International Journal of Modern Physics C, Vol. 12, No. 7 (2001) 1043–1053

MULTIPLE BASE SUBSTITUTION CORRECTIONS IN DNA SEQUENCE EVOLUTION

M. KOWALCZUK*, P. MACKIEWICZ*, D. SZCZEPANIK*, A. NOWICKA*, M. DUDKIEWICZ*, M. R. DUDEK†, and S. CEBRAT*

*Institute of Microbiology, University of Wrocław, ul. Przybyszewskiego 63/77, 54-148 Wrocław, Poland

†Institute of Physics, Pedagogical University of Zielona Góra 65-069 Zielona Góra, Poland

Received 18 June 2001
Revised 25 June 2001

We discuss the Jukes and Cantor's one-parameter model and Kimura's two-parameter model unability to describe evolution of asymmetric DNA molecules. The standard distance measure between two DNA sequences, which is the number of substitutions per site, should include the effect of multiple base substitutions separately for each type of the base. Otherwise, the respective tables of substitutions cannot reconstruct the asymmetric DNA molecule with respect to the composition. Basing on Kimura's neutral theory, we have derived a linear law for the correlation of the mean survival time of nucleotides under constant mutation pressure and their fraction in the genome. According to the law, the corrections to Kimura's theory have been discussed to describe evolution of genomes with asymmetric nucleotide composition.

We consider the particular case of the strongly asymmetric *Borrelia burgdorferi* genome and we discuss in detail the corrections, which should be introduced into the distance measure between two DNA sequences to include multiple base substitutions.

Keywords: DNA Evolution; Replication; Tables of Substitutions.

1. Introduction

Measuring the evolutionary distance between two DNA sequences requires the knowledge of the substitution rates of the nucleotides. Each DNA sequence is composed of four different nucleotides, Adenine (A), Guanine (G), Thymine (T) and Cytosine (C). A specific sequence of these nucleotides determines the information, which is transferred by the DNA molecule. In particular, the information, which is translated for proteins is coded by the genetic code, a specific set of triplets of nucleotides (codons) each of which codes for one amino acid.[1] Although there are 64 possible triplets, the number of amino acids is twenty. This means that the genetic code is degenerated because a given amino acid could be coded by more than one codon. In fact, it is the third nucleotide position in the codon, which is the most degenerated. Therefore, two different organisms for the same function can use the

same protein coded by different DNA sequences. Furthermore, it is also possible that substitution of one amino acid by some other amino acid does not change the function of the protein. By comparing two organisms, we can estimate the phylogenetic distance between them simply counting the number of substitutions between the homologous sequences.

Kimura's neutral theory[2] 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. In this case, the mutations follow the Poisson statistics:

$$g(t) = \frac{1}{\tau}\, e^{-t/\tau}\,, \tag{1}$$

where τ is the mean survival time for the nucleotides and $g(t)$ is the decay probability density at time moment $t \geq 0$. Thus, the number of nonmutated nucleotides at time moment t is equal to:

$$N(t) = N(0)e^{-t/\tau}\,. \tag{2}$$

If we knew the mean survival time τ, we could set a proper time scale for genomes. However, we do not know the ancestral DNA sequences and the mean survival time of nucleotides remains unknown. Therefore, in order to examine DNA sequence evolution, we need to consider the accumulated substitutions between two homologous DNA sequences, which have a common ancestor. In case of DNA sequences, there are four possible nucleotides and therefore a model of nucleotide substitutions has to be described with the help of twelve substitutions, where four of them represent transitions A $\longleftrightarrow$ G, C $\longleftrightarrow$ T, and the remaining eight substitutions represent transversions, A $\longleftrightarrow$ C, A $\longleftrightarrow$ T, G $\longleftrightarrow$ C, G $\longleftrightarrow$ T. In order to correct for multiple nucleotide substitutions, which appear in the evolving DNA sequences, one usually considers a statistical model of DNA evolution in terms of the system of linear differential equations[5,6]:

$$\frac{dP_{\alpha\beta}(t)}{dt} = \sum_{\gamma} M_{\alpha\gamma} P_{\gamma\beta}(t)\,, \tag{3}$$

where α, β, γ = A, T, G, C, the symbol $P_{\alpha\beta}(t)$ represents the substitution probability from nucleotide α at time 0 to β after time t, and $M_{\alpha\gamma}$ is a 4×4 substitution rate matrix, which includes the substitution rates from one nucleotide to another. By construction, the sum by row of M is always zero and therefore there are twelve unknown substitution rates to be found.

