KFKI-1990-50/C

B. LUKÁCS
SZ. BÉRCZI
I. MOLNÁR
G. PAÁL

EVOLUTION: FROM COSMOGENESIS TO BIOGENESIS

Hungarian Academy of Sciences
CENTRAL
RESEARCH
INSTITUTE FOR
PHYSICS

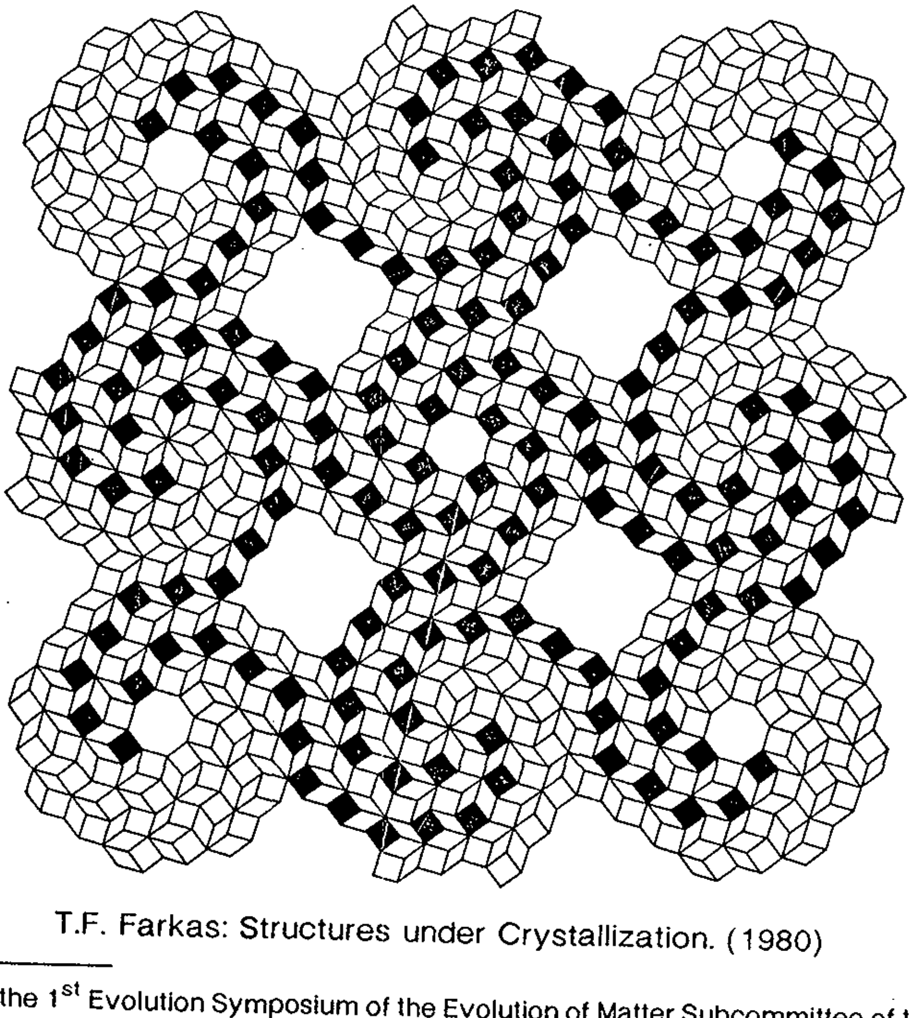
BUDAPEST

KFKI-1990-50/C
PREPRINT

EVOLUTION:
FROM COSMOGENESIS TO BIOGENESIS

B. LUKÁCS, SZ. BÉRCZI, I. MOLNÁR, G. PAÁL (eds.)

Central Research Institute for Physics
H-1525 Budapest 114, P.O.B. 49, Hungary



T.F. Farkas: Structures under Crystallization. (1980)

The material of the 1st Evolution Symposium of the Evolution of Matter Subcommittee of the Geonomy Scientific Committee of the Hungarian Academy of Science, 28-31 May, 1990.

HU ISSN 0368 5330

B. Lukács, Sz. Bérczi, I. Molnár, G. Paál (eds.): Evolution: from cosmogenesis to biogenesis.
KFKI-1990-50/C

ABSTRACT

The Volume contains the material of an interdisciplinary evolution symposium. The purpose was to shed some light on possible connections between steps of evolution of matter on different levels of organisation. The topics involved are as follow: cosmogenesis; galactic and stellar evolution; formation and evolution of the solar system; global atmospheric and tectonic changes of Earth; viral evolution; phylogeny and evolution of terrestrial life; evolution of neural system; hominization. The material also includes some discussions of the underlying phenomena and laws of Nature.

Б. Лукач, С, Берци, И. Молнар, Г. Паал (ред.): Эволюция от космогенезиса до биогенезиса. KFKI-1990-50/C

АННОТАЦИЯ

Сборник содержит материалы интердисциплинарного симпозиума по эволюции. Целью симпозиума являлось выяснение связи между шагами эволюции материи на разных уровнях организации. Были затронуты следующие темы: космогенезис, эволюция звезд и Солнечной системы, глобальная тектоника Земли, эволюция атмосферы Земли, эволюция вирусов, филогенезис, эволюция жизни, эволюция нервной системы, гоминизация, термодинамические аспекты эволюции. Материал содержит также некоторые соображения относительно законов, скрытых за процессами природы.

Lukács B., Bérczi Sz., Molnár I., Paál Gy. (szerk.): Evolúció: kozmogenézistől biogenézisig.
KFKI-1990-50/C

KIVONAT

A kötet egy interdiszciplináris evolúciós szimpózium anyaga. A szimpózium célja az volt, hogy megvilágítson esetleges kapcsolatokat a szerveződés különböző szintjein folyó evolúciós lépések közt. A témák: kozmogenézis; galaktikus és csillagevolúció; a Naprendszer kialakulása és története; globális változások a földkéregben és a légkörben; a vírusok evolúciója; filogenézis és a földi élet evolúciója; az idegrendszer evolúciója; emberréválás. Az anyag szintén tartalmaz bizonyos meggondolásokat a folyamatok mögötti természeti törvényekről.

INTRODUCTION

This Volume is the written material of the symposium "Evolution:Cosmogenesis...Biogenesis...", held between 28th and 31st May, 1990, in Budapest. This symposium had been organised under the aigis of the Geonomical Scientific Commitee of the Hungarian Academy of Sciences, by the Evolution of Matter Subcommittee of the said committee. The Geonomical Committee is an interdisciplinary body for general overview of the natural sciences to develop new long-range connections among different disciplines; the Subcommittee started in 1989, tries to follow (or trace back) the steps of self-organisation of matter from the Beginning to the present, and the result of its first symposium is this Volume.

However, no constructive method is known to view "Natural Science" as a unit; the global picture is built up from blocks constructed by the separate disciplines. Still, an overview may give answers to old questions. For example, on all levels of organisation the actual path of evolution may be and is regarded alternatively either as accidental or as predetermined. Such questions have connotations not only professional but philosophical, religious, eschatological &c. as well, but we do not want to leave the grounds of strict natural science. Our present purpose is to show up the available most coherent picture as a starting point of any deduction.

As an example, for terrestrial evolution of matter we mention that life uses mainly the most abundant elements of the Universe. These abundances may or may not be accidental, but the answer can come only from other disciplines. Similarly, the conditions resulting in Earth had been produced by cosmologic, galactic and (presolar) stellar evolutionary steps, which, therefore, should be simultaneously studied.

Generally the lecturers of the symposium are experts of their own fields. The Organising Committee (the Editors) took the responsibility of selecting the actual persons, but the texts are the lecturers' sovereign works. The plan of the symposium tried to establish some links or overlaps. Some separated extra comments of the Editors may appear after texts (in the gothic italics of the present Introduction), with a signature E. Of course, for these comments the lecturers are not responsible.

We hope that the Volume will promote some further multidisciplinary activities. The Editors thank the Theory Department of the Particle and Nuclear Institute of the Central Research Institute for physics for technical help. The edition was partly supported by the OTKA fund N° 1772.

Editors (and Organising Committee)
B. Lukács (president), Sz. Bérczi, I. Molnár, G. Paál
Evolution of Matter Subcommittee of the Geonomical Scientific
Committee of the Hungarian Academy of Sciences

CONTENT

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

OPENING ADDRESS

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

I have learnt with pleasure that the Working Commission for Matter Development of the Geonomical Commission of the Hungarian Academy of Sciences will organize a symposium with the title "Evolution: Cosmogenesis .....Biogenesis". I remember that nearly a quarter of a century ago we conducted talks with the geologist Szádeczki-Kardoss-Elemér and the meteorologist Béll Béla about the problem how the Earth Sciences distributed in many branches could be grouped around a unified resarch conception not mentioned the fact that they have an objects from different domains of our Earth to be studied. During these discussions it has been conceived that the flow of matter and of energy could be the central nucleus around which various branches of science (geography, geology, geophysics, meteorology, geodesy etc.) should be grouped, or even all the branches of natural sciences would be interested. So it has come to the organization of the first conference on matter-and-energy-flow in 1967 and three more have followed.

The idea has developed more and more in an interdisciplinary way, since every branch of science is trying to change the static way of looking into a dynamic one, which includes also the concept of evolution. Taking into account this idea the Geonomic Commission has been organized in the frame of the tenth section of the Hungarian Academy of Sciences under the leadership of Szádeczki-Kardoss Elemér, the Commission having an interdisciplinary character from the beginning on. In the frame of this commission we wanted to develop the concept of changing-developing into a link connecting the sciences with one another and so to counterbalance the disadvantages involved in the frittering away as a consequence of excessive specialization. This objective has been followed by the Commission for more than a decade and although it happened sometimes that we came into an impass in the course of our investigations, the present symposium is just showing that the combining force and activity of the science does exist inflexibly and the young organizing staff possesses the ambition to further develop this idea.

Of course we do not claim a full ownership of this idea, since the world-wilde development of sciences shows that the boundaries of the various branches of sciences are more and more fading away and just even the frontier zones develop into the most interesting domains of successful research.

May I express my best wishes and hope for a successful reach of the goal set or at least for an approach to it.

Barta György
President of the Geonomical Commission
Fellow of the Hungarian Academy of
Science

PHYSICAL BACKGROUND TO EVOLUTIONARY STUDIES

KFKI-1990-50/C B. Lukàcs & al. (eds.): Evolution: from Cosmogenesis to Biogenesis

B. Lukàcs+ & G. Paál*
+Central Resarch Institute for Physics, H-1525 Bp. 114. Pf. 49,
Budapest, Hungary
*Konkoly Observatory, H-1525 Bp. 114. Pf. 67, Budapest, Hungary

Received on 25th July, 1990

ABSTRACT

In the lack of the fundamental unified theory of all fundamental physical phenomena, here we outline only some very general physical frames for interdisciplinary evolutionary studies from the fundamental constants and the elementary particle data.

We would like to understand and describe the evolving Universe. Taking literally, this goal is impossible to reach, since the Universe is more complicated than the mind trying to understand it. So we will try to understand its global features.

1. FUNDAMENTAL SCALES

As far as we know the Universe [visible part: 10^{55} g mass and 10^{28} cm size] contains the following scales, going downward in size (in parentheses the characteristic data)

galaxy clusters [~10^{4} galaxies, 10^{25} cm] -> galaxies [10^{11} stars, 10^{23} cm] -> stars [10^{57} particles, 10^{11} cm] (and possibly their planets) -> living organisms [between 10^{9} and 10^{31} particles, 10^{-5} cm (mycoplasma) and 10^{3} cm (whale), judged from terrestrial experience] -> *molecules and atoms* [~10^{-24} g, 10^{-8} cm] -> elementary particles [masses: 10^{-24} g (baryons &c.), 10^{-27} g (electron), 10^{-31} g ?? (neutrinos ?), <<10^{-31} g (photon, gluon, neutrinos?); sizes: 10^{-13} cm (baryons), <<10^{-13} cm (leptons)].

Of course, all the larger structures *consist of the last ones*, but this fact is trivial. Observations indicate that baryons are composite, consisting of "point-like" quarks. The "elementary particles" *are either neutral, or charged by the* "elementary charge" e:

$$e^2 = 2.3 * 10^{-19}\ gcm^3/s^2$$

(or by its double; the quarks seem to have one or two-thirds of it).

- *Why those particular objects?*
- Are there any laws behind the above data?

2. *FUNDAMENTAL PHENOMENA AND LAWS*

According to our present knowledge there are 3 kinds of general phenomena, to which *all objects, processes etc. are subjected*. Each possesses its own characteristic constant, and influences everything. The traditional names come from ages when *they were believed independent of each other. Here comes a very* short recapitulation of history. For further study cf. e.g. Eddington (1987), Sen (1968).

Gravitation; constant: $G=6.67*10^{-8}\ cm^3/gs^2$.

Weight is known from prehuman ages. The connection between *weight and mass was already known to Aristotle of Stageira*. Complete quantitative laws from Newton (1687). The constant was first measured by Cavendish (1798).

Relativity; constant: $c=3.00*10^{10}$ cm/s.

Its constant, as velocity of light, first successfully measured by Römer (1675). First signal for the existence of the fundamental phenomenon was the negative result of Michelson's experiment (1881) to measure Earth's velocity in the ether. The complete theory is Special Relativity (Einstein, 1905).

Quantization; constant: $\hbar=1.05*10^{-27}\ gcm^2/s$.

First (unnoticed) appearance of the constant in Wien's blackbody radiation distribution (1896). First consciously introduced by Planck (1900). Recognition of general relevance by Einstein (photoelectric effect, 1905, and specific heat of crystalls, 1907).Complete theory is Heisenberg's matrix mechanics or Schrödinger's wave mechanics (later called by the common name quantum mechanics) (1926).

The three original "complete theories" can (and, as will immediately be seen, indeed do) mutually contradict each other, because all three claim general validity but each theory contains only one constant. So the next step required was partial unification.

Gravity vs. Relativity.

Gravity influences *photons; if they accelerate downwards*, then the light velocity may exceed the velocity of light c, forbidden in relativity. Solution: **General Relativity** (Einstein, 1916). In it gravity is no more a force, but a consequence of the nontrivial space-time geometry.

Quantum mechanics vs. Relativity.

Quantum mechanics gives velocity uncertainty for a mass point closed into a box: $m\Delta v\Delta x \approx h/2$. For an electron localised *better than* 10^{-11} *cm, the uncertainty would require* $v>c$, forbidden by relativity. Apparent solution: relativistic quantum mechanics (Dirac, 1928), but that *theory is not quite well defined*. Namely, in quantum mechanics a system has its wave function depending on 3N variables, where N is the number of particles *present. In relativistic situations pair creation* <-> annihilation is possible, so N is not fixed. Solution: **Quantum Field Theory** (Dirac & al., 1930), where the wave function is replaced by operators ("second quantization").

Quantum mechanics vs. Gravity

Quantum mechanics prevents the sharp measurement of the Newton potential of gravity; the limitation is a fundamental uncertainty, so the potential cannot exist in exact sense. Solution would be (*nonrelativistic*) **Quantum Gravity**, but in 1926, with quantum mechanics established, the individual Gravity was no more a valid theory, so this partial unification was ignored until recently. (See Diósi, 1990 and literature therein.)

Any of the three partial unifications would contradict to *the third individual theory or to another partial unification*, because they ignore different effects. We give here one example. General Relativity contradicts Quantum Field Theory in the following way. The right hand side of the Einstein equation is the energy-momentum tensor T_{ik}, which is an *operator* in Quantum Field Theory. However, the left hand side of the equation

$$R_{ik} - \tfrac{1}{2}g_{ik}R = \kappa T_{ik} \tag{2.1}$$

contains pure curvature terms which are classical numbers in general relativity. So the equality cannot hold. Possible solution: e.g. quantization of geometry. The episthemologic situation is visualized on Fig. 1: individual theories are the visible peaks of the tetrahedron, partial unifications are the edges of the front triangle face; present day possibilities lie on this face, while the "physical reality" is at the hidden fourth peak.

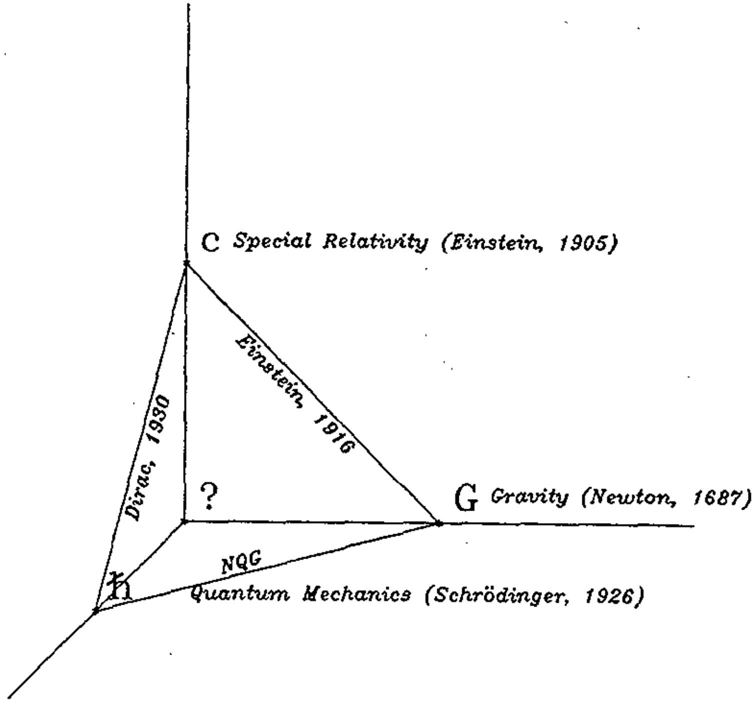


Fig. 1: The scheme of fundamental phenomena and their theories. NQG=(Newtonian) Quantum Gravity

The correct theory could only be the complete unification: a theory differing from all previous ones for structure, containing all the three fundamental constants, and yielding the 3 partial unifications in proper limits. Such a theory is not at reach, but there are promising attempts as supergravity and superstring theory, containing all the 3 constants.

3. UNDERSTANDING "ELEMENTARY" PARTICLES?

If our elementary particles are really fundamental, then the true fundamental theory must describe them, predicting masses, charges, &c. While the unified fundamental theory is not at reach, the 3 constants uniquely give its characteristic scales (up to numerical factors). For example, consider the elementary electric charge e:

$$\dim[e^2] = gcm^3/s^2 = \dim[\hbar c] \qquad (3.1)$$

Therefore the (still unknown) theory must give

$$e^2 = (\text{Number constant})*\hbar c. \qquad (3.2)$$

The observed value of the number constant is 1/137.04, not impossible to be obtained from a theory. Now, the situation is not so promising for *masses* and *sizes* of elementary particles, because of the following reason. Purely from G, ħ and c one can derive only *one* mass, length and time, since $[G]=cm^3/gs^2$, $[\hbar]=gcm^2/s$ and $[c]=cm/s$, all different. Therefore the proper physical dimensions, g, cm and s, can be obtained only by the combinations discovered by Planck:

Planck mass $M_{Pl} = \sqrt{\hbar c/G} \sim 10^{-5}$ g
Planck length $L_{Pl} = \sqrt{\hbar G/c^3} \sim 10^{-33}$ cm
Planck time $t_{Pl} = \sqrt{\hbar G/c^5} \sim 10^{-44}$ s

which are the characteristic scales of the (unknown) theory. These Planck data represent the only physical units defined uniquely by the laws of Nature themselves without any arbitrary convention. Visualizing the unknown "complete unified theory" via "quantum fluctuations of the space-time geometry", the probable scales for correlation length and time are L_{Pl} and t_{Pl}, while the probable energy of one correlated volume is $M_{Pl}c^2$.

Now observe that the masses, sizes and life times of our elementary particles are completely foreign to these scales. The ratio of Planck and proton masses *e.g.* is

$$M_{Pl}/M_p = (\hbar c/GM_p^2)^{1/2} = 1.3*10^{19}.$$

(A basic large dimensionless number of Nature?) There is a possible alternative:

1) The elementary particles of our physics are not truely elementary but complicated (accidental?) objects.
2) There are further fundamental phenomena with new scales.

The second possibility seems promising but is very problematic. To demonstrate this, let us introduce a fourth fundamental phenomenon (of still unknown nature) with its fundamental constant $\lambda=10^{-13}$ cm. Then the hadronic sizes are "explained" and via $m \sim \hbar/\lambda c \sim 10^{-24}$ g masses as well. However, the electron size is $<<10^{-13}$ cm, according to high energy scattering experiments (a kind of electron microscopy). Then again we have an alternative: i) $\lambda >> R_e >> L_{Pl}$; ii) $R_e \sim L_{Pl}$. In the first case a *fifth* scale (and fundamental phenomenon) would be needed, but limitless procreation of fundamental phenomena are not favoured. In the second case the electron would pick up the effects of quantum fluctuations of one cell (not averaged in its volume), i.e. would get M_{Pl} (20 orders of magnitude higher than the observed electron mass), not seen.

So: present elementary particles clearly exist but are uninterpretable in the light of present fundamental phenomena.

4. UNDERSTANDING STRUCTURES

As soon as we simply accept the existence and a few properties of some elementary particles as additional empirical facts, many of the composite objects (structures) become at once interpretable, mostly by simple stability considerations. In so doing we may attempt to theoretically reproduce the design of Nature from elementary particles to the entire observable Universe in a simple and aesthetic way. The result is briefly explained in the present comment and summarized in the enclosed Figure where mass M vs. size R are displayed in logarithmic scales, with an origo at the proton (Lukács & Paál, 1981; Paál & Lukács, 1988).

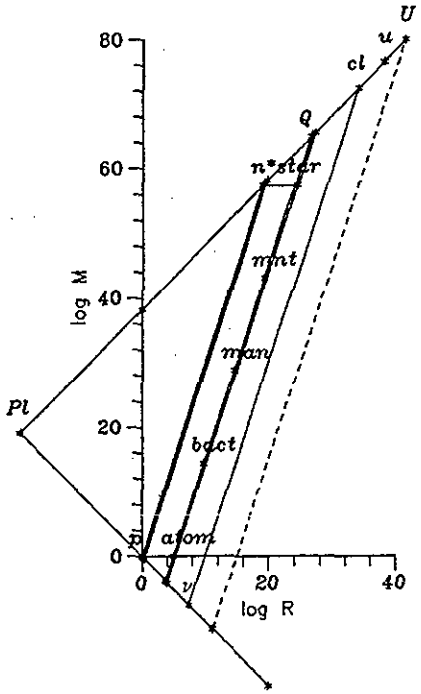


Fig. 2: Cosmic mass-size diagram. Legends: p: proton, ν: neutrino, n: neutron star, Q: quasar, cl: protocluster, u: Universe at decoupling, U: Universe at present, Pl: Planck data, mnt: mountain, bact: bacterium. Heavy, double, single, dashed lines: nuclear, atomic, leptonic, cosmologic densities, resp.*

Even if the evolution of objects may be very complicated, depending on wide variety of unknown initial conditions, the stable *final* stages of evolution (or slowly evolving quasistable ones) could in principle be prescribed by elementary physical considerations. Anyways, only at least partially stable objects, with some degree of persistence can be recognised as objects at all. *This stability implies (approximate) time independence.* Therefore it is easy to check if the basic physical phenomena and their constants tell us something about stable existence.

Out of the 3 basic physical measures (mass, size and time) the last one is then to be omitted. On the (M, R) plane of *Fig. 2* this means finding functional relations of the type

$$F(M,R)=f(G,h,c).$$

Taking into account the above physical dimensions (Duncan, 1955) of the constants one can find only 3 algebraically independent combinations (without permitting time dependence), namely omitting always one constant,

$$[G/c^2] = [M/R] = g/cm$$
$$[h/c] = [MR] = gcm$$
$$[h^2/G] = [M^3R] = g^3cm.$$

It is easy to see that the first two of them have obvious immediate connection to known stability considerations, while the third will be dealt with in Diósi's paper in the present Volume. Indeed, taking into account that *c* is a universal limitation for propagation of signals, it is clear that *c* cannot be overrun by any velocity coming from other disciplines. Combining therefore the predictions of Newton's gravity theory with this velocity limitation one could naïvely argue that even the escape velocities cannot exceed *c*, i.e.

$$GM/R=v^2<c^2 \rightarrow M\leq Rc^2/G \rightarrow \lg M\leq \lg R+\lg(c^2/G)$$

and similarly the quantum velocity uncertainty (coming from Heisenberg's uncertainty principle) is limited by the same value, i.e.

$$\delta x\delta p\geq h \rightarrow RMc\geq h \rightarrow \lg M\geq -\lg R+\lg(h/c).$$

Therefore one obtains 2 straight lines on the (lgM, lgR) plane of *Fig. 2, limiting the places of possible stable configurations on* the diagram from the left. On the ascending line each mass defines a length (the Schwarzschild radius) below which irresistible gravitational collapse leads to black hole formation; on the descending line one finds to each mass a length (its Compton length), *below which our naïve notions of space and time become* obscure and even meaningless because of quantum uncertainty. Remarkably these results can be verified from the partial unifications General Relativity and Quantum Field Theory.

Note that we have found only *limits* of possible stable configurations, whatever the cause of their stability. This means that the 3 basic constants do *not* prescribe the properties of stable objects without some independent additional information about their constituents, the elementary particles. (We have seen that these constants cannot give e.g. the particle *masses*.)

Anyhow, the objects of the real world are "well aware of the law", and apparently respect the above limitations. On Fig. 2 the points representing astronomical objects lie close to the upper boundary, while microphysical ones are near the lower one, in the permitted region. Objects built up only from nucleons, atoms and (presumably) neutral leptons (e.g. neutrinos) are also roughly of nuclear, atomic (and "leptonic") density and therefore aligned along equidensity lines (of slope +3, since $M\sim R^3$) attached to the respective objects. So matter can form stable equilibrium configurations only with masses and sizes corresponding to these lines. The natural laws and the building blocks determine the basic features of the structures.

The above "construction of the world" is surprisingly regular: *the equidensity lines happen to be just equidistant. These* lines reveal the intimate connection between the micro- and macrocosmos. The neutron star (n^*) is the "sign of the neutron in the sky"*; the ordinary star and the quasar are those of the

---

* According to astrophysical considerations indeed $M_*/M_p = (\hbar c/GM_p^2)^{3/2} = (M_{Pl}/M_p)^3$ just as in our graphical construction.

atom; while the protocluster (cl), galaxy cluster now, is possibly that of a neutral lepton (a kind of neutrino or other weakly interacting particle)+. So the astronomical macrocosmos is just the microcosmos "projected to the sky".

The central equidensity line is the most populated one according to our knowledge. Here one finds the particles of cosmic dust, meteorites, biological and geological formations, moons, planets. In case of *stable* equilibrium the two extremes on this line are the atoms (purely electrically bound objects) and stars (purely gravitationally bound objects), lying on horizontal lines across points p and n*, respectively, 5 orders of magnitudes away from the latter ones. Therefore, loosely speaking stars represent "gravitational atoms", while atoms mean "electric stars". Between these two extremities man represents *aurea mediocritas* - a "gravitationally limited electric being" - who is therefore larger than the H atom by just the same factor as smaller than the star. (Otherwise he would be broken in pieces if fallen down to the ground.) Calculating this geometric mean between star and atom directly from basic natural constants, one gets about 78 kg (!) just like our typical human mass indeed. Considering, therefore, orders of magnitude, one finds that the "measure of Nature" is "anthropocentric" both in mass and in size, but this fact has nothing to do with any kind of subjective wishful thinking.

Furthermore note on the Figure that the series *atom* (hydrogen), *bacterium* (primitive living), *man* (most evolved living), *mountain* (highest still stable) and *star* is also equidistant in a good approximation.

5. UNDERSTANDING THE UNIVERSE

The above all-embracing order controls not only each individual, but also the "totality", i.e. the Universe as well. This can be made obvious by Fig. 2 which shows that not only the above series of "ordinary" objects but also the series of extreme astrophysical objects (neutron star, minimal quasar, protocluster, observable Universe) is just equidistant. This implies that the series of equidistant equidensity lines can be extended to include the density line of the *Universe* (of $\sim 10^{-29}$ g/cm^3). The intersection point of this line with the black hole line of slope +1 correctly gives the mass and size of the entire observable Universe ($\sim 10^{80}$ proton mass and $\sim 10^{42}$ proton radius). This coincidence demonstrates that the Universe is also an extreme gravitating system. (So gravity effects do not cancel out at large as suggested by the Newtonian picture.) It is worth mentioning that the point "Universe" equally well characterizes both the part of the totality observationally known at present and the part which can in principle become known via ideal observations, because signals from essentially more distant regions have not yet reached us during the entire past of the Universe beginning from its "Big Bang birth" till now.

A further beautiful expession of the all-embracing regularity is that the geometric mean between the size of the *Universe* and that of the *atom* is just about the size of the *star*. The geometric mean between the size (or mass) of the *star* and that of the *atom* is the *man* while the geometric mean between the *man* and the *atom* is the *bacterium*.

In algebraic language one may write

$$R_U/R_{star}=R_{star}/R_{atom}=(R_{man}/R_{atom})^2=(R_{bact}/R_{atom})^4$$

and similar formulae for M_{star}, M_{man}, M_{atom} as well. Furthermore, probably $M_{cl}/M_{star}=(M_p/M_\nu)^2$

As a consequence of the well known expansion of the Universe the position of the point U is time dependent, and, according to the suggestion of Fig. 2, it should move just along the limiting line of slope +1 (otherwise its present precise fitting to this line would be highly improbable). This is indeed true according to the details of modern cosmology, so that the whole past history of the Universe can also be read off the diagram. Its evolution clearly ought to have started from the intersection point of the limiting lines (Lukács & Paál, 1988), corresponding to the Planck length and mass. The cosmologic evolution itself will be trated in some more details in the next paper of the present Volume.

All this is not merely a magic of numbers. Many of the found regularities are straightforward consequences of simple physical arguments, mostly of stability considerations, while others indeed "depend rather delicately on apparent *coincidences* among physical constants" which in turn prove to be prerequisites of our existence (Carr & Rees, 1979). A tiny distuning of initial data or strengths of interactions or particle masses or asymmetries would be enough to completely destroy our comfortable Universe, which seems as a "suit tailored just to our human measure". We may therefore be surprised to find ourselves in an "anthropomorphous" cosmos. The fully unexpected message of these considerations appears to be that both the "anthropocentric" and the "anthropomorphous" characters are properties of Nature herself, so that these attributes begin to lose their purely pejorative meaning.

Such very general physical considerations are just appropriate for interdisciplinary studies, because they connect astronomical, geological, biological and microscopical objects, and also for evolutionary studies because they fix the possible stable endpoints of evolutions.

REFERENCES

Carr B. J. & Rees M. J. (1979): Nature 278, 605

Diósi L., 1990: in this Volume

Duncan W. J., 1955: Physical Similarity and Dimensional Analysis. E. Arnold & Co., London

Eddington A., 1987: Space, Time & Gravitation. Cambridge University Press, Cambridge

Lukács B. & Paál G. (1981): Csillagászati Évkönyv 1982, Gondolat Budapest. p. 250 (In Hungarian)

Lukács B. & Paál G. (1988): Astroph. Space Sci. 146, 347

Paál G. & Lukács B., 1989: in Symmetry of Structures, eds. Gy. Darvas & D. Nagy, Budapest, p. 402

Sen D. K., 1968: Fields and/or Particles. The Ryerson Press, Toronto

---

+ Recently there have appeared indications for protoaggregates catalysed by heavy neutral leptons as well, corresponding to neutron star mass, size and density.

COSMOGENESIS

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

G. Paál[*] & B. Lukács[+]
[*]Konkoly Observatory, H-1525 Bp. 114. Pf. 67., Budapest, Hungary
[+]Central Research Institute for Physics, H-1525 Bp. 114. Pf. 49.,
Budapest, Hungary

Received on 12th Oct., 1990

ABSTRACT

The seemingly esoteric statements of modern cosmology are brought within reach of nonspecialist. The simplest well known physical laws, including Einstein's $E=mc^2$ with c as universal limiting velocity, suffice to grasp modern cosmogenesis as the most natural possibility. Modern particle physics seems to realize this possibility.

A. PHENOMENOLOGY

A.1. GENESIS OF SPACE

Astronomical observations strongly support a uniform (homogeneous & isotropic), expanding and gravitating Universe. Surprising conclusion is that all these already imply that the space itself is expanding (not the cosmic matter expands in fixed space). We demonstrate (or rather visualize) this statement for 2-dimensional uniform spaces (surfaces) as examples.

a) *Sphere*. Uniformly expanding matter is possible only on a *properly expanding sphere*; on a *fixed* sphere matter smeared out isotropically around one pole along radial lines cannot expand uniformly at the same rate along transversal directions (parallel circles). This is a purely geometric constraint.

b) *Plane*. On an infinite *fixed* plain "pastry-board" a uniformly expanding infinite "paste" is impossible because maintaining uniformity of paste requires an expansion rate proportional to distance, which in turn would lead to v>c outwards of some distance, excluded by relativity. Uniformly expanding paste (matter) is possible only *without* unexpanding board (representative of 2-dimensional absolute space). Of course, coexpanding space, with coexpanding coordinate system, remains possible.

Detailed calculations [1] verify the statement that all these remain true for 3 dimensions and any constant curvature and even if relativistic corrections are taken into account for distances, times and velocities (except for vanishing gravity, irrelevant in our gravitating Universe).

So *uniformly* expanding motion must lead to expanding (or increasing) space, by other words it generates new space.

KINESIS OF MATTER LEADS TO GENESIS OF SPACE.

Uniform expansion of space can be mathematically described by *one* function of time R(t), e.g. the changing distance between any two fixed spots on an expanding baloon or plane.

We have seen that matter drives space, therefore we have to determine the dynamics of matter in order to describe the dynamics of comoving space, i.e. to determine the laws for R(t).

A.2. EQUATIONS FOR SPATIOGENESIS

Classical Newtonian theory of gravity in itself is inadequate to describe the motion of an infinite uniform Universe: it gives infinite forces from all directions whose sum is indefinite [2], while their naïvely supposed mutual cancellation by symmetry would mean an effectively forceless nongravitating Universe, in disagreement with the observational indication of the previous paper of the present Volume. On the other hand Newtonian theory can give gravitational acceleration only relative to a fixed inertial system while we already know that in the Universe there can be no rigid frame of reference at all. There is however one more possibility for getting a guess for the dynamics of an infinite Universe from the Newtonian theory, which is free of the above contradiction, inconsistency and indefiniteness.

THE BEHAVIOUR OF AN INFINITE MAXIMALLY SYMMETRIC SYSTEM IS TO *BE REGARDED AS A LIMITING CASE OF THAT OF MAXIMALLY SYMMETRIC FI*NITE CONFIGURATIONS.

Each point of a finite sphere moves with nonvanishing gravitational acceleration relative to its center according to the equation:

$$\tfrac{1}{2} m\ddot{r} = -GmM(r)/r^2, \qquad (1)$$

where the dot denotes time derivative, r is the radius of an auxiliary sphere extending from the center to the prefixed point, while M is the mass inside it. By symmetry, the accelerations from any outside spherical shell cancel. So increasing the initial sphere by inclusion of further outer shells up to infinity does not change the limit which remains (1), as if the outer shells were absent. By homogeneity, the relative acceleration of any two points of any sphere is also given by (1).

Note that eq. (1) gives a first integral of motion:

$$\tfrac{1}{2}\dot{r}^2 - GM/r = const. \qquad (2)$$

If we look for the analogon of the above equations for the motion of the *space* rather than that of the matter, the rôle of distance r is taken by R, and then we get:

$$\tfrac{1}{2}\dot{R}^2 - GM/R = \mathrm{const.} \qquad (3)$$

This is indeed exactly an equation of Einstein's General Relativity for uniform Universes [1]. This result is not surprising because the irrelevance of the outer shells for the inner situation *is also true in General Relativity via Birkhoff's* Theorem [3]. On the other hande for small spheres (weak gravity) the predictions of the two gravitational theories coincide. So the above agreement is necessary for a self-similar Universe, which cannot possess any prefixed scale.

This gravitational equation of the *whole* Universe clearly shows that contrary to the naïve expextation an unchanging static Universe is impossible. *Namely in an unchanging Universe M must* be constant, then eq. (3) implies

$$\ddot{R} = -GM/R^2 \neq 0. \qquad (4)$$

So space should be generated (created or destroyed) if matter is not. (Remember the classical experience: a purely gravitational system of bodies is either expanding with deceleration or contracting with acceleration; and a *homogeneous* pressure does not acts as counterforce.)

GRAVITY GENERATES MOTION, MOTION GENERATES SPACE.

Then consider an expanding Universe still with constant M. Eq. (3) shows that smaller size R corresponds to larger $\dot{R}$; below some R the right hand side is negligible compared to *any* term on the left hand side, and then the solution to eq. (3) (with a suitable 0 point of time) is

$$\dot{R}^2/t^2 \sim 1/R \dashrightarrow R^3 \sim t^2 \qquad (5)$$

just like in Kepler's Third Law (for circular orbits).This equation would suggest "nothing" (i.e. no space, R=0) at the "beginning" (t=0). However the previous paper of the present Volume indicates that unquantized physics and usual geometry breaks down at least at $R_{Pl} \sim 10^{-60}R_U \sim 10^{-33}$ cm, where R_U is the size of the present observable Universe. This is then a genesis of space from "almost nothing" (multiplying the scales by a factor of "only" 10^{60}.

A.3. GENESIS OF MATTER

In principle a successful cosmogenesis scenario ought to explain the genesis of both space and matter. Indications for such a "materiogenesis" indeed come from physical considerations as well. Simplest of them goes as follows. Let us go back into the suspected initial state. There $R \sim R_{Pl}$, but $\rho \sim \rho_{present}(R_U/R_{Pl})^3$, which is horribly above ρ_{Pl}, the maximal interpretable density. So there must still be something basically wrong with the suspected early evolution, and the density change is obviously far overestimated. Instead, when already near ρ_{Pl}, this density ought to remain constant. But hence $M \sim R^3$, a drastic creation of matter (out of almost nothing, M_{Pl}). As much matter ought to be created as needed to fill up the space being created.

SYNCHRONOUS GENESIS OF SPACE AND MATTER IS THE PRICE OF POSSIBLE EARLY UNIVERSES.

A.4. EQUATIONS FOR MATERIOGENESIS

Another very well known classical equation also retains its form for expansion of space in General Relativity, and this equation clearly offers the possibility of genesis of matter. This is the law of adiabaticity of expansion: which must hold true, since, because of homogeneity, there is no net exchange of heat or anything else between large parts. This coincides with the second and last of the components of Einstein equation for a homogeneous isotropic Universe and reads as

$$\dot{E} = -P\dot{V} \qquad (6)$$

where $E=Mc^2$, P is the dynamic pressure and V is the comoving volume$\sim R^3$. In familiar circumstances this equation is interpreted as describing the change of the internal energy of a gas in a cylinder behind a piston as a consequence of the work of the pressure of the gas eaten up by the expansion. In such cases we usually tacitly assume that the internal volume expands at the expense of the outer one, and then the same is true for the energies (with opposite sign). But for a homogeneous Universe there is no "outside", the same relative increase happens everywhere. Therefore there is no other part for balancing the changes for either space or energy. Under such circumstances the very eq. (6) which permitted conservation by compensation *generally* will prescribe the change of energy (matter) via genesis of space in the lack of compensation.

Expansion means $\dot{V}>0$, so sign of $\dot{E}$ = -sign of P. For usual matter with $p>0$ expansion leads to *destruction* of matter, while $p<0$. Therefore the very early Universe must have had *negative* pressure (for mutual genesis of space and matter, inavoidable, as seen, for possible initial states).

MATTER SO GENERATED DOES NOT COME FROM ANYWHERE, AND IS MADE NOT BY NOTHING BUT BY THE MATTER JUST THERE.

Constant density, required above, leads to

$$P = -\rho c^2 \qquad (7)$$

via eq. (6). (Being $E=\rho c^2 V$.) This is by certain an unusual pulling pressure instead of the familiar pushing ones, but not impossible. Note from

$$E = TS - pV. \qquad (8)$$

of thermodynamics that a surplus of energy leads to defect of pressure (T is temperature, S is entropy). Just this fact eats up the metastable phase in phase transition.

A.5. GENESIS OF EXPANDING MOTION

There still remains an important question. *What* does maintain the permanent expansion resulting in permanent genesis of space and thus of matter? The answer again comes from eqs. (3) & (6). For R they combine as

$$\ddot{R} = -(4\pi/3)G(\rho+3Pc^{-2})R \qquad (9)$$

which reveals the important fact that the source of gravity is not only $M=\rho V$ but also PVc^{-2} (which is negligible in present laboratory circumstances). So negative P yields *antigravity* effects, and in case of eq. (7) attraction is overbalanced by repulsion. This means that even without gradient counterterms, impossible in a homogeneous Universe, the bodies are *repelled* by gravity in the very early Universe.

This repulsion can also be seen from eq. (3) alone, if ρ=const.; then $\ddot{R} \sim R$, so $R \sim e^{\alpha t}$.

EXPANSION IS MAINTAINED, EVEN ACCELERATED, BY ANTIGRAVITY.

Thus in this early Universe, gravitational "action at a distance" is repulsive (via $\rho+3P/c^2<0$) leading to expanding motion, while the local pressure is attractive (via $P<0$). Between neighbouring parts the direction of force is opposite to that of motion, resulting in negative work, so in increase of internal energy. This situation reminds one to the case of friction or to the spontaneous motion of a spring in compression phase. (But there the negative work produces heat or strain, not additional matter in the very same form as the preexisting one.)

A.6 SCENARIO FOR COSMOGENESIS

On the basis of all said above a recipe of a complete cosmogenesis out of *almost* nothing (i.e. a state with Planck data) is formally fairly easy, and contained already in the basic eqs. (3-6). At the start one needs a slightly expanding medium with $P \approx -\rho c^2$; this generates antigravity, which in turn generates more expansion, i.e. more space, which generates more matter with $P \approx -\rho c^2$, which again generates additional antigravity, &c. in a complete circle of genetic process. There is no need for any external force or energy, the creation of motion, space and matter happens completely free of charge. So, if this scenario is indeed realised, at first we have a cosmic era of "self-building", a mutual creation of space and matter. This should, however, be stopped, and followed by an era with $P \geq 0$, an era of "self-dilution", a creation of space without new matter, leading to the present situation with islands of ordinary cosmic matter separated by huge almost empty interstellar space. "More space than needed for matter." The presently observed nonnegative pressure is therefore needed for the evolution of present cosmic structure.

NEGATIVE PRESSURE CREATES RAW MATTER FOR BUILDING STRUCTURES; POSITIVE ONE STOPS THIS CREATION TO PRODUCE THE ELBOW ROOM FOR BUILDING.

A.7 QUANTUM BIRTH OF COSMOS

The outlined scenario does not start from *strictly* nothing; an embryonal Planck Universe is required. But such an object is already a single microstate or particle in view of Heisenberg's uncertainty principle which reads as

$$\Delta E \Delta t \approx h \qquad (10)$$

meaning that any date of it is at cca. one "fluctuation" from its absence. E.g., by virtue of eq. (10) such an entity cannot be divided into distinguishable parts. But then it is the *par excellence* quantum object, and therefore the eqs. (3-9) belonging to the unquantized limit are not valid for it. So until complete "relativistic quantum gravity" will be at reach, we cannot speak of a cause or antecedent of this Planck state of the Universe, and it is quite possible that such notions will not have any meaning even in the full theory. Guessing from the uncertainty principle itself, one expects that such a state may spontaneously appear and disappear (just as in the creation-annihilation processes of virtual particles) with a Planck time scale. However, if it appears with $P \approx -\rho c^2$, then there may start a spontaneous energy creation with a Plack time scale too, so the energy may exceed the limit of quantum fluctuation back to 0. Then some energy or matter survives, and can initiate the irreversible circle of self-amplifying creation of energy, space and motion. With some luck the road is open from a virtual microworld to a real causal macroworld.

EVEN CREATION NEED NOT BE A VIOLATION OF ETERNAL LAWS, BUT JUST A COMPILENCE WITH THEM.

According to the above considerations, we have a good chance of uncovering the secrets of cosmogenesis even on the basis of our present science. Remember: "God is sly, but not malicious" as Einstein put it.

B. MICROSCOPY

Now we are going to try to realize the program of the scheme of Part A according to the present status of quantum field theory, general relativity, particle physics, &c. We start from Planck data (or just with generating them), and go forward with the expansion.

B.1 TO THE PLANCK UNIVERSE

Relativistic Quantum Gravity as an operative theory does not yet exists. However some characteristics of it can be guessed from Quantum Field Theory and General Relativity, including their inconsistencies.

In a time-dependent geometry the Hamiltonian is time-dependent too. So, starting with a vacuum of no energy, later the same vacuum may have some energy density and pressure. For details see [4]. Now, this is a pure Quantum Field effect, so it is not clear, what is to be written into the Einstein equation as the *classical* energy-momentum tensor. But using some expectation value of the energy-momentum operator one gets a self-consistent system of equations, which may be a kind of the "classical" limit of the "true" "Quantum Gravity" equations.

In a model calculation it turned out [5] that the cosmologic solution of these equations cannot be extended back in time until the curvature singularity (Big Bang). The solution appears without preliminaries at a radius $\sim R_{Pl}$, in expansion. Before it the solution does not exist. This may be interpreted as the first distinguishable appearance of anything just above the quantum fluctuation level. The model system realizes the result of dimensional analysis of the previous paper.

B.2 SELF-CREATION NEAR PLANCK DATA

For the circle of generation of matter, space and motion a serious negative pressure was anticipated. This is possible in the spirit of eq. (7) if the energy density of the system has a positive lower bound. To demonstrate this, let us take a system with

$$\epsilon = as^{3/4} + b \qquad (11)$$

with a and b constants, where s is the entropy density. By pure thermodynamics

$$p = (\epsilon - b)/3 - b \qquad (12)$$

So near $\epsilon = b$ the $p \approx -\epsilon$.

Quantum Field Theories sometimes predict such a positive lower bound. By simpler words, in such situations even the ground state makes the space-time curved. From observations we know that our present vacuum (the idealized interstellar space) causes at best a negligible curvature (compatible with 0), therefore up to observational margin this state puts a convenient zero point for the energy density. Now, our case appears, if a ground state of a Quantum Field theory (i.e. a relative minimum) has an energy surplus compared to the empty interstellar space.

The simplest example is a quartic potential with two minima of different depths, with $\epsilon = 0$ in the absolute minimum. If during the cosmic expansion the cooling state, by any chance, happens to enter the shallower minimum, then a positive minimal energy is present even at T=0, and then the machinery starts to work.

Such quartic potentials appear in any Quantum Field theory using spontaneous symmetry breaking for generating masses for the vector bosons (in order to get rid of unrenormalizable divergences). The theory of highest known energy scale with such property is Grand Unification, where the energy scale is 10^{-4} Planck energy, $\sim 10^{15}$ GeV, $\sim 10^{28}$ K [6]. So at this temperature, if something goes wrong, a creation cycle can start.

The problem is that in our Universe the matter later obviously came out of this ground state. It may seem easy to imagine such a scenario, e.g. in such a way that the actual state leaves the positive energy "false" vacuum by tunnelling through the central peak of the quartic potential. However, there is a tendency to have serious problems when realizing any scenario in which the state is only transiently trapped in a false vacuum: e.g. it is difficult to tunnel back during an exponential expansion [7].

A sequence of ingenious scenarios were invented for transient trapping; each of them needs additive assumptions, coincidences or at least a tuning of parameter values of the theory. (For a review of peculiarities see Ref. 8.) However, working scenarios are *possible*; here we do not have to be too curious about the actual scenario which had been successful.

B.3 MATTER DURING THE GRAND UNIFICATION ERA

The present knowledge about particle physics shows that the symmetry group of interactions is SU(3)xSU(2)xU(1). Here the first symmetry belongs to the quantum chromodynamics (QCD), which is the interaction of quarks via gluons (and whose peripherial effect is the "strong interaction"): the symmetry is a consequence of the 3 possible "colours" of quarks, not influencing the interaction. The remaining group is that of the Weinberg-Salam theory of the "unified" electroweak interaction [6]. Although not quite correctly, one can visualize it as an SU(2) symmetry of weak interaction (same for neutron and proton or for electron and neutrino), and an U(1) symmetry of electromagnetism.

Now, the smallest group, which is simple and contains the above group is SO(5). Assuming that this minimal group is the fundamental group of physics, we arrive at the simplest Grand Unification, and for our present goals that is enough. One can classify the elementary particles of this theory as follows.

Scalars: Spin=0. the Higgs bosons of the theory producing the quartic potentials needed in the theory. *Not yet seen.*

Fermions: Spin=½. They are: quarks, 6 flavours and 3 colours, *not seen because of confinement* (see later); leptons, 3 charged, e, μ and τ, *all seen*, 3 corresponding neutral neutrinos, ν_τ *not yet seen.* They are subjects to some conservation laws.

Vector bosons: Spin=1. They are: γ, W and Z, *seen*, gluons, *not seen because of confinement*, X and Y, *unseen*. They mediate the interactions of fermions; γ establishes electromagnetism, W and Z belong to weak interaction, gluons to chromodynamics, while X and Y work between quarks and leptons.

Because of the complete symmetry SO(5), *if* the actual state is so symmetric too, then quarks and leptons are equivalent except for different charge &c. All the interactions have the same strengths, and quarks and leptons can be converted into each other in some processes. Then only two types of combinations of particle numbers are conserved: net electric charge Q and baryon-lepton numbers for 3 families B_α-L_α. (For the first family the baryons considered are neutrons and protons, while the leptons are the electron and its neutrino.)

Since these combinations are strictly conserved in the theory, it is worthwhile to count the conservation constants in the present observable Universe. The matter turns out to be neutral, so Q=0. As for B_α-L_α, the third family would involve baryons with top and bottom quark, and τ lepton with its neutrino. Such particles are not found outside of laboratory, so B_3-L_3=0. Practically the same holds for the second family (strange and charmed baryons, muons and muon neutrinos). For the first family the counting is not quite possible. From neutrality proton and electron numbers are equal. Neutrons appear in nuclei, and according to cosmic abundance of elements (cf. the contribution in this Volume) the neutron number is roughly 15% of the proton number. But electron neutrino numbers cannot be taken from observations (most cosmic neutrinos cannot be detected.) Some numerical simulations of cosmic evolution would suggest the present neutrino density in the order of the photon density, which, in turn, is 9 orders of magnitude higher than the electron density. However, in L_1 the difference of the neutrino and antineutrino numbers appears, which may practically be anything. With a luck this difference *may* be so slight as the neutron number, and therefore B_1-L_1=0 is *permitted* by observations, however it is impossible to prove it.

Then one has arrived at a *possible* Universe whose *all* conserved quantities are 0. Only such a Universe can appear via quantum fluctuations. Henceforth we assume the 0 values of the above conserved quantities.

The present Universe does not exhibit particle-antiparticle symmetry: while particles are abundant, antiparticles are practically absent (except for neutrinos). This may imply either a primordial symmetric state with CP-violating laws or a CP asymmetric primordial state with symmetric laws. Here we will not choose between the possibilities, only note that very weak CP violations are observed in the present environment (most clearly for Kaons).

With a quartic potential of Higgs self-interaction the actual state is very different for high and low temperatures. When the temperature is high enough ($T>E_0\sim10^{15}$ GeV, which is the Grand Unification energy scale), the actual state is well above the central peak of the potential, so in expectation value it is mirror-symmetric. In such a state each particle, except for the Higgses, is massless (for details see [7] or [9]).

On the othe hand, at low temperatures the actual state is located in one of the side valleys near the minimum. Then the expectation values of some Higgs fields differ from 0 (the symmetry of the actual state is less than that of the theory). Therefore vector bosons coupled to these Higgses (X and Y) get effective masses in the order of E_0. At temperatures far below E_0 the massive bosons can appear only virtually, for a very short time, so with a very short range. Therefore the interactions mediated by them become practically absent. Henceforth the symmetry SO(5) is broken to our *observed* symmetry SU(3)xSU(2)xU(1), and baryon and lepton numbers are practically separately conserved henceforth. They take just the values valid during the transition era.

The symmetry breaking is expected to be in connection with a phase transition; for cosmic phase transition scenarios see the review [9].

B.4 THROUGH THE DESERT OF PARTICLE PHYSICS

By solving the dynamical equations (3-6) with a radiation field $P=\epsilon/3\sim T^4$ one obtains $T\sim1/\sqrt{t}$, where the numerical factors contain G and the number of particle species light compared to T. For details henceforth we refer to [9] and citations therein.

Hence one can get that the breakdown of SU(5) symmetry happened at cca. 10^{-35} s after Big Bang or the primordial fluctuation. Afterwards the matter cools and nothing serious happens until T crosses the next fundamental energy scale. Now, no energy scale is observed, known or strongly predicted in particle physics between 10^{14} and 1000 GeVs. (The desert.) Therefore it seems that there was an uneventful cooling until $t\sim10^{-10}$ s.

B.5 THE SEPARATION OF ELECTROMAGNETISM AND WEAK INTERACTION

The energy scale of the Weinberg-Salam theory is somewhere between 100 and 1000 GeVs. Reaching that, the W and Z bosons get masses from interactions with a Higgs field whose symmetry just has broken down. The resulting masses are *experimentally measured* to be 88 and 95 GeVs. So below such temperatures these bosons will appear only virtually, with a shortening range. This is the reason why the "weak interaction", mediated by W and Z, is indeed very weak in the present Universe.

B.6 HADRONISATION FROM QUARKS

The vacuum state of Quantum Chromodynamics seems to have a *positive energy density* $B^4/(hc)^3$, *where B is the energy scale of* the theory. So for $T<B$ the QCD vacuum is energetically dispreferred. The possibility to reach our present vacuum is a neutralisation of QCD forces. Since QCD contains 3 different charges ("colours"), this can be achieved in two ways. Either 3 quarks, or a quark and an antiquark must form a close group; then force lines are restricted to a small volume, with our familiar vacuum outside. According to measured quantum numbers of particles, the *first formation is a baryon* (the lightest is the proton with 938 MeV rest energy) and the second is a *meson* (the lightest is the pion with 140 MeV rest energy).

From a separated quark the "force lines" would extend to infinity, with infinite energy. Therefore free quarks cannot be produced at any finite energy, which is the phenomenon called confinement, and is conform to the negative experience. This is *the reason that quarks and gluons are not observed*.

However, the theory predicts the quark content of hadrons, so from the measured hadronic masses the parameters of QCD can be deduced. ("Hadron spectroscopy".) According to various models, 145 $MeV<B<260$ MeV. Hence the temperature, below which the *quarks* form hadronic configurations is cca. 160 MeV.

This temperature was reached at 8 μs. Then hadronic blobs started to appear; this bubbling lasted for cca. 7 μs (according *to model calculations*). At the end of this period all our familiar particles have appeared.

THE PRESENT UNIVERSE CONSISTS OF PROTONS WHICH WERE ONCE CREATED BY THE UNIVERSE.

B.7 THE LEPTONIC ERA

Since the temperature of hadronisation is almost the same as the pion rest energy, all the freshly created hadrons rapidly decay *or annihilate*, except for protons and neutrons, saved by baryon conservation. However T is still far above the electron mass, so electron-positron and neutrino-antineutrino pairs are still abundant. This remains so until $T \sim 0.5$ MeV, $t \sim 1$ s.

B.8 PRIMORDIAL NUCLEOSYNTHESIS

At $T \sim (m_n - m_p - m_e) \approx 1$ MeV the neutrons start to decay. However the *neutron half lifetime is 12 minutes*, therefore they still have chances to build up stable nuclei as $d=(pn)$, $He^3=(ppn)$ and $He^4=(ppnn)$. Of them the most weakly bound is d with 2.2 MeV binding energy. Since T is already below this value, some light nuclei can survive.

The competing processes are the nucleosynthesis and neutron decay. For order of magnitude the cooling scale time is just t (the evolution equations have a power solution when radiation *dominates the energy-momentum tensor*), therefore it is short compared to the neutron life time. So the produced nuclei are not destroyed by photons at the end of the period. From the present *observed mass density numerical simulations* indicate cca. 20% He^4 created in this primordial nucleosynthesis. Deuterons and heavier nuclei are negligible.

B.9 THE BIRTH OF ATOMS

Not far after 1 s $T \approx m_e c^2$. There the e^+e^- pairs annihilate and we are left only with the slight electron surplus. The leptonic era is over.

Henceforth the story depends on the possible existence of a further mass scale ~30 eV. (See the previous article of this Volume.) If the neutrino has such a mass, then something drastic happens at the corresponding temperature. Namely there the *neu*trinos become nonrelativistic, and with some more cooling they become bound by their own self-gravity. At such a temperature the critical mass is about 10^{15} solar mass; above that the configuration is bound and starts to collapse.

If the neutrinos are massless, then the critical temperature is cca. 10 eV, which is the binding energy of a hydrogen atom. Below this energy electrons are captured by protons and they form a bound neutral atom. When *the atoms dominate, the matter becomes* transparent for photons because of the lack of free electric charges. After this moment the photons cannot efficiently destroy the random denser blobs of matter if their mass is sufficient (again cca. the above critical mass is obtained). These blobs start to collapse as above.

This mass roughly corresponds to a galaxy cluster, now observable; the corresponding time is ~100000 ys after Big Bang. *Therefore we can conclude that the formation* and separation of proto-galaxy clusters happened at that time.

B. 10 BIRTHS OF GALAXIES AND SMALLER ASTRONOMICAL OBJECTS

A completely spherical collapse of the protoclusters is a matter of initial conditions and highly improbable. Therefore the expected fate is a collapse+rotation, ending with breakdown of the original blobs *into parts. It is quite possible that local* galaxy groups and individual galaxies were born in this way. Later the protogalaxies could also collapse, and again initial asymmetries led to fragmentation. Such a process can stop at the mass which is stable against further collapse. As it was mentioned in the previous paper this mass scale is cca. 1 solar mass.

NOW THE UNIVERSE IS SEEN TO BE BUILT UP FROM GALAXIES AND STARS *BUT HAS NOT BEEN BUILT UP FROM GALAXIES AND STARS BUT GALAXIES* AND STARS HAVE BEEN "BUILT DOWN" FROM THE EXPANDING UNIVERSE.

C. THE PRESENT PICTURE

All this above is theory or theoretical calculation. No doubt, according to the known physical laws the Universe could generate space, matter and motion itself, but the question is, whether it generated them (therefore itself) in the "proper", i.e. *whether the generated Universe* is the present one or not.

Obviously a lot of details is *not* obtained, partly because some details of the laws are still unknown, partly because the later evolution happened in inhomogeneous local environments, where the calculations need different methods. Still we can ask about the general features.

The largest coherent structures seen by astronomers are the galaxy clusters of 1000-10000 galaxies. This means ~10^{15} solar masses, in accordance with the above story. Then galaxies, spherical clusters, &c. are seen, which may be the decay products of the protoclusters. In addition, stars are seen with ~1 solar mass, as predicted, and planets, whose formation belongs to later articles of this Volume, similarly to other objects on the planets between planetary and atomic sizes.

Neutral atoms are around us, and the scenario predicted them, at least the H and He atoms. With serious difficulties the cosmic abundances can be deduced by comparing spectroscopic data with our knowledge about physical situations on the surface of stars [10]. One such curve is Fig. 1. Indeed, H and He are predominant, but heavier atoms are also present. The simplest explanation of the discrepancy is that heavier elements have been formed in the stars, and the subsequent articles will show that it is indeed so. Still, Fig. 1 is not without interest for us. First, observe the roughly exponential decrease with atomic weight. This is quite conform with the gradual building of heavier elements in a relatively *young* Universe (no saturation); the local minima and maxima can easily be explained from details of nuclear physics and astrophysics. Second, such abundances are the initial conditions for (new) stars, planets, living organisms, &c. And indeed, the curve shows local maxima at the *fundamental* constituents of (known) planetary bodies and organisms. Their building atoms were already present.

REFERENCES

[1] H. P. Robertson & T. W. Noonan: Relativity and Cosmology. Saunders, 1969
[2] Four Letters from Sir Isaac Newton to Doctor Bentley, Containing Some Arguments in Proof of a Deity. London, 1756
[3] R. C. Tolman: Relativity, Thermodynamics and Cosmology. Clarendon Press, Oxford, 1934
[4] G. W. Gibbons & S. Hawking: Phys. Rev. D15, 2738 (1977)
[5] L. Diósi & al.: Astroph. Space Sci. 122, 371 (1986)
[6] P. Langacker: Phys. Rep. 72C, 185 (1981)
[7] T. W. B. Kibble: Monopoles in Quantum Field Theories, ed. M. S. Craigie, World Scientific, Singapore 1982, p. 341
[8] B. Lukács: Proc. 4th Marcel Grossmann Meeting Rome, 1985, ed. R. Ruffini, North-Holland, p. 1393
[9] B. Kämpfer, B. Lukács & G. Paál: KFKI-1989-64, in print in ECHAYA
[10] Eva Novotny: Introduction to stellar atmospheres and interiors. Oxford University Press, New York, 1973

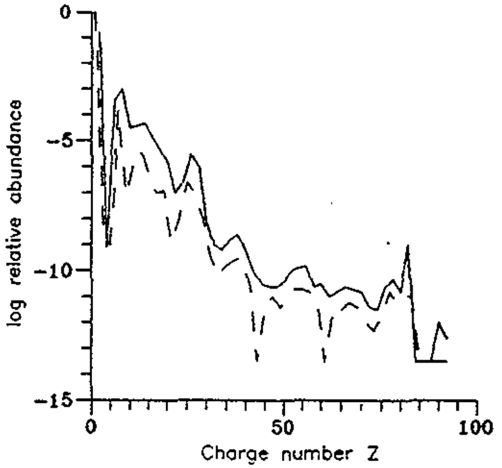


Fig. 1: Logarithms of cosmic abundances of elements, separately for even (solid) and odd (dashed) atomic numbers. −13: not seen.

Now the next article is about stability of nuclei, therefore some peculiarities of the cosmic abundance curves will be explained. But remember that in earlier stages of the Universe this curve was steeper, because the hot Universe itself practically did not produce elements beyond He. So the first stars had a different composition, consequently they may have been different for luminosity &c.

E

ON THE STABILITY OF NUCLEI

KFKI-1990-50/C B. Lukács & al. (eds.): Evolution: from Cosmogenesis to Biogenesis

L. Jéki
Central Research Institute for Physics
of the Hungarian Academy of Sciences
H-1525 Bp. 114. Pf. 49., Budapest, Hungary

Received on 2nd Oct., 1990

NUCLEAR FORCES

The simplest composite nucleus is deuterium which consists of one proton and one neutron. It is known from experiment that fusion of these components to form a nucleus of deuteron is accompanied by the release of about 2.2 MeV energy in the form of gamma radiation. This is the binding energy of deuterons. The question arises what kind of force keeps nucleons together? This cementing force cannot be gravitation for then the binding energy would be 36 times less in order of magnitude. An electromagnetic interaction - more exactly a magnetic one (as the neutron is electrically neutral) - would also lead to a binding energy two orders of magnitude lower. It is obvious that an interaction stronger than those of classical physics is active in nuclei. We shall call this strong interaction. Our aim is to gather information on it and on this basis we shall try to answer the question: What is the cementing force keeping nucleons together and how does this force stabilize nuclei? After many decades of studying nuclear forces, and having collected a considerable amount of experimental and theoretical data we are still unable to tell exactly what nuclear forces are or what we really mean by strong interaction. This can be attributed, among other reasons, to the fact that the interaction between nucleons and between nuclei - the nuclear forces - actually appear as the resultant of fundamental strong interactions between the constituents of nucleons, these constituents being even "more elementary" particles.

In nature, 287 stable or very long lived isotopes are found. About 2400 further isotopes have been produced artificially and it is conceivable that we shall be able to study another few thousand in the future. It depends on the measuring techniques, which of the increasingly unstable isotopes - with shorter and shorter lifetimes will be produced and investigated.

The cohesion that binds nuclei is strongest in the middle of the periodic table and the binding energy per nucleon decreases towards both lighter and heavier elements. In other words, the latter - light and heavy - elements are energetically unstable. The smaller ones can build up larger nuclei by fusion, which process is accompanied by the release of energy. The reverse process also involves release of energy: the fission or decay of large nuclei leads to an energetically more advantageous situation, too. Theoretically this means that, on a purely energetic basis only the elements around iron might exist and all remaining elements - if they ever existed - should have become transformed into these. We know, however, how many stable isotopes exist in nature. When looking for the factors and conditions which lead to stabilization, we also have to take into account the Coulomb interaction between positively charged protons. During fusion, the identically charged nuclei approaching each other have to overcome Coulomb repulsion so that the nuclear forces can exert their influence. On the other hand, the fission of heavy nuclei into two parts presupposes that one fragment moves out of the force field of the other thereby shifting from the range of attraction of nuclear forces to the range of Coulomb repulsion. As we shall see, the stability of nuclei depends on the joint effect of not just two but of several factors, actually on the resultant of attractive and repulsive forces. The decisive factor is the balance of nuclear and Coulomb interactions.

As a first approach, we endow these attractive nuclear forces with one single property first formulated by Wigner: they have a short range of action (within 1 to 2 fermi). Starting from this assumption, we can try to calculate the binding energy. When calculating the total kinetic energy of the nuclei, it is considered that they form a Fermi gas with the nucleons moving freely in the nucleus, without any force field. (Since any nucleon in the nucleus is surrounded by further, uniformly distributed nucleons, the actions of force of the latter neutralize each other which means that none of them in the nucleus feels any force. This condition can obviously be met inside a big nucleus only and not on its surface.) When calculating the potential energy, we have to make assumptions on the two-nucleon potential. Let the interaction be an attractive and distance dependent one of purely Wigner type. The contribution from the surface energy also has to be taken into account. In order to calculate the Coulomb energy we assume a uniform proton distribution with the electrostatic interaction between protons being proportional to the square of their charge. We also take into account the so-called "pair energy". This is the contribution from nucleon pairs of identical state but opposite spin which may be formed in the Fermi gas of nucleons otherwise obeying Pauli's exclusion principle. The calculated binding energy is a function of mass number A and nuclear charge Z where the coefficients of the individual terms are to be determined by fitting to experimental data. The formula obtained is therefore a semi-empirical one:

$$\varepsilon_k = C_T A - C_F A^{2/3} - C_C Z^2 A^{-1/3} - C_S (A-2Z)^2 - C_P \delta A^{-3/4}$$

It is the second term in the formula – the contribution from surface energy – that is essential for light nuclei and the Coulomb term for the heavier ones. These terms describe that the curve of binding energy per nucleon changes into decreasing. This semi-empirical formula of the binding energy well describes the general trend but it does not reflect the fine structure of the experimental curve – namely that adjacent nuclei may have considerably different binding energies. In view of this, we should make further assumptions on nuclear forces.

When calculating the formula of binding energy, we made suppositions which partly contradict each other: nuclear forces were conceived of as attractive and distance-dependent but the nucleon density in a nucleus was supposed as being constant. Under purely attractive nuclear forces the nucleus would collapse and could not become saturated. Nuclear forces must be more complex than has been supposed up to now: the purely attractive Wigner term should be completed by a repulsive core. This was introduced by Jastrow . This repulsion is active up to about one third of the range of action of nuclear forces. The picture is even further complicated by the Majorana potential which yields an attractive or repulsive contribution on the basis of quantum mechanical considerations which cannot be detailed here.

NUCLEAR MODELS

Let us try to construct a model of the nucleus by using this more exact, modified description of nuclear forces. This task will not be an easy one bearing in mind that nuclei show quite varied properties. Although heavy nuclei contain a number of nucleons but this number is too small for them to be treated by statistical methods. The perturbation approach, well proven in atomic shell physics, is not applicable, either. Instead, we can compare nuclei to well known physical systems. The models obtained in this way describe certain domains of the nucleus and certain experimental observations. At present none of the models is able to supply a full description: each particular model describes another property and/or another domain of the nucleus.

The liquid drop model was developed in the early forties as a means of interpreting the fission of heavy nuclei. In view of the short range of action of nuclear forces it can be assumed that nuclei interact with their immediate neighbours only, similarly to molecules in a liquid. According to this model, proposed by Bohr, , the nucleus behaves as a liquid drop: upon absorbing energy it begins to vibrate which may cause it to change into two droplets connected by an elongated neck. As the distance between these smaller droplets increases, the nuclear force acting between them is gradually reduced until the droplets can penetrate through the potential wall. Fission takes place as soon as the total energy of the fission fragments becomes lower than the energy of the deformed nucleus. Bohr's liquid drop model is well able to explain fission phenomena but it does not answer the general questions of the stability of nuclei.

In the shell model, postulated by Goeppert Mayer and Jensen in 1949, nucleons also obey the laws of quantum mechanics. They can occupy discrete states of well defined energies and motions. Within the nucleus, nuclei are ordered according to Pauli's exclusion principle. Each proton is at a separate energy level from the lowest one, and the number of filled states equals that of the protons. Neutrons, similarly, fill another series of energy states. If the nucleons occupy the states at the lowest energy levels then the nucleus is in its ground state.

The shell model describes the nucleus by analogy with the atomic shell's structure.The elements whose electrons fill each position in a closed shell are marked out by their stability. The elements with completed electron shells are the noble gases. Empirical facts show that the stability of nuclei is a function of the completedness of the proton and neutron shells. The nucleon numbers indicating the closure of the shell are the so-called magic numbers:

2, 8, 20, 28, 50, 82, 126, 184.

The twofold magic nuclei, in which both protons and neutrons form a closed shell, are especially stable. Such nuclei are, for example, the nuclei of calcium-48 (with 20 protons and 28 neutrons) or lead-208 (with 82 protons and 126 neutrons). The shell model explains the stability of nuclei; it becomes clear why calcium has so unexpectedly many (6) stable isotopes for Z=20. The twofold magic calcium-48 is stable in spite of its high relative excess of neutrons. The same can be said of tin with its 10 stable isotopes for Z=50.

The collective model was worked out by Aage Bohr and Mottelson in 1952. This unites certain features of the shell and liquid drop models. According to this the motion of nucleons in the unfilled outer shells of the nucleus is determined by the inner, completed shells. On the other hand, the nucleons outside the closed shell also influence the nucleons of the core and the shape of the latter. Magic nuclei and those adjacent to them are hard, rigid. Nuclei lying farther from magic ones are easily deformable, the core may take the shape of a cigar (elongated sphere) or disc (flattened sphere). Another version of the shell model (developed in another direction by Greiner et al. in 1969 is the so-called two-centre shell model. This improved the description of the fission process by showing how the pre-nuclei of the fission poducts are formed in heavy nuclei. This theory also predicted new modes of radioactive decay (which we shall discuss later). In addition, this model is well applicable in designing superheavy elements by describing how and which nuclei are able to unite into superheavy elements.

Let me here briefly advert to the matter of superheavy elements. The stability, i.e. the half-life, of unstable transuranic elements rapidly decreases with increasing atomic charge. The half-lives, which are of about several thousand years in the vicinity of uranium, shrink to a split second beyond the atomic charge of 100. At the same time one can expect, based on the success of the shell model, that magic numbers will exert their stabilizing effect in the range of superheavy elements as well. After lead, the next expected twofold magic nucleus would contain 114 protons and 184 neutrons. In fact, this stabilizing effect, expected for the atomic charge of 114, is already effective for much lower values. At present the synthesis of superheavy elements has reached atomic charge 109. Experimental data show that the quantum mechanical effects described in the shell model are already active for atomic numbers 106 to 109. *Surprisingly, some of the nuclei in this range were observed to* "resist" spontaneous fission; these nuclei transform by alpha decay. The production of superheavy elements, their investigation, and the study of different combinations of nuclei whose fusion can yield the wanted superheavy element are the subject of one of the most intensively studied, highly informative branches of nuclear physics today.

Even such a brief outline of nuclear models reveals how manifold and diverse the underlying ideas are. We have treated nuclei as the free particles of a gas, as the molecules of a liquid – in the droplet model –; and calculations have also been carried out in which a nucleus is conceived as a solid, crystalline structure. This crystalline model is also able to reproduce the fundamental statements of shell and droplet models whereas it considers the nucleus as a lattice consisting of alpha particles. This model also makes use of the discovery that nucleons themselves have a composite structure, being built of quarks. Here we would only refer to the fact that efforts towards a deeper understanding of nuclear forces have long been linked with the achievements and problems of particle physics. It was as long ago as in the thirties that Yukawa supposed the existence of a mediating particle of strong interaction, similarly to photon mediating electromagnetic interaction. These predicted particles - pi mesons or pions - were, indeed, discovered by experimental physics. The meson theory of nuclear forces and the consideration of quark-gluon problem attempt to describe nuclear forces at a deeper level than the phenomenological models.

DECAY MODES

Up to now we have dealt with nuclei "at rest" when nucleons and the whole of the nucleus are in their states of lowest energy; this is the ground state of the nucleus. The nucleon may absorb energy from the surrounding space or from a particle colliding with it, whereby it becomes excited, i.e. it shifts to a higher energy state. The nuclear models enable these excited states to be calculated in good agreement with experimental data. From the *unstable excited state the nucleus "falls back" to its ground* state by releasing energy in the form of radioactive radiation. The emission of electromagnetic rays, i.e. gamma quanta, does not change the composition of the nucleus, only its energetic state. The possible gamma transitions between the individual excited states are governed by selection rules based on sophisticated quantum mechanical considerations.

When emitting beta radiation, the nucleus is transformed: the element (isotope) appearing in its end state differs from that in the initial state before the decay. A neutron may be transformed into a proton, then an electron and an antineutrino can be observed in addition to the proton. If a proton transforms into a neutron, the process is accompanied by the emission of a positron and a neutrino. The proton of the nucleus may capture an electron from the shell, the emitted particles are then a neutron and a neutrino. Consequently, the different versions of beta decay involve a transition of neutrons and protons into each other in both directions, which means that the composition of the nucleus changes. This type of transition depends upon basic energetic conditions, the possible processes can be calculated from the binding energies. The difference between the masses of the initial nucleus and the one that is finally obtained has to cover the mass of the electron (positron) and the kinetic energy of the electron and neutrino. Should the mass difference be too low to meet this condition, no decay will take place. In the weak interaction governing beta decay sophisticated quantum mechanical relationships accompany the energetic conditions to make certain transitions allowed or forbidden to different extent. Correspondingly, a number of different transitions are possible between two adjacent nuclei. For instance a nucleus may show a preference to transform from a highly excited state into an adjacent isotope rather than to its ground state.

