ZAGADNIENIA BIOFIZYKI WSPÓŁCZESNEJ

tom 4 1979 s. 119–128

Werner Ebeling, Reinhard Mahnke

KINETICS OF MOLECULAR REPLICATION AND SELECTION*

1. INTRODUCTION

In the last twenty years we learned from molecular biology that the nucleic acid molecules in a living cell and especially the DNA molecules are the stable source of all the information which is necessary in order to guarantee the structure and function of living organism and in particular for the synthesis of proteins. DNA is a double stranded biopolymer consisting of four types of monomers: A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). Since the biopolymers are string-shaped heteropolymers (sequences) with a length of about 10^2 to 10^8 monomers their information content is enormeous. The total number of sequences of a length ν and a basis λ (number of different monomers) is

$$N = \lambda^{\nu} \tag{1}$$

with a length $\nu = 1000$ and $\lambda = 4$ the possibilities of different DNA sequences are about $N = 4^{1000} \approx 10^{600}$. On the other hand, if we consider protein biopolymers of the same length consisting of $\lambda = 20$ different units the number of different proteins is about $N = 20^{1000} \approx 10^{1300}$. Since these numbers are much more greater than the total number of molecules on earth (which is much smaller than 10^{100}) a fundamental question arises: How can coded sequences as carries of structure and information arise spontaneously on earth? What are the differences between a random sequence and an ordered sequence? What kind of measures can be used to measure the "value" of a given macromolecule? Following the hypothesis of Eigen that the evolution of biopolymers is a result of a molecular selection process, first we consider the main ideas of Eigen's theory (Chapter 2). Than we give a short survey about the stochastic description of sequence kinetics by master equations (Chapter 3). In the following paragraph we focused our attention on the problem of complexity showing its increase (Chapter 4). Finnaly in Chapter 5 we introduce a computer model of a selection-mutation process.

2. MAIN IDEAS OF THE THEORY OF EIGEN

In the theory of selection and evolution of biopolymers [8] the main question which Eigen considered is how self-organization of matter can come about. The objects of Eigen's theory are macromolecules (sequences) $i = Z_1 \ldots Z_\nu$, where, Z_k standing on the k-th

*Rozszerzony wykład wygłoszony w Letniej Szkole nt. *Biochemiczne i biofizyczne podejścia w biologii molekularnej*, Jabłonna, 25 IX – 1 X 1977. (*Przyp. red.*)

position in the string is an element of the collection of monomers. The fundamental properties of the molecules are the ability to produce identical copies (replication) and error copies (mutation):

(a) $a + i \rightarrow i + i$ autocatalytic replication process

(b) $a + i \rightarrow i + j$ mutation process

(a – building stone; j – error copy of template i)

In this so-called quasi-linear system there is up to now no internal coupling between the different molecular species. If we consider certain enviromental contraints (e.g. constant overall number of molecules) the kinetic rate equations become nonlinear due to the arising coupling between the species. For the rate equations of such processes we find following Eigen [8]:

$$\dot{x}_i = (A_i - D_i)\, x_i + \sum_{j \neq 1} (A_{ij}\, x_j - A_{ji}\, x_i) - k_o\, x_i, \tag{2}$$

where $x_i(t)$ represents the concentration of molecules of species i (i = 1, 2, ..., S). The first term of the right side of eq.(2) describes the growth and the decay processes, the second – the transition process corresponding to stochastic error copies with the mutation rates A_{ij} and the last term – the dilution flux to control the concentration in order to keep the total number of molecules constant.

$$\sum_{i=1}^{s} x_i = n = \text{const} \tag{3}$$

With the selection constraint (3) eq. (2) can be rearranged to a new set of stochastic differential equations.

$$\dot{x}_i = (E_i - < E >)\, x_i + g_i(t) \tag{4}$$

with

$$\begin{aligned} E_i &= A_i - D_i && \text{rate of reproduction} \\ < E > &= (\sum_i E_i\, x_i)/n \\ g_i(t) &= \sum_j (A_{ij}\, x_j - A_{ji}\, x_i) && \text{mutation rate.} \end{aligned} \tag{5}$$

The solution of eq.(4) has been studied by several authors [8, 14, 20]. In the case of small error terms (A_{ij} being small compared to E_i) the species with the highest reproduction rate becomes dominant. If the mutation rates are of the same order as the reproduction rate several species may coexist [4].

