ELSEVIER

Physica A 314 (2002) 646–654

PHYSICA A

www.elsevier.com/locate/physa

Replication associated mutational pressure generating long-range correlation in DNA

Paweł Mackiewicz[a], Maria Kowalczuk[a], Dorota Mackiewicz[a], Aleksandra Nowicka[a], Małgorzata Dudkiewicz[a], Agnieszka Łaszkiewicz[a], Mirosław R. Dudek[b], Stanisław Cebrat[a,*]

[a] *Department of genetics, Institute of Microbiology, ul. Przybyszewskiego 63/77, Wrocław 51-148, Poland*
[b] *Institute of Physics, University of Zielona Góra, Zielona Góra 65-069, Poland*

Abstract

There are many biological mechanisms which introduce long-range correlations into the DNA molecule. One of the most important is replication of chromosomes, its mechanisms and topology. Replication associated mutational pressure, defined as specific preferences in nucleotide substitutions during replication, generates asymmetry in the genome. On the other hand, substitution rates, which determine the evolutionary turnover time of a nucleotide, are highly correlated with the fraction of that nucleotide in the genome. Assuming the Azbel hypothesis that the number of mutations per genome per generation is invariant and universal, a general rule for mutational pressure can be formulated: the half-time of a nucleotide turnover in the genome is linearly dependent on the number of this nucleotide in the genome.
© 2002 Elsevier Science B.V. All rights reserved.

PACS: 87.14.G; 05.10.−a

Keywords: DNA evolution; Replication; Tables of substitutions

1. Introduction

The DNA molecule is a sequence of four different nucleotides: Adenine (A), Guanine (G), Thymine (T), and Cytosine (C). Fifty years ago, Chargaff [1,2] noted

[*] Corresponding author. Tel.: +48-71-3247303; fax: +48-71-3252151.
E-mail address: cebrat@microb.uni.wroc.pl (S. Cebrat).
URL: http://smORFland.microb.uni.wroc.pl

PII: S0378-4371(02)01167-6

a species invariant component of DNA composition: [A]=[T] and [G]=[C], called Parity Rule 1 (PR1). This rule helped Watson and Crick to construct the double stranded model of DNA which assumed the complementarity rule stating that A at one strand corresponds to T at the other strand and G corresponds to C. Thus, PR1 turned out to be a deterministic rule. If DNA sequence had random nucleotide composition this parity rule should be also in force for each of the two DNA strands. In such a case the rule [A] = [T] and [G] = [C] is a stochastic one and it is called the Parity Rule 2 (PR2) [3]. Actually, natural DNA does not have random sequence of nucleotides and there are many mechanisms which introduce local disturbances and deviations from the state resulting from the second parity rule. These deviations are called DNA asymmetry (see [4]). The most important mechanisms introducing asymmetry into DNA molecules are:

- selection, keeping the coding sequences very asymmetric, and
- mutational pressure associated with replication.

About 10 years ago, Li et al. [5,6], Peng et al. [7] and Voss [8], who analyzed DNA walks, found long-range base–base correlations in DNA sequences. They noted [5–7] that the strongest long-range correlations are seen in intergenic sequences, while coding sequences have only short-range correlations [8]. Further studies have shown that correlations could be also observed in protein coding sequences if they are read in a specific order [9–12]. Some sequences, out of protein coding ones, code for rRNA or tRNA or have regulatory functions, but most of them are not coding, and are not under selection forces. Thus, they should reflect the mutational pressure and the observed correlations would result from this pressure. In this paper we describe how the specific asymmetry observed in intergenic sequences is generated by replication associated mutational pressure, and some universal properties of the mutational pressure operating in many genomes.

2. Materials and methods

DNA sequence of the *Borrelia burgdorferi* genome was downloaded from www.ncbi.nlm.nih.gov. The DNA walks were performed in different variants (see [12] for details). Two-dimensional variants were described previously by Mizraji and Ninio [13], Gates [14], Berthelsen et al. [15], Lobry [16], and Cebrat et al. [9]. The shifts of the walker are: (0,1) for G, (1,0) for A, (0,-1) for C and (-1,0) for T. The walks can be used to picture any DNA sequence, or in a specific version to picture the composition of each codon position separately in a protein coding gene. In other versions of DNA walks, the walker shows directly the deviations from the PR2 rule. In the [A–T] version of the walk the walker checks each position on chromosome (x-axis coordinates) and its shifts are: (1,1) for A, (1,-1) for T and (1,0) for G and C. In the [G–C] version of the walk, the corresponding shifts are: (1,1) for G, (1,-1) for C and (1,0) for A and T [17]. Thus, the plot shows the relative abundance of A over T or G over C (y-axis coordinates) on one DNA strand. Matrix of nucleotide substitutions for the *B. burgdorferi* genome was found experimentally as described by Kowalczuk et al. [18].

