How gene survival depends on their length

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

[1] Department of Genomics, Institute of Genetics and 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. Gene survival depends on the mutational pressure acting on the gene sequences and selection pressure for the function of the gene products. While the probability of the occurrence of mutations inside genes depends roughly linearly on their length, the probability of elimination of their function does not grow linearly with the length because of the intragenic suppression effect. Furthermore, the probability of redefinition of the stop and start codons is independent of the gene length while shortening of gene sequences by generating stop codons inside gene sequences depends on gene length.

1 Introduction

The primary mechanism introducing biodiversity into populations is mutagenesis. Further diversification of genomes is the result of recombination between genomes which are already different. One of many different mechanisms introducing mutations into genomes are single nucleotide substitutions which happen during DNA replication. There are four different kinds of nucleotides Adenine (A), Thymine (T), Guanine (G), and Cytosine (C) and substitution of one of them by any of the three others are random but highly biased. Some nucleotides are more often substituted than others and the substituting nucleotides are also unevenly "chosen" [1]. Thus, for each of the twelve possible kinds of nucleotide substitutions a specific probability of the event can be experimentally estimated and put into the "matrix of substitutions" (Tab. 1) [2].
Besides the twelve values in the matrix describing the specific substitution probabilities, there are four values in the diagonal, describing the stability of nucleotides against mutations. Thus, it is possible to talk about the stability of the nucleotide positions in the genes and genomes. The most stable genes should

* To whom all correspondence should be sent.

	A	T	G	C
A	0.300	0.103	0.067	0.023
T	0.066	0.480	0.035	0.035
G	0.164	0.116	0.149	0.015
C	0.070	0.261	0.047	0.073

Table 1. Frequencies of substitutions in the leading strand of the *B. burgdorferi* genome (nucleotide in a column is substituted by a nucleotide in a row). All substitution frequencies between different nucleotides sum up to 1.

be built of the most stable nucleotides. On the other hand, the selection for gene function demands rather specific composition of the gene products which restricts not only the nucleotide composition of genes but, which is more important, the proper length of the coding sequence. The stability of the gene as a coding unit depends on both its nucleotide composition and its length. A substitution inside the coding sequence can exert very different effects on the amino acid sequence of its product. There are silent mutations which do not change the sense of the coding sequence (due to the degeneracy of the genetic code), some substitutions change one amino acid in the gene product for another one though very similar, but some substitutions can change the properties of the coded amino acid significantly and such substitutions are potentially dangerous - they can lead to the nonfunctional gene product. Nevertheless, two consecutive mutations can complement their effects. For example, while in one position an amino acid with an acidic residue is replaced by a neutral one, in another position a neutral amino acid could be replaced by an acidic one and the final product does not change its isoelectric point, which seems to be one of the critical values for the protein activity. Such complementation of the mutation effects inside a gene is called intragenic suppression. If the former mutation is not lethal, the consecutive one can be introduced even many generations later. But there are still some other, much more dangerous point mutations - substitutions which eliminate the start or stop codons. These codons are responsible for initiation and termination of the protein synthesis, respectively (there is one start codon in the universal genetic code - ATG, and three stop codons: TAA, TAG, TGA). While elimination of the stop causes additional elongation of the coding sequence, elimination of the start could shorten it. The latter seems to be more deleterious for the gene product. Note that the frequency of these mutations does not depend on the length of the genes since each gene has one start and one stop codon. Furthermore, start and stop codons could be generated inside the coding sequence. In that case, frequency of generation of starts and stops depends on the length of genes and particularly generation of stops is dangerous because it shortens the length of the gene product. The effect of generation of a start codon can be considered as another amino acid substitution, since the start codon inside the gene codes for methionine (ATG in the universal code). In this paper we analyze, using the Monte Carlo methods, the stability of the

real genes of different length found in the *Borrelia burgdorferi* genome under the mutational pressure experimentally described for this genome.