The second step of Eigen's theory is the so-called hypercycle. This model of a cyclic system which combines the complementary introduction with catalytic couplings has been worked out by Eigen and Schuster [8, 9, 19]. It consists of a number of

polynucleotide information carriers I_i like DNA sequences (each of which is able to reproduce itself) and on the other hand each is reproduced with the catalytic help of the protein E_{i-1} whose functional properties are encoded in the precursor nucleic acid I_{i-1}.

Numerical calculations by Schuster [19] and analytical results by Jones [15] have shown that this model demonstrates how collections of different kinds of species may coexist through internal couplings in the presence of constraints. This work is still in progress. Different models for the evolution of biopolymers at all levels of biological organization have been worked out by Goldbeter [10] for regulatory processes at the subcellular level and Babloyantz and Hiernaux [12] for processes during the embryogenesis.

3. STOCHASTIC DESCRIPTION OF SEQUENCE KINETICS BY MASTER EQUATIONS

Now we switch from the description by stochastic differential equations to the description by master equations [6, 7]. The sample space of our problem is the S-dimensional occupation number space [6]

$$\Omega := \{y\} = \{N_1, N_2, \ldots, N_s \mid N_i = 0, 1, 2, \ldots\}. \tag{6}$$

$N_i(t)$ denotes the number of molecules of sequence i. The state of our system at time t is defined as the probability distribution

$$P(y, t) = P(N_1, N_2, \ldots, N_s, t), \tag{7}$$

which time evolution is given by the master equation

$$\frac{\partial}{\partial t} P(y, t) = \sum_{y' \in \Omega} \{W(y \mid y') P(y', t) - W(y' \mid y) P(y, t)\}. \tag{8}$$

According to eq. (3) we keep the total number of molecules constant

$$\sum_{i=1}^{s} N_i = N = \text{const.} \tag{9}$$

Assuming that the transitions are one-step-processes the transition probabilities

$$W(N_1 \ldots N_i + 1 \ldots N_j - 1 \ldots N_s \mid N_1 \ldots N_i \ldots N_j \ldots N_s) = A_{ij} N_j + \frac{1}{N} E_i N_i N_j, \tag{10}$$

describe the properties of the molecules, i.e. the term $A_{ij} N_j$ – the mutation process and the second order term-the replication process with the replication rate E_i. Then the following master equation is derived [6, 7]

$$\frac{\partial}{\partial t} P(N_1 \ldots N_i \ldots N_j \ldots N_s) = \sum_{i \neq j} \left\{ \frac{1}{N} E_i (N_i - 1)(N_j + 1) \right.$$

$$\times P(N_1 \ldots N_i - 1 \ldots N_j + 1 \ldots N_s)$$

$$-\frac{1}{N} E_j N_i N_j P(N_1 \ldots N_i \ldots N_j \ldots N_s)\Big\} \tag{11}$$

$$+ \sum_{i \neq j} \Big\{A_{ij}(N_j + 1)\, P(N_1 \ldots N_i - 1 \ldots N_j + 1 \ldots N_s)\Big\}$$

$$- A_{ji} N_i P(N_1 \ldots N_i \ldots N_j \ldots N_s)\Big\}.$$

In order to find solutions we may use the method of generating functions. If we define the S-dimensional generating function as follows

$$F(U_1, \ldots, U_s, t) = \sum_N P(N_1, \ldots, N_s, t)\, U_1^{N_1} \ldots U_s^{N_s}, \tag{12}$$

we get from eq. (11) a multi-dimensional Fokker-Planck equation

$$\frac{\partial F}{\partial t} = \sum_{i \neq j} K_{ij} \frac{\partial^2 F}{\partial U_i \partial U_j} + \sum_i K_i \frac{\partial F}{\partial U_i}. \tag{13}$$

The drift coefficients $K_i = f(A_{ij}, U_1, \ldots, U_s)$ and the diffusion coefficients $K_{ij} = g(E_i, U_i, U_j)$ are in general nonlinear functions. Many attempts have been made to discuss the properties of eq. (13), see e.g. [11, 13, 18]. But this work is still in progress.

