Available online at www.sciencedirect.com

Future Generation Computer Systems 21 (2005) 1033–1039

www.elsevier.com/locate/fgcs

Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution

M. Dudkiewicz[a], P. Mackiewicz[a], A. Nowicka[a], M. Kowalczuk[a], D. Mackiewicz[a], N. Polak[a], K. Smolarczyk[a], J. Banaszak[a], M.R. Dudek[b], S. Cebrat[a,*]

[a] *Institute of Genetics and Microbiology, University of Wrocław, ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland*
[b] *Institute of Physics, University of Zielona Góra, ul. Szafrana 4A, 65-516 Zielona Góra, Poland*

Available online 2 April 2004

Abstract

There are three main elements deciding about the effect of mutations on the protein coding sequences—the type of the substitution of nucleotide, the selection for the function of the gene product and the nature of the genetic code itself. Selection used to be considered as the only directional process among the evolutionary mechanisms. In fact the mutational pressure is also "directional" which means that the rates of particular nucleotide substitutions tend to produce a DNA molecule with a specific nucleotide composition. Using Monte Carlo simulations we have shown that the genetic code plays the central role in buffering the effect of mutations and that all three elements are optimised in the generation of the genetic diversity in such a way that deleterious effects of mutations are substantially reduced.

Keywords: Mutation; Selection; Evolution; Genetic code; Monte Carlo simulation

1. Introduction

Genetic information is coded in the nucleotide sequences of the double-stranded DNA molecules. The sequence of nucleotides in one DNA strand corresponds to the sequence of RNA which is the "messenger" of information in the process of its translation into amino acid sequences of proteins. This strand is co-linear to RNA and it is organised in codons—tri-nucleotide sequences corresponding to amino acids. There are 64 codons of which 61 code for 20 amino acids and three for stop translation signals. Because there are much more different codons than coded amino acids, the genetic code is called degenerate. Since the discovery of the genetic code, the way how it is degenerated is one of the most fascinating problems of genetics. Is it the best of all possible codes?

The hypotheses trying to explain the evolution of the genetic code can be divided into two groups (see, for review of hypotheses on the origin of the genetic code, Refs. [10,16,17])—one called mechanistic, assuming structural and physicochemical relationships between codons or anticodons and amino acids [5,23,27,28].

[*] Corresponding author. Tel.: +48 71 3756303; fax: +48 71 3252151.

E-mail addresses: mdudek@proton.if.uz.zgora.pl (M.R. Dudek), cebrat@microb.uni.wroc.pl (S. Cebrat).
URL: http://smORFland.microb.uni.wroc.pl.

0167-739X/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.future.2004.03.003

The other group of hypotheses, called stochastic, assumes that initially the codon assignment could vary and it was the selection pressure exerted on the early organisms which left the code optimal from the selection point of view [1,2,8,9,11,24,27,28]. The optimisation of the genetic code was based on reducing the harmful effects of the mutations. It is very likely that organisms which reduced the deleterious effects of mutations won the competition. That is why not only the degeneration of the genetic code is important but also the way how it is degenerated. During further evolution connected with an increase of genomes the genetic code was frozen—it was not possible to re-interpret the meaning of any codon in a large genome [3]. Some small corrections were possible in the small genomes giving "dialects" of the genetic code known as deviations from the universal code [6,12–14,22,26].

Advocating stochastic hypotheses, we expect that if the genetic code was "frozen" at the early stage of evolution when the genomes were rather small, it is the code itself which dictates the parameters of directional mutational pressure which have to cooperate with the selection pressure to minimise the deleterious effects of substitutions in contemporary genomes. There are two premises which indicate that in fact both the selection and the mutational pressures are fitted to the genetic code—one that the prevalence of the amino acids in the proteins is correlated with the number of codons representing a given amino acid in the genetic code stated by King and Jukes [15] and confirmed nowadays by huge amount of genomic data, and the second that the most "mutable" codons in the genome correspond to the least-represented amino acids and that these amino acids are the most watched by selection [21]. In this paper we show directly, using Monte Carlo simulations, the changing parameters of any of the three counterparts of the coding functions: relative substitution probabilities of the directional mutational pressure, the way how the genetic code is degenerated or the amino acid composition of proteins increases the deleterious effects of mutations.

