Information amount threshold in self-replicating RNA-protospecies: branching processes approach

Krzysztof A. Cyran

Abstract**—**The paper addresses the problem of information content threshold in the early stage of RNA-World. This terms refers to the hypothetical stage of the evolution of Life which assumes that before emergence of organisms whose genome was based on DNA molecules and enzymatic activities were performed by proteins there existed world of RNA-protospecies in which RNA molecules constituted both the genetic material and enzymes. According to this theory the RNA enzymes, called rybozymes, were required for metabolism and for self-replication. However, as it was already shown basing on information loss - selection balance approach, and as it is presented in the paper using branching processes approach, the replication error-rate is a crucial quantity for the maximum information content of the RNA-protospecies. Therefore, one hypothetical rybozyme called RNA replicase is required in the early phase of RNA-World, since it can reduce the mutation rate and thus allow for development of genomes with increasing information content. Otherwise, the information would have been lost, and the error catastrophe would have taken place. However, the information preserved in the RNA replicase itself is strongly limited, because in the phase of evolution proceeding the emergence of this rybozyme the replication could not take the advantage of the low mutation rates and yet the evolution of RNA-strands leading finally to the "invention" of replicase had to satisfy the information limiting constraints.. Therefore, RNA replicase would have never been able to evolve if its function could appear only in RNA chains containing large amounts of information. In the paper this problem is considered using model proposed by Demetrius and Kimmel. This model draws the conclusions relaying on the criticality property in branching processes. While utilizing this approach, the originality of the paper lies is the introduction into the model the parameters which can be experimentally measured in a test tube. Therefore the estimations of the maximum information content of the primordial RNA-based RNA replicase can be determined using data from biochemical experiments. Last but not least, the paper can encourage biochemists for experiments yielding results helpful in the estimation of the probability of the break of phosphodiester bonds in RNA molecules under conditions feasible on the early Earth.

Keywords**—**Criticality in branching processes, origins of Life, RNA-world, RNA-based RNA replicase, complexity threshold, information amount, probability of phosphodiester bond break, computational modeling

I. Introduction

WHAT is Life? How did it originate? Is replication required for reproducing, heredity and thus natural selection? Can replication and metabolism be (at least logically) separated?

Depending on the answers to these questions, the origin of Life can be considered as the subsequent genesis replication and metabolism or simultaneous occurrence of both these processes. The latter consideration is referred to as hypothesis of a single-origin, the former is a case of double-origin hypothesis in which however metabolism appears before replication.

In the theories which assume the following sequence of events

- *first metabolism - then replication*,

the origin of self-replication is explained at first glance quite similarly as in

- *first replication - then metabolism*

theories.

However, there is one crucial difference. It concerns the type of the environment of self-replicating macromolecules which can be treated as precursors of modern genes. Theis problem of occurrence of replication in already biotic conditions versus origin of replication in pre-biotic environment is discussed in more detail in the subsequent sections of the paper basing on the results of the author's experiments with criticality in Demetrius/Kimmel branching process model [1,2].

Aforementioned experiments suggesting substantial limitation of the complexity threshold in the early Life resulting in the limiting amount of information which could be preserved in the evolution of early RNA molecules constituting proto-species, have given some favor to the origin of replication in biotic conditions. These experiments indicated difficulties with self-replication in pre-biotic environment with accuracy of replication comparable to that present in biochemical experiments of RNA molecules evolution in a test tube.

The scientific work financed by Ministry of Science and Higher Education in Poland from funds for supporting science in 2008-2010, as a research project number N N519 319035

K. A. Cyran is with the Institute of Informatics, Silesian University of Technology, Gliwice, 44-100 Poland, phone: +48-32-237-2500; fax: +48-32- -237-2733; e-mail: krzysztof.cyran@ polsl.pl.

Perhaps it is also worth to notice that these conclusions corroborate with results obtained with the use of others methods. The method based on a balance between information loss and Darwinian selection predicts equally clearly the difficulties with the origin of self-replicating macromolecules in abiotic environment.

The novelty of this current article is the introduction to the model the parameter associated with probability of the phosphodiester bond break, the parameter which can be experimentally measured in a test tube. Therefore presented here results can be easily refined after a series of biochemical experiments yielding the estimates of the feasible values of this probability under the geological conditions presumable present on the young Earth.

Despite the results which limit the complexity threshold and thus the information content in the replicating macromolecules in abiotic conditions, the view that Life originated twice, first as metabolism and then as replication, is not very popular. More fashionable, perhaps due to its elegant simplicity, view states that Life started with replication only.

The origin of metabolic apparatus within such theories is treated as the milestone in the evolution, but not as the second origin *per se*. This neat picture is nowadays represented by many theories of RNA world, introduced formally in 1986 by Gilbert [3] and then in 1989 by Joyce [4] who refers to hypothetical early stage of the evolution when both genetic and structural/metabolic functions were performed by macromolecules of RNA.

There are, however, some serious problems with this elegant model. The reasoning included in a classical paper by Niesert et al. [5] suggesting serious problems with stability of proposed by Eigen [6] hyper cycles, as well as computer modeling of metabolic self-sustaining cycles done by Sagre and Lancet [7] and estimations of conditions required for avoiding error catastrophe caused by the loss of information in the RNA world signal only some of these difficulties. When trated seriously, they suggest that this neat picture should be perhaps changed to sort of garbage-bag world despite of its less elegant structure.

The Garbage-bag world used by Dyson [8] for representation of the life after the beginning of metabolism assumes the second origin, that of replication, in already biotic conditions. Numerical experiments performed by the author of this paper with criticality and extinction of branching processes seem to confirm Dyson's view for some values of the probability of phosphodiester bond break, or at least they give the strong limits within which the replication should operate in order to avoid extinction of RNA- macromolecule-based proto-species.

In known to us cellular form of life the relation between replication and metabolism is always defined by the inherent circular interconnections. Within a cell the replication of the genome is possible only with the help of protein enzymes. Not only do they catalyze metabolic activity of the cell, but also, which is even more crucial, the production of them is directed by the information encoded in a genome.

In such circular dependencies it is impossible to separate one phenomenon from the other and based solely on the observation of cellular life it is difficult to say which function originated first. However, despite some difficulties it is possible to imagine other forms of life which are not only logically coherent but they can also provide a clue to determination of the sequence of events leading to modern Life.

