Properties of the Genetic Code under Directional, Asymmetric Mutational Pressure

Małgorzata Dudkiewicz[1], Paweł Mackiewicz[1], Aleksandra Nowicka[1], Maria Kowalczuk[1], Dorota Mackiewicz[1], Natalia Polak[1], Kamila Smolarczyk[1], Mirosław R. Dudek[2], and Stanisław Cebrat[1⋆]

[1] Department of Genetics, Institute of Microbiology, University of Wroclaw, ul. Przybyszewskiego 63/77, PL-54148 Wroclaw, Poland
{malgosia, pamac, nowicka, kowal, dorota, polak, smolar, cebrat}@microb.uni.wroc.pl
http://smORFland.microb.uni.wroc.pl

[2] Institute of Physics, University of Zielona Góra, ul. A. Szafrana 4a, PL-65516 Zielona Góra, Poland
mdudek@proton.if.uz.zgora.pl

Abstract. We have used the Monte Carlo method for simulating the evolution of protein coding sequences from the *Borrelia burgdorferi* genome under the directional mutational pressure, described by the nucleotide substitution matrix experimentally found for this genome. Since the mutational pressure is asymmetric – different for each of the two DNA strands, and the coding sequences are also asymmetric, the mutagenic effect depends on the topology of the coding sequence on the chromosome. While the direct effect of the directional mutational pressure on the codon usage is predictable, the effect on the amino-acid composition depends on the degeneracy of the genetic code, initial composition of the protein sequence and codon usage. Assuming additional degeneracy of information connected with the structure of amino-acids, we have found that the best strategy for the evolution of some genes in the *B. burgdorferi* genome is to change the mutational pressure by inversions, which corresponds to changing the substitution matrix to the mirror one. It mimics the behavior of a stock market player who can gain by investing only in two stock sets even if during that time both sets have lost.

Introduction

The DNA molecule is composed of two anti-parallel sequences (strands) of four different nucleotides: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). The complementarity rule predicts that A in one strand corresponds to T in the other one and G corresponds to C [1]. As a consequence, the number of A in the whole DNA molecule equals the number of T and the number of G equals the number of C [2]. These rules, called parity rules 1 ($PR1$) are deterministic. One could predict that for a random DNA molecule the same rules, but for each of

⋆ To whom all correspondence should be sent.

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2657, pp. 343–350, 2003.

the two DNA strands, should be in force. These parity rules for single strands are stochastic rather than deterministic and are called parity rules 2 – $PR2$ [3], [4]. In the natural DNA sequences there are many mechanisms which introduce deviations from $PR2$ and sequences with such deviations are called asymmetric. The most important mechanisms introducing asymmetry into the DNA molecule are the replication-associated directional mutational pressure (see for review: [5], [6], [7] and selection for a proper amino-acid composition of proteins coded by genes [8], [9]. Due to the degeneracy of the genetic code, the same amino-acid can be coded by many (up to six) different codons – tri-nucleotide sequences. It means that DNA sequences with different nucleotide composition can code for the same protein. Thus, a coding nucleotide sequence under directional mutational pressure could adapt to this pressure by using preferred or more stable codons without any change in the amino-acid composition. The interesting question is how the degeneracy of the genetic code ameliorates the viability of coding sequences and, assuming a declared tolerance for amino-acid composition – what could be the best strategy for genes. Genes are pieces of DNA composed of two complementary strands, one strand is called coding (or sense) sequence and it is usually richer in purines (A and G) while the other – anti-sense strand is relatively rich in pyrimidines (T and C) [10], [11], [12], [13]. Since the composition of the two DNA strands is different [14], [15], [16], [17], [18], it is important if the sense strand of a gene lies on the leading DNA strand replicating continuously or on the lagging DNA strand replicating discontinuously through intermediate Okazaki fragments. Knowing the real mutational pressure exerted on each DNA strand of the *B. burgdorferi* genome described by the matrix of nucleotide substitutions as well as sequences and positions of all genes in the *B. burgdorferi* genome, we have used Monte Carlo simulations for the estimation the of susceptibility of coding sequences to the mutational pressure.

