Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 336 (2004) 63–73
PHYSICA A
www.elsevier.com/locate/physa

Simulation of gene evolution under directional mutational pressure

Małgorzata Dudkiewicz[a], Paweł Mackiewicz[a], Maria Kowalczuk[a], Dorota Mackiewicz[a], Aleksandra Nowicka[a], Natalia Polak[a], Kamila Smolarczyk[a], Joanna Banaszak[a], Mirosław R. Dudek[b], Stanisław Cebrat[a,*]

[a] *Department of Genomics, Institute of Genetics and Microbiology, Wrocław University ul. Przybyszewskiego 63/77, 51-148 Wrocław, Poland*
[b] *Institute of Physics, University of Zielona Góra ul. Szafrana 4a, 65-069 Zielona Góra, Poland*

Received 3 November 2003

Abstract

The two main mechanisms generating the genetic diversity, mutation and recombination, have random character but they are biased which has an effect on the generation of asymmetry in the bacterial chromosome structure and in the protein coding sequences. Thus, like in a case of two chiral molecules—the two possible orientations of a gene in relation to the topology of a chromosome are not equivalent. Assuming that the sequence of a gene may oscillate only between certain limits of its structural composition means that the gene could be forced out of these limits by the directional mutation pressure, in the course of evolution. The probability of the event depends on the time the gene stays under the same mutation pressure. Inversion of the gene changes the directional mutational pressure to the reciprocal one and hence it changes the distance of the gene to its lower and upper bound of the structural tolerance. Using Monte Carlo methods we were able to simulate the evolution of genes under experimentally found mutational pressure, assuming simple mechanisms of selection. We found that the mutation and recombination should work in accordance to lower their negative effects on the function of the products of coding sequences.

PACS: 87.10.+e; 02.70.Lq; 07.05.Tp

Keywords: Sense strand; DNA asymmetry; Coding; Evolution; Mutation

* Corresponding author. Tel.: +48-71-3756303; fax: +48-71-3252151.
E-mail address: cebrat@microb.uni.wroc.pl (S. Cebrat).

0378-4371/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2004.01.011

1. Introduction

This year we are celebrating the 15th anniversary of the greatest discovery in biology of the 20th century—the description of the elegant model of the DNA molecule [1]. DNA is composed of only four different elements—nucleotides: A—(Adenine), T—(Thymine), G—(Guanine) and C—(Cytosine). These nucleotides form two long sequences—strands which are interconnected in one double-strand helix. There is the complementarity rule, called Chargaff's rule or parity rule 1 (PR1) [2], stating that A on one DNA strand corresponds to T on the other strand and similarly G corresponds to C. Thus, the number of A is equal to the number of T and the number of G is equal to the number of C in the whole molecule. If we imagine a random DNA sequence, the number of a particular nucleotide in one strand should roughly correspond to the number of that nucleotide in the other strand (the difference should be statistically insignificant), as implications—the number of A should be statistically equal to the number of T in the same strand and the number of G should equal to the number of C. This rule is called Chargaff's parity rule 2 (PR2) Refs. [3–5] and usually, for the natural DNA sequences it is very roughly satisfied only in the scale of the whole chromosome.

Thus, there must be some mechanisms which introduce the deviations from the PR2 into DNA molecules found in nature. In fact there are many such mechanisms but two of them are the most important—directional mutational pressure associated with the DNA replication and the coding functions of DNA sequences (see for review Refs. [6,7]).