II. Metabolism, Replication and Reproduction: How Are they related?

It is generally accepted view that two unlikely events, like the birth of metabolism and that of replication in abiotic conditions, are more probable to occur sequentially than at the same moment. This becomes even more evident if we realize that if these events are to occur in a sequence then the second unlikely event would occur in already biotic environment organized by the first.

Therefore minority of scientists (if any) believe the origin of the cells as we know them can happen as one extremely unlikely event. Instead, they suggest separation of the events in time, with majority advocating the birth of replication before the start of metabolism [3, 4, 9-11] and minority believing the contrary [8].

In this paper the majority's view will be referred to, after Dyson [8], as single-origin theories, since the origin of self-replication is considered in them not only as the first origin, but as the only origin. Contrary to that, the minority's view states that the first sort of life started with the birth of metabolic (autocatalitic) activity within a protocell and the second origin of replicative life appeared when self-replicating nucleotide strands became the parasites within the protocells. In the double-origin hypothesis we consider therefore two distinct forms of life:

- autocatalitic but not self-replicating life of protocells which reproduce distributing the molecules in a statistical fashion

being the first hypothetical form, and

- purely parasitic life of self-replicating but not metabolizing nucleotide strands

constituting the second form of living organisms emerging only in the environment prepared by the first form.

The evolution of both kinds of life by millions of years could have led to the symbiosis of host autocatalitic protocells and their replicative parasites. In subsequent millions of years the complete interdependence of both species might have occurred, which is visible in contemporary life where it is virtually impossible to separate metabolism from replicative activities of the cells.

After this discussion, the next two issues should be clarified. The first is

- the difference between notion of replication and that of reproduction

while the second is

- the meaning of the term metabolism.

Very often reproduction and replication are used as synonyms. It is justified, since always in observable cellular life the reproduction of the cell is performed together with the replication of the genomic DNA or RNA molecules. Yet there is the fundamental difference between the two.

Reproduction is a term denoting an action of the cell dividing into two daughter cells having similar properties to parental cells. Even if today this is always accompanied by process of replication of cell's genome, the latter is not logically necessary for the first.

One can also imagine reproduction performed in less deterministic way. The level of inheritance in such a system would of course be lower but statistical reproduction would not completely suppress inheritance [8]. And it is inheritance resulting from reproduction and not the exact replication of genomic macromolecules what is really important for the evolution driven by Darwinian natural selection.

In fact, Darwin had no idea about replication of nucleotide polymers when he proposed his theory of evolution. Therefore, replication is neither an assumption, nor a consequence of his theory. It is rather a very efficient way of reproduction, but not the only one leading to inheritance.

Even though it is not logically required for inheritance, replication has been treated as the basis for reproduction from the very beginning in all single-origin theories. On the contrary, in double-origin hypotheses replication of a genome had evolved from a parasitic, self-replicating form of life.

According to such hypotheses it happened long after the invasion of replicating parasites on statistically reproducing protocells performing metabolic activities for themselves and for parasites.

Molecular biologists would be astonished by such conjecture as long as they mean the term metabolism as genetically driven activity of a cell while it is not the only one. Metabolism can also be considered as self-sustained autocatalitic activity of a cell capable for extraction of negentropy from the environment. Such meaning was prevailing in the times when the nature of the replication was unknown [12] and it is still present especially in German language literature [8].

III. Formulation of the Problem

After clarification of the terms and their meaning in previous section, the problem tackled in this paper can be posed.

The amount of information in hypothetical RNA-protospecies can be considered in several stages of the RNA-World. Here the author is interested only in the early (but not the forst) stage of this hypothetical world directly proceeding the first phase of short oligonucleotides of the length not exceeding 30-50 units. The radical change of the possible ways of reconstructing the RNA polinucleotides in these two phases is the reason of the differentiation between the two. This issue will be explained in detail in the next section, however it should be stated now that in both phases the protospecies are considered to be as simple as possible, i.e. they are single strands of RNA macromolecules.

Because of this very simple form of the protospecies in the considered stage of the RNA-world, the amount of information preserved in such organisms can be directly correlated with the length of the RNA strand. The four letter alphabet of adenine (A), cytosine (C), guanine (G) and uracil (U) is used to strore the genetic information and the classical information theory can be used to compute the amount of information carried by this molecule.

However, with the length of the molecule the notion of the complexity threshold comes on the scene. That latter defines a maximum length of self-replicating RNA strands which could avoid error catastrophe. In this and subsequent parts of the paper there are summarized the main theoretical results of Demetrius/Kimmel model [1, 2] which can be used for computation of complexity threshold for different mutation rates and probabilities of RNA hydrolysis.

It is a well known fact that reproduction involving replication of genetic material produces almost identical copies of parental cells. The word *almost* is of great concern in the whole history of Life since it reflects possibilities of rare changes caused by mutations on one hand, and relative constancy of the genotype of the given species on the other. The exact replication of nucleotide strands could not have led to the whole variety of the living creatures. On the other hand, too large mutation rate would have led to error catastrophe and the process of evolutionary organization of Life could not have proceeded.

Genetic experiments indicated that the rate of mutation in contemporary organisms is influenced by a lot of factors, such as the DNA repairs performed by protein enzymes called DNA helicases coded by such genes as RECQL, BLM, WRN [13 – 14] and many others. They are involved in surprisingly many phases of DNA metabolism, including

- transcription,
- recombination,
- accurate chromosomal segregation,

and various mechanisms of DNA-repair, such as [15]

- mismatch repair,
- nucleotide excision repair, and
- direct repair.

These mechanisms have evolved because genomes are often subject to damage caused by chemical and physical agents present in the environment, or by [15]

- endogenously generated alkylating agents,
- free radicals, and
- replication errors.

Therefore effectiveness of genome repair is one of the crucial factors determining the fitness of species.

However, species capable of DNA repair must have long genomes used for coding many complex enzymes, including mentioned helicases. Hence, to assure more accurate replication (i.e. the smaller mutation rate), the longer nucleotide chain is required. This is of course reflected in the growing amount of information needed for the coding so many functions.

Yet for longer chains there is smaller probability that they are replicated without error for given mutation rate due to almost independent replications of separate nucleotides in a chain. The conclusion of this discussion is the existence of maximum length of a poly-nucleotide strand that will not (almost surely, i.e. with probability one) become extinct. This length is called the complexity threshold and its value is surely dependent on the mutation rate per nucleotide, as well as on ability of the poly-nucleotide strand to survive for subsequent replication. This length also defines the maximum amount of the information content in the early RNA-protospecies which could have replicated without help of the RNA-replicase rybozme.