2. Materials and methods

All simulations were performed on DNA sequences of the *Borrelia burgdorferi* genome [7] downloaded from ftp://www.ncbi.nlm.nih.gov.

Table 1
Frequencies of substitutions in the leading strand of the *B. burgdorferi* genome

From	To			
	A	T	G	C
A	–	0.103	0.067	0.023
T	0.065	–	0.035	0.035
G	0.164	0.116	–	0.015
C	0.070	0.261	0.047	–

The parameters of the replication-associated directional mutational pressure for the *B. burgdorferi* genome have been used as described in the substitution matrices (Tables 1 and 2) [18–20]. Note that the substitution matrix describing the mutational pressure for the leading DNA strand is a mirror one for the matrix describing the mutational pressure for the complementary lagging DNA strand.

The protein coding sequences of the *B. burgdorferi* genome were divided for two classes: (1) lying on the leading strand (564 sequences of the total length of 560 550 nucleotides) and (2) lying on the lagging strand (286 sequences of the total length of 291 933 nucleotides). If the genes from the leading strand are under mutational pressure characteristic for them it means that the sense strands of these genes (collinear to RNA) is under the mutational pressure for the leading DNA strand. In one Monte Carlo step (MCS) each nucleotide of the sequence is drawn with a probability $P_{\mathrm{mut}} = 0.01$ and then substituted by another nucleotide with the probability described by the corresponding value in the substitution matrix. Then, all codon substitutions and corresponding amino acid substitutions introduced into the coded proteins, resulting from the nucleotide substitutions, are counted.

We have performed simulations using the simple model of selection for the global amino acid compo-

Table 2
Frequencies of substitutions in the lagging strand of the *B. burgdorferi* genome

From	To			
	A	T	G	C
A	–	0.065	0.035	0.035
T	0.103	–	0.023	0.067
G	0.261	0.070	–	0.047
C	0.116	0.164	0.015	–

sition of gene products as described by Dudkiewicz et al. [4]. When the viability of coding sequences was studied, the selection parameter tolerance (*T*) for the amino acid composition of individual sequences was introduced. It describes the maximum allowed deviation in the amino acid composition of the protein coded by a given gene. It is expressed by the sum of absolute values of differences between fractions of amino acids as follows:

$$T = \sum_{i}^{20} |f_i^0 - f_i^t|$$

where f^0 is the fraction of a given amino acid in the original sequence (before mutations) and f^t is the fraction of a given amino acid in the sequence after mutations.

Arbitrarily, we have assumed as the value of tolerance for amino acid composition $T = 0.3$, the average difference in fractions of amino acids between 442 pairs of orthologs belonging to two related genomes: *B. burgdorferi* and *Treponema pallidum*. Orthologs are sequences from different species which evolved by vertical descent and are usually responsible for the same function in different organisms. These orthologs were extracted from COGs database [25] downloaded from ftp://www.ncbi.nlm.nih.gov/pub/COG. If the number of substituted amino acids in a given gene product overpasses the declared tolerance $T = 0.3$, the coded sequence is “killed” and replaced by the corresponding one from the parallel evolving genomic sequence. During the simulation we have counted the number of genes eliminated by selection and the number of accumulated substitutions in the surviving genes.

3. Results of the gene evolution simulation

The results of all simulations were compared with the so-called standard simulations in which we used nucleotide substitution matrices as described in the methods section, the standard genetic code and the sequences of protein coding sequences of the real *B. burgdorferi* genome, separately for those located on the leading DNA strand and the lagging DNA strand.

