DNA–sequence analysis using Markov chain models

Boris Ryabko
Siberian University of Telecommunication an Informatics,
Institute of Computational Technologies, Siberian Branch of RAS
Novosibirsk
E-mail: boris@ryabko.net

Natalie Usotskaya
Novosibirsk State University
Novosibirsk
E-mail: usotskaya@gmail.com

Abstract— **The statistical structure of DNA–sequences is of a great interest to molecular biology, genetics and the theory of evolution (see Chen and others,** *GIW-99,* **1999, Aktulga and others,** *EURASIP J. of Bioinformatics and Systems Biology,* **2007, Li,** *Computers and Chemistry* **, 1997). One of the approaches is a sequence modeling using Markov processes of different orders, and further statistical estimation of their parameters (see Simons and others,** *JSPI* **, 2005). In this paper we use firstly the test for the serial independence from Ryabko, Astola (***Stat. Methodology,* **2006) to estimate the "memory" (or connectivity) of genetic texts and secondly we apply the homogeneity test for solving the DNA–based problem connected to the phylogenetic system of various organisms.**

I. Introduction

The problem of investigation of the DNA–sequence structure became interesting when a large amount of data was accumulated using new methods of DNA sequencing. Nowadays, several areas of research, such as molecular biology, genetics, theory of evolution, pharmacology, etc., are interested in the diverse investigation of the DNA structure. There are several approaches to analyze DNA–sequences. One of the most widespread is to describe them using Markov processes of different orders ([1], [2]). This approach is evolved in this paper, using the test suggested in [3], which gives the opportunity to estimate the "memory" of DNA–sequences. It is possible to determine the depth of interconnection between symbols within one sequence of letters, using this method.

In molecular biology it is often necessary to compare different parts of genetic texts, for example, while constructing phylogenetic trees for various organisms ([4], [5]). In this paper we use the test for homogeneity, suggested in [3] and estimate the measure of "relatedness" between DNA–sequences. At first we investigate experimentally the efficiency of suggested tests. And then the algorithms are applied to analyze the genetic texts of different biological organisms.

The obtained results coincide with many quantitative and qualitative characteristics known from literature. It demonstrates the efficiency of the method. Furthermore, we obtain several new results, interesting for bioinformatics.

II. Experimental efficiency of theoretic-information tests

Several statistical tests for hypothesis testing were suggested in [3], such as goodness-of-fit, serial independence, homogeneity testing and the others. The input data for these tests are considered to be generated by a stationary and ergodic Markovian source over a finite alphabet (see [6]). Non-parametric tests for such problems were unknown before, and obtained results are asymptotic, that is why we have to estimate experimentally the efficiency of suggested algorithms. In order to do this we carried out experiments over the generated data.

We will consider two tests — the test for serial independence and the test for homogeneity. Testing of serial independence helps us to determine the "memory" (or the order) of the Markovian source by any generated sequence of letters. Homogeneity testing allows to check whether two sequences are generated by the same source or by two different sources. (To find the formal description of the considered algorithms see [3].)

Before using these algorithms one has to choose some universal code (or the method of data compression). As the samples of such universal codes it is possible to consider widely known data compression methods (or so-called archievers), for example *WinRAR*. But it turned out that the best results were received while using *Gencompress* — a special archiver for genetic data compression, which is among the best archivers for genetic data until nowadays (see [7]). While using *Gencompress* the correct hypothesis was accepted when the input data was near 8 times less than the corresponding data while using the common–purpose archivers. That is why all the results of testing for the generated sequences and for the real genetic texts are presented only in case of usage *Gencompress* as the universal code.

Let us formulate the definitions of the empirical Shannon entropy of the m-th order and of the universal code, which will be used later (see [3]).

Given sample X is presented by r sequences $x^1 = x_1^1 \dots x_{t_1}^1, \dots, x^r = x_1^r \dots x_{t_r}^r$ and $t = \sum_{i=1}^r t_i$. The empirical m–order Shannon entropy ($0 \le m \le t$) for given $x^1, \dots, x^r$ is defined as following:

$$h_m^*(X) = -\sum_{v \in A^m} \frac{\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v)}{(t - mr)} \sum_{a \in A} \frac{\nu_{x^1 \diamond \dots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v)} \log \frac{\nu_{x^1 \diamond \dots \diamond x^r}(va)}{\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v)},$$

where $\bar{\nu}_{x^1 \diamond \dots \diamond x^r}(v) = \sum_{a \in A} \nu_{x^1 \diamond \dots \diamond x^r}(va)$, $\nu_{x^1 \diamond \dots \diamond x^r}(v) = \sum_{i=1}^r \nu_{x^i}(v)$, and $\nu_{x^i}(v)$ denotes the number of occurrences of the word v in the word x^i.