2 Material and Methods

Simulations have been performed on 850 genes taken from the *B. burgdorferi* genome [3], whose sequence and annotations were downloaded from GenBank (*ftp://ftp.ncbi.nih.gov*). The gene sequences were subjected to the replication-associated mutational pressure (RAMP) described by the matrix of nucleotide substitution frequencies - Table 1 [2]. Since in this genome the RAMP is significantly different for the two differently replicating DNA strands: leading and lagging [4], we have applied two different matrices respectively for the genes located on these strands. The matrix describing RAMP of the lagging strand is the mirror reflection of the RAMP for the leading DNA strand. In one Monte Carlo Step (MCS) each nucleotide of the gene sequence was drawn with a probability $p_{mut} = 0.001$ then substituted by another nucleotide with the probability described by the corresponding parameter in the substitution matrix. We have applied two kinds of selection for gene survival: selection for the amino acid composition and selection for start and stop codons. After each round of mutations, we translated the nucleotide sequences into the amino acid sequences and compared the resulting composition of the proteins with the original one. For each gene we calculated the selection parameter T for the amino acid composition as follows:

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

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. It describes the difference in the global amino acid composition of a protein coded by a given gene after mutations and its original sequence from the real genome. If T was below the assumed threshold, the gene stayed mutated and went to the next round of mutations (the next MC step). If T trespassed the threshold - the gene was "killed" and replaced by its allele from the second genomic sequence, originally identical, simulated parallely.

We have applied three variants of selection for start and stop codons. Gene was killed when:

- its start codon was substituted by a non-start codon,
- its stop codon was substituted by a non-stop codon,
- a stop codon was generated inside the gene sequence.

In *B. burgdorferi* genes ATG, TTG, GTG are used as start codons and TAA, TAG, TGA - as stop codons. We have assumed that substitutions between these start codons and between these stop codons are neutral. Similarly to the previous simulations the "killed" gene was replaced by its allele from the second genomic sequence simulated parallely. After each MCS the number of gene replacements

(the number of killed genes) was counted. All simulations were performed for 1000 Monte Carlo steps, repeated 10 times and averaged.

3 Results and Discussion

Selection for the amino acid composition
In the first simulations we have assumed that genes are "killed" only because of changes in the amino acid composition of their products. The results of the simulations done for the whole set of genes from the *B. burgdorferi* genome are shown in Fig. 1.

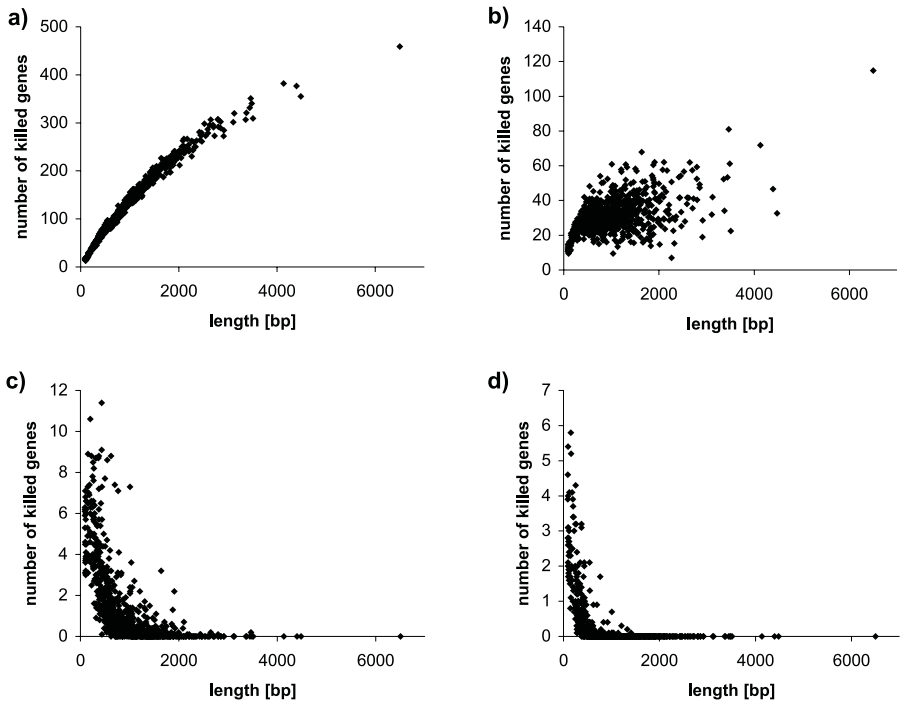

Fig. 1. Elimination of genes from *B. burgdorferi* genome by the selection pressure on their global amino acid composition. In all four cases the genes were under their specific mutational pressure but under different strength of selection pressure - different threshold T: a) $T = 0.01$; b) $T = 0.1$; c) $T = 0.25$; d) $T = 0.33$.