2. DNA asymmetry introduced by replication

The two DNA strands forming the double helix are polar and pair in opposite directions. On the other hand, the synthesis of any new nucleotide strand can be performed only in the direction $5' \rightarrow 3'$ on the oppositely oriented matrix strand. It is obvious that the mechanism of replication of one strand has to be different than that of the other strand. In fact one strand is synthesized continuously (leading strand) and the other one is completed from fragments (lagging strand) [9]. Besides any further implications of the topology of replication, these differences in replication accompanied by some different preferences in the introducing errors during synthesis (mutations) generate a bias into the nucleotide composition of the two strands, called the DNA asymmetry. Such a DNA asymmetry, in whole bacterial genomes, was observed for the first time by Lobry [8]. The best way to show the asymmetry is a cumulative, detrended DNA walk [10,11] shown in Fig. 1. In this figure two extremes are visible. These extremes correspond to the points where the role of the DNA strand changes from the leading to the lagging one or vice versa (see description of Fig. 1.), or just the points where replication starts or terminates. In the most of bacterial genomes the leading strand is richer in G and T (see Ref. [7] for review). DNA replication is a very sophisticated process of information copying. Usually the number of errors is of the order of one per genome replication (there are millions or even billions of nucleotides in

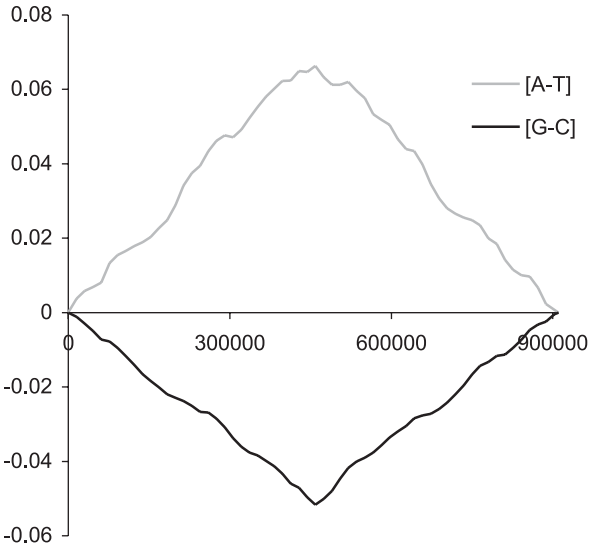

Fig. 1. Cumulative, detrended DNA walk for the *B. burgdorferi* genome. The values on *y*-axis represent cumulative difference between complementary nucleotides. The values on *x*-axis represent positions on chromosome in base pairs. Two visible extremes correspond to the points where the DNA strand changes from leading to lagging one (the point in the middle is the origin of replication).

Table 1
Table of nucleotide substitution rates of *B.burgdorferi* genome (leading strand)

	A	T	G	C
A	0.807	0.065	0.164	0.070
T	0.103	0.865	0.116	0.261
G	0.067	0.035	0.705	0.047
C	0.023	0.035	0.015	0.622

The element M_{ij} of the table is representing the relative probability that nucleotide j (in column) will mutate to the nucleotide i (in row). $\sum_i M_{ij} = 1$.

one genome) [12]. Nevertheless, accumulation of errors in the evolutionary time scale eventually leads to the DNA asymmetry seen in the contemporary bacterial genomes [4,10,13,14]. To reproduce the process of accumulation of mutations (substitutions of one nucleotide by another during replication) the relative frequencies of all the 12 possible substitutions on the leading DNA strand of the *Borrelia burgdorferi* genome were estimated [15–17]. The substitution matrix describing the directional mutational pressure in the leading DNA strand of this genome is shown in Table 1. The matrix has some interesting properties; any DNA sequence during the simulation of evolution under such a mutational pressure tends to reach the nucleotide composition of the third codon positions (the most degenerate) of the genes positioned on the leading

DNA strand, the frequency of nucleotides in the sequence in equilibrium with its mutational pressure is highly linearly correlated with their half-time of substitution [15], which means that very mutable nucleotides are less represented in the DNA sequence. The most important property of this mutational pressure is that the frequencies of the substitution of the same nucleotide on different DNA strands are different—there are different substitution matrices for leading and lagging DNA strands. We call them "the mirror matrices" since the values of the substitution rates for the lagging strand are the same as in Table 1 but the respective symbols of the nucleotides should be substituted by the complementary ones, i.e., *A* for *T*, *T* for *A*, *G* for *C* and *C* for *G*. A further consequence of this feature is that DNA in equilibrium is asymmetric,—it does not obey the PR2.

