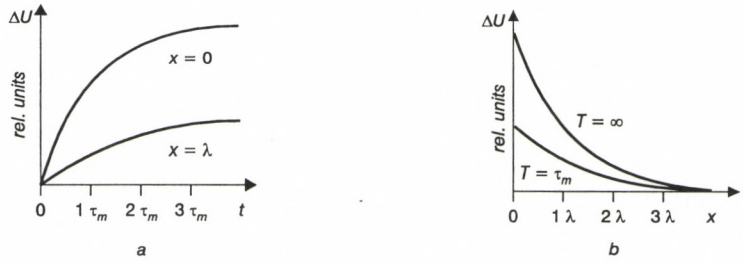


Fig. 7.8. Time course (*a*) and spatial distribution (*b*) of the local change of the membrane potential. ΔU is the change of the membrane potential with respect to the resting potential value U_0 $x = 0$ is the site of excitation, $x = \lambda$ is the membrane length constant distance from the site of excitation; $t = \infty$ relates to the development of the maximum membrane potential change, and $t = \tau_m$, to time τ_m after the excitation

the maximum value of the voltage, whereas the other curve shows the situation at the time $t = \tau_m$, after the start of the exciting pulse.

The electric quantities of the model can be determined experimentally. In Table 7.3 the quantities ρ_m, and γ_m are the cross-membrane resistance and capacity of a membrane with a surface area of 1 cm^2. ρ_i and ρ_e denote the resistivities in the longitudinal direction for the intracellular and the extracellular space. The curves constructed with these data display a satisfactory fit to the experimental curves describing the voltage change.

Table 7.3. Some characteristic data of excitable cells (20 °C)

Fibre type	ρ_i (Ω cm)	ρ_e (Ω cm)	ρ_m (Ω cm^2)	γ_m (μF/cm^2)	Time constant (ms)	Fibre diameter (μm)	Membrane length constant (cm)
Squid axon	30	22	700	1	0.7	500	0.5
Lobster nerve	60	22	2000	1	2	75	0.25
Crab nerve	60	22	5000	1	5	30	0.25
Frog muscle	200	87	4000	6	24	75	0.2

7.3. Excited cells

Curves *a, b* and *c* in Fig. 7.9 refer to responses to square current pulses for a frog muscle fibre. Curves *a* and *b* correspond to local depolarization, whereas curve *c* represents the response to excitation above the stimulus threshold. This latter phenomenon is related to the *excitation processes of the fibre.*

7.3.1. Electric properties

A change in membrane potential similar to curve *c* in Fig. 7.9 is obtained whenever an excitation process is triggered in a single fibre. This phenomenon is called an *action potential.* It is generally true that, independently of its intensity, each stimulus reaching or exceeding the depolarization threshold level produces identical action potentials.

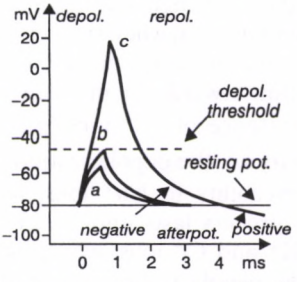


Fig. 7.9. Local depolarization response (curves *a* and *b*) and action potential (curve *c*) due to a square-wave pulse on frog skeletal muscle
The zero point on the time axis indicates the start of the action potential

The action potentials can be characterized by various data, for instance by the maximum voltage change, the *potential peak,* which in the case of a frog muscle fibre extends from – 80 mV to +20 mV as demonstrated in Fig. 7.9. Thus, the total change in this case amounts to 100 mV. The rising part of the peak is the *depolarization,* and the descending part the *repolarization.* It generally holds that the former of these two processes is the faster. From the duration of the depolarization, which is within 1 ms, and the height of the potential peak, the *depolarization rate* is found to be 10^2–10^3 V/s. Further characteristics are the *duration* of the action potential, and the presence, magnitude and duration of *after-depolarization* and *after-hyperpolarization potentials.* The action potentials of various cell types differ in the height and duration of the peak and in the re- and depolarization processes.

Of the quantities associated with the action potential, the stimulus threshold level should be mentioned; its value (in the diagram ca. 30 mV) is not constant even for a given system: it depends strongly upon the state of the system and, since it changes continuously within the physiological limits, the threshold level fluctuates at about an average value.

The stimulus threshold changes characteristically in the course of the action potential too. Of these changes, only the most striking ones are discussed here. During the period of the potential peak the stimulus threshold becomes infinitely large, which means that a new stimulus cannot induce an excitation process; from the viewpoint of excitability this is the absolute refractory period. After the peak the stimulus threshold is higher for some time than the normal one (the relative refractory period); afterwards it reaches its resting value through a strongly damped oscillation.

7.3.2. The action potential and its modelling

The development of the action potential is an extremely complex process. This complexity is due to the different changes of the membrane permeability for various ions, and also to the special modifications of the migration conditions of the individual ions. In the following section the processes will first be described qualitatively, and the possibilities of a quantitative description will subsequently be shown.

1. Action potential–membrane permeability. It is observed that every change of the membrane potential is followed by a change of its permeability. Let us follow this process in the course of the action potential. First, when the depolarization threshold is exceeded, the permeability of the membrane increases mainly to Na^+ ions. As a consequence, a large number of Na^+ ions will flow towards the intracellular space in accordance with the concentration gradient. The presence of Na^+ ions brings the negative potential of the intracellular space nearer to zero, i.e. the depolarization increases. However, this further increases the membrane permeability to Na^+ ions, and a self-amplifying process is induced *(Hodgkin cycle).* The process lasts until other effects terminate the depolarization, e.g. the migration of K^+ and Cl^- ions, which likewise increases during depolarization. The termination of the depolarization is helped also by the inactivation of the Na^+ channels. These latter phenomena are somewhat delayed with respect to the increase of the Na^+ flux, and their predominance is indicated by the decrease of the

action potential subsequent to the peak. The depolarization-decreasing effect of the K^+ and Cl^- ion fluxes can be understood easily by considering the fact that, corresponding to the concentration gradient, there is a K^+ efflux and a Cl^- influx. Thus, the effects of the two types of ions finally cause the potential of the intracellular space to become more negative. The described mechanism operates until the development of the resting state and even somewhat longer, overshooting it thereby producing an after-hyperpolarization potential (cf. Fig. 7.9). Subsequently the initial position is restored. The occurrence of this process therefore indicates that the regulating system which restores the resting potential operates by a negative feedback process (cf. section 8.2).

2. Quantitative characterization of ionic fluxes. The main question associated with the process of the action potential is how the membrane permeability and the current vary with time for different ions, and how these values depend upon the *actual value of the membrane potential.* Answers to these questions are given by the empirical results; [7.8] provides a full quantitative description of the action potential. The measurement of the time dependence of the ionic fluxes is carried out by the *voltage clamp* technique. The essence of the method is to keep the membrane potential at a fixed value during measurement, which can be achieved easily with a suitable regulating circuit.

In the voltage clamp method, besides the measuring electrodes a second electrode pair is used; the voltage setting the actual membrane potential (U) must be applied to these. If the set voltage reaches or exceeds the depolarization threshold level, an action potential will be induced, and an ion movement characteristic of this will be produced. Of course, this would change the set voltage U. The change is compensated by a voltage applied to the second electrode pair and in this case the compensating current intensity is always equal to the actual ionic flux to be measured. Besides the total ionic flux, the method allows measurement of the individual ionic fluxes. For this purpose some suitable substance must be added to the extracellular fluid, which blocks the membrane permeability for the Na^+ or the K^+ ion (e.g. tetrodotoxin or tetraethylammonium ion).

Figure 7.10*a* presents as an example the time dependence of the ionic fluxes (or more exactly the current densities) for a single constant membrane voltage, when $U = 0$ mV. Figure 7.10*b* depicts the kinetics of the membrane conductivity. This can be calculated from the results of the measurements. The diagram shows that whenever the membrane voltage rises suddenly from its resting value to above the depolarization threshold level (in the present case from – 60 mV to 0 mV) the conductivity of the cell membrane suddenly increases for the Na^+ ion and returns to the initial value only gradually. For the K^+ ion, on the other hand, the sudden change of the voltage results in a gradual increase of the conductivity; during depolarization, the conductivity stays at the increased level. Thus, in accordance with the conductivity change, the resultant flux initially consists mainly of a Na^+ influx, followed subsequently by a K^+ efflux.

The time dependences of the ionic current densities can also be determined at membrane voltages different from the case shown in Fig. 7.10. If the maximum ionic current density produced (the saturation value) is assigned to each voltage value, Fig. 7.11 is obtained; this depicts the dependences of the two most important ionic current densities on the membrane voltage. It may be seen from the curves that the Na^+ and K^+ ions differ considerably from each other in behaviour. It is obvious that there exists a

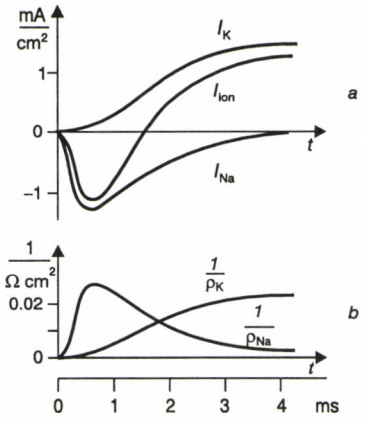


Fig. 7.10. Kinetics of ion current (*a*) and conductivity (*b*) changes produced in response to a voltage change (ΔU = 60 mV) applied to the membrane
I_{Na}, I_K and I_{ion} denote the Na^+, K^+ and total ionic current densities.
Negative values on the ordinate denote a positive ion influx, and positive values an efflux; conductivity of the membrane for Na^+ and K^+ ions denoted by $1/\rho_{Na}$ and $1/\rho_K$, respectively

membrane potential at which the movement of the Na^+ ions ceases. This voltage, the *equilibrium potential,* is approximately 60 mV in our case. Its value is equal to that calculated from the intra- and extracellular Na^+ concentrations via the Nernst equation (cf. section 5.4.4). The sign of the equilibrium voltage is opposite to that of the e.m.f. due to the concentration difference of the Na^+ ions. In other words: the equilibrium voltage compensates the e.m.f. originating from the concentration difference of the Na^+ ions, and hence impedes the movement of these ions. It should be mentioned that the equilibrium voltage for Na^+ ions in the case of the rat muscle is 62 mV, for the frog muscle 44 mV and for the squid axon 45 mV. Consequently, if the membrane voltage approaches or reaches the value of the equilibrium potential, the movement of Na^+ ions decreases and finally stops. Therefore, the equilibrium potential is an important factor in the reversion of the

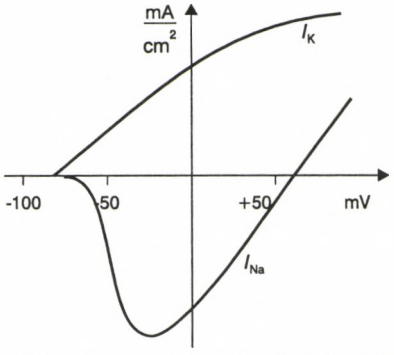


Fig. 7.11. Dependence of the maximum ionic current densities on the membrane voltage, kept at a constant value by the voltage clamp technique

self-amplifying process associated with the movement of the Na^+ ions, and it also plays an important role in promoting the predominance of the processes (the movement of the K^+ ions) leading to the original state, i.e. to the resting membrane potential.

3. The electric model of the action potential. The electric processes discussed in the previous point can be interpreted by the further development of the model considered in section 7.2.4. Let us begin with the circuit presented in Fig. 7.7, which will be amended on the basis of the previous discussion. The *ionic movement* across the membrane associated with the excitation processes can no longer be treated in a uniform way, as was the case with the local hyper- and depolarization, for during the excitation process the permeability of the membrane changes in different ways for the more important ions (Na^+, K^+, Cl^-). Accordingly, the original model circuit is supplemented by the division of both the resistance R_m and the e.m.f. E in the RC circuit. The divisions are shown in Fig. 7.12. The membrane permeabilities for Na^+, K^+ and Cl^- ions are represented by the resistances R_{Na}, R_K and R_{Cl} and E_{Na}, E_K and E_{Cl} are the e.m.f.'s produced by the concentration differences of the respective ions. In this case too C_m denotes the membrane capacitance and U is the actual membrane voltage. The movement of a certain ion is induced by the difference of the membrane voltage and the e.m.f. due to the concentration difference of the given ion between the intra- and extracellular spaces (e.g. $U - E_{Na}$). Experience shows that the resistances in the model can be varied independently of one another.

In an excited cell, similarly as for hyper- and depolarization discussed in section 7.2.4, a current flows along the fibre in the intra- and extracellular spaces as well as across the membrane. In this case too the membrane current (I_m) consists of the ionic (I_{ion}) and the capacitive current (I_C) which means that [7.8] also holds for excitation. However, in order to characterize the excitation process quantitatively, [7.8] must be supplemented with the empirical results relating to the ionic current densities. This expresses the fact that the difference between the actual membrane voltage (U) and the equilibrium potential (E_{Na} and E_K) of the given ion and the time dependence of the membrane resistances, which are different for the two most important ions (cf. Fig. 7.10*b*) together determine the values of the ionic current densities.

As in the case of the resting potential, the obtained relation has two possible solutions. One characterizes the time dependence of the action potential, and the other relates to

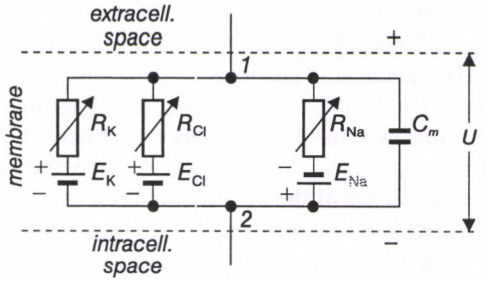


Fig. 7.12. Block diagram of one unit of the electric model suggested by Hodgkin, Huxley and Katz

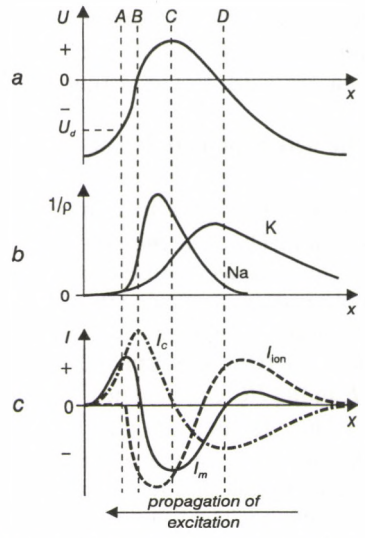


Fig. 7.13. One possible solution of the electric model of the action potential: the spatial distribution (x) of the data characterizing the state of the excited fibre at a given moment
U: actual membrane potential; $1/\rho$: conductivity; I_m = membrane current; I_C = capacitive current; I_{ion} = ionic current densities related to unit membrane length

its *propagation* (spatial dependence). In both cases the solutions yield results in agreement with experience. As an example, without giving in detail the solution of the differential equation characterizing the action potential, Fig. 7.13 demonstrates graphically the propagation of the excitation. The solution was obtained with the assumption that the rate of propagation along the fibre is constant. This assumption means that the abscissa of the coordinate system representing distance is equivalent to the time axis. Figure 7.13 depicts the spatial changes of the data characterizing the excitation process in a fibre section at a given moment. The nerve fibre in this case runs parallel to the X-axis and the excitation propagates from right to left. The dashed line C indicates the plane of maximum excitation in the fibre. To the left of this plane the excitation develops while to the right the repolarization process occurs. Diagram *a* represents the spatial change of the membrane potential. At the moment shown, the action potential peak lies in the plane C, while the planes B and D denote the sites of the steepest depolarization and repolarization changes, respectively. Diagram *b* gives the change of the membrane permeability (conductivity $1/\rho$) for Na^+ and K^+ ions, and diagram *c* depicts the change of the resultant membrane current density (I_m) and its components (I_{ion} and I_C) in the course of depolarization and repolarization, respectively. As previously, the negative current direction again indicates the influx of the positive charges.

The Figure leads to the following conclusions. An appreciable ion current is observed only from the point (A) where the membrane voltage reaches the depolarization threshold level (U_d). Before this site the ion current is very low and a considerable capacitive current can be observed. These results are in agreement with phenomena discussed in

section 7.2.4. At the potential peak (C) a large ionic influx is observed, which corresponds to the resultant of the maximum Na^+ influx and the simultaneously increasing K^+ efflux. During the repolarization the K^+ efflux predominates. To summarize, the Hodgkin–Huxley–Katz model yields results in agreement with the empirical facts, and this model can be used successfully to interpret the action potential.

4. Solid state physical models. The electrodiffusion as well as the electric models – as already pointed out – do not relate any molecular picture to the development of the resting and action potentials, respectively. The solid state physical model helps to eliminate this shortcoming. In the following an attempt is made to sketch the role of the structural factors in the development and preservation of the different intra- and extracellular ion concentrations, further on the importance of the structure in the operation and regulation of the ionic transport processes will be explored.

The results obtained with *model membranes* used to play and still are playing an important role in the development of the solid state physical approach. The model membranes are lipid–water systems where the lipid double layer is either a single lipid membrane, or the double layer forms vesicles (liposomes; cf. section 1.5.5). In these systems the lipid components, the quality of these components, and in the case of more than one lipid their proportions can be properly changed according to the experimental purpose. Many physical characteristics of these model membranes (for instance their thickness, their specific electric capacity, their permeability to water) are quantitatively the same as the corresponding data of the genuine cellular membranes. Due to the effect of suitable modifying substances (e.g. proteins, oligopeptides), the permeability of the model membranes to ions becomes similar to that of the cellular membranes. The model membranes provide a well-defined experimental system which allows to carry out physical measurements in exact (controlled) conditions and also make possible the (e.g. statistical physical) interpretation of the experimental results on a molecular level.

a) One sphere of the problems interpreted by the solid state physical models refers to the *structural factors* which contribute to the preservation of the concentration difference. The more important of these factors are the structure of the cell membranes and their constituents, respectively, further on the free or bound state of water and the ions. At present it appears that both factors participate to a certain degree in the development of the phenomenon, though the extent of their participation is still not clear. The model attributes some importance to the different behaviour of the *bonding sites* at the two sides of the membrane in the preservation of the concentration differences. This reasoning is supported by the experience that *the structure of the cell membranes is not symmetrical* since the intra- and extracellular solutions are built up to different lipids and proteins. For instance in the case of the membranes of the red blood cells at the external part of the lipid layers of the cell lecithin and sphingomyelin, whereas on the intracellular side phosphatidylethanolamine and phosphatidylserine are found. In this latter layer the proportion of cholesterol is smaller than on the extracellular side. Asymmetries are revealed also by the membrane proteins of the red blood cells.

The *differences between the proteins in the intra- and extracellular space* found on the two surfaces of the membrane contribute to the differences between the two membrane surfaces, which together result, for example, in that the ion-binding capacities concerning the sodium and potassium ions are different on the internal and the external membrane surfaces, respectively. The bound ions participate only to a small degree in the develop-

ment of the concentration gradient, or do not participate in it at all. Accordingly, the chemical potential driving the ions toward the smaller concentration is actually smaller than the value of the chemical potential as calculated from the mean value of the ion concentration in the intra- and extracellular space.

The results of structural investigations (mainly NMR and microcalorimetry) also demonstrate that a large part of the *water* in the cells should be present in *a bound state,* i.e. its structure falls between the structure of ice and free water (cf. section 1.5.1). The solubility and the diffusion velocity of the ions are smaller in the bound water than in free solutions. It should be noted here that the decrease of solubility and diffusion velocity proved to be selective, thus for instance both are larger for Na^+ due to its larger hydrate shell than for K^+. Consequently also the bound water contributes to the decrease of the energy of the active transport required to maintain the migration, and to help preserving the concentration gradient.

In the course of stimulation the changes occurring within the cell and its environment lead to the transformation of the conformation (phase transition) of the constituents, mainly the proteins. As a result, the ion-binding capacity of these constituents as well as the structure of the water undergo an abrupt change. Such effects may be produced for instance by the potential changes of 10^2–10^3 V/s during depolarization. These sudden changes (jumps) explain the phenomena related to the depolarization branch of the action potential. The conformation characteristics of the resting state and the concomitant restitution of the ion-binding capacity satisfactorily explain the repolarization period.

b) In the phenomenon of the resting and especially the action potential the *ion transport,* mainly the sodium and potassium transport, and the *molecular regulation of transport* play an important role. Interesting information can be obtained about the most important molecular mechanisms of these processes from the model membranes, since, according to experience, the permeability of the lipid membranes for individual ions modified by suitable oligo- and polypeptides (as for instance the antibiotic Gramicidin) increases considerably. The increase in permeability is related to the so-called *channel-producing* property of the protein (peptide). According to the model a specific channel operates for every ion, which may be either *passive* or *active.* The active channels are closed in resting state, ion transport does not take place through them. Transport of K^+ and Na^+ ions is going on through passive channels even in the resting state in accordance with the concentration gradient tending to the equalization of concentration differences. This equalization is prevented by the *$Na^+ - K^+$ pump,* which extrudes Na^+ from the cell while taking in K^+ using chemical energy of the hydrolysis of ATP (cf. section 5.5.5). Thus, the concentration gradients of these two ions across the membrane are maintained by the passive and the ATP-dependent active transport. – The active channels may exist in two states: permeable or impermeable. External effects (e.g. changes in the electric field, interaction with so-called activator/mediator substances) can transform these two states into each other. In resting cell membranes both the sodium and potassium channels are with high probability in the impermeable state. In case of the change of the electric field due to the stimulatory processes, however (cf. section 7.3.1), the channels open, i.e. attain a permeable state resulting in the increased transport of both ions. At the molecular level

the permeability changes of the active channels are interrelated with the conformation changes of their constituent proteins (peptides). For the examination of the functioning of the ion-permeable channels the model membranes and the membrane parts obtained by the patch clamp method (cf. section 7.2.1) are especially suitable. The task in this case is the measurement of very low (1 to 10 pA) intensities or intensity changes. As a result of such measurements it is known that some 10 to 100 sodium channels are found on a membrane surface of 1 μm^2, which means that the distance between these channels is approximately 0.1 μm. When one single channel opens, an electric current of the magnitude of picoamperes may be measured, which corresponds to the transport of 10^7 ions per second.

7.3.3. Propagation of the action potential

The action potential propagates with a definite velocity and a nearly unchanged amplitude from the site of triggering along the muscle (nerve) fibre. In the propagation of the action potential a part is played by the fact that at a given point of the membrane an action potential is produced whenever the resting potential reaches or exceeds the depolarisation threshold level. The action potential maximum appearing at a given point of a muscle fibre is reduced to about one-third of its initial value at a distance of the membrane length constant. This is quite sufficient to trigger an action potential at this more distant site too.

Table 7.3 shows that from the viewpoint of the propagation mechanism the situation for the squid giant axon, with a diameter of ca. 0.5 mm, is more favourable; if the length constant is sufficiently large, the velocity of propagation may attain even 10–20 m/s. In higher living organisms, the fibres generally have smaller diameters. However, the propagation of excitation may still be faster, due to the well-insulating myelin sheath of the nerve fibres. This is explained by the resistance of the membrane which, due to the presence of the myelin layer (between the Ranvier nodes), is extremely high and consequently the action potential propagates with practically no time loss ($\lambda \sim \sqrt{R_m/R_i}$) and is delayed only in the area of the nodes. This is the so-called saltatory conduction.

If a nerve fibre is excited in the middle, the excitation propagates in both directions from the point of excitation. However, the normal function of the nervous system is to transport the information carried by the excitation from a given point in the organism to

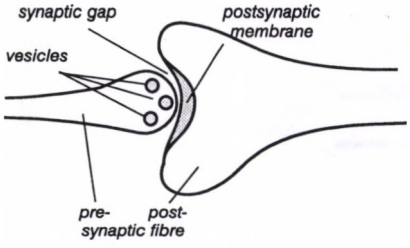


Fig. 7.14. Diagram relating to the synapse

another one. This function in the nervous system is based on the presence of a rectifying system. The rectification is carried out most frequently mediated by means of chemical substances, in morphological and functional units, the *synapses,* which are specialized to interlink the excitable cells. The synapse between two nerve cells is shown in Fig. 7.14. The presynaptic fibre establishes a connection to the dendrite of the neuron through the *postsynaptic membrane.* Between the presynaptic fibre and the postsynaptic membrane there is the synaptic gap, with a width of approximately 20 nm. In the presynaptic end are the *synaptic vesicles.* In the transmission of excitation the *neurotransmitter substance* (e.g. acetylcholine, norepinephrine) produced in the presynaptic fibre plays an essential role. The membrane of a given postsynaptic fibre is sensitive only to its respective neurotransmitter molecule. As regards their functions, two types of synapses exist: *excitatory* and *inhibitory* synapses.

The mechanism of synaptic transmission is the following: under the effect of the action potential reaching the presynaptic end the neurotransmitter passes into the synaptic gap. From here its molecules diffuse to the postsynaptic membrane and change the membrane conductivity to Na^+ ions. As a result of the Na^+ ion flux, the potential of the postsynaptic membrane changes: this is called *postsynaptic potential.* In the case of excitatory synapses this involves depolarization and in inhibitory ones hyperpolarization. According to experience the postsynaptic membrane contains a particularly high amount of proteins. The transmitter substance changes the conformation of these proteins and thus enhances the permeability of the ion channels. As a consequence of the break-down of the transmitter substance the original state of the channels is restored.

7.3.4. Action potential of fibre bundles. Dipole model

Figure 7.15 presents a situation with both measuring electrodes on the *surface* of the fibre. (The stimulating electrodes are not shown.) The direction of excitation propagation is denoted by the arrow. The recorded signal is shown in Fig. 7.15*b*. The ordinate gives the potential of electrode *A* relative to *B.* Section I–II of the curve corresponds to the propagation of the excitation from *A* to *B,* and point II is associated with the case when the excitation has already arrived at *B*. Section II–III depicts the restoration of the original situation. The action potential curves presented in Figs 7.6 and 7.9 are so-called *monophasic* action potential curves, whereas that in Fig. 7.15*b* is *biphasic* action potential curve.

Figure 7.15 depicts an experiment carried out on a single fibre, though the voltage

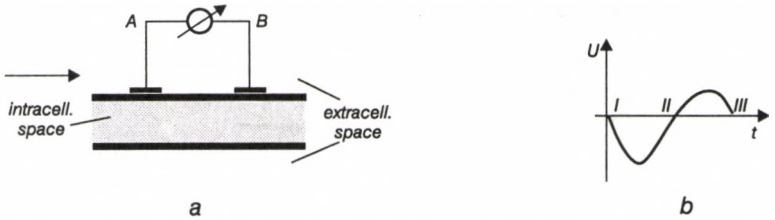


Fig. 7.15. Diagram relating to recording of the biphasic action potential (*a*) and the recorded action potential (*b*) in case of excitation propagating in the direction of the arrow

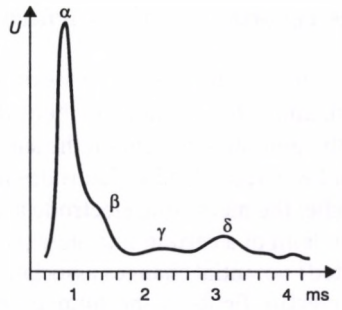


Fig. 7.16. Action potential recorded on the n. saphenus of the cat at a distance of 6 cm from the site of stimulation

signal associated with the action of a nerve bundle can be studied in a similar way. In this latter case the recorded curve usually contains several maxima, as a consequence of the various propagating velocities of the action potentials in the individual fibres of the bundle. Figure 7.16 shows the action potential recorded from cat n. saphenus, which consists of about 2600 fibres. These can be divided into four main groups, depending on their diameters. Since the velocity of propagation varies (among others) with the fibre diameter, instead of one action potential maximum four different maxima (denoted by α, β, γ and δ) can be detected on an electrode placed at a sufficiently large distance from the site of excitation. The curve may be regarded as the *resultant* of the action potentials propagating with different velocities in the individual fibres.

The biphasic action potential can easily be modelled with an electric dipole. Let us consider a single functioning fibre. The surface of its active part is at a negative potential with respect to the interior of the cell; in the resting part, on the other hand, the potential difference is of opposite sign. If the fibre surface is studied, a varying electric field can be observed on it and in its environment. Figure 7.17 illustrates the moment when the voltage change associated with the excitation process propagating in the direction of the arrow reaches the *A–A'* plane. An electric field exists between the two sides of the fibre divided by the plane. The lines of force are also depicted. The field is similar to the field of a dipole, the dipole moment pointing in the direction of excitation propagation. In the process of restoration the fibre behaves as a propagating dipole, but the direction of the dipole moment is now reversed.

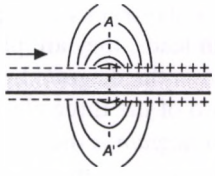


Fig. 7.17. An active nerve fibre as a dipole
In the presented case the excitation propagating in the direction of the arrow has reached the *A–A'* plane. The diagram also shows the lines of force of the dipole electric field

7.4. Voltages recorded on the surface of the body

In the living organism many electric phenomena arise in connection with the physiological activity of some organ, and can be studied without disturbing the integrity of the living organism by means of the potential present on the surface of the body. In the above discussion the action potential was recorded by electrodes placed directly on the surface of an active fibre or fibre bundle; the measuring electrode and the organ to be studied will now be separated by some medium of a certain thickness (various tissues). Below we deal only with (periodically repeated) potential changes associated with the function of certain organs. The sources of the electric fields in the human body are electric dipoles with varying moments, produced by functioning organs inside the body. On the surface of the body, only *the resultant field of many dipoles* can be studied and never a single dipole field.

In medical practice the recording of the voltages associated with the function of the heart, the central nervous system, the skeletal muscles and with vision are of the greatest interest.

7.4.1. Electrocardiography

The title of this chapter refers to a procedure used in medical diagnosis, which is based on the measurement of electric potentials resulting from the function of the heart on the surface of the body. In this chapter the more important physical aspects of electrocardiography will be dealt with (cf. also section 6.6).

The state of excitation of the heart changes in a rather complex way in both space and time. Close to the heart (e.g. over the epicardium) the spatial distribution of the potential reveals as potential sources the individual parts of the myocardium which are characteristic components of the resulting total potential. Somewhat removed from the heart (at a distance comparable with the dimensions of the heart) the details revealing the components reflecting the individual parts of the myocardium are blurred, and a potential distribution characteristic of the resultant dipole moment becomes predominant. Figure 7.18 demonstrates the field and potential distribution as developed in a given section and at a given time around the heart simulated by a single so-called *equivalent dipole.* The figure refers to a simplified situation, where the dielectric constant of the surrounding medium is regarded as equal everywhere and the conductance is considered negligible. This does not tally with reality, since the dielectric constants of lung tissues, muscles and bones are different and the conductance of these tissues is not negligible either.

