CLASSIFICATION OF SYMBOL SEQUENCES OVER THEIR FREQUENCY DICTIONARIES: TOWARDS THE CONNECTION BETWEEN STRUCTURE AND NATURAL TAXONOMY

A.N.Gorban[¶], T.G.Popova[¶*], M.G.Sadovsky[*¶]

[¶]*Institute of Computing Modelling of SD of RAS, Krasnoyarsk 660036*
[*]*Institute of Biophysics of SD of RAS, Krasnoyarsk, 660036*

The classifications of bacterial 16S RNA sequences developed over the real and transformed frequency dictionaries have been studied. Two sequences considered to be close each other, when their frequency dictionaries were close in Euclidean metrics. A procedure to transform a dictionary is proposed that makes clear some features of the information pattern of a symbol sequence. A comparative study of classifications developed over the real frequency dictionaries vs. the transformed ones has been carried out. A correlation between an information pattern of nucleotide sequences and taxonomy of the bearer of the sequence was found. The sites with high information value are found, that were the main factors of the difference between the classes in a classification. The classification of nucleotide sequences developed over the real frequency dictionaries of the thickness 3 reveals the best correlation to a gender of bacteria. A set of sequences of the same gender is included entirely into one class, as a rule, and the exclusions occur rarely. A hierarchical classification yields one or two taxonomy groups on each level of the classification. An unexpectedly often (in comparison to the expected), or unexpectedly rare occurrence of some sites within a sequence makes a basic difference between the structure patterns of the classes yielded; a number of those sites is not too great. Further investigations are necessary in order to compare the sites revealed with those determined due to other methodology.

Key words: nucleotide sequence, word, frequency, dictionary, function, classification, dynamic kerns method

Introduction

A study of the relation between a structure of symbol sequences (S) and the meaning of them encrypted in the interlocation of symbols is a key problem for molecular biology, biophysics and many other fields of sciences. Usually, researchers meet no problem in understating the function of sequences studied; at least, they may discuss it and elaborate a common opinion on that subject. A structure is much more complicated matter to understand. When studying nucleotide sequences,

they often tell about the intron-exon structure [1], or about the structure determined by operons, etc. [2]. Further, we should understand the structure of a sequence as the frequency dictionary of that latter, either real one [3 – 6], or the reconstructed one [7, 8], or the transformed one [8]. Such understanding of the structure of a sequence enables a researcher to introduce easily the idea of a closeness of two (or several) structures. Namely, two (or several) sequences are considered to be close each other, when their frequency dictionaries are close. The real frequency dictionary W_q (of the thickness q) is defined as the list of all the strings of the length q occurred in the given sequence accompanied with the frequency of their occurrence [3 – 6]. It has been shown that sufficiently big family of nucleotide sequences could be split into groups, according to closeness of the frequency dictionaries of those sequences. All the sequences within a group are close each other with respect to the Euclidean metrics between their frequency dictionaries. We have observed the correlation between a function encoded and the classification of sequences developed over their frequency dictionaries, for the *Ca*–dependent peptides; also, such correlation has been observed for the classification mentioned and the taxonomy of the organisms bearing those sequences [9].

The sequence of the frequency dictionaries W_1, W_2, ... W_q, corresponded to the same text yields a relation, namely all the foregoing dictionaries could be obtained from the succeeding ones due to a simple summation. In other words, a thinner dictionary could always be obtained from a thicker one. An inverse statement does not hold true. The exact reconstruction of a thicker dictionary from a given one does not exist always, in general case. The exact reconstruction of W_k over W_q for $k > q$ is possible iff all the words in W_q have the unique continuation. Otherwise, the single-valued reconstruction is impossible: each frequency dictionary W_q of the smaller thickness yields the set of thicker frequency dictionaries. One could nevertheless seek for the most probable continuation of a thicker dictionary $\widetilde{W}_k$ that corresponds to the given dictionary of smaller thickness W_q. To get this most probable dictionary $\widetilde{W}_k(q)$ one must choose among all possible continuations of the given dictionary W_q that one with the least determinacy, i.e. that one, which yields the maximal entropy. The exact solution of this extreme problem looks very close to the well-known Kirkwood approximation in statistical physics. If a reconstructed frequency dictionary $\widetilde{W}_k(q)$ coincides entirely with the original one W_k, then it means that the entire information on the original text is contained in the dictionary W_q. Deviations of the reconstructed dictionary $\widetilde{W}_k(q)$ from the real one W_k show what new is introduced by k–tiples, in comparison to q–tiples $k > q$.

