A LARGE-SCALE COMPARISON OF GENOMIC SEQUENCES: ONE PROMISING APPROACH

Valery Kirzhner[1], Eviatar Nevo, Abraham Korol and Alexander Bolshoy

Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel. Email: [1]valery@esti.haifa.ac.il

Received 12-IX-2002

ABSTRACT

We introduce a novel, linguistic-like method of genome analysis. We propose a natural approach to characterizing genomic sequences based on occurrences of fixed length words from a predefined, sufficiently large set of words (strings over the alphabet {A, C, G, T}). A measure based on this approach is called compositional spectrum and is actually a histogram of imperfect word occurrences. Our results assert that the compositional spectrum is an overall characteristic of a long sequence i.e., a complete genome or an uninterrupted part of a chromosome. This attribute is manifested in the similarity of spectra obtained on different stretches of the same genome, and simultaneously in a broad range of dissimilarities between spectral representations of different genomes. High flexibility characterizes this approach due to imperfect matching and as a result sets of relatively long words can be considered. The proposed approach may have various applications in intra- and intergenomic sequence comparisons.

KEYWORDS: DNA linguistics; sequence analysis; statistical geometry; rank correlation.

1. INTRODUCTION

Methods of measurement of relatedness between two genetic sequences without performance of pairwise alignment have appeared since the early eighties. These methods are called linguistic because they are analogous to the formal linguistic analysis of human texts. There are methods based on calculations of observed occurrences (frequencies) of oligomers and their distribution, and those based on deviations between the observed and the expected occurrences (contrast words, genome signatures) in biological sequences as well. Brendel *et al.* (1986) introduced the concept of a meaningful word as an element of an organism-specific vocabulary in the DNA language. The authors tried to use an analogy to human languages in identifying words and compilation of vocabularies. As a method of comparison between two sequences it was proposed to construct the complete deviation sets of the sequences (vocabularies of contrast values) and determine a distance between these vectors instead of assessing distance based on sequence aligment. The measure of the deviation of the observed frequency of a word from its expected occurrence in the given sequence was introduced in Brendel *et al.* (1986) and slightly corrected in

Acta Biotheoretica **51**: 73–89, 2003.

Pevzner *et al*. (1989). Pietrokovski *et a l.* (1990) proposed usage of the linguistic similarity measure for fast and simple preliminary estimation of relatedness between two sequences. Definition of linguistic similarity was related to the notion of contrast words and vocabularies. The contrast value for each word could be calculated as the difference between its observed and expected frequencies in the given sequence. The contrast values of all the words of a given length k form the contrast k-vocabulary. The authors proposed to use a correlation coefficient C_k formula to compare two such vocabularies of the same word length k. Pietrokovski *et al*. (1990) claimed that usage of relatively small k ($2 \leq k \leq 6$) could satisfy all practical needs of a researcher. They proposed to use an integral value $S_{2\text{-}5} = (C_2 + C_3 + C_4 + C_5)$ for quantitative description of the linguistic similarity between sequences (Pietrokovski *et al*., 1990; Pietrokovski, 1994). Following this approach, Karlin introduced the measure of sequence similarity based on dinucleotide relative abundance extremes (Karlin and Burge, 1995), and widely used it for sequence taxonomy (Karlin, 1998; Karlin and Mrazek, 1997).

Recently, we proposed a natural approach to characterizing genomic sequences, based on imperfect occurrences of fixed length words from a sufficiently large set (Kirzhner *et al.*, 2000; Kirzhner *et al.*, 2002). Following this approach, any genomic sequence can be characterized by a histogram of frequencies of imperfect matching of fixed length words from the selected set that we have named compositional spectra (CS). In general, this approach presumes parameter adaptation to the concrete problem formulation and the targeted set(s) of species.

2. COMPOSITIONAL SPECTRUM OF DNA SEQUENCE

Definition

Let us take a set W of n different oligonucleotides (words) w_i of length L. Clearly, $n \leq 4^L$, where 4^L is the total number of different oligonucleotides of length L. For each word w_i of the set W and any chosen large target sequence S one can calculate the observed number of matches $m_i = m(w_i)$, allowing for a preset number r of replacements of symbols for each matching location, say 0, 1, or 2 mismatches (i.e., $r = 2$). This approximate matching can be denoted as '*r-mismatching*'. Now let $M = \sum m_i$. The frequency distribution $F(W; S)$ of $f_i = m_i/M$ will be referred to as '*compositional spectrum*' of the sequence S relative to the set W. To produce the word sets, we employed a random generator assuming equal probabilities of including each of the four nucleotides at any current position of any word. It is noteworthy that the foregoing approximate matching, or r-matching, allows usage of relatively long words (say, with $L = 10 - 15$) that otherwise would be impossible for even very large genomes.

The results obtained in this work are mainly concerned with CS for parameters $L = 10$, $n = 200$, $r = 2$. The number of different words when $L = 10$ and the alphabet contains four letters is equal to 1024×1024. When the error is $r = 2$, every word of the set W effectively corresponds to the number of different words of length 10, not greater than $1 + 10 + 10(10 - 1)/2 = 56$. Thus, the number of different words of length 10 that effectively corresponds to a set of words with $n = 200$ does not exceed

$56 \times 20 = 11,200$, i.e. not greater than 1% of all possible words. In much the same way we obtain $11 \times 200 = 2,200$ for $r = 1$ and $176 \times 200 = 35,200$ for $r = 3$.

For the word length $L = 10$ and $n = 200$, the characteristics of a sequence cover depending upon the value of r are presented in Table 1.

Table 1. Characteristics of a sequence cover depending upon the value of r for $L = 10$ and $n = 200$.

r	Min %	Max %	Mean %	75%
1	0.17	0.76	0.45	0.25 – 0.5
2	3.1	13.1	7.2	3.4 – 9.0
3	31.0	87.0	57.0	45.0 – 60.0

The first column of Table 1 contains the values of $r = 1,2,3$. In the second column, "Min %" means the minimum percent of cover by a random set of words for all sequences from Appendix A. In much the same way, "Max %" is determined as the maximum percent of cover for the same set of sequences, and "Mean %" is the mean value of cover for all sequences. However, for the majority (75% of the sequences analyzed), we may find a narrower cover interval (Min % – Max %), which is shown in the last column of the table.