3. Results and discussion

In Fig. 1 a two-dimensional DNA walk performed on the whole *B. burgdorferi* genome is shown. The walker steps was as follows: for G (0,1), for A (1,0), for C (0,-1) and for T (-1,0). Another plot in the Fig. 1 shows the same type of DNA walk and in the same scale performed on intergenic sequences from leading part of Watson strand spliced together (all protein coding sequences were cut off the analyzed genome). There are very strong compositional trends in both the whole genomic sequence and the intergenic sequences. Furthermore, there is a specific point where these trends change their signs for reciprocal ones. Comparing this plot with the topology of the *B. burgdorferi* chromosome one can notice that this specific point is the origin of replication of the chromosome (Fig. 2). Each arm of the same DNA strand is replicated by different mechanisms. One is replicated continuously and it is called the leading strand, while the other one is replicated discontinuously by joining fragments and it is called the lagging strand. If we perform walks of the [A–T] or [G–C] type, keeping the chromosome scale, the position of the origin of replication can be localized very precisely [19,20], DNA walk on protein coding sequences alone performed

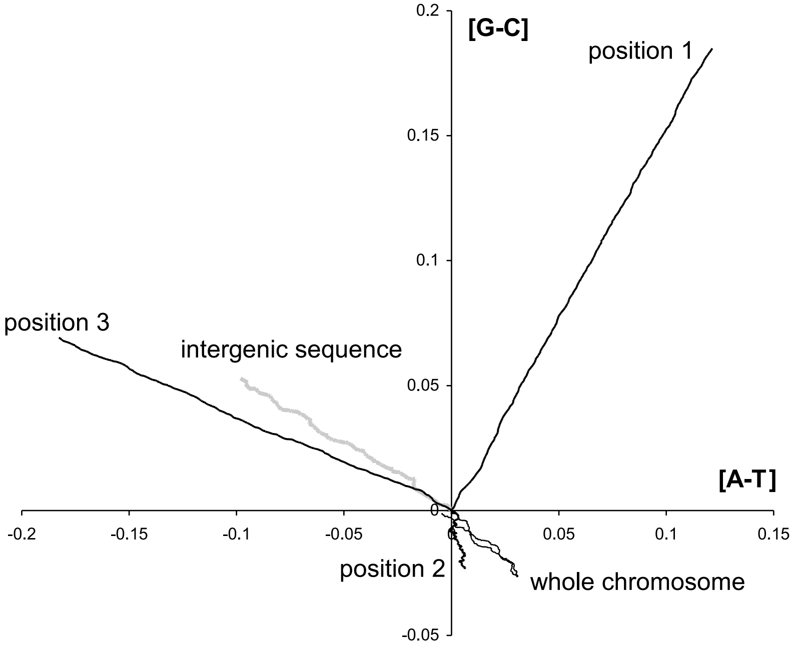

Fig. 1. DNA walks in the two-dimensional space. One walk was performed on the whole Watson strand of the *B. burgdorferi* chromosome, and the other one on intergenic sequences of the leading part of the Watson strand (walker visited each nucleotide of the analyzed sequence). Walks describing positions 1, 2 or 3 represent “jumps” of the walker separately for the first, the second and the third positions in codons of all spliced ORFs from leading strand of the genome.