Jukes and Cantor[3] assumed in their model that all substitutions occur at the same rate, $M_{\alpha\gamma} = u$ (where $\alpha \neq \gamma$), and they concluded that the number of substitutions per site between two homologous DNA sequences originating from the common ancestor is equal to:

$$K = -\frac{3}{4} \ln\left(1 - \frac{4}{3}\, D\right), \tag{4}$$

where D is equal to the proportion of different nucleotides between the DNA sequences under examination. The second model, which is commonly used in constructing phylogenetic trees, is Kimura's two-parameter model,[4,5] which permits different types of substitution rates (transitions and transversions). In Kimura's two-parameter model, the number of substitutions between homologous sequences is given by the following expression[5]:

$$K = -\frac{1}{2}\ln(1 - 2D_1 - D_2) - \frac{1}{4}\ln(1 - 2D_2)\,, \tag{5}$$

where D_1 and D_2 are equal to a fraction of the substitutions recognized as transitions and transversions, respectively, at which the two sequences differ from each other.

The problem of the estimation of divergence K between two species is even more complicated, because some mutations are lethal, they lead to the death of organism in which they have occurred. All these phenomena, which should be taken into consideration when the phylogenetic distances are estimated, make the task very complicated. The main problem with the above simplifications of the description of DNA sequence evolution is that they fail to predict the asymmetry of the DNA molecule. Recent progress of genome sequencing programmes has brought many complete genome sequences, and their analyses have shown that the simplifying assumptions of one-parameter model, two-parameter model and even six-parameter model[5,6] are unable to describe the mutational pressure imposed on these genomes. The nucleotide composition of one DNA strand is different from the nucleotide composition of the complementary strand. One of the explanations of this observation is that there is a different mutational pressure on each DNA strand, which results from different mechanisms involved in replication of the two DNA strands. The models including the mutational pressure have to be considered to clear up its role in the resulting asymmetry observed in many genomes.

2. Tables of Substitution Rates

The topology of the replication fork requires different enzymatic mechanisms for the synthesis of leading and lagging fragments of the DNA molecule with different error rates.[9–11] Usually, the leading fragment of each DNA strand is richer in Guanine than in Cytosine and it is richer in Thymine than in Adenine.[8,12,13,15] Hence, the replication, which is asymmetric, is responsible for introducing strong trends into the nucleotide distribution along the DNA strands.[14] The trends can be observed both in intergenic sequences and in genes (see, e.g., Refs. 8 and 16). The example of such asymmetry in genes is shown in Fig. 1 (curve (a)), where there is plotted a cumulative walk on number of Guanine and Cytosine for a DNA sequence (Watson strand) consisting of spliced nucleotides from the third positions in codons of the genes of the *Borrelia burgdorferi* genome. In this case, the walker follows the DNA sequence and it goes "up" if it meets Guanine and it goes "down" if it meets Cytosine, otherwise, it waits. The ORI and TER in Fig. 1 determine the switching

