Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie

Krynica Morska, 23rd–27th September 2012

INFLUENCE OF MUTATIONAL PRESSURE ON PROTEIN CODING SEQUENCES IN BACTERIAL GENOMES

Małgorzata Wańczyk[1], Paweł Błażej, Paweł Mackiewicz

[1]Department of Genomics, University of Wrocław
ul. Przybyszewskiego 63/77, 51-148 Wrocław,
[1]malgorzata.wanczyk@smorfland.uni.wroc.pl

ABSTRACT

Composition of protein coding sequences is a result of compromise between selective constraints on their coding function and the mutational pressure, which can modified to some extent the nucleotide composition preferred by the selection. To analyze the influence of mutational pressure on protein coding sequences, we simulated evolution of protein coding sequences from two bacterial genomes, *Borrelia burgdorferii* and *Chlamydia muridarum*. The sequences were subjected to mutational substitution matrices and the selection process was modeled by the modified version of the commonly used algorithm GeneMark for finding protein coding sequences in prokaryotic genomes. We considered the effect of three types of selection mechanisms ('negative', 'stabilizing' and 'positive') on changes in the coding potential during the simulations. We also studied the influence of mutational and selection pressures on the nucleotide composition of three codon positions in protein coding genes.

INTRODUCTION

Protein coding sequences show the characteristic triplet structure, which can be described by the specific codon usage, different nucleotide composition in three codon position and other compositional measures [1–6]. The gene structure is shaped not only by the strong selection pressure resulting from the coding and functional requirements but also by the mutational pressure related usually with replication or transcription processes. Among three codon positions, the third ones are strongly influenced by mutations because nucleotide substitutions in these sites usually do not change coded amino acid residues or their properties. In turn, the second position in codons are very conserved because some mutations in these positions can change hydrophobicity of coded amino acid residues. Therefore the structure and composition of protein coding sequences is a superposition of both mutations and selections.

To study the influence of mutational pressure on protein coding sequences, we elaborated a simulation model in the example of two bacterial genomes. The model was previously considered in the context of the influence of different mutational pressures on gene evolution [7–10]. Here we have tested this model using different selection mechanisms on coding potential of studied gene sequences. Moreover, we have analyzed the influence of mutational pressure and selection constraints on the nucleotide composition of three codon positions in protein coding sequences.

Table 1. The substitution matrix describing mutational pressure in the DNA strands for the *B. burgdorferi* genome. A nucleotide in the first column is substituted by a nucleotide in the first row.

leading DNA strand				
	A	T	G	C
A	0.81	0.10	0.07	0.02
T	0.07	0.87	0.03	0.03
G	0.16	0.12	0.71	0.01
C	0.07	0.26	0.05	0.62

lagging DNA strand				
	A	T	G	C
A	0.87	0.07	0.03	0.03
T	0.1	0.81	0.02	0.07
G	0.26	0.07	0.62	0.05
C	0.12	0.16	0.01	0.71

Table 2. The substitution matrix describing mutational pressure in the DNA strands for the *C. muridarum* genome. A nucleotide in the first column is substituted by a nucleotide in the first row.

leading DNA strand				
	A	T	G	C
A	0.73	0.05	0.19	0.04
T	0.05	0.74	0.04	0.16
G	0.15	0.05	0.78	0.02
C	0.05	0.18	0.02	0.75

lagging DNA strand				
	A	T	G	C
A	0.76	0.05	0.14	0.05
T	0.05	0.71	0.03	0.21
G	0.13	0.04	0.81	0.02
C	0.06	0.20	0.02	0.72

MATERIAL AND METHODS

Simulation model

Similarly to [10] we considered a population of 72 individuals. Each individual was composed of protein coding sequences with annotated function, which came from one of two bacterial genomes: *Borrelia burgdorferi* or *Chlamydia muridarum*. The gene sequences and their annotations were downloaded from NCBI database (www.ncbi.nlm.nih.gov). The set of protein coding sequences was divided into two subsets according to their location on differently replicated DNA strands: the genes lying on the leading and the lagging DNA strand. Each individual of *B. burgdorferii* genome consisted of 333 leading strand genes and 142 lagging strand genes whereas an individual representing *C. muridarum* contained 315 genes from the leading strand and 249 genes from the lagging strand.