1. Electric leads. The electrodes shown in Fig. 7.18 denote the usual standard limb electrodes, the so-called Einthoven leads; they are placed on the right and left arm and on the left foot. These are denoted in the diagram by the abbreviations RA, LA and LF, respectively. Let us connect any two of the three electrodes on the limbs to the voltage-recording apparatus. The *electrocardiogram* is the curve showing the change of voltage in time. In the course of a single heart cycle, maxima (P, R, T) and minima (Q, S), are recorded as depicted in Fig. 7.19. The former are the positive, and the latter the negative waves. In the first case the heart apex is at a positive potential relative to the heart base, while the situation is reversed in the second case.

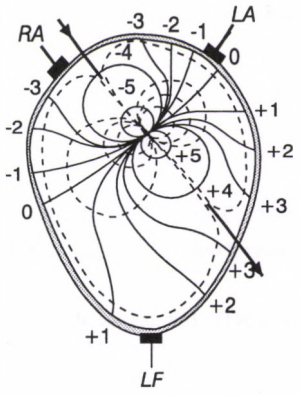


Fig. 7.18. Electric field developing in the environment of a functioning heart as an electric dipole (frontal section of trunk model)
The dotted lines denote the lines of force, and the continuous ones the equipotential surfaces. The direction of the dipole moment is shown by the arrow. The potential of the equipotential surface perpendicular to the dipole moment is arbitrarily taken as zero; the potential values of the other surfaces are denoted by positive or negative numbers (Katz, 1937)

Of the three standard limb electrodes, two can be connected in three different ways. Lead I connects the left and right arms, lead II the right arm and the left foot, and lead III the left arm and the left foot. In all three cases curves essentially similar to the first one are obtained, though the three curves are not independent of one another, since the algebraic sum of the voltages is always zero. It follows that if two of the three curves are known, the third can always be constructed.

In the evaluation of electrocardiograms, concepts defined by certain geometric consideration are frequently used. Figure 7.20 demonstrates *Einthoven's triangle.* This is an equilateral triangle whose apices denote the positions of the electrodes. The voltages U_1, U_2 and U_3 at a given time (associated with the R wave for instance), are drawn on the sides of the triangle. By convention the arrows point to the positive electrode. With a healthy heart (for the R wave) the LA electrode is at a positive potential with respect to the RA electrode, and consequently the arrow on the side of the triangle representing lead I will point towards LA. The LF electrode is at a positive potential with respect to both LA and RA, and for this reason the arrows relating to leads II and III point towards LF. The sum of the lengths of the line sections relating to leads I and III is equal to the length of the

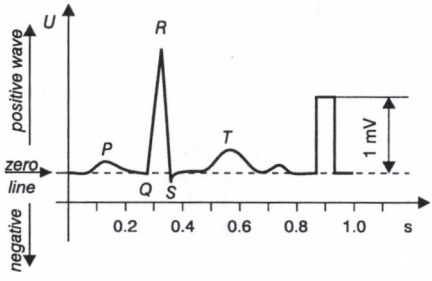


Fig. 7.19. A typical electrocardiogram

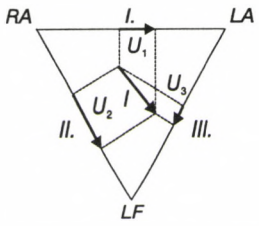


Fig. 7.20. Einthoven's triangle with the integral vector *I*

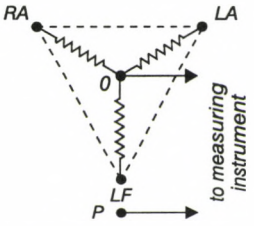


Fig. 7.21. Unipolar lead according to Wilson

line section relating to lead II, which expresses the fact that the algebraic sum of the voltages is zero. With the aid of the diagram, the heart *integral vector* (*QRS* vector), denoted by *I*, can be constructed in a simple way. This is a "vector", whose projections on the triangle sides are equal to the "voltage vectors" drawn on the triangle sides. The integral vector associated with the *R* wave, i.e. the largest vector, is the *main electric axis* of the heart.

The knowledge of two voltage vectors is sufficient for the construction of the integral vector, and the projection of this vector on the third side is equal to the third voltage vector. Without proof, however, it should be noted that this situation holds only for equilateral triangles, and this is clearly the reason why Einthoven's considerations relate to an equilateral triangle.

Besides the standard limb leads, use is frequently made in medical practice of one electrode placed at an active point of the chest, while the other electrode is connected to a point of *constant potential.* The former is the *active,* and the latter the *indifferent* electrode. In the standard limb leads both electrodes are active electrodes and in this case the leads are *bipolar.* On the other hand, if one electrode is kept at a constant potential, the lead will be *unipolar.* Similar electrode arrangements with the same nomenclature are also used in cases other than electrocardiography.

The indifferent electrode in electrocardiography is usually the *Wilson central terminal.* This is denoted in Fig. 7.21 by the point O, which is obtained if the RA, LA and LF electrodes are connected through equal resistances (5000–10,000 Ω). From experience and also from the relevant theoretical considerations, the value of the potential at O is practically constant. In electrocardiography the so-called *12-lead system* is frequently used. In this system besides the three standard limb leads, three further unipolar limb leads (aVR, aVL, aVF) as well as six chest leads are applied.

2. Special methods. The methods discussed in the following section extend the possibilities of the standard limb, and the 12-lead systems.

a) *Vectorcardiography* – as already explained by its name – determines the resultant of the dipole moment vectors produced by the heart. This vector changes continuously, both in space and time. Placing the heart in a

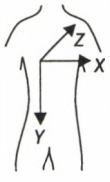


Fig. 7.22. Three-dimensional coordinate system fitted to the body

Fig. 7.23. Vectorcardiograms in the frontal (*XY*), sagittal (*YZ*) and horizontal (*XZ*) planes. The larger loop relates to the *QRS* waves and the smaller one to the *T* wave

three-dimensional coordinate system in space (Fig. 7.22) the x, y and z components of the resultant vector are measured. For the purposes of the measurement several kinds of electrode arrangements have been developed which perceive the projections upon the coordinate planes *XY*, *YZ* and *XZ* as well as their temporal changes. The result of the measurement may be displayed on the screen of a cathode-ray tube. Figure 7.23 shows such typical *vectorcardiograms*: the three closed plane curves (loops) give information about the projections of the dipole moment onto three planes.

According to the experience, vectorcardiography is considerably more sensitive than the conventional or 12-lead electrocardiography, at least in some cases. The digital processing taking the place of the direct evaluation of the records is expected to promote the propagation of the method.

b) The knowledge of the *potential distribution on the surface of the body* gives a nearly total information about cardiac action. The potential values are determined by a large number of electrodes (60–250) placed at various parts of the trunk. The results are recorded in the form of computer processed potential distribution maps.

Figure 7.24 shows a potential distribution map, representing the whole body laid out in a single plane. The horizontal line on top of the figure is the clavicle (manubrium sterni), the bottom line corresponds to the umbilical region. On the left side (ab) of the figure are represented the potential levels as measured on the frontal part, and on the right side (cd) those on the posterior part of the trunk. The thin, broken line on the ECG curve in the right of the figure indicates the time when the potential map has been determined. The contour lines connect the points of identical potential. The dotted line represents the zero potential level, while the other contour lines indicate the symmetrical potential increments in 120 μV steps, as related to the Wilson reference lead. Similar potential distribution maps record the other phases of cardiac function. The physician establishes his diagnosis on the basis of the potential maps.

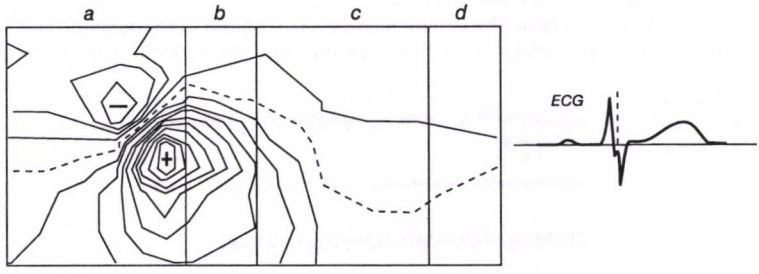


Fig. 7.24. A potential distribution map on the body surface
The time of determination is indicated by the thin dashed line in the ECG curve.
Signs + and – denote the sites of highest positive and negative potentials, respectively

7.4.2. Potentials connected with cerebral and muscular functions and with light sensation

1. Potential changes produced by the brain or its various regions can also be recorded. These changes are the *macrorhythm,* which differs considerably from the *microrhythm* produced by the action potentials appearing in the individual functioning nerve cells. Actually, the macrorhythm is the resultant of the action potentials of the cerebral neurons. The method of examining the macrorhythm associated with the function of the central nervous system is *electroencephalography (EEG),* and the recorded curves are *electroencephalograms.*

For this type of examination one active and one indifferent electrode are used. The active electrode is placed either on the appropriate point of the skull or (in surgical intervention) directly on the cerebral cortex (this latter method is *electrocorticography*). For the indifferent electrode some inactive site, e.g. the ear-lobe or the Wilson central terminal, is selected.

The recorded electric signals (waves) associated with the cerebral functions are characterized either by the *frequency* of the potential change, or by the value (amplitude) of the potential difference. Under physiological conditions, the frequency of the waves lies in the range 2–40 Hz, with amplitudes of 20–80 μV, depending upon the activity of the nervous system (wakefulness, sleep, etc.). In pathological cases the characteristic frequencies are rather low, whereas the wave amplitude may amount to several hundred μV.

Experience shows that, in the evaluation of electroencephalograms, information about the cerebral functions can be obtained, mainly from the differences between the macrorhythms of the various areas. For a correct assessment of the differences, simultaneous electroencephalograms are generally recorded for several cerebral areas. (In clinical practice 20–30-channel EEG equipments are used.) In a normal state, the records in the individual channels generally consist of the superposition of several waves of different frequencies and amplitudes (Fig. 7.25). The separation of the individual components with definite frequencies and amplitudes requires appropriate mathematical analysis by computer.

2. The voluntary function of the skeletal muscles depends upon the functions of the motor nerve and the nerve–muscle junction. Depending upon the delicacy of the function of a given muscle, a single motor nerve innervates several muscle fibres. One motor neuron together with its innervated muscle fibre constitutes a motor unit. *Electromyography* (*EMG*) studies functions connected with the motor units. This method of examination is important in the functional diagnosis of the peripheral nerves.

For the purpose of this examination either unipolar or bipolar leads are used. In the case of a measuring electrode (or electrodes) placed on the *skin surface,* the resultant of the action potentials of several motor units is recorded; with a needle electrode, on the other hand, the propagation of the excitation in one nerve fibre and its transmission from the motor neuron to a muscle fibre may be examined separately in a single motor unit. The excitation processes of the motor nerve are frequently triggered by a square-wave generator connected to the recorder (cf. section 6.3.3), which also allows the determination of the propagation velocity of the excitation.

3. The action potentials induced by illuminating the retina are examined by *electroretinography* (*ERG*). The recording electrode is usually placed directly on the frontal surface of the eye, on the cornea. In this way voltage amplitudes of 20–300 μV are obtained. The electroretinogram is produced as the resultant of several voltage components varying in time; the maxima obtained are usually called the *a, b, c, d* waves. The shape of the

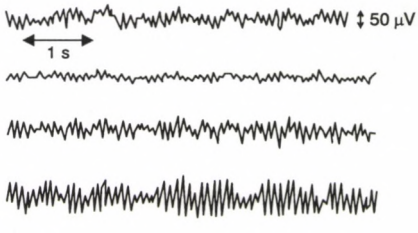


Fig. 7.25. Electroencephalograms

electroretinogram (mainly of the *a* and *b* waves), i.e. the amplitude of the waves, their duration and appearance after the start of the illumination are found to depend strongly upon the conditions of the examination (degree of dark adaptation, duration and intensity of illumination, etc.) which raises the necessity of standardization.

Table 7.4 Characteristic bioelectric potential data

Action potential	Frequency range (Hz)	Voltage (mV)	Notes
One single cell	0–10,000	50–130	Monophasic action potential
Electrocardiography	0.1–200	0.1–3	
Electroencephalography	1–70	0.001–0.1	
Electrocorticography	10–100	0.01–0.1	
Electromyography	10–1000	0.1–5	Surface electrode
Electromyography	10–10,000	0.05–5	Needle electrode
Electroretinography	0.1–100	0.02–0.3	

In this section, only certain, special problems associated with the recording of the action potentials (e.g. the construction of the measuring electrodes) have been dealt with, since the technical problems of signal shape analysis were discussed in more detail in section 6.6.1. Here we give only some informative data which may help in the selection of the electric equipment to be used to record the action potentials. Table 7.4 summarizes some of the more important data characteristic of the action potentials discussed. For the sake of comparison, the Table contains the same data on the action potential of a single cell. The frequency data were obtained by Fourier analysis of the respective curves.

7.5. Biophysical aspects of the sensory functions

The sensory functions consist of the vision, hearing, conscious and subconscious sensations of pressure and position, etc., i.e. all those processes in general by which the living organism processes the stimuli arriving from the outside world or from inside the organism. Special systems have been developed for the perception and processing of different stimuli. For instance, the system perceiving and processing light is the eye, the optic nerve and the visual centres. However, the various systems are specific only as concerns the primary processing of stimulus energies, i.e. their transformation. Subsequent to this process, every type of stimulus uniformly produces action potentials in the respective sensory nerves. The type of the energy (mechanical, electromagnetic, chemical, etc.) appearing as a stimulus, the intensity of the stimulus and its change and spatial distribution are expressed only in the localization of the action potential and in the parameters characterizing its changes.

7.5.1. Sensory functions in general

The sensory functions are discussed in this section without emphasizing the specific details; only the general aspects are surveyed. As regards the principles of operation, every system associated with the sensory functions may be modelled as an *analogue signal processing system* (cf. section 6.6.1). In this survey, only the aspects of signal processing are considered and the discussion is restricted to processes occurring in the periphery. With

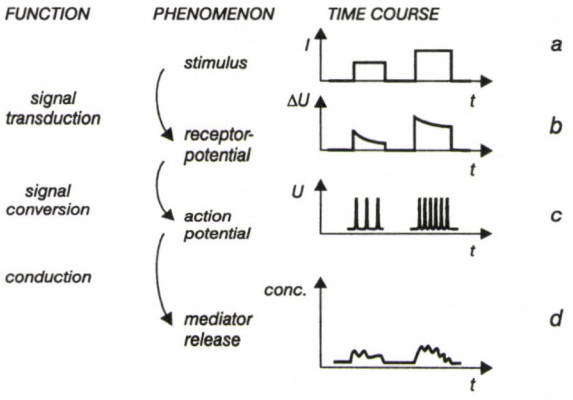


Fig. 7.26. Diagram relating to the function of the receptor cell
Time course of the processes: *a*: change of the stimulus intensity; *b*: change of the receptor potential; *c*: action potential pulses; *d*: change of the synaptic mediator substance concentration

the aid of Fig. 7.26 let us follow the chain of processes induced in the receptor cell by the stimulus, and the quantitative relations between the individual elements of the chain.

1. Signal transduction. Physical or chemical stimuli manifest their effects primarily in the *receptor cells* (or in a part of them). The receptor cell is either directly exposed to the stimulus, or is localized in a structure which makes full use of the stimulus energy and transmits it without loss. In the latter case the stimulus affects the cells indirectly. The direct case is observed, for instance, in connection with the stretch receptors of the muscles, whereas the indirect case is typical for the sense organs. The *transduction of stimulus signals into electric* signals is a general function of receptor cells, or more exactly, of the specialized membrane parts of these cells. Thus, light is transduced in the rods and cones of the retina, and sound in the hair cells of the cochlea. Accordingly, the receptor cells act as transducers. In response to the effect of the stimulus, the resting potential of the receptor cells changes considerably; this change is called the *receptor potential.* The transducer function of the receptor cell, and with it the relation between the receptor potential and the resting potential, will be demonstrated by a simple example for the case of the *stretch receptors of the muscle.* For these receptors an adequate stimulus is the stretching of the muscle fibres, which, as already mentioned, has a direct effect. As a consequence of mechanical stretching, a local depolarization is induced on the membrane of the receptor cell; with regard to the electric model of the membrane (cf. Fig. 7.7) this may be interpreted in the following way. The stretching increases the membrane surface area, and decreases its thickness (the volume of the membrane remains constant). Both dimensional changes result in an increase of the capacitance. The same dimensional changes simultaneously increase the membrane permeability, which in turn decreases the membrane cross-resistance (as concerns only the dimensional changes, the increase of the permeability is aspecific, since it holds for each ion). According to the model the capacitance increase induces a charge flow. The charges are supplied by the e.m.f. of the membrane potential, and the induced current flows through a decreased resistance. It is

quite clear that during this process the potential difference between the two plates of the capacitor is smaller than at rest, which corresponds to the depolarization observable in the given case. The resting potential is restored only when the capacitor is recharged.

It may be observed in this example, and it is also generally true, that the transducer function consists of two steps. In the first step the stimulus affects only the molecules specific for it, and in the course of this step the *energy of the stimulus* undergoes primary *transformation.* In the stretch receptor, for instance, this means simply a mechanical change in the receptor membrane structure, while in the case of the light receptor a conformational change of a photochemical nature takes place in the photolabile pigment (cf. section 2.7). The second step of the signal transduction is an aspecific one and in every case involves a *permeability change* following the structural change of the membrane. This permeability change (frequently a permeability increase) generates ionic fluxes and the receptor potential.

Lines *a* and *b* on the right-hand side of Fig. 7.26 depict the kinetics of the stimulus and the induced potential. It should be noted that the duration of the receptor potential coincides with that of the stimulus, and its value depends upon the stimulus intensity. It is generally true that the receptor potential varies with the stimulus intensity either linearly or approximately logarithmically. In the latter case it may be said that the change of the receptor potential is approximately proportional to the relative change of the stimulus intensity.

2. Signal conversion. The electric signal induced by the stimulus, i.e. the receptor potential, is converted as the next step in the processing. This transformation is an *analogue–analogue conversion* (cf. section 6.6.1, point 2). The new signal is also an electric one; it is the *action potential* of the sensory nerve associated with the receptor cell. The action potential is initiated by the receptor potential in such a way that it affects the corresponding membrane part of the sensory neuron as a stimulus. Consequently, in the sensory function the receptor potential plays the same role as, for instance, the square-wave current pulse in the artificial excitation of a nerve. Because of this function the receptor potential is also called the *generator potential.*

The generator potential may either increase or decrease the resting potential of the sensory nerve, and may generate hyper- or depolarization, respectively. The electric properties of the membrane are locally changed by a hyperpolarization and by a depolarization which does not attain the threshold level. For the generation of the action potential and for signal transmission it is necessary that the value of the generator potential should be at least as high as the threshold level of the respective membrane part. The stimulus must be strong enough to induce a sufficiently large generator potential. The smallest stimulus intensity whose generator potential can induce an action potential on the sensory nerve is the *threshold intensity.*

3. Conduction. Consider Fig. 7.26 once more. The diagram depicts a long-lasting generator (receptor) potential which generates several action potentials in succession in the sensory fibre. It is seen that the frequency of the action potential series is the higher, the larger the generator potential, i.e. the stronger the stimulus. All this can be understood

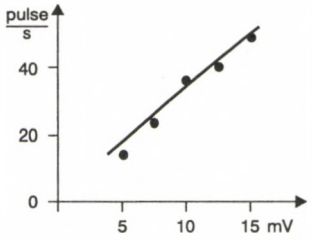


Fig. 7.27. Relation between the magnitude of the generator potential and the frequency of the action potentials generated in the stretch receptor of crab muscle

on the basis of the electric properties of the excitable cell, since in the course of the action potential the depolarization threshold level decreases after the potential peak from a relatively large value to the value associated with the resting state (cf. section 7.3.1); consequently, the larger the generator potential, the sooner the next action potential can be started. The possible upper limit of the action potential frequency, as determined by the duration of the absolute inexcitability of the nerve, is approximately 1000 Hz.

The relation between the value of the generator potential and the frequency of the action potential of the sensory nerves is demonstrated by an example in Fig. 7.27. The frequency of the action potential changes in proportion to the change in value of the generator potential. With certain receptors (e.g. the hair cells in the labyrinth) the action potentials are generated with a constant frequency on the nerve, even without the effect of a generator potential. This frequency is increased or decreased by the depolarization or the hyperpolarization caused by the generator potential. However, it is true in these cases too that the higher the change of the generator potential, the higher the change of the frequency of the action potential. If the relations between the generator potential and the stimulus intensity and between the generator potential and the frequency of the action potential are compared, it is found, in agreement with experience, that the frequency change of the action potential is nearly proportional to the relative intensity change of the stimulus. This finding, which is generally valid for the sensory functions, is the basis of the Weber–Fechner law (cf. section 6.4.1).

Let us return to Fig. 7.26. The last chain-link of the receptor cell function is the synaptic transmission of the stimulus by the liberation of the relevant synaptic mediator substance. The further processing of the action potentials developing on the postsynaptic fibre (to produce sensation) occurs at the appropriate sites in the central nervous system, where the action potential series propagating along the sensory nerve fibres finally arrive.

It should also be mentioned that in signal processing the *amplification* of the signal is an important intermediate process. Amplification connected with the sensory function is carried out in the course of the signal transduction and conversion. The energy necessary for the signal amplification is provided by the metabolic processes of the cells.

7.5.2. Hearing, as an example of sensory function

In the following, a relatively well-understood sensory function, hearing, is presented in detail as an example. The discussion will include only those physical phenomena which

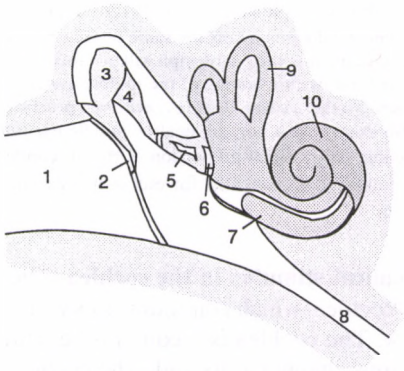


Fig. 7.28. The structure of the ear
1: outer auditory canal; *2*: tympanic membrane (eardrum); *3*: hammer; *4*: anvil; *5*: stirrup; *6*: oval window; *7*: round window; *8*: Eustachian tube; *9*: semicircular canals; *10*: cochlea

are associated with the development of the sound sensation by the processes occurring in the ear. The organ of hearing is the ear (Fig. 7.28), which perceives mechanical vibrations (sounds) in a well-defined frequency range (20–20,000 Hz). The special structural units of the ear fulfil the functions of receiving the energy of the sound stimulus and transmitting it almost without loss to the receptor cells. The ear also contains receptor cells with their respective nerve fibres. These latter structures act as energy transducers and converters.

1. Energy transmission. The *tympanic membrane* separates the *middle ear* from the outer ear. The middle ear is separated from the inner ear by an oval and a round window covered by thin membranes. The middle ear contains three small bones, the middle ear ossicles: these are the *malleus* (hammer), the *incus* (anvil) and the *stapes* (stirrup), whose base is attached to the oval window.

The sound vibrations arriving from the air are transmitted by the middle ear with only minimal energy loss to the inner ear, or more exactly to the fluid (endo- and perilymph) in the coiled tube-like organ, the *cochlea*. For a given sound wave the vibrational amplitude is smaller in a liquid than in the air. This holds for the ear too. Under these circumstances the energy transmission will be favourable if a higher pressure is transmitted to the inner ear than the pressure actually present in the open air or on the tympanum. In fact, an amplification occurs in the middle ear in which both the tympanic membrane and the ossicles participate. The pressure increase results partly from the fact that, while the surface area of the tympanic membrane part to which the handle of the hammer is rigidly attached is ca. 55 mm^2, the stirrup footplate has a surface area of only 3.2 mm^2. The same force is distributed first on the greater, and subsequently on the smaller surface area, which corresponds to an approximately 17-fold pressure amplification. Another reason for the pressure increase is the lever system formed by the ossicles; the ratio of the lever arms is 1.3:1. As a result of the middle ear function, the pressure on the stirrup footplate is 20–22 times that on the tympanic membrane. These favourable circumstances of energy transmission in the middle ear result in the more sensitive hearing of *air-conducted* sound than that conducted through the skull bones.

Pressure changes reach the cochlear fluid through the tympanic membrane–ossicles system and through the tympanic cavity as well. The vibrations of the tympanic membrane are transmitted to the air in the tympanic cavity, and through this to the round window, which intercommunicates with the fluid of the inner ear. However, the pressure amplitudes of the vibrations arriving at the round window are approximately 20 times smaller than those on the oval window, and consequently their role can be neglected in the case of a healthy middle ear. Nevertheless, the situation is quite different if the tympanic membrane and the auditory ossicles are missing. In this case the vibrations arrive at both windows through the air. Under these conditions, not only the earlier-mentioned pressure

increase is absent, which clearly results in impaired hearing, but an additional decrease will occur due to the vibrations reaching the oval and round windows in nearly the same phase (the two windows are compressed almost simultaneously, for instance), and consequently the endolymph will be moved only by the small pressure difference arriving from the two windows. The total dysfunctioning of the middle ear (with an otherwise healthy inner ear) leads to a hearing loss of 40–60 phons. In these cases hearing is achieved practically only through bone conduction.

Naturally, bone conduction also exists with a normal, healthy middle ear function. For example, we hear our own voice mainly by bone conduction; only a small proportion of the air-conducted sound reaches our own ear, which is shielded from the voice sounds. Consequently, for our own voice the loss of air conduction results in a sound decrease of only 5–10 phon.

2. Analysis of the mechanical stimulus in the cochlea. The transducer and converter functions take place in the cochlea, which communicates with the middle ear through the oval and the round window. The cochlea is a coiled tube with two and a half turns and with a narrowing membranous channel at its end. The cochlea is divided into three parts by a partly bony, partly fibrous wall (Fig. 7.29). The elastic fibrous separating wall is the *basilar membane,* and the thinner wall is *Reissner's membrane.* The former is mainly important for hearing. Its width increases from 0.04 mm to 0.5 mm from the oval window towards the apex of the cochlea. On the basilar membrane is the *organ of Corti,* which contains the endings of the auditory nerve fibres. These are connected with elongated cells covered on top with hair, the hair cells, which are further covered by the *tectorial membrane.* The vibrations arriving from the middle ear to the cochlear fluid are transmitted from here to the tectorial membrane and Reissner's membrane.

In connection with the functions of the sense organs, including those of hearing, it is essential to understand the analysis of the physical effect. In the present case this analysis involves the recognition of the physical parameters of the sound stimulus and its processing in the air. The sound analysis relates basically to the frequency and intensity, i.e. it is connected with the formation of the pitch level and sound intensity sensation. The basilar membrane participates in the first step of the analysis (Table 7.5).

In model and cadaver experiments, *Békésy* investigated the vibrations of the basilar membrane and hence gave a satisfactory explanation of the frequency and intensity analysis. He found that the movement of the stirrup, with the mediation of the cochlear fluid, induces *travelling* waves in the basilar membrane, with a frequency equal to the sound frequency. The shape of the travelling waves is influenced not only by the frequency, but also by the elasticity of the membrane, the connections between the fibres,

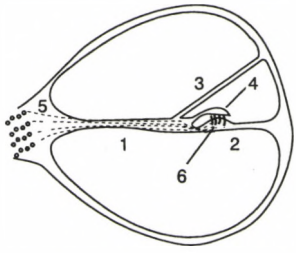


Fig. 7.29. Schematic diagram of the cochlea
1: lamina spiralis ossea; *2*: basilar membrane with the organ of Corti; *3*: Reissner's membrane; *4*: tectorial membrane; *5*: fibres of auditory nerve; *6*: hair cells

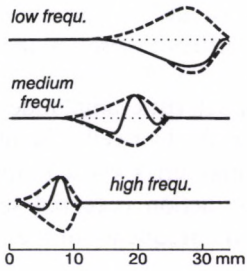


Fig. 7.30. Travelling waves developing on the basilar membrane at various frequencies
The numbers on the horizontal axis denote the distances to the oval window;
the dashed lines show the amplitude distribution (envelopes)

the friction between the basilar membrane and the surrounding medium, etc. The overall result is that the amplitude of the travelling wave varies along the membrane, even at a fixed intensity. Figure 7.30 depicts vibrations at a given time, and also the amplitude distributions. At low frequencies the maximum amplitude is formed close to the apex of the cochlea, while at sufficiently high frequencies it lies near the oval window. Thus, Békésy's experiments indicate that the frequency dependence of the location of the maximum amplitude forms the basis of the frequency analysis, though the maximum is not sharp. This, in brief, constitutes the *place theory of hearing.* The amplitude of the displacements of the basilar membrane is relatively small; it is about 10^{-11} m at sound intensities corresponding to the auditory threshold. The analysis of the sound intensity also takes place in the cochlea, since the amplitude of the mechanical vibrations and the area of the vibrating surface of the basilar membrane depend upon the sound intensity. It is of decisive importance in the hearing process that the structures (including the organ of Corti) on the basilar membrane are deformed to various degrees by the vibration of this membrane.

3. Transducer and converter function. The receptor cells are the hair cells, which thus perform the transducer function. The *shear forces* acting on these cells induce the receptor potential characteristic of the hearing process. This is the *microphone potential.*

All general statements on the characterization of the receptor potential are also valid for the microphone potential. It is found in practice that the frequency of the microphone potential is equal to the frequency of the sound stimulus. A microphone potential is produced in every hair cell located on the vibrating part of the basilar membrane. The *amplitude* distribution of the potential changes follows the amplitude distribution of the travelling waves produced on the basilar membrane. This means that the maximum of the microphone potential is found at the maximum vibrating amplitude. Thus, in the case of the microphone potential, similarly as for the basilar membrane, the excitation frequency is expressed partly by the frequency of the potential and partly by the *position of the maximum amplitude.* The *intensity* of the sound stimulus is manifested partly via the *amplitude* of the produced microphone potentials and partly via the *size of the area* where the microphone potential is actually produced.

The microphone potentials generate the action potentials in the auditory nerve endings at the hair cells. The properties of these action potentials are related to the sound stimulus analysis in the following way. For a sound stimulus of given frequency the microphone potential has a characteristic amplitude distribution; accordingly, in the nerve fibres in the environment of the hair cells, the microphone potentials of different amplitudes generate action potential series whose frequency changes from fibre to fibre. The *frequency distribution* of the action potential series and the *localization of the maximum frequency* characterize the frequency of the sound stimulus (curves 1 and 3 in Fig. 7.31).

It has already been mentioned that the intensity of the sound stimulus influences both the size of the area where the microphone potential develops and the magnitude of the microphone potential. Consequently, the *number* of active auditory nerve fibres is characteristic of the generation range of the microphone potential, while the magnitude of the microphone potential is expressed by the frequency of the action potential series (curves 1 and 2 in Fig. 7.31). As a final result, these two parameters of the action potential characterize the intensity of the sound stimulus at the auditory nerve level.

In order to measure the potentials associated with hearing, one electrode is usually placed on the auditory nerve, and the other on some indifferent area, e.g. the petrous bone. If the active electrode is sufficiently close to the cochlea, the resultant of the two potentials (microphone and action potential) is measured. However, if the electrode is at some distance, the action potential can be recorded separately. The microphone potential can then be determined by comparing the two experimental results (from the difference curves).