A code φ is called universal if for any stationary and ergodic source τ

$$\lim_{t \to \infty} t^{-1}(-\log \tau(x_1 \dots x_t) - |\varphi(x_1 \dots x_t)|) = 0$$

978-1-4244-2271-5/08/$25.00 ©2008 IEEE

with probability 1. So, informally speaking, universal codes estimate the probability characteristics of the source τ and use them for efficient "compression".

A. Test of serial independence

Let us start from the description of the test for serial independence. Let the sample X is presented by r sequences $x^1 = x_1^1 \dots x_{t_1}^1, \dots, x^r = x_1^r \dots x_{t_r}^r$, which are generated by some unknown source, and let $t = \sum_{i=1}^{r} t_i$. The symbol $\varphi(X)$ denotes some uniquely decodable code (or the lossless method of data compression). In addition $h_m^*(X)$ denotes the empirical Shannon entropy of the m-th order.

The main hypothesis H_0^{SI} is that the source, which generated all the sequences from the sample, is Markovian, whose order is not greater than m, ($m \geq 0$), and the alternative hypothesis H_1^{SI} is that the sample X is generated by the source whose order is greater than m.

The test is as follows (see [3]): *Let φ be any code. By definition, the hypothesis H_0^{SI} is accepted if*

$$(t - mr)h_m^*(X) - |\varphi(X)| \leq \log(1/\alpha),$$

where $\alpha \in (0, 1)$. Otherwise, H_0^{SI} is rejected.

It was proved that for any code φ the Type I error is less than or equal to α, and for the universal code φ the Type II error goes to 0, when t tends to infinity.

To estimate the efficiency of the given test we consider two families of stochastic Markovian processes of the first and second order over the 2-letter and 4-letter alphabets, respectively (the case of 4-letter alphabet corresponds the case of the genetic texts). The probability distribution of generated sequences is presented in Table I, where A is an alphabet, $0 \leq \delta \leq {}^1/_{|A|}$ and m is the order of the source. Furthermore, for limit probabilities the following claim is true: for the binary alphabet $P(0)$=$P(1)$=${}^1/_2$, and for the 4-letter alphabet $P(0)$=$P(1)$=$P(2)$=$P(3)$=${}^1/_4$. Let us use the test of serial independence for analyzing the processes from Table I in order to estimate the power of the test over the generated sequences. In order to estimate experimentally the size of the input data (which is necessary to find given divergences through analyzable sequences) we vary the value of parameter δ. (It is obvious that the less is the value of δ — the closer the examined Markovian source is to the Bernoulli distribution with equal probabilities of symbols, so it is hard to determine the correct order of the source, which is greater than 0.)

Here and below the required level of significance is equal to 0.01. The results are presented in Table II. While testing in each case we generate 50 sequences according to the distributions from Table I, and the lengths vary as 2^n, $8 \leq n \leq 28$. Furthermore, the value of δ varies: for the 2-letter alphabet δ takes on the value 0.3, 0.2, 0.1, 0.05, 0.025, and for the 4-letter alphabet — 0.2, 0.1, 0.05, 0.025, 0.01. The cells of the table include the length of generated sequences for which the test determines the correct order of Markovian source for all 50 samples for the first time. For example, in case of the binary alphabet (δ=0.2) and the real order of the source $m = 1$ the test correctly determines the order for all 50 samples, when the length of the generated sequences is equal to 2^{11}. "No" denotes the case, when for all 50 sample the correct hypothesis is not accepted.

