ELSEVIER

Physica A 273 (1999) 116–131

PHYSICA A

www.elsevier.com/locate/physa

# Optimization of gene sequences under constant mutational pressure and selection

M. Kowalczuk[a], A. Gierlik[a], P. Mackiewicz[a], S. Cebrat[a], M.R. Dudek[b,*]

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

[b] *Institute of Theoretical Physics, University of Wrocław, pl. Maxa Borna 9, 50-204 Wrocław, Poland*

Received 25 July 1999

**Abstract**

We have analyzed the influence of constant mutational pressure and selection on the nucleotide composition of DNA sequences of various size, which were represented by the genes of the *Borrelia burgdorferi* genome. With the help of MC simulations we have found that longer DNA sequences accumulate much less base substitutions per sequence length than short sequences. This leads us to the conclusion that the accuracy of replication may determine the size of genome.


*PACS:* 87.14.G; 87.23.Kg

*Keywords:* DNA; Mutation; Selection; Molecular clock hypothesis

## 1. Introduction

There are a few statistical models discussing evolutionary relationship between species, like Markov stochastic models (e.g., [1,2]), models based on multifractal dimension of DNA walks [3] or the observation of long-range correlation in DNA sequences [4,5], and models exploiting the idea of self-organization (e.g., [6,7]). In some of the models, the phylogenetic trees are reconstructed with the help of different sorts of distance measures between DNA sequences regardless of their length. The widely accepted molecular clock hypothesis [8] also does not include directly the genome length. Therefore, if short and long DNA sequences would behave in a different way

* Corresponding author. Fax: +48-71-214-454.
*E-mail address:* mdudek@mirek.ift.uni.wroc.pl (M.R. Dudek)

0378-4371/99/$ - see front matter 
PII: S0378-4371(99)00346-5

during evolution, one should revise some results concerning distances between species. In the following, we considered DNA sequences of various sizes and we showed with the help of MC simulations that they behave differently during constant mutational pressure and selection. Shorter sequences accumulate more mutations per sequence length than longer ones. This means that if the fidelity of replication is better, the mutational pressure on long DNA sequences is smaller and therefore the genome can be longer. Perhaps this is why it can possess longer genes. In bacterial genomes the accuracy of replication is lower and we do not observe long genes. If dependence of mortality on the length of DNA sequences found by us with the help of computer simulations was confirmed experimentally this would have serious consequences also to the molecular clock hypothesis. The suggestion of strong relation between genome size and mutation rate can be found in the paper by Azbel [9] contained in this issue.

## 2. Model

We constructed a simplified model of gene evolution in which a population of genes of various sizes is subject to constant mutational pressure and selection. In the model, selection is applied independently to each gene and therefore it is more proper to think of them as separate genomes. The population of genes would cease to exist after some time if the point mutations and selection were the only mechanisms of evolution. Therefore, we introduced two populations of genes $A$ and $B$, which initially are identical. During evolution run the populations $A$ and $B$ experience mutational pressure and selection in alternate time steps. In every time step, if a gene from one population is killed by a lethal mutation then a new gene – its allel from the other population – substitutes it. This ensures that the accumulated mutations, if not lethal, are inherited and the distribution of gene length in every population is time invariant. In our case, populations $A$ and $B$ consist of non-overlapping *open reading frames* (ORFs), which are at least 100 codons long, originating from leading fragments of Watson (W) and Crick (C) strand of the *Borrelia burgdorferi* genome [10], downloaded by us from *www.ncbi.nlm.nih.gov*.

We decided to choose the B. burgdorferi genome because it has very strong asymmetry between leading and lagging DNA strands. The asymmetry has been discussed by us in our previous paper [11]. In this case genes lying on the leading and lagging DNA strands belong to two separate groups with respect to their composition. Therefore, one can think of the *B. burgdorferi* genome as being at steady state.

### 2.1. Table of substitution rates

We assume that the frequencies of nucleotides $P_A, P_T, P_G, P_C$ (A – adenine, T – thymine, G – guanine, C – cytosine) in the intergenic sequences of the *B. burgdorferi* genome are the equilibrium ones. We use them to find the substitution rates from one base to another one. By Markovian theory, the base frequencies, $P_A(t), P_T(t), P_G(t)$,

$P_C(t)$, at any discrete time moment $t$ should satisfy the following equations:

$$P_\alpha(t+1) = P_\alpha(t)\left(1 - \sum_{\beta\neq\alpha} W_{\alpha\beta}\right) + \sum_{\beta\neq\alpha} W_{\alpha\beta}P_\beta(t)\,, \tag{1}$$

where $\alpha,\beta = A,T,G,C$ and symbols $W_{\alpha\beta}$ represent elements of the substitution matrix – the substitution rates. Then, from the steady-state condition we obtain the set of four equations

$$P_\alpha \sum_{\beta\neq\alpha} W_{\alpha\beta} = \sum_{\beta\neq\alpha} W_{\alpha\beta}P_\beta(t) \tag{2}$$

with 12 unknown substitution rates $W_{\alpha\beta}$. Thus, we have infinite number of solutions with the same frequences $P_A(t), P_T(t), P_G(t), P_C(t)$. In particular, the solution of (2) with respect to $W_{AG}, W_{CT}, W_{AT}, W_{CG}$ leads to the following expressions:

$$W_{AG} = (-const * P_C + P_G * W_{GA} + P_G * W_{GC})/P_A\,, \tag{3}$$

$$W_{CT} = (-const * P_C + P_G * W_{GC} + P_T * W_{TC})/P_C\,, \tag{4}$$

$$W_{AT} = (const * P_C - P_G * W_{GC} + P_T * W_{TA})/P_A\,, \tag{5}$$

$$W_{CG} = const\,, \tag{6}$$

where *const* can take an arbitrary value. Notice, that four substitution rates $W_{CA}, W_{AC}$, $W_{TG}$, and $W_{GT}$ are not included in the above expressions. Only some of the solutions can have biological meaning. In particular, the rate of transition substitutions ($A \leftrightarrow G$, $C \leftrightarrow T$) should dominate over the rate of transversion substitutions ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$, $G \leftrightarrow T$). The particular case of unequal transitions and transversions has been used in Kimura two-parameter model [2] where only two unknown variables representing the transition and transversion rates were used instead of 12 unknowns. For the purpose of computer simulations we took at random one realization of (3)–(6) which satisfied the above biological constrains and in addition we assumed that the rate of substitution of cytosine by thymine should dominate over all other substitution rates. This random choice we realized with the help of computer random number generator by defining: $const = rnd_1$, $W_{GC} = rnd_2$, $W_{GA} = rnd_3$, $W_{TC} = rnd_4$, $W_{TA} = rnd_5$. The remaining four substitution rates $W_{CA}, W_{AC}, W_{TG}, W_{GT}$ which do not appear in the above expressions we reduced to two by the assumption of the detailed balance principle

$$W_{CA} = P_A W_{AC}/P_C\,, \tag{7}$$

$$W_{TG} = P_G W_{GT}/P_T\,, \tag{8}$$

where $W_{AC} = rnd_6$, $W_{GT} = rnd_7$. Thus, we have a 12-parameter model of substitution rates but in fact only seven parameters are independent and generated by computer random number generator.

Our computer simulations are based on one trial substitution matrix:

| | $A$ | $T$ | $G$ | $C$ |
|---|---|---|---|---|
| $A$ | $W_{AA}$ | 0.041966 | 0.0774995 | 0.0187382 |
| $T$ | 0.0371477 | $W_{TT}$ | 0.000964454 | 0.13171 |
| $G$ | 0.134202 | 0.00287175 | $W_{GG}$ | 0.0252094 |
| $C$ | 0.0446251 | 0.484372 | 0.000693648 | $W_{CC}$ |

, (9)

where the diagonal elements are defined in (1). One may find a better substitution matrix by making its optimization with respect to number of lethal mutations per one generation. The question arises: is there a unique optimum substitution matrix $W$ for which the genome mortality is the smallest or are there many of them? However, we do not consider the optimization problem in the paper.

It is important to notice the strong asymmetry of the substitution rates. Thus, the unequal substitution rates determine the specific biased mutational pressure. In case when all mutation rates are equal there is no DNA asymmetry between leading and lagging DNA strands.

### 2.2. Selection tables

If we partitioned genes in the genome into small fragments of the same length, say equal to $d$ nucleotides, then some of the fragments would be over-represented and some would be under-represented. Some sequences of length $d$ might never appear in genome. The usage of these fragments in the genome becomes important when one discusses genome properties. The small values of $d$ correspond to very coarse-grained information about the genome but larger values of $d$ become gene specific and therefore can be used to model selection rules. We constructed tables of the fragment usage for each of the three base positions in codons separately. This is justified by observation of strong asymmetry of the base positions in codons (e.g. [12]). The asymmetry makes possible to construct methods for ORF discrimination with respect to coding properties [13,14]. The three tables of gene fragment usage were prepared as follows. All non-overlapping ORFs (at least 300 bases long) of the *B. burgdorferi* genome originating from the leading DNA strand were partitioned into segments with the help of sliding window which had width equal to $d = 3k$ bases and sliding step equal to 3 bases in order to preserve codon structure. The $d$-segments containing codon STOP ($TAA, TAG, TGA$) were left out. Next, we splitted every $d$-fragment of ORF into three subsegments of $k$ bases long separately for base positions (1), (2), and (3) in codons. For each base position in codons the $k$-sequences were stored in a separate table consisting of $4^k$ elements representing all possible sequences $k$ symbols long built with the help of four-component alphabet $A, T, G, C$. Hence, we got three banks of gene fragments which we used in computer simulations. Each table of $k$-sequences was constructed as in Fig. 1 where $k$ iteration steps are necessary to represent a sequence consisting of $k$ symbols. If $k = 1$ we have only four different 1-sequences, $A, T, G, C$. If

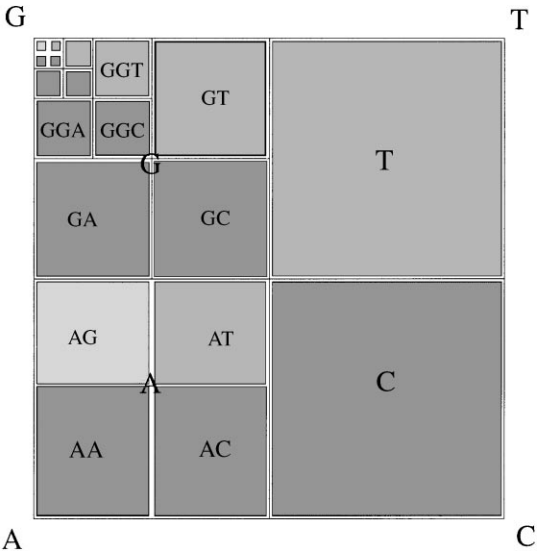


Fig. 1. Construction of table of $4^k$ symbol sequences using alphabet consisting of $A, T, G, C$.