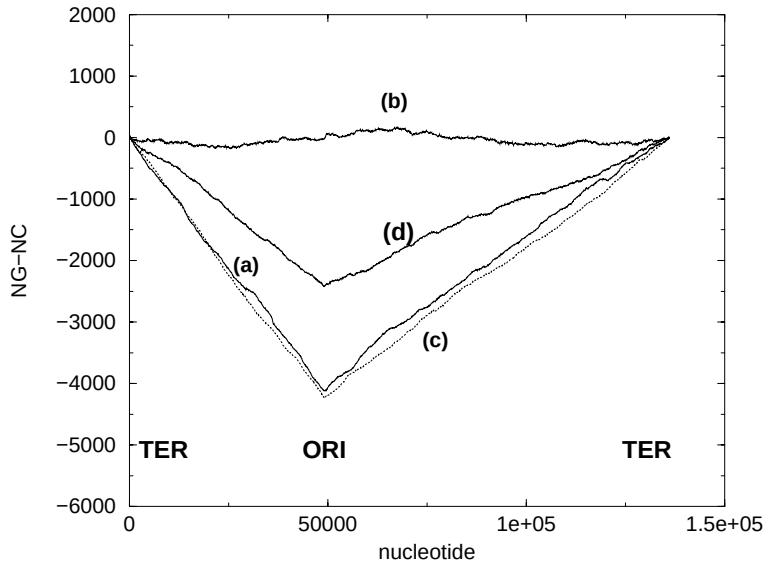

Fig. 1. A cumulative walk G–C at third nucleotide position in codons of the genes of Watson DNA strand in *Borrelia burgdorferi* genome. The curve (a) represents the original DNA sequence, (b) the sequence in (a) mutated according to the one-parameter Jukes–Cantor model of evolution after $t = 5000$ time steps, (c) the sequence in (a) mutated according to empirical table of substitution rates [Eq. (9)] after $t = 5000$ time steps, (d) the sequence in (a) evolved according to corrected one-parameter Jukes–Cantor model after $t = 5000$. The ORI is not located at the middle because of the different number of genes at leading and lagging fragments of DNA strand.

mode of replication between the leading fragment and lagging fragment of the DNA strand (the ORI–TER fragment and TER–ORI fragment, respectively).

It is known that even in intergenic sequences, there is a relatively strong triplet signal in FFT power spectrum as it has been discussed in papers by Voss,[17] Peng *et al.*,[18] Buldyrev *et al.*[19] and Gierlik *et al.*[20] Thus, we assumed that some intergenic sequences have derived from coding sequences and could freely accumulate mutations with frequencies determined by the replication-associated mutational pressure.

By comparing intergenic sequences with homologous sequences of genes, we were able to construct an empirical table of substitution rates describing the mutational pressure for the leading strand of the *Borrelia burgdorferi* genome. We could expect that the properly constructed tables of substitution rates should retain both the DNA sequence composition and the strand asymmetry. Usually, one tries to find the unknown twelve substitution rates $M_{\alpha\beta}\,(\alpha \neq \beta)$ with the help of the four equations describing the steady-state condition:

$$F_\alpha \sum_{\beta\neq\alpha} M_{\alpha\beta} = \sum_{\beta\neq\alpha} F_\beta M_{\beta\alpha}\,, \tag{6}$$

where α, β = A, T, G, C, and F_α represents the fraction of nucleotide α. The problem of the infinite number of solutions is avoided by an additional assumption

that some of the substitution rates are equal. However then, the reconstructed DNA sequence loses the asymmetry of the leading and lagging strands (see curve (b) in Fig. 1). Another way to avoid the problem is using a computer random number generator for generating the substitution rates $M_{\alpha\beta}$.[21] Although these tables satisfy the balance equations (Eq. (6)), they do not have any biological meaning. In our recent paper,[22] an empirical table M of substitution rates has been constructed for *Borrelia burgdorferi* genome (its DNA sequence has very strong leading-lagging asymmetry):

$$M = \begin{pmatrix} a & uW_{\mathrm{AT}} & uW_{\mathrm{AG}} & uW_{\mathrm{AC}} \\ uW_{\mathrm{TA}} & b & uW_{\mathrm{TG}} & uW_{\mathrm{TC}} \\ uW_{\mathrm{GA}} & uW_{\mathrm{GT}} & c & uW_{\mathrm{GC}} \\ uW_{\mathrm{CA}} & uW_{\mathrm{CT}} & uW_{\mathrm{CG}} & d \end{pmatrix} \tag{7}$$

basing on the intergenic sequences, where by definition[5,6] the values of the parameters a, b, c, d are such that the sum by rows in the matrix M is equal to zero, the parameter u represents the mutation rate, $W_{\alpha\beta}$ denotes the weight factor for the substitution from nucleotide α to nucleotide β and the sum of the weights over all different values of α and β is equal to one:

$$\sum_{\alpha\neq\beta} W_{\alpha\beta} = 1\,. \tag{8}$$

In this specific representation, the factors $W_{\alpha\beta}$ are nothing else but the relative substitution rates. In the case of the *Borrelia burgdorferi* genome, the numerical values of $W_{\alpha\beta}$ for the leading DNA strand are the following:

$$\begin{aligned} &W_{\mathrm{GA}} = 0.1637 \quad W_{\mathrm{GT}} = 0.1157 \quad W_{\mathrm{GC}} = 0.0147 \quad W_{\mathrm{AG}} = 0.0667 \\ &W_{\mathrm{AT}} = 0.1027 \quad W_{\mathrm{AC}} = 0.0228 \quad W_{\mathrm{TG}} = 0.0347 \quad W_{\mathrm{TA}} = 0.0655 \\ &W_{\mathrm{TC}} = 0.0350 \quad W_{\mathrm{CG}} = 0.0470 \quad W_{\mathrm{CA}} = 0.0702 \quad W_{\mathrm{CT}} = 0.2613\,. \end{aligned} \tag{9}$$

