CMST 13(1) 5-12 (2007) DOI:10.12921/cmst.2007.13.01.05-12

COMPUTATIONAL METHODS IN SCIENCE AND TECHNOLOGY **13(1)**, 5-12 (2007)

Information Weights of Nucleotides in DNA Sequences

M. R. Dudek[1], S. Cebrat[2], M. Kowalczuk[2], P. Mackiewicz[2], A. Nowicka[3], D. Mackiewicz[2], M. Dudkiewicz[3]

[1]*Institute of Physics, University of Zielona Góra, ul. A. Szafrana 4a, 65-516 Zielona Góra, Poland*

[2]*Division of Genomics, University of Wroclaw, ul. Przybyszewskiego 63/77, 54-148 Wroclaw, Poland*

[3]*Faculty of Agriculture, Department of Biometrics, SGGW, Warszawa, Poland*

(Rec. Juny 22, 2005)

Abstract: The protein sequence is coded with the help of the triplets of nucleotides, each corresponding to one amino acid in a protein sequence. The triplet code of the coding sequences possesses some informative redundancy. Some triplets are more probable than others. The analogous redundancies appear in all natural languages. The non-equal frequency of the characters in plain text makes possible that entire words can be predicted given the context of the plain text. This is typical problem in cryptanalysis that a plain text is compressed before encrypting it in order to reduce the language redundancies. The nucleotides represent the natural units to discuss the redundancies in the coding sequences of natural genomes. The mutation pressure and selection pressure are the main factors responsible for the observed redundancies. Then, the nucleotide frequency in DNA seems to be the natural information weight. We show, that the probability of a nucleotide to stay non-mutated becomes another, the more efficient information weight. It has smaller redundancy although it is correlated with the nucleotide frequency. We have found the values of probability for nucleotide to stay non-mutated in the particular case of the *Borrellia burgdorferi* genome. In order to examine the usefulness of the new frequencies we used them in a problem of bit-string packing in a channel with a given capacity. We performed a computer experiment, in which we have generated all possible oligomers consisting of *k* nucleotides and we have shown, that if the number of bits of the information carried out by the oligomers does not exceed a given threshold value, the same as calculated for genes of the *Borrelia burgdorferi* genome, then the distribution of the generated oligomers resembles the one used by these genes.

Key words: protein sequence, *Borrelia burgdorferi* genome, oligomers, codon

I. INTRODUCTION

Since the famous paper by Crick et al. [1] it is generally accepted that there exists, in all living organisms, a code that makes possible information transfer from DNA to proteins. Namely, the protein sequence is coded by codons in DNA sequences, which are the triplets of nucleotides, each corresponding to one amino acid. There are 64 possible triplets and only 20 amino acids. Hence, the genetic code is degenerated. Crick et al. [1] proved that codons are always read from the start, codon after codon, they do not overlap, there are no commas between them. Therefore, each strand of the coding region of a DNA molecule could be read in three different reading frames and all the statistical analyses of the coding properties of the DNA sequences should be done in a specific reference system consistent with the triplet nature of the genetic code [2-4].

The coding DNA sequence can be thought as a message necessary to be transferred from source to receiver through a noisy information channel, e.g. [5, 6]. Then, the four letter alphabet, *A*, *T*, *G*, *C*, which denotes DNA nucleotides representing the source, should be translated into a 20 letter alphabet of amino acids in protein representing the receiver. The nucleotides have non-equal frequencies, f_A, f_T, f_G, f_C, of their occurrence, which is the result of the mutational pressure and selectional pressure. Thus, these frequencies express the natural redundancy of genetic code in the genome under consideration. Shannon [7] considered the generation of a message to be a Markov process, subject to a noise, and he introduced an expression

$$H = -k \sum_i p_i \log f_j , \tag{1}$$

known as information entropy, which is a measure of information in a message, and j denotes letters of the alphabet under consideration, f_j represents the probability of the symbol j, and the coefficient k is for the purpose of a unit of measure. In the case of binary alphabet, one usually chooses logarithm base 2 in the above expression. Then, the entropy of a message consisting of symbols *A*, *T*, *G*, *C* reads as

$$H = \sum_{j=A,T,G,C} f_j \log_2 \left(\frac{1}{f_j} \right) , \tag{2}$$

where

$$B_j = \log_2\left(\frac{1}{f_j}\right) \qquad (3)$$

represents the number of bits necessary to code nucleotide j in optimal way. If there is no additional information about the frequency of the occurrence of the nucleotides, then two bits are necessary for each nucleotide. On the other hand, the codons (triplets) need six bits of information for each codon, whereas the amino acids need four bits. The sense of the Shannon's channel capacity theorem [7] is that if one wants to send a message from source to receiver, with as few errors as possible, then the code of the message should posses some redundancy, because the channel adds noise to the transmitted message. For example, in coding sequences of the *Borrellia burgdorferi* genome, which we will address at the below we have the following nucleotide fractions and numbers of bits:

$$f_A = 0.3060, f_T = 0.4901, f_G = 0.1365, f_C = 0.0675, \qquad (4)$$

$$B_A = 1.7084,\ B_T = 1.0289,\ B_G = 2.8730,\ B_C = 3.8890. \qquad (5)$$

Then, the total number of bits representing nucleotide frequencies is equal to $B = B_A + B_T + B_G + B_C \approx 9.5$, i.e., there is necessary about 2.4 bits per nucleotide. In the following section, we will show that the information carried by the probability that these nucleotides will stay non-mutated is much less redundant because in this case we have got only about 2.01 bits per nucleotide.