With the semi-empirical formula of binding energy we mentioned the balancing role of nuclear forces and Coulomb interaction. For heavy nuclei with mass numbers over 200, alpha decay (the emission of a nucleus of He) becomes energetically favourable. Still more favoured is the decay - fission - of the nucleus into two parts. With elements occurring in nature, spontaneous fission without excitation takes place with a very low probability, in the form of a tunnelling process. In the case of far transuranic elements the preferred, natural way of transformation is spontaneous fission. Energetically these nuclei are so likely to decay that any rearrangement of their components reduces the potential energy thereby promoting the decay.

Consequently, nuclear transformations take place by radioactive radiation (alpha, beta, gamma) and nuclear fission. In addition to these well-known and frequent processes some further forms of decay were also observed which may be called exotic. These have also enriched our knowledge of nuclei though they play a practically negligible role in the formation and transformation of nuclei.

EXOTIC DECAY MODES

As long ago as in 1914 it was assumed that a nucleus may also emit one or more protons during radioactive decay, but this process was not actually observed until 1981. During the fusion of two nuclei of medium mass number in a particle accelerator, the resulting nucleus was observed to transform into a stable isotope through several steps. It was in this decay series that the proton emission took place. Two-neutron decay was also observed: the artificially produced lithium-11 isotope transforms into the beryllium-11 isotope by emitting two neutrons.

The recently discovered nuclei rich in neutrons and having an anomalously large radius cannot be described simply by the available nuclear models. In certain light nuclei which abound in neutrons the two most loosely bound neutrons form a "neutron halo" around the nuclear matter of normal density. Experimental data published by CERN in 1990 confirm this result; the weakly bound neutron halo transforms into a deuterium particle. Thus a new type of radioactive decay has been found: the beta-delayed deuteron emission. It was previously observed in nuclei abounding in neutrons that the daughter nucleus forming in beta decay (possibly even in excited state) is not stable and immediately disintegrates into its components. At first the emission of neutrons, tritons and alpha particles was observed during this decay. Recently further data have augmented these findings. The process, in which the nucleus of He-6 emits a deuteron after the beta decay, is the following:

$$^{6}He \rightarrow {}^{6}Li^{*} + \beta^{-} + \bar{\nu}$$

$$^{6}Li^{*} \rightarrow \alpha + d$$

The decay of the daughter nucleus is governed by strong interaction, whereas the initial beta decay involves weak interaction only, so the latter seems to be slow for a process of nuclear physics. The half-life is 10^{-2}s in the initial beta decay whereas it is 10^{-21}s in the second step, so the overall half-life of the whole decay process is determined by the slow beta decay. It is supposed that one of the neutrons of the halo undergoes beta decay, i.e. it transforms into a proton which is then attached to the second neutron to form a deuteron. The bound proton-neutron system has a very weak binding - only a few times 10 keV - even so it is able to escape since the strong interaction already weakens outside the core of the nucleus, in this case in the halo.

When studying elements heavier than lead, in certain cases the emission of a particle heavier than the alpha particle has been observed. The new process takes place simultaneously with the alpha emission, and competes with it. Its probability is by orders of magnitude lower: a single decay of this new type occurs for every one milliard alpha decays. In the experiment the unstable radium-223 isotope produced in the decay chain of uranium-235 was investigated. Radium-223 disintegrates by alpha emission, too, but theory also allows the emission of carbon-12, -13, -14 and -15. The end product is one of the isotopes of lead. The theoretically calculated most probable emission is that of carbon-14, this probability is by several orders of magnitude higher than those for other carbon isotopes. The expectation was confirmed by experiment, the emission of carbon-14 could indeed be observed. Later it was shown that radium-222 and radium-226 also emit nuclei of carbon-14 but this is about ten times less likely than radium-223. This new type of decay is known as cluster radioactivity.

The emission of neon-24 was first observed in 1985: in Dubna from protactinium-231, uranium-233 and thorium-230; in Berkeley from uranium-232. In the case of uranium-232 there is one neon emission to 500 milliard alpha decays. The emission of magnesium was observed with the decay of uranium-234. Uranium-234 shows three types of cluster radioactivity: it emits He- (alpha), neon and magnesium nuclei.

THE BIRTH OF ELEMENTS

On the basis of studying several hundred stable and several thousand unstable isotopes a concept has crystallized on the stability of nuclei, on nuclear forces, on the transformation of nuclei into each other. This knowledge may also help in trying to answer the question when and where the various isotopes bornduring the history of the Universe. According to the standard model of the Universe, about one minute after the Big Bang it was cold enough for the formation and survival of bound deuterium nuclei. In the course of several steps, almost the entire quantity of hydrogen transformed into helium through successive nuclear reactions. The appearance of the remaining elements is already closely connected with the evolution of the stars, with the history of different types of stars, with the birth and formation of galaxies. Astronomy is called upon to answer the questions when and in which stars the conditions prevailed that enabled the heavier and heavier elements to be built up. During fusion reactions helium burns to form carbon and oxygen, carbon to form neon, sodium, magnesium, etc. Then, over the burning cycles of neon and oxygen (not detailed here) and through the burning of silicon one can arrive at the formation of manganese, chromium, iron, cobalt and nickel - i.e. of the heaviest elements obtainable by fusion. In the stars, following different patterns of evolution, the burning cycles may take place slowly, quasi-steadily, but also in an explosion-like manner. For the formation of elements heavier than iron neutrons are needed. Neutrons are always released in fusion processes in which light elements are formed. With the capture of a neutron a new isotope of the given element is produced. If this isotope is unstable, then it transforms by beta decay into the isotope of an adjacent element with the same mass number. This process may repeat itself several times, thereby gradually leading to the formation of heavier and heavier elements. This is the slow way of the formation of elements (s-process). Under special conditions - e.g. at the beginning of a supernova explosion - the elements around iron, formed by fusion, may be exposed to an extremely strong neutron flux as well. Then the nucleus captures a number of neutrons before starting to transform through one or more beta decays towards its stability valley. Through such multiple neutron captures (the so-called fast or r-process) isotopes far from stability are formed.

In connection with nuclei, a number of quesitons arises: Why is one nucleus stable while one that is only slightly different being unstable? When, where and in what kind of physical processes were the different isotopes formed? Is their abundance in the Universe uniform or not, and why is it like weobserve it. We have partial answers to these questions, but many new discoveries are needed in various branches of physics and astronomy in the coming decades before we can really clearly see the history of the birth and later life of elements.

ON A POSSIBLE LINK BETWEEN NONRELATIVISTIC GRAVITATION AND QUANTIZATION

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

Lajos Diósi
Central Research Institute for Physics
H-1525 Budapest 114, P.O.B.49, Hungary

Received on 1st Oct., 1990

Abstract

It has been shown that nonrelativistic gravitational and quantum effects become comparable at length scales about 10^{-5}cm. We have conjectured that, in this intermediate region, a certain nontrivial unification of gravitation and quantum theories is needed which would, *furhtermore*, anticipate peculiar new effects and objects.

Learning three disjoint domains of phenomena, three fundamental physical constants have become necessary to beintroduced. They are, in turn:

G – Newton's gravitational constant *for celestial motion*,
c – velocity of light for electromagnetic fenomena and
$\hbar$ – Planck's constant for atomic effects.

Later on, more complex phenomena have also been investigated where two ones of the above three fundamental *phenomena have been* simultaneously involved. The corresponding mathematical equations contain two fundamental constants, accordingly:

G and c – in cosmology,
$\hbar$ and c – in particle physics,
$\hbar$ and G – well, that is just the issue to be discussed below.

Let us add that, *for the recent decades*, a peculiar domain of physicshas also been considered where all the *three fundamental constants must be included.*
G and c and $\hbar$ – in the cosmology of the very early universe.

Coming back to the combination $\hbar$+G we have to admit: we do not know any phenomenon where both gravitation (G) and quantization ($\hbar$) would simultaneously play essential roles, while relativity (c) could not. Nevertheless, there does exist a logical motivation to assume *such phenomena* as *well* as to guess their domain[1-3].

Let us consider two bodies, each of mass m, at a distance r from each other. The energy of their gravitational interaction is about

$$U_{grav} \approx Gm^2/r. \qquad (1)$$

At the same time, the relative velocity of the two bodies has an uncertainty of about $\hbar/rm$, due to Heisenberg's uncertainty relations. Thus Quantum Mechanics predicts the following kinetic energy for the relative motion:

$$U_{quan} \approx \hbar^2/r^2m \; . \qquad (2)$$

Let us assume the gravitational and quantum kinetical energies are of the same order. The proper condition is the following:

$$rm^3 \approx \hbar^2/G \approx 10^{-47} cmg^3 \; . \qquad (3)$$

This condition does not allow us to determine both the distance r and the mass m. Assume, however, the two bodies are of normal density (i.e. cca. $1gcm^{-3}$) and, furthermore, their sizes are of the order of their distance r. Then the above mathematical condition will lead to the following critical mass and size, respectively:

$$r_{crit} \approx 10^{-5} cm \; , \qquad (4)$$

$$m_{crit} \approx 10^{-15} g \; . \qquad (5)$$

Bodies of the above ranges of size and mass, getting close to each other, would be influenced by a competition between nonrelativistic gravitational and quantum effects. In practice, such a system is dominated by other, e.g. van der Waals forces, nevertheless, the above critical length and mass scales possess principal sense. One can not exclude the possibility that, in a certain domain of phenomena, the Newtonian gravitation and the quantization show up together indeed and they do that just on the critical scales calculated above.

Physical systems, smaller than the scales (4)-(5), are expected to satisfy the usual quantum mechanics, i.e. the Schrödinger equation. On larger scales, however, quantum coherence gets blurred by gravitational uncertainties and, as a result, classical mechanics will govern the dynamics of the system in question.

In a sense, the scales (4)-(5) represent a borderline between micro- and macroworld. Objects with scales about (4) and (5) belong to the intermediate region where gravitational and quantum effects are in competition. Their equations are to be matched together. However, history of physics teaches us that such a unified theory must be more than a simple coupling the two equations, Newton's and Schrödinger's ones, actually.

Still we might not have a firm theory for the intermediate region, therefore unexpected effects or objects may turn out to be related to that region. On Fig. 1 we show a collection of various natural scales. It is noticable that the smallest known autonomous biological beings, i.e. mykoplasms, have sizes and masses about (4) and (5), respectively. Whether it is merely an accidental coincidence or something more, we admit it is open question to discuss.

References

[1] L.Diósi: Phys.Lett. 105A, 199 (1984)

[2] L.Diósi and B.Lukács: Annln.Phys. 44, 488 (1987)

[3] L.Diósi and B.Lukács: Newtonian Quantum Gravity, in Proc. of Balatonszéplak Relativity Workshop (CRIP, Budapest, 1985)

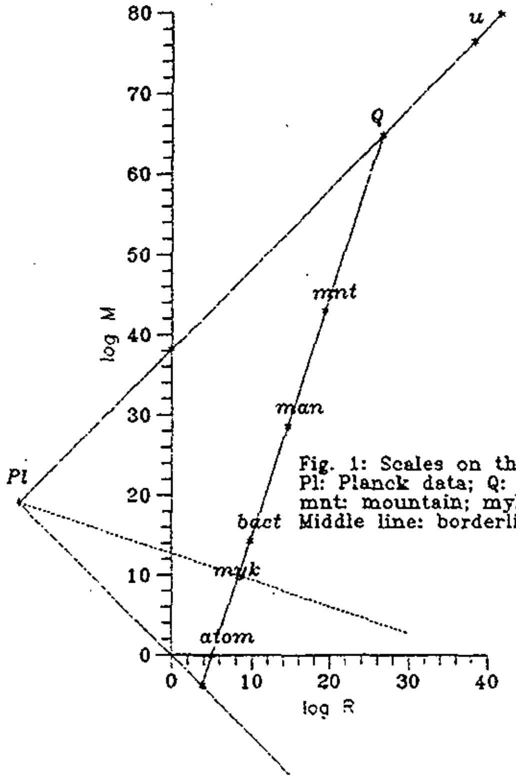


Fig. 1: Scales on the (M,R) diagram.
Pl: Planck data; Q: quasar; U: Universe;
mnt: mountain; myk: mykoplasma.
Middle line: borderline from eq. (3).

The quantum gravity effects seem to introduce a stochastic element into the purely deterministic quantum evolution equations. This may perhaps lead to a new type of "indeterminism", or to a greater degree of freedom of behaviour. If this new freedom is important for life (for which we do not want to argue here), then Life cannot be described with the well-established physical theories available now, and then further theoretical work is needed.

E

Evolution of the Galaxy

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

L.G. Balázs
Konkoly Observatory
Budapest, XII. Konkoly-Thege út 13-17
1121 Hungary

Received on 17th Sept., 1990

ABSTRACT

Our world is rich in structures. Our deep past was significantly different from the observable face of our world. The process stretching between the present world, rich in structures, and the ancient world, poor in it, was progressed on two strongly connected parallel branches, on the cosmic and on the chemical-biological branches. The formation of the chemical elements stopped at the light atoms (lithium) in the homogeneous stage and the building up of the heavier elements, up to the uranium, was probably going on within the galaxies.

INTRODUCTION

We are living in a world, rich in structures. Our body consists of cells and the cells themselves are set up from very complicated organic compounds, the proteins, containing many thousands of atoms. One may split these very complicated organic compounds into atoms, and the atoms into electrons, protons neutrons. These particles, however, are not the final building blocks of the matter, if one may speak from the smallest ones at all. These series of structures flow somewhere into the so-called intensive infinity. Looking towards the macro- and mega-size phenomena on the other side, we see series of structures building one on the other. Our life is connected with a cosmic body, the Earth, and the energy necessary to maintain it given by an other cosmic body , the Sun. The Sun and the Earth, together with its eight 'planet-brothers', form the Solar System. The Sun, however, is only one of the 10^{11} stars of our Galaxy, the Milky Way. The galaxies are concentrated into clusters of different sizes and this series disappears for our eyes somewhere in the so-called extensive infinity. Taking into account the fact that no information can proceed faster than the light looking into the extensive infinity brings us deeper and deeper into the past as well. The modern observing astronomy and the related theoretical investigations revealed that our deep past was significantly different from the presently observable face of our world. What we observe today is a snapshot of a long cosmic history starting with a hot ancient stage, poor in structures, and producing the presently observable face of the Cosmos originated from the early phase by means of an explosion-like expansion producing a high variety of cosmic structures. The process stretching between the present world, rich in structures, and the ancient world, poor in it, was progressed on two strongly connected parallel branches. The first was the cosmic branch yielding the condensation of the matter into galaxies, stars and planets. The other branch is the chemical-biological branch producing the organization of matter into atoms, molecules and biological structures. From the philosophical point of view it is appropriate to use the term of evolution for the very complicated process starting with the ancient stage and leading to the human being, organized into society. Nevertheless, some particular aspects of this evolution are only irreversible processes. In the astronomical terminology, however, they are also called evolution. According to our present knowledge the matter was microscopically and macroscopically rather homogeneous before its fragmentation into cosmic bodies. The formation of the chemical elements stopped at the light atoms (lithium) in this homogeneous stage and the building up of the heavier elements, up to the uranium, was probably going on within the galaxies. The molecules, in particular the organic compounds, were synthesized only in certain types of galaxies, like our Milky Way, and mostly at special places, in the dust and gas clouds, strongly concent- rating along the plane of symmetry of these galaxies. The formation of biological matter shows even much greater scatter in space and time. We know for instance that life does exist on the Earth but not on the Moon which is our next door neighbor on cosmic scale and probably we may say the same thing for the whole Solar System. Taking into account the fact that a significant part of the cosmic and chemical processes, required for the origin of the life, were going on within our Galaxy it seems to be reasonable to study the galactic part of these processes in more detail.

THE GALACTIC STORY

The story undoubtedly started with the formation of our Galaxy. There are many misty details in this story at the present but it does not mean that we do not have scientific ideas on the formation of the Milky Way. It means rather that we have more answer for one question and we do not know, based on the available observations, which of them is the correct one. We know, however, that the globular clusters are the known eldest cosmic objects in our stellar system forming a slightly flattened, slowly rotating subsystem in our Galaxy. The chemical abundance of the elements heavier than the carbon is in most cases less than 1%, such elements can not be synthesized, the low abundance of the heavier elements indicate, in general, an early epoch of formation after the birth of the Galaxy. Nevertheless, the presence of these heavy elements in the globular clusters shows that they do not represent the primordial objects of the Milky Way because their heavy element content was already processed by some preceding stellar population. There were several attempt to find objects containing no heavy elements but without any results until now. Despite all of these one thinks that the globular clusters and those stars having similar chemical composition and kinematical properties as the globular clusters display the primordial form of the Milky Way. If we try to answer the question how our stellar system got its presently observable face we have to answer the question how the stars were born (where and when), what kind of dynamical effects the stars were experienced bringing them to the place we observe now and how they were enriched with different heavy elements. It is a characteristic feature of presently observable face of our Galaxy that the motion of a considerable fraction of the stars significantly differs from that of the globular clusters. These stars form a rapidly rotating strongly flattened system. This is the galactic disk and the system which the globular clusters belong to is the halo. The abundance of heavy elements of the disk stars is similar to that of our Sun which is also a member of this subsystem. The youngest stars are strongly concentrated to the symmetry plane of the disk. The interstellar matter , consisting of gas and dust particles of half micron diameter, is also concentrating on the same way. These youngest objects do not populate uniformly the disk but display characteristic spiral arms. The concentration of the interstellar matter to the symmetry plane of the Galaxy and the similar distribution of youngest stars together with the ordering of these objects into spiral arms led to the obvious assumption that there is a basic relationship among the birth of the stars, the interstellar matter and the presence of the spiral arms.

THE BIRTH OF THE STARS

It is a basic observational fact that the young stars always appears in interstellar dust and gas clouds. One may ask, of course, how can we decide which stars are young. In 1947 Ambarcumjan sovjet astronomer pointed out that the trapezium system in the Orion constellation was dynamically unstable. It means that the stars of the trapezium have to be very young because since their formation only a short time could pass due to the instability of the system. The stars of the trapezium were hot blue stars. Ambarcumjan showed furthermore that these hot blue stars, the so-called O and B type stars always appeared in smaller-greater groups, in stellar associations. The stellar associations are gravitationally weakly bounded systems. Moreover, according to some measurements, there are among them gravitationally unbounded systems as well. Therefore, if we observe them presently it passed only a short time on cosmic time scale since their formation. In the second half of the 50th and especially starting from the 60th due to the rapid development of the electronic computers it became possible to modelize the birth and evolution of the stars. These models confirmed that the O and B type stars finished their careers within a few million years and it was justified on this way Ambarcumjan's idea on the young age of these objects also theoretically. The close relationship between the spatial distribution of the young stars and the interstellar matter obviously suggested between them some genetic connection. A widely accepted idea for this relationship was the instabilities and the subsequent gravitational contraction of the inter- stellar matter. Trying to make some quantitative prediction based on this hypothesis one encounters serious difficulties. The basic physical quantity of the gravitational fragmentation of the interstellar matter is the Jeans length indicating the charac- teristic size at which the diffuse media become unstable against gravi- tational contraction at a given density and temperature. Taking into account the average density and temperature of the interstellar matter then this characteristic size equals a few times 10^4 solar mass. It means that a cloud having the mass of a stellar association could be formed on this way but it remains open the question how proceeds the fragmentation until the birth of the stars. A possible way for it if we assume that the Jeans length is also changes during the subsequent contraction and cooling of the clouds formed by the fragmentation of the homogeneous primordial diffuse matter. This situation changes drastically by taking into account other instabilities and processes which are capable of producing conditions in the interstellar matter being significantly different from the average. Involving processes which are helping the condensation of diffuse matter to overcome the 'initial difficulties' and changes its temperature and density then the formation of individual stars might be possible. There are many discussions in this context on the role of different shock processes propagating in the interstellar medium and making suitable conditions for star formation. There are several triggering mechanisms for exciting shock waves of different sizes. Such triggering mechanisms could be the cloud-cloud collisions, stellar wind bubbles around hot luminous stars and dynamical disturbances in the disk of our Galaxy. At this point we turn back to the problem of origin of spiral arms characterizing certain type of galaxies, like ours. The dynamical response of the disk of tis type of stellar systems is a spiral form disturbance propagating as a density wave. The gravitational potential well changes the streaming direction of the interstellar matter. Since this streaming supersonical this change triggers a shock wave. The shock wave makes the necessary compression of the interstellar matter providing suitable conditions for star formation. The stars born in the wave front and having short lifetimes, e.g. the hot blue giant stars, we observe near to their birth places and displays beautifully the spiral arms. The formation of the stars from the diffuse matter is probably not a strictly gravitational process. The presence of the magnetic force significantly changes the hydrodynamical behavior of the matter. In any case the star formation is completed by the gravitation.

DINAMICAL EVOLUTION

The primordial Galaxy differed mostly from its present face in that respect that it was a slightly flattened, slowly rotating spheroid and did not contain the disk. The first generation of the stars conserved this feature. After the formation of the stars their motion was strictly governed by the gravitational force. The fate of the remaining matter was different. Due to cooling down it was continuously loosing energy while its angular momentum remained unchanged. As a consequence it could contract along the axis of rotation but in the perpendicular direction it much more difficult. This difference in the contraction along the rotational axis and in the perpendicular direction led to the formation a gaseous disk and the star formation continued in the disk. The timetable of the star formation was different in the different subsystems of our Galaxy. The birth rate of the stars in the halo is very low recently. The gas in the halo produced the globular clusters and the density of the rest of the matter is in general to low at the present time for the formation of new stars. In the disk of the Galaxy the conditions are completely different. Since a lot of gas was concentrated after the initial contraction, leading to the formation of the disk, the birth of the stars has been continued until now. The initial dynamical evolution made an important by-product, the central condensation of the Galaxy, the bulge. The large spatial number density of the stars may indicate that the initial gas density and the stellar birth rate was very high in this region. The present value of this birth rate is very low, together with the low density of the interstellar gas. According to some hypotheses the gas is continuously swept out by a galactic wind or by some explosions going on in the central part of our stellar system. The central region of the Galaxy is not simply a region populated by stars more densely. High resolution radio measurements indicated that there is an object of less than 10 astronomical unit diameter and of 4 million solar masses. Some people believe that this central body is a massive black hole and the release of the energy, which is also observable in the centre of the quasars, proceeds by means of the matter falling into the black hole. The riddle of the quasars and the active galactic nuclei are in the highlight of the astronomical research since 25 years but we are far from understanding the way of the formation of those objects and the source of the energy released.

CHEMICAL EVOLUTION - THE DEATH OF THE STARS

As we have mentioned earlier the fusion of the atomic nuclei was stopped somewhere at the lithium, after the Big Bang and before the fragmentation of the matter into clouds. The subsequent chemical evolution took place already within the Galaxy. We have mentioned earlier that the primeval stellar population conserved the initial form of our Galaxy. To study the process of the chemical enrichment of the Milky Way we have to investigate their spatial distribution. As to the heavier elements, the Earth, Sun, meteorites, the interstellar matter and the stars in the solar neighborhood have a very similar chemical abundance in respect to the relative abundance of the elements heavier than the hydrogen and the helium. The subsystems of our Galaxy differ from each of the other in relative abundance of the hydrogen and the Helium to the heavier elements. The chemical composition of the Sun is usually called as the cosmic abundance which is $X=.59$, $Y=.39$ and $Z=.02$ where the numbers represent the weight fraction of the hydrogen, Helium and the heavier elements, respectively. We can find stars in the Solar neighborhood which contain much less heavy elements than our Sun. These metal poor stars also differ in their kinematical properties. They do not belong to the disk of our stellar system like the globular clusters and their heavy element abundance is 1/10 and 1/100 that of the Sun. There is strong correlation between the kinematical properties and the chemical abundance of the stars. The stars in the highly flattened disk have nearly circular orbits around the centre of the Galaxy and their chemical abundance is similar to our Sun. The stars in the halo, at the contrary, have orbits of large excentricity and low heavy element content. In the context of the dynamical evolution we mentioned already that the galactic disk probably born by the collapse of the diffuse matter not condensed into stars during the formation of the halo. The spatial distribution of the heavy elements indicate that the chemical enrichment took place during the formation of the disk. It means that some process had to exist producing heavy elements and recycle them into the inter- stellar matter and the newly born stars inherited already this abundance of elements. There is an other gradient in the abundance of the heavy elements pointing towards the galactic center, very well pronounced by the globular clusters. According to this gradient the heavy element content of the globular clusters increase in the direction of the centre of our galaxy. For this gradient probably the processes are responsible producing the heavy elements during the formation of the central condensation. In the light of all these facts we have to consider what kind of processes could be responsible for the heavy element formation. If we intended to synthesize these elements from lighter components (the Big Bang produced only light nuclei) we can get some idea on the possible processes and the necessary physical conditions, by comparing the initial and the final chemical abundances. The burning of the hydrogen into Helium is going on at few ten million degrees. This process is the major source of the energy production of the vast majority of the stars. The burning of Helium into heavier elements requires some hundred million degrees and the building up of the elements on this way can proceed to the iron, energetically the most stable among the elements. In the reality this process can not get the iron because the star becomes unstable before reaching it. The synthesis of elements heavier than the iron requires some non-equilibrium process. This non-equilibrium process may take place in the supernova explosions. Among the most precious records in the old chronicles are those reporting on the unexpectedly appeared new stars. The most famous among them perhaps the celestial phenomenon appeared in 1054 and recorded by chinese astronomers having at present a characteristic nebular patch, the Crab nebula, in its place. This nebula emits observable electromagnetic radiation from the gamma to the radio range indicating the presence of energetic particles moving in strong magnetic field. According to theoretical computations these stellar explosions, called also supernova explosions, sign the death of stars heavier than a certain critical threshold. There are controversial values for this critical mass but it probably larger than 1.6 solar masses, the Chandrasekhar limit. The corresponding investigations revealed that the lifetime of the stars having masses lower than our Sun comparable with the age of our Galaxy. We mentioned earlier that the heavier elements represent only a few percent of the mass of stars having the Solar chemical abundance. Their estimated total mass equals about 10^{11} solar masses. It means that about 10^{9} solar masses had to be converted into heavier elements. Analyzing the frequency of supernovae in external galaxies one concluded that in the average there is one flare-up in every few decades in galaxies similar to ours. As a matter of fact the last supernova in our galaxy was observed by Kepler in 1604. If we compared these frequencies with the 10^{10} years age of the galactic disk we conclude that every supernova had to produce 2-3 solar masses heavier elements. Of course, not this whole amount was built up by explosive processes but still a significant fraction should be produced by explosions. Taking into account the large uncertainties of these estimations the supernovae could be responsible for all of the heavy elements but it is worth of discussing other possible resources. There are very energetic processes in some galactic nuclei which appears dramatically pronounced in the quasars. Unfortunately, until now we are far from having detailed models from these energetic phenomena and therefore we are not in a position to make quantitative estimations concerning the heavy element production in these processes. The next step in the chemical evolution was the synthesis of the molecules in the interstellar space. In the last 20 years by means of the radio astronomical observations one discovered a number of organic molecules, predominantly in dense interstellar dust clouds. Namely, inside these clouds the molecules are protected against the destroying interstellar radiation field. So the chemistry and the survival of molecules is localized at certain particular places in our Galaxy. There is a very complicated process towards the birth of biological structures and the appearance of intelligence. The cosmic phenomena outlined obviously give a general frame work for them but the details of these processes are far beyond the scope of the present paper.

About the mentioned possible black hole in the galactic center. it is worthwhile to remember the reader to the paper "Physical Backgrounds..." in this Volume. There Fig. 2 gives interesting constraints between masses and sizes.

Now, an object with 4 million solar masses can be a black hole with a size cca. 1% of the mentioned observed upper limit.

E

EVOLUTION OF STARS

Attila Grandpierre

Konkoly Observatory, Budapest
1525 Budapest, P.O.Box 67

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

Received on 5th Oct., 1990

ABSTRACT. It is presented here the main nuclear reactions producing energy in stellar interiors. By these reactions the stars are working as alchimic factories in transmutating the elements building up the periodic system of elements from the hydrogen until the iron and more. Then it is given an outline of stellar evolution on the Hertzsprung-Russell diagram. I point out, that the widely propagated belief of the exclusiveness of the contraction at stellar formation has to face with basical challenges confirmed by new observations. The early stages of stars seem to show behaviours contradicting the generally held views of stellar birth. I give here some ways to test the possibility that some unknown, basically new reaction is at work at young stellar cores (e.g. a phase transition of Higgs-bosons).

DETAILS. The stars are starring by their own intinsic nature. As far as we understand it now, mostly gravitational and nuclear energies are liberated in the stellar bodies. Until today, it is not made any systematical test about any basical need to look for some new fundamental process supplying stellar energies. What we know is that most of the stars which are old enough, we can interpret with a more or less coherent picture in which these stars are born from a contraction of a protostellar nebula when the two main energy producing process are the gravitational contraction and the nuclear fusion.

Nevertheless, the young stars are presenting some principal difficulties to this picture. In this way it seems that our understanding of the stars can get a radical innovation just from the study of the evolution of the young, just born stars. I want to show here some ways opening before us in this field.

By the classical contraction hypothesis the protostellar nebula starts to contract by a gravitational instability or with the help of some outer effect e.g. shock-waves from supernovae or high-velocity stellar winds from neighborhood young stars. In this first stage the star is contracting and the liberating gravitational energy is used to rise the temperature of the star. Half of the energy produced is radiated away from the surface of the star. During the contraction the first energy producing nuclear reaction starts to work at about million Kelvin degree. This is the so-called deuterium-burning (Stahler,1983). At this stage the star has a large radius and a low surface temperature, so the star is above the main sequence in the temperature-luminosity diagram (Hertzsprung-Russell diagram, see Fig.1.) where our Sun is now.

$p + {}^2H \rightarrow {}^3He + \gamma \quad T \sim 10^6$ K

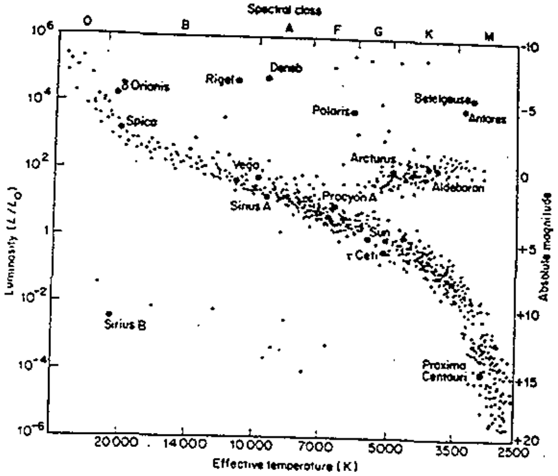


Fig.1.
Hertzsprung-Russell diagram of the solar neighborhood stars

After the exhaustion of the deuterium-fuel the star begins to contract again and the energy is supplied again mainly from the gravitation. When the temperature at stellar cores reaches four million Kelvin degree, a new reaction starts to work, the proton-proton reaction (Salpeter,1952, Bethe,1939).

$$\begin{tikzcd}[column sep=small]
p + p \rightarrow {}^{2}H + e^{+} + \nu_e ,\ T \geqslant 4\times10^{6}\ K & \\
{}^{2}H + p \rightarrow {}^{3}He + \gamma \arrow[d] \arrow[dr] & \\
{}^{3}He + {}^{3}He \rightarrow {}^{7}Be + \gamma \arrow[d] \arrow[dr] & {}^{3}He + {}^{3}He \rightarrow {}^{4}He + 2p \\
{}^{7}Be + e^{-} \rightarrow {}^{7}Li + \nu_e & {}^{7}Be + p \rightarrow {}^{8}B + \gamma \\
{}^{7}Li + p \rightarrow {}^{4}He + {}^{4}He & {}^{8}B \rightarrow {}^{8}Be + e^{+} + \nu_e \\
 & {}^{8}Be \rightarrow 2\,{}^{4}He
\end{tikzcd}$$

As a result, the proton-proton chain produces helium from hydrogen,

$4H \rightarrow {}^{4}He$ and liberates 6.7 MeV per nucleon energy.

During the proton-proton chain the energy production goes with the *fourth power of the temperature,*

$\varepsilon_{pp} \propto T^{4}$.

The next fusion reaction, the CNO cycle begins to work when the central temperature reaches $T > 16\times10^{6}$ K,

$^{12}C + p \rightarrow {}^{13}N + \gamma \qquad \varepsilon_{pp} \propto T^{15}$

$^{13}N \rightarrow {}^{13}C + e^{+} + \nu_e$

$^{13}C + p \rightarrow {}^{14}N + \gamma$

$^{14}N + p \rightarrow {}^{15}O + \gamma$

$^{15}O \rightarrow {}^{15}N + e^{+} + \nu_e$

$^{15}N + p \rightarrow {}^{12}C + {}^{4}He$.

As a result, $4\,{}^{1}H \rightarrow {}^{4}He$ and the CNO elements are catalizators. When the hydrogen is also exhausted, the star begins to contract until its central temperature reaches 10^{8} K when the He burning starts.

$3\,{}^{4}He \rightarrow {}^{12}C$

$4\,{}^{4}He \rightarrow {}^{16}O$.

At this stage the star turns off the main sequence and becomes a red giant because the huge energy production blows up the envelope of the star and its radius becomes so large that the surface temperature decreases significantly.

The star then exhausts its He content too and gravitational contraction of stellar cores periodically alternates with new types of nuclear reactions while the central temperature of the star is continuously arising,

$2\,{}^{12}C \rightarrow Mg, Ne, Na, Al$

$2\,{}^{16}O \rightarrow S, Si, P$

When the central temperature reaches 2000×10^{6}K, the stellar core begins to produce the most stabil element, the iron. With the development of the iron core, the star can easily be unable to exert enough thermodynamical pressure in its core against the hydrostatic pressure of the envelope and the star can collapse and explode which we interpret as a supernova explosion. During the explosion all *the heavier elements are easily produced* and are thrown into the interstellar space.
In this way the supernovas are making us the favour to produce the earthly elements and to develop the background of our existence here (now the reader is expected to feel her/himself as a supernova remnant). Recently it appeared new theories acclaiming that two or three or more supernovas are necessary for explaining the details of the chemical composition of the Earth.

As a first check of the contraction hypothesis we can use the results of the neutrino detections from the solar core(Davis, 1970-). The standard theoretical solar models suggest that we should have to measure a *solar neutrino flux with the proton-proton* cycle (the central temperature of the Sun is about 14 million K) 7.8×10^{-36} counts/sec, L = 7.8 solar neutrino unit(SNU). The solar neutrino problem, which is a basical challenge of our understanding the Sun, consists in the fact that the measured neutrino flux is less than the third of this value,

L = 2.1 SNU (Davis et al.,1990).

What means more, recently there are new and independent measures of the solar neutrino flux strengthening this result and simultaneously suggesting strongly that the solar neutrino flux is changing in (anti)correlation with the solar magnetic activity *cycle of eleven years (see Fig.2. and Lande,1990, Davis,1989).*

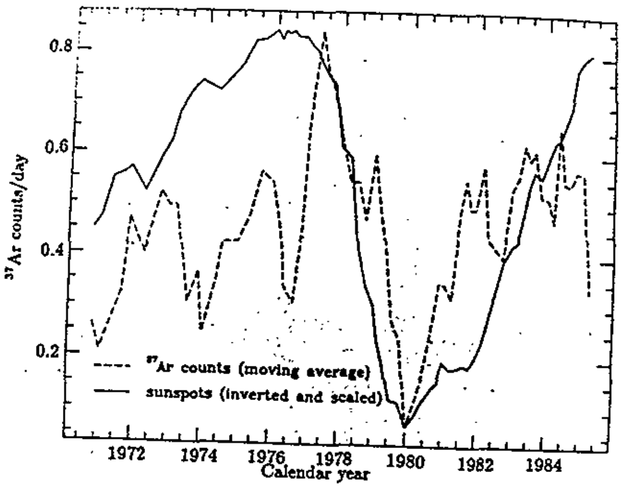


Fig.2.
The anticorrelation of the solar neutrino flux as measured with the ^{37}Ar counts with the sunspot number

It presents a fundamental problem with the standard solar models. And if we do not understand our Sun, what can we think about our understanding of the much less known other stars?

One possibility to explain the neutrino discrepancy is to assume that the Sun does not work as a regulated fusion reactor and the observed anticorrelation of the solar neutrino flux with the solar activity indices reflects real changes at the solar core in the energy producing reactions with the 11 years cycle. Recently I made a suggestion, that (e.g.tidal) flows at the solar core with an amplitude of 1 - 10^3 cm/s can produce in a local magnetic field assumed to be present with a strength of 10^3 Gauss can generate local thermonuclear runaways which makes the energy production to be unbalanced (Grandpierre,1990a,b). Öpik (1972) has shown that the gravitational effects of the planets Jupiter, Earth, Venus and Saturn generate a tidal flow in the Sun with an amplitude of 100 cm/s. In regarding the magnetic fields, it is usual to assume much larger strength at the solar core and the 10 Gauss value seems to be quite plausible. The mechanism works in the following way. The flows mean that electrically charged particles of the ionised solar plasma are moving in a magnetic field. This means that a dynamo effect is at work and so the flows generate electric fields locally. The electric fields are accelerating the particles to a large speed during which they suffer continuously many collisions so they heat the plasma to a very high temperature locally, because the nuclear reactions are extremely sensitive to the temperature, and because of a positive feedback (the larger the temperature, the larger the rate of the energy production, which rises the more the temperature etc.). In this way from time to time a local thermonuclear runaway develops which then generates exlosive cells. These cells are hotter then their surroundings, so they are accelerated by the buoyant force outwards towards the surface. They can reach the surface where they may produce secondary explosions which lead to flare phenomena what is the most spectacular phenomena of solar and stellar activity.

It is interesting to note here that Wood(1972) calculated the planetary tides exerted to the Sun and he found that the tides run in close correlation with the sunspot numbers for two hundred years having a cycle of 11.2 years (Figs.3,4.). With this mechanism I was able to interpret the dependence of stellar activity cycles on rotation rates(Grandpierre,1990b).

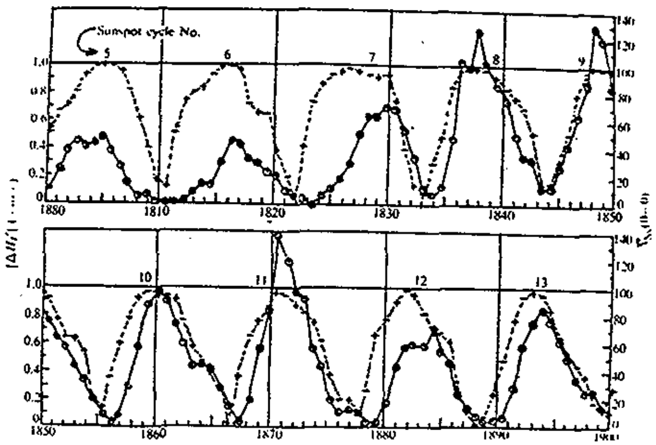


Fig.3.
Tidal height fluctuations (dashed line) and sunspot numbers(solid line)

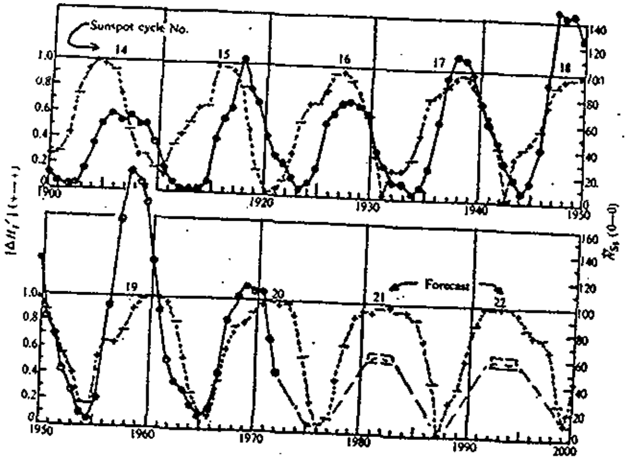


Fig. 4.

The most significant character of the Hertzsprung-Russell diagram is that the stars are not populating it homogeneously but they are definitely concentrated in it. As one can see in Fig.1., most of the stars are on the main sequence. This means that the stars are spending most of their lifetimes in the state when they are using the proton-proton cycle and the CNO cycle to maintain their states and radiating in a constant level. When a star is contracting into the main sequence, it exhausts its hydrogen fuel the more quickly the larger its mass. In a globular cluster, which contains roughly similar aged stars, we will expect that the stars with larger mass will evolve sooner off the main sequence. That is exactly what we observe in measuring the globular clusters with different ages (see Fig.5.). The older the globular cluster, the lower the temperature (smaller the mass) where the stars are turning off the main sequence.

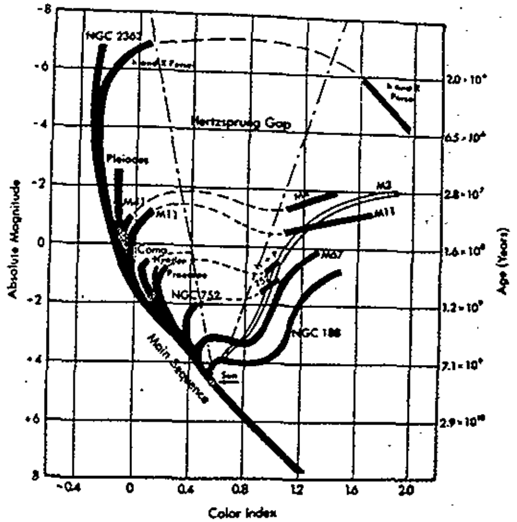
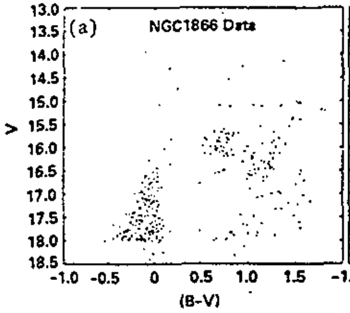


Fig.5.
Color-magnitude diagram of globular clusters

So it seems that for older stars and clusters the contraction hypothesis works satisfactorily. Now let us look after the young clusters. In Figs. 6,7 (from Becker et al.,1984) we can see *synthetic*, calculated temperature luminosity diagrams and compare them with the observed ones. While the agreement is good for the main sequence and giant branch stars, we do not see points above the main sequence with low luminosity and low surface temperature in a number what occurs in the observed diagram. We can *speculate, that* the more point density we see on the diagram when the more concentrated energy supply is working in the star. So the contraction hypothesis can miss some energy production mechanism.



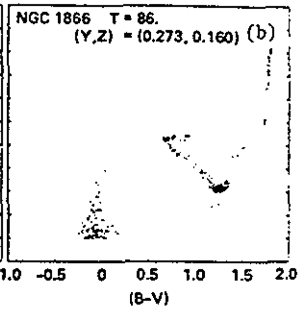


Fig.6.

Observed color-magnitude diagram of NGC 1866

Fig.7.

Calculated synthetic color-magnitude diagram of NGC 1866

Now let us see an interpretive diagram for the young stellar clusters (Fig.8., from Cohen and Kuhi,1979). Plotted are the evolutionary tracks for stars with different masses (expressed in terms of solar masses). In these the stars begin to evolve from up towards the main sequence. First the tracks are almost vertical,which means that the stars are gaining their energies in these stages mainly from gravitational contraction. Stellar models are telling us that in these stages the stars are completely convective, i.e. buoyant forces are mixing the material through the stars. For larger masses, these vertical lines turns after a time into almost horizontal. These horizontal tracks belongs to a radiative state of stellar interiors when the convective flows are ceasing at the stellar cores. Plotted are the isochrones which show the times necessary to reach the point on the tracks with contraction. The contractive process can be followed with the help of the constant stellar radius lines(dashed).

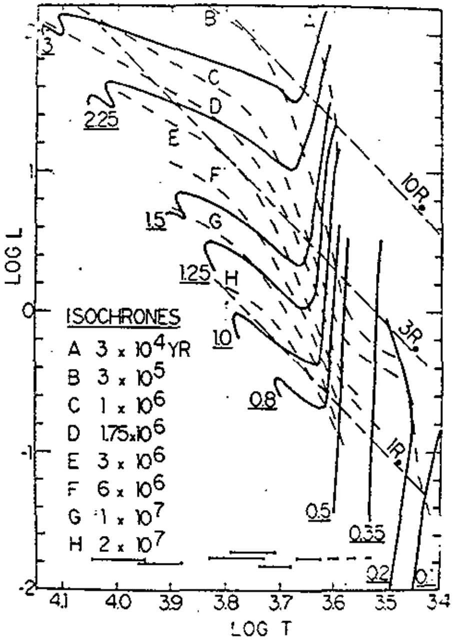


Fig.8.
Interpretive diagram of young stellar clusters

With the help of Fig. 8. it is easy to interpret now the Hertzsprung-Russell diagrams(HRDs) of the young stellar clusters. First let we see the HRD of the few million years old NGC 2264 cluster on Fig.9.(Walker,1956). One can recognize that the cluster is so young,that the stars with spectral types later then B9 (log T 4.1) had not the occasion during their lives to reach the main sequence (the smaller the mass of a star, the later its spectral type, the slower its evolution). This means that the turn on to the main sequence did not happened for all the stars. What is also very interesting is that some stars are below the main sequence. By the contraction hypothesis this fact cannot got any interpretation. The area below the main sequence is a forbidden region for the young stars by this hypothesis.

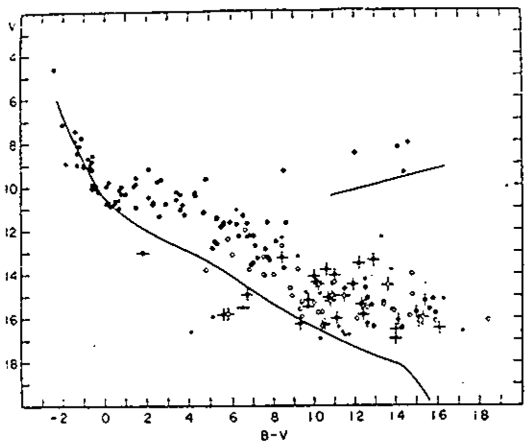


Fig.9.
Color-magnitude diagram of NGC 2264(observed)

A more recent diagram for the same young cluster for more faint stars are made by Adams et al.,1983 (Fig. 10). What we can see here is the shocking fact that a significant amount of the faint stars are below the main sequence! It is characteristic to the popularity of the contraction hypothesis within the astronomers that the authors of this work constructing this diagram even did not try to interpret this fact.

The thing is that the two diagram is easily be united in our mind (I will do it in an other paper). Then it is clear that the tendency of the stars towards later spectral types (lower surface temperatures) to getting closer to the main sequence is really continued at the fainter stars so that around log T =3.60 already a large percent of stars are below the main sequence(BMS). It is really surprising that the tendency of the lower temperature stars to be more far away from the main sequence, which is also prescribed by the contraction hypothesis, is completed with an opposite tendency which broadens the main sequence into a fan-shaped area.

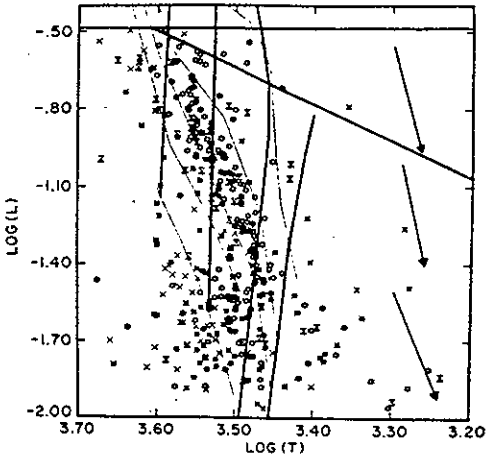


Fig.10.
Observed Hertzsprung-Russell diagram of the faint stars in NGC 2264

This characteristic of the lower main sequence is also shown in Fig.11. for the faint flare stars of the few hundred million years old Pleiades young cluster (Parsamyan, Oganyan 1989) in their spectrum-luminosity diagram. It seems that from their 333 stars more than the half is below the main sequence.

One of the most young stellar clusters is the IC 1396 in the Orion which is a few million years old only. The recent measurements of 276 hydrogen emission stars (which are the ones thought widely to be pre-main-sequence PMS stars) is shown in the luminosity-color diagram of Kun(1990),see Fig.12.. This diagram is practically the same as the HRD. As one can see, there are a lot of medium-mass stars with a color B-V=(-0.5)-(+0.5) (which is an equivalent of a spectral type B5-G0 or a temperature 30 000 K - 5 900 K) below the main sequence. One can speculate that as the stars become older, they lose mass enough to turn into a later spectral type while still preserving their BMS state (Fig.12.).

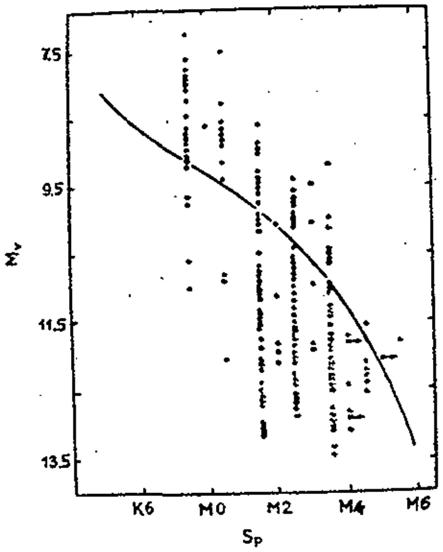


Fig.11.
Spectrum-luminosity diagram of the 333 faint flare stars in the Pleiades cluster

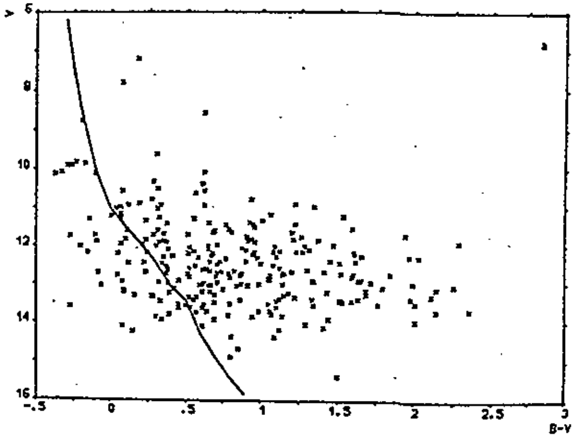


Fig.12.
Color-magnitude diagram of the H-alpha emission stars in the Orion young cluster

When looking after a HRD for larger mass young stars, we can see the diagram of Finkenzeller and Mundt(1974) ,Fig.13. It is very strange that while at the old stellar aggregates and clusters we do not see any objects below the main sequence, just at the young stars, which should have to be high above the main sequence, do we see objects to be BMS stars. There are some efforts made to interpret these strange diagrams with observational errors, selectional effects or by abnormal circumstellar absorption. Nevertheless, we have to keep in our mind, that Adams et al.(1983) proved that any abnormal circumstellar absorption is measured to be play a role orders of magnitudes less than 0.1 percent. From the work of Parsamyan(1986) it is clear that foreground stars, observational errors and spectral type uncertainties while are playing a role, their sum effect is significantly less than the observed effect.

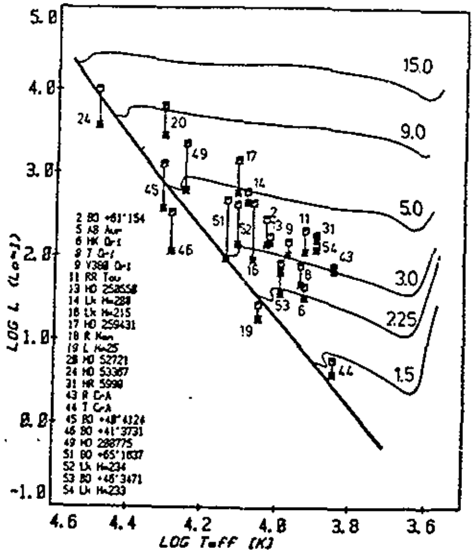


Fig.13.
Hertzsprung-Russell diagram for Herbig Ae/Be stars

It seems that we have enough observational basis to look after some other conception of stellar birth and early evolution. Already in 1954 Ambartsumian developed an alternative idea for stellar origin trying to interpret the observations showing that the expansion processes are dominating at the formation of stars. It is still true that the contraction hypotheses did not got any direct observational support. At the young stars by the contraction hypotheses we should have to observe generally mass infall to the stars. What is observed, is the opposite. All the young stars are showing signs of mass outflows, without any unambiguous sign of mass inflow. The most young the star,the most pronounced are the signs of mass outflows. Just at the youngest stars the so-called bipolar mass outflows are the most developed. The observations are telling us that the stars are born in associations of early-type OB stars and later-type T Tauri type stars and these associations are expanding (see also Grandpierre,1986).

Ambartsumian suggests that a superdense prestellar material is expanding and its transformation at lower densities to normal stellar material is the process creating the stars. For a long time this hypotheses did not got much attention because the theory was not developed enough in detail to be able to model the star formation process. Nevertheless, recently the Nobel-prize winner Lee and his co-workers developed more concretely the idea that superdense Higgs-bosons in coherent state can form stars which are able to produce energies much more effectively than the ordinary matter, i.e. with almost 100 percent effectivity (Friedberg, Lee, Pang 1987 and Chiu,1990).

Haro(1976) proposed a view by which the youngest stellar objects are the so-called Herbig-Haro objects. He assumed that there exist an evolutionary line from the Herbig-Haro objects from below the main sequence through the Herbig Ae/Be stars or at lower masses trhrough the T Tauri stars to the flare stars. It is interesting to note here that recently Movsesyan and Magakyan(1989) published a paper about a new object with optical bipolar outflow CoKu Tau/1. This object is very faint and its spectrum seems to be similar to the Herbig-Haro objects. This means that this star is deeply below the main sequence on the HRD while at the same time it is very young since it has a strong bipolar outflow. Just these characteristics should have to possess the youngest stellar objects formating by expansion, having unusually high surface temperature (keeping left in the HRD) and being very faint(keeping down from the main sequence).

There are other signs showing that there are two basically different ways of star formation. Some T Tauri type stars are clearly formed in regions far from any nebula, these are the isolated T Tauri stars (De La Reza et al.,1989). It seems to be hardly interpreted the origin of stars with contraction of a nebula without signs of a nebula which are normally easily observed around a star.

It is clear since a long time that some basical difficulties are present for the contraction hypotheses to interpret the HRD evolutionary tracks of young stars. There are some age-parameters like the variability, the Li-abundance, the nebulosity, the infrered excesses of the circumstellar dust, the presence of bipolar outflows, the hydrogen emission line strength which when shown in the HRD of young clusters, does not show a clear tendency to correlate with the age derived from the position of the stars in the HRD (Walter et al.,1988,see Fig.14.). In this diagram the T letters represents T Tauri stars with early activity characteristics and the filled circles are for the so-called naked T Tauri stars which are much less active by the above age-parameters. Nevertheless, it is a significant overlapping on the HRD for these two different kinds of objects.

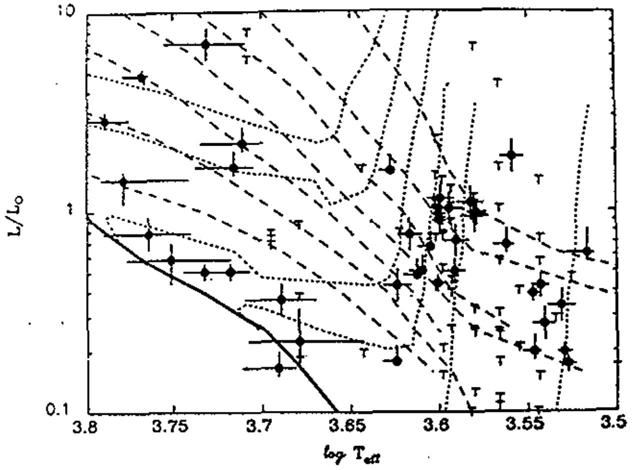


Fig.14.
Hertzprung-Russell diagram of the pre-main-sequence stars in the Taurus-Auriga young cluster

On Fig. 15. it is presented the decay of chromospheric emission with stellar age. The crosses stand for main-sequence solar-mass stars, the letters T mark T Tauri stars, the asterisks represent naked T Tauri stars in regarding their Mg II h + k fluxes. The open circles stand for naked T Tauri stars, the filled circles for solar-mass stars by their Ca II H + K surface fluxes. The solid line is the exponential-decay solution for Mg II from Simon, Herbig and Boesgaard. The classical T Tauri stars are clearly anomalous. I think that this diagram suggest that the T Tauri stars show an activity supplied by an extra source.

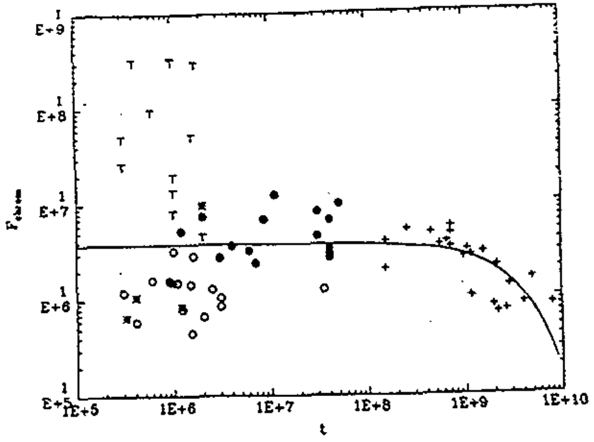


Fig.15.
The dependence of the strength of chromospheric activity of young and main sequence stars

Appenzeller(1983) remarked that the activity of T Tauri stars may be 10^4-10^6 times stronger than in the Sun relative to the photospheric continuum. Because the chromospheric emission in the Sun is 6.77×10^{-5} part of the photospheric value(see e.g. Athay,1976), this means that the chromospheric emission of the T Tauri stars may be ten times stronger then the photospheric continuum for these stars. I suggest that in this case the T Tauri stars behave like a continuously erupting instable objects rather than possess with a stable photosphere. At T Tauri stars the larger part of energy may liberate in form of flares which makes it a hard task to locate them in the HRD. If one count the excess energy of the flares as radiated from a normal photosphere, it is clear that we locate these stars more high above the main sequence than they as stars are. This means again; that some T Tauri stars are much closer or below the main sequence.

By my hypothesis of tidally triggered stellar activity, the level of activity is strongly influenced by planetary tides. Because the planet formation is going on a time scale 10^5-10^6 years, one can expect the rise of the planetary induced flare activity with this shift in stellar ages. There are observations showing this effect. Herbig, Vrba and Rydgren(1986) made a suggestion, that e.g. the T Tauri star FQ Tau is only slightly above its presumed point on the main sequence, yet it has a very strong emission spectrum with a line width of hydrogen alpha 114 Å. The similar spectral type Pleiades stars are showing only 3 Å in comparison.

I suggest that in completing the classical contraction hypothesis it is in work at stellar formation an alternative expansion-born process. During the evolution these expanding objects can quickly or slowly intersect the main sequence from below, depending on their masses. Their activity characteristics are also influenced by the planetary tides effecting on them when they turn to be 10^5 - 10^6 years old. In this way the confusion on relating the age characteristics with the localization on the HRD can be dismissed with the careful analysis of the individual objects.

If it is in work a new type of energy producing mechanism, the position anomalies can give us some physical insight into its physics. All the young stars below the main sequence are extremely interesting objects to reveal this new physics. It is highly promising to separate the two class of young objects in the HRD. In making this, we would be able to derive synthetic HRDs for the contracting and expanding objects separately. With the help of these HRDs showing the localization of the two classes of objects we could derive separately their evolutionary tracks from observations in comparing the different young clusters diagrams.

It is worthwile to note that the existence of the fan-shaped area at young clusters gives us a little physical insight of the nature of the stars located below the main sequence. It is well known, that metal-poor dwarfs, the so-called subdwarfs populates the same area in very old globular clusters. These subdwarfs possess very high space velocity characteristic of Population II stars which is a decisive sign of their old ages. The BMS objects found in young stars does not have this high space velocity, consistently with the age of its parent clusters. In this way we can think that these dwarf stars are originating in a process which produces no metals, just as in the case of the creation events (see Arp, Burbidge,Hoyle,Narlikar,Wickremasinghe,1990). I can suggest to interpret the fanshape in a way that expansion-born stars with larger initial masses can pull on more metal-rich material from their environment by their larger gravitation in their earliest stages when they are very hot but very faint. The larger the mass of the star, the more material it is attracted to the surface of the star, the more close it will be to the main sequence.

It seems to be an urge need of efforts constructing observational HRDs for young clusters and calculate synthetic HRDs with a special care for unusual objects. The evolution of very young stellar objects promises new and challenging excitements by predicting that basically new physical processes are at work and we can observe and understand their nature.

References.

Adams,M.T.,Strom,K.M. and Strom,S.E.1983,Astrophys.J.Suppl.Ser. Vol.53,893.

Ambartsumian,V.A.1958, in "La Structure de l'Univers" (ed. Stoops, R.), Solvay Conference Reports, Bruxelles.

Appenzeller,I.1983,Rev.Mexicana Astron.Astrof.,Vol.7,151.

Arp,H.C.,Burbidge,G.,Hoyle,F.,Narlikar,J.V.and Wickremasinghe,N.C. 1990,Nature Vol.346,807.

Athay,R.G.1976,in "The Solar Chromosphere and Corona :Quiet Sun", Astrophys. and Space Sci. Libr. Vol.53,D.Reidel,Dordrecht,423.

Becker,S.A.,Mathews,G.J.and Brunish,W.M.1984,in"Observational Tests of the Stellar Evolution Theory",(ed.Renzini,A.),83.

Bethe,H.A.1939,Phys.Rev.Vol.55,434.

Chiu,H.-Y.1990,Astrophys.J.Vol.354,302.

Cohen,M. and Kuhi,L.V.1979,Astrophs.J.Suppl.Ser. Vol.41,743.

Davis,R.,Lande,K.,Lee,C.K.,Cleveland,B.T. and Ullman,J.1990,in "Inside the Sun", Astrophys. Space Sci.Libr.Vol.159,(eds. Berthomieu,G. and Cribier,M.),171.

De La Reza,R.,Torres,C.A.O.,Quast,G.,Castilho,B.V.and Vieira,G.L. 1989, Astrophys.J.Letters,Vol.343,L61.

Finkenzeller,U. and Mundt,R.1984,Astr.Ap.Suppl.,Vol.55,109.

Friedberg,R.,Lee,T.D. and Pang,Y.1987,Phys.Rev.D,Vol.35,3640.

Grandpierre,A.1986,Acta Philosophica,Tomus 13,189.(in Hungarian).

Grandpierre,A.1990a,Sol.Phys.(in press).

Grandpierre,A.1990b,Mem.Soc.Astron.Ital.Vol.61.(in press).

Haro,G.1976,Bol.Inst.Tonantzintla,Vol.2,3.

Herbig,G.H.,Vrba,F.J.and Rydgren,A.E.1986,Astron.J.91,575.

Lande,K.1990,cited in Science News,Vol.137,245.

Kun,M.1990,Ph.D.Thesis,Konkoly Observatory.

Movsesyan,T.A.and Magakyan,T.Yu.1989,Astrofizika,Tom.31,Vyp.3,501.

Opik,E.1972,Irish Astron.J.Vol.10,298.

Parsamyan,E.S.1986,in "Star Clusters and Associations",(eds. Balazs,B. and Szecsenyi-Nagy,G.),Budapest,115.

Parsamyan,E.S.and Oganyan,G.B.1989,Astrofizika,Tom.30,220.

Salpeter,E.E.1952,Phys.Rev.88,547.

Stahler,S.W.1983,Astrphys.J.Vol.274,822.

Walker,M.F.1956,Astrophys.J.Suppl.Ser.Vol.2,365.

Walter,F.M.,Brown,A.,Mathieu,R.D.,Myers,P.C. and Vrba,F.J.1988, Astron.J.Vol.96,297.

Wood,K.D.1972,Nature 240,91.

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

COSMIC ABUNDANCES

S. BARCZA
Konkoly Observatory
1525 BUDAPEST Pf 67
HUNGARY

Received on 17th Oct., 1990

The roots of the idea for universal abundances of the chemical elements go back to the thirties or forties of our century when fine spectral analysis of stellar spectra became available. With increasing amount of material surprising uniformities were found: the abundances of the different chemical elements were the same in the photospheres of the Sun or in remote stars, interstellar clouds or the celestial objects. The differences in the abundances did not exceed the error of their determination.

The most precise method to determine cosmic abundances is a spectral fine analysis which is described in many textbooks [1], [2]. The essence of the procedure is that we have to construct the model of the radiating region of the celestial body (distribution of temperature, pressure, velocity, turbulence, chemical composition etc.), we have to know the relevant atomphysical data of the radiating species (spectral line strengths, population of the atomic energy levels, line broadening). With this knowledge we compute the frequency distribution of the emerging radiation and we compare it with the observed spectra. The parameters of the radiating region must be varied until the best coincidence of the observed and computed spectra is found. *One* of the parameters is the chemical abundance of the elements, thus, we can determine them. This procedure is complicated, because of its uncertainties the abundances can be determined in the most favourable cases by an accuracy of factor 2-3. The photosphere of a star is hot, we can determine abundances of elements since the molecules are destroyed except for some stable molecules in cool stellar atmospheres.

The abundances of the light elements from H to Li are surprisingly uniform in very different celestial bodies and the same is true for the heavier elements. The simple picture of mixing the observed cosmic abundances from two uniform mixtures of light and heavy elements is very close to the reality. We have, however, to mention the numerous exceptions to this simple picture.

The following theoretical background can explain this simple picture. The abundances of the light elements from H to Li depend critically on the early evolution of the universe, it is now generally accepted that D, He, Li were produced in the big bang while the heavier elements were produced in stellar interiors and they were expelled in the interstellar space by supernova explosions of massive stars in the early phases of the evolution of galaxies. These are the main lines of our knowledge on the creation of chemical elements. The numerous exceptions in the observed abundances must be explained individually. The explanations are usually acceptable, thus our simple picture on the creation of the chemical elements has survived the challenge by the observations which indicate very peculiar chemical compositions in the atmosphere of some stars.

Finally we give a typical composition for stars of Population I, the abundances of some chemical elements in the photosphere of the Sun [2]. The abundance is the logarithm of the number density of the given species in a volume which contains 1012 hydrogen atoms. Population I indicates that the stars belong to the younger stellar population of our Galaxy.

Chemical abundances of some elements in the solar photosphere [3]

H	He	C	N	O	F	Ne	Na	Mg	Al	Si	P	S
12.0	?	8.5	7.9	8.8	4.6	7.6	6.3	7.6	6.4	7.7	5.4	7.2

References

[1] A. Unsöld, Physik der Sternatmosphären, Springer, Berlin, 1968
[2] D. Mihalas, Stellar Atmospheres, W. H. reeman, San Francisco, 1978
[3] A. Unsöld, B. Baschek, Der neue Kosmos, Springer, Berlin, 1981

SOLAR SYSTEM EVOLUTION:
CRYSTALS AND PLANETARY BODIES
/A REVIEW/

KFKI-1990-50/C B. Lukács & al. (eds.): Evolution: from Cosmogenesis to Biogenesis

Szaniszló Bérczi
Eötvös Loránd University
Department of General Technics, Budapest,
H-1088 Rákóczi ut 5. Hungary

Received on 10th Oct., 1990

The path of the evolution of matter can be described - as a first approximation - by the series of stable structures /Bérczi, 1978./. This series represents a hierarchy of embedding. The series can be divided to two parts, not only from antropocentric point of view: to a cosmic series and to a microworld series. /Stable structures are results of equivalency-classifications on structural levels: they are the representatives of the structural levels.//Fig. 1./

Solar System, which belong to a star, i.e. the Sun, can be divided to two basic types of subsystems in this stable structure series. Description of the evolution of the Solar System by these two structural units: by crystals and planetary bodies will be in focus in our paper. These units make it possible to give two independent models: one which describes chemical structure and differentiations, and one which describes the "gravitational crystallization" in the size-frequency spectrum of preplanetary bodies./Table.1./

Three principles form the basic steps in this review:

1. Reduction: which is used to correspond the main steps of evolution to two stable structures of the Solar System: to crystals and to planetary bodies;
2. Differentiation of a set: the set of the two basic stable structures will be considered with common origin, and evolutionary history will be shown as global /and local/ differentiation events followed on spectra of structures;
3. Layered planetary bodies: they can be divided to belts of

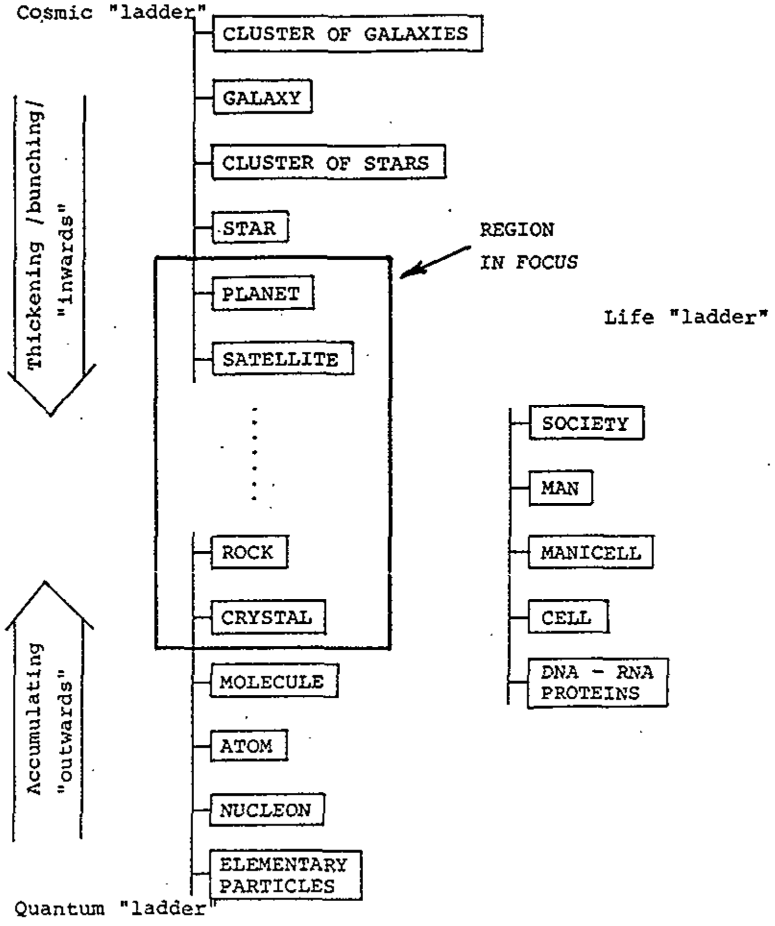


Fig.1. The hierarchy of "stable structures"
The concept "stable structure" refers a relative stability of the structures during decomposition or building. Layer-representatives of this "stable structure" hierarchy form multisets which later undergo on a crystallization period which produce the following layer of "stable Structure" /Bérczi, 1978./.

Three kinds of multiple-particle system-approaches in the Solar System Evolution Models

Authors	Multiset considered	Parameter distinguished	Regardless of	Phenomena which preserved the events
Kepler Newton	orbiting mass-points	orbital elements	a. chemic. compos. b. size-frequency	remnant orbits of celestial bodies in the Solar System
Larimer Lewis- -Barshay Grossman	orbiting minerals	chemical composition	a. orbital element b. size-frequency	composition of minerals in smaller and larger bodies of the Solar System
Hartmann Wilhelms Wood Chapman	orbiting bodies and "particles"	spectrum of size-frequency of "particles"	a. orbital element b. chemic. compos.	crater statistics of surface layers of different age from planetary bodies

Table.1. The "parameter distinguished" column shows those regions where reductions over the "regardless of" column were carried out in model-building.

different material phases /and their circulation/. Different layers, especially those on the surface, preserved prints of events from inside and from outside. Evolution of Solar System during planetary ages can be partly deciphered from the correspondence of inner planetary transformations /in the form of new layers/ to the outer deformations with interplanetary origin.

REDUCTION

Reduction is classification of different phenomena according to an essential, common parameter. This parameter collects together the phenomena, which were distinct and disjunct earlier. Table 1. shows three different reductions in description of Solar System. The first one is classical: it is according to the Keplerian-Newtonian mass-point concept. The other two are according to mineralogy /or chemical composition/ and size-spectrum: these reductions are the results of the last 20 years of studies /among others/ of planetary and meteoritic materials with petrological, chemical and optical methods /i.e. IR spectra/ on one hand; and interplanetary particle spectra and planetary surface crater frequencies and their correspondence on the other hand. Both mineralogy and size-spectrum of Solar System will be considered as two independent variables, which determined two independent periods - one after the other - in Solar System evolution. This independence can be explained by the equal amount of bonding energy of an Earth-like planetary body both from mineral-chemical and from gravitational work of interactions during "chemical-mineraligical" and "gravitational" precipitation and accretion;

Estimation of the chemical bonding energy of minerals of an Earth-like body:

$$E_{\text{chem.}} = \frac{\text{mass of the Earth}}{\text{mean gram-equivalent of Earth}} \times \text{Avogadro number} \times$$

$$\times \text{ mean bonding energy of one "mean molecule"}$$

The values are as follows: Mass of the Earth 6.10^{24}kilogramm, let the mean gram-equivalent of the Earth's material equal to 30 gramm, Avogadro number is 6.10^{23}, and let the mean bonding energy of one "mean molecule" equal to 5 electronvolts.