IV. Solution based on criticality in branching processes

As it was already stated, the goal of this study is estimation of the complexity threshold in the early phase of RNA-World, after the stage of very short RNA oligonucleotides (up to 20-30 units). The latter phase, extensively simulated by Ma et al. [16], is characterized by non-negligible probability of restoring the sequence of oligonucleotide strand from scratch, i.e. for oligos of the length λ being considerably less than 50 (the probability of restoring from scratch the sequence of 50 nucleotides is equal $4^{-50} = 2^{-100}$ i.e. it is smaller than 10^{-30}, thus it can be safely considered as negligible).

Therefore it can be safely concluded that the sequences composed of more than 100 nucleotides must have occurred in the continuous evolution of shorter sequences rather than by *ab-initio* creation. The consequence of this fact is that once a given lineage of sequences (proto-species) becomes extinct it practically cannot be brought back to existence (unless highly improbable process of random setting of required nucleotides would have happen).

With aforementioned assumptions, consider the RNA-species with the RNA chain of the length λ, where $\lambda > 50$. Let such species replicate with the mutation rate per nucleotide equl μ.

Then, the probability that the single nucleotide in a strand is copied without an error is given by $p = 1 - \mu$. This yields in a model of independent nucleotides replications the probability of correct replication of the whole polinucleotide strand equal $v = p^{\lambda}$.

Consider also situations designated by S_0, S_1 and S_2 yielding in the next generation 0, 1, and 2 individuals respectively. Denoting the probability that RNA strand is not hydrolyzed by w, it is obvious that

- situation S_0 takes place when, with the probability 1-w, individual does not survive to replication stage (next generation) because of hydrolysis.

Similarly,

- S_1 denotes the situation when, with the probability $w(1- v)$, the individual is not hydrolyzed at least until next generation, but it produces a copy of itself with an error.

Finally,

- situation S_2 denotes the case when, with the probability wv, the individual not only is not hydrolyzed but also it replicates without error yielding two identical strands.

In the further analysis it is assumed that the Demetrius/Kimmel model is used. It proposes that the population of error-free RNA strands follows the Galton-Watson branching process with the number of individuals Z_t at time t (Fig. 1).

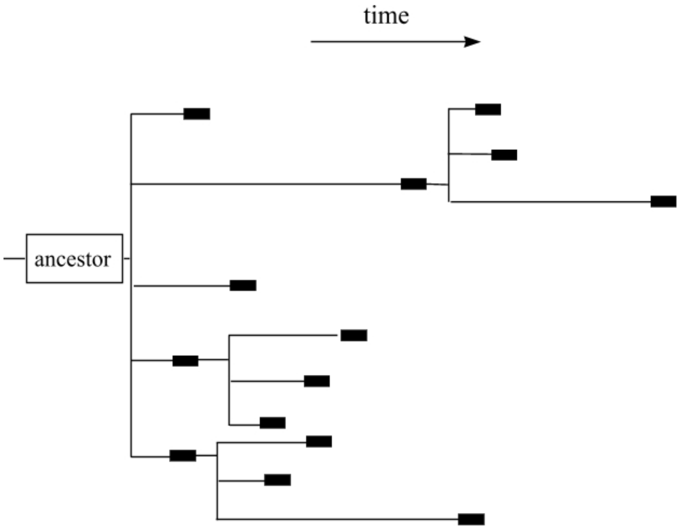

Fig. 1. The schematic diagram of the branching process. Each progeny acts like a new ancestor

Formally, the branching process is defined by [2]

- Doubly infinite family of iid rv's $\{X_{i,n}\}$, "potential" numbers of progeny of i-th individual in generation n
- Z_t = number of individuals in generation t, obtained by summation

$$Z_0 \qquad = 1$$

$$Z_{n+1} \quad = \begin{cases} X_{1,t} + \ldots + X_{Z_t,t}; & Z_t > 0, \\ 0; & Z_t = 0, \end{cases} \qquad n \geq 1,$$

Such a process is said to be supercritical when the probability of its eventual extinction q satisfies inequality $q < 1$. This happens only when $E(X) > 1$, where random variable X (denoting the number of descendants of given individual chain) is equal zero, one or two with probabilities of situations S_0, S_1, and S_2 respectively.

Interestingly, even when $\lim_{t\to\infty} E(Z_t) = 1$ for $E(X) = 1$, the probability of eventual extinction $q = 1$ in this critical case, despite the result looks somewhat counterintuitive. Consider now the probability generating functions of variables X in branching processes

Let $f(s)$ be the progeny probability generating function. Then the asymptotic behavior of $f_t(s)$ determines the limit theorems for the process $\{Z_t\}$. Since $f(s)$ is a power series with non-negative coefficients $\{p_k\}$, $p_0 + p_1 < 1$, and $f'(1^-)=m$ therefore

- $f(s)$ is strictly convex and increasing in [0,1]
- $f(0) = p_0$ and $f(1)=1$
- If $m \leq 1$ then $f(s) > s$ for $s \in [0,1)$
- If $m > 1$ then $f(s) = s$ has a unique root in $s \in [0,1)$.

It follows that the probability generation function $f(s)$ of the progeny numbers in the process modeling the evolution of the early RNA-protospecies is given by

$$f(s) = 1 - w + w(1-v)s + wvs^2. \qquad (1)$$

Lemma 1: The iterates of probability generating function $f(s)$ converge to the smallest root q of the equation $f(s) = s$ for $s \in [0,1]$.

Proof:

If $s \in [0,q)$ then $f_t(s) \uparrow q$ as $t \to \infty$
If $s \in (q,1)$ then $f_t(s) \downarrow q$ as $t \to \infty$
If $s=q$ or $s=1$ then $f_t(s)=s$ for all t

Let us consider the special case of the Lemma 1 when $s=0$. It follows that $f_t(0) \uparrow q$ as $t \to \infty$.