The problem is closely related with designing DNA codes, which is important for biotechnology applications, e.g., in storing and retrieving information in synthetic DNA strands or as molecular bar codes in chemical libraries. This has been discussed recently by Marathe et al. [12] (see also very rich literature within). The necessity of the suitable weight given to nucleotides, in such a way that they be consistent with the amino acid structure, is notified also in papers dealing with DNA symmetry consideration in terms of group theory [11].

II. INFORMATION WEIGHTS WITH DECREASED REDUNDANCIES

In papers [13-15], we concluded that in natural genome the frequency of occurrence f_j of the nucleotides ($j = A, T, G, C$), in the third position in codons, is linearly related to the respective mean survival time τ_j,

$$f_j = m_0 \tau_j + c_0, \qquad (6)$$

with the same coefficients, m_0 and c_0, for each nucleotide. The coefficient m_0 is proportional to mutation rate u, experienced by the genome under consideration. This means, that in natural genome, with balanced mutation pressure and selectional pressure, the nucleotide occurrences are highly correlated. This observation does not contradict to the Kimura's neutral theory [16] of evolution, which 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. Actually, the mutations are random but they are correlated with the DNA composition. Thus, the frequency f_j contains information specific for genome and therefore it seems to be a natural candidate to model information weight for nucleotides. However it possesses a large informative redundancy, mainly due to the mutation pressure.

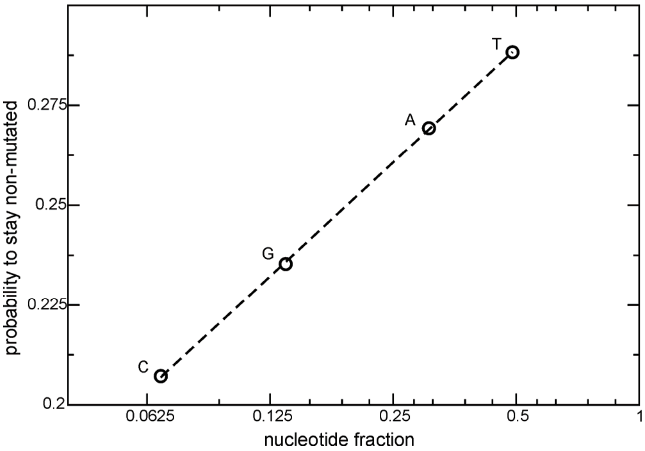


Fig. 1. Relation between fraction of nucleotides representing third position in triplets of the coding sequences of the *Borrellia burgdorferi* genome and the probability that they stay non-mutated. The lower bound for the probability is presented, when the mutation rate $u = 1$

The question rises, whether we could estimate a nucleotide fraction $\tilde{f}_j$, which is not influenced by nucleotide substitutions. To answer the question, we considered the probability that nucleotide j becomes non-mutated, which is equal to

$$P_j = 1 - uW_j = 1 - u\sum_{i \neq j} W_{ji}, \qquad (7)$$

where W_j is the relative mutation probability of nucleotide j, W_{ji} is the relative probability that the nucleotide j is changed to the nucleotide i, and $W_A + W_T + W_G + W_C = 1$. The parameter u represents mutation rate. In the case of the *Borrellia burgdorferi* genome, we have found [13-15] an empirical mutation table, applying for genes of leading DNA strand (it is transpose matrix for genes of lagging DNA strand), where the values of W_{ij} are the following:

$$\begin{array}{lll} W_{AG} = 0.0667 & W_{TG} = 0.0347 & W_{CG} = 0.0470 \\ W_{GA} = 0.1637 & W_{TA} = 0.0655 & W_{CA} = 0.0702 \\ W_{GT} = 0.1157 & W_{AT} = 0.1027 & W_{CT} = 0.2613 \\ W_{GC} = 0.0147 & W_{AC} = 0.0228 & W_{TC} = 0.0350 \end{array} \qquad (8)$$

