ELSEVIER

Physica A 312 (2002) 447–457

PHYSICA A

www.elsevier.com/locate/physa

Compositional spectrum—revealing patterns for genomic sequence characterization and comparison

Valery M. Kirzhner *, Abraham B. Korol, Alexander Bolshoy, Eviatar Nevo

Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

Received 13 February 2002

Abstract

In this paper we propose a natural approach to characterizing genomic sequences, based on occurrences of fixed length words (strings over the alphabet $\{A, C, G, T\}$) from a sufficiently large set W of arbitrary (in general case) words. According to our approach, any genomic sequence can be characterized by a histogram of frequencies of imperfect matching of words from the set W that is called a compositional spectrum (CS). The specificity of CSs is manifest in a reasonable similarity of spectra obtained on different stretches of the same genome and, simultaneously, in a broad range of dissimilarities between spectral characteristics of different genomes. The proposed approach may have various applications in intra- and intergenomic sequence comparisons.

Keywords: DNA sequences; Set of words; Sequence comparisons; Compositional spectra; Imperfect matching

1. Introduction

Significant progress in genome structure and evolution has been made in sequencing full genomes of many viruses, bacteria and some eukaryotes, including humans, *Drosophila*, yeast, nematode, and significant parts of some plants (in particular *Arabidopsis* and rice) (http://www.ncbi.nlm.nih.gov/). The era of comparative genomics is advancing rapidly. The overriding question is how do we deal with, interpret and use all this new information? Can it be used by sequence comparison to highlight enigmas and mysteries of genome organization dynamics and evolution?

* Corresponding author.
E-mail addresses: valery@esti.haifa.ac.il (V.M. Kirzhner), korol@esti.haifa.ac.il (A.B. Korol).

0378-4371/02/$ - see front matter
PII: S0378-4371(02)00843-9

Several types of genome analysis could be developed on the basis of frequency of words. For example, within the framework of linguistic analysis, one can estimate (derive) a set of certain words (dictionary) from given sequences [1–6] and its various characteristics, in particular, from the information viewpoint [7–9]. Many linguistic attempts aimed to characterize genomic sequences through sets of all used words with lengths not exceeding some threshold [2,4,10,11]. Note, that the typical word size of the nucleotide language is found to be 3–5 [2,4,12].

The formal-statistical approach characterizes a DNA sequence by frequencies of words of the given class [13,14]. In particular, in this way the frequencies of words that are, by some measure, over- or under-represented in relatively large target sequences can be evaluated [15,16]. An overview of statistical and probabilistic properties of words, as occurring in the analysis of biological sequences, is given in Ref. [17].

In this study, our purpose is to introduce a natural measure for genome comparisons that could be called "linguistic" as well, because it also deals with "words", albeit in a different way from previous studies. For a relatively small arbitrary set of fixed-length words, some specific pattern, referred to as a "compositional spectrum", could be defined. A compositional spectrum is the distribution of frequencies of imperfectly matching selected words calculated over large genomic stretches (e.g. $N \approx 5 \times 10^5$–10^6 bp). For construction of a spectrum we propose to choose words length not < 8 (in fact, 8–15). Our analysis showed that compositional spectra are informative for intragenomic and intergenomic comparisons and highlight in-depth evolutionary patterns and processes.

2. Basic definitions

2.1. Compositional spectra of sequences

In the text of this paper, we use the terms 'biosequence' or 'genomic sequence' to denote unspaced text over the four-letter alphabet $D = \{A, C, G, T\}$. A string of length L over the alphabet D will be referred to as word (oligonucleotide) of length L. If x is a substring of a string S, we will assume that word x has a perfect occurrence in the target sequence S. We assert that word y has an imperfect occurrence in S if there exists a substring x of S, and the distance (in a given metric space) between x and y is less than a threshold (in the given metrics). We employed two distance metrics for counting imperfect occurrences of a word:

(1) Word x is an imperfect occurrence of w_i in S if *Hamming distance* between word w_i and word x is less than a given r. This approximate matching can be denoted as "r-mismatching" (see also Refs. [19,20]).
(2) Word x is an imperfect occurrence of w_i in S if the *smallest weighted sum* of mismatches, insertions, and deletions is less than a given r. This sequence metrics is well known in sequence-alignment applications [18].