In the first stage we estimated the importance of the genetic code degeneracy. The best measure of that is the comparison between the number of substituted codons and the number of substituted amino acids. The results are shown in Fig. 1. It is trivial that each substitution of a nucleotide causes a substitution of a codon but it could happen that the mutation is silent—the sense of the new codon is the same as before the mutation. As an effect, the number of codon substitutions is higher than the number of amino acid substitutions.

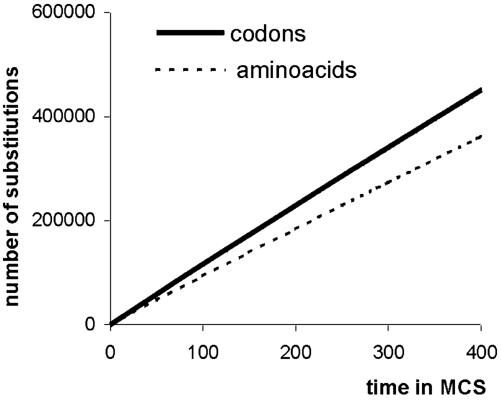

Fig. 1. Accumulation of codon and amino acid substitutions in protein coding sequences located in the leading strand of the *B. burgdorferi* genome. *Y*-axis scaled in the accumulated number of substitutions (codon or amino acid). There was no selection pressure imposed during the simulation and the multiple substitutions and reversions were also counted ($P_{\mathrm{mut}} = 0.01$ in these simulations).

However, the important feature of the genetic code degeneracy is the way how it is degenerated. To show that, we have produced the other version of the genetic code keeping the same level of degeneracy of the codon positions, replacing one kind of the amino acid by another one with the same level of degeneracy, i.e. since valine is coded by four codons it can be replaced for example by threonine which is also coded by four codons. After such a transformation of the genetic code, all amino acid sequences coded by the *B. burgdorferi* genome were re-translated into new nucleotide sequences, keeping the same codon usage as for the original *B. burgdorferi* genome which means (for the above example) that for each position of threonine new codons are drawn with the same relative probability as they occurred in the set of valine codons in the real *B. burgdorferi* genome. Note that after this genome transformation codons have changed their meanings but the amino acid composition and the level of coding degeneracy for particular amino acids have not changed. Simulations of evolution of

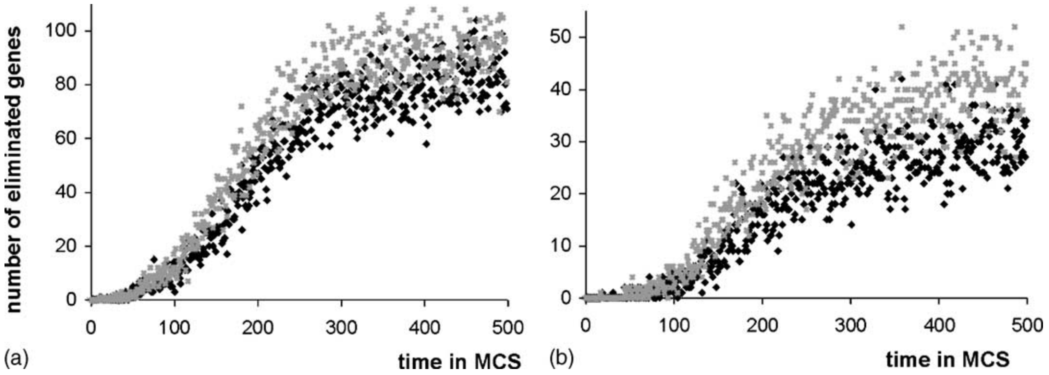

Fig. 2. Elimination of genes after re-definition of the codon meanings: (a) genes from the leading DNA strand; (b) genes from the lagging DNA strand. Black dots for standard simulation and grey crosses for the one with the changed code. The amino acid composition of gene products was as those coded by the *B. burgdorferi* genome, but the meaning of codons was changed in such a way that the level of degeneracy of the code for each amino acid stayed unchanged. See text for details.

