Open Sys. & Information Dyn. **7**: 1–17, 2000

Classification of Symbol Sequences over Their Frequency Dictionaries: Towards the Connection Between Structure and Natural Taxonomy

A. N. Gorban[1], T. G. Popova[1,2] and M. G. Sadovsky[1,2]

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

[2] *Institute of Biophysics of SD of RAS, Krasnoyarsk 660036*

(Received: December 6, 1999)

Abstract. The classifications of bacterial 16S RNA sequences developed over the real and transformed frequency dictionaries have been studied. Two sequences are considered to be close, when their frequency dictionaries are 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 real frequency dictionaries vs. the transformed ones has been carried out. A correlation between information patterns of nucleotide sequences and taxonomy of the bearer of the sequence was found. The sites with high information value are found to be the main factors of the difference between the classes in a classification. The classification of nucleotide sequences developed over real frequency dictionaries of 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 exclusions occur rarely. A hierarchical classification yields one or two taxonomy groups on each level of classification. An unexpectedly often, 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 to large. Further investigations are necessary in order to compare the sites revealed with those determined due to other methodology.

1. Introduction

A study of the relation between the structure of symbol sequences (S) and their meaning encrypted in the interlocation of symbols is a key problem for molecular biology, biophysics and many other fields of science. 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, one often talks 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 its frequency dictionary, either a real one [3–6], or a reconstructed one [7, 8], or a 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 to each other, when their frequency dictionaries are close. The real frequency dictionary W_q (of

thickness q) is defined as the list of all strings of length q occurring in the given sequence accompanied by the frequency of their occurrence [3 – 6]. It has been shown that sufficiently large family of nucleotide sequences can be classified into groups, according to the closeness of their frequency dictionaries. All the sequences within a group are close with respect to the Euclidean distance between their frequency dictionaries. We have observed a correlation between the 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 frequency dictionaries $W_1, W_2, \ldots, W_q$, corresponding to the same text yields a relation, namely all the foregoing dictionaries can be obtained from the succeeding ones by a simple summation. In other words, a thinner dictionary could always be obtained from a thicker one. An inverse statement does not hold true. An exact reconstruction of a thicker dictionary from a given one does not always exist, in general. The exact reconstruction of W_k over W_q for $k > q$ is possible if and only if all the words in W_q have unique continuation. Otherwise, the single-valued reconstruction is impossible: each frequency dictionary W_q of smaller thickness yields a set of thicker frequency dictionaries. One could nevertheless seek for the most probable continuation $\widetilde{W}_k$ of a given dictionary of smaller thickness W_q. Namely, to get the most probable dictionary $\widetilde{W}_k(q)$ one must select from all possible continuations of the given dictionary W_q the one with the least determinacy, i.e. the one which yields the maximal entropy. The exact solution of this extreme problem resembles the well-known Kirkwood approximation in statistical physics. If a reconstructed frequency dictionary $\widetilde{W}_k(q)$ coincides exactly with the original one W_k, then it means that the entire information carried by the original text is contained in the dictionary W_q. Differences between the reconstructed dictionary $\widetilde{W}_k(q)$ and the real one W_k show additional information contributed by k-tuples, in comparison to q-tuples, $k > q$.

From that point of view, the difference between the real frequency dictionary W_k and the most probable continuation $\widetilde{W}_k(q)$ are of the greatest interest. To measure these differences, one introduces a new object, so-called *transformed dictionary*, where each word is assigned the ratio of its real frequency and the frequency obtained in a reconstruction from the thinner one. Such a value shows how much the real frequency of a word differs from the expected one. This transformation of the frequency dictionary allows one to explain some peculiarities of the information structure of nucleotide sequences (NS).

The study of correlations between the structure and the function of nucleotide sequence requires the determination of some other relations among them (e.g., a classification). A set of genes always allows at least two independent classifications of such a type: one over the taxonomy of gene bearer, and another over the function of these genes. This paper is aimed to study the relation between the structure of NS (that is assumed to be a frequency dictionary, either real one, or transformed),

and the taxonomy of the gene bearer. Thus, only the pair *structure vs. taxonomy* is studied here, rather than the 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 for our investigation those nucleotide sequence which determine the same function in various organisms.

