Applications of information–theoretic tests for analysis of DNA sequences based on Markov chain models

N. Usotskaya
Novosibirsk State University
Novosibirsk
usotskaya@gmail.com

B. Ryabko[1]
Siberian State University of Telecommunication
and Informatics, Institute of Computational Technologies,
Siberian branch of the RAS, Novosibirsk
boris@ryabko.net

Abstract

The statistical structure of DNA sequences is of great interest to molecular biology, genetics and the theory of evolution. One of the popular approaches is sequence modeling using Markov processes of different orders, and further statistical estimation of their parameters. To continue the investigations according this approach tests for hypothesis testing are used to estimate the "memory" (or connectivity) of genetic texts and to solve the DNA–based problem connected to the phylogenetic system of various organisms.

Keywords: universal coding, data compression, hypothesis testing, Shannon entropy, Markov processes, DNA, genetic texts, genes, procaryote, eukaryotes, phylogenetic trees.

1 Introduction

The DNA sequence structure investigation became of an interest after large amount of data was accumulated using new methods of DNA sequencing (see [1]–[3], [5]–[8]). The completed in 2003 project gave the whole human DNA–sequence and the opportunity to obtain genomes of various organisms other than human. Nowadays, several areas of research, such as molecular biology, genetics, theory of evolution, pharmacology, etc, are interested in diverse investigations 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 (for example, [3], [5], [9]). This approach is evolved in the paper, using a test, suggested in [10], 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 ([5], [6]). Different approaches are used to construct a distance matrix between sequences in order to solve this problem. In this paper we use a test for homogeneity (see [10]), which gives an opportunity to estimate the measure of "relatedness" between DNA–sequences, for example, between different chromosomes or whole genomes of several organisms.

We first estimated experimentally the efficiency of the suggested tests, and then we applied them to analyze genetic texts of various biological organisms. The obtained results coincide with many quantitative and qualitative characteristics known from the literature, which demonstrates the efficiency of the method. Furthermore, we obtained several new results, interesting for bioinformatics.

The paper is organized in the following way. In the next section we present experimental results of efficiency research on simulation sequences. In section 3 we describe the

[1] Address: 86, Kirova street, Novosibirsk, Russia. Tel: 7383 269 8204. Fax: 3832698272.

notions of molecular biology and the application of the above methods for the analysis of DNA–sequences of various organisms. Subsection 3.1 contains the results of "memory" estimation for several genetic texts of various organisms. Subsection 3.2 is devoted to second considered problem of bioinformatics — estimating the measure of relatedness of various DNA–sequences and constructing phylogenetic tree according to the obtained results.

2 Experimental efficiency of information–theoretic tests

Such problems as the goodness-of-fit testing, homogeneity testing and others did not have non-parametric decisions, which were suggested in [10]. According to the paper the results for these tests are asymptotic and the exact efficiency of the tests is unknown. That is why we had to estimate experimentally the algorithms efficiency. To solve it we carried out several experiments over the simulation data.

We considered the sequences over the finite alphabet which were generated by the Markov process of finite memory and discrete time. In other words, it means that for the Markov process of the order m the probability that the next appearing symbol depends only on the m previous symbols (see [4]):

$$P(x_i = a|x_1 = a_1, \dots, x_{i-1} = a_{i-1}) = P(x_i = a|x_{i-m} = a_{i-m}, \dots, x_{i-1} = a_{i-1}),$$

for all i, $a, a_1, \dots a_{i-1}$.

The first examined problem is the estimation of the source "memory" using the test for serial independence (see [10]). The second one — the problem of estimating the measure of relatedness between two sequences — is solved using the test for homogeneity (see [10]).

Before using these tests one has to choose some method of data compression. The symbol $\varphi(X)$ denotes some uniquely decodable code (or the lossless method of data compression), where X is the set of sequences. As encoders φ we used *GenCompress* — a special archiver for genetic data compression, which is among the best archivers for genetic data (see [2]). We use the denotation $h^*_m(X)$ for the empirical Shannon entropy of the m-th order. (The formal definitions of the code and the empirical Shannon entropy are given in the appendix.)

2.1 Estimation of the source "memory"

Let us start from the description of the test for serial independence. Let there be a sample X presented by r sequences $x^1 = x^1_1 \dots x^1_{t_1}, \dots, x^r = x^r_1 \dots x^r_{t_r}$, generated independently by some unknown source, and let $t = \sum_{i=1}^r t_i$. Two hypotheses are considered about the source, which generates the sequences from the sample. The main hypothesis H_0^{SI} is that the source is Markov, 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 source whose order is greater than m.

The suggested test is as follows (see [10]): *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 if the code φ is universal then the Type II error goes to 0, when t tends to infinity.

To estimate the efficiency of the given test we considered two families of stochastic Markov processes of the first and second order over the 2-letter and 4-letter alphabets respectively (the case of 4-letter alphabet corresponds to the case of genetic texts). The probability distributions are presented in Table 1, where A is an alphabet, $0 \leq \delta \leq {}^1/_{|A|}$.

Table 1: The distributions which were used to generate simulation sequences. Binary and 4-letter alphabets were considered: $A = \{0, 1\}, A = \{0, 1, 2, 3\}$. The "memory" m took on the values $m = 1, 2$.

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

We decided to estimate the power of suggested tests. We analyzed the sequences, which were generated by the source with the distribution from Table 1. In order to estimate experimentally the size of the input data (which is necessary to find given divergences through being analyzed sequences) we varied the value of the parameter δ from Table 1. We considered only several values of δ as some example of decreasing sequence. First experiments were carried out for big δ–values in order to obtain the distribution, very far from the Bernoulli source with equal probabilities of symbols. And then the value of δ was decreasing to obtain more similar sources. (It is obvious that the less is the value of δ, the closer the examined Markov 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.)

