Optimisation of Asymmetric Mutational Pressure and Selection Pressure Around the Universal Genetic Code

Paweł Mackiewicz, Przemysław Biecek, Dorota Mackiewicz, Joanna Kiraga, Krystian Baczkowski, Maciej Sobczynski, and Stanisław Cebrat

Department of Genomics, Faculty of Biotechnology, University of Wrocaw, ul. Przybyszewskiego 63/77, 51-148 Wrocaw, Poland
pamac@smorfland.uni.wroc.pl
http://www.smorfland.uni.wroc.pl/

Abstract. One of hypotheses explaining the origin of the genetic code assumes that its evolution has minimised the deleterious effects of mutations in coded proteins. To estimate the level of such optimization, we calculated optimal codes for genes located on differently replicating DNA strands separately assuming the rate of amino acid substitutions in proteins as a measure of code's susceptibility to errors. The optimal code for genes located on one DNA strand was simultaneously worse than the universal code for the genes located on the other strand. Furthermore, we generated 20 million random codes of which only 23 were better than the universal one for genes located on both strands simultaneously while about two orders of magnitude more codes were better for each of the two strands separately. The result indicates that the existing universal code, the mutational pressure, the codon and amino acid compositions are highly optimised for the both differently replicating DNA strands.

Keywords: genetic code, error minimization, adaptation, asymmetric mutational pressure, amino acid usage, leading strand, lagging strand.

1 Introduction

There are three main groups of hypotheses trying to explain the origin and evolution of the genetic code: chemical, historical and adaptive (see for review [1,2,3]). The first one assumes some structural and physicochemical relationships and interactions between stretches of RNA (codons, anticodons, reversed codons, codon-anticodon double helices etc.) and coded amino acids [4,5,6]. So far, a well-confirmed relationship has been found for seven of eight amino acids (see for review: [7]). The second hypothesis states that codons in the simpler, ancestral genetic code coded for only a small subset of amino acids and later, along with the evolution of biochemical organization of primary cells, newly synthesised amino acids took over the codons from the amino acids to which they were related in biosynthetic pathways [8,9,10,11,12]. The third group of hypotheses assumes that the codon assignments could initially vary and it was the selection pressure

M. Bubak et al. (Eds.): ICCS 2008, Part III, LNCS 5103, pp. 100–109, 2008.

which optimized the code to reduce harmful effects of mutations occurring during replication and transcription (lethal-mutation hypothesis) and to minimize errors during translation process (translational-error hypothesis), [6,13,14,15,16,17,18]; see for review: [19].

Primordial organisms whose code reduced the deleterious effects of errors won eventually the competition and survived. During further evolution connected with the increase of genome size, the genetic code was frozen [20] and it was not possible to re-interpret the meaning of any codon because the whole complex translational machinery was already adapted to the code and every such change would have catastrophic consequences for the organisms. Nevertheless, some optimization took place already in the first stages of the genetic code evolution, probably before ,,freezing". One optimization results directly from the simple structural relationships between nucleotides in the double helix - one large and one small nucleotide fit better to form a pair. Thus, transitions which happen with much higher frequency than transversions have much less deleterious mutational effect than transversions. Actually, it was shown that the genetic code is well adapted to the transition/transversion bias [16].

Assuming the adaptive hypothesis of the genetic code evolution we expect that if the genetic code was ,,frozen" at an early stage of evolution when genomes were relatively small, it is the code itself that imposes further restrictions on the mutational pressure, amino acid and codon usage, and the translational machinery in order to minimize the deleterious effects of mutations. Thus, the mutational pressure cannot be completely random, as one could claim, but it is highly biased and it cooperates with the selection pressure on amino acid and codon usage to minimise the harmful effects of nucleotide substitutions. One of the premises is that the most ,,mutable" codons in the genome correspond to the least-represented amino acids [21,22]. Monte Carlo simulations showed that changing of parameters of any of the three counterparts of the coding functions: relative nucleotide substitution rates in the mutational pressure, the way the genetic code is degenerated or the amino acid composition of proteomes increases the deleterious effects of mutations in studied genomes [23].

