Correlation between Mutation Pressure, Selection Pressure, and Occurrence of Amino Acids

Aleksandra Nowicka[1], Paweł Mackiewicz[1], Małgorzata Dudkiewicz[1], Dorota Mackiewicz[1], Maria Kowalczuk[1], Stanisław Cebrat[1], and Mirosław R. Dudek[2]*

[1] Department of Genetics, Institute of Microbiology, University of Wroclaw,
ul. Przybyszewskiego 63/77, PL-54148 Wroclaw, Poland
{nowicka, pamac, malgosia, dorota, kowal, 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. With the help of the empirical mutation table for nucleotides in the *Borrelia burgdorferi* genome we have performed Monte Carlo simulation of the pure mutation pressure experienced by the genes of the genome. We have examined the divergence of the mutated genes from the ancestral ones and we have constructed MPM1 matrix (Mutation Probability Matrix) of the substitution rates between amino acids of the diverging genes. The results have been compared to mutation data matrix PAM1 PET91 representing mutation and selection data of 16130 homologous genes od different organisms. We have found that the effective survival time of amino acids in organisms follows a power law with respect to frequency of their occurrence in genes. This makes possible to find the effect of the pure mutational pressure and the selection on the amino acid composition of genes. The results are universal in the sense that the survival time of amino acids calculated from the higher order PAMk matrices ($k > 1$) follows the same power law as in the case of PAM1 matrices.

1 Introduction

Determining the evolutionary distances between two protein sequences requires the knowledge of the substitution rates of amino-acids. It is generally accepted that the more substitutions are necessary to change one sequence into another, the more unrelated they are and the larger their distance to the common ancestor. The most widely used method for the calculation of distances between sequences is based on the mutation data matrix, M_{ij}, published by Dayhoff et al. [1], where i, j represent amino acids, and an element M_{ij} of the matrix gives

* corresponding author

P.M.A. Sloot et al. (Eds.): ICCS 2003, LNCS 2658, pp. 650–657, 2003.

the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary time interval. The interval corresponding to 1 percent of substitutions between two compared sequences is called one PAM (Percent of Accepted Mutations), and the corresponding matrix is denoted as PAM1 matrix. There is assumed a Markov model of sequence evolution and a simple power M^k of the PAM1 matrix (multiplied by itself k times) denotes a matrix, PAMk, that gives the amino acid substitution probability after k PAMs. Today, a much more accurate PAM matrix is available, generated from 16130 protein sequences, published by Jones et al. [2]. The large number of compared genes guarantees that the matrix has negligible statistical errors and it can be considered to be the reference matrix during the calculations of the phylogenetic distances. The matrix is also known as PET91 matrix.

Recently, by comparing intergenic sequences being remnants of coding sequences with homologous sequences of genes, we have constructed an empirical table of the nucleotide substitution rates in the case of the leading DNA strand of the *B. burgdorferi* genome [3],[4],[5]. We have found that substitution rates, which determine the evolutionary turnover time of a given kind of nucleotide in third codon positions of coding sequences, are highly correlated with the frequency of the occurrence of that nucleotide in the sequences. There is a compositional bias produced by replication process, introducing long-range correlation among nucleotides in the third positions in codons, which is very similar to the bias seen in the intergenic sequences [6].

We have used the empirical table of nucleotide substitution rates to simulate mutational pressure on the genes lying on the leading DNA strand of the *B.burgdorferi* genome and we have constructed MPM1 matrix (Mutation Probability Matrix) for amino acid substitutions in the evolving genes. Thus the resulting table represents the percent of amino acid substitutions introduced by mutational pressure and not by selection. Next, we compared the survival times of the amino acids in the case without any selection with the effective survival times of the amino acids, counted with the help of the PAM1 PET91 matrix.

2 Mutation Table for Nucleotides

DNA sequence of the *B.burgdorferi* genome was downloaded from the website *www.ncbi.nlm.nih.gov.* The empirical mutation table for nucleotides in third positions in codons, which we used in the paper, is the following [3],[4]:

$$M = \begin{pmatrix} 1-uW_A & u\,W_{AT} & u\,W_{AG} & u\,W_{AC} \\ u\,W_{TA} & 1-uW_T & u\,W_{TG} & u\,W_{TC} \\ u\,W_{GA} & u\,W_{GT} & 1-uW_G & u\,W_{GC} \\ u\,W_{CA} & u\,W_{CT} & u\,W_{CG} & 1-uW_C \end{pmatrix} \qquad (1)$$

where [1]

$$\begin{array}{llll} W_{GA} = 0.0667 & W_{GT} = 0.0347 & W_{GC} = 0.0470 & W_{AG} = 0.1637 \\ W_{AT} = 0.0655 & W_{AC} = 0.0702 & W_{TG} = 0.1157 & W_{TA} = 0.1027 \\ W_{TC} = 0.2613 & W_{CG} = 0.0147 & W_{CA} = 0.0228 & W_{CT} = 0.0350 \end{array} \qquad (2)$$