TABLE I

THE DISTRIBUTIONS WHICH ARE USED TO GENERATE SIMULATION SEQUENCES

	$A = \{0, 1\}$	$A = \{0, 1, 2, 3\}$
m=1	P(0|0)=${}^1/_2$+δ P(0|1)=${}^1/_2$-δ	P(0|0)=${}^1/_4$+δ P(0|1)=${}^1/_4$-δ P(1|0)=${}^1/_4$+δ P(0|1)=${}^1/_4$-δ P(2|0)=${}^1/_4$-δ P(0|1)=${}^1/_4$+δ P(0|2)=${}^1/_4$+δ P(0|3)=${}^1/_4$-δ P(1|2)=${}^1/_4$+δ P(0|3)=${}^1/_4$-δ P(2|2)=${}^1/_4$-δ P(0|3)=${}^1/_4$+δ
m=2	P(0|00)=${}^1/_2$+δ P(0|11)=${}^1/_2$+δ P(0|01)=${}^1/_2$-δ P(0|10)=${}^1/_2$-δ	P(0|00)=P(0|22)=P(0|13)=P(0|31)=${}^1/_4$+δ P(1|00)=P(1|22)=P(1|13)=P(1|31)=${}^1/_4$+δ P(2|00)=P(2|22)=P(2|13)=P(2|31)=${}^1/_4$-δ P(0|01)=P(0|10)=P(0|23)=P(0|32)=${}^1/_4$-δ P(1|01)=P(1|10)=P(1|23)=P(1|32)=${}^1/_4$-δ P(2|01)=P(2|10)=P(2|23)=P(2|32)=${}^1/_4$+δ P(0|11)=P(0|02)=P(0|20)=P(0|33)=${}^1/_4$+δ P(1|11)=P(1|02)=P(1|20)=P(1|33)=${}^1/_4$+δ P(2|11)=P(2|02)=P(2|20)=P(2|33)=${}^1/_4$-δ P(0|03)=P(0|30)=P(0|12)=P(0|21)=${}^1/_4$-δ P(1|03)=P(1|30)=P(1|12)=P(1|21)=${}^1/_4$-δ P(2|03)=P(2|30)=P(2|12)=P(2|21)=${}^1/_4$+δ

TABLE II

EFFICIENCY OF THE TEST FOR SERIAL INDEPENDENCE

$\mid A \mid$=2			$\mid A \mid$= 4		
	m = 1	m = 2		m = 1	m = 2
δ	identification length		δ	identification length	
0.3	2^{10}	2^{10}	0.2	2^{10}	2^{10}
0.2	2^{11}	2^{11}	0.1	2^{13}	2^{13}
0.1	2^{13}	2^{13}	0.05	2^{15}	2^{15}
0.05	2^{18}	2^{18}	0.025	2^{18}	2^{18}
0.025	*no*	*no*	0.01	*no*	*no*

Thus we see the experimental efficiency of the test for serial independence, because the correct determination of the order takes place for the sequences of moderate lengths.

B. Homogeneity testing

Let us investigate the experimental efficiency of the test for homogeneity to determine the measure of relatedness between different sequences. First of all we formulate the algorithm (see [3]). As above $\varphi(X)$ denotes some uniquely decodable code, X is a being analyzed sample from r sequences $x^1 = x_1^1 \dots x_{t_1}^1, \dots, x^r = x_1^r \dots x_{t_r}^r$. Besides, it is preliminarily known that all these sequences are generated by Markovian sources whose orders are not greater then m, ($m \geq 0$). Let $t = \sum_{i=1}^{r} t_i$, and $h_m^*(X)$ is an empirical Shannon entropy of the m-th order.

The main hypothesis H_0^{hom} is that all sequences are generated by the same source, and the alternative hypothesis H_1^{hom}

is that there exist two sequences $x^i \neq x^j$ from the sample X that are generated by two different sources.

The test is as follows (see [3]): *Let φ be any code. By definition, the hypothesis H_0^{hom} is accepted if*

$$(t - mr)h_m^*(X) - \sum_{i=1}^{r} |\varphi(x^i)| \leq \log(1/\alpha),$$

where $\alpha \in (0, 1)$. Otherwise, H_0^{hom} is rejected.

It was proved that for any code φ the Type I error is less than or equal to α, and for a universal code φ the Type II error goes to 0, when t tends to infinity, so that the constant $c > 0$ exists and $c < t_j/t$ for all j's.

In order to estimate the power of suggested algorithm over generated sequences we consider the pairs of sequences: one is generated by the source from Table I and another is generated by the Bernoulli source with equal probabilities of symbols. The value of δ is decreasing, because it is obvious that while δ is decreasing, the sequence generated by the Markovian source is closer to the sequence generated by the Bernoulli source with equal probabilities of symbols. So, it is hard to distinguish them as generated by two different sources.