However, it is not simply to optimise the mutational pressure. The mutational pressures acting on the differently replicating (leading or lagging) DNA strands show different patterns of nucleotide substitutions and leads to the strong bias (asymmetry) in nucleotide composition between the two DNA strands observed in almost all bacterial chromosomes [24,25,26,27,28,29] and long regions of eukaryotic chromosomes [30,31,32]. Therefore, genes are subjected to different mutational pressures depending on their location on the differently replicating DNA strands, which affects their codon usage and amino acid composition of coded proteins [33,34,35,36].

Although several simulation studies about the optimization of the genetic code were carried out [14,15,16,17,18], none of them considered the real and global genomic aspect of this optimization, i.e. the real mutational pressure, gene content, codon and amino acid usage. Zhu *et al.* [37] found that the universal genetic code appears to be less optimised for error minimization when specific codon

usage for particular species was considered. However, the authors applied the same and simple mutation pattern for all species in this analysis, which do not fit to the specific codon usage and they concluded that the specific mutation pattern should be taken into account. In this paper we considered the optimization of the genetic code in the context of the two different mutational pressures specific for the leading and lagging DNA strands acting on the asymmetric genome of *Borrelia burgdorferi*. This genome shows the strongest asymmetry between the leading and lagging strands detected so far [33,34,38] thus, it is suitable for such studies.

2 Materials and Methods

All our analyses were performed on the *B. burgdorferi* genome [39] whose sequence and annotations were downloaded from GenBank [40]. Based on these data we calculated the content of codons, amino acids and codon usage for 564 leading strand genes and 286 lagging strand genes. The mutational pressure characteristic for this genome was found by Kowalczuk *et al.* [41]. The pressure was described by the nucleotide substitution matrices (M_n) as follows:

$$M_n = \begin{bmatrix} 1-pR_A & pR_{AC} & pR_{AG} & pR_{AT} \\ pR_{CA} & 1-pR_C & pR_{CG} & pR_{CT} \\ pR_{GA} & pR_{GC} & 1-pR_G & pR_{GT} \\ pR_{TA} & pR_{TC} & pR_{TG} & 1-pR_T \end{bmatrix}$$

where: p is the overall mutation rate; R_{ij} for $i, j = A, C, G, T$ and $i \neq j$ is the relative rate of substitution of the nucleotide i by the nucleotide j; R_i (in the diagonal) for $i = A, C, G, T$ represents the relative substitution rate of nucleotide i by any of the other three nucleotides.

$$R_i = \sum_{i \neq j} R_{ij}$$

and $R_A + R_C + R_G + R_T = 1$. For $p = 1$ the matrix describing the leading strand mutational pressure is:

$$M_n^{leading} = \begin{bmatrix} 0.808 & 0.023 & 0.067 & 0.103 \\ 0.070 & 0.621 & 0.047 & 0.261 \\ 0.164 & 0.015 & 0.706 & 0.116 \\ 0.065 & 0.035 & 0.035 & 0.865 \end{bmatrix}$$

The matrix represents the most probable pure mutational pressure associated with replication acting on the leading strand. Because DNA strands are complementary, the mutational pressure acting on the lagging strand is a kind of the mirror reflection of the pressure exerted on the leading strand, e.g. R_{GA} for the leading strand corresponds to R_{CT} for the lagging strand etc. In our analyses we have assumed p = 10^{-8} which approximately corresponds to the observed number of substitutions in a bacterial genome per nucleotide per generation [42].

The codon substitution matrix (M_c) containing relative rate of substitutions of one codon by another one was derived from the nucleotide substitution matrix (M_n). The M_c is the Kronecker product of three M_n matrices: $M_c = M_n \otimes M_n \otimes M_n$. For example, the substitution rate of codon GCA to codon CTA equals $R^c_{GCA\to CTA} = p^2 R_{GC} R_{CT}(1 - pR_A)$. In the M_c each row contains the substitution rates for one of 64 codons to another one:

$$M_c = \begin{bmatrix} R^c_{AAA\to AAA} & R^c_{AAA\to AAC} & R^c_{AAA\to AAG} & \cdots & R^c_{AAA\to TTT} \\ R^c_{AAC\to AAA} & R^c_{AAC\to AAC} & R^c_{AAC\to AAG} & \cdots & R^c_{AAC\to TTT} \\ R^c_{AAT\to AAA} & R^c_{AAT\to AAC} & R^c_{AAT\to AAG} & \cdots & R^c_{AAT\to TTT} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R^c_{TTT\to AAA} & R^c_{TTT\to AAC} & R^c_{TTT\to AAG} & \cdots & R^c_{TTT\to TTT} \end{bmatrix}$$

where: $R^c_{n\to m}$ for indices of codons $n, m \in \{1..64\}$ represents the relative rate of substitution of codon n by codon m.

Each row of M_c was multiplied by the codon usage of a given codon U_n (i.e. relative frequency of a codon among other synonymous codons coding the same amino acid or stop codon) giving the M_u matrix:

$$M_u = \begin{bmatrix} U_{AAA}R^c_{AAA\to AAA} & U_{AAA}R^c_{AAA\to AAC} & \cdots & U_{AAA}R^c_{AAA\to TTT} \\ U_{AAC}R^c_{AAC\to AAA} & U_{AAC}R^c_{AAC\to AAC} & \cdots & U_{AAC}R^c_{AAC\to TTT} \\ U_{AAT}R^c_{AAT\to AAA} & U_{AAT}R^c_{AAT\to AAC} & \cdots & U_{AAT}R^c_{AAT\to TTT} \\ \vdots & \vdots & \ddots & \vdots \\ U_{TTT}R^c_{TTT\to AAA} & U_{TTT}R^c_{TTT\to AAC} & \cdots & U_{TTT}R^c_{TTT\to TTT} \end{bmatrix}$$

where: U_n stands for codon usage of codon n, where $n \in \{1..64\}$.

To obtain the amino acid substitution matrix (M_a) containing relative rates of substitutions of one amino acid or stop by another, the respective elements of M_u matrix were summed up, which gives the matrix of amino acids (and stops) substitution:

$$M_a = \begin{bmatrix} R^a_{Ala\to Ala} & R^a_{Ala\to Arg} & R^a_{Ala\to Asn} & \cdots & R^a_{Ala\to Val} \\ R^a_{Arg\to Ala} & R^a_{Arg\to Arg} & R^a_{Arg\to Asn} & \cdots & R^a_{Arg\to Val} \\ R^a_{Asn\to Ala} & R^a_{Asn\to Arg} & R^a_{Asn\to Asn} & \cdots & R^a_{Asn\to Val} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R^a_{Val\to Ala} & R^a_{Val\to Arg} & R^a_{Val\to Asn} & \cdots & R^a_{Val\to Val} \end{bmatrix}$$

where: $R^a_{p\to q}$ for $p, q \in \{1..21\}$ represents the relative rate of substitution of amino acid (or stop) p by amino acid q.

The sum of each row of M_a gives the rate of substitution of amino acid p (or stop) p to another:

$$R^a_p = \sum_{q\neq p} R^a_{p\to q}.$$

Such calculations were carried out for the leading strand data and for the lagging strand data separately.

3 Results and Discussion

In order to estimate how the genetic code and the mutational pressures are optimized for differently replicating strands, we considered the number of substituted amino acids (and stops). Therefore we multiplied each rate of substitution of a given amino acid R_p^a by the number A_p of this amino acid in the coded proteins and summed the products:

$$S_A = \sum_{p=1}^{21} (A_p R_p^a).$$

In our consideration we applied the number of substituted amino acids instead of fraction because we wanted to analyse the genetic code optimization in the context of the whole genome including the bias between the numbers of the leading and lagging strand genes. For constant R_p^a and A_p, S_A reaches the minimum if $A_p < A_{p+1} < A_{p+N}$ and simultaneously if $R_p^a > R_{p+1}^a > R_{p+N}^a$ i.e. when A_p and S_A are negatively correlated. In other words the total cost of mutations is lower if the rate of mutation is higher for the less frequent residues than for the more frequent ones. Interestingly, A_p and S_A calculated for the real genome data show statistically significant negative correlation (Fig. 1) that suggests a tendency to minimization of amino acid substitutions in the real genome.