From that point of view, the deviations of the real frequency dictionary W_k from the most probable continuation $\widetilde{W}_k(q)$ are of the greatest interest. To measure these deviations, one should introduce a new object, so called "transformed dictionary". In that latter, each word is assigned with the ration of the real frequency of the word to the frequency obtained due to a reconstruction from the thinner one, i.e. the value is introduced which displays how much the real frequency of the word differs from the expected one. This transformation of the frequency dictionary allows to explicate some peculiarities of the information structure of nucleotide sequences (NS).

A study of correlation between the structure and function of NS requires to determine some other relations among them (e.g., a classification). One could easily see that a set of genes provides always at least two independent classification of that type: over a taxonomy of gene bearer, and over the function of those genes. This paper is aimed to study a relation between the structure of NS (that is assumed to be a frequency dictionary, either real one, or transformed), and taxonomy of the gene bearer. Thus, only the couple "structure vs. taxonomy" is studied here, from a tripod pattern "structure ÷ taxonomy ÷ (biochemical) function". Since the structural variations may result both from differences in the functions encoded, and from the differences in taxonomy of the gene's bearer, we have chosen those NS that determine the same function in various organisms, for our investigation.

Currently, a huge amount of the sequenced genes is obtained, that allows to carry out the comparative investigation described above over the set of genes that would provide a valid results. We have used 16S RNA of bacteria of various species. All the sequences of that type realise the same function, not only in bacteria, but in other organisms with higher taxonomy position.

Frequency Dictionary

Let correspond to each gene sequenced a symbol sequence from four–letter alphabet, of the same length (i.e., number of nucleotides) N — a (genetic) text. Let call a word any continuos subsequence of the length q from the genetic text. Assign each word with its frequency (that is the number of its copies within the genetic text under investigation divided by the total number of words within the text); such list of all the words of the length q occurred within the text accompanied with their frequency would be called the frequency dictionary W_q [3 – 8].

If n is the power of the alphabet, then total number of all the words of the length q is n^q. Obviously, not every word of the length q could be met within a text, as soon as q is large enough. Let complete the frequency dictionary of the text studied to the entire one (which contains all the

words of the given length q) adding the words with zero frequency. Then every frequency dictionary could be represented as a point $F(f_1, f_2, \ldots, f_{n^q})$ in n^q–dimensional space with the coordinates presented by the frequencies of the relevant words f_j: $0 \le f_j \le 1$, $j = 1, 2, \ldots, n^q$. Then, a set of genes yields the set of points in n^q–dimensional space, according to such representation.

Transformation of frequency dictionary

Let consider a set of frequency dictionaries (of various thickness) corresponded to the same genetic text: W_1, W_2, … , W_q, … , W_N. Then the question arises what part of the information about the original text contains the dictionary of the thickness q. It is well-known, that all the words within a dictionary occur in a single copy, for some specific thickness d^* of the dictionary [3, 4]. Hence, all the words in W_q have a unique continuation, for $q > d^*$ and any dictionary W_k could be unambiguously reconstructed from the dictionary W_q, as q becomes greater than d^*, including the original text [4, 6]. Thus, any frequency dictionary W_q with $q > d^*$ contains the entire information about the original text.

Any thinner frequency dictionary is obtained from the given one W_q, by summation, and a part of the information about the text is eliminated, when $q < d^*$. It makes an inverse transformation (from the given dictionary to a thicker one) ambiguous. For every frequency dictionary W_q with $q < d^*$ there exists a set of different frequency dictionaries W_k of the same thickness k for $k > q$, and any of them might be considered as a continuation of the original dictionary. We should select the dictionary $\widetilde{W}_k(q)$ which is the most probable continuation of W_q; let's call that dictionary the reconstructed one. A methodology of implementation of the reconstructed dictionary is based on the maximum entropy principle [7, 8].