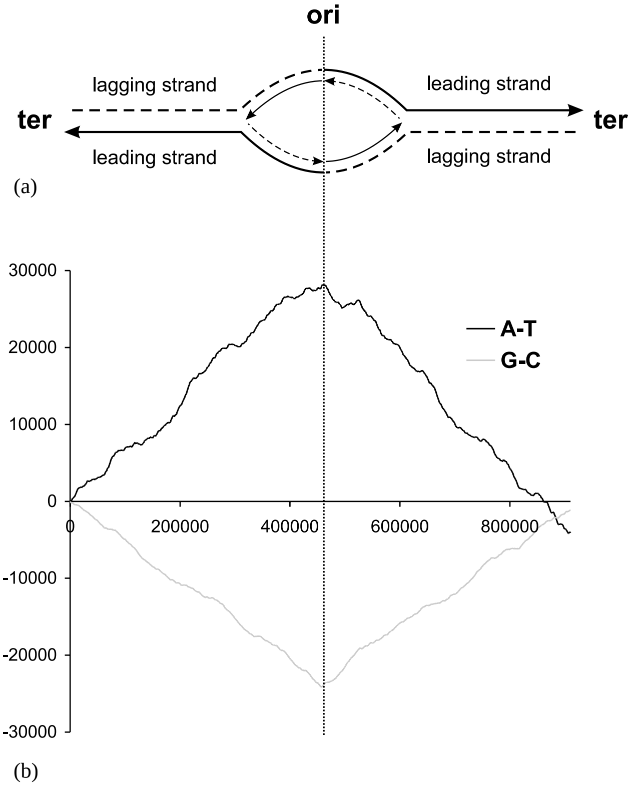

Fig. 2. Topology of replication of the linear *B. burgdorferi* chromosome (a), walks of [A–T] and [G–C] type performed on Watson strand (the upper one) (b), x-coordinates correspond to the number of nucleotides in the chromosome.

on eukaryotic genes produces a plot which resembles random Brownian movements rather than the walk on intergenic sequences, unless we introduce some order into these sequences. For example, we can splice together only the genes whose coding sequences are located on one DNA strand, e.g. the leading strand, and instead of walking we will jump every three nucleotides. That means that the walker performs three different walks, one for each position in codons separately. Now, very strong correlations are seen for each walk (Fig. 1). Note that the compositional bias in the third positions in codons is similar to the bias seen in the intergenic sequences. It is known that the third codon positions are the most degenerated ones and many types of mutations in these positions do not change the meaning of the codon. These mutations are called silent. This suggests that the similarity between the composition of the third positions and intergenic sequences results from accumulation of mutations during replication of the chromosome, and is not influenced by selection. We have assumed that at least some of the intergenic sequences were generated by recombination mechanisms transferring coding sequences or their parts into intergenic space. Thus, such “pseudogenes” as

Table 1
Table of substitution frequencies in leading strand of the *B. burgdorferi* genome (BbTS)

	A	T	G	C	Σ
A	—	0.1027	0.0667	0.0228	0.1922
T	0.0655	—	0.0347	0.0350	0.1352
G	0.1637	0.1157	—	0.0147	0.2941
C	0.0702	0.2613	0.0470	—	0.3785

The nucleotide indicated in the first column is substituted by the nucleotide indicated in the first row. The last column is the sum of frequencies of substitution of a given nucleotide by any other nucleotide. Note that all frequencies are normalized to 1.

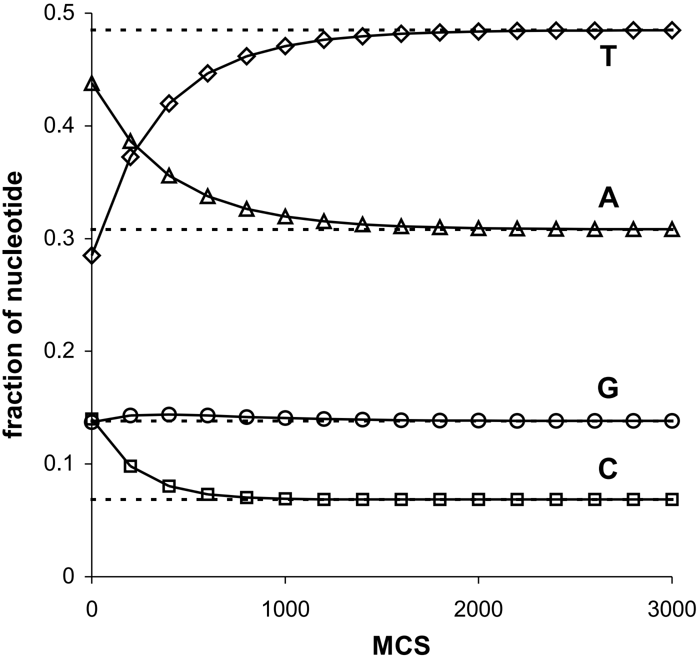

Fig. 3. Evolution of the sequence composition under the BbTS mutational pressure. The initial sequence had the composition of protein coding sequences from the lagging strand of the *B. burgdorferi* chromosome. After about 2000 Monte Carlo steps fraction of each nucleotide reached the composition of the third positions (corresponding dotted lines levels).

remnants of duplicated genes should accumulate mutations freely. By comparing the original gene to its pseudogene counterpart we were able to estimate the frequency of each type of substitution in the *B. burgdorferi* genome. These frequencies, representing the mutational pressure, are presented in Table 1 (BbTS).