After choosing the parameter δ for current experiment we began to vary the length of considered sequences. We started from rather short ones, like 2^8, and than increased them as the power of 2. It happened so that for short sequences the test accepted the main hypothesis H_0^{SI} for $m = 0$. But when the sequence length increased — the test accepted the main hypothesis only for larger values of m, greater then 0. We knew a priory the order of the source — it was equal to 1 or 2 according to the considered source from Table 1. And we stopped our experiments when the test detected the correct source order for all sequences

Table 2: Experimental estimation of the efficiency for the test of serial independence, the order of the source is 1 ($m = 1$). H_0^{SI} claims that all 50 sample sequences are generated by the source whose order is not greater then m, and H_1^{SI} is accepted if the order of the source is greater then m. The cells contain the number of correct results among 50 carried out experiments.

	$\lvert A\rvert = 2$					$\lvert A\rvert = 4$				
Len.$\backslash\delta$	0.3	0.2	0.1	0.05	0.025	0.2	0.1	0.05	0.025	0.01
2^8	0	0	0	0	0	0	0	0	0	0
2^9	24	0	0	0	0	25	0	0	0	0
2^{10}	**50**	3	0	0	0	**50**	0	0	0	0
2^{11}	50	**50**	0	0	0	50	0	0	0	0
2^{12}	50	50	1	0	0	50	45	0	0	0
2^{13}	50	50	**50**	0	0	50	**50**	0	0	0
2^{14}	50	50	50	0	0	50	50	5	0	0
2^{15}	50	50	50	1	0	50	50	**50**	0	0
2^{16}	50	50	50	9	0	50	50	50	0	0
2^{17}	50	50	50	47	0	50	50	50	15	0
2^{18}	50	50	50	**50**	0	50	50	50	**50**	0
2^{19}	50	50	50	50	0	50	50	50	50	0
2^{21}	50	50	50	50	0	50	50	50	50	0
2^{23}	50	50	50	50	0	50	50	50	50	0
2^{25}	50	50	50	50	0	50	50	50	50	0
2^{28}	50	50	50	50	0	50	50	50	50	0

from the sample, accepting the main hypothesis H_0^{SI} for the correct value of m.

The results of testing are presented in Tables 2 – 3. (Here and below the required level of significance α is equal to 0.01.) In each case we generated 50 sequences according to the distributions from Table 1, and their lengths were equal to 2^n, $8 \le n \le 28$. Besides, the value of δ was varied: for the 2-letter alphabet δ takes on the values 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 tables include the amount of sequences (from 50 generated ones) for which the test correctly determined the order of Markov source. The bold style indicates the correct order determination for the first time for all 50 sequences. For example, the cell in the intersection of the row 2^9 and the column 0.3 in Table 2 corresponding to the 2-letter alphabet, includes the value 24, and this means that the test correctly determines the order of Markov source 24 times among 50 considered samples for the sequences of the length 2^9 ($\delta = 0.3$).

Thus we see the experimental efficiency of the suggested algorithm for hypothesis testing, because the correct determination of the order takes place for the sequences of moderate lengths.

2.2 Homogeneity testing

Let us turn to the experimental investigation of the efficiency of the test for homogeneity in order to estimate the measure of relatedness between different sequences. Let us formulate the algorithm, suggested in [10].

Similarly to the test for serial independence X is a sample presented by r sequences $x^1 = x_1^1 \dots x_{t_1}^1, \dots, x^r = x_1^r \dots x_{t_r}^r$, $\varphi(X)$ denotes some uniquely decodable code. Besides, it

Table 3: Experimental estimation of the efficiency for the test of serial independence, the order of the source is 2 ($m = 2$). H_0^{SI} claims that all 50 sample sequences are generated by the source whose order is not greater then m, and H_1^{SI} is accepted if the order of the source is greater then m. The cells contain the number of correct results among 50 carried out experiments.

	$\|A\| = 2$					$\|A\| = 4$				
$Len.\backslash\delta$	0.3	0.2	0.1	0.05	0.025	0.2	0.1	0.05	0.025	0.01
2^8	0	0	0	0	0	0	0	0	0	0
2^9	26	0	0	0	0	12	0	0	0	0
2^{10}	**50**	2	0	0	0	**50**	0	0	0	0
2^{11}	50	**50**	0	0	0	50	0	0	0	0
2^{12}	50	50	1	0	0	50	43	0	0	0
2^{13}	50	50	**50**	0	0	50	**50**	0	0	0
2^{14}	50	50	50	0	0	50	50	4	0	0
2^{15}	50	50	50	0	0	50	50	**50**	0	0
2^{16}	50	50	50	8	0	50	50	50	0	0
2^{17}	50	50	50	46	0	50	50	50	10	0
2^{18}	50	50	50	**50**	0	50	50	50	**50**	0
2^{19}	50	50	50	50	0	50	50	50	50	0
2^{21}	50	50	50	50	0	50	50	50	50	0
2^{23}	50	50	50	50	0	50	50	50	50	0
2^{25}	50	50	50	50	0	50	50	50	50	0
2^{28}	50	50	50	50	0	50	50	50	50	0

is known a priory that all these sequences are generated by Markov 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. Two hypotheses are considered about the sample: 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 suggested test is as follows (see [10]): *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 if the code φ is universal then 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.

