On the Probability of Codon−Codon Mutational Replacements

D. Pumpernik, D. Lukman, and B. Borštnik*

National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia

Received April 14, 1997[⊗]

The content of the data base of blocks of protein sequences (Henikoff, S. ; Henikoff, J. G. *Genomics* **1994**, *19*, 97−107) were used in order to locate the homologous gene sequences. Aligning these sequences the 64 × 64 matrix of codon−codon interchanges was constructed. Matrices were generated for various phylogenetical groups of organisms and for various degrees of relatedness of source sequences. The resulting matrices provide the information about the codon usage: the frequency of silent and the frequency of recognizable mutations. An attempt was made to interpret the codon−codon interchange frequencies in the spirit of the supposition that the evolutionary events are of predominantly stochastic character.

1. INTRODUCTION

The explosive growth of DNA sequence information offers the possibility to analyze the existing data and to produce new information which can provide insight into the molecular structure and processes in the living cell. Such results are especially valuable when the phenomenon of molecular evolution is under investigation. In biological evolution the spontaneous generation of point mutations, insertions, and deletions is the fundamental driving force. In this work we shall treat one−nucleotide interchanges (transitions and transversions) as the basic mutational events. Further, we shall only consider the mutations in coding regions of genes (exons). One can treat the mutational process as a two-step process: the submittal followed by the acceptance. In between there is a two level filtering procedure: less stringent filtering due to the constraints imposed on the DNA and RNA level (control of silent mutations) and more stringent filtering (control of recognizable mutations) due to the constraints imposed on the amino acid sequence level. The most appropriate method to investigate the nature of the above mentioned natural filtering of submitted point mutations is to study quantitatively the mutational interchanges. This can be done by aligning orthologous or paralogous pairs of segments of protein coding genes and counting aligned nonidentical pairs of codons which herald the codon−codon interchanges in the evolutionary history of the particular gene segment. The results of this kind of numerical procedures have already been reported in the early history of biological computing. Dayhoff[1] produced the amino acid mutation probability matrices for amino acids at the time when the predominant type of sequential information was in the form of amino acid sequences. Recently, as the amount of DNA sequential data is sharply increasing, the determination of the mutation probability matrices for amino acids was redone.[2] One can also try to determine the codon−codon transition probabilities. On a limited basis such an attempt was performed by Schoeniger *et al.*[3] The detailed knowledge of codon−codon transition probabilities enables the insight into the evolutionary processes as well as better understanding of molecular mechanisms.

2. NUMERICAL METHODOLOGY

Statistical interpretation of mutational processes can proceed via the master equation formalism,[4] which means that the mutational process can be considered as a rate process. An issue which needs to be addressed first refers to the question of existence or nonexistence of detailed balance. According to the present understanding of evolutionary processes it is not possible to defend any decisive opinion regarding this question. To remain on the safe side it is most opportune, due to nonexistence of molecular fossil data, to suppose that the principle of detailed balance is fulfilled. In this case the mutational process can be described in terms of two quantities: the mutation probability matrix **W** and the corresponding vector of composition **f**. The entities which are the subject of mutational changes can be either nucleotides, codons—these are triplets of nucleotides, or amino acids. We shall deal with codon−codon replacements. The quantity of our interest will be besides **W** and **f** also the matrix denoting the codon−codon replacement frequencies **A**. In the case that the regime with detailed balance is operative, the above mentioned quantities are connected with the expressions[1,5]

$$W_{ij}/W_{ji} = f_i/f_j \tag{1}$$

and

$$A_{ij} \propto W_{ij} f_j \tag{2}$$

The two matrices **A** and **W** depend upon *PAM* (percentage of accepted mutations) parameter which means that the mutation probability matrix in the limit $PAM \rightarrow 0$ is the generic quantity from which higher *PAM* matrices can be generated by rising $\mathbf{W}(PAM = 1)$ to an arbitrary power

$$\mathbf{W}(PAM) = [\mathbf{W}(PAM = 1)]^{PAM} \tag{3}$$

The purpose of this paper is to provide the information about the properties of **A** matrices for codon−codon interchanges by analyzing the sequences available in genetic sequence databanks. The **A** matrix has the dimension 64 × 64 and is symmetric. We shall disregard nonsense mutations (amino acid coding triplets mutating to *stop* codons or vice versa) since there is no information about their frequency. This means that the subject of our treatment will be one half

[⊗] Abstract published in *Advance ACS Abstracts,* November 1, 1997.