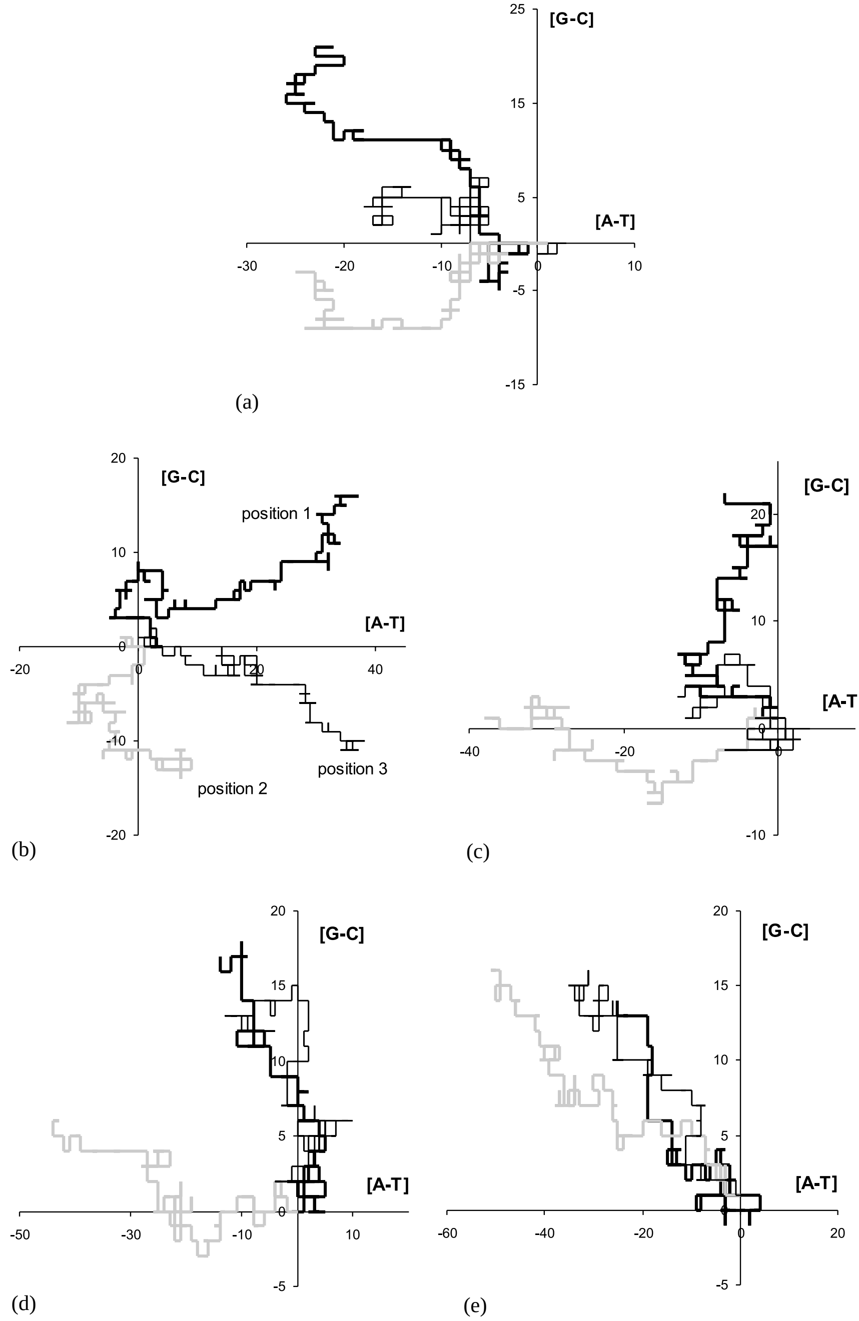

Fig. 4. Two-dimensional DNA walks performed separately for each position in codons for: (a) sequence homologous to the gene of signal peptidase 1, (b) sequence of the gene coding for signal peptidase 1, and (c)–(e) the sequences of the gene after simulated evolution under the BbTS mutational pressure after 300, 600 and 900 MCS, respectively.

