Вісник ХНУ, № 1075, серія «Фізика», вип. 18, 2013. с. 46-49

PACS: 71.45.Gm , 87.14.gk
УДК: 530.16, 53.072.11

Analysis of long-range correlations in DNA molecules: A new approach to biological classification

S.S. Melnik[1], S.V. Denisov[2], A.A. Maystrenko[3], E.Yu. Butova[4], and O.V. Usatenko[1]

[1)] *A. Ya. Usikov Institute of Radiophysics and Electronics, 12, Proskura st., Kharkov, 61085, Ukraine*
[2)] *Institute of Physics, University of Augsburg, Universitätsstr. 1, 86159 Augsburg*
[3)] *V.N. Karazin Kharkiv National University, Svobody Sq. 4, 61022, Kharkiv, Ukraine*
[4)] *Kharkiv National Medical University, 4 Lenin Avenue, Kharkiv, 61022, Ukraine*

We analyze the structure of DNA molecules of different organisms by using the additive Markov chain approach. By transforming nucleotide sequences into binary strings, we perform statistical analysis of the corresponding «texts». The scaling of the DNA texts on the scale of 10^5 base pairs is different for organisms belonging to different taxonomy domains. We calculate the memory function for the DNA of the X-chromosome of Drosophila melanogaster, and show that this function represents a useful toolkit to study statistical patterns of nucleotide sequences. Our method can be used for a computer-aided classification of living organisms.

Keywords: DNA molecule, additive Markov chain, random sequence, domain, memory function.

Анализируются молекулы ДНК различных организмов с помощью подхода аддитивных цепей Маркова. Преобразуя нуклеотидные последовательности в бинарные цепи, мы проводим статистический анализ соответствующих «текстов». Скейлинговые свойства ДНК текстов на расстояниях порядка 10^5 нуклеотидов различны для организмов, относящихся к разным доменам жизни. Мы рассчитываем также функцию памяти для ДНК Х-хромосомы дрозофилы, и показываем, что эта функция представляет собой полезный инструмент для изучения статистических закономерностей нуклеотидных последовательностей. Наш метод может быть использован для автоматизированной классификации живых организмов.

Ключевые слова: молекулы ДНК, аддитивная цепь Маркова, случайная последовательность, домен, функция памяти.

Аналізуються молекули ДНК різних організмів за допомогою підходу адитивних ланцюгів Маркова. Перетворюючи нуклеотидні послідовності в бінарні ланцюги, ми проводимо статистичний аналіз відповідних «текстів». Скейлінгові властивості ДНК текстів на відстанях близько 10^5 нуклеотидів відрізняються для організмів, що відносяться до різних доменів. Ми розраховуємо також функцію пам'яті для ДНК Х-хромосоми дрозофіли, і показуємо, що ця функція є корисним інструментом для вивчення статистичних закономірностей нуклеотидних послідовностей. Наш метод може бути використаний для автоматизованої класифікації живих організмів.

Ключові слова: молекули ДНК, адитивний ланцюг Маркова, випадкова послідовність, домен, функція пам'яті.

Introduction

The classification of living organisms – taxonomy – has a long and fascinating history. The first approaches were based on pure phenotypic criteria, like shape, number of legs, etc. Then came Charles Darwin and proposed to sort all organisms by tracing them back to their common ancestors. Finally, genetics rose to its glory and classical phenotypic criteria were replaced by molecular ones.

The genetic information is encoded in DNA and RNA molecules. The building blocks of DNA/RNA are nucleotides, two purines (adenine, A, and guanine, G) and two pyrimidines (cytosine, C, and thymine, T). The nucleotides are linked end to end, forming a long polymer chain. There are different kinds of DNA and RNA even in a single cell (chromosomal DNA and mitochondrial DNA, messenger RNA and transfer RNA etc). All these macromolecules not only bear the information about an organism itself but also encode its genealogy, and the analysis of nucleotide sequences can reveal the evolutionary relationships in a way that phenotype cannot. Moreover, the attempt to calibrate the "level of organization" of an organism may lead to completely opposite results when performed on the level of DNA/RNA structure and on the level of phenotype. By using the nucleotide sequence homology, C. R. Woese et al. proposed a new classification of organisms, by introducing three main 'domains' or 'urkingdoms': Bacteria (eubacteria), Archaea

(archaebacteria), and Eucarya (eukaryotes) [1,2]. One of the main features of new taxonomy was the separation between bacteria and arhaea, which were sharing the same kingdom before. After more than thirty years, the domain-based classification remains a subject of hot debates, and some biologists and genetics are still keep to the old, five-kingdom scheme.