The numbers are the same for the lagging DNA strand, but the respective symbols of the nucleotides should be substituted for the complementary ones, i.e., A for T, T for A, G for C and C for G. This empirical table of substitution rates retains the asymmetry of the DNA strands. It is evident in Fig. 1 (curve (c)), where the results of computer simulations with the mutation events according to the table, Eq. (9) are presented.

If we introduce the following four relations:

$$W_\alpha = \sum_{\beta\neq\alpha} W_{\alpha\beta}\,, \tag{10}$$

where $\alpha = \mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}$ (note that $W_{\mathrm{A}} + W_{\mathrm{T}} + W_{\mathrm{G}} + W_{\mathrm{C}} = 1$), then the fraction of nonmutated nucleotides α at time moment t can be written as:

$$F_\alpha(t) = F_\alpha(0)(1 - uW_\alpha)^t = F_\alpha(0)e^{t\ln(1-uW_\alpha)}\,. \tag{11}$$

According to Eq. (2), the same can be written as follows:

$$F_\alpha(t) = F_\alpha(0)e^{-t/\tau_\alpha}\,, \tag{12}$$

and hence we can relate the substitution rate W_α to the mean survival time, τ_α, of the nucleotide α:

$$\tau_\alpha = -\frac{1}{\ln(1-uW_\alpha)} \approx \frac{1}{uW_\alpha}\,, \tag{13}$$

where the last equation is true for small value of the substitution rate u.

We can relate the probability $p(t)$ of a mutation event after time t to the probabilities $P_{\mathrm{A}}(t)$, $P_{\mathrm{T}}(t)$, $P_{\mathrm{G}}(t)$ and $P_{\mathrm{C}}(t)$ that, respectively, nucleotide A, nucleotide T, nucleotide G or nucleotide C has been substituted for another one, i.e.,

$$p(t) = F_{\mathrm{A}}(0)P_{\mathrm{A}}(t) + F_{\mathrm{T}}(0)P_{\mathrm{T}}(t) + F_{\mathrm{G}}(0)P_{\mathrm{G}}(t) + F_{\mathrm{C}}(0)P_{\mathrm{C}}(t)\,, \tag{14}$$

where

$$P_\alpha(t) = \int_0^t g_\alpha(t')\,dt' = 1 - e^{-t/\tau_\alpha}\,, \tag{15}$$

and $g_\alpha(t')$ is the decay probability density of nucleotide α at time moment $t \geq 0$, defined in Eq. (1). Equation (14) reduces to the following relationship for the mean survival time of the nucleotides:

$$e^{-t/\tau} = F_{\mathrm{A}}(0)e^{-t/\tau_{\mathrm{A}}} + F_{\mathrm{T}}(0)e^{-t/\tau_{\mathrm{T}}} + F_{\mathrm{G}}(0)e^{-t/\tau_{\mathrm{G}}} + F_{\mathrm{C}}(0)e^{-t/\tau_{\mathrm{C}}}\,, \tag{16}$$

which becomes very simple for the vanishingly small value of t:

$$\frac{1}{\tau} \approx F_{\mathrm{A}}(0)\frac{1}{\tau_{\mathrm{A}}} + F_{\mathrm{T}}(0)\frac{1}{\tau_{\mathrm{T}}} + F_{\mathrm{G}}(0)\frac{1}{\tau_{\mathrm{G}}} + F_{\mathrm{C}}(0)\frac{1}{\tau_{\mathrm{C}}}\,. \tag{17}$$