Currently, a huge amount of genetic data is available that allows to carry out a comparative investigation described above. We have used 16S RNA of bacteria of various species. All sequences of this type realise the same funetian, not only in bacteria. but in other organisms with higher taxonomy position.

2. Frequency Dictionary

To each sequenced gene there corresponds a symbol sequence of the same length (i.e., number of nucleotides) N, a genetic text over a four-letter alphabet. Any continuous subsequence of length q of the genetic text is called a word. We assign to each word its frequency, that is the number of its copies within the genetic text divided by the total number of words within the text; such list of all q-letter words occurring within the text together with their frequencies is called the frequency dictionary W_q [3–8].

If n is the cardinality of the alphabet, then total number of words of length q is n^q. Obviously, not every word of length q is likely to occur in a text, especially when q is large enough. Let us complete the frequency dictionary of a given text to the entire one (i.e., the one which contains all the words of length q) adding the words with zero frequency. Then every frequency dictionary can be represented as a point $F(f_1, f_2, \ldots, f_{n^q})$ in a n^q-dimensional space with the coordinates representing the frequencies of the corresponding words, $0 \leq f_j \leq 1$, $j = 1, 2, \ldots, n^q$. Then, a set of genes yields a set of points in n^q-dimensional space, according to such representation.

3. Transformation of Frequency Dictionary

Let us consider a set of frequency dictionaries (of various thickness) corresponding to the same genetic text: $W_1, W_2, \ldots, W_q, \ldots, W_N$. Then the question arises what part of the information contained in the original text is represented by the dictionary of thickness q. It is well-known that for some specific thickness d^*, all the words within a dictionary occur in a single copy [3, 4]. Hence, for $q > d^*$ all words in W_q have a unique continuation, and any dictionary W_k (including the original text) can be unambiguously reconstructed from W_q, as q becomes greater than d^* [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 a given one W_q by summation, and a part of the information about the text is lost when $q < d^*$. It makes an

inverse transformation (from a 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 $k > q$, and any of them may 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 us call such a dictionary the reconstructed one. The method of reconstruction of a dictionary is based on the maximum entropy principle [7, 8].

The entropy of a frequency dictionary W_q is defined as

$$S_q = -\sum_{j=1}^{n^q} f_j \ln f_j \,. \tag{1}$$

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 unique solution

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

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

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

The formulae (2) and (3) coincide with well-known expressions for transition probabilities in a symbol sequence obtained as a realization of a 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 are independent of any peculiar structure of a sequence. These formulae present the most likelihood hypothesis on the frequency dictionary of thickness $q + s$, resulting via reconstruction from a dictionary of thickness q. One should consider the original symbol sequence to be Markovian if and only if the expressions for real (but not for the reconstructed ones) frequencies would be valid in the limiting case of an infinitely long original sequence.

Let us now consider dictionaries reconstructed from a one symbol thinner dictionary. The formulae for the reconstructed frequencies of $\widetilde{W}_q(q-1)$ are:

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

The reconstructed frequencies will be denoted by $\tilde{f}$.

The dictionary $\widetilde{W}_q$ 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 one to mark explicitly the peculiarities of the information structure of nucleotide sequences, since the maximal differences between the real and reconstructed frequencies, for the dictionaries of a given thickness, are the most "unexpected" events in a transition from the dictionary of thickness $q-1$ to the one of thickness q.