such genomes were performed under the original mutational pressure and selection as described in Section 2. The results in Fig. 2 show that the sensitivity of the coding sequences to the mutational pressure measured by the rate of gene elimination was much higher than in the standard simulation. It means that not only the level of degeneracy is important but also "how" the code is degenerated. This observation is true for coding sequences from both the leading and the lagging DNA strands. Furthermore, divergence measured by the accumulation of amino acid substitutions of such sequences stays roughly at the same level as for real sequences (Fig. 3). Thus, it could be concluded that the cost of selection pressure under the changed genetic code was higher for producing roughly the same level of variation it has to eliminate much more genes from the pool.

To check how the structure of the substitution matrix influences gene elimination rate we used a random mutational matrix in simulations (normalised as the original substitution matrices used in other simulations). In this version of simulations the genetic code and the selection pressure were the same as in standard simulations, but the preferences in the nucleotide substitutions were changed for uniform substitution rates for all nucleotides—so-called one-parameter substitution matrix. In Fig. 4 we have shown the number of accumulated substitutions in the coding sequences and the

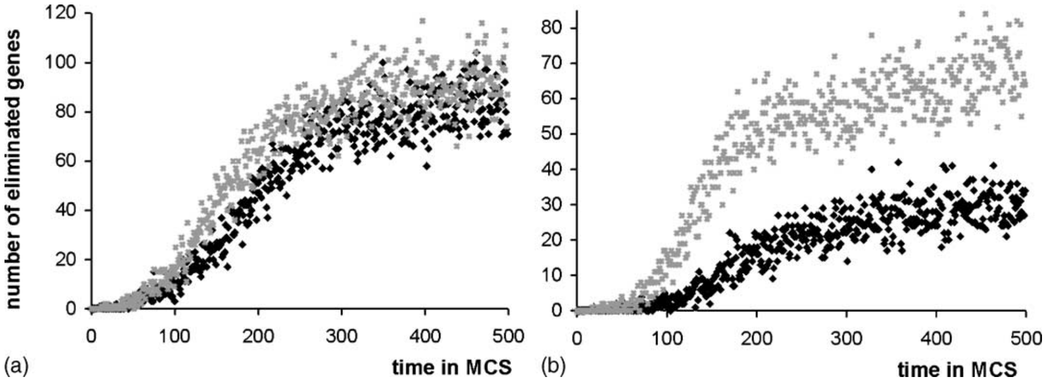

Fig. 3. The elimination rate of genes under mutational pressure with the uniform substitution rates for all nucleotides (one-parameter substitution matrix): (a) genes from the leading DNA strand; (b) genes from the lagging DNA strand. Black dots for standard simulation and grey crosses for the one with one-parameter substitution matrix. The amino acid composition of gene products was as in the *B. burgdorferi* genome, and the standard genetic code was used for simulations.