We have examined τ_α ($\alpha = \mathrm{A}, \mathrm{T}, \mathrm{G}, \mathrm{C}$) with the help of substitution tables, which have been published for various genomes (e.g., Refs. 23 and 24) both analytically and in computer simulations, and we have found that all of them share a unique feature (see also Ref. 22): if the genome under consideration is in equilibrium with respect to the mutation events, the fraction F_α of each type of nucleotide $\alpha =$ A, T, G, C is linearly related to the respective mean survival time τ_α:

$$F_\alpha = m_0\tau_\alpha + c_0\,. \tag{18}$$

The correlation coefficient is as high as 0.999. This property of DNA composition does not held for tables of substitution rates, M, constructed for DNA sequence under selection pressure. We observed that even computer generated tables, M, which satisfy the steady-state condition (Eq. (6)), typically will not satisfy the property in Eq. (18). It is shown in Fig. 2, where the relation between the mean survival time and F_α is plotted. In the case of the computer-generated table, the following substitution rates have been used:

$$\begin{aligned}
&W_{\mathrm{GA}} = 0.0219 \quad W_{\mathrm{GT}} = 0.0254 \quad W_{\mathrm{GC}} = 0.0426 \quad W_{\mathrm{AG}} = 0.0211\\
&W_{\mathrm{AT}} = 0.0621 \quad W_{\mathrm{AC}} = 0.0403 \quad W_{\mathrm{TG}} = 0.0071 \quad W_{\mathrm{TA}} = 0.0458\\
&W_{\mathrm{TC}} = 0.0564 \quad W_{\mathrm{CG}} = 0.0346 \quad W_{\mathrm{CA}} = 0.1825 \quad W_{\mathrm{CT}} = 0.4604
\end{aligned} \tag{19}$$

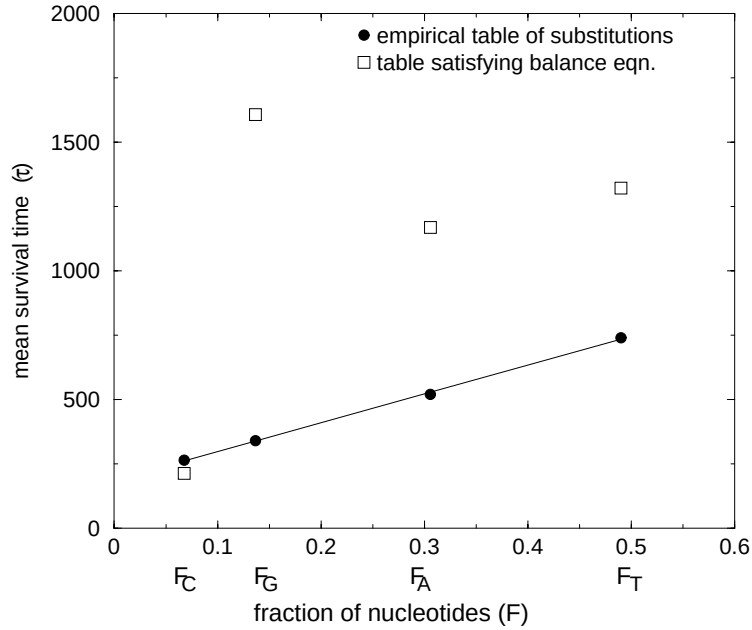

Fig. 2. Mean survial time of nucleotide α = A, T, G, C versus occurrence of them, F_α, in third positions in codons of leading strand ORFs of *Borrelia burgdorferi* genome. Two cases are plotted: the one corresponding to the empirical table [Eq. (9)] and another one corresponding to the computer generated table of substitution rates [Eq. (19)].

and the same nucleotide fractions as in the case of the empirical tables, i.e., $F_{\rm A} = 0.306$, $F_{\rm T} = 0.4901$, $F_{\rm G} = 0.1365$, $F_{\rm C} = 0.0675$.

What are the consequences of the linear evolution law for genomes, which have an equilibrium composition? The steady state condition (Eq. (6)) ensures that the mean value of the respective nucleotide fraction $F_\alpha(t)$ should be the same at each time moment t. In such a genome, the mean survival time τ of a nucleotide should be a conserved quantity and it is clear that it tends to satisfying an equation analogous to Eq. (17) (which is always true for the sequences being nondistant from their predecessor) at any time moment t, i.e.,

$$\frac{1}{\tau} = F_{\rm A}(t)\frac{1}{\tau_{\rm A}} + F_{\rm T}(t)\frac{1}{\tau_{\rm T}} + F_{\rm G}(t)\frac{1}{\tau_{\rm G}} + F_{\rm C}(t)\frac{1}{\tau_{\rm C}}\,. \tag{20}$$

Once the left-hand side of the equation is a constant, the mean survival time of each nucleotide α = A, T, G, C should be proportional to the fraction F_α it belongs to in the genome. Moreover, in a natural genome, these dependences are correlated by the same linear rule, defined in Eq. (18).