Another aspect coming from coding theory which is very useful for the study of molecular sequences is the concept of metric spaces. The metric distance between two sequences of a length ν is defined by

$$d(i, j) = \sum_{k=1}^{\nu} (1 - \delta_{Z_k^i Z_k^j}). \tag{14}$$

This so-called Hamming distance between the sequences i and j is the number of events required to convert one sequence into the other. Generalizations of the Hamming distance between strings of different length can be found in [7], mathematical contributions to the geometry of Hamming spaces in [1]. But let us consider the four nucleobases as members of a finite ordered alphabet

$$u = (\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}) \tag{15}$$

The set of all possible sequences (words) $W(u)$ is constructed over this alphabet u, and we are able to make a numeration in the following sense

$$\begin{aligned} W(u) &: \Lambda, \mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{AA}, \mathrm{AB}, \mathrm{AC}, \mathrm{AD}, \mathrm{BA}, \ldots \\ g &: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, \ldots \end{aligned} \tag{16}$$

..., BD, ..., DD, AAA, AAB, ...

..., 12, ..., 20, 21, 22, ...

(Λ = empty sequence)

The number $g(i)$ is called Gödel number of the word i. For every i there exist a $g(i)$ and vice versa, e.g. for a natural number n there exist a word $g^{-1}(n)$. To construct the mapping the following expression gives the Gödel number of an arbitrary word $\{Z_{i_1} Z_{i_2} \ldots Z_{i_\nu}\}$,

$$g(\{Z_{i_1} Z_{i_2} \ldots Z_{i_\nu}\}) = \sum_{k=1}^{\nu} i_k \cdot 4^{\nu-k} \tag{17}$$

e.g. $g(\mathrm{A\,C\,D}) = 1 \cdot 4^2 + 3 \cdot 4^1 + 4 \cdot 4^0 = 32.$

4. THE CONCEPT OF COMPLEXITY

After introducing the distance between two sequences $d(i, j)$ (see Chapter 3, eq.(14)) we want to discuss the question what measure can be used to decide if anarbitrary sequence is complex or not. Complexity is one of those concepts which we know well by intuition but find difficult to explain. If we consider for example the following sequences

i = AAAAA CCCCCC

j = ACCDBBCACD

it is a matter of fact, that j is more complex than i. For i we find a short description like five A's to the left of six C's, but for j we do not find a correspondingly short rule. Various complexity measures have been suggested for areas like information theory, computer science, automata theory, mathematics etc. [2, 16, 15, 22]. Some of the most important new concepts go back to K o l m o g o r o v [16]. In general, a system is deemed complex (or complicated) when the interconnection or arrangement is difficult to trace or understand. The complexity of an arbitrarily long randomized sequence is much more greater than a regular or ordered one with high redundancy. In other words, we identify complexity of a sequence with the length of its shortest description. We use the following definition of the complexity $K(i)$ of a sequence i [16, 17]

$$K(i) \quad := \quad l[S(i, L)]$$

and note the inequality

$$K(i) \leqslant l(i) + C$$

L : description language

$S(i, L)$: shortest description of i in L (18)

$l[S(i, L)]$: length of $S(i, L)$

$l(i)$: length of sequence i

C : constant independent of i.