3. Asymmetry of the coding sequences

A protein coding gene is a fragment of the double stranded DNA, but only one strand of this fragment is a matrix for synthesis of the messenger RNA molecule which is translated into amino acid sequence. Since "the sense of the message" is transferred by the RNA molecule, the matrix strand is called anti-sense and the strand complementary to the matrix is called the sense strand. Thus, describing the position of the gene and its "direction" we are referring to the position and direction of the sense strand of the gene which is the same as the direction of its transcription into RNA. The sense strand codes for the amino acid sequence of the coded protein. There are 61 codons (trinucleotide sequences) in the genetic code which code for 20 amino acids and the three other stand for stop signals, where the protein synthesis stops. Thus, the genetic code is degenerate. Only two amino acids are coded by single codons. Each of the rest of amino acids is coded by more than one codon. In half of codons, any substitution in the third position does not change the meaning of the codon. These are four-fold degenerate codons and they are situated in separate "boxes". The rest of codons are organised in the so-called semi-boxes where substitution at the third position in codons of one purine by the another one or one pyrimidine by another pyrimidine does not change the meaning of the codon. The third position is the most degenerate. But the other positions in codons decide often about the structural property of the coded amino acid, not only about the amino acid itself, (i.e., T at the second position codes for hydrophobic amino acids, while A for polar amino acids). The structure of the genetic code and the selection pressure exerted on the proteins are responsible for the specific asymmetry of the protein coding sequences. Usually the sense strands of these sequences are richer in purines (A and G) than in pyrimidines (T and C) [19,20]. Furthermore, each position in the codon has its own specific asymmetry [22] which can be depicted in the diagrams of two-dimensional DNA walks of Berthelsen type [11,21] (see the results of DNA walk on one coding sequence in the *B. burgdorferi* genome and on the all coding sequences positioned in this genome on the leading and lagging strand, Fig. 2(a) and (b)). It can be seen from these DNA walks that the nucleotide compositions of the coding sequences are highly correlated. Taking under consideration the fact that protein coding sequences are asymmetric and that the frequencies of

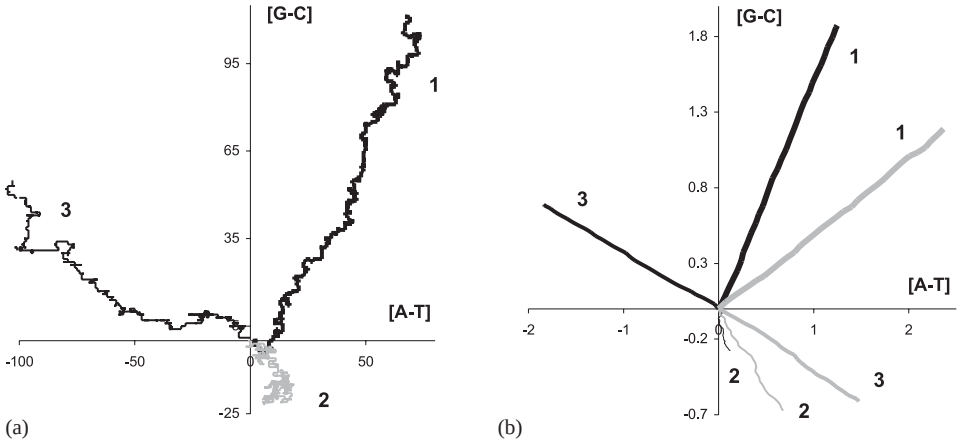

Fig. 2. Berthelsen's type walks [21] for (a) ORF BB0020 (556 bp) and (b) for all the coding sequences positioned in the genome of *B. burgdorferi* on the leading (black lines) and lagging (grey lines) strand. The numbers 1, 2, 3 indicate the positions in codons.

substitutions of the same nucleotide on different strands are different it is very important for the overall frequency of mutation of a given gene how it is oriented on the bacterial chromosome.

4. Materials and methods

All simulations were performed with DNA sequences of the *B. burgdorferi* genome [18] downloaded from *www.ncbi.nlm.nih.gov*. The replication associated directional mutational pressure for the *B. burgdorferi* genome is described in terms of the substitution matrix presented in Table 1 for the leading DNA strand and its mirror representation for the lagging DNA strand [15–17].