We tried to determine the power of the test for homogeneity. As a sample we considered a pair of sequences: one of them was generated by the source from Table 1 and another was generated by the Bernoulli source with equal probabilities of symbols. So the main hypothesis H_0^{hom} was that two sequences from the sample were generated by the same source. But according to the being analyzed sequences, the alternative hypothesis H_1^{hom} was correct — that the sequences from the sample were generated by two different sources.

The value of δ was varied during experiments, because it is obvious that while δ is decreasing, the sequence, generated by the Markov source from Table 1, becomes closer to the sequence, generated by the Bernoulli source with equal probabilities of symbols. After

Table 4: Experimental estimation of the efficiency of homogeneity testing, the order of the source is equal to 1 ($m = 1$). H_0^{hom} claims that all sequences are generated by one source and the alternative one H_1^{hom} is that there are two sequences that are generated by different sources. The cells of the table contain the amount of samples, for which the test determined the sequences as generated by two different sources among 50 carried out experiments.

	$\|A\| = 2$				$\|A\| = 4$			
$Len.\backslash\delta$	0.3	0.2	0.1	0.05	0.2	0.1	0.05	0.025
2^8	0	0	0	0	0	0	0	0
2^9	0	0	0	0	0	0	0	0
2^{10}	1	0	0	0	0	0	0	0
2^{11}	**50**	0	0	0	**50**	0	0	0
2^{12}	50	45	0	0	50	20	0	0
2^{13}	50	**50**	0	0	50	**50**	0	0
2^{14}	50	50	0	0	50	50	0	0
2^{15}	50	50	**50**	0	50	50	0	0
2^{16}	50	50	50	0	50	50	**50**	0
2^{17}	50	50	50	0	50	50	50	0
2^{18}	50	50	50	0	50	50	50	0
2^{19}	50	50	50	0	50	50	50	0
2^{20}	50	50	50	0	50	50	50	0
2^{23}	50	50	50	0	50	50	50	0
2^{25}	50	50	50	0	50	50	50	0
2^{28}	50	50	50	0	50	50	50	0

choosing the value of δ for current experiment we began to vary the length of considered sequences from the sample. It happened that for short sequences the test accepted the main hypothesis H_0^{hom}, which was wrong. But as the length of considered sequences was increasing the correct hypothesis H_1^{hom} was accepted more often. We finished our experiments when the test accepted the correct hypothesis for all being analyzed samples.

The results are presented in Tables 4–5. We generated 50 sequences of lengths 2^n, $8 \leq n \leq 28$, according to the distributions from Table 1, moreover δ was varied 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. Moreover, we generated one sequence of the length 2^n for every n by the Bernoulli source with equal probabilities of symbols. So we analyzed 50 pairs of sequences in each case: the first one was generated according to the distribution from Table 1 and the second was generated by the Bernoulli source with equal probabilities of symbols. The cells of the tables contain the number of pairs for which the test determines correctly that they are generated by two different Markov sources. The value of the source order m is decided to be known a priory. The bold type indicates the case when for all 50 samples among 50 generated ones the test for the first time distinguished the sequences as generated by two different sources. 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 testing results presented in Tables 2 – 5, the tests correctly determine the order of the source and also distinguish sequences, generated by two different sources, if the divergence of the sequence from one generated by the source with equal probabilities of symbols is more then 0.025 over the 4-letter alphabet. Furthermore, the required amount of input data for this analysis is moderate.

Table 5: Experimental estimation of the efficiency of homogeneity testing, the order of the source is equal to 2 ($m = 2$). H_0^{hom} claims that all sequences are generated by one source and the alternative one H_1^{hom} is that there are two sequences that are generated by different sources. The cells of the table contain the amount of samples, for which the test determined the sequences as generated by two different sources among 50 carried out experiments.

	$\lvert A\rvert = 2$				$\lvert A\rvert = 4$			
$Len.\backslash\delta$	0.3	0.2	0.1	0.05	0.2	0.1	0.05	0.025
2^8	0	0	0	0	0	0	0	0
2^9	0	0	0	0	0	0	0	0
2^{10}	0	0	0	0	0	0	0	0
2^{11}	**50**	0	0	0	**50**	0	0	0
2^{12}	50	34	0	0	50	0	0	0
2^{13}	50	**50**	0	0	50	0	0	0
2^{14}	50	50	0	0	50	**50**	0	0
2^{15}	50	50	**50**	0	50	50	0	0
2^{16}	50	50	50	0	50	50	**50**	0
2^{17}	50	50	50	0	50	50	50	0
2^{18}	50	50	50	0	50	50	50	0
2^{19}	50	50	50	0	50	50	50	0
2^{20}	50	50	50	0	50	50	50	0
2^{23}	50	50	50	0	50	50	50	0
2^{25}	50	50	50	0	50	50	50	0
2^{28}	50	50	50	0	50	50	50	0

3 Applications for the analysis of genetic texts

3.1 Genetics notions

Before the examination of molecular biology problems devoted to the analysis of DNA–sequences let us consider several biological notions, which are used in this section. 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 [6]).

Then the DNA–sequence is divided into triplets of symbols, which are called *codons.* Codons are common units of the genetic code, because they are used to encode the insertion of one amino acid, in turn the sequence of amino acids forms genes. *Genes* are sections of DNA–sequence bearing consistent information about one protein or ribonucleic acid molecule. The succession of codons within one gene determines the succession of amino acids in the protein chain, which is encoded by this gene. The genes of highly organized creatures consist of two parts (those parts are called exons and introns). Exons are the coding sites of the gen, so the sequence of letters here corresponds to some sequence of amino acids of the protein, whereas introns are the part of the gene that do not contain information about the amino acids of protein.