Let us transform the frequency dictionary of a nucleotide sequence in the following manner: for each length q of words, starting from $q=2$, build the dictionary $\tilde{W}_q(q-1)$ reconstructed from the preceding one. As before, each nucleotide sequence is represented with a point $P(p_1, p_2, \ldots, p_{n^q})$ in a n^q-dimensional space, where the coordinates of the point are the ratios of the real and the reconstructed frequencies: $p_j = f_j/\tilde{f}_j$, for $\tilde{f}_j \neq 0$, and $p_j = 1$ for $\tilde{f}_j = 0$, $j = 1, 2, \ldots, n^q$. The values p_j show how much the real frequencies differ from the expected ones. If $p_j \approx 1$ for some word within the genetic text, then the information value of this word is not high: its most probable expected frequency almost coincides to the real one. The words whose frequency ratios p_j differ from the real one significantly present the most valuable sites of a given length within the nucleotide sequence studied. We take the threshold value to be 15–20% in our study. It should be stressed that the length of a site is rather essential, since any low-valued site of length q may be incorporated into a high-valued site of length $q+1$, which in turn may be incorporated into a longer site of 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 this value from 1 (that latter might occur simultaneously). The difference mentioned above becomes significant when it reaches the level of 15 to 20% as it will be shown later on.

4. Algorithms of Automatic Classification

The implementation of classification of objects requires the definition of a proximity measure among them. For symbol sequences, such a measure can be introduced in several different ways [2, 14]. Here we have used the following one:

Given the length q of words, any genetic texts can be represented as a point in n^q-dimensional space, corresponding to the frequency dictionary of thickness q (either real or the transformed one). Two sequences would be considered close, if the corresponding points in n^q-dimensional space are close. The distance between two points in this space is taken to be the Euclidean metrics. The study of the distribution of points in n^q-dimensional space allows one to split the original set of sequences into a number of classes, where the sequences are close to each other within a class.

Automatic classification algorithms have been used to perform the task mentioned above. We have used a dynamic kern method in our investigation [12, 15]. In brief, it looks as follows. Consider a set $\{F^i\}$, $i = 1, 2, \ldots, M$ of M points that should be split into several classes. Let the initial number of classes and the initial distribution over these classes be given. First, for each class k, the centre

$C^k(c_1^k, \ldots, c_{n^q}^k)$ is calculated according to

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

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

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

is calculated and the classification of each point is re-determined. A point is assumed to belong to the class which yields the least distance (6). As soon as all the paints are processed, the group centers are re-calculated. This procedure — the calculation of centers and the rearrangement of points — is run until no point is moved further from one class to another.

If all the classes obtained are disjoint, the classification is done; otherwise, two closest classes can be merged into one, and the entire procedure must be run again. Two classes are presumed to be different if the distance between their centers exceeds the 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 the indices of points belonging to the k-th class.

The number of classes to be distinguished by a classification procedure is unknown *a priori*. Initially, the set of points should be split into sufficiently large number of classes. Due to the consecutive merges, the maximal possible number of classes is found, that still satisfy the separation condition. The algorithm implemented here is entirely similar to the cluster analysis provided by Cohonen neural networks [15].

5. Results and Discussion

We have studied 1730 different bacterial 16S RNA sequences [17]. Tab. 1 shows the taxonomy composition of the set fo nucleotide sequences considered. It is evident, that the taxae are rather diverse but inhomogeneous with respect to the number of sequences within the same taxa.

5.1. Transformed frequency dictionary

For each entity from the set of 16S RNA sequences the transformed frequency dictionary of thickness 3 has been obtained. The specific distribution densities of

N	Taxa	Number of NS
1	Chloflexaccae/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	Spirochactales; Spirochaetaceae	35
14	Others	56

Tab. 1. Taxonomy composition of the nucleotide sequences of 16S RNA studied.

the values of the transformed frequencies are shown in Fig. 1. Obviously, the total number of such distributions is 64, according to the number of possible triplets. The distribution density of the triplets *TAT*, *CCT*, *AAA*, and *GAA* take the marginal positions (extreme left, extreme right, extreme upper, and extreme down, respectively), while the distribution densities of the triplets *TTT*, *CAA*, and *AGC* are the most typical for the set of nucleotide sequences studied.