TABLE III

EFFICIENCY OF THE TEST FOR HOMOGENEITY

$\mid A \mid$=2			$\mid A \mid$= 4		
	m = 1	m = 2		m = 1	m = 2
δ	identification length		δ	identification length	
0.3	2^{11}	2^{11}	0.2	2^{11}	2^{11}
0.2	2^{13}	2^{13}	0.1	2^{13}	2^{14}
0.1	2^{15}	2^{15}	0.05	2^{16}	2^{16}
0.05	*no*	*no*	0.025	*no*	*no*

The results are presented in Table III. We generate 50 sequences of lengths 2^n, $8 \leq n \leq 28$, according to the distributions from Table I, moreover δ varies as 0.3, 0.2, 0.1, 0.05 for the 2-letter alphabet, and as 0.2, 0.1, 0.05, 0.025 for the 4-letter alphabet. Besides, we generate one sequence of length 2^n for every n by the Bernoulli source with equal probabilities of symbols. So we analyze 50 pairs of sequences in each case. The cells of the table contain the length of generated sequences for which the test determines correctly for the first time that all pairs of samples are generated by two different sources. The value of the source order is decided to be known preliminary. For example, for the 4-letter alphabet (δ=0.05) the test determines all 50 pairs of sequences as generated by two different sources when the length of sequences is equal to 2^{15}. "No" also denotes the case, when for all 50 sample the correct hypothesis is not accepted.

Thus we see the experimental efficiency of the test for homogeneity, because it can effectively distinguish two rather "close" to each other sequences.

Summarizing the results of testing presented in Tables II — III, we see that if the divergence of the sequence from one generated by the source with equal probabilities of symbols is more then 0.025 over 4-letter alphabet, than tests detect it. And they correctly determine the order of the source and also distinguish sequences, generated by two different sources. Furthermore, the required amount of input data for these analysis is moderate.

III. APPLICATIONS FOR THE GENETIC TEXT ANALYSIS

This section includes the results of the experiments over DNA–sequences of several biological organisms. We use theoretic–informational tests, suggested in [3], as the tool for analysis. It is well known that the DNA–sequence of any biological organism contains the genetic information about it. The DNA molecule is a long double helix consisting of two strands. Each helix is a chain of bases, the chemical units of four types: A, C, G, T. So we can consider the DNA–sequence as generated by some source over the 4-letter alphabet $\{A, C, G, T\}$ (for example, see [4]) and can use algorithms from [3] for the hypothesis testing to investigate the statistical structure of various genetic texts.

A. Experimental investigation of the genetic text "memory"

Some works are mentioned in [8], where several suggestions about the depth of interconnection between symbols within one DNA–sequence are mentioned. These suggestions vary from 3–6 bases up to 1–10000 bases. As a result the attempts to model genetic texts by Markovian processes applied only to the sources of low orders — zero, first and second (see, for example, [1], [2]). Besides, the problems of dependencies in DNA sequences were considered in [5] using the mutual information estimation. So the question about the depth of interconnection between symbols within the DNA–sequence was not finally solved. In order to estimate the "memory" of genetic texts we carried out several experiments using theoretic–informational tests for the hypothesis testing which were considered in the previous section.

We received several earlier unknown results while carrying out the analysis of genetic texts. In particular we found that the value of "memory" varies greatly even among organisms biologically close to each other. And the obtained results show the dispersion of the value from 2 up to 9 for considered genetic texts.

To investigate the DNA–sequence "memory" of various species we considered several procaryotes and eukaryotes. The genomes of 38 archaebacteria and 43 bacteria were considered among procaryotes (all the chromosomes were considered if there were any). All DNA–sequences were taken from the database [11]. We will present the results of experiments only for the subset of considered organisms in order to make the text more comfortable for reading and not too complex with details. So we consider the following *archaebacteria*: Aeropyrum pernix K1 (u_1), Archaeoglobus fulgidus (u_2), Picrophilus torridus DSM 9790 (u_3), Pyrobaculum aerophilum str. IM2 (u_4), Pyrobaculum arsenaticum DSM 13514 (u_5), Pyrobaculum calidifontis JCM 11548 (u_6), Pyrobaculum islandicum DSM 4184 (u_7), Pyrococcus abyssi (u_8), Pyrococcus furiosus DSM 3638 (u_9), Pyrococcus horikoshii OT3(u_{10}), Sulfolobus acidocaldarius DSM 639 (u_{11}), Sulfolobus solfataricus P2