Derivative spectra

Consider variations of the foregoing procedure to compare spectra for sets of direct and transformed words. For any chosen set W, one can produce a related set of reverse complementary words W^*: for example, if w_i = ATCCGACGGT then w_i^* = ACCGTCGGAT. Application of the above procedure with the set W^* to a sequence S will produce its own spectrum $F^*(W; S) = F(W^*;S)$. Other related spectra of the chosen set W, for example, $F^{**}(W;S) = F(W^{**};S)$, where W^{**} is a set of mirror sequences w_i^{**} (i.e., w_i^{**} = TGGCAGCCTA for w_i = ATCCGACGGT), exist as well.

Visualization of compositional spectra

The application of the proposed concept will be illustrated on a series of examples where CSs are employed for large-scale genome comparisons. Let S be a genomic sequence that we want to characterize relative to a given set of words W. We use different sets, composed of 200 words each, including a set of decamers (W_{10}), a reverse set of decamers W^*_{10} constructed of reverse complementary words to given W_{10} and a set of 15-mers (W_{15}).

A given order of words w_i in W predisposes a shape of the compositional spectrum of S relative to W. A compositional spectrum, which is a vector of frequency distribution, will be presented as a distribution plot, where the abscissa corresponds to the running index i of w_i, and the ordinate presents frequencies of w_i in S. For better visualization of multiple spectra related to a few sequences S_1, S_2, …, S_k, an order of words w_i in W is not chosen randomly, but instead, directly related to a descending order of f_i frequencies in a certain sequence S. Let us call such an order a "reference ranking" and denote it as Ord(W, S). This order of words w_i in W related to the

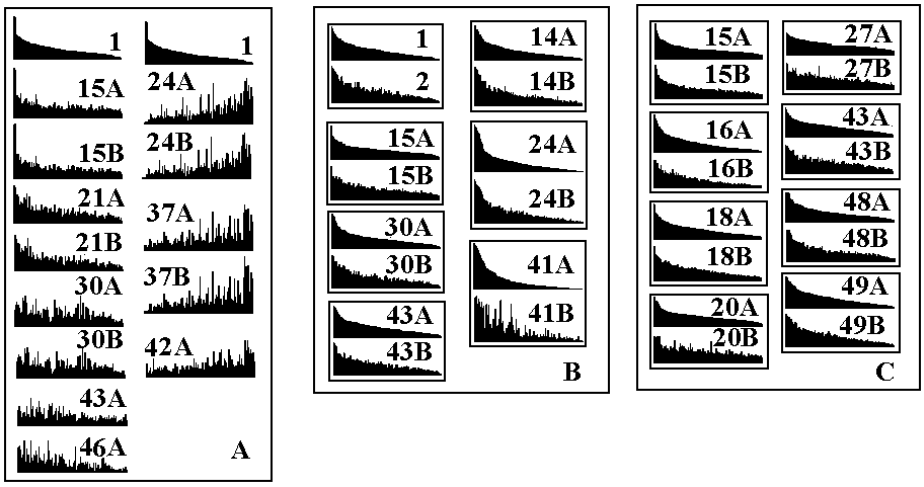


Figure 1. Compositional spectra of long stretches of genomic DNA of various species. The abscissa represents the words of the set *W* placed in some order, whereas the ordinate shows the observed frequencies $F(W, S)$ of the words in the sequence *S*. Compositional spectra (CS) of selected species are shown, where two distantly located fragments represent each organism, named *A* and *B*. For designations of the species see Appendix. (**A**). CS are ordered using as a reference sequence a contig of human *X* chromosome (No. 1). All spectra are calculated using the same set of words W_{10}. (**B**). Pairs of CS of the two DNA fragments of the same genome are shown. We ordered the set *W* in descending order of the observed number of matches in one of the genomic stretches. This order of words (denoted by *A* in each group of contigs) was used to present a spectrum of the second segment of the same genome (denoted by *B*). All compositional spectra are calculated using the same set of words W_{15}. (**C**). Comparison of spectra based on direct (*W*) and complementary words *W**. One stretch of a genome is represented by a spectrum $F(W_{10}, S)$, whereas another stretch is characterized by a spectrum $F(W^*_{10}, S)$. The order of complementary words strongly corresponds to the order of direct words.

sequence S_0 - notation $\mathrm{Ord}(W, S_0)$ - makes differences in spectrum $F(W, S_j)$ in comparison with $F(W, S_0)$ clearly observable (see compositional spectra of a few species in Figure 1). Figure 1A shows CSs of a few selected species *S* characterized by the set of decamers W_{10}. As a sequence producing the reference ranking we took the human *X*-chromosome. Compositional spectra of the sequences selected for Figure 1A present diversity of possible cases. For example, the spectra of *Drosophila melanogaster* (No. 15) and the bacteria *Streptococcus pyogenes* (No. 21) seem similar to a triangle histogram of *Homo sapiens* chr. *X* (No. 1). Similarity of spectra means that more frequent words in one sequence occur frequently in the other sequence as well. In other words, there is correlation between their compositional spectra. On the contrary, the spectra of the bacteria *Aquifex aeolicus* (No. 30) do not demonstrate any similarity to the reference spectrum no. 1 of *Homo sapiens*. It means that these sequences are unrelated. The bacteria *Mycobacterium tuberculosis* (No. 24) and *Pseudomonas aeruginosa* (No. 37) provide examples of spectra with anti-correlation

to the spectrum No. 1 of *Homo sapiens*. These examples show that there are all kinds of possible intergenomic relations. However, negative or close to zero correlations in intragenomic comparisons have not been observed (see Figure 1B). As a rule, the spectra of two stretches from the same genome look very similar. Even in the bacteria *Borrelia burgdorferi* (No. 41), where the spectra are considerably different, correlation between them is definitely positive. In the following section we will describe these observations in a formal manner.

A worthy observation is that when a set of words W is sufficiently large it equally characterizes both strands of DNA. Indeed, in Figure 1C one stretch of a genome is represented by a spectrum $F(W_{10}, S)$, whereas another stretch is characterized by a spectrum $F(W^*_{10}, S)$. It is interesting to note that corresponding spectra proved very similar. This effect seemingly may be explained for the prokaryotic genomes by apparently uniform distribution of genes on both strands but it is correct for Eukarya as well.

A measure of genome similarity

The intuitive impressions of intergenomic or intragenomic similarities and dissimilarities can be supported by distance metrics obtained, based on the Spearman rank correlation coefficients ρ and τ (Kendall, 1970). Both coefficients are based on the number of pairwise transpositions of words that transform one order into another in such a way that words of the same rank occupy identical positions in the two spectra. Consequently, quantities $d_1' = 1 - \rho$ and $d_2' = 1 - \tau$ $(0 \le d^* \le 2)$ can be considered as distances between two spectra. The maximal distance $d^* = 2$ corresponds to strictly reverse compositional spectra, whereas the minimal distance of zero corresponds to identically ordered spectra of the fragments. We use a notation CS-distance for this measure.