Now consider what is the distribution of the transformed frequencies within the dictionary (of thickness 3) of the same taxa. To do so, we develop the frequency dictionary averaged over all sequences listed in Tab. 1. The density distribution of the transformed frequencies averaged over these 13 taxon groups is shown in Fig. 2. The numbers on the vertical axis show how many triplets from 64 have the transformed frequency in the given range. 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 that the transformed dictionaries possess higher information value due to a wider expanded distribution pattern on the one hand, and due to the fact that they present obvious differences among the taxae, on the other.

Family *Firmicutes Actinomycetes* is the most abundant in the original set of sequences (335 entities). Tab. 2 shows the sites of length 3 of high information value and their transformed frequencies averaged over the family. The table consists of two parts: the leftmost one presents the triplets with occurrence frequency higher than predicted $p_j > 1$, and the rightmost one presents those with the frequency

Real frequency is greater than expected		Real frequency is lower than expected	
Triplet	$f/\tilde{f}$	Triplet	$f/\tilde{f}$
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

Tab. 2. The sites of length 3 with high information value, averaged over *Firmicutes; Actinomycetes* taxae.

Real frequency is greater than expected		Real frequency is lower than expected		Zero frequency, with non-zero expected	
4-tupple	$f/\tilde{f}$	4-tuple	$f/\tilde{f}$	4-tuple	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

Tab. 3. The sites of length 4 with high information value, averaged over *Firmicutes; Actinomycetes* taxae.

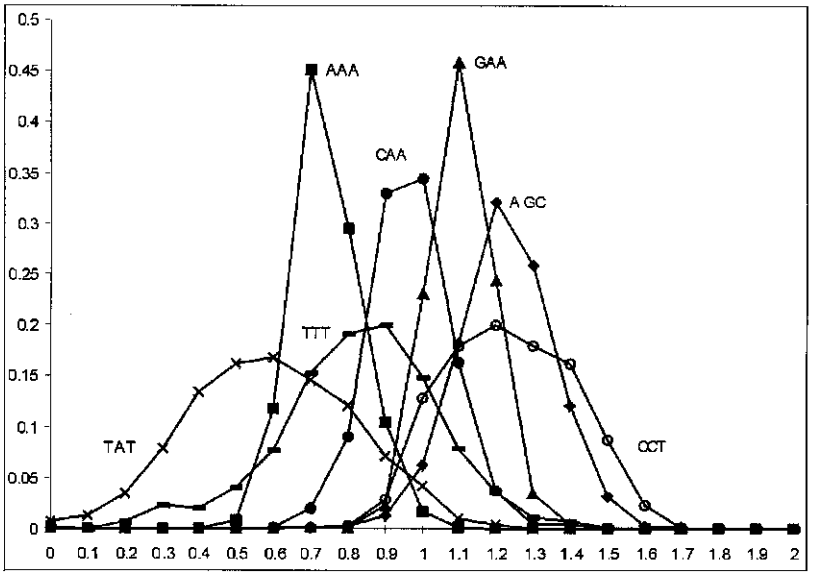

Fig. 1. The densities of the transformed frequencies of triplets. Vertical axis shows the value of the transformed frequencies and the horizontal one shows the density of the latter for a given triplet obtained for the entire set of all nucleotide sequences under investigation. The charts for TAT, AAA, GAA and CCT triplets take the marginal positions, while the charts for TTT, CAA, and AGC triplets seems to be the most typical for the set of sequences studied.

under the predicted $p_j < 1$. Tab. 3 shows the sites of length 4 with high information value. This table consists of three parts, and the first one and the second one are similar to those of Tab. 2, while the third part shows 4-tuples which do not occur in the real dictionary, while their expectancy is above zero ($f_j = 0$; $\tilde{f}_j \neq 0$). In this Table, n denotes the number of copies predicted for a given 4-tuple 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 or may not be incorporated into a similar type of longer sites. And vice versa: the 4-symbol long sites with high information value may or may not include highly information-valued triplets. One can observe a significant non-monotonicity in the succession of information valued sites, as their length grows up.