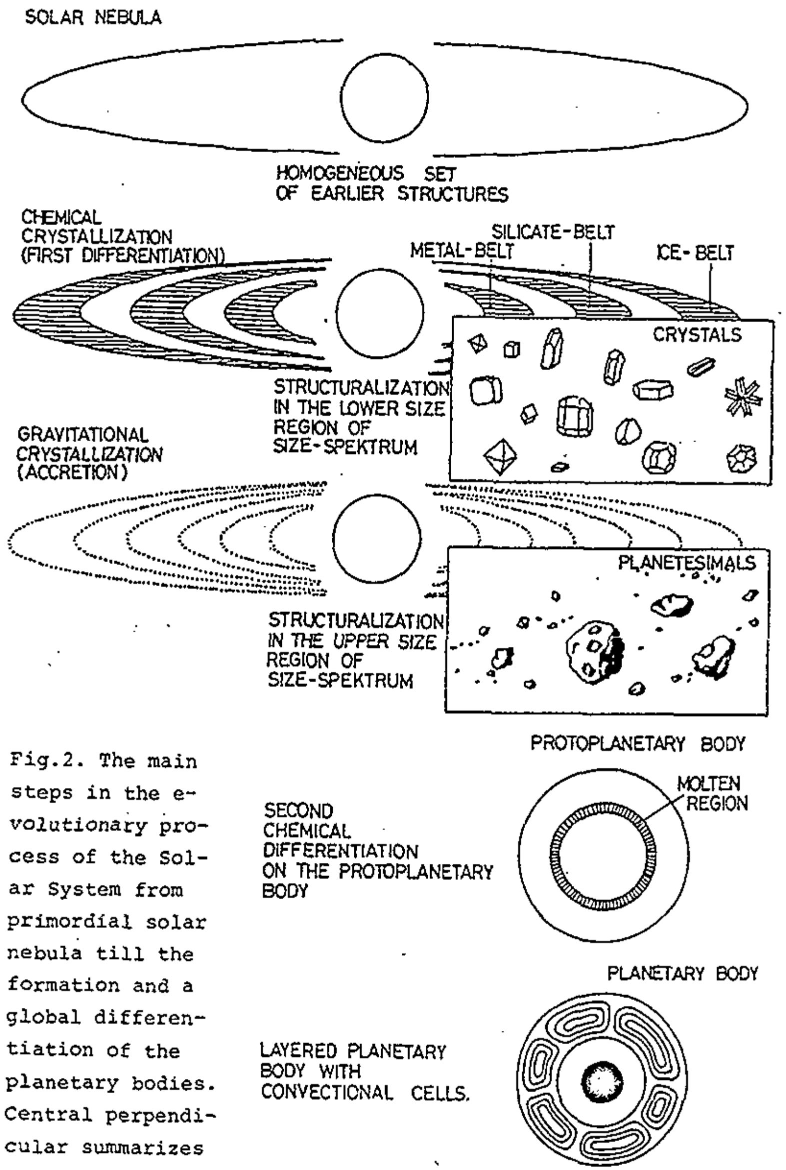


Fig.2. The main steps in the evolutionary process of the Solar System from primordial solar nebula till the formation and a global differentiation of the planetary bodies. Central perpendicular summarizes the events on the system level of Solar System /the three stages with solar nebula/ and the right side perpendicular region summarizes events referring planet formation: a planetary evolutionary point of view focused in the next paper /Illés, E.,this volume/.

With the given data the chemical energy of the Earth is:

$$E_{chem.} = 6.10^{50}\ \text{eV} \approx 10^{32}\ \text{joule.}$$

The gravitational bonding energy is:

$$E_{grav.} = \frac{1}{2}\, G\, \frac{M^2}{R}$$

where M is the mass of the Earth, R is the radius of the Earth, and G is the gravitational constant. With the given data the gravitational energy of the Earth is:

$$E_{grav.} = 10^{32}\ \text{joule.}$$

Within the error of the estimations the two bonding energy types are equal.Therefore we show the evolution of the Solar System in two, independent episodes: chemical crystallization and "gravitational crystallization"./Bérczi, 1978. Fig.2./

DIFFERENTIATION

Mechanical models of stellar accretion deduce the initial conditions of the two later "precipitating" periods. The conservation of the moment of momentum in the contracting stellar nebula constraints outer equatorial layers to remain left out from further accretion when solar nebula contracted to the measure of Neptun's orbit /Alfvén, 1970./ This nebula left out from the further accretion is the object of the crystallization processes to be shown.

CHEMICAL CRYSTALLIZATION

Model of Lewis and Barshay /1975/ describes the series of minerals which precipitate from the Solar Nebula with cosmic /solar/ elementary abundances. In their summarizing map, on a p-T diagram of solar gas / solid phase borders they can deduce precipitating minerals which are in equilibrium with the slowly cooling gas nebula. The map /Fig.3./ shows that cooling temperature /with the distance from the Sun/ differentiates the nebula. This temperature-differentiation forms chemical belts around the sun, which can be characterized by a dominating mineral phase. There is a great three part division of precipitating belts: metal - silicate - ice./Fig.2./. The inner belts are: refractory minerals, metal Ni-Fe, and the three main rock forming silicates: feldspar, pyroxene and olivine. Troilite divides this hot region from the outer belts, where H_2O begins to play a dominant role. First H_2O trans-

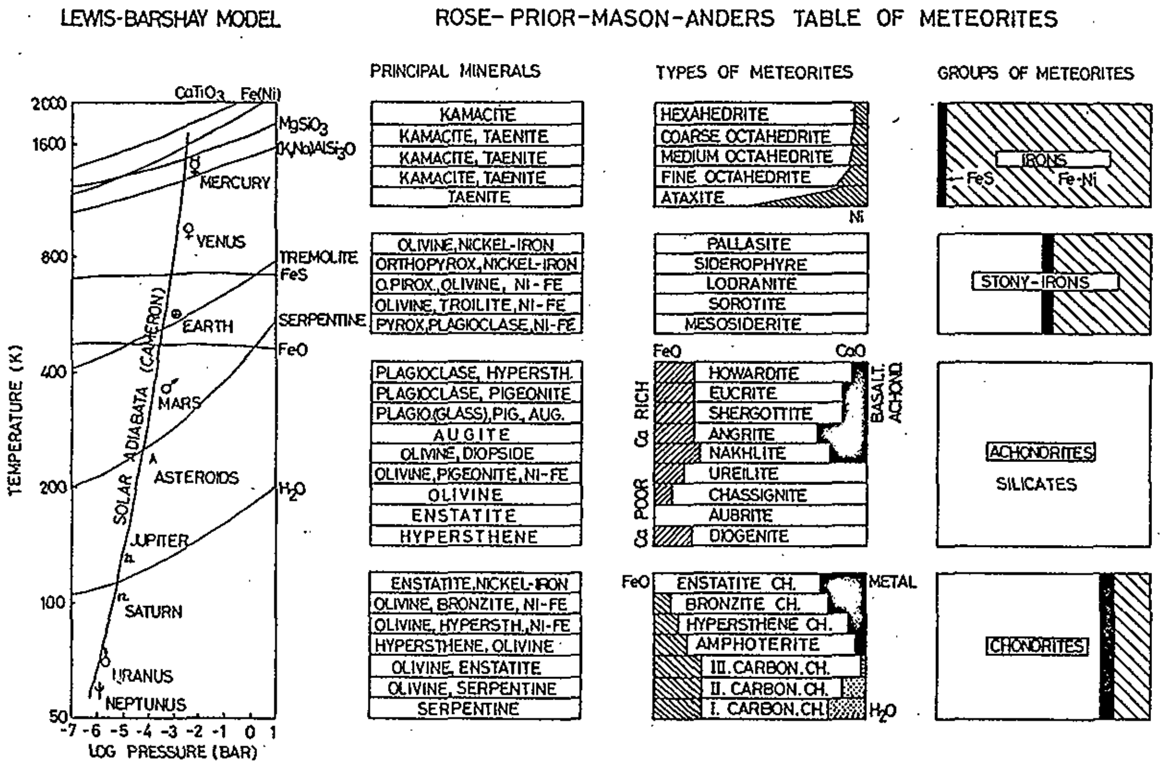


Fig.3. Summary of the chemical crystallization in the Solar System according to Lewis and Barshay /1975/ and meteoritic correspondence to these mineral bands./Bérczi, 1991/

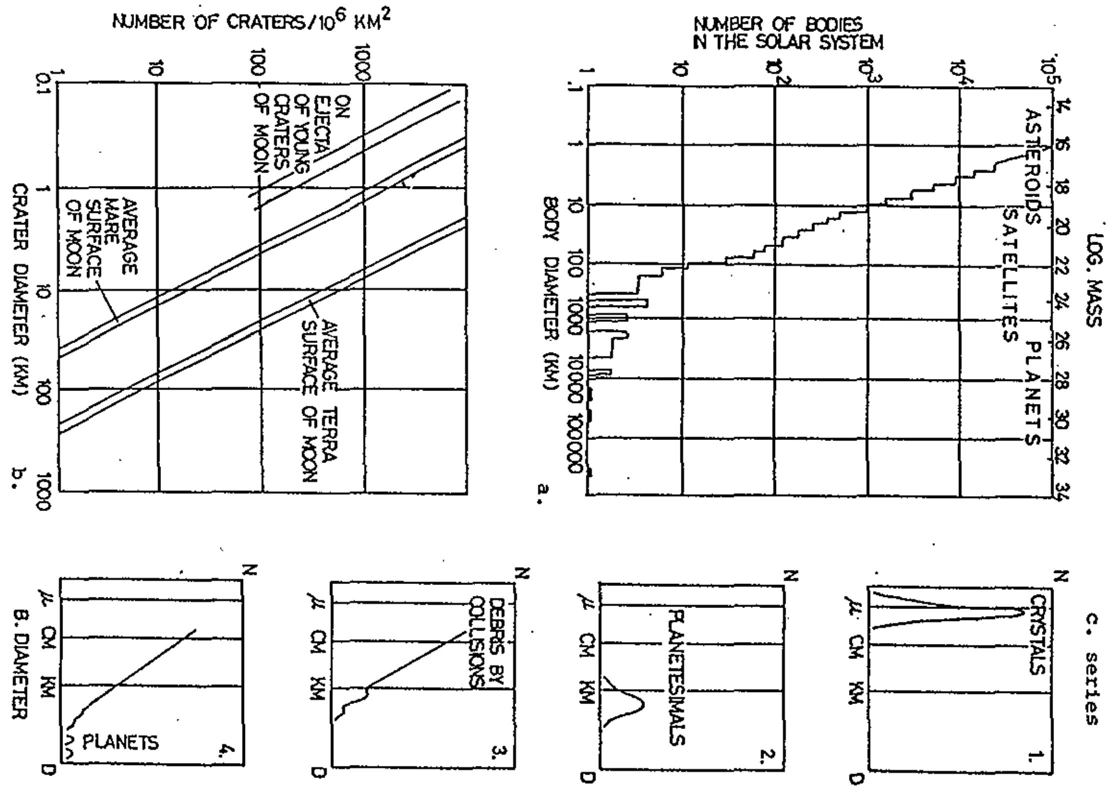


Fig.4. The size-spectrum of the Solar System recently /a./, and its print on lunar surfaces of different ages /b./, and theoretical changes of this size-spectrum during the formation of planetary system around the Sun./a. and c. from Hartmann, 1975./

forms silicates into their hidrated variants /Serpentine and tremolite/ and later crystallizes in ice and forms clathrates with ammonia and methane. The minerals of this precipitating model of Lewis and Barshay can be found in meteorites as shown on the Rose-Prior-Mason-Anders Table part of Fig.3.

"GRAVITATIONAL CRYSTALLIZATION"

The process of accumulation into planetary bodies can be followed on the theoretical size-spectrum of the Solar System /hartmann, 1971./. Precipitating minerals appear on this diagram when they grow to a micrometer size /Fig.4.c.1./ Slow collisions collect them to planetesimal measures /Fig.4.c.2./, which later became fragmented by destructive large-speed collisions of growing planetesimals /Fig.4.c.3./. At least some planetary bodies grow up large enough to survive destructions, but fragmentic part of the spectrum remains a remnant of the accretional period of planetary formation /Fig.4.c.4./ This fragmentation spectrum has been preserved on different planetary surface layers in the form of crater frequency of these surfaces. /Fig.4.b./.

Remnants of both periods of "crystallizations" can be found in the smaller bodies of the Solar System. Their recent size-frequency spectrum is shown on Fig.4.a. Their optical and near infrared spectrum was compared in the last two decades: especially those of asteroids and meteorites can give good correspondence /Fig.5./. Asteroid/meteorite spectral similarities proove not only the presence of primordial materials of carbonaceous chondrites in the asteroid belt, but the presence of differentiated asteroidal bodies as well. /Chapman,1975., Gradie, Tedesco, 1982./

LAYERED PLANETARY BODIES

Materials has been transformed after their accretion into large planetary bodies. Accretional and radiactivity heat resulted in partial meltings /at least/ on these bodies and this process rearranged mineral composition of planetary bodies. Density layered belts were formed. But the degree of of heating and differentiation depended on the measure /or equally the mass/ of the planet. Degree of partial melting or remelting can be characterized by the abundance of REE content of planetary rocks. The scale of

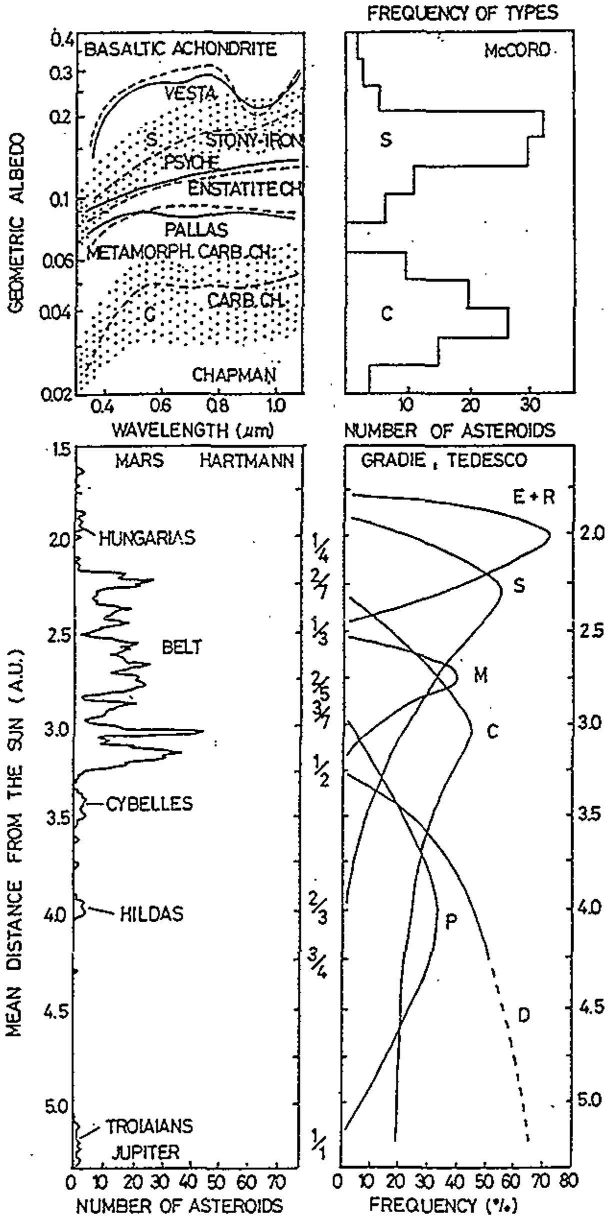


Fig.5. Optical and near-infrared spectral correspondences between peculiar meteorites and asteroids; frequency and distribution of characteristical spectral types in the asteroid belt /Chapman, 1975., McCord, 1975., Hartmann, 1975., Gradie and Tedesco, 1982./

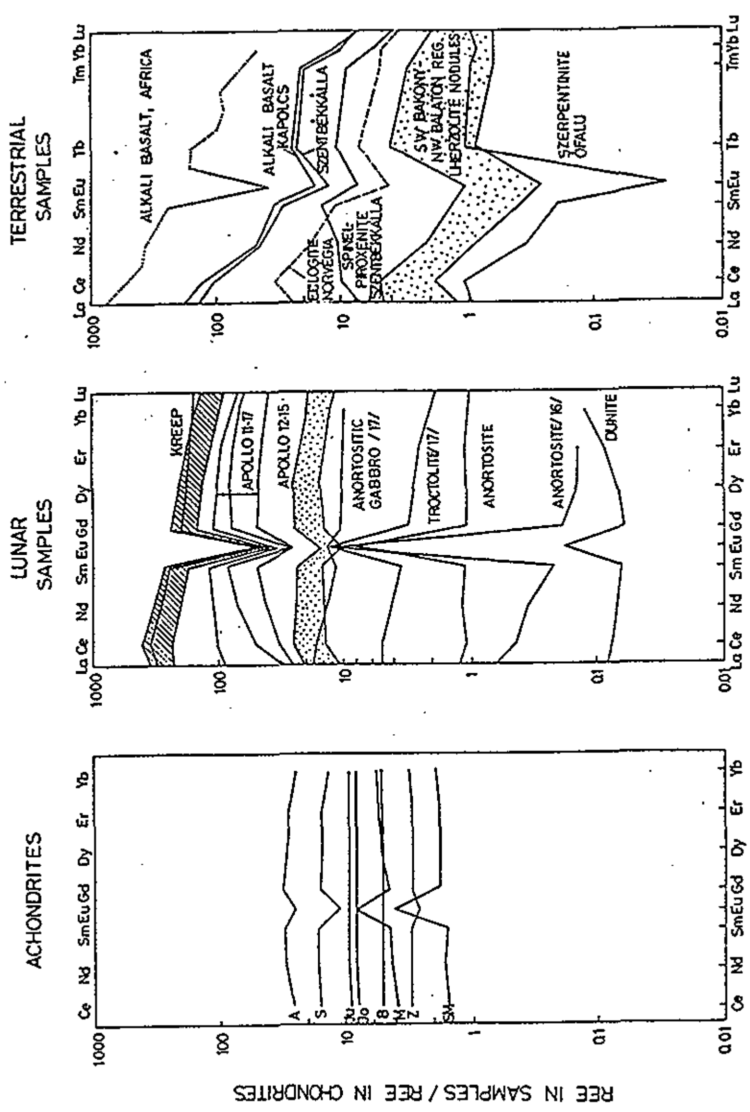


Fig.6. The degree and scale of differentiation depend on the mass of a planetary body. REE abundances of achondrites with supposed asteroidal origin, those of lunar samples and terrestrial samples show well the opening of the range between enrichment and depletion from asteroids to Earth-like planets./a.,b. from Allegré et al,1977. c.Bérczi Sz., Bérczi J., 1986./

separation of REE and their enrichment in other partially melted rocks is the largest on the Earth, smaller on the Moon and the less on the meteorites with asteroidal origin /achondrites/. This is the characteristical feature of the Fig.6. /Allegré et al.1977, Bérczi, Sz., Bérczi, J. 1986./.

This short review intended to focus on the basic principles and the basic events in description of the Solar System structural evolution. The concise historical summary was helped with selected and fitted figure-molecules, which were taken from a new atlas of the evolution of matter under construction. The rich literature of the topic offers more details, as follows:

LITERATURE AND REFERENCES

Adams, J.B., McCord, T.B. /1972/: Electronic spectra of pyroxenes and interpretation of telescopic spectral reflectivity curves of the Moon. Proc. Third Lunar Sci. Conf. pp. 3021-3034.

Alfvén, H., Arrhenius, G. /1976/: Evolution of the Solar System. NASA, Washington. NASA-SP-345.

Allegré, C.J. et al. /1977/: Comparative chemical history of the Earth, the Moon and parent body of achondrite. Phil. Trans. R. Soc. London. A 285. pp. 55-67.

Anders, E. /1964/: Origin, age and composition of meteorites. Space Science Review 3. pp. 583-714.

Anders, E. /1971/: Meteorites and the early Solar System. Ann. Rev. Astron. Astrophys. 9. pp 1-34.

Barshay, S.S., Lewis, J.S. /1975/: Chemistry of solar material. in.: The Dusty Universe /eds.: Field, G.B., Cameron, A.G.W./ Neale Watson Acad. Publ.

Basilevsky, A.T. /1976/: On the evolution rate of small lunar craters. Proc. Seventh Lunar Sci. Conf. pp. 1005-1020.

Bérczi Sz. /1978/: Planetológia /Planetology/. Tankönyvkiadó, Budapest.

Bérczi Sz. /1978/: Cyclicity in the history of matter and its application to the elementary structure-building processes of the Solar System. Lect. on Symp. Cyclicity and Evolution; Publ.: Proceedings of the Symposium. Szádeczky-Kardoss E./ed./ Acta Geol. Acad. Sci. Hung. 23. pp 163-171 /Budapest, 1980/

Bérczi Sz., Bérczi J. /1986/: REE Content in the Szentbékkálla series of peridotite inclusions. Acta Min.Petr. Szeged.28.p.61-74.

Bérczi Sz. */1989/: Rendező elvek a Planetológiában /Ordering principles* in Planetology/ Term. Vil. 120. No.7. pp. 290-298.

Bérczi Sz. /1991/: Kristályoktól Bolygótestekig /From Crystals to Planetary Bodies/ Akadémiai Kiadó, Budapest /in press/.

Bunch, T.E. /1974/: Meteorites: Clues to the origin of the Solar System./in: Greeley, Schultz, eds.: A Primer in Lunar Geology/ pp. 69-88.

Burns, J.A., Matthews, M.S. /eds./ /1986/: Satellites. University of Arizona Press, Tucson.

Cameron, A.G.W. /1974/: Accumulation processes in the primitive solar nebula. Icarus. 18. pp. 407-451.

Chapman, C.R., McCord, T.B., Johnson, T.V. /1973/: Asteroid spectral reflectivities. Astronomical Journal. 78. pp. 126-140.

Chapman, C.R., Salisbury, J.W. /1973/: Comparisons of meteorite and asteroid spectral reflectivities. Icarus. 19. pp 507-522.

Chapman, C.R. /1975/: The nature of asteroids. Sci.Am.232.pp. 24-33.

Feierberg, M.A., Drake, M.J. /1980/: The meteorite-asteroid connection: the infrared spectra of eucrites, shergottites, and Vesta. Science. 209. pp. 805-807.

Gaffey, M.J. /1976/: Spectral reflectance characteristics of the meteorite classes. J. Geophys. Res. 81. pp. 905-920.

Gast, P.W. /1971/: The chemical composition and the structure of the Moon. The Moon. 5.

Gast, P.W. /1973/: Lunar magmatism in time and space. /in: Lunar Science IV. eds. Chamberlain, Wattkins/ Houston

Gault, D.E. et al. /1975/: Some comparisons of impact craters on Mercury and Moon. J.Geophys. Res. 80. pp. 2444-2460.

Gradie, J., Tedesco, E. /1982/: Compositional structure of the asteroid belt. Science. 216. pp. 1405-1407.

Grossmann, L. /1972/: Condensation in the primitive solar nebula. Geochim. Cosmochim. Acta. 36. pp. 597-620.

Harrison, W.J. /1981/: Pertitioning of REE between minerals and coexisting melts during partial melting of a garnet lherzolite. American Mineralogist. 66. pp. 242-259.

Hartmann, W.K., Wood, C.A. /1971/: Moon: origin and evolution of multiring basins. The Moon. 3. pp. 1-55.

Hartmann, W.K. /1976/: The smaller bodies of the Solar System. Sci. Am. 236. pp. 143-159.

Haskin, L.A., Frey, F.A., Schmitt, R.A., Smith, R.H. /1966/: Meteoritic, solar and terrestrial rare-earth distributions. /in: Physics and Chemistry of the Earth. Vol. VII.; eds.: Ahrens, L.H., Press, F., Runcorn, S.K., Urey, H.C./ Pergamon Press, Oxford.

Kerridge, J.F., Matthews, M.S. /1988/: Meteorites and the Early Solar System. University of Arizona Press, Tucson.

Kushiro, I., Mysen, B.O. /1977/: Compositional variations of coexisting phases with degree of melting of peridotite in the upper mantle. American Mineralogist. 62. pp. 843-865.

Larimer, J.W. /1967/: Chemical fractionation in meteorites. I.:Condensation of the elements. Geochim.Cosmochim.Acta. 31. p.1215-1238.

Lewis, J.S. /1972/: Metal/silicate fractionation in the Solar System. Earth Planet. Sci. Lett. 15. pp. 286-290.

Lewis, J.S. /1975/: The chemistry of the Solar System. Sci.Am. 230. pp. 50-65.

Lewis, J.S. /1974/: The temperature gradient in the solar nebula. Science. 186. pp. 440-442.

Mason, B. /1962/: Meteorites. Wiley, New York.

Matson, D.L. et al. /1976/: Asteroids and comparative planetology. Proc. Seventh Lunar Sci. Conf. pp. 3603-3627. Houston.

McCord, T.B., Gaffey, M.J. /1974/: *Asteroids: surface composition* from reflectance spectroscopy. Science. 186. pp. 352-355.

The Moon - A new appraisal from space missions and laboratory analyses. The Royal Society, London, 1977.

Mysen, B.O., Holloway, J.R. /1977/: Experimental determination of rare earth fractionation patterns in partial melts from peridotite in the upper mantle. Earth Planet. Sci. Lett. 34. pp. 231-237.

Nixon, P. /1989/: Mantle Xenoliths. John Wiley and Sons, New York.

Ringwood, A.E. /1975/: Some aspects of the minor element chemistry of lunar mare basalts. The Moon. 12. pp. 127-157.

Szádeczky-Kardoss, E. /1975/: Geonómia. Budapest.

Wilhelms, D.E. /1987/: The Geologic History of the Moon. U.S. Geological Survey Professional Paper No. 1348. Washington D.C.

Yanai, K., Kojima, H. /1987/: Photographic Catalog of the Antarctic Meteorites. National Institute of Polar Research, Japan. Tokyo.

EVOLUTION OF PLANETARY BODIES

E. Illés-Almár
Konkoly Observatory
Budapest, H-1525 Box 67 Hungary

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

Received on 4th Oct., 1990

ABSTRACT

A concised treatment of different conditions influencing the evolution of planetary bodies is given on the basis of the literature — following the way of thinking of the author.

INTRODUCTION

Thanks to the results of space research in the past 30 years, there are at least 25 crusty planetary bodies (instead of one) investigated directly i.e. geophysics could turn into an experimental science. Geophyisical models created to one single body, the Earth, may be tested now through different initial conditions of many others.

Meanwhile it has been discovered that the eternal fields of ice on the Antarctic collect meteorites like a giant telescope. The slow motion of its glaciers integrates them in time and some slopes bring them into focus en masse, hereby having multiplied the number of accessible samples of heavenly material. Moreover the white icecover helps to collect ancient meteorites without any selection effect.

Both new kinds of investigation, the in situ analysis of izotop abundances in particular, made a scientific evaluation of earlier hypotheses in planetary cosmogony possible, rejecting unrealistic ideas and leading to a more or less consistent theory of the evolution of the Solar System.

What are the most important new characteristics of such a theory? First of all accident through collisions played a more important role in the evolution than supposed earlier, not only at the beginning, during the accumulation of small particles, but also later at the end and after the accretion period. The impact of large projectiles gave sometimes rise to the birth of new planetary bodies, sometimes left important marks (i.e. craters) on the surfaces. Collisions influenced considerably the conditions of the origin and evolution of terrestrial life even at a later phase. Besides the more or less continuously acting influence of corpuscular and UV radiation - guaranteeing useful mutations of life-forms - collisions proved to be a genaral and important ongoing phenomenon responsible for the mass extinction of different species. Its significance comes from the fact that in a state-of-the-art planetary cosmogony the effect of catastrophic collisions is a natural consequence of the existing conditions and not an ad-hoc hypothesis as in many previous theories. This new conception needs, however, further verification.

It is evident namely that impact features are present on every planetary body with a more or less stable crust -- independently of its heliocentric distance. The scale varies from mu to a few thousand km. Moreover at the end of the accretion period large impacting projectiles left multiring basins on the surface of many planetary bodies. Consequently we have to suppose that accretion took place through collision not only by condensation and collision was a general phenomenon influencing the evolution even in cases when impact craters are missing.

One of the interesting results of space probes in the outer part of the Solar System is that contrary to the lunar crust the majority of these satellites have an icy composition. This came as a surprise although it was known previously that the density of giant planets decreases with distance. The fact, however, that the satellites follow the same rule (see Fig. 1) led us to the conclusion that the bulk of the satellites accreted mainly from the circumplanetary matter (the composition of which was determined dominantly by the T(r) function) and modified only slightly by projectiles originating from other parts of the Solar System.

Another surprizing observation was the intensity and influence of tidal heating. Traces of activity have been found in every satellite system and even small icy satellites, without such "traditional" heat sources as gravitational contraction or radioactivity, indicated melting periods by their spherical shape. Moreover, the innermost small bodies proved to be the more active in every satellite system, pointing to the fact that the "belt of life" is not necessarily restricted to a narrow heliocentric distance zone, hence life conditions may be more wide-spread than supposed earlier.

Greenhouse effect has also vital importance in planetary evolution. If the partial pressure of CO_2 is significantly larger than in the terrestrial atmosphere then heat-escape is limited also in other wavebands (e.g. runaway green-house effect on Venus). It means also that there is a serious danger burning fossile fuels or even releasing other kinds of latent energy (e.g. nuclear) on the surface of the Earth since it will contribute to the terrestrial heat balance by warming up the atmosphere and releasing CO_2. The only solution of this ecological problem is given by any kind of transformation of the solar energy available on the surface by devices like solar panels, wind-engines or hydroelectric stations.

Finally magnetic fields play also an important role in the evolution. A planet is shielded by its own magnetic field from high-energy particles which represent a serious danger for life in interplanetary space. But strong magnetic fields of a planet may influence significantly the radiation hazard for near-by satellites and electric current of their flux tube may generate "hot spots" on their surface (see the case of Jupiter and Io).

Numbers in parenthesis [n] refer to the literature, in /n/ to the figures, in (n) to remarks.

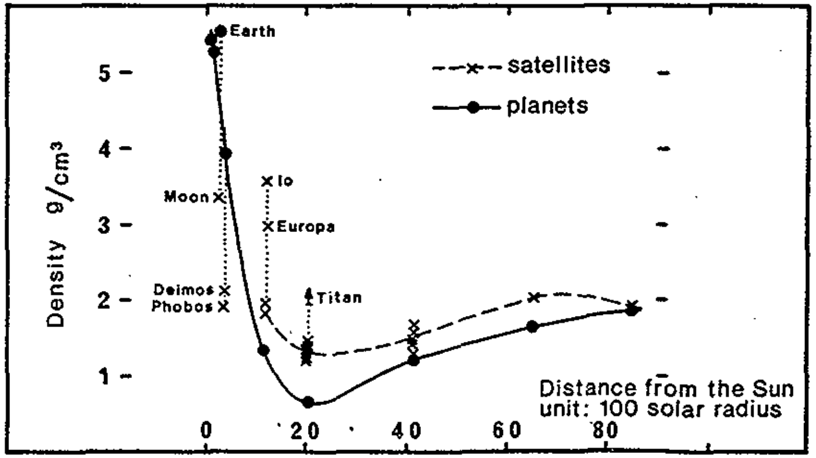


Fig. 1

DEVELOPMENT OF A PLANETARY BODY AFTER ACCRETION:

thermal history + collisional history =
heating/cooling as a function of time

Heat sources for a planetary body

- solar radiation	surface heating,	asymmetric,	continuous
- gravitational separation	volume heating,	symmetric,	long term, decays
- radiogenic heating	volume heating,	symmetric,	long term, decays
- impact heating (1)	surface heating,	asymmetric,	episodic, $\Delta t \sim 0$
- tidal heating (2)	volume heating,		episodic, $\Delta t \gg 0$
- magnetospheric heating (3)	volume heating,	asymmetric,	continuous

Cooling of a planetary body:

- interior to surface by
  - conduction
  - convection → volcanism
- surface
  - radiation
  - sublimation — escape
  - in the presence of an atmosphere by
    - evaporation
    - atmospheric
      - convection
      - radiation (4)
      - escape

Remarks:

(1) Impact energy is transformed partly into
- heat energy → surface melting
- kinetic energy of target ejecta
  - rays/halo around craters /1/
  - escape of material, erosion loss of atmosphere
- seismic energy: makes tectonism to work
  - fracturing around the impact site
  - antipodal fracturing /2/
  - the impact site will be a local centre of activity
- acoustic energy

(2) Tidal heating can be extremely important in regular, interacting satellite systems at times of resonances between evolving orbits.
(Violent ongoing geologic activity on Io /3/.
Ancient geologic activity on Ganymede , Enceladus, Miranda, Ariel.
In the case of Triton the orbit circularization after capture resulted in tidal heating.
Pluto-Charon system is completely coupled → there is no tidal heating now. No trace of geologic activity is expected if there was no orbital evolution during their previous history.
Tidal heating plays some role in the large density and the prolonged geologic activity of the Earth as well.)

(3) If a satellite without an atmosphere is orbiting inside a planetary magnetosphere then the electric current flowing in the magnetic flux tube may enter into the body of the satellite:
- the place of the entrance is heated by the current giving rise to
  - higher temperature of the body (Amalthea by 5°)
  - continuous volcanic activity site (Io [10])
- volumetric electrolysis of water-ice giving rise to the possibility for burning/detonation of electrolytic gas [16,17]

(4) greenhouse effect can slow down the cooling
(Earth [20] , significant on Venus, may be important on Titan)

WHAT WAS THE TEMPERATURE MAXIMUM REACHED BY A PLANETARY BODY?

- not enough for melting: irregular shape } { surface alteration
- only just enough for melting: spherical shape } { by impact only
- more than enough for melting: spherical shape } gravitational separation
  + geological activity }

If melted, convection can be induced
by inhomogenities in composition/temperature
within the core/mantle/liquidosphere/atmosphere

Consequences of the convection within the
- core/mantle: magnetic field
  (core: Mercury, Earth, Mars, Jupiter, Saturn
  mantle: Uranus, Neptune)
- mantle: volcanism/tectonism → outgassing
- liquidosphere/atmosphere:
  - erosion of the crust
  - redistribution of energy
- *atmosphere:*
  - *formation of precipitation*
  - changes in electrical condition of the atmosphere (lightening, recombination)

Diameter limits for melting:
- in the case of radiogenic heating in
  - rocky *bodies: ~ 800* km
  - icy bodies : ~2000 km
- if tidal heating is switched on : can be as low as ~400 km
  (Hyperion has not but Mimas do has a spherical shape /4/)

NUMBER OF PLANETARY BODIES OF THE SOLAR SYSTEM KNOWN UP-TO-NOW

grouped according to the

orbit	orbit and size	composition	phase of its material (temperature)
helio-centric: 9 planets x asteroids x comets	4 giant planets 4 large planets 1 middle-size pl. x small planets (=asteroids) x comets	4 gaseous planets 6 rocky p.b. x rocky debries: asteroids outgassed com. meteoroids	4 gaseous planets 5 crusty p.b. with substantial amount of atmosphere 20 crusty p.b. without atmosphere x debries without atmosphere
planeto-centric 60 sat.	7 large sat. 15 middle-size sat. 38 small sat.	19 icy p.b. x icy debries: comets	

pl.: planet: p.b.: planetary body: sat.: satellite: com.: comet: x: many

CLASSIFICATION OF PLANETARY SATELLITES

planet's name	total number of satellites	number of large sat.	number of middle-size sat.	number of small sat.
Mercury	0	-	-	-
Venus	0	-	-	-
Earth	1	1	-	-
Mars	2	-	-	2
Jupiter	16	4	-	12
Saturn	17	1	7	9
Uranus	15	-	5	10
Neptune	8	1	2	5
Pluto	1	-	1	-

WHAT KIND OF MARKS MAY BE LEFT ON THE SURFACES BY THE DIFFERENT EVENTS?

Heating: expansion in general } except in the case of special material
Cooling: contraction in general} (e.g. water densest at 4°C)
Phase change: expansion or contraction
Impact:

- scars - crater (on every crusty planetary body)
    - centre of fracturing
    - antipodal fracturing (Mercury antipodal to Caloris Basin /2/)
    - crust break-through → volcanism /5/ (maria on the Moon, Triton?, white spots on Umbriel? light? or dark? material on Japetus?)
- rays/halo around craters as target ejecta spread /1/
- too large core (part of Mercury's crust splashed down by a huge impact and escaped[1])
- moon (part of Earth's crust splashed down by a huge impact and a part of it accreted into the Moon [11,24,25]
    after explosion of proto-Pluto by a large-body impact the debries accreted into a binary planet Pluto-Charon [9])
- mascon (Moon)
- blow off the atmosphere (Argyrae /6/ impact on Mars [12,13,22,23])
- implantation of the impactor's material into the atmosphere ("nuclear winter": extinction of living species on Earth)
- which slowly settles down forming an "anomalous layer" on the surface (e.g. iridium-rich layer on the C-T boundary on Earth)
- deposit of flood caused by tsunamies along the continent's margins if the impactor reached the ocean (East coast of Africa, C-T boundary)
- shocked quatz grains if the impactor struck continental crust[14] (North-America, C-T boundary)
- carbon deposit layer settled down after world-wide fire caused by the impact-heat (Earth, C-T boundary)
- acid rain [15]

The effect of the impact may be modified by the condition of the

- impactor: - coming from planetocentric or heliocentric orbit (giving rise to a smaller /7/ or a larger /8/ crater)
    - having an angle of impact
        - ~ perpendicular: - spherical crater (common everywhere)
            - total explosion of the target
        - flat: - butterfly shaped crater /9/ (Mars)
            - escape of intact boulders (SNC meteorits from Mars)
            - splash down of crust: escape of material
- target: - in the presence of atmosphere/liquidosphere
        - the impactor is melted/vaporized/broken/exploded (lack of small craters on Venus [29])
        - the crater erodes
    - in the presence of volatile rich terrain
        - lobate craters /10/ can be created (Mars)
        - and if afterwards the surroundings is flooded by volcanism of magma -- poorer in volatiles -- then negative ringed craters /11/ can be formed by erosion (Mars)
    - the strength of the surface material can influence
        - the existence of the central peak in the crater (Jovian ice-satellites are softer → no central peak.)
        - the relaxation rate of the emerging relief (Curved crater floor on Tethys /12/ ← at time

WHAT KINDS OF DEFORMATION CAN BE PRESENT ON THE SURFACES OF CRUSTY PLANETARY BODIES?

Crust fissures because of change of curvature caused by
- tidal deformation (Europa linear features /13/)
- dome formation above mantle upflow (Earth, East-African graben)
- uplift on places of converging mantle flow if there is no subduction (Venus parquet terrain units [2]/14/)

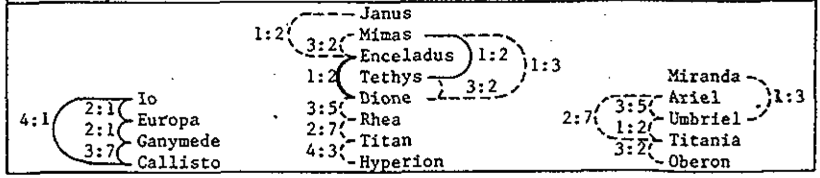
Expansion:
- global: rift valley
    - because of heating
    - because of freezing of the interior of an icy body the crust cracks through owing to volume increase /15/ (Tethys, Titania, Dione?, Ariel?)
- local:
    - rift system because of local heating /16/ (Enceladus, Miranda?)
    - global rift network because of global mantle-circulation /17,18,19/: crust-pieces are spread away by mantle convection (Earth, mid-oceanic ridges: huge polygonal units on Ganymede, Ariel [6], Umbriel [7], Triton)

Compression
- global: drying up (contraction) of a planetary body: thrust faults /20/ (Mercury, Miranda[5])
- local: material towering
    - on sites of converging mantle flows:
        - lift off of crust blocks /21/ (Tibetian Plateau on Earth, Lakshmi Planum on Venus, Miranda?)
        - chains of mountains /22/ (Earth, Venus)
        - continents /23/ (Earth, Venus)
        - subduction /24/ (deep oceanic trenches on Earth)
    - parallel grooves when new crust is formed and older one is thrusted (Ganymede, Enceladus, inside the expansion graben on Ariel, mid-oceanic ridges on Earth /25/)
    - collision of slipping material with stable crust fragments around
        - mountain slopes: grooves around mountains /26/ (Mars, circular grooves around large shield volcanoes)
        - the slopes of bulges (Mars, around Tharsis bulge)
        - the slopes of uplifts formed by converging flows (Venus parquet units and parquet terrain inside a parquet unit [2]/27/)

Mixed: expansion and compression
- transform fault /19/ (Earth, Ganymede, Enceladus, Triton?)
- global plate tectonism /28/ (only on Earth)

Impact may cause some modification of all kinds of deformations.

RESONANCES IN SATELLITE SYSTEMS OF JUPITER, SATURN AND URANUS.
(Recent resonances are represented by heavy lines, the earlier possible ones by dotted lines.)



WHAT LEVEL OF GEOLOGICAL ACTIVITY WAS REALIZED ON THE VARIOUS PLANETARY BODIES?

No traces of activity
- only targets of impacts /29/
  (all the debries-satellites, Mimas, Hyperion)
- albedo features:frost along cracks /30/
  (trailing side of Dione and Rhea)

Freezing of interior of an icy body: rift valley of planet-size /15/
(Tethys, Titania, Dione?, Ariel?)

Drying up (contraction of the core): thrust fault /20/
(Mercury, Miranda? [5])

Traces of differenciation on the surface
- rays/halo of albedo feature around impact craters /1/
  (Mercury, Venus, Moon, Ganymede, Callisto, Titania, Oberon, Ariel)
- mare /5/: low viscosity lava flow
  -- probably in connection with break-through of young crust
  (Mercury, Moon, Mars, Triton, Japetus?, Umbriel?, Oberon?)
- linear albedo features at rift network /13/ because of crust fissures in connection with change of curvature (Europa, light and dark)

Traces of mantle circulation
- rift valley of somewhat smaller than planet-size /31/
  (large canyon on Venus, Valles Marineris on Mars)
- local groove-system becuse of
  - local heating /16/ (Enceladus) or
  - rising or sinking of boulders /32/ (Miranda)
- huge polygonal units of ancient crust /33/
  (Ganymede, Ariel, Umbriel [7], Triton, Earth)
  - together with global rift network /17/
    (Ganymede, Ariel, Triton, Earth)
  - with parallel groove system inside the grabens of the rift network /18/ (Ganymede, Earth, Ariel [6])
  - and transform faulted parallel grooves /19/ (Earth)
- shield-volcano (hot spot volcanism) /34/: local mantle upflow
  (volcanoes on Venus and Mars, Hawaiian type volcanoes on Earth)
- material outflow along cracks: new crust forming, thrusting because of compression: parallel groove system /18/
  (Ganymede, Enceladus, Ariel, Earth)
- traces of crust vanishing:
  (half craters /35/ on Enceladus and Ganymede
  deep oceanic trenches /36/ on Earth)
- transform faults /19/: displacing of crust-pieces on surface of a sphere
  - local (Ganymede, Enceladus, Triton?)
  - global: global plate tectonism (Earth)
- lift off of crust pieces /21/ (Miranda, Venus, Earth)
- crease of crust: chains of mountains /22/
  (Earth,Freya and Akna mountains on Venus)
- continents /23/ (Earth, Ishtar Terra on Venus)

Traces of small-scale terrain circulation within the crust:
terrain polygons /37/ (Triton?, Earth)

Traces of slipping on slopes in connection with
- mountains /26/ (Mars)
- bulge (Mars around Tharsis bulge)
- *dome /27/ (parquet terrain on Venus)*

VOLCANISM

Kind of volcanism:
- along cracks: the largest amount of volcanic material is emerging to the surface by this kind of volcamism
  (Mid-oceanic ridges on Earth /25/,
  Moon, Mars, Europa, Ganymede, Enceladus, Ariel, Triton)
- hot spot volcanism /34/(Venus, Mars, Io, Earth:Hawaiian-type volcanoes)
- subductional volcanism with volatile-rich lava /38/
  (only on Earth from volcanoes along the deep oceanic trenches)

Material of "lava":
- silicate magma (Mercury, Venus, Earth, Moon, Mars)
- sulphur (Io)
- water (Europa?, Enceladus?)
- ice /39/ (Ariel [18])
- nitrogen (Triton [19])

Result of volcanism:
- ovoid /40/ (unsuccessful volcanism) (Earth, Venus)
- resurfacing (flooding, erosion)
- outgassing

Recent active volcanism
(Earth to ~10 km high, Io to ~250 km high, Triton to ~8 km high, material from Enceladus forms the E-ring?)

ATMOSPHERES OF CRUSTY PLANETARY BODIES

name of the planetary body	surface pressure (atm)	composition	corresponding percentages
Mercury	10^{-15}	He, H_2	98%, 2%
Venus	90	CO_2, N_2	96,4%, 3,4%
Earth	1	N_2,O_2,H_2O, Ar	78%, 21%, 0,1%, 0,9%
Mars	0,007	CO_2, N_2, Ar	93,3%, 2,7%, 1,6%
Io	10^{-12}	SO_2, S, Na	
Titan	1,6	N_2, Ar, CH_4	85%, 12%, 3%
Triton	10^{-7}	N_2, CH_4	
Pluto	?	CH_4[8], N_2?, Ar?	

TRACES OF RUNNING LIQUID ON THE SURFACES /41/

Traces of running liquid's erosion: Earth, Mars (water), Io (sulphur)
River beds on the surface: only on Earth and Mars
Recently running water: only on Earth
Liquid ocean:
- on the surface:
  Earth : H_2O
  Titan?: metan, etan, propen, propan? [26,27,28]
- under an ice-crust:
  Europa: H_2O (100 km deep, ice crust 20-30 km thick [21])
  Enceladus?: H_2O
  Triton?: nitrogen (from the depth of 20-30 km?)

VOLATILE ELEMENTS (ATMOSPHERE, LIQUIDOSPHERE)

Sources: only giant planets are able to capture the surrounding gases, in all other cases the bulk of the volatile content directly inherited from gases occluded in the solid planetesimals from which the planetary body accreted
- outgassed from the interior of the planetary body (active volcanism speeds it up)
- volatile content of an impactor body may also contribute
- implantation from solar wind (generally negligible except in the case of Mercury)

Losses:
- escape (slow mechanism)
  - growing with temperature
  - decreasing with the mass-increase of the planetary body (growing escape velocity)
  - the escaped volatiles of satellites are exhausted gravitationally/electromagnetically by their own planet (Apparent at the orbit of Mimas, Enceladus, Tethys, Dione, Rhea, Titan. Spectacular in the case of Io.)
  - in the presence of an own magnetic field ions can be accelerated by electric fields to escape velocities (polar wind at Earth)
- solar wind erosion (slow mechanism) especially strong in the lack of own magnetic field (Venus, at high solar activity in the case of Mars)
- blow off in connection with impacts (episodic, quick mechanism) it may occasionally be significant (Mars, Argyrae impact /6/)
- fixed into the soil (slow mechanism)
  CO_2 into charbonates:
  H_2O into hydrated silicates:
  H_2O, CH_4, NH_3, N_2 into ices/clathrats
- freezing onto the surface /42/ (slow/quick mechanism depending on the temperature and its variation) (glaciers only on Earth /42/)

Regain: given material is regained only in certain temperature ranges by
- rain-fall (Venus: sulphuric acid, hydrochloric acid
  Earth: water, sulphurous acid, water with hydrochloric acid)
- snow-fall/hoar-frost (Earth: water
  Mars: water, carbon dioxide
  Io: sulphur, sulphur dioxide
  Titan: hydrocarbon aerosols of larger molecular weight [27]
  Triton: nitrogen, metan)

Recycling: in substantial amount only in the case of Earth because of global plate tectonism

POLAR CAPS IN THE SOLAR SYSTEM /43/	
Earth	H_2O
Mars northern polar cap:	H_2O
southern polar cap:	CO_2 / CO_2-H_2O clathrat
Triton	N_2
Pluto	N_2?

WHAT ARE THE MOST IMPORTANT PARAMETERS FOR THE ORIGIN OF LIFE [3]?

- temperature distribution
- chemistry

} { - in the atmosphere
- in the ocean
- on the surface of the crust

Both depends strongly whether
1./ the source material was already emplaced when planetesimal formation began
2./ during the accretion there was a continuous infall of interstellar material into the circumsolar region

Temperature
in case 1./ continuously high (run away accretion)
in case 2./ continuously low after cessation of early runaway accretion (the impact hot spots were distributed in time and space, they could cool down before the next impact)

Surface composition
in case 1./
- magma-ocean on the surface
- water - as solution in melt → decomposed by metallic iron
  - as wapor in the atmosphere dissociated by UV photons
  } {hidrogen escaped
  oxigen accumulated
- CO_2 in the atmosphere → greenhouse effect strong
- biomolecules destroyed (if built up)

in case 2./
- always a water ocean on the surface with present day mass at the end of the accretion
- water bound in sediments
- water loss and oxigen production slow (present day level)
- continental territory small
- ocean shallow
- many islands because of extensive volcanism →
  - many places for biomolecules being screened from UV radiation under an overlying layer of water
  - many places for accumulation of rust and clay minerals

Composition of the primordial atmosphere

(supposing that the bulk of it is identical with the volatile content of ordinary chondrites — since they could be the end-product of interstellar dust aggregates — a minor part of it came from carbonaceous chondrites and the rest is atmospheric by-product)

- H_2O, CO_2, CO, N_2, SO_2
- hydrocarbons come from carbonaceous meteorites
- no NH_3/CH_4 neither in ordinary nor in carbonaceous case
- hydrogen (maximum 1%) comes from meteoritic carbids and water (through photoinduced oxidation of dissolved Fe^{2+})
- NH_3 formation by
  - lightening in H_2 rich atmosphere
  - photoreduction of water by nitrogen on rutile (TiO_2) sands in the intertidal and wave zone of the ocean

LITERATURE

A: REFERENCES

1. Nature 335,496,1988.
2. Earth, Moon and Planets 45,101,1989.
3. Earth, Moon and Planets 37,187,1987.
4. Rev. Geophys. 23,391,1985.
5. Earth, Moon and Planets 42,293,1988.
6. Nature 327,201,1987.
7. Nature 338,324,1989.
8. Nature 336,452,1988.
9. Nature 335,240,1988.
10. Nature 282,811,1979.
11. Acta Astronautica 17,5,463,1988.
12. Icarus 71,203,1987.
13. Nature 326,30,1987.
14. Sky and Telescope July 1987, p. 13.
15. Nature 330,108,1987.
16. Moon and Planets 23,339,1980.
17. Earth, Moon and Planets 44,7,1989.
18. Sky and Telescope Dec. 1988, p. 606.
19. Sky and Telescope Febr. 1990, p. 136.
20. Fizikai Szemle 1989. Oct. p.363.
21. Bioastronomy - The Next Steps. ed. G. Marx, p. 21, Kluver Academic Publishers, 1988.
22. Compositional Clues to the History of the Terrestrial Planet Atmospheres. ed. R. Pepin, p. 44, LPI Techn. Rep. No 83-01 Lunar and Planetary Institute, Houston, 1982.
23. Evolution of Giant Gaseous Protoplanets Embedded in the Primitive Solar Nebula. A.G.W. Cameron et al., Harvard-Smithsonian Center for Astrophysics, Cambridge, Mass. 1982.
24. Nature 338,29,1989.
25. Earth, Moon and Planets 42,201,1988.
26. COSPAR XXVII.4.3.1.,1988.
27. Rev. Geophys. 25,299,1987.
28. Rev. Geophys. 25,264,1987.
29. Sci. Am. Aug. 1980, p. 56.

B: REVIEW PAPERS IN SPECIAL ISSUES

Pis'ma Astron. Zh. 5,No1-5,1979 (Venera 11-12)
8,No7,1982 (Venera 13-14)
12,No1,1986 (Vega 1-2)
COSPAR XXVI.8.3.1,1986 (Venera 15-16)
Nature 292,20.Aug.1981 (Voyager-1 at Saturn)
296,13,1982 (Venus Conference after Pioneer)
Science 194,17.Dec.1976. (Viking)
205,6.July.1979. (Pioneer Venus)
206,23.Nov.1979. (Voyager-2 at Jupiter)
207,25.Jan.1980. (Pioneer at Saturn)
212,10.Apr.1981. (Voyager-1 at Saturn)
213,18.Sept.1981. (Voyager-2 at Saturn)
233,4.July 1986. (Voyager-2 at Uranus)
246,15.Dec.1989. (Voyager-2 at Neptune)
J.G.R. 82,30.Sept.1977. (Viking Mission)
84,30.Dec.1979. (Viking Extended Mission)
VENUS ed. Hunten et al. The Univ. of Arizona Press, Tucson, Arizona, 1983.
SATURN ed. T. Gehrels, The Univ of Arizona Press, Tucson, Arizona, 1984.
PROTOSTARS AND PLANETS ed. T. Gehrels, The Univ. of Arizona Press, Tucson, Ariozona, 1978.
THE NASA MARS CONFERENCE ed. D.B. Reiber, Sci. and Techn. Series 71, Univelt, Incorporated, P.O. Box 28130, San Diego, Calif. 92128.

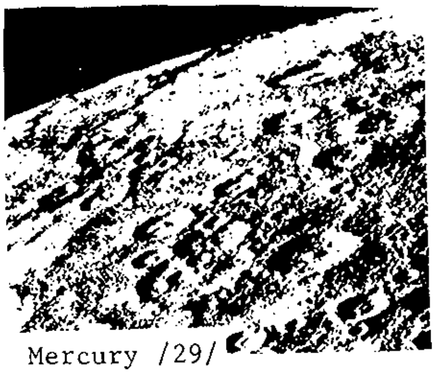
C: GENERAL PUBLICATIONS

COSPAR XXVII. 1988. 4.P.1.2.;4.1.6.;4.3.7.;4.4.2.
Rev. of Geophys. and Space Phys. 21,153,1983; 22,101,1984;22,131,1984.
Rev. of Geophys. 23,61,1985; 25,615,1987.
Geophys. Res. Lett. 9,499,1982; 9,1073,1982.
J.G.R. 86,7679,1981; 86,8513,1981.
Icarus 67,88,1986; 76,295,1988; 78,63,1989; 78,90,1989.
A.P.J. 238,1,1980.
Planet. Space Sci. 33,109,1985; 34,585,1986.
Nature 280,5725,1979; 282,811,1979; 283,815,1980; 289,17,1981; 290,759,1981; 298,44,1982; 301,666,1983; 302,663,1983; 326,30,1987; 334,145,1988; 335,127,1988; 335,240,1988; 335,587,1988; 336,143,1988; 336,714,1988.
Soviet Space Res. XXV,307,1987.
Earth, Moon and Planets 35,117,1986; 36,23,1986; 36,97,1986; 36,167,1986; 36,275,1986; 37,1,1987; 37,17,1987; 37,287,1987; 38,137,1987; 38,183,1987; 38,285,1987; 38,299,1987; 39,129,1987; 39,195,1987; 39,207,1987; 39,225,1987; 39,239,1987; 39,251,1987; 39,275,1987; 40,1,1988; 41,1,1988; 41,127,1988; 41,173,1988; 41,295,1988; 42,107,1988; 44,219,1989; 45,237,1989.
Science 206,1071,1979; 207,181,1980; 216,181,1982; 216,630,1982; 238,170,1987.
Sci. Am. 1980/Aug.p.46; 1982/Jan.p.73; 1987/May,p.40.
Mercury 1980/March-April,p.41; 1980/May-June,p.53.
Spaceflight 31,340,1989.
Journal of the British Interplanetary Society 41,41,1988; 41,49,1988.
Sky and Telescope 1981/Jan.p.14; 1983/Nov.p.405; 1984/Febr.p.110; 1985/Jun.p.507; 1986/Oct.p.336; 1989/Febr.p.149; 1989/Oct.p.358; 1990/Febr.p.136; 1990/Febr.p.146.
Second COSPAR Colloquium on Environmental Model of Mars, Sopron, Jan.22-26. 1990.
AGU CHAPMAN CONFERENCE, Balatonfured, June 4-8. 1990.
Fizika 1981-82,p.87, Gondolat, Budapest.
Fõld és Ég 1979/Nov.p.328; 1980/Jan.p.3; 1980/May,p.130; 1982/March,p.66; 1982/Apr.p.98; 1982/May,p.130; 1990/Jan.p.3; 1990/Febr.p.38.
Élet és Tudomány 1981/Aug.14,p.1034; 1981/Aug.21,p.1070.
Természet Világa 1984/Aug.p.353; 1989/July,p.299.

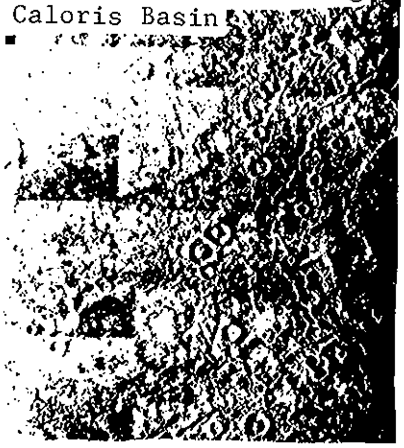
	No IAU number of satellites	NAME OF PLANETARY BODY	DISCOVERY YEAR DISCOVERER	P sideral orbital period day	a mean distance from the central body km	a unit: radius of the central body	i orbital inclination with respect to the equator of the central body degrees	e orbital eccentricity	T TEMPERATURE ΔT because of e estimated T Kelvin	T estimated T (on the basis of the visual albedo) Kelvin	measured T Kelvin	M mass kg	M mass unit: lunar mass	D diameter km	D diameter unit: lunar diameter	ρ density g/cm^3	g surface gravity unit: terrestrial g	m visual magnitude	A albedo %	smallest distance of highest-resolution-space-photos km	highest resolution km	SURFACE relief km	SURFACE age unit: 10^9 years	SURFACE composition	atmosphere	ocean	polar cap	magnetic field	is rotation captured?	shape (O: sphere, ø: irregular)
E	1	Moon		27,322	384000	60,27	5,1	0,055	8	187	40)-123	7,35 E22	1,0000	3476	1,000	3,34	0,16	-12,7	7	40	0,001	8	4,6	Si(100%).	-	-	-	-	+	O
MARS	1	Phobos	1877 Hall	0,319	9380	2,76	1,02	0,018	29	315	230	9,6 E15	1,3E-7	27	0,007	1,9 +0,5	0,0009	11,6	5	88	0,01	0,1	>4	C?	-	-	-	-	+	ø
	2	Deimos	1877 Hall	1,262	23500	6,91	1,82	0,002	29	315		2,0 E15	2,7E-8	12	0,003	2,1	0,0004	12,7	5	28	0,01	0,1	>4	C?	-	-	-	-	+	ø
JUPITER	16	Metis (79J3)	1980 Voyager	0,295	127960	1,79	~0	~0	6	145				~40	0,011		0,002	17,4	~5											
	14	Adrastea (79J1)	1979 Voyager	0,298	128980	1,81	~0	~0	6	145				<40	0,011		0,001	18,9	~5	3000										
	5	Amalthea	1892 Barnard	0,489	181300	2,54	0,455	0,003	6	145				240	0,069		0,009	14,1	6	417000	8		>4	Si,S?	-	-	-		+	ø
	15	Thebe (79J2)	1980 Voyager	0,675	221900	3,12	0,9	0,013	6	145				<80	0,023		0,003	15,5	~5											
	1	Io	1610 Galilei	1,769	412600	5,78	0,027	0,000	5	143	133	8,916E22	1,21	3632	1,045	3,57	0,18	5	60	18800	1	10-1	<E6years	Si,S	+		-	?	+	O
	2	Europa	1610 Galilei	3,551	670900	9,40	0,468	0,000	5	125	110- 85	4,873E22	0,66	3126	0,899	2,97	0,14	5,3	60	204200	4	0,2	E7years	Si(97%),H_2O	-	+?	-		+	O
	3	Ganymede	1610 Galilei	7,155	1070000	14,99	0,183	0,001	6	156	145 85	1,490E23	2,03	5276	1,518	1,93+0,01	0,15	4,6	40	59660	1	0,1	4-3,1	Si(50%),H_2O	-		-		+	O
	4	Callisto	1610 Galilei	16,689	1880000	26,33	0,253	0,007	6	166	155- 80	1,064E23	1,45	4820	1,386	1,83+0,01	0,12	5,6	20	124000	2		4,6-3,8	Si(47%),H_2O	-		-		+	O
	13	Leda	1974 Koval	240	11110000	155,61	27	0,147	7	170				~14	0,004		0,0003	20,2												
	6	Himalia	1904 Perrine	250,6	11470000	160,65	28	0,158	7	170				~170	0,049		0,004	14,8	~3											
	10	Lysithea	1938 Nicholson	260	11710000	164,01	29	0,12	7	170				<32	0,009		0,0006	18,4												
	7	Elara	1905 Perrine	260,1	11740000	164,43	26	0,207	7	170				~80	0,023		0,002	16,7	~3											
	12	Ananke	1951 Nicholson	617	20700000	290	147	0,169	7	170				~28	0,008		0,0005	18,9												
	11	Carme	1938 Nicholson	692	22350000	313	163	0,207	7	170				~40	0,011		0,0008	18,0												
	8	Pasiphae	1908 Mellotte	735	23300000	326	147	0,40	7	170				~46	0,013		0,0009	17,7												
	9	Sinope	1914 Nicholson	758	23700000	332	156	0,275	7	170				~36	0,010		0,0008	18,3												
SATURN	17	Atlas	1980 Voyager	0,606	138200	2,28	0,3	0,002	6	112				<50	0,014		0,0008	18	~40	219000	13									ø
	16	Prometheus (80S27)	1980 Voyager	0,613	139400	2,31	0	0,004	6	112				<200	0,057		0,006	16,5	~40	246400	7									ø
	15	Pandora (80S26)	1980 Voyager	0,629	141700	2,35	0,1	0,004	6	112				<140	0,040		0,005	16	~40	107000	8									ø
	10	Janus (80S1)	1966 Dollfus	0,696	151400	2,51	0,1	0,007	6	112				<220	0,063		0,005	14,5	~40	121000	3		>4						+	ø
	11	Epimetheus (80S3)	1977 Fountain,Larson	0,696	151450	2,51	0,3	0,009	6	112				<140	0,040		0,004	15,5	~40	222700	3,2		>4						+	ø
	1	Mimas	1789 Herschel	0,942	185500	3,08	1,517	0,020	6	101		3,76 E19	0,0005	394	0,113	1,43+0,10	0,007	12,9	60	88240	2	9	>3,9	Si(31%),H_2O	-	-	-		+	O!
	2	Enceladus	1789 Herschel	1,370	238000	3,95	0,023	0,004				7,4 E19	0,0010	502	0,144	1,20+0,6	0,008	11,8	100	86890	2		E8years	Si(9%),H_2O	-	?	-		+	O
	3	Tethys	1684 Cassini	1,888	294700	4,88	1,093	0,000	6	88		6,26 E20	0,0085	1060	0,305	1,25+0,06	0,015	10,3	80	92470	2	3	3,8-4	Si(13%),H_2O	-	-	-		+	O
	13	Telesto (80S13)	1980 Voyager	1,888	294670	4,88			6	112				<30	0,009		0,0004	19	~40	153500	12									ø
	14	Calypso (80S25)	1980 Voyager	1,888	294670	4,88			6	112				<24	0,007		0,0003	18,5	~40	237320	5									ø
	4	Dione	1684 Cassini	2,737	377400	6,26	0,023	0,002	6	101		1,05 E21	0,0142	1120	0,322	1,43+0,1	0,022	10,4	62	160000	3		>4	Si(18%),H_2O	-	-	-		+	O
	12	Helena(DioneB,80S6)	1980 Voyager	2,737	377400	6,26	0,2	0,005	6	112				<34	0,010		0,0008	10,5	~40	225000	6									ø
	5	Rhea	1672 Cassini	4,518	527000	8,73	0,35	0,001	6	112	99- 73	2,28 E21	0,0310	1530	0,440	1,33+0,1	0,023	9,7	60	72000	1		>4	Si(12%),H_2O	-	-	-		+	O
	6	Titan	1655 Huygens	15,945	1222000	20,3	0,33	0,029	6	114	94	1,36 E23	1,8503	<5150	1,481	>1,9	0,014	8,4	35	3920	1		cloud	Si(42%),H_2O	+!	+?	?	?	+	O
	7	Hyperion	1848 Bond	21,277	1481000	24,6	0,4	0,104	8	120		1,1 E20	0,0015	610	0,118		0,006	14,2	20	470640	9		>4	ices	-	-	-		-!	ø!
	8	Japetus	1671 Cassini	79,331	3560000	59	14,7	0,028	8	122		1,93 E21	0,0262	1460	0,420	1,2 +0,5	0,02	11,9	50-5	908340	17		>4	Si(7%),ices	-	-	-		+	O
	9	Phoebe	1898 Pickering	550	12930000	215	150	0,163	8	126				220	0,063		0,005	16,5	5	2075500	38			C?					+	O!
URANUS	13	Cordelia (86U7)	1986 Voyager	0,334	49700	1,94			4	87	80			24	0,007				<10											
	14	Ophelia (86U8)	1986 Voyager	0,377	53800	2,10			4	87	80			32	0,009				<10											
	15	Bianca (86U9)	1986 Voyager	0,435	59200	2,31			4	87	80			48	0,013				<10											
	9	Cressida (86U3)	1986 Voyager	0,464	61800	2,41			4	87	80			72	0,021				<10											
	12	Desdemona (86U6)	1986 Voyager	0,474	62700	2,45			4	87	80			48	0,013				<10											
	8	Juliet (86U2)	1986 Voyager	0,496	64600	2,52			4	87	80			72	0,021				<10											
	7	Portia (86U1)	1986 Voyager	0,513	66100	2,58			4	87	80			88	0,025				<10											
	10	Rosalind (86U4)	1986 Voyager	0,558	69900	2,73			4	87	80			48	0,014				<10											
	11	Belinda (86U5)	1986 Voyager	0,624	75300	2,94			4	87	80			145	0,042				<10											
	6	Puck (85U1)	1985 Voyager	0,761	86000	3,36			4	87	80			170	0,049				7	500000	9									ø
	5	Miranda	1948 Kuiper	1,414	129900	5,07	4,22	0,027	4	81	84	3,4 E19	0,0004	484	0,139	1,26+0,39	0,004	16,5	24	29000	0,6	20	<4	Si,ices	-	-	-		+	O
	1	Ariel	1851 Lassell	2,520	190900	7,46	0,31	0,003	4	79	86	6,7 E20	0,0091	1160	0,334	1,65+0,30	0,01	14,4	31	130000	2,5	1	<<4	Si,ices	-	-	-		+	O
	2	Umbriel	1851 Lassell	4,144	266000	10,39	0,36	0,005	4	85	86	7,6 E20	0,0103	1190	0,342	1,44+0,28	0,008	15,3	50-19	325000	11		4	Si,ices	-	-	-		+	O
	3	Titania	1787 Herschel	8,706	436300	17,04	0,14	0,002	4	83	86	1,2 E21	0,0163	1610	0,463	1,59+0,09	0,02	16	28	364000	7	5	<4	Si(50%),ices	-	-	-		+	O
	4	Oberon	1787 Herschel	13,463	583400	22,79	0,10	0,001	4	84	86	8,2 E21	0,1115	1550	0,446	1,50+0,10	0,1	16,2	26	470000	12	20	>4	Si(50%),ices	-	-	-		+	O
NEPTUNE		89N6	1989 Voyager	0,296	48000	1,95	0	0						54	0,014				6		23,6									ø
		89N5	1989 Voyager	0,313	50000	2,02	4,5	0						80	0,026				6		17,4									ø
		89N3	1989 Voyager	0,333	52500	2,12	<1	0						150	0,040				6		16,9									ø
		89N4	1989 Voyager	0,43	62000	2,51	<1	0						180	0,046				5		18,4									ø
		89N2	1989 Voyager	0,554	73600	2,97	<1	0						190	0,058				6		4,1									ø
	3	89N1	1989 Voyager	1,121	117600	4,75	<1	0						400	0,120				6	146000	1,3									ø
	1	Triton	1846 Lassell	5,877	354800	14,36	157	0,001	0	65	38			2705	0,782	2,02		13,6	90	38500	0,8	0,2	<E6years	N_2,CH_4 ices	+	?	+		+	O
	2	Nereida	1949 Kuiper	359,88	5510000	222,63	29	0,75	0	65				340	0,098			18,7	14	4700000	43,3								+	O
PL	1	Charon	1978 Christy	6,387	17000	15,45	105		15	55		1,28 E21	0,0174	1119	0,316	1,84+0,19	0,02	17	42					ices	?	?			+	
PLANETS		Mercury		88	57,9E6	83		0,2056	130	625	703- 88	3,3 E23	4,49	4880	1,40	5,43	0,37	- 0,2	56	328	0,05	3	4,6	Si(100%)	-	-	-	+	?	O
		Venus		224,7	108,2E6	156		0,0068	3	324	773-773	4,87 E24	66,26	12104	3,48	5,24	0,90	- 3,6	72	1000	1,5	10	1	Si(100%)	+	-		-	-	O
		Earth		365,3	149,6E6	215	7,25	0,0167	8	349	288	5,98 E24	81,31	12756	3,67	5,52	1,00		39	200	0,01	10	4,6- <E6	Si(100%)	+	+	+	+	-	O
		Mars		687	227,9E6	328		0,0934	27	306	293-233	6,4 E23	8,74	6787	1,95	3,93	0,38	1,8	16	1500	0,1	25	>4	Si(100%)	+	-	+	+	-	O
		Jupiter		11,86 y	778,3E6	1120		0,0485	6	131	143	1,899E27	25836	142796	41,08	1,326	2,53	- 1,7	70	42800	1		cloud	Si-core,H,He	+			+	-	O
		Saturn		29,45 y	1427 E6	2053		0,0556	7	95	123	5,69 E26	7741	120000	34,52	0,686	1,07	1,3	75	101100	3		cloud	Si-core,H,He	+			+	-	O
		Uranus	1781 Herschel	84,01 y	2869 E6	4128		0,0472	4	67	59	8,7 E25	1183	51200	14,73	1,267	0,90	6	90	81400	2		cloud	Si-core,H,He	+			+	-	O
		Neptune	1846 Leverrier,Galle	164,79 y	4496 E6	6469		0,0086	0	53	59	1,03 E26	1401	49500	14,24	1,64	1,14	7,9	82	4905			cloud	Si-core,H,He	+			+	-	O
		Pluto	1930 Tombaugh	250,3 y	5900 E6	8489		0,250	13	49	55	8,2 E22	1,12	2245	0,63	1,86	0,09	15	61					ices	+		+		-	
		Sun										1,99 E30	2,7 E7	1392000	400	1,41	27,93	-26,86						H(90%),He(9%)	+			+		O

WHAT LEVEL OF GEOLOGICAL ACTIVITY WAS REALIZED ON CRUSTY PLANETARY BODIES WITH DIAMETER LARGER THAN 240 KM

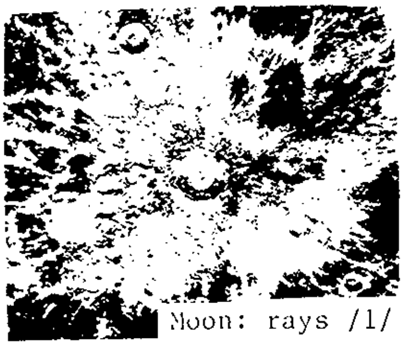
	uniform surface	surface saturated with craters	larger craters (population I)	smaller craters (population II)	multiring basin	rays/haloes around impact craters	bright crix-crax albedo features	rift valley of planet size	contraction (thrust fault)	any traces of out-flowing material	mare (lava of low viscosity)	expansion	rift system	parallel grooves	transform faults	shield volcanoes	calderas	recent volcanism	mountain chains	continents	erosion of flowing material	river bed
Moon	-	+-	+	+	+	+	-	-	-	+!	+!	-	-	-	-	-	-	-	-	-	-	-
Amalthea	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Io	+	-	-!	-!	-	-	-	-	-	+	-	-	-	-	-	-	+!	+!	-	-	+	-
Europa	+	-	-!	~!	-	-	-	-	-	+?	-	-	+!	-	-	-	-	-?	-	-	-	-
Ganymede	-!	+-	+	+	+	+	-	-	-	-	-	+?	+!	+!	+!	-	-	-	-	-	-	-
Callisto	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Mimas	+	+	+!	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Enceladus	-	+-	-!	+	-	-	-	-	-	-	-	+?	+!	+!	+!	-	-	-?	-	-	-	-
Tethys	-	+-	+	+	-	-	-	+!	-	-	-	+?	-	-	-	-	-	-	-	-	-	-
Dione	-	+	+	+	-	-	+!	+	-	-	-	+	-	-	-	-	-	-	-	-	-	-
Rhea	-	+	-	+	-	-	+!	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Titan	Nothing visible on the surface because of hydrocarbon smog/snow.																					
Hyperion	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Japetus	-!	+	+	+	-	-	-	-	-	+?	+?	-	-	-	-	-	-	-	-	-	-	-
Miranda	-	+-	-	+	-	-	-	-	+?	+	-	-	+!	+!	-	-	-	-	-	-	-	-
Ariel	-	+-	-	+	-	+	-	+	-	-	-	+!	+!	+	-	-	-	-	-	-	-	-
Umbriel	-!	+	+	+	-	-!	-	-	-	+?	+?	-	-	-	-	-	-	-	-	-	-	-
Titania	-	+-	-	+	-	+	-	+	-	+?	+?	+?	-	-	-	-	-	-	-	-	-	-
Oberon	+	+	+	+	-	+	-	-	-	+?	+?	-	-	-	-	-	-	-	-	-	-	-
89N1	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Triton	-	-	-!	-	-	-	-	-	-	+!	+!	-	+!	-	+?	-	+!	+!	-	-	-	-
Nereid																						
Charon																						
Mercury	+	+	+	+	+	+	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-
Venus	-	-	-	+	-	+	-	-	-	+	-	+	?	+	?	+!	+	+?	+!	+!	-	-
Earth	-	-	-	+	-	-	-	-	-	+	-	+	+	+	+	+	+	+	+	+	+	+
Mars	-	+-	+	+	-	-	-	-	-	+	-	+!	-	+	-	+!	+	-	-	-	+	+!
Pluto	-																					