On the basis of sampling distribution of the Spearman correlation coefficient one can test different hypotheses regarding the similarity of compositional spectra, namely by means of approximate test statistics $d/\sqrt{V_d}$, where $V_d = V_R$ is the sampling variance of d (note that $V_R = 1/(n-1)$: see Kendall and Stuart (1967)). The finite length of the analyzed sequences S and a limited size of the source genome(s) cause the approximate nature of these statistics. Assuming a normal distribution of the test statistics, deviations of d from either 0 or 2 will be significant at the 0.05 level if d or 2 - d exceed $1.96/\sqrt{(n-1)}$, where n is the number of words in the set W. Thus, for $n = 200$, $d \le 0.139$ can be considered as a non-significant difference. Clearly, corrections for multiple comparisons should be made in the case where many pairs of species are considered simultaneously. For example, distances among different spectra presented in Figure 1 are the following. Intergenomic distances between a fragment of *Homo sapiens* (No. 1) and other fragments shown in Figure 1A calculated using the d_1 measure are: d_1(1,15A)=0.20; d_1(1,21A)=0.20; d_1(1,30A)=0.44; d_1(1,24A)=1.68; d_1(1,24A)=1.59. As expected, intragenomic distances are smaller: d_1(1,2)=0.03; d_1(14A,14B)=0.03; d_1(15A,15B)=0.04; d_1(24A,24B)=0.02; d_1(30A,30B)=0.02; d_1(41A,41B)=0.15; d_1(43A,43B)=0.09. In case of the mosaic genome of *Borrelia burgdorferi* (No. 41) we show all intragenomic calculations (Table 2). The complete genome *B. burgdorferi* that has a length of 910724 bp, was divided into eight

overlapping segments, each 200K in size: the first fragment is from 1 to 200,000, the second fragment is from 100,000 to 300,000, and so on, up to the eighth, which is from position 700,000 to 900,000. One can see that the whole genome consists of two patches with relatively small internal distances.

Table 2. Distances d_1 among contiguous overlapping fragments of *Borrelia burgdorferi* each 200K in size.

Con-tigs	1	2	3	4	5	6	7	8
1	0.00	0.03	0.07	0.10	0.20	0.27	0.28	0.28
2	0.03	0.00	0.03	0.10	0.21	0.28	0.29	0.28
3	0.07	0.03	0.00	0.05	0.18	0.27	0.27	0.27
4	0.10	0.10	0.05	0.00	0.09	0.19	0.19	0.20
5	0.20	0.21	0.18	0.09	0.00	0.05	0.08	0.08
6	0.27	0.28	0.27	0.19	0.05	0.00	0.03	0.05
7	0.28	0.29	0.27	0.19	0.08	0.03	0.00	0.03
8	0.28	0.28	0.27	0.20	0.08	0.05	0.03	0.00

For the sake of comparison of CSs, we also used Euclidean distance d_3 (S; S') = $(\Sigma(f_i(\mathrm{S}) - f_i(\mathrm{S}''))^2)^{1/2}$. However, as shown below, it is less effective.

Compositional spectra as random objects

We will consider the set W as a sample from some class of sets. Namely, we call the set of words W uniformly random, if for each word at each position every letter of the alphabet $\{A, T, C, G\}$ has an identical probability of appearance (0.25). To characterize the robustness of the obtained results, we will explore 100 such uniformly random sets W.

3. SOME STATISTICAL CHARACTERISTICS OF CS-BASED DISTANCES

The definition of CS-distance derived by using a random set of words is correct only in the case that the numerical values of these distances will not depend (or will depend only slightly) on the sampled set of words W.

Distance variation among contigs of different species measured by different sets of words

Let us take an arbitrary set $W = W_1$. Consequently, CSs can be generated for each species. One could then calculate the CS-distance for each pair of species and this procedure could be repeated many times for other sampled sets W. The natural expectation is that the CS-distances generated by various W_i will be similar to that generated by W_1. The results presented in Figure 2 demonstrate that this is indeed the case when the size of the sets W is sufficiently large. Clearly, with an increasing size of W, variation of $d(S, S')$ decreases at any pair of sequences (S, S'). Thus, the chosen

set parameters (L = 10, r = 2) for this study provides a sufficiently narrow variation of the pairwise distances with respect to a sampling variation of the word sets.

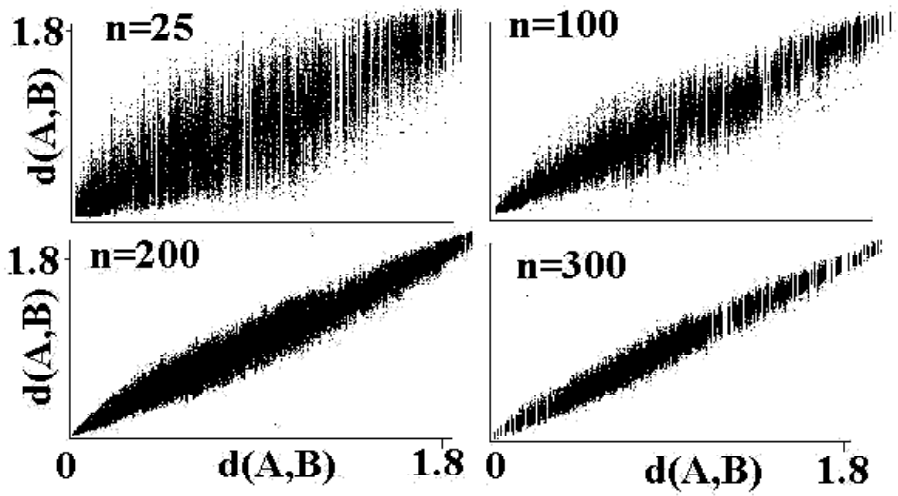


Figure 2. The effect of the vocabulary size (n) on CS-distance. On the axis X all pairs of sequences are ordered values by increase of CS-distance, on an axis Y corresponding values of distances for the same pairs obtained on 100 random sets are shown. With increasing size of W, variation of CS decreases for any pair of sequences.