$k=2$ there are $4^2=16$ possible 2-sequences, $GG, GT, \ldots, CC$, arranged in order set by the previous iteration step as in Fig. 1. It is similar for the higher values of $k$. Every possible $k$-sequence is represented by a point in the table of symbols.

In Figs. 2, 4 and 5, we presented the usage of 6-sequences originating from ORFs of the *B. burgdorferi* genome separately for the first, second and third base position in codons. Fig. 2 can be compared with Fig. 3 for the *Escherichia coli* genome and one can notice strong occupation of the vertical lines in the symbolic tables. The similar results were obtained for other genomes. The patterns much resemble the results of Jeffrey's analysis of DNA sequences in terms of Chaos Game representation [15]. The different patterns in Figs. 2–4 corresponding to three nucleotide positions in codons mean that the role of the positions in coding is also different. We made the same analysis of symbol sequences up to $k=9$ and we got the same structure of the tables of $k$-sequences.

### 2.3. Simulation algorithm

If a mutation happens at some position in a gene, it is clear that it may change the logical meaning of its neighborhood (not only the sense of the mutated codon). We decided to extend the neighborhood to $k$ nucleotides. Thus a single mutation can be associated with at most $k$ sequences of the length of $k$ bases and containing the mutation. We analyze all of them and check if they are present in the proper bank of gene fragments. We choose this bank of gene fragments which corresponds to the same base position in codons as the position of the mutation. If at least one of the affected $k$-sequences is not present in the bank, the gene is killed. Otherwise, the decision whether the affected gene should be killed or not is made with the help

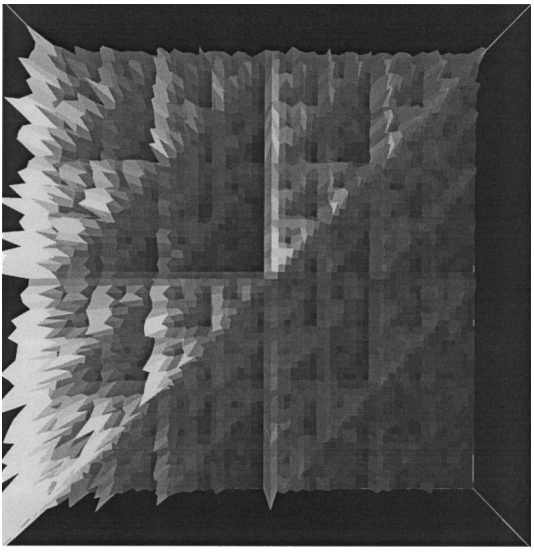

Fig. 2. Perspective view of the usage of 6-sequences in the first base position of codons originating from the non-overlapping ORFs (at least 100 codons long) of the *B. burgdorferi* genome. The right diagonal consists of 6-sequences containing symbols *A* and *T* only, whereas the left diagonal consists of symbols *C* and *G* only.

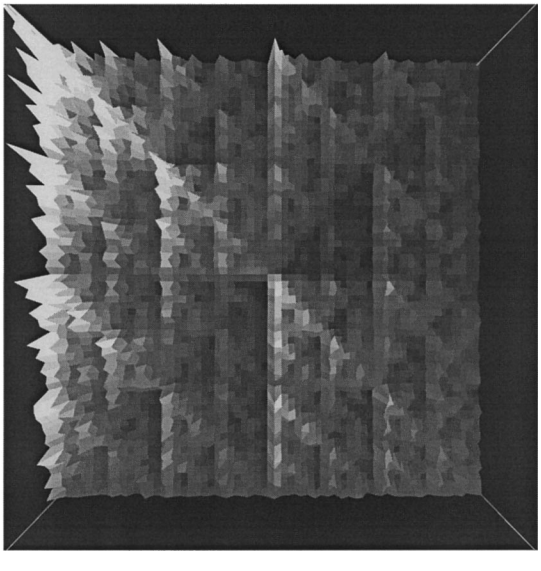

Fig. 3. Perspective view of the usage of 6-sequences in the first base position of codons originating from the non-overlapping ORFs (at least 100 codons long) of the *Escherichia coli* genome. The dominant usage along the vertical lines is common for other genomes, also for the eukaryotes.

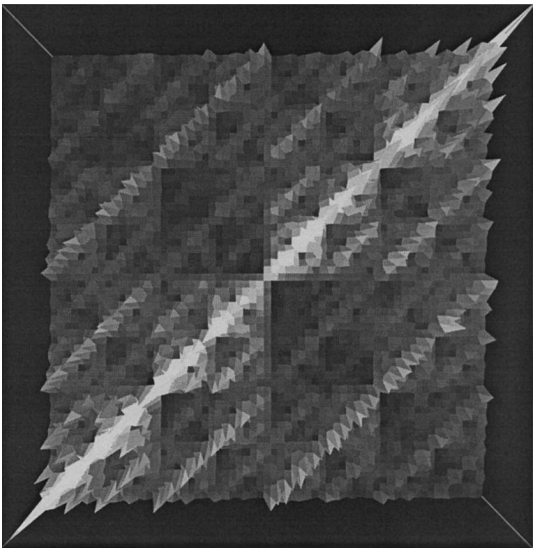

Fig. 4. The same as in Fig. 2 but for the second base position in codons.