During one simulation step, a particular nucleotide in the gene sequences was chosen for mutation using the Poisson process assuming one mutation per genome on average. Then, the selected nucleotide was substituted by another according to a transition probability matrix appropriate for the given DNA (leading or lagging) strand and the genome (*B. burgdorferi* or *C. muridarum*). The mutational matrix for *B. burgdorferi* was taken from [11] and the matrix for *C. muridarum* from [12]. These two matrices were created in different ways and are shown in Tab. 1 and Tab. 2.

In the selection stage, an individual was eliminated from the population when a stop codon appeared in its gene sequence. In addition, the coding potential of a mutated protein coding sequence in a given individual was checked. If the sequence was recognized as coding, the nucleotide substitution was accepted, otherwise the individual was eliminated and replaced by other one with a proper coding potential. The coding potential was calculated by the gene finding algorithm GM [13] which is based on commonly used GeneMark algorithm for finding protein coding sequences [14]. It assumes that protein coding sequences can be modeled by a non-homogeneous three periodic Markov chain. During the learning step the initial probabilities and also six transition probability matrices for genes are calculated. The model of non-protein coding DNA sequences is described by a homogeneous Markov chain. During the testing step for a given DNA sequence the conditional probability of being in a given reading frame under the protein coding condition

and *a posteriori* probability are calculated [13]. A studied sequence is coding when *a posteriori* probability (coding potential) achieves the highest value in its first reading frame. This approach achieved very good results in recognition of protein coding genes in many genomes [13, 15].

We considered three types of selection for the coding potential. A given individual was eliminated when at least one of its mutated gene sequences had its coding potential lower than:

A. 0.5 according to the GM algorithm ('negative' selection);
B. the original value at the beginning of the simulation ('stabilizing' selection);
C. the value for the same sequence in the previous simulation step ('positive' selection).

Graphical representation of protein coding sequence by DNA walks

To visualize the influence of mutational and selection pressures on nucleotide composition of three codon positions of protein coding sequences, we used a graphic representation of coding DNA sequences performed by three DNA walks for each of three codon positions separately [5, 6, 16] - see Fig. 3. The walk (called a 'spider') starts at the first nucleotide in the fixed codon position and jumps every third nucleotide to the last one. Every jump starts at the origin of a Cartesian plane and is associated with a unit shift, which depends on the nucleotide visited during the walk. The shift is $(0, 1)$ for guanine, $(1, 0)$ for adenine, $(0, -1)$ for cytosine, and $(-1, 0)$ for thymine.

RESULTS AND DISCUSSION

Coding potential change under different types of selection

Fig. 1 shows the change of the average coding potential during simulations for gene sequences not subjected and subjected to the selection of type A (a 'negative' selection). Independently of the studied genome, the coding potential felt dramatically till 2 million simulation steps and then stabilized to the value of 0.5 for the selection conditions. As expected, in the simulation without selection, this parameter decreased faster and reached value close to 0. However, when the selection of type B (a 'stabilizing' selection) was applied, the average coding potential stayed during the simulation almost the same as the original one (Fig. 2). Interestingly, when the selection of type C (a 'positive' selection) was used, the coding potential increased rapidly during the first 2 million MCS of the simulations and stabilized to the value near 1.0 (Fig. 2).

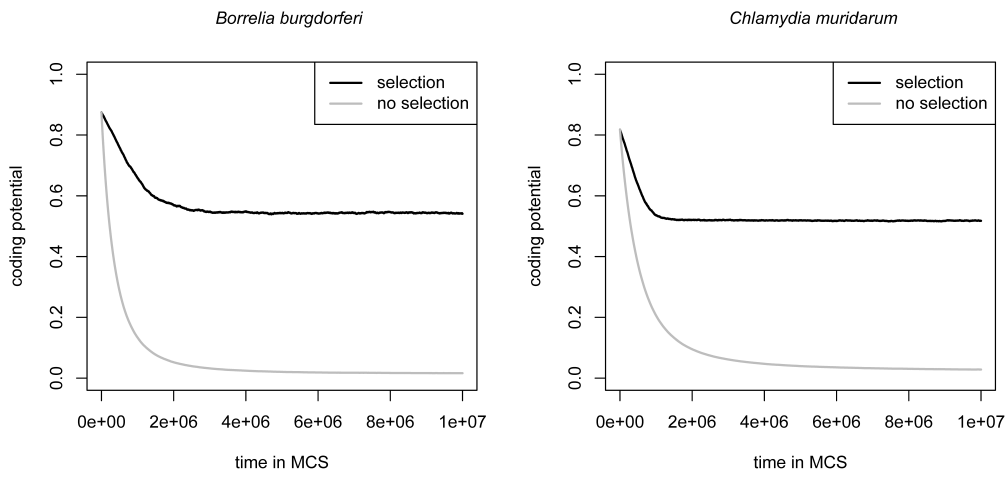

Figure 1. The comparison of the average coding potential in simulations without and with selection of type A for two bacterial genomes.