It is obvious that the number of mutations which hit the genes roughly linearly depends on their length. If the assumed tolerance is very low - even one substituted amino acid can eliminate the gene, the intragenic suppressions are very rare and are slightly more probable for longer genes. That is why in such

conditions the probability of elimination of genes depends almost linearly on their length only slightly decreasing with length (Fig. 1a). When increasing the tolerance, the probability of intragenic suppression grows (that could be compared with the buffer capacity) and the probability of killing the gene decreases (Fig. 1b-d). The longer genes can deal with the mutational pressure more successfully than shorter ones. However, one can suspect that the effect of higher sensitivity of shorter genes is caused by the biased nucleotide composition of the short and long genes. To eliminate this effect, we have constructed artificial genes composed of different numbers of repetitions of the same unit being the coding sequence of one short *B. burgdorferi* gene. The observed effect was similar to that for the real set of genes though much more regular because the effect of nucleotide composition of different genes has been eliminated (Fig. 2). Results of our simulation studies have been supported by the analysis of closely related genomes. It has been found that the divergence rate of genes is negatively correlated with their length. Furthermore, the negative correlation is more evident for highly conserved genes (very low tolerance in our simulations).

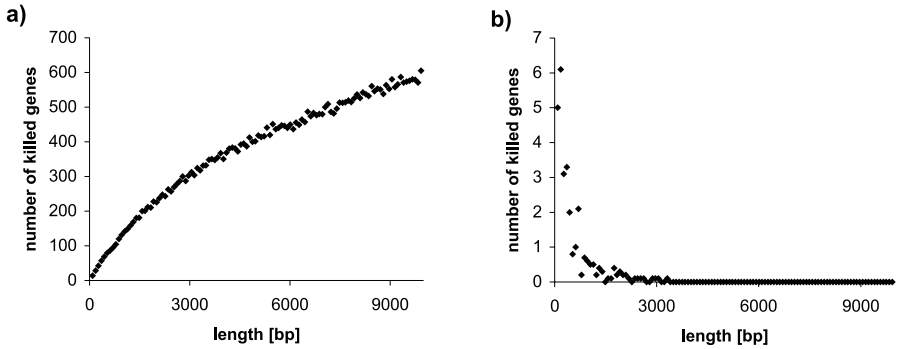

Fig. 2. Elimination of virtual genes of different length by the selection pressure on their global amino acid composition. All sequences - virtual genes - were produced by repeating different number of times one coding sequence of 31 codons long (deprived start and stop codons). These sequences were under their specific mutational but under different strength of selection pressure - different threshold T: a) $T = 0.01$; b) $T = 0.25$.

Selection on the start and stop codons

As it was explained in the introduction section, some amino acid substitutions in the gene products could be complemented by intragenic suppression. That is why the amino acid sequences of different homologous sequences can differ significantly and still fulfill the same function. The situation is quite different for the stop codons generated inside the coding sequence or the eliminated start codons. Both kinds of mutations can significantly shorten the coded protein with deleterious effect on its function, which leads to elimination of the gene and in consequence the genome if the function of this gene can not be complemented.