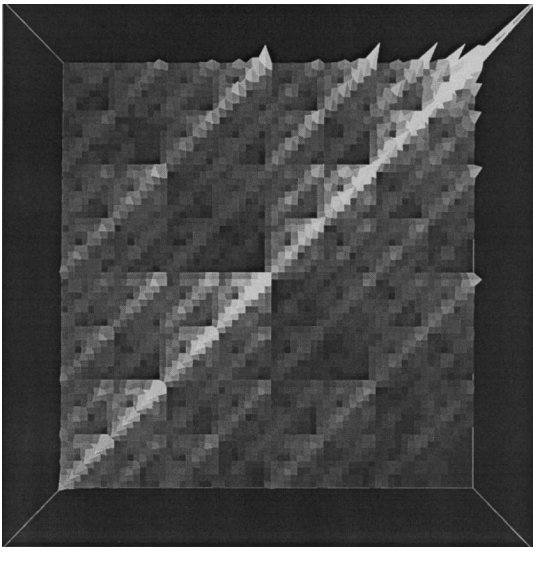

Fig. 5. The same as in Fig. 2 but for the third base position in codons.

of the Metropolis rule [16]. We defined the conditional probability $p_{ij}$ that the new $k$-sequence $(j)$, the one affected by mutation, will be substituted for the old one $(i)$ as follows:

$$p_{ij} = \delta(1, \mathrm{e}^{-G(N_i - N_j)/N_{max}}) , \tag{10}$$

where $\delta$ is Kronecker delta, $N_i$ and $N_j$ are the multiplicities of $k$-sequences of the type $i$ and $j$ in the bank of gene fragments, $N_{max}$ is the maximum multiplicity in the bank of gene fragments, $G = \mu/k_B T$ is a parameter controlling the noise level in accepting mutations, and $\mu$ plays the role of a chemical potential for the appearance of $k$-sequence, $T$ is temperature, $k_B$ is Boltzmann constant. Notice, that according to (10) high temperature (small value of $G$) corresponds to the situation where almost every substitution is accepted. On the other hand low temperature (large value of $G$) corresponds to the situation where the mutated $k$-sequence is almost always rejected if its multiplicity is smaller than the multiplicity of the $k$-sequence before the substitution. In the simulations we apply the selection rules (Metropolis rules) only to the first and the second base position in codons. The third base position in codons is left without selection to control the accuracy of the simulations – it simply accumulates substitutions. We also decided to have the same value of $G$ for the selection rules in every base position in codons.

The simulation algorithm is as follows. At time $t = 0$ there are two identical populations of genes, $A$ and $B$.

(i) apply a constant mutational pressure to population $A$ (with the probability $p_{mut}$ we apply substitution matrix rules (Eq. (9)) to every base of genes in population $A$).

(ii) if a mutation happens, i.e., one base is substituted for another one, and all affected $k$-neighborhoods of it are represented in the bank of gene fragments then the selection rules (Eq. (10)) are applied to accept the mutation or reject it. All $k$-sequences containing the mutation are subject to the selection rules. If at least one of the $k$-sequences fails the selection rules, the gene which contains the mutation is replaced by its counterpart (allel) from the other population. The same is if at least one of the affected $k$-sequences is not represented in the respective bank of gene fragments – the one corresponding to base position of the mutation in codon. Otherwise the mutation is accepted.

(iii) $t = t + 1$. Go to step (i) of the algorithm and replace the population $A$ by $B$ and $B$ by $A$. Thus, populations $A$ and $B$ experience mutational pressure and selection in alternate time steps.

## 3. Discussion of results

In order to discuss the results of computer simulations we introduce DNA walks in two-dimensional space ($[A-T][G-C]$) separately for each base position in codons [12]. We have spliced together all ORFs in population $A(B)$ and the DNA walks represent the entire population and not a single gene only. The populations $A$ and $B$ at $t = 0$ are identical and in Fig. 6 they are represented by three DNA walks labeled (1),(2),(3) for the respective base positions in codon. The other DNA walks (1′),(2′),(3′) on the same figure represent population $A$ (population $B$ looks the same with small fluctuations) after 100 000 time steps during which the ORFs were exposed to constant mutational

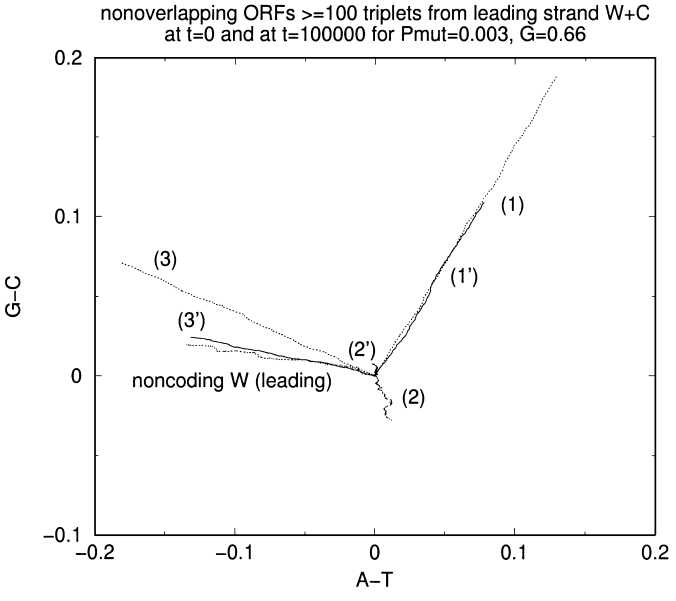


Fig. 6. DNA walks in $[A-T, G-C]$ space for non-overlapping ORFs from leading strand W and C (ORFs are longer or equal to 100 triplets) – legs (1),(2),(3), and walks for the ORFs at $t = 100\,000$ subject to constant mutation rate $p_{mut} = 0.003$, and $G = 0.66$.