The protein coding sequences of the *B. burgdorferi* genome were divided into two classes: (1) the ones lying on the leading strand (564 sequences of the total length of 560,550 nucleotides) and (2) the ones lying on the lagging strand (286 sequences of the total length of 291,933 nucleotides). When we say that 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 the probability $p_{mut} = 0.01$ and then substituted by another nucleotide with the probability described by the corresponding parameter 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. The selection parameter—tolerance T for the codon and the amino acid composition of individual sequences is introduced. It describes the maximum

allowed deviation in the codon composition of the gene after the mutation in comparison to its original sequence (occurring in the genome) or 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 codons or amino acids as follows:

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

where $f_i(0)$ is the fraction of a given ith codon/amino acid in the original sequence (before mutations) and $f_i(t)$ is a fraction of a given codon/amino acid in the sequence after mutations in t MCS.

Arbitrarily, we have assumed as the value of tolerance T for amino acid composition the average distance between 442 pairs of orthologs belonging to two related genomes: *B. burgdorferi* and *Treponema pallidum*. Orthologs are sequences from different species which have evolved by vertical descent and are usually responsible for the same function in different organisms [23]. These orthologs were extracted from COGs database [24] downloaded from *ftp://www.ncbi.nlm.nih.gov/pub/COG* in September 2001. The tolerance estimated by this method is equal to 0.3. If the number of substituted elements in the protein coded by a given gene overpasses the declared tolerance, the coding sequence is "killed" and replaced by the corresponding one from the second, parallely evolving, genomic sequence. We have performed simulations using the simple model of selection for the global amino acid composition of gene products [25]. During the simulation we counted the number of substitutions introduced into genes, the number of genes eliminated by selection and the number of accumulated substitutions in the surviving genes (divergence from the original sequence). We simulated simultaneously the evolution of two, initially identical genomes. In one MCS, the first genome was mutated and each gene of this genome was checked for survival, if a gene was killed by selection it was replaced by its homolog from the second genome. When the mutation of the first genome was completed, the second genome was mutated and if any gene in this genome was killed it was replaced by its homolog from the first genome. In this paper we have examined the average number of gene replacements for the whole *B. burgdorferi* genome—genes located on the leading strand and genes located on the lagging strand.

5. Results and discussion

In Fig. 3 we have shown the prolonged simulation of evolution of genes both, from the leading and the lagging DNA strands, under stable mutational pressure. The frequency of replacement of genes grows in time. It can be explained by the accumulation of amino acid replacements in the gene product during evolution. At the beginning of the simulation the fraction of substituted amino acids is too small to eliminate a gene. When the fraction of substitutions overpasses the assumed tolerance T, the gene is killed and replaced by its homolog, which already has accumulated some mutations. Since the mutational pressure is "directional"—it has preferences for introducing some types of substitution, as a consequence the composition of the gene products tends to

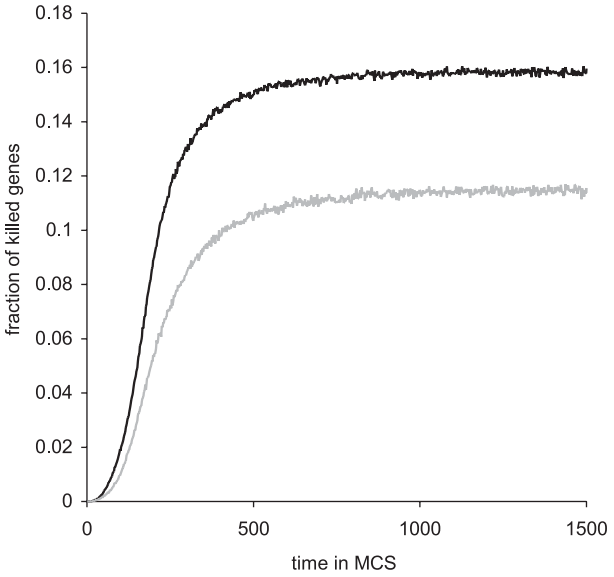

Fig. 3. Fraction of killed genes from the leading (black line) and the lagging (grey line) DNA strands under stable mutational pressure. The results were averaged from 300 simulations.