1 Material and Methods

All simulations were performed with DNA sequences of the *B. burgdorferi* genome downloaded from *www.ncbi.nlm.nih.gov*. The replication associated directional mutational pressure for the *B. burgdorferi* genome has been described in the so called substitution matrices by [19], [20], [21]. Note that the substitution matrix describing the mutational pressure for one strand i.e. the leading DNA strand has a mirror matrix describing the mutational pressure for the complementary – the lagging DNA strand. In this paper we have performed our studies with 292 genes lying on the leading strand of the *B. burgdorferi* genome of the total length of 309024 nucleotides. If the genes from the leading strand are under the mutational pressure characteristic for them it means that the sense strands of these genes are under the mutational pressure for the leading DNA strand. If such a gene is inverted, it means that its anti-sense strand is under the mutational pressure characteristic for the leading strand. In one Monte Carlo Step (MCS) each nucleotide of the sequence is drawn with a probability $p_{mut} = 0.001$ and then substituted by another nucleotide with the probability described by

the corresponding parameter in the substitution matrix. Then, all nucleotide substitutions, codon substitutions and corresponding amino-acid substitutions introduced during simulation into the coded proteins are counted. When the viability of coding sequences is studied, two genomic sequences (composed of 292 genes), originally identical, are consecutively mutated. Moreover, 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 a protein coded by a given gene after mutations in comparison to its original sequence (occurring in the genome). It is expressed by the sum of absolute values of differences between fractions of codons or amino acids as follows:

$$\sum_{i=1}^{20} |f_i(0) - f_i(t)|, \tag{1}$$

where: $f_i(0)$ is a fraction of a given amino-acid in the original sequence (before mutations) and $f_i(t)$ is a fraction of a given amino acid in the sequence after mutations in t MCS. Arbitrarily, we have assumed as a value of tolerance for amino-acid composition the average distance between 442 pairs of orthologs belonging to two related genomes: *B. burgdorferi* and *Treponema pallidum*. These orthologs were extracted from COGs databse downloaded from *ftp://www.ncbi.nlm.nih.gov/pub/COG* in September 2001. COGs contain protein sequences which are supposed to have evolved from one ancestral protein [22]. Orthologs are sequences from different species which have evolved by vertical descent and are usually responsible for the same function in different organisms. If the number of substituted elements in a given gene overpasses the declared tolerance $T = 0.3$, the coded sequence is "killed" and replaced by the corresponding one from the second genomic sequence.

2 Results and Discussion

Directional mutational pressures found for many prokaryotic as well as eukaryotic genomes generate DNA sequences in equilibrium whose fractions of nucleotides are highly correlated with the turnover rate of nucleotides [19]. It means that if any DNA sequence stays under a stable directional mutational pressure for long enough time, its nucleotide composition reflects the mutational pressure described in a relation:

$$F_a \sim P_{mut} t_a + const, \tag{2}$$

where: F_a – fraction of a nucleotide in the DNA strand, t_a – half time of survival of a given nucleotide. The DNA sequence which is not in equilibrium with the mutational pressure and which is richer in nucleotides with higher turnover rate than the sequence in equilibrium (the case of coding sequences in many genomes) would tend to reach the equilibrium if there are no selection forces. Furthermore, during this process the total mutability of the sequence decreases. It is trivial, since nucleotides with higher substitution probability are replaced by nucleotides