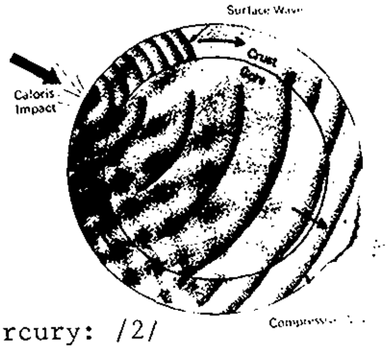

Mercury /29/

Mercury: the multiring Caloris Basin




Mercury /20/

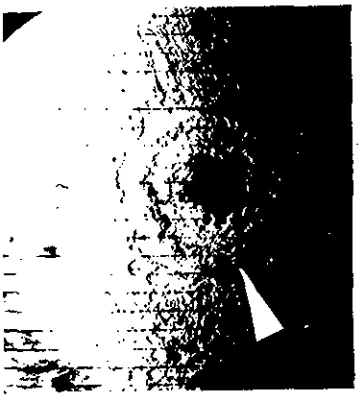

Moon: rays /1/



Mercury: /2/
how a multiring basin was born


Deimos /29/

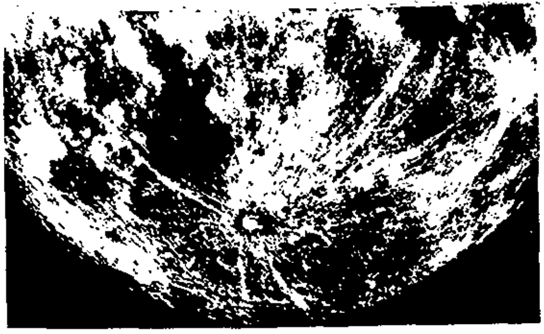

Moon: maria /5/


Moon: Mare Orientale
multiring basin

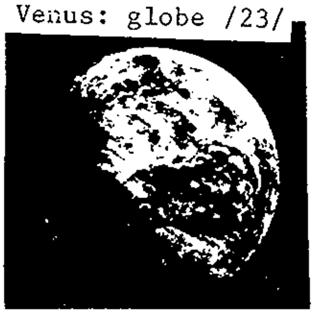

Phobos /29/

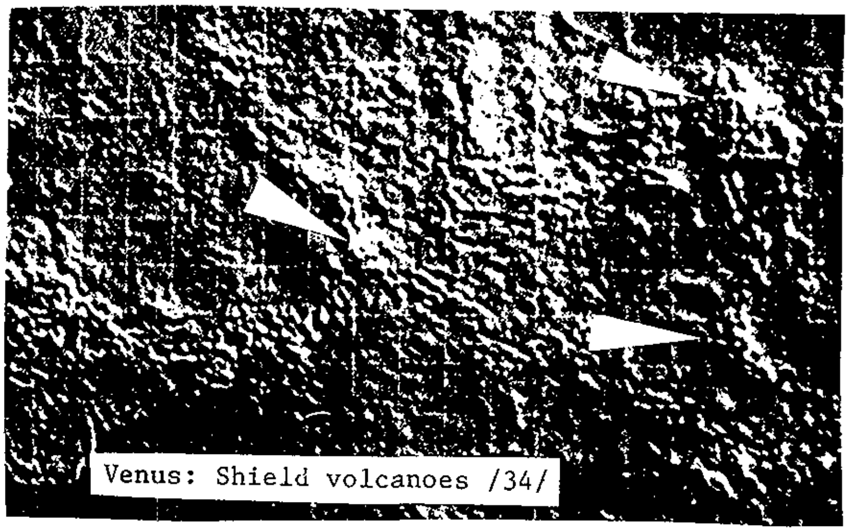

Moon: boundary of a mare /5/

Moon: rays around impact crater /1/



Venus: globe /23/



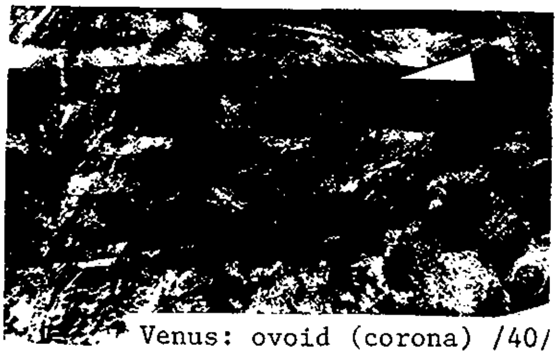


Venus: Shield volcanoes /34/

Venus: mountain chains /22/

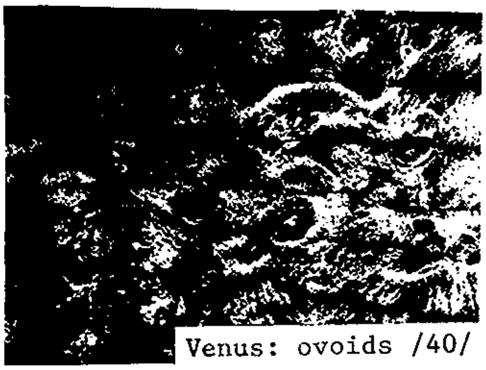
Venus: parquet terrain units /14/

Venus: Lakshmi Planum /21/



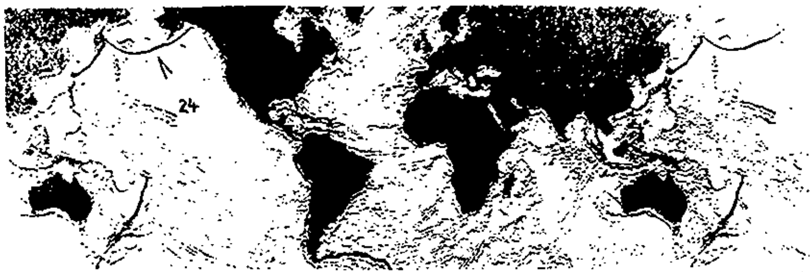
Venus: ovoid (corona) /40/

Venus: parquet terrain /27/

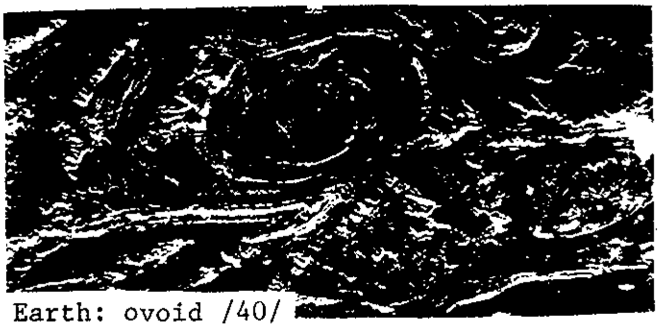


Venus: ovoids /40/

Earth:
continents /23/



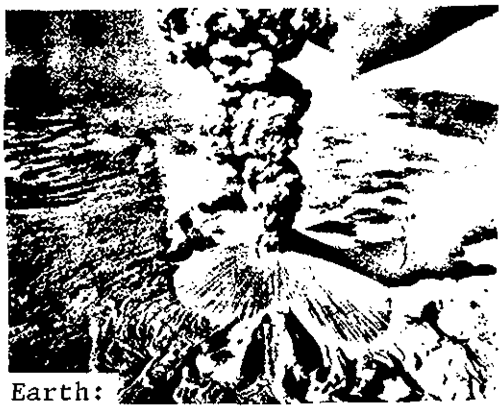
Earth: /28,33,17,18,19/ deep oceanic trenches /24/



Earth: ovoid /40/

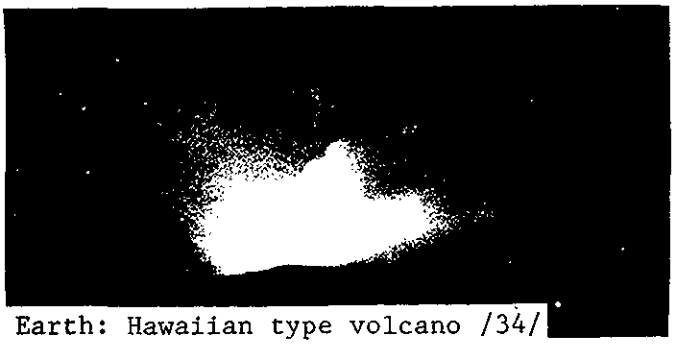
Earth:
mid oceanic ridges /25,18/
with transform faults /19/

Earth: mountain chains /22/



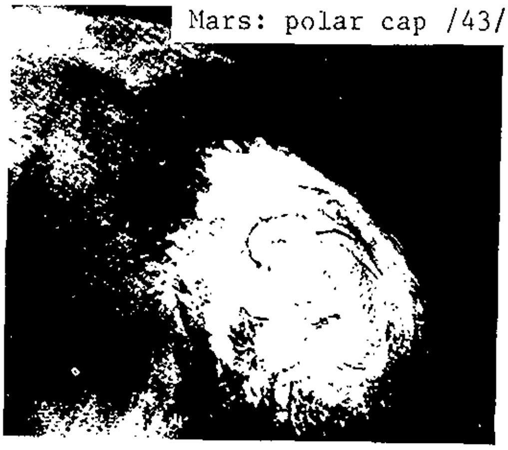
Earth:
subductional type volcano /38/

Earth: glaciers /42/

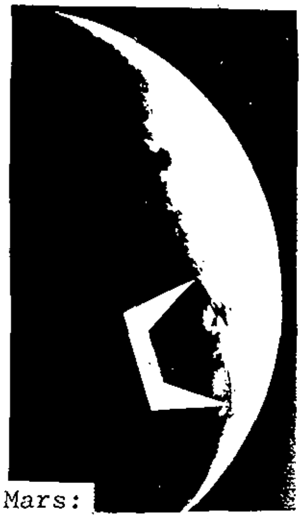


Earth: Hawaiian type volcano /34/

Mars:
shield volcanoes /34/



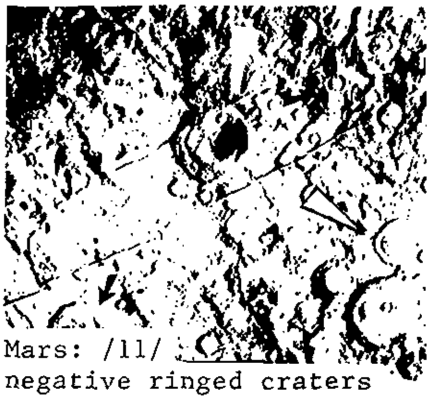
Mars: polar cap /43/



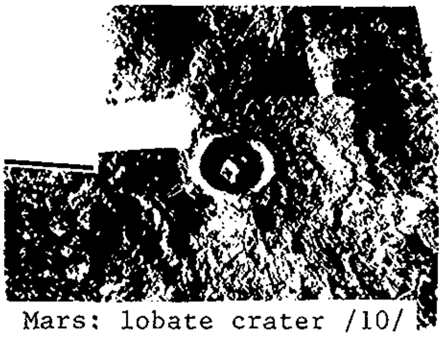
Mars:
rift valley /31/
(Valles Marineris)

Mars:
Argyrae Basin /6/

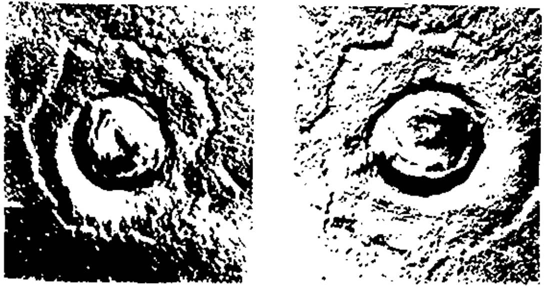
Mars:
butterfly shaped crater /9/



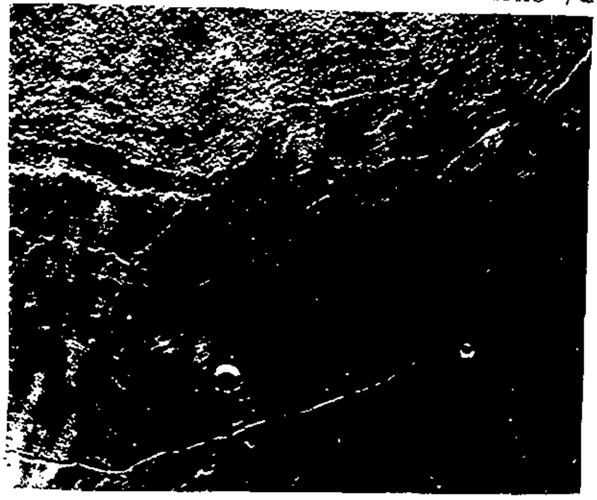
Mars: /11/
negative ringed craters



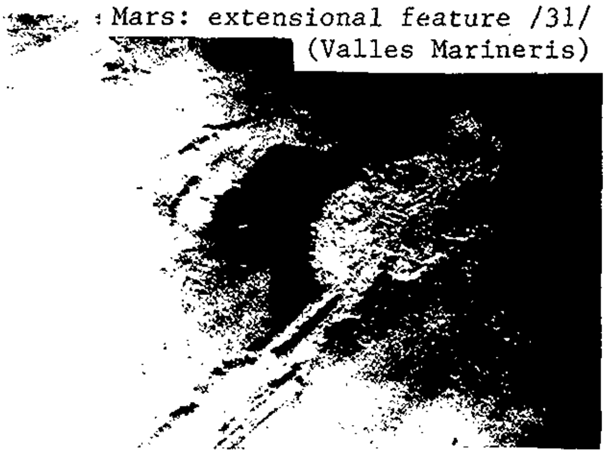
Mars: lobate crater /10/



Mars: lobate craters /10/

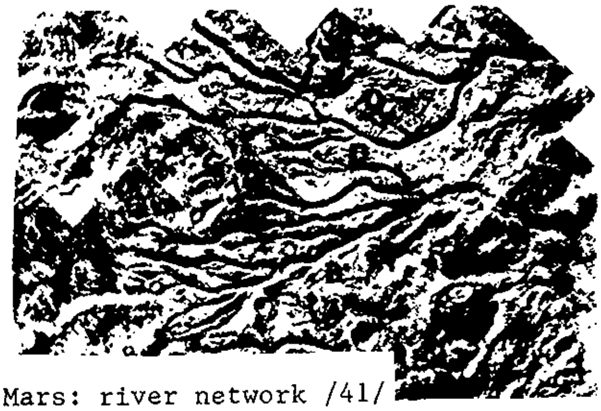


Mars:
grooved terrain around Arsia Mons /26/



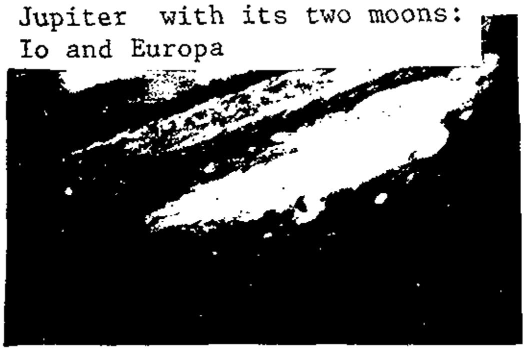
Mars: extensional feature /31/
(Valles Marineris)

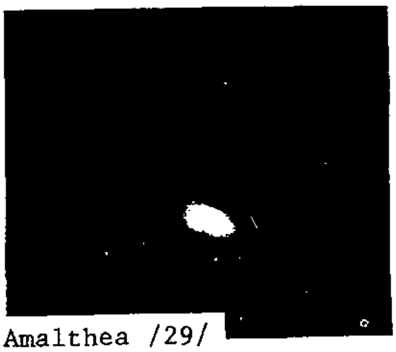
Mars: traces of flowing water /41/



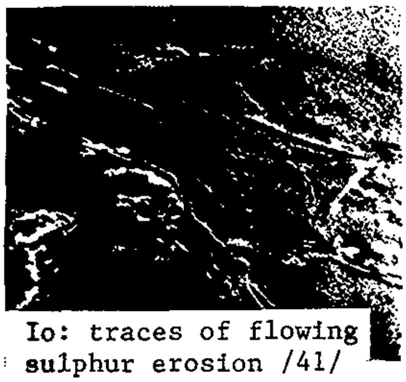
Mars: river network /41/

Jupiter with its two moons: Io and Europa





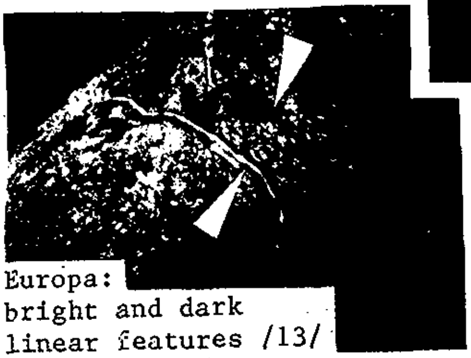
Amalthea /29/



Io: traces of flowing sulphur erosion /41/

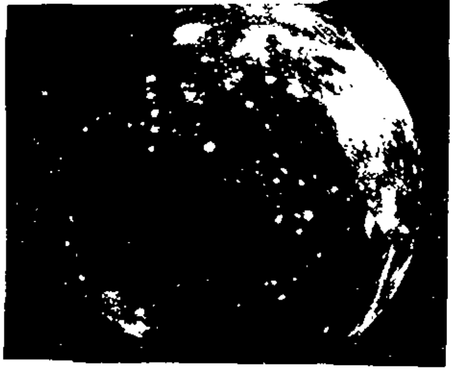
The two sides of Io /3/

Io: Pele volcano

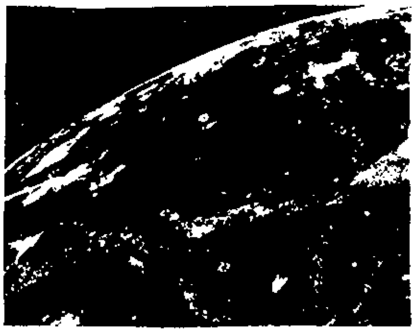


Europa: bright and dark linear features /13/

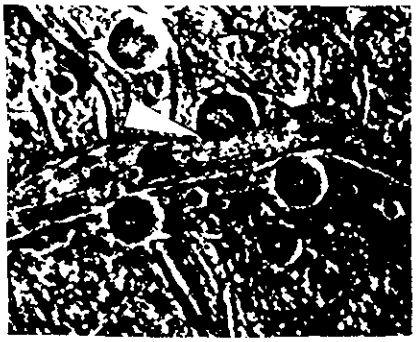
Ganymede /1,17,33/



Ganymede : large dark polygonal surface elements /33/ with brighter linear features /17/



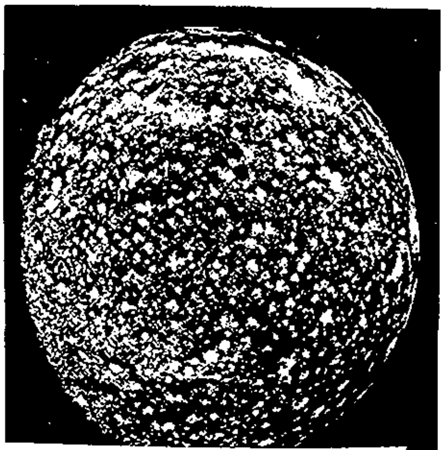
Ganymede : half craters /35/



Ganymede : transform fault /19/

Ganymede : grooved terrain /18/

Ganymede : parallel grooves /18/



Callisto: saturated with impact craters /29/

Callisto: Valhalla multiring basin

Janus /29/

Dione /29/

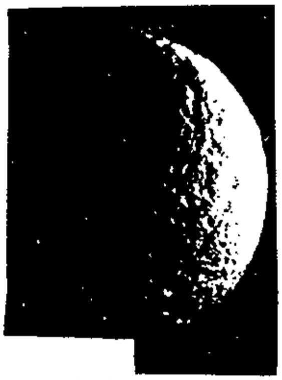
Mimas: /29/
deep crater floor.
The excavated
material escaped.

Rhea: full with
impact craters /29/

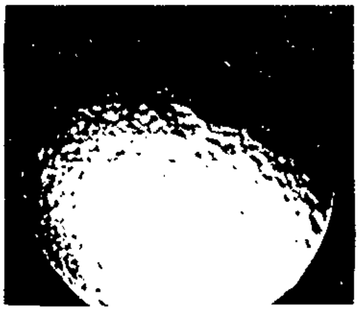
Tethys:
rift valley /15/

The cloudy Titan

Dione: /30/ bright
albedo features

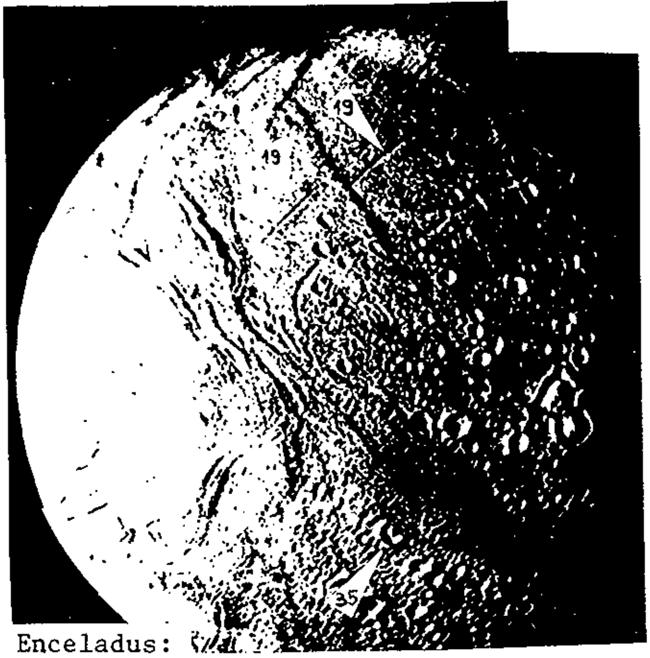


Mimas /4/:
saturated with
impact craters /29/

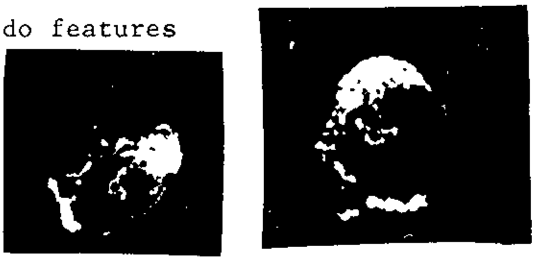


Tethys: Odysseus crater
(curved crater floor) /12/

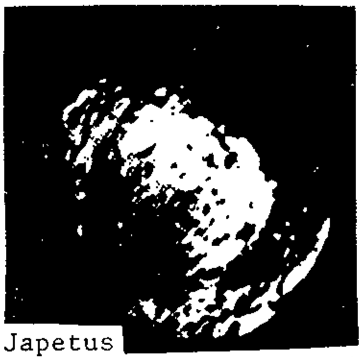
Rhea: /30/
bright albedo features



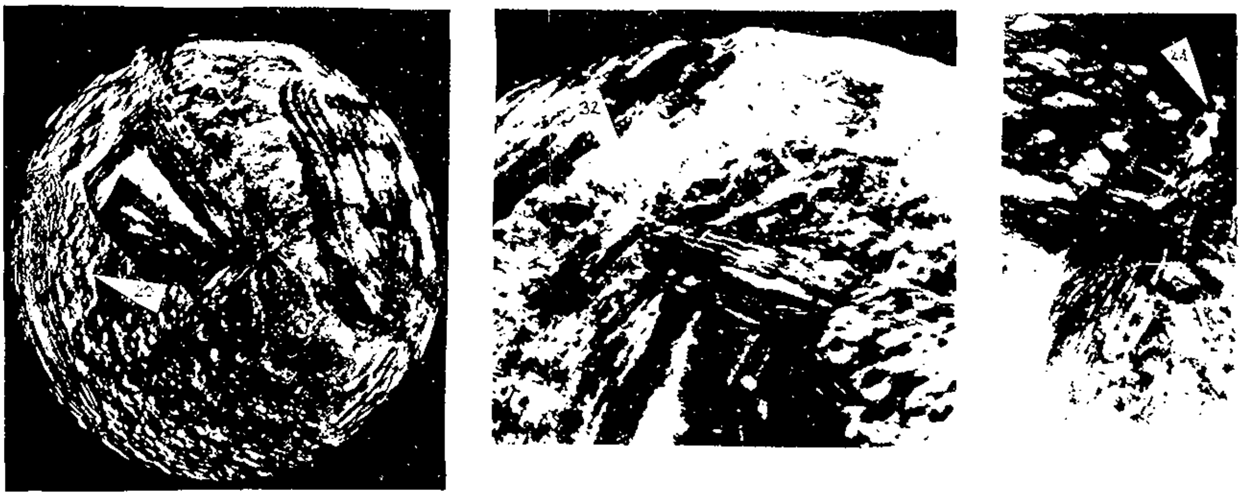
Enceladus:
rift system /16/, half craters /35/,
transform faults /19/, parallel grooves /18/.



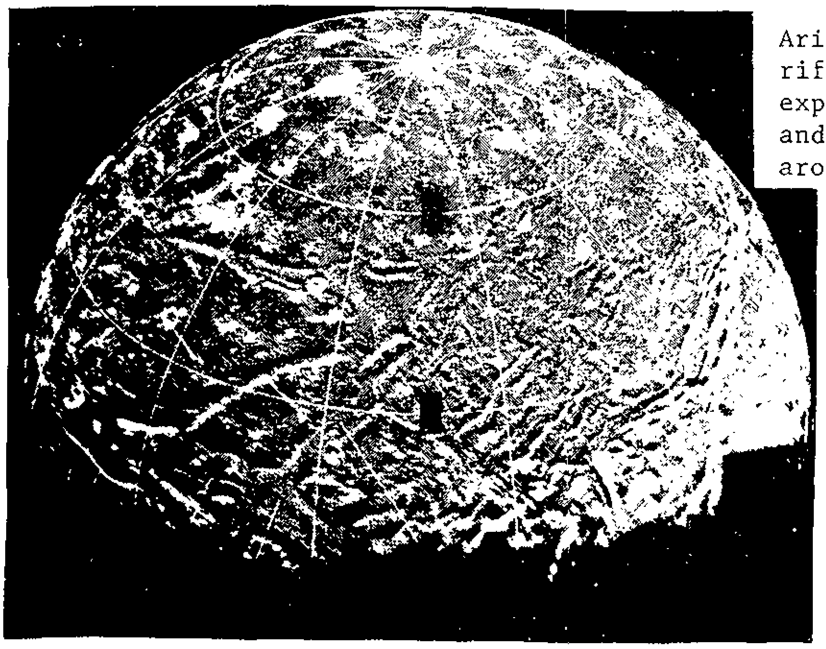
Two sides of Hyperion /4,29/



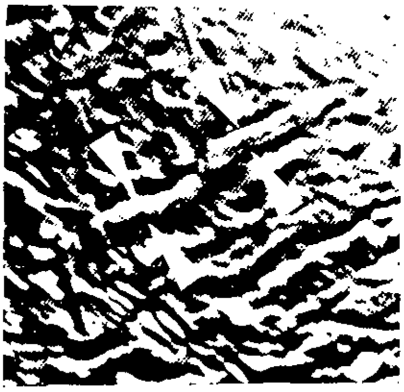
Japetus
in full-moon phase /5/



Miranda: ovoids, thrust fault /20/, groove system /32/, lift off /21/

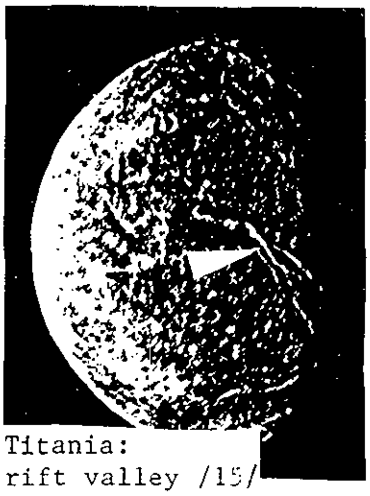


Ariel: rift valley /15/,
rift network /17/,
expansion feature /18/
and bright haloes
around impact craters /1/

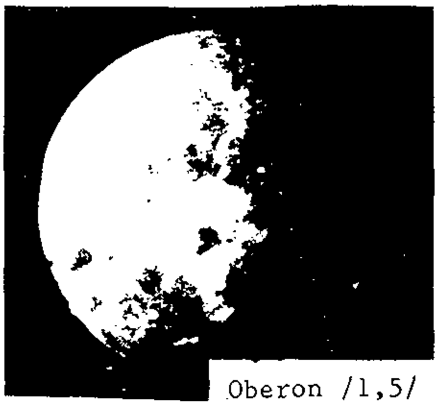


Ariel: ice volcanoes /39/

Umbriel /33/:
bright features /5/

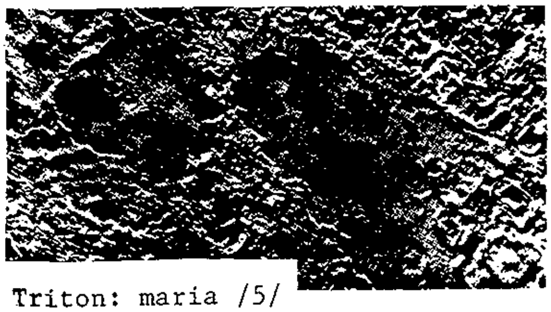


Titania:
rift valley /15/

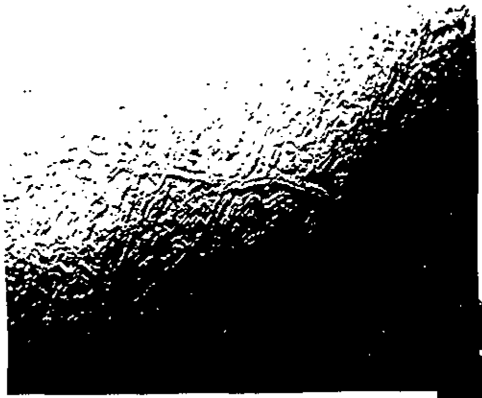


Oberon /1,5/

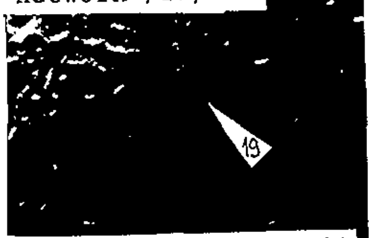
89N1 /29/



Triton: maria /5/



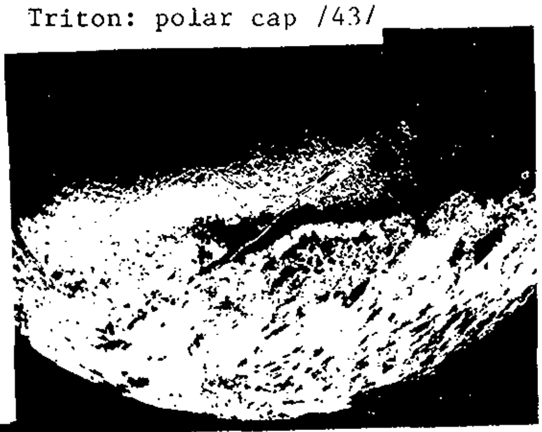
Triton: rift network /17/



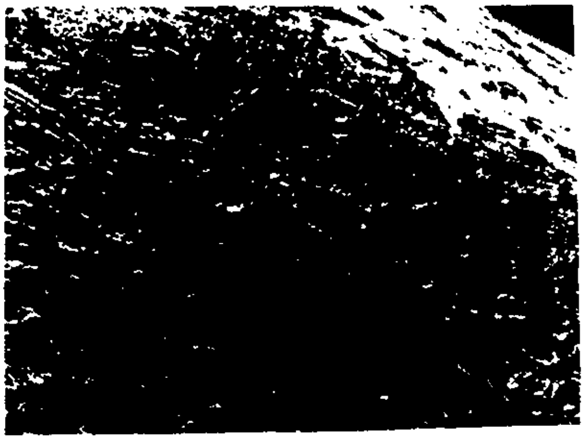
transform fault? /19/

Triton: calderas?

Triton: polar cap /43/



Triton: terrain polygons /37/



KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

CONVECTION IN THE EARTH'S MANTLE: MOTOR OF THE GEOLOGIC EVOLUTION

L. Cserepes

Geophysical Department of Eötvös University,
1083 Budapest, Kun Béla tér 2, Hungary

Received on 20th June, 1990

Thermal convection in the Earth's life

"Pantha rhei", everything flows in the different shells of the Earth's body, starting from the atmosphere down to the central core. There is no solid motionless matter above our head or under our foot. For the atmosphere this is quite evident, but on the geological time scale the Earth's crust deforms also with surprising ease. The next shell, the most voluminous one in the Earth with its depth of 2900 km, the mantle, circulates at a speed of several centimetres a year and drives in this way the wandering of continents. The fluid metallic core of the Earth convects much more rapidly and produces the geomagnetic field. This global system of circulations is driven primarily by the thermal buoyancy force in every shell, therefore it may be called thermal convection. The heat source of the atmosphere is the solar radiation, while the internal layers of the Earth are heated by the radioactivity of their constituents or by the original heat which accumulated at the beginning of the Earth's history.

These circulations make the Earth a living body (in the planetological sense) in contrast to such frozen planets as the Moon.

On the surface of the Earth, continents drift, oceanic basins are born and later disappear. According to the recent global tectonic theory, the primary tectonic units of the Earth's surface are the plates (Figure 1) which behave almost rigidly and move with respect to each other. These blocks carry continents and their motion leads to the growth or disappearance of oceans. The plate tectonic motion gives the frame and motive force to practically every endogenic geological process like orogenesis or magmatism.

The drift of the plates is the surface manifestation of convection in the mantle. Consequently one can say that the internal motor of the geologic evolution of the Earth's surface is the thermal convection of the mantle. This evolution shows a certain cyclicity, e.g. the recurrence of orogenic periods in the Earth's

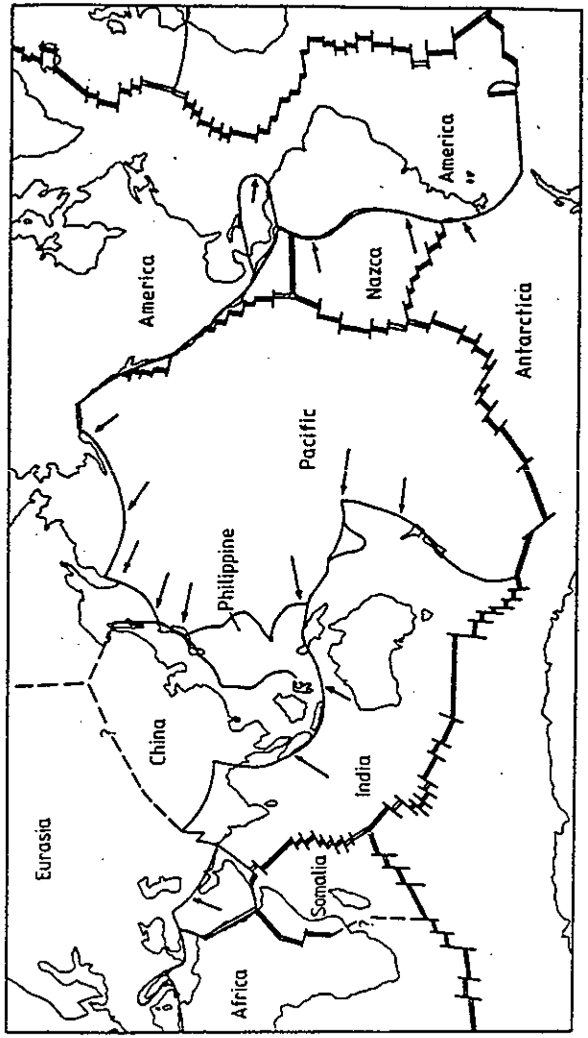


Fig.1. Lithospheric plates on the surface of the Earth. Divergent plate boundaries are drawn by double lines, convergent boundaries and boundaries with antiparallel plate motion (transform faults) by single thick lines. Arrows show examples of direction of plate convergence.

life, on the other hand it leads to irreversible, unidirectional processes like the permanent growth of continental crust, a result of the continuous crust-mantle differentiation.

The plates are said to be separate blocks of the lithosphere, coldest outer shell of the Earth, which consists of the crust and the uppermost part of the mantle. This shell, 100 km deep approximately, behaves rather rigidly, at least in comparison with deeper, less viscous parts of the mantle. Its relative rigidity justifies the usage of the term "plate" for the lithospheric blocks. This paper will now summarize recent views about the origin of the lithosphere and its plates in the light of the convection theory.

The lithosphere as a convective boundary layer

A simple model case for studying free thermal convection is the circulation maintained in a horizontal fluid layer by bottom heating. The most important dimensionless hydrodynamical parameter describing this flow is the Rayleigh number

$$Ra = \frac{\alpha g \Delta T d^3}{\nu \kappa}$$

where α is the coefficient of thermal expansion, g is the gravitational acceleration, ΔT is the temperature difference between the lower and upper boundaries, d is the layer depth, ν is the kinematic viscosity and κ is the thermal diffusivity. Convection starts when Ra exceeds the critical Rayleigh number Ra_{cr} which is in the order of 10^3. The actual Rayleigh number of the Earth's mantle is about 10^7-10^8, therefore this shell of the Earth must be in vigorous circulation. In this range of Ra the convection shows a sharp boundary layer structure (Figure 2) which means that the significant temperature gradients are confined to thin boundary layers at the top and bottom as well as to thin jet-like structures in the centre of upgoing and downgoing flow. The strong circulation and mixing in the centre of the convective "cells" eliminates any large temperature variation.

The Earth's mantle differs in many respects from the simple case of Figure 2, but the boundary layers must be present in the Earth since they are necessary elements of the high Rayleigh number flow. For example, the top boundary layer with its high vertical temperature gradient is a consequence of high vertical heat flux produced by the vigorous convection. Near the horizontal upper surface this high heat flux cannot be transferred by vertical convection, therefore strong conduction and high gradients must develop.

The main complication present in the real Earth is due to the temperature-dependence of its effective viscosity. Decreasing temperatures induce rapidly increasing viscosity. Therefore the upper boundary layer of the Earth is not only a thermal boundary layer but also a mechanical one, which differs from the deeper fluid layers by its very high viscosity. This thermo-mechanical upper boundary layer of the Earth is called conventionally the "lithosphere". The viscosity contrast between the lithosphere and the underlying layers amounts to several orders of magnitude. This appears as a relative rigidity of the lithosphere. (It is to be noted that with increasing depth the pressure-dependence of viscosity becomes important, probably dominant, leading to increasing viscosities. Therefore a zone of viscosity minimum is expected below the lithosphere. This weak zone is called the "asthenosphere".)

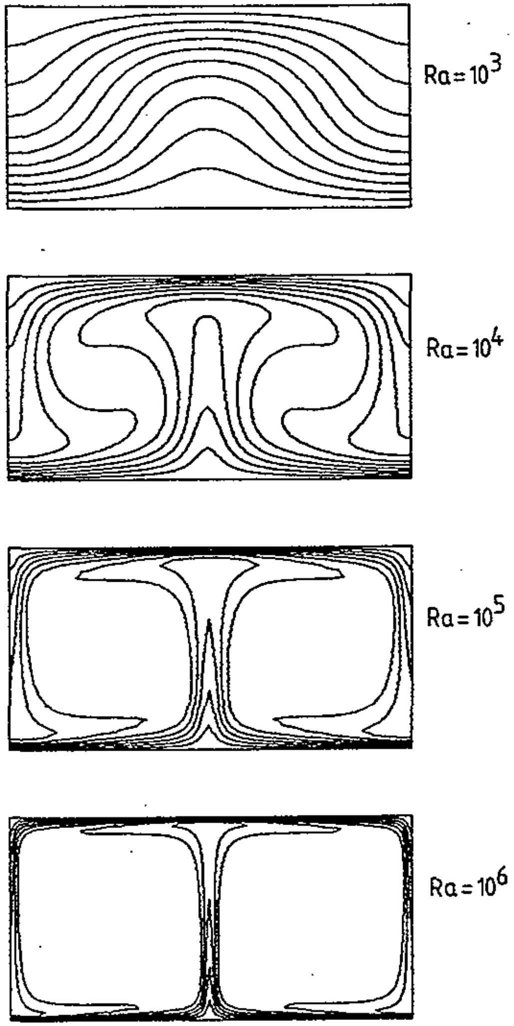


Fig.2. Development of boundary layers at high Rayleigh numbers. The figures show a *vertical* cross-section of two counter-rotating convective cells in an infinite horizontal layer. The contours are isotherms drawn at equal intervals. The horizontal boundaries of the fluid are isothermal.

The boundary layer theory predicts a power law for the thickness of the boundary layer δ:

$$\frac{\delta}{d} \sim (Ra/Ra_{cr})^{-m}$$

where $m = 0.28$ for temperature-dependent viscosities (Richter et al., 1983). Using this law, a theoretical thickness of around 100 km can be estimated for the lithosphere. This result agrees very well with the geophysical observations.

Of course, a lower boundary layer can also be present in the mantle at the core-mantle boundary. This is even a necessity if the bottom heating, i.e. the heat flux from the core into the mantle is significant as is assumed by the dynamo theory of the geomagnetic field.

The breakup of the lithosphere into plates

The quasi-rigid behaviour of the lithosphere complicates the numerical problem of convection modelling. If this behaviour is not taken into account in the numerical solution of the equations of convection, e.g. if the temperature-dependence of viscosity is weakened, then the typical three-dimensional pattern of convection is similar to that of Figure 3. The convective cell of this figure shows

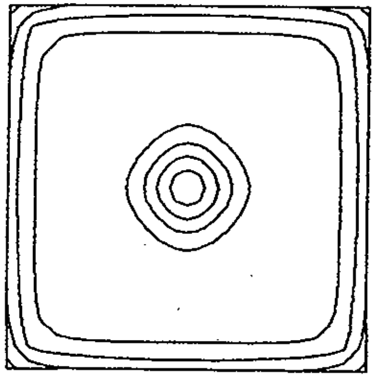


Fig.3. Structure of convection with temperature-dependent viscosity: isotherms in a *horizontal* plane, in the low viscosity zone ("asthenosphere") below the surface. Temperature is decreasing from the centre towards the periphery of the cell. An infinite horizontal layer may be filled up with the repetition of such square cells.

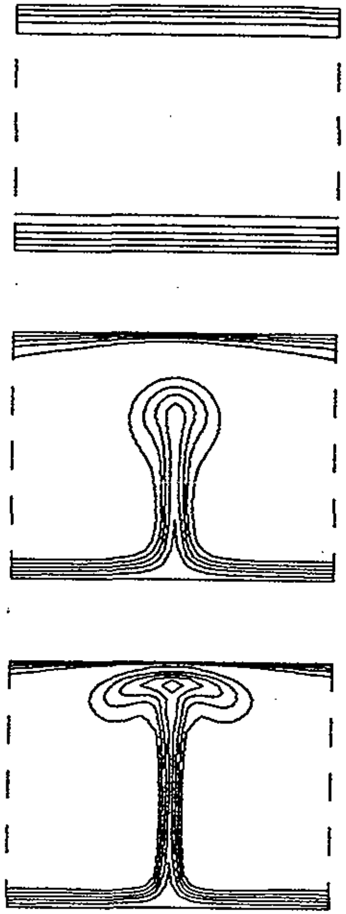


Fig.4. Three stages of development of a new ascending jet: isotherms in a *vertical* cross-section. The uppermost figure shows the interior of a previously existing convective cell, the lower boundary layer is unstable. A small perturbation triggers the new upwelling (second and third stage). The Rayleigh number is $2.5 \cdot 10^7$.

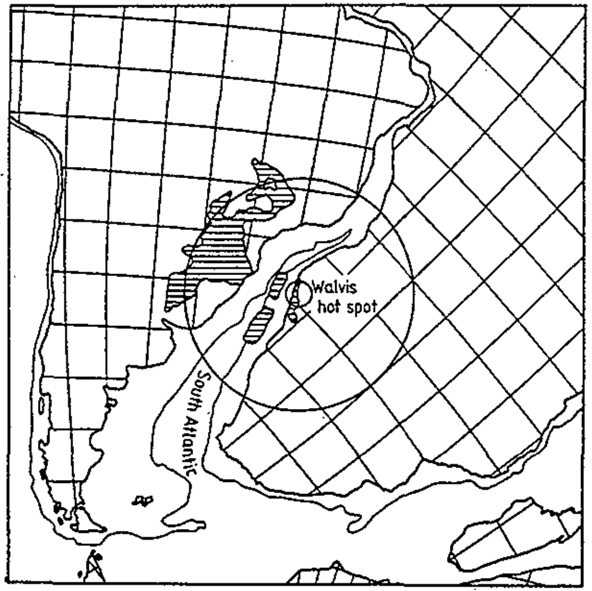


Fig.5. Reconstruction of the South Atlantic, 120 million years ago. Shaded is the region of flood basalts, production of the Walvis hot spot (White and McKenzie, 1989). The outer circle shows the extent of the mushroom-shaped hot temperature anomaly caused by the mantle upwelling (cf. Figure 4).

a thin columnar (jet-like) upwelling in the centre and a contiguous, sheet-like downwelling all around the periphery of the cell. The sheet-like nature of the descending current corresponds very well to the observed geometry of subducting oceanic slabs (Figure 1), but the long linear structure of oceanic ridges suggests a sheet-like character for the ascending current as well. The discrepancy between the simple three-dimensional models (Figure 3) and the observed reality is explained mostly (e.g. Bercovici et al., 1989) by assuming that the sheet-like upwelling beneath oceanic ridges is only a shallow and passive feature, due to the breakup of the rigid lithosphere. The real deep-rooted upgoing current may then be concentrated, jet-like and not necessarily situated beneath oceanic ridges. Indeed, high-heat-flux volcanic "hot spots" like Hawaii, Cape Verde Islands or Iceland are usually attributed to ascending jets coming from the deep mantle and they are not always connected with oceanic ridges.

On the other hand, the initial breakup of a lithospheric plate into two blocks seems to be a consequence of the arrival of a new deep-rooted jet at the base of the lithosphere (White and McKenzie, 1989). In the convection theory, this newly born jet is due to the instability of the lower boundary layer. The instability emits a rapidly rising high-temperature blob and a tail (Figure 4). The blob, arriving at the top, warms up the lithosphere, producing very high heat flow, extensive partial melting and the thinning of the lithosphere. The newly formed upgoing current can break up the thinned lithosphere by its dynamic stresses. These events can be traced at several continental breakups (e.g. Figure 5). The high heat content of the newly born jet and the consequent melting below the lithosphere causes lava extrusions in big volumes over large areas.

The scenario is then the following: A new boundary layer instability produces a rapid ascending jet which breaks up the overlying lithospheric plate into two plates, thus initiating the evolution of a new oceanic basin. The jet gradually calms down but remains a deep-rooted upwelling centre of the circulation for a long time. The surface signature of it is a long-lived hot spot in the new ocean, not necessarily at the spreading centre of the oceanic ridge. The oceanic ridge overlies a shallow sheet-like ascending current which is only a passive consequence of spreading. On the other hand, the slab-type subduction corresponds very well to the sheet-like character of downwelling found in simulations with temperature-dependent viscosity (Figure 3). This scenario is proved to work in numerical experiments with explicit presence of plates in the models (Cserepes and Christensen, 1990).

References

Bercovici, D., G. Schubert and G.A. Glatzmaier: Three-dimensional spherical models of convection in the Earth's mantle, Science, **244**, 950-955, 1989.

Cserepes, L. and U. Christensen: Three-dimensional convection under drifting plates, Geophys. Res. Lett., in press, 1990.

Richter, F.M., H.C. Nataf and S.F. Daly: Heat transfer and horizontally averaged temperature of convection with large viscosity variations, J. Fluid Mech., **129**, 173-192, 1983.

White, R. and D. McKenzie: Magmatism at rift zones: the generation of volcanic continental margins and flood basalts, J. Geophys. Res., **94**, 7685-7729, 1989.

EFFECT OF MEGATECTONIC PROCESSES ON THE EVOLUTION OF THE LIVING WORLD

B. Lukács & al. (eds.): Evolution: from Cosmogenesis to Biogenesis KFKI-1990-50/C

Árpád JUHÁSZ
Hungarian Television, Budapest

Received on 16th Oct., 1990

It has been well-known for long times that the large movements and changes of the Earth's surface considerably affected the evolution of the living world. At the end of the Paleozoic the Variscan orogeny coincided in time with the large extinction periods of the flora and fauna and with the progress of types, respectively. These phenomena were accompanied with large-scale climatic changes, as well. A similar turn can be observed at the Cretaceous/ /Paleogene boundary in the evolution of the living world, this is the period when the folding of young mountain chains has become intense. It is also known that 450 to 400 million years BP the Caledonian orogeny allowed the marine fauna and flora to adapt themselves to the terrestrial conditions on the temporarily subsiding and rising lands and to occupy the marginal parts of the continents.

It would be rather attractive to consider these surficial rearrangements as decisive reasons in the evolution of the living world. The situation, however, is by all means not so simple. The Earth was inhabited already 3 billion years BP, before the Caledonian orogeny but these populations were restricted to the oceans. The occupation of the land required the change of the atmosphere's composition, too. During the several billion years the oxygen production of marine algae increased the free oxygen content of the atmosphere and generated the ozone shield protecting the Earth surface from the ultraviolet radiation, respectively. Obviously, the life could occupy the land only in an orogenic period when it was already protected also in the lands from the ultraviolet radiation. Since that period the erosion processes of continents have been changed by the plants, the vegetation generated the soils and preserved the surface from erosion that had been very strong in the earlier periods. Subsequently, the vegetation occupied the terrestrial parts within a relatively short time in geological sense and generated flora realms during very short time span.

Undoubtfully, during the orogenic movements not only simple vertical topographic changes proceed. In these cases the boundary between the sea and land is considerably displaced, certain seas become closed, in other regions new oceans open, the salinity of the separated part basins changes, too. The Tethys fairly well exemplifies this phenomenon. Since we have known that the shore-line displacements, the rise or subsidence of the crust have proceeded also in the geological past only with a velocity of a few centimeter per year, it has become obvious that from the aspect of the living world these surficial changes could not have taken place with dramatic speed but have been slow enough for most of the species to adapt themselves to the conditions. Though in comparison with the Earth's history of 4.5 billion years the extinction phenomena proceeded suddenly, these changes involved great time intervals. E.g. at the time of the Variscan orogeny first the pteridophytes were replaced by the gymnosperms but the fundamental change of the fauna, i.e. extinction and progress of new types followed only seceral ten million years after this process. Similar time displacement can be observed in the Cretaceous in the case of both the flora and fauna, i.e. about 100 million years BP the gymnosperms were replaced by the angiosperms and the extinction of different terrestrial animal groups followed only subsequently by 30 to 40 million years, among others the extinction of dinosaurs. Their living space has been occupied very rapidly by mammals that had been subordinate till that period.

One may believe that in this case the change of vegetal food sources affected the animal kingdom. But, contemporaneously with the changes in the terrestrial animals, enormous extinction waves followed also in the marine regions. E.g. at the end of the Paleozoic the unicellular foraminifers of greater size, the trilobites, certain corals and echinoderms disappeared. In the Cretaceous the mollusc carnivorous ammonites predominating so far in the world ocean, certain wide-spread foraminifers and numerous shallow marine reef-building organisms also disappeared. The changes in the oceans can be first of all explained by the totel rearrangement of shallow marine biotopes. At the end of the Paleozoic the Pangea possessed a special position since in places of seas separating the former continental parts mountains arose. The Late Plaozoic glaciation, first of all in the southern hemisphere followed in relation with the continental drift. The closing of continents changed the former oceanic current systems, the wind systems also changed and the position of continents as compared to the poles was also modified.

The interactions between surface changes and life can be analyzed with more troubles in the early billions of years when life was restricted solely to the oceans. Uncertainities are growing what about the period of the initiation of life. It is sure that in the early life of our planet the temperature conditions did not allow the presence or formation of life. Though the first solid crust developed relatively rapidly subsequently to the total melting but at the beginning it was often broken by the glowing magma material of the deeper seated strata. In this period no fluid primitive ocean can be spoken about and the composition of the primitive atmosphere considerably differred from that of the recent one. Based on the radioactive age determinations of the first sedimentary rocks the presence of fluid water can be presumed to exist around 3.7 billion years ago. The age of the oldest pillow lava like basaltic rocks relating to ocean bottom as also 3.8 to 3.6 billion years. Since that time all the geological observations prove that the overwhelming majority of the Earth's surface was covered by oceans during billions of years. Life was generated in this medium and as mentioned above this has been its birthplace for about 3 billion years.

In this early period two factors can be taken into account that decisively affected the marine living world: the ultraviolet radiation and the crustal movements. The latter one formed continents from time to time, changed the depth of oceans, changed their salinity, modified the ocean currents, produced new shallow marine biotopes and allowed the development and evolution of reef-like assemblages close to the shores. The two effects occurred many times together since in lack of the ozone shield only thick water layer was the only protection against the ultraviolet radiation. The fixed reef-like biocenoses, however, got considerable dose of radiation during the rise of sea-shores. The oldest known alga reefs are 3.5 billion years old.

In this initial period the most decisive factor is that not only the surficial processes acted on the living world, but the living world re-acted on the evolution of the Earth generating a special atmosphere and new types of rocks, e.g. limestone and dolomite. In the Paleozoic it changed the surficial weathering and erosion and due to the decomposition of the dead organisms it produced coal and hydrocarbon beds.

It would be an error to derive the evolution, crises or flourishing of the living world mechanically from the surficial changes and to neglect its inclination to the perpetual regeneration.

We are not sure how to judge the effect of changes of the Earth's magnetic field on the living world. Paleomagnetic studies revealed that the northern and southern magnetic poles of our planet interchange within geologically very short intervals, i.e. by several hundred thousand or one-two million years. The magnetic field of the Earth decreases to the tenth at the moment of the interchange, the Van Allen belt surrounding the Earth will be deformed and does not protect the surface from the cosmic rays /the ozone shield, however, provides the the protection against the ultraviolet radiation/. These cosmic ray showers, however, do not correlate in time with the extinction or mutation of the species. Moreover, mankind itself had to undergo several cosmic ray showers, mainly when taking into account that the evolutionary way of man can traced as back as to the Miocene. Similarly, no unambiguous relationship exists between the volcanic activity and the climate. It was previously presumed that the intense volcanic activity of the Carboniferous period transported many carbon dioxide into the atmosphere and this would result in a green-house effect. The warming up together with the abundance of carbon dioxide produced extraordinarily favourable conditions for the vegetation. The rich vegetation producing the greatest coal reserves of the world is explained by this fact.

Based on instrumental measurements it seems so that in case of volcanism the fine dust erupted into the atmosphere has decisive role which as against the carbon dioxide results in the decrease of solar radiation and leads to sooling. Similar double tendencies may prevail in the course of the human activity concerning the energy production since the thermal power plants emit not only carbon dioxide but also large quantities of fine dust. By all means in the age of scientific-technical revolution mankind affects the atmosphere and the living world much more rapidly than the geological conditions of the Earth's change. The surficial changes due to mining and building activities, the rearrangement of the river network on a continental scale represent environmental changes that exceed in scale the spontaneous changes of the Earth.

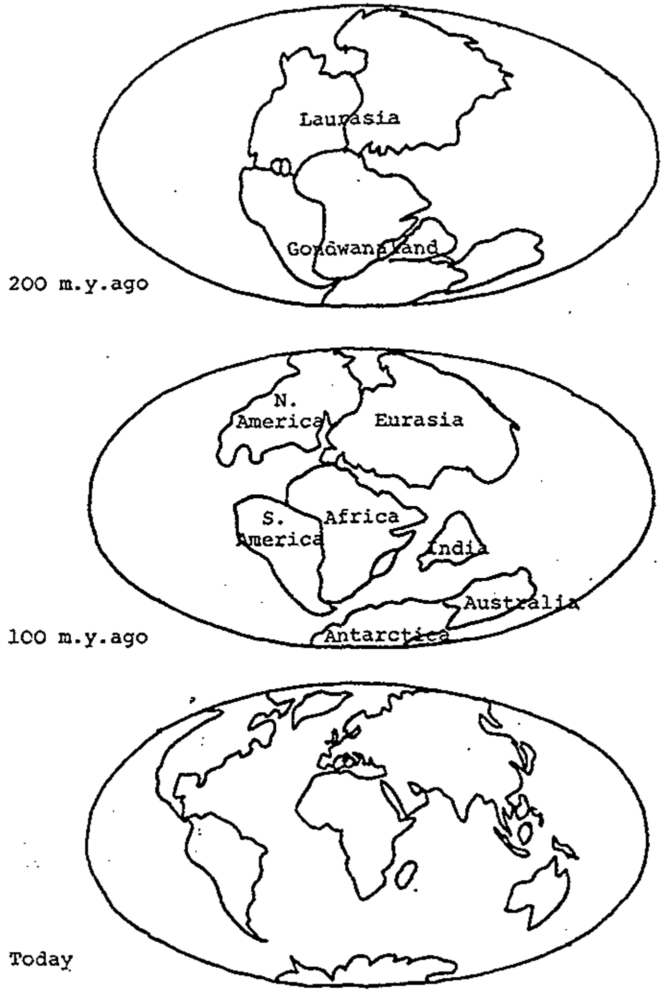


Plate movements after the breakup of Pangea /reconstruction/.

The megatectonic effects influencing the course of evolution were expressed both in macroevolution and microevolution. Megatectonic effects caused bot aggregations and isolations between ecological communities, species and populations. The megatectonic perturbations of ecological communities reorganized the chains of trophic interactions, inducing altered ecological interactions. The changing ecological contexts in turn affected species-specific and/or populational interactions, e.g. the historical sequences of the networks of predation, competition, parasitism, mutualism. symbiosis etc.

As regards the macroevolutionary effects, megatectonic motions operated in the set up of reproductive isolations between populations. It is well known that the reproductive isolation has been the overture of generation of new species. By expanding or contracting the number of individuals of populations, tectonic effects were able to affect the level of genetic and phenotypic variation in microevolution of local populations.

In brief, the emergence and extinction ,moreover the interspecific and intraspecific ecological interactions have changed during megatectonic deformations, creating renewed diversity, complexity and stability patterns. Therefore, megatectonic effects served as constraints on evolutions, opening up new evolutionary possibilities in many cases.

ε

EVOLUTION OF THE TERRESTRIAL ATMOSPHERE

Sz. Bérczi[1] & B. Lukács[2]
[1]Dept. of General Technics, Roland Eötvös University, H-1088 Rákóczi
út 5., Budapest, Hungary
[2]Central Research Institute for Physics, H-1525 Bp. 114. Pf. 49.,
Budapest, Hungary

Received on 26th Oct., 1990

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

According to several indepent evidences and arguments the primordial terrestrial atmosphere must have seriosly deffered from the present one. The reasons are as follows.

1) Earth and Sun have originated from the same matter (Bérczi, in this Volume), so, even with differential evaporation &c. molecules containing hydrogen, such as water vapour, methane and ammonia would be expected as dominant.

2) Free O_2 is not expected in equilibrium, being too aggressive.

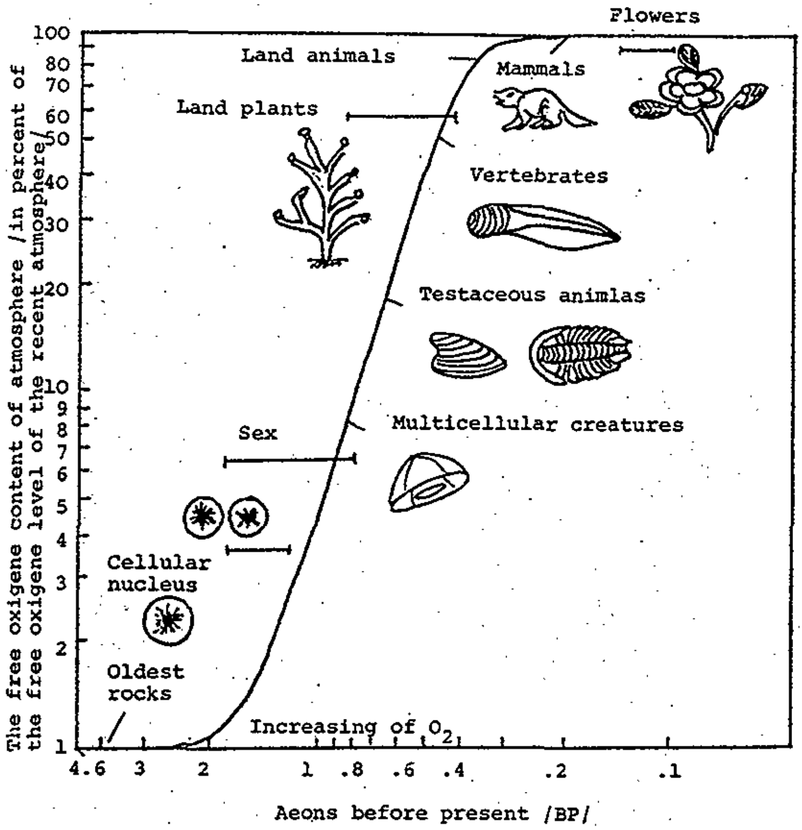
3) The present atmosphere is not friendly with organic macromolecules, so the development of early life would have been impossible.

Present living organisms can e.g. deliberate O_2 or eat up methane or ammonia. Thermodynamic considerations show (Lukács, at the end of this Volume) that Life can convert the matter content of the atmosphere on a time scale much shorter than the age of Earth. Even without organisms UV radiation produces some O_2, and astronomical observations show reducing or neutral atmospheres around the other planets.

Model calculations starting from possible primordial atmospheres can result in the present one, and some palaeonthologic data suggest the changes. Fig. 1 is such a compilation. For more explanation and further details see the standard literature, e.g. Refs. 1 & 2.

REFERENCES

[1] C. W. Montgomery: Physical Geology. Wm. C. Brown Publ., Dubuque, Iowa, 1987
[2] Hutchinson Encyclopaedia of the Earth. (P. J. Smith, ed.). Equinox, Oxford, 1981



Increasing free oxigene content of the terrestrial atmosphere.

Fig.1.

THE EVOLUTION OF VIRUSES?

A.S. Koch

Semmelweis University, Medical School, 2nd. Pathological Anatomy Department

Budapest, Hungary.

Received on 23rd July, 1990

KFKI-1990-50/C   B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

One usually tries to start a presentation with a "bon mot" to arouse the interest of the audience. Why not try once the opposite? Let us start with a platitude.

Viruses do not exist because we have discovered them but rather we could discover them because they existed. Moreover, their discovery started the other way round: Viruses did make themselves discover. They caused diseases in man, animals, plants (and even bacteria), thus attracting the attention of doctors, veterinarians, farmers and bacteriologists to their existence. And so they did in an epoch when everybody was hunting for microscopic, live pathogens, the new paradigm of medicine and biology at the turn of the past century.

No wonder, therefore that the first viruses emerged as agressive, acutely pathogenic "filterable" agents, causing epidemic outbreaks with great losses either in human lives or in animal or plant production. These viruses received a well deserved attention and it was soon discovered that as pathogenic agents they can be controlled by the classical epidemiological measures (Burnet, 1945; Nobel Prize 1960), including the active immunization of their potential hosts. In a medical-epidemiological sense they were therefore included in the group of pathogenic microbes with invasiveness, virulence and possibly toxigenicity. The classification of viruses, too, was based on their pathogenecity and host range, and they were accordingly categorized as neurotropic, enterotropic, dermotropic, etc. viruses, pathogenic for plants or animals (including humans). Some hesitation existed as to the potential viral nature of the bacteriophages.

The crucial breakthrough came when Luria (1953), Delbrück (1950) and Hershey (1952) (Nobel Prize 1969) discovered that all true viruses essentially represent an infectious genetic information. At that time it was already known that the material manifestation of inheritable information was in all biological entities the desoxyribonucleic acid, -DNA- (or in some viruses ribonucleic acid -RNA-) macromolecule. Therefore the seemingly tautologic statement of Lwoff (1960; Nobel Prize 1965) "the virus is a virus" had at that time a profund meaning. Lwoff and Tournier proposed a classification of all viruses essentially on the basis of the chemical nature of their genome and the geometry of their extracellular form, the virion. These authors also defined the characteristic biological features distinguishing viruses from all other biological entities as shown in a somewhat modified form in Table 1.

The fundamental change of paradigm in virology (in genetics and cell biology) has led to an exponential increase of knowledge in the last 40 years. This change resulted in a much less speculative and romantic approach to the problem of virus evolution. The basis of storage, coding and reading of the genetic information being universal in all biological entities, viruses were soon found to be optimal and highly versatile tools of molecular biology (including genetics and evolution).

As shown in Table 2, at present various mobile and extrachromosomal autonomous genetic entities are known and the viruses seem to represent a special evolutionary product of such "normal" structures. A somewhat modified scheme originally proposed by Jacob (Nobel Prize 1965) and Wollman (1961) gives an idea on that concept (Fig.1.)

At present state of knowledge the classical concept of viruses being on the "boderline between live and non-live" seems to be obsolete. It is known that no principle or substance exists that would confer "life" to non-live matter. It is a minimum level of structural and regulatory complexity that causes life to evolve from coacervates of regulated, cooperative catalytic functions and structures. No part of a cell is live, only the cell as an entity does live. It appears therefore that virus evolution could not start before cells had existed (but apparently started quite soon after cells came into being).

The viruses as autonomous extrachromosomal units of genetic information may potentially evolve from an already existing autonomous live unit (essentially a cell) in at least three possible ways, as follows:

1) As a result of prolonged symbiotic coexistence with successive loss of "unnecessary" genes and/or as a result of gene exchanges and mutations. Typical examples of that type of evolution are the eukaryote-prokaryote synbionts Cyanophora paradoxa and Pelomyxa palustris, and as extreme forms the mitochondria of animal cells and the chloroplasts in plant cells. That type of evolution hardly applies to viruses because of their strictly parasitic coexistence with the host cell. Certain authors nevertheless believe that evolution in this way cannot yet be definitely excluded in the case of the largest DNA viruses (e.g. members of the Poxvirus family).

2) Evolution of an independent entity from a mobile element (either transposon or episome) of a genome. The close resemblance of replicative mechanisms and some homologies of certain sequences suggest a possible evolutionary relationship between transposons and the DNA viruses replicating via reverse transcription (Hepatitis B Virus, HBV; Cauliflower Mosaic Virus, CaMV). Both HBV and CaMV would either replicate or integrate (like true episomes) and differ from a true transposon not only by their ability to integrate as multiple DNA copias (cDNA), but also, by producing virion progeny having a direct cytopathic effect on the host cell. In this light it does not seem too far-fetched to speculate that a virus emerges when a given DNA sequence looses its integrating capacity and transforms to a true, virion producing, non-integrating cytopathic virus (distantly related to a true plasmid). The possible mechanisms of transformation may be multiple mutations, frameshifts (not only in transcription, but also in splicing), gene (or sequence) losses or gains (insertion and deletion), recombinations and other natural (or occasionally induced) modifications of the genome. It should be noted that a HBV infection is nearly always chronic, with continous production of non-infectious virus-specific surface antigen (HBsAg). The integrated viral genome may occasionally induce a primary hepatocellular cancer.

Most true animal DNA viruses except those of the Poxvirus family would rarely cause a fatal illness, being mainly involved in chronic persistent infections with prolonged shedding of antigen; such infections occasionally cause (infrequently malignant) transformation of the cells. The DNA of the Epstein-Barr virus (EBV) exists in the African Burkitt lymphoma cells as an episome, replicating autonomously and producing some EB nuclear antigen, but never complete virions (abortive cycle). Transforming polyoma- or adenoviruses produce no virions either in the transformed cells (non-permissivity). Their genome (polyoma complete, adeno defective) is randomly integrated into the host genome and codes for the production of specific tumor (T) antigens. However, in many cases the integration of the papova- or adenovirus genome has no obvious biological consequences. The human adenoviruses do not appear to have any causel relationship with any human tumor, and produce a malignant transformation only in newborn rhodents. Investigations into the papilloma viruses of Australian cattle disclosed the surprising fact that the virus abundantly produced new virions while growing in the benign outgrowth (papilloma), whereas in the hairless or non-pigmented cutaneous regions of the sick animal the same virus occasionally acted as an inducer of cancer. The cancer cells, however, dit not produce virions and contained the viral genome in the form of episomal elements. The facts listed here as examples strongly suggest an evolutionary relationship of many DNA viruses with the so-called "normal" mobile genetic elements, such as transposons (yeast Tyl; Dorsophila Copia; humans Alu), processed genes (e.g. the globin gene in humans) rearranged genes (e.g. antibody diversity) and the episomal elements.

As to the possible origin of RNA viruses, let me refer to recent advances in the study of heterogeneous nuclear RNA (hnRNA), small nuclear ribonucleoprotein particles (snRNP) and processing mechanisms involved in the maturing of mRNA. It should be noted that while the prokaryotes have only type II RNA polymerase (transcribing mRNS) eukaryotes additionally have type I RNA polymerase for the synthesis of large ribosomal RNA, as well as type III RNA polymerase for the production of a great variety of diverse small RNAs, including rRNA and tRNAs.

The average viral RNA is single stranded and has a MW of 2,5-7,5 x 10^6 daltons, very much like the average classes of mature polycistronic mRNA. The viral RNA may have positive,- negative,- or ambi-sense. Positive-sense RNAs resemble mRNA, being capped at the 5'end and polyadenylated at 3'end. The negative-sense RNA viruses carry in their virions specific RNA polymerase, that makes possible the synthesis of a positive-sense (mRNA) strand. The purified RNA of such viruses is not infectious. Some members of the Arenavirus family have ambisense RNA in their genome. Some RNA viruses have a double stranded genome.

The single stranded (but diploidic) genome of the RNA tumor viruses replicated by reverse (RNA--DNA) transcription, also deserves mention. The DNA trascripts of these viruses are integrated into the host genome. If the viral genome is complete, progeny RNA tumor viruses are produced. A copy of the viral promoter, with or without some residual viral genes (defective virus) will, if inserted close to a protooncogen (cellular oncogen), give rise to malignant transformation without virion production.

Since the central step of virus replication is the production of mRNA, Baltimore (and Temin; Nobel Prize 1975) classified all viruses on this basis as shown in Fig.2. The synthesis of viral mRNA being in most cases performed by the host's machinery, it is very likely that viral structures are "familiar" to the host. In view of this it does not seem impossible that some viral RNAs may have originated from tinkering with the genetic "junk" resulting from splicing or some other processing of the hnRNA.

As alreadly noted, the genetic functions and structures of the cells and viruses are essentially similar and to some extent even "interchangeable". Thus the hypothesis lies close at hand that viruses originated from, or rather in, the cells as by-products of the maintainance and translation of genetic information. Once a self-replicating autonomous unit had come into existence, its genetic changes could easily led to the evolution of a primordial virus. Mutations, insertions, deletions, genetic shifts and drifts could than readily do the rest of the work required for the production of the entity known as virus.

The viroids (Diener, 1976), autonomously replicating small (MW = 1,1 X

10^5) naked infectious units consisting of single-stranded RNA, can be regarded as the classical type of a "primordial virus". Up to now they have been found only in plants and are not true viruses, because they do not have an extracellular (virion) form. They accumulate as protein-nucleic acid complexes in the nucleolus like the "normal" small nuclear ribonucleoproteins (snRNP) and are replicated entirely by pre-existing (or activated) host enzymes. The viroids are not translated and exist in a covalently closed circular form with several short base paired regions with alternate mismatched loops. They have some regions homologous to the 5'end of an snRNA (e.g. U1) involved in splicing. Thus it is thought that viroids may come from aberrant introns, the more so as spliced introns are themselves often covalently closed single stranded RNAs with base paired regions. Sequence similarities were also found between viroids and the ends of transposable elements of genes, as well as of some retroviral proviruses. Thus viroids may represent transcripts of such elements that have become independent of the DNA phase in their replication.

Even this brief and superficial review of current knowledge on molecular genetic events and gene transcription permits a deep insight into the complexity and abundance of those events and phenomena wich may serve as sources of autonomously replicating genetic elements. Although the details are obscure, there seems to be little doubt that episomes (dsDNA), viroids (ssRNA) and various present viruses may have evolved from "normal" elements of DNA and from various functional and "vaste" RNAs.

Last but not least it should be noted that since the laws of molecular genetics are universal, a "coated" autonomous genetic element - as is a virus - may in principle easly move across any taxonomic barrier. Evidence has been presented that viruses harboured by rhodents or other wild or domesticated animals are being transferred by arthropodes to another animal or bird. Thus enrichment or modification of the viral genome during its horizontal and vertical migration in populations of varied susceptibility or resistance appears to be another factor for guiding viral evolution. Such effects are well demonstrable for example by the continuous antigenic "evolution" of human influenzaviruses (Orthomyxovirus) and AIDS viruses. Viruses having scanty genetic markers (in comparison to bacteria or Drosophila) are or rather will become excellent tools for investigations into the molecular genetic bases of evolution.

Conclusions

The viruses have no fossile ancestors, although they seem to have evolved parallel to the begining of life on earth. Considering the extraordinary conservativism of genetic tinkering ("Evolution remembers only successes"- F. Jacob), as well as its extreme versatility ("there is nothing impossible in biology" - A. Lwoff), one would wonder rather than comprehend why virus families are relatively innumerous (Table 3). However, the surprising variability of some virus families (e.g. Myxoviruses, AIDS viruses) and the intricate interactions of the genetic machinery of viral and host genome in other families (Retroviruses, Hepadnaviruses, viroids) help the investigator to divine the possible mechanisms of the evolution of autonomous, encapsidated infectious genetic elements (viruses) from the "normal" mobile and autonomous genetic elements of the cell or even from the genetic junk available in nucleus and nucleous as a by-product of splicing or some other processing of the hnRNA.

REFERENCES

Alberts B., Bray D., Lewis J., Raff M., Roberts K. and Watson J. D.: Molecular Biology of the Cell. Garland Publ. Inc., N.Y., London (1983) pp.232-240.

Baltimore D.: Expression of animal virus genomes. Bact. Rev. 35:235 (1971).

Boyd J. S. K.: Mutation in a bacterial virus. Nature 168:994-995 (1951).

Burnet F. M.: Virus as organism. Cambridge, Harward Univ. Press (1945).

Burnet F. M., Fraser K. B. and Lind P. E.: Genetic interaction between influenza viruses. Nature 171:163-165 (1953).

Delbrück M.: Viruses. Pasadena, Cal. Inst. Technol. Bookstore. (1950).

Diener T. O.: Viroids. Trends in Biochem. Sci. 9:133-136 (1984).

Doolitte R. F., Feng D. F., Johnson M. S. and McClure M. A.: Origins and evolutionary relationships of enteroviruses. Quart. Rev.Biol. 64:1-30 (1989).

Fenner F.: The evolution of Viruses. Search 4:477-485 (1973).

Hershey A. D. and Chase M.: Independent function of viral protein and nucleic acid in growth of bacteriophage J. Gen. Physiol. 36:39-56 (1952).

Jacob F. and Wollman E. L.: Sexuality and the genetics of bacteria. Acad. Press, N.Y. and London (1961).

Lewin B.: Gene expression. Vol. 3. Plasmids and Phages. J. Wiley and Sons N.Y., London, Sydney, Toronto (1977).

Luria S. E.: General Virology. John Wiley and Sons Inc. N.Y., Chapman and Hall, Ltd., London (1953).

Lwoff A., Horne R. and Turnier P.: A system of viruses. Cold Spring Harbour Symp. Quant. Biol., Vol. 27. (1962).

Novick R.P.: Extrachromosomal inheritance in bacteria. Bact. Rev. 33:210-263 (1969).

Rapp F. and Westmoreland: Cell transformation by DNA-containing viruses. B.B.A. 458:167-211 (1976).

Sandberg A. A. ed.: The chromosomes in human cancer and leukemia. Elsevir, Holland (1980).

Shapiro J. A. ed.: Mobile genetic elements. Acad. Press N.Y. (1983).