Table 6: Testing of serial independence for archaebacteria in order to determine their genetic text "memory" (Part I). The value of "memory" is presented in the last column.

N	Name	Chromo- some	Length	Number of genes	Memory
1	Aeropyrum pernix K1		1669696	1752	3
2	Archaeoglobus fulgidus		2178400	2486	3
3	Haloarcula marismortui	I	3131724	3186	3
4	ATCC 43049	II	288050	285	4
5	Halobacterium sp. NRC-1		2014239	2127	3
6	Haloquadratum walsbyi DSM 16790		3132494	2875	7
7	Hyperthermus butylicus DSM 5456		1667163	1672	3
8	Metallosphaera sedula DSM 5348		2191517	2341	3
9	Methanocaldococcus jannaschii DSM 2661		1664970	1772	3
10	Methanococcoides burtonii DSM 6242		2575032	2497	8
11	Methanococcus maripaludis C5		1780761	1880	6
12	Methanococcus maripaludis S2		1661137	1772	5
13	Methanocorpusculum labreanum Z		1804962	1819	6
14	Methanoculleus marisnigri JR1		2478101	2555	4
15	Methanopyrus kandleri AV19		1694969	1729	3
16	Methanosaeta thermophila PT		1879471	1781	7
17	Methanosarcina barkeri str. Fusaro, chr. I		4837408	3811	9
18	Methanosarcina mazei Go1		4096345	3436	8
19	Methanosarcina acetivorans C2A		5751492	4721	9

3.2 Experimental investigation of the genetic text "memory"

Several papers are mentioned in [7], which contain suggestions about the depth of interconnection between symbols within one DNA–sequence. One of the suggestion is that the depth of interconnection varies only from 3 to 6 bases, while the other assumes that the variation is from 1 to 10000 bases. So the question about the depth of interconnection between symbols within the DNA–sequences was not finally solved. As a result the attempts to model genetic texts by Markov processes applied only to the sources of low orders — zero, first and second (see, for example, [3], [9]).

In order to estimate the "memory" of genetic texts we carried out several experiments by using information–theoretical tests for hypothesis testing, which were considered in the previous section. We obtained 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 biologically close organisms. In addition the obtained results show the dispersion of the "memory" value from 2 up to 9 for considered genetic texts (see Tables 6 – 12). To investigate the DNA–sequence "memory" of various species we considered several

Table 7: Testing of serial independence for archaebacteria in order to determine their genetic text "memory" (Part II). The value of "memory" is presented in the last column.

N	Name	Chromosome	Length	Number of genes	Memory
20	Methanosphaera stadtmanae DSM 3091		1767403	1588	7
21	Methanospirillum hungatei JF-1		3544738	3304	8
22	Nanoarchaeum equitans Kin4-M		490885	582	3
23	Natronomonas pharaonis DSM 2160		2595221	2726	3
24	Picrophilus torridus DSM 9790		1545895	1581	3
25	Pyrobaculum aerophilum str. IM2		2222430	2706	3
26	Pyrobaculum arsenaticum DSM 13514		2121076	2407	3
27	Pyrobaculum calidifontis JCM 11548		2009313	2200	3
28	Pyrobaculum islandicum DSM 4184		1826402	2062	5
29	Pyrococcus abyssi		1765118	1993	3
30	Pyrococcus furiosus DSM 3638		1908256	2228	6
31	Pyrococcus horikoshii OT3		1738505	2005	3
32	Staphylothermus marinus F1		1570485	1646	3
33	Sulfolobus acidocaldarius DSM 639		2225959	2329	3
34	Sulfolobus solfataricus P2		2992245	3031	9
35	Sulfolobus tokodaii str. 7		2694756	2874	7
36	Thermococcus kodakarensis KOD1		2088737	2358	3
37	Thermofilum pendens Hrk 5		1781889	1879	3
38	Thermoplasma acidophilum DSM 1728		1564906	1530	3
39	Thermoplasma volcanium GSS1		1584804	1548	6

procaryote and eukaryotes. Procaryote is a kingdom of single-celled alive organisms without an arranged cell nucleus. Bacteria and archaebacteria are procaryote. Eukaryotes are cell nucleus organisms (see [11]).

So to determine the "memory" of various genetic texts we considered the main hypothesis H_0^{SI} that the "memory" of the being analyzed DNA–sequence was equal to m. We tested several values of m in order, in other words, first we suggested that the DNA–sequence "memory" was equal to 0 and we considered the main hypothesis H_0^{SI} that $m = 0$. After the test rejected this hypothesis we considered the new main hypothesis H_0^{SI} that $m = 1$, and so on. We stopped our experiments when the main hypothesis was accepted for some m. This value of m is considered as the obtained "memory" of the genetic text.

The genomes of 38 archaebacteria and 43 bacteria were analyzed among procaryote (all the chromosomes were considered if there were any). All the DNA–sequences were taken from the database [11]. We considered only whole genetic texts during experiments. The results of calculation for archaebacteria are presented in Tables 6–7 and for bacteria in Tables 8–10. Such popular objects of biological research were taken as samples of eucaryotes: cryptomonad alga *Guillardia theta nucleomorph*, budding yeast *Saccharomyces cerevisiae S288C*, fission yeast *Schizosaccharomyces pombe* and microsporidian parasite *Encephalitozoon cuniculi*, for each the whole amount of chromosomes was considered — 3, 16, 3 and 11, respectively. The results are presented in Tables 11–12.