drift to one direction rather than fluctuate randomly around the original composition. The situation changes drastically when the directional mutational pressure is changed into the mirror one. In nature it corresponds to the gene inversion or its translocation from the leading strand to the lagging strand or vice versa. In Fig. 4 we have shown the accumulation of amino acid substitutions in genes from the leading and the lagging strands and their replacement rates in simulations when the substitution matrices were changed for the mirror ones every 200 MC steps. There are a few interesting observations. The most important—the replacement rates decrease after switching the substitution matrices. It can be explained by the inverted directions of substitutions, i.e., if the directional mutational pressure tends to replace A by T (if in the second position in codons, it changes polar amino acid into the hydrophobic one), the inversion would prefer changes of T into A which invert also the trend in amino acid substitutions. The second important observation is the increase of the divergence rate measured as the fraction of accumulated mutations. There are two reasons for that, one is the lower replacement rate, thus the substitutions are accepted because the gene is not replaced and the second, because the absolute rate of substitutions is growing. The nucleotide composition of DNA sequences adapts to the stable directional mutational pressure even under selection pressure and the probability of a substitution is decreased in time. When the substitution matrix is changed during the simulation, the DNA sequence is not adapted to the new type of mutational pressure and the substitution rate is higher. Finally, the third observation is that the genes positioned by the evolutionary mechanisms on differently replicating DNA strands differently respond to

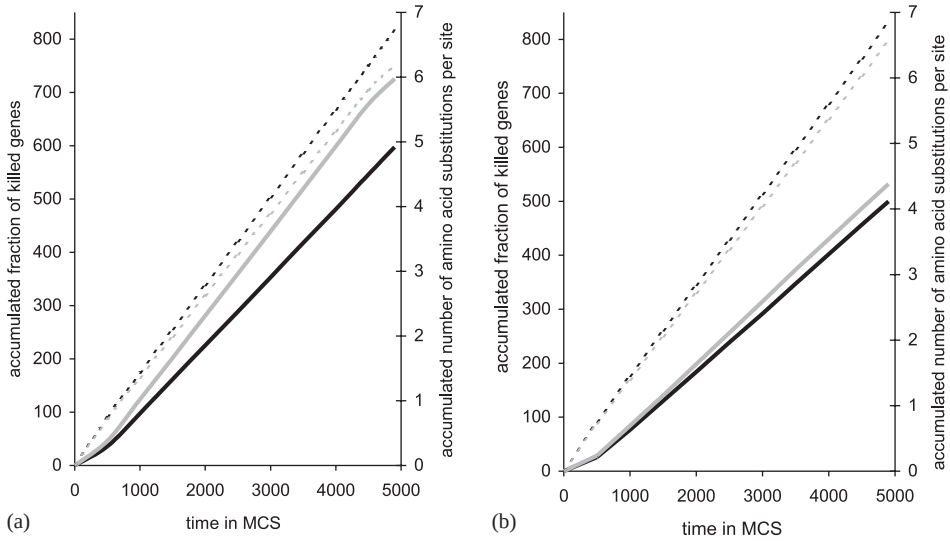

Fig. 4. Accumulated amino acid substitutions in genes from the leading (a) and the lagging (b) strands and their elimination. Solid lines—killing rates (light one for the simulation with stable mutational pressure, dark one for the pressure changing every MC step), dashed lines for cumulated substitutions (colours as above).

the directional mutational pressure operating on the two strands and differently respond to the periodical changes in that pressure.