However, since

$$\begin{aligned} &\lim_{t\to\infty} f_t(0) = \\ &\lim_{t\to\infty} P(Z_t = 0) = \\ &\lim_{t\to\infty} P(Z_i = 0 \ \text{ for some } 1 \geq i \geq t) = \\ &P(Z_i = 0 \text{ for some } i \geq 1) = \\ &P(\lim_{t\to\infty} Z_t = 0) \end{aligned}$$

which by the definition is the probability that the process ever becomes extinct, therefore the extinction probability of the process $\{Z_t\}$ is the smallest non-negative root q of the equation $f(s)=s$. It is equal to 1 if $m \leq 1$ and it is less than 1 if $m > 1$.

In considered in the paper case, the probability of extinction q, being the smallest positive root of the equation $f(s) - s = 0$, with roots $q_1 = 1$ and $q_2 = (1-w)/wv$, is obviously equal to q_2. Of course in order to avoid the necessity of extinction (with probability one), q_1 must be greater than q_2, which yields the inequality

$$\frac{1-w}{w} < v. \qquad (2)$$

Even if inequality (2) is satisfied there is a chance of extinction, which can happen with probability $P(ext)=q_2$ and long-term survival of species is expected only with probability $P(surv)$ given by

$$P(surv) = 1 - \frac{1-w}{wv}. \qquad (3)$$

The illustration of extinction of supercritical branching processes modeling species not exceeding complexity threshold is presented in Fig. 2. Simulations were performed by the software using sophisticated random number generators described in [17, 18]. This software, written by the author, was used also in [19] to make inferences about interactions of Neanderthals and archaic *H. sapiens*, based on the mitochondrial DNA record. It was also used to estimate the age of the root of mitochondrial DNA polymorphism of modern humans in [20] and in extended version in [21].

Result described by (2) can be obtained also directly from the criticality condition expressed as inequality $E(X) = f'(1) = w\,(1+v) > 1$. Indeed, the last statement is satisfied only when formula $v > (1-w)/w$ given by inequality (2) holds.

Substituting $v = (1-\mu)^{\lambda}$ and solving with respect to λ there is obtained the complexity threshold satisfying

$$\lambda < \frac{\ln(1-w) - \ln w}{\ln(1-\mu)}. \qquad (4)$$

The above formula does not take into consideration the fact that probability w of avoiding by RNA strand the hydrolysis at least to the subsequent replication event is also dependent on the length of the strand λ.

To introduce this dependency, consider more detailed model in which parameter r denotes the probability of breaking the phosphodiester bond between nucleotides in the RNA strand in the time between successive replications. Since in a strand of λ nucleotides there are λ – 1 bonds, therefore $w = (1-r)^{\lambda-1}$.

In the new model it is impossible to obtain explicit formula for λ so possible values should be computed numerically from inequality

$$1 < (1-r)^{\lambda-1}\left(1+(1-\mu)^{\lambda}\right). \qquad (5)$$

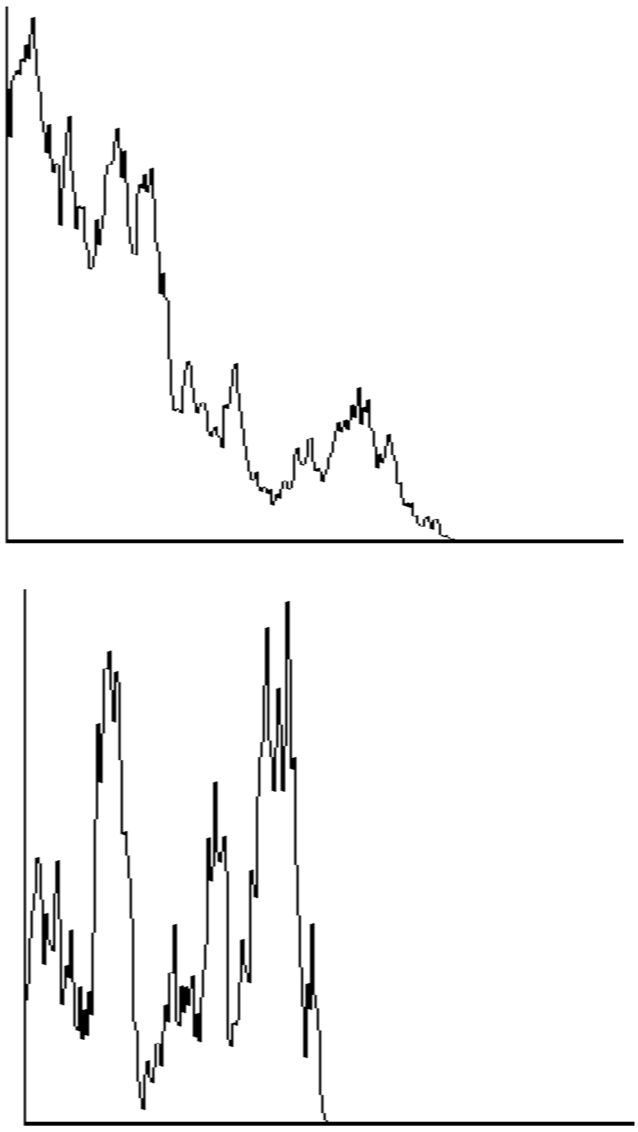

Fig. 2. Two extinct realizations of the evolution based on a branching process started from 100 individuals with length slightly below complexity threshold (adapted from [22]).

Assuming that the complexity threshold, denoted as $\lambda_{critical}$, is defined as such λ for which formula (5) modified to be an equation holds, the critical mutation rate $\mu_{critical}$ is given by

$$\mu_{critical} = 1-\left(\frac{1}{(1-r)^{\lambda_{critical}-1}}-1\right)^{\frac{1}{\lambda_{critical}}} \qquad (6)$$

For all mutation rates larger than $\mu_{critical}$ RNA species become extinct with probability one. In the Fig. 3 the 3D plot of the border function for $\mu_{critical}$ is presented for range of parameter $\lambda_{critical}$ from 1 to 10^3 and range of parameter r from 10^{-4} to 10^{-3}. Fig. 4 presents similar plot for parameter r ranging from 10^{-5} to 10^{-4}.

It is clearly visible that for larger values of $\lambda_{critical}$ the critical mutation rate $\mu_{critical}$ must be smaller. Not surprisingly, the slope of this function decreasing with $\lambda_{critical}$ is steeper for higher probabilities of the phosphodiester bond break r.