Table 8: Testing of serial independence for bacteria in order to determine their genetic text "memory" (Part I). The value of "memory" is presented in the last column.

N	Name	Chromo-some	Length	Number of genes	Memory
1	Acidobacteria bacterium Ellin345		5650368	4834	4
2	Acidothermus cellulolyticus 11B		2443540	2217	3
3	Anaplasma marginale St Maries		1197687	1005	8
4	Anaplasma phagocytophilum HZ		1471282	1411	8
5	Aquifex aeolicus		1551335	1580	3
6	Bacillus anthracis Ames		5227293	5630	7
7	Bacillus anthracis str Sterne		5228663	5415	7
8	Bacillus cereus ATCC 10987		5224283	5772	8
9	Bacillus cereus ATCC 14579		5411809	5476	8
10	Bacillus cereus ZK		5300915	5269	8
11	Bacillus clausii KSM-K16		4303871	4204	7
12	Bacillus halodurans		4202352	4171	9
13	Bacillus licheniformis ATCC 14580		4222334	4290	7
14	Bacillus thuringiensis Al Hakam		5257091	4883	8
15	Bacillus thuringiensis konkukian		5237682	5261	8
16	Bacteroides fragilis NCTC 9434		5205140	4347	7

Tables 6–10 are organized uniformly. The column "Name" contains the Latin name of the organism. The column "Chromosome" points to the number of considered chromosome if there is one. The column "Length" contains the number of bases in the DNA chain. There is the number of different genes in the DNA chain in the column "Number of genes" (the data were taken from [11]). The column "Memory" contains the results of the "memory" calculation for the considered organism genetic texts using the tests from [10]. Tables 11–12 differ only in the form of the organism enumeration. The name of the current organism is indicated through the whole table and the numbers of considered chromosomes are presented in the column "Chromosome".

Let us present several observations according to the obtained results. Considering the data from tables 6–12 we note that bacteria and eucaryotes have the relatively large values of "memory" though the length of each chromosome or the whole genome is small enough. So 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 such noncoding sites of the DNA–sequence as introns (they may add some long correlations within considered DNA–sequence) and the increasing amount of duplicating genes.

The preliminary analysis let us suggest that the length and the amount of genes of the DNA–sequence is statistically associated with the genetic text "memory". The coefficients of correlation between the pairs of the data samples were calculated to check this hypothesis: between the "memory" and the length, between the "memory" and the number of genes. The results are presented in Table 13. Thus we see that the "memory" characteristic is of standalone biological interest, because the correlation with other standard parameters of the DNA–sequence 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 9: Testing of serial independence for bacteria in order to determine their genetic text "memory" (Part II). The value of "memory" is presented in the last column.

N	Name	Chromo- some	Length	Number of genes	Memory
17	Bacteroides fragilis YCH46		5277274	4670	7
18	Bacteroides thetaiotaomicron VPI-5482		6260361	4864	8
19	Bartonella bacilliformis KC583		1445021	1375	8
20	Bartonella henselae Houston-1		1931047	1665	8
21	Bartonella quintana Toulouse		1581384	1356	8
22	Baumannia cicadellinicola Homalodisca coagulata		686194	651	2
23	Bdellovibrio bacteriovorus		3782950	3623	3
24	Bifidobacterium adolescentis ATCC 15703		2089645	1700	6
25	Bifidobacterium longum		2256640	1798	6
26	Bordetella bronchiseptica		5339179	5072	3
27	Bordetella parapertussis		4773551	4467	4
28	Bordetella pertussis		4086189	3867	8
29	Borrelia afzelii PKo		905394	894	6
30	Borrelia burgdorferi		910724	875	6
31	Borrelia garinii PBi		904246	869	6
32	Bradyrhizobium ORS278		7456587	6818	4
33	Brucella abortus 9-941	I	2124241	2200	4
34		II	1162204	1156	3
35	Brucella melitensis	I	2117144	2107	4
36		II	1177787	1157	3
37	Brucella melitensis biovar	I	2121359	2236	4
38	Abortus	II	1156948	1182	3
39	Brucella suis 1330	I	2107794	2231	4
40		II	1207381	1220	3

Let us mention such unexpected fact that the "memory" of DNA–sequences even for the biologically related organisms (belonging to one genus) can vary greatly. Archaebacteria from the genus *Sulfolobus* and bacteria from *Bordetella* are the examples. Archeobacteria *Sulfolobus acidocaldarius DSM 639*, *Sulfolobus solfataricus P2* and *Sulfolobus tokodaii str.7* have the comparable length of genomes: from 2.1MB to 2.8MB though the determined by the test "memory" differs considerably — 3, 9 and 7, respectively. In regard to bacteria *Bordetella bronchiseptica*, *Bordetella parapertussis* and *Bordetella pertussis*, the size of genomes varies from 4MB to 5.3MB, but the obtained "memory" has 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 organisms of one genus.

According to the literature Markov processes of the order not greater than 2 are usually used to model DNA–sequences. But according to the obtained results the genetic text "memory" is usually more than 2. Therefore it is better to use the models of higher orders to analyze the dependencies within DNA–sequences. In order to find the most suitable Markov process it is possible to use the test for serial independence, suggested in [10].

Table 10: Testing of serial independence for bacteria in order to determine their genetic text "memory" (Part III). The value of "memory" is presented in the last column.