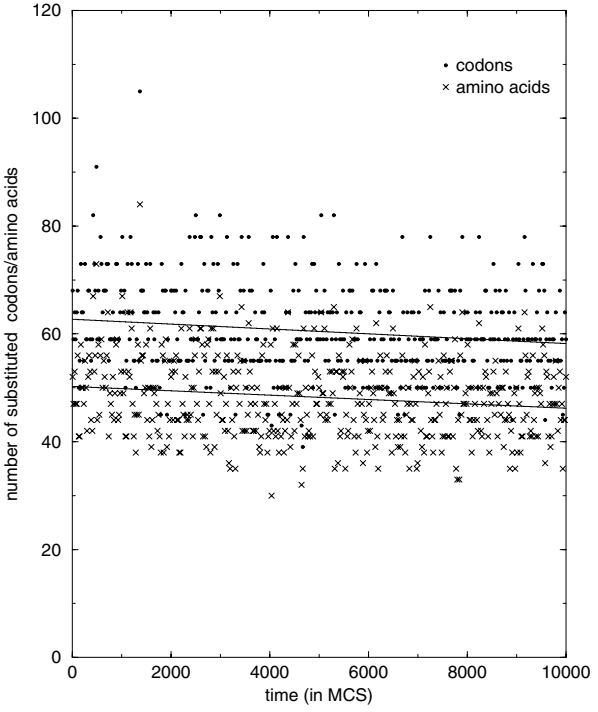

Fig. 1. Time dependence of the number of substituted codons and amino acids under a stable mutatational pressure without selection.

with lower substitution probability. This effect directly translates into the substitution rates of codons – each nucleotide substitution simultaneously leads to the change of codon. But, due to the degeneracy of the genetic code, there are many codons which code for the same amino-acid. Thus, not every substitution of nucleotide in the codon changes the sense of this codon – not every mutation is a miss-sense mutation. The effect of the evolution of the composition of DNA sequences towards equilibrium and the effect of the degeneracy of the genetic code is seen in Fig.1. Comparing different genomes it is very easy to find genes which code for proteins fulfilling the same functions in different organisms. Furthermore, these genes very often share the same amino-acid residues at the corresponding positions which seems to prove that they have a common ancestor sequence. Such genes are called orthologs. In fact, the degree of homology between two orthologs could be low (much lower than 0.5). This means that there is another level of degeneracy in the coding genetic information. Not only different codons can code for the same amino-acid, but it is possible to not disturb the protein function by placing different amino-acids at the same position in the protein sequences. We will call this effect "tolerance" in the next simulations. We have assumed the tolerance 0.3, meaning that the maximum overall composition

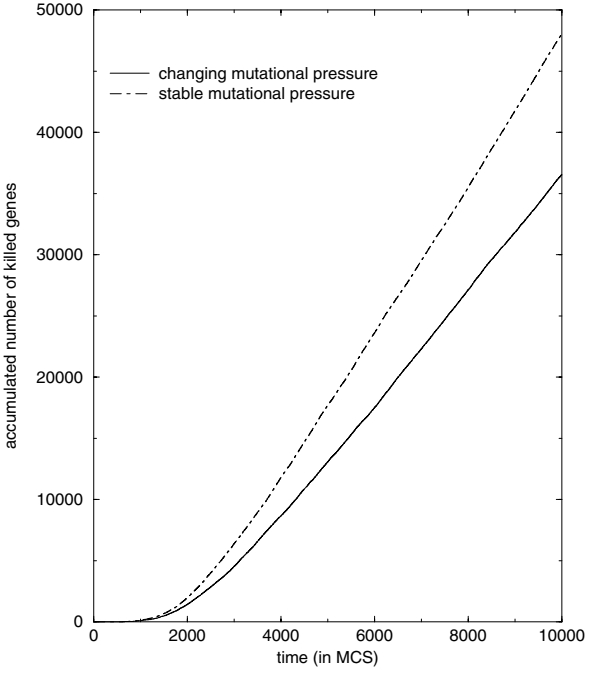

Fig. 2. The accumulated number of killed genes (since the start of simulation) in the case of the same mutational pressure in each Monte Carlo step t, and in the case when the mutational pressure is switching periodically in time.