In the extreme case of $u = 1$, we obtain the lower bound for the probability P_j (Eq. 7) that nucleotide j becomes unchanged by mutation at some instant of time t. We used these values to construct information weights for nucleotides, which are free of the mutation pressure. To this aim, we normalized the probabilities, P_j, and each nucleotide has been assigned a value

$$B_j = \log_2 \left(1/P_j\right), \tag{9}$$

being an average number of bits necessary to code this nucleotide in optimal way. We have got the following values for the information capacity of the particular nucleotides,

$$B_A = 1.89278, B_T = 1.79457, B_G = 2.08743, B_C = 2.27122 \tag{10}$$

This means that the average number of bits per nucleotide is equal to

$$\frac{1}{4} \sum_{i=A,T,G,C} B_i \approx 2.01. \tag{11}$$

In Fig. 1, there is presented in the log-normal plot the probability that nucleotide in the third position in codons stays non-mutated versus the nucleotide fraction in this position. The value of $u = 1$. We can observe that $P_j \sim \log\left(f_j\right)$. The consequence of this exponential law is that the frequency, $\tilde{f}_j$, of the non-mutated nucleotide j, which is defined as

$$\tilde{f}_j = P_j f_j. \tag{12}$$

is, in practice, correlated linearly with their fraction f_j.

The numbers of bits in Eq. 10 can be compared to the numbers of bits in Eq. 5 representing the nucleotide fraction f_j in the position (3) in codons of genes which have about 2.4 bits per nucleotide. The same calculations for the normalized fractions $\tilde{f}_j$ yield even larger number of bits per nucleotide which is about 2.5 bits. We may conclude that the probabilities P_j represent the weights with the smallest redundancies from among the considered examples.

In the case of the *B. burdorferi* genome we have succeeded with the knowledge of the mutation table. We have not such table for amino acids, even in this genome. However, there is known a table of amino acid substitutions published by Jones et al. [8], which results from statistical analysis of 16130 protein sequences from few species. Unluckily, this table includes both the mutational pressure and selectional pressure. The table represents so called PAM1 matrix, corresponding to 1 percent of substitutions between two compared sequences. For each amino acid we calculated the probability $P(j)$ ($j = 1, 2, ..., 20$) that it is unchanged, in the same way as for nucleotides in Eq. 7. Next, we calculated for them the respective number of bits B_j. They take the values presented in Table 1. For the data in the table we have got about 4.5 bits per amino acid.

Table 1 Average number of bits representing 20 amino acids calculated with the help of the PAM1 substitution table published by Jones et al. [8]

Amino acid	Average number of bits
A	3.9382
R	3.3455
N	4.9298
D	4.9286
C	4.9177
Q	4.9273
E	4.9307
G	3.9266
H	4.9230
I	4.3469
L	3.3503
K	4.9301
M	5.9207
F	4.9221
P	3.9233
S	3.9344
T	3.3570
W	5.9157
Y	4.9228
V	3.9362

III. BIT-STRING PACKING IN AN INFORMATION CHANNEL

The DNA sequence representing chromosome under consideration is very long even if we would have restrict ourselves only to genes. Therefore, we have decided to consider the fragments of genes represented by oligomers consisting of k nucleotides. Because the coding sequence has to be read in triplets of nucleotides the number k is a multiple of 3, $k = 3n$, where n is some integer. We analyze independently three subsequences of nucleotides from position (1), (2) and (3) in codons, read codon after codon, instead of one sequence of the total oligomer. Thus, each subsequence consists of n nucleotides and therefore there are possible at most 4^n different sequences because there are four nucleotides. In the case of non-equal nucleotide frequencies we can expect that the value of the total number of bits representing each subsequence

$$B = \sum_{i=1}^{n} B_i \tag{13}$$

could be different for unlike subsequences. We decided to test the usefulness of the information weights introduced in the previous section in the problem of pattern recognition, where we have compared the computer generated sequences of n nucleotides to the ones taken from genes in the *Borrellia* *burgdorferi* genome. Therein we have partitioned all genes under consideration into non-overlapping *k*-oligomers and their nucleotides have been assigned the information weight according to Eq. 10. The nucleotides in all 4^n sequences generated with the help of computer also have assigned the same values of the weight as those which are defined in Eq. 10. In natural genome we have many repetitions of the particular sequences but the number of unlike classes of sequences associated with the particular position in codons is always less than or equal to 4^n.

Now, let us imagine a computer experiment in which we have two banks of unlike sequences consisting of $n = 6$ nucleotides, e.g., the first bank has been obtained for position (3) of codons in genes of the *Borrellia burgdorferi* genome and the second bank possesses all $4^n = 4096$ subsequence experiment rest on the choosing from these two banks only these sequences for which the total number of bits defined in Eq. 13 does not exceed a given threshold value. We have called this threshold value the channel capacity on account of the analogy with the Shannon's channel capacity theorem. We change the assumed value of the channel capacity and each time when we do this we count the number of sequences from these two banks which coincide with the help of an overlap parameter being defined as follows:

$$q = \frac{x}{x + y + z}, \tag{14}$$

where x denotes the number of sequences, which are exactly the same in two banks, y denotes the number of sequences from the second bank, which are different from those in the first bank, and z denotes the number of these sequences in natural genome (the first bank) where the total number of bits exceeds the given channel capacity.