Let us consider a set W of n different words w_i of length L, $n \ll 4^L$, where 4^L is the total number of different words of length L. By m_i we denote the number of *imperfect*

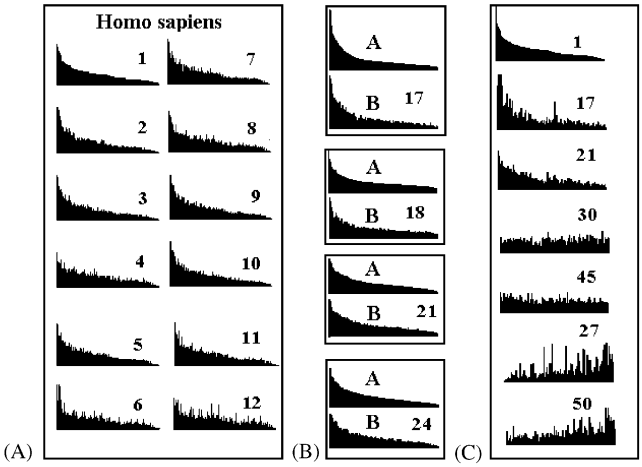

Fig. 1. Examples of compositional spectra of various species, based on any set of words W. Length of a word in W is 10 ($L=10$). Number of words is 200 ($n=200$), r-mismatching with $r=2$. (A) Shows spectra of long contigs from human chromosomes with range of $C+G$ contents from 0.36 to 0.50; ordering by chromosome X. (B) Shows spectra of pairs of different contigs from four genomes, ordered in each pair by the first contig (A). (C) Shows spectra of 7 contigs, ordered by human X-chromosome with different level of similarity to X-chromosome. For designations of the species see Table 1.

occurrences of word w_i of the set W in a target sequence S: $m_i = \mathrm{occ}(w_i|S)$. Now let $M = \Sigma m_i$. The frequency distribution $F(W,S)$: $\{f_i = m_i/M\}$ will be referred to as a compositional spectrum of the sequence S relative to the set W.

2.2. Visualization of compositional spectra

The main application of the method of compositional spectra is in sequence comparison. Let $S_1, S_2, \ldots, S_k$ be genomic sequences that we want to compare relative to a given set W. Clearly, a given order of words w_j in W predetermines the shape of the compositional spectrum of S_i relative to W. Indeed, a compositional spectrum, which is actually a frequency distribution on set W, may be presented as a distribution plot, where X-axis corresponds to the running index j of words w_j, and Y-axis presents frequencies f_{ij} of w_i. For better visualization of multiple spectra, one can choose to order the words $w_i \in W$ non-randomly, but naturally related to a descending order of f_{ij} frequencies relative to any S, say $S = S_1$. In other words, we can denote such an arrangement $\mathrm{Ord}(W,S)$, that word w_i is followed by word w_j $(i > j)$ if and only if $m_i \geqslant m_j$ $(m_i = \mathrm{occ}(w_i|S_0))$. In case of equality $m_i = m_j$ each relative ordering of w_i and w_j is permitted. This order of words w_i in W is non-random and derives from the target sequence composition. Such ordering facilitates comparisons of a given set of spectra relative to any chosen one (see compositional spectra of a few species in Fig. 1).

2.3. Distance in the space of compositional spectra

Based on any chosen set of words W it is possible to measure the difference between any two sequences. Namely, by definition, a compositional spectrum of the sequence S relative to the set W is a frequency distribution $F(W,S)$: $\{f_t = m_t/M\}$ on set W. Then, clearly, standard Euclid distance may be used, but our analysis showed this is not the best choice from the viewpoint of sensitivity in detecting sequence dissimilarities (unpublished results). We propose to use two measures, d_1 and d_2, related to the number of permutations in a vector and based on rank correlations ρ and τ (for definition see Ref. [21]). Let us denote $d_1 = 1 - \rho$, and $d_2 = 1 - \tau$, $0 \leqslant d_1,\ d_2 \leqslant 2$. These distances correspond to heuristically acceptable understanding of proximity between two orderings. Therefore, if distance between two sequences $d(S_i, S_j) = 0$ (with d being either d_1 or d_2), then their spectra f_{ti} and f_{tj} are identically ordered and we can say that "S_i is compositionally congruent to S_j". In case of the maximal distance $d(S_i, S_j) = 2$, spectra S_i and S_j are ordered in the strictly reversed order. It seems intuitively obvious that distance between such sequences should be larger than between two unrelated (random) sequences (when $\rho, \tau = 0$ and $d = 1$).

3. Compositional spectra as random objects

For any given parameters L and n we shall consider a set W as a random sample from the set of all possible words, with a sampling procedure that can be represented as a stochastic procedure of words generation. Let us consider a few examples.

(1) Let the words be produced by adding sequentially new letters (out of the four-letter alphabet) with equal probability of appearance of each letter at the current position. We shall name such stochastic procedure *uniformly* random and any random set W of words produced by this procedure will also be referred to as a uniformly random set. Then, a compositional spectrum $F(W,S)$ for each S is a random spectrum. Clearly, the distance between any two random spectra is also a random variable.