[1] The transpose matrix convention has been chosen in [3].

and the elements of the matrix give the probability that nucleotide in column j will mutate to the nucleotide in row i during one replication cycle. The symbols W_{ij} represent relative substitution probability of nucleotide j by nucleotide i, and u represents mutation rate. The symbols W_j in the diagonal represent relative substitution probability of nucleotide j:

$$W_j = \sum_{i \neq j} W_{ij}, \tag{3}$$

and

$$W_A + W_T + W_G + W_C = 1. \tag{4}$$

The expression for the mean survival time of the nucleotide j depends on W_j as follows (derivation can be found in [3])

$$\tau_j = -\frac{1}{ln(1 - u\ W_j)} \approx \frac{1}{u\ W_j}. \tag{5}$$

The above approximated formula is true for small values of the mutation rate u.

In papers [3],[4],[5], we concluded that in a natural genome the frequency of occurrence f_j of the nucleotides, in the third position in codons, is linearly related to the respective mean survival time τ_j,

$$f_j = m_0\ \tau_j + c_0, \tag{6}$$

with the same coefficients, m_0 and c_0, for each nucleotide. The Kimura's neutral theory [7] of evolution assumes the constancy of the evolution rate, where the mutations are random events, much the same as the random decay events of the radioactive decay. However, the linear law in (6) is not contrary to the Kimura's theory. Still, the mutations represent random decay events but they are correlated with the DNA composition.

3 Mutational Pressure MPM1 Matrix Construction for Amino Acids

In order to compare an effect of the pure mutational pressure and the selection pressure on amino acid composition of genes we used the results of Monte Carlo simulation of the mutational pressure applied on 564 genes from the leading DNA strand of the *B. burdorferi* genome. This enabled us to calculate amino acid substitution rates which, next we could compare with the ones originating from the PAM1 PET91 substitution rates matrix [2]. The way, the experimental PET91 matrix has been constructed, determined our simulation algorithm, which consisted of the following steps:

(i) for each gene, considered to be an ancestral one, make two copies of the gene at $t = 0$,

(ii) increase time step t ($t = t + 1$) and with frequency u mutate nucleotides of the two gene copies with the probability distribution defined by the elements of the mutation matrix in (1),

(iii) goto (ii) unless the number of amino acid substitutions between the homologous protein sequences reaches 1%,

The applied value of the mutation pressure was $u = 0.01$. The steps (i)–(iii) have been repeated 10^5 times in order to calculate the averaged values of the substitution rates between the homologous protein sequences. These values we used to construct a mutation probability matrix MPM1 according to the procedure of Dayhoff et al.[1] and Jones et al. [2]. The resulting mutation table, with substitution probabilities M_{ij}, the amino acid mutability m_j, and the fraction f_j of amino-acid in the compared sequences have been presented, respectively, in Table 1 and Table 2.

The elements M_{ij} of the MPM1 matrix in Table 1 have been scaled with the parameter λ, which related them to the evolutionary distance of one percent of substitutions and it is equal to 0.00009731 in our simulations.

4 Discussion of Results

The major qualitative difference between the MPM matrix introduced in the previous section and the PAM1 PET91 matrix published in the paper by Jones et al.[2] is that the first one is a result of pure mutational pressure whereas the second one is a result of both mutational and selection pressures. Thus, we have two evolutionary mechanisms responsible for the resulting PAM matrices.

With the help of formula (5) (extended to amino acids) we have calculated effective survival times of amino acids in the case of the MPM1 matrix (Table 1) and the mutational/selectional PAM1 PET91 matrix ([2]). The value of the parameter λ is a counterpart of u in (5). In Fig.1, we presented the relation between the calculated survival time o amino acids and their fractions in the *B. burgdorferi* proteins, in the pairs of diverged genes, in a log-log scale. One can observe that the data are highly correlated and in both cases the dependence of the mean survival time of amino acid on the fraction of the amino acid represents a power law:

$$\tau_j \sim F_j^{\alpha} \tag{7}$$

with a negative value of $\alpha \approx -1.3$ in the case of selection and a positive value of $\alpha \approx 0.2$ in the case of mutation pressure on the leading DNA strand of the *B. burgdorferi* genome. The value of α for the analogous mutational PAM1 matrix calculated in the case of the lagging DNA strand of the *B. burgdorferi* genome is about twice as small. It is worth to underline that the slopes α are the same for the matrices PAMk with high values of k, and thus, they are universal with respect to evolution.