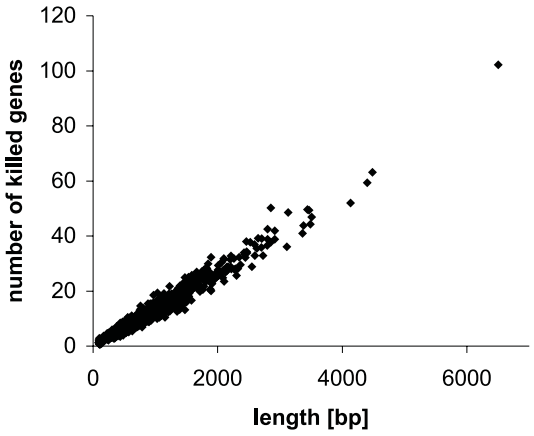

Fig. 3. Killing effect of the stop codon generation inside the coding sequences. Any other mechanisms of selections were switched off.

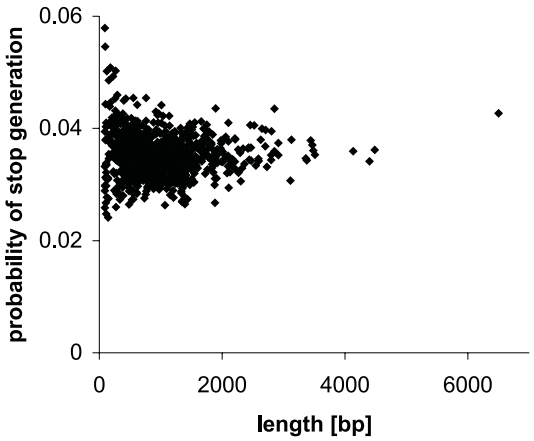

Fig. 4. Relationship between probability of the stop codon generation in genes of *B. burgdorferi* and their length. For each codon the probability of transformation into the stop codon were counted according to the matrix of nucleotide substitution. Then, for each gene the probabilities were weighted by the fractions of codons and summed.

In the next simulations we have eliminated genes when their start codon was substituted by a non-start codon. Since each gene has one start codon, it is obvious that the rate of gene elimination did not depend on the gene size (results not shown). A similar effect was observed for the elimination of stops. Elimination of the stop does not necessarily lead to the gene elimination because these mutations elongate the gene products. We have found that the frequency of stop codons usage in the *B. burgdorferi* genes corresponds almost exactly to the usage counted from the nucleotide composition of DNA in the equilibrium with the mutational pressure. This suggests that it is the mutational pressure which structures the nucleotide composition of stop codons and there is no selection pressure on specific stops usage. More dangerous for the gene function could be the shortening of the coding sequences by generation of stop codons inside the genes. Simulations of this effect have shown that the killing effect depended strongly on the length of the genes (Fig.3). Analytical calculations of the probability of generation of the stop codons inside genes (normalized per length unit) have shown that the generation of stops is not correlated with the gene length (Fig. 4). These results suggest that the selection pressure for the longer genes has not resulted in the decreasing the probability of generating stops - or the longer genes do not avoid codons which could be mutated to the stop codons with higher probability.

4 Conclusions

Simulations of the relationships between the genes' length and their survival have shown that while the short genes and the long ones are equally sensitive for killing by elimination the stop or start codons, the killing effect by amino acid substitutions seems to be relatively stronger for shorter genes, because of the lower probability of intragenic suppression. This effect can be compensated by the effect of stops generation, which can not be suppressed by other intragenic mutations and the longer genes are more susceptible to such mutations. Since the effect of the mutations at the borders of the coding sequences concerns also mutations at the border of introns and exons in the eukaryotic genes, it seems that the price which these genes have to pay for increasing the probability of defects has to be compensated by other profits from using this risky and complicated coding strategy.

Acknowledgements
The work was done in the program COST Action P10 and supported by the grant number 1016/S/IMi/03. M.K. was supported by the Foundation for Polish Science.

References

1. Frank, A.C., Lobry, J.R., Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene **238** (1999) 65–77

2. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: High correlation between the turnover of nucleotides under mutational pressure and the DNA composition. BMC Evol. Biol. **1** (2001) (1):13
3. Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K. et al.: Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390** (1997) 580–586
4. Mackiewicz, P., Kowalczuk, M., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Laszkiewicz, A., Dudek, M.R., Cebrat, S.: Replication associated mutational pressure generating long-range correlation in DNA. Physica A **314** (2002) 646–654