Distribution of the intragenomic and intergenomic distances

Results of the intragenomic and intergenomic comparisons of spectra on the basis of the three above considered measures (see Section 2) are presented in Figure 3. The first column shows distributions of intragenomic distances (between sequences A and B from one genome, taken for all 38 pro- and eukaryote species, generated repeatedly using 100 random sets W_{10} each with n = 200 words). The second column shows the histograms of 37x38/2 intergenomic distances (using the same procedure as with intragenomic distances). The distances d_1 (Figure 3A) and d_2 (Figure 3B) proved to yield quite similar results. It would be natural to prefer d_1 based on the fact that the ratio of its ranges of variation for intergenomic comparisons is R_1 =1.84/0.20=9.2 is two-fold of the corresponding ratio for d_2: R_2 = (1.70/0.40) = 4.5. It is noteworthy that both measures are similar with respect to the range of variation of intraspecific and interspecific distances relative to the potentially possible range 0 - 2. Namely, in both, only a small left part of the total range is not empty for the intraspecific comparisons, and only a small right part of the total range is empty for the intraspecific comparisons. We are very satisfied with the last fact because it means that the maximal distances between the analyzed genomes actually approached the maximum possible value, corroborating thereby the extremely high diversity of chosen species (representing all three major sections of life {*Archaea*, *Eubacteria* and *Eukaria*}). One more comment will be useful regarding the comparative properties of d_1 and d_2. The

above-mentioned two-fold difference between R_1 and R_2 is a result of two effects: lower values of d_1 than d_2 intragenomic comparisons, and opposite ranking for intergenomic comparisons. Both effects vote for d_1. For all these reasons the measure d_1 was chosen as the basis for our further study (which will be referred to as CS-distance).

It would be of interest to compare the chosen CS-distance with the more widespread Euclidean measures d_3 (Figure 3C). The inferiority of this measure compared to the proposed CS-distance can be clearly seen from the comparison of the obtained intergenomic distances with the highest value possible for this distance. Indeed, unlike the CS-distance distribution where only a small right hand-part of the total range is empty, the obtained values using Euclidean distance leave empty the predominant (about 6/7) part of the total range.

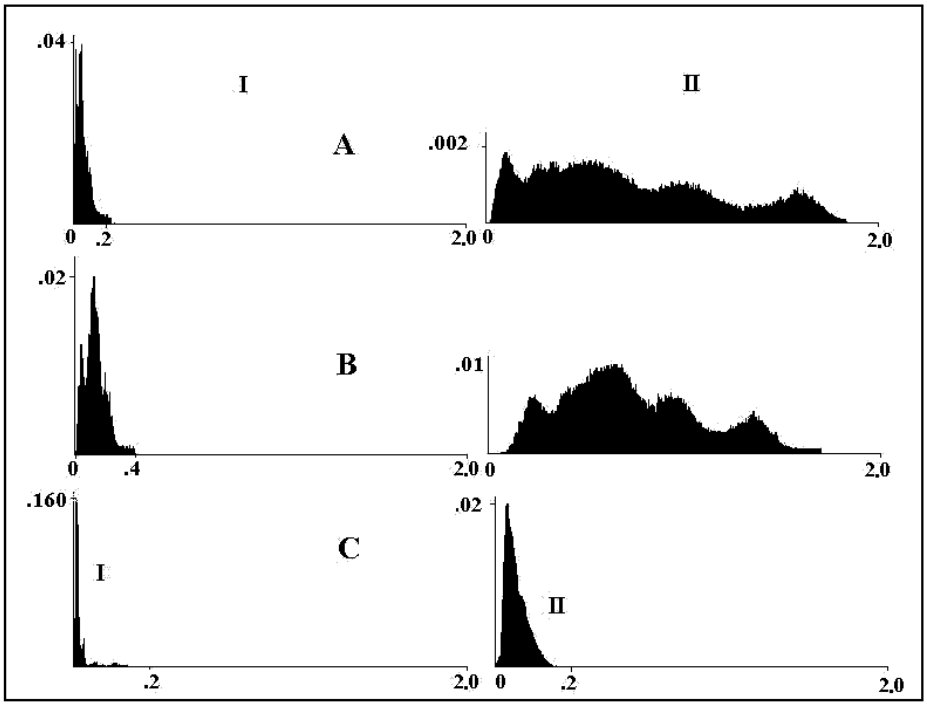


Figure 3. Distribution of intragenomic and intergenomic CS-distances based on three considered measures. (**A**) using Spearman rank correlation ρ; (**B**) using Spearman rank correlation τ; (**C**) using Euclidian distance. The first column shows the distributions of intragenomic distances, the second column shows the histograms of intergenomic distances.

Low sampling variation of relative ordering of species

The possibility of using d_1 as a measure of intergenomic distances depends on the stability of the interspecific relationships revealed by CS-comparisons, relative to the sampling variations of the sets *W*. In other words, we are interested in the ordering of species relative to each other rather than to the numerical values of the distances.

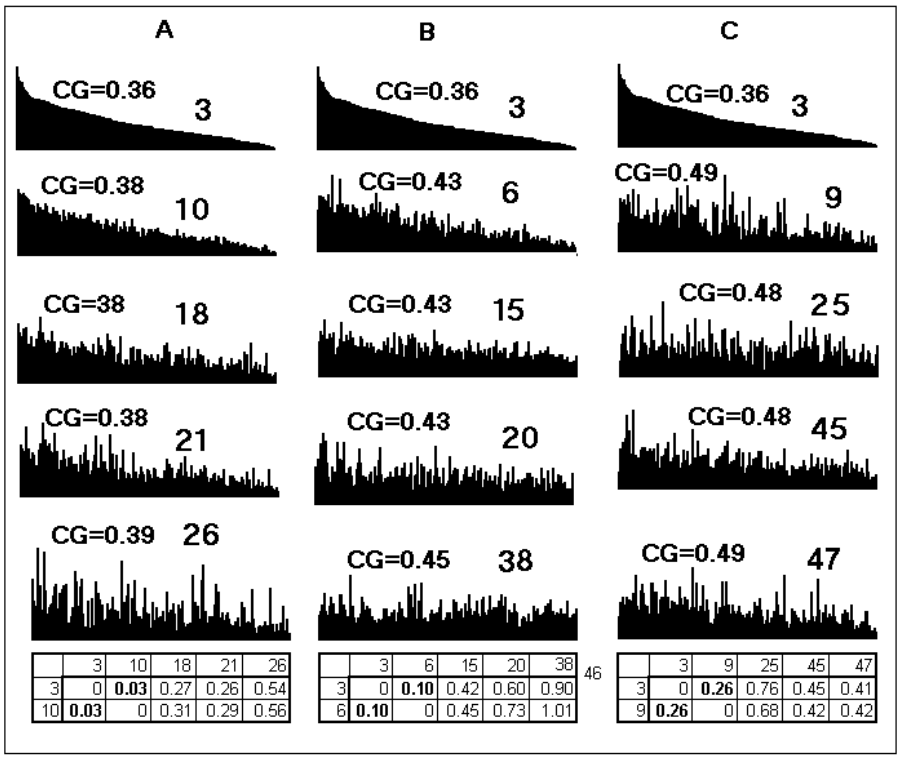
Consequently, we would like to test the stability of some characteristics of species ordering with respect to variations of *W*.