N	Name	Chromo-some	Length	Number of genes	Memory
41	Buchnera aphidicola Cc Cinara cedri		416380	397	3
42	Buchnera aphidicola Sg		641454	619	2
43	Buchnera aphidicola str. Bp		615980	550	3
44	Buchnera sp		640681	607	2
45	Burkholderia mallei NCTC	I	2284095	2215	7
46	10229	II	3458208	3409	7
47	Burkholderia mallei NCTC	I	2352693	2412	7
48	10247	II	3495678	3553	8
49	Burkholderia mallei	I	1734922	1763	7
50	SAVP1	II	3497479	3532	7
51	Helicobacter pylori 26695		1667867	1630	6
52	Helicobacter pylori J99		1643831	1535	4
53	Staphylococcus aureus RF122		2742531	2665	8
54	Staphylococcus epidermidis ATCC 12228		2499279	2495	8
55	Staphylococcus haemolyticus JCSC1435		2685031	2753	8
56	Streptococcus agalactiae A909		2127858	2136	8
57	Streptococcus pyogenes M1 GAS SF370		1852455	1805	7
58	Streptococcus pyogenes MGAS315		1900535	1951	8
59	Streptococcus thermophilus LMG18311		1796846	1974	8

3.3 Homogeneity testing for genetic texts

In molecular biology and genetics the problem of genome comparison or comparison of its parts is often risen. The solution of this problem allows us to find the same or related genes, to build the phylogenetic trees, etc. ([1], [6]). Let us consider the problem of estimating the measure of relatedness between various organisms, trying to understand, whether two DNA–sequences are "generated" by one source or by to different sources. The obtained results were used to construct the example of the phylogenetic tree. In this section the attempt to estimate the measure of relatedness between various organisms is undertaken using the test for homogeneity (see [10]).

The binary logarithm of the shortest initial fragmentation length (on which we were able to distinct two sequences as generated by different sources) was the indicator of closeness between two DNA–sequences (Tables 14–16). The initial fragmentation of the DNA–sequence was increasing as a power of 2. That is if we considered the initial fragmentation of the length 2^n, then the length increased to 2^{n+1} and so on. When we found the value of n, on which the hypothesis of homogeneity was rejected, then this n was supposed to be the measure of relativeness between considered sequences. If the sequences vary greatly then the measure of closeness is small. But if the sequences are very close to each other then the distinguishing of sequences may not take place even when one considers the whole genome.

Table 11: Testing of serial independence for eukaryotes in order to determine their genetic text "memory" (Part I). The value of "memory" is presented in the last column.

N	Chromosome	Length	Number of genes	Memory
Saccharomyces cerevisiae S288C				
1	1	230208	101	7
2	2	813178	420	6
3	3	316617	173	6
4	4	1531918	785	8
5	5	576869	297	7
6	6	270148	137	3
7	7	1090946	563	7
8	8	562643	293	7
9	9	439885	227	6
10	10	745745	382	7
11	11	666454	329	3
12	12	1078175	532	8
13	13	924429	482	7
14	14	784333	409	7
15	15	1091289	558	7
16	16	948062	483	7
Guillardia theta nucleomorph				
17	1	196216	160	7
18	2	180915	126	7
19	3	174133	163	7
Schizosaccharomyces pombe				
20	1	5566797	4643	8
21	2	4467299	3856	7
22	3	2455984	1913	8

Thus the larger is the value corresponding to the pair of sequences the more close these sequences are to each other.

So for all pairs of genetic texts we considered the main hypothesis H_0^{hom} that they were "generated" by the same source, and the alternative hypothesis that they were "generated" by two different sources. If the distinguishing of the DNA–sequences happened only when whole genomes were considered then we denoted the average value of lengths for considered organisms with the symbol (‡) over it. And "*no*" denotes the case when even the whole genome consideration did not give us an opportunity to distinguish two sequences.

Let us present the results of homogeneity testing for several groups of organisms. In Table 14 there are the results of the test for 7 archaebacteria: *Archaeoglobus fulgidus (u_1)*, *Methanococcus maripaludis C5* (u_2), *Methanococcus maripaludis S2* (u_3), *Pyrococcus abyssi* (u_4), *Pyrococcus furiosus DSM 3638* (u_5), *Pyrococcus horikoshii OT3* (u_6), *Thermoplasma volcanium GSS1* (u_7). These samples were chosen to form two groups of biologically close organisms (pair u_2, u_3 is from the genus *Methanococcus*, triplet u_4, u_5, u_6 — from the genus *Pyrococcus*) in order to compare them with each other and with the organisms from other genera — u_1 and u_7. As it is seen from Table 14, the sequences u_2 and u_3 were not determined by the test as generated by different sources even when one considered the whole genomes, just like the triplet u_4, u_5 and u_6, whereas the other pairs were differed in

Table 12: Testing of serial independence for eukaryotes in order to determine their genetic text "memory" (Part II). The value of "memory" is presented in the last column.

N	Chromosome	Length	Number of genes	Memory
		Encephalitozoon cuniculi		
23	1	209982	166	6
24	2	197426	158	3
25	3	194439	159	3
26	4	218328	173	4
27	5	211018	176	3
28	6	220294	178	4
29	7	226573	195	3
30	8	238147	213	4
31	9	250202	211	6
32	10	262796	196	3
33	11	267509	215	5

Table 13: Coefficients of correlation between the genetic text "memory" and length or the number of genes.

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

smaller initial fragmentation. It is predictable because these combinations of the organisms are taxonomically related. This result is especially interesting if one remembers that the length of sequence for archaebacteria is relatively small.