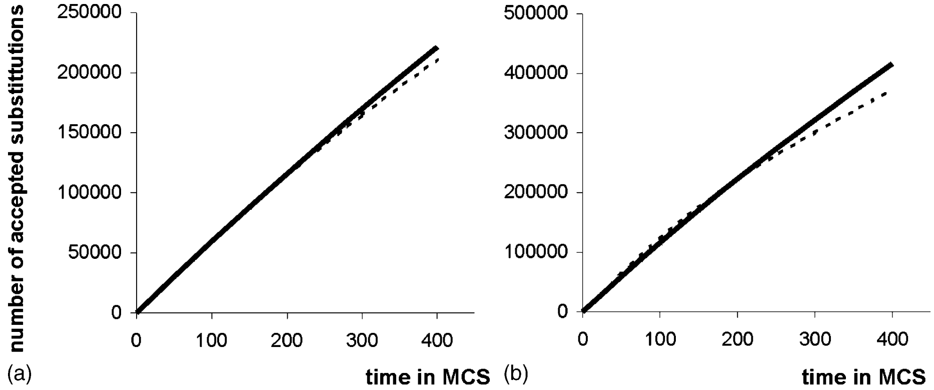

Fig. 4. The accumulation of amino acid substitutions in the products of the surviving genes for the simulations described in Fig. 3: (a) genes from the leading DNA strand; (b) genes from the lagging DNA strand. Solid line is for standard simulation and dotted one for simulation with one parameter substitution matrix. The *Y*-axis shows the accumulated number of accepted mutations which did not kill the coding sequence, since multiple substitutions and reversions are also counted the numbers do not show directly the divergence of sequences.

elimination rate of genes. The killing effect of such mutational pressure was much higher than for the standard simulation. These results suggest that the nucleotide composition of genes, the universal genetic code and the mutational pressure are optimised to give relatively low effect of gene elimination.

There is another counterpart producing the effect of mutagenesis—the amino acid composition of proteins coded by the genome. This composition is a result of selection choosing among the products of mutations. To check the effect of the global amino acid composition on the gene elimination we have constructed a "virtual genome" of *B. burgdorferi* which coded for a random sequence of amino acids according to the uniform distribution in the coded sequences (statistically the same fraction of each amino acid), the same codon preferences for the amino acids and the same distribution of the length of open-reading frames as in the natural genome. Simulations were performed under the *B. burgdorferi* mutational pressure using the universal genetic code. The results of simulations are shown in Fig. 5. Again the deleterious effects of substitutions increased significantly. This is in accordance with our previous finding that the frequencies of amino acids in proteins are negatively correlated with their stability under the mutational pressure. The most

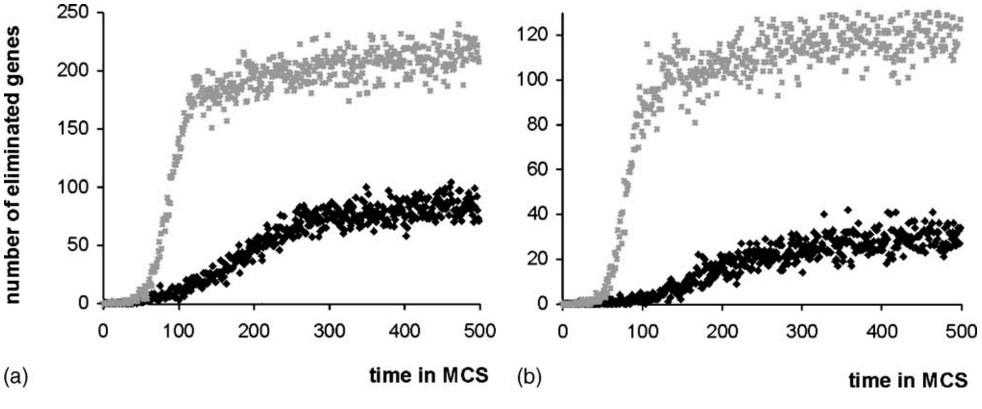

Fig. 5. The elimination rate of genes coding proteins with uniform amino acid composition: (a) genes from the leading DNA strand; (b) genes from the lagging DNA strand. Black dots for standard simulation and grey crosses for uniform amino acid composition. In the simulation the standard genetic code was used and the length distribution of open-reading frames and nucleotide substitution matrices were as in *B. burgdorferi* genome.

mutable amino acids—cysteine and tryptophan—are the least-frequent amino acids in proteins and the most conserved in their positions—the best watched by selection [21]. This is the same rule which concerns the frequency of nucleotides—the least-frequent nucleotide in the sequence in equilibrium with the mutational pressure is the least stable and undergoes the most frequent substitutions [18–20].