S0095-2338(97)00227-8 CCC: $14.00

(lower left triangle) of the **A** matrices with 61 × 61 elements. The matrix elements along the diagonal belong to pairs of identical codon triplets and provide us with the information about the codon usage. Off diagonal matrix elements belong either to silent or to recognizable mutations. Silent mutations are those where one codon is replaced by another synonymous codon. Synonymous codons can differ in one, two, or three nucleotides. Two and three nucleotide silent replacements can take place only in the case of 6-fold degenerate coding (serine, leucine and arginine). Recognizable mutations will be divided into classes according to the number of silent (n) and recognizable (m) nucleotide replacements. The most informative are the counts of the replacements with $(n, m) = (0, 1)$ and $(n, m) = (1, 1)$. The mutations with $m > 1$ were not treated quantitatively because the statistics are too weak, while some $n > 1$ replacement counts will be reported directly. The analysis of relative frequency of occurrence of codon−codon interchanges belonging to the above mentioned classes will help us to elucidate the question regarding the partition of the constraints between DNA/RNA level and amino acid level. Let us suppose that the frequency of codon−codon replacements can be parameterized in the following way

$$A_{ij} \propto \alpha^n \beta^m \tag{4}$$

where α and β are two parameters to be determined. This ansatz is constructed on the basis of the supposition that there is an uniform rate of mutational submissions, while the mutation rejection probability can be expressed in the product form with α and β being the reduction factors due to the constraints imposed by silent and recognizable mutations, respectively.

The **A** matrices can be only generated through a large scale alignment of protein coding regions of genes. In principle it is possible to perform a global alignment within the entire pool of known sequences—this is to align each sequence with each other sequence. If one is interested in protein coding regions one can only process those entries of the genetic sequence data base where the locations and proper reading frame of exons are determined with sufficient certainty. To this point a valuable piece of work was performed by by S. Henikoff and J. G. Henikoff[6,7] who constructed the data base of blocks. From our point of view the practical utility of protein blocks database is due to its richness of homologous motifs of amino acid sequences. Although we need homologous nucleotide sequences the blocks database still possesses its full utility for our purpose. We worked out the numerical methodology which enables us to construct the matrices of codon−codon replacement frequencies. From GenBank nucleotide sequence database the amino acid coding regions were translated into amino acid sequences, and the realizations of homologous blocks sequences were searched in translated amino acid sequences. In the last step the nucleotide sequences coding for individual amino acid blocks sequences were aligned, and the homology for each aligned pair was determined. If the homology was high enough the aligned pairs of nucleotide sequences were used to increment the codon−codon replacement frequency matrix **A**. Such a matrix contains all the details of the mutational pattern, while the matrices corresponding to high *PAM* values contain less information since ultimately, as a consequence of eq 3, in the limit $PAM \rightarrow \infty$ the W matrix attains the form[1] $\mathbf{W} = [\mathbf{f}, \mathbf{f}...\mathbf{f}]$. However, to determine a low *PAM* value **A** matrix one needs high homologies between the pairs of sequences, and in this case it is difficult to get good enough statistics. We decided to accumulate mutational data from the pairs of sequences with the homologies on the nucleotide level equal or better than 90%. This yields the average *PAM* value in the interval $5 < PAM < 8$.

Table 1[a]

	primates	rodents	invertebrates	plants	bacteria
α/β	10	17	34	50	57
$[\alpha\beta]_{pred}/[\alpha\beta]_{obs}$	3.3	2.5	3.5	40	7

[a] Results of analysis of the resulting *A* matrices. Upper line represents the quotient of rates of one-nucleotide silent versus one-nucleotide recognizable mutations for five phylogeny entities. Lower line represents the ratio of predicted and observed $n = 1$, $m = 1$ transitions.

3. RESULTS AND DISCUSSION

By sequence processing we generated five **A** matrices and the corresponding composition vectors for five GenBank (release 96, October 1996)[8] phylogeny entries: primates, rodents, invertebrates, plants, and bacteria. The diagonal elements of **A** matrices provide us with the information about the codon usage and, subsequently, about the amino acid composition of the encoded proteins. As it is known for decades since King and Jukes[9] published the paper which became a cornerstone of the theory of neutral evolution, the amino acid composition of proteins is roughly determined by the number of codons by which a particular amino acid is coded. This rule is blurred to some extent by the interference of codon usage and nonuniformity of dinucleotide frequencies which can be observed also in introns and flanking regions of genes, such as the underrepresentation of *CG* and *AT* dinucleotides in eukaryotes. Four out of six codons of arginine, for instance, contain *CG* dinucleotide and therefore arginine does not give an above average contribution to the protein composition in spite of 6-fold coding. The 61 diagonal elements of **A** matrices should have, if the neutral theory of evolution is supposed to hold, comparable intensity. It turns out that the values can differ by two orders of magnitude. Amino acid composition results as a sum over the codons belonging to a particular amino acid. This summation reduces the scatter so that the composition vector can be roughly approximated by $q_i/61$ where q_i is the degeneracy of the genetic code for ith amino acid. The maximum discrepancy between observed and predicted amino acid composition components is a factor of 2 in either direction.