5.2. Classification

The set of 165 RNA has been split into classes with the help of the dynamic kern method. The classification has been elaborated both over the real frequency dictionaries, and over the transformed dictionaries. The classifications obtained

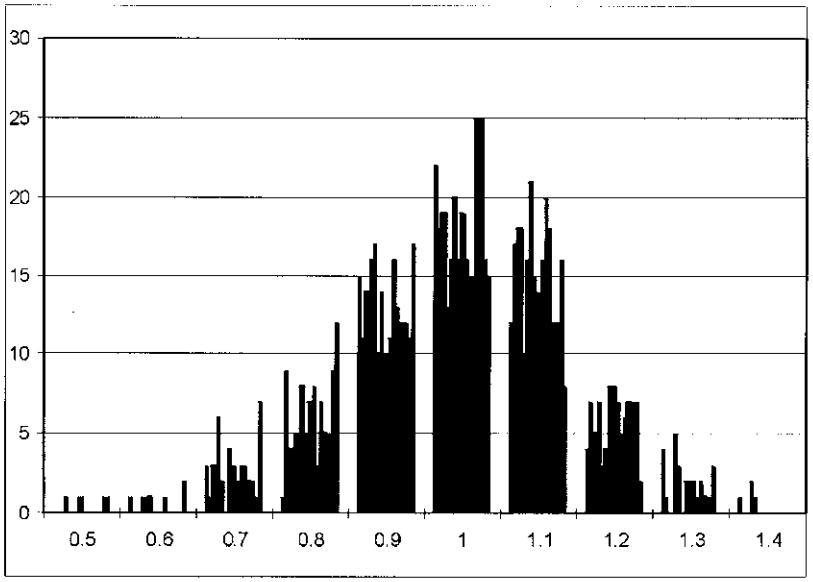

Fig. 2. The density of the transformed frequencies. The horizontal axis shows the value of the transformed frequency, and the vertical one shows a number of triplets (of 64 totally available) with the given transformed frequency. Each narrow bar in the chart corresponds to the frequency dictionary averaged over the entire body of sequences within one of 13 isolated taxonomy groups (as of Tab. 1).

differ, as it has been expected. Nevertheless, both classifications yield a reasonable correlation with the taxonomy classification of the gene bearer.

The classifications were performed for the dictionaries of thickness 3, since a reliable method implication is possible only when the number of objects to be classified exceeds significantly the dimension of space. Besides, the dictionary of thickness 3 represents more structural entities of a nucleotide sequence (in comparison to the dictionaries of thickness 2 and also to thickness 1); at least one structure is presented completely in the dictionary of the thickness 3, that is the genetic code structure.

The classification of the set of 16S RNA sequences performed on the real frequency dictionaries is shown in Fig. 4. The original set of sequences is 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. Non-randomness of such a split is evident. Moreover, even though the sequences of some genders occupy both classes, a significant irregularity of their distribution among two classes is obvious. A correlation between the statistical structure of nucleotide sequence and the taxonomy of its bearer is evident.

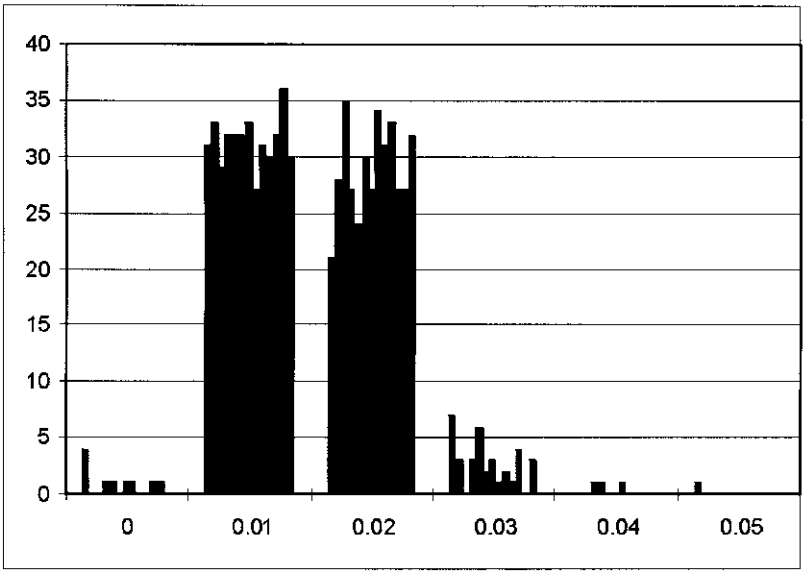

Fig. 3. The density of the real frequencies (cf. Fig. 2).