Since all experiments of evolution of RNA strands in abiotic conditions in a test tube have yielded mutation rates greater than 10^{-2}, this value can be treated as the cutoff level for the surfaces presented in Fig. 3 and 4. Only points with such coordinates ($\lambda_{critical}$, r) for which the surface of the function $\mu_{critical}$ is above this cutoff represent conditions feasible for long-lasting evolution of the RNA protospecies which avoid the error catastrophe.

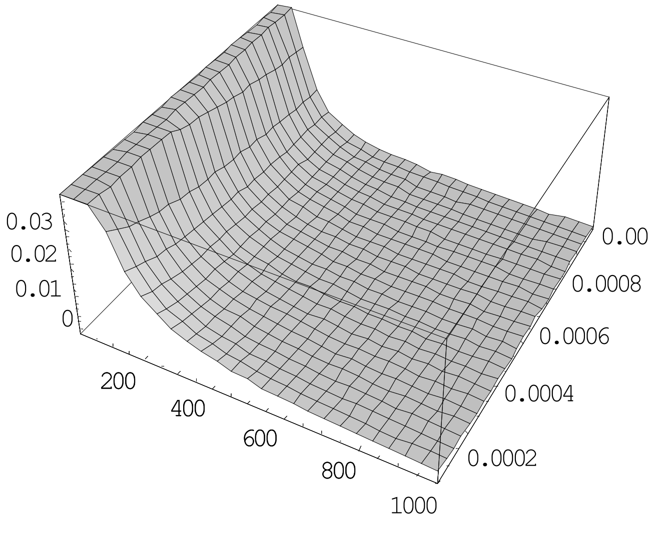

Fig. 3. Surface of the function $\mu_{critical}(\lambda_{critical}, r)$ for r ranging from 10^{-4} to 10^{-3} and $\lambda_{\text{critical}}$ ranging from 1 to 10^3 .

The surfaces presented in Fig. 3 and 4 provide a lot of qualitative information about the character of function representing the critical mutation rate with respect to complexity threshold and probability of the break of phosphodiester bond. One of the most interesting features which can be studied from these figures is the monotonicity of the two-dimensional function.

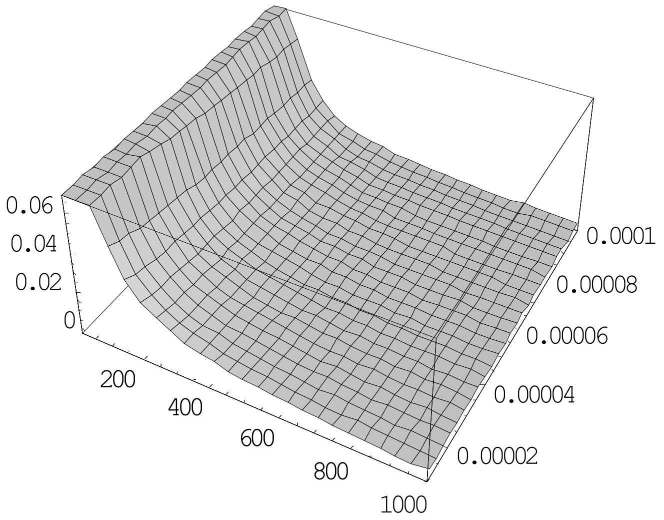

Fig. 4. Surface of the function $\mu_{critical}(\lambda_{critical}, r)$ for r ranging from 10^{-5} to 10^{-4} and $\lambda_{\text{critical}}$ ranging from 1 to 10^{3} .

However, it is impossible to read from these charts the quantitative characteristics. Therefore, having in mind the monotonic course of the function $\mu_{critical}$ with respect to r, instead of presenting two-dimensional surfaces, Fig. 5, 6, and 7 show one-dimensional curves for values of parameter r equal to 10^{-3}, 10^{-4}, and 10^{-5} respectively, assuming $\mu_{\text{criticail}} = 10^{-2}$.

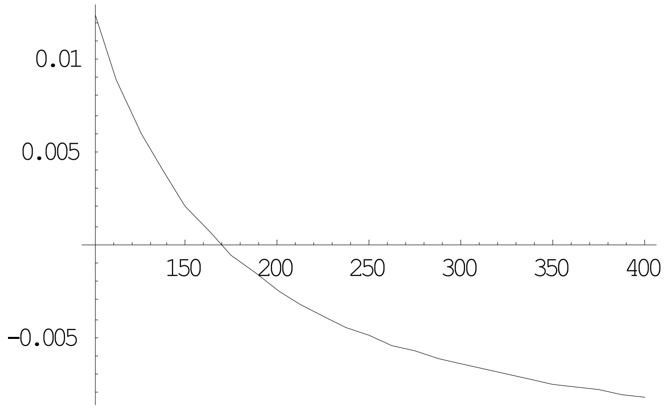

Fig. 5. The complexity threshold for $r = 10^{-3}$ and $\mu_{\text{criticail}} = 10^{-2}$.

The choice of such three values of this parameter is based on [16] and will be discussed further in the Conclusions. Here it is sufficient to mention that these values are representative for the wide range of this parameter varying from 10^{-5} to 10^{-3} because of monotonic character of the function with respect to this parameter. Therefore for all values between these three points the resulting one-dimensional curves would yield the values of the complexity threshold between those presented in the figures.

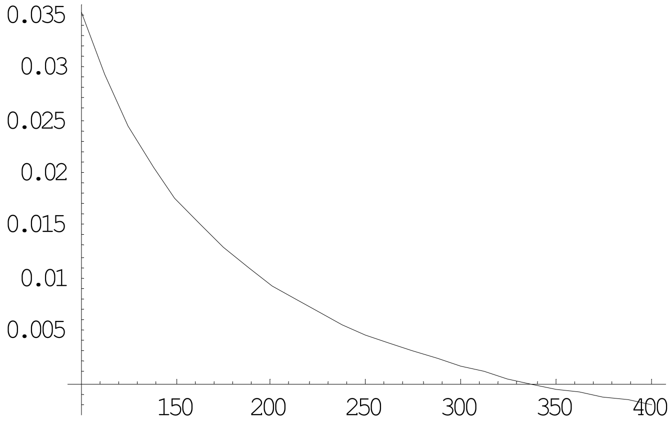

Fig. 6. The complexity threshold for $r = 10^{-4}$ and $\mu_{\text{criticail}} = 10^{-2}$.