It is obvious that the selection parameters introduced into the model are simple and do not correspond accurately to all the parameters important in the estimation of the biological activity of gene products. Nevertheless, taking into account that the selection forces are distributed along the single coding sequences and the whole chromosomes very unevenly, at present it is impossible to model them at the genomic level of studies. However, some trials at the level of single genes (proteins) are possible. But such trials on single sequences cannot address the general question of the genetic code property.

These preliminary results indicate that there is a peculiar symmetry between the directional mutational pressure and the selection pressure. The “axis” of the symmetry is the universal genetic code, probably the oldest being existing on Earth. Knowing the relations between the directional mutational pressure exerted on the nucleotide sequences and the selection pressure directly eliminating genomes with the deleterious effects of substitutions in the gene products would make it possible to predict some evolution events and better understand evolutionary processes.

References

[1] C. Alff-Steinberger, The genetic code and error transmission, Proc. Natl. Acad. Sci. U.S.A. 64 (1969) 584–591.

[2] D.H. Ardell, On error minimization in a sequential origin of the standard genetic code, J. Mol. Evol. 47 (1998) 1–13.

[3] F.H. Crick, The origin of the genetic code, J. Mol. Biol. 38 (1968) 367–379.

[4] M. Dudkiewicz, P. Mackiewicz, A. Nowicka, M. Kowalczuk, D. Mackiewicz, N. Polak, K. Smolarczyk, M.R. Dudek, S. Cebrat, Properties of genetic code under directional, asymmetric mutational pressure, in: P.M.A. Sloot, et al. (Ed.), Proc. ICCS 2003. Lecture Notes in Computer Science, vol. 2657, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 343–350.

[5] P. Dunnill, Triplet nucleotide-amino-acid pairing a stereochemical basis for the division between protein and non-protein amino-acids, Nature 210 (1966) 1265–1267.

[6] M. Ehara, Y. Hayashi-Ishimaru, Y. Inagaki, T. Ohama, Use of a deviant mitochondrial genetic code in yellow-green algae as a landmark for segregating members within the phylum, J. Mol. Evol. 45 (1997) 119–124.

[7] C.M. Fraser, S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey, et al., Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*, Nature 390 (1997) 580–586.

[8] S.J. Freeland, L.D. Hurst, The genetic code is one in a million, J. Mol. Evol. 47 (1998) 238–248.

[9] S.J. Freeland, R.D. Knight, L.F. Landweber, L.D. Hurst, Early fixation of an optimal genetic code, Mol. Biol. Evol. 17 (2000) 511–518.

[10] Di.M. Giulio, On the origin of the genetic code, J. Theor. Biol. 187 (1997) 573–581.

[11] D. Haig, L.D. Hurst, A quantitative measure of error minimization in the genetic code, J. Mol. Evol. 33 (1991) 412–417 [Erratum in J. Mol. Evol. 49 (1999) 708].

[12] Y. Hayashi-Ishimaru, M. Ehara, Y. Inagaki, T. Ohama, A deviant mitochondrial genetic code in prymnesiophytes (yellow-algae): UGA codon for tryptophan, Curr. Genet. 32 (1997) 296–299.

[13] Y. Hayashi-Ishimaru, T. Ohama, Y. Kawatsu, K. Nakamura, S. Osawa, UAG is a sense codon in several chlorophycean mitochondria, Curr. Genet. 30 (1996) 29–33.

[14] P.J. Keeling, W.F. Doolittle, Widespread and ancient distribution of a noncanonical genetic code in diplomonads, Mol. Biol. Evol. 14 (1997) 895–901.

[15] J.L. King, T.H. Jukes, Non-Darwinian evolution, Science 164 (1969) 788–797.

[16] R.D. Knight, S.J. Freeland, L.F. Landweber, Selection, Trends Biochem. Sci. 24 (1999) 241–247.

[17] R.D. Knight, The origin and evolution of the genetic code: statistical and experimental investigations, Ph.D. Thesis, Department of Ecology and Evolutionary Biology, Princeton University, 2001.