The entropy of a frequency dictionary W_q is defined as

$$S_q = -\sum_{j=1}^{n^q} f_j \ln f_j . \qquad (1)$$

A reconstruction of the dictionary $\widetilde{W}_k(q)$ must be provided with no additional information, hence the reconstructed dictionary must yield the maximal possible value of the entropy. The extreme problem $S_{q+s} \to \max$ with the bound condition for dictionaries $W_q \leftarrow \widetilde{W}_{q+s}(q)$ has a single solution

$$f_{i_1\ldots i_q i_{q+1}\ldots i_{q+s}} = \frac{f_{i_1\ldots i_q} f_{i_2\ldots i_{q+1}} \ldots f_{i_{q-s+1}\ldots i_{q+s}}}{f_{i_2\ldots i_q} f_{i_3\ldots i_{q+1}} \ldots f_{i_{q-s+1}\ldots i_{q+s-1}}} \text{ for } q>1, \text{ and} \tag{2}$$

$$f_{i_1\ldots i_{q+s}} = f_{i_1}\ldots f_{i_{q+s}} \text{ for } q=1, \tag{3}$$

here $i_1\ldots i_q i_{q+1}\ldots i_{q+s}$ is the word of the length $q+s$ and index i corresponds to a nucleotide. The expressions (2) and (3) looks very similar to the well-known Kirkwood approximation, for some problems of statistical physics [13].

The formulae (2) and (3) coincide with the well-known expressions for the transitional probabilities in a symbol sequence obtained as a realisation of the Markov random process, for $s=1$ (there are some specific differences for $s>1$). It should be stressed, that the formulae (2) and (3) for the reconstructed dictionary is yielded with respect to neither hypothesis on a peculiar structure of a sequence. These formulae present the most likelihood hypothesis on the frequency dictionary of the thickness $q+s$, that could be implemented from a consideration of the dictionary of the thickness q. One should consider the original symbol sequence to be Markovian one if and only if the expressions for the real (but not the reconstructed ones) frequencies would hold true in the limit case of infinitely long original sequence.

Here we considered the reconstructed dictionaries over the dictionaries of one symbol thinner. Below are the formulae for the reconstructed frequencies of $\widetilde{W}_q(q-1)$, for a convenience:

$$f_{i_1\ldots i_q} = \frac{f_{i_1\ldots i_{q-1}} f_{i_2\ldots i_q}}{f_{i_2\ldots i_{q-1}}}. \tag{4}$$

The reconstructed frequencies will be denoted as $\widetilde{f}$.

The dictionary $\widetilde{W}_q$ yielded is the most probable continuation of the dictionary W_{q-1}. A comparison of the reconstructed dictionary $\widetilde{W}_q$ with the original one W_q, of the same thickness q, allows to make explicitly the peculiarities of the information structure of nucleotide sequences, since the maximal differences between the real and reconstructed frequencies, for the dictionaries of the given thickness, are the most "unexpected" events in a transition from the dictionary of the thickness $q-1$ to the dictionary of the thickness q.

Let's transform the frequency dictionary of a nucleotide sequence in the following manner: for each length q of words, starting from $q=2$, develop the dictionary $\widetilde{W}_q(q-1)$ reconstructed from the preceding one. As before, each nucleotide sequence would be corresponded with the point $P(p_1, p_2, ..., p_{n^q})$ in n^q–dimensional space, where the coordinates of a point are the ratios of the real

frequency to the reconstructed one: $p_j = f_j / \widetilde{f}_j$, for $\widetilde{f}_j \neq 0$, and $p_j = 1$, for $\widetilde{f}_j = 0$, $j = 1, 2, \ldots, n^q$. The values p_j show how much the real frequencies of words differ from the expected ones. If $p_j \approx 1$ for some word within the genetic text, then the information value of that word is not high: its most probable expected frequency almost coincides to the real one. The words, whose frequency ratio p_j differs to the greatest extent (either upward, or downward) from the real one present the most valuable sites of the given length within the nucleotide sequence studied. We presume the value threshold to be equal to 15 – 20 %, in our study. It should be stressed that the length of a site is rather essential, since any low-valued site of the length q might be incorporated into the high-valued site of the length $q+1$, which, in turn, might be incorporated into a longer site of the low information value. To compare various nucleotide sequences, one should consider the differences between p_j observed for different sequences rather than the deviations of that value form 1 (that latter might occur simultaneously). The difference mentioned above becomes significant as it reaches the 15 to 20 % level, as it would be shown further.