The individual steps of the peripheral stimulus analysis occurring in the ear are summarized in Table 7.5.

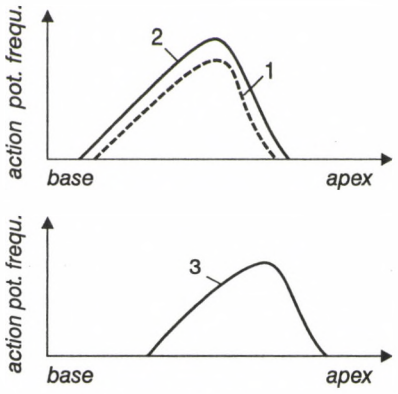


Fig. 7.31. Representation of the intensity and frequency of the sound stimulus reaching the ear in the frequency distribution of the action potentials propagating via the auditory nerve
The horizontal axis represents the positional coordinates of the fibres supplying the basilar membrane (from the base of the cochlea to its apex). *1* and *2*: sounds of the same frequency but different intensities; *1* and *3*: sounds of the same intensity but different frequencies

Table 7.5. Appearance and representation of the frequencies and intensities of sound stimuli reaching the ear in the course of peripheral signal processing

	Transducer function ↓		*Converter* function ↓
Middle ear	Basilar membrane	Hair cells	Auditory nerve
The *frequency of forced vibration* agrees with that of the stimulus	The *frequency* of the strongly damped wave agrees with that of the stimulus	The *frequency* of the *microphone potential* agrees with that of the stimulus	
	The position of the *amplitude maximum* depends upon the frequency	The *position* of the *maximum* amplitude of the microphone potential depends upon the frequency	The *frequency distribution* and the *position of the maximum frequency* are characteristic of the stimulus frequency
The *product of the intensity and the area* remains constant	The magnitude of the *amplitude* increases with the intensity of the stimulus	The *amplitude* of the microphone potential increases with the intensity of the stimulus	The *frequency* of the action potential increases with the intensity of the stimulus
	The size of the *vibrating region* increases with the intensity	The *region* in which the microphone potential appears increases with the intensity of the stimulus	The *number of active fibres* increases with the intensity of the stimulus

REFERENCES

Books

Békésy, G., Experiments in Hearing. McGraw-Hill, New York (1960)

Hoppe, W., Lohman, W., Markl, H., Zeiger, H., Biophysics. Springer-Verlag, Berlin (1983)

Kandel, E. C., Schwartz, J. M. (eds), Principles of Neural Science. Edward Arnold, London (1981)

Kozmann, Gy., Cserjés, Zs., Rochlitz, T., Szlávik, F., Data presentation problems of body surface potential mapping. In: van Dam, R. Th., van Oosterom, A. (eds): Electrocardiographic Body Surface Mapping. Nijhoff Publ., Dordrecht, pp. 127–139 (1986)

Pilkington, T. C., Engineering Contributions to Biophysical Electrocardiography. IEEE Press, New York (1982)

Sybesma, C., Biophysics, An Introduction. Kluwer Academic Publishers, Dordrecht–Boston–London (1989)

Tobias, J., Foundations of Modern Auditory Theory. Vol. 1. Academic Press, New York (1970)

Papers

Adrian, R. H., Rectification in muscle membrane. *Prog. Biophys. Molec. Biol., 19,* 341–369. Pergamon Press, Oxford (1969)

Aldoroty, R. A., April, E. W., Donnan potentials from striated muscle liquid crystals. A-Band and I-Band mesaurements. *Biophys. J., 46,* 769 (1984)

Aldoroty, R. A., April, E. W., Donnan potentials from striated muscle liquid crystals. Sarcomere length dependence. *Biophys. J., 47,* 89 (1985)

Cope, F. W., A primer of water structuring and cation association in cells. I. Introduction: the big picture. *Physiol. Chem. Phys., 8,* 479–483 (1976)

Cope, F. W., Solid state theory of competitive diffusion of associated Na^+ and K^+ in cells by free cation and vacancy (hole) mechanisms with application to nerve. *Physiol. Chem. Phys., 9,* 389–398 (1977)

Fogh-Anderson, N., Bjerrum, P. J., Siggaard-Andersen, O., Ionic binding, net charge, and Donnan effect of human serum albumin as a function of pH. *Clin. Chem., 39,* 48 (1993)

Hudspeth, A. J., The cellular basis of hearing: the biophysics of hair cells. *Science,* 230, 745–752 (1985)

Job, A., Buland, F., Maumet, L., Picard, J., VISAUDIO: a Windows software application for Bekesy audiogram analysis and hearing research. *Comput. Methods. Programs Biomed., 49,* 95 (1996)

Katz, L. N., Concerning a new concept of the genesis of the electrocardiogram. *Am. Heart J., 13,* 17–19 (1937)

Kozmann, Gy., Lux, R. L., Green, L. S., Sources of variability in normal body surface potential maps. *Circulation, 79,* 1077–1083 (1989)

Ling, G. N., The cellular resting and action potentials: interpretation based on the association-induction hypothesis. *Physiol. Chem. Phys., 14,* 47–96 (1982)

8. THE ELEMENTS OF BIOCYBERNETICS COMMUNICATION AND CONTROL

Cybernetics, one of the most rapidly developing branches of science, at the same time affects the development of practically all other sciences. It has particularly close connections to biology and medicine, and these ties may be expected to become even stronger in the future. *Cybernetics* deals with the problems of *communication (information transmission)* and *control* in highly organized systems (electronic computers, living organisms, factories, etc.), and studies the common structural and functional principles to be found in the different systems. This makes cybernetics an interdisciplinary science; it forms a connecting link between such fields as mathematics, technical sciences, biology, psychology, linguistics, ecology, etc.

Biocybernetics may be thought of as a scientific border territory in which a deeper understanding of biological phenomena and processes is sought with the aid of cybernetics, and in which the discovery of new relations in biological systems is attempted.

8.1. Information transmission

Technical communication or information-transmitting systems include the telegraph, radio and television, while biological examples are the processes of seeing, hearing, etc. In the following sections the common features of communication systems with various structures (information-transmitting chains) will be discussed. Attention will be given to the concept of information as a measurable quantity.

8.1.1. Information-transmitting systems

General structure. The schematic diagram of an information-transmitting system is presented in Fig. 8.1. Information arrives from the information source to the transmitter unit, which not only transforms the information into *signals* suitable for further trans-

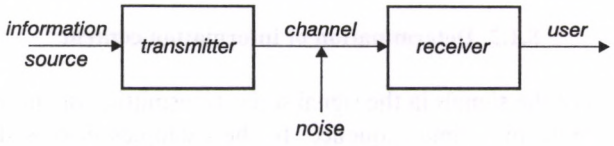


Fig. 8.1. Block diagram of a communication system

mission, but also transmits the signals which carry the information. The assignment of an unequivocal signal series to the information is *coding.* The *code* is the unequivocal system of signals (symbols) assigned by agreement. The signals are transmitted to the receiver through a *channel.* Every system which transmits information serves as a channel.

Besides the signals carrying information, signals from the environment may also reach the channel. These latter signals comprise *noise,* which may distort the original meaning of the information, thereby disturbing its correct interpretation. Information processing in which signals are transformed into directly understandable information is carried out in the *receiver.* This latter process is *decoding.*

Some examples of communication systems. As a technical example the radio may be mentioned, where the structures performing the various functions can be well illustrated, and the processes occurring in the system can easily be followed. In a radio broadcasting system the information source is the voice of the speaker, and the transmitter unit corresponds to the radio broadcasting transmitter. The coding includes all those processes which transform the sound vibrations into modulated radiowaves. The radiowaves form the channel and the radio receivers play the role of the decoding unit. This latter device transforms the modulated electromagnetic waves into sound vibrations. The user of the information is the listener.

As a biological example, let us consider hearing (cf. section 7.5.2). In this case the information source is the sound arriving at the ears. The hair cells of the organ of Corti correspond to the transmitter, for in these cells the mechanical vibration is transformed into a microphone potential inducing the action potential of the fibres of the auditory nerve. The coding consists of these processes. Subsequently, the information is carried by action potential signals, the channel being the fibres of the auditory nerve, and the receiver the hearing cortex. The information on the frequency and intensity of the sound stimulus is coded in two ways, involving the position of the active auditory nerve fibres, and also the frequency of the action potential signal of a single fibre.

A further example is from the field of biological processes at a molecular level. In the case of genetic information the information source is the DNA, whose base sequence (cf. section 1.5.4) indicates the information required by the protein-synthetizing apparatus concerning the structure of the protein. The transmitter is the DNA section (gene) corresponding to the protein in question, and the information is coded in the process of transcription. The signal combination suitable for a transmission in the given case is the base sequence of the messenger RNA, which arrives through the cytoplasm (the channel) to the receiver, the ribosome. The decoding, i.e. the translation, is carried out here, resulting in a protein with primary (and higher order) structure corresponding to the information induced by the gene.

8.1.2. Determination of information content

The configuration of the signals in the signal series transmitting the information is either a spatial arrangement or a time sequence. In the examples discussed above the base sequence of the DNA molecule represents the first case, and the modulation of the

carrier electromagnetic waves in a radio the second possibility. In the process of hearing associated with action potential signals, both arrangements are present.

The quantity of information can be determined exactly in two steps; these are discussed in points 1 and 2 below. Here only one basic idea will be emphasized: a configuration contains the more information, the more unexpected it is. In a different formulation: the information content is the greater, the smaller the probability of occurrence of a given configuration.

1. The uncertainty of the experimental results. The entropy of the experiments. In any given case there are generally various possibilities for the content of a communication, the result of some observation or the outcome of an experiment; the actual content, result or outcome (in general the outcome) is usually not known in advance: the outcome is uncertain. For instance, one cannot tell in advance the number of the lottery ticket which will be drawn, and similarly it is impossible to know in advance in which way the actual DNA molecule will be built up from the four different nucleotide bases. The uncertainty concerning the outcome of the individual experiments and observations (subsequently: experiments) may be characterized quantitatively in the following way.

a) First we are dealing with an experiment which has k outcomes of *equal probability.* An example of this is lottery ticket drawing, since each ticket participates in the game with the same probability. Clearly, the uncertainty is the greater, the higher the number of possibilities of the outcome. If k equal outcomes of an experiment α are possible, the uncertainty of the experimental result, denoted by $H(\alpha)$, can be characterized by $\log k$, i.e.

$$H(\alpha) = \log k \qquad [8.1]$$

Definition [8.1] expresses the empirical finding that the uncertainty increases with the number of possible outcomes, and the fact too that for one possible outcome the result is beyond doubt, i.e. there is no uncertainty at all: according to [8.1] $H(\alpha) = 0$ when $k = 1$.

Let us calculate the uncertainty of the nucleotide base sequence for a DNA molecule consisting of 10^6 bases. At any site of the molecule any one of the four different bases may occur, and consequently the number of possible sequences will be $4^{(10^6)}$ (repeated variation). Assuming that the probabilities of every possible sequence are equal, [8.1] may be used. In this case $k = 4^{(10^6)}$, so that $H(\alpha) = 10^6 \log 4$.

b) We now discuss experiments whose possible outcomes occur with *different probabilities.* We start from the above special case (equal probabilities of outcomes) and transform [8.1] to make it suitable for generalization (the outcomes have different probabilities).

It is clear that

$$\log k = \frac{1}{k}\log k + \frac{1}{k}\log k + \ldots + \frac{1}{k}\log k \qquad [8.2a]$$

where the right-hand side consists of the sum of k terms. Since

$$\log k = \log \left[\frac{1}{k}\right]^{-1} = -\log \frac{1}{k}$$

the right-hand side of [8.2a] can also be written in the form

$$\log k = -\frac{1}{k}\log\frac{1}{k} - \frac{1}{k}\log\frac{1}{k} - \ldots - \frac{1}{k}\log\frac{1}{k} \qquad [8.2b]$$

The right-hand side of [8.2b] may be conceived in the following way. Let us denote the possible outcomes of an experiment by $A_1, A_2 \ldots, A_k$, or briefly by A_i, where $i = 1, 2, \ldots, k$, and the probabilities of their occurrence by $P(A_1)$, $P(A_2),\ldots,$ $P(A_k)$, or briefly by $P(A_i)$, where $i = 1, 2, \ldots, k$. If the outcomes A_i occur with equal probabilities, $P(A_i) = 1/k$ ($i = 1, 2, \ldots, k$), and [8.2b] can be written as

$$\log k = H(\alpha) = -P(A_1)\log P(A_1) - P(A_2)\log P(A_2) - \ldots - \ldots - P(A_k)\log P(A_k)$$

or more concisely

$$H(\alpha) = -\sum_{i=1}^{k} P(A_i)\log P(A_i) \qquad [8.3]$$

The right-hand side of [8.3] can also be calculated when the probabilities of the various outcomes are different. The value obtained is termed in every case (on certain physical analogies) the entropy of the experiment. By definition the *uncertainty of the outcome of an experiment* is characterized by the *entropy of the experiment.* Obviously [8.3] includes [8.1] as a special case.

It was assumed in the above DNA example that the possible base sequences occur with equal probabilities. In reality the probabilities of the possible sequences are not equal, and consequently the data obtained via [8.1] do not give the entropy correctly. Our knowledge on the probabilities of the individual configurations is at present still far from complete, but it seems to be obvious that the actual uncertainty is smaller than that obtained from [8.1], since the following statement is true in general: *the most uncertain of all experiments with k outcomes will be that one whose possible outcomes have equal probabilities.*

The correctness of this statement will be demonstrated in the simplest case, when only two outcomes (A_1 and A_2) are possible. Let us denote the probabilities of the outcomes by $P(A_1)$ and $P(A_2)$. Since one of the two outcomes is sure to occur, $P(A_1) + P(A_2) = 1$. Thus, for any value of $P(A_1)$, $P(A_2)$ can be calculated, and from these two data the entropy $H(\alpha)$ of the experiment in question can be obtained from [8.3]. The results of the calculations are depicted in Fig. 8.2. The uncertainty is indeed maximum at $P(A_1) = P(A_2) = 0.5$, when the value of the uncertainty is 0.3 (cf. left-hand ordinate).

2. Quantity of information. If the expected result is obtained in an experiment, the resulting information is considered to be less than in the event of an unexpected outcome. Consequently, the information is characterized by a quantity which for some outcome A_i of the experiment is the smaller, the larger the probability $P(A_i)$ of the occurrence of A_i, and vice versa. By definition: *the information obtained on the occurrence of an outcome A_i of an experiment is characterized by the quantity* $\log [1/P(A_i)]$ i.e. by $-\log P(A_i)$.

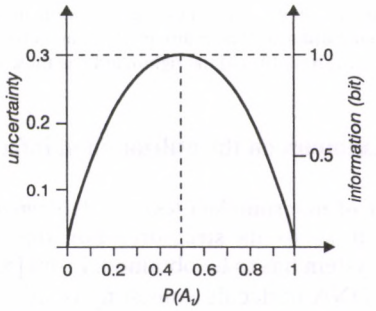


Fig. 8.2. The uncertainty of an experiment with two possible outcomes as a function of the probability of occurrence of one outcome $P(A_1)$
The right-hand ordinate shows the quantity of information to be obtained from experiments with two outcomes

In practice, not only the individual information quantities of the *individual outcomes* of an experiment are important; the calculations frequently involve the information content of the *respective experiment.* The information content of an experiment is characterized by the average (expected) value of the individual information quantities, which will be shown to be equal to the entropy of the experiment.

Let us denote, as previously, the possible outcomes of an experiment by $A_1, A_2, ..., ..., A_k$, and the associated probabilities by $P(A_1), P(A_2), ..., P(A_k)$. Let us assume that if the experiment is repeated N times, the number of occurences of outcomes $A_1, A_2, ..., A_k$ will be $n_1, n_2, ..., n_k$, respectively $\left[\sum_{i=1}^{k} n_i = N\right]$. The average (expected) value of the information obtained in the individual experiments will clearly be

$$\frac{-n_1 \log P(A_1) - n_2 \log P(A_2) - ... - n_k \log P(A_k)}{N}$$

If N is sufficiently large, then by the interpretation of probability n_i/N can be considered equal to $P(A_i)$; consequently, the above relation can be written in the form

$$-\sum_{i=1}^{k} P(A_i) \log P(A_i) \qquad [8.4]$$

which is identical with the entropy of the experiment. This means that the entropy of the experiment is *equal to the average value of the information contents of the individual outcomes of the experiment.*

Thus, the information content of the previously mentioned DNA molecule is $10^6 \log 4$.

In practice, instead of logarithms to the base 10 calculations are made with logarithms to the base 2. Consequently, an experiment has unit information content if its entropy calculated in a logarithmic system to the base 2 is 1. The unit information content obtained with logarithms to the base 2 is called one *bit.* One bit of information is obtained from an experiment which has two equally probable outcomes. The information content of the above DNA molecule will thus be 2×10^6 bits.

Figure 8.2 also demonstrates the information content of the experiment with two possible outcomes. This is indicated in bit units on the right-hand ordinate. The right- and left-hand ordinates differ only in their scales; the left-hand one gives the logarithms to the base 10, and the right-hand one those to the base 2, of the same numbers.

8.1.3. Examples on the utilization of information

1. The information content of macromolecules. If a *biological macromolecule* is known, there is no uncertainty left as to its structure. For this, information equal to the information content of the system has to be obtained. From [8.4], the information content of the base sequence of a DNA molecule consisting of 10^6 nucleotide bases is at most 2×10^6 bits. This is the amount of information to be collected by suitable physical and chemical methods to establish the base sequence of the DNA molecule, which determines the higher order structure of the molecule, too. The information content of the amino acid sequence of a protein molecule can be calculated similarly. Proteins are built up of 20 different amino acids, and thus the information content of the structure of a protein molecule consisting of, for example, 500 amino acids will be approximately 2×10^3 bits. This value is lower by three orders of magnitude than that found in the example of DNA, but the collection of even this information still presents a formidable task, because the determination of the sequence requires a number of experimental steps of the same order of magnitude (or only slightly less) as the number of bits of information to be collected.

Our example explains the well-known fact that research work succeeded first only in the sequencing of relatively small proteins and nucleic acids or nucleic acid fragments. One of the first structures determined was that of insulin, consisting of 51 amino acids, whose information content is merely approximately 200 bits.

Recently the base sequencing has considerably quickened up. The already revealed DNA segments consist of 10^4–10^5 nucleic acid bases, e.g. they contain 10,000–100,000 bits of information. For example a large number of small-size viral genomes and numerous important chromosome parts have been base sequenced. It should be mentioned here that the human chromosome consists of 3×10^9–4×10^9 bases. With the application of the "traditional" chemical methods such an amount of information could have been obtained only in a very long time, in several decades. Thus up to 1991 the sequences of only 3000 genes of the human genome were known. The introduction of a combined genetic–biochemical method meant a great breakthrough; it increased considerably the amount of information yielded by the individual experiments. At the beginning of 1995 the sequences of about 5000 genes became known, at the end of the same year their number increased to 25,000. The base sequences of 25,000 genes represent the sequences of roughly 80 million (8×10^7) bases of the human genome.

2. The estimation of genetic information. As already mentioned, *genetic information* is stored by the nucleotide base sequence, and the protein-synthetizing system of the cells synthetizes the required protein in accordance with the base sequence. In the case of various viruses, the information necessary for the building-up of the virus proteins is stored in a single nucleic acid. Let us calculate for *how many different proteins* is information carried by the nucleic acid of a simple virus, bacteriophage MS2, consisting of 3.3×10^3 bases. The calculations are based on the following reasoning: the number of types of proteins that can be produced will be the number of times the information content of a molecule of average size to be found in the nucleic acid in question. Since an average protein (consisting of 500 amino acids) contains approximately 2×10^3 bits of

information, phage MS2 with its approximately 6.6×10^3 bits allows the production of 3 types of proteins. Empirical results indicate that this calculation is correct. It is found in experience that for small viruses containing only a few bases (e.g. bacteriophage ΦX174), a given DNA part may contain information relating to two different proteins, i.e. the codes of the two proteins overlap each other in the nucleic acid. The above estimation of course does not consider this case; it yields only the *lower limit* of the possible number of proteins.

3. Information flow. The *capacity of an information communication channel* is characterized by the maximum information transmitted without noise per unit time. This is the information flow.

The capacity of a *single nerve fibre as an information channel* is less than 10^3 bits/s. This estimation is based on the fact that at most 10^3 action potential signals are transmitted through one fibre per second (the duration of an action potential is approximately 1 ms), and that the information is carried by the presence or absence of the action potential.[1] This alternative means a maximum of one bit of information. The above capacity value is obtained in this way. The capacity of a fibre bundle is given by the sum of the capacities of the individual fibres. The human organism receives about 10^{10} bits of information per second from the external world via the sense organs. This at first sight vast information flow is obtained by the following estimation. We consider only the information received by the eye. Each receptor cell (rod) ensuring twilight vision is found to be able to distinguish 32, i.e. 2^5 different brightness grades. Thus, on the appearance of a single image, one rod yields 5 bits of information. Since the human eye has to perceive one image for at least 1/16 s, and since the retina contains approximately 10^8 rods, the total information flow via the rods amounts to 10^{10} bits/s.

It may be mentioned (without going into details) that the cones contribute to this quantity with an information flow of approximately 10^9 bits/s. (It can be shown as well that in case of hearing the information flow is about 5×10^4 bits/s.)

Only ca. 100 bits/s of the information reaching the organism is consciously perceived, the rest being selected out. (Of this consciously perceived information 10 bits/s are stored for a short time in the central nervous system and 1 bit/s for a long time.) The selection may be achieved in various ways in a given organ. In the case of the eye, it begins with the fact that the number of *n. opticus* fibres is only of the order of 10^6, while the order of the number of receptor cells is 10^8. Since the channel capacity of one nerve fibre is less than 10^3 bits/s, the total capacity of the *n. opticus* is less than 10^9 bits/s. (Another comparison: in case of analogous postal telecommunication lines the forwarded information current is maximum 56 kbits/s, in digital ones 128 kbits/s, while the data transmission rate between personal computers is 100 Mbits/s.)

[1] The action potential in the present case is treated as a digital signal, which is not inconsistent with our treatment in section 7.5.1, where the same action potential was considered from a different aspect as an analogue signal.

8.2. Control

A considerable proportion of the processes in the various technical, biological, etc. systems occur in an ordered, i.e., in a coordinated and controlled way. For instance, the adaptation of living organisms to their environment or the production in a factory is accomplished by control. This control is carried out either via conscious elements or without them (automatic control). In the examples considered both types occur, and in practice we are usually concerned with such cases.

The two basic functional elements of a control system are the controlled and the controlling units (controller). In the previously discussed examples both the controlled units and the control centres may be of many types. Thus, in the more highly developed living organisms the control centre for adaptation is the central nervous system, while in the factory the centre is the director, who may carry out production control with the aid of a computer centre.

Each type of control is based on the *transmission of information.* Information arrives at and departs from the control centre. The arriving information may originate from sources outside the control system, but from inside the system itself, too. However, information of the opposite direction always flows from the centre toward the controlled units.

The concept of control includes two types of processes: *simple control* (without feedback) and *control with feedback* or *regulation.* In control without feedback no information reaches the control centre from the controlled unit, that is the process does not react upon the control. In the case of regulation, on the other hand, there is feedback (negative feedback) and it is this feedback which enables the controlled system to serve the preset goal. In the following we shall deal with regulation in detail, since this process plays a fundamental role in the function of biological systems.

8.2.1. Regulation. The functional scheme of regulating systems

The structure and function of regulating systems. The task is to ensure that a given parameter of a system (e.g. temperature, pressure, the concentration of some chemical component) should always assume a prescribed value. We speak of *constant value control* if the value of the parameter has to be kept at the same level. For instance, the maintenance of the constant temperature of a thermostat is carried out by this type of regulation. The regulation of the body temperature, blood pressure, pH value, etc. of living organisms is of the same type. In *sequential control* the parameter value changes in accordance with the actual requirement. As a biological example, repressive regulation of enzyme synthesis may be mentioned. In the case of a *time-schedule control,* the parameter changes according to a preset programme. A time-schedule control operates to ensure for example the heating rate of a thermostat with gradually increasing temperature, or the progress in the ontogenesis of a living organism.

The simplest control systems contain a single feedback system (these are the *one-loop regulating systems*). The general structure of these systems and the functions of the individual functional units (the functional block scheme) are depicted in Fig. 8.3. The arrows denote the connections between the units and the directions of information

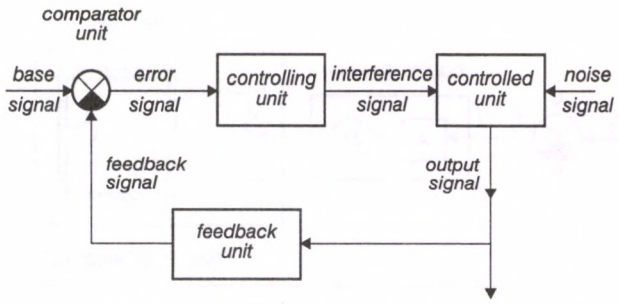


Fig. 8.3. Functional diagram of a simple regulating system

transmission. Communication is achieved through the input and output signals of the units. (In biological regulating systems, information is always carried by analogue signals.) The parameter characterizing the actual state of the regulated system is the *output signal,* while the *base* or *setting signal* gives the preset parameter value to be maintained by the system. Besides the *controller* and *controlled* (and *feedback)* units, every control system also contains a *comparator unit,* into which the base signal and the *feedback* or *control* signal (changing together with the signal of the controlled parameter) arrive. The comparator unit forms the difference of these two quantities[2] (negative feedback), and if it is different from zero the difference will pass back into the controller unit as an *error signal;* this in turn influences the controlled system by means of an *interference signal.* Besides this signal, other not negligible noise signals may affect the system. Naturally, the latter also influence the change of the output or feedback signals.

Examples of regulation. In this section we investigate the operation of some technical and biological regulating systems in order to demonstrate the actual performance of the functional units of the schematic diagram and the information transmitted and processed by them. Figure 8.4 depicts the functional diagram of an *electric thermostat* operating as a *constant value* regulating system. The actual temperature of the thermostat is measured with a thermocouple *T,* for instance. The thermovoltage (U) supplied by the thermocouple is the output signal, which is compared by feedback with the voltage U_0 corresponding to the required constant temperature. This function is performed by the comparator unit (C) which produces the difference of the two voltages ($U_0 - U$). After suitable amplification (A), the voltage difference operates an electric motor (M). (The amplifier and the electric motor together constitute the controller unit.) The rotation of the motor supplies the interference signal, which results in a displacement of the sliding contact of a variable resistor (R) connected in series with the thermostat heater (H). The motor turns in one or the other direction, depending on whether U is larger or smaller than U_0, and accordingly the contact slides to the required position. If $U > U_0$, the sliding contact moves to increase the circuit resistance, resulting in a decrease of the heating current and

[2] By convention, the difference-forming function of the comparator unit is indicated in the drawing by blackening a quarter of the circle representing the unit.

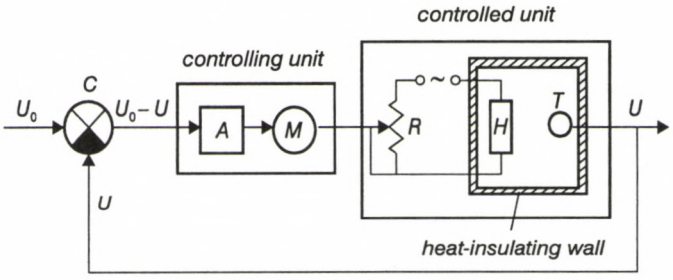


Fig. 8.4. Block diagram of an electric thermostat

together with this the temperature of the thermostat. If $U < U_0$, however, the reverse process takes place. The sliding contact moves to and fro as long as the temperature deviates from the preset value.

Regulation as an organizing principle can be found at every level of the organism, from a molecular level up to sophisticated organ systems. Figure 8.5 depicts the functional diagram of a biological regulation at a molecular level, the regulation of enzyme synthesis. For the sake of better understanding, the more refined details of the mechanism are omitted, and only the essential, general features important from the aspect of regulation are emphasized. The controlled unit is the protein-synthetizing system (R), whose functional state may be characterized by the concentration within the cell of the enzyme of interest. This concentration is the output signal. The base signal is the *repressor substance level,* which determines the *optimum enzyme level* required by the life conditions of the cell. With changing life conditions the base signal changes too, and the regulation must follow these changes. Thus, the present case represents sequential control. The feedback signal is the concentration of the substrate of the enzyme in question, which is determined by the respective enzyme concentration. More enzyme molecules break down more substrate molecules, and consequently a higher enzyme concentration is associated with a lower substrate level and vice versa. The feedback signal and the base signal, i.e. the substrate level and the repressor substance level, are compared in the following way. If the appropriate operator gene interacts with the repressor substance, no transcription

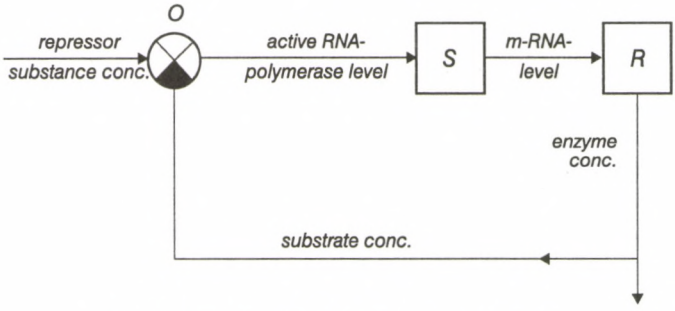


Fig. 8.5. Functional block diagram of enzyme synthesis

takes place on the associated structure gene. However, if the substrate molecules inhibit the coupling of the repressor substance with the operator gene, DNA polymerase becomes activated and mRNA will be synthetized according to the structure gene base sequence. Thus, the operator gene (*O*) is the comparator unit and the error signal will be the active RNA polymerase level. It is clear that the structure gene (*S*) plays the role of the controller unit. The enzyme synthesis in the protein-synthetizing system (ribosome) proceeds in accordance with the mRNA copied from the structure gene. This means that the larger the error signal, the more mRNA will be synthetized, and as a consequence the synthesis of the enzyme in question will increase. The interference signal is the level of mRNA encoding the enzyme in question.

Another relatively simple example of biological regulation at a molecular level is the control mechanism acting in the *development of phage MS2.* This phage consists of a single RNA molecule surrounded by an envelope containing ca. 180 identical protein molecule. The regulation is related to the template and messenger functions of the RNA molecule. The former function means that in the course of development the single RNA molecule of the phage, having entered the bacterium cell, serves as a template for the synthesis of further phage RNA molecules. A fundamental role in this process is played by the own RNA polymerase of the phage; this too is formed in the intrabacterial phage development and one gene of the phage contains the information for its base sequence. The RNA synthesis stops when no more RNA polymerase is synthesized on the basis of this gene (and the already existing molecules decompose). The messenger function of the RNA is the synthesis or the phage coat-protein; however, this becomes important only after the production of a sufficient number of RNA molecules. The genetic information necessary for the coat-protein synthesis is also supplied by the RNA, or more exactly by its appropriate gene.