If we apply the linear rule to the one-parameter Jukes–Cantor model, every element $M_{\alpha\beta}$ of the respective table of substitution rates will be expressed in terms of two parameters, m_0 and c_0, from Eq. (18), instead of only one value u — the

mutation rate, i.e.,

$$M = \begin{pmatrix} a & 1/(m_0F_{\rm A}+c_0) & 1/(m_0F_{\rm A}+c_0) & 1/(m_0F_{\rm A}+c_0) \\ 1/(m_0F_{\rm T}+c_0) & b & 1/(m_0F_{\rm T}+c_0) & 1/(m_0T_{\rm A}+c_0) \\ 1/(m_0F_{\rm G}+c_0) & 1/(m_0F_{\rm G}+c_0) & c & 1/(m_0F_{\rm G}+c_0) \\ 1/(m_0F_{\rm C}+c_0) & 1/(m_0F_{\rm C}+c_0) & 1/(m_0F_{\rm C}+c_0) & d \end{pmatrix}, \tag{21}$$

where we have used the relation from Eq. (13):

$$uW_\alpha = \frac{1}{m_0F_\alpha + c_0}\,, \tag{22}$$

and the meaning of the parameters a, b, c, d is the same as previously. In an analogous way, the extention of the two-parameter Kimura's model to the one including the linear law corrections is straightforward.

3. Discussion of the Corrections to One-Parameter Model

According to the one-parameter model, the substitution probability $P_{\alpha\beta}(t)$ of having nucleotide β after time t instead of nucleotide α at time $t = 0$ is expressed with the help of the formula:

$$P_{\alpha\beta}(t) = \frac{1}{4} - \frac{1}{4}\,e^{-4ut}\,, \tag{23}$$

if $\alpha \neq \beta$ (see details in Ref. 5). Hence, the decay probability of a nucleotide α is equal to $P_\alpha(t) = \sum_{\beta\neq\alpha} P_{\alpha\beta} = (3/4)(1 - e^{-4ut})$. The geneticists consider this decay probability as the probability of accumulation of mutations after time t. The last formula could be also concluded from Eq. (15), where the same $P_\alpha(t)$ is expressed in terms of the mean survival time τ_α. In Fig. 3, one can compare the time dependence of this decay probability (represented by dot-dashed curve) with the analogous result of computer simulations for which the empirical table Eq. (9) of substitution rates has been used. The divergence of both results is evident. Thus, Kimura's model predictions, e.g., concerning genetic distances between two species can be very far from the expected ones, especially in the case of asymmetric genomes. We could experience this feature of the one-parameter model already from the analysis of the G–C DNA walk (curve (b)) in Fig. 1. The linear evolution law we have discovered, Eq. (18), for mean survival times of nucleotides introduces an additional requirement for the evolution rate — the substitution rates and the fractions of nucleotides in the genome become correlated. Another consequence of the linear evolution law is that of the numbers $N_\alpha^{\rm ret}(t)$ of returns after time t to the same nucleotide type α = A, T, G, C at a site as the original sequence are correlated. The return substitutions are not simply the back mutations (reversions) but they may represent a long sequence of various substitution events (at most t after time t) ending with the substitution leading to the same type of the nucleotide at the