[18] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, High correlation between the turnover of nucleotides under mutational pressure and the DNA composition, BMC Evol. Biol. 1 (2001) 13, http://www.biomedcentral.com/1471-2148/1/13.

[19] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, Multiple base substitution corrections in DNA sequence evolution, Int. J. Modern Phys. C 12 (2001) 1043–1053.

[20] P. Mackiewicz, M. Kowalczuk, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, A. Laszkiewicz, M.R. Dudek, S. Cebrat, Replication associated mutational pressure generating long-range correlation in DNA, Physica A: Stat. Mech. Appl. 314 (2002) 646–654.

[21] A. Nowicka, P. Mackiewicz, M. Dudkiewicz, D. Mackiewicz, M. Kowalczuk, S. Cebrat, M.R. Dudek, Correlation between mutation pressure, selection pressure, in: P.M.A. Sloot, et al. (Eds.), Proc. ICCS 2003. Lecture Notes in Computer Science, vol. 2658, Springer-Verlag, Berlin, Heidelberg, 2003, pp. 650–657.

[22] S. Osawa, Evolution of the Genetic Code, Oxford University Press, Oxford, 1995.

[23] S.R. Pelc, M.G. Welton, Stereochemical relationship between coding triplets and amino-acids, Nature 209 (1966) 868–872.

[24] T.M. Sonneborn, Degeneracy in the genetic code: extent, nature and genetic implications, in: V. Bryson, H.J. Vogel (Eds.), Evolving Genes and Proteins, Academic Press, New York, 1965, pp. 297–377.

[25] R.L. Tatusov, D.A. Natale, I.V. Garkavtsev, T.A. Tatusova, U.T. Shankavaram, B.S. Rao, B. Kiryutin, M.Y. Galperin, N.D. Fedorova, E.V. Koonin, The COG database: new developments in phylogenetic classification of proteins from complete genomes, Nucleic Acids Res. 29 (2001) 22–28.

[26] A.B. Tourancheau, N. Tsao, L.A. Klobutcher, R.E. Pearlman, A. Adoutte, Genetic code deviations in the ciliates: evidence for multiple and independent events, EMBO J. 14 (1995) 3262–3267.

[27] C.R. Woese, D.H. Dugre, W.C. Saxinger, S.A. Dugre, The molecular basis for the genetic code, Proc. Natl. Acad. Sci. U.S.A. 55 (1966) 966–974.

[28] C.R. Woese, The Genetic Code: The Molecular Basis for Genetic Expression, Harper & Row, New York, 1967.

Małgorzata Dudkiewicz, Joanna Banaszak, Natalia Polak, Aleksandra Nowicka, Kamila Smolarczyk, Dorota & **Daria** & **Paweł Mackiewicz, Maria Kowalczuk, Stanisław Cebrat and Mirosław R. Dudek**. Małgorzata, Joanna, Natalia and Kamila are PhD students. Aleksandra, Dorota, Paweł and Maria are postdocs. Mirosław, professor of Zielona Góra University, and Stanisław, professor, playing the role of PapaSmORF. Daria (the smallest one in the centre) is the Future Generation of Computing Scientists (FGCS). All the members of the group but Mirosław are geneticists, Mirosław is the only physicist. The whole group is working on the genome analyses and simulations of the gene and genome evolution, basing on the real DNA sequences. Our previous works were connected with the recognition of the protein coding sequences, genome coding capacity and estimation of the total number of coding ORFs in genomes, including Small Open-Reading Frames (SmORFs). Some of us are also simulating the population dynamics—age structured populations without the internal structure of genes in the individual genomes. Recently, in co-operation with the group of Dietrich Stauffer from Cologne University we are introducing the gene structure into these coarse models and we plan to meet our models of the real gene and genome evolution in the near future. More information about our group and papers can be found at our web site: http://SmORFland.microb.uni.wroc.pl.