TABLE IV

TESTING OF SERIAL INDEPENDENCE FOR THE GENETIC TEXTS

Archae-bacteria	Len, Mb	Mem	Bacte-ria	Len, Mb	Mem	Euca-ryote	Len, Mb	Mem
u_1	1.6	3	u_{16}	5.6	4	u_{31}^1	0.2	6
u_2	2.1	3	u_{17}	2.4	3	u_{31}^2	0.19	3
u_3	1.5	3	u_{18}	1.1	8	u_{31}^3	0.19	3
u_4	2.2	3	u_{19}	1.4	8	u_{31}^4	0.2	4
u_5	2.1	3	u_{20}	1.4	8	u_{31}^5	0.2	3
u_6	2.0	3	u_{21}	1.9	8	u_{31}^6	0.22	4
u_7	1.8	5	u_{22}	1.5	8	u_{31}^7	0.22	3
u_8	1.7	3	u_{23}	0.6	2	u_{31}^8	0.23	4
u_9	1.9	6	u_{24}	2.0	6	u_{31}^9	0.25	6
u_{10}	1.7	3	u_{25}	2.2	6	u_{31}^{10}	0.26	3
u_{11}	2.2	3	u_{26}	5.3	3	u_{31}^{11}	0.26	5
u_{12}	2.9	9	u_{27}	4.7	4			
u_{13}	2.6	7	u_{28}	4.0	8	u_{32}^1	5.5	8
u_{14}	1.5	3	u_{29}	1.6	6	u_{32}^2	4.4	7
u_{15}	1.5	6	u_{30}	1.6	4	u_{32}^2	2.4	8

(u_{12}), Sulfolobus tokodaii str. 7 (u_{13}), Thermoplasma acidophilum DSM 1728 (u_{14}), Thermoplasma volcanium GSS1 (u_{15}) and *bacteria*: Acidobacteria bacterium Ellin345 (u_{16}), Acidothermus cellulolyticus 11B (u_{17}), Anaplasma marginale St Maries (u_{18}), Anaplasma phagocytophilum HZ (u_{19}), Bartonella bacilliformis KC583 (u_{20}), Bartonella henselae Houston-1 (u_{21}), Bartonella quintana Toulouse (u_{22}), Baumannia cicadellinicola (u_{23}), Bifidobacterium adolescentis ATCC 15703 (u_{24}), Bifidobacterium longum (u_{25}), Bordetella bronchiseptica (u_{26}), Bordetella parapertussis (u_{27}), Bordetella pertussis (u_{28}), Helicobacter pylori 26695 (u_{29}), Helicobacter pylori J99 (u_{30}). As a sample of eucaryotes such popular objects of biological research as cryptomonad alga *Guillardia theta nucleomorph*, budding yeast *Saccharomyces cerevisiae S288C*, microsporidian parasite *Encephalitozoon cuniculi* (u_{31}) and fission yeast *Schizosaccharomyces pombe* (u_{32}) were taken, for each of them the whole amount of chromosomes was considered — 3, 16, 11 and 3 respectively. We will present only last two organisms to the simplicity of the presentation. The denotation is u_i^j for the j-th chromosome of the i-th organism (i = 31, 32).

So to determine the "memory" of various genetic texts we consider the main hypothesis H_0^{SI} that the "memory" of the being analyzed DNA–sequence is equal to m, m varies from 0 till the moment of accepting the main hypothesis. We consider only whole genetic texts during our experiments.

The results are presented in Table IV. Columns "Len" indicate the length of the DNA–sequence, and "Mem" — the obtained value of "memory". Considering the obtained data we noted that bacteria and eucaryotes had the relatively large values of "memory" though the length of each chromosome or the whole genome was small enough. On the basis of this results it is possible to assume the existence of large interconnections between symbols within DNA–sequences for these species. The possible reason could be the appearance of introns (they are the noncoding sites of the gene which do not contain the information about the amino acids of protein) and the increasing amount of duplicating genes.