In our example the *regulation* means that the phage RNA synthesis should stop at the right moment, while the production of coat-protein molecules in the required quantity and intensity is ensured. In the example, two different kinds of mechanism operate: one of them controls the template function and the other the messenger function. However, it is sufficient to consider the template-forming system as a controlled system, since the messenger function is determined by the template function. The number of produced phage RNA molecules provides information about the state of the template-forming system; this is the output signal. Depending upon the physiological state of a bacterial cell, more or fewer phages may be synthetized. The number of phage RNA molecules is the *base signal.* An appreciable quantity of coat-protein can obviously be synthetized only after the formation of a sufficient number of RNA molecules, though a small amount is produced earlier, since the number of protein molecules increases together with that of RNA molecules. The *feedback signal* is not the number of produced RNA molecules, but the number of protein molecules increasing together with that of RNA. The *comparison* of the base and feedback signals is carried out by the starting region of the RNA polymerase gene, since the coat-protein molecules interact with the starting region of the RNA polymerase until their number is small, thereby inhibiting the synthesis of RNA polymerase. Consequently, the number of regions in the bacterium which permit polymerase synthesis gradually decreases, and this decrease serves as the error signal. Through the decrease in the number of polymerase molecules, the RNA polymerase gene influences the template-forming, i.e. the controlled unit. From the above it is quite obvious that the decrease in the number of polymerase molecules acts as an *interference signal,* and the RNA polymerase gene operates as a *controller unit.*

8.2.2. The study of regulating systems. Transition functions

The concept of dynamic analysis. Technical control systems are known in advance, since they are usually constructed from elements with well-known properties, and the elements are put together according to purpose. Under these circumstances it is understandable that the response of the system (output signal) to known external effects (input signals) can be determined quantitatively in advance. The reverse case is also conceivable, when

conclusions concerning the effect produced on the system are drawn from the response. The situation is usually more difficult with biological control systems. The details of these systems are generally little known, and the main task is to establish the structure of the system and the properties of its elements. On the other hand, the object of our investigation may be restricted to determining the response of the system to some external effect under given circumstances, or to determining what effect produced a given response. For instance, the consideration of the possible effects of some therapeutic intervention involves studying the possible responses of the biological system to a known external effect. In medical diagnostics, on the other hand, from the change in a characteristic parameter of a biological system with well-known properties conclusions are drawn on the effect inducing this change.

In the study of a system the black box approach is a frequently used model. Any system may be treated by this model if its internal structure is not taken into account (or even if it is not known), and only the relations between the input and output effects (input and output signals) are investigated. This type of treatment is the *dynamic analysis.* It follows from the earlier discussion that the black box method, i.e. dynamic analysis, is quite frequently applied in the case of biological systems.

Example of dynamic analysis. Dynamic analysis is generally a rather complex method, since the testing of the system is carried out in various ways, using different input signals. Figure 8.6 shows some of the most frequently used testing signal forms. The horizontal axis of the diagram denotes time, while the ordinate relates to the interfering effect, that is to the change of some parameter ($f(t)$) of the control system. In diagram *a* the investigated parameter changes suddenly from one value to another, after which it keeps the new value for a prolonged period. In diagram *b* the parameter returns to its original value after a sudden rapid change. Diagram *c* demonstrates the testing when (disregarding a short initial period) the value of the parameter changes at a constant rate. The procedure outlined in diagram *a* is followed, for instance, in the study of a thermostat, if the temperature to be regulated is changed to a prescribed higher value. Case *b* is encountered when vessels stored in a thermostat are exchanged for colder ones. In case *c* the thermostat temperature increases according to a linear program.

The concept of the transition function. Let us discuss case *a* in some detail. Dynamic analysis (especially its mathematical treatment) is simplified if the rate of change of the parameter is so fast that the time interval ε may be regarded as zero ($\varepsilon \to 0$). In this case the parameter will change not as in Fig. 8.6*a*, but rather according to Fig. 8.7. Considering by definition a unit change, the function depicted in Fig. 8.7 is the *unit-step function.* In this case the dynamic analysis consists in studying the response of the system to the unit-step effect.

The function describing the response, i.e. the time course of the output signal, is the *transition function.* Blood pressure regulation is a good example to demonstrate a biological transition function. On the basis of the black box model, this regulation may be discussed without describing the functional structure of the system. The object of the actual investigation is the response developed by the organism when, for instance, a

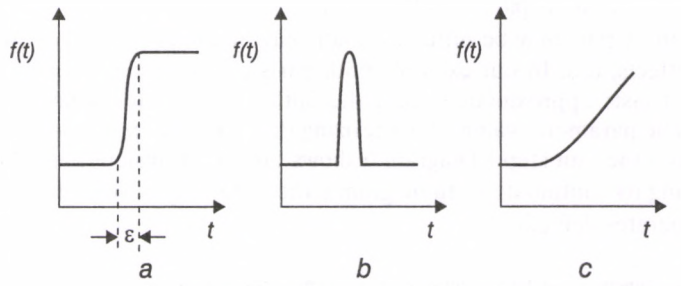


Fig. 8.6. Input signals of various shapes

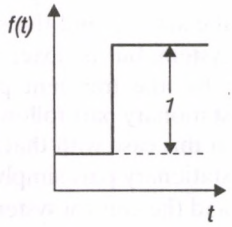


Fig. 8.7. Unit-step function

person suddenly stands up from a lying position. The positional change corresponds to the unit step. The response is a sudden decrease of the arterial mean pressure (approximately 13 kPa), which subsequently usually returns to its original value within a few seconds. The return process may follow different time courses; the three most characteristic types are depicted in Fig. 8.8. Diagram *a* shows an aperiodically damped response, *b* a periodically damped one, and *c* an undamped response. This latter case is always associated with a defective regulation. It should be noted that in the case of the voltage clamp technique discussed in section 7.3.2 the sudden setting of the membrane voltage corresponds to the unit-step signal, and the separated ionic fluxes to the aperiodically damped response of the system.

The response of the system, i.e. the transition function, consists of two parts. One part changes with time (this is the transient part), whereas the other remains constant (this

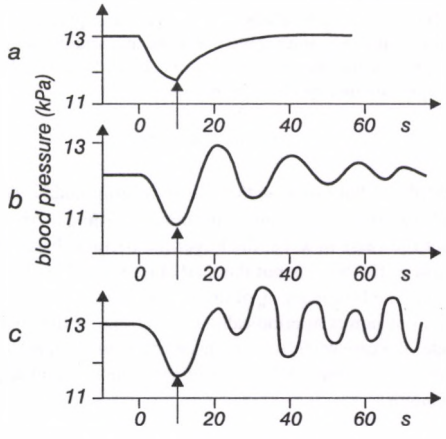


Fig. 8.8. Variation of the output signal of the blood-pressure regulating system in time (transition function), due to a sudden change in state (unit-step function). The diagram depicts the results of actual measurements
The zero point on the time axis corresponds to the time of the position change; the response of the control system begins at the time indicated by the arrow

is the stationary part). The transient part is determined by the internal properties of the system, and the stationary part may be influenced not only by the properties of the system, but by external effects, too. In our example both parts can be observed in Fig. 8.8*a*: the transient period lasts approximately until the 30th second, after which the stationary part follows. (The parameter value characterizing the stationary period agrees in this case with that before the unit step.) Diagram *b* shows only the transient part, the stationary part simply being its continuation. In diagram *c* the stationary part is missing, and the control system operates defectively.

The transient part of transition functions can be well illustrated by a simple physical analogy. Let us consider the motion of a body fastened to a spiral spring. (The mass of the spring can be neglected with respect to the mass of the body.) Let us displace the body in the vertical direction from its equilibrium position and then leave it alone. It is found by experience that the motion of the body is a damped oscillation. The damping may be so strong that the displaced body will not move in the other direction beyond its equilibrium position, but approaches the equilibrium position from the direction in which it was originally displaced.

In the interpretation of this phenomenon two forces are considered. The first is the elastic force (X_e) inducing the motion, while the second is the frictional force (X_f) impeding the motion. The elastic force may be taken as proportional to the displacement (x) and its direction is opposite to that of the displacement. The frictional force, on the other hand, may be taken as proportional to the velocity (dx/dt) of the motion, this force also being opposite in direction to the displacement. The following relations hold

$$X_e = -m\omega_0^2 x \quad \text{and} \quad X_f = -2m\kappa \frac{dx}{dt} \tag{8.5}$$

where m denotes the mass of the body, and ω_0 and κ are constants characterizing the elastic and frictional force, respectively. (ω_0 is the angular frequency of frictionless vibration.) The resultant force is given by the following differential equation

$$m \frac{d^2x}{dt^2} = -m\omega_0^2 x - 2m\kappa \frac{dx}{dt} \tag{8.6}$$

where the resultant force has been written as the mass of the body multiplied by its acceleration. The motion of the body is described by the functions $x(t)$ which satisfy the differential equation (cf. Appendix, B5). Omitting the details of the solution of the differential equation, only the more important results are given.

The body moves with a damped oscillation if $0 < \kappa < \omega_0$, i.e. if the friction is larger than zero, but smaller than a certain value. In this case [8.6] is satisfied by the relation

$$x = ae^{-\kappa t} \sin \omega t \tag{8.7}$$

where a is a constant. [8.7] describes a harmonic oscillation whose amplitude is determined by the product before the sine function. The value of this product decreases with time according to an exponential function. The rate of decrease is the faster, the larger the value of κ, i.e. the larger the friction (Figs 8.9*b–c*). If $\kappa = 0$, i.e. if there is no friction, the harmonic oscillation of the body is not damped (Fig. 8.9*a*). If $\kappa \neq 0$, the angular frequency ω of the damped oscillation is smaller than the frequency ω_0 of the undamped oscillation: $\omega^2 = \omega_0^2 - \kappa^2$.

The body does not oscillate, but moves aperiodically if $\kappa \geq \omega_0$, i.e. if the friction is larger than or equal to a certain value. First let us consider the case when $\kappa = \omega_0$; this is the *aperiodic limiting case.* Let v_0 denote the velocity with which the body at rest is pushed at time $t = 0$. After being pushed and subsequently left alone, the motion of the body is described by the relation

$$x = v_0 te^{-\kappa t} \tag{8.8}$$

According to [8.8] the displacement first increases, then reaches a maximum and subsequently decreases; the body gradually approaches its equilibrium position (which in principle is reached after an infinitely long time; Fig. 8.9*d*). The particulars of the case $\kappa > \omega_0$ will not be discussed; it should be mentioned only that the approach to the equilibrium position will be the slower, the larger is κ compared to ω_0.

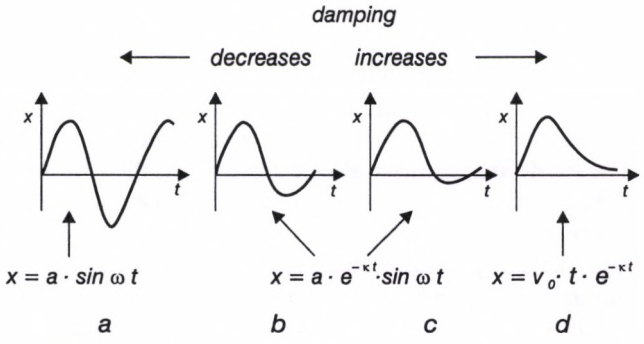


Fig. 8.9. The motion of an oscillating body with various dampings

The analogy between the discussed physical example and the regulating system is obvious. The state of the regulating system corresponds to the body at rest when the controlled parameter has the prescribed value. The displacement of the oscillating body is analogous to the difference between the output signal and the base signal. The elastic force inducing the oscillation represents the tendency of the regulating system to maintain the prescribed parameter value. The elastic force is proportional to the displacement; in the case of a regulating system the restoring tendency is the stronger, the larger the deviation of the output signal from the base signal. Friction damps the motion of the oscillating body. Similar tendencies are operative in the regulating systems. It follows from the analogy that regulation occurs only in their presence. It is expected from a *stable* regulation that the output signal will approach the prescribed value by means of aperiodic motion.

The aperiodic limiting case is the most favourable one, since the approach is then the fastest. Figure 8.9 demonstrates the relation between the stability of regulation and the transient part of the transition functions as well.

8.3. Computers

Computers are electronic instruments for the storage, processing and forwarding of data; they serve for the fast and automatic solution of problems in different fields of life. From the construction of the first computers up to the present about half a century has passed. During this time computers have become indispensable tools in science and practice alike.

Naturally, the use of the computers has spread also in the medical science and practice. In the following first the principles of the structure and operation of the computers will be outlined, then the significance of their application in medicine will be discussed.

8.3.1. The structure and operation of computers

1. The assembly of the material components of the computer is called the *hardware*, while the intellectual products ensuring the accomplishment of the tasks are the *software.* Hardware and software have always been closely interrelated and influence the progress of each other. The hardware consists of electronic units; their development is closely related to that of electronics; in the first computers, from the middle of the 40s, electron tubes operated which were followed by the transistors from the middle of the 50s, while from the 60s larger and larger scale integrated circuits has served as basis for their operation. The outlined development has entailed the increase in the operating speed of the computers and the considerable decrease in their size and energy requirements. From the 70s a direct man–machine contact has become possible by the appearance of the personal computers (PC). The development in our days is marked by the organization of PCs into networks and their connection to central computer(s).

The basic structure of computers is seen in Fig. 8.10. The *peripherals* serve for the information transfer (input and output) between the computer and the external world. Data and instructions may be entered into the computer via the peripherals, and the results of the tasks performed by the computer also appear in the peripherals. Peripherals are the generally known keyboards and monitors or displays of the computers, however, in a wider sense a PC may be considered the peripheral of a central machine of a network if it is supplied with a *modem* unit (modulator-demodulator) and with a telecommunication line for information transfer.

The *processor* unit is the centre of the system which executes the prescribed tasks and supervises the whole system. The third essential unit is the *memory* or *store.* This stores the data and instructions necessary for the execution of the tasks. One part of the memory is the Read Only Memory (ROM), while the other is the Random Access Memory (RAM). The ROM part is the fixed memory, where the information is stored permanently and cannot be erased. The RAM part is called also read/write memory. The name indicates the function: after the writing-in, the information may be read as long and as many times as necessary. If the information is not required any more, new information can be written in its place. The size of the storage capacity is usually characterized with the amount of the information which may be stored.[3] This part of the memory is in a direct contact with the processor, therefore it is called *primary memory.* In the *background memories (secondary memories)* a great amount of information may be stored but the processor does not have a direct access to it. Background memories are e.g. the magnetic disks or compact disks (CD). The information may be stored electronically, by means of

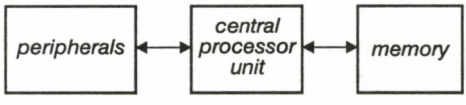


Fig. 8.10. Block diagram of a digital computer

[3] For practical reasons the storage capacity of the memory of the computers is given in bytes; 1 byte = 8 bit.

semiconductor circuits; nowadays, however, attempts are made at the exploration and elaboration of information storage at the "molecular" level based on other principles. In this respect we refer on one hand to the formation of molecular layers with ordered structure, on the other hand to the molecular information storage accomplished in the nucleic acids by Nature: the average information content of 50,000 base pairs (cf. section 8.1.3) is about 100 kilobits, i.e. approximately the same as that of certain semiconductor circuits. However, this amount of information occupies a volume of only about 10^{-16} cm^3 as opposed to some tenth of cm^3 in the case of semiconductor memories.

2. *The operation of the units of the computer* and their functional contact with each other is demonstrated by a simple example.

Consider the following quadratic equation: $3x^2 + 2x - 5 = 0$. For its solution we need the solution formula on one hand:

$$x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

and on the other hand the values of *a, b* and *c,* i.e. *data.* With the help of a pocket calculator the solution may be performed by first considering the radicant: multiply $a = 3$ by 4, then multiply it by the value of *c*, i.e. 5. Since in the example *c* is negative, the sign is changed to get the right sign for $4ac$. Then the sign is changed to get $-4ac$. To this the square of $b = 2$ is added. The square root of the sum is extracted. First $b = 2$ is subtracted from the positive root, the remainder is divided first by 2, then by $a = 3$. This is one of the solutions of the equation. The other solution is obtained by repeating the steps following the extraction of roots with the negative root.

The series of the above elementary steps is an *algorithm.* The processor of the computer solves such step-by-step procedures. It is revealed from the example that for the solution partly data, partly appropriate algorithms are necessary. Naturally, a given task may be also solved with another algorithm. On the other hand, an already tested algorithm may be repeated for the solution of a given type of tasks as often as necessary, only the data have to be changed. This *ad libitum* applicability is the advantage of algorithms. The *data* necessary for the operation of the processor are given in the form of numbers, i.e. they are *coded in the form of numbers.* However, not only the data but also the algorithms may be coded as numbers: this is the so-called *Neumann principle*, the machines operating according to this principle are *Neumann-type* computers.

The data and the algorithms are stored in the *segments* of the memory. The processor successively reads and interprets them and executes the instructions. The sums of the appropriately coded algorithms are the programs which constitute the software of the computer. Actually, the programs are inserted between the hardware and the user to promote the solution of the tasks.

The writing of a program means the appropriate coding of the algorithm. In the case of the *machine language* the algorithms and data coded as numbers are loaded in the segments of the memory; these informations can be directly interpreted by the processor. If instead of numbers other, suitable symbols are used, the program is written in an *assembler language*. In this case a compiler program converts the symbols to numbers for

the processors. The *algorithmic languages* consist of a complexity of such symbols in which certain algorithm parts are coded by the appropriate symbols. Naturally here again suitable compiler programs make the algorithms interpretable for the processor. The best-known algorithmic languages are e.g. ALGOL and its advanced version, PASCAL, and also BASIC, COBOL and FORTRAN.

The "user-friendly" application of computers is made possible by the *operational systems.* These are programs with the help of which the computer carries out many routine operations "automatically": e.g. it searches for a certain program in the background memory, loads it into the primary memory, and copies it if necessary. Such operational systems are DOS and WINDOWS.

8.3.2. Some possibilities for the application of computers in medicine

As already mentioned, computers are important parts of our everyday life. In addition to their general use, the increasing possibilities of computers are also utilized in the medical practice. In the following some applications will be dealt with which have played an important part in medical activity so far and the significance of which will be increased in the future, considering the prospective development.

1. Data recording. It is one of the traditional uses of computers. A computer with a large memory capacity enables the handling, storing and systematization of a very large number of data. The required data are obtained within a few seconds. The data in the recording system may be changed, supplemented or deleted. Medical data recording makes use of these possibilities. The computerized handling of the filing system of a medical district as well as of large medical institutions (hospitals, university hospitals, outpatient departments), furthermore urban, county and even central national data banks work this way. In this case the medical data of every citizen (e.g. birth, sex, children's diseases, vaccinations, diagnostic results, treatments) are stored and can be used whenever required. For the general practitioner pharmaceutical catalogues may be also useful which contain the description of the active principles, price, availability and indication of the drugs.

Nowadays it is not necessary to restrict ourselves to the memory capacity of the memories of our PC or to the data stored in them. By means of *computer networks*, the huge databases of the world are practically available for computers with relatively small memories through central computers. In addition to catalogues of libraries and museums, there are catalogues in these databases which are useful also for medicine, e.g. the data of the gene banks, the protein catalogue with the amino acid sequences, lists of medical journals and books.

An important group of the databases are the *bibliographical databases* processing the medical literature which contain every data on the basis of which the original article, monograph, dissertation, etc. may be found (title of the article and the journal, date of publication, name of the authors) and usually also the abstract of the publication. Such databases are e.g. *Medline* and *Excerpta Medica* which comprise the whole of medicine together with its borderlands. Another one is *HealthSTAR*, a bibliographical database concerning the nonclinical aspects of public health (organization, planning, management, etc.).

Another group consists of *full text databases* which contain complete documents. An example is *Justis,* the full-text legal database of the European Community which includes recommendations also for the public health and for the products of the food and pharmacological industries.

A further type are the *directory-type and factographic databases* containing numerical data, addresses, etc. One of them is *Handynet*, the catalogue of appliances for the handicapped.

This list is far from being complete, there are several thousand databases around the world. The full texts of newspapers, journals, information of companies, patents, the date and program of congresses, etc. may be all looked up. The information have its price, depending on the type. The medical databases are usually cheaper, while those having some relation to the business life are expensive. Through the continuously expanding *Internet* free databases are available, too. Nevertheless, it has to be mentioned that the information obtained from the Internet have to be dealt with a certain reservation and critic, because everybody may place any information on it and at present nobody is obliged to update and maintain the placed database.

The Internet offers various other possibilities: through it correspondence groups, panel discussions, free softwares, information leaflets, anatomical and surgical atlases, educational films are accessible, caritative or nursing organizations may be contacted, consultations may be carried out with experts, etc.

The computerized network can make a dialogical contact between machines or persons far away from each other. A modified version of this possibility has a practical medical application, e.g. in the form of emergency consultations. Namely with the help of the network an advice may be asked from an expert without the physical presence of the expert at the given place: nowadays every information characterizing the disease may be sent everywhere without delay, even in the form of images, and the answer may arrive in a similar way.

2. Medical diagnostic equipments. The constantly developing diagnostic procedures introduced in the past decades could not exist without computers. Such are the computerized tomography (CT), SPECT, PET, etc. Their common characteristic is that they employ a special apparatus which collects measurement values of certain parameters (cf. section 6.7) according to a previously determined program, point by point, layer by layer, then, after processing the collected data presents them in the form of images. On the basis of the data the image may present the segments of a given part of the body from any direction, and even the three-dimensional reconstruction of certain, especially important parts is possible. If necessary, the selected graphic information may be stored for a set time (usually in background memories) and therefore the later retrieval of the images and graphic information is also possible. A medical equipment comprises the apparatus necessary for the measurement and a special computer which carries out automatically the measurement, the controlling of the measuring apparatus, furthermore the collection and processing of the data. In the computers of the above medical systems the data storing– data processing function plays therefore such an important role that the development and increasing power of these equipments is closely related to the development of the computers in this respect.

3. Expert systems. A further medical utilization of computers is the application of expert systems in diagnostics. These systems collect and utilize all the information concerning a given disease available at the time being and by means of this information help the physician to establish the diagnosis. The expert system contain rules based on the most up-to-date knowledge available. By using these and the patient's data (e.g. age, symptoms, results of laboratory tests) the computer establishes its diagnosis, or, more frequently, in complicated cases gives several diagnostic possibilities together with the pertaining probabilities. Naturally the expert systems do not take the place of the physician, but play rather a consulting role so that the physician and the computer are in a dialogue with each other. In the course of this dialogue the physician may give instructions and information concerning the patient, may ask questions and may also direct the operation of the computer by emphasizing and weighting the importance of the data and by presenting new aspects. The computer gives advices, asks for more information and, if asked, explains its decision. The more advanced versions of the system have also "learning" capabilities which enable the entering of new rules or the modification of the existing rules on the basis of the experience gained from the actual cases. The computerized network may also considerably increase the "expertise" of the expert systems, since other expert systems may be involved in the diagnosing, the CT, MRI, etc. images may be transmitted through the network and thus the opinions of other "experts" in the network may also participate in the therapeutical decision-making.

REFERENCES

Books

Drischel, H., Einführung in die Biokybernetik. Akademie Verlag, Berlin (1972)

Goldberger, F. (ed.), Biological Regulation and Development. Vol. 1. Gene Expression. Plenum Press, New York–London (1980)

Norton, P., Inside the IBM PC. Brady Communications Co., Prentice-Hall, London (1983)

Rose, J. (ed.), Progress of Cybernetics, Vol. 1. Gordon and Breach, London (1970)

Vorndran, E. P., Entwicklungsgeschichte des Computers. KOE-Verlag, Berlin (1982)

Woollard, B. G., Digital Integrated Circuits and Computers. McGraw-Hill Hook Company, London (1978)

Papers

Davis, R., Bachanan, B., Shortliffe, E., Production rules as a representation for a knowledge-based consultation program. *Artif. Int., 8,* 15–45 (1977)

Mercando, A. D., Computer help with medical diagnosis and procedure codes. *Pacing Clin. Electrophysiol., 20,* 349 (1977)

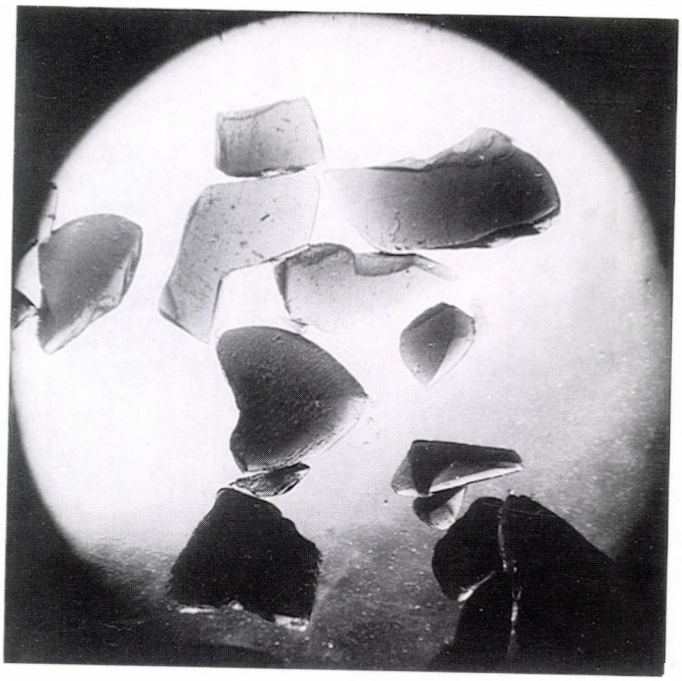
Patterson, D. A., Microprogramming. *Scientific American, 248,* 50–57 (1983)

Poole, C. J., Millman, A., ABC of medical computing. Adaptive computer technology. *British Medical J., 311,* 1149 (1995)

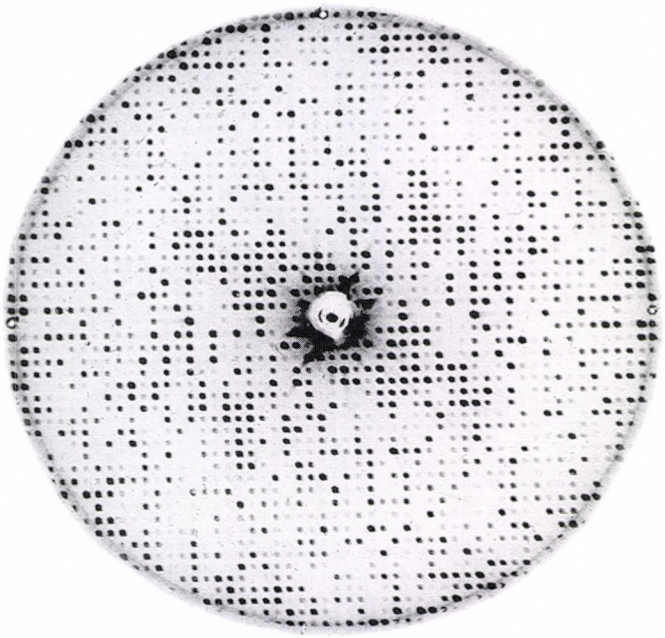
Szolovits, P., Pauker, S. C., Categorical and probabilistic reasoning in medical diagnosis. *Artif. Int., 11,* 115–144 (1978)

Thull, B., Janssens, U., Rau, G., Hanrath, P., Approach to computer-based medication planning and coordination support in intensive care units. *Technol. Health Care, 5,* 219 (1997)

Toong, H. D., Gupta, A., Personal computers. *Scientific American, 247,* 89–99 (1982)