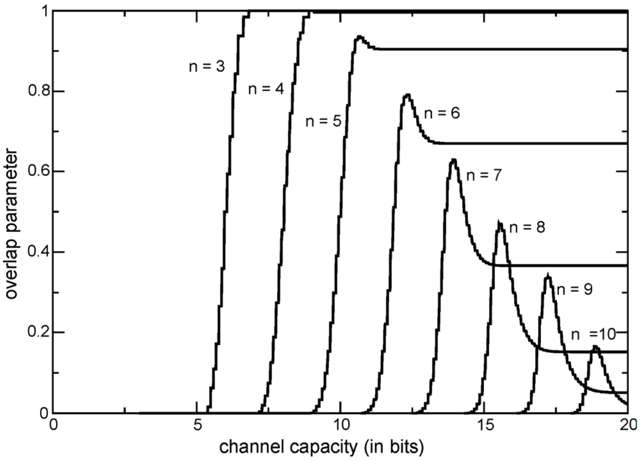


Fig. 2. Dependence of the overlap parameter q in Eq. 14 on the given channel capacity in the case of two banks of n-oligonucleotides, the first bank represents the position 3 in codons of the genes in the *Borrellia burgdorferi* genome and the second bank represents all possible sequences consisting of n nucleotides. The plots have been done for the values of $n = 3, 4, 5, 6, 7, 8, 9, 10$. The nucleotides have been assigned information weights defined in Eq. 10

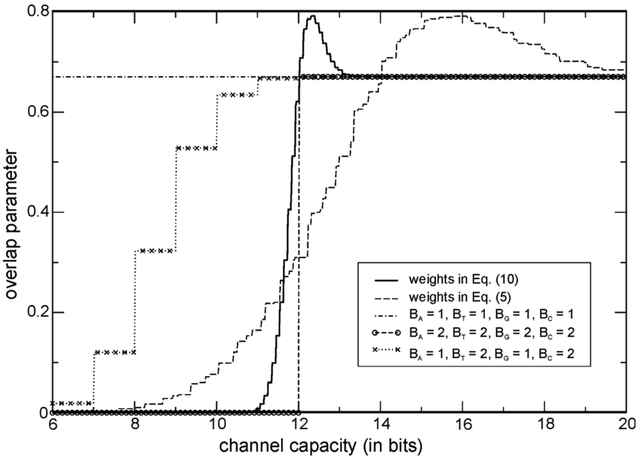


Fig. 3. There is shown the dependence of the overlap parameter q on the channel capacity for the 6-oligonucleotides in the case of two banks of sequences as in Fig. 2 and different values of the information weights

The longer the oligonucleotides the more gene specific they become. From our observations the best choice of the length of the oligonucleotides for statistical analyses is equal to $n = 6$. The result presented in Fig. 2 could seem to be trivial – the maximum of the curves for the respective values of n shifts to the right direction with the increasing value of n. However, in Fig. 3 we can observe that the maximum value of q appears only for the proper choice of the number of bits associated with the nucleotides. There is no sharp maximum in the case of the loss of information when all nucleotides have assigned the same weight $B_j = 1$. On the other hand, when all nucleotides have assigned a weight equal to two bits there is only a signal that below some value of the channel capacity there is no coincidence between the sequences of the two banks under consideration. There are, in the figure, two curves with the largest value of maximum value of q, which are corresponding to two different sets of information weights, defined in Eqs. 5 and 10. In general, one could find more representations $\{B_j\}$ ($j = A, T, G, C$) of information weight leading to the same result (the same maximum value). The trivial ones are those which correspond to other values of mutation rate u. We expect that the optimum representation should be that which requires the less amount of redundant bits.

We have analyzed all possible 6-oligonucleotides in each nucleotide position in codons of genes, and we have found the lower bound and the upper bound of the information capacity of these sequences for the particular choice of the weights B_A, B_T, B_G, B_C defined in Eq. 10. Besides, we have translated the triplets of nucleotides (codons) into the amino acids and the lower and upper bounds have been found also for the amino

Table 2. The lower and upper bound for the average number of bits representing 6-tuples of nucleotides in position (1), (2), and (3) in codons of 18-oligomers cut off from the genes of the *Borrellia burgdorferi* genome, and the 6-tuples of the corresponding amino acids

	Amino acids	Nucl. (1)	Nucl. (2)	Nucl. (3)
Lower bound	20.0778	10.7674	10.7674	10.7674
Upper bound	31.5625	13.4435	13.6273	13.0651

acids. In Table 2 we have presented the results for $n = 6$. Notice, that the range of the differences among the numbers of bits representing oligonucleotides specific for each position in codons of the examined 18-oligomers is of the order of maximum 3 bits whereas the corresponding range for the amino acids is about four times larger. We would like to emphasize that if we had measured the nucleotides with the values of B_A, B_T, B_G, B_C introduced in Eq. 5 being constructed with the help of the nucleotide occurrences f_A, f_T, f_G, f_C, then the range of the corresponding differences would be almost four times larger.