(2) Every probability vector $[p_1, p_2, p_3, p_4]$, $(p_1 + p_2 + p_3 + p_4 = 1,\ p_i \geqslant 0)$ present allows generation of a set of words W with fitting frequencies of letters A, C, G, T. We shall call such a random set W *compositionally* random. In these notations uniformly random set $(0.25, 0.25, 0.25, 0.25)$ is a compositionally random set.

(3) *Markovian* random set W arises, if each word w_i is generated by a predefined Markov chain model (probability of a letter in a given position of a word is function of k previous letters of the word).

(4) A random set of words sampled from any fixed genome sequence can also serve as a random set W, but here the sampling is from a universe of a more complex origin. Such set W can also be used for analyzing relationships between other genomes.

Selection of such W-generating procedures produces a corresponding class of compositional spectra.

4. Results and discussion

4.1. Data

Genomic sequences of 48 species of Eukaryota, Eubacteria, and Archaea were used in this study (Table 1). All sequences are rather long (≈ 2–5×10^5 letters) and represent substantially large parts of genomes. These included either continuous stretches (contigs) sampled from databases or summed-up composite sequences combined from a few contigs to produce two different target sequences, each of 200–500 kb for each species. In a few cases, the available material was only sufficient to build one such target.

4.2. Compositional spectra of DNA sequences

From the fact that one employs a random set of words W, it immediately follows that the compositional spectra $F(W,S)$ and $F(W,S')$ of sequences S and S' are also random constructs, hence the distance $d(S,S')$ is a random value itself. Let us analyze the distribution of random variable d_1 (or d_2) produced by uniformly random set W. The main empiric result is that for d_1 (as well as d_2) the measure $d(S,S')$ displayed statistical stability, when S and S' are genomic sequences. Indeed, distribution of $d(S,S')$ appeared to be close to normal, and its standard deviation decreases when the number of words n in set W rises. For 100 uniformly random sets W and for every pair of species, i and j, (i,j from our collection of 48 species, $i \neq j$) we calculated standard deviation σ_{ij} of distances d_1 and d_2 over all tested sets of W. Averaged standard deviation across all possible pairs i,j

$$\bar{\sigma} = \frac{1}{N} \sum \sigma_{ij} \tag{1}$$

can be considered an indicator of robustness of the measure for given L, n, r. Taking $L = 10$, and $r = 2$, and varying n, we obtained the results shown in Table 2.

These results show that the consistency of the genome comparisons by means of compositional spectra increases with n, albeit till some saturating point corresponding to $n \sim 200$. Thus, the size $n = 200$ of the set W is a reasonable asymptotic choice. Similar results were obtained for compositionally random variables W and Markovian random sets (data not shown). Thus, distance estimations between sequences obtained on the basis of a certain choice of W-generating procedure are sufficiently consistent.

Table 1
Mean standard distance (d_1) deviation between two genomic sequences as a function of number of words in W