In Table 15 there are the results of testing for 10 bacteria: *Acidobacteria bacterium Ellin 345* (u_8), *Helicobacter pylori 26695* (u_9), *Helicobacter pylori J99* (u_{10}), *Staphylococcus aureus RF122* (u_{11}), *Staphylococcus epidermidis ATCC* 12228 (u_{12}), *Staphylococcus haemolyticus JCSC1435* (u_{13}), *Streptococcus agalactiae A909* (u_{14}), *Streptococcus pyogenes M1 GAS* SF370 (u_{15}), *Streptococcus pyogenes MGAS315* (u_{16}), *Streptococcus thermophilus LMG 18311* (u_{17}). These organisms were chosen using the same criteria just for the archaebacteria — several samples were taken from the same genus to form the group of close organisms. These groups were compared between each other and with u_8, which does not belong to any group. The obtained data give the opportunity to make a conclusion that the bacterium u_8 differs a lot from all others which corresponds to its position in the hierarchy of the bacteria. Moreover if one considers the following combinations of genetic texts: pair u_9, u_{10} from *Helicobacter*, triplet u_{11}, u_{12}, u_{13} from *Streptococcus*, quadruple u_{14} — u_{17} from *Streptococcus*, then it is possible to mention that for each combination the distinguishing took place either for the large initial fragmentation of the DNA–sequence or did not take place at all. So these organisms are close to each other according to the test for homogeneity.

In Table 16 there are results for 7 procaryote: *Archaeoglobus fulgidus (u_{18})*, *Pyrococcus abyssi* (u_{19}), *Pyrococcus horikoshii OT3* (u_{20}), *Escherichia coli K-12 MG1655* (u_{21}), *Haemophilus influenzae* (u_{22}), *Helicobacter pylori 26695* (u_{23}), *Helicobacter pylori J99* (u_{24}). This set of organisms was chosen because it was analyzed in [2], where the phylogenetic tree (see Figure 1a) was built. The phylogenetic tree (see Figure 1b) was obtained according to

Table 14: Homogeneity testing for archaebacteria. 7 archaebacteria of various genera were considered in order to determine - whether the test distinguished the genetic texts of close organism or not, and what would be the corresponding length of distinguishing. The cells contain n — the power of 2, which notes that the distinguishing took place when the length of sequences was equal to 2^n. Symbol ”no” notes that the test did not distinguish even the whole genomes.

Archaea	u_1	u_2	u_3	u_4	u_5	u_6	u_7
u_1	—	16	19	18	17	17	17
u_2	16	—	*no*	15	15	15	16
u_3	19	*no*	–	19	19	19	19
u_4	18	15	19	—	*no*	*no*	17
u_5	17	15	19	*no*	—	*no*	17
u_6	17	15	19	*no*	*no*	—	16
u_7	17	16	19	17	17	16	—

the results of Table 16. We used the method of the ”nearest-neighbor” to construct it. It means that we chose the most close to each other sequences and then we considered them as one element. To continue the procedure we re-counted the items of the distance matrix: if the i-th and j-th sequences were the most close ones, than for any sequence k $(m_{ki} + m_{kj})/2$ was assumed to be the distance between it and the ”glued” sequence, corresponding to i-th and j-th, where m_{kl} — was the item of the initial distance matrix. It is easy to mention that these trees are the same except the position of u_{21}. Perhaps this position is the result of the original u_{21}–sequence length that is 2.5 times larger than for other samples.

Therefore, the information–theoretic test for homogeneity can be used to determine the ”measure” of relatedness between genomes of various organisms or between chromosomes of the same organism.

4 Conclusion

The problems of DNA–sequence modeling and estimating the measure of relatedness between genetic texts of various organisms lie in the field of interest of molecular biology, genetics and other areas of research. The suggested tests for the serial independence and homogeneity (see [10]) can help to find the new useful methods for solving these problems.

There are several approaches to analyze the statistical structure of DNA—sequences. One of the most famous is to model them using Markov processes of different orders. But previously the Markov models with order less or equal to 2 were considered in the literature as more simple ones. Although according to the obtained results the ”memory” of genetic texts is usually more than 2. So it is better to use the processes of higher order, because these Markov models could give the opportunity to reveal more delicate regularities in DNA–sequence structure.

In molecular biology it is often necessary to compare different parts of genetic texts, for example, while constructing phylogenetic trees for various organisms. The test for homogeneity can become a tool for estimating the measure of relatedness between genomes or chromosomes of various organisms, because it is possible to determine, wether two given sequences are generated by the same source or by two different sources.

The obtained results of the test for serial independence and the test for homogeneity coincide with the known biological data, which demonstrates the efficiency of the considered method.

Table 15: Homogeneity testing for bacteria. 10 bacteria of various genera were considered in order to determine - whether the test distinguished the genetic texts of close organism or not, and what would be the corresponding length of distinguishing. The cells contain n — the power of 2, which notes that the distinguishing took place when the length of sequences was equal to 2^n. Symbol "no" notes that the test did not distinguish even the whole genomes. Symbol ‡ notes the cells which contain the average length of two being analyzed sequences ($*10^4$), because for these pair the distinguishing took place only while one considered the whole genomes.