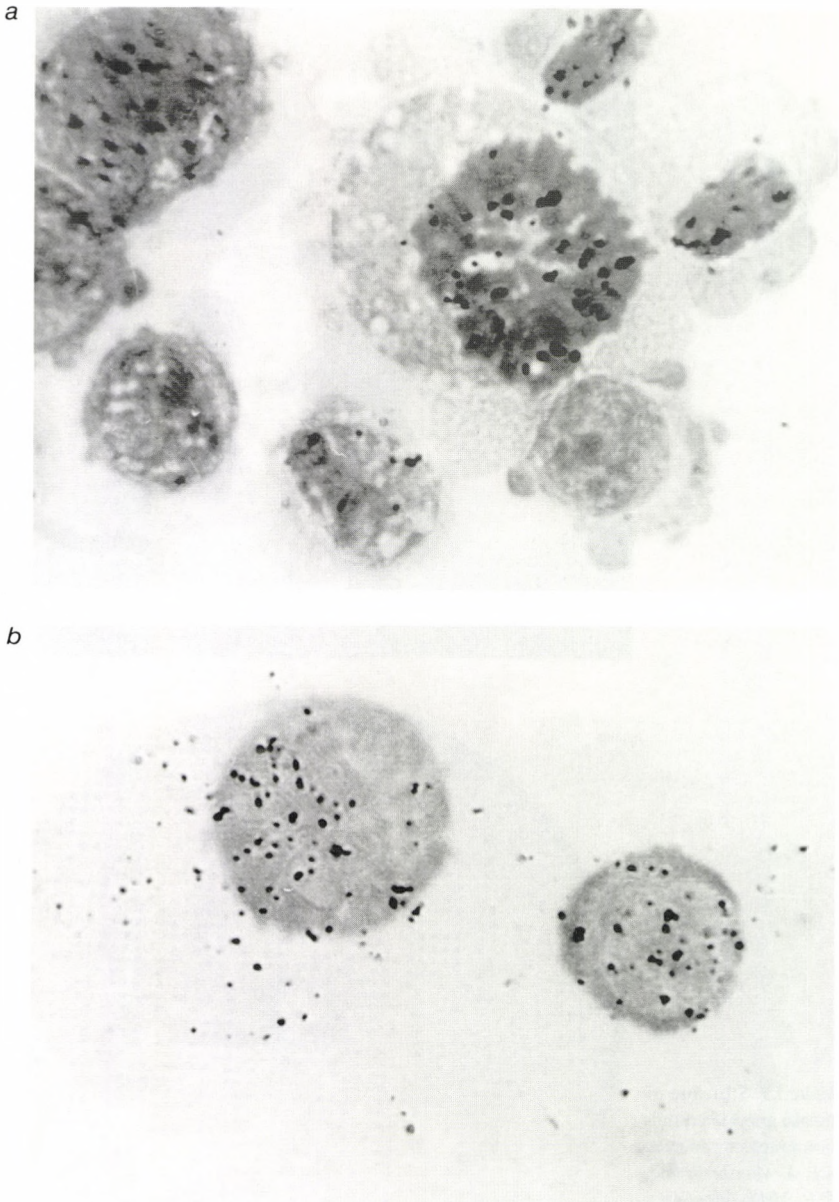


a

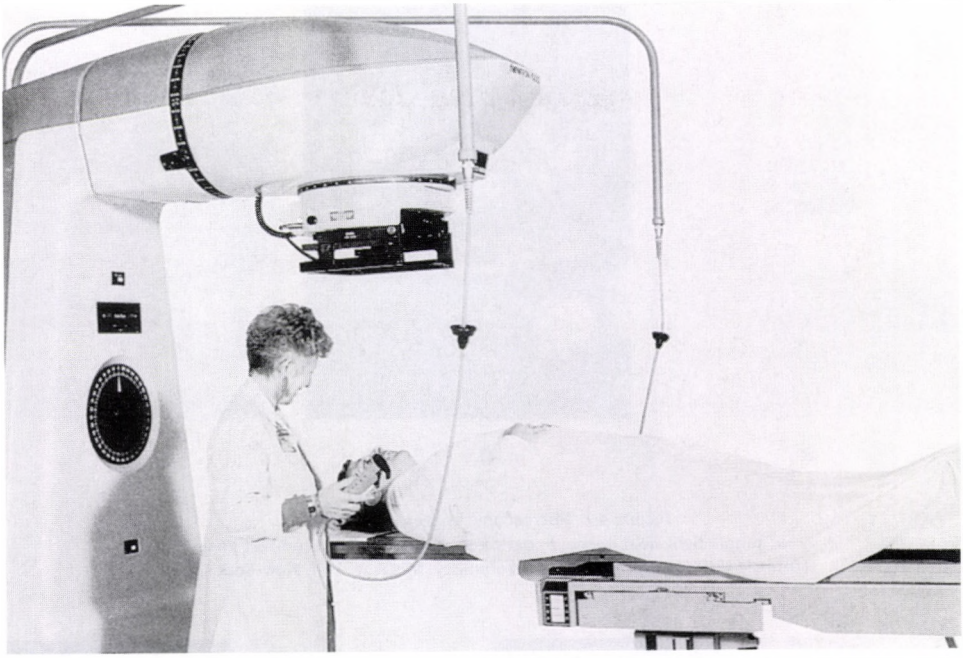


b

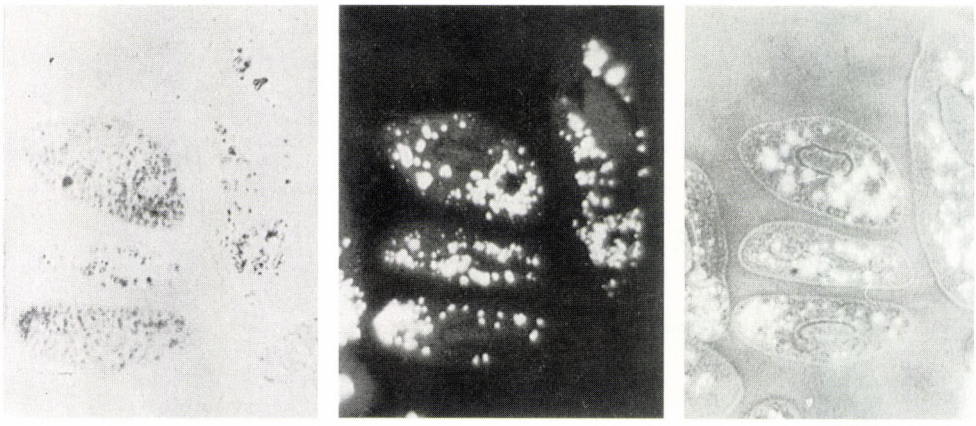
Picture 1.1. Structure of catalase enzyme crystals (photographs by courtesy of *B. K. Vainshtein* and *W. R. Melik-Adamyan*)
a: microscopic image of small magnification; the size of one crystal is 0.3–0.5 mm;
b: X-ray diffraction photograph



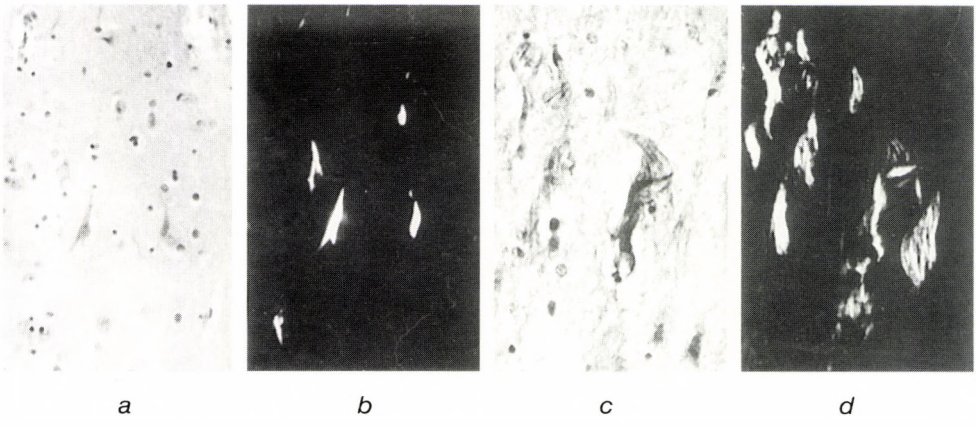
Picture 3.1. Microautoradiogram of mouse-ascites lymphoma (NK) cells labelled with ^{3}H-thymidine (*a*) and ^{14}C-thymidine (*b*) (photographs by *L. Varga*)
a: Ilford G5 emulsion, Giemsa staining, magnification 1500×,
b: Ilford G5 emulsion, methyl green-pyronine staining, magnification 1500×



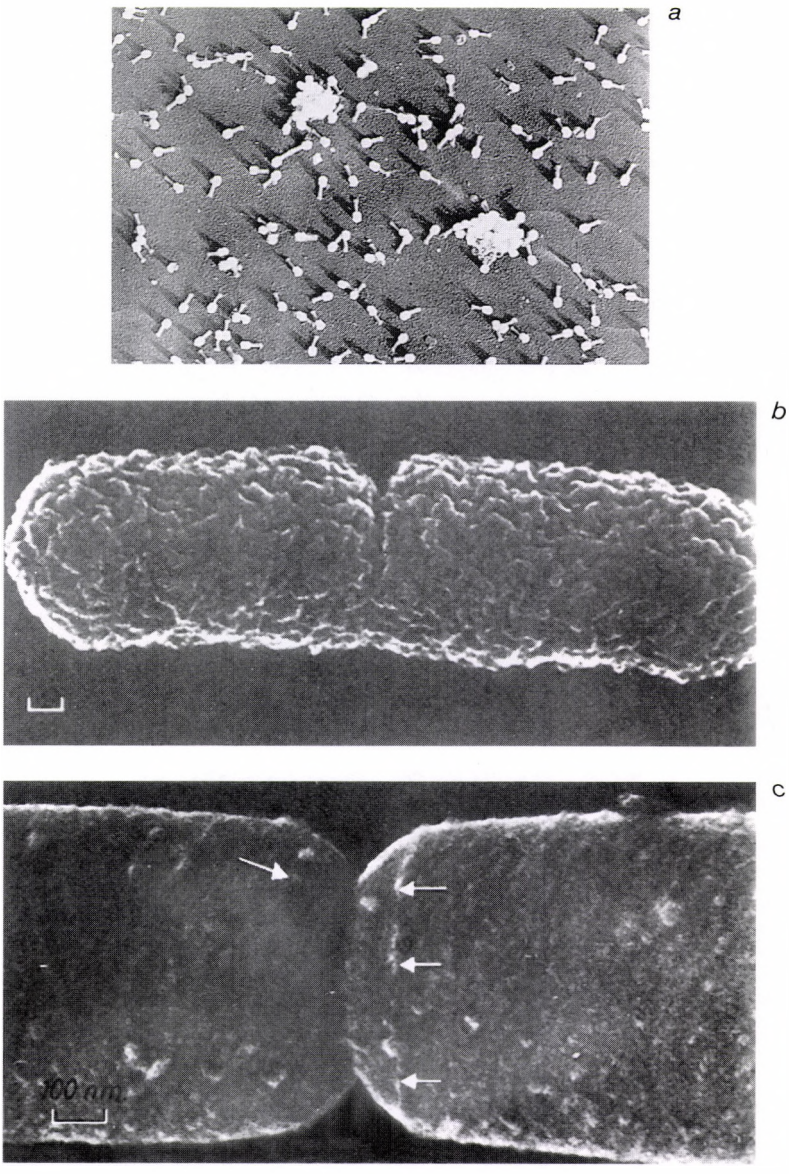
Picture 3.2. Cobalt unit with rotatable irradiation head and treatment table



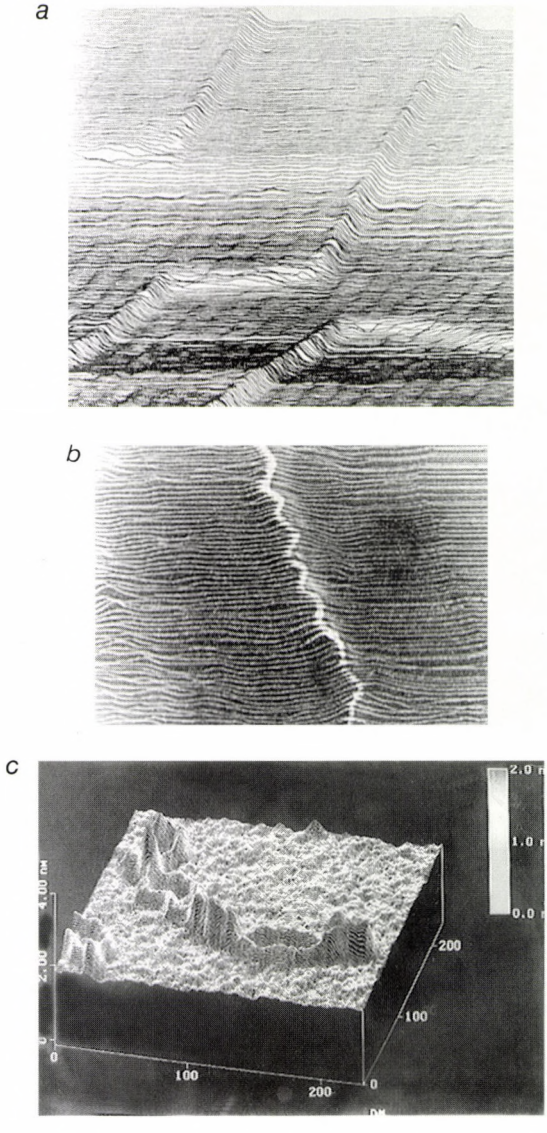
Picture 4.1. Microscopic images of protozoa
a: simple light field image; *b*: dark field image; *c*: phase contrast image
(from *M. J. Pelczar, R. D. Ried*, Microbiology. McGraw-Hill, New York 1965)



Picture 4.2. Pictures made from brain sections in Alzheimer's disease
A and **C** with conventional light microscope; **B** and **D** with polarization microscope
On the latter the birefringent Alzheimer fibres resembling locks are well visible (*L. Módis*, Organization of the Extracellular Matrix: a Polarization Microscopic Approach, CRC Press, Boca Raton 1991)

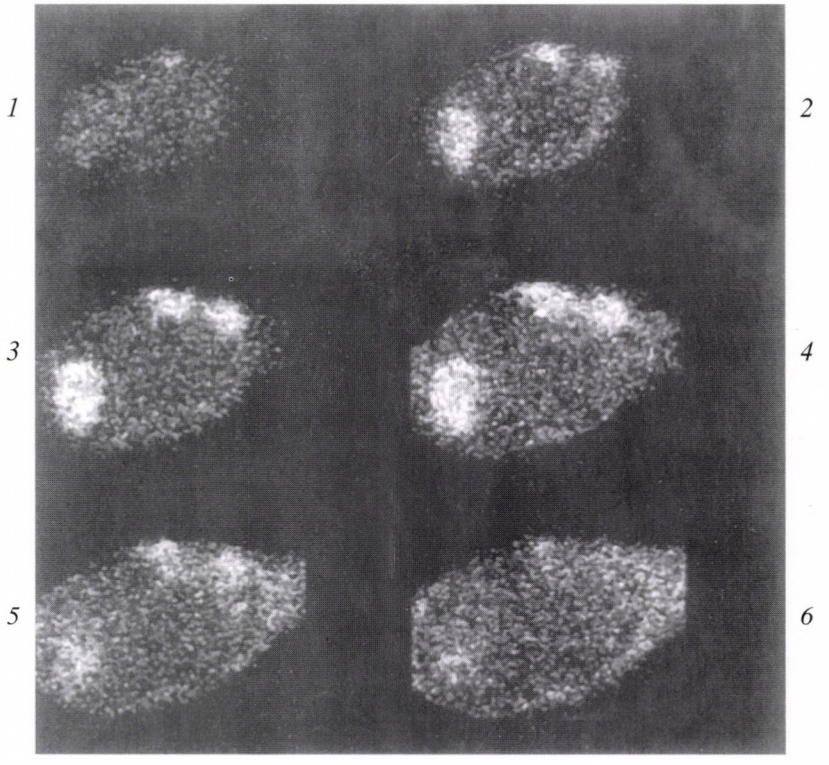


Picture 4.3. Electron micrograms
a: electron microgram of T2 phages, obtained with a traditional electron microscope but using special contrast and plasticity-improving technique (magnification: 25,000×; *R. M. Herriott, J. L. Barlow,* J. Gen. Physiol., 36, 17, 1952); *b–c*: scanning electron micrograms of dividing bacteria (*E. coli* and *B. subtilis*). The arrows indicate the ridges separating the old and new surface areas produced in the division (*K. Amako, A. Umeda,* J. Gen. Microbiol., *98*, 297, 1977)

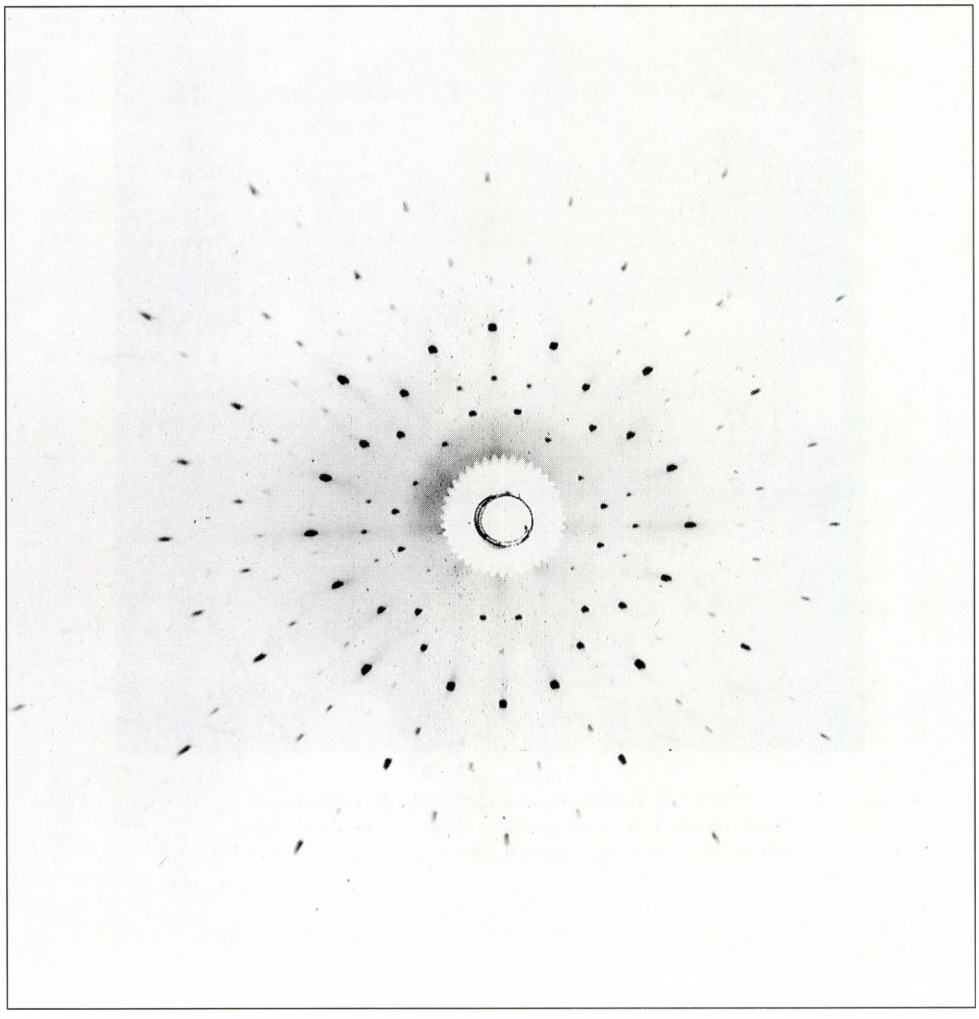


Picture 4.4. Pictures made with modern scanning microscopes

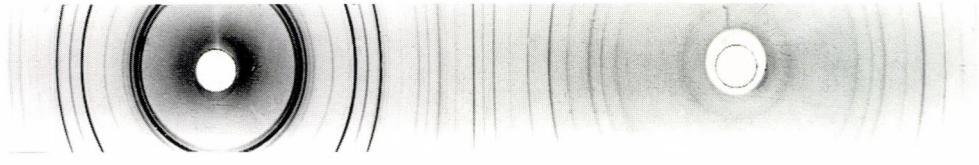
a: STM image of an Au surface with a few tenth nm high monatomic steps; width of a scan » 0.15 nm (*G. Binnig* et al., Surf. Sci., *144*, 321, 1984); *b*: STM image of DNA on a carbon substrate; the wavelength of the zigzag is 3.5 ± 0.15 nm which corresponds to the twisting periodicty of the double helix (*G. Binnig*, Bull. Am. Phys. Soc., *31*, 217, 1986); *c*: Picture of a part of a DNA molecule made with an AFM; the protuberances represent enzyme molecules participating in the transcription (*K. Rippe, J. Langowski*, Heidelberg, 1995)



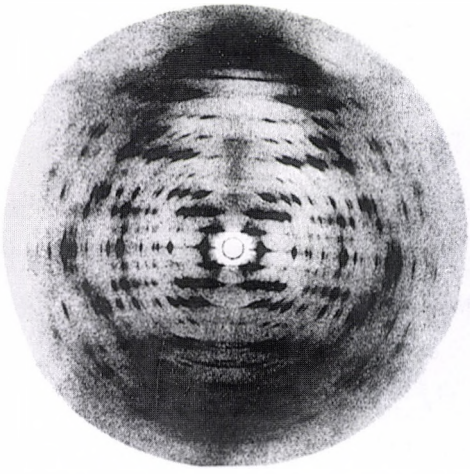
Picture 4.5. Picture made with a confocal microscope from the downwardly successive planes of the nucleus of a Hela cell, length: approximately 15 mm (*A. Kurz, P. Lichter,* Heidelberg, 1995)



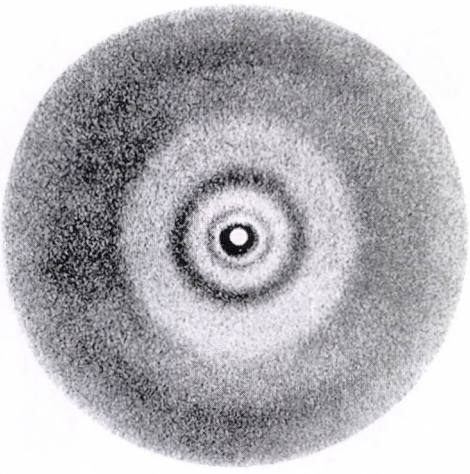
Picture 4.6. Laue diagram of a NaCl crystal (photograph by courtesy of *L. Varga*)



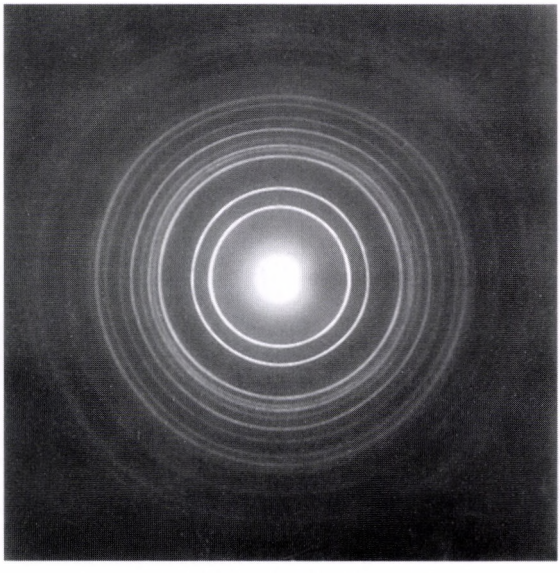
Picture 4.7. Debye–Scherrer diagram of crystalline ZnS powder (photograph by courtesy of *M. Jahnke*)



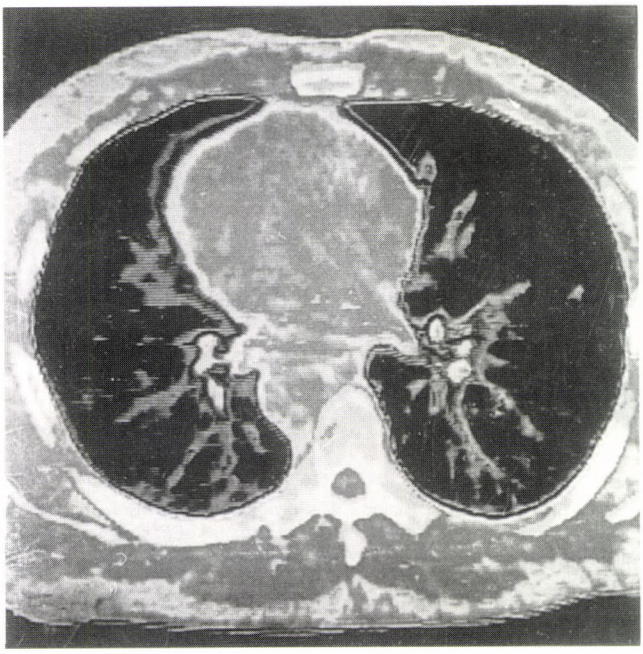
Picture 4.8. X-ray diffraction pattern of a single DNA fibre (by *M. H. F. Wilkins;* taken from *R. B. Setlow, E. C. Pollard:* Molecular Biophysics. Addison–Wesley Publ. Co., 1962)



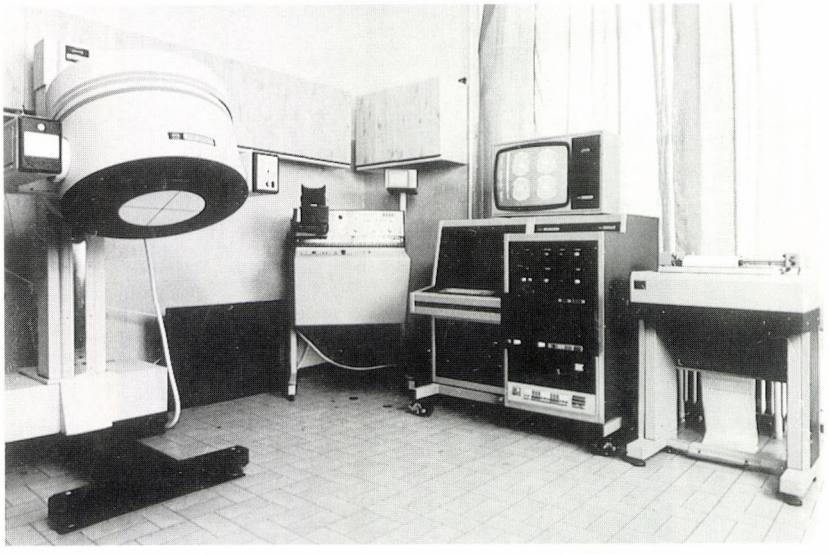
Picture 4.9. X-ray diffraction pattern obtained from the helical domains of *E. coli* tyrosine-tRNA (by kind permission of *W. Fuller*)



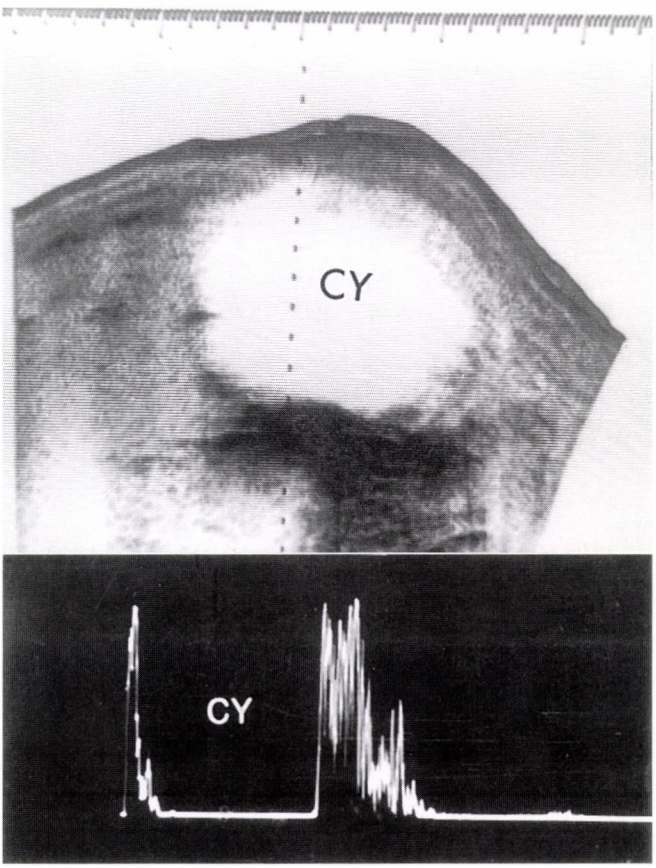
Picture 4.10. Diffraction of an electron beam on a thin gold foil (photograph by courtesy of *A. Barna*)



Picture 6.1. Densitogram of a section through the chest. The vertebra, heart, great arteries, lung and bronchi are clearly seen



Picture 6.2. Gamma-camera with accessories (GAMMA Művek, Hungary). Left to right: wide-field scintillation detector; electronic signal processor, information processing computer and display unit; graphic printer



Picture 6.3. Echogram of a cyst (CY) Above: a two-dimensional grey-scale *B* image; below: an *A* image. The latter relates to the dashed line on the *B* image (Department of Radiology, Semmelweis University of Medicine)

9. UNIVERSAL TABLES

Table 9.1. The International System of Units (Système International d'Unités; notation: SI) Base Units

Base quantity	Base unit	
	name	symbol
Length	metre	m
Mass	kilogram	kg
Time	second	s
Electric current intensity	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

Table 9.2. Derived SI units with special names

Quantity	SI unit		
	name	symbol	expressed in other SI units
Plane angle	radian	rad*	$m\ m^{-1}$
Solid angle	steradian	sr	$m^2\ m^{-2}$
Frequency	hertz	Hz	s^{-1}
Force	newton	N	$J\ m^{-1}$
Pressure	pascal	Pa	$N\ m^{-2}$
Energy, work, quantity of heat	joule	J	N m
Power	watt	W	$J\ s^{-1}$
Electric charge	coulomb	C	A s
Electric potential, electric voltage	volt	V	$W\ A^{-1}$
Electric capacitance	farad	F	$C\ V^{-1}$
Electric resistance	ohm	Ω	$V\ A^{-1}$
Electric conductance	siemens	S	$A\ V^{-1}$
Magnetic flux	weber	Wb	V s
Magnetic flux density	tesla	T	$Wb\ m^{-2}$
Inductance	henry	H	$Wb\ A^{-1}$
Luminous flux	lumen	lm	cd sr
Illuminance	lux	lx	$lm\ m^{-2}$
Activity of a radioactive source	becquerel	Bq	s^{-1}
Absorbed dose	gray	Gy	$J\ kg^{-1}$
Equivalent dose, effective dose	sievert	Sv	$J\ kg^{-1}$

* Should not be mistaken for the traditional unit of the absorbed dose, also denoted by rad

Table 9.3. SI prefixes

Prefix		Multiplier
name	symbol	
exa	E	10^{18}
peta	P	10^{15}
tera	T	10^{12}
giga	G	10^{9}
mega	M	10^{6}
kilo	k	10^{3}
hecto	h	10^{2}
deca	da	10
deci	d	10^{-1}
centi	c	10^{-2}
milli	m	10^{-3}
micro	μ	10^{-6}
nano	n	10^{-9}
pico	p	10^{-12}
femto	f	10^{-15}
atto	a	10^{-18}

Table 9.4. Interconversion of traditional and SI units

Table 9.4.1 Force units

	dyne	newton	pond	kilopond
dyne	1	10^{-5}	1.02×10^{-3}	1.02×10^{-6}
newton	10^{5}	1	1.02×10^{2}	1.02×10^{-1}
pond	981	9.81×10^{-3}	1	10^{-3}
kilopond	9.81×10^{5}	9.81	10^{3}	1

Table 9.4.2 Energy units

	erg	joule	metre-kilopond	kilowatt-hour	litre-atm	calorie	electron-volt
erg	1	10^{-7}	1.02×10^{-8}	2.78×10^{-14}	9.87×10^{-10}	2.39×10^{-8}	0.624×10^{12}
joule	10^{7}	1	0.102	2.78×10^{-7}	9.87×10^{-3}	0.239	0.624×10^{19}
metre-kilopond	9.81×10^{7}	9.81	1	2.72×10^{-6}	9.68×10^{-2}	2.34	0.612×10^{20}
kilowatt-hour	3.6×10^{13}	3.6×10^{6}	3.67×10^{5}	1	3.55×10^{4}	8.6×10^{5}	2.25×10^{25}
litre-atm	1.013×10^{9}	1.013×10^{2}	10.33	2.815×10^{-5}	1	24.22	0.633×10^{21}
calorie	4.187×10^{7}	4.187	0.427	1.16×10^{-6}	4.13×10^{-2}	1	2.63×10^{19}
electron-volt	1.6×10^{-12}	1.6×10^{-19}	1.63×10^{-20}	4.45×10^{-26}	1.58×10^{-21}	3.83×10^{-20}	1

Table 9.4.3 Pressure units

	dyn cm^{-2}	pascal (Pa)	p cm^{-2}	tech. atm. (at)	phys. atm. (atm)	torr (mm Hg)	bar
dyn cm^{-2}	1	10^{-1}	1.02×10^{-3}	1.02×10^{-6}	9.87×10^{-7}	7.5×10^{-4}	10^{-6}
pascal (Pa)	10	1	1.02×10^{-2}	1.02×10^{-5}	9.87×10^{-6}	7.5×10^{-3}	10^{-5}
p cm^{-2}	981	98.1	1	10^{-3}	9.68×10^{-4}	0.736	9.81×10^{-4}
tech. atm. (at)	9.81×10^{5}	9.81×10^{4}	10^{3}	1	9.68×10^{-1}	736	0.981
phys. atm. (atm)	1.013×10^{6}	1.013×10^{5}	1033.23	1.03323	1	760	1.01325
torr (mm Hg)	1333	133.3	1.36	1.36×10^{-3}	1.32×10^{-3}	1	1.333×10^{-3}
bar	10^{6}	10^{5}	1.02×10^{3}	1.02	9.87×10^{-1}	750	1

Table 9.4.4 Supplement to the use of measurement units

1. The plane angle may be also expressed in terms of the *degree* (denoted by °), its 60th part the *minute* (denoted by ') and its 3600th part the *second* (denoted by ")

$$1^\circ = \frac{\pi}{180} \text{ rad}$$

2. Time units used without restrictions are the *minute* (denoted by min), the *hour* (denoted by h), the *day* (denoted by d) and the calendar units, i.e. the week, month and year

$$1 \text{ d} = 24 \text{ h} = 1440 \text{ min} = 86{,}400 \text{ s}$$

3. In atomic and nuclear physics the *atomic mass unit* (denoted by u) may be used, which is one-twelfth of the mass of the carbon 12 atom

$$1 \text{ u} = 1.66 \times 10^{-27} \text{ kg}$$

4. The energy unit *watt-hour* (notation: Wh) can be used without restriction

$$1 \text{ Wh} = 3600 \text{ J}$$

The energy unit which may be used in atomic and nuclear physics is the *electron-volt* (denoted by eV)

$$1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$$

5. The temperature unit *Celsius degree* (denoted by °C) may be used without any restriction. The temperature 0 °C is equal to 273.16 K (Kelvin degrees). The Celsius degree as a temperature difference is equal to the Kelvin.