In Fig.1 we may observe a kind of evolutional scissors acting on amino acids. Once the less frequent amino acids, like W (tryptophane), C (cysteine), have much shorter turn over time compared with other amino acids (as can be seen from the lower line) the selection pressure (upper line) counteracts with the effect. On the other hand, the most mutable amino acids, like L (leucine) or I (isoleucine) , which are very frequent in genes, seem to be much weakly influenced by selection.

Table 1. Simulated Mutation Probability Matrix for an evolutionary distance of 1 PAM (splitted into two parts). Values of the matrix elements are scaled by a factor of 10^5 and rounded to an integer. The symbols in the first row and the first column represent amino-acids and numbers following colons -number of codons representing a given amino-acid in the universal genetic code.

	A:4	R:6	N:2	D:2	C:2	Q:2	E:2	G:4	H:2	I:3	L:6	K:2	M:1	F:2	P:4	S:6	T:4	W:1	Y:2	V:4
A:4	99027	0	0	60	1	0	44	56	0	0	0	0	0	0	79	94	304	0	0	230
R:6	0	98784	1	0	168	91	1	151	123	39	21	129	93	0	54	101	35	334	0	0
N:2	0	1	98925	255	3	1	1	1	217	126	0	213	1	0	0	118	93	0	241	1
D:2	77	0	220	98935	2	0	232	157	129	0	0	1	0	0	0	0	0	0	203	145
C:2	0	33	0	0	97443	0	0	64	1	0	0	0	0	50	0	66	0	199	100	0
Q:2	0	51	0	0	0	99243	32	0	350	0	25	41	0	0	47	0	0	0	0	0
E:2	63	1	1	258	0	99	99132	173	1	0	0	191	1	0	0	0	0	1	1	131
G:4	69	227	0	151	483	0	149	99089	0	0	0	0	0	0	0	82	0	282	0	143
H:2	0	38	37	25	1	194	0	0	98313	0	35	0	0	0	60	0	0	0	110	0
I:3	1	103	180	1	2	0	0	1	1	99025	106	98	557	178	1	74	366	0	1	261
L:6	1	61	0	0	3	130	0	0	322	116	99267	0	228	412	342	94	1	304	1	124
K:2	0	325	295	1	0	184	276	1	1	95	0	99245	231	0	0	1	97	1	1	1
M:1	0	47	0	0	0	0	0	0	0	106	40	46	98686	0	0	0	38	1	0	38
F:2	1	0	0	0	467	0	0	0	2	128	268	0	1	98928	1	153	1	1	221	123
P:4	42	35	0	0	1	54	0	0	126	0	77	0	0	0	98948	107	77	0	0	0
S:6	166	219	140	1	717	0	0	118	1	61	71	0	1	177	359	98951	261	47	85	1
T:4	225	31	46	0	1	0	0	0	0	127	0	35	69	0	108	109	98724	0	0	1
W:1	0	43	0	0	127	0	0	24	0	0	13	0	0	0	0	3	0	98827	0	0
Y:2	0	0	153	150	580	1	0	0	414	0	0	0	0	137	0	46	0	1	99033	0
V:4	329	1	1	162	1	0	132	166	0	175	76	0	133	115	0	1	2	1	1	98802

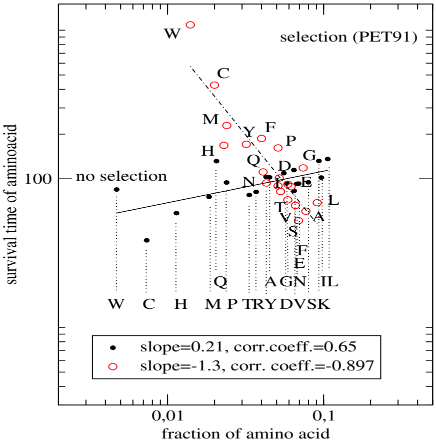

Fig. 1. Relation between survival time of amino-acids and their fractions in compared pairs of diverged homologous genes in the case with selection (PET91) and in the case without selection (simulated mutational pressure of the *B. burgdorferi* genome).