4. Properties of the substitution matrix

We have tested the matrix of substitutions found for the *B. burgdorferi* genome in computer simulations. When coding sequences of the *B. burgdorferi* genome were allowed to freely accumulate mutations introduced with the frequency described by the BbTS matrix, after several generations both the first and the second positions reached the nucleotide composition of the third positions (Fig. 3). Furthermore, we have found some pseudogenes in this genome which seem caught in the process of disintegration by mutations (Fig. 4). The BbTS matrix has other interesting properties. Let us imagine the set of adenines in the genome. Knowing the frequency of substitutions of adenine by any other nucleotide it is possible to count the time period when half of all adenines are substituted. Such evolutionary turnover rate of nucleotides is highly correlated with their fraction F_N in the DNA with correlation coefficient close to 1 (Fig. 5). This property is universal for any substitution matrix found in other genomes unless they are "contaminated" with the effect of selection. The correlation between half-time of nucleotide turnover and F_N is not an intrinsic feature of any table of substitution. We have generated by computer thousands of substitution matrices which were able to generate DNA in equilibrium with the same composition as BbTS does. Some of these matrices give correlations between turnover rate and the fraction of nucleotides but many of them do not possess such property. After analyzing these matrices we

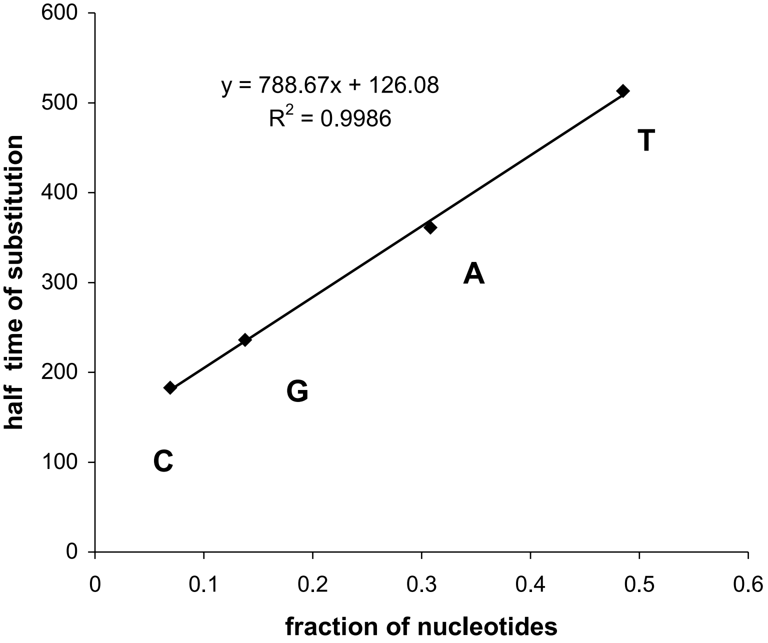

Fig. 5. Relation between the fraction of nucleotides in third codon position of coding sequences from the leading strand of the *B. burgdorferi* chromosome and the half-time of their substitution.