$K(i)$ depends on the description language, a general algorithm for the calculation of $K(i)$ for given i is not known. In terms of Turing machines and partial recursive functions there exist theorems which give upper and lower bounds for $K(i)$ [16]. The concept of complexity is also associated with the problem of information in biology [1]. If the amount and value of information stored in the sequence is increasing the richness of biological functions and properties of the carries of information is also increasing (the link is given by the genetic code); so to say the higher the $K(i)$-value, the richer is the object in contents, i.e. high complexity is associated with high potentiality of behavior [5, 7]. That is why evolving biological macromolecules must have the ability to produce something more complicated than themselves. The sequences are more complicated than the elements which can be made of them and during the evolution process the sequences become more and more complicated because of the increase in their complexity which corresponds to the aperiodicity in the sense of Schrodinger [21]. For this reason we assume that for an evolutionary process (see the selection and mutation process in Chapter 2) the selection value E_i of a given sequence i must increase with the complexity $K(i)$ of i or in other words the reproduction rate

$$E_i = f_i\,(K(i)) \tag{19}$$

is assumed to be an increasing function of the complexity of i. We believe that the complexity is one of the most important quantitative characteristics of biopolymers; there are of course, many other qualitative characteristics [21]. For the mutation rates A_{ij} (see eq.(4),(5)) we do the following assumption

$$A_{ij} = F_{ij}(d(ij)), \tag{20}$$

where $F(x)$ is a decreasing function, that means the mutation probability between the sequences i and j decreases with increasing metric distance between i and j. Point mutations occur with greater probability, simultaneous mutations of two or more monomers are improbable.

5. MODELS AND DISCUSSION

In order to give an example for these rather abstract concepts we discuss special models of the evolution of sequences. We have to choose the two sets of parameters – the reproduction rates, eq.(19), and the mutation rates, eq.(20) – in a special way. Since no algorithm for the Kolmogorov complexity was available we have used in our model another simple measure of complexity given by an arbitrarily chosen algorithm based on a doublet valuation.

The calculation of this complexity of a given sequence was done by a computer program. The algorithm was described in detail elsewhere [3]. We only want to mention that the algorithm works in the following way: The elements on position 1 of the sequence get the value $K = 1, 2, 3$ or 4 if there is A, B, C or D. Then all following doublets are valued after a tabular function. The connection between the complexity and the reproduction rate was given by

$$E_i = \ln K(i). \tag{21}$$

On the other hand our assumption about the mutation rates is very easy. We consider only point mutations and destruction processes

$$A_{ij} = \mu \delta_{1,d(i,j)} + C\nu_j, \tag{22}$$

where the last term describes a destruction rate proportional to the length.

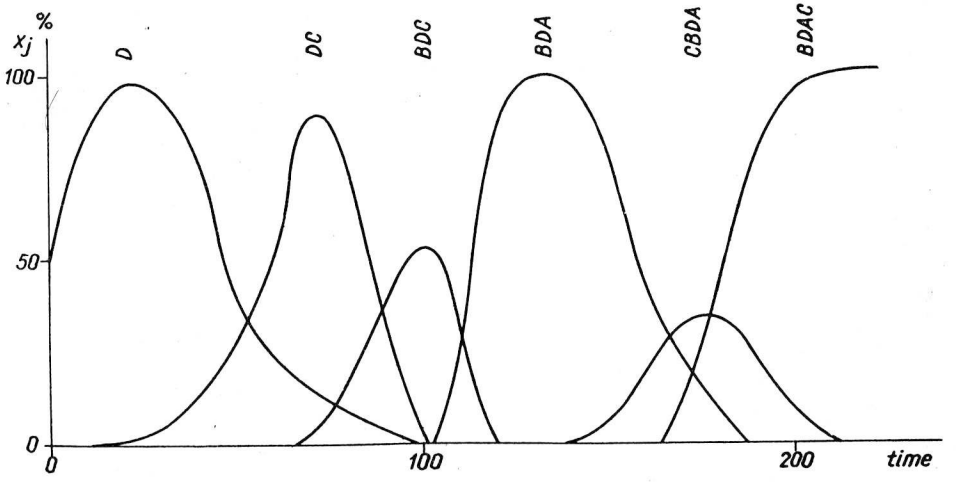

Fig. 1. Model for the stochastic replication of sequences of the units A, B, C and D by integration of the stochastic differential equations