Additionally, for better illustration, the mutation rate cutoff value 10^{-2} was subtracted from the function $\mu_{critical}$, defined by (6), so the resulting plots cross the horizontal axis exactly at the point indicating the complexity threshold $\lambda_{critical}$.

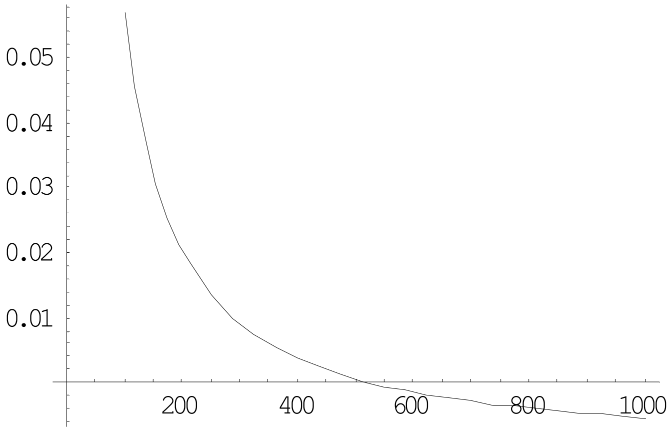

Fig. 7. The complexity threshold for $r = 10^{-5}$ and $\mu_{\text{criticail}} = 10^{-2}$.

While Figures 5, 6, and 7 assume the mutation rate per nucleotide $\mu_{\text{criticail}} = 10^{-2}$ such rate is reported rather as a limit of accuracy in replication without help of the RNA replicase than the actual inaccuracy. Compare for example [8]:

”All the experiments that have been done with RNA replication under abiotic conditions give error rates of the order of 10^{-2} at best”

The vast majority of experiments yield this rate to be as big as 2×10^{-2} or even 5×10^{-2}, as it is reported in [10]:

”The error rate depends on the medium, the temperature, and so on, but very roughly the wrong base pairs with a G once in 20 times”.

Having this in mind the set of Figures 8, 9, 10 and set of Figures 11, 12, 13 are the counterparts of the set of Figures 5, 6, 7 where the mutation rates per nucleotide $\mu_{\text{criticail}}$ are 2×10^{-2} and 5×10^{-2} respectively.

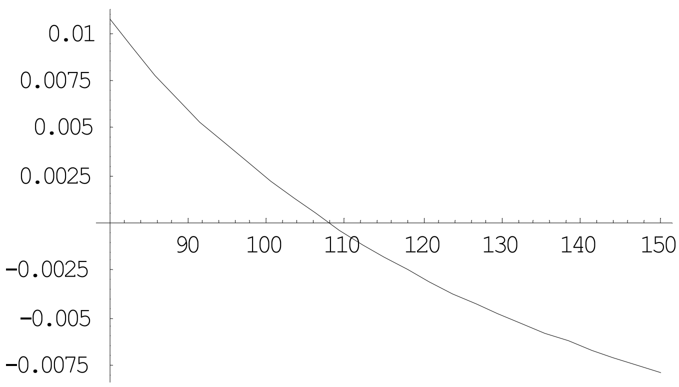

Fig. 8. The complexity threshold for $r = 10^{-3}$ and $\mu_{criticail} = 2\times10^{-2}$.

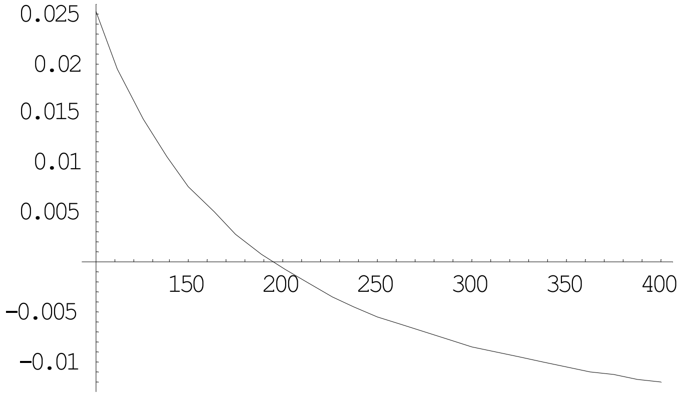

Fig. 9. The complexity threshold for $r = 10^{-4}$ and $\mu_{criticail} = 2\times10^{-2}$.

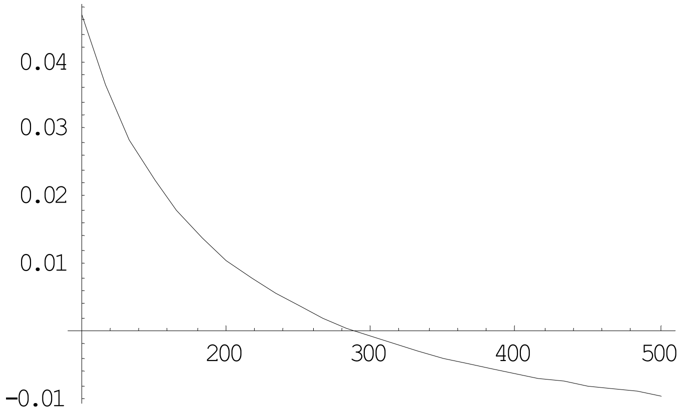

Fig. 10. The complexity threshold for $r = 10^{-5}$ and $\mu_{criticail} = 2\times10^{-2}$.

V. Discussion and Conclusions

The problem with RNA-world is that the rybozyme, crucial for replication of any RNA-species, called RNA-replicase is yet to be discovered. Many advocates of RNA-world do not consider this lack seriously, believing that it is only the matter of time when experimental confirmation happens. This belief is based on strong foundations, since some enzymatic activity exhibited by RNA molecules has been already demonstrated. Extending the range of discovered rybozymes to RNA-replicase is only one step further.

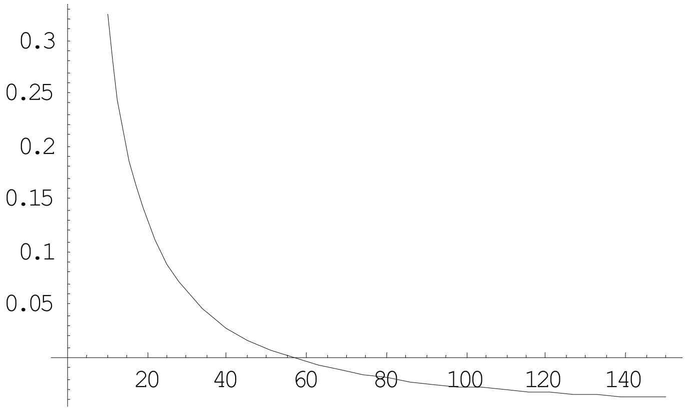

Fig. 11. The complexity threshold for $r = 10^{-3}$ and $\mu_{criticail} = 5\times10^{-2}$.