The preliminary analysis lets us to suggest that the length and the amount of genes of the DNA–sequence is statistically associated with the "memory" of the genetic texts. To test this hypothesis the coefficients of the correlation between the pairs of the data samples were calculated: between the "memory" and the length, between the "memory" and the number of genes. The results are presented in Table V. Thus we see that the "memory" characteristic is of standalone biological interest, because the correlation with standard parameters of DNA–sequences exists but its module is not too close to 0 or 1. So the "memory" of the DNA–sequence may give new information about the organization of the DNA structure.

TABLE V

THE COEFFICIENTS OF CORRELATION BETWEEN "MEMORY" AND STANDARD PARAMETERS OF THE GENETIC TEXTS

Type	Memory/length	Memory/number of genes
Archaebacteria	0.63	0.53
Bacteria	0.37	0.355
Eukaryotes	0.457	0.384

Let us mention such unexpected fact that the "memory" of DNA–sequences even for the biologically related organisms (belonging to same genus) can vary greatly. As the example we cite an instance the archaebacteria from the genus *Sulfolobus* and bacteria from *Bordetella*. Archeobacteria have the comparable length of genomes: from 2.1MB to 2.8MB though the determined "memory" differs considerably — 3, 9 and 7 respectively. In regard to bacteria, the size of genomes in this case varies from 4MB to 5.3MB, but the determined "memory" have the values 3, 4 and 8 respectively, moreover the largest memory (8) is for the smallest genome — *Bordetella pertussis*. Thus these samples show that the depth of interconnection between symbols in the DNA–sequence can vary even for the biologically close genera.

Thus we can make a conclusion that this method of "memory" determination of the genetic text can be useful for choosing the appropriate model when modeling the DNA–sequence. According to the results of the test for serial independence the "memory" of genetic texts is usually more than 2, whereas Markovian models of the low order were used earlier for modeling the DNA–sequences. Therefore we have to use the models of the higher order while analyzing dependencies within DNA–sequences.

B. Homogeneity testing for genetic texts

In the molecular biology and genetics the problem of comparison the genomes and their parts is often risen. The solution of this problem allows us to find the same or related genes, to build phylogenetic trees, etc. ([9], [4]). Let us consider the problem of estimation the measure of relatedness between different organisms trying to understand — whether two DNA–sequences are "generated" by one source or by to different sources. To construct the phylogenetic tree usually the matrix of distances between the DNA–sequences of various

organisms is built ([7], [10]). In this section the attempt to estimate the measure of relatedness between various organisms was undertaken using the test for homogeneity from [3].

The binary logarithm of the shortest length of initial fragmentation (on which we distinct two sequences as generated by different sources) was the indicator of closeness for two DNA–sequences (tables VI–VII). The initial fragmentation of the DNA–sequence increased as the power of 2. That is we consider the initial fragmentation of the length 2^n, then the length increases to 2^{n+1} and so on. When we find the length of sequences 2^m such that the hypothesis of homogeneity is rejected then given m is the measure of relativeness for analyzable sequences.

If the distinguishing of the DNA–sequences happens only when the whole genome are considered then the symbol (‡) is used to indicate this case in the tables. "No" denotes the case when even the whole genome consideration does not give us an opportunity to distinguish two sequences. If the sequences vary greatly then the "measure" of closeness is small. But if the sequences are very close to each other than the distinguishing of sequences will not take place even when one considers the whole genome. Thus the larger is the value corresponding to the pair of sequences — the closer these sequences are to each other.

TABLE VI

HOMOGENEITY TESTING FOR ARCHAEBACTERIA

	u_2	u_{33}	u_{34}	u_8	u_9	u_{10}	u_{15}
u_2	—	16	19	18	17	17	17
u_{33}	16	—	no	15	15	15	16
u_{34}	19	no	–	19	19	19	19
u_8	18	15	19	—	no	no	17
u_9	17	15	19	no	—	no	17
u_{10}	17	15	19	no	no	—	16
u_{15}	17	16	19	17	17	16	—