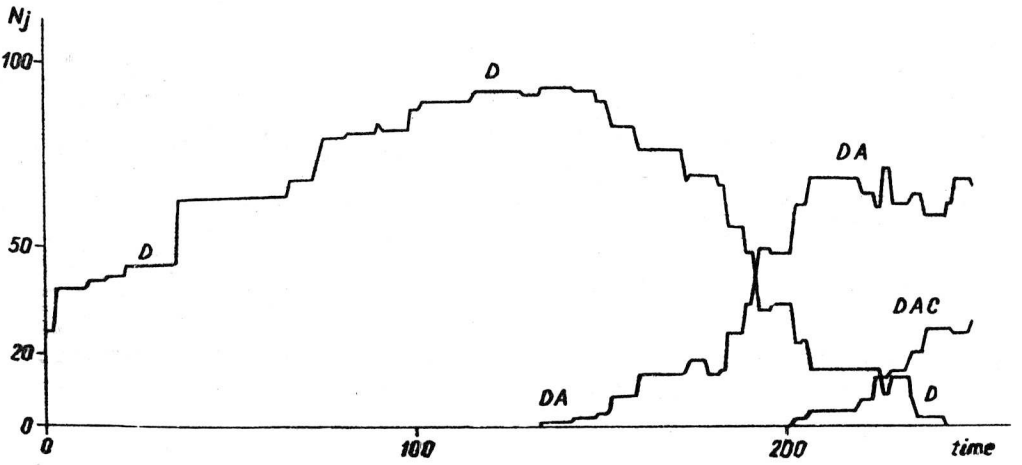

Fig. 2. Stochastic replication of A, B, C, D-sequences described by master equations with a low mutation frequency

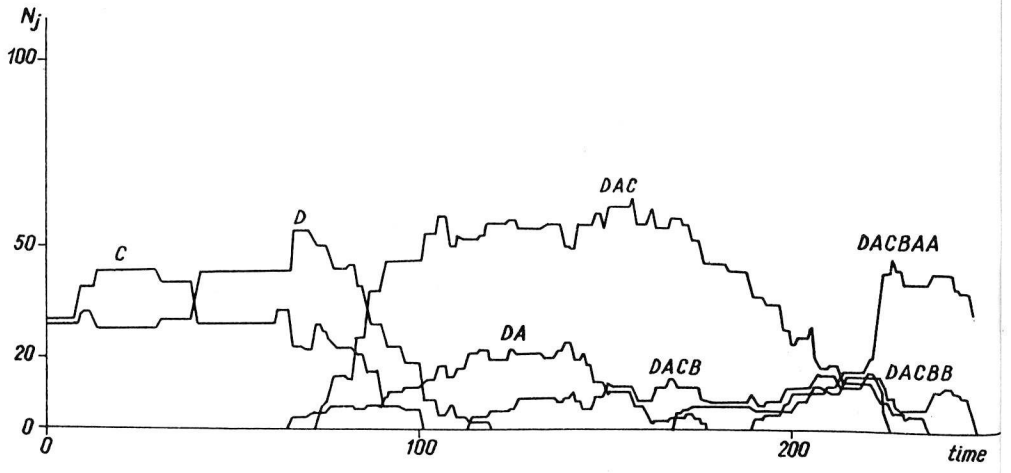

Fig. 3. Another numerical realization with a relative high mutation frequency (master equation model)

Mutations are possible between sequences with the metric distance equal to one with the mutation frequency μ, e.g. one monomer is changed or one monomer is added. They appear at random and discrete times t_1, t_2, t_3, ... Between this time intervals the evolution of the system is given by the deterministic eq. (4) or in a stochastic model by the master eq. (11).