pressure and selection. The chosen value of $G$ ensures that the "directions" of the DNA walks (1) and (1′) coincide. In case of Fig. 6 the amount of 100 000 time steps is so large that populations $A$ and $B$ achieve the steady state. Position (3) in codons was not considered for selection. It was only a test if the simulations were correct. In this case this DNA walk should coincide with the DNA walk for the intergenic (non-coding) sequences from the leading DNA strand. We defined the intergenic sequences as these fragments of DNA which do not contain ORFs longer or equal to 100 codons. In Figs. 7 and 8 we show how the base composition at position (1) and (2) of codons in ORFs from population $A$ depends on parameter $G$. Notice that the DNA walks are still strongly anisotropic even if their composition is varying. The two extreme cases are present: when the mutational pressure is much stronger than selection (small value of $G$) and when selection is much stronger than mutational pressure (large value of $G$). One could think of the improved simulation algorithm where selection is applied to all three base positions in codons with different values of $G$ to obtain the same effect of common "directions" for DNA walks as it is presented in Fig. 6 for the first base position in codons.

The values of the base frequencies at every base position in codons are varying during evolution as in Figs. 9–11. One can notice that already after 20 000 generations the populations are in equilibrium. This new equilibrium weakly depends on the parameter $p_{mut}$ representing mutation rate.

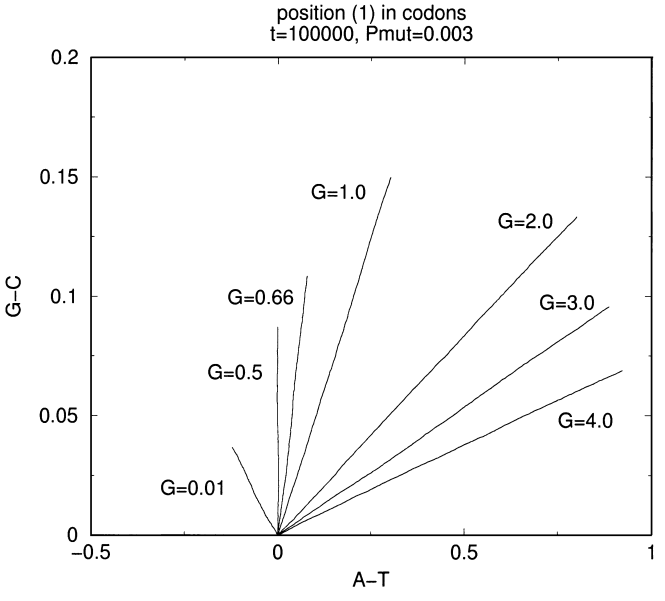


Fig. 7. DNA walk representation of the 1st base position in codons for various temperatures $G = 0.01$, $0.5, 0.66, 1.0, 2.0, 3.0, 4.0$ and constant mutation rate $p_{mut} = 0.003$.

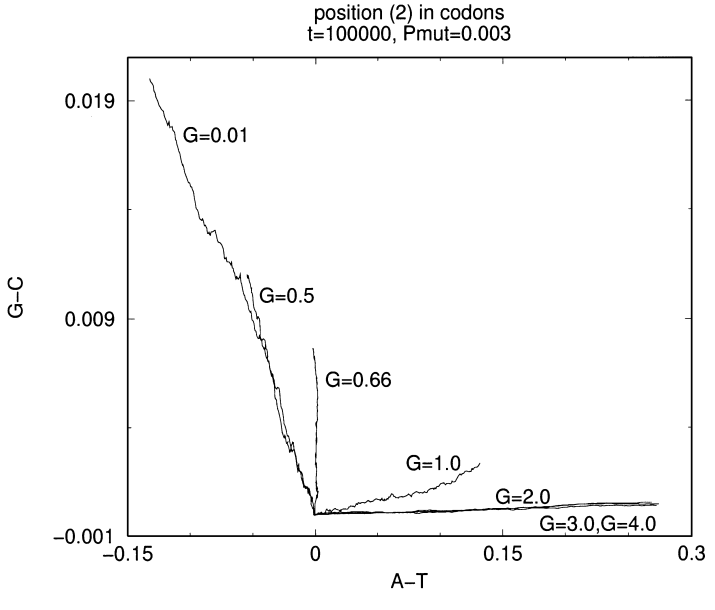


Fig. 8. DNA walk representation of the 2nd base position in codons for various values of $G = 0.01$, $0.5, 0.66, 1.0, 2.0, 3.0, 4.0$ and constant mutation rate $p_{mut} = 0.003$.