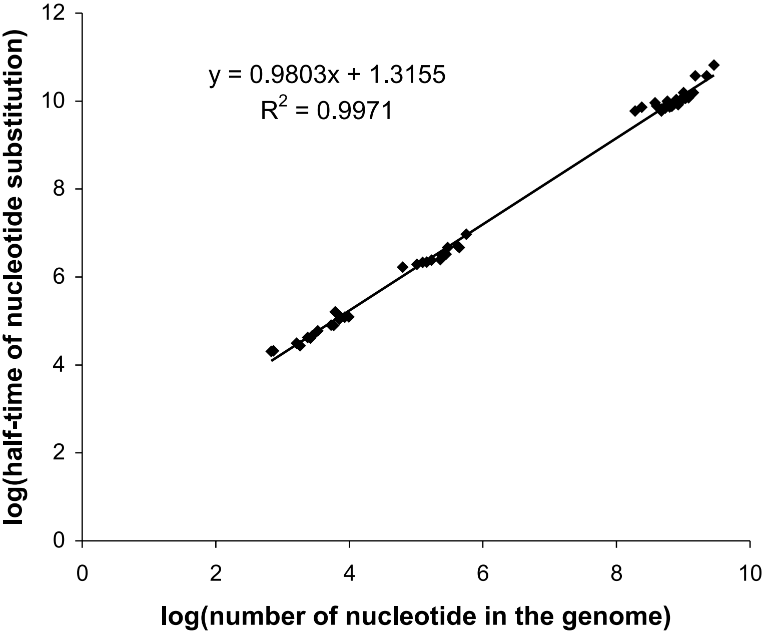

Fig. 6. Relation between the number of nucleotides in a genome and their turnover rates measured by half-time of their substitutions.

have found that matrices introducing relatively low number of mutations into DNA in equilibrium have this property. Moreover, they resemble the natural BbTS matrix. Azbel [21] suggested that the optimal rate of mutations should be around one mutation per genome per generation, independently of the size of the genome. Many experimental results have shown that at least the order of this estimation is correct [22]. We have tested mutational matrices found by other authors for genomes of different sizes and counted the composition of DNA in equilibrium with the corresponding mutational matrix. Then, assuming Azbel's suggestion we have plotted the half time of turnover of nucleotides against their numbers in the genome. The result is shown in Fig. 6. It is obvious that the mutational pressure is the result of evolution. It seems trivial that to keep the frequency of substitutions in the genome at a low level it is best to lower the probability of making an error for this very nucleotide whose fraction in the genome is the highest. Unfortunately, such an understanding does not make trivial the observed linear correlations.

Acknowledgements

This work was supported by The State Committee for Scientific Research, grant numbers 6 P04A 025-18 and 6 P04A 016-20. P.M. was supported by Foundation for Polish Science.

References

[1] E. Chargaff, Experientia 6 (1950) 201.

[2] E. Chargaff, Fed. Proc. 10 (1951) 654.

[3] R. Rudner, J.D. Karkas, E. Chargaff, Proc. Natl. Acad. Sci. USA 60 (1968) 921.

[4] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, J. Appl. Genet. 42 (2001) 553.

[5] W. Li, Int. J. Bifurc. Chaos 2 (1992) 137.

[6] W. Li, K. Kaneko, Europhys. Lett. 17 (1992) 655.

[7] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons, H.E. Stanley, Physica A 221 (1995) 180.

[8] R. Voss, Phys. Rev. Lett. 68 (1992) 3805.

[9] S. Cebrat, M.R. Dudek, P. Mackiewicz, M. Kowalczuk, M. Fita, Microb. Comp. Genomics 2 (1997) 259.

[10] S. Cebrat, M.R. Dudek, A. Rogowska, J. Appl. Genet. 38 (1997) 1.

[11] S. Cebrat, M.R. Dudek, Eur. Phys. J. B 3 271.

[12] S. Cebrat, M.R. Dudek, A. Gierlik, M. Kowalczuk, P. Mackiewicz, Physica A 265 (1999) 78.

[13] E. Mizraji, J. Ninio, Biochimie 67 (1985) 445.

[14] M.A. Gates, J. Theor. Biol. 119 (1986) 281.

[15] Ch.L. Berthelsen, J.A. Glazier, M.H. Skolnick, Phys. Rev. A 45 (1992) 8902.

[16] J.R. Lobry, Biochimie 78 (1996) 323.

[17] P. Mackiewicz, A. Gierlik, M. Kowalczuk, M.R. Dudek, S. Cebrat, J. Appl. Genet. 40 (1999) 1.

[18] M. Kowalczuk, P. Mackiewicz, D. Szczepanik, A. Nowicka, M. Dudkiewicz, M.R. Dudek, S. Cebrat, Int. J. Mod. Phys. C 12 (2001) 1043.

[19] A. Grigoriev, Nucl. Acid Res. 26 (1998) 2286.

[20] A. Zawilak, S. Cebrat, P. Mackiewicz, A. Krl-Hulewicz, D. Jakimowicz, W. Messer, G. Gociniak, J. Czerwika-Zakrzewska, Nucl. Acid Res. 29 (2001) 2251.

[21] M.Y. Azbel, Physica A 273 (1999) 75.

[22] J.W. Drake, B. Charlesworth, D. Charlesworth, J.F. Crow, Genetics 148 (1988) 1667.