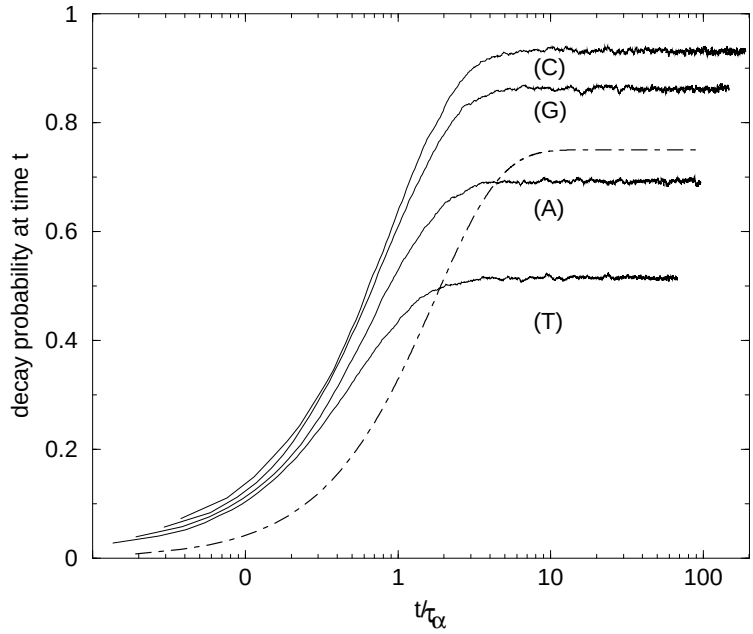

Fig. 3. Time dependence of the mulation probability for the nucleotides A, T, G, C. The result of the one-parameter Jukes–Cantor model has been represented with the help of dot-dashed curve. In the latter case, the mean survival time for a nucleotide has been used. The other curves represent the result of computer simulations, in which we used the empirical table of substitution rates [Eq. (9)].

site as in the original DNA sequence. In Fig. 4, we present the time dependence of the ratio:

$$R_\alpha(t) = \frac{N_\alpha^{\mathrm{ret}}(t)}{N_\alpha(0)}\,, \tag{24}$$

where $N_\alpha(0)$ denotes the number of nucleotides α at the original DNA sequence. This ratio defines the site return ability of each type α of nucleotide in the genome. We can observe that the lower the fraction of some kind α of nucleotide, the lower the respective return ability, R_α. It is never the case in Kimura's one-parameter model for which there is the same return ability for each nucleotide after time $t \gg \tau$.

If we correct the values of Kimura's model substitution rates, according to Eq. (21) so that they all satisfy the linear evolution law Eq. (18), the DNA sequence evolving under mutation pressure will keep the asymmetry of its composition. It can be observed in Fig. 1 in the case of curve (d). This result is very optimistic because this means that even the simple models of evolution, like one-parameter and two-parameter Kimura's models can still be used to estimate distances between species if the respective corrections to substitution rates are introduced. However, we should remember that the biological meaning of the two unknown parameters, m_0 and c_0, in the linear law introduced in Eq. (18) should be cleared up.

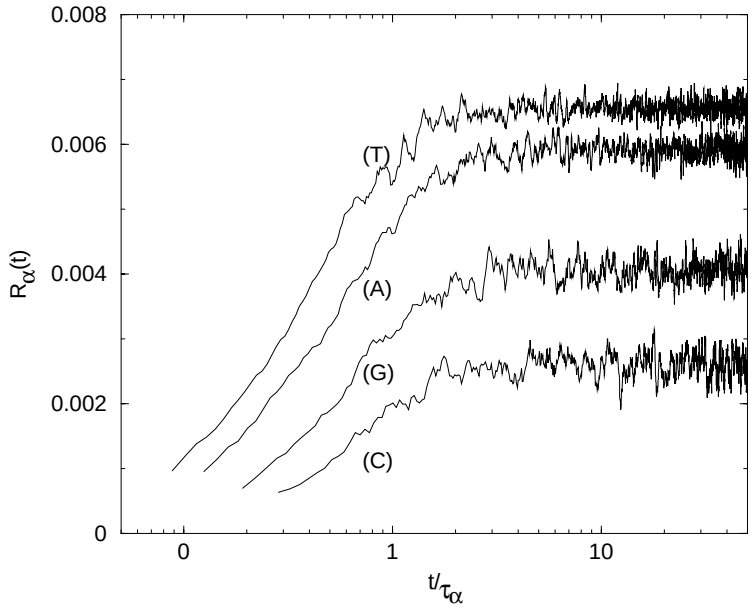

Fig. 4. Dependence of the ratio, $N_\alpha^{\rm ret}(t)/N_\alpha(t=0)$, on time for the number of returns after time t to the same nucleotide type α at a site as in original sequence and the number of this type of nucleotides at original sequence.