We have selected a few genes which differently responded to the directional mutational pressure and analysed their stability under different frequencies of inversions. Results are shown in Fig. 5. There are some genes which prefer to stay on the same strand, some genes prefer to change the mutational pressure very often but with the same time spent on each strand and some genes are more stable with allowed inversions but with different time spent on each strand. The results of simulations indicate that the rate of inversion is a very important element in the overall rate of gene evolution. Since inversions, as the other recombination events (they can be connected with distant translocations), can be considered as mechanisms destabilizing the structure of prokaryotic chromosomes, there should be some other evolutionary mechanisms increasing the inversion frequency with low probability of destabilising the function of the gene. In fact, in prokaryotic genomes there are many transposons which encompass a block of genes responsible for performing the function which can be translocated around the chromosome with relatively low risk of the deleterious effect [26]. Such translocations are possible and frequent in the prokaryotic genomes but they are forbidden in the eukaryotic ones, because of the strategy of reproduction. In eukaryotes translocations lead to the production of unbalanced gametes. That is why genes have to stay at least roughly at the same positions on chromosomes in every genome of the cross breeding populations to enable sexual reproduction. On the other hand there are a lot of transposable elements in the eukaryotic genomes which constitute a significant fraction of whole genomes and do not code for

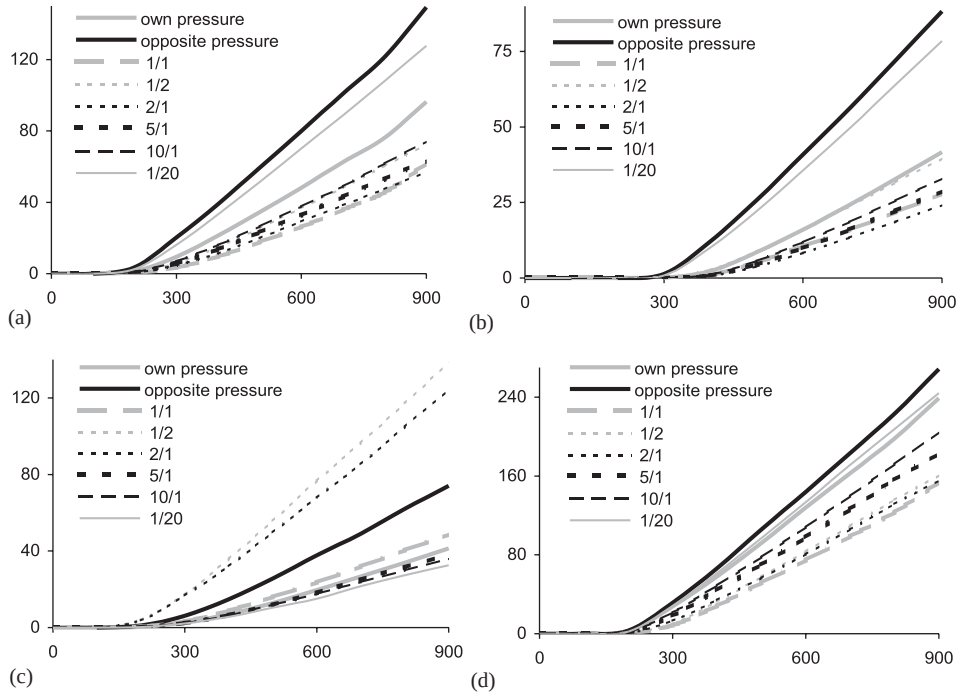

Fig. 5. Elimination rate of some selected *B. burgdorferi* genes under different frequencies of inversions: (a) predicted coding region BB0806 from the leading strand; (b) acriflavine resistance gene (acrB) from the leading strand; (c) predicted coding region BB0009 (lagging strand); (d) DNA-directed RNA polymerase (rpoB) (leading strand). The values on y-axis represent accumulated number of killed genes. The values on x-axis represent simulation time in MCS. Legend description: own pressure—simulation without inversions, under stable mutational pressure proper for the strand on which analysed gene is located, opposite pressure—stable mutational pressure, but opposite to the previous one (proper for the complementary strand), numeric characters: x/y—simulation with inversions where: x—the number of MCS under mutational pressure proper for the strand on which analysed gene is located, y—the number of MCS under mutational pressure opposite to the previous one. The ratio x/y is repeated till the end of the simulation. The results were averaged from 100 simulations.