Szathmáry E.: Towards evolution of ribozimes: Nature 344:90 (1990).

Temin H. M. and Baltimore D.: RNA-directed DNA synthesis and RNA tumor viruses., Adv. Virus Res. 17:129 (1972).

Tovarnitsky V. J. and Tikhonenko T. I.: The infective properties of viral nucleic acid. Russ. Rev. Biol. 49:15-30 (1960).

Watson J. D., Tooze J. and Kurtz D. T.: Recombinant DNA-A short course. W. H. Freeman and Co., N. Y. and Oxford (1983).

Table 1.

Differentiating features of microbes and viruses

	Bacteria[1]	Mycoplasms[1]	Rickettsia[2]	Chlamidia[2]	Virus[2]
Growth on media	+	+	+	0	0
Binary fission	+	+	+	+	0
DNA+RNA	+	+	+	+	0[3]
Ribosomes	+	+	+	+	0
Metabolism	+	+	+	+[4]	0
Antibiotic sensitivity	+	+	+	+	0
Interferon sensitivity	0	0	0	+	+
Organism	+	+	+[5]	+[5]	0[5]

1 obligate, 2 facultative intracellular parasite

3 except for hepatitis B virus (HBV)

4 various exigencies

5 filterable

Table 2.

Mobile and extrachromosomal genetic elements

Element	Nucleic acid	Autonomous replication	Reverse transcription	Integration	Capsid
Transposon	l.ds. DNA	0	+	+	0
Episome	c.ds. DNA	+	0	+	0
Plasmid	c.ds. DNA	+	0	0	0
Virus	l.ss. DNA	+	0	0[1]	+
	c.ss. DNA	+	0	0	+
	l.ds. DNA	+	0	0[2]	+
	c.ds. DNA	+	0[3]	0[4]	+
	l.ss. RNA[5]	+	0[6]	0[6]	+
	l.ds. RNA[7]	+	0	0	+
Covirus	l.ss.RNA	+	0	0	+[8]
Viroid	l.ss.RNA	+	?[9]	?[10]	-

Notes: 1 Adeno-associated viruses (AAV) integrate

2 Temperate phages (e.g.) integrate

3 Hepatitis B (HBV) and Caulimo Viruses (CaMV) replicate via reverse transcription

4 Some Papovaviruses (when causing cancer), HBV and CaMV may integrate

5 May have + or - sense genome. Some of them have segmented genome (2-8 segments)

6 Retroviridae (Oncorna viruses) have in their virions a specific reverse transcriptase and their DNA copias integrate into the host genome.

7 Single family (Reoviridae) with segmented (10-12) genome.

8 Different genes of the virus in separate capsids

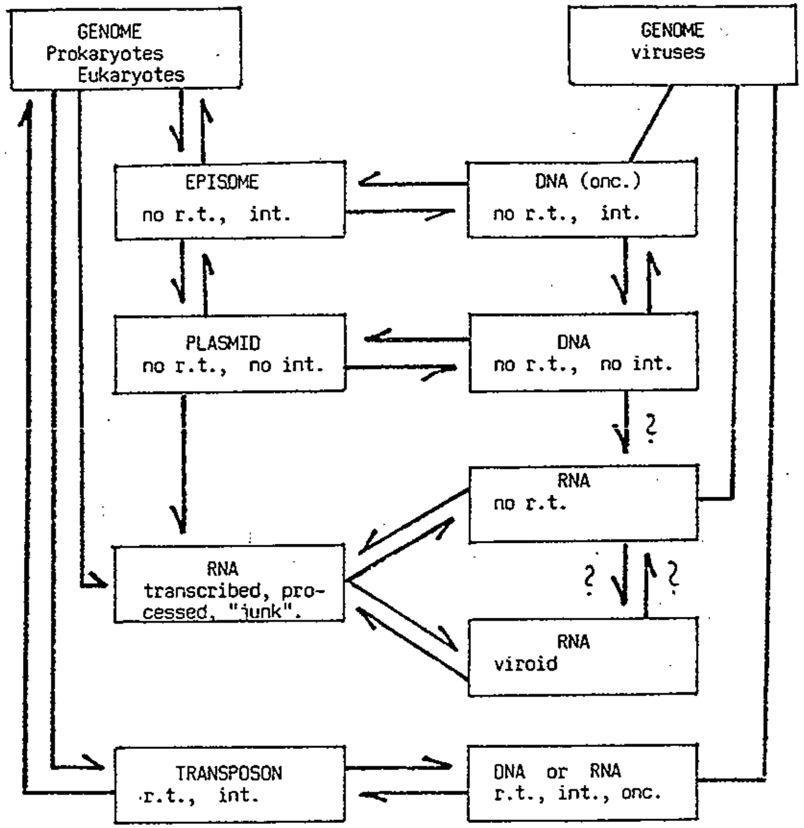
9 Not yet identified

10 Complementary DNA segments found in the host genome.

l = linear, c = circular

FIG. 1.

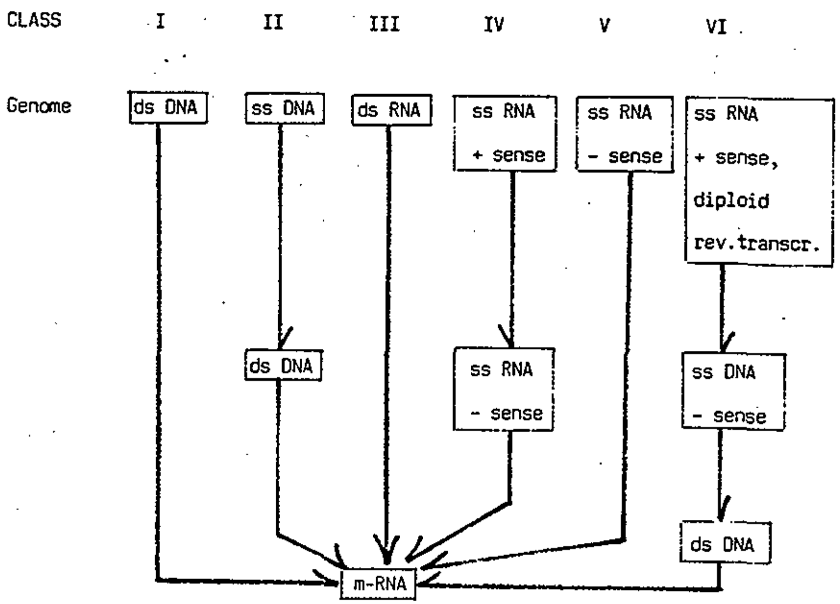
A possible flow chart of events in virus evolution.



Notes:
r.t. = reverse transcription
int. = integration
onc. = oncogenic

FIG. 2.

Classification of viruses as templates for m-RNA synthesis



ds = double stranded

ss = single stranded

Table 3.

The known Families of viruses

Total number of virus Families:			52
	Out of them:	DNA	18
		RNA	34
Number of virus Families pathogenic (also) for vertebrates:			19
	Out of them:	DNA	6
		RNA	13

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

Useful Coding Before Translation: the Coding Coenzyme Handle Hypothesis for the Origin of the Genetic Code

Eörs Szathmáry

Received on 14th Aug., 1990

Ecological Modelling Research Group, Department of Plant Taxonomy and Ecology, Eötvös University, H-1083 Budapest, Kun Béla tér 2, Hungary

Abstract. The hypothesis postulates that amino acids were first used as coenzymes of ribozymes. Similar to many contemporary coenzmyes, amino acids are thought to have been equipped with nucleotide "handles". In order to reduce the error in ribozyme-amino acid coenzyme association, different type amino acids should have been assigned to different mono- or oligonucleotides, which could have been the first appearance of a coded assignment between amino acids and nucleotides *without* translation. The three-letter code must have become fixed before the appearance of oligopeptide tracts. The emergence of oligopeptides could have triggered a bifurcation of RNA function: some ribozymes have become better messengers while losing catalytic activity; whereas others have remained as shrinking cores of ribonucleoprotein enzymes. Stereochemical bias in oligonucleotide-amino acid interaction could have facilitated the fixation of an unambiguous code, though absolute specificity (predestination) was not necessary. The assumption that coding had appeared before translation in an RNA world to improve the repertoire of amino acid cofactors renders a primordial error catastrophe loop inefficient.

As with many questions in biology, one can take here as well either a bottom-up or a top-down route of enquiery. The bottom-up strategy tries to get from elements with known properties to systems, whereas the top-down approach attempts to figure out the details considering the pattern of behaviour at the system level. In evolution, these approaches often have an additional twist since they may (but do not necessarily) refer to possible temporal order as well: it is understood that rudimentary elements precede systems not only in the structural but also in the temporal sense. Since we have a reasonably rich description of the present state of the genetic code without having ideas of commensurate clarity and detailedness about the earliest stages, I shall concentrate on the bottom-up route which, unfortunately, will not take us right to the top.

We still do not have a clue how all the complicated machinery of protein synthesis with ribosomes, tRNAs, and synthetases, could arise. While we are sure about the end of the process leading to the code, the beginning is far less definite. It is clear that basically we have two options: either an RNA world or a protein world; here I will assume the existence of an RNA world (Gilbert, 1986). Although we do not have a full picture of the origin of translation, some ingenious ideas have been put forward. Some of these are compatible with each other, others are mutually exclusive. In what follows I will outline my approach to the problem which does not include a full scenario either. Some elements of the argument will be borrowed from other hypotheses.

Coding Coenzyme Handles

I think with Gánti (1983) that the origin of translation and the genetic code must have followed a route leading from ribozymes to *isofunctional* protein enzymes. A crude picture of the process passing via ribonucleoprotein enzymes has been formed by White (1982). The details of this transition are far from trivial. The distance between two nucleotides in nucleic acids is about 0.34 nm, that between two amino acid residues in proteins is about 0.36 nm. This correspondence implies that if we have a functional ribozyme, it could not act as a messenger for an isofunctional protein using the present three-letter genetic code, since critical distances within the molecules would not be conserved (and vice versa: mRNAs of today's proteins are unlikely to behave as ribozymes). Note, however, that White's scheme does not rely on such immediate isofunctionality; short oligopeptides might have first aided the ribozymes.

I think that the first amino acids used regularly by RNA organisms were acting as coenzymes of ribozymes. Many contemporary coenzymes have nucleotide "handles" by which they could have easily been positioned in the active sites of ribozymes (Fig. 1). Amino acids cofactors must have had similar handles (Fig. 2). Thus I think that contemporary amino acid activation is a relic of the final step of an ancient process synthesizing amino acid cofactors, and the ribozymes catalyzing it were the first assignment catalysts.

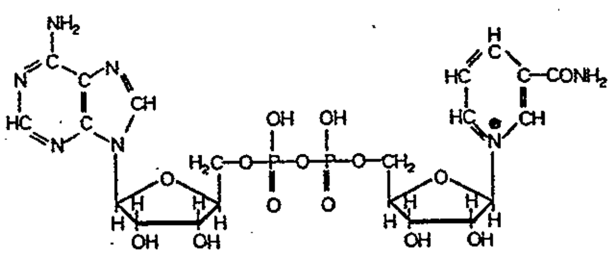


Fig. 1. The coenzyme NAD$^+$. Note the adenine residue and the 5',5' pyrophosphate anhydride bond between the two nucleotides.

The final statement can be clarified as follows. Amino acids have a wider variety of chemical functions than nucleotides, so it must have paid to use as many cofactor types as possible. Given the assumption that amino acids had nucleotide handles, four amino acid types could have been assigned to four nucleotide types. One can immediately raise the question about the advantage of such an assignment. Since a single nucleotide cannot generally be expected to be fixed by base pairing alone for a long time, both the nucleotide handle as well as the attached amino acid must have been bound to the ribozyme by secondary chemical bonds (hydrogen bridges and Van der Waals forces).

Three error types in amino acid cofactor binding are possible: (1) wrong amino acid with good handle, (2) good amino acid with wrong handle, and (3) wrong amino acid with wrong handle. While type 1 and 3 errors would severely impair catalysis, type 2 errors would be not so harmful, although the configuration of the active site would be somewhat distorted. If the four amino acid types occurred in all the 16 assignment (in which every amino acid can be paired with every handle), selective and strong binding by the ribozymes would be difficult. If each amino acid type is assigned to only one nucleotide type, distinction becomes easy, and catalytic activity is reliably increased by distinguishable coenzymes. Therefore, the most ancient code could have been a one-letter one, but it was not used to synthesize proteins in translation. Amino acid cofactors could have been used at critical points only, to aid acid-base catalysis or the formation of hydrophobic pockets, for example.

The emergence of a specific assignment can be aided by affinities between amino acids and nucleotides. It is known, for example, that hydrophobic amino acids favour A, the most hydrophobic base (see Lacey and Mullins, 1983 for review, and Lacey *et al.* 1984; Mullins *et al.* 1984; Lacey and Mullins, 1985 for further data). Any such affinity could have been amplified by the active site of the cofactor-synthesizing ribozymes. (It seems likely that a group of chemically similar amino acids was assigned to each nucleotide, in the spirit of the ambiguity reduction hypothesis; cf. Woese, 1965, 1967; Fitch, 1966).

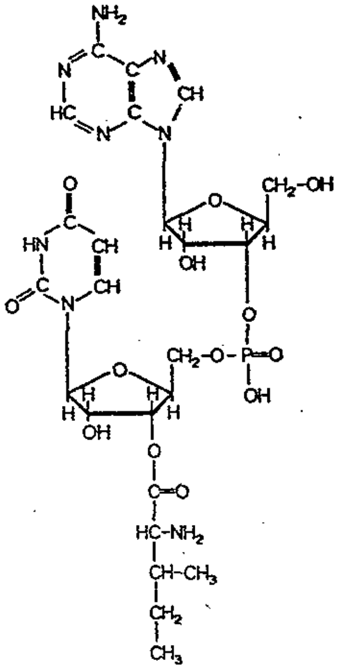


Fig. 2. The "coenzyme" isoleucine with an AU handle. Note the 3',5' phosphodiester bond between the two nucleotides.

Further increase in catalytic potential could have been achieved only through the synthesis of more coenzymes with specific handles. Handles made of two nucleotides instead of one *allow for 16 non-overlapping assignments; a dramatic increase indeed.* The essence of the former one-to-one assignments must have been preserved; if Phe had been assigned to A, it should have become assigned to AA later, and similarly for the other assignments. At this stage nearly every function of ribozymes could have been fulfilled better by proteins, at least in principle.

Although it is clear that the final configuration with a maximum of 16 *assignments is selectively advantageous,* the details of the transition need elucidation. Doublets rather than singlets can be used to encode for either entirely novel amino acids, or to reduce ambiguous assignments in a given vocabulary. In what follows I will give an example of the latter process; vocabulary extension can be visualized along quite similar lines.

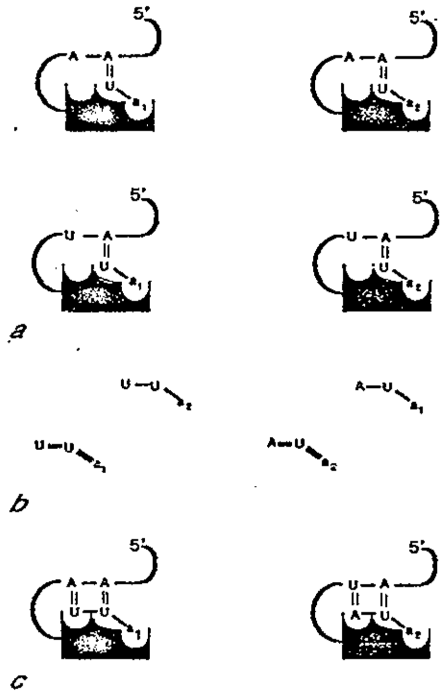


Fig. 3. The usage of amino acid coenzymes in ribozymes and the ambiguity reduction process. (*a*) U is assigned ambiguously to amino acids a_1 and a_2. (*b*) The appearance of longer handles. (*c*) Favourable nucleotide usage at the coenzyme binding sites and in the handles leads to unambiguous assignment and refined catalytic potential.

Imagine that in the one-letter "code" U is assigned to two different, chemically related amino acids a_1 and a_2 (Fig. 3*a*). One way for passing to a two-letter code is to reduce ambiguity as follows. By mutation, the relevant coenzyme-synthesizing ribozyme could charge related dinucleotides as well rather than only U with the amino acids (Fig. 3*b*). If there is place for the novel coenzymes with their larger handles to fit into the ribozymes (Fig. 3*c*), then this can happen with the continuity of function.

How one can get from ambiguous two-letter assignments (Fig. 3*b*) to unambiguous ones (Fig. 3*c*) is influenced by three conditions: (1) the relative production rate of specific assignment catalysts; (2) the usage (analogous to codon usage in present-day organisms) of dinucleotides in the ribozymes at the cofactor-binding sites; and (3) the bias (stereochemical, for example) in the assignment itself (Fig. 3*b*). I will discuss these in turn, using *mutatis mutandis* some arguments by Bedian (1982). Assume that there is no intrinsic bias in the production rate of the mutant assignment catalysts, which leaves us with factors (2) and (3), to which I will refer to as doublet and antidoublet usage, respectively, for the sake of simplicity. The crucial question is the relationship between these two, presumably biased, usages.

Table 1

Possible relationships between a fixed coenzyme-handle antidoublet and coenzyme-binding doublets.

(a)

	a_1	a_2
U		+
A	+	

(b)

	a_1	a_2
U		+
A	+	

(c)

	a_1	a_2
U	+	+
A		

(d)

	a_1	a_2
U		
A	+	+

(e)

	a_1	a_2
U	+	
A		+

(a) Antidoublet configuration. (b–e) Doublet configurations. For doublets, + indicates a bias towards the novel coenzyme-binding nucleotide in ribozymes working better with amino acid a_1 and in those performing better with amino acid a_2. For the sake of simplicity, only U and A are included in this example.

Let us imagine that antidoublet usage is fixed. This is justified because it is known that the bias in amino acid-dinucleotide interaction is more pronounced than that in amino acid-mononucleotide interaction (reviewed by Lacey and Mullins, 1983; Shimizu, 1987) There are four possible doublet usage patterns (Table 1). Case *b* is in complete agreement with the antidoublet usage *a*, hence fixation of unambiguous coding is easily driven by improved catalytic potential. Cases *c* and *d* are partly in conflict with antidoublet usage, since either ribozymes working better with a_1 or a_2 have the *same* nucleotide in the novel doublet position. The assignment (say, for a_1), where the two usages agree can be utilized easily, the other antidoublet cannot compete successfully beacuse of mispairing at the novel position. This is not so with the other type of ribozymes needing a_2 in their active site. There the handle of the correct amino acid pairs poorly, whereas that of the wrong amino acid pairs perfectly. The situation is not hopeless, however. Continuity of function is maintained through two facts: one, that the initial simple handle with the ambiguous assignment can still pair; second, that one amino acid used for the other has been customary in the initial phase itself. Nevertheless, mutations altering unfavourable doublets at the critical site will be at a selective advantage and thus will ultimately get fixed.

Completely opposite biases for doublet (Table 1e) *versus* antidoublet usage leads into an impasse; evolutionary progress becomes possible when either one is altered, the first by mutation pressure, the second perhaps by other mutant synthesizing ribozymes modifying the initial assignment bias.

From Amino Acid Coenzymes to Proteins

Longer tracts of neighbouring amino acids would have later been formed on ribozymes. Ribozymes which could bind these tracts only through the handles would be at an advantage since they would be shorter. The enclosing cell would replicate faster, too. It could have been at this stage where handles must have been transformed to ancient adaptors (primordial tRNA analogues), which could bind not only to the ribozyme by base pairing but also to each other (cf. Orgel, 1989; Smith and Yarus, 1989). In this way parts of ribozymes binding the amino acids rather than the handles directly became dispensable. The fixation of the triplet code presumably occured at this stage: two nucleotides were required for coding, and the third nucleotide served as a kind of "spacer" between adjacent adaptors, to be used for coding only later (the "two out of three" hypothesis; Lagerkvist, 1986). This coding frame would have become frozen by the formation of the first oligopeptide-like tracts.

The origin of adaptor molecules, and hence the transition from at most trinucleotide handles to ur-tRNAs, is not clear. They may have been recruited from molecules serving other roles previously. An interesting suggestion of Weiner and Maizels (1987) is of this genre: they think that ancient tRNA-like molecules tagged the genomic RNA molecules for replication. Removal of a tag led to molecules acting as catalysts only; free tags could have been charged with amino acids. But apart from this particular example, it seems very likely that an organism based on RNA catalysis contained all sorts of smaller and larger oligoribonucleotides for various reasons; some of them could have been only debris. Fortuitous charging of any suitable molecule would have led to the emergence of adaptors.

At this stage specificity of activation by different ribonucleoside-triphosphates served no coding function any more; today amino acids are activated by ATP only (note the difference in the binding of amino acids to AMP *versus* tRNA: the former is on the 5', the latter on the 3' C atom of ribose). If self-sustaining oligopeptides could have been formed from these aligned amino acids, this would have served with the benefit of making adaptors reusable. Here a differentiation of ribozyme genes is to be expected: some versions of the original gene produced RNAs which became more efficient messengers while losing catalytic ability, whereas other descendants were still involved as subunits in functional ribonucleoprotein enzymes. Such a division of labour must have been advantageous, since assembly of functioning enzymes would have been faster and messengers would have become reusable. Ribozymes catalyzing peptide synthesis would have been advantageous, and become the ancestors of ribosomal RNAs. (It should be noted that workers in the field begin to think that crucial steps in protein synthesis are catalyzed by ribosomal RNAs rather then proteins, and in particular the formation of the peptide bond occurs spontaneously, once the two tRNAs are positioned suitably, see Moore, 1988; Dahlberg, 1989). Thus we have got coding, messengers, adaptors, and RNAsomes (primordial ribosomes). The problem is that oligo- (or poly-) peptides are synthesized in stretches, and no complete messengers exist. In order to replace RNA entirely the cell must have solved two problems: the size problem as explained above (complete messengers must be considerably longer than ribozymes), and the problem of forming a single polypeptide chain acting as an enzyme.

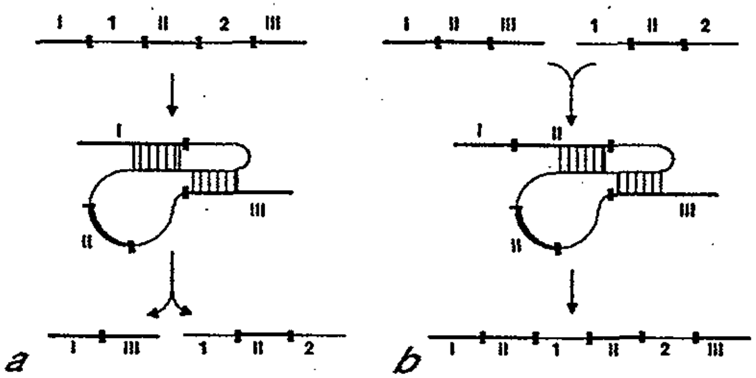


Fig. 4. Exon duplication by reverse self-splicing of an RNA transposon. Roman and arabic numerals stand for exons and introns, respectively. (*a*) Alternative self-splicing leads to a deficient spliced product and an RNA transposon containing exon II. (*b*) Reverse self-splicing of the transposon into a fully spliced gene product, resulting in the duplication of exon II. See Augustin *et al.* (1990) for experimentally demonstrated reverse self-splicing.

Ribozymes themselves are likely to have been formed only after the regular self-excision of introns. Moreover, exon shuffling was presumably possible through the activity of RNA transposons, utilizing the ability of reverse self-splicing (Gilbert, 1986). Exons in the RNA world could have been functional cassettes like stem and loop structures (hairpins; cf. Cedergren and Grosjean, 1987). As shown on Fig. 4, these component processes can lead to exon duplication, which in turn results in longer potential messengers, while keeping the overall chemical character of the single exon in the duplicated region. Repeats of oligonucleotide modules can have peculiar features from the point of view of translation. If the reading frame (having no interrupting stop

Fig. 5. The growth and coding capacity of oligomeric repeats (Ohno, 1987). (*a*) The pentamer is elongated by unequal pairing and primed replication. (*b*) The elongated RNA codes for a polypeptide chain with the pentapeptidic repeat Ser-Lys-Ala-Lys-Gln in *all three* reading frames. This is due to the fact that the initial oligomer is a pentanucleotide. (*c*) The one nucleotide longer hexanucleotide repeat can code only for dipeptidic repeats, or or may be aborted by terminator codons (ter). Thus, ancient RNAs could have been made of oligoribonucleotide modules whose sizes are not integer multiples of three.

codons) is made of triplets, then concatenated modules can be correcly read in any frame provided the length of one module differs from an integer multiple of three (Ohno, 1987; Fig. 5).

Before the appearance of messengers, starts and ends of sequences harbouring amino acid tracts were marked by the entire structure of ribozymes. However, translation on messengers needs start and stop signals. While I have no idea why the codon of methionine was chosen as an initiation signal, choice of stop codons may have had two different reasons. There is the possibility that the assignment of any amino acid to triplets now serving as terminators was unfavourable for some physico-chemical reason. On the other hand, it is peculiar, as Senapathy (1988) has noticed, that the splicing signals in eukaryotic pre-mRNAs (and group II introns; cf. Cech, 1986) are closely related to terminator codons. He therefore argues that splicing signals emerged from stop codons. Considering, however, the likely possibility that self-splicing and exon shuffling preceded protein synthesis, I think that a better interpretation is just the opposite: stop codons evolved from ancient splicing signals. This had two advantages: useless introns were not translated, and sequences prone to frequent enzymatic cutting were not included as messengers for protein synthesis. The contiguity of coding tracts of messengers could have been achieved by mutations of nonsense to sense codons, and aided by the information-preserving feature of oligonucleotide repeats (cf. Fig. 5) .

The crucial assumption in my incomplete scenario is that of a continuous development of enzymatic activity: because each gene was present in several copies, experimentation with new variants was possible without the risk of losing function entirely. The original ribozyme served as a shrinking core or scaffold, onto which the novel protein structures assembled. The ribonucleoprotein world has been a transient one, but memories of it are kept in the few contemporary examples of functional ribonucleoprotein complexes, including as important ones as ribosomes (Moore, 1988).

Discussion

Amino Acid and Other Coenzymes

As it is apparent from Figs 1 and 2, there are two crucial differences between the structure of existing nucleotide coenzymes and that of the postulated amino acid coenzymes. First, the coupling of the two nucleotides in NAD, for example, differs from that in nucleic acids. Second, the non-nucleotide part is often bonded to the 5' rather than to the 3' C of ribose. The second feature is true for the aminoacyl-AMP (not a coenzyme but the activated form of amino acids) molecules and S-adenosylmethionine.

The current absence of 3',5' phosphodiester dinucleotide coenzymes can be a ghost of the past competition on behalf of amino acid coenzymes with coding handles; no molecules but these were allowed to mimic the structure of nucleic acids and thus have been enabled to associate with ribozymes through conventional base pairing; an option not available to NAD, for example. However, one must be careful with this argument since one cannot substantiate a theory by a predicted *absence* of a phenomenon.

The binding of the amino acid to the 3' (ester bond) rather than to the 5' (anhydride bond) C of ribose could have served a distinguishing role between activation and coenzyme usage.

Although amino acid coenzymes from contemporary organisms are not known, it is interesting that Gly-tRNA and Ser-tRNA are used for peptidoglycan synthesis (see Inouye and Delihas, 1988) and Glu-tRNA is required for chlorophyll synthesis in chloroplasts (Schön *et al.* 1986). It may well be true that these non-coenzymatic activities have been present in ribo-organisms also (cf. Benner *et al.* 1989) utilising amino acid coenzymes with coding handles.

Note that the coding handle hypothesis is independent from the assumed intermediate stages in my scenario of the ribozyme-protein enzyme transition; it is compatible with the deletion-replacement alternative (Benner *et al.* 1989) as well.

Frozen Accident versus Stereochemistry

The nature of the amino acid-codon assignments would, according the above scenario, not follow the "fozen accident" hypothesis (Crick, 1968). Sometimes people think that it states that the assignments are completely random, which is wrong. In reality, the hypothesis maintains that, while codon configurations should be resistant to mutations, and similar codons should code for similar amino acids, there is no chemical relationship between the cognate amino acid and the corresponding anticodons (or codons). The fact that aminoacyl-tRNA synthetases rarely recognize the identity of the correct tRNA by the anticodon (Yarus, 1988) seems to be in accord with this. Historical accidents in the choice properties of ancient assignment catalysts would thus have been decisive in shaping the code. This view was challenged repeatedly.

Woese (1965, 1967) coupled the ambiguity reduction hypothesis with the *so-called stereochemical one in thinking that the anticodons of ancient tRNAs* stereochemically recognized the cognate amino acid, similar to the lock-and-key relation between substrate and enzyme. Obviously, this fit could not have led to very accurate disrimination, hence the ambiguity. According to this view, at least overall chemical similarity, if not steric complementarity, between amino acids and anticodons should be found. There exists such a significant relationship between anticodon and amino acid polarities (Weber and Lacey, 1978; Jungck, 1978).

The usual objection is that the measurements (reviewed by Lacey and Mullins, 1983) have been taken under highly non-biological conditions. The snag is that for amino acids, measurements in various 'non-biological' systems used to correlate nicely with each other and with the "mutation ring" of Swanson (1984). The most promising approach is due to Shimizu (1982), who demonstrated that there is in fact a lock-and-key relationship between space-filling molecular models of a complex consisting of the anticodon and the so-called disriminator base of tRNA, and the cognate amino acid. *In vitro* measurements made for some of the suggested complexes (Shimizu, 1987; Yoneda *et al.* 1985) show that the binding is real, although not very strong, and seems to be discriminating enough. The fact that the anticodon and the discriminator base are on opposite ends of tRNAs is obviously difficult to explain on this ground. Perhaps ur-tRNAs were considerably shorter (as in the scenario presented previously), and the tail folded back on to the anticodon, to form the necessary complex.

It may be surprising that such a smallish complex of ribonucleotides could discriminate among the amino acids. Note, however, that due to steric relations, even in a sizeable ribozyme the number of bases directly repsonsible for binding is unlikely to exceed four anyway. The rest is needed to reduce the fluctuation of the binding site configuration. I have proposed a crucial experiment to test (Szathmáry, 1989) based on the suggestion (Szathmáry, 1984; 1990) that different ribozymes could be produced by *in vitro* RNA breeding. Breeding RNAs able to bind amino acids seems equally possible. After sufficient experimentation, a whole array of different RNAs, each specific for a particular amino acid, is to be expected. Different RNAs binding the same amino acid could also emerge. Then one could test for the significant occurrence of anticodon-like structures in the selected molecules. Until then, one is left with the feeling that the complementarity demonstrated by Shimizu (and by others for related models) could have hardly arise by chance alone.

Yet is must be emphasized that although stereochemical bias enhances the evolution of an unambiguous code, strict "biochemical predestination" is by no means necessary for this (Bedian, 1982).

The Inverse Problem of the Error Catastrophe

Usage of amino acid coenzmyes with a dominantly ribozyme background has an important consequence for the 'error catastrophe' problem. Orgel (1963, 1973) has *called attention to the problem that translation errors, if going beyond* a certain frequency, can initiate a runaway process resulting in ultimate death of the cell, since erroneously produced translation proteins will make more translation errors. Hoffman (1974) realized that there was an *inverse* problem in the origin of life: given a crude enzymatic activity and low selectivity, how can a rudimentary system get off the ground? Note that, in agreement with Bedian's (1982) observation, the number of critical sites should not be too high. This condition is easily satisfied in an RNA world, where amino acids are used in important cases as coenzymes, but there is only a very faint, if any, error loop since proteins as such are lacking at this stage.

Acknowledgement. I thank István Scheuring and Mária Újhelyi for useful discussion. This work was supported by the Hungarian National Research Fund (OTKA).

References

Augustin, S., Müller, M.W. and Schweyen, R.J. (1990): Reverse self-splicing of group II intron RNAs *in vitro*. *Nature* **343**, 383-386.

Bedian, V. (1982): The role of assignment catalysts in the origin of the genetic code. *Orig. Life* **12**, 181-204.

Benner, S.A., Ellington, A.D., and Tauer, A. (1989): Modern metabolism as a palimpsest of the RNA world. *Proc. Natl Acad. Sci. USA* **86**, 7054-7058.

Cech, T.R. (1986): The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* **44**, 207-210.

Cedergren, R. and Grosjean, H. (1987): On the primacy of primordial RNA. *BioSystems* **20**, 175-180.

Crick, F.H.C. (1968): The origin of the genetic code. *J. Mol. Biol.* **38**, 367-379.

Dahlberg, A.E. (1989): The functional role of ribosomal RNA in protein synthesis. *Cell* **57**, 525-529.

Gánti, T. (1983): The origin of the earliest sequences (in Hungarian). *Biológia* **31**, 47-54.

Gilbert, W. (1986): The RNA world. *Nature* **319**, 618.

Hoffman, G.W. (1974): On the origin of the genetic code and the stability of the translation apparatus. *J. Mol. Biol.* **86**, 349-362.

Inouye, M. and Delihas, N. (1988): Small RNAs in prokaryotes: a growing list of diverse roles. *Cell* **53**, 5-7.

Jungck, J.R. (1978): The genetic code as a periodic table. *J. Mol. Evol.* **11**, 211-224.

Lacey, J.C.Jr. and Mullins, D.W. (1983): Experimental studies related to the origin of the genetic code and the process of protein synthesis. *Orig. Life* **13**, 3-42.

Lacey, J.C.Jr., Mullins, D.W.Jr. and Khaled, M.A. (1984): The case for the anticode. *Orig. Life* **14**, 505-511.

Lacey, J.C.Jr. and Mullins, D.W. (1985): Genetic coding catalysis. *J. theor. Biol.* **115**, 595-601.

Lagerkvist, U. (1986): Evolutionary aspects of unconventional codon reading. *Chemica Scripta* **26B**, 85-89.

Moore, P.B. (1988): The ribosome returns. *Nature* **331**, 223-227.

Mullins, D.W.Jr., Senaratne, N. and Lacey, J.C.Jr. (1984): Aminoacyl-nucleotide reactions: studies related to the origin of the genetic code and protein synthesis. *Orig. Life* **14**, 597-604.

Ohno, S. (1987). Evolution from primordial oligomeric repeats to modern coding sequences. *J. Mol. Evol.* **25**, 325-329.

Orgel, L.E. (1963): The maintenance of the accuracy of protein synthesis and its relevance to ageing. *Proc. Natl Acad. Sci. USA* **49**, 517-521.

Orgel, L.E. (1973): Ageing of clones of mammalian cells. *Nature* **243**, 441-445.

Orgel, L.E. (1989): The origin of polynucleotide-directed protein synthesis. *J. Mol. Evol.* **29**, 465-474.

Schön, A., Krupp, G., Gough, S., Berry-Lowe, S., Kannangara, C.G. and Söll, D. (1986): The RNA required in the first step of chlorophyll biosynthesis is a chloroplast glutamate tRNA. *Nature* **322**, 281-284.

Senapathy, P. (1988): Possible evolution of splice-junction signals in eukaryotic genes from stop codons. *Proc. Natl. Acad. Sci. USA* **85**, 1129-1133.

Shimizu, M. (1982): Molecular basis for the genetic code. *J. Mol. Evol.* **18**, 297-303.

Shimizu, M. (1987): Specific interactions of dinucleoside monophosphates with their cognate amino acids. *J. Phys. Soc. Japan* **56**, 43-45.

Smith, D. and Yarus, M. (1989): tRNA-tRNA interactions within cellular ribosomes. *Proc. Natl Acad. Sci. USA* **86**, 4397-4401.

Szathmáry, E. (1984): The roots of individual organization (in Hungarian). In: G. Vida (ed.) *Evolution IV. Frontiers of Evolution Research.* pp. 37-157. Natura, Budapest.

Szathmáry, E. (1989): The emergence, maintenance, and transitions of the earliest evolutionary units. In: P.H. Harvey and L. Partridge (eds): *Oxford Surveys in Evolutionary Biology Vol. 6.* pp. 169-205. Oxford Univ. Press.

Szathmáry, E. (1990): Towards the evolution of ribozymes. *Nature* **344**, 90.

Swanson, R. (1984): A unifying principle for the amino acid code. *Bull. Math. Biol.* **46**, 187-203.

Weber, A.L. and Lacey, J.C.Jr. (1978): Genetic code correlations: amino acids and their anticodon nucleotides. *J. Mol. Evol.* **11**, 199-210.

Weiner, A.M. and Maizels, N. (1987): tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: implications for the origin of protein synthesis. *Proc. Natl Acad. Sci. USA* **84**, 7383-7387.

White, H.B.III (1976): Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* **7**, 101-104.

White, H.B.III (1982): Evolution of coenzymes and the origin of pyridine nucleotides. In: *The Pyridine Nucleotide Coenzymes.* pp. 1-17. Academic Press, New York.

Woese, C.R. (1965): On the evolution of the genetic code. *Proc. Natl Acad. Sci. USA* **82**, 1160-1164.

Woese, C.R. (1967): *The genetic code.* Harper & Row, New York.

Yoneda, S., Shimizu, M., Fujii, S. and Go, N. (1985): Induced fitting between a complex of four nucleotides and the cognate amino acid. *Biophys. Chem.* **23**, 91-104.

ON PHYLOGENESIS

Sz. Bérczi[1], B. Lukács[2] & I. Molnár[3]
[1]Dept. of General Technics, Roland Eötvös University, H-1088
Rákóczi út 5, Budapest, Hungary
[2]Central Research Institute for Physics, H-1525 Bp. 114. Pf.
49., Budapest, Hungary
[3]Dept. of Genetics, Roland Eötvös University, H-1088 Múzeum krt.
4a., Budapest, Hungary

Received on 26th Oct., 1990

ABSTRACT

A general scheme is outlined for the evolutionary history of organisms. For this history evidences are of double origin: dated fossils, and biochemical, genetical and morphological inferences from extant organisms.

1. INTRODUCTION

An interdisciplinary symposium may open and reveal new directions of explorations although it focus on classical problems. One of the oldest framework of description in biological evolution is the phylogenetic tree. In our paper we intend to switch on new lamps to illuminate this classical framework in order to find a more detailed structure of it from different viewpoints.

One important viewpoint is: hierarchy. Hierarchy emphasizes a basic background information of investigations: the structures, as WHOLE-s are formed on multi-layered (multi-embedded) series of lower structures. The second important viewpoint follows from the first one: it is the distinction and separation of evolutionary informations belonging to different hierarchy levels. The consequence of this viewpoint-distinctions is the assumption: different layers of structural hierarchy may exhibit different dominant evolutionary mechanisms and tendencies.

Hierarchy levels are the results of equivalency classifications on structures. Structures in a common range according to their space, time and some other basic parameters form a class which has a representative in the hierarchy table of "stable structures" (Bérczi, this volume). The main trends of evolution of matter can be comprehended by these "stable structures" in the "Cosmo", "Micro", and "Life" blocks (ladders) of hierarchy. The hierarchy structure of evolutionary series suggested two different inter-hierarchy processes which alternately follow each other. One is the differentiation, the competition period between equal-hierarchy structures; the other, the building together into a more complex structure is the structuralization period during evolution.

If we use these principles to the evolution of living world we may separate three basic structural levels of hierarchy. These three levels are the following ones: unicellular level, multicellular level and the level of creatures living together in ecological systems. Different equivalency-groups on these hierarchy levels are: KINGDOMS on unicellular level, PHYLA on multicellular level and ECOSYSTEMS on multi-creatures level.

In our paper two of these three hierarchy levels are discussed: those of phyla and Kingdoms. Overview of classical trees of phyla and discussions of the basic principles and results from the correlations of fossil remnants according to their morphology, sites of finding and communities over correspondences between layers and fossils, layers and their radioactive material content till the parallel exact background time scales and their mutual correspondences, and finally the division of phylogenetic and Kingdom-genetic main processes will be given.

We shall see that on different levels of hierarchy the dominant processes follow different main evolutionary tendencies. During the formation and development of unicellular creatures symbiotic processes, the building together into a more complex unit programs were the dominant tendencies of evolution, while later in the multicellular evolutionary period the morphological-functional-topological differentiation of the phylogenetical units (the branching of the tree of phyla) was the dominant evolutionary trend. During the period of the (here not discussed) ecological organizational level also the building together tendency became dominant.

It is important to note that focusing on hierarchical parting and the corresponding tendencies of evolutionary processes is not a strict and exclusive characteristic, of course. Embedding of structures in a hierarchy implies that the different tendencies operate parallel with each other on the different working hierarchy levels, but with different evolutionary speeds. Hierarchy levels and the corresponding dominant evolutionary tendencies represent a method which helps to emphasize those aspects of evolution, which can not be shown in the classical frameworks. Fig. 1. shows both the classical phylogenetic tree and the division of Kingdom evolutionary region with building together tendencies (and the two neighbourhood hierarchy levels) and gives a map to connect different evolutionary regions discussed.

2. THE EVIDENCE OF FOSSILS

The data of fossils, as we are going to see, establish an ensemble of biological events in a 3 dimensional pseudo-Riemannian (surface-time) continuum. To see this, remember that the actual fossil, which is now being unearthed at the geographic coordinates θ, φ at depth z, died at some time t, when the geographic coordinates of the same site were the different values Θ, Φ (cf. the motion of litosphere block earlier referred as "wandering of poles"). This seems to suggest a coordinate transformation law

$$\begin{aligned} \theta &= \theta(\Theta, \Phi, t) \\ \varphi &= \varphi(\Theta, \Phi, t) \\ z &= z(\Theta, \Phi, t) \end{aligned} \qquad (2.1)$$

in the domains

$$0 \le \theta, \Theta < \pi, \quad 0 \le \varphi, \Phi < 2\pi, \quad t_0 \le t \le t_{present}, \quad z \le 0 \qquad (2.2)$$

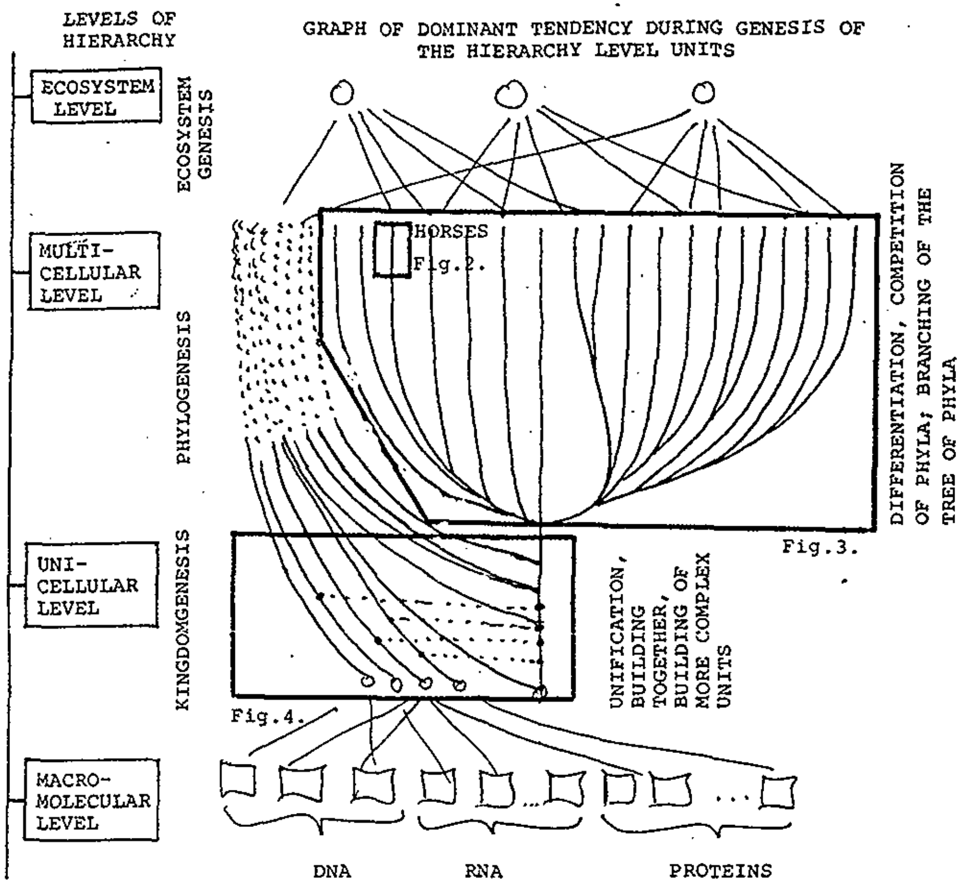
where $t_0 \approx -4.6$ Gys.

We shall see that the functions (2.1) do not exist on the *whole domain*. However, they may exist on finite domains, and this existence will be discussed. Taking first the existence granted, now we are going to outline the principles of the determination of the functions.

First the motion of the litosphere blocks. Consider a rigid block. On it we identify a set of sites

$$\theta_\alpha, \ \varphi_\alpha, \ \alpha = 1 \ldots \nu$$

Between the sites we can measure $\frac{1}{2}\nu(\nu+1)$ distances $s_{\alpha\beta}$.
These are the *present* data; however on a *rigid* block all the distances must be preserved. These constraints seriously restrict



Summarizing map to the Phylogenesis. The summary may be shown by using the structural hierarchy of "stable structures" /relative stability according to the building up of the next hierarchy level/, left column. In this paper two levels of hierarchy are discussed in more details: that of classical phyla and that of kingdom level of unicellular creatures. Tendencies of evolution are different in these two levels, but this alternation of the tendencies between "building together" and "differentiation" is characteristical feature of the evolution of metter on an universal scale.

Fig.1.

the possible motions, and the *differential* laws are simple. Consider a motion in the coordinate space x^i:

$$x^{i'} = x^i + K^i(x^k)dt \tag{2.3}$$

where K^i is the vectorial field describing the shifts at different places. The distance to be kept is that of the sphere

$$ds^2 = R^2(d\theta^2 + \cos^2\theta d\varphi^2) = \\ = g_{rs}dx^r dx^s \tag{2.4}$$

where g_{ik} is the metric tensor [1]; the Einstein convention means summation for indices occurring twice, up and down. (R is the radius; in the usual but not geographic convention θ differs with an additive $\pi/2$.) Then the condition that motion (2.4) preserve the distances is

$$K_{i;k} + K_{k;i} = 0 \tag{2.5}$$

which is the Killing equation, and ; stands for covariant derivatives [1]. Solving the equation for the above geometry one obtains

$$d\theta = \{a(t)\sin\varphi + b(t)\cos\varphi\}dt$$
$$d\varphi = \{[a(t)\cos\varphi - b(t)\sin\varphi]tg\theta \\ + c(t)\}dt \tag{2.6}$$

So 3 independent possible motions exist, with SO(3) commutation law [2], obvious for the 2-sphere. The global law is more complicated, as will be immediately seen.

As for the depth z, except for exceptional circumstances there is continuous sedimentation, e.g. the depth is monotonous with time. One can write

$$dz = \zeta(\theta, \varphi, t; p)dt \tag{2.7}$$

where ζ is the rate of sedimentation, and p stands for the parameter(s) of some local circumstances. The dependence on geographic coordinates is for climate, snow, &c.

Now we are going to look for the ways of determining the actual transformations.

3. THE MOTION OF LITOSPHERE BLOCKS

Assume first that there is no problem with dating, and that we can identify a rigid block. The first problem is relegated to the next Section; the second is automatically checked by the results. We have a set of *synchronous* sites from t_0, with the actual coordinates

$$x_\alpha{}^i = \{\theta_\alpha, \varphi_\alpha\}, \quad \alpha = 1\ldots\nu \tag{3.1}$$

and we are looking for

$$\{\Theta_\alpha(t_0), \Phi_\alpha(t_0)\}$$

In the motion $\frac{1}{2}\nu(\nu+1)$ distances are invariant:

$$s_{\alpha\beta}(\theta, \varphi) = s_{\alpha\beta}(\Theta, \Phi) \tag{3.2}$$

For definiteness' sake we give the distances on the sphere:

$$s_{\alpha\beta} = R^2\{\mathrm{arctg}[\sin^2\Theta_\alpha - C_{\alpha\beta}{}^2]^{-1/2}\cos\Theta_\alpha - \\ - \mathrm{arctg}[\sin^2\Theta_\beta - C_{\alpha\beta}{}^2]^{-1/2}\cos\Theta_\beta\} \tag{3.3}$$

where $C_{\alpha\beta}$ is determined via

$$2(\Phi_\beta - \Phi_\alpha) = \\ = \arcsin\{(1-C_{\alpha\beta}{}^2)^{-1}[(1+C_{\alpha\beta}{}^2) - 2C_{\alpha\beta}{}^2\sin^{-2}\Theta_\beta]\} - \\ - \arcsin\{(1-C_{\alpha\beta}{}^2)^{-1}[(1+C_{\alpha\beta}{}^2) - 2C_{\alpha\beta}{}^2\sin^{-2}\Theta_\alpha]\} \tag{3.4}$$

For ν sites the number of variables to be determined is 2ν, while that of equations is $\frac{1}{2}\nu(\nu+1)$. Therefore for $\nu>3$ one would expect overdetermination, i.e. no solution at all. However we have seen that there exist motions preserving all the distances: the rigid motions which have 3 degrees of freedom, being this the number of SO(3) generators [2]. The explicit formula is obtained by integrating eqs. (2.6):

$$\Theta=\arcsin\{\cos\beta\sin\theta+\sin\beta\cos\theta\sin(\varphi+\gamma)\} \quad (3.5a)$$

$$\Phi=\mathrm{arctg}\{\{-\cos\alpha\sin\beta\sin\theta+\cos\theta[\sin\alpha\cos(\varphi+\gamma)+ +\cos\beta\cos\alpha\sin(\varphi+\gamma)]\}/\{\sin\alpha\sin\beta\sin\theta +\cos\theta[\cos\alpha\cos(\varphi+\gamma)-\cos\beta\sin\alpha\sin(\varphi+\gamma)]\} \quad (3.5b)$$

The formula is rather complicated, but contains indeed 3 free parameters of the motion of the whole block, α, β and γ.

Therefore all the original coordinates can be reconstructed by measuring the present coordinates + 3 combinations of the original ones. This can be achieved if we can observe 3 original *directions* on the present sites. Neglecting the moderate difference between geographic and magnetic poles one can make use of palaeomagnetism.

If ferromagnetic materials (iron, nickel, cobalt; for compound e.g. the magnetit) are cooling through the so called Curie point (for iron 770 C°) in an external magnetic field, then by interaction the magnetic momenta of a fraction of the electrons take a common direction parallel to that of the external field [3]. Later this magnetisation remains, except for *strong* external fields or reheating. Since e.g. volcanic lava contains some ferromagnetic components, the original north direction can be traced back from some ancient rocks [4].

A direction is the relation

$$d\varphi/d\theta = a \quad (3.6)$$

The *north* direction is

$$d\varphi = 0. \quad (3.7)$$

Therefore the ancient magnetism shows $d\Phi_\alpha(t_0)=0$, whence one gets

$$\{(\partial\Phi/\partial\theta)(\partial\Phi/\partial\varphi)^{-1}\}_\alpha = a_\alpha \quad (3.8)$$

So one direction is one restriction for the transformation (3.5). Because of the 3 free parameters we need measurement at least at 3 sites; if we want to distinguish between geographic and magnetic poles 2 more data are needed. For more sites again the relations are theoretically overdetermined. Therefore if there is a solution, then it is a check for the original assumptions. Obviously because of measurement errors &c. there will be *moderate* deviations, but that problem can be overcome by fitting procedures.

If the differences are substantial, then there is somewhere a serious error with the fundamental assumptions. The most obvious possibility is that the sites originally belonged to more than one rigid blocks; then the incompatible parts are to be handled separately.

We do not go into the further details of the procedure. The moral of this Section is that, with a sufficient quantity of palaeomagnetic data, from dated synchronous sites the ancient geographic coordinates can be reconstructed. Now let us see, how to date.

4. RADIOACTIVE DATING

The firmest dating methods use radioactivity. Radioactive atoms decay with the rule

$$N(t) = N_0e^{-\lambda t} \quad (4.1)$$

where λ is characteristic to the matter (connected to the *half lifetime* τ by $\tau=0.693/\lambda$) [5], and, according to numerous experiences, is highly independent of external circumstances. The reason for this is that radioactivity happens in the *nucleus* (a slight exception is the β-decay by K-capture, where the electron shells participate too), and there the density is cca. 10^{15} times the average terrestrial one, while the specific energy corresponds to 10^{10} K. Such concentrations cannot be affected from outside.

If one knows the original quantity of atoms N_0, then from the measured N(t) one gets t using λ, determined in laboratory. This is the situation for C^{14} dating [6], since the cosmic radiation maintains a roughly constant concentration of the atmospheric radioactive carbon, continuously building into the living organisms, but decaying after death. Unfortunately for C^{14} τ=5730 ys, so the method is ineffective after *cca.* 50000 ys.

If the mother and daughter nuclei can be found together, and the daughter one cannot be expected originally (being e.g. a noble gas) then the total number of them shows N_0. The simplest such case is $K^{40} \rightarrow Ar^{40}$, where the mother nucleus is an isotope of the very abundant potassium [7].

Unfortunately abundant nuclei with $\tau<1$ Gy cannot be expected anymore in the present Earth. So substantial errors appear for the *near* past, say below 1 My. However these are technical problems.

So one can calculate the ages of *some* identifiable layers. Therefore transformation (2.4)is possible *on a lattice*; between lattice points some smooth interpolation is needed, whose details are also immaterial now.

5. STRATIGRAPHY

The solid surface of the planets consists of units: blocks and strata with 3-dimensional extension. They are surveyed by stratigraphy which puts them into proper order drawing correspondences and relations between them, identifies the course of events, the course of devastating and building up effects which formed them and roughly states their relative ages too.

On Earth's surface the various strata of deposition and sedimentation can be marked off and put in order according to their contents of fossil remnants (biostratigraphy) or up to their general physical (chemical, mineralogical, petrological) characteristics (rock stratigraphy). A series of strata identified in this way forms a time stratigraphical sequence, where the units are stratigraphical units of the geological ages. Each unit can be corresponded to a geological period on the time scale.

The basic principles of stratigraphy can be summarized as follows. For the order of the (not overturned) depositions it is true that the younger rocks are laying on the older ones, the intrusive rocks are younger than the penetrated ones and the unit cut by a fault is older than the fault itself. The extensions of recent geological layer-forming-events to the past and to the whole planetary surface are also important basic initial conditions in the use of the recognised surface forming processes in stratigraphy.

If two strata are touching, one overlaps the other, so we can easily determine their *relative* ages. If we can find a wide-spread unit, then smaller units laying on it or outlooking from it can also be dated relative to this wide-spread layer. This opportunity gives a hierarchical program to stratigraphy: let us look for such wide-spread strata on the planetary surfaces -we may call them *data-planes*- which can be followed on a great part of the whole planetary surface.

The result of the stratigraphical (global and local) mapping is an ideal sequence of stratification, in which succession of strata form a "bore log", where a relative age can be corresponded to each strata. On the other hand, independent series of morphologic evolution of living beings also can be corresponded to several strata. *Cross* correspondence of these independent processes, on different levels of evolutionary hierarchy, results in a complex dating method to stratigraphical series of planetary surface depositional units. The mutual correspondence enables us to refine absolute ages, and, more generally, to get the abstraction of time as independent variable. Feedback with this time concept gives the tool of the general time sequence on which the fossils and other fixed states of evolution of matter can be studied: the episodical remnant can be projected on the idealised background of time scale. This methodological series of abstraction, generalisation and development of concepts is an important condition of the reconstructions of evolutionary scenarios. Generalized principles of stratigraphy form one of the most important conceptional constructional background to evolution: both in the local history of the Earth and in the evolution of the Universe by extension of the stratigraphical principle to a general embedding principle for structures in the series of "stable structures", which build a skeleton to the global evolutionary history of matter.

6. THE GEOLOGICAL ERAS

An era is primarily a set of subsequent identifiable layers, and, therefore, is not *a priori* synchronous on the whole Earth. Therefore it is much more certain, which is the layer or era of the fossil than its age. However with suitable identifier organisms the synchronisation can be made probable and it can be checked by datings. The following Table gives the well established eras, some of their characteristic events or organisms, and the average dates. Not all the systems are identical (Ordovician is e.g. often classified as part of Silurian). For further details see e.g. Ref. 8.

Era or subera	Event	Beginning, Mys
Cenozoic		
Quaternary	Ice ages	-1.6
Pliocene	Permanent ice on Antarctic	-5.2
Miocene	Andesithes in Carpathians	-23.3
Oligocene	Cooling starts	-35.4
Eocene	Nummulites	-56.5
Paleocene	Divergence of mammals	-65.0
Mesozoic		
Cretaceous	End of Dinosaurs	-145.6
Jurassic	Birds	-208.0
Triassic	Early mammals	-245.0
Paleozoic		
Permian	Red Sandstone	-290.0
Carboniferous	Carbon accumulation	-362.5
Devonian	Amphibia	-408.5
Silurian	Land plants	-439.0
Ordovician	Bryozoa	-510.0
Cambrian	Outer shell appears	-570
Precambrian		
Vendian?		
Ediacara?	Ediacarans	-600?
Riphean?	Dalsland orogeny	-1650?

Table 1: Geological eras

Before Cambrian the picture is somewhat obscure. It seems that the main reason is that the hard outer shell appeared at the beginning of Cambrian, so previous organisms had much less chance to be preserved. Thus, first the sequence of fossils becomes discontinuous in time, and, second, the lack of clear identifiers makes the earlier tentative "eras" rather local. However, work is in progress in this field; some precambrian candidates are included in the Table with question marks. Possibly the best founded such one is Ediacaran.

For later purposes we give 2 more data. Earth had developed $4.6*10^9$ years ago, according to radioactive dating of different Pb ores *and* meteorites. The oldest layer in which more or less clear fossils of life forms was found is the Onverwacht and Fig Tree series. In the Barberton mountain, on the border of Transvaal and Swaziland, the Onverwacht seems to start at -3.7 Gys; the microfossils were formed maybe between -3.4 and -3 Gys [9].

So the well and continuously documented part of the 3 dimensional space-time of fossils covers cca. sixth of the history of terrestrial life, and not more.

7. PHYLOGENETIC RECONSTRUCTION BY MORPHOLOGIC ANALYSIS

Now we return to the picture where the 3-dimensional space-time continuum is filled with outward appearances of living organisms (reconstructed from fossils). Our knowledge about these organisms is restricted. We cannot reconstruct biochemical details, DNA sequence, &c., and we are lucky if any information is preserved about the nature of skin, outer shell and such. Most of the information is anatomical; roughly speaking we have a surface-time continuum of 3-dimensional photographs. But *that* we have, so let us start with this most direct *morphologic* set of data.

It was learnt in the first half of last century that life had been changing during the available past. Departing from the present the forms are becoming more and more different, and this is as well true with any other starting point [10]. So there is an evolution of terrestrial life, and it is not cyclic.

Recent organisms are classified into hierarchic taxonomical classes according to the decreasing degree of similarities. The hierarchy is very complicated, and here only the main points are to be mentioned. The fundamental unit is *species*, within which unrestricted hybridisation is possible. Therefore (without extreme geographic &c. separation) one expects serious overlaps of any characteristics of any two groups; if not, hybridization will see about in some generations. Then similar species are classified into a common *genus*, and so on; the main higher taxa are *familia*, *ordo*, *classis*, *phylum* and *regnum*. In such higher units hybridization cannot work, therefore differences between different species are not expected to be washed out.

The similarities and differences should be quantified in order to proper classification into proper taxonomical units. E.g. a smaller difference would classify two species congeneric, while bigger differences would put them into separate genera, families &c. Instead, generally decision is made on semiquantitative basis: with a sufficient earlier experience accumulated, one takes the "important" or "characteristic" differences, and estimates their degree. Of course, the method could be made (almost) purely quantitative *by a construction giving distances in the space of* morphologic data.

Consider a set of animals of the same general structure, when all the differences are those of degree. (The length of the dolphin leg may be infinitesimal, when 0 is a good approximation. However, one cannot compare limb lengths of a dolphin and an octopus up from fifth limb; there is no possibility to say that the length of the fifth dolphin limb is 0.) Let us have agreed, which are the relevant morhologic data; then for a pair αβ of species we have a pair of parameter sets

$$\{p_\alpha{}^I, p_\beta{}^I\}, \quad \alpha=1\ldots\mu; \quad I=1\ldots N$$

N being the number of parameters and μ that of species. Then for a distance $s_{\alpha\beta}$ [11]

$$s_{\alpha\beta} = s(p_\alpha{}^I, p_\beta{}^I) \tag{7.1}$$

in such a way that

$$s(p_\alpha{}^I, p_\beta{}^I) \geq 0 \tag{7.2}$$

$$s(p_\alpha{}^I, p_\alpha{}^I) \equiv 0 \tag{7.3}$$

$$s(p_\alpha{}^I, p_\beta{}^I) \equiv s(p_\beta{}^I, p_\alpha{}^I) \tag{7.4}$$

$$s(p_\alpha{}^I, p_\beta{}^I) + s(p_\beta{}^I, p_\gamma{}^I) \geq s(p_\beta{}^I, p_\gamma{}^I) \tag{7.5}$$

Now, such morphologic "distances" are indeed in use; for a review see Ref. 12. However, the problem is that intuitive distance definitions generally result in something, which is in best case only a *deviation*, not a distance, not fulfilling e.g. the triangle inequality (7.5). So here we proceed rigorously. First we concentrate on *neighbouring* points, and the result shall show its own generalisation.

Henceforth we omit the vectorial index ^{I} and instead the vectorial variables will be boldfaces. For immediate neighbouring points instead of $\mathbf{p}_\alpha$ and $\mathbf{p}_\beta$ one can use $\mathbf{p}$ (some average) and $d\mathbf{p} \equiv \mathbf{p}_\beta - \mathbf{p}_\alpha$. By eq. (7.3) $s(\mathbf{p}, d\mathbf{p})=0$ at $d\mathbf{p}=0$, and by eq. (7.2) it is a minimum. Then for $d\mathbf{p} \to 0$ s depends on a quadratic form of $d\mathbf{p}$. Then eq. (7.4) is automatically fulfilled. Now, if we want a distance *linear* in small parameter differences, then the only remaining possibility is

$$ds^2 = g_{RS}(\mathbf{p})dp^R dp^S \tag{7.6}$$

This infinitesimal distance is just the distance in a Riemann space (whose points are the possible *forms*), therefore for finite distances one can integrate up the infinitesimal distances along the shortest lines (geodesics), for the details e.g. Ref. 1.

Now, *if* we have a set of species with *reliable* distance values, *then* g can be calculated back from $s_{\alpha\beta}$ via eq. (7.6). If we have more distances than metric components, then even not quite reliable s values may give a decent g by averaging and fitting, and afterwards g gives the better distances.

Of course, lots of technical problems might arise if one tried to realize this program; e.g. the number of components of g may be quite remarkable. However, serious simplifications may also occur. We know that some quantitative parameters may exhibit serious variations within a low taxonomic unit. E.g. dogs can vary in almost any *individual* parameter, but still they are dogs as subspecies and wolves as species. Therefore, if a morphologic distance can correspond to taxonomy at all, then g_{ik} must be very special to give s<1 for the frequent variations. This observation helps to reduce the number of independent coordinates and metric tensor components, as will be demonstrated on an extremely idealised toy example.

Consider a hypothetic extremely primitive phylum of Metazoa, which consists of shapeless cylindric animals, with the only measurable data mass M, length L and radius R. Then at first the space is 3 dimensional, therefore the metric tensor contains 6 functions of 3 variables.

However, animals have roughly the density of water, therefore

$$M \approx \pi^2 \rho_0 R^2 L \tag{7.7}$$

with $\rho_0=1$ g/cm^3. The the coordinate M would be redundant, so can be ignored. We remain with

$$ds^2 = g_{RR}dR^2 + 2g_{RL}dRdL + g_{LL}dL^2 \tag{7.8}$$

i.e. with 3 functions of 2 variables.

Now we can observe that for one species both R and L change during ontogenesis, in a correlated way. For such a change $L=L_{tr} \equiv L_0 f(R/R_0)$, ds must be 0. Then one can write

$$R \equiv x, \quad L \equiv L_{tr} y \tag{7.9}$$

and hence, requiring that ds^2 be quadratic in dy, we get

$$g_{RL} = -g_{LL}(L_0/R_0)f'y$$

$$g_{RR} = +g_{LL}(L_0/R_0)^2 f'^2 y^2 \tag{7.10}$$

which means one independent function (g_{LL}) of two variables; a substantial final simplification.

Here we stop; we wanted only to demonstrate that in principle one can build up a *morphologic* distance, where the resulted structure is a Riemannian geometry with correct triangle inequality and all other conditions. We must admit that the *definition* of $s_{\alpha\beta}$ is not sufficiently operative, clear and objective yet. The present taxonomic units inherit the intuitive elements of their construction as well. But such a programme can be adopted.

It would be interesting to see how the distances change during evolution. For some cases, as e.g. for Equidae, a number of quantitative parameters are used, and then the evolution could be drawn in the morphologic Riemannian space. However, *nothing seems* to suggest that even natural changes would happen on geodesic lines, just as in thermodynamics they are not. [13], [14]

8. PHYLOGENETIC TREES

Let us now draw a diagram of evolution in the morphologic space. The diagram should be of N+1 dimensions, being the last one the evolution parameter, either time, or "the degree of evolution". But of course one cannot *draw* a diagram above 3 dimensions, and even that is difficult. So let us make a 2 dimensional diagram.

Such a diagram is necessarily false from two reasons. One is the tremendous dimensional reduction ("projection"), the other is that we are going to map a *pseudo*-Riemann geometry (including time) on a Riemannian one. (See the same problem when the space-*time* of the Minkowski world is depicted on a sheet of paper.) However, one cannot help in this. Maybe the minimal distortion is obtained if we act as follows.

Select an axis of evolution, i.e. the "mainstream", which will be the vertical coordinate axis x=0. Measure the distances in any time from this "mainstream" (anyhow), and measure the x distances accordingly. Then a familiar 2-dimensional tree is obtained. For the ancestry of horses such a simplified diagram is Fig. 2; for further details see e.g. Ref. 15.

Admitting that this is only one example, we use this one to demonstrate the graph structure deduced from fossils. According to tremendous amount of data accumulated, up from Cambrian the evolution is fundamentally divergent (before it the data are insufficient to draw a *solid conclusion*). *To be sure, morphologic* convergence is not unknown: good examples are the piscine appearances of reptilian Ichtiosaurs or mammalian dolphins, or the marsupalian counterparts of some placentalians. Still, comparing fossils from subsequent layers one deduces that divergence is the primary character. Namely:

1) The convergence generally does not hold for the fundamental structures. I.e. Ichtyosaurs had become aquatic as their far Crossopterigian ancestors. However they had retained the reptilian structures inside. Similarly, dolphins remained mammals with lactation, &c.

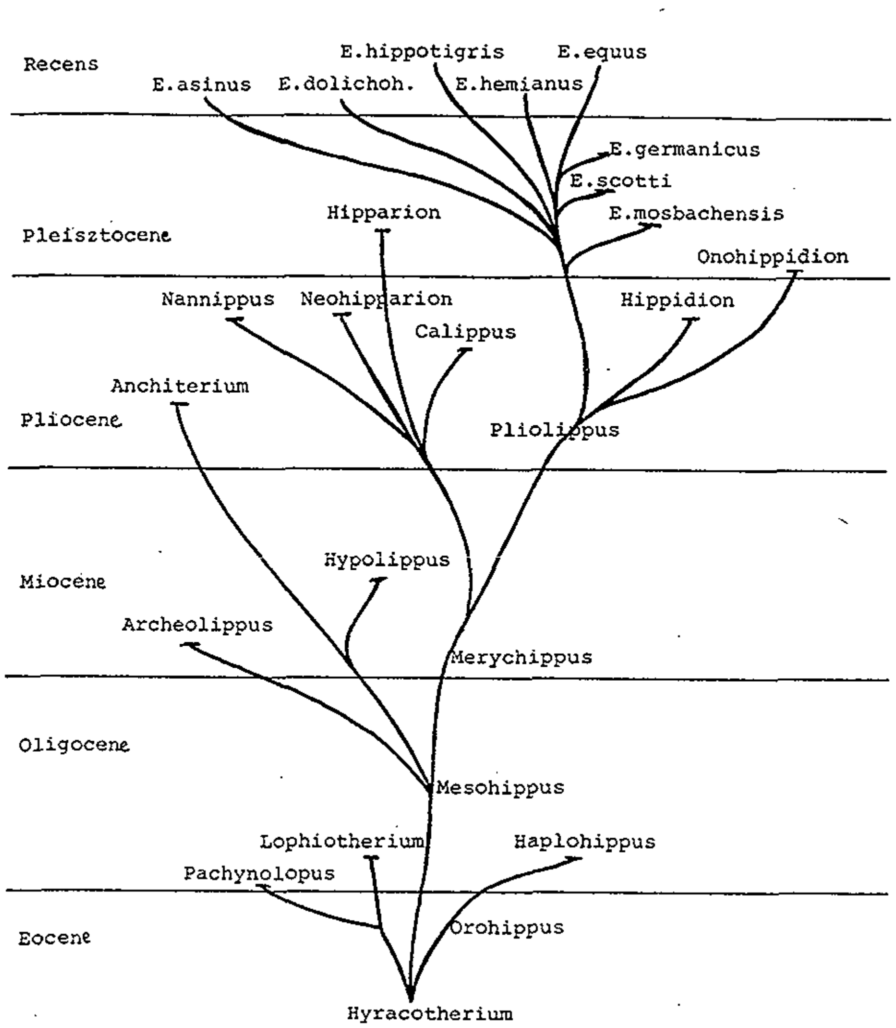


Fig.2. Evolution of horses.

2) The evolutionary changes are irreversible (Dollo's Law) Dolphins still have degenerate legs, e.g., and have *not* reconverted quadrupedal limbs into Crossopterigian fins.

3) Cases can be observed when evolution has a branching point with two diverging descendants for a single ancestor. However, from Cambrian to present no case is known when two substantially different ancestors would have been seen to be converging for a while and then merging into a single descendant species.

We feel this last point the most important. According to it one can distinguish two different vertexes:

```
   \ /                           |
 1  \/   quite familiar     2   /\   never seen
    |                          /  \
```

To be sure, vertex structure 1 is now regarded as the consequence of some fundaments of genetics of the eucariote cells, and in this sense the phenomenologic rules are explained from deeper ones. However, this vertex rule was known well before the advent of modern genetics, and can clearly be seen on Fig. 2.

From the great number of observations all conform to the vertex rule one can deduce the rule as something valid for the whole period available for morphological analysis in the 3 dimensional space-time continuum (i.e. again from Cambrian to present). Therefore when fossils are not quite continuous, one should use Vertex 1 for interpolation, but not Vertex 2. By this method one can get a continuous graph representing kinships of species, genera, or higher taxonomic units. This is a "family tree" of living organisms.

For higher taxonomic units the situation is less and less clear, because the junctions are farther and farther in the past. For classes of Vertebrata the branching is still roughly clear; e.g. the origin of mammals is well clarified among Therapsida reptiles [16]. However, going to *phyla* the situation changes. Rougly speaking, it is the fundamental morphology which changes between 2 phyla [17], therefore we lose all the possibilities to characterize such differences by differences of quantitative data.

Of course, different taxonomic systems distinguish different numbers of phyla; however, from morphology, the number of animal phyla is always between one and half and two dozens. Fig. 3 is compiled from textbooks, and works with 22 ones; with the exception of Archaeocyatha all recent. It is quite possible that there existed others too, now extinct, but it is hard to decide whether the lack of success to find the proper taxonomic location of a fossil is caused by objective or subjective difficulties. The problem is that *all* these phyla are either hardly able to be fossilized or known from Cambrian. Therefore *no forking or branching is observed for phyla.* From the empirical data the phyla of the regnum Animalia might even all be disconnected from the beginning. (**And similarly, there were some ideas that plants would be triphyletic, according to monopodial, dichotomic and verticillate branching structures [18].**) However this is highly im*probable* e.g. because of serious biochemical, cytological &c. similarities. So by intuitive guesses and embryologic &c. observations there are some possible graph diagrams, one of which is Fig. 3.

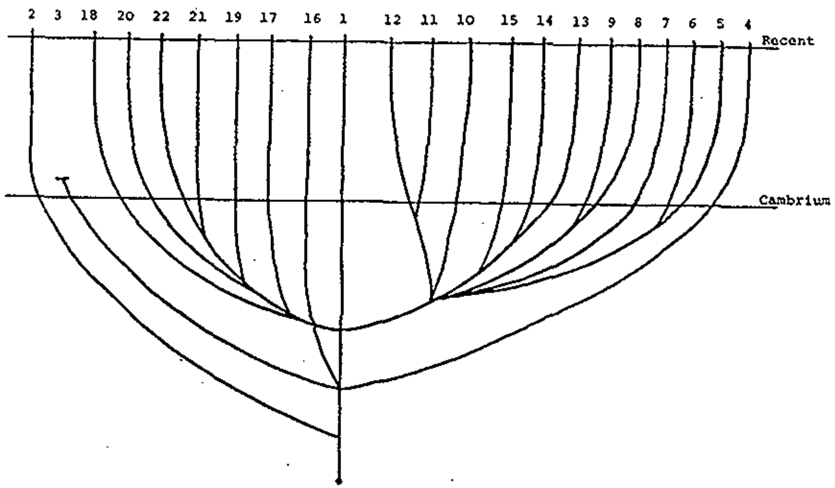


Fig. 3

Numbers for phyla: 1:Protozoa, 2:Porifera, 3:Archaeocyatha, 4:Cnidaria, 5:Platyhelminthes, 6:Nemertini, 7:Kamptozoa, 8:Nemathelminthes, 9:Priapulida, 10:Sipunculida, 11:Mollusca, 12:Echiurida, 13:Annelida, 14:Onychophora, 15:Arthropoda, 16:Tentaculata, 17:Chaetognatha, 18:Pogonophora, 19:Hemichordata, 20:Echinodermata, 21:Tunicata, 22:Vertebrata. For higher groups: 2+3=Parazoa, 4=Radiata, >4=Bilateralia; 5-15=Protostomata (5-9=Acoelomata, 10-12=Oligomer Coelomata, 13-15=Articulata), 17-22=Deuterostomata (17-20=Oligomer Deuterostomata, 21-22=Chordata).

If we want to argue for this graph, or we are interested in the structure for higher taxa (as e.g. for *regni*), then we must go before Cambrian, where the data are not continuous and where anyways the ancestral animals were too simple for clear morphologic differences. There fossils can practically not be used. One must turn to biochemical or genetic data of *recent* organisms, and extrapolations from the recent ones back.