IV. DISCUSSION OF THE RESULTS

The presence of the lower and upper bound for information packing in DNA sequences and protein sequences imposes natural selection criterion for the nucleotide sequences which can be considered as the coding sequences. Therefore we generated with the help of computer random number generator as many of k-oligomers as there are in genes of natural genome and all the oligomers were fulfilling the following three conditions:

- the frequency of occurrence of nucleotides was the same as in coding sequences of natural genome, separately in position (1), (2) and (3) in codons,
- each nucleotide j is assigned a value B_j (Eqs. 10),
- the lower bound and the upper bound for the selected n-nucleotides in each position in codons of k-oligomers ($k = 3n$) could not exceed the value of the lower and upper bound for genes in natural genome as well as the triplets of nucleotides from k-oligomers cannot exceed the lower and the upper bound for amino acids after translation them into a sequence of amino acids.

We have found, that the distribution of the generated by computer k-nucleotides in all nucleotide positions in codons coincide up to the noise introduced by the over-represented and under-represented oligonucleotides with the corresponding distributions in natural genome. This could be seen in the Figs. 4-6, where we placed all generated oligomers and natural ones in a space [A, T, G, C] with the help of IFS (*Iterated Function System*) transformation [18]. In the case of $k = 6$, the points of the space [A, T, G, C] represent all possible 4096 k-tuples of nucleotides (oligomer classes) and the hills represent the numbers of the same sequences in the class. The detailed construction of the [A, T, G, C] space can be found in our paper [23]. The representation is similar to the chaos game representation of DNA sequences in the form of fractal images first developed by Jeffrey [19] and followed by others, e.g., [20]. The size of the hills is closely related to the mutation pressure and selection. The particular case of the statistical properties of short oligonucleotides have been discussed recently by Buldyrev et al. [22]. In particular, they showed

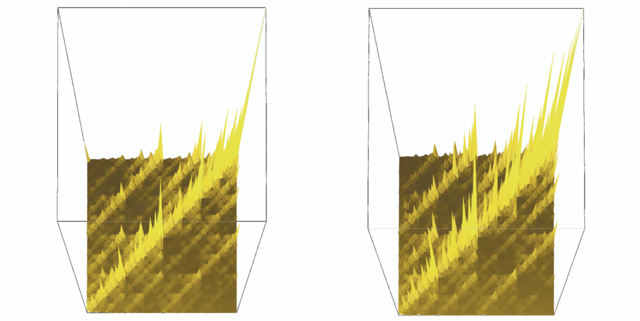


Fig. 4. Left: distribution of 6-oligonucleotides in position 3 in codons of genes from the leading strand of the *Borrellia burgdorferi* genome in [A, T, G, C] space. Right: the same for computer generated sequences

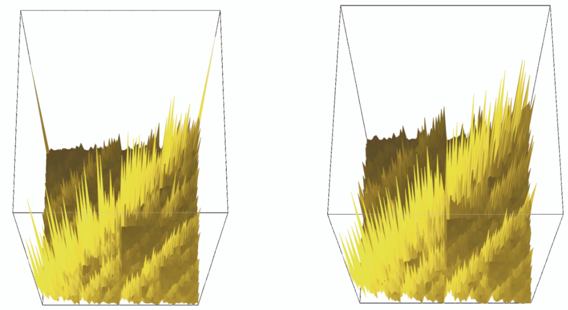


Fig. 5. The same as in Fig. 4, but for position 2 in codons

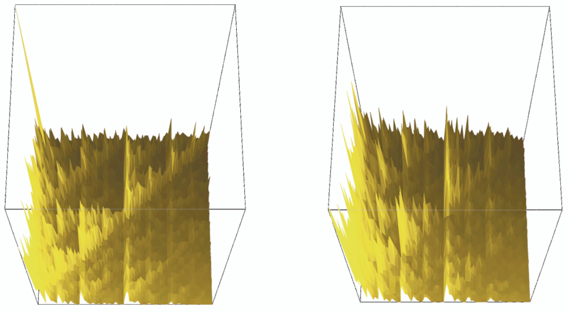


Fig. 6. The same as in Fig. 4, but for position 1 in codons

that the number of dimeric tandem repeats in coding DNA sequences is exponential, whereas in non-coding sequences it is more often described by a power law. Other analysis of the *k*-oligomers, like Zipf analysis, can be found elsewhere, e.g., [24, 25].