Algorithms of automatic classification

The implementation of a classification of objects require a definition of a closeness measure among them. For symbol sequences, those measure could be introduced in several different ways [2, 14]. Here we have used the following measure.

For some length q of words, any genetic texts has been represented as a point in n^q – dimensional space, corresponded to the frequency dictionary of the thickness q (either real one, or transformed). Two sequences would be considered to be close each other, if two points in n^q – dimensional space are close, corresponded to their frequency dictionaries of the thickness q. A distance between two points in this multidimensional space is determined due to Euclidean metrics. A study of the distribution of the points in n^q –dimensional space allows to split an original set of sequences into a number of classes, where the sequences are close each other within a class, with respect to a closeness used. Otherwise, it allows to verify an absence of such splitting.

The automatic classification algorithms have been used to develop the classification mentioned above. We used a dynamic kern method in our investigation [12, 15]. In brief, looks as follows. Consider a set $\{F^i\}$ $i = 1, 2, \ldots, M$ of M points in space that should be split into several classes. Let the initial number of classes, and the initial distribution over those classes be provided. Firstly, for each k –th class, the centre $C^k(c_1^k, \ldots, c_{n^q}^k)$ is calculated according to

$$c_j^{\,k} = \frac{1}{l_k}\sum_{i=1}^{l_k} p_j^i, \quad j = 1, 2, \ \ldots, n^q, \tag{5}$$

where l_k is the number of elements in k–th class. Then, for each point of the original set F^i the distance

$$d_i^{\,k} = \rho\left(C^k, F^i\right), \quad i = 1, 2, \ \ldots, \ M \tag{6}$$

is calculated from that latter to the centre of each class, and the outfit of each point is re-determined. A point is presumed to belong to the class which yields the least distance to that point. As soon as all the points are run, the group centres are re-calculated. This procedure — calculation of centres and rearrangement of points — is run until no one point was replaced from one class to another.

If all the classes obtained differ each other, the classification is done; otherwise, two closest classes should be merged into one, and the entire procedure must be run again. Two classes are presumed to differ each other, if the distance between their centres exceeds a maximal average radius of the classes to be distinguished. The average radius of the k-th class is defined as

$$R^k = \frac{1}{l_k}\sum_i d_i^k, \tag{7}$$

with d_i^k determined according to (6), and i runs all the values so that the point belongs to the k-th class.

A number of classes to be observed due to a classification implementation is unknown *á priori*. Initially, the set of points should be split into sufficiently big number of classes. Due to the consecutive merges, the maximal possible number of classes occurs, that still satisfy the separation condition. The algorithm implemented here is absolutely similar to the cluster analysis provided by Cohonen neural networks [15].

Results and discussion

We have studied the bacterial 16S RNA sequences, their total number was 1730 [17]. Table 1 shows the taxonomy composition of the set of nucleotide sequences studied. It is evident, the taxae are rather diverse, but inhomogeneous with respect to a number of sequences within the same taxa.

I. Transformed frequency dictionary

For each entity form the set of 16S RNA sequences the transformed frequency dictionary has been developed, of the thickness 3. The specific distribution density of the values of the transformed frequencies are shown in Fig. 1. Obviously, the total number of such distributions is 64, according to a number of possible triplets. The distribution density of the triplets *TAT, CCT, AAA,* and *GAA*

take the marginal positions (ultimate left, ultimate right, ultimate upper, and ultimate down, respectively), while the distribution density of the triplets *TTT, CAA,* and *AGC* are the most typical for the set of nucleotide sequences studied.

Now consider what is a distribution of the transformed frequencies within the dictionary (of the thickens 3) of the same taxa. To do so, we develop the frequency dictionary averaged over all sequences included in a line from Table 1. The density distribution of the transformed frequencies averaged over those 13 taxon groups is shown in Fig. 2. The numbers over the vertical axis show how many triplets from 64 have the transformed frequency from the range given. Each narrow bar corresponds to a single taxa. To compare with, the similar data are shown in Fig. 3 presenting the real frequencies of the same taxae. It is evident, the transformed dictionaries possess higher information value, due to a wider expanded distribution patter, on one hand, and since they present the clear differences among the taxae, on the other.