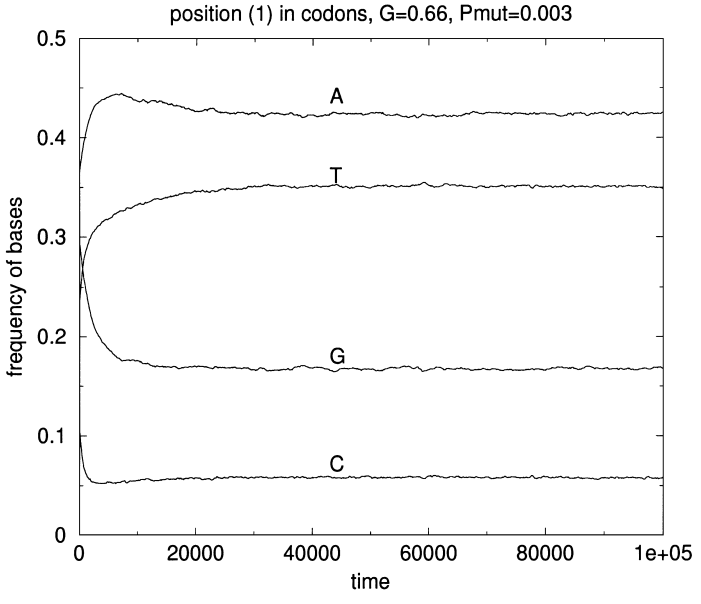


Fig. 9. Composition of the 1st base position in codons vs. time at $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$.

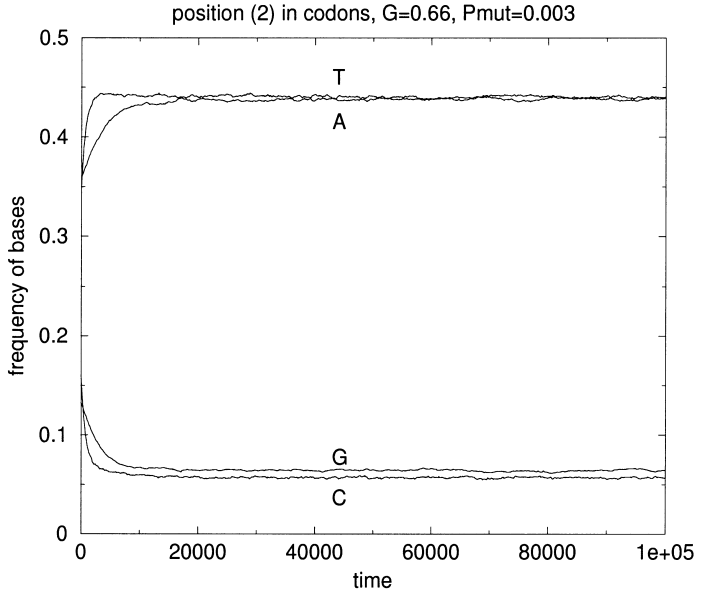


Fig. 10. Composition of the 2nd base position in codons vs. time for $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$.

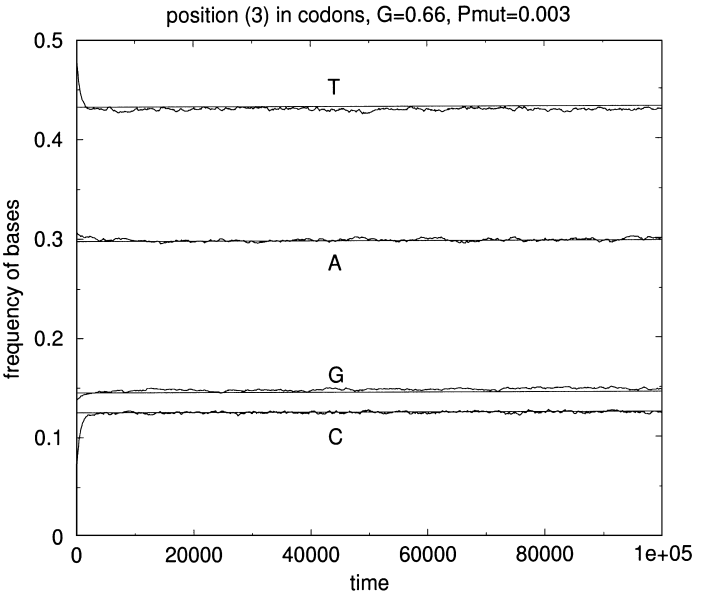


Fig. 11. Composition of the 3rd base position in codons vs. time for $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$.

The number of all substitutions (both lethal and those accepted by selection) depends on the base position in genes. We presented the dependence of the number of all substitutions on time in Fig. 12. One can notice that at equilibrium the mutations at third position dominate over the mutations at other base positions in codons. The different composition of the base positions in codons is responsible for that. It is interesting to compare the result with the fraction of all substitutions which are accepted by selection (Fig. 13). One can observe a strong protection against substitutions in the second base position in codons. This result we could expect from biological observations. We should remember that in our model the selection rules are applied independently to each gene in population $A$ and $B$ and therefore the discussed evolution of genes should be interpreted as evolution of genomes. This interpretation applies also to the results presented in Figs. 12 and 13. The other result of our computer simulations, in Fig. 14, that the number of substitutions accumulated in genes of $A$ and $B$ depends on the genes length, also should be related to genomes. The computer simulations suggest that the number of accumulated substitutions per genome length is less for longer genomes than for short ones. If this observation was confirmed experimentally, one should revise the molecular clock hypothesis. The distances between species might change. Since replication is the major mechanism introducing the mutational pressure, when the fidelity of replication is higher the mutational pressure on long DNA sequences is smaller and therefore the genome can be longer [17–19]. This may also influence the length of genes. In bacterial genomes the accuracy of replication is lower and we observe no long genes.