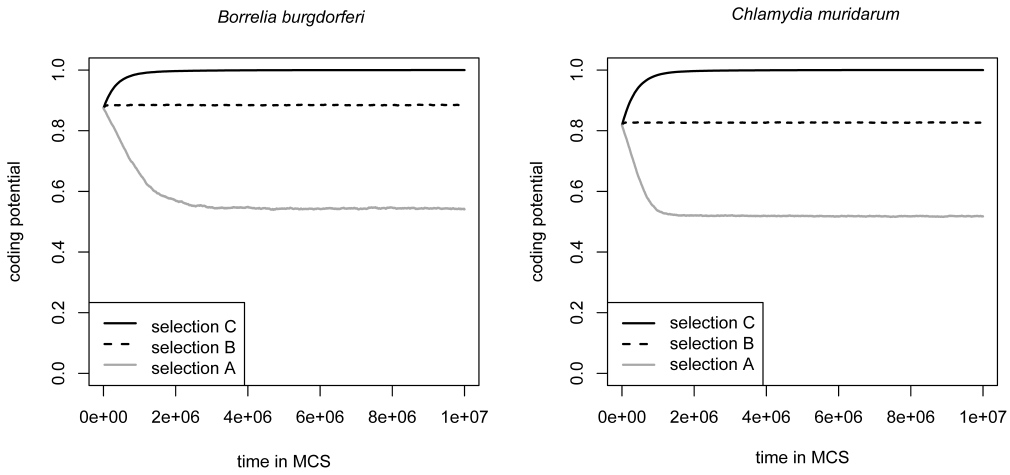

Figure 2. The comparison of the average coding potential in simulations with three different types of selections for two bacterial genomes.

These results indicate that the mutational pressure acts strongly against the preservation of the coding potential and quickly destroys the coding structure of analyzed genes. However, it is possible to improve the coding potential if some kind of 'positive' selection is applied.

Influence of mutational and selection pressures on composition of protein coding genes

To visualize the influence of mutational and selection pressures on compositional structure of protein coding sequences, we joined all such sequences from the leading DNA strand for the given genome and performed DNA walks (see Material and Methods) - Fig. 3. This graphical representation of the sequence shows the trends in the nucleotide composition of particular positions in codons. Original, non-mutated protein coding genes in both genomes (Fig. 3A) are characterized by the clear excess of the adenine over thymine and guanine over cytosine in the first codon positions whereas the third codon positions have more thymine than adenine and are also richer in guanine than cytosine. The second codon positions show much weaker compositional trends with some surplus of cytosine over guanine.

Interestingly, the trends in the nucleotide compositions changed during the simulations despite applying the 'negative' selection pressure of type A (Fig. 3B). The first codon positions of the *B. burgdorferi* genes changed their composition becoming richer in thymine than adenine whereas the second codon positions became similar to the third ones. The *C. muridarum* genes had modified only compositional trends in their second positions also towards the third codon positions. The changes of composition in the second positions are very intriguing because these positions are considered very conserved during evolution. The obtained results indicate that the applied nucleotide selection constraints described by GM algorithm are not strong enough to maintain their composition.

The third codon positions showed the weakest changes in their composition in both genomes, which seems expected because these positions usually show the highest tendency to accumulate mutations and should not change significantly under the applied mutational pressure. In agreement with that, the first and the second codon positions become similar to the third ones, especially in the *B. burgdorferi* sequences, when only mutational pressure was used without selection constraints (Fig. 3C).