Nonetheless, a similar diagram for a higher taxon level shows significantly more uniformity in the occupation of both classes by the genes from the same taxon group. This effect may follow from the well-known fact that of higher taxanomic levels of prokaryote seems to be rather artificial [16].

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

A hierarchic classification of the original set of 16S RNA based on the transformed dictionaries is shown in Fig. 5. One can obviously see that some specific taxonomy units are separated on each level of the classification. In spite of a 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 performed on the transformed dictionaries on each hierarchy level in more detail, drawing special attention to the features that differentiate between the sequences from separate classes. As one

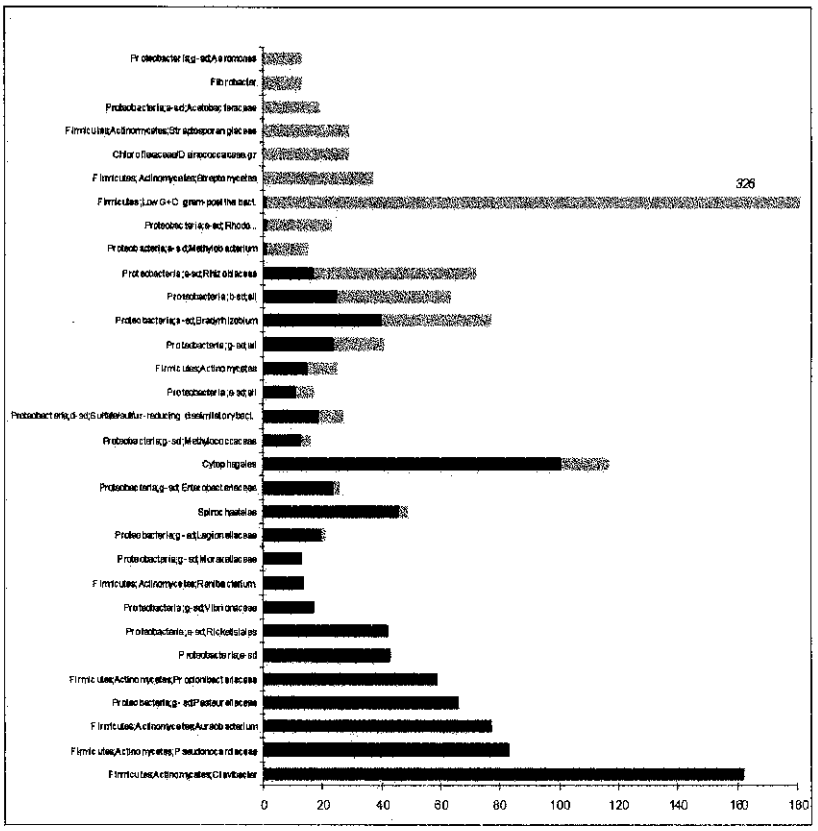

Fig. 4. The classification of the set of bacterial 16S RNA over the real frequency dictionaries of thickness 3. Horizontal axis shows the number of sequences, and the vertical one shows the taxonomy group. The number of sequences belonging to the first class is shown in black, while number of the sequences belonging to the second class is shown in grey.