First let us see if the morphologic diagram (Fig. 2) is compatible with genetic data. Ref. 19 gives the evolutional tree of Metazoa *deduced* from the *recent* 18S rRNA sequences. For the graph structure the tree is fairly similar to Fig. 2, except that some Tentaculata and the Pogonophora is transferred to among Protostomata. So one can conclude that the origins of animal phyla still can be *guessed* from morphologic data. However, it is no more true for the origin of Metazoa, or, further, for the connection among *regni*.

It seems trivial that Metazoa originated from some Protozoans. Recent examples even may shed some light on the intermediate steps. The Volvox colony of Euglenae show that after dividing the daughter cells still can remain together. The Mesozoa (if not utter degenerates) are quantitatively ideal intermediate beings. And Porifera (=Parazoa) consist of individual but cooperating cells. However the results obtained in such ways are very limited. E.g., hence one cannot know if Metazoa are monophyletic or not, if they are, from which Protozoan group they started, &c.

For different regni the common ancestors are even farther in the past. Originally only two regni were distinguished (animals vs. plants), and now the traditional systems contain five (say, Animalia, Fungi, Plantae, Eubacteria, Archaebacteria [20]). If, e.g., genera differ in *quantitative* morphology, while phyla do in morphologic *structure*, then regni differ in fundamental biochemistry (e.g. autotrophy vs. heterotrophy).

For extrapolation hypotheses are needed. We know that enzymatic sequences are determined by DNA ones, which in turn are subjects of continuous mutation, caused by e.g. cosmic radiation hits. As far as we know, mutations are random, and mutation rates are constant for *moderate* times (although for longer times they can change because of changes in the environment or intracellular defense mechanism). *Vital* mutations are subjects of selection, so rare to survive, but neutral ones survive also randomly. (See e.g. the argumentation in Refs. 21-22.) Therefore sequence differences are monotonously growing with the length of independent evolution, (against reversible changes see the probability arguments of Ref. 23) and can be used to trace back the branching structure. (Clearly the differences do not measure simply *time*, as seen e.g. from the tree in Ref. 19. There for 18S rRNA changes the "farthest evolved" Deuterostomata is Tunicata with 5.45% sequence changes from the common ancestor of Proto- and Deuterostomata, while on the Protostomata side it is Sipunculida, with 9.46%, both synchronous and seemingly not too evolved.) Of course, one must be careful to select the specific enzyme or RNA part, appropriate to the taxonomic level considered.

Before finally leaving the field of morphology, let us mention the only very slight indication of fossil morphology for connection of two regni. *The matter is very hypothetic, and we* are not arguing for it, but no harm to list it as a possibility. Some authors classify the phylum Archaeocyatha, together with Receptaculitoida, into a new *regnum* Archaeata *between* Metazoa and Metaphyta. The idea is roughly the following. We can observe permanent loss of chloroplasts in some Protistae. Therefore the usual guess is that the divergent evolution of Animalia and Plantae started from the protozoan stage. However fossils permit the *alternative that the divergence happened in multicellular stage*, in the young Precambrian, or even between Ediacara and Cambrian. The suspicion can be based on the difficulties to find a mouth or even a ventrodorsal difference on many Ediacaran fossils [24], which are of quite large size, causing anyways a problem when interpeting them as ancestral to Cambrian organisms.

A multicellular organism of 0.5 m size without any mouth cannot be anything else than an autotrophe. *If* it is not empty inside (*as a blastula*), then the central cells must eat the food produced by the outer layers, so then the whole organism is a myxotrophe, whence specialisation can go in both directions. On this grounds it is not impossible to classify together several problematic fossils, as Petalonamae [25], mouthless Ediacarans and Archaeocyatha as ancient myxotrophes. For further *possible* support observe that in some 18S rRNA analyses Cnidaria go with Plantae instead of Animalia [26].

An alternative possibility is that Ediacarans *were* empty inside [24], when the above arguments for myxotrophy are seriously weakened. We, however, do not want and do not have to settle this question here.

As told above, we cannot go further into the past via morphology, and must turn to biochemistry and genetics, extrapolating from extant organisms. But first let us see how to get *quantitative* measures of *separation* from biochemical or genetical differences. As seen in Sect. 7, "geometrisation" needs two steps: i) an operative definition, how to *measure* something for distance; ii) the check if it is indeed a distance (meaning, roughly Rels. (7.2-5)). If so, then metricisation follows (almost; fine mathematical details are omitted here). But the check should be done on a lot of species, and the present status of art is rather before this stage. So here we mention well-accepted *suggestions* for a biochemical and a genetic distance; for further details see Ref. 27 and citations therein.

Consider two species with the same multiple representations of proteins but of course with different probabilities. Then the probability distributions give a correlation coefficient $R_{\alpha\beta}$, and

$$s_{\alpha\beta} \approx -\ln R_{\alpha\beta} \qquad (8.1)$$

For DNA sequences, form a hybrid DNA, and compare the dissociation temperature to that of pure ones. There is a decrease $\Delta_{\alpha\beta}T$, and then one can try with

$$s_{\alpha\beta} \approx Q\Delta_{\alpha\beta}T \qquad (8.2)$$

where Q sets the scale.

These quantities are more directly connected to phylogenetic changes than the outward appearance. But, as seen, even for phyla the results are roughly compatible. However, this is not necessarily so for all the details. An important example for the profit of the use of biochemical or genetic distances to establish taxonomy is the rearrangement in the superfamilia Hominoidea. Its *recent* members are the *gibbons* (2 species), the orang, gorilla and chimpanzees (2 species), and man. Henceforth we ignore the gibbons.

The traditional classification is based on the obviously expressed morphologic differences between man and the others. Therefore they were classified into two families: Hominida (man) and Pongida (orang, gorilla and chimpanzees). Since fossils classifiable into this family Pongida exist from -13 Mys, the traditional picture is Fig. 6a.

On the other hand, the man-chimpanzee genetic distance has been found extremely *low*, 0.62, *below* the average of that for close congeneric species. This problem suggests a reclassification of the 4 genera (species?) as: family Pongida=orang; family X (still no proper name is accepted)=man, chimpanzees and gorilla. Then the earliest divergence led to orang; the new picture is Fig. 6b.

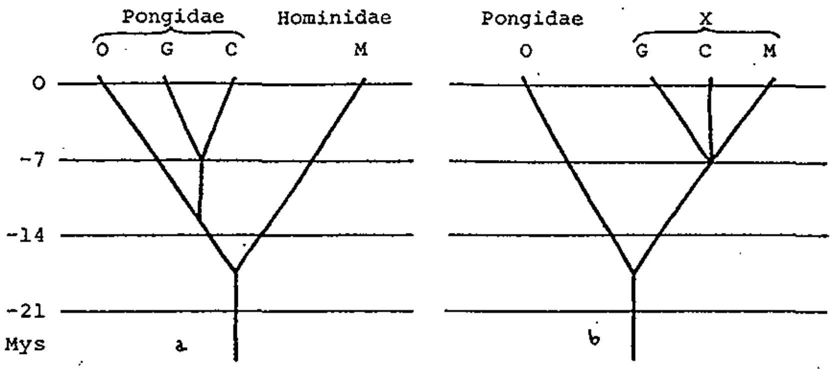


Fig. 4
Traditional and new pictures about recent Hominoideae.

Now we can try to use the biochemical and genetic differences to explore the relations among regni.

9. AN OUTINE OF THE ORIGINS OF BIOLOGICAL KINGDOMS

The origins of biological Kigdoms are the origins of a diversity of basic cell structure. The origins of cells are in turn a symbiosis between membranes, catalists and genes ([29], [30]) embedded into the context of the evolutionarily changing cytoplasm. Following and modifying the works of Cavalier-Smith ([28], [29], [30], [31]) and others (e.g. Ref. 32), first the characteristic features of the Kingdoms of organisms (cf. the Table of Fig. 5), secondly a compact scenario for the generations of Kingdoms are presented.

There are two classes of Kingdoms : Procaryota and Eucaryota. In Procaryotes the DNA is not separated from ribosomes by an envelop, instead attachted to "envelope skeleton"; the endomembrane system and the cytoskeleton are absent. The Procaryota has two Kingdoms.

1. *The Eubacteria possesses murein and acylglycerol membrane lipids.*

2. *The Arhebacteria* posesses isoprenoid ether lipids, and outer membrane is absent (plasma membrane is present), no murein.

3. *The Archezoa* are characterized by the formation of a nucleus, while organelles as mitochondria &c. are absent.

The Eucaryota cells are nucleated, and have endomembrane system, cytoskeleton, cytosis, aglicerol membrane lipids, at least one of mitochondrion, peroxisome and chloroplast, there is no peptidoglican cell wall. The Eucaryota can be classified into five Kingdoms on the basis of molecular and cell structure, and into a sixth one otherwise, discussed later.

4. *The Protozoa* contains chloroplast envelope of three membranes, includes mycetozoa, and it is predominantly phagotrophic.

5. *The Fungi* shows food absorption, typically chitinous walls, and no chloroplast or phagocytosis.

6. *The Chromista* cells have (periplastidal compartments) tubular ciliary mastigonemes and/or chloroplast endoplasmatic reticulum.

7. In *the Planta* cells plastids are always present, the plastids are not in the endoplasmatic reticulum, there is photosynthesis with chloplasts having two membranes.

8. *The Animalia* are phagotropic organisms.

Up to this point this is the most recent scheme of Cavalier-Smith [31]. If chloroplasts are relative easy to acquire, this scenario is the most probable one, as far as we know. The corresponding phylogenetic tree is displayed on Fig. 5. For the details and the explanations see the cited works of Cavalier-Smith and Margulis.

However up to now it is not quite clear if the inclusion of cyanobacteria to form chloroplasts was one event or happened in several times independently. In some time this question may be answered via comparative analyses of chloroplasts in Euglenozoa, Plants and Chromista. Until that there is a possibility that the acquiring of chloroplasts was a single event, and the Ophisthokonta have lost them later [29]. This assumption (i.e. that it is much more difficult to acquire a chloroplast than to lose it, and some losses are indeed seen), together with a second one that the Ediacara animals have not had empty internal space, plus some (admittedly not unambigous) fossil data may suggest a different scenario, with a ninth regnum. As said previously, there are guesses for mixotrophy of Ediacarans. Indeed, it is possible that these creatures were of mixotrophic foraging , being the external cells photosynthetic (no mouth to eat), the internal cells heterotrophic, loosing their photosynthetic ability in darkness. If so, such a multicellular being is a possible common ancestor of Metazoa and Metaphyta. This kingdom is:

9. *the Archeata*, discussed extensively in many paleontological publications (cf e.g. Pflug's works).

A possible unified phylogenetic scenario is presented in Fig. 6.

10. CONCLUSIONS

According to a general overview of evolution of matter based on stable structures there are two types of periods in evolution considering its main tendencies. One is a structure forming (building together) or structuralizationary period, the other is a differentiation period when the earlier formed structures compete both in multiplication and structure formation of the next level of structures. ([33], [34]) The main branching structure of the phylogenetic tree represents the differentiation and competition period of evolution on multicellular level of living organisms. The other evolutionary tendency, building together, seems to have been dominant on the unicellular level of evolution: this led to the evolution of kingdoms.

The question of kingdoms or regni is still open, and nobody has enough and unambigousevidence for any scenario. However the evolution of life; the existence of *irreversible* changes seems to be a well established *fact*. From the evidences it is not quite clear, which one is the unique direction. This question will be briefly investigated later in this Volume.

ACKNOWLEDGEMENTS

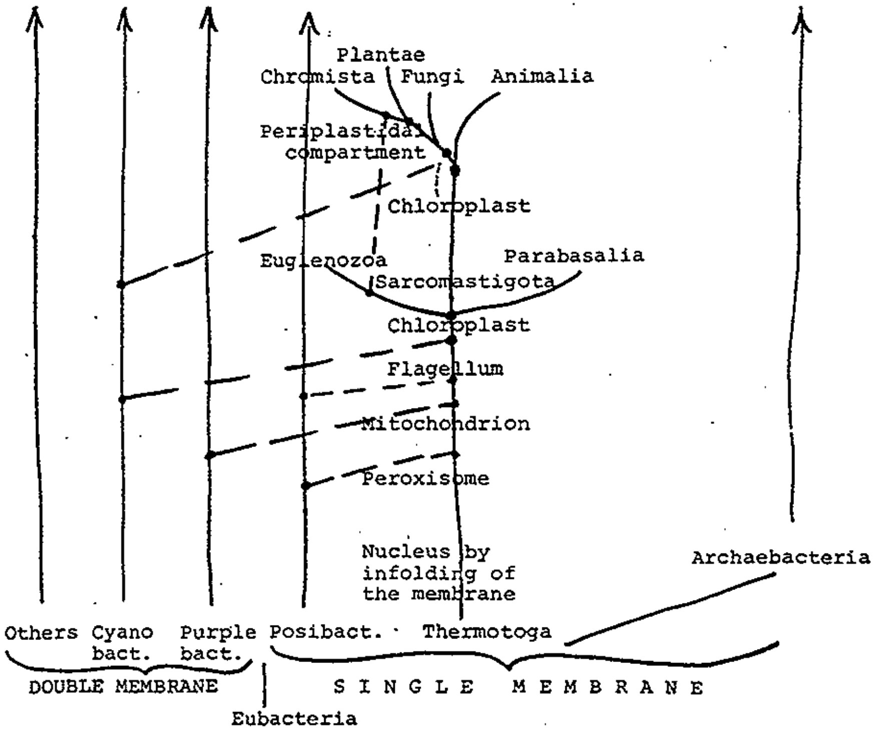
One of the authors (B. L.) would like to thank Drs. F. Glück and I. Racz for illuminating discussions.

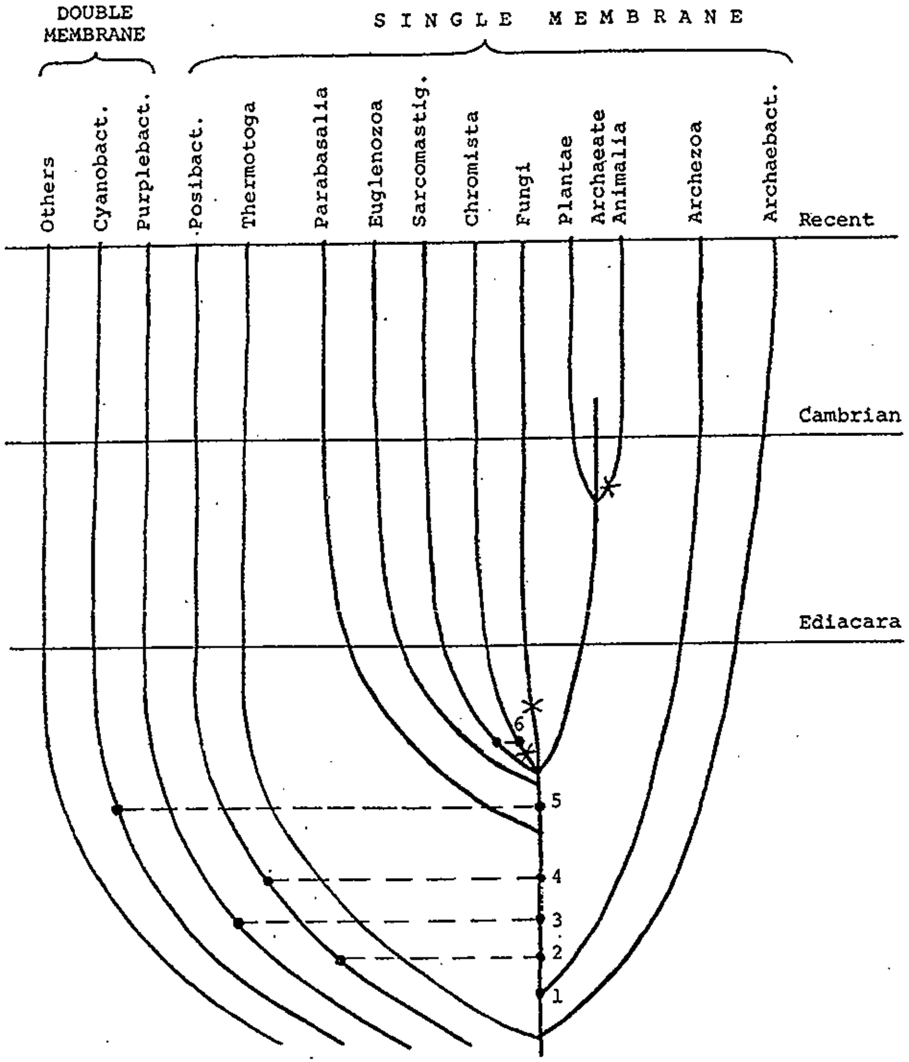
Symbiotae

	Nucleus	Mito-chondrion	Peroxi-some+deriv.	Chloro-plast	others as	Flag-ellum
Eubacteria	-	-	-	-		-
Archaebacteria	-	-	-	-		-
Archezoa	x	-	-	-		?
Protozoa	x	x	x	.	hydrogeno-soma	x
Plantae	x	x	x	x		x
Animalia	x	x	x	-		x
Fungi	x	x	x	-		x
Chromista	x	x	x	x/?/	periplastidal compartment	x

- not found
x primarily existing
. in some phyla

Fig. 5





1 Autogeneous formation of nucleus
2 Inclusion of peroxisome
3 Inclusion of mitochondrion
4 Inclusion of flagellum /Spirochaeta?/
5 Inclusion of chloroplast
6 Inclusion of periplastidal compartment
X Loss of chloroplast

Fig.6.

REFERENCES

[1] L. P. Eisenhardt: Riemannian Geometry. Cambridge University Press, Cambridge, 1950
[2] A. Z. Petrov: Novye metogy v obshei teorii otnositel'nosti. Nauka, Moscow, 1966
[3] Ch. Kittel: Introduction to Solid State Physics. J. Wiley & Sons, New York, 1961
[4] A. Brynjólfsson: Advances in Physics 6, 247 (1957)
[5] L. Eisenbud, G. T. Garvey & E. P. Wigner: General Principles of Nuclear Structure (in Handbook of Physics, eds. E. U. Condon & H. Odishaw), McGraw-Hill, New York, 1967
[6] C. Renfrew: Before Civilization. Penguin, Harmondsworth, 1979
[7] M. J. Aitken: Physics and Archaeology. Oxford University Press, Oxford, 1974
[8] W. B. Harland & al.: A Geologic Time Scale, Cambridge University Press, Cambridge, 1989
[9] J. Brooks & G. Shaw: Origin and Development of Living System. Academic Press, London, 1973
[10] G. Cuvier: Recherches sur les ossemens fossiles. Paris, 1812
[11] C. Lanczos: Space through the Ages. Academic Press, London, 1976
[12] L. M. Cherry & al.: Evolution, 36, 914 (1982)
[13] P. Salamon & al.: J. Chem. Phys. 73, 1001 (1980)
[14] L. Diósi, B. Lukács & A. Rácz: J. Chem. Phys. 91, 3061 (1989)
[15] G. G. Simpson: The Major Features of Evolution. Columbia University Press, New York, 1982
[16] T. S. Kemp: Mammal-like Reptiles and the Origin of Mammals. Academic Press, London, 1982
[17] M. A. Fedonkin: in A. Hoffman & M. H. Nitecki (eds.) Problematic Fossil Taxa, Oxford University Press, Oxford, 1986, p. 54
[18] P . Greguss: Beih. z. Bot. Zentralbl. XXXVI, 229 (1918); P. Greguss: A szárazföldi növényvilág háromirányú (trifiletikus) fejlődéstörténete. Szeged, 1965 (in Hungarian)
[19] J. A. Lake: Proc. Natl. Acad. Sci. USA 87 763 (1990)
[20] S. A. Benner, A. D. Ellington & A. Tauer: Proc. Natl. Acad. Sci. USA, 86, 7054 (1989)
[21] M. Kimura: Nature 217, 624 (1968)
[22] J. L. King & T. H. Jukes: Science 164, 788 (1969)
[23] M. Eigen: Naturwissenschaften 58, 465 (1971)
[24] A. Seilacher: Lethaia 22, 229 (1989)
[25] H. D. Pflug: Paleontographica A134, 153 (1970); A139, 134 (1972)
[26] K. G. Field & al.: Science 239, 748 (1988)
[27] Mary-Claire King & A. C. Wilson: Science 188, 107 (1975)
[28] T. Cavalier-Smith: Endocytobiology II, 1027 (1983)
[29] T. Cavalier-Smith: Ann. N. Y. Acad. Sci. 503, 17 (1987)
[30] T. Cavalier-Smith: CSH. Symp. Quant. Biol. LII, 805 (1987)
[31] T. Cavalier-Smith: Endocytobiology IV, 515 (1989)
[32] L. Margulis: Symbiosis and Cell Evolution. Freeman, San Francisco, 1981
[33] Sz. Bérczi: Acta Geol. Hung. 23, 163 (1980)
[34] V. Csányi: Az evolúció általános elmélete. Akadémiai, Budapest, 1979 (in Hungarian)

A MECHANOCHEMICAL MODEL FOR SPATIOTEMPORAL SELF-REPRODUCTION

J. Verhás and I. Molnár

Department of Chemical Physics, Technical Univesity,
Budapest, Hungary, H-1521

Received on 11th July, 1990

Department of Genetics, Lorand Eotvos University,
Muzeum krt. 4/a. Budapest, Humgary, H-1088

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

Abstract

A two dimensional liquid crystal cell membrane model storing bending energy is examined. The model displays an automatic 'cell' division process determining the stationary size of the cells, which is stable against perturbations. On the basis of the kinetics of metabolic processes, a time scale for division is derived. The different rates of reproduction turn out comparable. The relationship between the bifurcation behaviour the model shows and the heterochrony of reproduction is discussed.

1. Introduction

One of the most ancient biological structures is the spherically symmetric membrane vesicle. Its significance is at least twofold. First, spherical vesicles are the candidate precursors of the origins of modern cells. Secondly, there are many modern cell components, which have retained this typical or generic shape, e.g. secretory vesicles, endosomes or receptosomes.

The aim of this paper is to propose a physico-chemical description of deformation and subsequent reproduction of vesicles or primeval cells. This problem is of paramount importance both for cell and developmental biology and evolutionary theory. As Goodwin (1984. p. 219) claims : "There can be no adequate evolutionary theory without a causal account of reproduction." This statement reflects our motivation. In general, the concept of reproduction is connected to different classes of biological entities with different complexity. We have chosen the study of vesicle reproduction because the interactions between mechanical deformations and chemical reactions can be made tractable in comparatively simple ways. We use the terms vesicle and primeval cell interchangeably.

Many excellent surveys are available on the physico-chemical processes of the biology of vesicle formation and reproduction (see especially Tanford, 1973; Luke, 1982; Koch, 1985; Oster et al, 1989; Chevalier & Zemb, 1990). These alternative approaches define the (continuum) mechanical and chemical properties of biological membranes, determining the dynamics and geometry of membrane bilayers. We follow a variational approach which is of the following specific properties that are different from the spirit of the above listed works :

1. Our model describing vesicle reproduction is design such as to generate other structures as well, including gastrulation and axially symmetric, periodical structures (Molnar & Verhas, 1990).

2. Therefore, we probably stress more strongly the evolutionary implications of the model.

This paper is organized in the following way. First, we describe a model capable of generating vesicle reproduction. Secondly, we show how vesicle size is regulated by reproduction. Thirdly, we introduce a time scale for the comparability of the different reproductive rates controlled by three classes of chemical kinetics.

2. A mechanochemical model of self-reproduction.

The vesicle model is based on the excess of the Helmholtz free energy due to curvature. Following Helfrich (1974) and Deuling and Helfrich (1976), the curvature energy is given as

$$F_c=\int[K_1(c_1+c_2-a)^2+K_2c_1c_2]d\Omega \tag{2.1}$$

where c_1 and c_2 are the principal curvatures, K_1, K_2 and a are material constants influenced by the metabolic processes. The minimization of the free energy with constant volume and surface area leads to the system of differential equations which, in non-dimensional form, for surfaces of revolution, reads

$$\frac{d^2\alpha}{ds^2}=\left(\frac{\sin\alpha}{x}-\frac{d\alpha}{ds}\right)\frac{\cos\alpha}{x}+\frac{\sin\alpha}{2\cos\alpha}\left[\left(\frac{\sin\alpha}{x}-A\right)^2-\left(\frac{d\alpha}{ds}\right)^2-K+K\frac{x}{\sin\alpha}\right] \tag{2.2}$$

(here K is the non-dimensional underpressure) and

$$\frac{dx}{ds}=\cos\alpha \tag{2.3}$$

Explanation of the notations

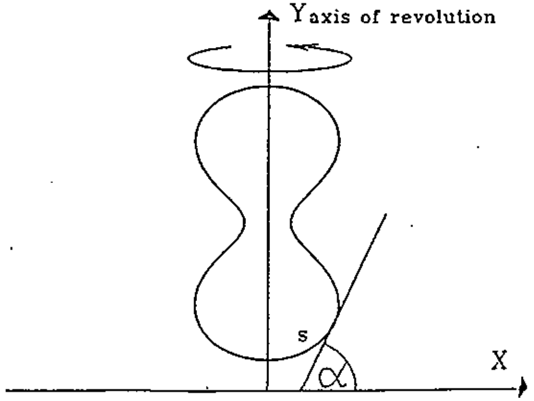
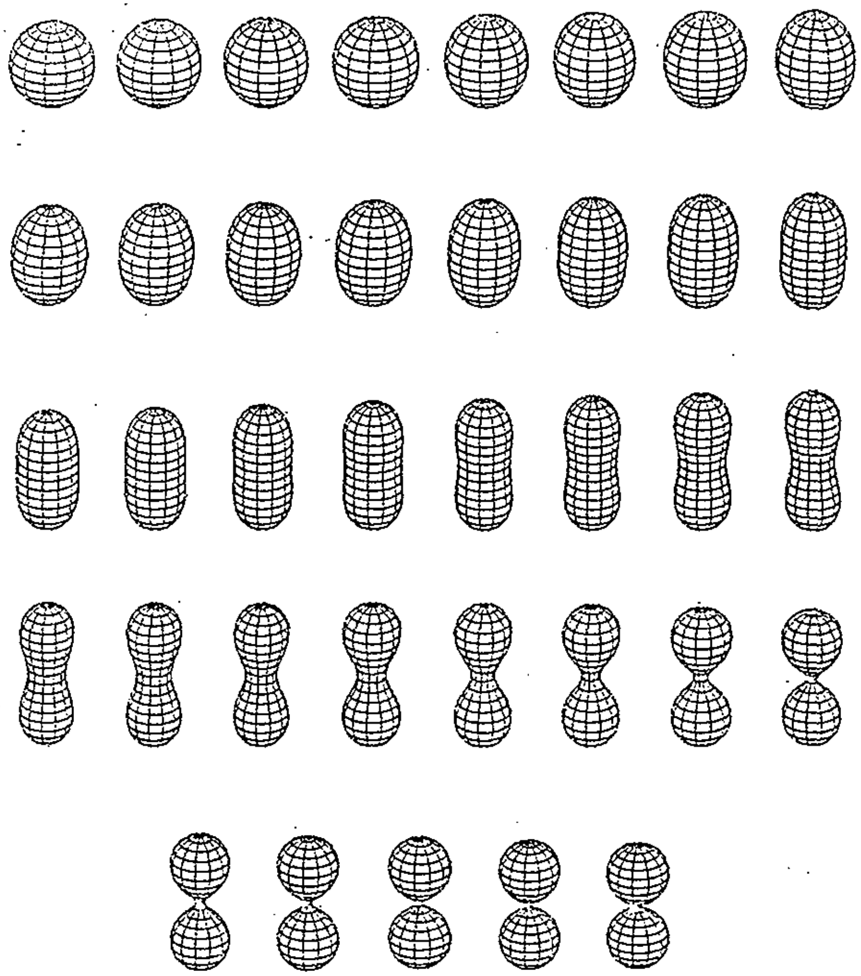


Figure 1.
The graphical illustration of the meaning of the x, y, s and alpha (α) quantities, which are used in the equations.



The degree of saturation (relative volume)

100%	100%	99.9%	9987	99.5%	99.2%	98.8%	98.4%
97.8%	97.2%	96.5%	95.7%	94.7%	93.6%	92.4%	91.0%
89.4%	87.8%	85.9%	83.9%	81.9%	79.8%	77.8%	76.0%
74.3%	72.8%	71.7%	70.8%	70.1%	69.7%	69.7%	69.9%
	70,0%	70.2%	70.3%	70.5%	70.6%		

Figure 2.
A trajectory of self-reproduction by vesicle fission.

The notation is explained in figure 1. The computations exhibit shapes leading to reproduction (figure 2). The parameters are given in figure 3. The computation starts at the bottom where the equations contain a removable singularity. (The first step of the Runge-Cutta procedure are done by hand.) This way the computations are able to perform any form of rotational symmetry. The symmetry breaking is simply a loss of stability. The variational calculus, when minimizing the Lagrangian, gives the same results as the complicated analysis of stresses, forces and torques in the membrane.

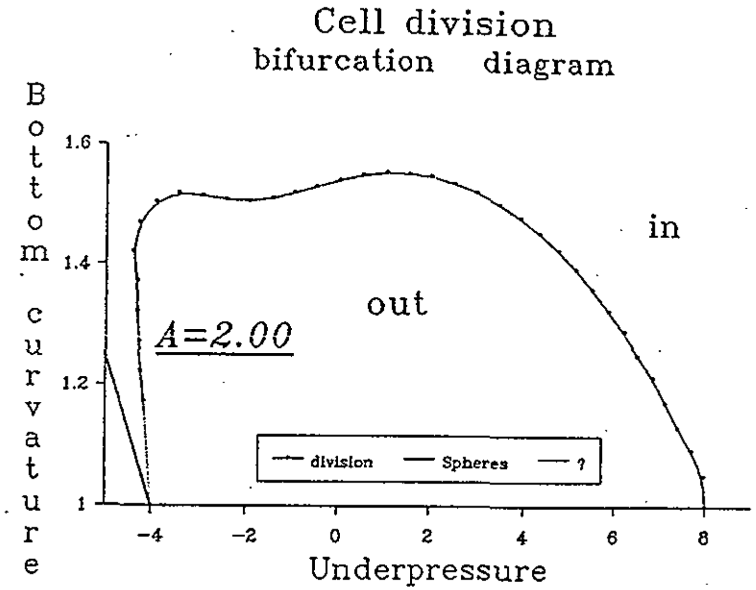


Figure 3.
The bifurcation diagram of self-reproduction.

3. A physico-chemical description of self-reproduction.

The above model shows that the membrane can have the natural property leading to cell division. The shape of the dividing cells are determined by the actual volume and the area of the surface. The pertaining quantity is the ratio of the volume to the maximal one that can be wrapped up into the same envelope:

$$\frac{V}{V_m} = 6\sqrt{\pi}\,\frac{V}{F^{3/2}}\,. \qquad (3.1)$$

Here V is the volume of the cell, F the surface, V_m the volume of the sphere with surface F. This ratio is always less then one and when it decreases under $\frac{1}{\sqrt{2}}$ the cell divides into two spheres. To determine the time scale, we have to know the kinetics of the metabolic processes that can be modeled in several ways. It seems plausible to suppose that the materials both of the membrane and of the plasma is produced inside the cell by the same system of chemical reactions. Hence, if the nutrition transport across the membrane is plentiful, the differential equations

$$\frac{dF}{dt} = K_1V \qquad \frac{dV}{dt} = K_2V \tag{3.2}$$

hold. If the nutrition is the slow process (it is the determining one) then they govern a growth rate proportional to the surface:

$$\frac{dF}{dt} = K_1'F \qquad \frac{dV}{dt} = K_2'F. \tag{3.3}$$

It is also possible that a single object inside the cell can produce the materials in question and this case, the rate of growth is constant i.e.

$$\frac{dF}{dt} = K_1'' \qquad \frac{dV}{dt} = K_2'' \tag{3.4}$$

or if the activity of this object varies with time

$$\frac{dF}{dt} = K_1^*G(\ldots) \qquad \frac{dV}{dt} = K_2^*G(\ldots). \tag{3.5}$$

The kinetics of the processes can be, of course, any other of a lot, nevertheless, these are rather simple for using as a basis for further argumentations and they cannot be far from reality.

3.1 Vesicle size regulation by reproduction.

Any of the above kinetics is accepted, the ratio of the two derivatives does not depend on time. This circumstance gives the opportunity to determine the volume as a function of surface, independent of the kinetics. According to the possibilities, we introduce the proper one of the notations

$$K = \frac{K_1V}{K_2V} \quad \text{or} \quad K = \frac{K_1'F}{K_2'F} \quad \text{or} \quad K = \frac{K_1''}{K_2''} \quad \text{or} \quad K = \frac{K_1^*G(\ldots)}{K_2^*G(\ldots)},$$

and obtain the differential equation

$$\frac{dV}{dF} = K, \tag{3.6}$$

which has the solution

$$V = V_o + K(F-F_o). \tag{3.7}$$

Here V_o is the initial volume and F_o the initial surface. The measure for the degree of saturation is

$$\frac{V}{V_m} = 6\sqrt{\pi}\,\frac{V_o+K(F-F_o)}{F^{3/2}}$$

depending on the actual value of the surface area. If the initial form is a sphere with radius r, the equations just derived get

$$V = V_o\left[1+\frac{3K}{r}\left(\frac{F}{F_o} - 1\right)\right] \tag{3.8}$$

and

$$\frac{V}{V_m} = \frac{1 + \frac{3K}{r}\left(\frac{F}{F_o} -1\right)}{\left(\frac{F}{F_o}\right)^{\frac{3}{2}}}. \tag{3.9}$$

According to the referred mechanism of division, when the degree of saturation has decreased under $\frac{1}{\sqrt{2}}$, the cell splits up into two spheres, the radius of which are not generally equal to that of the initial sphere r, the new cells can be either smaller or bigger. Denote the radius of the spheres just after division by R. The surface

$$F=2*4\pi R^2 \quad \text{or} \quad \frac{F}{F_o} = 2\frac{R^2}{r^2} \tag{3.10}$$

from where, we obtain

$$\frac{V}{V_m} = \frac{1}{\sqrt{2}} = \frac{1+ \frac{3K}{r}(2\frac{R^2}{r^2} -1)}{2\sqrt{2}\,\frac{R^3}{r^3}}. \tag{3.11}$$

This leads to

$$\frac{r}{3K} = \frac{2\left(\frac{R}{r}\right)^2 -1}{2\left(\frac{R}{r}\right)^2 -1}. \tag{3.12}$$

This formula gives the size of the child cells depending on the size of the original one. The equation is of third order, so the solution is not available in an explicit form, it is plotted in figure 4. The scale in the picture takes the radius r=3K 100%. The minimum of the curve is at 66.7%, the straight line is the asymptote, which has the equation

$$R- \frac{1}{\sqrt[3]{2}} = \frac{1}{\sqrt[3]{2}}\,(r- \frac{1}{3})$$

$$(R=0.794r+0.069).$$

We see from the equation that

and

$$R<r \quad \text{if} \quad r>3K \qquad\qquad (3.13)$$

or

$$R>r \quad \text{if} \quad r<3K$$

$$R=r \quad \text{if} \quad r=3K.$$

The inequalities say that during the consequent divisions the sequence of the radii of the spheres tend to 3K, whatever the initial radius was. The size r=3K is stable against the perturbations coming from the outer world. The environmental effects, of course, may change the value K, involving, this way, the change of the stable cell size.

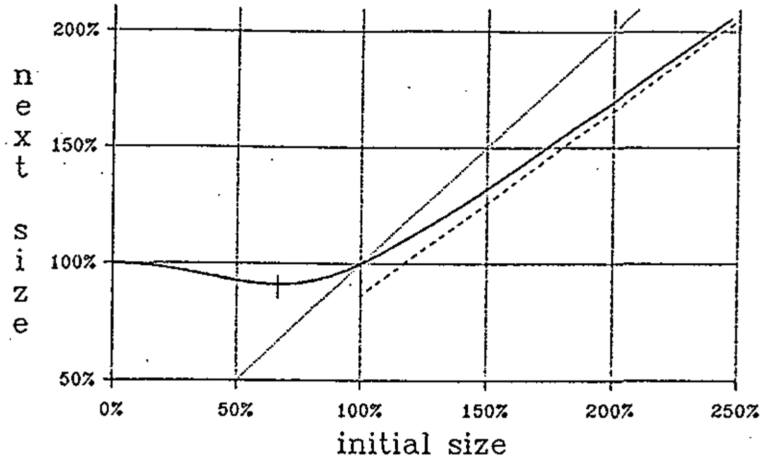


Figure 4.

The relationship between the sizes of subsequent vesicles.

So far, we have not taken care whether the ratio of saturation given by equation (3.9) is less then one. Because its initial value is 1, the initial derivative can not be positive. Regarding the quantity $\frac{F}{F_o}$ as an independent variable, the derivative

$$\frac{d\frac{V}{V_m}}{d\frac{F}{F_o}} = \frac{\frac{3K}{r}\frac{F}{F_o} - \frac{3}{2}\left[1+\frac{3K}{r}\left(\frac{F}{F_o}-1\right)\right]}{\left(\frac{F}{F_o}\right)^{5/2}} \qquad (3.14)$$

is obtained, for the initial value of which we get

$$\left.\frac{d\frac{V}{V_m}}{d\frac{F}{F_o}}\right|_1 = \frac{3K}{r} - \frac{3}{2} \leq 0. \qquad (3.15)$$

From here, we obtain the inequality

$$r \geq 2K. \qquad (3.16)$$

According to this, in any case the radius is not less then 2/3 times the stable size, the above argumentation holds, but else, some modification is necessary. If the initial size is too small the processes of growth produces volume having no room inside the envelope. The membrane cracks, the superfluous material of the plasma pours out. The events from here depend on the forces closed the membrane originally. If they are able to close the leak the process of pouring out goes on until the radius reaches the critical size (r=2K) and then the above sequence of divisions proceeds. Otherwise, the damage is very serious and the cell dies. To illustrate what has been said, here are two sequences of consecutive sizes, the first is with increasing radii:

66.7% 91.1% 96.0% 98.1% 99.1% 99.5% 99.8% 99.9%
99.9% 100% ... ,

the other with decreasing ones:

1000% 801.1% 543.4% 518.5% 419.6% 314.4% 279.7% 231.4%
193.7% 164.8% 143.0% 127.2% 116.3% 109.2% 105.0% 102.6%
101.3% 100.7% 100.3% 100.2% 100.1% 100% ...

3.2. A time scale for reproduction and the heterochrony of a morphogenetic bifurcation

There is a strong suspicion that a dominating mode of shape changes in evolution is the heterochrony (Alberch et al, 1979). The heterochrony is the change in the timing and rate of developmental processes at the different hierarchical levels of the organisms.

It is shown already elsewhere (Molnar & Verhas, 1990), that when the dynamics generated by the model discussed in this paper passes through different curvature domains, displays a set of qualitative behaviours. We have also proposed that purely spatial branchings (bifurcations) possess different rate and timing, generating alternative spatiotemporal developmental dynamics in evolution. In the next step we shall point out how heterochronic division rate changes can control the time of a spatial bifurcation (i.e. division). For this purpose, a time scale of division is to be derived, which makes possible the comparison of the different rates of reproduction caused by the bifurcation behaviour.

First, we treat the time scale of the division for a cell of stable initial size. It, of course, does depend on the mechanism of the kinetics, nevertheless, those given in equations (3.2) and (3.3) give rise to similar formulae.

For cells of stable size,

$$\frac{V}{V_o} = \frac{F}{F_o} = 2$$

at fission and the saturation ratio is

$$\frac{V}{V_m} = \left(\frac{F}{F_o}\right)^{-\frac{1}{2}} = \left(\frac{V}{V_o}\right)^{-\frac{1}{2}}. \qquad (3.17)$$

According to the kinetics given in (3.2)

$$V=V_0 \exp K_2 t \tag{3.18}$$

while for the other kinetics, in accordance with (3.3)

$$F=F_0 \exp K_1' t \tag{3.19}$$

From here, the duration of a period of division is

$$\tau = \frac{1}{K_2} \ln 2 \quad \text{and} \quad \tau = \frac{1}{K_1'} \ln 2 \tag{3.20}$$

respectively. Applying them and the equation (3.17), we obtain:

$$\frac{V}{V_m} = \exp \left(- \frac{\ln 2}{2} \frac{t}{\tau}\right) . \tag{3.21}$$

From here

$$\frac{t}{\tau} = - \frac{2 \ln \frac{V}{V_m}}{\ln 2} \tag{3.22}$$

follows, which gives the relative time scale of the division process. The numerical values belonging to the shapes in figure 2 are given in table 1. If the kinetics is given by (3.4), the time scale is a bit different. To determine it, we solve the differential equation system (3.4):

Table 1.
Relative timescale of shape transformation and division.
(Kinetics: eqs. 3.2 or 3.3)

0%	0%	0.3%	0.8%	1.4%	2.3%	3.4%	4.7%
6.3%	8.1%	10.3%	12.8%	15.8%	19.0%	22.9%	27.2%
32.2%	37.7%	43.9%	50.6%	57.6%	65.0%	72.4%	79.3%
85.9%	91.5%	95.9%	99.6%	100%	100%	100%	100%
	100%	100%	100%	100%	100%		

$$F=F_0+K_1''t \quad ; \quad V=V_0+K_2''t \tag{3.23}$$

As this time we are interested in cells with stable initial size, according to the relations (3.4) and (3.6) we write

$$\frac{F}{F_0} = \frac{V}{V_0} = 1+ \frac{K_1''^3}{36K_2''^3} t \tag{3.24}$$

The period of the division

$$\tau = \frac{36\pi K_2''^2}{K_1''^3} , \tag{3.25}$$

which yields

$$\frac{F}{F_0} = \frac{V}{V_0} = 1+ \frac{t}{\tau} . \tag{3.26}$$

The time scale of the division is

$$\frac{t}{\tau} = \frac{1}{\left(\frac{V}{V_m}\right)^2} -1 , \tag{3.27}$$

the numerical values belonging to the shapes in figure 2 are given in table 2. Here we notice that for the last nine form, 100% is written as we suspect that these forms belong to unstable equilibrium and the process of division proceeds fast after reaching these shapes. Nevertheless, this suspicion seems rather trivial from the physico-chemical point of view, and is in accordance with the customary structure of bifurcations, to prove it some further mathematical investigations are needed.

Table 2.
Relative timescale of shape transformation and division.
(Kinetics: eq. 3.4)

0%	.0%	0.2%	0.6%	1.0%	1.6%	2.4%	3.3%
4.4%	5.8%	7.4%	9.3%	11.6%	14.1%	17.2%	20.7%
25.0%	29.9%	35.6%	42.0%	49.1%	57.0%	65.2%	73.3%
81.3%	88.5%	94.5%	99.5%	100%	100%	100%	100%
	100%	100%	100%	100%	100%		

Now, we turn to other sizes. As each kinetics perform its own outcome, they are analyzed one by one. We do not take care of too small sizes ($r<2K$). Regard, first, the kinetics (3.2). From equations (3.7), (3.9) and (3.18), we get

$$\frac{V}{V_m} = \frac{\exp(K_2 t)}{1+ \left[\frac{r}{3K} \exp(K_2 t) -1\right]^{3/2}} . \tag{3.28}$$

At the very time of division this fracture has the value $\frac{1}{\sqrt{2}}$, which gives the relation between time and size in an implicit form. Introducing the stable size division time to the formula — by using equation (3.20) — and rearranging equation (3.28), we get

$$\frac{r}{3K} = \frac{2^{\left(\frac{1}{3} + \frac{2t_d}{3\tau}\right)} -2}{2^{\left(\frac{t_d}{\tau}\right)}-1} ; \qquad \frac{t_d}{\tau} = 1+3\frac{\ln R/r}{\ln 2} \tag{3.29}$$

In a very similar way, for kinetics (3.3), we obtain

$$\frac{r}{3K} = \frac{2^{\left(\frac{t_d}{\tau}\right)}-1}{2^{\left(\frac{3t_d}{2\tau} - \frac{1}{2}\right)}-1} ; \qquad \frac{t_d}{\tau} = 1+2\frac{\ln R/r}{\ln 2} . \tag{3.30}$$

The kinetics in (3.4) leads to

$$\frac{r}{3K} = \frac{\frac{t_d}{\tau}\sqrt{2}}{\left(1+\frac{t_d}{\tau}\right)^{3/2} - 2} \quad ; \qquad \frac{t_d}{\tau} = -1+2\frac{R^2}{r^2} . \qquad (3.31)$$

The ratio $\frac{t_d}{\tau}$ is plotted against $\frac{r}{3K}$ in figure 5. for all the three kinetics.

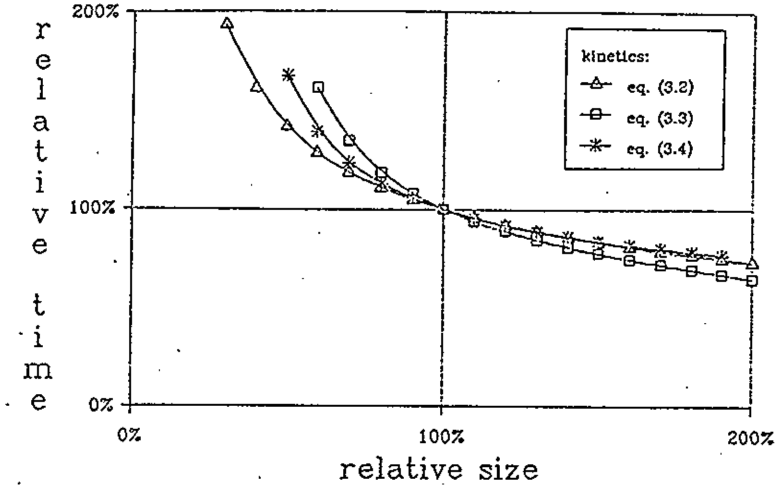


Figure 5.

The relationships between the relative vesicle size and the relative division time for different kinetics.

4. Discussion

In this paper we have discussed a mechanochemical model driven by three classes of chemical kinetics using a variational approach. The model displays self-reproduction of which the rate can be controlled by polymorphic kinetics. Thus it was possible to connect heterochronic and bifurcation behaviour, which are probably the alternative modes for the generation of spatial structures in the evolution of development.

We think that this study suggests for us that slow mechanochemical events serve as a source of developmental homeostasis or stability under the transformations of fast state transitions.

References

Alberch, P. - Gould, S. J. - Oster, G. F. and Wake, D.B. (1979). Size and shape in ontogeny and phylogeny. Paleobiology 5, 296..332.

Chevalier, Y. - Zemb, T. (1990). The structure of micelles and microemulsions. Rep. Prog. Phys. 53 279..371.

Deuling, H. J. - Helfrich, W. (1976). Red blood cell shapes as explained on the basis of curvature elasticity. Biophys. J. 16/8, 861..868;

Evans, E. - Rawicz, W. (1990). Entropy driven tension and bending elasticity in condensed-fluid membranes. Phys. Rev. Lett. 64, 2094..2097.

Goodwin, B. C. (1984). A relational or field theory of reproduction and its evolutionary consequences. In: Beyond neo-Darwinism. ed. Ho, M. W. - Saunders, P. T. pp. 219..241. Academic Press, London.

Helfrich, W. (1973). Elastic properties of lipid bilayers: theory and possible experiments. Z. Naturf. (c) 28 693..;

Helfrich, W. (1974). The size of bilayer vesicles generated by sonication. Phys. Lett. A 50, 115..116.

Koch, A. L. (1985). Primeval cells; possible energy generating and cell division mechanisms. J. Mol. Evol. 21, 270..277.

Luke, J. C. (1982). A method for the calculation of vesicle shapes. SIAM J. Appl. Math. 42, 333..345.

Molnar, I. - Verhas, J. (1990). Mechanical transformations of development in evolution. In: Organizational constraints on the dynamics of evolution. ed. Maynard Smith, J. - Vida, G. pp. 89..105. Manchester University Press, Manchester.

Oster, G. F. - Cheng, L. Y. - Moore, H. H. - Parelson, A. S. (1989). Vesicle formation in the Golgi apparatus. J. Theor. Biol. 141, 463..504.

Szatmary, E. (1989). The emergence, maintenance and transitions of earliest evolutionary unites. Oxford Surv. Evol. Biol. 6, 169..205.

Tanford, C. (1980). The hydrophobic effect: Formation of micelles and biological membranes. Wiley, New York.

Tarumi, K. - Shwegler, H. (1986). A non-linear treatment of the protocell model by a boundary layer approximation. Bull. Math. Biol. 49, 307..320.

Verhas, J. (1988). A liquid crystal model for early cell division. Liquid Crystals 3, 1183..1190.

ORGANIZATIONAL RULES, MORPHOGENETIC TRANSFORMATIONS AND EVOLUTION

Istvàn Molnàr
Eötvös Lorand University, Department of Genetics, Budapest,
Múzeum krt. 4/a. H-1088

Received on 31st Oct., 1990

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

1. AN INTRODUCTION TO THE EVOLUTION OF DEVELOPMENTAL SYSTEMS

The neo-Darwinian theory of evolution has been described in terms of three properties ever since Darwin.These properties are heredity, variation and reproduction (cf. Maynard Smith, 1986 for a more detailed discussion of this picture of evolution). The component processes of evolution can be classified at least into mutations, adaptations, constraints.frequency changes, rate-and direction determining processes (Endler & McLellan, 1988).

The traditional developmental processes include the growth generating body size, the cell differentiation and its spatiotemporal organization, the pattern formation,generating ordered heterogeneity of the organisms, and morphogenesis, governing the shape, form and in general the morphology of organisms.

A central problem of evolution is to reveal the connections of evolutionary and developmental component processes. Although there exists a vast body of work discusing the relationships between development and evolution (Gould, 1977;Bonner, 1982 ; Goodwin et al 1983; Raff & Kaufman 1983 ; Ho & Saunders, 1984 ; Molnar, 1984; Buss, 1987; Arthur, 1988 ; Thomson, 1988; Wake & Roth, 1989 ; Maynard Smith et al 1985 ; Kauffman, 1987 ; Alberch et al, 1979 ; Horder, 1989 ; etc.), there is no general agreement even on what problems must be solved, or what class of facts is to be explained.

In my view three general problems need to be address in order to get a deeper insight into the nature of the evolution of development. These are as follows : 1. The nature of the variation of developmental dynamics ; 2. The developmental control of evolution ; 3. The ecological and evolutionary control of development.(Molnar, in prep.).

The traditional view of evolution has been concentrated on the gene frequency changes of the populations.The reason was that the evolutionary play work in ecological theatre of organisms. The evolution, however, can be regarded as a process of collective developmental transformations in populations .If so, then the morphological transformations in populations may be derived from developmental transformations. Since the topic to be discussed here is the origins and transformations of the biological forms, I shall discuss the morphogenetic transformations

A brief survey of the theoretical aspects of the evolution of development include the following alternative point of views

Buss'theory of development can be briefly summarized : "The thesis developed here is that the complex interdependent processes which we refer to as development are reflections of ancient interactions between cell lineages in their quest for increased replication " (Buss, 1987, p.29). The dynamic competitive and cooperation of the coexistence of cell lineages do not incorporate the important physicochemical aspects of morphogenesis and pattern formation. The symmetry conservation and symmetry breaking mechanisms of cell states and cell lineages is neglected, except the inductive interactions.It has been proposed (Maynard Smith, 1990; Molnar, 1990) that dual inheritance consisting of the DNA and its patterns of methylation may be a plausible symmetry breaker, of which the amplification or suppression may account for a competitive .cellular diversity (see Figs. 1, 2.).

The bifurcation theory of morphogenetic fields describe the generations of forms in a mechanochemical framework (Oster et al, 1980; 1988 ; Murray & Oster 1984;Goodwin 1990).This research line is inspirated mainly by the classic paper of Turing (1952), which is reaction-diffusion theory of ×morhogenesis .The morphogenetic field theory is unable to incorporate the collective, populational behaviour of ecologically interacting populations.Therefore, in this view of evolution the morphogenetic fields are closed important *transformations, such as the error* correcting and variation generating sexual processes.

A similar evolutionary picture has been proposed, in its spirit, in which the organisms are regarded as the variations of infolding of shell/membrane systems (Molnar & Verhas, 1990).

A fundamental aspect of the evolution of organisms is the interactive dissipative dynamics and the (sequential and parallel) information processing coupled via the genetic code (Maynard Smith, 1986).

The ensembles of (epistatically interacting) genes can be represented by a random network of 0 (inhibited) and 1 (activated) elements and the interactions of these random genetic networks retain generic, ordered behavior(Kauffman, 1987). The developmental logic immanent in evolution suggest a deep structure for biology, in which the hereditary variations are themes on generic, frozen organizational configurations (Kauffman, 1987 ; Goodwin, 1990).At this moment this evolutionary concept seems to violate the evolutionary continuity of descent by modifications, unless we assume that the basins of the attractors of genetic networks and of morphogenetic fields are leaked.In the latter case the generic, frozen organizational features can not be frozen.Notice, for instance, that even the genetic code is changing.

Arthur (1988) in a different theory of evolution of development proposed the ×morphogentic tree set of hypothesis.The morphogenetic tree describes the causal structure of development, which he represented by a tree consisting of binary branchings, describing the tree-like propagation of the developmental and genetic heterogeneity.The structure of the tree can be transformed by selection.This clear view of developmental transformations does not contain real ontogenetic mechanisms.Also, it is build up a reification, being the ×morphogentic tree is a representation, and the selection reality.Similar reifications can be observed in those cases, when it is suppose that "adaptive evolution is governed by interactive, rugged adaptive landscapes, or the ×morhogenesis and its variations are governed by the solutions of morphogenetic field equations.Although I sympathize with all these views, this short analysis of their content may suggest the acceptance of their joint merits.

2. DEVELOPMENTAL CONTROLS IN EVOLUTION

Traditionally, by developmental controls genetic, epigenetic and environmetal controls are meant.

As regards the genetic controls, the general view is that there exist a genetic program, governing development and its evolution (Davidson, 1990 ; Cavener, 1989 ; Raff & Kaufman, 1983 ;but see Oster et al 1988, and Goodwin 1990 for counter-examples.

It is not implausible to regard the genetic effects as governors of developmental parameters, such as the rate, time of differentiation, movements, death of the cells and their surface properties which act in the spatiotemporal organization of cell populations.The nuclear and cytoplasmatic gene effects act via protein synthesis .The mentioned developmental parameters are in turn *regulators of gene effects via induction*.The genetic regulatory architectures are organized in a combinatorial superpositional way (*garcia-Bellido, 1986).These are specific to each form od ontogenesis.Common regulatory rules and principles are to be found only at the lower levels of the hieararchically organized gene controls (Davidson, 1990)

The cell surface contains coupling regulatory molecules *teinberg &Poole, 1981 ; Edelman, 1986). Edelman described cell and subtrate adhesion molecules, for instance, showing regularities in their spatiotemporal appearance.The mechanochemical cell structures respond by mechanical deformations or by chemical changings to concentration gradients.The local and global order in some cases, like in the fruit fly, can be described by spatial harmonics (Goodwin, 1990).The progressively finer subdivision of embryos uses huge amount of variation, based on a relatively few developmental parameters.

As to the environmental controls of development, an important *characteristic of a genotype is its reaction norm*.The reaction norm is the phenotypic response of a genotype to an environmental parameter gradient (Lewontion, 1983), providing information on the possible range of phenotypic variation (plasticity), and stability.

The polymorphism of the environmental controls is legio. Ex*amples from algal *morhogenesis reveal* (Trainor, 1970) that nutritional controls are capable of producing such morphogenetic changes that even the species-specific traits may be lost (Fig. 3). Environmental controls are capable of reorganizing food chain reactions by morphogenetic changes. Certain rotifers react to vi*tamin E treatment by increasing body size*, capturing larger preys (Gilbert, 1989.)

Instead of listing the unexhaustible cases of alternative environmental controls, how can we summarize the current ecological tendencies in the understanding of the connections of individual, inter-and intraspecific interactions ? Tilman (1990, p.5.) has proposed a promising, particularly succint idea : "(T)he development of predictive ecological theory requires 1. the determination of the major environmental constraints, 2. the determination of the tradeoffs that organisms face in dealing with these constrains, and 3. the explicit inclusion of these constraints and tradeoffs as the mechanisms of intraspecific and interspecific interactions ."These ecological ideas can be useful in the exploration of morphological evolution.

3. MORPHOGENETIC TRANSFORMATIONS

The transformations are relations, by which one structure, function or process is *converted into an another one .By morphogenetic* transformation I mean a change of a developmental pathway, in which the generation of form by the altered developmental pathway is also altered.Here a brief survey of the variations of the morphogenetic transformations in the evolution of bacteria, protists, plants and animals *show* a number of regularities, constraints and tradeoffs. The forms of the bacteria is controlled by the mechanical equilibrium between the cytoplasmatic pressure and the surface tension-like force of the cell wall (Koch, 1990). The *bacteria show* a relatively limited form diversity (see Fig. 4). The world of bacteria are dominated by a biochemical functional diversity. The protists have cytoskeleton (a spatial filamental meshwork), which is capable of generating jointly with the cell membrane or different kinds of internal or external inorganic skeleton (cf.Goodwin, 1989). The essential form *generator* of eucaryotes is the variable controlled cytoskeletal deformations, including ionic, mainly calcium-regulated controls.According to May (1988), the recorded number of protist species is 260000.

As the Fig. 5 illustrates, the plant form must satisfy antagonistic design requirements. A *specific plant characteristic* is their rigid cell wall.The assembly of plant cells can be described in terms of the theory of cellular solids (Niklas, 1989). The plant body consists of coupled population of physiologically and mechanochemically integrated cell populations or modules (cf. Gottlieb, 1984 ; Harper et al 1986; Klekowski, 1988 ; Groff & Kaplan, *1988).Plants posses* opened growth systems.Their modul number, shoot and growth length, flowering nodes, presence, absence and the angle of branchings, etc.can be manipulated genetically.The plant form change occurs at the level of the metamere (leaf-axillary bud-internode). A fundamental plant developmental strategy is the spatiotemporal allocation of the meristema governing the plant geometry within the phyllotactic and other constraints, of which the genetic and environmental perturbations are not well understood. As Cullis (1988, p.60) demonstrated, "higher plants have a genetically controlled variation system.This can be activated by a number of shocks.This genomic variation can be manifest as photypic variation *from which better* adapted lines can be selected.In the absence of any "shocks" the variation system is not active, or active at a very low level, and it is the exposure of the organism to the "shock" which causes variation.The limitation of the variation to a subset of the genome, which is controlled by the physiological *state of the* cell, gives the variation a Lamarckian dimension in that repeated exposures to the same shocks generate the same range of variants." The estimated number of plant species is 235000 (Barnes, 1989). The evolutionary variations of the shoo-root systems can be classifiable (Groff & Kaplan, 1988), into the following classes : 1.Neither *shoot-borne roots nor root-borne shoots* develop; 2. The plant forms shoot-borne roots but not root-borne shoots. 3. The plant forms root-borne shoots but not shoot-borne roots; 4. The plant form both shoot- borne roots and root-borne shoots. The important ecological implications of this classification was recognized by Harper (cit. Groff & Caplan 1988, p.410) : " *I would expect* plants forming roots from their shoots to be limited by light, and those forming shoots from their roots to be lited by water and nutrients."

Figs. 6-8 show the evolutionary transformations of the roots, leaves, and flowers.

The evolution of animal and plant forms are dominated by cylidrical geometry, with other plate-or blob-or-spherical forms. The animals can have 1. stiff, branching cylinder form; 2.fluid-filled, fiber-wound hydrostats ; 3. kinetic design with flexible joints.

My own view of the evolutionary transformations of the biological forms can be briefly summarized in the following form. The origins, maintenance, and transitions of biological forms are driven by three coupled processes : 1. By a variety of physicochemical properties of cells and cell populations ; 2. By selective mechanisms characterizing the components and processes of developing organisms ; 3. By a historical sequence of environmental controls. (Molnar, in prep).

The outline of the argument is as follows . It is demonstrated that the physico-chemical dynamics of the development can explain a number of ontogenetic processes (Oster et al, 1980; Oster & Murray 1984 ; Goodwin, 1990 ; Thompson, 1917, Huzell 1933).

Tremendous evidence show that selective cell death in the generation of vertebrate limb, in the dveloping nervous system, in insect and vertebrate immune system, in the aspecific and specific effector system, in developmental selection of somatic mutations in plants do contain selective components and processes, driven by mechanochemical drive (Klekowski, 1988, Michaelson, 1987, Sach, 1988, Hinchliffe, 1981., Buss 1987, 1988 etc.).

4. ORGANIZATIONAL RULES AND EVOLUTION

Organisms, their parts, populations, communities seem to show a set of regularities in their origin, maintenance and transitions (Huxley, 1932 ; Harper et al, 1986, Bell, 1984; Goldwasser 1988; Lindenmayer, 1982; Oster et al 1988; Mitchison & Wilcox, 1972 ; Green, 1987 ; Edelman, 1986 ; Drake, 1990 ; French et al, 1976 ; Horder, 1989 ; Hinchliffe & Johnson, 1980 ; Tomlinson, 1983 ; Raup, 1966; Kauffman, 1987 ; Lewis, 1981 ; Goodwin, 1990 ; Davidson, 1990 ;Hadorn, 1978; Wake & Roth, 1987 ; Ellison & Harvell, 1989 ;Londsdale, 1990 ;).

There are many concepts of the rule of transformations at different levels of organization a brief survey of these rule concepts clarify their meaning. After an overview some useful rules, the following question will be address : What are the developmental and evolutionary relevances of these rules ?

The genetic coding rules and the Mendel's rules are the most familiar organizational rules. It is perhaps surprising that the genetic coding rules are not universal, "frozen" rules, but these are changing in the evolution (Osawa et al, 1990). Also, there exist genes that violate Mendel's rules (Crow, 1979). If such fundamental rules are of dynamic nature, it is expectable that the lesser fundamental rules show much more "wobbling" in the self-assembly of biological structures.

The developmental rules are a subcase of the organizational rules, because they are restricted to the level of developing organisms. In the pre-Darwinian comparative embryology von Baer (1828) had already discovered "laws of development". Gould (1977, p.56.) put these laws into English in the following way ("as probably the most important words of in the history of embryology, "according to Gould). "1. The general features of a large group of animals appear earlier in the embryo than the special features. 2. Less general characters are developed from the most general, and so forth, until finally the most specialized appear. 3. Each embryo of a given species,..., instead of passing through the stages of other animals, departs more and more from them. 4. Fundamentally therefore, the embryo of a higher animal is never like the adult of a lower animal, but only like its embryo."

The validity of these rules are of statistical nature, because there are counter-examples. Von Baer's laws of development have been reinterpreted in the context of the bifurcation theory (Molnar & Verhas, 1990). By the iterations of symmetry breaking instabilities a progressively finer spatiotemporal subdivision can be generated in model embryos. The governing equation for the case of gastrulation has been described in this volume, too (see Verhas & Molnar).

In the current developmental biology we can observe cellular, tissue and organ rules. This is carefully analysed for the case of phyllotaxis, in general for the organization of the shoot of plants (Green, 1987). Green analysed the three major philotactic (leaf placement) patterns. The leaves may be in a helical arrangement, alternating in a plane, or whorled. Greens' definition of rule is as follows : "The most efficient rules have the instantaneous character found in a time-based differential equation. When integrated over time and space, the rule produce the /developmental/ behaviour (Green, 1987, p.657)."

The three major phylotactic patterns are variation on a single generative theme. (See Jacob, 1982, that evolution is tinkering of old structures). Green's explanation of phyllotactic patterns is based on purely mechanical activity of the iterative activity of cellulose reinforce pattern of the shoot surface. This activity can be reduced through the cytoskeletal cell behaviour to microtubules and associated proteins, finally to RNA and DNA.

In this attractive scenario the explicit reference to direct genetic form controls, photosynthetically or nutritionally induced, leave-forming meristematic excitability is lacking. The lack of environmental control of the leaf-forming or other functional dynamics of the rule-equipped system contains necessary, but not sufficient developmental description.

For the description of dynamic plant morphology and foraging Bell proposed constructional rules (see Bell, 1984, for a review). The plant construction can be described by the probability of branchings, the lengths of the modules (internodiums), and the angles of the branchings. The iterations of this parameters simple rules of growth lead to a variety of plant architecture.

The limited set of basic tree architectures can also be treated by the iterations of simple rules of growth (Tomlinson, 1983). Halle et al. (1978) have reduced the basic tree forms to 23 classes or "models". As Tomlinson (1983 p.142.) claims, "trees develop according to a precise genetic ground plan-the architectural model-that determines the form of the tree." Tomlinson also notes that the demography of tree populations is of very important role on the determination of tree forms. Niklas (1986, 1989) put forward the general view of plant architecture, stressing the trade-offs (measured in negative correlations whenever possible) of the competitive requirements between the components of the plant design (see Fig. 5).

5. CONCLUSION: CELLULAR AND MODULAR COEVOLUTION

There are two major classes of developmental units with a number of transients between them, of which the spatiotemporal iterations generate the diversity of organisms. The first is the cell characterizing unitary ("solitary") organisms, and an integrated cell population, the modul, characterizing modular individuals.

The interactions between the different forms of selection and the iterative aggregations of the cellular and modular units suggest an evolutionary view, which may throw light on a unified treatment of the evolution of unitary and modular organisms. The selection for cellular and modular aggregations is driven by polymorphic symmetry conserving and symmetry breaking physico-chemical mechanisms. These mechanisms may be expressed as phenomenological rules at many levels of the biological organization. It is reasonable to suppose that the symmetry breaking and conserving mechanisms are also under selection, for the generated structures by them posses selectable functions.

The most probable element in the initiation of the evolutionary divergence of the unitary and modular organisms may have been the emergence and subsequent evolutionary changes of the stem cells. The stem cells are capable of producing themselves ×mainteningcell and tissue homeostasis, and other cell types, creating cellular diversity within and between organisms.

A plausible *proposition is* that the *interactions of iterative* cellular and modular developmental units, respectively, have resulted in coevolution of cells and modules by the selective modification of the rules and boundary conditions of the iterations of the developmental units. There is a vast amount of evidence, that the main ecological·interactions, such as competition, parasitism, predation, mutualism, etc. can be observed in bacteria (especially their multicellular forms selected for mainly foraging), protists, animals, fungi and plant.It is a commonplace that the plants and modular (mainly sessile) animals can be regarded as a population of modules.The dynamics of these modules are governed by joint mechanochemical forces, and ecological, mainly nutritional and demographical controls.

In summarising, in this paper I hope to demonstrate that at many levels of the biological organization there exist an ordered set of regular events, which are manifested at genetic, developmental, ecological and evolutionary scales.These regularities can be described by *organizational rules*. A deductive theory of biological organization, of which the predictive one only a special case, must contain the clearer relationships of the organizational rules, the boundary conditions or constraints within which the rule are operating, and the effect of historical contingencies making a too regular biological world more realistic.

REFERENCES

Alberch,P.,S.J.Gould,G.F.Oster,&D.B.Wake. 1979. Size and shape in ontogeny and phylogeny. Paleobiology 5, 296-317.

Arthur,W. 1988. A Theory of Evolution of Development.Wiley,London

Barnes,R.D. 1989. Diversaity of organisms : how much do we know ? Amer.Zool. 29,1075-1084.

Black,I.B. 1986. The trophic molecules and the evolution of nervous system. Proc.Natl.Acad.Sci.USA. 83,8249-8252.

Bonner,J.T. 1988. The Evolution of Complexity by means of Natural Selection. Princeton University Press,Princeton

Brown,N.A.& L.Wolpert. 1990. The development of handedness in left/right asymmetry. Development 109 : 1-9.

Buss,L.W. 1987. The Evolution of Individuality.Princeton University Press,Princeton

Buss,L.W. 1988. Diversification and germ-line determination.Paleobiology 14,313-321.

Cavener,D.R. 1989. Evolution of developmental regulation.Amer.Nat. 134,459-473.

Cullis,C.A. 1988. Control of variation in higher plants.In:Evolutionary Processes and Metaphors.(M.-W.Ho & S.W.Fox,eds.) pp.49-62. Wiley,New York

Davidson,E.H. 1990. *How embryos work : a comparative view of cell* fate specification. Development 108 : 365-390.

Ebbesson,S.O.E. 1980. The parcellation theory and its relation to interspecific variability in brain organization, evolutionary and ontogenetic development,and neuronal plasticity. Cell Tissue Res. 213,179-212.

Endler,J.A. & T. McLellan. 1988. The processes of evolution:toward a newer synthesis. Ann.Rev.Ecol.Syst. 19 : 395-421.

Galloway,J.W. 1989. Reflections on the ambivalent helix. Experientia 45,859-872.

×Garcia-Bellido,A. 1986. Genetic analysis of morphogenesis. XVII.Stadler Genetic Symp. 187-209.

Gould,S.J. 1977. Ontogeny and Phylogeny. Harvad University Press, Cambridge

Goldsmith,T.H. 1990. Optimization,constraint,and history in the evolution of *eyes. Q.Rev.Biol.* 65,281-322.

Goodwin,B.C. 1988. Morphogenesis and heredity. In: Evolutionary Processes and Metaphors. (M.-Ho & S.W.Fox,eds.) pp.145-162. Wiley,New York

Goodwin,B.C. 1989. Unicellular morphogenesis. In: Cell Shape : Determinants,Regulation, and Regulatory Role. pp.365-391.Academic Press,New York

Goodwin,B.C. 1990. Morphogenesis : gene action within the context of cytoplasmic order. In : G. Malacinski,ed. pp.3-22. McGrow-Hill

Gottlieb,L.D. 1984. Genetic and morhological evolution in plants. Amer.Nat. 123,681-709.

Green,P.B. 1987. Inheritance of pattern : analysis from phenotype to gene. Amer.Zool. 27,657-673.

Groff,P.A. & D.R.Kaplan. 1988. The relation of root systems to shoot
systems in vascular plants. Bot.Rev. 54,387-422.

Gurdon,J.B. 1988. A community effect in animal development. Nature 336,772-774.

Hastinngs,I.M. 1989. Potential germline competition in animals and its evolutionary implications. Genetics : 123, 191-197.

Holliday,R. 1987. The inheritance of epigenetic defects. Science 238,163-170.

Jacob,F. 1982. The Actual and the Possible.Pantheon Books,New York

Katz,M.J. 1983. Ontophyletics : studying evolution beyond the genome. Persp.Biol.Med. 26,323-333.

Kauffman,S.A. 1987. Developmental logic and its evolution. BioEssays 6: 82-87.

Kirk, D.L. 1988. The ontogeny and phylogeny of cellular differentiation in Volvox. Trends Genet. 4,32-36.

Klekowski,E.J.Jr. 1988. Mutation, Developmental Selection, and Plant Evolution. Columbia University Press, New York

Lewontin,R.C. 1983. The organism as the subject and object of evolution. Sciencia 118,56-82.

Lauder,G.F. 1981. Form and function : structural analysis in evolutionary mormphology. Paleobiology 7,430-442.

May,R.M. 1988. How many species are there on Earth ? Science 241,1441-49.

Maynard Smith,J. 1990. Models of a dual inheritance systems. J.Theor.Biol. 143,41-53.

Maynard Smith J.,et al. 1985. Developmental constraints in evolution. Q.Rev.Biol. 60 : 265-287.

Michaelson,J. 1987. Cell selection in development. Biol.Rev.62,115-139.

Molnar I. 1990. The evolutionary aspects of dual inheritance in modular and unitary organisms. Symp.Biol.Hung. 39 , 355-362.

Molnar I. & J. Verhas. 1990. Mechanical transformations of development in evolution. In: (J.Maynard Smith & G.Vida,eds.) Organizational Constraints on the Dynamics of Evolution. pp. 89-105. Manchester University Press, Manchester

Murray, J.D. & G.F.Oster. 1984. Generation of biological pattern and form. J.Math.Med.Biol. 1, 51-75.

Oster, G., G.Odell, & P.Alberch. 1980. Mechanics, morphogenesis and evolution. Lect.Math.Life Sci. 13, 165-255.

Oster, G.F., N.Shubin, J.D.Murray & P. Alberch. 1988. Evolution and morphogenetic rules : the shape of the vertebrate limb in ontogeny and phylogeny. Evolution 42, 862-884.

Niklas, K.J. 1986. Evolution of plant shape : design constraints. TREE 1, 67-72.

Niklas K.J. 1989. The cellular mechanics of plants. Amer.Sci.77, 344-349.

Palmer, A.R.& C. Strobeck. 1986. Fluctuating asymmetry : measurment, analysis, patterns. Ann.Rev.Ecol.Syst. 17, 391-421.

Raff, R.A. & T.C.Kaufman. 1983. Embryos, Genes, and Evolution. Macmillan, New York

Rosen, J. 1990. Fundamental manifestations of symmetry in physics. Found.Phys. 20, 283-307.

Sachs, T. 1988. Epigenetic selection : an alternative mechanism of pattern formation. J.Theor.Biol. 134, 547-559.

Sackville Hamilton, N.R., B.Scmidt, & J.L.Harper. 1987. Life-history concepts and the population biology of clonal organisms. Proc.R.Soc.Lond.B. 232, 35-57.

Steinberg, M.S. & T.J.Poole. 1981. Strategies specifying form and pattern: adhesion-guided multicellular assembly. Phil.Trans.R.Soc.Lond.B. 295, 451-460.

Steinberg, P.W. & H.R.Horvitz. 1984. The genetic control of cell lineage during nematode development. Ann.Rev.Genet. 18, 489-524.

Thompson, D' Arcy, W. 1917. On Growth and Form. Cambridge University Press, Cambridge

Trainor, F.R. 1970. Algal morphogenesis : nutritional factors. Ann.New York Acad.Sci. 175, 749-756.

Turing, A.M. 1952. The chemical basis of morphogenesis. Phil.Trans.R.Soc.Lond.B. 237, 37-72.

Wainwright, S.A. 1988. Axis and Circumference.Harvard University Press, Cambridge

Wake, D. & A.Larson. 1987. Multidimensional analysis of an evolving lineage. Science 238, 42-48.

Mitchison J. H. Wilcox 1972. Rule governing cell division. Nature

Webster, G.C. & B.C.Goodwin. 1982. The origin of species : a structuralist approach. 5, 15-47.

Tilman, D. 1990. Constraints and tradeoffs : toward a predictive theory of competition and succession. Oikos 58, 3-15.

Goldwasser, L. 1988. Branching patterns, generating rules, and astogenic trajectories in arborescent bryozoans. J.Theo.Biol. 132, 179-201.