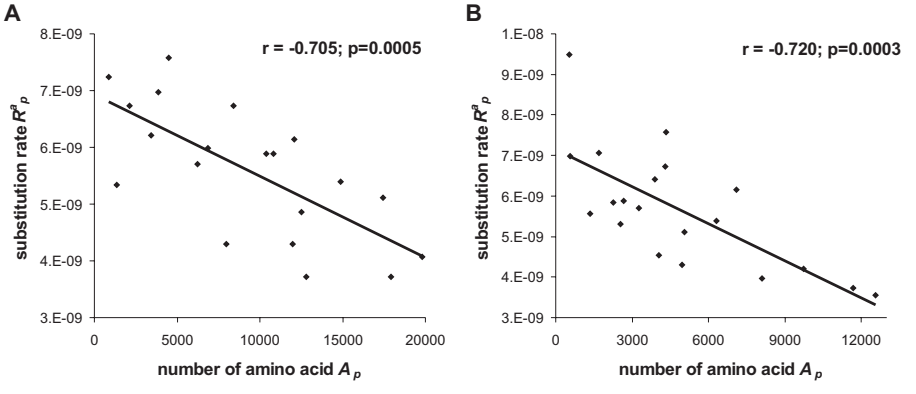

Fig. 1. Correlation between substitution rates R_p^a and the number of amino acids A_p for the leading strand (A) and for the lagging strand (B) data

However, it is possible to find such ascription of amino acids to codons (i.e. to elaborate a new genetic code) which is better optimized than the universal code - according to the minimization of the number of amino acid substitutions. The best way is to rank reversely R_p^a versus A_p. Similarly, one can obtain the worst code giving the highest number of amino acid substitutions by the ranking of A_p and S_A accordingly. The results of such transformations made separately for the leading and for the lagging strand cases are shown in Table 1. Such a

transformation did not change the global structure of the genetic code, retains the same degeneracy level and redundancy of the canonical genetic code. It assumes that the groups of codons which code for a single amino acid are fixed and we changed only the assignments between amino acids and the codon blocks. For example, in the case of the leading strand, tryptophane is now coded by four proline codons and proline by one methionine codon. The ascriptions of some amino acids were changed but some of them retained their positions. Because stop codons have special meanings and are represented by only one per gene, we did not change the position of the stops in this transformation.

Table 1. The ascription of codons of a given amino acid that minimizes the number of substitutions separately for the leading and lagging strand cases. Amino acids that have not changed their position are in **bold**.

amino acid	A	**R**	**N**	D	C	Q	E	G	H	**I**	L	**K**	M	**F**	P	S	T	**W**	Y	V
leading strand	G	**R**	**N**	S	W	A	K	T	Q	**I**	F	L	H	Y	M	V	E	P	D	C
lagging strand	H	A	L	R	G	D	T	S	M	**I**	N	**K**	C	**F**	E	Y	Q	**W**	V	P

Table 2 shows the expected number of substituted amino acids (including stops) S_A - calculated for the universal genetic code, for the best one and for the worst one for the giving DNA strand. It is possible to find the optimal code for protein coding sequences located on one DNA strand but such a code is not simultaneously the optimal one for genes located on the other DNA strand. In fact it is worse than the universal one. Nevertheless, the S_A values for the universal code and the both classes of genes are much closer to the best code than to the worst one. The value of S_A for the universal code fall between the values for optimal codes for DNA strands.

Table 2. The expected number of missense mutations (including stops) S_A calculated for the universal genetic code, for the best one and for the worst one for the leading and lagging strand proteins

		Code optimal for:		The worst code
DNA strand	Universal code	leading strand	lagging strand	for the giving strand
leading	0.000955	0.000921	0.000974	0.001168
lagging	0.000482	0.000488	0.000465	0.000645

As it was shown above it is easy to find the optimal code for each strand separately but it is difficult to calculate the code that would be optimal for the two strands simultaneously. To solve the problem we have generated 20 million random genetic codes replacing one amino acid by another one as described previously, i.e. retaining the global structure of the genetic code retaining the same degeneracy level and redundancy. Such a transformation corresponds to the method widely used in other studies [14,15,16,17]. For each generated code we calculated the number of substituted amino acids (excluding stops) S_A separately for genes located on the leading and lagging strands. Next we counted for these

two sets of genes how many codes produce the S_A value smaller than the value for the universal code and we counted how many random codes are better for both sets. In the last case we considered two conditions:

1. total number of substitutions (i.e. the sum of the S_A for the leading and for the lagging strand) produced by a generated code is smaller than under the universal code;
2. generated code is better simultaneously for each of the two strands.