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 which specific triplets determine the difference between the classes. The values of the transformed frequencies are shown on the vertical axis; dashed line represents the group (I), dotted line represents the group (II), and solid line

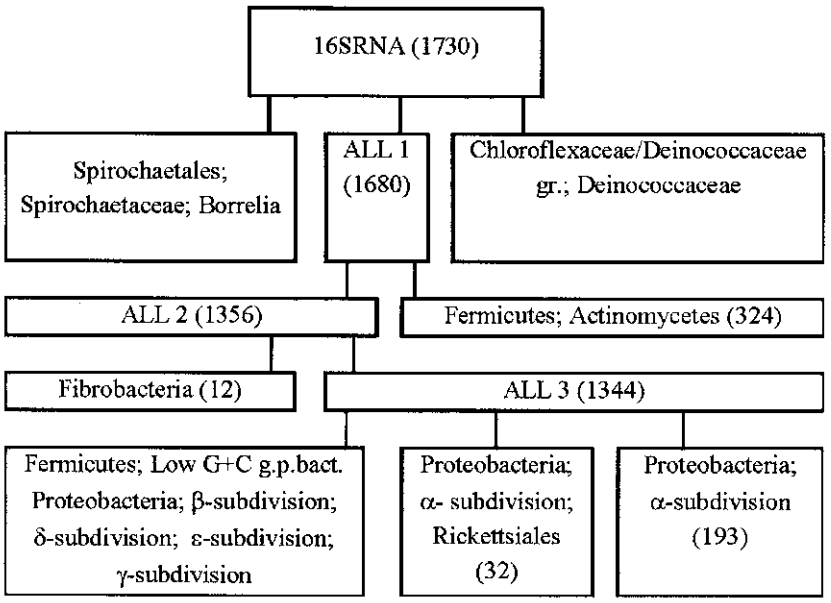

Fig. 5. Hierarchy classification of the set of bacterial 16S RNA over the transformed frequency dictionaries of thickness 3. The figure shows four levels of the classification.

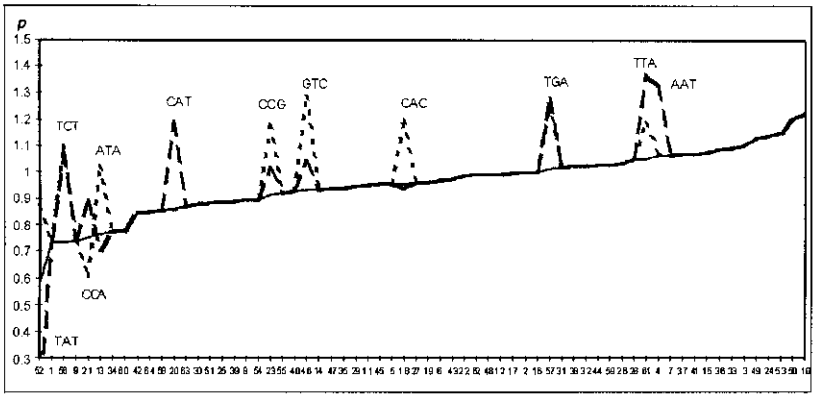

Fig. 6. The largest deviations of the transformed frequencies which determine the difference among three groups at the first level of classification shown in Fig. 5. The solid line corresponds to the group *All* 1, the dashed line corresponds to the group *Chloroflexaceae/Deinococcaceae; Deinococcaceae* and the dotted line corresponds to the group *5pirochaetales; Spirochaetaceae; Borrelia*. The horizontal axis shows the number of triplets enumerated lexicographically and ordered in the increasing value of the transformed frequency in the group *All* 1. The vertical axis shows the value of the transformed frequencies; small deviations are smoothed.

represents the group (III). The horizontal axis shows the number of 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. Tab. 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 the impact of non-randomness in the distribution of nucleotides within sequences. This is why we have also calculated the values of the centre coordinates determined for the real frequency dictionaries, as well. Tab. 5 shows these data for the same triplets. A comparison of these two tables makes it clear that the main factors underlying 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 based on the real dictionary vs. the latter performed on the transformed ones.

A classification performed on the transformed dictionaries yielded a rather abundant group of sequences *All* 1 (see Fig. 5); this allowed us to develop the next level of classification and the group has been in turn split into several classes. It should be said that the classes obtained failed to satisfy the separation condition. Nevertheless, in the separation of this group into two classes the family *Firmicutes Actinomycetes* becomes isolated, which is rather an interesting fact by itself. The absence of a split satisfying the separation condition can 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 the separation into 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. Tab. 6 (similar to Tab. 4) presents the main factors of the difference of the structures. Tab. 7 (similar to the Tab. 5) shows the coordinates of the centre of these 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 implemented classification. 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 in this case. Besides, it is rather interesting to trace the sequences of some taxonomy group into the same class.