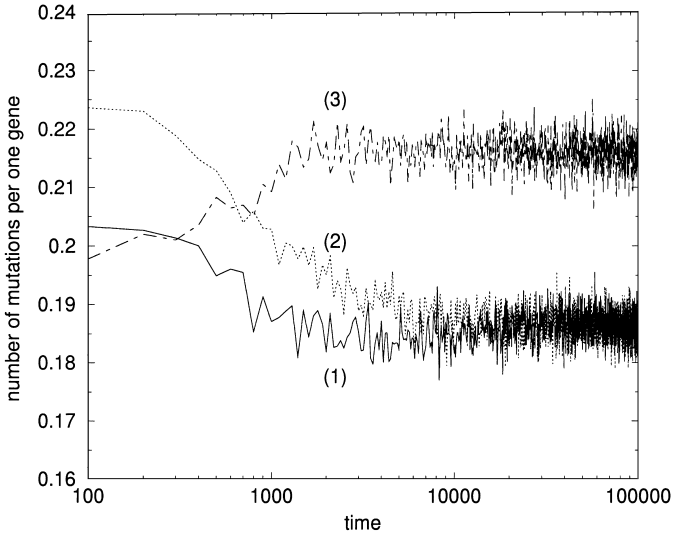


Fig. 12. Number of all mutations at time $t$ per one gene and per one generation vs. time $t$ for $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$ for different base positions in codons. Both the lethal mutations and those which are accepted by selection are included.

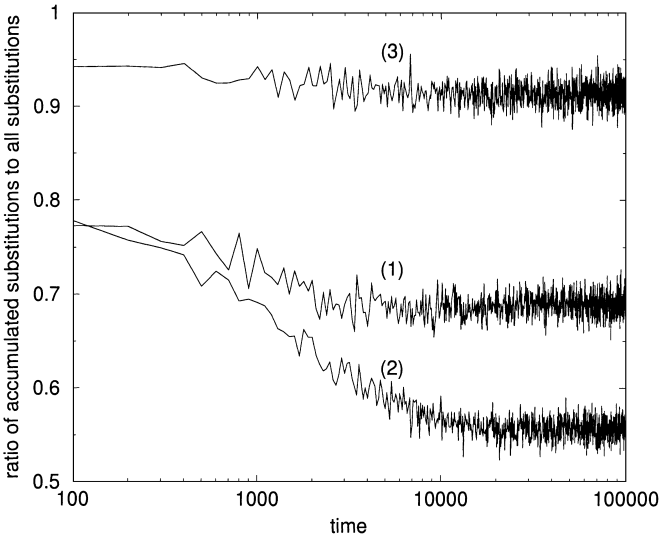


Fig. 13. Ratio of mutations accepted by selection to all mutations per one generation vs. time for $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$.

The selection rules conserve the gene structure even if they are under strong mutational pressure. It is evident from Figs. 15 and 16 and, where the $k$-sequence usage has been shown for the first and the second base position in codons (compare with Figs. 2 and 3).

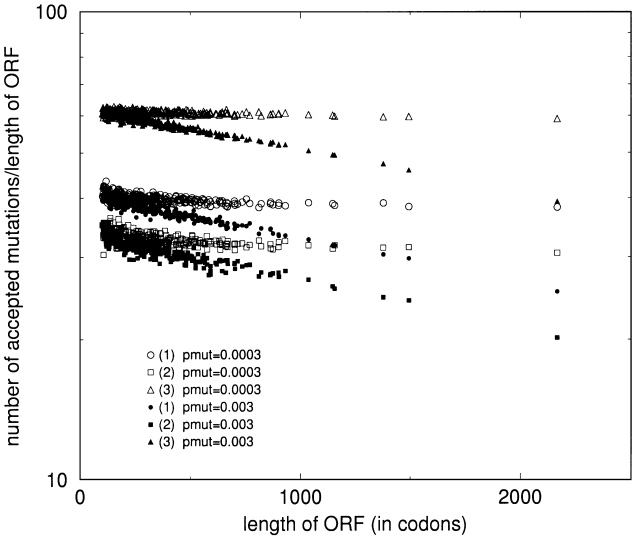


Fig. 14. Dependence of the amount of mutations (per length of ORF) accepted by selection on length of ORFs in population $A$ at $t = 100\,000$. $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$ (empty symbols) and $p_{mut} = 0.0003$ (filled symbols)

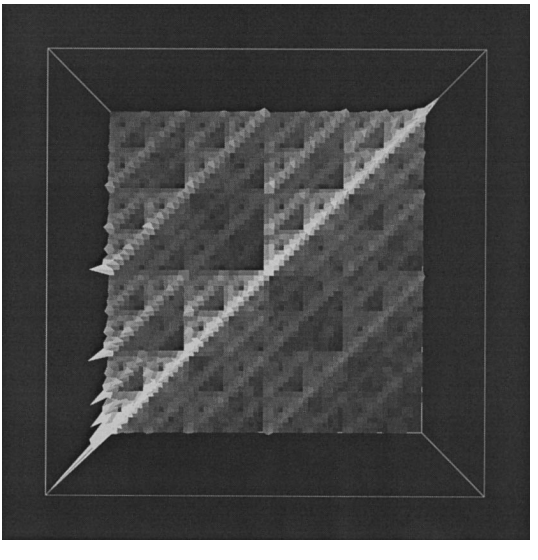

Fig. 15. The usage of the 6-sequences for the first base position in codons for population $A$ of ORFs after 100 000 time steps, when $G = 0.66$ and constant mutation rate $p_{mut} = 0.003$ is applied.