For any reference sequence *i*, one could determine the relative ordering of all other sequences based on their CS-distance to *i*. We denote such relative ordering as *i*-order. As could be seen from the above-mentioned results, CSs of two sequences of the same genome are, as a rule, closer to each other than to any other genome. This fact does not depend on the sampled set *W*. Hence, we could interpret an ordering of sequences as a species ordering. Let us check robustness of an *i*-order. We will first check the stability of appearance of the first element in the *i*-order (a species with minimal distance to the reference one). It turned out that in all species considered, the first element did not depend on the sampled set in more than 70% of the *W* sets, while in 22 of 38 species the corresponding range of consistency was even 90 - 100% of the cases (of course, with every species *i* we identified its closest neighbor in *i*-order). In all of the cases only one alternative to the conservative first element was found. These results indicate that close proximity measured by CS-distance does not depend on *W* content.

Let us check the variation of each species *j* in the *i*-order upon *W*-variation $P^i(i,W)$ is a denotation of a rank of *j* in *i*-order, $D_j = \mid P^i(i,W) - (P^i(i,W') \mid$ is the absolute value of the difference between two ranks $P^i(i,W)$ and $P^i(i,W')$. It appears that the mean value of D_j taken over all pairs W_t, W_s $(t \neq s;\ t,s = 1,...,100)$ and all $i\,(i = 1,...,38)$, for each $j = 1,...,38$ belongs to the interval [0.5, 2]. This means that the positions of any species *j* in any two *i*-orders (generated by two arbitrary sets *W* and *W'*) will differ in average for not more than two steps. Correspondingly, the close similarity of orders obtained using different word sets is reflected in a high Spearman rank correlation (0.96 when averaged across all pairs of *W* and *W* ' and all species).

4. COMPOSITIONAL SPECTRA AND GC CONTENT

One of the objections to the proposed method may be related to a strong influence of GC-content on the measure, both within and between species. It could be speculated that the observed variation CS-distance is definitely related to a G+C genomic composition but is not completely predefined by it. We performed two tests to demonstrate this statement. The results of the first test are presented in Figure 4. This figure is organized as a table with three columns and six rows. The top and the bottom rows are arranged differently from the other rows. The first (top) row consists of a spectrum related to the fragment of the *H.sapiens* chromosome 1 (No. 3) repeated three times to appear in all three columns, because this spectrum serves as a reference ranking. The second row consists of spectra related to three different human chromosomes. The next three rows are constructed from spectra related to various organisms, which were selected according to the G+C content of these sequences fitting corresponding human fragments from the same column. The bottom row presents the d_1-distances between appropriate human and other sequences. All the numbers are related to the notation of sequences in the Appendix A. Remarkably, a distance between two human fragments is always smaller than a distance between a human fragment and an unrelated sequence, in spite of the latter possessing G+C composition identical to those of the human segment.



	3	10	18	21	26
3	0	**0.03**	0.27	0.26	0.54
10	**0.03**	0	0.31	0.29	0.56

	3	6	15	20	38
3	0	**0.10**	0.42	0.60	0.90
6	**0.10**	0	0.45	0.73	1.01

46

	3	9	25	45	47
3	0	**0.26**	0.76	0.45	0.41
9	**0.26**	0	0.68	0.42	0.42

Figure 4. Comparison of CS-distance from different parts of the Human genome and other species with the same GC content. **A** First two spectra belong to Human contigs 3 and 10 accordingly. These sequences have approximately identical CG content. The other three sequences from another species also have the same value of GC content. Values CS-distance between everyone Human contigs up to all others also is given in the table; **B** The first spectrum in this group is the same as in group **A**. However following (Human) contigs has the greater contents GC and the same contents have others contigs of this group; **C** Similarly in group about the First spectrum in this group same as in group **A** and **B** and others contigs have higher GC content than in group **B**.

We also performed randomization tests to study the dependence of a d_1-distance upon G+C content. Every organism in our database is presented by two fragments: *A* and *B*. For the purpose of this test, fragments *B* were randomized by partial reshuffling. GC- or AT-reshuffle means that only nucleotides *G* and *C* or *A* and *T*, respectively, participated in the reshuffling. Distributions of such "pseudo-intragenomic" distances between an original segment *A* and reshuffled sequence *B* are presented in Figure 5 in the upper part of it (part I). Distributions of "pseudo-intergenomic" distances are presented in the part II of Figure 5. Figure 5 is constructed analogously to Figure 3. Comparison of Figures 3 and 5 leads us to the conclusion that even partial reshuffling, while not only the G+C content is preserved but also all A and