6. Conclusion

We have studied relations between the structure of a nucleotide sequence and the taxonomy of its bearer. An extended group of 16S RNA has been studied to answer this question. The proximity of structures was understood as the proximity of frequency dictionaries, either real or transformed ones, in Euclidean metrics. From

	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

Tab. 4. The main factors behind the difference among three classes.

	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

Tab. 5 Real frequencies corresponding to the main factors determining the difference among 3 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

Tab. 6. The main factors behind the difference between two classes.

	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

Tab. 7. Real frequencies corresponding to the main factors determining the difference between 2 classes.

the point of view of the molecular aspects of the selection theory, the most visible thing here is that the classification performed on the real frequency dictionaries of 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 a 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. the usage of the reconstructed frequency dictionary in order to detect a non-random component in a distribution of k-tuples, allows one 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 of 16S RNA studied, the classes obtained contain the sequences of specific taxonomy.

A classification of sequences performed on the real frequency dictionaries differs basically from that performed on 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 the 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 rare or, contrary, frequent (in comparison to the expected one) occurrence of some words within a sequence; a number of these words is not too large (see Tables 4 to 7). The sites determined according to their information characteristics do not necessarily coincide with other structure entities determined by other 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 other methods.

Bibliography

1. Ph. A. Sharp, Cell **77**, No. 6, 805 (1994).
2. H. P. Vockey, *Information Theory and Molecular Biology*, Cambridge Univ. Press, N.Y., 1992.
3. A. N. Gorban, E. M. Mirkes, T. G. Popova, M. G. Sadovsky, Biofizika **38**, 762 (1993) (in Russian).
4. A. N. Gorban, E. M. Mirkes, T. G. Popova, M. G. Sadovsky, Genetika **29**, 1314 (1994) (in Russian).
5. A. N. Gorban, T. G. Popova, M. G. Sadovsky, Molekulyarnaya biologiya **28**, 313 (1994) (in Russian).
6. V. D. Gusev, V. A. Kulichkova, T. N. Titkova, *Empirical prediction of Images*, Comp. Systems **83**, Novosibirsk Inst. of Math. of SD of Acad. Sci. USSR, pp. 11–33, 1980 (in Russian).
7. N. N. Bugaenko, A. N. Gorban, M. G. Sadovsky, Molekulyarnaya biologiya **30**, 529 (1996) (in Russian).

8. N. N. Bugaenko, A. N. Gorban, M. G. Sadovsky, Open Sys. Information Dyn. **5**, 265 (1998).
9. A. N. Gorban, T. G. Popova, M. G. Sadovsky, *Proc. of First Int. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk, 1998*, vol. 2, pp. 314–317.
10. M. S. Gelfand, J. Comput. Biol. **2**, 87 (1995).
11. J.-M. Claverie, I. Sauvaget, L. Bougueleret, in: *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*, R. F. Doolittle, ed., (Meth. Enzymol. **183**), pp. 252–281, 1994.
12. E. B. Baum, D. Boneh, DIMACS Ser. in Discrete Math. and Theor. Computer Science **44**, 77 (1999).
13. J. Kirkwood and E. Boggs, J. Chem. Phys. **10**, 394 (1942).
14. P. Bork, Trends Genet. **12**, 425 (1996).
15. A. N. Gorban, D. F. Rossiev, *Neural networks on PC*, Novosibirsk, Nauka Pbls., 1996, (in Russian).
16. H. G. Schlegel, *Allgemeine Mikrobiologie. 6 überarbeiten Auflage*, Georg Thieme Verlag, Stuttgardt, N.Y., 1985.
17. ftp://ccrv.obs-vlfr.fr/pub/christen/16S/