"reasonable" functions. These elements can "push" slightly a coding sequence by inserting upstream or downstream of the gene. Such a shift of the gene in one direction might be unimportant from the point of view of the gamete production but it changes distance of the gene from ARSes (Autonomously Replicating Sequences playing the role of origins of replications). In consequence it changes the relative frequencies of replication of the sense strand of a gene as a leading strand or lagging strand. Thus, intergenic sequences in the eukaryotic genomes can tune the directional mutational pressure to the optimal level for a given gene just by putting it at the proper position between two ARSes. Could it be the role of intergenic sequences in the eukaryotic genomes?

6. Conclusions

Protein coding sequences under stable directional mutational pressure accumulate nucleotide substitutions which eventually cause the elimination of the gene by selection. Gene inversions, changing the "direction" of the mutational pressure decrease the lethal effect of mutations while increasing the rate of accumulation of substitutions. The effect of gene inversions in prokaryotic genomes can be replaced in the eukaryotic ones by the relatively short distance "shift" of coding sequences between two neighbouring ARSes. This mechanism of avoiding the deleterious effects of directional mutational pressure could explain the role of intergenic sequences in eukaryotic genomes and different roles of the genetic transposable elements in organisms with different reproduction strategies.

Acknowledgements

M.K. was supported by Foundation for Polish Science.

References

[1] J.D. Watson, F.C.H. Crick, Nature 327 (1953) 169.
[2] E. Chargaff, Experientia 6 (1950) 201.
[3] E. Chargaff, Fe. Proc. 10 (1951) 654.
[4] J.R. Lobry, Mol. Biol. Evol. 13 (1996) 660.
[5] N. Sueoka, J. Mol. Evol. 40 (1995) 318–325.
[6] A.C. Frank, J.R. Lobry, Gene 238 (1999) 65.
[7] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, J. Appl. Genet. 42 (2001) 553.
[8] J.R. Lobry, J. Mol. Evol. 40 (1995) 326–330;
J.R. Lobry, J. Mol. Evol. 41 (1995) 680.
[9] R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto, A. Sugino, Proc. Natl. Acad. Sci. USA 59 (2) (1968) 598.
[10] J.M. Freeman, T.N. Plasterer, T.F. Smith, S.C. Mohr, Science 279 (1998) 1827.
[11] S. Cebrat, M. Dudek, Eur. Phys. J. B 3 (1998) 271.
[12] M.Y. Azbel, Physica A 273 (1999) 75.
[13] P. Mackiewicz, A. Gierlik, M. Kowalczuk, M.R. Dudek, S. Cebrat, Genome Res. 9 (1999) 409.
[14] E.P. Rocha, A. Danchin, A. Viari, Mol. Microbiol. 32 (1999) 11.
[15] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, BMC Evol. Biol. 1 (2001) 1.
[16] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, Int. J. Modern Phys. C 12 (7) (2001) 1043.
[17] P. Mackiewicz, M. Kowalczuk, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, A. Laszkiewicz, M.R. Dudek, S. Cebrat, Physica A: Stat. Mech. Appl. 314 (1–4) (2002) 646.
[18] 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., Nature 390 (1997) 580.
[19] G. Gutierrez, L. Marquez, A. Martin, Nucleic Acids Res. 24 (1996) 2525.
[20] S. Cebrat, M.R. Dudek, A. Rogowska, J. Appl. Genet. 38 (1997) 1.
[21] Ch.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Phys. Rev. A 45 (1992) 8902.
[22] S. Cebrat, M.R. Dudek, P. Mackiewicz, M. Kowalczuk, M. Fita, Microbiol. Comp. Genomics 2 (1997) 259.
[23] W.M. Fitch, Syst. Zool. 19 (1970) 99.

[24] 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, Nucleic Acids Res. 29 (2001) 22.
[25] M. Dudkiewicz, P. Mackiewicz, A. Nowicka, M. Kowalczuk, D. Mackiewicz, N. Polak, K. Smolarczyk, M.R. Dudek, S. Cebrat, Lecture Notes Computer SC 2658 (2003) 650.
[26] P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, S. Cebrat, Genome Biol. 2 (12) (2001) 1004.1–1004.4.