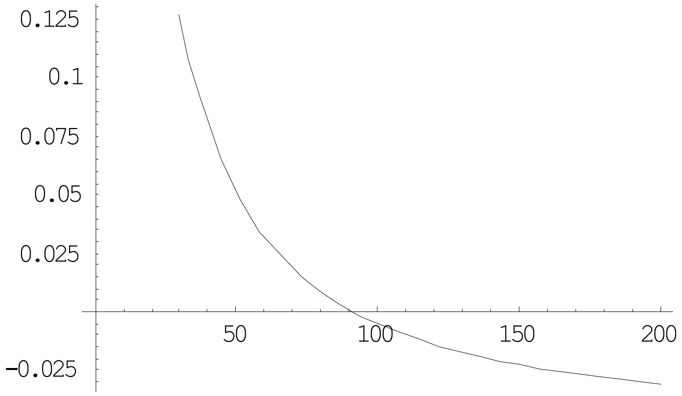

Fig. 12. The complexity threshold for $r = 10^{-4}$ and $\mu_{criticail} = 5\times10^{-2}$.

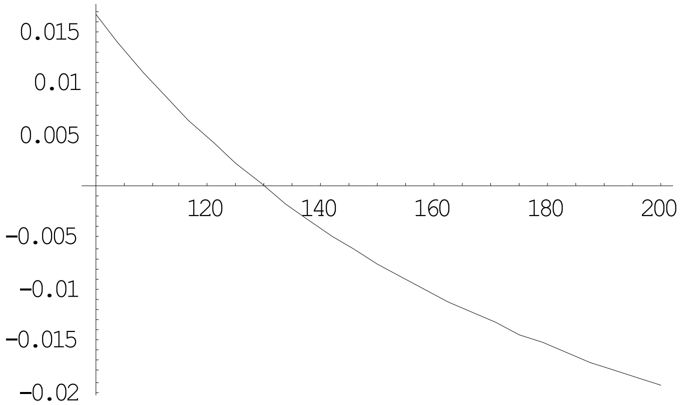

Fig. 13. The complexity threshold for $r = 10^{-5}$ and $\mu_{criticail} = 5\times10^{-2}$.

Perhaps the above line of argument is correct, however, except for the only hypothetical existence of RNA-based RNA-replicase there exists at least one more serious problem and the paper focuses on it. This problem is caused by the possible error catastrophe which can easily occur when RNA strands try to replicate in abiotic conditions. All experiments of the evolution of RNA performed in a test tube indicate that without help of replicase enzyme the error of replication is larger than 10^{-2}. Even if we assume that instead of enzyme the rybozyme could be used in the RNA-world, there had to be a period when even primordial replicase had not yet evolved.

Extensive simulations of the very first phase of RNA-world have proved that it is feasible to create short RNA strands of the length not exceeding 30 nucleotides [16]. It is however hard to believe that such short oligonucleotides could catalyze its own replication in selective way. And such selective primordial replicase is required to amplify the growth of itself and not all unrelated strands. What is the maximum length of selective replicase? In the light of experiments reported in the paper this critical length is dependent on probability of phosphodiester bond break. What are feasible values of this parameter then?

Certainly r is dependent on the environmental conditions like temperature or the concentration of nucleotides in a solution (the smaller concentration the longer time between replications and thus larger w for equal values of temperature and other environmental parameters such like pH of the solution for example).

However, it is possible to obtain experimentally the value of r for given environment. The conditions existing on early Earth, feasible from geological point of view, can be therefore simulated in a test tube and then the model proposed in a paper can be applied with the reliable value of parameter r, as it is already in the case of parameter μ of the order 0.01. In lack of such experiments, the wide range of r from 10^{-5} to 10^{-3} was treated as plausible.

The author has the hope that after performing aforementioned chemical experiments the refinement of the limiting information content can be achieved using presented in the paper methodology. Until there is a lack of reliable estimates of the phospodiester bond break probability, the discussion with a broad range of its possible values is of some worth.

The extremely low value of this probability such as 10^{-5} represents the situation of substantial concentration of nucleotides and other environmental conditions supporting fast replication. Whether such conditions are feasible on the early Earth is an open question, but if so, they simplify imagining the evolution of hypothetical primordial RNA replicase to selective replicase catalyzing only its own replication. The complexity threshold for such rybozyme exceeds 500 nucleotides which is probably enough to activate the proposed function. However, if r proves to be as large as 10^{-3} or more, then the complexity threshold for selectively working replicase is considerable less and only 170 nucleotides must have been sufficient to constitute such rybozyme.

While this is not impossible, the result would have limited the domain of hypothetical replicases to sequences shorter than 200 nucleotides. Perhaps it would also suggest the double-origin hypothesis in which the replication occurs in biotic condition of metabolizing proto-cells. Such conditions could easier reduce the mutation rate to 10^{-3}, the value that yields complexity threshold well above 10^3 for any considered value of parameter r.

And last but not least the issue of the amount of information. In the paper it is shown that the before the emergence of the primordial RNA-replicase rybozyme the amount of information which could been preserved in self-replicating RNA protospecies had to be limited to 10^3 bits. This is twice (because one nucleotide can code for two bits of information) the complexity threshold limit for parameters $\mu = 10^{-2}$, $r = 10^{-5}$. Most probably, i.e. for $\mu = 2\times10^{-2}$ and $r = 10^{-4}$ the amount of information could not exceed 4×10^2 bits

If the RNA-world had ever existed the Nature had to find very information efficient system being able to encode the complex function of emerging primordial RNA-replicase rybozyme having not more than 200 nucleotides i.e. using probably only less than 400 bits of information.

Acknowledgments

The author would like to thank Prof. M. Kimmel from Department of Statistics at Rice University in Houston TX, USA, for advice and long discussions concerning application of branching processes to biological research.

References

[1] L. Demetrius, P. Schuster, K. Sigmund, “Polynucleotide Evolution and Branching Processes,” *Bull. Math. Biol.*, vol. 47, 1985, pp. 239-262.