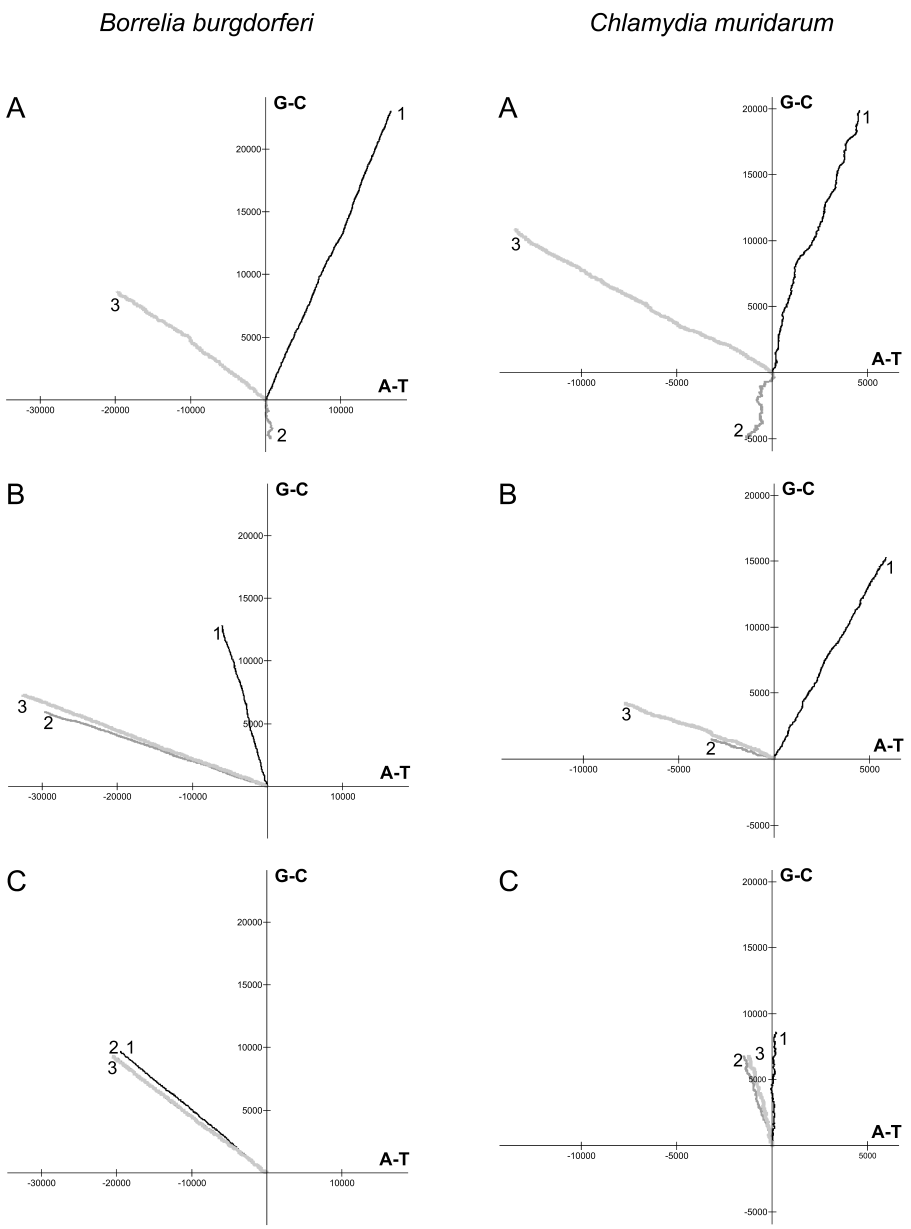

Figure 3. Graphical representation of nucleotide composition of protein coding genes from the leading DNA strand in two bacterial genomes: A - original sequences at the beginning of the simulation; B - sequences after 10 million MCS subjected to selection of type A; C - sequences after 10 million MCS without selection. Three DNA walks for particular codon positions were correspondingly numbered.

However, compositional trends of all codon positions in the *C. muridarum* genes did not become identical because the applied 10 million simulation time was too short to obtained the stationary. Moreover, the trends in *C. muridarum* genes subjected only to the mutational pressure did not resemble the third codon positions of the original gene sequences in contrast to the *B. burgdorferi*. It may indicate that the *B. burgdorferi* substitution matrix better describes the mutational pressure than that of *C. muridarum*. Actually, these two matrices were constructed in two different ways. The *B. burgdorferi* matrix was received by the comparison of original genes with their potential pseudogenes [11] whereas the *C. muridarum* matrix by the comparison of four fold degenerated sites in orthologous genes [12]. It cannot be excluded that the third codon positions in the *C. muridarum* genes do not reflect pure mutational pressure because are subjected to the selection for translation efficiency reflected in synonymous codon usage [17, 18]. It is also possible that this matrix is not in equilibrium with the current composition of the third codon positions.