6. For the amount of substance the elementary entity must be specified: atom, molecule, ion, electron, etc.

7. The SI unit of the *concentration of a substance* is mol/m^3: the quantity of the component in question in a mixture of 1 m^3 is 1 mol.

The SI unit of the *molality* is mol/kg: in a solvent of 1 kg the quantity of the component in question is 1 mol.

8. For the determination of the pressure of a fluid and of a gas, besides the pascal also the *bar* may be used (denoted by bar)

$$1 \text{ bar} = 10^5 \text{ Pa}$$

Table 9.5. Some important material constants

Table 9.5.1. Material constants of solids

Material	Density at 20 °C $\frac{kg}{dm^3}$	Linear thermal expansion coefficient* $\times 10^{-6}\frac{1}{K}$	Specific heat** $\frac{kJ}{kg\,K}$	Melting point °C	Heat of fusion $\frac{kJ}{kg}$	Tensile modulus $\frac{kN}{mm^2}$	Tensile strength $\frac{N}{mm^2}$
Aluminium	2.7	24	0.9	660	398	70	60–160
Brass	≈8.5	19	0.39	≈920		105	330–530
Bronze	8.8	16	0.39	1083	209	126	400–450
Glass	2.4–2.8	3–10	0.75–0.80			50–80	30–90
Gold	19.3	14	0.13	1064	63	80	110–130
Iron and steel types	7.7–8.9	9–12	0.46–0.54	1200–1540		115–195	140–350
Lead	11.3	29	0.13	328	25	16	15–18
Platinum	21.5	9	0.13	1772	101	160	130–200
Quartz	2.7	0.6	0.73	≈1700		60	800
Silver	10.5	19	0.24	961	105	76	140–380
Tungsten	19.3	5	0.13	3410	192	360	1000–4000

* Relative change of length for 1 K variation of temperature; the volume expansion coefficient is about three times this value

** At ordinary temperature and pressure

Table 9.5.2. Material constants of fluids

Material	Density at 20 °C $\frac{kg}{dm^3}$	Surface tension* $\frac{mJ}{m^2}$	Viscosity at 20 °C mPa s	Volume expansion coefficient* $\times 10^{-5} \frac{1}{K}$	Specific heat* $\frac{kJ}{kg\,K}$	Melting point °C	Heat of fusion $\frac{kJ}{kg}$	Boiling point °C	Critical temperature °C
Acetic acid	1.05	28	1.22	107	1.97	16.6	192	117.9	322
Acetone	0.79	24	0.32	149	2.22	−94.8	98	56.1	236
Benzene	0.88	29	0.65	124	1.72	5.5	128	80.1	289
Chloroform	1.49	27	0.57	127	0.96	−63.5	75	≈61	262
Diethyl ether	0.71	17	0.24	166	2.26	−116.3	98	34.6	194
Ethanol	0.79	23	1.19	112	2.43	−117.3	108	78.5	244
Glycerol	1.26	63	1.49	50	2.39	20	201	290	452
Mercury	13.55	≈476	1.55	18	0.13	−38.9	12	357	1460
Olive oil	0.91	33	84	72	1.97				
Water common	0.998	73	1.0	≈20	4.18	0.0	334	100	374.2
Water heavy	1.105	68	1.25	230	4.21	3.8	318	101.4	371.5

* At common temperature and pressure

Table 9.5.3. Material constants of gases

Gases	Density* at 0 °C $\frac{kg}{m^3}$	Viscosity* at 20 °C μPa s	Boiling point* °C	Critical	
				temperature °C	pressure MPa
Air	1.29	18.2	−193	−140.6	3.8
Carbon dioxide	1.98	14.8	−78.5**	31	7.4
Nitrogen	1.25	17.4	−195.8	−146.8	3.4
Oxygen	1.43	20.2	−182.9	−118.4	5.1

* At atmospheric pressure
** Sublimation point

Table 9.6. Electric resistivity of some metals and resistor materials at 20 °C
($\Omega\ mm^2\ m^{-1}$)

Aluminium	0.03	Bronze	0.01
Iron	0.1–0.15	Constantan	0.49
Lead	0.21	Kanthal	1.1–1.45
Platinum	0.10	Manganine	0.43
Silver	0.02	Nickel-silver	0.3–0.36
Tungsten	0.06		

Table 9.7. Electric conductivity of NaCl solution at 20 °C

Concentration (c) mol/litre	Specific conductivity (κ) $\Omega^{-1}\ m^{-1}$	Equivalent conductivity $\left(\Lambda = \frac{\kappa}{c}\right)$
5	22.2	4.4
3	17.8	5.9
2	13.5	6.8
1	7.7	7.8
0.5	4.2	8.5
0.1	1.0	9.7
0.05	0.5	9.6
0.01	0.1	10.0
0.005	0.05	10.4
0.001	0.01	10.7

Table 9.8. Refractive indices of some materials for light of 589 nm wavelength (Na D line) at 20 °C

Ethanol	1.360
Glass	1.517–1.890
Rock salt	1.544
Silica glass	1.459
Water	1.333

Table 9.9. Some data on biological substances

Density (kg dm^{-3})	
Bone	1.7–2.0
Cartilage (average)	≈1.1
Fatty tissue	0.92–0.94
Blood cells	≈1.1
Plasma	≈1.03
Blood (average)	≈1.06
Urine	1.001–1.035

Viscosity relative to water (at 20 °C)	
Plasma	1.8–2.0
Blood (average)	4.2–6
Cytoplasm	≈45
Endolymph	≈1.8

Specific heat (kJ kg^{-1} K^{-1})	
Blood	≈3.9
Compact bone	1.3–1.7
Fatty tissue	≈3
Body tissue (average)	≈3.5

Specific electric conductivity at 30 MHz frequency (Ω^{-1} m^{-1})	
Blood	≈1.1
Muscle	≈0.8
Spleen	≈0.6
Liver	≈0.5
Brain	≈0.45
Fatty tissue	≈0.05

Dielectric constant at 30 MHz frequency	
Blood	≈140
Muscle	≈110
Spleen	≈200
Liver	≈140
Brain	≈160
Fatty tissue	≈ 12

Table 9.10. Fundamental physical constants

Velocity of light in vacuum	$c = 2.998 \times 10^{8}$ m s^{-1}
Universal gas constant	$R = 8.314$ J mol^{-1} K^{-1}
Avogadro constant	$N_A = 6.02 \times 10^{23}$ mol^{-1}
Boltzmann constant	$k = 1.38 \times 10^{-23}$ J K^{-1}
Electron rest mass	$m_e = 9.11 \times 10^{-31}$ kg
Proton rest mass	$m_p = 1.67 \times 10^{-27}$ kg
Elementary charge	$e = 1.6 \times 10^{-19}$ C
Planck constant	$h = 6.62 \times 10^{-34}$ J s
Faraday constant	$F = 96{,}485$ C mol^{-1}

Table 9.11. Characteristic data on some important radionuclides

Chemical element and its atomic number		Isotope symbol	Physical half-life	Type of decay	Maximum particle energy (MeV)	γ-energy (MeV)
Hydrogen	1	^{3}H	12.33 years	β^-	0.0186	—
Carbon	6	^{11}C	20.4 min	β^+	0.96	—
		^{14}C	5760 years	β^-	0.155	—
Nitrogen	7	^{13}N	10 min	β^+	1.19	—
Oxygen	8	^{15}O	2 min	β^+	1.73	—
Fluorine	9	^{18}F	109.8 min	β^+	0.633	—
Sodium	11	^{24}Na	15.02 hours	β^-, γ	1.392	2.754 1.369
Phosphorus	15	^{32}P	14.28 days	β^-	1.710	—
Sulfur	16	^{35}S	87.2 days	β^-	0.167	—
Potassium	19	^{40}K	1.28×10^9 years	β^-, K (10%)	1.31	1.46 after K
		^{42}K	12.36 hours	β^-, γ	3.52 (75%) 1.99 (25%)	1.525
Calcium	20	^{45}Ca	163 days	β^-	0.257	—
Chromium	24	^{51}Cr	27.7 days	K, e^-, γ	0.315 (e^-)	0.320
Iron	26	^{52}Fe	8.2 hours	β^+, γ	0.8	0.5
		^{59}Fe	44.6 days	β^-, γ	1.566	1.30 1.10
Cobalt	27	^{60}Co	5.272 years	β^-, γ	0.318	1.33 1.17
Copper	29	^{64}Cu	12.74 hours	β^- (39%) β^+ (19%) K (42%) γ (1%)	β^-: 0.575 β^+: 0.656	1.34
Krypton	36	^{85}Kr	10.73 years	β^-, γ	0.687	0.514
Rubidium	37	^{81}Rb	4.7 hours	β^+, γ	0.99	1.93 0.95
		^{86}Rb	18.65 days	β^-, γ	1.78	1.078
Strontium	38	^{90}Sr	29 years	β^-	0.546	—
Yttrium	39	^{90}Y	64 hours	β^-, γ (0.4%)	2.29	1.761
Technetium	43	^{99m}Tc	6.02 hours	γ	—	0.140
Indium	49	^{113m}In	1.658 hours	γ	—	0.391
Iodine	53	^{123}I	13.3 hours	K, γ	—	0.16
		^{125}I	59.7 days	K, γ	—	0.0355
		^{131}I	8.04 days	β^-, γ	0.606 0.25 0.81	0.364 0.080 0.723
Xenon	54	^{133}Xe	5.29 days	β^-, γ	0.346	0.081
Caesium	55	^{137}Cs	30.1 years	β^-, γ	0.512 (92.6%) 1.173 (7.4%)	0.661
Gold	79	^{198}Au	2.695 days	β^-, γ	0.961	0.411
Mercury	80	^{203}Hg	46.6 days	β^-, γ	0.212	0.279
Radon	86	^{222}Rn	3.824 days	α	5.489	—
Radium	88	^{226}Ra	1600 years	α, γ (6%)	4.784	0.186 0.260
					4.598	0.609
Uranium	92	^{238}U	4.47×10^9 years	α, γ	4.2	0.048

APPENDIX

A. Some physical examples

A1. The Boltzmann distribution

Consider the following phenomena:

- the pressure and density distribution of air in the gravitational force field in case of thermodynamic equilibrium as the function of height;
- the altitude distribution of particles floating in gases and liquids at equilibrium;
- sedimentation equilibrium distribution during centrifugation;
- temperature dependence of the pressure of saturated vapours;
- temperature dependence of the rate and equilibrium constansts of chemical reactions;
- Maxwellian velocity distribution;
- temperature dependence of the concentration of thermal defects in ordered atomic systems (crystals, macromolecules).

A common feature of the above cases is that a multiparticle system is simultaneously influenced by effects causing order and disorder. Thermal motion promotes disorder, while the various force fields (e.g. gravitational field, interactions between molecules) contribute to order. The simultaneous existence of the two opposite "effects" produces the distribution of particles in thermal equilibrium, the *Boltzmann distribution.*

a) The most comprehensible example is the pressure and density distribution of air or of one its components at a given temperature as the function of height. The qualitative answer of the question is well known: both the pressure and the density decrease rapidly with increasing height.

To obtain quantitative relations, let us denote the pressure, density and molecular concentration at height h_1 by p_1, ρ_1 and n_1, and those at height $h_2 = h_1 + h$ by p_2, ρ_2 and n_2, respectively. From simple considerations we obtain:

$$\frac{p_2}{p_1} = \frac{\rho_2}{\rho_1} = \frac{n_2}{n_1} = e^{-\frac{\Delta\varepsilon}{kT}}, \quad \Delta\varepsilon = \mu g h_2 - \mu g h_1$$

where T is the absolute temperature, μ the mass of a molecule, k the Boltzmann constant, and g the gravitational acceleration. $\Delta\varepsilon$ denotes the potential energy difference between molecules at heights h_2 and h_1. Thus, from the formula, the pressure of the gas decreases with altitude. Since the concentration of molecules and with it the density of the gas too is proportional to pressure, the formula gives the variation of these quantities as well. As the relation is applicable to the atmosphere of the Earth, it is called the *barometric altitude formula.*

To derive the formula, let us consider a thin horizontal layer between heights h and $h + dh$ in a vertical gas column at constant temperature, and let us denote the corresponding pressure values by p and $p + dp$, respectively. If dh is sufficiently small, the density of the gas can be considered constant; this is denoted by ρ. The pressure on the bottom of the layer will be evidently higher than that on its top surface due to the weight of the layer; thus:

$$dp = -\rho g dh$$

The negative sign reflects the fact that a positive dh value is associated with a negative dp. Substitution of ρ from [1.32] yields the following differential equation:

$$\frac{dp}{p} = -\frac{\mu g}{kT} dh$$

Integration between the limits h_1, h_2 and p_1, p_2 respectively, gives the formula in question.

The formula correctly describes the altitude distribution of particles suspended in air and the distribution of colloidal particles floating in liquids, so long as the temperature in the studied volume can be considered constant. Naturally, in the calculation of the potential energy of the particles both the gravitational force and the Archimedean buoyancy must be taken into account. Further, this formula can be used to determine the sedimentation equilibrium distribution (cf. section 4.6.5).

b) The barometric altitude formula is a special case of a more general distribution law, the *Boltzmann distribution.*

According to the relative number (relative concentration n_2/n_1) of molecules whose energy differs by $\Delta\varepsilon = \varepsilon_2 - \varepsilon_1$:

$$\frac{n_2}{n_1} = e^{-\frac{\Delta\varepsilon}{kT}} \quad \text{or} \quad \frac{n_2}{n_1} = e^{-\frac{\Delta E}{RT}}$$

The second equation is obtained from the first by multiplying the numerator and denominator of its exponent by the Loschmidt number. Thus, k is substituted by $kL = R$ (general gas constant), and $\Delta\varepsilon$ by $\Delta\varepsilon L = \Delta E$, the energy difference per mole. The power expressions in these formulas are frequently called the Boltzmann factor.

If the Boltzmann distribution is applied to the freely (randomly) moving particles of a perfect gas, ε is their kinetic energy. From the energy distribution the velocity distribution of the particles can be derived as well (cf. Maxwellian velocity distribution, section 1.4.1(c)). If the particles are not moving freely but in some force field (e.g. in the gravitational field or in the force field of the neighbouring particles), ε contains besides the kinetic energy the potential energy as well. In the cases mentioned in **a)** just the potential energy plays a role.

c) On the basis of the Boltzmann distribution the following relation can be reached which gives the variation of the vapour tension of a substance with temperature (T) to a good approximation:

$$p \sim e^{-\frac{\Delta E}{RT}}$$

Here ΔE is the heat of evaporation (related to 1 mole). This is the energy difference of 1 mole gas and 1 mole condensed substance in thermodynamic equilibrium at temperature T.

The *rate of a chemical reaction* is proportional to the number of activated molecules, i.e. molecules having sufficiently high energy to react with each other. In *thermal activation* the concentration c^* of molecules with sufficient excess energy, the *activation energy*, satisfies the relation

$$c^* \sim e^{-\frac{\Delta E}{RT}}$$

where ΔE is the activation energy related to 1 mole. Since the reaction rate is proportional to the number of activated molecules, the rate constant k will be

$$k \sim e^{-\frac{\Delta E}{RT}}$$

The latter relation also demonstrates that the rate of a chemical reaction increases exponentially with increasing temperature.

It follows from the above relations that the temperature dependence of the *equilibrium constant of a chemical reaction* ($K = \vec{k}/\overleftarrow{k}$) is given by

$$K \sim e^{-\frac{\Delta H}{RT}}$$

where ΔH is the heat of reaction ($\Delta H = \vec{\Delta E} - \overleftarrow{\Delta E}$). ΔH is negative in exothermic reactions, and consequently for this reaction type K decreases with increasing temperature. For endothermic reactions, on the other hand, K increases with increasing temperature.

Other applications of Boltzmann distribution are the Maxwellian velocity distribution (cf. section 1.4.1), and relation for the concentration of Schottky defects (cf. section 1.4.3), and we shall meet it again in connection with the temperature dependence of the concentration of thermal defects in macromolecules (cf. sections 1.5.2–1.5.5).

A2. Light refraction on spherical surface

1. Optical power or refractivity. Diopter. The simplest imaging system is a spherical cap-shaped boundary surface on both sides of which there are homogeneous media having different refractive indices. The optical lenses and the eye as an optical system are also built like this.

Consider Fig. A/1. In all four parts of the figure, n and n' are the refractive indices of the media, r is the radius of curvature of the boundary surface.

The straight line passing through the acme of the cylinder A and the centre of the curvature O is the *principal* or *optical axis*. The rays starting from object-point P are either *convergent* or *divergent* after refraction. In the first case the rays meet in one point, while in the second one they proceed as if coming from one point. Thus in the latter case only the rays extended backwards meet. Figures *a* and *b* are examples of the first case, *c* and *d* of the second, respectively. Point P' is the image of point P in each case. Figures *a* and

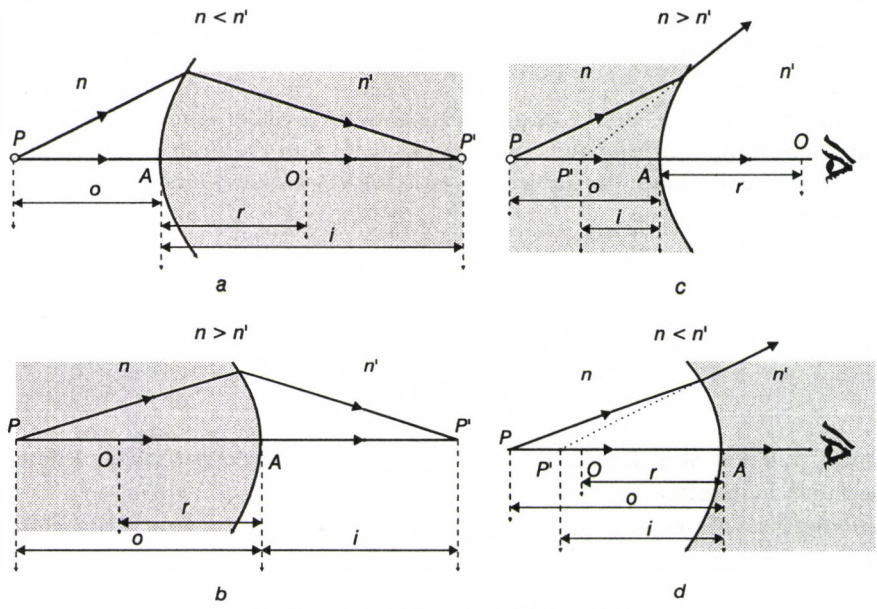


Fig. A/1. Light refraction on spherical surfaces

b show the formation of *real* images, while Figs *c* and *d* that of *virtual* images. The image is pointlike only if the rays coming from the object point reach the boundary surface almost perpendicularly (*paraxial rays*). In what follows only such cases will be dealt with.

The position of object point P is determined by the *object distance* o, that of image point P' is determined by the *image distance* i. Both of them are measured from acme A. *We assign signs to the distances. The distances which are in the direction of the incidence measured from A are considered positive, while those which are in the opposite direction negative.*

By changing o, i *also* changes, according to the following relation:

$$\frac{n'}{i} - \frac{n}{o} = \frac{n' n}{r} \qquad [1]$$

[1] can be obtained by simple geometrical considerations, using the rule of refraction, if we bear in mind that in case of paraxial rays the angles are small enough so that instead of their sines or tangents the angles themselves may be used in the calculations (or vice versa).

Deduction. According to the rules of geometry, in Fig. A/2 $\alpha = \varepsilon + \varphi$ and $\varphi = \alpha' + \varepsilon'$. Therefore

$$\frac{\alpha}{\alpha'} = \frac{\varepsilon + \varphi}{\varphi - \varepsilon'}$$

According to the rule of refraction $n'/n = \sin \alpha / \sin \alpha'$. However, as mentioned above, in case of paraxial rays the angles are small enough so that instead of their sines the angles may be used in the calculation, thus $\sin \alpha / \sin \alpha' = \alpha / \alpha'$. Therefore instead of the above equation we may write

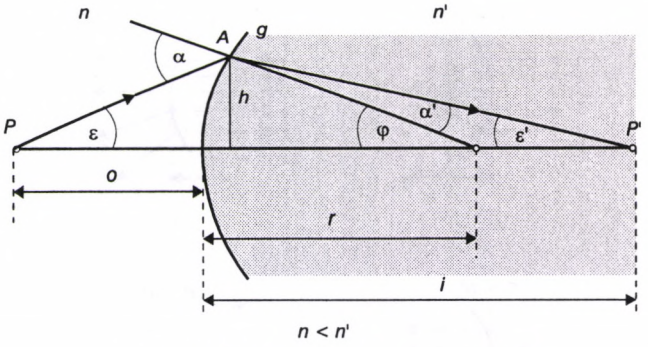


Fig. A/2. Scheme for the deduction of [1]

$$\frac{\varepsilon+\varphi}{\varphi-\varepsilon'}=\frac{n}{n'}$$

In case of paraxial rays also the following relations are valid to a good approximation: $\varepsilon \approx h/o$, $\varepsilon' \approx h/i$, $\varphi \approx h/r$. By using these the above formula takes the form of [1], which was the object of the deduction.

The right side of [1] depends only on the refraction of the media and the radius of curvature of the boundary surface, thus it is constant in case of given media and boundary surfaces. The quantity

$$D=\frac{n'-n}{r} \qquad [2]$$

is the *optical power* or *refractivity* of the spherical surface forming the boundary of the refractive media. Its unit used in practice is m^{-1}, called *diopter* (its sign: dptr).

The refractivity of a refractive surface may be positive or negative. In case of Fig. A/1*a* $n' > n$ and $r > 0$, consequently the signs of the numerator and the denominator in [2] are the same (both are positive), thus the refractivity is positive. In case of Fig. A/1*b* $n' < n$ and $r < 0$, the numerator and the denominator have the same sign here too, therefore the refractivity is positive. It may be shown in a similar way that in the cases of Figs A/1*c* and *d* the signs of numerator and the denominator are different, thus these boundary surfaces are characterized by a negative refractivity.

Considering [2], [1] takes the following form:

$$\frac{n'}{i}-\frac{n}{o}=D \qquad [3]$$

If the refractivity is positive, the boundary surface is converging, if it is negative, the boundary surface is diverging. Thus Figs A/1*a* and *b* show converging boundary surfaces, while Figs A/1*c* and *d* diverging boundary surfaces, respectively.

2. Focal points and focal distances. Let a beam of rays parallel to the principal axis fall on the refractive surface. If the refractivity is positive, the refracted rays are convergent, their common point of intersection is the second focus, F', on the side of the image (Fig.

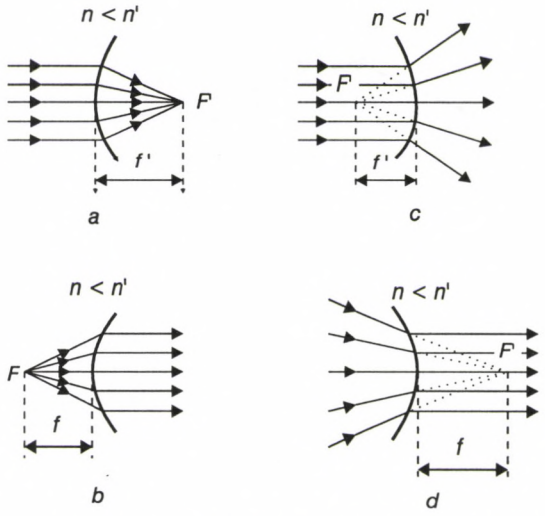


Fig. A/3. Real and virtual points of focus. The positions of the first (F) and second (F') focal points

A/3*a*). Point F, producing rays which become parallel to the principal axis after refraction, is the first focus, and is on the side of the object (Fig. A/3*b*). The foci are on the principal axis, their distances from the acme of the surface f and f' are the *focal distances* of the boundary surface.

Boundary surfaces with negative refractivity also have focal points. However, while in the case of positive refractivity the focal points are real, in case of negative refractivity they are virtual. Here the second focus is the virtual image point (F') from which the rays falling parallel to the principal axis seem to start after refraction (Fig. A/3*c*). The first focus is the point F. Rays passing through this point are parallel to the principal axis after refraction (Fig. A/3*d*). In case of converging boundary surfaces f is negative, f' is positive, while in case of diverging surfaces the signs are reversed.

By using [3], a simple relation may be obtained with respect of the focal distances. Namely, if the substitution $o = \infty$ is carried out, the resulting image distance is equal to f'. Thus

$$f' = \frac{n'}{D} \quad \text{and} \quad D = \frac{n'}{f'} \qquad [4a\text{–}b]$$

If, on the other hand, the substitution $i = \infty$ is performed, the resulting object distance is f.

$$f = -\frac{n}{D} \quad \text{and} \quad D = -\frac{n}{f} \qquad [5a\text{–}b]$$

Consequently: *the product of the reciprocals of both focal distances and the pertaining refractive indices* (disregarding the signs) *is the refractivity.*

Example. Let us calculate the focal distances and the refractivity if $r = 8$ mm, $n = 1$ and $n' = 1.336$. According to [2], $D = 42$ dptr; according to [4] and [5], $f \approx -24$ mm,

$f' \approx 32$ mm. – The example gives data approximately corresponding to those of the lentectomized eye if for the sake of simplicity the eye is considered a medium with a homogeneous refractive index (n').

3. The position and quality of the image. Image construction. With the help of spherical boundary surfaces images can be made not only of luminous points but also of bodies consisting of an infinite number of luminous points. The practical significance of these systems lies just in this. In case of paraxial rays the image of a straight line is practically straight, the image of a plane is plane. The question is, what kind of an image is formed and where at a given object distance. The image, considering its quality, may be real or virtual, enlarged or reduced, erect or inverted with respect to the object.

With respect to the position of the object, an easily manageable relation may be obtained if [3] is transformed by using [4] and [5]. This is the frequently applied *lens formula:*

$$\frac{f'}{i} + \frac{f}{o} = 1 \qquad [6]$$

The image is real if o and i have opposite signs, and virtual if they have the same signs. This statement may be read already from Fig. A/1*a–d.*

The answers to the remaining questions concerning the degree of magnification and the position of the image in the different positions of the object will be given later, in [7], the most important statements, however, are summarized here.

In case of *positive refractivity (converging boundary surface):*

a) If the object is outside the double focal distance, the image is formed on the other side of the refractive surface, between the single and double focal distances; it is real, inverted and reduced.

b) If the object is at the double focal distance, the object is also formed in the double focal distance on the other side; it is real, inverted and has the same size as the object.

c) If the object is between the single and double focal distances, the image is on the other side outside the double focal distance; it is real, inverted and enlarged.

d) If the object is inside the single focal distance, the image is formed on the same side; it is virtual, erect and enlarged.

In case of *negative refractivity (diverging boundary surface):*

The image is always formed on the side of the object; it is virtual, erect and reduced.

The position and quality of the image in a given case can be determined also by geometrical construction. Figure A/4 shows examples for this. Its part *a* demonstrates the formation of a real image, while part *b* that of the virtual image, respectively. For the sake of simplicity the object is always illustrated by an arrow AA' standing on the principal axis normal to it. Its image BB' is also normal to the principal axis. Considering that the image of the base point of the arrow is on the principal axis, we will confine ourselves to the construction of the image of the arrow point. Any of the rays starting from the arrow point may be used for the construction, but the following two ray paths may be easily followed:

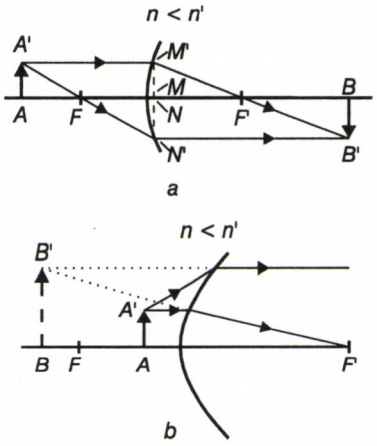


Fig. A/4. Image construction

a) the ray which falls to the boundary surface from a path parallel to the principal axis and after leaving the surface travels towards the second focus;

b) the ray which falls to the boundary surface coming from the focus and after refraction travels parallel to the principal axis.

The above two ray paths suffice for the construction of the image point.

Although Fig. A/4*a–b* show converging boundary surfaces, the construction may be carried out also for diverging boundary surfaces according to the same consideration.

4. Linear magnification. It is the relation of a linear dimension of the image to the *corresponding* linear dimension of the object. Thus, in Fig. A/4*a–b* the magnification is given by the quotient $N = BB'/AA'$. The quotient is larger than 1 in case of enlarged images and is smaller than 1 in case of reduced images. A *sign* is also assigned to the magnification: according to the convention, *the magnification is positive if the position of the image is erect and negative if it is inverted*. Thus, magnification is e.g. –1/2, if the longitudinal dimensions of the image are half of those of the object and the position of the image is inverted.

The magnification N may be expressed in several ways by the data characteristic of the system, with the help of f and i, and also considering the signs:

$$N = \frac{i - f'}{f'}, \quad N = \frac{f}{o - f} \quad N = \frac{n}{n'} \times \frac{i}{o} \qquad [7a–c]$$

From [7c] it follows that N is positive, i.e. the image is in the erect position if the object distance and the image distance have the same signs, i.e. the image is on the same side of the refractive surface as the object; N is negative, i.e. the image has an inverted position if the signs of the image distance and the object distance are different, i.e. the image and the object are on the opposite sides.

Deduction. It is sufficient to deduce only one of the above relations, e.g. [7a], since the others may be obtained from this by a simple transformation by using the known relations.