Family *Firmicutes*; *Actinomycetes* is the most abundant in the original set of sequences (335 entities). Table 2 shows the sites of the length 3 of high information value and their transformed frequencies averaged over that family. The table consists of two parts: the former presents the triplets with occurrence frequency higher than predicted $p_j > 1$, and the latter presents those with the frequency under predicted $p_j < 1$. Table 3 shows the sites of the length 4 with high information value. This table consists of three parts, and the first one and the second one are similar to those from the Table 2, the third part shows 4-tipples which do not occur in the real dictionary, while their expectancy is above zero ($f_j = 0;\ \widetilde{f}_j \neq 0$), in this Table *n* denoted the number of copies predicted of the given 4–tipple according to the entire set of entities analysed. A comparison of these two tables shows explicitly that the sites of high information value of various lengths may be incorporated into that type of longer sites, and may not be incorporated. And vice versa, the 4–symbol long sites with high information value may include the highly information valued triplets, and may not include. One can see a significant non-monotony in the succeed of the information valued sites, as the length of these latter grows up.

II. Classification

The set of 16S RNA has been split into the classes according to the dynamic kern method. The classification has been elaborated both over the real frequency dictionaries, and over the transformed dictionaries. The classification obtained differ, as it has been expected. Nevertheless, both classifications yield a reasonable correlation to the taxonomy classification of the gene bearer.

The classifications were developed over the dictionaries of the thickness 3, since a reliable method implication is possible only when the number of the objects to be classified exceeds

significantly the dimension of the space. Besides, the dictionary of the thickness 3 represents more structural entities of a nucleotide sequence (in comparison to the dictionaries of the thickness 2 and, moreover, to the thickness 1); at least one structure is presented completely at a dictionary of the thickness 3, that is the genetic code structure.

The classification of the set of 16S RNA sequences developed over the real frequency dictionaries is shown in Fig. 4. The original set of sequences split into two classes. The vertical axis on the diagram presents the taxae, and the horizontal axis presents the number of sequences from the given taxon that belong to a class. A separation of each taxon group into these two classes is shown in grey and black colour. A non-randomness of that split is evident. Moreover, inspite of the sequences of some genes occupy both classes, a significant irregularity of their distribution among two classes is obvious. A correlation between a statistical structure of nucleotide sequence and taxonomy of its bearer is evident. Nonetheless, a development of a similar diagram for the higher taxon level results in a significant growth of a equity of an occupation of both classes by the genes from the same taxon group. This effect may follow from a well-known fact that the taxonomy of higher taxonomic levels of prokaryote seems to be rather artificial [16].

The nucleotide sequences of some organisms, and prokaryotes especially, yield a significant preference in the occurrence of some peculiar nucleotides that effects quite poorly on the enzyme structure encoded by those nucleotide sequences. Then the correlation between the classification implemented over the real frequency dictionaries and taxonomy follows from a diversity of nucleotide composition of those sequences. Here the classification over the transformed frequency dictionaries seems to be more fruitful, both from the point of view of a methodology of a detection and isolation of the structures in nucleotide sequences, and the classification *per se* of the specific group of sequences to be studied.

An hierarchic classification of the original set of 16S RNA studied implemented due to the transformed dictionaries is shown in Fig. 5. One can obviously see, that some specific taxonomy units are separated on the each level of the classification. Inspite of rather moderate number of the units separated, they contain almost all sequences of this peculiar taxonomy group, from the original set of 16S RNA sequences.