Lonsdale, W.M. 1990. The self-thinning rule : dead or alive ? Ecology 71, 1373-1388.

Ellison, A.M & Harvell, C.D. 1989. Size hierarchies in Membranipora membracea : do colonial animals follow the same rules as plants ? Oikos 55, 349-355.

Raup, D.M. 1966. Geometric analysis of shell coiling : general problems. J.Paleontology 40, 1178-1190.

Koch, A.L. 1990. Growth and form of the bacterial cell wall. Amer. Sci. 78, 327-341.

Huxley, J.S. 1932. Problems of relative growth. Dover, New York

Leppik, E.E. 1977. The evolution of capitulum types of the Compositae in the light of insect-plant interactions. In : The Biology and Chemistry of the Compositae. vol.1. (V.H.Heywood, J.B.Harborne & B.L. Turner, eds.) pp.61-89..Academic Press, New York.

Lindenmayer, A. 1982. Developmental algorithms : lineage versus interactive control mechanisms. In : Developmental Order : Its Origin and Regulation. pp.219-245. New York, Alan R.Liss.

Mitchison, G.J. & M.Wilcox. 1972. Rule governing cell division in Anabena. Nature 239, 110-111.

Edelman, G.M. 1986. Cell adhesion molecules in the regulation of animal form and pattern. Ann.Rev.Cell Biol. 2, 81-116.

Osawa, S., A.Muto, T.H.Jukes & T.Oham. 1990. Evolutionary changes in the genetic code. Proc.R.Soc.Lond.B. 241, 19-28.

Crow, J. 1979. Genes that violate Mendel'rules. Sci.Amer.

Gilbert, J.J. 1980. Developmental polymorphism in the rotifer Asplanchna sieboldi. Amer.Sci. 68, 636-646.

Horder, T.J. 1989. Syllabus for an embryological theory. In : D.B.Wake & G.Roth, eds.: Complex Organismal Functions pp.315-348. Wiley, New York

Bonner, J.T.ed. 1982. Development and Evolution. Springer, Berlin

Goodwin, B.C.et al eds. 1983. Evolution and Development. Cambridge Univ.Press.

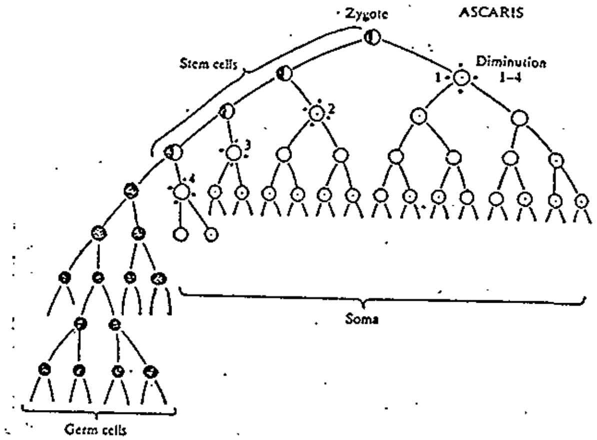


Fig. 1. Relationships between stem cells, germ cells and soma in Ascaris (After Wilson, 1925).

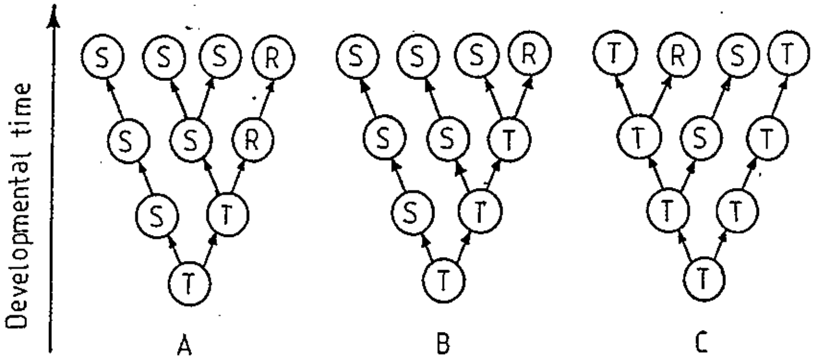


Fig.2 Preformistic /A/, epigenetic /B/ and somatic embryogenetic/C/ basic developmental modes,describes by Buss /1987/.
T denotes totipotent cells,generating themselves and S somatic and R reproductive cells.

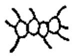
STRAIN	FORM	NUTRIENTS	RESPONSE
S. obliquus		calcium	
N. 46		thiamine	
		soil extract	
276-4e		phosphorus	
culture 16		NH$_4$ at pH 8.5	
		10 µg Fe/l	

Fig. 3 The effects of different nutrients on the morphogenesis in several Scenedesmus strains (After Trainor, 1970).

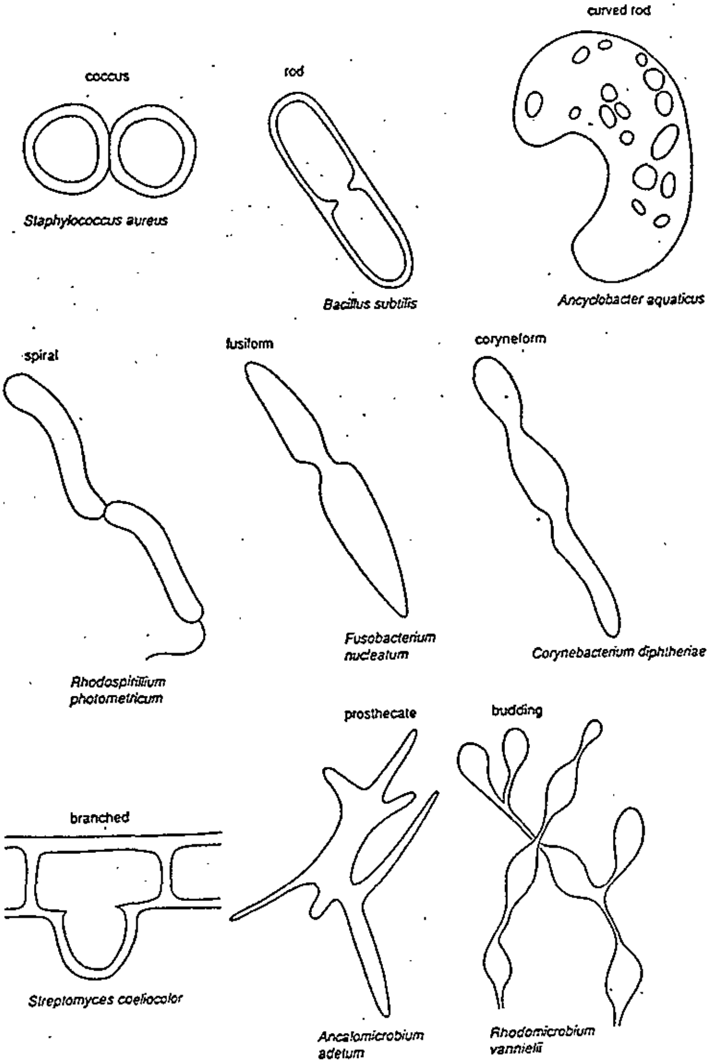


Fig. 4 A representative set of basic bacterial forms (After Koch, 1990).

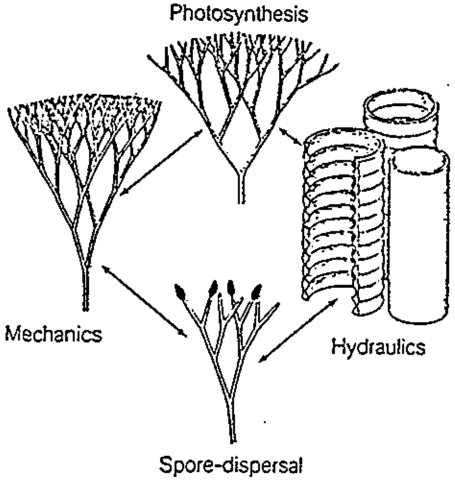


Fig. 5 Antagonistic design requiements in plant growth and reproduction. A stylized vascular plant shows tradeoffs in the plant architecture /After Niklas, 1986).

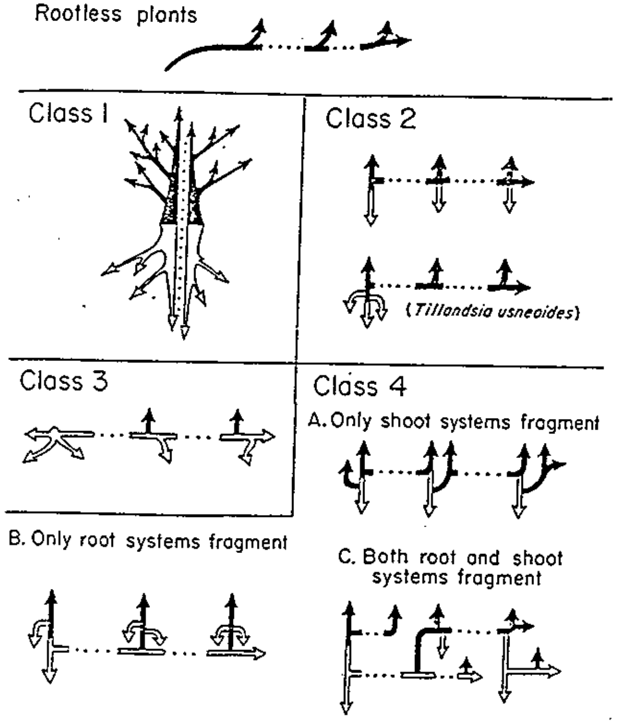


Fig. 6 The variability of the modes of clonal growth among the structural classes of coupled root-shoot systems (After Groff and Kaplan, 1988).

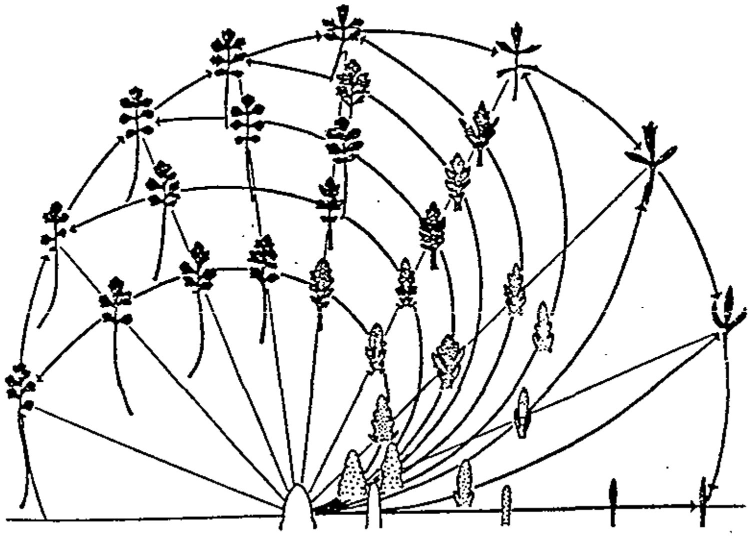


Fig. 7 Generative transformations of the morphogenetic sequences of leaves from a garden cress plant (Lepidung sativum). All leaf buds emerge from the same basic shape, progressing along curving form trajectories from the initial bud to the final leaf (After Goodwin, 1988).

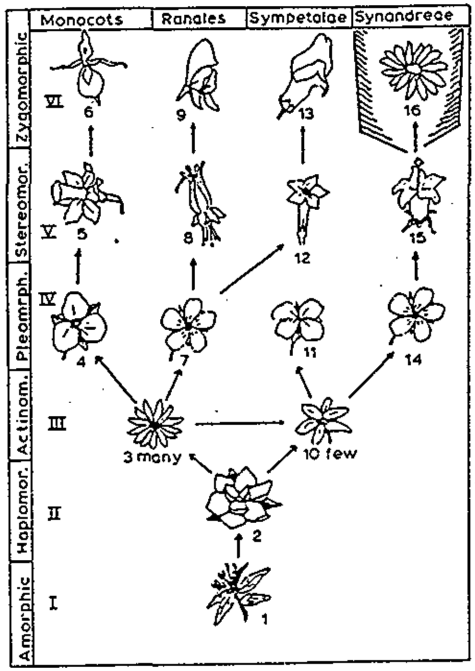


Fig. 8 The transformations of flower forms during the evolution of angiosperm. The set of transformations generated relatively few basic forms having simple symmetries in the last 100 million years (After Leppik, 1977).

ON THE RISE AND DECLINE (R&D) OF BRAINS

Received on 15th Oct., 1990

E. LÁBOS

Semmelweis Medical School, 1st Dept. of Anatomy, Neurobiology Unit. Budapest,

Hungary

Postal address:

E.Labos

Semmelweiss Medical School

1st Department of Anatomy

Neurobiology Unit of Hungarian Academy of Sciences

1450 Budapest IX

Tuzolto-u. 58

HUNGARY

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

ABSTRACT

A survey of the principal stages of neural evolution and cephalogenesis is presented. The animal taxons are specifically clustered according to comparative data and hypothetical trends. At last, a special approach to evolution is proposed of which the key-concept is called" phylogenetic forest". It means that various ordering criteria may determine various phylogenetic trees of which the correlation analysis or comparisons are additional descriptive steps. It may happen that these trees cannot be unified into a single one.

INTRODUCTION

The emergence of the so called intelligent, conscious human behavior together with its substrate, the human brain is an unsolved problem in the theory of evolution.

The answer to the other question, whether the theory of evolution itself is tenable or untenable, in this special case depends mainly on the philosophical or speculative background. A theory which is unsufficiently supported by facts and thus it is refuted - as Spencer says - is replaced often by nothing.

The question of brain evolution seems exceptionally difficul, since it is regarded as a 'jump', unexpected from the foregoings. No satisfactory transitional cases are demonstrable with assurance and in all respect. Particular examples are yet collectable. For example, the man as a tool-handling and tool-making subject is not unique among the animals. Formal equivalents or archetypes of several human social behavioural patterns also exist.

However, not the 'evolutionary jump' represents only the difficulty. *Jumps occur everywhere. However, this jump resulted in man is outstanding.*

What is usually regarded as evolutionary approach concerning any kind of biological substrate, not only in the case of brain? Jerison (1973) claims that an investigation is regardable 'evolutionary' if beyond comparisons and/or classifications it is historical. It would mean that it takes into account e.g. fossil brains. Regrettable enough, the actually disposable items are rather endocasts and not brains. Intact fossil brains practically are not available, only their rather empty places which at most limit the size and volume and rarely very residual soft parts. This holds in cases when skulls existed at all. More often, even no solid envolopes can be found or whole animals are lacking. At 'lower' phyla the lack of solid coverings makes reconstruction of fossil nervous systems practically impossible. The *Haeckelian law, the recapitulation of phylogenesis by the ontogenesis* is *not* reliable and this supposition gives only hints but does not replace solution.

Moreover, even in favorable cases, paleological findings alone does not determine uneqivocal partial ordering immanent in evolution. Furthermore, it is always unreliable to claim that a recent species is a follower of an extinct item. That is why, among others - as Jerison (1973) correctly sees - the so called recognized rules of brain evolution frequently proved to be erronous. Nevertheless, comparison of various neural ontogenetic pathways (if available at all) are useful stock of evolutionary principles.

Nervous systems and still more the brains are not ubiquitous in the Animal Kingdom. Nevertheless, the excitability is regarded as a universal property of living objects. In an extremly broad sense the 'responsivness' as representing the archetype of excitability can be even extended to all exsisting objects. Thus, a non-living item could say that 'respondeo ergo sum'.

The enigma finished with the appearence of human brain. It is a conscious item, capable of observing its own activity, recognizing its own existence. At the present time, no methods exist by which the activity of a conscious and a unconscious brain can be certainly distinguished from each other neither in natural nor artificial items. This dilemma is not solved by the conversational test of Turing (1951). However, in human medicine, the diagnosis of unconscious state of a patient works fairly well.

PRENEURAL EVOLUTION AND THE EXCITABILITY

While only excitable cells occur both in animals and plants, the nervous system (NS) is restricted to Metazoan species. Various NSs are specialized for quick control of all parts of the organism including itself. It may be surprizing that out of NS and muscle cells the 'non-excitable' stigma is also used as a professional term... The excuse of this attribute is that in neurons and muscle cells the signs of excitations are more spectacular than elsewhere: electrical signs and quick mechanical motions.

The nervous control includes responses to external stimuli and endogenous activity which is remote from ad hoc external agents. Moreover, some differentiated forms of nervous control appears to be purposeful or goal-oriented e.g. due to their preventive character. However, in a so called neutral (i.e. non teleological) jargon this means a high level of adaptive machinery.

Do exist unicellular systems (1CNS) in various Protozoan species which might correspond to nervous systems (NS) of multicellular animals or not? How the NS appears in primitive Meso-, Para- and Metazoan animals like those of Moruloidea, Porifera, Cnidaria or Ctenophora? When a cell may be called a neuron? How the sensory, transmitter and efferent functions become separated: from one cell or not (Kleinberg, 1872; Hertwig and Hertwig, 1878)? A multitude of similar problems are only partially solved mainly by suppositions which are hardly verifiable.

In Protozoa the electrical signs of excitation and resting state are demonstrable. Coordination of flagellary, ciliary and protoplasma motions is present in different unicellular organisms. Two problems are of paramount interest: (a) the autonomy versus centralization of these control processes and (b) the manner of coordination. Are these actions realized through protoplasma or by events restricted into the surface membrane? The preference is given mostly to the second alternative. Recent answers fails to confirm the presence of intracellular conductive fiber system inside any Protozoan species. It is also the negative response is preferable with respect to the presence of specialized NS in Moruloidea (Bullock and Horridge, 1965).

The various forms of taxis are observable in Protozoan behavior and remains present further in various multicellular non-primitive organisms. Taxes already may give the impression of 'intelligence'. This strictly suggests carefulness concerning the definition of 'intelligent behaviour'. Nevertheless, many taxis phenomena are well interpretable in terms of non-living control terms.

PRIMITIVE METAZOAN INVERTEBRATE NERVOUS SYSTEMS.
(IVNS-1; Bullock and Horridge, 1965; Lentz, 1968).

The existence of NS in 'higher' animal phyla becomes less and less questionable, becomes evident. The term 'higher' is based here on traditional (usually morphological and paleozoological) criteria out of neurosciences.

The invertebrate nervous systems might be categorized into different phylogenetic (or comparative) categories:

(a) Diffuse Nervous System (DNS) in Mesozoa, Parazoa and Acoelomata (Cnidaria, Ctenophora). There are differences: (1) In Moruloidea the existence of true nervous system is not well documented; (2) In Porifera the distribution of putative nerve cells is not regular; (3) In Cnidaria and Ctenophora the gross architecture is not completely diffuse.

(b) NS with ganglia (GNS) including cerebral one, segmentation, cords, comissures, i.e. an 'orthogon-like (ladder)' NS is present in various flatworms. Some orthogon-like NS-s can be regarded as archaic segmental or presegmental cases. The segmentation apppears in Annelida, Arthropoda and Archipodiata. Presegmental architecture or segmentation is not explicit in Mollusca with some exception: in Loricata and Monoplacophora (Neopilina).

(c) Differentiation of cerebral part (CGNS) arises from Platyhelminthes and further phyla. It does not associated with a definite segmentation.

(d) Special cases: e.g. Echinodermata ring-nervous system (RNS) or non-segmental Molluscan cases. The central symmetry of Echinodermata seems enigmatic until we know about their bilateral symmetry in larval form.

(e) Epithelial-ectodermal NS with local thickenings. Differentiation of peripheral and autonomous NS. Appears in various primitive or reduced worms. It is not a peculiarity even in Deuterostomia.

Diffuse systems are especially characteristic to Porifera and Cnidaria. However, no reliable data speaking for the appearence of nervous system are available concerning Moruloidea.

In Porifera classes Lentz (1973) reports about primitive NS, while Bullock and Horridge (after the analytical survey of Jenifer, 1962) do not give the affirmative response to the problem of existence of NS in sponges. Nevertheless, syncitium with cholinesterase activity of sponges or with the detection of other neurospecific substances argue for the emergence of neural specialization. DNS-s consist of syncitial set of cells identifiable either by neurospecific or by non-specific substances. In this sense numerous Porifera and still more Cnidaria species have already DNS (a reticulum or syncitium). Differentiation of sense organs also arise.

It is here necessary to define four terms of neurobiology mainly after Bullock and Horridge (1965) and other textbooks:

RETICULUM = Anastomosing (often keeping protoplasma continuity) fibers without specified connections;

SYNCITIUM = A multinuclear cell usually with elongated processes resembling to network. Often NS said to have syncitium consisting of disjoint cells.

PLEXUS = A general term of nerve fibers. The cells are either synaptically connected or contiguous components. The term is often used for peripheral NS;

NETWORK = Usually synaptically connected nerve cells. Sometimes applied for unspecified connected machinery.

Remark that these - in a part historical - terms are still used for the description of neuronal or pseudoneuronal connections in lower phyla. In the context of DNS all of these terms are applied. The essential differences -

included in the used terms - can be reduced to two properties:

(a) Continuity of protoplasma versus contiguity (= separation by external membrane);

(b) Cells with separated (disjoint) protoplasma may touch each other and the interactions might be either uni- or bidirectional with respect of possible influences through chemical and/or electrical signals.

When the cerebralization or encephalization reaches the degree at which the name of brain at least as a metaphora is justified? This name is applicable by convention. In textbooks it appears first at Platyhelminthes. Thus the attribute of 'cerebral' is used for otherwise rather primitive invertebrates. It labels the concentrated and separated group of cells at the 'cranial end' of the animal, where cranial is applied also as a metaphora taken from animals holding a 'cranium' (neolatin name of skull) or well-distinguished head.

The 'orthogon' and/or ladder-like NS architecture which occur in numerous phyla we regard as a 'preludium of segmentation'. Segmentation might have evolutionary advantage. Segments are capable of responding autonomously by local reflexes or control. At the same time the extensive longitudinal and transversal connections points to the possibility of coordination, occasionally synchronization or timing. These NS-s are multicentral consisting of interconnected centers. Such or similar architecture becomes quite general.

The cerebralization means that the first ganglion becomes dominant or at least different from other ones if they exist. Often the cerebral ganglion is not required to carry out specialized function. Analogous phenomenon is recognizable even at Vertebrata. Functions, regeneration and survival without 'brain' are heterogenous.

The emergence of 'orthogon' NS and the signs of mono- or multi-centralization in the form of ganglia cannot be sharply separated.

EXACT DESIGNS OF SOME METAZOAN INVERTEBRATE NERVOUS SYSTEMS.

(IVNS-2; collected from Bullock and Horridge, 1965 Vol 1, pp. 599-609).

While the number of cells is undefined and the significance of single neurons becomes smaller in 'higher' NS-s, there are certain 'lower' taxons in which the number of neurons is fairly constant or at least some neurons are well identified as homologous to each other.

Outstanding examples are as follows:

NEMATODA:

(1) Ascaris lumbricoides 162 cells in central NS: 62-23-77 sensory/internuncial/motor respectively; 92 neurons in tail and enteric system (Goldschmidt, 1908; Chitwood and Chitwood, 1940);

(2) Rhabditis anomala has the same number of central neurons as Ascaris has: 162;

(3) Anguilla aceti: 279 neurons;

(4) Rhabditis longicauda: about 200 cells;

It is here interesting to remark that in certain computer companies a research work is processed with respect of such better-identifiable nervous systems like that of Caenorhabditis elegans.

ROTIFERA (plus about 200 in other ganglia)

(5) Epiphanes senta: 183+34+23=240 central nerve cells (Martini, 1912);

(6) Asplancha priodonta: about 225+50+48=323 central neurons (Nachtwey, 1925);

(7) Synchaeta triophtalma: 223+38+20=281 in three main ganglion (Peters, 1931);

ACANTOCEPHALA:

(8) Hamaniella microcephala: 80 cells and 60 output lines for cerebral ganglion;

(9) Macracantorhynchus echinodiscus: 86 cells, 56 outgoing wires;

(10) Bolbosoma turbinella: 73 cells;

ANNELIDA:

(11) Pheretima communissima: for a typical ganglion of the 98 segments 1000-4500 cells, 400-1000 fibers going to about 50000-60000 muscle fibers were counted (Ogawa, 1939 cit. by Bullock and Horridge, 1965);

The individuality seems to be gradually lost when reaching the level of Annelida species. However, numerous Gastropoda and Cephalopoda nerve cells are fairly well identifiable.

The neuropile, the core of ganglia consisting mainly of fibers is characteristic in Mollusca. Its internal organization is essentially unknown. Some ganglia of Annelida shows much simpler structure with few neurons, like Hirudo.

MORE COMPLEX INVERTEBRATE NERVOUS SYSTEMS.

(IVNS-3; Bullock and Horridge, 1965)

Most probably, Arthropoda, Mollusca and Echinodermata species represent special special lines in evolution. Concerning of nervous system it appears that links to higher phyla are missing or still not correctly recognized. The emergence of Deuterostomia seems to be a jump.

Insect performances include even 'social behavior' (bees, ants). Cephalopoda are capable of differentiated learning.

Echinodermata display peculiar circular gross geometry of NS because of the more or less (pseudo)central symmetries of their body. It is necessary to remark here, that the larval symmetry of Echinodermata is bilateral, while the adult bilaterality is most often an obscure property. It is an outstanding funnctional problem here is related to the autonomous versus central coordination and the interplay of these two kinds of control. The principle of quick autonomous control together with centralized NS remains a conservative achievement which occurs in previous and further taxons too.

It is disturbing that neither neurochemistry, nor finer internal organization can point to the specific substrates of new achievements. Thus not well reproducible demonstrations are available with respect of molecular substrates of memory or learning which is already clearly present in Cephalopoda. Sperry says that:

'no one has yet succeeded in demonstrating anatomically a single fiber or fiber connection that could be attributed with assurance to have been implanted by learning' (page 8, Sperry, 1964).

His statement is not out of time and it is valid for non-antomical, e.g.

molecular substrates too.

The best we can do is the recording of differences, the reproducible *phenomena and give careful interpretation. Comparisons and classifications* represent steps forward. However, as soon as any kind of - otherwise possible - ordering is regarded as a foundation of evolution we usually speculate only. Nevertheless: 'speculatio necesse est'.

DEUTEROSTOMIA-1.

Two peculiar phyla: Homalopterygia(Chaetognatha; 50 species) and Pogonophora (Brachiata) merit special attention. In Homalopterygia two longitudinal cords are observable interconnecting ganglia and one larger ventral ganglion (See in Dudich, 1975; Bullock and Horridge, 1965 vol 2 p. 1565). Brachiata show more peculiar properties. The ganglia are claimed to be absent (*Dudich, Bullock and Horridge*), intraepidermal plexus is however present. Nerve cells are distributed in the epithelium. In the anteriormost part a brain-like group of cells may occur.

NERVOUS SYSTEM OF INVERTEBRATE CHORDATA. DEUTEROSTOMIA-2.

Concerning cords, two superficial analogies emerge: (a) homology with the cords of worms, Arthropoda etc. or (b) homology to spinal cord of Vertebrata or cords of invertebrate Chordata. A supplementary possibility is the lack of 'true' correspondance of longitudinal structures of these phyla.

The main problem is to find the satisfactory ontogenetic foundation of the homologies of cord and ganglion formations with comparison to other phyla.

The classification followed here regards the Vertebrata, Prochordata, Hemicordata and Cephalocordata as separated phyla. By this convention the name of 'invertebrate chordata' is justified. This is not unique, e.g. Ariens (1960) claims about Amphioxus that 'classed as a chordate and not a true vertebrate..'.

In such a framework the immediate -formal or real - ancestors of Vertebrata *are most probably the Hemi-, Pro- and and Cephalochordata with* 90, 1900 and 13 recent species respectively. The true ancestors are most probably fossilized and the classification of recent items does not reflect the true evolution. This is a frequent and hardly avoidable dilemma of evolutionary arrangements of animals concerning other taxons too. In the chordate phyla numerous candidates of homologies with Vertebrata brain parts could be listed. However, a given correspondence is usually speculative.

It does not exist a general agreement about homologies of spinal cords of Vertebrata and the elongated cords originating from subepithelial nervous tissue e. g. elongated nervous structures of Hemi-, Pro- and Cephalochordata. Only unreliable parallelism is founded among Hemichordata and cords of Cephalocordata or spinal cord of Vertebrata. It is a different problem to accept or to refuse a correspondence between elongated support tissues in these phyla and columna vertebralis of Vertebrata of 'narrow sense'.

Many authors without hesitation (e.g. Ariens, 1960) speaks about the 'spinal cord' of Amphioxus following suggestions of authors of the 19th century (Retzius, Rhode etc).

Nevertheless, the principal question remains to establish more strictly the true homologies. Superficial analogies of elongated parts can be supposed either with more primitive phyla and Vertebrata. In case of Amphioxus the frontal part is regarded as brain or archencephalon and the caudal elongation as spinal cord.

In Tunicata species beyond the ganglion formation the NS is sophisticated, including also a cerebral part.

VERTEBRATE NERVOUS SYSTEM. DEUTEROSTOMIA-3.

(Sarnat and Netsky, 1974; Jerison, 1973; Masterson et al., 1976; Sepp, 1949).

At this point the degree of evolution is based mainly on the arising of newer and newer cerebral parts beyond spinal cord: the size and relative size or significance of Rhombencephalon(REC), Mesencephalon (MEC), Diencephalon(DEC) and finally the Telencephalon (TEC) with neocortex and more archaic basal part increase.

A peculiar theory of existing three brain complex was introduced by MacLean (1973; see also in Sagan, 1977). According to this scheme the Vertebrata brain would roughly consist of a so called archaic R(=Reptilian) complex, the limbic system and neocortex.

The number of neurons increases in different classes or their ratio is also different in different parts of brain. The white and grey substance are separated.

Concerning synaptic organization layers and glomeruli are frequent.

A very special - essentially Mammalian - formation is the corpus callosum and the lateralization (Sperry, 1964).

In Vertebrata the vascular and NS gets into a special relation and the blood-brain barrier develops.

THE HUMAN BRAIN (Past, Presence and Future).

The (pre)hominid animals display well-known paleological features (Passingham, 1982). The majority of the accounted signs are geometrically mesurable properties or the mass of brain. It is an almost general belief that the 'information-processing capacity' is in direct correlation to the size and enlargement of brain. Most probably this is not an absolute rule.

The problem is not separable from the definition of man. Numerous - equally insufficient - solutions have been proposed. The question of emergence of Homo sapiens with its so called intelligent and cognitive behavior is simply unsolved. The problem start with the hard definition of these behavioral properties. We attribute intelligent behavior also to apes (Premack, 1973; van Lavick-Goodall, 1971) and fragments or traces or illusion of intelligent behavioral patterns in lower animals too. Some decades ago this was strongly denied.

Probably, the human brain is not the only existing conscious brain, but this statement is almost undecidable. Nowadays, it is believed that the speech or other sophisticated forms of communication are not restricted to man. *Primates also display speech-like* communication forms and are capable of handling tools, display special 'sociobiological' behavior. Many attempts were made to distinguish these properties from those of the man. Thus for example it is usually refuted to regard any kind of animal social behaviour and human society as homologous behaviours. The animal behavior is often regarded 'mechanical', 'non-intelligent', 'closed', 'instinctive' etc.

What does transform the ancestors into man with respect of nervous system? Self-consciousness and that of the environment, language, social factors, working ability, capacity to preview, religion and a *multitude of* other factors are non-separable entities. Their role in evolution is extensively discussed since ancient times.

The especially evolved learning ability and associative memory are also regarded as *human characteristics. Until now,* the memory research failed in finding the substrate of memory or learning. Remind the quoted statement of Sperry (1964). The same lack of evidence holds for candidate molecular mechanisms of memory and learning too. At the same time it is evident that the ontogenesis of human brain is connected to maturation by learning.

Learning is a part of human ontogenesis and it is not only human. After Edelman a selectionist theory is directly applied to the nervous system. However, nothing speaks against the presence of a kind of ontogenetic (not *phylogenetic..) Lamarckian* (instead of or in parallel with ontogenetic Darwinism) concept. Cells, which divide after having differentiated shows that genetical machinery works in the generation of cells when the acquired cell characteristics are propagated in a cell line. However, it is *harder to* explain the passing on of acquired neural properties if the cell division was broken off. Thus a conventional Lamarckism applied to neuron generations is applicable only at the early ontogenetic periods. Unfortunately any kind of such ideas *are like a red rag to a bull.* Nevertheless, the neural ontogenesis seems to be more complicated and perhaps not explained by traditional (i.e. Darwinian) phylogenetic terms restricted to ontogenesis. E.g. special sorts of metamorphosis may mean the radical metamorphosis of the NS too.

Let us pose finally the question of *future evolution of brain. Here,* a plenty of room is given for free speculations and the answer is dependent on the ambiance of the responding persons. The 'future of the brain' is directly related to the future of the man. Rose (1973) and Sagan (1977) deal with the *problem. Their discussions* demonstrate that this question is partly beyond biology. Moreover, it may fall into moralization or sentimentalism. In this respect the short essay of Lorenz (1973) is offered - among numerous similar ones - as an instructive reading facing reality. *No guarantee exists against* the stop of evolution as a direct consequence of the so called conscious and intelligent brain and human activity. It is not impossible that intelligent beings are in fact self-desorganizing systems (SdOS) without essential external help.

There is a peculiar concept favorized by Kemeny (1972) and others, criticized by Roszak (1986). Kemeny claimed that beyond Man a non-organic evolution of computers could take place which would replace the Homo as a 'top-species'.

I do not share this technophil idea and mostly agree with Roszak, the opponent. I think that in case of a possible decay of man, the computer species will not survive alone. Computers may become ubiquitous and the life *of Mankind becomes* strongly dependent on a very vulnerable technical environment. Thus if it is not organized well, the technical evolution could cause the 'Untergang' of the Homo sapiens without the survival of technical evolution. The idea of such a survival is not scientifically supported. No independent 'Computer *faber' or 'Computer sapiens'* exist. Even if it is not true, is it consolation for anybody to believe in the survival of machines?

CONCLUSIONS: *PHYLOGENETIC FOREST.*

The evolutionary approach in relation with the nervous system is often applies conjectures and not clear-cut evidences. It cannot be claimed in relation with nervous systems after Bates (as Poulton quotes with respect of butterfly wings in Mason, (1928)), that 'the nature writes as on the tablet the story of modification of species'. The story of neural evolution seems at least as obscure as the evolution of other *partial systems of the organisms,* like e.g. biochemical evolution, evolution of energy-houseold, that of the mechanical motions, evolution based on macromolecular nucleic sequences etc.

Demonstration of orderly changes or even something like a 'progress' based on 'neural items' requires the establishment of a theory of neural evolution separated from other evolutionary theories. This is not an easy task. E.g. the monograph of Ariens-Kapers and coworkers (1960; 3 volumes, about 2000 pages) lists the morphological and crude histological data available *for Vertebrata, including Man (and the chordate* Amphioxus). Details on spinal cord, medulla oblongata, ..., neocortex suggest 'a quantitative progress', variations in shapes, changes in the ratios of distinguishable parts). The last means that specific parts become dominant in 'higher' species which are regarded more progressed than others. Such observations lead to the conclusion that certain parts of brain are more significant than others to specific functions. Nevertheless, no detailed correlations were explored between morphological and behavioural findings. A construction or recognition *of an evolutionary order* in the organization of network wiring and its development seems presently only as a serious challenge only with some hints for the solution and it is not a zet solvable.

What can be said ('in general' or 'in details')?

(0) Protozoan animals may have already sensory and motor functions *together with coordination. The three primary function of NS* is present in lack of NS.

(1) The specialization of neural cells emerges. The simplest reflex or responding machinery may consist and/or develop from one two or even three different cells. It seems to be generally accepted that *the site of origin* is ectodermal for sensation and some of the conductions but the effector cells take the origin from other layer. Thus e.g. two ecto- and one mesodermal cell is satisfactory for a complete reflex arc. Some decades ago the 'monoaxonism' *and polydendritism' (Ariens, 1960, vol I., p. 86.)* seemed a rule based on developmental forces, but now it is not regarded exclusive and ubiquitous.

(2) A progressive (multi)centralization of nerve cells takes place.

(3) The first ganglion, the 'cerebral one' appears later. The intrinsic organization of ganglia sometimes shows exact design, sometimes only separation into surface layer and core (neuropile).

(4) The number of neurons increases.

(5) The density of wiring, the arborization of cell processes becomes higher or richer. However, I underline again that now it is still illusory to recognize a so called 'progression' in network wiring.

(6) The absolute or the relative mass (weight) of 'brain' increases.

(7) Intuitively speaking, the complexity of behaviour increases but *the* complexity is not an easily measurable property.

(8) The shape of comparable, homologous parts (e.g. nuclei or ganglia) display rather a variation than a progress. It is theoretically difficult to claim a certain shape to be more evolved than other ones.

(9) It is believed that memory capacity increases, learning patterns (paradigms) becomes more and more sophisticated. However, it remains a problem that neither of these properties are well-quantifiable.

(10) Manipulation with instruments and sophisticated inter-individual transfer of information is regarded as signs of progress. The fragments of such behaviours are present even in Insects.

Thus more than one criterium can be taken into account for the foundation of neural evolution even if it is based solely on neuromorphological factors.

Neurochemistry and humoral control show more obscure evolutionary trends. Almost all of the putative transmiters are archaic, present already in organism regarded traditionally primitive ones (see in Lentz, 1968). Moreover, some of them might participate in the chemical control of ontogenesis itself.

The evolution of behaviour may be described as an enhancement of recognition, integration and coordination capacities. In comparisons of development for behaviour, the 'learning' ability is generally regarded as a 'higher nervous activity'. Now, we could describe the non-living mechanisms of adaptive formations of an automaton. Thus learning is not as enigmatic as it seemed before. The actual details of economic learning in living species still remains hidden. The ability of working or communication - at least in rudimentary or imperfect forms - are recognizable in lower species too. It is claimed usually that these performances are not comparable with externally similar actions of the Man.

Two outstanding properties of NS are difficult to explain: (1) economy, i.e. functioning in complicated way with relatively few cells; (2) tolerance to local vulnerability, i.e. local redundancy. The two achievements seems contradictory to each other.

The final step of evolution is usually based on the emergence of 'intelligence' and this clearly differs from the evolution based on e.g. energetic aspects (see the thermodynamic approach to evolution presented by B. Lukacs in this volume). This underlines that a multiplex kind of evolutionary theory is at least desirable. It follows that to derive 'phylogenetic trees' by single, sometimes obscure criteria is untenable. That is why a forest of phylogenetic trees is required where each tree of this fictitious forest is based on clear criteria, a definite partial ordering of taxons.

The method of ordering of various species is most often based on 'youngest common ancestors' and a sequence of recent species does not mean a direct derivation of one recent species to an other living one. On the contrary most frequently a time interval - the life of the ancestor species - is associated to usually two recent species.

Thus to list more than one apparent evolutionary line by different partial systems looks a suitable strategy. This is a multi-evolutionary approach even if it is restricted to neural substrates and control. It encloses a searching for partial ordering of recent and fossil taxons based on neuroanatomical, physiological, behavioural and other neural criteria. It seems desirable to deal first with taxonomy of the nervous systems separately and after look for correlation with other 'evolutionary lines and trees' if such correlations exists at all... The possible 'neural classifications' and various possible orderings should not be necessarily identical to equivalence relations and orders based on other (morphological, biochemical, ontogenetical etc.) aspects. The unfortunate reality is that very few detailed data are available about fossil NS-s since the soft parts are rarely fossilize well (Sarnat and Netsky, 1974; Harman, 1954). For 'lower animals' even 'endocasts' are not available since solid envolopes for brains do not exist. The research of fossil human 'intelligence' or even IQ of dead men (criticized by Gould (1981) and others) approaches the criteria of pseudoscience in which almost any desired conclusion can be drawn.

Nevertheless, the intelligence and its emergence remains a good criterium of progress with the addition that it is not easy to define 'intelligent' behaviour. This should not be necessarily in correlation with gross anatomical, size, histological, neurochemical or other behavioural criteria.

At last, it is believed that a 'total harmony' of various phylogenetic trends never will arrive. The evolution is supposed to be not expressed by a single tree. It is better to describe it by a metaphora of a fictitious forest.

SUMMARY

After a survey of various nervous systems it looks evident, that without solving basic questions of neurobiology itself, the evolution of nervous systems cannot be described in a satisfactory manner.

Concerning the concept of evolution itself I prefer to be evolutionist believing that the evolutionary speculations are tenable despite of the appearence of Homo sapiens. However, on cosmic scale, the terrestrial evolution which resulted in man is most probably an ephemeric, local, historical, finite and non-generic category.

TABLE OF SOME OUTSTANDING NEURAL CHARACTERISTICS IN VARIOUS TAXONS
(81 ANIMAL CLASSES)

01-01.ZOOMASTIGOPHORA	Flagella control;
02-01.RHIZOPODA	Resting and action potentials; Amoeboidal protoplasma motions, pseudopodia;
03-01.HELIOZOA	In some cases flagellary control
04-01.SPOROZOA	Control of motion;
05-01.RADIOLARIA	Filopodia control; Hydrostatic-muscular control;
06-02.PROTOCILIATA	Ciliary integration;
07-02.EUCILIATA	Control of ciliari movements; Is there a fiber system?
08-02.SUCTORIA	Suction control if exists;
09-03.PLANULADAE	*No data available*
10-04.CALCEREA	Syncitial connections; Primitive neuron like cells; Myogenic motions?
11-04.TRIAXONIA	Is there a nervous system in sponges?
12-04.TETRAXONIA	Is there a nervous system in sponges?
13-04.CORNACUSPONGIAE	Is there a nervous system in sponges?
14-04.CEROSPONGIAE	Is there a nervous system in sponges?
15-05.HYDROZOA	Is there a nervous system in Hydra? Diffuse nervous system;
16-05.SCYPHOZOA	Rhythmicity; pacemakers;
17-05.ANTHOZOA	90-100 cm/sec conduction; Fast and slow conduction;
18-06.TENTACULIFERA	Swimming control
19-06.ATENTACULATA	Reflexes; Luminescence is nervously controlled;
20-07.TURBELLARIA	Orthogon;Nervous tone; Brain removal: intact locomotion but not always; Conditioned responses(Planaria);
21-07.TREMATODES	Orthogon; Unicellular photoreceptors;
22-07.CESTODES	Longitudinal cords, comissures;
23-08.NEMATOIDEA	Ascaris: 162+92 neurons, the best known species;
24-08.NEMATOMORPHA	Some giant cells;
25-08.ACANTOCEPHALA	Hamaniella: 80 cerebral and 30 genital ggl. cells; 1-2 fibers to receptors;
26-09.ANOPLA	Linaeus longissimus: 30 m;
27-09.ENOPLA	Brain of two lobes; Case of 200 ocellum
28-10.ROTATORIA	Epiphanes: 183+34+23 nerve cells,959 cells Asplancha: 323; Synchaeta 281;
29-10.GASTROTRICHA	Large brain, no sign of peripheral plexus;
30-10.KINORHYNCHA	Ribbon-like brain with cell zones;
31-11.KAMPTOZOA	Some ganglia; A peripheral plexus (nerve net?); Special motions;
32-12.PRIAPULOIDEA	Brain, ventral cord, plexus, circular arrangements;
33-12.SIPUNCULIDEA	Brain, ventral cord, plexus, circular arrangements;
34-12.ECHIUROIDEA	Traces of segmentation;
35-13.ARCHANNELIDA	Central brains;
36-13.POLYCHAETA	Giant fiber system in Annelida; Reciprocal inhibition and excitation; Orthogonal nervous system; Complex brain with centers;
37-13.OLIGOCHAETA	Pheretima: 1000 or 6241 cells; Segmental septa in giant fibers; Sensory-motor connection; Reciprocal excitation and inhibition;
38-13.MYZOSTOMOIDEA	Pharyngeal ring;
39-13.HIRUDINOIDEA	Cords, segments, clusters
40-14.SOLENOGASTRES	Ladder-like, orthogonal system;
41-14.LORICATA	Subradular organ; Orthogonal arrangement;
42-14.MONOPLACOPHORA	Two cerebral ganglion;
43-14.SCAPHOPODA	Typical: ggl, comissures, connectives;
44-14.GASTROPODA	Giant neurons and spontaneous activity; Segmentation is not explicit;
45-14.BIVALVIA	Periodic and rhythmic motions; No segmentation; Larval and adult nervous system;
46-14.CEPHALOPODA	Giant fibers and synapses; Ability to learn;
47-15.PHORONOIDEA	Entirely epithelial nervous system; Giant fiber system;
48-15.BRYOZOA	Nerve nets between or within individuals; Hollow cerebral ganglion;
49-15.BRACHIOPODA	Only(?) statocyst as sense organ;
50-16.ONYCHOPHORA	Lobes, brain nerves, ladder, cords, segments;
51-16.TARDIGRADA	Row of paired ganglia; 20-50 cells/ganglion;
52-16.LINGUATULOIDEA	Both cord or a single mass of neurons occurs;
53-17.CRUSTACEA	Various statocysts, eyes;
54-17.DIPLOPODA	Locomotion patterns;
55-17.PAUROPODA	Parts of cerebrum (deuto,trito-,proto-);
56-17.CHILOPODA	Strategic position between insects and myriopods; Fused ganglia;
57-17.SYMPHILA	Locomotor coordination by any pieces of body;
58-17.INSECTA	Dance of honey-bees; Feromones; Fly control; 19584 sense organs on antenna of Apis;
59-17.MEROSTOMATA	Limulus eyes, Hartline on the lateral eyes;
60-17.ARACHNOIDEA	Segmentation; Various eyes; Cobweb making control;
61-17.PANTOPODA	4 ventral and 1 supraesophageal ganglion;
62-18.CHAETOGNATA	Rapid movements, special sensory equipment;
63-19.POGONOPHORA	Intraepidermal nervous system;
64-20.CRINOIDEA	Basiepithelial plexus; Swimming unit;
65-20.HOLOTHURGIDEA	Radial cord; Various sensory endings;
66-20.ECHINOIDEA	Peripherally regulated system of motion;
67-20.ASTEROIDEA	Integration and individuality of arms; Light sensitivity;
68-20.OPHIUROIDEA	Local reflexes, central control; Nerve ring;
69-21.ENTEROPNEUSTA	Intraepidermal plexus with local thickenings; Two intraepidermal cords;
70-21.PTEROBRANCHIA	Cords, rings, ramifications, epidermal parts;
71-22.COPELATA	Cerebral ggl. of 3 parts;
72-22.TETHYOIDEA	Cerebral ggl, brain nerves, neurosecration;
73-22.THALIACEA	Large photic sense organs;
74-23.LEPTOCARDII	Spinal cord (?);
75-24.CYCLOSTOMATA	Neurosecretory organ;
76-24.CHONDRYCHTYES	Mesencephalon; Cerebellum;
77-24.PISCES	Diencephalon; Electrical lobes; Mauthner-cell; Telencephalon with olfactory function;
78-24.AMPHIBIA	Tectum;
79-24.REPTILIA	Neopallium; R-complex;
80-24.AVES	Neostriatum;
81-24.MAMMALIA	Cerebral neocortex; Intelligence;

Main secondary sources: Bullock and Horridge, 1965; Kostojanc, 1957; Various zoological handbooks; Sepp, 1949; Dudich, 1975.

The approximate number of phyla, classes, ordo, families or genera are 24, 81, 400, 5000, 50000(?), 10000000(? Herbert Spencer). No detailed, 'complete' comparative and systematic data of nervous systems and their performences are available. This would be urgent before representative species become fossil, dying out.

REFERENCES

Ariens Kapers, C. U., Huber, C. and Crosby, E. C. (1960): The Comparative Anatomy of the Nervous System of Vertebrates, Including Man. 3 vols. Hafner Publ. Co., New York.

Bullock, T. H. and Horridge, G. A. (1965): Structure and Function in the Nervous Systems of Invertebrates. 2 vols. Freeman, San Francisco.

Edelman, G. M. (1987): Neural Darwinism. Basic Books, New York.

Discussion on: Is there a Nervous System in Hydra? In The Biology of Hydra. University of Miami Press. pp 69-76.

Dudich, E. (1975): Allatrendszertan (Classification of Animals). Tankonyvkiado, 8p. 3rd edition.

van Lawick-Goodall, J. (1971): In the Shadow of Man. Collins, London.

Gould, S. J. (1981): The mismeasure of man. Norton, New York.

Harman, P. J.: Paleoneurologic, Neoneurologic and Ontogenetic Aspects of Brain Phylogeny. The American Museum of Natural History. New York. 1957.

Hertwig, O. and Hertwig R. (1878): Das Nervensystem und die Sinnesorgane der Medusean . Vogel. Leipzig. 186 pp.

Jerison, H. J. (1973): Evolution of the Brain and Intelligence. Academic Press, New York.

Katz, M. J. and Lasek, R. (1978): Evolution of the nervous system: Role of ontogenetic mechanisms in the evolution of matching populations. Proc. Natl. Acad. Sci. USA 75: 1349-1352.

Katz, M. J. Lasek, R. J. and Kaiserman-Abramof, I. R. (1981): Ontophyletics of the nervous system: Eyeless mutants illustrate how ontogenetic buffer mechanisms channel evolution. Proc. Natl. Acad. Sci. USA 78:397-401.

Kenyon, C. J. (1983): Pattern, symmetry and surprises in the development of Caenorhabditis elegans. Trends in Biochemical Sciences, 8(10): 349-351.

Kemeny, J. (1972): Man and the Computer. Scribers, New York.

Kleinberg, N. (1872): Hydra. Eine anatomisch-emtwicklungsgeschichtliche Untersuchungen. Leipzig. 90 pp.

Kostojanc, H. S.: Basic Comparative Physiology. Vol II. SSSR Academy Publ. 1957.(In Hungarian 1961).

Lagutenko, Ju., P. (1981): Structural Organization of Brain Cord of Annelida. Nauka, Leningrad, (In Russian).

Lenz, T. L. (1968): Primitive Nervous Syystems. Yale UP, New Haven.

Lorenz, K.: (1965): Evolution and Modification of Behavior, Chicago UP.

Lorenz, K. (1973): Die Acht Todsunden der Zivilisierten Menschheit. Piper, Munchen.

Lorenz, K. (1978): Vergleichende Verhaltensforschung. Grundlagen der Ethologie. Springer, Berlin.

Lentz, T. L. (1968): Primitive Nervous System. Yale UP, New Haven.

MaclLean, P. D. (1973): A Triune Concept of the Brain and Behaviour. U. of Toronto Press, Toronto.

Marler, P. and Hamilton, W. J. (1966): Mechanisms of animal behavior. Wiley, New York.

Mason, F. (Ed.; 1928): Creation by Evolution (A Consensus...). Duckworth, London (392 pages, 27 contributors).

Masterson, R. B., Campbell, C. B. G., Bitterman, M. E. and Hotton, N. (1976): Evolution of Brain and Behavior in Vertebrates. Wiley, New York.

NEC Corporation Research and Development. A Bulletin, 1989.

Passingham, R. E. (1982): The Human Primate. Freeman, San Francisco.

Premack, D. (1977): Intelligence in Ape and Man. AP, New York

Rose S.: Concious Brain. Pelican Books. 1976.

Roszak, T. (1986): The Cult of Information. - The Folklore of the Computers and the True Art of Thinnkking. Pantheon, New York.

Sagan, C. (1977): The Dragons of Eden. Ballantine Books, New York.

Sarnat , H. B. and Netsky, M. G. (1974): Evolution of the Nervous System. Oxford U. P., London.

Sepp, E. K. (1949): History of Evolution of Nervous System in Vertebrates. Medgiz., Moscow. (In Russsian).

Spencer, H. (1868,1854): The Development Hypothesis. Reeprinted in Essays - Scientific, Political and Speculative, I. (cit. by. Gregory in Mason (1928)).

Sperry, R. W. (1964): Problems Outstanding in the Evolution of Brain Function. The American Museum of Natural History. New York. 1964.

Turing, A. M. (1950): Computing machinery and intelligence. Mind 59:381-408. Reprinted in The World of Mathematics 4:433-460 (1956).

Welker, W. (1976): Brain Evolution in Mammals: A Review of Concepts, Problems, and Methods. In Mason (Masterson et al. Rds.).

THE HUMAN EVOLUTION

KFKI-1990-50/C B. Lukács & al. (eds.): Evolution: from Cosmogenesis to Biogenesis

L. Kordos

Hungarian Geological Instutute
H-1143 Budapest, Népstadion ut 14,Hungary

Received on 15th Oct., 1990

Formation of man, in close connection with the evolution of Primates, has a past of cca. 60 million years. The first primitive monkeys of insectivore relation sundered 40-45 million years ago, both in geographical and in evolutionary sense. At the beginning of the Eocene Africa and Europe were still connected, and also Europe with North America through the North Atlantic. The so called New World Monkeys /Platyrhini/ occupied this continent, splitting afterwards. The continents separated in the Middle Eocene, this group survived in North America, migrated into South America, while extinct out in Africa and Europe. The Holaectic was populated by Old World Monkeys /Catarhini/ /Fleagle and Kay, 1985/. Hence several branches had started as e.g. the ancestral groups of recent Cercopithecoid monkays and apes /Hominoidea/. The separation of these branches happened in North Africa, 30-35 million years ago. From the ape ancestral branch first the evolutionary line of gibbon became deteched, 20-25 million years ago, then that of the orang 14-15 million years ago, and finally that leading to gorilla, chimpanzee and man, 6-7 million years ago. These data of separation have been based on fossils with other animals and plants, on radiometric dating and on calculations in molecular evolution /Pilbeam, 1985; Holmes, 1989; Hasegawa et al., 1989/. The human line separated from apes 5-6 million years ago. Henceforth we can speak of Hominid evolution. The first Hominids can be found in Africa, and very probably there were developed. Their radiation happened at the beginning of the Pleistocene Ice Age, cca. 2 million years ago, when the species of genus Homo spread in Eurasia. Australia and America become occupied by the already developed Homo sapiens, only several ten thousands years ago.

The background of Human Evolution

In the 30 million years of human evolution there were two global events which seriously influenced the way, rote and spatial extension of the evolution, namely /1/ the global plate tectonic events determining the relative positions of Africa and Eurasia, and /2/ the effects of global climatic changes on the environment.
The Tethys Ocean between Africa and Eurasia is continuously shrinking after the Oligocene, the Africa plate is approading Europe. In South Europe and Near East seberal microcontinents are sliding, connected in cogwheel fashion. In the Neogene period, so important in the development of apes and man, it has been possible to finds three periods when Africa was in transient connection with Eurasia via Arabia. The first such period was 18-20 million years ago, when the African proboscideans migrated into Eurasia. In this period no ancient monkey left Africa; they were developing in compact populations in East Africa. Most abundant were the Proconsuls, Rangwapitheci and smaller ancient apes. It is still an open question, whence did the hominid evolution start, possible are some groups of the Proconsul, the recently discovered Afropithecus or Turkanopithecus.

The second connection between Africa and Eurasia had been established cca. 15 million years ago. Then the African early apes migrated into Turkey via Arabia, and thence to the western border of the Mediterranean. Of them the Pliopithecus was very widespread, its most complete skeleton fragments were found in Dévényujfalu /Devinská Nová Ves/ near Pressburg. The Pliopithecus was extremely arboreal, similarly to the recent gibbon, to which it was a collateral ancestor. The other ancient ape group has the common name Dryopithecus. This group courtested of species which lived in closed population and exhibited verY strong sexual dimprhism; their closest kinship is of African origin. Still open, whether the South Asian ancient apes migrated into this wave or later.

The third Africa-Eurasia contact existed 10-11 million years ago. On element of the extensive intercontinental faunal interchange was the rapid Eurasian conquest of the Hipparion of American origin. In the same time a great number of African Hominoids reached the southern fringe of Eurasia via the above mentioned route. On the Siwalik Hills at the base of Himalaya the most frequent was the Sivapithecus, which can be regarded as ancestor of orangutan, and the South Chinese Lufengopithecus is probably also related to this species. It is possible that the Sivapithecus reached also Europe, however here a group of monkeys of smaller stature, classified into Dryopithecus, was more frequent. It is possible that the 10 million year old Rudapithecus hungaricus belonged to this group. Recent investigations seem to indicate the close relationship between Sivapithecus and Dryopithecus.

The other important factor of early hominization was the gradud climatic change strat at the end of Paleogene. Being the Antarctic, as other continental plates as wll, in continuous shifting, in the Oligocene /30-35 million years ago/ the neighbourhoud of the South Pole had become continental. Then started the accummulation of permanent ice on the Antarctic, and with this began inb climatic sense the ice age in which we still live. This ice age has changed the climatic zones and vegetation. For the Pliocene the permanent ice had appeared even in the maritime neighbourhood of the North Pole. The continuous advance of ice broke up the zone of the tropic forests occupying even South Europe at the beginning of Neogene. At the Equator the dominance of jungles was followed by more open savannah. The juicy leaves and fruits were succeeded by plants of arid zones and hard seeds /Andrews, 1981; Shipman andWalker, 1989/. In East Africa the serious environmental changes of global origin coincided with the global tectonic event creating the East African Rift Valley. The volcanic activity increased, some regions had shifted as blocks.

These very serious events of million year scale showed common tendencies and, through several intermediate steps, influenced the evolution og Neogene Hominoids. Many theories see the reason of acceleration of the evolution of early hominization in these changes or in adaptation to them. Anyways, it is a fact that, with generally increasing diversity, the 30 million years of human evolution exhibits 3 peaks in numbers of individuals and species /Vrba, 1985; Hill, 1987/. The first peak occurs between 16 and 18 million years when the great environmental changes started in East Africa. The socond is located at 10 million years, when, after the migration into Eurasia the Hominoids spread on a very large geographic region. Finally, the third peak starts with the Pleistocene when again serious environmental changes began. That Homo migration, starting from Africa and involving the whole Holarctic, already helped immediately the formation of contemporary man.

Skeletal charcteristics of human evolution

In the evolution of apes, and later of Hominids, one can observe several rends, which have different velocities and therefore appear on one species in different extent.

One of the most important regions of changes is the skull.One can start from the first, Oligocene ape, the Aegyptopithecus. Behind the elongated facial part the cranium is also elongated. The orbits look forward. Later for Neogene Hominoids /Proconsul, Afropithecus, Turkanopithecus, Rudapitecus, Ouranopithecus/ the facial part is continuously shortening, while the cranium is broadening . On they few available Hominois skull common features are: the forward-looking orbits, the large interorbital distance, the straight zygomatic bone, the underdeveloped brow ridge of females. From these characterisatics the Sivapithecus seriously deviates. Its face, although reduced, remained elongated at the maxillo-nasal part, and the interorbital distance was very small. Its cranium fragment was very similar to that orangutan. For Neogene Hominoids it is striking that the facial part changed plastically, in various directions, while the cranium evolved slowly in uniform way. For primitive Hominoids the facial reduction was very strong, while the cranium remained flat.

Only the frontal lobes were extending, resulting in on expressed brow ridge. The chimpanzee and gorilla are in this stage, while the Hominid cranium became vaulted at the appearence of Homo, for Homo spaiens even the temporal region has increased. The further increase of frontal bone has gone onto the brow ridge, seemingly eliminating it. The reformation of face is accompained by the evolution of the maxillos-nasal region and the upper dental arch. Recently two Hominoid types have been distinguished in the morphology of subnasal region, an African and an Asian one.

The first group includes the African Neogene Hominoids, the Rudapithecus, the recent gorilla and chimpanzee, and, probably, the Chinese Lufengopithecus. The Asian group contains the Sivapithecus and its possible descendant, the orang.

The morphology of the upper dental arch can be reived from the ancestral type of elongated V-shape. Its evolved variant is the U-shape /Rudapithecus/, a further variant is the backwards convergent type /Turkanopithecus, Afropithecus/. For recent apes the U-shape has become cornered at the canines, and for Hominids it has continuously been changing into a parabolic arc. The evolution of the pattern of the dentition is very slow, and it is manifested on levels above individual species. Its most important factor is the continuous disappearence of the cingulum, and the simplification of the pattern on the chewing surface. But it is remarkable that the dentition of primitive Hominids keeps ancient features compared to the older evolved Hominoids. A characteristic feature of human evolution is the very strong reduction of the size of the canines, the decrease of diasteme, and its final disappearence at Homo. For Hominoids as well as for early Hominids, the upper central incisors are generally substantially larger than the lateral ones. Here the Rudapithecus is an exeption, being the two incisors roughly of similar size.

The increase of brain volume is parallel to the vaulting of the cranium, and for early Homo sapiens it already reached the present value.

Changes in conduct and the hominization

The "classical" change of conduct in the human evolution is bipedalism. Extended investigations of locomotion on fossils show that bipedalism had not yet appeared at Neogene Hominoids, only arboreal motions and quadeipedal ones adapted to arboreal envronment. Unambigous bipedalism can be seen at the earliest Hominids /Australopithecus afarensis/, 3.5-3.8 million years ago. This is indicated by not only fossil bones /"Lucy"/ but footprints in Laetolil /Tanzania/ as well. Bipedalism appeared at least 1 million years before that vaulting of cranium, and no strong causal connection is seen berween them.

Occasional use of natural objects is observed even for recent apes. The same phenomenon can be recognised on Hominoid sites from dislocated throwing stones or manuports. The first worked stone tool appears cca. 3.5 million years ago in East Africa; however numerous and standardized types appear only 2 million years ago and then they develop exponentially. The speed of tools in parallel with the vaulting of the cranium.

The collective hunting, when early men brought down animals larger than thay, is considered a very important biological and social criterion of the appearence of the first Homo. The beginning of this business coincides with the stabilisation of using tools.The first built camp site, in the form of a tent built from branches and hides and reinforced with stones, is known from the same time.

Tools and hunting drastically changed the constituents and quantity of the diet. While early Hominoids had lived predominantly on soft fruits and lives of evergreens, in the Neogene this was succeded by chewing hard seeds. Mixed, animal and vegetable diet is unambiquously seen even for the early Hominids. The diet determined the way of mastication, hence changed the face and the chewing surface of teeth.

The dominance of animal protein in diet must have had an important role in the entranced development of the neural system.

A characteristic feature of human activity is the use of fire. It is possible that some Australopithecines were able to posses or keep fire already 1.7-2 million years ago. But firemaking, as usual activity, can be seen only from the chinese Homo erectus /0.7 million years ago/.

According to the majority of opinions, the most important criterion of the function of Homo sapiens is the appearence os speech. This phenomenon can be deduced from the expansion of the skull at the temples, the conclusing of the network of blood-vessels. It is necessary, however, to note that the appearence of speech, probably several tens of thausands years ago, does not mean the formation of a new species. The Homo sapiens, as a species, is continuous at least for the last 200 000 years. The speech is only one of the last important steps of the evolution of this species.

The causes

In the several million years of history of human evolution one can recognise a lot of synchronous factors, which probably acted in small steps and in interaction. One most, however, confuss that the causal mechanism is not yet known for any important step in hominization. The only thing which we have is some well-documented groups of causes.

The increase of body size and reformation of the chewing surfaces is general for Hominoids, and can be observed for other groups of mammalians as well. What is unique, however, that is the reduction of the facial part of skull without domestication, and the latter appearence of bipedalism.

In the early evolution of apes the most important cause is anyways the global environmental change, which changed the biotic background, the sources of food. The global changes played some role also in the intercontinental migrations. The transient connections and longer separations between Africa and Eurasia were determining factors of phylogeny in the Neogene.

The fundamental factors of human evolutions can be classified into four groups /Potts, 1989/: /1/ locomotion and environment, /2/ social and reproductive factors, /3/ the connection of brain, technology, language and culture, and /4! diet, foraging. In addition, there were developing generally rather slowly, and with different rate, in most cases independently of each other. The process from the earliest ancient ape to Homo sapiens depends less and less on environmental factors, since the enhanced development of the neural system, unique in the animal world, has limited the way of human evolution.

REFERENCES

ANDREWS, P./1981/: Soecies diversity and diet in monkeys and apes during the Miocene.- In.: Stringer, C.B./ed./ Aspects of Human Evolution. Taylor and Francis.pp.25-61.

FLEAGLE, J.G. and R.F. KAY /1985/: The Paleobiology of Catarrhines.- In.: Delson, E./ed./: Ancestrors: The Hard Evidence.,Alan R. Liss.pp.23-36.

HASEGAWA, M., KASHINO, H. and Taka-aki YANO /1989/: Estimation of branching dates among primates by molecular cloks of nuclear DNA which slowed down in Hominoidea.- J. Human Evolution, 18:461-476.

HOLMES, E.C. and G. PESOLE, C. SACCONE /1989/: Stochastic models of molecular evolution and the estimation of phylogeny and rates of nucleotide substitution in the hominoid primates.-J. Human Evolution, 18: 775-794.

HILL, A./1987/: Causes of perceived faunal change in the later Neogene of East Africa.- J. Human Evolution, 16: 583-596.

PILBEAM, D./1985/: Patterns of Hominoid Evolution.-In.: Delson, E./ed./ Ancestors: The Hard Evidence, Alan R.Liss.,pp.23-36.

POTTS, R./1989/: Ecological Context and Explanations of Hominid Evolution.- Ossa.,14:99-112.

SHIPMAN, P. and A. WALKER /1989/: The costs of becoming a predator.- J. Human Evolution, 18: 373-392.

VRBA, E.S./1985/: Ecological and Adaptive Changes Associated With Early Hominid Evolution.- In.: Delson, E./ed./ Ancestors: The Hard Evidence, Alan R. Liss.,pp.63-71.

WALKER, A./1976/: Splitting Times among Hominoids Deduced from the Fossil Record.-In.: Molecular Anthropology, M.Goodman and R. Tashian /eds./, Plenum Press, pp.63-77.

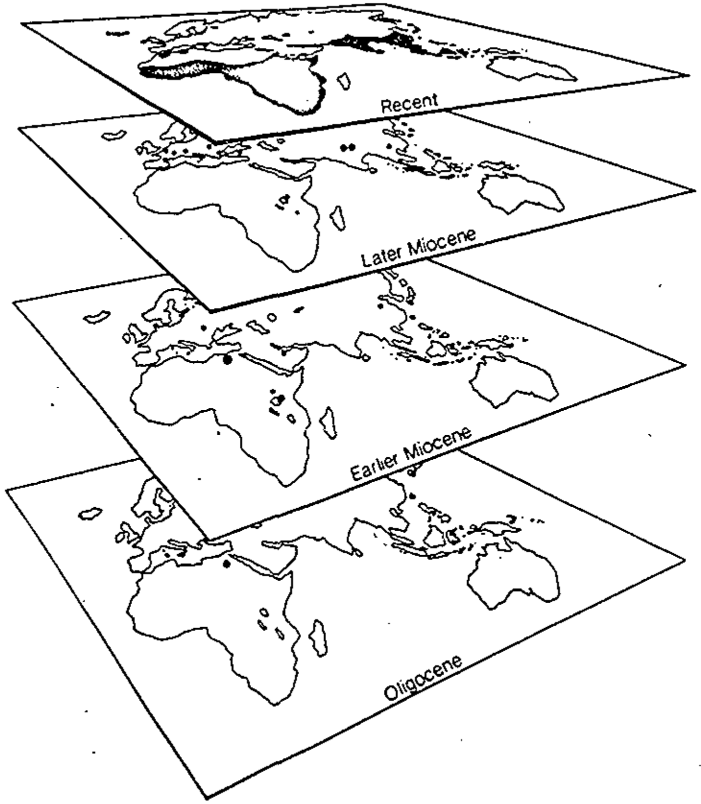


Fig. 1. Known distribution of catarrhine primates during the Oligocene, earlier Miocene, later Miocene and present /After Fleagle and Kay, 1985/.

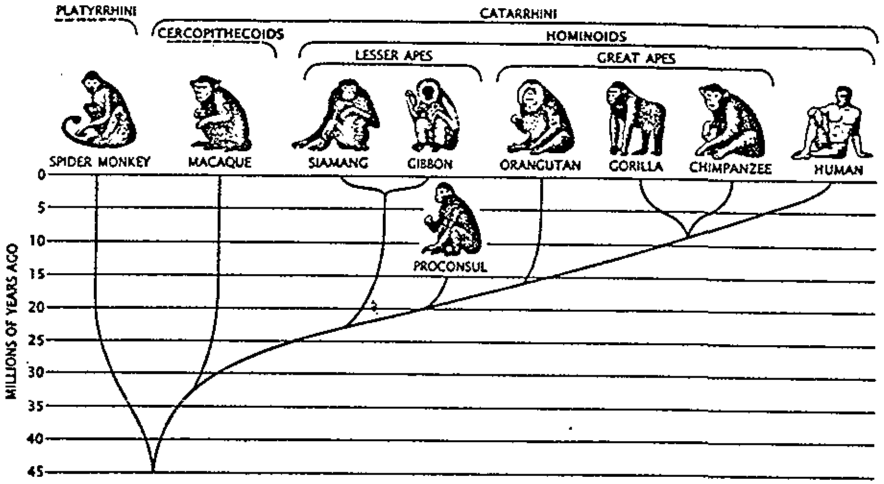


Fig. 2. Generalized phyletic tree of catarrhini primates

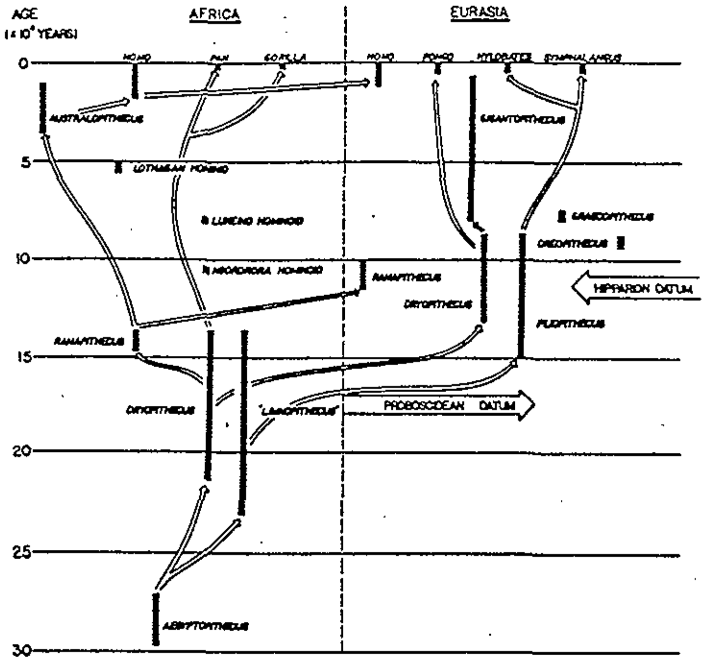


Fig. 3. Diagram to indicate the temporal and geographic ranges of hominoid primate species groups /After Walker, 1976/.

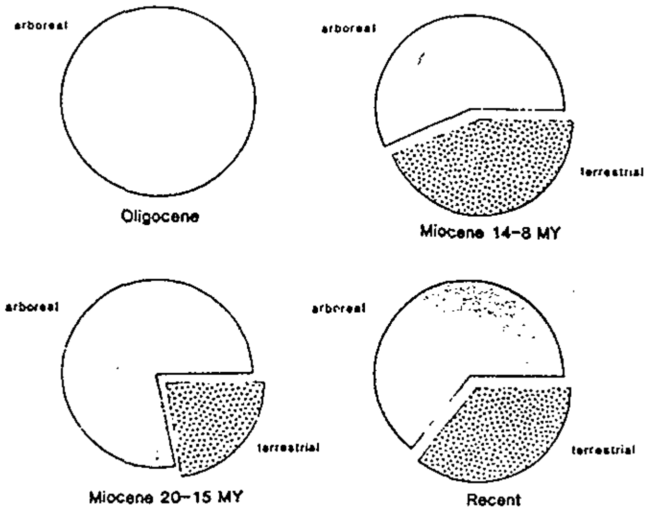


Fig. 4. Substratum preference of catarrhine primates /After Fleagle and Kay, 1985/.

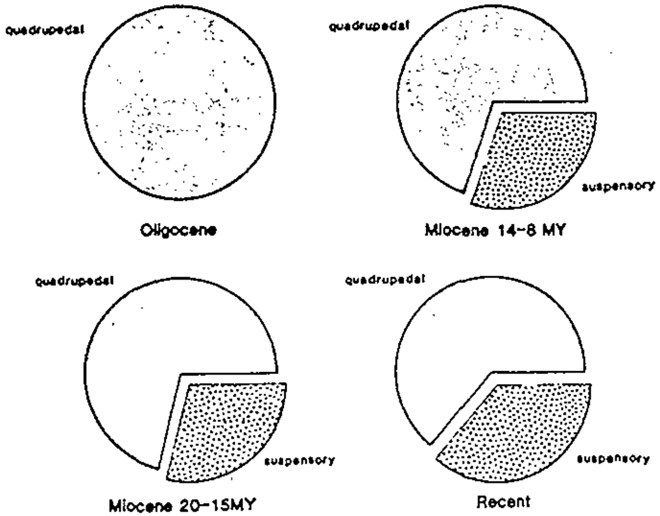


Fig. 5. Locomotor preference of catarrhine primates /After Fleagle and Kay, 1985/.

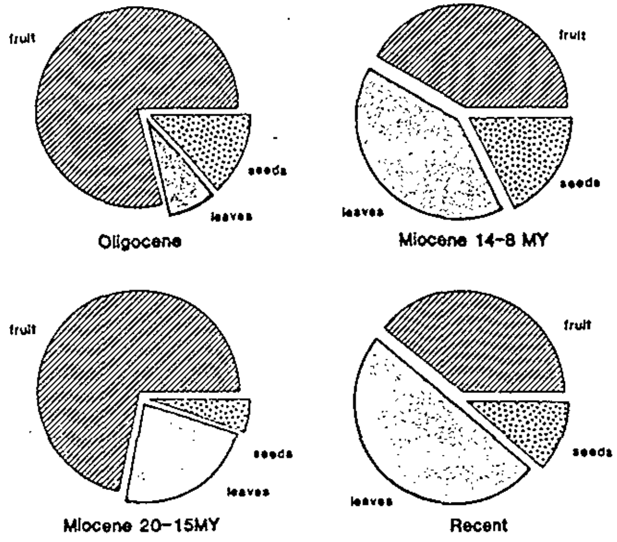


Fig. 6. Diets of catarrhine genera through time /After Fleagle and Kay, 1985/.