The evolutionary history of the species is written in their DNA/RNA sequences. Being consisting of four nucleotides only, they may be considered as texts written in a language with a four-letter alphabet. The correlation structure of these texts can be characterized by using different correlation functions. As early as the 1960s, there were attempts to study the statistical properties of nucleotide sequences by trying to estimate the correlations between the nearest-neighbor base pairs [3]. It has became clear that DNAs and RNAs do not represent pure random sequences, where all correlations are equal to zero and the fluctuations of the density of symbols obey Gaussian law. Statistical studies of correlation structures of nucleotide sequences have been carried then during the next several decades (see [4] for the historical overview).

The ultimate record of organism evolutionary history is encoded in ribosomal RNA, a key part of the replicating system, which is well isolated and has relatively low mutation rate. All this allows us to use ribosomal RNAs for the detection of relationships between distant species. In fact, the domain-based classification has been introduced on the base of the analysis of ribosomal RNAs [2]. However, RNAs are bad objects for the statistical analysis since they are short (the typical length of RNA sequences is about several thousands pairs). Long nuclear DNA molecules are much better in this respect. During last two decades a number of genomes were completely sequenced and their entire DNA structures becomes available for the analysis. It has been shown then that most of DNA sequences exhibit long-range correlations [4, 5], fractal structures [6], *1/f* noise features [7] etc. Several authors have noticed the close connection between the correlation structure of DNA and the evolution history of the corresponding species [8].

In this paper we report the results of application of a new toolbox, the multistep Markov modeling [9, 10], to the statistical analysis of nuclear DNA sequences. Namely, we show that the proposed approach is capable to distinguish members of different domains. We also demonstrate that an alternative of the standard correlation function, the memory function [11], provides a new measure of the organizational complexity of nuclear DNAs.

DNA as a binary sequence: additive multistep Markov chain approach.

Following the ideology of [5], we transform a DNA into a binary sequence, *a(i)*, such that *a(i)=0* if the nucleotide at the position i is a member of the pair "0"=$\{\alpha_1, \alpha_2\}$, and *a(i)=1* if the nucleotide belongs to the set pair "1"=$\{\beta_1, \beta_2\}$. There are three possible partitions gates:

$$"0" = \{A, G\}, "1" = \{C, T\}, \quad (1.a)$$

$$"0" = \{A, T\}, "1" = \{C, G\}, \quad (1.b)$$

$$"0" = \{A, C\}, "1" = \{G, T\}. \quad (1.c)$$

The first partition is based on the purine pyrimidine dichotomy, the second follows the hydrogen-bond energy classification (adenine always forms three bonds with timine, while guanine forms only two bonds with cytosine). The third coding completes the set of all possible partitions.

One of the basic characteristics of a binary sequence is the correlation function,

$$K(r_1, r_2) = \overline{(a_{r_1} - \overline{a})(a_{r_2} - \overline{a})}. \quad (2)$$

In the case of statistically homogeneous and translationary invariant sequence, the corresponding correlation function depends only on a single argument,

$$K(r) = \overline{(a_n - \overline{a})(a_{n+r} - \overline{a})} = \overline{a_n a_{n+r}} - \overline{a}^2 \quad (3)$$

For each binary sequence we can built corresponding sequence k_n, where k_n is a number of symbols "1" in L-length word,

$$k_n = \sum_{l=1}^{L} a_{n+l}. \quad (4)$$

The variance is the function, which measures the deviation of a random value k from its average,

$$D(L) = \overline{(k - \overline{k})^2} = \frac{1}{M - L} \sum_{n=1}^{M-L} (k_n - \overline{k})^2. \quad (5)$$

Let consider the simplest possible case, a completely random Bernoulli-like random sequence. The probability to encounter either of the symbols, "1" or "0", is independent of the position and equal to *1/2*. It is evident that the corresponding correlation function is uniformly equal to zero, $C(r) \equiv 0$, for all $r>0$. The variance obeys the normal diffusion law and scales like $D(L) \propto L/4$.

There is a simple relation between two functions,

$$K(r) = \frac{1}{2}\left(D(r-1) - 2D(r) + D(r+1)\right), \quad (6)$$

or

$$K(r) = \frac{1}{2}\frac{d^2 D(r)}{dr^2}. \quad (7)$$

A binary sequence *a(i)* is Markovian when the probability to encounter at the position *i* a given symbol *s*, *s*="0" or "1", depends on the finite number of previous symbols only, *a(j)*, $i-N<j<i$. If $N=1$ then we deal with

a one-step Markov chain, otherwise the sequence is a *N*-step Markov chain [12]. A binary *N*-step Markov chain is fully specified by the set of conditional probabilities, $P(a_i|a_{i-N}, a_{i-N}, \ldots a_{i-1})$. One needs only two values to specify an unbiased one-step Markov chain. The number of parameters grows exponentially with the memory length N, so that in order to generate a binary *N*-step Markov chain one has to specify 2^N probabilities. For $N \sim 10^3 - 10^6$ (which is the case of nuclear DNAs), one should store an astronomically large number of parameters. This is practically impossible, so that there is a certain need for more realistic models.