Let us consider the results of a classification implementation over the transformed dictionaries, on each hierarchy level, in more detail, drawing special attention to the features that make difference between the sequences from different classes, and to what extent. As one can see in Fig. 5, the original set of sequences has split into three classes, on the first level of classification: (I) — *Chloroflexaceae/Deinococcaceae group*; *Deinococcaceae;* (II) — *Spirochaetales*;

Spirochaetaceae; *Borrelia*; (III) — *All* 1 (all the other sequences). Fig. 6 shows what specific triplets determine the difference between the classes. The values of the transformed frequencies are shown on vertical axis of that chart; dashed line represents the group (I), dotted line represents the group (II), and solid line represents the group (III). The horizontal axis shows the numbers of the triplets ordered in the growth value of the transformed frequency within the group (III). All the triplets are enumerated in the lexicographic order. The slight differences between the groups are diminished. Table 4 corresponds to Fig. 6. It contains the coordinates of classes (and the triplets relevant to them) that yield the maximal difference between the classes. The transformed frequency dictionary represents an impact of non-randomness in the distribution of nucleotides within sequences, that is why we have also calculated the values of the centre coordinates determined for the real frequency dictionaries, as well. Table 5 shows these data for the same triplets. A comparison of these two tables makes it clear that the main factors of the difference between the classes may be the same for real and transformed dictionaries, and they may differ significantly. There is no evident correlation of the classification implemented over the real dictionary vs. that latter implemented over the transformed ones.

A classification implemented over the transformed dictionaries yielded a rather abundant group of sequences *All* 1 (see Fig. 5); this allowed to develop the next level of the classification and the group has been split into several classes, in turn. It should be said that the classes obtained failed to satisfy the separation condition. Nevertheless, a separation of this group into two classes the family *Firmicutes; Actinomycetes* becomes isolated, that is rather interesting itself. An absence of a split satisfying the separation condition could result from several reasons. Probably, the most important for that is the taxonomy diversity of the nucleotide sequences observed within this group that makes a variation of information characteristics rather smooth, thus hiding a separation onto the explicit classes.

The third level of classification yielded a split of the group of nucleotide sequences into two classes, with the family *Fibrobacteria* concentrated in one of them. Table 6 (similar to the table 4) presents the main factors of the difference of the structures. The Table 7 (similar to the table 5) shows the coordinates of the centre of those classes calculated for the real dictionaries, in order to provide a comparison.

The classes obtained at the fourth level of the classification failed to satisfy the separation condition, and we show them just to finalise the classification implemented. The split into three classes here is optimal, in the following sense: the best relation between the average radii of the

classes obtained is achieved at the case of a splitting into three classes. Besides, it is rather interesting to trace the sequences of some taxonomy group into the same class.

Conclusion

We have studied an interrelation between a structure of nucleotide sequence and a taxonomy of its bearer. An extended group of 16S RNA has been studied to answer this question. A proximity of structures was understood as a proximity of frequency dictionaries, either real or transformed ones, in Euclidean metrics. From the point of view of the molecular aspects of the selection theory, the most imprint thing here is that the classification implemented over the real frequency dictionaries of the thickness 3 correlates best of all to the generae. A gender is included entirely either into the first class, or into the second one, and the exclusions are rarely met. An association of nucleotide sequences into the taxonomy groups of family range, or higher results in significant decay of the correlation of the taxonomy and the classification implemented over the statistical properties of the relevant sequences.

A transformation of the frequency dictionary of a nucleotide sequence, i.e. a usage of the reconstructed frequency dictionary in order to outfit a non-random component in a distribution of *k*–tipples, allows to compare the nucleotide sequences over their information characteristics. Automatic classification of a set of nucleotide sequences over their transformed dictionaries yields the classes of proximal sequences. In the case studied, where 16S RNA were studied the classes obtained contain the sequences of specific taxonomy.

A classification of sequences developed over the real frequency dictionaries differs basically from that latter developed over the transformed ones. A classification over the real frequency dictionaries represent mainly the difference in nucleotide composition of the sequences. A decomposition of the original set of sequences into two classes with a good correlation between a class occupation and the taxonomy of the bearer of a sequence proves this idea clearly. A classification over the transformed dictionaries isolates one or two groups of sequences of the same taxonomy, on each level of the classification. The difference in structure among the classes obtained manifests itself in a seldom or, contrary, frequent (in comparison to the expected one) occurrence of some words within a sequence; a number of those words is not too great (see Tables 4 to 7). The sites determined according to their information characteristics do not obligatory coincide with the other structure entities determined by another methods [10, 11]. It is rather important to carry out a comparative study of the sites determined by the information characteristics with those determined by the other methods.