of a protein can be changed due to the accumulated mutations in the coding sequence by this factor, as described in the Method section. This tolerance corresponds to the average difference between orthologs belonging to two related genomes: *B. burgdorferi* and *Treponema pallidum*. If during the simulation of mutagenesis the fraction of replaced amino-acids in the coded sequence exceeds the tolerance the gene is replaced by its copy from the parallely simulated genome. Note that this copy could have also many mutations accumulated during the previous MCSs of simulations. The dynamics of elimination of genes and their replacement is shown in Fig.2. The solid bold line represents the accumulated number of replaced sequences of genes located originally in the genome on the leading DNA strand. At the beginning of the simulation there is no killing effect because we have assumed that all genes stay at their maximally stable composition. As the simulation proceeds, genes accumulate mutations and eventually surpass the declared tolerance. Then they are replaced by corresponding copies. Since the replacement copies already have some accumulated substitutions, the probability of their elimination in the next MCSs is much higher than the initial probability, that is why the elimination rate increases with time reaching the constant value after a few hundred of MCSs. In fact, one Monte Carlo Step in our simulation corresponds to about 100000 generations in Nature (the muta-

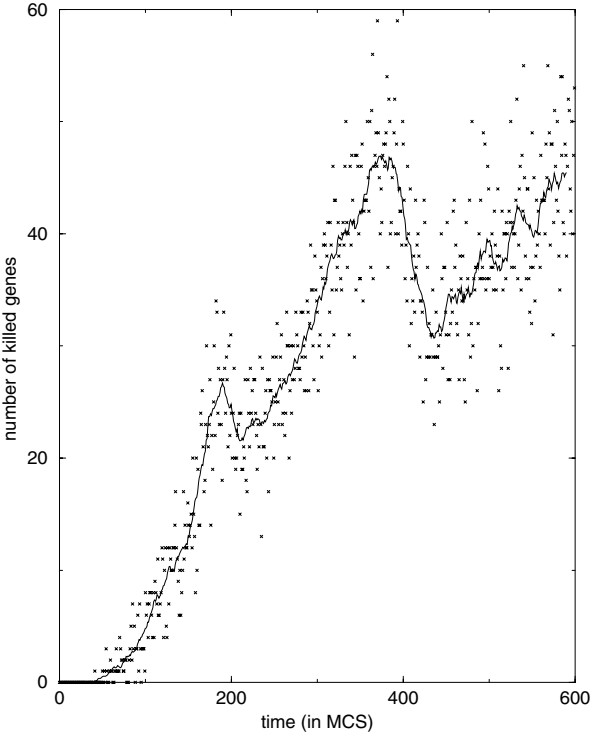

Fig. 3. The number of genes killed by selection in the course of evolution during which mutation pressure switches every 200 MCSs from one specific for the leading DNA strand to the one specific for the lagging DNA strand and *vice versa*.

tion rate in Nature is of the order of 1 per genome per generation). During so many generations recombination processes can translocate or invert the coding sequences inside the genome. Inversion of a gene implies the change of the directional mutational pressure into the mirror one. That is why we checked what would be the effect of switching the mutational pressure from the one characteristic for the leading strand into the mutational pressure characteristic for the lagging DNA strands, which mimics the inversion of a gene. In Fig.2 we have shown the dynamics of gene elimination under the changing mutational pressure (light line). After every 100 MC steps we exchanged the substitution matrix for the mirror one. It is evident that the "killing" rate decrease after inversion. The rate of nucleotide substitution is growing at these conditions (Fig. 3). There is a paradox – to escape the elimination by mutations, the best strategy for a gene is to increase the mutation rate but under the mirror asymmetric substitution matrix, what is possible just by inversion. In Fig. 4 we have shown how the elimination rate depends on the substitution accumulation when inversion is allowed and when it is forbidden.

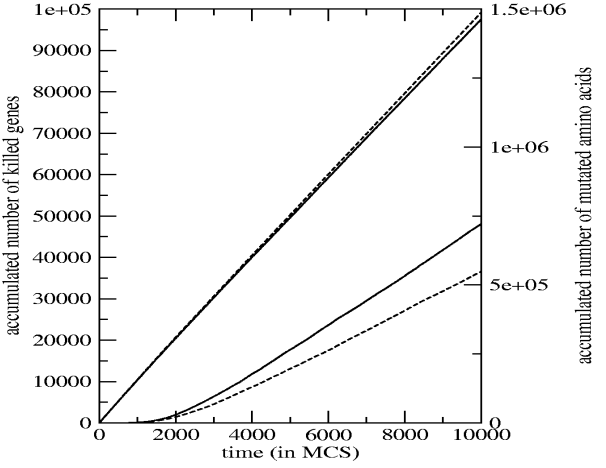

Fig. 4. Accumulated number of genes killed by selection and the accumulated number of mutated amino acids (two upper lines). Solid lines without, and dashed lines with switching mutation matrices.