Then we have to solve the master equation numerically. Figure 1 shows one realization with numerical integration of the stochastic differential equation (adapted from [3], Fig. 2 and Fig. 3 show results of the master equation model (adapted from [18]). In both cases the behavior of new species is characterized by the following relation. If the reproduction rate E_n of the new sequence n is greater than the rate E_m of the equilibrated (selected) species m, the new mutand will outgrow the former distribution. On the other hand, if $E_n < E_m$ holds, the new mutant dies out because of the missing selection advantage. Detailed investigations of the behavior of new species can be found in [6]. Another aspect is the influence of the mutation frequency (the number of mutations per unit time) comparing Fig. 2 and Fig. 3. The computer realizations states that there must be a so-called optimal mutation frequency. If the mutation frequency is too small, the evolution progress is lengthy. In the opposite case no species is able to survive because there are too many new mutants. Finally we want to mention a slightly different version of our model [18]. The information stored in the sequence is read in codons. If we consider the allocation

A – Adenine, B – Guanine, C – Uracil, D – Cytosine

the beginning is indicated by the codon ACB, the end by the codon CBA (so-called nonsense codons). Further we use the codon ABB as (artificial) active center, e.g.

DC ACB ADB DCC ABB BDA CBA CBDC.

We assume that the reproduction rate is given by

$$E_i = ln\,[K(i) + a_i] \tag{23}$$

with

$$a_i = \begin{cases} \text{const} > 0 & \text{for sequences with special triplets} \\ 0 & \text{otherwise} \end{cases}$$

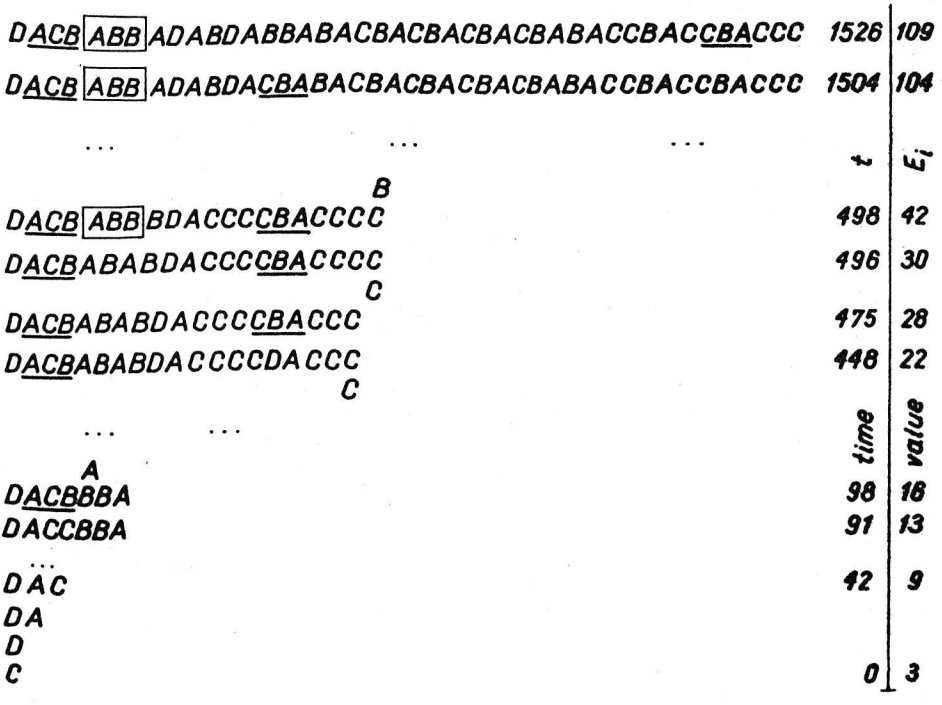

Fig. 4. Selected sequences with increasing reproduction rates during the replication and mutation process showing the appearance of special triple ts
The following (arbitrary) choice for the additional value a_i was used: Sequences containing a start codon ACB get $a_i = 2$. Sequences containing a start codon followed by m triplet codons, between them $n < m$ active centers ABB, and furtheron a termination codon CBA get $a_i = (2 + n + 15m)$

Figure 4 gives the main steps of one realization showing the selected sequences during the evolution. In this process more and more complicated sequences are formed.