Additive multi-step Markov chains are very useful in this respect [9, 10]. Their probability functions *P* are given by

$$P(a_n \mid a_{n-N}, a_{n-N+1}, \ldots, a_{n-1}) = \\ = \overline{a} + \sum_{r=1}^{N} F(r)(a_{n-r} - \overline{a}), \qquad (8)$$

where *F(r)* is the memory function. There is a certain relation between the memory function and the corresponding correlation function,

$$K(r) = \sum_{r'=1}^{N} F(r')K(r-r'), \qquad r>0, \quad (9)$$

which allow us to generate a binary sequence with a given correlation function.

We start the analysis with the nuclear DNA of Bacilus subtilis. The variance *D(L)* for different partitions, is shown in Fig.1. The correlations are maximal for the purine-primidine coding scheme, Eq. (1.a). Henceforth, we keep this scheme for the analysis of other DNA sequences.

The results of analysis of genome DNAs for three

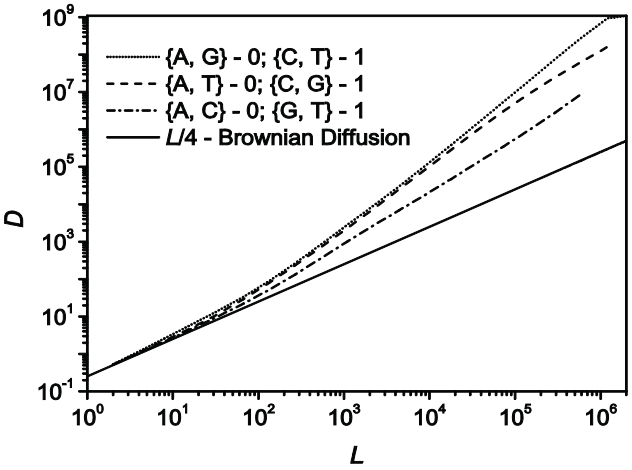

Fig. 1. The dependence of the dispersion *D(L)* vs *L* for the coarse-grained DNA text Bacillus subtilis (complete genome) for three different codings. The thick dashed line corresponds to the linear dependence *L/4*.

organisms belonging to different domains are reported in Fig.2. The memory length *N* can be identified as the inflection point of the curve *D(L)* [9]. Since this point is absent for the DNA of Bacillus subtilis, we therefore conclude that the characteristic length of correlations in this case is of the order of the length of the whole molecule.

The memory length of the Drosophila melanogaster

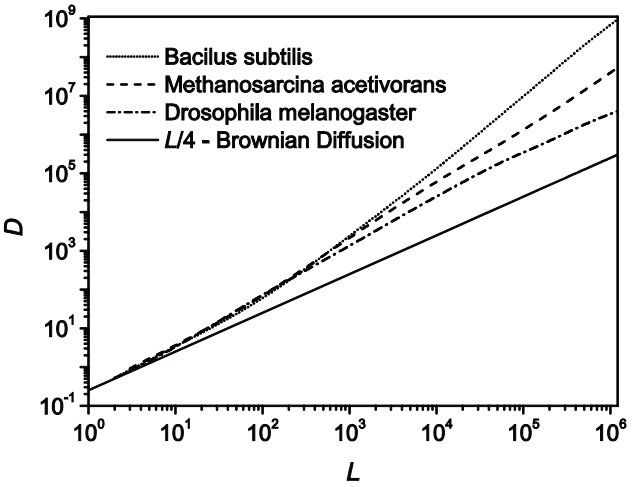

Fig.2. The dependence of the dispersion *D(L)* vs *L* for the coarse-grained DNA text of Bacillus subtilis (complete genome), Methanosarcina acetivorans (complete genome), and Drosophila melanogaster (X-chromosome), for the coding *{A,G}* = "0", *{C,T}* = "1". The thick dashed line corresponds to the linear dependence L/4.

DNA is smaller than that of Methanosarcina acetivorans. All these results are in a perfect agreement with the well-known among geneticist fact: the fraction of coding regions in DNA decreases with the degree of biological complexity, or "organization". The Eukariots are more complex than Archae, while the lasts are more complex then the Bacteria. We find that all three curves lie above the line *L/4*. It is the consequence of the persistence of correlations – after a purine with a high probability follows a purine, after a pyrimidine follows a pyrimidine. This result is in a contrast with our results for written texts, where the observed antipersistence on the scale $L \leq 10^2$ has been attributed to the grammar rules. Therefore, we can concede that such rules are either absent completely or play a minor role in DNA texts.