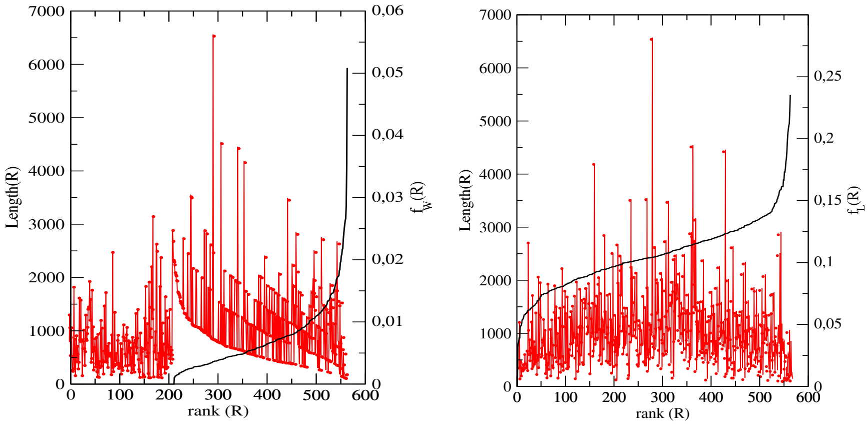

Fig. 2. Gene length versus gene rank, where each gene has assigned a rank with respect to fraction of tryptophane (left graph) and with respect to leucine (right graph). In each graph there are two plots: the dots represent gene's size vs. rank, whereas the second plot represent the fraction vs. rank.

Table 2. Relative mutabilities and fractions of 20 amino acids in the compared simulated sequences. We used the convention that mutabilities are relative to alanine and it is arbitrarily assigned a value of 100.

amino acid	relative mutability (m_j)	fraction (f_j)
A	100.00	0.0449
R	126.09	0.0369
N	110.42	0.0671
D	109.29	0.0579
C	262.91	0.0073
Q	77.67	0.0206
E	89.88	0.0644
G	92.21	0.0582
H	173.18	0.0114
I	100.12	0.0964
L	75.23	0.1060
K	77.51	0.0930
M	135.01	0.0184
F	110.12	0.0690
P	108.02	0.0238
S	107.78	0.0796
T	131.05	0.0333
W	127.44	0.0047
Y	99.25	0.0427
V	123.56	0.0645

In genes the fraction of the amino acids most protected by selection strongly depends on gene size, it is diminishing when gene's size is increasing. The effect weakens if we go into right the direction of the evolutionary scissors in Fig.1. To show this, we ordered all 564 genes under consideration with respect to fraction of an examined amino acid and the genes have been assigned a rank number. Next, we plotted the dependence of both the gene size on the rank and the dependence of the amino acid fraction in the gene on the rank. The resulting plots in Figs.2 correspond to two evolutionary extreme cases, representing tryptophan and leucine. It is evident that in the case of tryptophan the fraction of that amino acid in genes is anti-correlated with the gene's size (notice, that about 1/3 genes do not posses tryptophan). In the case of leucine there is a crossover and the effect of selection is evident only for the genes which have more than 10% of that amino acid. When the fraction of leucine is less than 10% there is even a reverse effect, i.e., the increasing fraction is correlated with the increasing gene's size. If we look at the evolutionary scissors in Fig.1, we can see that in the case of leucine the survival time, which originates from pure mutational pressure is longer than its selectional counterpart. Recently, there has appeared a paper by Xia and Li [8] discussing which amino acid properties (like polarity, isoelectric point, volume etc.) affect protein evolution. Thus, there is a possibility to relate these properties to our discussion of the evolutionary scissors.

5 Conclusions

With the help of computer simulations of the mutational pressure experienced by genes in the *B.burgdorferi* genome we have shown that the amino acids which experience the highest selectional pressure have the shortest turn over time with respect to mutation pressure. The fraction of these amino acids in genes depends on the gene's size. Much different is the selectional role of the amino acids, like leucine, from the right hand side of the selectional scissors. Although they have long turn over time with respect to the mutational pressure, their fraction cannot be too high. This could be considered as an effect of optimisation of the genetic information on coding with processes of mutagenesis and phenotype selection for protein functions.

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

References

1. Dayhoff, M.O., Schwartz, R.M., and Orcutt, B.C.: A Model of Evolutionary Change in Proteins. In: Atlas of Protein Sequence and Structure , Vol. 5 Suppl. 3 (1978) 345–352
2. Jones, D.T., Taylor, W.R., and Thornton, J.M.: The rapid generation of mutation data matrices from protein sequences. In: CABIOS, Vol. 8 no. 3, (1992) 275–282
3. Kowalczuk, M., Mackiewicz, P., Szczepanik, D., Nowicka, A., Dudkiewicz, M., Dudek, M.R., and Cebrat, S.: Multiple base substitution corrections in DNA sequence evolution. Int. J. Mod. Phys. C **12** (2001) 1043–1053
4. 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
5. 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 Evolutionary Biology **1** (2001) (1):13
6. Cebrat, S., Dudek, M.R., Gierlik, A., Kowalczuk, M., Mackiewicz, P.: Effect of replication on the third base of codons. Physica A **265** (1999) 78–84
7. Kimura, M.: The Neutral Theory of Molecular Evolution, Cambridge University Press, Cambridge (1983)
8. Xia, X., Li, W-H.: What Amino Acid Properties Affect Protein Evolution. J. Mol. Evol. **47** (1998) 557–564