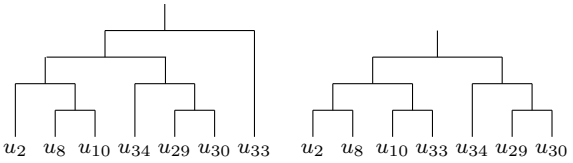
In Table VI there are the results of the test for 7 archaebacteria: u_2, Methanococcus maripaludis C5 (u_{33}), Methanococcus maripaludis S2 (u_{34}), u_8, u_9, u_{10} and u_{15}. These samples were chosen to form two groups of biologically close organisms (pair u_{33}, u_{34} is from the genus *Methanococcus*, triplet u_8, u_9, u_{10} — from the genus *Pyrococcus*) in order to compare them to each other and to the organisms from another genera — u_2 and u_{15}. According to Table VI, u_{33} and u_{34} were not determined by test as generated by different sources even when one considered the whole genomes, just like the triplet u_8, u_9 and u_{10}. It is predictable because these combinations of the organisms are taxonomically related whereas the other pairs are differed on the smaller initial fragmentation. This result is especially interesting if one remembers that the length of sequence for archaebacteria is relatively small and less than for the considered below simulation sequences.

In Table VII there are results for 7 procaryotes: u_2, u_8, u_{10}, *Escherichia coli K-12 MG1655* (u_{33}), *Haemophilus influenzae* (u_{34}), u_{29}, u_{30}. This set of organisms is chosen because it was analyzed in [7], where the phylogenetic tree 1b was built. The phylogenetic tree 1a was obtained according to the results of Table VII. It is easy to mention that these trees are the same except the position of u_{33}. Perhaps the reason is the length of the original DNA–sequence — it is 2.5 times larger than for other samples.

TABLE VII

HOMOGENEITY TESTING FOR PROCARYOTES FROM [7]

	u_2	u_8	u_{10}	u_{33}	u_{34}	u_{29}	u_{30}
u_2	—	18	17	17	17	17	17
u_8	18	—	no	14	15	14	15
u_{10}	17	no	–	14	15	15	15
u_{33}	17	14	14	—	15	14	14
u_{34}	17	15	15	15	—	20	‡
u_{29}	17	14	15	14	20	—	no
u_{30}	17	15	15	14	‡	no	—



a) b)

Phylogenetic trees, a) — according to the data from Table VII, b) — from [7].

Therefore, the test for homogeneity can be used to estimate the measure of the relatedness between genomes of vfrious organisms or between chromosomes of the same organism.

ACKNOWLEDGMENT

Research was supported by Russian Foundation for Basic Research (grant no. 06-07-89025).

REFERENCES

[1] M. Farach, M. Noordewier, S. Savari, L. Shepp, A. Wyner, A. Ziv, "On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence", *Proc. 6 Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 48–57, 1994.

[2] G. Simons, Y.-Ch. Yao, G. Morton, "Global Markov models for eukaryote nucleotide data", *JSPI*, vol. 130, pp. 251–275, 2005.

[3] B. Ryabko, J. Astola, "Universal codes as a basis for time series testing", *Stat. Meth.*, vol. 3, pp. 375–397, 2006.

[4] R.M. Karp, "Mathematical Challenges from Genomics and Molecular Biology", *Notices of the AMS*, vol. 49, N 5, pp. 544–553, 2002.

[5] J. Hagenauer, Z. Dawy, B. Goebel, P. Hanus, J.C. Mueller, "Genomic analysis using methods from information theory", *IEEE Information Theory Workshop (ITW 2004)*, pp. 55–59, 2004.

[6] M.J. Weinberger, A. Lempe, J. Ziv, "A sequential algorithm for the universal coding of finite memory sources", *Information Theory, IEEE Transactions*, vol. 38, N 3, pp. 1002–1014, 1992.

[7] X. Chen, S. Kwong, M. Li, "A Compression Algorithm for DNA Sequences and Its Applications in Genome Comparison", *Proc. 10 Workshop on Genome Informatics (GIW-99)*, pp. 51–61, 1999.

[8] W. Li, "The Study of Correlation Structures of DNA–sequences" // *Computers and Chemistry*, vol. 21, N 4, pp. 257–271, 1997.

[9] H.M. Aktulga, I. Kontoyiannis, K.A. Lyznik, L. Szpankowski, A.Y. Grama, W. Szpankowski, "Identifying statistical dependence in genomic sequences via mutual information estimates", *EURASIP Journal on Bioinformatics and Systems Biology*, accepted 25.09.2007.

[10] I. Oprea, S. Pasca, V. Gavrila, "Method of DNA Analysis Using the Estimation of the Algorithmic Complexity", *Leonardo Electronic Journal of Practices and Technologies*, vol. 3, N 5, pp. 53–66, 2004.

[11] National Center for Biotechnology Information: **www.ncbi.nlm.nih.gov**.