Consider Fig. A/4*a*. It follows from the similarity of triangles $BF'B'$ and $MF'M'$ that $\overline{BB'} : \overline{MM'} = (o - f') : f'$. When writing the proportion it was also taken into consideration that in case of paraxial rays the point M practically coincides with the acme of the cylinder. If in addition it is also considered that according to the construction $\overline{MM'} = \overline{AA'}$ and the image is in the inverted position, the written proportion takes the form of [7a]. The same result may be also obtained naturally on the basis of Fig. A/4*b*.

A3. System of centred surfaces. Optical lenses

In the following such systems will be dealt with which consist of several different spherical boundary surfaces and in which the surfaces are separated from each other by media with different refractions (cf. section A2). If the centres of the curvature are on the same straight line, the system is *centred*, and the line is the principal axis of the system. In Fig. A/5*a* only the first and last boundary surfaces are presented, and n stands for the refraction of the *first*, n' for that of the *last* medium. The rays travelling parallel to the principal axis go through F' after leaving the system. This is the *second focus of the system.* On the other hand, point F, rays starting from which become parallel upon leaving the system, is the *first focus of the system.* Centred systems are e.g. the optical lenses made usually of glass, and our eyes, too.

1. Thin systems (thin lenses). Systems in which the distance between the first and the last refractive surface is negligible as compared to the distance of the focal points from the system of the refractive surfaces are called thin lenses. Graphically the thin system is illustrated by a single straight line which represents the practically coincident boundary surfaces (Fig. A/5*b*).

The following fundamental statements can be made concerning the thin systems:

a) All relations and statements valid for a single refractive surface (cf. section A2) are also valid for the system of refractive surfaces.

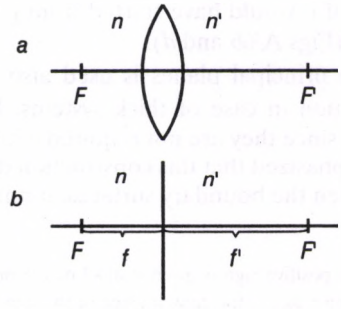


Fig. A/5. Centred system

b) The refractivity D of a system of several refractive surfaces is equal to the algebraic sum of the refractivities of the individual refractive surfaces D_1, D_2, ..., D_n, i.e.:

$$D = D_1 + D_2 + \ldots + D_n \qquad [1]$$

Obviously, in case of a usual optical lens there are two boundary surfaces, thus a sum of two refractivities is considered.

c) The "resultant" refractivity may be positive or negative. In the first case the lens is *converging* or *convex,* in the second one it is *diverging* or *concave.*

If the media are the same in front of and behind the lens, then by combining $n = n'$ with A2 [4] and A2 [5] $f = -f'$. This means that the absolute values of the focal distances are the same, only their signs are different.

In such cases some relations become simplified, e.g. the lens equation (cf. section A2) may be written in the following forms:

$$\frac{1}{i} - \frac{1}{o} = \frac{1}{f'} \quad * \quad \text{and} \quad \frac{1}{o} - \frac{1}{i} = \frac{1}{f} \quad * \qquad [2a\text{–}b]$$

A further simplification is if the lens is surrounded by air. In this case $n = n' = 1$, $D = 1/f'$ and $D = -1/f$. Therefore: If the *optical lens is surrounded by air, the optical power of the lens is equal to the reciprocal of the focal distance (disregarding the signs).*

2. Thick systems (thick lenses). The statements concerning the thin systems (thin lenses) are also valid for the thick systems (thick lenses) if the focal points and the positions of the object and the image are measured from appropriately selected and easily definable planes, the so-called *principal planes.* (Only [1] has to be corrected in case of thick systems, but this will not be discussed here.) Every thick system has two principal planes which are normal to the principal axis (Fig. A/6). The points of intersections K and K' of the principal planes (H and H') are the *principal points.* Thus the focal distances are represented by the distances FK and $F'K'$, having the proper sign. The object distance is measured from the principal point K, while the image distance from principal point K'.

The principal planes are characterized by their most important feature: each ray which falls to the first boundary surface travelling towards a point P of principal plane H, leaves the last boundary surface as if it would have started from point P' of principal plane H', which is opposite to point P (Figs A/6*b* and *d*).

The above feature of the principal planes is used also in Fig. A/7, which gives an example for image construction in case of thick systems. In the figures the boundary surfaces are not even shown, since they are not required with respect of the construction. Nevertheless, it has to be emphasized that this construction does not give any information about the path of light between the boundary surfaces, it only reveals how the rays travel

* In the literature frequently the positive sign is given instead of the negative one in the lens equation. In the present case the use of the negative sign is the consequence of the sign convention mentioned in point 1 of section A2.

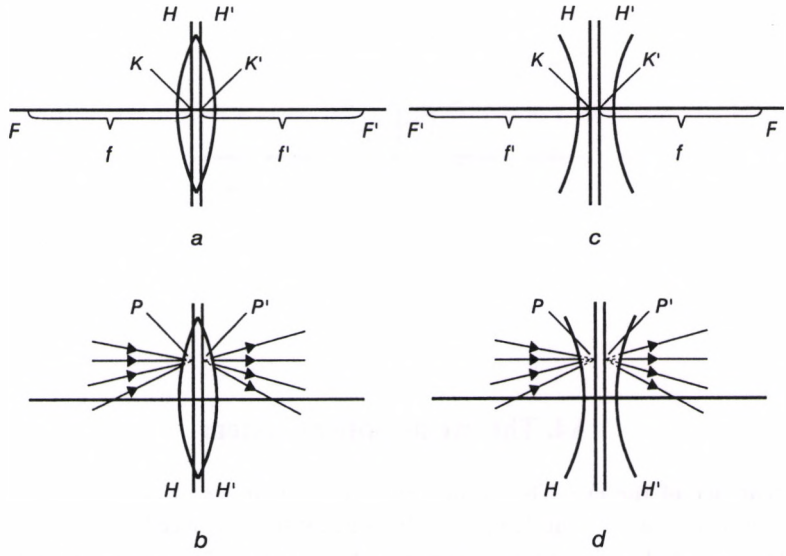


Fig. A/6. Principal planes and principal points in convex lenses *(a–b)* and concave lenses *(c–d)*

before and behind them. For the practical constructions this is usually satisfactory. If Fig. A/7 is compared with Fig. A/6, it may be also said that the thin system (even a single boundary surface!) is a special case in which the two principal planes and therefore also the two principal points coincide.

Finally another remark. There are always rays which, coming from an object point, fall on the system in a way that they leave it without changing their direction, undergoing only parallel displacement. This ray path is shown in Fig. A/8. The points in which the extensions of the ray falling on the system and leaving it meets the principal axis are the *nodal points*. In the figure the nodal points are represented by points C and C'. If the media on the two sides of the system have the same refractivity, the nodal points coincide with the principal points.

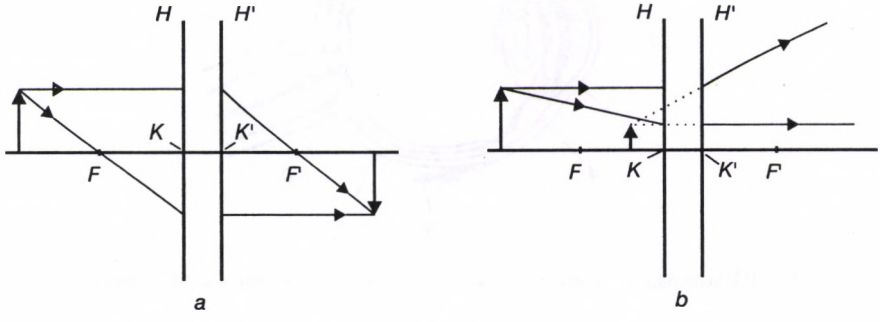


Fig. A/7. Image construction for thick lenses *a*: convex lens, *b*: concave lens

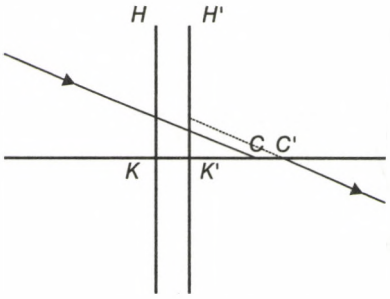


Fig. A/8. Nodal points

A4. The eye as optical system

1. The structure of the eye. The human eye has a diameter of about 24 mm, it is an approximately spherical organ[1] (Fig. A/9). Its wall consists of three layers. The outermost of them is the *sclera (1)*; its anterior, somewhat protruding, transparent part is the *cornea (2).* The middle, *vascular layer (uvea)* consists of three parts. Its posterior two-thirds is the *choroid (3)*, which becomes gradually thicker in the anterior third and continues in the *ciliary body (4)*. The latter contains the *ciliary muscle* which regulates the optical power of the *crystalline lens (5).* In the very front the vascular layer becomes thin again and is called *iris (6).*The iris defines the colour of the eye. In the middle of the iris there is a round opening, the *pupil.* Its variable diameter determines the amount of light getting into the eye. The innermost, only a few tenths of mm thick, transparent layer is the *retina (7).* This is the screen of our eye on which the reduced, real and inverted image of the objects are formed. The retina contains the end-organs of the optic nerve in which the visual stimuli

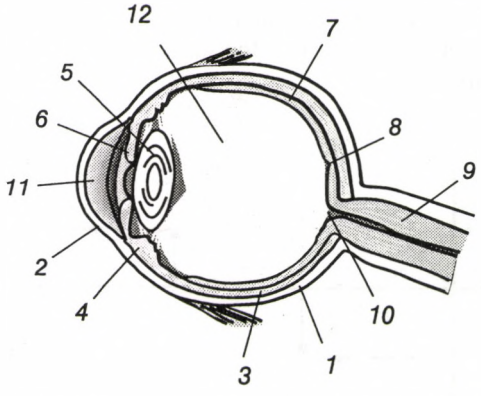


Fig. A/9. Schematic structure of the eye (see the text for the meaning of the numbers)

[1] Data listed are referred to healthy (emmetropic) eye.

are transformed into neural stimuli. The photosensitive elements of the retina are the *cones* and *rods*. They are most densely located in and around the *macula lutea (yellow spot; 8)* of the retina which contains a yellow pigment (cf. also point 4). When we want to see an object sharply, we bring our eyes in such a position that the image will be on the macula lutea. About 4 mm nasally from the macula lutea there is the *optic nerve (9)*. Its site of entrance is the *optic papilla (papilla nervi optici; 10)*. Here there are no photosensitive end-organs (*blind spot*).

The rays of light travel through several refractive media before reaching the retina. The cornea has already been mentioned. Between the iris and the cornea there is the anterior chamber (*11*) which is filled with a watery solution, the *humor aqueous*. Behind the iris is the biconvex lens (*5*), its diameter is about 10 mm, its thickness is about 4 mm. Its anterior surface is at a distance of approximately 3.6 mm, while the posterior is at a distance of about 7.2 mm from the cornea. Its surface facing the iris is less convex than the posterior one. The radii of curvature are 10 mm and 6 mm, respectively. The inner portion of the eye is filled with the *vitreous body (corpus vitreum; 12)* which has a gelatinous consistence.

The rays of light entering the eye are refracted first on the cornea having a radius of curvature of 7.8 mm. The refractive index of the cornea is 1.376. The next refraction takes places when the light enters the humor aqueous from the cornea. The refractive index of the humor aqueous is 1.336. The vitreous body has the same refractive index, thus before and behind the lens there are media the refractive indices of which are practically identical and also similar to that of water.

The lens itself has an onion-like layered structure. Its refractive index increases towards its centre: 1.386–1.406. Due to its peculiar structure the lens behaves as if it would consist completely of a substance with a refractive index of 1.41, i.e. a substance the refractive index of which is even larger than that of the lenticular nucleus.

2. Refractivity of the eye. According to the above, refraction takes places on four boundary surfaces in the eye: on the anterior and posterior surfaces of the cornea and on the anterior and posterior surfaces of the lens. The refractivity of the individual surfaces may be easily computed on the basis of A2[2], by using the above data:

the refractivity of the anterior surface of the cornea:

$$D_1 \approx 48.3 \text{ dptr},$$

the refractivity of the posterior surface of the cornea:

$$D_2 \approx -5.1 \text{ dptr},$$

the refractivity of the anterior surface of the lens:

$$D_3 \approx 7.4 \text{ dptr},$$

the refractivity of the posterior surface of the lens:

$$D_4 \approx 12.3 \text{ dptr}$$

If the eye is considered a thin system, its refractivity calculated according to A3[1] is approximately 63 dptr, from which the lens makes out only about 20 dptr. Thus the refractivity of the eye is determined mainly by the anterior surface of the cornea and less by the lens.

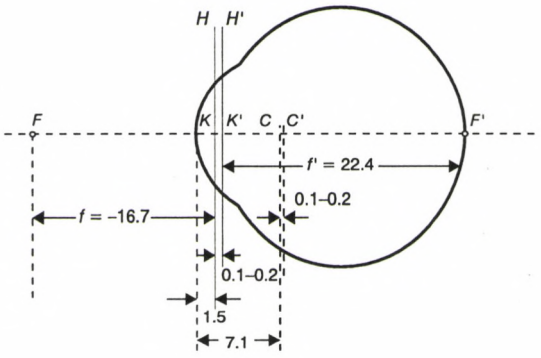


Fig. A/10. The positions of the focal points (*F* and *F'*), principal planes (*H* and *H'*), principal points (*K* and *K'*) and nodal points (*C* and *C'*) of the eye. The distances are given in mm

According to more exact examinations, our eye cannot be considered a thin system. The *average* data of the normal (emmetrope) eye are shown in Fig. A/10.

In the *reduced eye* the principal planes close to each other are substituted with a single principal plane on the object side of which there is air ($n = 1$), on the image side a medium with a refractive index somewhat larger than that of water ($n' = 1.336$). The common principal plane is 1.6 mm from the anterior surface of the cornea. The nodal points of the reduced eye also coincide. The united nodal point is 7.2 mm from the surface of the cornea.

3. Distance adjustment (accommodation). In the accommodation the main role is played by the lens. The non-accommodating eye sees sharply the object being in infinite distance. The essence of accommodation is that at a decreasing object distance the lens becomes more convex due to the contraction of the ciliary muscles. The above data referred to the average eye looking in the infinite (not accommodated). During accommodation mainly the radius of curvature of the anterior surface of the lens decreases (the curvature increases), namely from 10 mm to about 6 mm. Due to the peculiar lamellar structure of the lens the average refractive index also increases, to a maximum of 1.424. Because of the accommodation the refractivity of the eye may be increase by about 10 dptr, thus it may reach 70 dptr.

The most distant point from which the rays still form a sharp image on the retina is the *far point (punctum remotum; R)* of the eye. The far point of the healthy eye is in the infinite. The nearest point from which the eye is able to form a sharp image is the *near point (punctum proximum; P).* At the age of twenty years the near point is about 10 cm from the eye and by increasing age it becomes more and more distant. At the age of fifty years it is about 40 cm from the eye. Thus the accommodation decreases with the age.

Looking at the near point takes up the complete accommodation capacity of the eye which is very tiring on the long run. The distance at which fine works, e.g. reading may be carried out for a long time without overstraining one's eye is the *distance of clear sight.* Its value in healthy eyes is about 25 cm.

The accommodation is measured by the optical power of the thin glass lens D_A which, placed in front of the eye accommodated to the far point (theoretically to the common principal plane), would adjust the eye to the near point. If the distances of the far point R, o_R and of near point P, o_P (from the common principal plane) are known, D_A can be easily calculated:

$$D_A = \frac{1}{o_R} - \frac{1}{o_P} \qquad [1]$$

The quantities $1/o_R$ and $1/o_P$ are usually given in m^{-1}, in other words in dioptres and are indicated as D_R and D_P, respectively. Thus [1] may be written also in the following form:

$$D_A = D_R - D_P \qquad [2]$$

Verification. Let the refractivity of the eye set to the far point be D. In this case the object distance is o_R, the image distance is the distance of the retina from the common principal plane, i. According to A2[3], $n'/i - n/o_R = D$. Let us place the thin lens with a refractivity of D_A immediately in front of the eye. According to A3[1] the refractivity of the combined system will be $D + D_A$. D_A should be selected so that at object distance o_P the image distance i be as previously. In this case, according to A2[3], $n'/i - n/o_P = D + D_A$. Subtracting the first equation from the second and considering that $n = 1$, relation [1] is obtained.

Examples

1. How many dioptres is the accommodation ability of the eye the far point of which is in the infinite, while its near point is at a distance of 10 cm? Since in this case $o_R = -\infty$ and $o_P = -0.1$ m, according to [1] the accommodation ability is 10 dptr. Approximately so much is the accommodation ability e.g. at the age of 20 years.
2. How much is the accommodation ability if $o_R = -\infty$ and $o_P = -0.4$ m? Calculating in the similar way as above, 2.5 dptr are obtained as result. This is the extent of the accommodation ability e.g. at the age of 50 years.
3. How much is the accommodation ability if $o_R = -1$ m and $o_P = -0.2$ m? According to [1], the result is 4 dptr.

4. Daylight and night vision (adaptation). As already mentioned, the photosensitive elements are the most densely located in and around the yellow spot. At the thin centre of the yellow spot, in the *central fovea (fovea centralis)* there are only cones in a dense arrangement, the rods are at the margin of the yellow spot and in other areas of the retina. In these places the rods are already in majority, their average proportion is 20 rods to 1 cone. The total number of the cones is estimated to be 7 million, that of the rods 130 million, respectively. The cones are about 30 μm long and have a diameter of 5–6 μm, while the rods are about 60 μm long and their diameter is about 2 μm.

The cones and the rods play different roles in the perception of light. The rods are more sensitive than the cones but they do not react to colours. Those areas of the retina where the density of the cones is high are more able to perceive fine details than its other areas. On the basis of all this, there are two types of vision. One of them is *daylight* vision which is connected to the excitation of the cones and functions only at an adequate illumination, but it makes possible the differentiation between colours and the perception of fine details. The other is the *night* vision which is connected to the excitation of the rods, it is less coloured and less fine. Generally the two processes are present together and the role of the rods comes in the forefront only at weak illuminations.

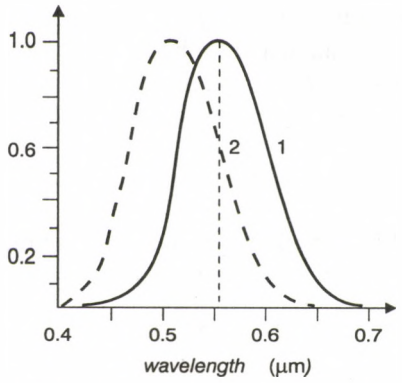


Fig. A/11. The sensitivity curve of the eye *1*: at daylight illumination, *2*: at twilight
On the ordinate the sensitivity is given in relative units (cf. section 2.3.2)

The sensitivity curve of the night vision is similar to that of the daylight vision but it is shifted towards the shorter wavelengths (Fig. A/11). The maximum of the curve of daylight vision is at about 555 nm, while that of the night vision is at about 510 nm.

The rods reach the maximum of their sensitivity only after a while; we may also say that the switch from daylight vision to night vision takes time. This process is the *adaptation.* The complete adaptation requires about 30 to 40 minutes.

5. Sensitivity of the eye. Our eye already perceives light which brings about approximately 10^{-9} lx illumination on the pupil. This means that light sensation may be evoked if in every 1/10 sec only 100 photons fall on the whole pupil. The great sensitivity of the eye becomes more evident if we consider that about 50% of the photons do not even reach the retina due to absorption and reflection. Moreover, not all the photons reaching the retina are absorbed by the cones and rods. Even in the optimal case, the light sensation comes from only about 10% of the incident photons, in our case from roughly 10 photons.

It is worth comparing the threshold of the illumination with the maximum illumination which is tolerated by the eye without pain or injury. If the illumination is gradually increased, some 100 lx is well tolerated, however, a weaker illumination may elicit unpleasant sensations if it appears abruptly.

6. Resolving power of the eye. Visual acuity. According to the experience, the smallest visual angle at which two object points can be differentiated from each other by the eye is about one minute. In this case the images of the two points are at a distance of about 5 μm from each other on the retina. This is the size of the diameter of one cone. Thus it may assumed that the eye is not able to resolve smaller distances because then the images of the two points fall on the same cone. However, differentiation is possibly only if the images of the object points are formed on different cones.

At first sight we may think that we might be able to see finer details if there would be finer photosensitive elements in our eye. Nevertheless, the calculations reveal that a greater resolving power could not be reached in this case either because of the light

diffraction following from the wave properties of the light. *The resolving power of our eye reaches even so the limit set by the light diffraction.* The limitation of the resolving power by diffraction was discussed in connection with the microscopes (cf. section 4.1.2). The argumentation presented there may be adapted also to the eye.

The physician speaks of *visual acuity (visus)* rather than resolving power. If the smallest visual angle at which the eye is still able to perceive details is about one minute, the visual acuity is one unit. If the perception of details takes places only at a larger angle, the visual acuity is smaller than 1. If, e.g. a visual angle of 2 min is necessary for the resolution, the vision is 1/2 = 0.5, in case of 4 min it is 1/4 = 0.25, etc.

7. Visual field. The visual field of the eye at rest extends to about the half of the total solid angle. The visual field is larger in the horizontal than in the vertical direction. Nasally we see to about 60°, temporally, on the other hand, we even perceive objects which are at 95° from the optical axis. Upwards and downwards the visual field is about 60°. – The visual fields of the two eyes partially overlap (Fig. A/12). The common area comprises horizontally approximately 120°. – Without moving the head, the eyes can be moved in each direction by about 30° to 40°. If a larger turn is required, the head is moved.

The visual acuity reaches its maximum only at a very small solid angle, it decreases towards the periphery of the visual field. At the periphery the visual acuity is so small that actually we are able to perceive only movements. The solid angle of the maximal visual acuity is determined by the size of the area of the central fovea. This means a solid angle by which the nail of our index finger may be seen while the arm is completely extended. Nevertheless, all this does not disturb our vision, because if something arouses our interest at the periphery, we move immediately our eyes in that direction and the image of the object already appears on the yellow spot. Under such conditions the blind spot does not hinder our vision either.

Although in our eye two-dimensional images are formed, we are still able to perceive depth. An important role in this is played by binocular vision. If we see a detail of an object sharply, its image is projected on the yellow spot in both eyes. If the detail is far enough, the angle included by the optical axes of the two eyes, the angle of convergence, is practically zero. The angle of convergence increases with decreasing distances (Fig.

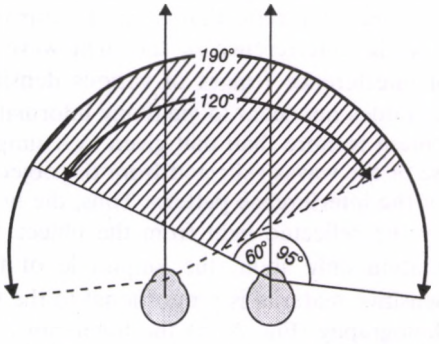


Fig. A/12. The visual field of the eye. The shaded area illustrates the binocular visual field

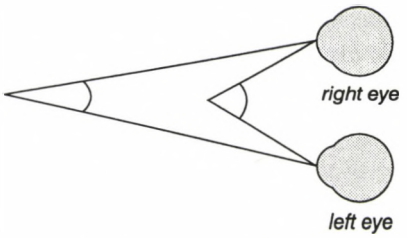


Fig. A/13. The convergence of the visual axes

A/13). We sense the angular change and on its basis we are able to estimate the distance of the objects. The estimation is more exact for near objects; this is understandable if we consider that in case of far objects the angle of convergence changes only slightly with changing distances.

The most substantial condition of depth perception is that slightly different images of the objects are formed in the two eyes. The distance between our eyes is 65 to 72 mm. Therefore the right eye sees the object somewhat from the right, the left eye somewhat from the left. The two images are united in the cerebral cortex and the depth is estimated from their differences.

The estimation of distances is made easier by the virtual size of the known objects (people, houses, etc.) and also by the changes in the colour of the objects caused by the light diffraction of the intermediate air layer. In monocular vision just the above-mentioned empirical knowledge plays a role in depth perception.

A5. Holography

Holography allows the production of three-dimensional enlarged pictures of various objects. The method consists of two steps. The first step *produces the hologram* of the object, and the second step *(reconstruction)* reconstructs the picture from the hologram. Every hologram is an *interference pattern* on a photographic plate or film (Fig. A/14). The light from the light source strikes the plate or film via two routes: after reflection from the object *(object wave),* and after reflection from the mirror *(reference wave).* The hologram is produced by the interference of the light waves arriving from the two directions. A system of interference fringes of various densities is produced on the photographic plate, which does not seem to yield any information about the object. In fact, the information content is more than that given by a simple photograph since the amplitude and the phase of the waves reflected from the object play equally important roles in the formation of the interference pattern. Thus, the hologram contains the full information collected by the reflected light from the object. The usual photographic picture provides information only about the amplitude of the reflected beam: the darkening of the light-sensitive material is proportional to the square of the amplitude. In the second step of holography (Fig. A/15) the hologram is transilluminated, which results in two images of the object, one *real* and the other *virtual.*

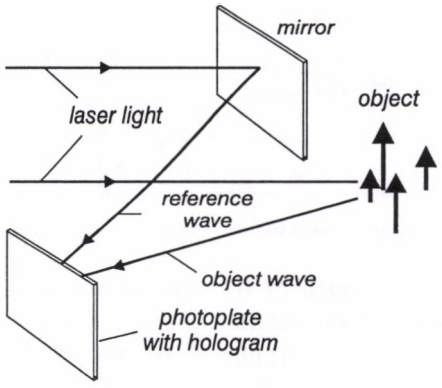


Fig. A/14. Production of a hologram

The production of a hologram requires light capable of interference over a long path, which means that the light should be a laser. The reconstruction does not necessarily demand the use of a laser, though a really good-quality image cannot be obtained without coherent light. The wavelength in the reconstruction need not be the same as that used in making the hologram.

The main properties of the holographic method and images may be summarized as follows.

– No lenses are required, either to make the hologram or for the reconstruction.

– The reconstruction is three-dimensional. If at inspecting the image the head is moved in the proper direction and to a proper extent, previously concealed details become visible.

– The image of the whole object can even be reconstructed from merely a part of the hologram with loss of only some of the finer details. The more details are lost, the smaller the domain of the hologram used to produce the picture. Dust, a faulty spot in the emulsion, or any other fault, usually covers only a small section of the hologram, and such imperfections do not seriously interfere with the quality of the reconstructions.

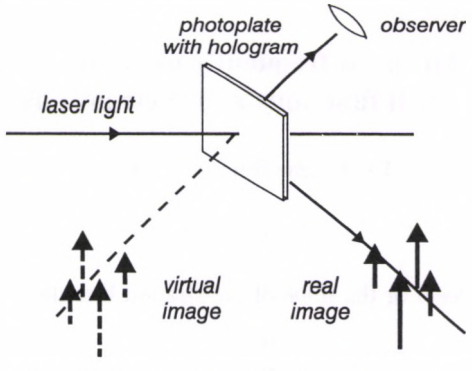


Fig. A/15. Image reconstruction from a hologram

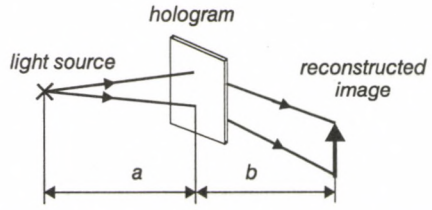


Fig. A/16. Magnification with the application of a divergent beam

– Several holograms can be made on the same photographic plate. For this purpose, only the angle of incidence of the reference beam must be altered each time a new hologram is made. In reconstruction the photographic plate is to be viewed from different angles; this can easily be achieved by the rotation of the plate.

In the reconstruction, *magnification* is obtained in two different ways. In one method the wavelength of the reconstructing light is longer than that used to produce the hologram. Thus, the hologram may be made with ultraviolet radiation for instance, and subsequently reconstructed with yellowish-red light. The magnification in this case is given by the ratio of the two wavelengths. Another method, similarly simple, puts the hologram in the path of a divergent beam. In the case outlined in Fig. A/16 the magnification is given by $(a + b)/a$. Combination of these two methods leads to magnifications of several hundredfold. The magnification can be further increased if the reconstructed image or its details are investigated microscopically. The resolution of the image is determined by the wavelength used to produce the hologram. The holographic method is particularly advantageous when moving objects difficult to follow with the normal traditional microscope are to be observed.

B. From the basics of differentiation and integration

In what follows we are dealing with mathematical knowledge which we referred to and relied on in the previous chapters of the book. Our objective is (without striving for completeness) to dissipate the reluctance to some mathematical concepts and connections and to refresh certain knowledge.

B1. The most frequently used one-variable real functions and their graphs

1.1. Linear function (Fig. B/1)

$$y = ax + b \; (a \neq 0)$$

Here a is called the *tangent* or the *slope* of the straight line since

$$\operatorname{tg} \alpha = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1} = a$$

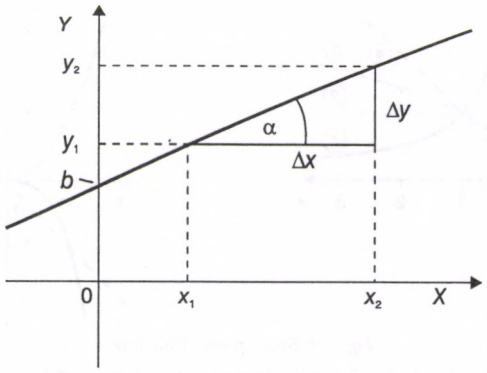


Fig. B/1. Linear function

and b is the *y-intersect* of the straight line. If $b = 0$, then we say that the variables x and y are directly proportional to each other; in this case a is called the *ratio of direct proportionality.*

1.2. Quadratic function (Fig. B/2)

$$y = ax^2 + bx + c$$

or

$$y = a(x - u)^2 + v \quad (a \neq 0)$$

Here u gives the minimum or maximum point of the function, and v the minimum or maximum values, respectively.