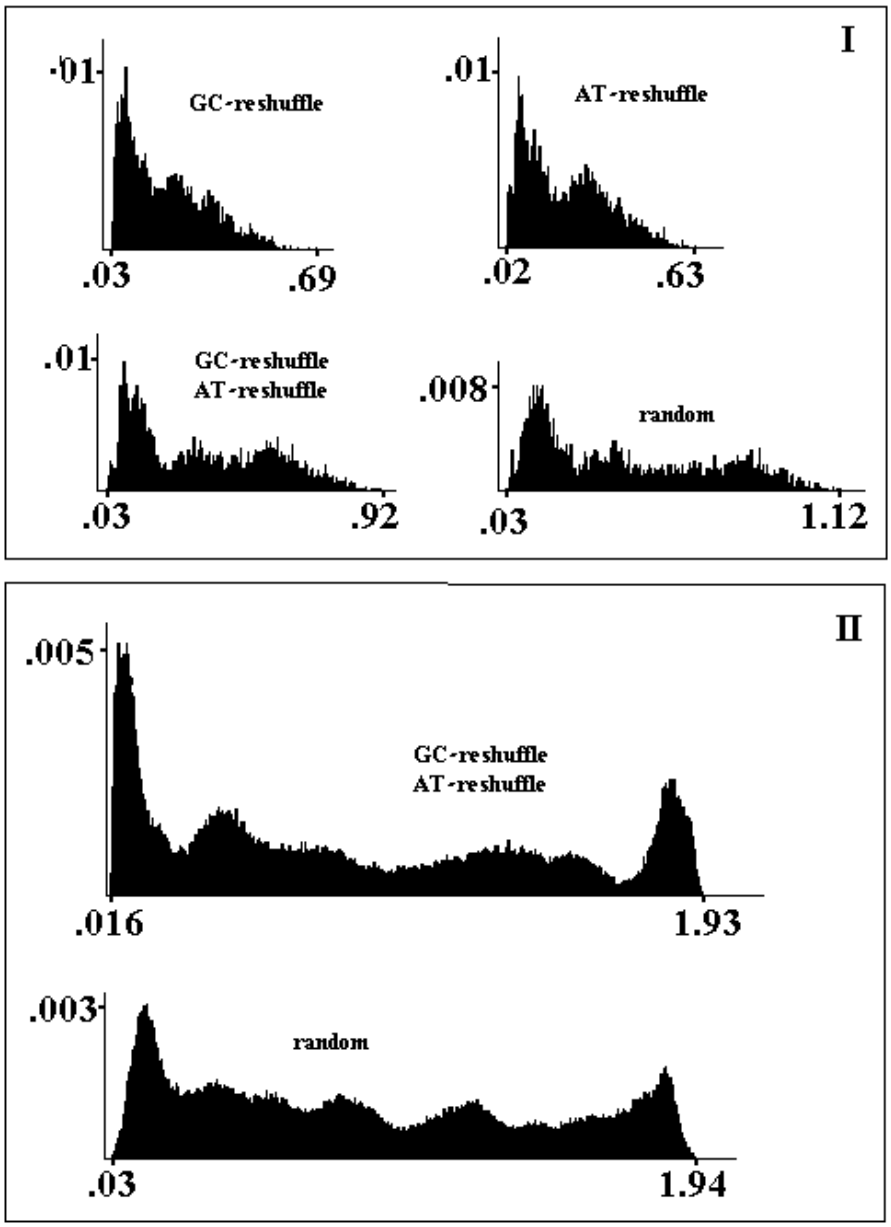


Figure 5. Distributions of pseudo-intragenomic and intergenomic CS-distances based on the Spearman rank correlation. Part I presents distributions of pseudo-intragenomic distances between an original fragment *A* and reshuffled fragment *B*. Part II presents distributions of pseudo-intergenomic distances between reshuffled fragments *B*.

T nucleotides hold their positions in sequence, substantially increases original CS-distance between fragments. For example, real intragenomic distances lay in the interval from 0 to 0.2 (Figure 3A), while pseudo intragenomic distances occupy the whole range from 0.04 to 0.66 in case of GC reshuffling, and from 0.04 to 0.9 in case of GC+AT reshuffling.

5. DISCUSSION

Compositional spectra - random set of words

Analyzing heterogeneity of DNA sequences is considered one of the main targets of the current stage of genomic studies. The natural and simple approach proposed in this study seems to be especially useful for examining intergenomic and intragenomic relations. We found that relative frequencies of appearance of individual words taken from an arbitrary set of words of a fixed length result in a species-specific pattern (referred to as a "compositional spectrum"). The pattern obtained appeared to be reproducible (among different samples from the same genome sequence) given a sufficiently large set of words and a sufficiently large genomic DNA sample. The proposed approach of genome characterization by the compositional spectra assumes some reasonable choice of three parameters: *length of words* (L), *number of words* (n), and *allowed mismatch* (r). From the biological point of view, the word length should be in the range of 8 to 20 bases, typical for protein binding sites, and the number of mismatches should correspond to about 20-30% of site length. The compositional spectrum is actually a histogram of imperfect word occurrences. The majority of histogram elements should have statistically significant values; otherwise, this approach loses its power. The length $L = 15$ would not be a good choice, for example, because it would require mismatch threshold $r = 8$, as we experimentally found examining genomic fragments of 500 Kb length. Our choice of the parameters ($L = 10$, $r = 2$) is a compromise between biological and statistical considerations (Reinert *et al.*, 2000).

Compositional spectra - non-random set of words

In our work we studied the peculiarities of the *compositional spectrum* that is based on the sets of random (to be more exact – evenly) distributed words. Therefore the parameters of the spectra, in particular distances between spectra, should be viewed as random quantities. For more details, see (Kirzhner *et al*., 2002). The same work refers to the fact that different probabilistic models of word sets generation (e.g. C+G rich words) provide a possibility to characterize correlations between sequences in different ways. Here, we will give one of the possible models of the spectrum on a *non-random* set of words, namely, the ones derived from a certain selected sequence. Simultaneously, we will further justify our use of words of length between 10 and 15.

Similarity of, or differences between, sequences may already be estimated at the level of letter frequencies. However, the question is: What is the correlation between the sequence itself and the frequencies mentioned? The answer is obvious: there is a tremendous number of sequences that are different in many ways. Yet they have the same letter frequencies. This is why the "distance" between two sequences based on

letter frequencies has a quite low "informational" quality. Such distances may sometimes obey the natural logic: the fragments of one genome will (as a rule) be sufficiently close to one another, while the genomes having very differing letter frequencies will be far from one another.

A similar situation arises when we try to characterize a sequence by frequency of words with lengths of 2, 3 or 4.

The situation is completely different for longer fragments. In particular, the frequencies of a full set of words of length L = 10 - 15, to a great extent, determine the sequence. Only in the case when the lengths of repetitions in the sequence exceeds the word length L then unique reconstruction is impossible. Dyer *et al*. (1994) determine the asymptotic limiting probability as $m \to \infty$ that a random string of length m over some alphabet σ can be determined uniquely by its substrings of length L. In our case, $\sigma = 4$, $m = 300,000 - 500,000$ and the estimate of L is approximately 10 - 12. Preparata *et al*. (1999) provide a simple algorithm which with high probability reconstructs sequences of length $O(4^k)$ (this is asymptotically optimal). If k = 10, than the length of a sequence that may be reconstructed practically equals 4^{10} = 1,000,000 letters. In our work, we used the fragments of genomes of length of 300,000 - 500,000.

Thus, the frequencies of a full set of sufficiently long words do represent the sequence well enough. Yet, how should these sets be compared to each other? How may the distance between them be determined when the frequency of every single word is very small. This task "*evaluation of distance between two sets based on the frequencies of the words"* is well defined and may be fulfilled in more than one way. One of the variants may be clusterization at the expense of identification of close (e.g. according to Hemming's metrics) words. After such a procedure, we obtain a pattern that in our work is referred to as a compositional spectrum. The word length L does not change, magnitude r equals Hemming's distance (or is calculated in some other way in the clusterization scheme). With such an approach, the set of words W is chosen not randomly, but rather guided by the method of clusterization, being "centers" of clusters.