If the genome under consideration is in equilibrium with respect to the mutation events, it experiences a constant mutational pressure and the mutation pressure is the same for every nucleotide. The latter is evident if we substitute, in Eq. (20), the nucleotide fractions F_α ($\alpha = {\rm A, T, G, C}$) for the respective mean survival times of the nucleotides according to Eq. (18). Then, we obtain the following simple relation between the mean survival time τ of the nucleotides and the mutation pressure u:

$$\frac{1}{\tau} = 4m_0 + c_0\left(\frac{1}{\tau_{\rm A}} + \frac{1}{\tau_{\rm T}} + \frac{1}{\tau_{\rm G}} + \frac{1}{\tau_{\rm C}}\right) \approx 4m_0 + c_0 u\,, \tag{25}$$

where the right-hand side we have used Eq. (13). This is important result since it means that even in asymmetric genomes, there is the same mutation pressure on each DNA strand (leading and lagging).

4. Conclusions

We discussed Kimura's models of evolution of DNA sequences and we showed the way to correct the substitution rates in the models to describe properly the evolution of asymmetric genomes. The linear evolution law we discovered for the mean survival times of nucleotides and their fractions in the genome, introduces a specific correlation between the frequencies of mutations at different sites of the evolving DNA sequence. In particular, we observed that the number of reversions in each site depends on the occurrence of the respective type of nucleotide in the

original DNA sequence. The method presented by us allows description of the asymmetric genomes, with respect to composition of the leading and lagging parts of DNA sequences.

Acknowledgments

The work was supported by KBN grants No. 6 P04A 025-18 and 6 P04A 016 20.

References

1. F. H. C. Crick, F. R. S. L. Barnett, S. Brenner, and R. J. Watts-Tobin, *Nature* **192**, 1227 (1961).
2. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
3. T. H. Jukes and C. R. Cantor, in *Mammalian Protein Metabolism*, ed. H. N. Munro (Academic Press, New York, 1969), p. 21.
4. M. Kimura, *J. Mol. Evol.* **16**, 111 (1980).
5. W.-H. Li and X. Gu, *Physica A* **221**, 159 (1995).
6. J. R. Lobry, *J. Mol. Evol.* **40**, 326 (1995).
7. S. Cebrat, M. R. Dudek, A. Gierlik, M. Kowalczuk, and P. Mackiewicz, *Physica A* **265**, 78 (1999).
8. P. Mackiewicz, A. Gierlik, M. Kowalczuk, M. R. Dudek, and S. Cebrat, *Genome Res.* **9**, 409 (1999).
9. A. Kornberg and T. A. Baker, *DNA Replication*, ed. Freeman (New York, 1992).
10. T. A. Kunkel, *BioEssays* **14**, 303 (1992).
11. S. Waga and B. Stillman, *Nature* **369**, 207 (1994).
12. J. Mrazek and S. Karlin, *Proc. Natl. Acad. Sci. USA* **95**, 3720 (1998).
13. A. Grigoriev, *Nucleic Acid Res.* **26**(10), 2286 (1998).
14. M. J. McLean, K. H. Wolfe, and K. M. Devine, *J. Mol. Evol.* **47**, 691 (1998).
15. J. R. Lobry, *Mol. Biol. Evol.* **13**(5), 660 (1996).
16. S. Cebrat and M. R. Dudek, *Eur. Phys. J. B* **3**, 271 (1998).
17. R. Voss, *Phys. Rev. Lett.* **68**, 3805 (1992).
18. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, *Nature* **356**, 168 (1992).
19. S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C.-K. Peng, M. Simons, and H. E. Stanley, *Phys. Rev. E* **51**, 5084 (1995).
20. A. Gierlik, P. Mackiewicz, M. Kowalczuk, S. Cebrat, and M. R. Dudek, *J. Mod. Phys. C* **10**, 1 (1999).
21. M. Kowalczuk, A. Gierlik, P. Mackiewicz, S. Cebrat, and M. R. Dudek, *Physica A* **273**, 116 (1999).
22. M. Kowalczuk, P. Mackiewicz, D. Szczepanik, A. Nowicka, M. Dudkiewicz, M. R. Dudek, and S. Cebrat, submitted to *Mol. Biol. Evol.*
23. K. Tamura, *Mol. Biol. Evol.* **9**, 814 (1992).
24. J. Adachi and M. Hasegawa, *Mol. Biol. Evol.* **13**, 200 (1996).