Although the variance and the correlation functions can tell a lot about the statistical properties of binary sequences, they have certain disadvantages. The most drastic one is that they cannot provide the recipe for the generation of a sequence with given statistical characteristics. Yet the benefit for having such a recipe would be tangible because it will give an operational description of the sequence. One of the possibilities to enter this qualitatively new level of the statistical analysis is to use the memory function. Equation (9) allows us to construct the memory

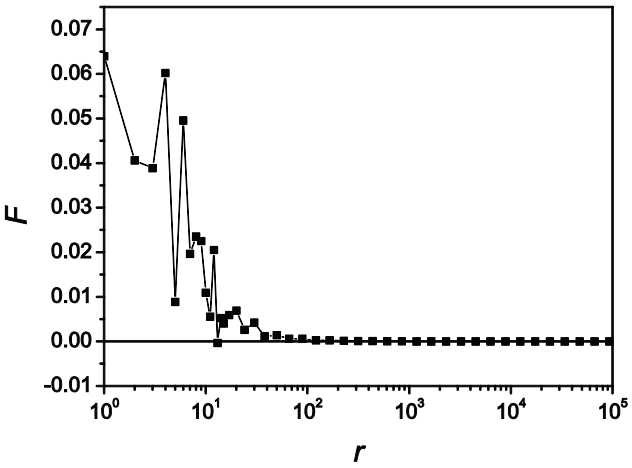

Fig. 3. Memory function *F(r)* of coarse-grained DNA text of Drosophila melanogaster (X-chromosome) for the coding *{A,G}* = "0", *{C,T}* = "1".

function of a multistep Markov chain having in hands the corresponding correlation function. We apply this procedure to the chromosomal DNA of Drosophila melanogaster. The so obtained memory functions are different from zero in the range $1 < r < 10^2$, in full accordance with the mosaic structure of DNA. Therefore, the memory function provides also a new quantitative measure of the DNA patchiness [13].

We want to underline that the calculation of the memory function is much more algorithmically complex procedure than the calculation of the variance or the correlation function. The former not only includes the information about the influence of a symbol a_{i-r} on a symbol a_i but quantifies the influence of all symbols, a_{i-N}, …, a_{i-1} , on a symbol a_i.

Conclusion.

The results of the multistep Markov modeling of nuclear DNAs reported in this paper are of interest for two primary reasons. First, they provide the compelling evidence for the domain-based classification of living organisms, which has already been developed by using ribosomal RNAs. Second, they demonstrate, how the multistep Markov method can be used to construct a memory function of nucleotide sequences, a "zipped" version of a given DNA molecule.

It might happen, however, that neither the correlation function nor the memory function is a good characteristic of the random binary sequence. For example, an integral correlation function (also called "differential variance"), which is intermediate between (3) and (5), is able to determine the memory length N more accurately than the location of the inflection point of the variance curve. One of intriguing perspectives is to calculate the memory functions of different organisms. The multistep Markov analysis of written texts has already shown that their memory functions have a peculiar structure. Namely, they seemingly drop to zero outside certain finite windows but, when being zoomed-in, reveal long power-law tails. Further studies in this direction will show whether this is the case for DNA texts.

1. C. L. Woese and G. E. Fox, Proc. Natl. Acad. Sci. USA **74**, 5088 (1977).
2. C. R. Woese, N. Goldenfeld, How the Microbial World Saved Evolution from the Scylla of Molecular Biology and the Charybdis of the Modern Synthesis, Microbiol. Mol. Biol. Rev. 2009; **73**(1):14-21.
3. J. Josse, A. D. Kaiser, and A. Kornberg, J. of Biol. Chem. **236**, 864 (1961); L. L. Gatlin, J. of Theor. Biology **10**, 281 (1966).
4. W. Li, Computers Chem. **21**, 257 (1997).
5. C. K. Peng *et al.*, Nature **356**, 168 (1992).
6. R. F. Voss, Phys. Rev. Lett. **70**, 1343 (1993).
7. W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
8. S. V. Buldyrev *et al.*, Biophys. J. **65**, 2673 (1993); R. Roman-Roldan, P. Bernaola-Galvan, and J. L. Oliver, Phys. Rev. Lett. **80**, 1344 (1998).
9. O. V. Usatenko and V. A. Yampol'skii, Phys. Rev. Lett. **90**, 110601 (2003).
10. O.V. Usatenko, S.S. Apostolov, Z.A. Mayzelis, S.S. Melnik, Random Finite-Valued Dynamical Systems: Additive Markov Chain Approach. Cambridge scientific publishers, Kharkov Series in Physics and Mathematics,Volume 1 (2009).
11. S. S. Melnyk, O.V. Usatenko, and V. A. Yampol'skii, Physica A **361**, 405 (2006); S.S. Melnyk *et al.*, J. Phys. A: Math. Gen. **39**, 14289 (2006).
12. I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).
13. G.M. Viswanthan *et al.*, Biophys. J. **72**, 866 (1997).