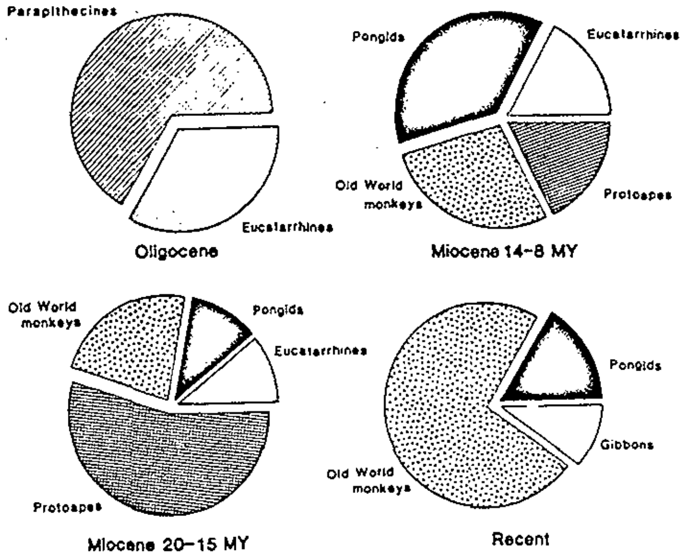


Fig. 7. Percentage distribution of informal catarrhines subunits to the overall pattern of catarrhine diversity at the generic level /After Fleagle and Kay, 1985/.

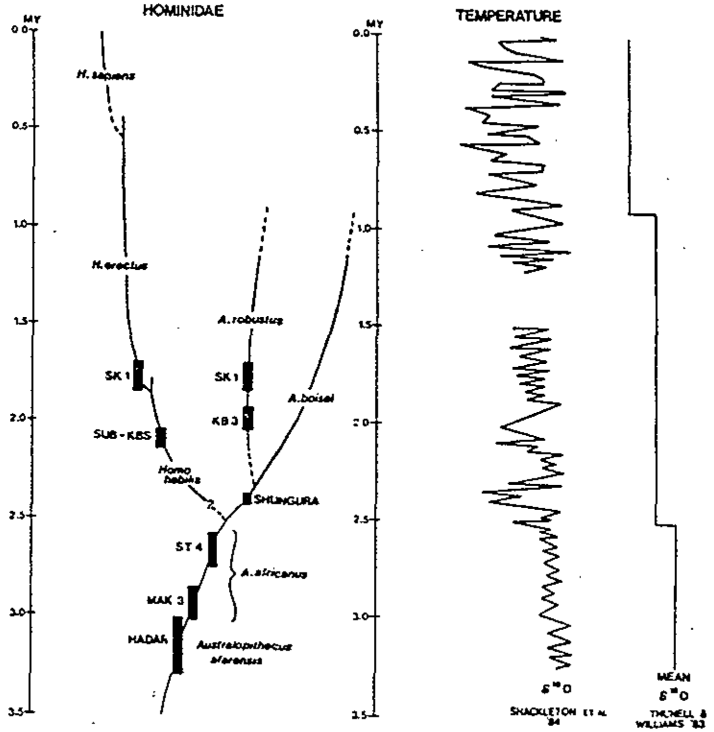


Fig. 8. A comparison, over the time range 3.5 m.y. to recent, of the hominid tree with oxygen isotope data from the deep sea record /After Vrba, 1985/.

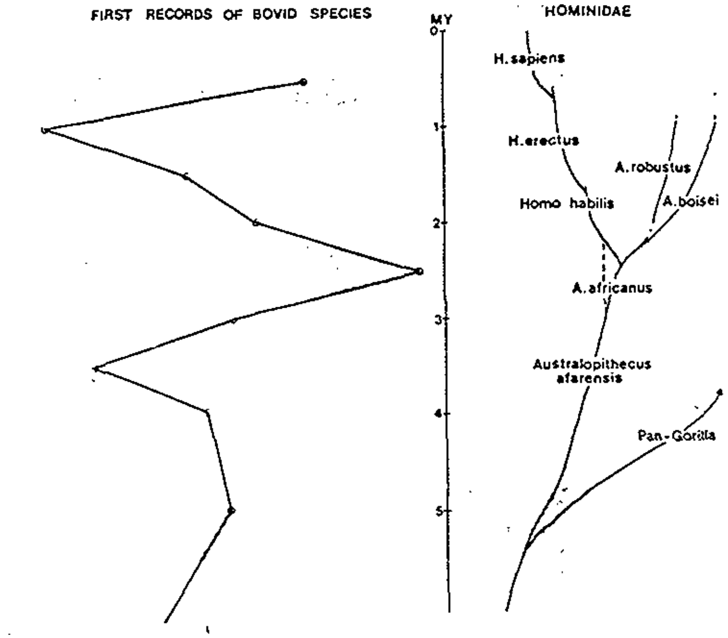


Fig. 9. Rates of origination of all species of Bovidae known from subsaharan Africa, Late Miocene - Recent, and a phylogenetic tree of Hominidae /After Vrba, 1985/.

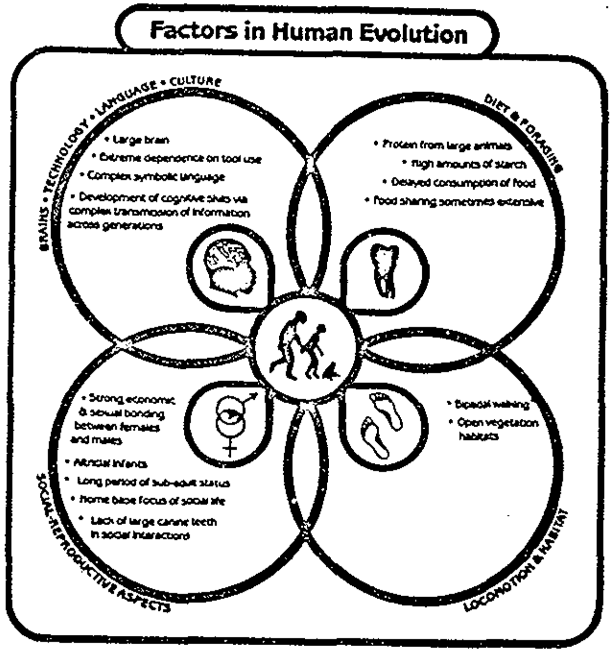


Fig. 10. The distinguishing features of human evolution may be considered as an interaction among four systems /After Potts, 1989/.

HOW TO JUMP INTO HUMANITY: A MATHEMATICAL RECONSTRUCTION

Ágnes Holba & B. Lukács
Central Research Institute for Physics, H-1525 Bp. 114. Pf. 49.
Budapest, Hungary

Received on 5th Oct., 1990

KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

ABSTRACT

Judged from the chromosome sets, one of the factors responsible for the appearance of a separate human line of evolution may have been a Robertson translocation. Here this step is analysed and some constraints are obtained for the parameters valid then and there.

1. INTRODUCTION

The previous paper (Kordos, in this Volume) extensively discussed the hominization process in the light of fossils and observed environmental changes, and the reader has got a coherent picture about an evolution of cca. 40 Mys leading to ourselves. However, about the last big "jump", separating the Hominids from their "ape" relatives (called colloquially Pongidae), fossils are silent: nothing has been found between 8 and 4 Mys. So this event can only be *reconstructed* from observed tendencies and recent data.

At the beginning of the separate hominid branch there must have been at least one very exceptional step as judged from some paradoxical features of the endproducts. For example: the genetic distance between Homo sapiens and the great African apes is in the order of that of close congeneric or even sibling species [1], [2], [3]; in contrast, taxonomically they are classified into separate families. The estimated time of separation changes between the 25 Mys of fossils put into traditional taxonomic context and 4 Mys of distance in immunology [4], [5]. It is hard to decide if the man-chimpanzee-gorilla separation was a trifurcation or two bifurcations [6], [7]. (Two subsequent bifurcations would anyways lead to strange result. If the human lineage had been detached first, then that would be expected to preserve the primitive features, i.e., say, bipedal locomotion would be ancestral to knuckle-walking of African apes. In any of the other two cases man would form a taxonomic unit with one ape against the other one.) And, in spite of the greater similarity in immunology between Homo and African Pongidae than between African and Asian ones [5] [8] all Pongidae have 23 pairs of autosomes while recent hominids have only 22.

The last fact indicates a Robertson translocation (or central fusion), which may be a natural explanation for some other peculiarities as well. E. g. such a translocation creates a fertility barrier between two subpopulations carrying the same genes, so can create two species even at very small genetical or biochemical distances. However, the same fertility barrier impedes the survival of such a mutation. Therefore the fate of a Robertson translocation always depends on a lot of parameters, and can be predicted only if all the parameter values are known. However, now we are in a topsy-turvy situation: we *know* that the fusion has propagated in the population, and try to find out *why* and *how*. By other word, the success of the mutation singles out a domain in the parameter space at 5 Mys ago, and we are looking for the actual domain.

In a fixed model this is a pure mathematical problem, and, indeed, we will try to restrain ourselves from *biologic* discussions as far as possible. However, it is not always possible because one has to select a specific model, with a restricted number of parameters, which involves technical simplifications &c., and this selection needs some biological considerations anyways. If the simplifications were too crude, the calculations could be repeated with less simplifications as well.

2. SOME ELEMENTARY FACTS

There are claims to have found the evidences for a Robertson translocation in hominisation. For a review see Refs. 2 & 7. According to it, human and chimpanzee chromosomes can be brought into correspondence by band technique, with the result that the difference is a number of pericentric inversions, and a central fusion of two chromosomes, originally small and acrocentric, resulting in the large metacentric human Chromosome 2. According to the genetic distances and estimated mutation rates, Refs. 5 & 6 places the separation of human and chimpanzee-gorilla lineages to -5 Mys, so this is the earliest possible date of the fusion. Ref. 7 shows 6 Robertson translocations in the catarrhyne evolution on the branch ending in us, which means roughly one in each 10 Mys. Therefore, while analogous mutations do happen in each generation (cf. the rare *inheritable* kind of Down syndrome), the survival of the mutation is rare indeed.

Since the fusion is absent in any Pongida, it must have happened in the separate human evolution. Ref. 6 estimates a *common* gorilla-chimpanzee branch for further 2 Mys, but the problem is that Ref. 7 sees some chromosome changes shared by man and chimpanzee, but *not* by gorilla, while some ones indeed characterize solely the chimpanzee and gorilla. Therefore it seems that there remained some restricted cross-breeding even after the separation of the human branch, and also the gorilla and chimpanzee subbranches were then already distinguishable. To be cautious, it is enough to state that the situation indicates a complicated genetic history.

3. ROBERTSON-TRANSLOCATION

Consider an animal whose chromosome set contains, besides arbitrary others, two pairs of acrocentric chromosomes (A A, B B). Assume a coincidence of *two* fragmentations: A loses a very small terminal fragment but retains the centromere, while B breaks losing just the centromere and the irrelevant terminal part behind. After *inverse* recombination the mutation changes the pair into

(A AB, B ba)

Here AB contains the full information of A *and* B, containing the centromere of A, while ba is blank, irrelevant, with the centromere of B. Therefore the mutant produces four kinds of haploid gametes, listed below. By pure combinatorics the four probabilities would be equally 1/4, but observations clearly show a moderate preference to the mutant descendents. The *simplest* way to describe the phenomenon is the meiotic scheme of Table 1, and this simplest meiotic scheme will be used in our model.

(A A;B B)	→	(A;B)	$p=1$
(A AB;B ba)	→	(A;B)	$p=\frac{1}{4}(1-k)$
		(A;ba)	$p=\frac{1}{4}(1+k)$
		(AB;B)	$p=\frac{1}{4}(1-k)$
		(AB;ba)	$p=\frac{1}{4}(1+k)$
(AB AB;ba ba)	→	(AB;ba)	$p=1.$

Table 1: The simplified meiotic scheme

Here $0\le k\le 1$ is the preference parameter (of any origin); for some *other* translocations of the recent human set [7] $k\sim 0.2$, and we will use this number in some numerical examples, but do not regard it anything more than an estimation.

4. THE PROPAGATION OF THE FUSION

Now we are going to define the model step by step. Consider a population with 3 different genotypes, their relative concentrations and fitness factors (meaning the ratio of life expectancies in reproductive age), as follows. Wild form: $(n_o,1)$, heterozygote mutant: (n_1,v), homozygote mutant: (n_2,w). Now we *assume* random mating, and mutation only in the ova, with probability ε. Then the concentrations change from generation to generation as

$$\begin{aligned}
n_o &\rightarrow Q(1-\varepsilon)[n_o+\tfrac{1}{4}(1-k)vn_1]^2 \\
n_1 &\rightarrow Q\{(2-\varepsilon)[\tfrac{1}{4}(1+k)vn_1+wn_2][n_o+\tfrac{1}{4}(1-k)vn_1]+\varepsilon[n_o+\tfrac{1}{4}(1-k)vn_1]^2 \\
&\quad +\tfrac{1}{8}(1-k^2)v^2n_1{}^2\} \qquad (4.1) \\
n_2 &\rightarrow Q\{[\tfrac{1}{4}(1+k)vn_1+wn_2]^2+\varepsilon[n_o+\tfrac{1}{4}(1-k)vn_1][\tfrac{1}{4}(1+k)vn_1+wn_2]\} \\
Q^{-1} &\equiv (n_o+\tfrac{1}{2}vn_1+wn_2)^2+\tfrac{1}{8}(1-k^2)v^2n_1{}^2
\end{aligned}$$

Fixing the parameters v, w and ε, the concentrations can be calculated step by step from an initial condition. Now we want to go back, calculating the parameters from the final result, but even then estimations for the suspected ranges are needed. For the analogous hereditary Down syndrome $\varepsilon\sim 10^{-4}-10^{-5}$ which may be so also for now, and we mentioned above a guess for k. However, v and w cannot be read off because they obviously depend on the gene content of the chromosomes; the considered fusion is over and cannot be anymore observed. As for the initial conditions, the obvious ones would be $n_o\simeq 1$, $n_1\simeq 0$, $n_2\simeq 0$ ("before fusion"). However, as we shall see, this state never existed, and such initial conditions *may* be incompatible with the known final result.

To reduce the number of unknown parameters, we *assume* that the change in fitness is proportional to the changes themselves:

$$w=1-2(v-1). \qquad (4.2)$$

With this now let us postpone the values of v and w for a moment.

The structure of eq. (4.1) is quite similar to that of systems of differential equations. In that case it is often observed that in asymptotic time the actual solution becomes independent of the details of the initial condition, but there are different asymptotic solutions for *very* different initial conditions. This is so in the present case too. E. g. for $\varepsilon=0$ Eq. (4.1) has 3 stationary solutions: i) $n_o=1$; ii) $n_2=1$; iii) a third solution obtained if we write the same concentrations on both sides (details later). The obvious expectation is that these final states belong to different domains of initial conditions. Then there is a stationary state reached asymptotically from the "natural" initial conditions. After some transient behaviour this state is well appreached and will not change *if the parameters are constant*. So if a change happened say 5 Mys ago, then there must have happened something either with the parameters, or with the actual state by a surprisingly high fluctuation. The first guess is $v\simeq w\simeq 1$, because the genes are unchanged. Then the asymptotic solution belonging to the initial condition $n_o=1$ is:

$$n_1\simeq\frac{2}{1-k}\varepsilon;\; n_2=\sigma(\varepsilon^2), \qquad (4.3)$$

So for any reasonable k the mutant concentrations remain very low forever, indeed. Then one is tempted to argue for $v\gg 1$, judged from the *present* serious advantage of man to chimpanzee. While it is strange *without* gene mutation, it is by no means impossible; Ref. 2 demonstrates that even now the genetic distance between the two species is very moderate and still the difference is serious. The last paragraph of Ref. 2 even suggests *position effect* as cause for the serious difference, and this does appear via Robertson translocation. However the fitness w of the human ancestor 5 Mys ago must not be identified with that of the present man, with his technology &c. And, by *assuming* high constant v and w values one gets the asymptotic solution $n_2\simeq 1$ in a few generations (details later), i.e. then the wild form practically cannot have existed at all.

5. TWO SCENARIOS

So there seem to remain only two possible scenarios (together with their combinations, of course). One is the *transient* increase of v. While one cannot wait *serious* increase from pure position effect, there can be a very special difference leading to definitely higher v under very special circumstances, which can exist, of course, only for limited time. Then we have to check if this transient increase is enough to produce the final homozygote mutant population by "the survival of the fittest". The other scenario is massive inbreeding (Adam & Eve), sometimes suggested. In it $v=1$, but the mutant must have been isolated from the population together with arbitrary number of the opposite sex. Then in the next generation

$$n_1/n_o\simeq(1+k)/(1-k). \qquad (5.1)$$

In the next step the fertility barrier of the heterozygotes already appears, so this scenario is equivalent with a very high initial n_1 value, with $n_2\simeq 0$.

6. RESULTS AND CONCLUSIONS

Consider the first scenario. Originally for an indefinite long time $v\simeq 1$; so we start from (4.3). At $t=0$ v increases (for a while). Let us see first the condition for the propagation of the mutation. Solving the evolution equations for the beginning ($n_1\ll 1$, $n_2\ll n_1$) one obtains that n_1 remains growing above a level $\sigma(\varepsilon)$ if

$$v>2/(1+k) \qquad (6.1)$$

So without the meiotic preference parameter $v>2$ would be needed, and the neighbourhood of $v=1$ can be reached only if $k\simeq 1$, which could happen only with improbably special meiotic mechanisms. The meiotic preference of Robertson-translocated mutants helps the propagation of the mutation, but in itself cannot do propagate it. Then in most cases the mutation dies indeed out, i.e. under *general* circumstances the mutant concentrations are cca. on the level (4.3). Now we are going to discuss the *special* circumstances compatible with propagation. For a short time v must have exceeded the threshold value (6.1). Assume, therefore, that after N generations v (and w) returns to 1; Fig. 1 is a calculation for v=2. One sees that if the favourable conditions remain for 30 generations, the new species dies out after some further 30 generations; however, raising N from 30 to 31, the new species survives and the old one will die out. The critical N value obviously decreases with increasing v; a numerical calculation leads to Fig. 2.

These results suggest that the sudden change in the fate of the two species happens at such a duration of the favourable conditions when the concentrations of the mutants have been able to reach some critical values. Here we have arrived at the position to decide the possibility of the Adam&Eve scenario as well. Since this problem is important enough, we are going to determine the critical n values as functions of k (for definiteness' sake, at $\varepsilon=0$, which is a good approximation at already substantial mutant concentrations).

Setting v=w=1, below a critical value the mutant concentrations decrease, above they increase. (For a stricter statement see the Appendix.) First we may look for the equilibrium concentrations which can remain unchanged. To get them one has to write the same n values on both sides of the evolution equations, which then give 2 algebraic equations for the two independent concentrations. The calculations show that the equations have only one root pair (if any) for any fixed $0<k<1$ in the possible domain $0<n_0,n_1,n_2<1$. The result is Fig. 3.

Now, this equilibrium may be stable or unstable; this question can be answered by starting a calculation in the neighbourhood of the critical concentrations. This perturbative calculation can be analytically performed (cf. Appendix), with the result that the equilibrium is always unstable (except for domains of zero measure). This result may be interpreted in the following way. If one of the homozygotes dominates the other then the minority is continuously being eaten up. For definiteness' sake, assume that n_2 is dominant. Then the heterozygote subpopulation (1) cannot maintain itself due to the reduced fertility. As for (0): (01) pairs result in (0) only in 50%, while all (02) pairs give (1). Being (2) dominant, (22) pairs are more frequent than (00) ones, so (2) wins. (Therefore after the new population is established, any border encounter of the populations ends without a mixed population; the two disjoint "species" remain disjoint, but with a limited possibility of gene exchange, which decrease the genetic distance, to be remembered.) The critical concentrations, whenceforth (2) is dominant, of course depend on the preference factor k of the meiotic processes, and just this dependence was calculated.

In our case it turns out that (except for domains of 0 measure) for $\varepsilon=0$, v=w=1 the plane of initial conditions (n_1,n_2) is divided into *two* regions: below the dividing line the final fate is $n_0=1$, above it is $n_2=1$.

Now we have explored our model system, so, with the utmost caution, are in the position to draw some conclusions. In the scenario "survival of the fittest" for k=0.2 the final result is Fig. 2; with higher k values the needed N decreases. It is definitely outside the scope of the present paper to guess if the unknown necessary environmental conditions helping the mutant could be present for the calculated time.

In the scenario "Adam&Eve" (or rather Adam or Eve) the conclusion can be obtained by putting the initial condition (5.1) onto Fig. 4. It turns out that for a complete isolation the critical k value is ≃0.35. Incomplete isolation is almost hopeless; with a second person of the same sex n_1/n_0 is roughly halved and the point gets out of the preferred domain except for k's close to 1.

Now, *this* possibility can in principle be checked by observations. If any fusion is seen in the present human population with k>0.35, then the scenario is possible, and the fusion is propagated just when the mutant is isolated and the subpopulation does not die out. For this, an isolated Adam is much better than an isolated Eve, of course. Being n_1 in the order of ε, this isolation is very improbable, but this fact may be conform with the observed rate 1/10 Mys.

If the meiosis cannot produce k>0.35, then the second scenario is ruled out in the present simplest model. More complicated meiotic schemes may help, but it is better not to hope in meiotic processes handmade directly for the propagation.

After the propagation of the mutation a new population has appeared with a chromosome number less by one pair. However this propagation was not necessarily worldwide. Some generations later there is again a possibility for mixing in peripherial encounters. This means a new initial condition $n_1=0$. The result is on Fig. 7 for k=0.2: one of the original constituents of the mixture dies always out in cca. 10 generations. However, during these generations a serious gene transfer is possible, thus keeping the genetic distance low.

ACKNOWLEDGEMENT

One of the authors (B. L.) acknowledges some discussions in the beginning of 70's with L. Végső about numerical simulations of the propagation of a central fusion.

APPENDIX: EVOLUTION IN THE NEIGHBOURHOOD OF A FINAL STATE

Here we give a detailed mathematical analysis of the possible asymptotics of paths governed by eq. (4.1) in the case $\varepsilon=0$, v=w=1.

Eq. (4.1) is *two* equations for *two* variables as can be seen by eliminating $n_1=1-n_0-n_2$. It is worthwhile to introduce two independent new variables x and y:

$$n_0=(x+y)/2;\quad n_1=1-x;\quad n_2=(x-y)/2 \tag{A.1}$$

and then for $\varepsilon=0$, v=w=1 (which condition will not be repeated henceforth) eq. (4.1) reads as

$$\begin{aligned} x_{i+1} &= \{Q(x)[(1+x)^2+(2y-k(1-x))^2]/8\}_i \\ y_{i+1} &= \{Q(x)(1+x)(2y-k(1-x))/4\}_i \end{aligned} \tag{A.2}$$

$$Q(x) = 8[2(1+x)^2+(1-k^2)(1-x)^2]^{-1}$$

Requiring equality of sides one gets the stationary solutions, i.e. the possible final states. The roots are as follows:

$$(x,y)= \quad (1,1); \qquad (1,-1) \qquad (x_0,y_0) \tag{A.3}$$

where

$$\begin{aligned} &\{(3-k^2)x_0^3+(1+k^2)x_0^2+(1+k^2)x_0-(1+k^2)\}\{(3-k^2)x_0+(1+k^2)\}^2- \\ &-8k^2(1+x_0)[(3-k^2)x_0^2+2k^2x_0+(1-k^2)] = 0 \end{aligned} \tag{A.4}$$

$$y_0 = 2k(1+x_0)/[(3-k^2)x_0+(1+k^2)]$$

The equation for x_0 depends only on k^2. For $k^2<1$ it has only one root in the domain $0\le x_0\le 1$, and in addition for $|k|>0.5$ this root gives such a y_0 which leads to the negativity of one of the n_i's. For k>0 the root is displayed on Fig. 3; the k<0 case can be obtained by the exchange of n_0 and n_2.

Now choose a possible final state and assume that the population is already in the neighbourhood. Then

$$x \equiv X + \xi; \; y \equiv Y + \eta$$

where ξ and η are small of first order and we calculate up to first order. Then eq. (A.2) reduces to a system of homogeneous difference equations. Because of the homogeneity, the changes slow down without limit going to the root. Then there one can substitute the difference equations by differential ones as

$$\xi_{i+1} = \xi_i + \dot{\xi}dt \qquad (A.5)$$

and we will measure the time in generations. A system of two homogeneous linear differential equations of constant coefficients has the solution

$$\xi = Ue^{\lambda t}+Ve^{\mu t}; \; \eta = We^{\lambda t}+Ze^{\mu t} \qquad (A.6)$$

where λ and μ are the solutions of a quadratic equation; either both are real, or they are complex conjugates (periodic solution). The differential equations give two relations among U, V, W and Z, and the remaining two can be determined from the initial conditions $\xi(t=0)$, $\eta(t=0)$.

For the first two final states of (A.3) one obtains

$$\lambda = (-1\pm k)/2; \; \mu \doteq -1 \qquad (A.7)$$

Both exponents are negative in both cases, therefore the states approach the final state with a half life time cca. 1 generation.

For the third final state (if exists) one gets

$$\dot{\xi} = (\alpha-1)\xi + \beta\eta$$

$$\dot{\eta} = \gamma\xi + (\delta-1)\eta \qquad (A.8)$$

with

$$\begin{aligned} A\alpha &= -2(3-k^2)x_o^2+4ky_o+2(1-k^2) \\ A\beta &= 4(kx_o+2y_o-k) \\ A\gamma &= -2(3-k^2)x_o y_o+4kx_o+2(1-k^2)y_o \qquad (A.9) \\ A\delta &= 4(x_o+1) \\ A &= (3-k^2)x_o^2+2(1+k^2)x_o+(3-k^2) \end{aligned}$$

Then for the exponents one gets

$$\lambda = [\alpha+\delta-2+\sqrt{(\alpha-\delta)^2+4\beta\gamma}\,]/2$$

$$\mu = [\alpha+\delta-2-\sqrt{(\alpha-\delta)^2+4\beta\gamma}\,]/2 \qquad (A.10)$$

Now, if both exponents are negative, then from anywhere in the neighbourhood the path approaches the equilibrium point; if both are positive, then all paths are repelled, so the equilibrium point is unstable. However, there is a third possibility, when $\lambda>0$ but $\mu<0$. In this case from most initial conditions the paths are repelled, but from the special ones, when U=W=0 they are attracted. These initial conditions lie on the *line*

$$(\mu-\alpha+1)\xi(0) - \beta\eta(0) = 0 \qquad (A.11)$$

which is a domain of 0 measure. Of course, these statements are valid only in the neighbourhood of the state.

So in general the plane of the independent initial conditions is divided into 3 domains: from "below" the paths go into $n_o=1$, from above into $n_2=1$, and from a neighbourhood of (x_o, y_o) they may go into that point. If both λ and μ are positive, this domain is absent, and if only μ is negative, it reduces to a domain of 0 measure (a section of a line). Fig. 4 displays the border of the "upper" domain; the signs of the exponents can be found numerically. In our case for 0<k<1 one exponent is always positive; the other is negative up to k≃0.85. However, from k≃0.5 the equilibrium point (x_o, y_o) is already outside the domain of possible concentrations ($n_i \geq 0$). Therefore the third domain exists whenever the equilibrium is possible at all; however it is always reduced to 0 measure, and can be ignored from practical point of view.

REFERENCES

[1] Elisabeth J. Bruce & F. J. Ayala: Nature 276, 264 (1978)
[2] M. C. King & C. A. Wilson: Science 188, 107 (1975)
[3] M. Blanc: La Recherche 15, 654 (1984)
[4] V. M. Sarich & J. E. Cronin: in Molecular Anthropology (eds. M. Goodman & R. E. Tashian), Plenum Press, N. Y. 1976, p. 141
[5] V. M. Sarich & A. C. Wilson: Science 158, 1200 (1967)
[6] N. Barton & J. S. Jones: Nature 306, 317 (1983)
[7] B. Dutrillaux: La Recherche, 12, 1246 (1981)
[8] M. Goodman: in Phylogeny of the Primates (eds. W. P. Luckett and F. S. Szalay), Plenum Press, N. Y. 1975, p. 219

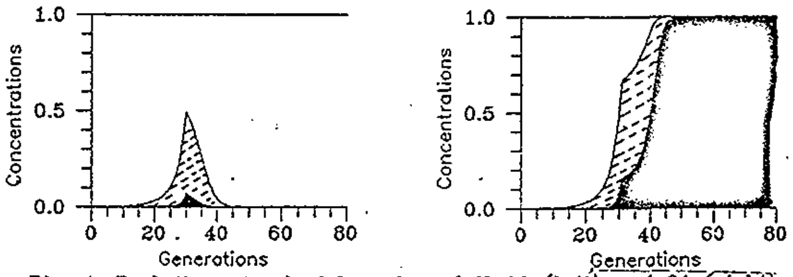


Fig. 1: Evolutions for k=0.2, v=2 and N=30 (left) and 31 (right). White: wild form, dashed: heterozygote, black: homozygote mutant.

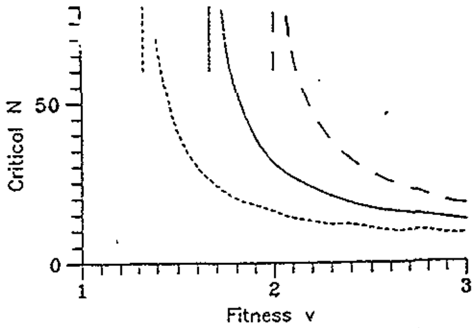


Fig. 2: Critical number of generations N vs. fitness v; k=.0 (dash), .2 (solid), .5 (dots)

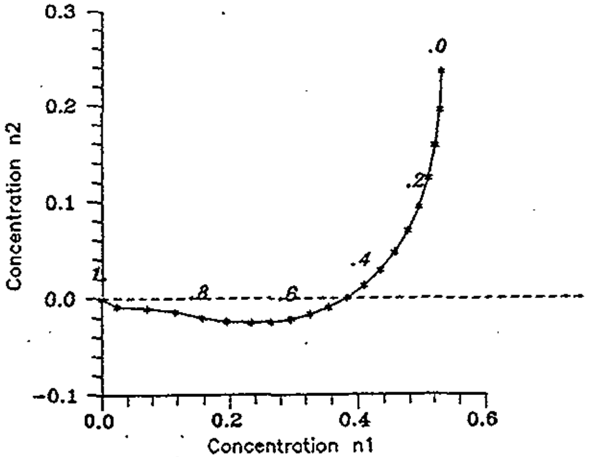


Fig. 3: Equilibrium points on the plane (n_1,n_2). Parameter k is given for each fourth star.

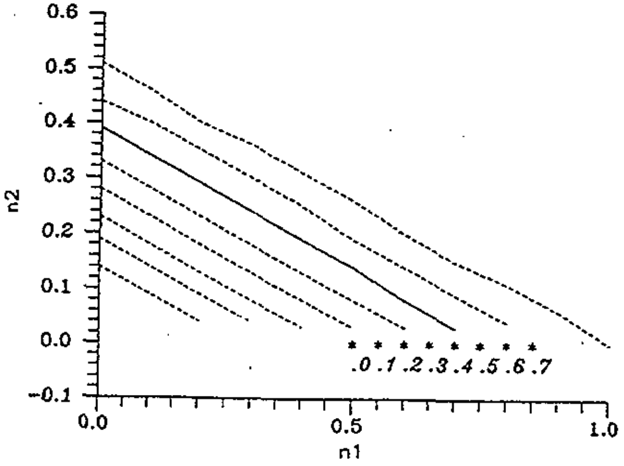


Fig. 4: Critical line on the (n_1,n_2) plane, for k values from 0.0 to 0.7 in 0.1 steps (line k=0.2 is solid). From above the line evolution ends in $n_2=1$. Stars: initial conditions of Adam&Eve scenario: k given.

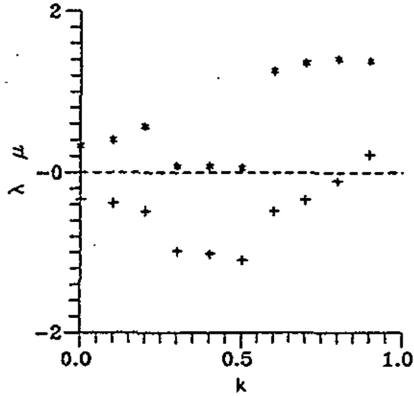


Fig. 5: Exponents λ (*) and μ (+).

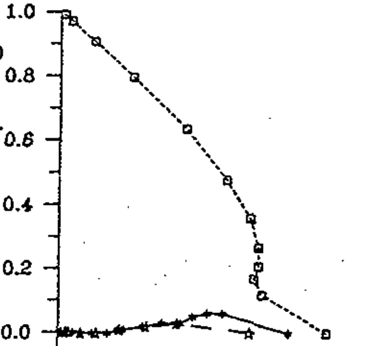


Fig. 6: Adam&Eve paths, steps by generations. k=.0 (☆); =.2 (+); =.4 (□)

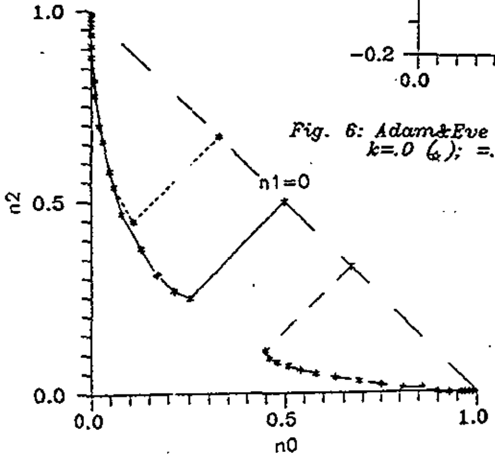


Fig. 7: Evolution from mixtures of two homozygote populations. Step: 1 generation.

ON THE THERMODYNAMICS OF THE EVOLUTION

B. Lukács
Central Research Institute for Physics, H-1525 Bp. 114. Pf. 49,
Budapest, Hungary

Received on 27th June, 1990
KFKI-1990-50/C B. Lukács & al. (eds.):
Evolution: from Cosmogenesis to Biogenesis

ABSTRACT

The problem of Time's Arrow is studied. Our question is what could be the physical quantity to measure the extent of evolution behind a living organism, species or system. Arguments are given that such a quantity may exist; if does, then belongs to thermodynamics and probably is the specific production of a potential related to entropy. Model calculations demonstrate that the global evolution of the whole Earth seems conform to the minimum principle of entropy production. Caution is invoked when trying to use such a parameter for one species.

1. INTRODUCTION

The evolution of terrestrial life is a complicated network of events, historically unique of course as every history already belonging to the past. In principle any minute details might be reconstructed from fossils. Therefore for any pair (A,B) of (recent or extinct) species one could decide which relations hold among the possible 3 ones:

1) A is ancestor of B.
2) B is ancestor of A.
3) Neither of them is ancestor of the other.

In case 1) one may tell, as definition, that B is *more evolved* than A; in case 2) oppositely, while in the remaining case such a statement cannot be done, or at least not on this basis. However, at this point such < > relations are merely historical facts, maybe important, maybe not.

Since the advent of the evolutionary idea there has been a continuous argumentation if the earlier/later or less/more evolved relations are mere historical facts, or reflect something else: a higher value of *something* for the more evolved species. In early times the second opinion was more fashionable. E.g. Haeckel emphasized [1] that as species substitutes species, there is a slow continuous progression in their constitutions. However, while *in average* the *prog*ression indeed wins, there are well documented cases of *re*gression as well. Today several authors regard the "survival of the fittest" as a tautology [2], and indeed, *if* fitness is measured by survival, it is so. Going down to molecular level, the evolution of protein molecules is believed by some authors to be driven by natural selection [3], while others think it to be a predominantly "neutral evolution" [4].

It was an important step forward in clarification Eigen's contribution [5] demonstrating that the evolution goes in the direction of the higher "selective value". His corresponding parameter was a combination of reproduction and decay rates and copy quality of a DNA or RNA "master copy". Indeed, the species of highest net multiplication rate will dominate finally the habitat. However, while such a quantity is a conceivable physical quantity for viruses which are almost pure nucleic acids, it would be hopeless to calculate it from more primitive data for a metazoon. (It could be *measured*, but then we are back again at the tautologic survival of survivors.) So, again: is there a simple physical (or chemical, &c.) quality or quantity which is better or greater in the more evolved species?

For simplicity's sake we will call this hypothetical quantity *evolutedness* to remain at a short and neutral term. Such a quality or quantity may or may not exist. Its existence is clearly not obvious. While lot of (subjective) arguments could be find in favour of humans compared to any older Primates, a consensus is quite possible against any advanced quality in the tapeworm or similar parasites, having descended from more autonomous organisms. However it is quite possible that our anthropocentric viewpoint is an obstacle to recognize the advanced properties of the tapeworm at the top of an evolutionary path completely different from ours. Until the onthologic question is settled, it is better to *suppose* that there exists a quantity monotonously increasing during evolution, only has not yet been found.

Now we try to define the problem in a formal way. Consider the full history of terrestrial life from the beginning $t_0 \approx -4 \cdot 10^9$ ys to the present t=0 y. Then let us make a sequence of snapshots (t=const. sections of the four-dimensional picture) at

$$t_0 < \ldots < t_{i-1} < t_i < t_{i+1} \ldots < 0$$

At any time t_i we can observe a great variety of living organisms $\{\mathbb{L}_\alpha\}$, each with population $N_{\alpha i}$. Now consider the sets $\{\mathbb{L}_{\alpha i}\}$ for a sequence of time labels $_i$, but in a shuffled order; the time period be sufficiently short not to show branching points. By standard palaeonthologic methods we can identify the group of species belonging to the same branches, but we cannot know *a priori*, which one is the ancestor and which is descendant. Now, the question is as follows. Is it possible to restore the true temporal order of species by investigating them in physical, biological etc. details and comparing their data?

Here we do not claim to have the answer. However, we have guesses for the direction in which the answer may be found. This paper intends to list arguments for the *probability* of the following statements.

1) The problem belongs to thermodynamics.
2) If a quantity measures the advance of evolution, it is intimately connected to (specific) entropy production.
3) The present status of thermodynamics is insufficient to *prove* the monotonous change of the possible candidates, but there are serious arguments.
4) Apart from the technical problem mentioned in Point 3), the picture is not too obscure for the *total* biosphere.
5) For individual species even the problem cannot be quantitatively formulated in this moment. However the situation may not be hopeless for small time intervals.

My only goal was to demonstrate that thermodynamics may be relevant in answering the specific question. Therefore what is presented here is not new either from biological or from physical viewpoint; what may be new is the particular context. Even so, I am afraid that, unfortunately, at some points this paper will be rather essayistic. Namely, the approach was triggered by a casual remark of a thermodynamic expert (years ago) from whom it later has turned out impossible to collect further information about thermodynamic details, and who has declined to give this lecture. However, at points obviously obscure (for the lecturer) the reader may turn to thermodynamic experts or to literature.

2. VERY SIMPLE EVOLUTION MODELS

Let us first see (in an admittedly oversimplified but quantitative model) how and why both progressive and regressive changes may happen in evolution. Consider two species, ancestor and descendant. Their population changes by both birth (β) and death (γ):

$$\dot{N}_\alpha = \beta_\alpha N_\alpha - \gamma_\alpha(N_\beta)N_\alpha, \quad \alpha=1,2. \tag{2.1}$$

With growing population there is shortage of some vital resource, so γ grows with N's; the simplest choice is $\gamma \sim N$. If there is no competition (*they do not eat each other's food*), *then*

$$\gamma_\alpha = c_\alpha N_\alpha . \tag{2.2}$$

Then eq. (2.1) can analytically be solved as

$$N_\alpha(t) = N_{\alpha o} f_\alpha(t)[1+(c_\alpha/\beta_\alpha)N_{\alpha o}(f_\alpha(t)-1)]^{-1},$$
$$f_\alpha(t) \equiv \exp\{\beta_\alpha(t-t_{\alpha o})\}. \tag{2.3}$$

I.e. the growth starts exponentially but then saturates at β_α/c_α.

If the second species appears in the saturation stage, then it is better to circumvent the saturation by substituting the scarce food. This is possible if the mutation enables the organism to synthetize the food from precursors. If so, then

$$\gamma_1=c_1N_1; \ \gamma_2=c_2(N_1+N_2), \ c_1<<c_2, \ \beta_1>\beta_2, \tag{2.4}$$

(the last unequality expresses that it is more difficult to reproduce the more complicated species). For $\beta_1=\beta_2$ there is analytic solution; anyways, the descendant does not disturb the ancestor, so Sol. (2.3) remains valid for N_1, and the descendant saturates at $N_{2\infty}=(\beta_2/c_2)[1-(c_2/c_1)]$. Now the new species survives by extra versatility; this is *progression*.

However, far from saturation (without food shortage) this strategy is unsuccessful. If both species are in the exponential regime, then for $\beta_1>\beta_2$ the second one dies out. But if the ancestor was able to synthetize a food, which is *abundant*, then this ability does not appear in γ, so can be lost without loss. Then the new species is simpler (also to reproduce):

$$\gamma_1=\gamma_2=c_1(N_1+N_2); \ \beta_1<\beta_2. \tag{2.5}$$

Introducing $\nu \equiv N_2/(N_1+N_2)$ one gets

$$\dot{\nu} = (\beta 2-\beta 1)\nu(1-\nu), \tag{2.6}$$

This equation can be solved in the form (2.3); the saturation value is $\nu_\infty=1$, i.e. the *more* versatile ancestor species dies out. (Note that this evolutionary process was observed in laboratory, leading to Spiegelman's harmless minimonster phages [6].) This is *regression*.

So the evolution can go indeed in both directions, therefore a quantity measuring the evolutedness cannot be connected with some intricatedness of the organism. Such a quantity must *monotonously* change with time, but it cannot be simply the number of mutation steps, because most point mutations are neutral, so irrelevant [4]. Because of (at least approximate) time-reversal symmetry in most parts of physics, such quantities must be looked for in thermodynamics. The most familiar example is *entropy*.

3. ENTROPY IN THERMODYNAMICS

It is well known that in a *closed* system the entropy S is continuously *growing*. Therefore by comparing states of the closed system in shuffled order, one can uniquely restore the true temporal sequence. This seems to be just what is needed, but wait a moment.

The nonnegativity of entropy production is a consequence of equilibrating nature of macroscopic interactions. Since, according to our present knowledge, this equilibrating nature is quite general (except for gravity, which is not an interaction in general relativity), the growth of entropy is general (except for, possibly, strong self-gravity, where the matter is not yet quite settled, but which situation is irrelevant now).

This monotonous property of the entropy S is a consequence of its probabilistic nature: its growth corresponds to the chance of the occurrence of more and more probable states. In this process the closed system becomes more and more homogeneous. The entropy maximum, as a variational principle, yields the correct equilibrium configurations for thermodynamic systems [7]. In the neighbourhood of the maximum of S (approached asymptotically) gradients of thermodynamic intensives will have become very moderate and processes will have slowed down. *(Heat Death.) In this* sense entropy growth is destructive.

However, Earth is not a closed system. While she fairly keeps her particle content and volume, energy is flowing in and out. Of course, by a heroic effort one can close up the open Sun-Earth system by reflecting walls at 5 billion light years, not yet reached by the solar radiation [8]. But then the total en*tropy of the system (growing of course)* consists of 4 parts: solar entropy (*growing* [9]); terrestrial entropy (in which we are interested); direct and reflected radiation entropy (growing with filling up) and secondary terrestrial radiation entropy (also growing). The increase of the sum does not tell anything about the terrestrial entropy, which *seems* decrease, judging from growing complicatedness of the biosphere. On the other hand, for an *open* system the increase of the entropy cannot be proven, because *it is quite possible to carry away entropy*.

Still, there is a special subclass of open systems, for which some variational principles are known or surmised. This is the case of stationary systems. A variety of model calculations suggests that some *production* is minimal in stable stationary states of an open system, at least as compared to *nearby* states, and hence principles are deduced. We will not formulate this statement in more details here, because the theorems proven up to now still need special conditions.

Near to equilibrium (when conduction coefficients are given at the equilibrium value) and without internal sources the quantity of minimal production is the entropy S, which fact has a transparent enough meaning: without the stationary fluxes S would be maximal with 0 production, so the production is just what must be. For more general cases I shift the responsibility upon the *shoulders of a more authentic expert who says* [10] (in my own translation): "...the [necessary] conditions generally do not hold, therefore the principle of minimal entropy production is not a universal principle in the non-equilibrium thermodynamics. There is no universal principle. Among the elaborated principles...the most general is Gyarmati's Governing Principle...". My guess is that here "there is no" was meant in Cuvier's sense ("l'homme fossile n'existe pas"), i.e. none has been found up to now. *Indeed in the restricted case, when heat conduction is the* only irreversible process, Prigogine and Glansdorff have found the proper potential of minimal production [11]; it is closely related to S and near equilibrium they coincide. In what follows, we, purely for simplicity, will speak and calculate *as if* the potential *were* the entropy. My personal guess is that this is a *relatively* minor simplification compared to *all others* done. Henceforth *this* simplification will not be mentioned; however remember it for not taking *details* in face value. For further details see the standard literature, chiefly the papers of Prigogine. We close this Section with the very simple case when the only flux crossing the system is the *energy* flux J, and no extensive quantity has source inside. The intensive conjugate to energy is 1/T, and the corresponding process is heat conduction

$$J \approx \kappa \, \mathrm{grad}(1/T). \qquad (3.1)$$

Therefore the entropy production σ can be written as

$$\sigma \approx \kappa [\mathrm{grad}(1/T)]^2 = J^2/\kappa. \qquad (3.2)$$

Now, in the stationary case, for a simple geometry, if, in addition, the dependence of κ is negligible on T, the total production is $J_0^2 V/\kappa$, and, according to principles and theorems, this is the minimum. For the transient initial evolution of the system, assume that the *incoming* flux is given and constant $J=J_0$. Now, within the system and behind it the flux may change, and 1/T changes as well, because the heat conduction redistributes the energy. It does just in such a way that σ go to the minimum value, whence one gets the steady state distribution of $J(x,t)$ and $T(x,t)$. The relaxation time τ depends on the particular conduction mechanisms, but the steady state does not.

Now we are going to apply the special minimum principle to the Earth-Sun system.

4. GAIAN EVOLUTION

Consider a steady state Sun (deviations postponed) and an Earth orbiting it at constant distance R_{SE}. Then Earth is the simplest open system: the only incoming flux is that of the energy, and constant. Integrated for the whole surface $4\pi r_E^2$

$$E_+ = W_0 (r_E/2R_{SE})^2 (1-a). \qquad (4.1)$$

Here a is the albedo; for the present Earth $a \approx 0.30$ [12], while for a barren rock Earth, judging from Moon, $a \approx 0.07$. The flux is of thermal distribution (however diluted by expanding propagation), with the data [8]

$$T_+ = 5704 \text{ K}; \; W_0 = 3.43 \times 10^{33} \text{ erg/s}. \qquad (4.2)$$

We now assume (approximate) steady state

$$E_- = E_+ \equiv E_0. \qquad (4.3)$$

For the present Earth observations show

$$T_- = 287 \text{ K}. \qquad (4.4)$$

In a steady state the internal entropy remains constant, thus

$$S_{prod} = S_- - S_+ = E_0((1/T_-) - (1/T_+)). \qquad (4.5)$$

Then, from observed data, we can calculate S_{prod} without knowing *anything* about the particular transfer mechanisms. (The only approximation here is the thermal shape of the outcoming radiation, but it can be checked and corrected from observations.)

Now, this entropy production is primarily not biological; a barren rock Earth would have *higher* S_{prod}. To see this, for an ideal grey Earth

$$T = (\sqrt{r_S/2R_{SE}})T_+ = 275 \text{ K}, \qquad (4.6)$$

(while an equivalent atmosphere *without* greenhouse effect would result in 253 K [13]). Then

$$S_{prod,bar} = 3.64 \times 10^{37} \text{ s}^{-1}; \; S_{prod,bar}/N = 1.02 \times 10^{-14} \text{ s}^{-1}, \qquad (4.7)$$

where N is the total particle number of the system. (The last date suggests a *shortest* possible time scale of substantial changes $\tau \sim 3 \times 10^6$ ys). The above entropy production would come purely from the processes converting shorter wavelength solar radiation into near infrared terrestrial one, which processes are analogous to heat conduction.

Observe the serious decrease of S_{prod} by the present atmosphere (higher reflection+greenhouse effect). So if originally Earth was devoid of atmosphere (or at least of atmospheric greenhouse effect), then the changes leading to present Earth are conform to the trend of minimizing S_{prod}. Of course, palaeoclimatology knows about cooling periods, but neither the solar energy output nor the albedo was constant. (For the initial Sun $W \approx 2.5 \times 10^{33}$ erg/s, $T \approx 5740$ K [8]; there are guesses for cyclic solar energy output with an amplitude ~5 % and cycle time 100-200 Mys [14]; at glacial periods the snow raises the albedo; &c.)

Smaller planets and moons are still barren. But their gravitations are weak to preserve the atmosphere, so for them *particle* flux is present as well, and then the constraints at which minimum is considered are different in their cases. Fig. 1 tries to display $S_{prod}/S_{prod,bar}$ for the planets, however temperature data are still uncertain for some outer planets, as shown by the width of the strip. Anyways, if the mass is sufficient, the entropy production is well below the "barren" value.

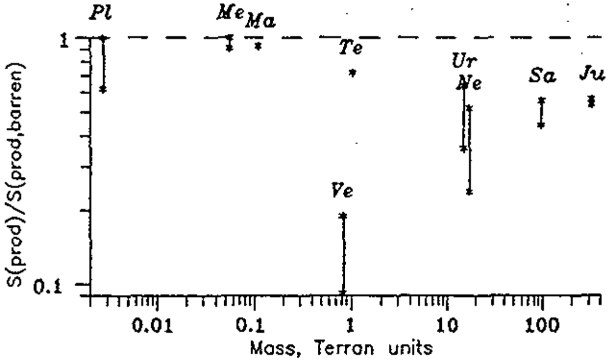


Fig. 1: Entropy production of planets compared to that of a barren rock planet. Some temperatures are still uncertain. Note the decrease from Venus.

Life, with its metabolism, is an excellent agent to change the chemical composition of an atmosphere. So living organisms can *decrease* S_{prod}, in spite of their high specific value, which in itself would show them noxious parasites of Gaia. As a characteristic date, a 75 kg man with 2000 kcal daily consumption produces

$$S_{prod,hum}=2.27\times10^{22}\ s^{-1};\ S_{prod,hum}/N=4.83\times10^{-7}\ s^{-1}, \qquad (4.8)$$

(thus having for substantial corporeal changes a shortest time scale some months, correctly), a specific production 10^7 times higher than that of Earth. So for the whole humanity S_{prod} is already 10^{-5} part of that of Earth, and perhaps the whole biosphere is already a nonnegligible *positive* contribution; however they may have *helped* a 30% decrease, and then the total balance points towards minimum. The spontaneous appearance and propagation of life can thus be conform to the minimum principle. And, similarly, the evolutionary steps can as well be, when the "fitter" or "of greater selective value" replaces its ancestor, as Eigen demonstrated [5] in the simplest case. (In the process when the surviving mutant RNA was replacing the master copy there appeared an integrated loss of total entropy production, even assumed the same affinities for the two RNA chains. The circumstances were such that pure DNA or RNA could replicate.)

5. INDIVIDUAL ORGANISMS AND SPECIES

One is tempted now to apply the principle on various species, saying that the lower specific entropy production belongs to the more evolved organism, and then the discussed problem is solved. Then one finds that woman's S_{prod}/N is smaller than man's one [15]. Not stopping here, it is obvious that S_{prod}/N is lower for cow than for woman, and for (androgyn) tapeworm than for cow. This sequence suggests that something may be misinterpreted.

Indeed, different organisms consume different kinds and quantities of materials (incoming particle fluxes), use different amounts of energy (incoming energy flux) and have different metabolisms (internal sources). So the different entropy productions appear at different constraint values, therefore not indicating at all the degrees of approaching the different minima. From the above mentioned unequalities the only possible conclusion is that woman lives more economically than man, cow than woman and tapeworm than cow, which is probably true. The interactions of interrelated organisms (eating or helping each other) are too strong to separate one for individual study (at the present state of art).

Except if *all* the circumstances remain constant. Of course, such an invariance is practically impossible to be proven, and I do not claim to know such a case. Still, it is possible that in a relatively short time period, when the species do not change their habits too much, ancestors and descendants could be distinguished by comparing the entropy productions. And I would like to close this Section with a demonstrative example, so now I will be deliberately overconfident.

There happened a continuous reptile-mammal transition in the Permian-Triassic period. Judging (maybe incorrectly) from *recent* representatives, this process started from a poikilotherm reptile (in average $T\approx14$ C°) and ended in a homoiotherm mammal ($T\approx37$ C° and fairly constant). The constancy is clearly favorable, it is economic, harmonizing the temperature-dependent enzyme activities, but the high temperature needs some explanation. Of course, good specific explanations exist, but let us look at the problem from the direction of the minimum principle. Compare imaginary very similar reptile and mammal (never existed), differing only in temperature and related properties, *caeteris paribus*. (One may hope, e.g., that the higher temperature does not need higher energy production because of the isolating fur, which is not absurd.) Then a higher T_- leads to lower S_{prod}, as seen in eq. (4.5), so the transition was conform to the minimum principle. Observe that the change happened not only on the mentioned Therapsida-Mammalia line, but on the distant Saurischia-Aves line as well, and roughly simultaneously.

So the development of a higher than atmospheric body temperature may have been a manifestation of the minimum principle, necessarily manifested in some form, anyways. For more detailed statements, I am afraid, specific biology-thermodynamics collaborations would be needed.

6. INTRODUCTION TO PHYSICAL ESCHATOLOGY

Since this is the last lecture of an evolution symposium, it would be proper now to contemplate about the endpoint of the evolution, but this question is, in addition, directly connected to the minimum principles. If evolution goes hand in hand with decreasing production of entropy (or anything else), what will happen when the minimum is reached? Surely, for a very simple case eq. (3.2) has shown the minimum value. Will the evolution stop at this point (or, if reached asymptotically, slow down without limits)?

There is a strong (and intimate) analogy with the so called Heat Death of the Universe, which has much more exhaustively been discussed. Therefore let us see first the analogy, and when we have learnt something, may venture to some guesses in the proper question too.

Consider first a *closed finite* system. In it entropy is growing, therefore changes are no more possible if it has reached its maximum. (If the maximum is being reached asymptotically, then changes are stopping asymptotically as well.) Then, in the last century, by a limiting procedure the Universe became visualized as a *closed infinite* system (i.e. the limit of growing closed ones) with the disturbing result that processes tend to stop asymptotically in complete thermal equilibrium, and Life tends to be impossible in the long run. To be sure, the idea was continuously criticized, and, indeed, the limiting process is not unique. However the discussion was by no means fruitless.

General Relativity made possible the correct formulation of the Heat Death problem. First, a homogeneous Universe is possible in General Relativity, while it is highly unstable in Newtonian gravity [16]. If the Universe is homogeneous on large scale, then infiniteness becomes irrelevant, because its distant parts do not affect each other. In addition, it turned out that the Universe can be closed finite too. On the other hand, it turned out as well that even the closed finite Universe is not a closed *thermodynamic* system, because of the change of the total volume. However even then the global behaviour has become calculable.

In the old cosmologic speculations the Universe generally had a guessed spectacular end, as Flood or Conflagration. Standard cosmologic models (cf. the cosmologic lecture of this Volume) now have a final alternative: Conflagration with infinite density or Frost with infinite boredom, depending on the fact that the Universe is gravitationally bound or unbound. Again, in standard models, the different behaviours are connected with the value of a conserved quantity appearing in the 3 dimensional curvature, i.e. finite Universes are bound and infinite ones are unbound. In the first case the total lifetime of the Universe is finite, and the problem of this Chapter does not exist, so we have to deal with only the second case. The alternative can be, in principle, settled by observing the ratio of expansion and gravity, but the uncertainties are still great, so the only possible answer is that the unbound Universe is more probable, and, if it is bound, the life expectancy is over 50 billion ys.

In the unbound case objects may have infinitely long existence. Will then every change asymptotically stop via Heat Dying? While there is always being a best available answer, we do not necessarily know the true one. Let us demonstrate it by means of a simple Gedankenexperiment. (Numerical data will be estimations only.) Prepare a container of several liters with infinitely strong and rigid outer walls and a realistic partition, and put into one part several moles of H_2 gas and several moles of O_2 into the other (with, if needed, some small amounts of catalyzers); start at room temperature in one half and at freezing point in the other. The initial specific entropy is some s_0. What is the history forward? (In general, a thermodynamic process equilibrates an intensive, which are, in our case, temperature T, pressure p and chemical potentials μ for each particle species.)

Stage 1): pure energy (heat) transfer via collisions with the partition. The endpoint of this process is homogeneous T in the whole container. It takes some time roughly in the order of a second. The specific entropy has increased by some ~0.1. Still p is different on the two sides of the partition, being the densities different too; because there are different particles in the halves, the chemical potentials cannot be equal at all.

Stage 2): the pressure difference is continuously trying to deform the partition. Depending on the strength, it bends in a period between, say, hours and centuries. At the end the volumes are changed so that p equalizes. Again s has gone up by ~0.1. The μ's are still different.

Stage 3): the agressive gases are eating away the partition in a period of a millenium. (Cf. archaeologic evidences.) Then both gases expand into the whole volume, and mix. Now $\mu\{H_2\}$ and $\mu\{O_2\}$ have become homogeneous throughout the whole container, and s has increased by ~1. Henceforth *everything* is homogeneous.

Stage 4): during several further millenia some parts of the gases react, resulting in H_2O. The reaction stops when $2\mu\{H_2\} + \mu\{O_2\} = 2\mu\{H_2O\}$. (Chemical equilibrium.) The temperature is several thousand K. Again s has gone up by ~1. The story ended here for a physicist at the end of last century.

Stage 5): nuclear fusion happens, with a velocity depending on (still unknown) catalyzers of cold fusion. The necessary time can be even above billion ys. The final temperature is several MeV (ten billion K). The theoretical endpoint of fusion is Fe^{56}, but the real endpoint is a mixture of electrons and different nuclei with the hadrochemical equilibrium

$$\mu\{H\}=\mu\{He/4\}=\ldots=\mu\{Fe\}/56=\ldots$$

These equalities determine the nucleus ratios. The specific entropy has gone up by ~1. Here ended the story for a physicist in 1960.

Stage 6): (if Grand Unification is correct, then) in a period ~10^{32} ys protons decay as $p \longleftrightarrow \pi^0+e^+$. Temperature is ~1 GeV. So most nuclei have been destroyed; the final ratios are determined by the above equalities and by the equilibrium condition of the decay $\mu\{H\}=\mu\{p\}=\mu\{e^+\}$. Again s has gone up by ~1. Here ends the story for us, so our opinion is that changes stop here.

However, e^+ *may* be unstable as well. According to our present knowledge even then the available energy is not too much henceforth, but one cannot exclude further processes. Similarly, the entropy went up in each stage and still *may* continue; since we seem to know that the available energy is finite, the final entropy is more probably finite than not, but remember that the *theoretical* final value went considerably up in a century.

And now back to life and entropy *production*. (In the meantime we have learnt that principles may show the direction but not the velocity. So the whole system, just now, may be at various stages of evolution.) In an unbound but cooling Universe each $J \longrightarrow 0$; so the processes are driven less and less. However, the organisms are more and more adapted, and, as we have seen, in deteriorating circumstances more and more versatile. This means new processes, which are able to go even with the less available flux. Of course, one expects the decrease of S_{prod}/N, so the characteristic time scale of substantial changes of the *whole* system will increase. (Cf. eq. (3.2), there the heat conduction coefficient κ represents the transfer mechanisms. On scales where $J \approx$const., κ grows to decrease σ, i.e. the *individual* processes are more and more efficient. Then, with decreasing J σ decreases faster than J^2.) However this is true for a whole gravitationally bound object, e.g. for a planet; we have seen that there is a possibility for the development of swift individual organisms if, e.g. the albedo of the whole planet goes to 1 or if its temperature is going to that of its star, &c. Everything depends on the relative velocities of cooling and evolution, provided that the total mass of the biosphere shrinks sufficiently fast compared to that of the planet. For terrestrial life the solar flux is fairly constant for the next 5 billion years [8] maintaining a convenient environment, but then Sun will destroy the biosphere when expanding into red (sub)giant

7. CONCLUSIONS

I wonder if any definite conlusion can be made. I would have liked to demonstrate that there is a chance to find a physical quantity measuring the advance of evolution of species; I think that the quantity is the specific production of an entropy-like potential. Indeed, we know that if a task is performed in redundant steps, then more entropy is dissipated than necessary. And what else is "fittest" than the one acting most accordingly to necessities and circumstances?

However, in the close future the quantity can be found only for chemically autarch systems (down from a planet to an island?). Even this goal would need resolute cooperation. As for the final fate of Life in Universe, it is not hopeless, but in the best case hard.

And it is better to distinguish between processes favoured by the minimum principle and by *us*. Fig. 1 shows that Aphrodite has decreased much more her entropy production than Gaia, she is quite advanced but, alas, inhabitable.

ACKNOWLEDGEMENT

The author acknowledges that the specific entropy production, as a promising parameter to measure the degree of the advance of the evolution, was mentioned him years ago by the author of Refs. 10 & 15.

REFERENCES

[1] E. Haeckel: Natürliche Schöpfungsgeschichte. 1868
[2] C. H. Waddington: The Nature of Life. Unwin, London, 1961
[3] M. Goodman & al.: Nature 253, 603 (1975)
[4] M. Kimura: Nature 217, 624 (1968); W-H. Li & al.: Nature 292, 237 (1981)
[5] M. Eigen: Naturwissenschaften 58, 465 (1971).
[6] R. Levisohn & S. Spiegelman: Proc. Nat. Acad. Sci. USA. 60, 866 (1968)
[7] H. B. Callen: Thermodynamics. J. Wiley, N.Y., 1960
[8] Eva Novotny: Introduction to Stellar Atmospheres and Interiors. Oxford University Press, N.Y., 1973
[9] This growth is a consequence of the $4H \longrightarrow He+2e$ fusion process. In a small isolated volume it is straightforward to calculate the resulted entropy growth. However, the change affects the gravitation balance, and thermodynamics is poorly defined for self-gravitating systems (because of problems with dividing and unifying them). So the increase is clear but the correct balance is not quite trivial.
[10] K. Martinás: Ph. D. Theses. Magy. Fiz. Foly. XXXIII, 182 (1985) (in Hungarian)
[11] P. Glansdorff & I. Prigogine: Thermodynamic Theory of Structure, Stability and Fluctuations. Wiley Interscience, London, 1971
[12] I. Asimov: Jupiter. Ace Books, New York, 1975
[13] J. Chamberlain & al.: Climate effects of Minor Atmospheric Constituents, in Carbon Dioxide Review (ed. Clarke), Clarendon Press, N.Y. 1982
[14] W. A. Fowler: Nature 238, 24 (1972)
[15] K. Martinás: private communication
[16] I. Newton: Four Letters from Sir Isaac Newton to Doctor Bentley, Containing Some Arguments in Proof of a Deity. London, 1756. (The letter written on 17th Jan., 1693.)

Let us try to understand now why the divergent branching structure of phylogenesis is possible. Simple model calculations would show that the "less successful" mutant is immediately suppressed; on the other hand, if it is "more successful", then it overgrows the original form, which one practically vanishes. (See e. g. Eigen's cited work.) Then one could not expect **two** *successors. However, there are interactions in an ecology, specialisations to different niches, &c. So the structure of the phylogenetic tree again demonstrates that thermodynamic considerations cannot be directly applied to one individual or species.*

However, we do not want to suggest the dominance of divergent branches **everywhere** *in biological evolution. Do not forget that Maynard Smith by geometric ideas on evolution was able to generate network type branches as well.*

E

CONCLUDING REMARKS

This Volume could not be anything else than a starting point for interdisciplinary synthesis, because it must have started with the individual disciplines. At the end of the Symposium, there was already a virtual possibility to recognize new connections between neighbouring or far topics. However, VERBA VOLANT, SCRIPTA MANENT. Now, looking at the Volume, one can ask about the possible connections. Some of them are very probable, and not quite new, at most it may be necessary to reemphasize them. However, there may be a class of intriguing possibilities, when the connection is not established beyond doubt, but is possible, and if it turns out real indeed, can explain two, seemingly independent and unexplained facts.

Let us proceed in a roughly chronologic order. Gravity theory can explain not only the expansion, creating space, but simultaneously also the generation of matter, raw material of further evolution. So, it is possible that the matter content is not an arbitrary initial condition, but determined by the laws. Cosmology and particle physics furnishes the simplest possible particles as building blocks; again expansion ensures the ressurection from premature Heath Death, and guarantees the prerequisites of later astronomical evolution.

The stars, with the existing laws of nuclear physics, necessarily build up just the elements which are indeed abundant. Carbon is abundant not by teleology, for life, and not by an accident, but by being the simplest stable nucleus consisting of integer number of He nuclei, which are both very bound and primordial products of the Universe. (Are laws teleological rather than the carbon?) And then observe that the easily created elements (C, N, O, Ca, Fe, &c.) are just the ones needed for planetological, geological and biological self-organisation. Furthermore, the present terrestrial atmosphere (quite impossible as static state from chemical reasons) is the product of geology and life, not a very lucky exceptional initial state. But again, all this evolution may have been possible only with very special physical laws, or at least very specially "chosen" values of constants in the laws.

Finally, our own emergence from the animal world seems to be driven (or at least promoted) by geological and meteorological events. Then still remains our future, the Arrow of Time in evolution and similar problems which sound slightly philosophical. It is interesting to see that even such problems can (maybe partially) formulated purely by natural sciences, specially by thermodynamics.

So far with the more or less well established connections. However now one can suggest some long-range connections, which may or may not exist, may or may not be important, but deserve mention, and maybe investigation as well. In the astrophysical lecture the solar neutrino problem was mentioned. It seems that the flux of solar neutrinos is less than necessary to produce the observed luminosity by fusion. Now, some hypotheses have been manufactured in which the neutrinos were produced but have later decayed, oscilleted &c. away in exotic (and unobserved) particle physics processes. However, as Fowler suggested in 1972, there is an alternative that the missing neutrinos were never produced because some percents of the energy comes not from fusion but from contraction. Then Sun is not stationary but cyclic. Contraction cannot maintain the energy output beyond the Kelvin-Helmholtz time, ~30 Mys, so then now there must be a slight cooling with this time scale. And such a cooling is seen on terrestrial climate, leading finally to a situation when cyclic ice ages became possible some 2 Mys ago. Maybe the reason is in the astrophysics of Sun, maybe it is terrestrial (but then the neutrino problem needs an independent explanation). It is intriguing to think that the final reason of our present emer-gence may be in the Sun.

But if the Sun has a cyclic behaviour, then earlier periodic ice ages are expected in the far past also; and indeed they happened in the Permian. We seem to know that these ice ages triggered the spread of mammal-like reptiles, whence the way was more or less straightforward to mammals, (so to us).

The extinction of dinosaurs, which seems to have been necessary to the flowering of mammals, has some dozens of insufficient explanations. Two of them is astronomical. One is the collision with an asteroid, when either the environment was poisoned by heavy metals, or the atmosphere was darkened by dust, so decreasing the temperature. The other is a nearby supernova explosion, producing hard radiation for mutations and extra heat for disturbing the thermal balance of reptiles.

We do not propose these explanations; we suggest them only as possibilities, for which the degree of probability cannot be determined without genuine interdisciplinary methods and collaborations. Also we do not tell that they are the only possible long-range connections, but the Gentle Reader may invent the futher ones by utilizing the Volume here.

E

The issues of the KFKI preprint/report series are classified as follows:

A. Particle and Nuclear Physics
B. General Relativity and Gravitation
C. Cosmic Rays and Space Research
D. Fusion and Plasma Physics
E. Solid State Physics
F. Semiconductor and Bubble Memory Physics and Technology
G. Nuclear Reactor Physics and Technology
H. Laboratory, Biomedical and Nuclear Reactor Electronics
I. Mechanical, Precision Mechanical and Nuclear Engineering
J. Analytical and Physical Chemistry
K. Health Physics
L. Vibration Analysis, CAD, CAM
M. Hardware and Software Development, Computer Applications, Programming
N. Computer Design, CAMAC, Computer Controlled Measurements

The complete series or issues discussing one or more of the subjects can be ordered; institutions are kindly requested to contact the KFKI Library, individuals the authors.

Title and classification of the issues published this year:

KFKI-1990-01/C A.J. Somogyi et al.: Particle acceleration in the plasma fields near Comet Halley

KFKI-1990-02/E S. Borbély et al.: Thermal stability of $Ti_{41}Zr_{41}Si_{18}$ metallic glass studied by small angle neutron scattering

KFKI-1990-03/A B. Kämpfer et al.: Description of the nuclear stopping process within anisotropic thermo-hydrodynamics

KFKI-1990-04/A S. Hegyi et al.: A prediction for multiplicity distributions at future hadronic colliders

KFKI-1990-05/N P. Jani: Reconstruction error analysis of sampling interferometers (in Hungarian)

KFKI-1990-06/A S. Hegyi et al.: Does KNO scaling hold at LEP energies?

KFKI-1990-07/A T. Gémesy et al.: Inclusive Ξ^- and $\bar{\Xi}^+$ production in 360 GeV/c pp interactions using the European Hybrid Spectrometer

KFKI-1990-08/G Th. Bandurski et al.: PMK: A programme for the off-line analysis of PMK-NVH experiments

KFKI-1990-09/A V.Sh. Gogokhia et al.: Determination of the pion decay constant and quark condensate in the chiral limit

KFKI-1990-10/E G. Ódor: Investigation of the defected Potts model by the density of states method

KFKI-1990-11/A Ágnes Holba et al.: Is the anomalous Brownian motion seen in emulsions?

KFKI-1990-12/G Baranyai G. et al.: KNI Sealing Detection. (in Hungarian)

KFKI-1990-13/M M. Barbuceanu et al.: Concurrent refinement: A model and shell for hierarchic problem solving

KFKI-1990-14/D U. Wenzel et al.: Messung der Elektronendichte in der Randschicht eines Tokamaks durch Atomstrahlinjektion

KFKI-1990-15/A L. Diósi: A relativistic theory for continuous measurement of qantum fields

KFKI-1990-16/E B. Újfalussy et al.: Electronic states of quasicrystals from the multiple scattering theory

KFKI-1990-17/D J.S. Bakos: High intensity far infrared laser with buffer gases

KFKI-1990-18/G S. Kiss et al.: Studies on vibration of fuel rods. II. Checking reliability of the mechanical model of the fuel rod vibration in a laboratory experiment (in Hungarian)

KFKI-1990-19/J Sz. Vass: Structure of sodium alkyl sulphate micelles. Results from density-neutron scattering and positron annihilation measurements

KFKI-1990-20/A S. Keszthelyi-Lándori et al.: Decay-time measurements on "pure" CsI scintillators prepared by different methods

KFKI-1990-21/A L. Diósi et al.: Statistical distance and the approach to KNO scaling

KFKI-1990-22/E L. Illgen et al.: Influence of impurities on the ductility and magnetic properties of iron-boron-silicon amorphous ribbons (in German)

KFKI-1990-23/C M.I. Verigin et al.: Ions of planetary origin in the martian magnetosphere (PHOBOS-2/TAUS experiment)

KFKI-1990-24/G F. Adorján et al.: EMERIS: An advanced information system for a materials testing reactor

KFKI-1990-25/A B. Lukács et al.: Whither has the nonthermal tail gone in nonrelativistic heavy ion collision?

KFKI-1990-26/C L. Földy et al.: On the rotation of comet P/Halley

KFKI-1990-27/C K. Szegő et al.: On the existence of a mantle region in the magnetosphere of nonmagnetic solar system bodies

KFKI-1990-28/E Dao Khac An.: Application of the thermodynamics of irreversible processes for the simultaneous diffusion of boron and arsenic and point defects in silicon material

KFKI-1990-29/E G. Konczos et al.: Precision density measurement on thin metallic foils with special reference to amorphous ribbons

KFKI-1990-30/A S. Hegyi et al.: The Feynmann fluid analogy in $e^+ e^-$ annihilation

KFKI-1990-31/E,J F.A. Hopf et al.: Modeling the line shape of a hollow cathode lamp

KFKI-1990-32/E I. Kovács et al.: Photoelectric response signal from lightly hydrated purple membrane in external electric field

KFKI-1990-33/A J. Réval et al.: Effect of the nuclear d—t resonance on muon sticking in μ-catalyzed fusion

KFKI-1990-34/J R. Schiller: Excess electron life history by dielectric relaxation

KFKI-1990-35/A N.L. Balazs et al.: Effects of asymmetry in string fragmentation

KFKI-1990-36/M Cs.J. Hegedűs: Generating conjugate directions for arbitrary matrices by matrix equations I.-II.

KFKI-1990-37/A *T. Csörgő et al.: Spacetime effects in reactivistic heavy ion collisions*

KFKI-1990-38/E A. Czitrovszky et al.: Measurement facilities and accuracy limits of sampling digital interferometers

KFKI-1990-39/E A. Czitrovszky et al.: Design and development of a laser airborne particle counter

KFKI-1990-40/D Gy. Bencze et al.: Two potential formalism, a general tool for separating nuclear and molecular dynamics

KFKI-1990-41/G S. Kiss et al.: Studies on vibration of fuel rods. III. Effects of detector size and position on vibration spectra of strings and rods (in Hungarian)

KFKI-1990-42/M P. Tollkuehn *et al.: Analysis of formal approaches in protocol conformance testing*

KFKI-1990-43/M L.Zs. Varga: Artificial Intelligence techniques and tools

KFKI-1990-44/G Tran Dinh Tri: Determination of components of the state vector in AR model

KFKI-1990-45/G Tran Dinh Tri: Determination of coefficient matrices for ARMA model

KFKI-1990-46/A T. Gémesy et al.: On quark diquark systems exited in "high mass" diffractive dissociation in proton proton collision at 360 GeV/c

KFKI-1990-47/A B.Kämpfer et al.: Fluid dynamics of anisotropic media

KFKI-1990-48/E A. Czitrovszky *et al.: Laser interferometers and their application in metrology* (in Hungarian)

KFKI-1990-49/E A. Czitrovszky et al.: Acoustooptical diffraction and its application (in Hungarian)

KFKI-1990-50/C B. Lukács et al.: Evolution: from cosmogenesis to biogenesis

Kiadja a Központi Fizikai Kutató Intézet
Felelős kiadó: Szegő Károly
Szakmai lektor: Vecsernyés Péter
Nyelvi lektor: Csörgő Tamás
Példányszám: 182 Törzsszám: 90-310
Készült a KFKI sokszorosító üzemében
Felelős vezető: Gonda Péter
Budapest, 1990. november hó