References

1. Sharp Ph. A. Split genes and RNA splicing // Cell, 1994. V.**77**, № 6. pp. 805 – 815.

2. Yockey H.P. Information Theory and Molecular Biology. Cambridge Univ.Press, N.Y. 1992.

3. Gorban A.N., Mirkes E.M., Popova T.G., Sadovsky M.G. A new approach to study the statistical properties of genetic sequences // Biofizika, 1993. Vol. **38**, pp.762 – 767. (in Russian)

4. Gorban A.N., Mirkes E.M., Popova T.G., Sadovsky M.G. Comparative redundancy of genes of some organisms and their viruses // Genetika, 1994. Vol. **29,** iss. 11, pp. 1314 – 1319. (in Russian)

5. Gorban A.N., Popova T.G., Sadovsky M.G. Redundancy of genetic texts and mosaic pattern of a genome // Molekulyarnaya biologiya, 1994. Vol. **28**, iss. 3, pp. 313 – 322. (in Russian)

6. Gusev V.D., Kulichkova V.A., Titkova T.N. Analysis of genetic texts. I. *l*–tipple characteristics // Empirical prediction of images (Comput. systems, iss. 83). Novosibirs: Inst.of math. of SD of Acad.Sci.USSR, 1980. pp.11 – 33. (in Russian)

7. Bugaenko N.N., Gorban A.N., Sadovsky M.G. Towards the determination of the information content of nuclotide sequences // Molekulyarnaya biologiya, 1996. vol. **30**, iss. 3. pp. 529 – 541. (in Russian)

8. Bugaenko N.N., Gorban A.N., Sadovsky M.G. Maximum entropy method in analysis of genetic text and measurement of its information content // Open System & Information Dynamics, 1998. V.5, № 3. pp.265 – 278.

9. Gorban A.N., Popova T.G., Sadovsky M.G. Automatic classification of nucleotide sequences and its relation to natural taxonomy and protein function // Proc. of 1st Int. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Aug., 24 – 27, 1998. V. **II**. PP. 314 – 317.

10. Gelfand M.S. Prediction of function in DNA sequence analysis // J. Comput. Biol. 1995. V. **2.** pp. 87 – 115.

11. Claverie J.-M., Sauvaget I., Bougueleret L. *k*-tuple frequency analysis: from intron/exon discrimination to T-cell epitope mapping. // Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences / Ed.by Doolittle R.F. (Meth. Enzymol. vol. **183**), 1990. pp. 252 – 281.

12. Baum E.B., Boneh D. Running Dynamic Programming Algorithms on a DNA Computer // DIMACS Ser. in Discrete Math. and Theor. Computer Science, 1999. V.44, AMS Press, pp.77 – 85.

13. Kirkwood, J. and Boggs, E.: The radial distribution function in liquids // J.Chem.Phys., 1942, **v**. 10, № 6б p.394.

14. Bork P. Go hunting in sequence databases but watch out for the traps location // Trends Genet., 1996. V. 12. pp. 425 – 427.

15. Gorban A.N., Rossiev D.F. Neural networks on PC. Novosibirsk: Nauka Pbls, 1996. 275 p. (in Russian)

16. Schlegel H.G. Allgemeine Mikrobiologie. 6 uberarbeiten Auflage, 1985. Georg Thieme Verlag, Stuttgardt, N.Y., 652 p.

17. ftp://ccrv.obs-vlfr.fr/pub/christen/16S/

Table 1

Taxonomy composition of the nucleotide sequences of 16S RNA studied

N	Taxa	Number of NS
1	Chloroflexaceae/Deinococcaceae group.	29
2	Cyanobacteria	9
3	Cytophagales	117
4	Fibrobacter	13
5	Firmicutes; Actinomycetes	335
6	Firmicutes; Low G+C gramm-positive bacteria	485
7	Proteobacteria; α subdivision	262
8	Proteobacteria; β subdivision	63
9	Proteobacteria; δ subdivision	47
10	Proteobacteria; ε subdivision	43
11	Proteobacteria; γ subdivision	216
12	Spirochaetales; Leptospiraceae	14
13	Spirochaetales; Spirochaetaceae	35
14	OTHERS	56