The first condition treats the leading and the lagging strand genes as one set whereas the second one treats them as separate, independent sets. The results are

Table 3. The number of random (generated) codes (among 20 million) which are better than the universal one according to the number of amino acid substitutions analysed in the aspect of the differently replicating strands. $S^{random}_{A\_leading}$ - the number of substituted amino acids in the leading strand proteins considering the random code; $S^{random}_{A\_lagging}$ - the number of substituted amino acids in the lagging strand proteins considering the random code; $S^{universal}_{A\_leading}$ - the number of substituted amino acids in the leading strand proteins considering the universal code; $S^{universal}_{A\_lagging}$ - the number of substituted amino acids in the lagging strand proteins considering the universal code.

Checked condition	The number of better codes
$S^{random}_{A\_leading} < S^{universal}_{A\_leading}$	6652
$S^{random}_{A\_lagging} < S^{universal}_{A\_lagging}$	733
$S^{random}_{A\_leading} + S^{random}_{A\_lagging} < S^{universal}_{A\_leading} + S^{universal}_{A\_lagging}$	160
$S^{random}_{A\_leading} < S^{universal}_{A\_leading}$ and $S^{random}_{A\_lagging} < S^{universal}_{A\_lagging}$	23

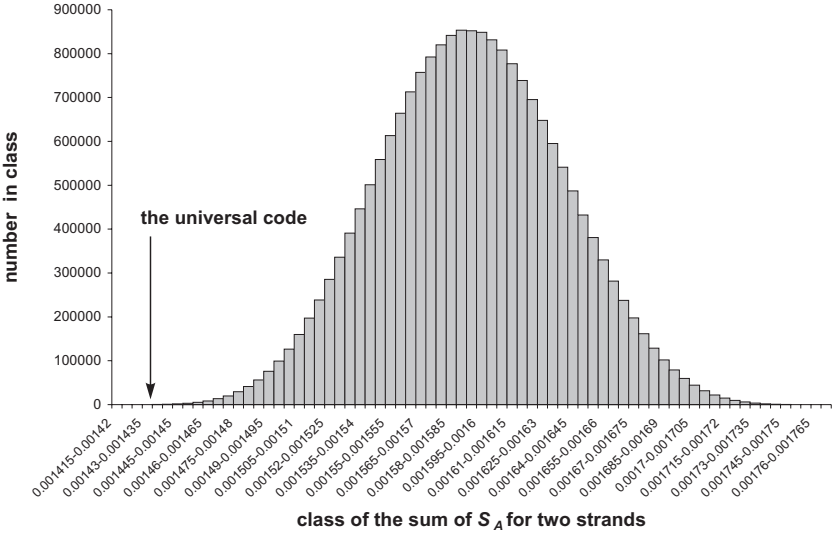

Fig. 2. Distribution of the sum of the number of substituted amino acids S_A for the leading and the lagging strands calculated for 20 million randomly generated genetic codes. The arrow indicates the S_A value for the universal code.

presented in Table 3. We have found that the probability of random generation of a code which would transmit fewer missense mutations in the set of genes located on the leading strand or the lagging strand is relatively high. Nevertheless, we have found much fewer codes which fulfil the first condition (160, i.e. 0.0008%) (Fig. 2) and even fewer, which fulfil the second condition (23, i.e. 0.000115%). The observed optimality of the code is very close to the results obtained by Freeland and Hurst [16], i.e. one per million.