To complete the simulation results we show in Fig. 17 how the number of killed genes (genomes) in our model depends on the value of $G$. One can observe almost linear behavior (the slope $\sim 1$) of the number of killed genes (genomes) in a wide range of parameter $G$ which controls the noise level in accepting mutated genes.

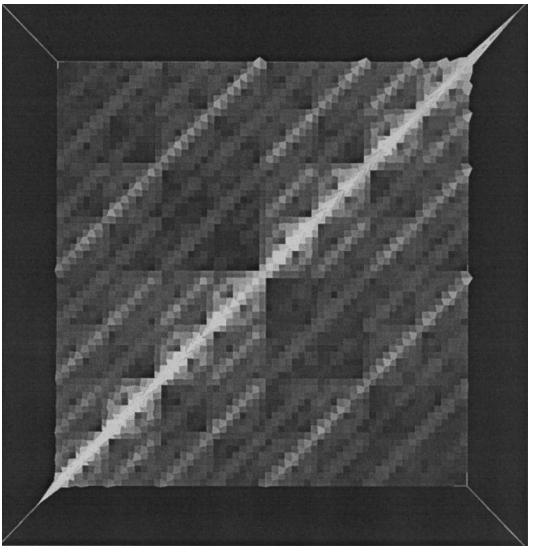

Fig. 16. The same as in Fig. 15 but for the second base position in codons.

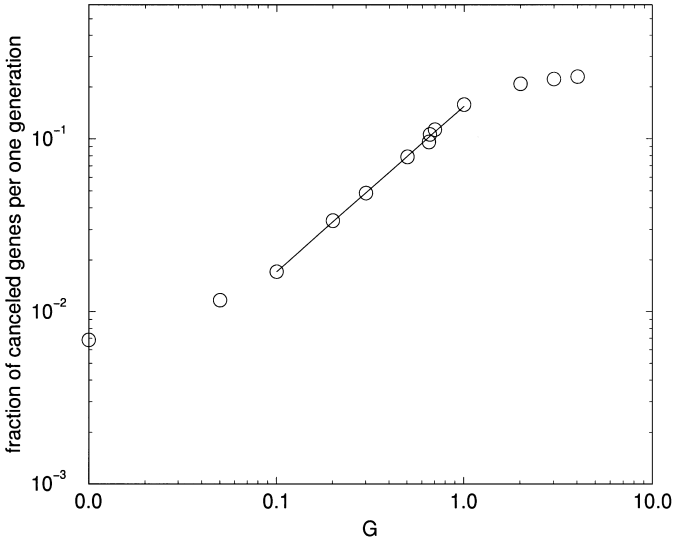


Fig. 17. Number of ORFs killed by selection rules vs. inverse temperature $G$.

## 4. Conclusions

We have constructed a model of DNA evolution, in which DNA sequences of various length are subject to constant mutational pressure and selection. We showed that they behave differently during evolution. The shorter sequences accumulate more

mutations per sequence length than the longer ones. In consequence, the rate of mutations determines genome length.

## Acknowledgements

This research has been supported by KBN Nr 6 PO4A 03014.

## References

[1] W.-H. Li, X. Gu, Physica A 221 (1995) 159.
[2] M. Kimura, J. Mol. Evol. 16 (1980) 111.
[3] J.A. Glazier, S. Raghavachari, Ch.L. Berthelsen, M. Skolnick, Phys. Rev. E 51 (1995) 2665.
[4] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, H.E. Stanley, M.H.R. Stanley, M. Simons, Biophys. J. 65 (1993) 2673.
[5] S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, H.E. Stanley, Phys. Rev. E 51 (1995) 5084.
[6] P. Bak, K. Sneppen, Phys. Rev. Lett. 71 (1993) 4083.
[7] K. Sneppen, Physica A 221 (1995) 168.
[8] C.I. Wu, W.H. Li, Proc. Natl. Acad. Sci. USA 82 (1985) 1741.
[9] M.Ya. Azbel, Physica A 273 (1999) 75–91 [these proceedings].
[10] C.M. Fraser, S. Casjens, W.M. Huang, G.G. Sutton, R. Clayton, R. Lathigra, O. White, K.A. Ketchum, R. Dodson, E.K. Hickey et al., Nature 390 (1997) 580.
[11] P. Mackiewicz, A. Gierlik, M. Kowalczuk, D. Szczepanik, M.R. Dudek, S. Cebrat, Physica A 273 (1999) 103–115 [these proceedings].
[12] S. Cebrat, M.R. Dudek, Eur. Phys. J. B 3 (1998) 271.
[13] S. Cebrat, M.R. Dudek, P. Mackiewicz, M. Kowalczuk, M. Fita, Microb. Comp. Genom. 2 (4) (1997) 259.
[14] S. Cebrat, M.R. Dudek, P. Mackiewicz, Theory Biosci. 117 (1988) 78.
[15] H. Joel Jeffrey, Nucleic Acid Res. 18 (1992) 2163.
[16] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21 (1953) 1087.
[17] J.W. Drake, Ann. NY Acad. Sci. 870 (1999) 100.
[18] J.W. Drake, B. Charlesworth, D. Charlesworth, J.F. Crow, Genetics 148 (4) (1998) 1667.
[19] J.W. Drake, Proc. Natl. Acad. Sci. USA 88 (16) (1991) 7160.