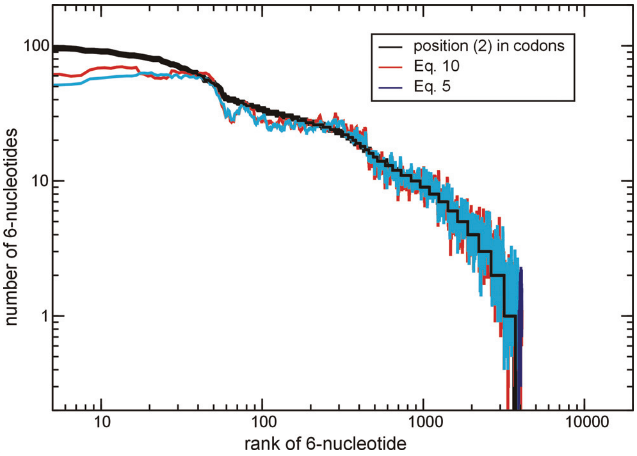


Fig. 7. Number of representants in 4096 classes of 6-sequences sorted with respect to their rank number in the case of: 6-nucleotides originating from position (2) in codons of the *Borrellia burgdorferi* genome (the hills in the left side Fig. 5), 6-nucleotides generated by computer (the hills in the right side Fig. 5) where information weights are defined in Eq. 10, 6-nucleotides generated by computer, where information weights are defined in Eq. 5. Every point in the data resulting from simulations has been aver aged over 10 elements

In the Figs. 4-6, the best result with respect to comparison of the natural sequences with the reconstructed oligomers we have got for the third position in codons, whereas the worst one we have got for the second position in codons. However, the reconstruction of the second position in codons and the first one also shares many features common with the natural genome. The choice of another set of $\{B_j\}$, the one originating from the nucleotide occurrence f_j (Eq. 4), leads to very similar results. We showed this in Fig. 7 for position (2) in codons. In the figure, all *k*-oligonucleotides have been assigned a rank with respect to their occurrence and there is plotted their number in each number 4096 classes versus the rank. The three cases are plotted in the figure: the number of *k*-nucleotides in natural genome, the number of *k*-nucleotides in generated sequences where the information weights have been defined in Eq. 10 and Eq. 5. As we can see, even in the case of the weakest reconstruction of DNA oligomers for the position (2) in codons, the number of representants of each of the 4096 classes well approximates the corresponding number in natural genome. Much better approximation we have got for the first position in codons and the third one. It seems, that suggested by us information weights B_j, in Eq. 10, basying on the mutation table for nucleotides, estimate information carried out by nucleotides better because they correspond to the smaller size of the information channel. This could be as in the example, given by Yockey [6], of bar code attached to packages items in stores that permits the cashier to record the price of the item. Namely, the amount of information estimated by us with the help of mutation table for nucleotides represents the sense code whereas the remaining part of it represents the redundant bits. Hence, the frequency of occurrence of nucleotides in natural genome represents two types of information being compromise between selection and mutation pressure. It is worth to add, that analogous information redundancy in protein sequences is very small.

The obtained by us reconstruction of DNA sequences with the help of only one rule, that the number of transferred bits cannot exceed some threshold value, suggests that the width of information channel is the basic mechanism of the information packing in DNA coding sequences.

There is different research done in the field of combinatorial DNA words design by Marathe et al. [12], in which they discuss some constrains imposed on constructed code, like Hamming constraint, free energy constraint etc. Their paper is one of the papers dealing with the study of biotechnological applications of DNA information. Our result, basying on the usage of the table of substitution rates for nucleotides and amino acids, could be also used as a new possibility of the designing the code of synthetic DNA for biotechnology purposes.

V. CONCLUSIONS

Our results suggest that the redundancies in the frequency of nucleotide occurrence in coding sequences substantially depend on mutation pressure. We have shown the method of the construction of the information weights with the reduced redundancies. It is possible to use it in some biotechnological applications dealing with designing synthetic DNA code.

References

[1] F. H. C. Crick, F. R. S. Leslie Barnett, S. Brenner and R. J. Watts-Tobin, Nature **192**, 1227-1232 (1961).

[2] B. Dujon, Trends Genet. **12**, 263-270 (1996).

[3] S. Cebrat and M. R. Dudek, Trends Genet. **12**, 12 (1996).

[4] B. Dujon and A. Goffeau, Trends Genet. **12**, Poster (1996).

[5] R. Román-Roldán, P. Bernaola-Galván and J. L. Oliver, Pattern Recogn. **29**, 1187-1194 (1996).

[6] H. P. Yockey, Computers and Chemistry **24**, 105-123 (2000)

[7] C. E. Shannon, Bell Syst. Tech. J. **27**, 379-424, 623-656 (1948).

[8] D. T. Jones, W. R. Taylor and J. M. Thornton, In CABIOS, **8(3)**, 275-282 (1992).

[9] S. Cebrat, M. R. Dudek, P. Mackiewicz, M. Kowalczuk and M. Fita, Microbial & Comparative Genomics **2(4)** 259-268 (1997).

[10] S. Cebrat and M. R. Dudek, Eur. Phys. J. **B3**, 271-276 (1998)

[11] Diana Duplij and Steven Duplij, Biophys. Bull. No **497**, 1-7 (2000), Visnyk Khark. Univ.

[12] A. Marathe, A. E. Condon and R. M. Corn, J. Comp. Biol. **8**, 201-219 (2001).