Such a set W of words derived for one of genomes (fixed genome) provides the possibility of analyzing the correlations with other genomes in relation to the fixed genome.

Note that Sandberg *et al.* (2001) proposed another way of using the complete set of words of length L for classifying genomes. It turned out that, in a certain statistical procedure, this complete set of words of length L = 9 from a short fragment of genome (400 bp) can efficiently establish whether two different fragments belong to the same genome, even if the test included closely related micro-organisms.

Compositional spectra - distances

Determination of the distance between spectra is the central point of the whole approach proposed. A "good distance" must reflect an informal understanding of spectra closeness. However, there are formal criteria of distance quality as well. Thus, it is advisable that, as a rule, parts of one genome were close enough to each other, as well as strains of one species. These requirements are natural. It is less obvious, but as well important that the distance distribution on a great amount of heterogeneous species is close to uniform and completely covers a possible range of distances.

Necessity of this condition becomes clear if we imagine the opposite. Namely, let the distance distribution be close to unimodal on a great amount of heterogeneous species. This essentially means that the distance between any pair of species is the same and equals the value of the mode and the distribution curve is described by dispersion around this. This distance scarcely reflects the genuine correlations between genomes. Euclidean distance (natural, if the set W is viewed as a vector) appeared to be one of this type. Euclidean distance depends on the values of word frequencies in a sequence, in other words, on the spectrum form. This characteristic has special peculiarities. Suffice to mention that in the case of the linguistic nature of words the frequency distribution of words obeys Zipf's law. In the case of CS this law does not hold. Nevertheless, the frequency parameter may bear a "latent" information being a "noise" in respect to the spectrum characteristics necessary for distance determination. In a certain sense, the closeness measure suggested in this work is the "minimum" one as it depends only on violation of word order in the spectra of two comparable sequences that are independently ordered. Its minimality lies in the fact that, as it seems, it is impossible to use less data about spectrum if the distance calculation is based upon all the words of the set W (and not on some kind of subsampling, for example). At the same time, this measure is clear: if in an independent ranking of two spectra all the words appear in the same places then the distance is zero. The more words in one spectrum shift their places in respect to the other spectrum and the greater the magnitude of shifts, the greater the distance.

Of course, other ways of distance determination that are not known to us by now are possible as well. Probably, other distances will provide a more effective use of the spectrum.

Prospects

Earlier, we had noted that for each sequence among those that were considered by us the closest is, as a rule, another one of the same genome. However, almost all considered genomes were monochromosomal. It is of interest to consider this result for a polychromosomal genome. Denote by D_{min} the minimum intergenomic distance d over all pairs of contigs for 100 different set W of words. The majority of pairs (87%) of contigs from different human chromosomes have the distances $d(h_i,h_j)$ less than D_{min}. For the other 13% of pair distances $d(h_i,h_j) > D_{min}$. But for every human chromosome h_i (h_j) entering into such a pair there always exist a closer chromosome h_{is} (h_{jt}) for which $d(h_i,h_{is}) < D_{min}$ and $d(h_j,h_{jt}) < D_{min}$.

We believe that the proposed approach may be useful in addressing various questions related to large-scale genome comparisons. For example, in our tests we have revealed genome compositional heterogeneity of several species, including *Borrelia burgdorferi*, which was found earlier using other methods (McInerney, 1998). Our CS-calculations also confirmed the recent findings that thermophillic eubacteria *Aquifex* and *Thermotoga* are closer to *Archaea* than to *Eubacteria* (Nelson *et al*., 1999).

ACKNOWLEDGEMENTS

We wish to thank Ed Trifonov for critical discussions; the Israeli Ministry of Immigrant Absorption and Ministry of Science for financial support of V.K. and A.K.,

A.B. is supported by the FIRST Foundation of the Israel Academy of Science and Humanities. The Ancell Teicher Research Foundation for Molecular Genetics and Evolution also financially supported the study.

APPENDIX A. LIST OF DEPICTED GENOMIC SEQUENCES

Notation of fragments in all figures mentioned corresponds to indexes in the Appendix. Every record in the list consists of the index, the name of the species, an accession number for some fragments is mentioned to avoid a wrong identification, and some data about a fragment in brackets. For a fragment of a complete genome we specify a starting point and a length of the fragment, for a fragment having an accession number we specify its length only. A real number with a value between 0 and 1 is related to the G+C content of the fragment. For example, the record "24 - Mycobacterium tuberculosis (A: 1199341 (358717), 0.65; B: 2579341 (358913), 0.65)" means that M. tuberculosis has an index of 24; it is presented by two fragments, A and B, from the complete genome, while the first segment starts at position 1199341 and has a length of 358717 bp, and the fragment B starts at position 2579341 and has a length of 358913 bp; both fragments have 65% G+C composition.

Eukaryota:

1 - *Homo sapiens* chr. X (NT 011528) (539188 , 0.40); 2 - *Homo sapiens* chr. Y (NT011864) (539595, 0.40); 3 - *Homo sapiens* chr. 1 (NT 004302) (539495, 0.36); 4 - *Homo sapiens* chr. 3 (NT 002444) (543554, 0.46); 5 - *Homo sapiens* chr. 4 (NT 006051) (535847, 0.44); 6 - Homo sapiens chr. 6 (NT 007122) (599072, 0.43); 7 - *Homo sapiens* chr. 7 (NT 007643) (447791 , 0.36); 8 - *Homo sapiens* chr. 11 (NT 008933) (506578 , 0.37); 9 - *Homo sapiens* chr. 12 (NT 009410) (532934 , 0.49); 10 - *Homo sapiens* chr. 13 (NT 009796) (537882 , 0.38); 11 - *Homo sapiens* chr. 20 (NT 011328) (540270 , 0.44); 12 - *Homo sapiens* chr. 22 (NT 001454) (701877 , 0.50); 13 - *Mus musculus* (A: chr7, AC012382, 276523, 0.44; B: chr.11, AL603707, 234182, 0.49); 14 - *Caenorhabditis elegans* (A: chr1, 1-438825, 0.36; B: chr2, 1-350000 , 0.36); 15 - *Drosophila melanogaster* (A: chr.2 AE003641, 299556, 0.42; B: chr. X, AE003506, 300000, 0.43); 16 - *Arabidopsis thaliana* (A: chr.1, NC 003071.1 100000-531319, 0.36; B: chr.1, NC 003075.1, 400000-727859, 0.36); 17 - A thaliana mitochondrial genome (A: NC 001284.1, 366923, 0.45); 18 - *Saccharomyces cerevisiae* (A: chrii, 1-800000, 0.38; B: chrxv,1-800000, 0.39); 19 - *Leishmania major* (A: AE001274, 1-171770, 0.62);