The computer game of realizations is practically inexhaustible. Possibly the study of the structure of evolution trees and of the kinetics of evolving systems of this kind discussed above may be helpful for the understanding of at least certain aspects of molecular replication and selection processes.

Maszynopis wpłynął do Redakcji 9 XII 1977;
po rewizji autorskiej otrzymano 22 IV 1978

REFERENCES

[1] Ahlswede R., Katona G. O. H., Discrete Math., 17, 1 (1977).
[2] Bremermann H. J., [w:] *Lecture Notes in Biomathematics*, Vol. 4, Springer-Verlag 1974.
[3] Ebeling W., Feistel R., studia biophysica, 46, 183 (1974).
[4] Ebeling W., Feistel R., Z. phys. Chemie, Leipzig, 257, 705 (1976).
[5] Ebeling W., *Strukturbildung bei Irreversiblen Prozessen*, Teubner Verlagsgesellschaft. Russian translation to appear, Leipzig 1976.
[6] Ebeling W., Feistel R., Ann. Phys. (Leipzig), 34, 81 (1977).
[7] Ebeling W., Feistel R., Jimenez-Montano M., [w:] *Nichtlineare Irreversible Prozesse*, Rostock 1977.
[8] Eigen M., Naturwiss., 58, 465 (1971).
[9] Eigen M., Ber. Bunsenges., 80, 1059 (1976).
[10] Goldbeter A., Lefever R., Biophys. J., 12, 1302 (1972); Goldbeter A.. Nicolis G., Biophysik, 8, 212 (1972).
[11] Haken H., Z. Physik B., 24, 321 (1976); Phys. Lett., 55A, 323 (1976).
[12] Hiernaux J., Babloyantz A., Nonequil J., Thermodynamics, 1, 33 (1976); Babloyantz A., Hiernaux J., Bull. Math. Biol., 37, 637 (1975).
[13] Horsthemke W., Bach A., Z. Physik B, 22, 189 (1975).
[14] Jones B. L., Enns R. H., Rangnekar S. S., Bull. Math. Biol., 38, 15 (1976).
[15] Jones B. L., J. Math. Biology, 4, 187 (1977).
[16] Kolmogorov A. N., Probl. Peredachi Inform., 1, 3 (1965); Zvonkin A. K., Levin L. A., Usp. Mat. Nauk, 25, 85 (1970).
[17] Löfgren L., Int. J. General Systems, 3, 197 (1977).
[18] Mahnke R., *About the theory of replication processes, unpublished results* (MS.-N°. 116 of Sektion Physik der Univ. Rostock).
[19] Schuster P., Chemie in unserer Zeit, 6, 1 (1972); Eigen M., Schuster P., Naturwiss., 64, 541, 1977; Küppers B., Progr. Biophys. Molec. Biol., 30, 1 (1975).
[20] Thompson C. J., McBride J. L., Math. Biosciences, 21, 127 (1974).
[21] Volkenstein M. V., Found. Phys., 7, 97 (1977); Zh. obshsei biologii, 37, 483 (1976).
[22] Von Neumann J., *Theory of Self-Reproducing Automata*. Ed. Burks A., University of Illinois Press, Urbana 1966.

Werner Ebeling, Reinhard Mahnke

KINETYKA MOLEKULARNEJ REPLIKACJI I SELEKCJI

Streszczenie

W artykule przedstawiono ogólne idee Eigena, dotyczące fizycznych podstaw powstawania na Ziemi sekwencji kodowych w procesie replikacji i selekcji makrocząsteczek oraz dyskusję szeregu nowych koncepcji. Ewolucję pewnego układu sekwencji (heteropolimery) w przestrzeni sekwencyjnej przedstawiono za pomocą równań typu „master".

Wprowadzone zostało pojęcie złożoności jako miara wartości danej sekwencji. Omówiono szereg przykładów ewolucji sekwencji.

Prof. WERNER EBELING, Dr. REINHARD MAHNKE
Sektion Physik, Wilhelm-Pieck-Universität
25 Rostock
GDR