[13] M. Kowalczuk, P. Mackiewicz, D. Szczepanik, A. Nowicka, M. Dudkiewicz, M. R. Dudek and S. Cebrat, Int. J. Mod. Phys. **C12,** 1043-1053 (2001).

[14] P. Mackiewicz, M. Kowalczuk, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, A. Łaszkiewicz, M. R. Dudek and S. Cebrat, Physica **A314**, 646-654 (2002).

[15] M. Kowalczuk, P. Mackiewicz, D. Mackiewicz, A. Nowicka, M. Dudkiewicz, M. R. Dudek and S. Cebrat, BMC Evolutionary Biology **1(1)**, 13 (2001).

[16] M. Kimura, *The Neutral Theory of Molecular Evolutio*n, Cambridge University Press, Cambridge (1983).

[17] A. Nowicka, P. Mackiewicz, M. Dudkiewicz, D. Mackiewicz, M. Kowalczuk, S. Cebrat and M. R. Dudek, in Computational Conference ICCS 2003, Melbourne and St. Petersburg, June 24, 2003, P. M. A. Sloot et al. (Eds.): Lecture Notes in Computer Science 2658, 650-657 (2003), see also cond-mat/0301214.

[18] M. F. Barnsley, *Fractals Everywhere*, Springer-Verlag. New York (1988).

[19] H. J. Jeffrey, Nucleic Acids Res. **18**, 2163-2170 (1990).

[20] B.-L. Hao, H. C. Lee and S. Zhang, Chaos, Solitons, Fractals **11**, 825-836 (2000).

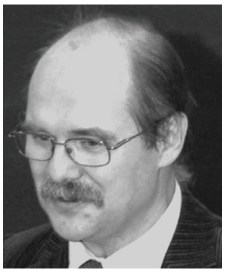
[21] A. Nowicka, M. R. Dudek, S. Cebrat, M. Kowalczuk, P. Mackiewicz, M. Dudkiewicz and D. Szczepanik CMST **6**, 65-71 (2000).

[22] S. V. Buldyrev, N. V. Dokholyan, S. Havlin, H. E. Stanley and R. H. R. Stanley, Physica **A273**, 19-32 (1999).

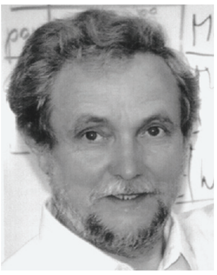
[23] M. Kowalczuk, A. Gierlik, P. Mackiewicz, S. Cebrat and M. R. Dudek, Physica **A273**, 116-131 (1999).

[24] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons and H. E. Stanley, Phys. Rev. Lett. **23** 3169-3172 (1994).

[25] N. Vandewalle, M. Ausloos, Physica A **268**, 240-249 (1999).



Mirosław R. Dudek (born in 1956), professor of University of Zielona Góra, physicist, specialized in methods of statistical physics, making research both in the field of physics (biased diffusion, phase transitions, sintering processes, percolation) and in life sciences (long range correlation in DNA, evolution of genomes, immune systems). Autor or co-author of over 60 scientific publications in prestigious journals including Surface Science, Physica A, Nuclear Acids Research, BMC Genomics, organizer of two scientific conferences. *Degrees:* Master of Science, University of Wrocław (1980), PhD (doctor of theoretical physics), University of Wrocław, Poland (1986), Habilitation thesis, University of Wrocław, Poland (1998). *Positions:* 1981-1985: PhD studies at Institute of Theoretical Physics, University of Wrocław, Poland, 1985-1986: Assistant lecturer at Institute of Theoretical Physics, University of Wrocław, Poland, 1986-2000: Permanent assistant professor position at Institute of Theoretical Physics, University of Wrocław, Poland, 1.10.2000: Professor position at Pedagogical University of Zielona Góra, Poland, 1.09.2001: Professor position at University of Zielona Góra, Poland, *Postdocs:* 1.09.1986-1.09.1987 – Post-doc position at Instituut-Lorentz voor Theoretische Natuurkunde, Leiden, The Netherlands, 1.04-9.12.1996 – Post-doc position at Chimie Theorique, Institute de la Catalyse, CNRS, Lyon, France.



Stanisław Cebrat (born in 1945), Professor at the University of Wrocław. Postdoctoral studies in French Academy of Sciences, Centre de Genetique Moleculaire in Gif-sur Yvette and at the Texas Tech University. Member of Polish Society for Genetics, Member of Network of Excellence Complex Systems EXYSTENCE, member of the COST Action P10 Physics in Risk (ESF), member of CA "General Integration of the Applications of Complexity in Science" (FP6), member of Assessment and Evaluation Committee, and DataBases Committee of GIACS. Interested in genetics, genomics, genome evolution, statistics, Monte Carlo simulations of population dynamics and applications of stochastic physics. Author and co-author of more than 100 scientific publications. Honoured with Distinguished Associate of Texas Tech University. Many times awarded by Polish Society for Genetics, Minister of Education and Rector of University of Wrocław for scientific achievements. Prize-winner of Polish Foundation for Science. Cooperates with Prof. D. Stauffer from University of Koln, Germany and P.M.C. de Oliveira from University Niteroi, Rio de Janeiro, Brazil. He spends his free times on carpentry and building homes of wood.