4 Conclusions and Perspectives

The results indicate that the existing universal code, the mutational pressure, the codon and amino acid composition are highly optimised in the context of the two differently replicating DNA strands minimizing the number of substituted amino acids in the coded proteins. In our studies we assumed quite simple measure of a code's susceptibility to errors - number of substituted amino acids - ignoring the differences in their physicochemical properties, e.g. hydrophobicity, polarity or isoelectric point. This simplification enabled to calculate analytically the optimal and the worst assignments of amino acids to codons and to compare them with the result obtained for universal genetic code considering mutational pressure, codon usage and amino acid composition specific for genes lying on differently replicating strands. However, considering of these physicochemical properties would probably decrease the number of random codes better than the universal one and could be further investigated. It would be also interesting to analyze genomic systems of other organisms in this aspect. A better code for one organism could be worse for another organism. If one wanted to look for the optimal code for all organisms, one should check each organism separately - its mutational pressure, amino acid composition and codon usage. It makes no sense to look for a genetic code that would be better for average codon usage or average mutational pressure. There are no average organisms in the biosphere. In the early stages of genetic code evolution the code optimised itself to minimizing harmful effects of various mutational pressures but after it was ,,frozen" the mutational pressure begun to tune to the universal code independently in different phylogenetic lineages.

Acknowledgements. The work was done in the frame of the ESF program COST Action P10, GIACS and UNESCO chair of interdisciplinary studies.

References

1. Di Giulio, M.: On the origin of the genetic code. J. Theor. Biol. 187, 573–581 (1997)
2. Knight, R.D., Freeland, S.J., Landweber, L.F.: Selection, history and chemistry: the three faces of the genetic code. Trends Biochem. Sci. 24, 241–247 (1999)
3. Knight, R.D.: The origin and evolution of the genetic code: statistical and experimental investigations. Ph.D. Thesis, Department of Ecology and Evolutionary Biology, Princeton University (2001)

4. Dunnill, P.: Triplet nucleotide-amino-acid pairing a stereochemical basis for the division between protein and non-protein amino-acids. Nature 210, 1265–1267 (1966)
5. Pelc, S.R., Welton, M.G.: Stereochemical relationship between coding triplets and amino-acids. Nature 209, 868–872 (1966)
6. Woese, C.R., Dugre, D.H., Saxinger, W.C., Dugre, S.A.: The molecular basis for the genetic code. Proc. Natl. Acad. Sci. USA 55, 966–974 (1966)
7. Yarus, M., Caporaso, J.G., Knight, R.: Origins of the genetic code: the escaped triplet theory. Annu. Rev. Biochem. 74, 179–198 (2005)
8. Wong, J.T.-F.: A Co-Evolution Theory of the Genetic Code. Proc. Natl. Acad. Sci. USA 72, 1909–1912 (1975)
9. Taylor, F.J.R., Coates, D.: The code within the codons. BioSystems 22, 177–187 (1989)
10. Di Giulio, M.: On the relationships between the genetic code coevolution hypothesis and the physicochemical hypothesis. Z. Naturforsch. [C] 46, 305–312 (1991)
11. Amirnovin, R.: An analysis of the metabolic theory of the origin of the genetic code. J. Mol. Evol. 44, 473–476 (1997)
12. Di Giulio, M., Medugno, M.: The Historical Factor: The Biosynthetic Relationships Between Amino Acids and Their Physicochemical Properties in the Origin of the Genetic Code. J. Mol. Evol. 46, 615–621 (1998)
13. Alff-Steinberger, C.: The genetic code and error transmission. Proc. Natl. Acad. Sci. USA 64, 584–591 (1969)
14. Haig, D., Hurst, L.D.: A quantitative measure of error minimization in the genetic code. J. Mol. Evol (Erratum in J. Mol. Evol. 49, 708 (1999)) 33, 412–417 (1991)
15. Ardell, D.H.: On error minimization in a sequential origin of the standard genetic code. J. Mol. Evol. 47, 1–13 (1998)
16. Freeland, S.J., Hurst, L.D.: The genetic code is one in a million. J. Mol. Evol. 47, 238–248 (1998)
17. Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D.: Early fixation of an optimal genetic code. Mol. Biol. Evol. 17, 511–518 (2000)
18. Sella, G., Ardell, D.H.: The impact of message mutation on the fitness of a genetic code. J. Mol. Evol. 54, 638–651 (2002)
19. Freeland, S.J., Wu, T., Keulmann, N.: The case for an error minimizing standard genetic code. Orig. Life Evol. Biosph. 33, 457–477 (2003)
20. Crick, F.H.: The origin of the genetic code. J. Mol. Evol. 38, 367–379 (1968)
21. Nowicka, A., Mackiewicz, P., Dudkiewicz, M., Mackiewicz, D., Kowalczuk, M., Cebrat, S., Dudek, M.R.: Correlation between mutation pressure, selection pressure, and occurrence of amino acids. In: Sloot, P.M.A., Abramson, D., Bogdanov, A.V., Gorbachev, Y.E., Dongarra, J., Zomaya, A.Y. (eds.) ICCS 2003. LNCS, vol. 2658, pp. 650–657. Springer, Heidelberg (2003)
22. Nowicka, A., Mackiewicz, P., Dudkiewicz, M., Mackiewicz, D., Kowalczuk, M., Banaszak, J., Cebrat, S., Dudek, M.R.: Representation of mutation pressure and selection pressure by PAM matrices. Applied Bioinformatics 3, 31–39 (2004)
23. Dudkiewicz, M., Mackiewicz, P., Nowicka, A., Kowalczuk, M., Mackiewicz, D., Polak, N., Smolarczyk, K., Banaszak, J., Dudek, M.R., Cebrat, S.: Correspondence between mutation and selection pressure and the genetic code degeneracy in the gene evolution. FGCS 21, 1033–1039 (2005)
24. Lobry, J.R.: Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665 (1996)
25. Mrazek, J., Karlin, S.: Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95, 3720–3725 (1998)