3 Conclusions

Mutation process is not a random process, it is rather highly correlated. If a coding sequence stay for a long time under the stable directional, asymmetric mutational pressure its composition adopt to this pressure but only in the limits set by selection forces. By inverting the asymmetric pressure it is possible to lower the probability of trespassing these limits set by selection. Our preliminary studies show averaged data for the set of genes from *B. burgdorferi* genome but it could be expected from the experimental data concerning the probability of inversion of genes and highly differentiated divergence rate that this effect could depend on the function and codon composition of genes. Further studies will probably show how important role in the process is played by the genetic code – its extremely sophisticated way of degeneracy. Is it the best?

Acknowledgements. The work was supported by the grant number 1016/S/IMi/02.

References

1. Watson, J.D. & Crick, F.C.H.: A structure for deoxyribose nucleic acid. Nature **327** (1953) 169–170
2. Chargaff, E.: Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia **6** (1950) 201–240

3. Sueoka, N.: Intrastrand parity rules of DNA bases composition and usage biases of synonymous codons. J. Mol. Evol. **40** (1995) 318–325, 42, 323
4. Lobry, J.R.: Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. **40** (1995) 326–330, 41, 680
5. Mrazek J., Karlin S.: Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95. (1998) 3720–3725
6. Frank A.C., Lobry J.R.: Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238** (1999) 65–77
7. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, A., Dudek, M.R., and Cebrat, S.: DNA asymmetry and the replicational mutational pressure. J. Appl. Genet. 42 (2001) 553–577
8. Karlin S., Blaisdell, B.E., Bucher, P.: Quantile distributions of amino acid usage in protein classes. Protein Eng. **5** (1992) 729–738
9. Karlin, S., Mrazek, J.: What drives codon choices in human genome? J. Mol. Biol. **262** (1996) 459–472
10. Shepherd J.C.: Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proc. Natl. Acad. Sci. USA 78. (1981) 1596–1600
11. Karlin S., Burge C.: Dinucleotide relative abundance extremes: a genomic signature. Trends Genet.**11** (1995) 283–290
12. Francino, M.P., Chao, L., Riley, M.A., Ochman, H.: Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science **272** (1996) 107–109
13. Cebrat, S., Dudek, M.R., Mackiewicz, P.: Sequence asymmetry as a parameter indicating coding sequences in Saccharomyces cerevisiae genome. Theory Bioscienc. **117** (1998) 78–89
14. Lobry, J. R.: Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13** (1996) 660–665
15. Blattner, F. R., Plunkett III, G., Bloch C.A., et al. (17 co-authors): The complete genome sequence of Escherichia coli K-12. Science **277** (1997) 1453–1462
16. Freeman, J. M., Plasterer, T.N., Smith, T.F., and Mohr, S.C.: Patterns of genome organization in bacteria. Science **279** (1998) 1827
17. Grigoriev, A.: Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. **26** (1998) 2286–2290
18. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., and Cebrat, S.: How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. **9** (1999) 409–416
19. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., and Cebrat, S.: High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol. Biol. **1** (2001) (1):13
20. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: Multiple base substitution corrections in DNA sequence evolution. Int. J. Modern Phys. C **12**(7) (2001) 1043–1053
21. Mackiewicz, P., Kowalczuk, M., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Łaszkiewicz, A., Dudek, M.R., Cebrat, S.: Replication associated mutational pressure generating long-range correlation in DNA. Physica A **314** (2002) 646–654
22. Tatusov, R. L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V.: The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29** (2001) 22–28