Off diagonal elements of **A** matrices give the evidence about the mutational events. Let us first comment on the suitability of the hypothesis that two or three nucleotide replacements at adjacent sites can be viewed as independent events of one nucleotide interchanges. This supposition cast in mathematical form is written in eq 4. If this equation is to hold the elements of **A** matrix should be grouped into categories defined by (n, m) pairs of indices defining the order of silent and recognizable mutations. The results of the analysis of **A** matrix elements belonging to (0, 0), (1, 0), and (1, 1) categories are presented in Table 1. The results are reasonable, although not entirely consistent with the stochastic hypothesis. We can see that recognizable point mutations run 10−50 times more slowly than the silent ones.

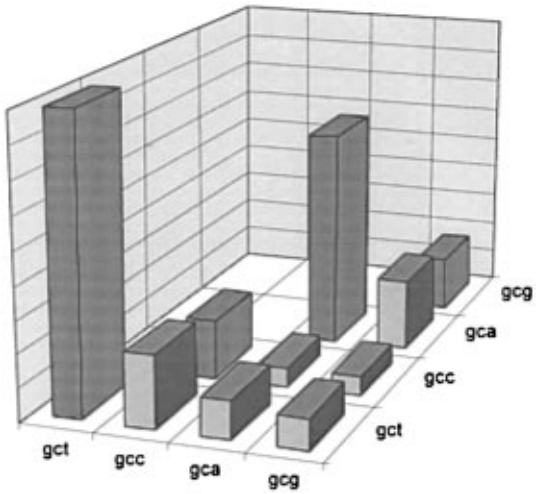

Figure 1. The illustration of the block of **A** matrix corresponding to Ala-Ala transitions in plants. The heights of the columns are proportional to codon usage (four diagonal entries) and one nucleotide silent transitions (off diagonal entries).

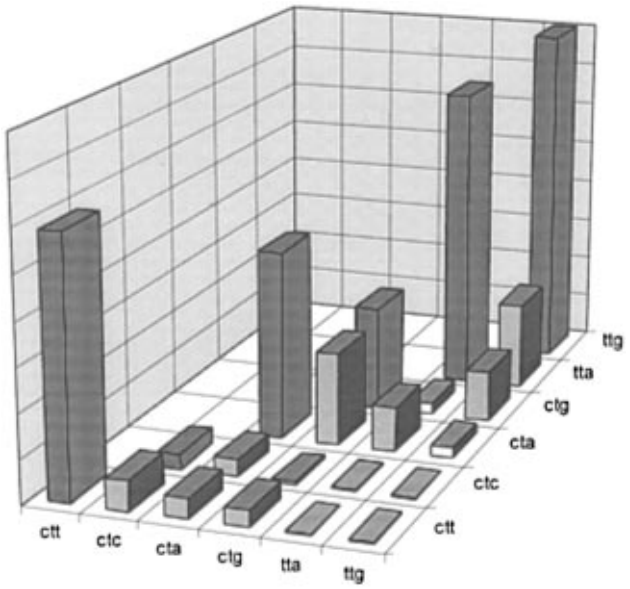

Figure 2. The same as Figure 1 except that for Leu-Leu block. Off diagonal elements belong to three categories of silent mutations with one, two, and three nucleotide interchanges as it can be inferred from the codon assignments along two horizontal directions of the graph. The heights of the columns which represent the A_{ij} values follow approximately the geometric progression according to eq 4.

The result of a cross checking of the stochastic hypothesis is in the lower line of Table 1 where it is shown that the (1, 1) mutations run with substantially lower pace as one would predict on the basis of eq 4. To get a pictorial insight into the nature of mutational events let us examine the pattern of A_{ij} elements belonging to Ala-Ala (Figure 1) and Leu-Leu (Figure 2) transitions in plants. In Ala-Ala case the heights of nondiagonal elements which represent the codon usage vary for a factor of five and also off diagonal elements which represent one nucleotide silent mutations are scattered for nearly the same amount. However, the ratio between the two categories of values is comparable to the corresponding ratio in other amino acid blocks. A more rich variety of categories of transitions with $(n, m) = (0, 0)$, (1, 0) (2, 0) and (3, 0) is encountered in the case of Leu-Leu which was also extracted from **A** matrix belonging to plants. It is depicted in Figure 2. The frequencies of three categories of transitions with $n = 1$, 2, and 3 should follow geometric progression. This is roughly the case, the scatter of the values within the same categories is nearly the same as in Ala-Ala case.