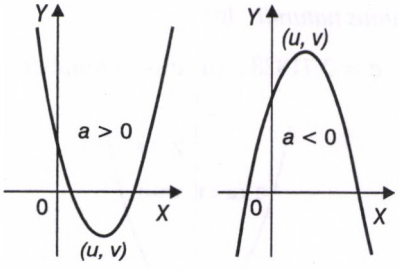


Fig. B/2. Quadratic function

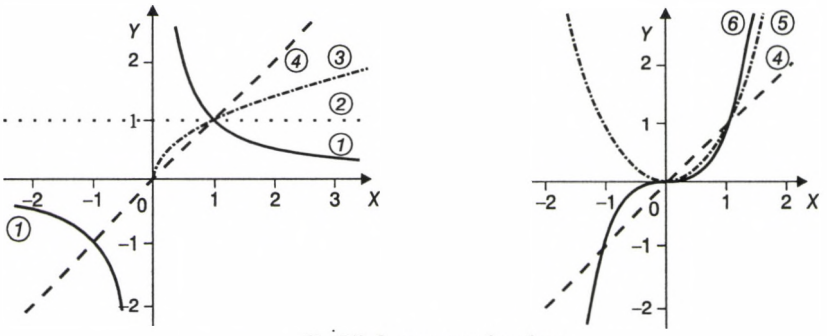


Fig. B/3. Some power functions

1: $\mu < 0$ $\left[y = \frac{1}{x} = x^{-1}\right]$; *2*: $\mu = 0$ $(y = x^0 = 1)$; *3*: $0 < \mu < 1$ $(y = \sqrt{x} = x^{1/2})$;
4: $\mu = 1$ $(y = x)$; *5*: $\mu = 2$ $(y = x^2)$; *6*: $\mu = 3$ $(y = x^3)$;

1.3. Power function (Fig. B/3)

$$y = x^{\mu}$$

The domain of the function consists of those real numbers whose μ-th power is defined. It is defined, for example, for all real numbers if $\mu = 3$, for all real numbers except $x = 0$ if $\mu = -1$, however, only for real numbers $x \geq 0$ if $\mu = 1/2$.

1.4. Exponential function (Fig. B/4)

$$y = a^x \quad (a > 0)$$

The graphs of the exponential functions $y = a^x$ and $y = (1/a)^x = a^{-x}$ are each other's reflections with respect to the y-axis. (If $a > 1$ then $0 < 1/a < 1$.)

An important special case is when the base is chosen to be the base number of the natural logarithm (*logarithmus naturalis;* ln),

$$e = 2{,}71828... \text{ (irrational number)}$$

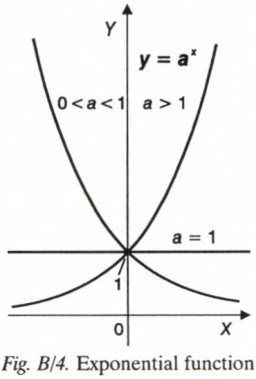


Fig. B/4. Exponential function

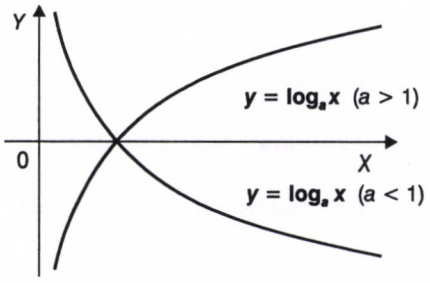


Fig. B/5. Logarithmic function

1.5. Logarithmic function (Fig. B/5)

$$y = log_a x \; (a > 0, a \neq 1)$$

The function is defined only for positive real numbers. By definition, $\log_a x$ denotes the power index for which, by raising the base a, we obtain x, i.e.

$$x = a^{\log_a x} = a^y$$

Therefore the difference between the logarithmic and the exponential functions is that x and y exchange their roles. In other words, the domain of the first function is the range of the second one, and vice versa. In such cases each function is called the *inverse* of the other one. Their graphs are each other's reflections with respect to the straight line drawn across the origin at 45 degrees to the x-axis (Fig. B/6).

In practice, mainly two kinds of logarithmic functions are used: to the base 10 (common logarithm) and to the base e (natural logarithm). Their usual notations are:

$$\log_{10} x = \lg x \quad \text{and} \quad \log_e x = \ln x$$

These logarithmic functions differ from each other only in coefficient.

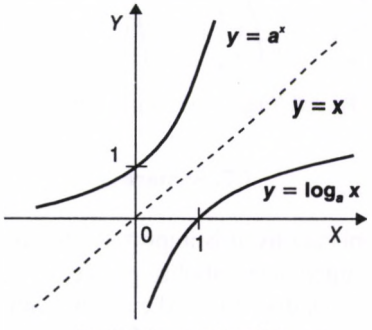


Fig. B/6. Inverse functions

1.6. Trigonometric functions

a) Sine and cosine functions (Fig. B/7)

$$y = \sin x \quad \text{and} \quad y = \cos x$$

These two functions are "similar", and by a simple translation they cover each other:

$$\cos x = \sin\left(x + \frac{\pi}{2}\right)$$

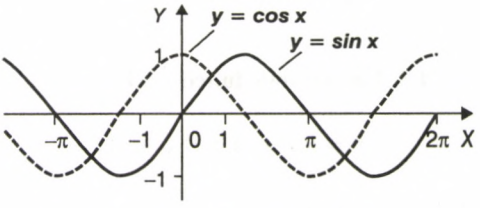


Fig. B/7. Sine and cosine functions

b) Tangent and cotangent functions (Fig. B/8)

$$y = \operatorname{tg} x \quad \text{and} \quad y = \operatorname{ctg} x$$

There is a reciprocal correlation between the two functions, thus, we can observe well the reflected symmetry in the figure.

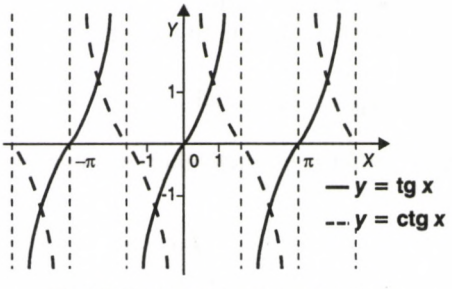


Fig. B/8. Tangent and cotangent functions

1.7. Remarks

a) Considering practical applications, it is important that if the independent variable of the exponential or the sine functions symbolize physical quantities, they necessarily may occur only in such combinations that finally they result dimensionless numbers. In such cases the usual forms of these functions are the following:

$$y = b\,a^{cx} \quad (a > 0), \quad y = \sin kt$$

From the mathematical point of view, these functions are such composite functions, where the inner functiong $g(x) = cx$ or $g(t) = kt$, respectively, is a linear function.

b) The functions listed previously and those which can be formed from these (for example, by addition or multiplication) are called *elementary functions.* As a counter-example: the function $y = |x|$ (absolute value function) is not elementary (Fig. B/9):

$$y = |x| = \begin{cases} x, \text{ if } x \geq 0 \\ -x, \text{ if } x < 0 \end{cases}$$

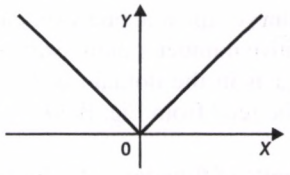


Fig. B/9. Absolute value function

B2. Limits

2.1. Limits of number sequences

A list of real numbers is called number sequence. In other words, a number sequence is such a real one-variable function whose domain is the set of the natural numbers. In this case the value a_n assigned to n is called the n-th value of the sequence ($n = 1, 2, 3, ...$).

For example, consider the number sequence $a_n = 1/n$ ($n = 1, 2, 3, ...$). One can see that proceeding in the sequence, the values are getting closer and closer to zero; e.g. after the tenth value each $a_n < 1/10$, after the hundredth value each $a_n < 1/100$.

We say that the limit value (the *limit* when n tends to infinity; $\lim_{n\to\infty}$) of the number sequence $a_1, a_2, ... a_n ...$ is the number A if for each positive number ε there exists a natural number n_0 such that $n > n_0$ implies $|a_n - A| < \varepsilon$. In case of the previous example, if ε is chosen to be 1/10 or 1/100, then $n_0 > 10$ or $n_0 > 100$, respectively, implies that the sequence is closer than ε to zero, the limit value.

2.2. Limits of functions at infinity

For example, consider the function $y = 1/x$ ($x > 0$) which in many aspects is similar to the example in section 2.1. Here one can see that if we increase x, then the values $f(x)$ are getting closer and closer to 0. If, e.g. $x > 10$, then the graph of the function remains in the strip between 0 and 1/10; however, if $x > 100$, then the width of this strip decreases to 1/100, etc.

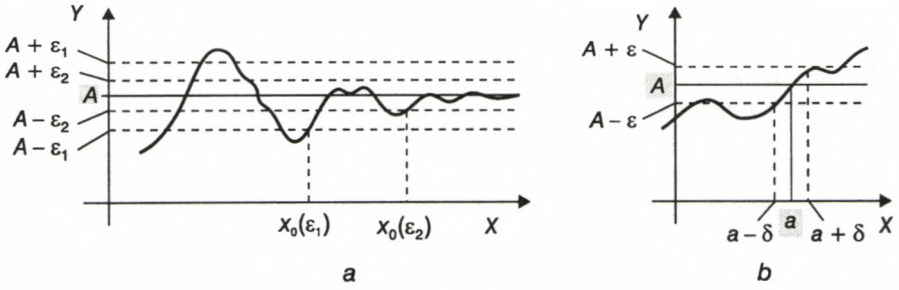


Fig. B/10. Function limit

We say that the limit value ($\lim_{n\to\infty}$) of the real one-variable function f at (positive) infinity is the number A if for each positive number ε there exists such a real number x_0 for which the following holds: if $x > x_0$ (x is in the domain of f) then $|f(x) - A| < \varepsilon$. The clear meaning of this definition can be read from Fig. B/10*a*.

2.3. Limits of functions at a finite point

Consider the following function:

$$y = \frac{2x^2 - 8}{x - 2}$$

Determine what value the function tends to when x tends to $x = 2$? This is an interesting question since $x = 2$ is not in the domain of the function because the denominator is 0 at this point.

At each point $x \neq 2$ we obtain the following:

$$\frac{2x^2 - 8}{x - 2} = 2\,\frac{(x+2)(x-2)}{x - 2} = 2\,(x+2)$$

However, this value tends to 8 when x tends to 2. Hence we can say that the limit value of the function at $x = 2$ is 8,

$$\lim_{x\to 2} \frac{2x^2 - 8}{x - 2} = 8$$

In general, we say that the *limit* of the real one-variable function f at a point a in the domain is the number A (Fig. B/10*b*) if for each positive number ε there is a positive number δ, for which $0 < |x - a| < \delta$ (x is in the domain of f) implies that $|f(x) - A| < \varepsilon$.

(Now if we consider the previous example, and e.g. ε is chosen to be 1/10, then the condition is satisfied for all $\delta < 5/100$. If $\varepsilon = 1/50$, then each value which is smaller than 1/100 corresponds to δ.)

2.4. Remark

Consider the following number sequence:

$$a_n = \left(1 + \frac{1}{n}\right)^n \quad (n = 1, 2, 3, ...)$$

One can see that this sequence is monotonously increasing, upper bounded, and thus it can be proved that its limit exists. Furthermore, one can see that this limit is the same number e as the *base* of the *natural logarithmic function* (ln). (The number e can be defined exactly as the limit of this sequence.)

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e$$

B3. Differentiation

3.1. Differential quotient (derivative)

A great deal of the functions describing real phenomena of Nature are so complicated that it is an almost unsolvable problem how to handle and determine them. In most cases, fortunately, this is not necessary because instead of the original functions, their appropriate approximations also satisfy the requirements.

Since the simplest functions are the linear functions (straight lines), the approximation with these is very important. One of the widespread methods for solving the problem is the so-called *linear interpolation,* when in an interval the function is approximated with the *secant line* belonging to the interval (Fig. B/11).

This approximation is exact at the endpoints of the interval; between them, however, it may sometimes cause errors of different extent. It seems that if we restrict the approximation to a neighbourhood of a single point, then the exactness can be improved.

Considering a function graph *in the neighbourhood of a given point,* among all straight lines passing through the point, the *tangent line* drawn across that point is the best approximation of the function (Fig. B/12). This "fits" most to the graph. Hence it is very

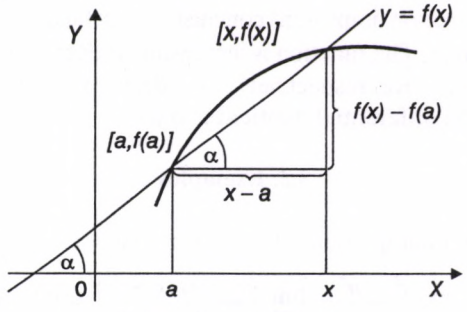


Fig. B/11. Linear interpolation

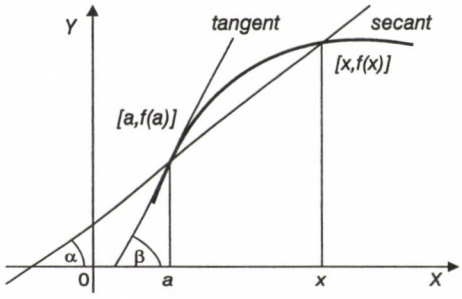


Fig. B/12. Differential quotient

important how to determine the tangent line. The tangent line is unambiguously determined by its *slope (tangent)* since one point of the tangent line is already known. This slope (tangent) is called the *differential quotient* or *derivative* at the given point.

Since the tangent line is the limit case of the secant line, the differential quotients can also be defined as the limit of the tangent of the secant lines. (This is understood in the sense that the tangents of the secant lines belonging to the given point a are real numbers depending on x, and we have to determine the limit value of the resulted function at the point a.)

The tangent of the secant line connecting the points $(a, f(a))$ and $(x, f(x))$ is

$$w_a(x) = \frac{f(x) - f(a)}{x - a} = \left[\equiv \frac{\Delta y}{\Delta x} \Big|_{x=a} = \operatorname{tg} \alpha \right]$$

hence the differential quotient at the point a is given by the following limit value:

$$\lim w_a(x) = \lim_{x \to a} \frac{f(x) - f(a)}{x - a} \equiv f'(a) \quad \left[\equiv \frac{dy}{dx} \Big|_{x=a} = \operatorname{tg} \beta \right]$$

(More of the usual notations are shown here.)

3.2. Remarks

a) The tangent of the secant line is usually called *difference quotient.*

b) Considering the function, important conclusions can be deduced from the behaviour of the differential quotient. The function is increasing or decreasing when the differential quotient is positive or negative, respectively. An extreme value (minimum or maximum) may occur only when the differential quotient is zero.

3.3. Examples

1. Determine the differential quotient of the function $y = x^2$ at the point $x = a$:

$$\lim_{x \to a} w_a(x) = \lim_{x \to a} \frac{x^2 - a^2}{x - a} = \lim_{x \to a} \frac{(x - a)(x + a)}{x - a} = \lim_{x \to a} (x + a) = 2a$$

2. To illustrate the concept of differentiability we also show a non-differentiable function.

Consider the function $y = |x|$ at the point $a = 0$.

$$w_0 = \frac{|x| - |0|}{x - 0} = \begin{cases} 1, & \text{if } x > 0 \\ -1, & \text{if } x < 0 \end{cases}$$

One can see that the limit of the difference quotient at the point 0 does not exist because if we tend to zero from the right or the left, the values $w_0(x)$ do not get closer to each other. Therefore, the function $y = |x|$ is non-differentiable at the point $a = 0$. Geometrically, this manifests itself by the fact that the considered point is a breaking point of the function graph, hence it has no sense to talk about the tangent line (cf. Fig. B/9).

In general we can say that a function is differentiable if its graph is "smooth", and non-differentiable if it is "broken" or "pointed".

3.4. Derivative function

Denote by D_f the set of those points in the domain of a real one-variable function f where the function is differentiable. Assuming that this set is not the empty set, we can define the derivative of f, denoted by f', such that to each point x in D_f we assign the differential quotient of f at the point x, $f'(x)$. Alternative notations are:

$$y', \frac{dy}{dx}, \dot{f}, Df, \frac{d}{dx}f$$

Hence, on the basis of example 3.3.1 one can easily see that the derivative function of the function $y = x^2$ is the function $y' = 2x$.

If the above conditions for f also hold for f', then similarly to the above method of the definition of the derivative function, we can define f'', the second derivative. Alternative notations are:

$$y'', \frac{d^2y}{dx^2}, \frac{d}{dx}\left(\frac{dy}{dx}\right), \ddot{f}$$

If the appropriate conditions are satisfied, this method can be continued in order to define the *higher-order* derivatives ($f^{(III)}$, $f^{(IV)}$, ...), too.

3.5. Remark

Here we mention the notion of *partial differential quotients.* This is related to *functions of more than one variable,* not treated in detail here.

For example, consider the real two-variable function ($f = f(x, y)$). If we fix e.g. the y-variable at a certain point $y = b$, then we get a real one-variable function, f_b, defined by the equation

$$f_b(x) = f(x, b)$$

The differential quotient of this latter function can be defined in a way as mentioned in section 3.1. (In other words, partial differentiation means that we take the derivative

according to a real variable while the other variables are all considered to be constant.) Notation:

$$f'_b(a) = \frac{df_b}{dx}\Big|_{x=a} = \frac{\partial f}{\partial x}\Big|_{(a,\,b)}$$

3.6. Rules for differentiation, derivatives of elementary functions

In the tables below, the derivatives of some elementary functions and the most important rules for differentiation can be found without proofs. By their help the most frequent differential problems can be solved relatively easily.

Table 1. Elementary functions and their derivatives

$y = c$	$y' = 0$
$y = x^n$	$y' = n\,x^{n-1}$
$y = a^x$	$y' = a^x \ln a$
$y = e^x$	$y' = e^x$
$y = \ln x$	$y' = 1/x$
$y = \sin x$	$y' = \cos x$
$y = \cos x$	$y' = -\sin x$

Rules for differentiation

$$(c\,f)' = c\,f'$$

$$(f_1 \pm f_2)' = f_1' \pm f_2'$$

$$(f_1 f_2)' = f_1' f_2 + f_2' f_1$$

$$\left(\frac{f_1}{f_2}\right)' = \frac{f_1' f_2 - f_2' f_1}{f_2^2}$$

Differential quotients of composite (intermediate) functions

Let f_2 be a function of x and f_1 a function of x^*, where $x^* = f_2(x)$. Then f_1 indirectly depends on x "via" f_2: $f_1(f_2(x))$. In such cases the following equation holds:

$$[f_1(f_2(x))]' = f_1'(f_2(x))f_2'(x)$$

or using alternative notations:

$$\frac{df_1}{dx} = \frac{df_1}{df_2}\,\frac{df_2}{dx}$$

This rule, the so-called *chain rule,* can of course be extended for compound composite functions, too.

3.7. Linear approximation

Let f, a real one-variable function, be differentiable at the point a. Then the increment of f corresponding to the point x can be written as follows (Fig. B/13):

$$\Delta f = f'(a)\,\Delta x + o(\Delta x)$$

where $o(\Delta x)$ denotes a function for which

$$\lim_{\Delta x \to 0} \frac{o(\Delta x)}{\Delta x} = 0$$

holds.

This means that in a small interval a differentiable function can be well approximated with a linear function.

For example, consider the linear approximation of the following functions according to this rule:

$$f(x) = f(a) + \Delta f \approx f(a) + f'(a)\Delta x = f(a) + f'(a)\,(x - a)$$

e^x in the neighbourhood of $a = 0$: $e^x \approx e^0 + e^0(x - 0) = x + 1$

$\ln x$ in the neighbourhood of $a = 1$: $\ln x \approx \ln 1 + \left[\frac{1}{1}\right](x - 1) = x - 1$

$\sin x$ in the neighbourhood of $a = 0$: $\sin x \approx \sin 0 + \cos 0\,(x - 0) = x$

Linear approximation is of great importance in natural sciences. In most cases the laws of Nature (e.g. elastic deformation, thermal expansion, etc.) are not substantially linear. Instead of the "real" but more sophisticated function actually describing the phenomena, we consider a linear approximation which is valid only in a certain domain and only with a certain exactness. This fact is related to such expressions as, for example, "thin lens", "small amplitude", "small expansion", "diluted solution", etc.

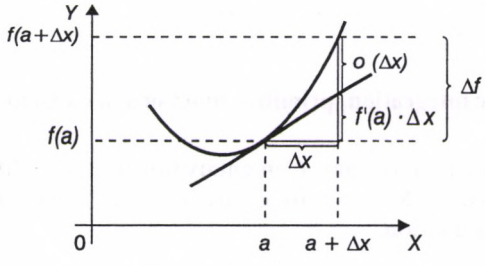


Fig. B/13. Linear approximation

B4. The integral

4.1. The indefinite integral (primitive function)

One of the most important problems of differential calculus is the following: Given a function, find the corresponding derivative function. In many cases, the following inverse problem is also important: Given a derivative function, find the corresponding "original" function, the so-called primitive function.

Denote by I the common part of the domains of the real functions f and F. F is supposed to be differentiable in the domain I. We say that F is an *indefinite integral* or *a primitive function* of the function f if at each point x in I

$$\frac{dF(x)}{dx} = F'(x) = f(x)$$

holds.

It follows from the definition that the function $F(x) + c$ (where c is an arbitrary real number, i.e. constant) is also a primitive function of the function f, as well as $F(x)$ itself, since the derivative of the constant function is zero (Fig. B/14).

To denote the general expression $F(x) + c$, we use the symbol $\int f(x)\,dx$, where the function $f(x)$ after the integral sign is called *integrand.*

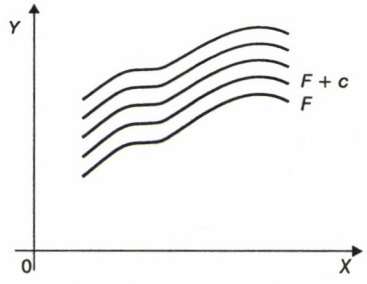


Fig. B/14. Primitive function

4.2. Rules for integration, primitive functions of elementary functions

Below the primitive functions of some elementary functions and the most important rules for integration are listed. Most of them are obvious from what has been told in connection with differentiation.

Table 2. Elementary functions and their primitive functions

$a\,x^n\ (n \neq -1)$	$\frac{a}{n+1}\,x^{n+1} + c$
$\frac{1}{x}$	$\ln \lvert x \rvert + c$
e^x	$e^x + c$
e^{cx}	$\frac{1}{c}e^{cx} + c$
a^x	$(\ln a)^{-1}\,a^x + c$
$\sin x$	$-\cos x + c$
$\cos x$	$\sin x + c$

Rules for integration

$$\int cf(x)\,dx = c\int f(x)\,dx$$

$$\int (f_1(x) \pm f_2(x))\,dx = \int f_1(x)\,dx \pm \int f_2(x)\,d(x)$$

$$\int f_1(x)\ f_2'(x)\,dx = f_1(x)\,f_2(x) - \int f_1'(x) f_2(x)\,dx$$

4.3. The definite integral (the formula of Newton and Leibniz)

First of all, instead of a definition we present a method for computing the definite integral applicable in many cases. If f is a real one-variable function and has a primitive function in the interval $[a, b]$ $(a \leq x \leq b)$, then the *definite integral* of the function f in the interval $[a, b]$ is given by the following formula:

$$\int_a^b f(x)\,dx = F(b) - F(a) = F(x)\Big|_a^b$$

Here F is an arbitrary primitive function of the function f in the interval $[a, b]$.

The definite integral is determined by the values of the primitive function of the integrand at points a and b, more precisely, by the difference of these two values.

The *geometric meaning* of the definite integral is shown, without proof, in Fig. B/15. Here the definite integral

$$T = \int_a^b f(x)\,d(x)$$

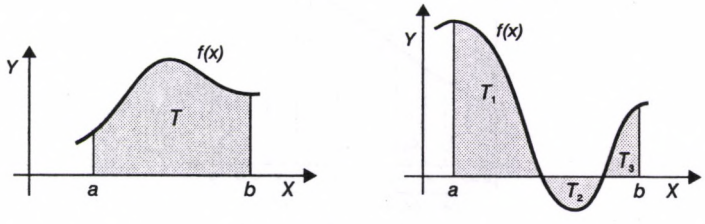


Fig. B/15. Definite integral

illustrates the measure of the signed area which is bounded by the x-axis, the graph of the function $f(x)$, and the straight lines $x = a$ and $x = b$. (In Fig. B/15, T, T_1, T_3 denote positive and T_2 negative areas.)

4.4. Remarks

a) The definite integral can be understood as a limit value taking into consideration the following.

Let the one-variable real function f be bounded in the interval $[a, b]$ (Fig. B/16). Divide the interval $[a, b]$ into n subintervals by means of points

$$a = x_0 < x_1 < x_2 < \ldots < x_{n-1} < x_n = b$$

Let furthermore

$$\Delta x_1, \Delta x_2, \ldots, \Delta x_n$$

denote the length of these subintervals. The fineness of the subdivision is understood as the maximal length of the subintervals; this is denoted by λ. Choose an arbitrary point ξ_i from each subinterval (x_{i-1}, x_i). Then the real number

$$K = \sum_{i=1}^{n} f(\xi_i)\, \Delta x_i$$

is called the Riemann sum for the function f with respect to the interval $[a, b]$. (Of course, any other choice of the points ξ_i and the subdivision usually results a different Riemann sum, K.) However, if the subdivision becomes even "finer", then the Riemann sums tend to the same real number, and this limit value is nothing else but the definite integral

$$\int_a^b f(x)\, d(x)$$

We say that the function f is (Riemann) integrable in the interval $[a, b]$, if there exists a real number, I, with the following properties: For any positive number ε there exists a positive number δ such that for an arbitrary subdivision of the interval $[a, b]$ with $\lambda < \delta$, the inequality

$$|K - I| = \left| \sum_{i=1}^{n} f(\xi_i)\, \Delta x_i - I \right| < \varepsilon$$

holds.

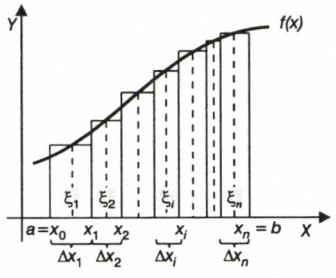


Fig. B/16. Riemann sum (Riemann integral)

This last statement can be considered as a definition of the definite integral. By this definition, the definite integral can be determined even for functions, which have no primitive function.

b) Further important rules for the definite integral are:

$$(1)\ \int_a^a f(x)\,d(x) = 0$$

$$(2)\ \int_a^b f(x)\,d(x) = -\int_b^a f(x)\,d(x)$$

$$(3)\ \int_a^c f(x)\,d(x) = \int_a^b f(x)\,d(x) + \int_b^c f(x)\,d(x) \quad (a < b < c)$$

$$(4)\ \left|\int_a^b f(x)\,d(x)\right| \leq \int_a^b |f(x)|\,d(x)$$

4.5. Examples

Calculate the following definite integrals (Fig. B/17):

(a) $\int_1^3 |x-2|\,dx = \int_1^2 (-x+2)\,dx + \int_2^3 (x-2)\,dx = \left[-\frac{x^2}{2}+2x\right]\Big|_1^2 + \left[\frac{x^2}{2}-2x\right]\Big|_2^3 =$

$$= \left[(-2+4) - \left[-\frac{1}{2}+2\right]\right] + \left[\left[\frac{9}{2}-6\right] - (2-4)\right] = \frac{1}{2}+\frac{1}{2} = 1$$

(b) $\int_0^\pi \cos x\,dx = \sin x\,\Big|_0^\pi = \sin\pi - \sin 0 = 0$

(c) $\int_a^b \frac{1}{x}dx = \ln x\,\Big|_a^b = \ln b - \ln a = \ln\frac{b}{a}$

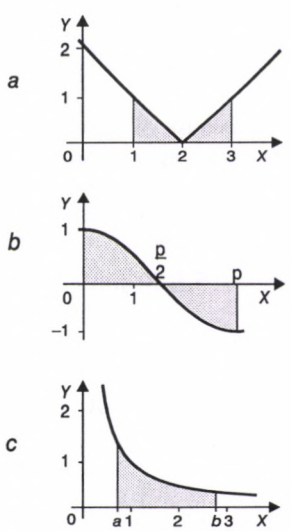


Fig. B/17. To the examples in section 4.5

Some definite integrals which are similar to the last example can be found many times in Chapter 5, for example, where we treated the isothermic work by gases, and in connection with the calculation of entropy (cf. section 5.3.4.), or in the Appendix,, when the barometric altitude formula is mentioned (cf. section A1).

B5. Differential equations

5.1. Definition

The differential equations are equations which contain a differentiable function and its derivative function (also higher-order derivative functions or partial derivative functions).

5.2. Solving a differential equation

It means that we have to determine all functions, which, along with their derivative functions, turn the equation into identity if we substitute them into the differential equation.

The simplest differential equations can be solved via integration. For example, if the differential equation can be written in the form

$$\frac{dy}{dx} = f_1(x) f_2(y) \qquad \text{[B1]}$$

where $f_1(x)$ and $f_2(x)$ are real functions in the intervals I_1 and I_2, respectively, and $f_2(y)$ never vanishes inside the corresponding interval, then the solution of the differential equation is given by the functions originating from the following equality:

$$\int f_1(x)\, d(x) = \int \frac{dy}{f_2(y)} \qquad \text{[B2]}$$

The functions obtained this way (because of integration) are not unambiguously determined (general solution). Hence, there appears an extra requirement for the unknown functions: They must satisfy certain complementary, so-called *initial* or *boundary conditions* (cf. example in section 5.4 below).

5.3. Remark

The differential equations play an important role in mathematics, physics and the technical sciences. Their application is also related to other sciences, especially to biology, but also to economics. Some important laws of the above-mentioned sciences can be expressed in the form of differential equations, the solution of which provides the description of the studied processes.

5.4. Example

Solve the differential equation $y' = ky$ where k is a constant. According to [B1], where $f_1(x) = k$, and $f_2(y) = y$, by using [B2] we obtain

$$\int k\, d(x) = \int \frac{dy}{y}$$

From this we get the following equation by using Table 2:

$$kx + c_1 = \ln y + c_2$$

(c_1 and c_2 are arbitrary real numbers) and

$$\ln y = kx + c \qquad (c = c_1 - c_2)$$

From this latter equality the following solution is obtained:

$$y = e^{kx + c} = Ce^{kx} \; (C = e^c) \qquad \text{(general solution)}$$

With the additional prescription that the value $y = y_0$ must correspond to the point $x = 0$ (initial condition), the solution becomes unambiguously determined:

$$y_0 = Ce^0 = C, \quad \text{therefore} \quad y = y_0\, e^{kx}$$

Some cases of this type are met with when treating the barometric altitude formula (cf. Appendix A1), the law of radiation attenuation (cf. section 2.3.1) and the law of radioactive decay (cf. section 3.1).

SUBJECT INDEX

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

AN INTRODUCTION TO BIOPHYSICS

WITH MEDICAL ORIENTATION

EDITED BY

G. RONTÓ AND I. TARJÁN

This book deals with various aspects of modern biophysics. In fact, it offers a great deal more than the mere fundamentals the title may suggest; beyond the stock of basic knowledge which usually forms the body of textbooks and handbooks for this branch of science, the readers made acquainted with the biological and medical applications as well.

The main topics discussed include the relationship between structure and function; radiations, physical methods in structure analysis; transport processes and thermodynamic principles; modelling in biology; medical electronics and biocybernetics. In the appendix some basic concepts in mathematics and physics are discussed which help in understanding.

The purpose of the book is to give an insight into the problems and perspectives of the rapidly developing, complex branch of biophysics. It is addressed primarily to physicians and biologists but may be of interest to physicists as well if they seek information on the current problems of biophysics and the related fields of application. It may be used both as a textbook and a handbook.

AKADÉMIAI KIADÓ, BUDAPEST

ISBN 963 05 8088 8