Bacteria	u_8	u_9	u_{10}	u_{11}	u_{12}	u_{13}	u_{14}	u_{15}	u_{16}	u_{17}
u_8	—	13	13	13	13	13	16	16	16	13
u_9	13	—	*no*	15	*no*	217‡	*no*	176‡	*no*	173‡
u_{10}	13	*no*	—	15	*no*	216‡	*no*	*no*	*no*	*no*
u_{11}	13	15	15	—	*no*	271‡	*no*	*no*	*no*	19
u_{12}	13	*no*	*no*	*no*	—	‡	*no*	20	19	214‡
u_{13}	13	217‡	216‡	271‡	259‡	—	240‡	226‡	229‡	224‡
u_{14}	16	*no*	*no*	*no*	*no*	240‡	—	*no*	*no*	214‡
u_{15}	16	176‡	*no*	*no*	20	226‡	*no*	—	*no*	*no*
u_{16}	16	*no*	*no*	*no*	19	229‡	*no*	*no*	—	20
u_{17}	13	173‡	*no*	19	214‡	224‡	214‡	*no*	20	—

5 Appendix. The empirical Shannon entropy

Let us formulate the definitions of the Shannon entropy, empirical Shannon entropy of the m-th order ant the universal code (see [10]).

Let τ be a stationary and ergodic source generating letters from a finite alphabet A. The m–order (conditional) Shannon entropy and the limit Shannon entropy are defined as follows:

$$h_m(\tau) = \sum_{v \in A^m} \tau(v) \sum_{a \in A} \tau(a|v) \log \tau(a|v), \qquad h_\infty(\tau) = \lim_{m \to \infty} h_m(\tau).$$

Given sample X for the analysis is presented by r sequences $x^1 = x^1_1 \dots x^1_{t_1}, \dots, x^r = x^r_1 \dots x^r_{t_r}$ and $t = \sum_{i=1}^r t_i$, then 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

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

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

Acknowledgment

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

Table 16: Homogeneity testing for the organisms from [2]. 7 procaryote were considered in order to determine - whether the test distinguished the genetic texts of close organism or not, and what would be the corresponding length of distinguishing. The cells contain n — the power of 2, which notes that the distinguishing took place when the length of sequences was equal to 2^n. Symbol "no" notes that the test did not distinguish even the whole genomes. Symbol ‡ notes the cells which contain the average length of two being analyzed sequences ($*10^4$), because for these pair the distinguishing took place only while one considered the whole genomes.

	u_{18}	u_{19}	u_{20}	u_{21}	u_{22}	u_{23}	u_{24}
u_{18}	—	18	17	17	17	17	17
u_{19}	18	—	*no*	14	15	14	15
u_{20}	17	*no*	–	14	15	15	15
u_{21}	17	14	14	—	15	14	14
u_{22}	17	15	15	15	—	20	173‡
u_{23}	17	14	15	14	20	—	*no*
u_{24}	17	15	15	14	173‡	*no*	—

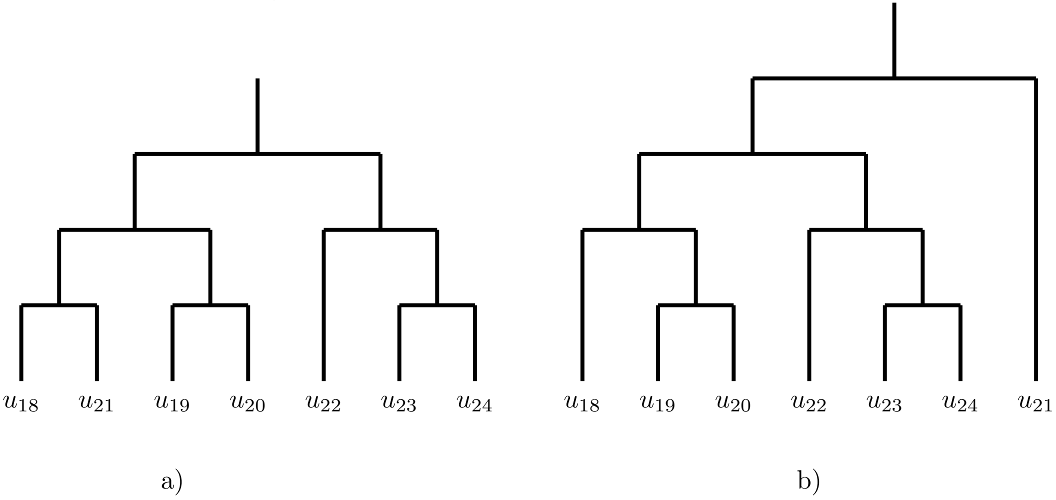

Figure 1: phylogenetic trees, a) — from [2],
b) — according to the data from Table 16.

References

[1] Aktulga, H.M., Kontoyiannis, I., Lyznik, L.A., Szpankowski, L., Grama, A.Y., Szpankowski, W., 2007. Identifying statistical dependence in genomic sequences via mutual information estimates. EURASIP J. Bioinformatics Systems Biology (accepted).

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

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

[4] Gallagher, R., 1968. Information theory and reliable communication, Wiley.

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

[6] Karp, R.M., 2002. Mathematical Challenges from Genomics and Molecular Biology. Notices of the AMS, 49(5), 544–553.

[7] Li, W., 1997. The Study of Correlation Structures of DNA–sequences: A Critical Review. Computers and Chemistry, 21(4), 257–271.

[8] Oprea, I., Pasca, S., Gavrila, V., 2004. Method of DNA Analysis Using the Estimation of the Algorithmic Complexity. Leonardo Electronic Journal of Practices and Technologies, 3(5), 53–66.

[9] Simons, G., Yao, Y-Ch., Morton, G., 2005. Global Markov models for eukaryote nucleotide data. J. Statist. Plann. Inference, 130, 251–275.

[10] Ryabko, B., Astola, J., 2006. Universal codes as a basis for time series testing. Statistical Methodology, 3, 375–397.

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