Krzysztof A. Cyran and Urszula Stańczyk

Institute of Informatics, Silesian University of Technology
Akademicka 16, 44-100 Gliwice, Poland

Stochastic Simulations of Branching Processes: Study on Complexity Threshold of RNA-World Species

The main reason of popularity of RNA-world hypothesis is the fact that it resolves by introduction of RNA-species the unsolved in another way problem of anteriority of DNA or proteins in the beginning of Life [1]. The problem is similar to that with an egg and a hen: the enzymatic proteins are needed for the replication of DNA strands (anteriority of proteins), however, DNA strands are required for forming enzymatic proteins (anteriority of DNA). The RNA-world eliminates both: the hen laying an egg and the egg giving the birth to a hen, and introduces the ancestral hybrid capable for performing roles of both a hen and an egg.

Similarly, in hypothetical RNA-species RNA plays the role of both genetic material and enzymes (called rybozymes) required for reproduction. This is a very attractive hypothesis, and perhaps it lacks only the experimental confirmation. The problem is that the enzyme, crucial for reproduction of any RNA-species, called RNA-based RNA-replicase is yet to be discovered. The advocates of RNA-world do not treat seriously this lack of experimental confirmation, believing that it is only the matter of time when such confirmation comes. Remarkably, some enzymatic activity exhibited by RNA molecules has been already discovered, and extending this limited activity to activities performed by all enzymes required for reproduction is only one step further.

Not going into hot debate between RNA-world proponents [1] and antagonists [2], we consider the problem of the complexity threshold of the RNA-species. This threshold defines a maximum sequence length, based on theory of branching processes and stochastic computer simulations. In the paper we summarize main theoretical results of Demetrius et. al. [3] and Kimmel and Axelrod [4], and further discuss them obtaining values of complexity threshold for different mutation rates and surviving probabilities. The stochastic simulations of slightly supercritical branching process for species not exceeding complexity threshold are also presented. They are performed by the software using sophisticated random number generators described in [5, 6]. This simulating branching processes software was used also in [7] to make inferences about interactions of Neanderthals and archaic *H* sapiens, based on the mitochondrial DNA record.

It is known that different species during reproduction produce almost identical copies of themselves. The word *almost* is of great concern in our study (and probably in the whole history of Life) since it is responsible for rare changes (caused by mutations) on one hand, and for keeping relatively constant genotype of the given species on the other. If the replication of nucleotide strands had occurred without any error then natural selection could not lead to whole variety of the living creatures. On the other hand if the mutation rate was to large, then instead of self-replication we would have random changes of the nucleotide sequence and the process of evolutionary organization of Life could not proceed.

Genetic observation indicate that the rate of mutation is influenced by a lot of factors, including the DNA repair performed by DNA helicases. They are involved in many aspects of DNA metabolism, including transcription, accurate chromosomal segregation, recombination, and repair. Helicase-dependent DNA repair include mismatch repair, nucleotide excision repair, and direct repair [8, 9]. Since genomes are subject to damage by chemical and physical agents in the environment, as well as by free radicals, endogenously generated alkylating agents or replication errors, the genetically determined effectiveness of repair is one of the important factors determining the fitness of corresponding phenotype.

Species which posses such complicated mechanisms of DNA repair must have long genome capable for coding many complex enzymes. This simple observation leads to the conclusion: the more accurate replication of the separate nucleotide (i.e. the smaller mutation rate per nucleotide), the longer nucleotide chain required. However, the longer chain, the smaller probability that it is replicated without error, given the value of a mutation rate. It is so because replication of the total chain exhibits not large departures from model of independent replications of separate nucleotides. Therefore, the question arises: what is the maximum length of a poly-nucleotide strand that will not (almost surely, i.e. with probability one) become extinct. This value is dependent on the mutation rate per nucleotide, and on ability to survive and perform reproduction.

Let us consider the RNA-species with the RNA chain length equal to λ. Assume also the mutation rate per nucleotide as equal to μ. Then the probability that the single nucleotide is copied without error is equal to $p = 1 - \mu$. In a model of independent replications of nucleotides, the resulting probability of replication of the whole strand P is equal to p^λ. There are three situations S_0, S_1 and S_2 yielding 0, 1, and 2 individuals respectively in the next generation. S_0 denotes the situation when (with the probability 1-w) the individual does not survive to replication stage (next generation). S_1 denotes the situation when (with the probability $w(1- p^\lambda)$) the individual does survive to next generation, but produce descendant with the error. Finally, S_s denotes the situation when (with the probability wp^λ) the individual does survive and replicates without error.

Let us assume, that the population of error-free RNAs follows the Galton-Watson branching process where Z_t denotes the number of individuals in generation t and the number of progeny of each individual is denoted by random variable X. Such a process is said to be supercritical when the probability of eventual extinction $q < 1$. This happens only when $E(X) > 1$ (even if $\lim_{t\to\infty} E(Z_t) = 1$ when $E(X) = 1$, counterintuitively, the probability of eventual extinction $q = 1$ in this critical case). In our study the $E(X) = w\,(1+p^\lambda)$, so it must hold: $w(1+p^\lambda) > 1$. The last statement is satisfied only if $p^\lambda > (1 - w) / w$, or (since both sides of inequality are less than 1) $\lambda < [ln(1\text{-}w) - ln(w)] / ln(p)$.