As one would expect, the sites of the **A** matrix which refer to recognizable mutations exhibits the strongest deviation from simplified picture which emerges on the basis of eq 4. The amino acid interchanges are controlled by the changes of protein function, and it is not likely that they would follow the pattern which would be characteristic for neutral mutational events. On the other hand, some codon−codon transitions with two nucleotide interchanges cannot be unambiguously characterized as type (0, 2) or (1, 1). These are the pairs of codons of which one belongs to the quadruplet and the other to the doublet. If the third nucleotide from the doublet mutates from *G* or *A* to *C* or *T*, then this is a recognizable mutation. If later the first nucleotide is mutated one would have another recognizable mutation, which means a (0, 2) case. If the above mentioned two mutations happen in reverse order, or if the first mutation takes place in the quadruplet, one obtains the (1, 1) case. It turns out that the transitions from the above mentioned ambiguous case are quite intense. If one puts them in the (1, 1) class, then the $m = 2$ class of replacements becomes significantly less populated than the classes of $m = 0$ and $m = 1$ replacements. This is to be expected since due to the evolutionary optimization of the genetic code,[5,10] majority of similar amino acids pairs with high similarity are in the $m = 1$ class.

Another criterium which helps us to judge the appropriateness of the stochastic hypothesis refers to the shape of the histogram which represents the distribution of A_{ij} elements with respect to their values. If the transition frequencies would conform to eq 4, then the histogram of A_{ij} values would exhibit a Gaussian shape with the width equal to $\sqrt{N_i}$ where N_i is the number of the elements in the class. Another extreme would be monotonously decreasing peakless distribution in the form of inverse power law.[11] Such a distribution can be expected to occur when the codon−codon interchanges are not obeying any simple algorithmic law such as the one postulated by eq 4 but are the product of very complex processes dictated by the paradigm of Darwinian evolutionary mechanisms. Our results show that $(n, 0)$ classes are more close to the Gaussian-like distribution, while (n, m) interchanges with $m > 0$ are more close to inverse power model. This finding diminishes the applicability of the stochastic model and tells us that it is not wise to try to oversimplify such a complex problem as is the biological evolution.

4. CONCLUSIONS

The major conclusions which can be drawn on the basis of this work refer to the presence of the stochastic component in point mutations, the elementary acts of molecular evolution. We got modest support for the thesis that

−two- and three-nucleotide mutations were accumulated as a series of independently occurring one-nucleotide replacements;

−there are two levels of mutational control: one on the DNA/RNA level and the other on the amino acid sequence

level. Such a structuring of the constraints enables the interpretation of the results in the spirit of eq 4;
−one can make no decisive conclusion regarding the question whether the principle of detailed balance is operative.

ACKNOWLEDGMENT

The work was supported by The Ministry of Science and Technology of Republic of Slovenia.

REFERENCES AND NOTES

(1) Dayhoff, M. O.; Schwartz, R. M.; Orccut, B. C. A Model of Evolutionary Changes in Proteins. In *Atlas of Protein Sequence and Structure;* NBRS, Washington, DC, 1978*;* Vol. 5, Supplement 3, pp 345−351.
(2) Jones, D. T.; Taylor, W. R.; Thornton, J. M. The Rapid Generation of Mutation Data Matrices from Protein Sequences. *Comp. Appl. Biosci.* **1992**, *8*, 275−282.
(3) Schoeniger, M.; Hofacker, G. L.; Borštnik, B. Stochastic traits of molecular evolution-Acceptance of point mutations in native actin genes. *J. Theor. Biol.* **1990**, *143*, 287−306.
(4) Haken, H. *Synergetics;* Springer: Berlin, Heidelberg, 1978; p 88.
(5) Borštnik, B.; Hofacker, G. L. Functional aspects of neutral patterns in protein evolution. In *Structure and Motion, Membranes, Nucleic acids and Proteins;* Clementi, E., Corongiu, G., Sarma, M. H., Sarma, R. H., Eds.; Adenine Press: Guilderland, 1985; pp 277−292.
(6) Henikoff, S.; Henikoff, J. G. Amino Acid Substitution Matrices from Protein Blocks. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 10915−10919.
(7) Henikoff, S.; Henikoff, J. G. Protein Family Classification Based on Searching a Database of Blocks. *Genomics* **1994**, *19*, 97−107.
(8) Benson, D.; Lipman, D. J.; Ostell, J. GenBank. *Nucleic Acid Res.* **1993**, *21*, 2963−2965.
(9) King, J. L.; Jukes, T. H. Non-Darwinian evolution. *Science* **1969**, *164*, 788−798.
(10) Borštnik, B.; Pumpernik, D.; Hofacker, G. L. Point mutations as an optimal search in biological evolution. *J. Theor. Biol.* **1987**, *125*, 249−268.
(11) Borštnik, B.; Pumpernik, D.; Lukman, D. Analysis of apparent l/f spectrum in DNA sequences. *Europhys. Lett.* **1993**, *23*, 389−394.

CI970227M