Table 2

The sites of the length 3 with high information value, averaged over *Firmicutes; Actinomycetes* taxae

Real frequency is greater than expected			Real frequency is lower than expected	
Triplet	$f/\tilde{f}$		Триплет	$f/\tilde{f}$
CCT	1,355		CCA	0,641
AGC	1,338		TAT	0,67
TAA	1,302		AAA	0,709
CTT	1,282		TAG	0,762
TCA	1,194		TCT	0,771
TAC	1,189		TTT	0,784
GAT	1,177		GAC	0,81

Table 3

The sites of the length 4 with high information value, averaged over *Firmicutes; Actinomycetes* taxae

Real frequency is greater than expected			Real frequency is lower than expected			Zero real frequency, with non-zero expected	
4-tipple	$f/\tilde{f}$		4- tipple	$f/\tilde{f}$		4- tipple	*n*
ATAT	2,035		CAAA	0,6		GTAT	487
TATC	1,918		TTCA	0,595		CATA	346
TTGT	1,874		GTTA	0,587		ATTT	321
ATAC	1,838		CACT	0,568		ACTT	293
ACTC	1,833		TAAG	0,561		TATA	273
ACAC	1,757		AGAG	0,56		ATAG	271
TTAT	1,729		TACT	0,555		TTCA	227
ATTA	1,698		TTAC	0,55		TTTA	138
CACA	1,678		GACA	0,516		CCTC	129
GCGA	1,66		ACAG	0,505		ATAA	125
AGTC	1,657		TGAT	0,483		AAAT	124
CATG	1,653		CGAG	0,482		CCAT	101
ATCA	1,633		AAAT	0,461		ATCT	92
CGAA	1,629		TTTA	0,452		TACT	89
CGCA	1,579		TTCT	0,43		TCTA	66
AGAT	1,565		CCTC	0,422		CCGA	66
TTCC	1,559		GGCA	0,418		TCTT	57
TCCA	1,505		CCGA	0,385		TGAT	55

Table 4

The main factors of a difference between three classes

	TAT	CAT	GTC	TCT	TTG	ATA	TCG	TTA	CCA	ATC
Chloroflexaceae/Deinococcaceae gr.; Deinococcaceae	**0,109**	**1,296**	0,998	1,06	**0,71**	**0,686**	0,916	**1,418**	0,952	1,028
Spirochaetales; Spirochaetaceae; Borrelia	0,978	**0,746**	**1,42**	**1,189**	0,958	1,099	**0,788**	1,224	**0,545**	**0,737**
All 1	**0,6**	0,995	1,019	**0,819**	0,99	**0,859**	1,088	1,14	0,808	1,019

Table 5

Real frequencies corresponded to the main factors of a difference

	TAT	CAT	GTC	TCT	TTG	ATA	TCG	TTA	CCA	ATC
Chloroflexaceae/Deinococcaceae gr.; Deinococcaceae	**0,0008**	0,0087	0,0132	0,0073	**0,0075**	**0,0047**	0,0110	0,0099	0,0134	0,0072
Spirochaetales; Spirochaetaceae; Borrelia	**0,0173**	0,0077	0,0173	**0,0125**	0,0141	**0,0237**	0,0090	0,0187	**0,0052**	0,0081
All 1	0,0059	0,0102	0,0140	0,0078	0,0144	0,0099	0,0140	0,0117	0,0102	0,0098

Table 6

The main factors of a difference between two classes

	CAA	TAA	CTA	CAT	CTG	TTC	GTA	TCT	TAC
All 3	0,973	1,23	0,945	1,001	0,998	0,948	1,048	0,836	1,144
Fibrobacteria	1,337	0,898	0,67	0,756	1,229	1,169	1,268	0,617	1,362

Table 7

Real frequencies corresponded to the main factors of a difference

	CAA	TAA	CTA	CAT	CTG	TTC	GTA	TCT	TAC
All 3	0,0153	0,0182	0,0121	0,0106	0,0173	0,0082	0,0173	0,0080	0,0139
Fibrobacteria	0,0221	0,01	0,0061	0,0091	0,0183	0,0096	0,018	0,0045	0,0129