In the Fig. 1 the 3D plot of the border function is presented. We can see that even for surviving probability $w = 0.7$ and replication probability of a single nucleotide $p = 0.999$ (that is mutation rate $\mu = 0.001$ as given in [1]), the RNA sequence length λ cannot be larger than 800 nucleotides. On the other hand it is hard to imagine the RNA-species of the genome length less than 800 nucleotides, capable for replication with such accuracy. Yet, if its length is larger, then such species will eventually extinct with probability 1.

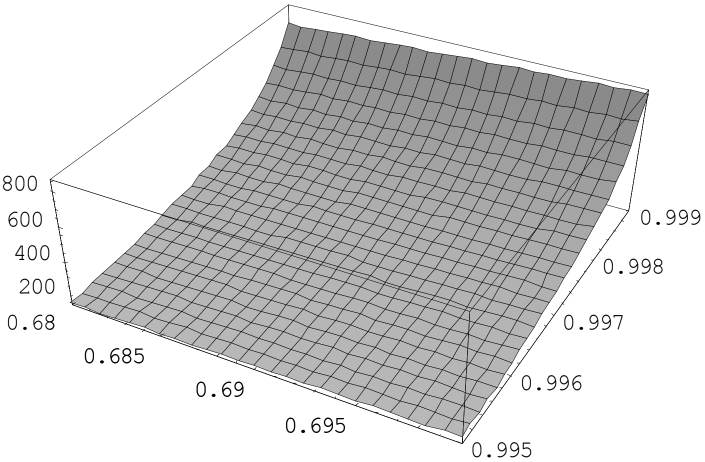

Fig.1 The plot of the function $\lambda = [ln(1\text{-}w) - ln(w)] / ln(p)$ for w changing from 0.68 to 0.7 and p changing from 0.995 to 0.999.

If we assume that $\lambda < [ln(1\text{-}w) - ln(w)] / ln(p)$, then the probability of non-extinction $q = 1 - (1 - w) / wp^\lambda$. In the paper we present simulations of Galton-Watson branching process, representing the evolution of considered RNA-species showing that even if above inequality if fulfilled, the extinction often happens (see Fig. 2). Therefore, the paper can help to grasp the values of parameters p, w, and the complexity threshold λ, feasible from biological perspective concerning the early Life, and in particular the moment, when *it all started* and (according to RNA-world hypothesis) first RNA-species of the length λ nucleotides began its self-replication.

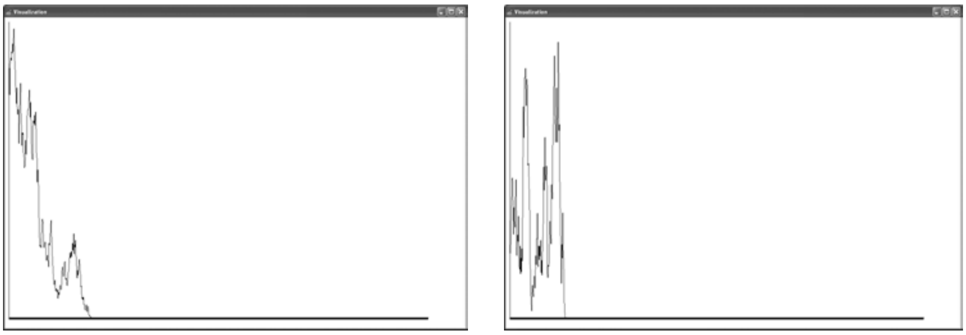

Fig. 2 Two extinct realizations of an evolution based on a slightly supercritical Galton-Watson branching process with $E(X)$=1.000184.

Acknowledgments. The authors would like to acknowledge the financial support of this work under the habilitation grant number BW/RGH-5/Rau-0/07 and under statutory activities BK2007 at Silesian University of Technology. One of the authors would like also to thank to Prof. Marek Kimmel from Department of Statistics at Rice University in Houston, TX, USA for long discussions and advise concerning the branching processes and their applications in biology and genetics.

References

1 Smith, J.M., Szathmary, E.: The Origins of Life. From the Birth of Life to the Origin of Language. Oxford University Press (1999)

2 Dyson, F.: Origins of Life. Revised Edition. Cambridge University Press (1999)

3 Demetrius, L., Schuster, P., Sigmund, K.: Polynucleotide Evolution and Branching Processes, Bull. Math. Biol. 47 (1985) 239-262

4 Kimmel, M., Axelrod, D.: Branching Processes in Biology. Springer-Verlag, New-York (2002)

5 Marsaglia, G.: Monkey Tests for Random Number Generators. Comput. Math. Appl. 9 (1993) 1-10

6 Marsaglia, G., Zaman, A., Tsang, W.W.: Toward a Universal Random Number Generator. Stat. Prob. Lett. 8 (1990) 35-39

7 Cyran, K.A., Kimmel, M.: Interactions of Neanderthals and Modern Humans: What Can Be Inferred from Mitochondrial DNA?. Math. Biosci. Eng. 2 (2005) 487-498

8 Cyran, K.A., Polańska, J., Kimmel, M.: Testing for Signatures of Natural Selection at Molecular Genes Level. Journal of Medical Informatics and Technologies 8 (2004) 31-39

9 K. Cyran, Rough sets in the interpretation of statistical tests outcomes for genes under hypothetical balancing selection, Lecture Notes in Artificial Intelligence, 716-725, 2007.