26. Frank, A.C., Lobry, J.R.: Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238, 65–77 (1999)
27. Tillier, E.R., Collins, R.A.: The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. 50, 249–257 (2000)
28. Kowalczuk, M., Mackiewicz, P., Mackiewicz, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., Cebrat, S.: DNA asymmetry and the replicational mutational pressure. J. Appl. Genet. 42, 553–577 (2001a)
29. Rocha, E.P.: The replication-related organization of bacterial genomes. Microbiology 150, 1609–1627 (2004)
30. Gierlik, A., Kowalczuk, M., Mackiewicz, P., Dudek, M.R., Cebrat, S.: Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? J. Theor. Biol. 202, 305–314 (2000)
31. Niu, D.K., Lin, K., Zhang, D.Y.: Strand compositional asymmetries of nuclear DNA in eukaryotes. J. Mol. Evol. 57, 325–334 (2003)
32. Touchon, M., Nicolay, S., Audit, B., Brodie of Brodie, E.B., d'Aubenton-Carafa, Y., Arneodo, A., Thermes, C.: Replication-associated strand asymmetries in mammalian genomes: Toward detection of replication origins. Proc. Natl. Acad. Sci. USA 102, 9836–9841 (2005)
33. McInerney, J.O.: Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proc. Natl. Acad. Sci. USA 95, 10698–10703 (1998)
34. Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., Wolfe, K.H.: Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. 27, 1642–1649 (1999)
35. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., Cebrat, S.: How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. 9, 409–416 (1999a)
36. Rocha, E.P., Danchin, A., Viari, A.: Universal replication biases in bacteria. Mol. Microbiol. 32, 11–16 (1999)
37. Zhu, C.T., Zeng, X.B., Huang, W.D.: Codon usage decreases the error minimization within the genetic code. J. Mol. Evol. 57, 533–537 (2003)
38. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Szczepanik, D., Dudek, M.R., Cebrat, S.: Mechanisms generating long-range correlation in nucleotide composition of the Borrelia burgdorferi genome. Physica A 273, 103–115 (1999b)
39. 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. (38 co-authors): Genomic sequence of a Lyme disease spirochaete Borrelia burgdorferi. Nature 390, 580–586 (1997)
40. GenBank, `ftp://www.ncbi.nlm.nih.gov`
41. 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, 13 (2001b)
42. Drake, J.W., Charlesworth, B., Charlesworth, D., Crow, J.F.: Rates of spontaneous mutation. Genetics 148, 1667–1686 (1998)