REFERENCES

[1] J. Tze-Fei Wong and R. Cedergren: *Natural selection versus primitive gene structure as determinant of codon usage*, European Journal of Biochemistry **159** (1986), 175–180.

[2] M. Borodovsky, Y. A. Sprizhitskii, E. I. Golovanov, and A. A Aleksandrov: *Statistical patterns in primary structures of the functional regions of the Genome in Escherichia coli*, Molecular Biology **20** (1986), 826–840, 1144–1150.

[3] J. W. Fickett and C. S. Tung: *Assessment of protein coding measures*, Nucleic Acids Research **20** (1992), 6441–6450.

[4] S. Karlin and J. Mrázek: *What drives codon choices in human genes?*, Journal of Molecular Biology **262** (1996), 459–472.

[5] S. Cebrat, M. R. Dudek, P. Mackiewicz, M. Kowalczuk, and M. Fita: *Asymmetry of coding versus non-coding strand sequences of different genomes*, Microbial and Comparative Genomics **2** (1997), 259–268.

[6] S. Cebrat, M. R. Dudek, and P. Mackiewicz: *Sequence asymmetry as a parameter indicating coding sequence in Saccharomyces cerevisiae genome*, Theory in Biosciences **117** (1998), 78–89.

[7] P. Mackiewicz, M. Dudkiewicz, M. Kowalczuk, D. Mackiewicz, J. Kiraga, N. Polak, K. Smolarczyk, A. Nowicka, M. R. Dudek, and S. Cebrat: *Differential gene survival under asymmetric directional mutational pressure*, Lecture Notes in Computer Science **3039** (2004), 687–693.

[8] M. Dudkiewicz, P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, A. Nowicka, N. Polak, K. Smolarczyk, J. Kiraga, M. R. Dudek, and S. Cebrat: *Higher mutation rate helps to rescue genes from the elimination by selection*, Biosystems **80** (2005), 192–199.

[9] D. Mackiewicz and S. Cebrat: *To understand nature computer modelling between genetics and evolution. In: Miekisz J., Lachowicz M. (eds.), From Genetics to Mathematics*, Series on Advances in Mathematics for Applied Sciences, vol. 79, World Scientific, Singapore, 2009.

[10] P. Błażej, P. Mackiewicz, and S Cebrat: *Simulation of bacterial genome evolution under replicational mutational pressures*, Proceedings of the BIOSTEC 2012, 5th International Joint Conference on Biomedical Engineering Systems and Technologies, Bioinformatics 2012, International Conference on Bioinformatics Models, Methods and Algorithms, Vilamoura, Algarve, Portugal, 1-4 February, 2012, pp. 51–57.

[11] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M. R. Dudek, and S. Cebrat: *High correlation between the turnover of nucleotides under mutational pressure and the DNA composition*, BMC Evolutionary Biology **1** (2001), 13.

[12] E. Rocha and A. Danchin: *Ongoing evolution of strand composition in bacterial genomes*, Molecular Biology and Evolution **18** (2001), 1789–1799.

[13] M. Wańczyk, P. Błażej, and P Mackiewicz: *Comparison of two algorithms based on Markov chains applied in recognition of protein coding sequences in prokaryotes*, Proceedings of the Seventeenth National Conference on Applications of Mathematics in Biology and Medicine. Zakopane-Koscielisko, 1-6 September 2011, pp. 118-123.

[14] M. Borodovsky and J. McIninch: *GeneMark: parallel gene recognition for both DNA strands*, Computers & Chemistry **17** (1993), 123–133.

[15] M. Wańczyk, P. Błażej, P Mackiewicz, and S. Cebrat: *How to deal with small Open Reading Frames*, Proceedings of the BIOSTEC 2012, 5th International Joint Conference on Biomedical Engineering Systems and Technologies, Bioinformatics 2012, International Conference on Bioinformatics Models, Methods and Algorithms, Vilamoura, Algarve, Portugal, 1-4 February, 2012, pp. 245-250.

[16] S. Cebrat and M. R. Dudek: *The effect of DNA phase structure on DNA walks*, The European Physical Journal B **3** (1998), 271–276.

[17] T. Ikemura: *Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes*, Journal of Molecular Biology **146** (1981), 1–21.

[18] J. L. Bennetzen and B. D. Hall: *Codon selection in yeast*, Journal of Biological Chemistry **257** (1982), 3026–3031.