[2] M. Kimmel, D. Axelrod, *Branching Processes in Biology*, Springer-Verlag, New-York, 2002.

[3] W. Gilbert, “The RNA World,” *Nature*, vol. 319, 1986, pp. 618-618.

[4] G.F. Joyce, “RNA Evolution and the Origins of Life,” *Nature*, vol. 338, 1989, pp. 217-224.

[5] U. Niesert, D. Harnasch, C. Bresch, “Origin of Life between Scylla and Charybdis,” *J. Mol. Evol.*, vol. 17, 1981, pp. 348-353.

[6] M. Eigen, W. Gardiner, P. Schuster, R. Winckler-Oswatitch, “The Origin of Genetic Information,” *Sci. Am.*, vol. 244, no. 4, 1981, pp. 88-118.

[7] D. Sagre, D. Lancet, “A Statistical Chemistry Approach to the Origin of Life,” Chemtracts–*Biochem. Mol. Biol.*, vol. 12, no. 6, 1999, pp. 382-397.

[8] F. Dyson, *Origins of Life*, *Revised Edition*, Cambridge University Press, 1999.

[9] E. Szathmary, L. Demeter, “Group Selection of Early Replicators and the Origin of Life,” *J. Theor. Biol.*, vol. 128, 1987, pp. 463-486.

[10] J.M. Smith, E. Szathmary, *The Origins of Life. From the Birth of Life to the Origin of Language*, Oxford University Press, 1999.

[11] K.E. McGinness, G.F. Joyce, “In Search of Replicase Rybozyme – Review,” *Chem. Biol.*, vol. 10, 2003, pp. 5-14.

[12] E. Schroedinger,“*What is Life*? *The Physical Aspect of the Living Cell*, Cambridge, Cambridge University Press, 1944.

[13] K.A. Cyran, “Rough Sets in the Interpretation of Statistical Tests Outcomes for Genes under Hypothetical Balancing Selection,” M.

Kryszkiewicz, J.F. Peters, H. Rybinski, A. Skowron (Eds.). *Lecture Notes in Artificial Intelligence*, vol. 4585, 2007, pp. 716-725.

[14] K.A. Cyran, ”PNN for Molecular Level Selection Detection,” N. Mastorakis (Ed.), *Lecture Notes in Electrical Engineering*, vol. 27, 2009, pp. 35-41.

[15] K.A. Cyran, J. Polańska, M. Kimmel, “Testing for Signatures of Natural Selection at Molecular Genes Level,” *J. Med. Informatics Technologies*, vol. 8, 2004, pp. 31-39.

[16] W. Ma, Ch. Yu, W. Zhang, “Monte Carlo Simulation of Early Molecular Evolution in the RNA World,” *BioSystems*, vol. 90, 2007, 28-39.

[17] G. Marsaglia, “Monkey Tests for Random Number Generators,” *Comput. Math. Appl.*, vol. 9, 1993, pp. 1-10.

[18] G. Marsaglia, A. Zaman, W.W. Tsang, “Toward a Universal Random Number Generator,” *Stat. Prob. Lett.*, vol. 8, 1990, pp. 35-39.

[19] K.A. Cyran, M. Kimmel, “Interactions of Neanderthals and Modern Humans: What Can Be Inferred from Mitochondrial DNA?,” *Math. Biosci. Eng.*, vol. 2, 2005, pp. 487-498.

[20] K. A. Cyran, “Mitochondrial Eve dating based on computer simulations of coalescence distributions for stochastic vs. deterministic population models,” in *Proc. the 7th WSEAS International Conf. on Systems Theory and Scientific Computation*, Athens, Greece, August 2007, pp. 107-112.

[21] K.A. Cyran, ”Simulating branching processes in the problem of Mitochondrial Eve dating based on coalescent distributions,” *International Journal of Mathematics and Computers in Simulation*, vol. 1, no. 3, 2007, 268-274.

[22] K.A. Cyran, U. Stańczyk, “Stochastic simulations of branching processes: Study on complexity threshold of RNA-world species,” Proc. XXXVI Ogólnopolska Konferencja Zastosowań Matematyki, Zakopane, Poland, 2007, pp. 19-22.

Krzysztof A. Cyran was born in Cracow, Poland, in 1968. He received MSc degree in computer science (1992) and PhD degree (with honours) in technical sciences with specialty in computer science (2000) from the Silesian University of Technology SUT, Gliwice, Poland. His PhD dissertation addresses the problem of image recognition with the use of computer generated holograms applied as ring-wedge detectors.

He has been an author and co-author of more than 80 technical papers in journals (several of them indexed by Thomson Scientific) and conference proceedings. These include scientific articles like: K. A. Cyran and A. Mrózek, “Rough sets in hybrid methods for pattern recognition,” *Int. J. Intel. Syst.*, vol. 16, 2001, pp. 149-168, and K. A. Cyran and M. Kimmel, “Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA?” *Math. Biosci. Eng.*, vol. 2, 2005, pp. 487-498, as well as a monograph: U. Stańczyk, K. Cyran, and B. Pochopień, *Theory of Logic Circuits*, vol 1 and 2, Gliwice: Publishers of the Silesian University of Technology, 2007. Dr. Cyran (in 2003-2004) was a Visiting Scholar in Department of Statistics at Rice University in Houston, US. He is currently the Assistant Professor and the Vice-Head of the Institute of Informatics at Silesian University of Technology, Gliwice, Poland. His current research interests are in image recognition and processing, artificial intelligence, digital circuits, decision support systems, rough sets, aviation and aeronautics, computational population genetics and bioinformatics.

Dr. Cyran has been involved in numerous statutory projects led at the Institute of Informatics and some scientific grants awarded by the State Committee for Scientific Research. He also has received several awards of the Rector of the Silesian University of Technology for his scientific achievements. In 2004-2005 he was a member of International Society for Computational Biology. Currently he is a member of the Editorial Board of Journal of Biological Systems, member of the Scientific Program Committee of WSEAS international conferences in Malta (ECC’08), Rodos (AIC’08, ISCGAV’08, ISTASC’08) and multiconference in Crete (CSCC’08) as well as a reviewer for Studia Informatica and such journals indexed by Thompson Scientific as: Optoelectronic Review, Mathematical Biosciences and Engineering, and Journal of Biological Systems.