Maria Kowalczuk, PhD. Assistant Editor, BioMed Central, responsible for managing the peer-review process for a subset of the BMC-series journals. *Previous position:* research associate at University of Wroclaw (2002-2006), Department of Genomics. *Research interests:* genetics, genomics, bioinformatics, computer simulations of gene, genome and population evolution. *Publications:* first author and co-author of over 30 scientific publications and over 50 national and international conference communications (full list and pdf version of the doctoral thesis written in English are available at http://smorfland.microb.uni.wroc.pl/paper.html). Maria Kowalczuk studied biology and specialized in microbiology. Her postgraduate research focused on substitution matrices used in simulations of gene and genome evolution. She obtained a PhD in 2002 and until 2006 she worked on a variety of topics centering around relationships between the genetic code and the mutational and selection pressures. At present she works for a publishing company BioMed Central where she manages peer review process for manuscripts submitted to a number of various journals including BMC Bioinformatics, BMC Genomics, BMC Evolutionary Biology, BMC Structural Biology, and BMC Systems Biology.

PAWEŁ MACKIEWICZ (born in 1971), received MSc degree in biology-zoology (1996) and MSc degree in biology-microbiology (1997) at the University of Wrocław. The PhD degree obtained in 2000 in biological sciences-genetics at the University of Wrocław. Habilitated in 2004 in biological sciences-genetics. Since 2006 Assistant Professor at the Faculty of Biotechnology University of Wrocław. Interested in genomics, bioinformatics, molecular phylogeny, evolution and biostatistics. Author and co-author of more than 50 scientific publications. Prizewinner of Foundation for Polish Science for young scientists in 2001 and 2002. Many times awarded by Polish Society for Genetics, Minister of Education and Rector of University of Wrocław for scientific achievements. Member of Polish Society for Genetics.

ALEKSANDRA NOWICKA. Master's degree: 1999; master's thesis: "Influence of replication connected mutation pressure on genomes asymmetry of prokaryotic genes"; December 2003, PhD thesis: „Evaluation of PAM amino acid substitutions matrices as a tool for philogenetic associations analyses" under supervisor of prof. M. R Dudek. During PhD course I used various methods of computer analysis and I conducted computer simulations of the different aspects of evolution. My current scientific subjects include issues in fields of genomics, evolutionism, epidemiology, immunology and ecology. Nowadays I am employed at Faculty of Agriculture Warsaw Agricultural University (SGGW). I attended domestic and foreign conferences (28 conference communicates) and I am the author or co-author of 20 publications.

DOROTA MACKIEWICZ (born in 1974), received MSc degree in biology-microbiology (1999) at the University of Wrocław. The PhD degree obtained in 2003 in biological sciences-genetics at the University of Wrocław. Since 2004 Research Associate at the University of Wrocław. Interested in genomics, bioinformatics, molecular phylogeny and population dynamics and genetics. Author and co-author of more than 20 scientific publications. Many times awarded by Polish Society for Genetics, Minister of Education and Rector of University of Wrocław for scientific achievements. Member of Polish Society for Genetics.

MAŁGORZATA DUDKIEWICZ, PHD (born 20.07.1976)**.** *Education:* 2000 – MSc in Protection of Environment Department of Genetics, Institute of Microbiology, Faculty of Natural Sciences, Wroclaw University, Thesis: Computer simulation of coding sequences evolution. 2004 – PhD in Molecular Biology, Department of Genomics, Institute of Genetics and Microbiology, Faculty of Natural Sciences, Wroclaw University, Supervisor: Prof. Stanislaw Cebrat, Thesis: Modeling of mutational and selectional pressure in prokaryotic genome. *Academic and Professional Experience*: 2004-2005 – Coordinator of National Polish Bone Marrow Donor Registry at Lower Silesian Center for Cellular Transplantation; 2005-present – Assistant Professor at Warsaw Agricultural University, Department of Biometrics; 2006-present – Consultant at Central Polish Bone Marrow Donor & CB Registry POLTRANSPLANT. *Research interest:* Four years of cooperation with genomic group from Wroclaw (Prof. St. Cebrat, Prof. M.R. Dudek at al.). Researches in the field of genomes analysis, the asymmetry of nucleotide composition in bacterial genomes, MC simulation of bacterial sequences evolution, looking for mutational pressure matrices, using internet bases, PAM matrices reconstruction from simulation results. One year of cooperation with young researchers group from Department of Biometrics, Warsaw Agricultural University, fields of interest: implementation of Markov Chains in comparative genome analysis, Markov set Chains as a method for mutational pressure analysis, modeling of rhizome plant development, molecular structures modeling.