Eubacteria:

20 - *Bacillus subtilis* (A: 1199941 (579647) , 0.43; B: 2219941 (399002), 0.41); 21 - *Streptococcus pyogenes* (A: 239941 (690238) , 0.38; B: 1079941 (696345), 0.39); 22 - *Mycoplasma genitalium* (A: 1 (287593), 0.33; B: 278581 (288000), 0.30); 23 - *Mycoplasma pneumoniae* (A: 239941 (199523) , 0.40; B: 539941 (199523), 0.40); 24 - *Mycobacterium tuberculosis* (A: 1199341 (358717), 0.65; B: 2579341 (358913), 0.65); 25 - *Synechocystis sp* (A: 719941 (349960), 0.48; B: 2699941 (350000), 0.47); 26 - *Helicobacter pylori* (A: 599941 (320335), 0.39; B: 1439941 (320387), 0.39); 27 - *Escherichia coli* (A:599941 (519942) , 0.51; B: 2999941 (542976) , 0.51); 28 -

Deinococcus radiodurans (A: 599941 (399971), 0.67; B: 1799941 (399983) , 0.66); 29 - *Thermotoga maritima* (A:59941 (370054), 0.46; B:1259941 (366191), 0.46); 30 - Aquifex aeolicus (A: 599941 (399976) , 0.43; B: 1199941 (400002) , 0.44); 31 - *Neisseria meningitidis* (A: 599941 (361259), 0.51; B: 1199941 (373905), 0.52); 32 - *Campylobacter jejuni* (A: 59341 (399984), 0.31; B: 1079341 (400002), 0.30); 33 - *Haemophilus influenzae* (A: 119941 (399863), 0.38; B: 1139941 (399981), 0.38); 34 - *Chlamydia trachomatis* (A: 59941 (399976), 0.41; B: 719941 (400002), 0.42); 35 - *Clostridium acetobutylicum* (A: 315781 (347567), 0.32; B: 3000001 (340105), 0.31); 36 - *Treponema pallidum* (A: 59941 (275933), 0.52; B: 719941 (275984), 0.53); 37 - *Pseudomonas aeruginosa* (A: 720001 (345206), 0.65; B: 1920001 (355163) , 0.67); 38 - *Actinobacillus actinomycetemcomitans* Strain HK1651 (A: 6000 (338681), 0.45; B: 800000 (344947), 0.45); 39 - *Rickettsia prowazekii* (A: 239941 (276000), 0.29; B:719941 (276000), 0.29); 40 - *Chlamydia pneumoniae* (A:360001 (300017), 0.39; B: 840001 (300000), 0.41); 41 - *Borrelia burgdorferi* (A: (1) 399967, 0.29; B: 400000 (399979), 0.29);

Archaea:

42 - *Halobacterium sp*. Plasmida NC 001869 (A: 191652, 0.58); 43 – *Pyrococcus horikoshii* (A: 240001 (399992), 0.42; B: 840001 (400002), 0.42); 44 - *Pyrococcus abyssi* (A: 360000 (360000), 0.45; B: 1200001 (360000), 0.45); 45 - *Archaeoglobus fulgidus* (A: 12061 (399986), 0.48; B: 1200061 (400002), 0.48); 46 - *Methanococcus jannaschii* (A: 120001 (399868), 0.32; B: 840001 (399977), 0.31); 47 - *Methanobacterium thermoautotrophicum* (A: 599941 (344374), 0.49; B: 1199941 (344455), 0.50); 48 - *Aeropyrum pernix* (A: 360061 (400002), 0.58; B: 960061 (400002), 0.57); 49 - *Sulfolobus solfataricus* AE006641 (A: 400001 (584947), 0.36; B: 1220002 (308779), 0.36);

REFERENCES

Brendel, V., J.S. Beckmann and E.N. Trifonov (1986). Linguistics of nucleotide sequences: morphology and comparison of vocabularies. Journal of Biomolecular Structure and Dynamics 4: 11-21.

Dyer, M., A. Frieze and S. Suen (1994). The probability of unique solutions of sequencing by hybridization. Journal of Computational Biology 1: 105-110.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. Current Opinion in Microbiology 1: 598-610.

Karlin, S. and J. Mrazek (1997). Compositional differences within and between eukaryotic genomes. Proceedings of the National Academy of Sciences of the United States of America 94: 10227-10232.

Karlin, S. and C. Burge (1995). Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics 11: 283-290.

Kendall, M. G. (1970). Rank Correlation Methods. Charles Griffin & Co., Ltd, London.

Kendall, M. G. and A. Stuart (1967). Inference and Relationship, 2. Charles Griffin & Co., Ltd, London.

Kirzhner, V.M., A.B. Korol, A. Bolshoy and E. Nevo (2000). Extensive Sets of Words Reveal Large-Scale Genome Organization. Poster in Genomes 2000: International Conference on Microbial and Model Genomes, Paris, France.

Kirzhner, V.M., A.B. Korol, A. Bolshoy and E. Nevo (2002). Compositional spectrum - revealing patterns for genomic sequence characterization and comparison. Physica A 312: 447- 457.

McInerney, J. O. (1998). Replicational and transcriptional selection on codon usage in Borrelia burgdorferi. Proceedings of the National Academy of Sciences of the United States of America 95: 10698-10703.

Nelson, K. E. R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. Mcdonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter and C. M. Fraser (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature 399: 323-329.

Pevzner, P., M. Borodovsky and A. Mironov (1989). Linguistics of nucleotide sequences. I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. Journal of Biomolecular Structure and Dynamics 6: 1013-1026.

Pietrokovski, S. (1994). Comparing nucleotide and protein sequences by linguistic methods. Journal of Biotechnology 35: 257-272.

Pietrokovski, S., J. Hirshonn and E. N. Trifonov (1990). Linguistic Measure of Taxonomic and Functional Relatedness of Nucleotide Sequences. Journal of Biomolecular Structure and Dynamics 7: 1251-1268.

Preparata, F., A. Frieze and E. Upfal (1999). Optimal reconstruction of a sequence from its probes. Journal of Computational Biology 6: 361-368.

Reinert, G., S. Schbath and M. S. Waterman (2000). Probabilistic and statistical properties of words: an overview. Journal of Computational Biology 7: 1-46.

Sandberg, R., G. Winberg, C-I. Branden, A. Kaske, I. Ernberg and J. Coster (2001). Capturing Whole-Genome Characteristics in Short Sequences Using a Naïve Bayesian Classifier. Genome Research 11: 1404-1409.