n	25	50	100	200	300
σ	0.15	0.10	0.09	0.05	0.05

Table 2
List of considered species, lengths of fragments (bp) and $C+G$ contents

Species name	Size (bp)	$C+G$	Species name	Size (bp)	$C+G$
Eukaryotes			34A—*Thermus thermophilus*	198,765	0.68
1—*Homo sapiens* chr. *X** (NT 011528)	539,188	0.40	35A—*Thermotoga maritima**	370,054	0.46
2—*Homo sapiens* chr. *Y** (NT 011864)	539,595	0.40	35B—*Thermotoga maritima**	366,191	0.46
3—*Homo sapiens* chr. 1* (NT 004302)	539,495	0.36	36A—*Aquifex aeolicus**	399,976	0.43
4—*Homo sapiens* chr. 3* (NT 002444)	543,554	0.46	36B—*Aquifex aeolicus**	400,002	0.44
5—*Homo sapiens* chr. 4* (NT 006051)	535,847	0.44	37A—*Neisseria gonorrhoeae*	350,020	0.53
6—*Homo sapiens* chr. 6* (NT 007122)	599,072	0.43	37B—*Neisseria gonorrhoeae*	355,192	0.54
7—*Homo sapiens* chr. 7* (NT 007643)	447,791	0.36	38A—*Neisseria meningitidis**	361,259	0.51
8—*Homo sapiens* chr. 11* (NT 008933)	506,578	0.37	38B—*Neisseria meningitidis**	373,905	0.52
9—*Homo sapiens* chr. 12* (NT 009410)	532,934	0.49	39A—*Campylobacter jejuni**	399,984	0.31
10—*Homo sapiens* chr. 13* (NT 009796)	537,882	0.38	39B—*Campylobacter jejuni**	400,002	0.30
11—*Homo sapiens* chr. 20* (NT 011328)	540,270	0.44	40A—*Haemophilus influenzae**	399,863	0.38
12—*Homo sapiens* chr. 22* (NT 001454)	701,877	0.50	40B—*Haemophilus influenzae**	399,981	0.38
13A—*Mus musculus*	395,579	0.48	41A—*Chlamydia trachomatis**	399,976	0.41
13B—*Mus musculus*	38,1917	0.47	41B—*Chlamydia trachomatis**	400,002	0.42
14A—Gallus gallus	94,266	0.56	42A—*Clostridium acetobutylicum*	347,567	0.32
15A—*Oryzias latipes* (*Medakafish*)	100,248	0.45	42B—*Clostridium acetobutylicum*	340,105	0.31
16A—*Xenopus laevis*	190,689	0.45	43A—*Treponema pallidum**	275,933	0.52
17A—*Caenorhabditis elegans**	438,825	0.36	43B—*Treponema pallidum**	275,984	0.54
17B—*Caenorhabditis elegans**	350,000	0.36	44A—*Pseudomonas aeruginosa**	345,206	0.65
18A—*Drosophila melanogaster**	594,587	0.43	44B—*Pseudomonas aeruginosa**	355,163	0.67
18B—*Drosophila melanogaster**	637,156	0.42	45A—*Porphyromonas gingivalis*	399,961	0.48
19A—*Arabidopsis thaliana*	431,319	0.36	45B—*Porphyromonas gingivalis*	399,972	0.47
19B—*Arabidopsis thaliana*	327,859	0.36	46A—*Actinobacillus actinomycetemcomitans*	338,681	0.45
20A—*A. thaliana* mitochondrial genome	365,493	0.45	46B—*Actinobacillus actinomycetemcomitans*	344,947	0.45
21A—*Saccharomyces cerevisiae**	800,004	0.38	47A—*Rickettsia prowazekii**	276,000	0.29
21B—*Saccharomyces cerevisiae**	1,200,006	0.39	47B—*Rickettsia prowazekii**	276,000	0.29
22A—*Leishmania major*	1,71,770	0.62	48A—*Chlamydia pneumoniae**	300,017	0.39

Eubacteria			48B—*Chlamydia pneumoniae**	300,000	0.41
23A—*Bacillus subtilis**	579,647	0.43	49A—*Borrelia burgdorferi**	399,967	0.29
23B—*Bacillus subtilis**	399,002	0.41	49B—*Borrelia burgdorferi**	399,979	0.29
24A—*Streptococcus pyogenes*	690,238	0.38			
24B—*Streptococcus pyogenes*	696,345	0.39			
25A—*Mycoplasma genitalium**	287,593	0.33	Archaea		
25B—*Mycoplasma genitalium**	288,000	0.30	50A—*Halobacterium sp. Plasmida**	191,652	0.58
26A—*Mycoplasma pneumoniae**	199,523	0.40	51A—*Pyrococcus horikoshii**	399,992	0.42
26B—*Mycoplasma pneumoniae**	199,523	0.40	51B—*Pyrococcus horikoshii**	400,002	0.42
27A—*Mycobacterium tuberculosis**	358,717	0.65	52A—*Pyrococcus abyssi**	360,000	0.45
27B—*Mycobacterium tuberculosis**	358,913	0.65	52B—*Pyrococcus abyssi**	360,000	0.45
28A—*Synechocystis sp**	349,960	0.48	53A—*Archaeoglobus fulgidus**	399,986	0.48
28B—*Synechocystis sp**	350,000	0.47	53B—*Archaeoglobus fulgidus**	400,002	0.48
29A—*Helicobacter pylori**	320,335	0.39	54A—*Methanococcus jannaschii**	399,868	0.32
29B—*Helicobacter pylori**	320,387	0.38	54B—*Methanococcus jannaschii**	399,977	0.32
30A—*Escherichia coli**	519,942	0.51	55A—*Methanobacterium thermoautotrophicum**	344,374	0.49
30B—*Escherichia coli**	542,976	0.51	55B—*Methanobacterium thermoautotrophicum**	344,455	0.50
31A—*Enterococcus faecalis*	399,976	0.37	56A—*Aeropyrum pernix**	400,002	0.58
31B—*Enterococcus faecalis*	399,956	0.37	56B—*Aeropyrum pernix**	400,002	0.57
32A—*Deinococcus radiodurans*	399,971	0.67	57A—*Sulfolobus solfataricus*	584,947	0.36
32B—*Deinococcus radiodurans*	399,983	0.66	57B—*Sulfolobus solfataricus*	308,779	0.36
33A—*Bacillus stearothermophilus*	378,629	0.53	58A—*Methanococcus maripaludis*	116,645	0.35
33B—*Bacillus stearothermophilus*	389,721	0.52	59A—*Methanosarcina mazeii*	60,692	0.45

*Contigs are marked by stars whereas sequences composed of several contigs are unmarked.

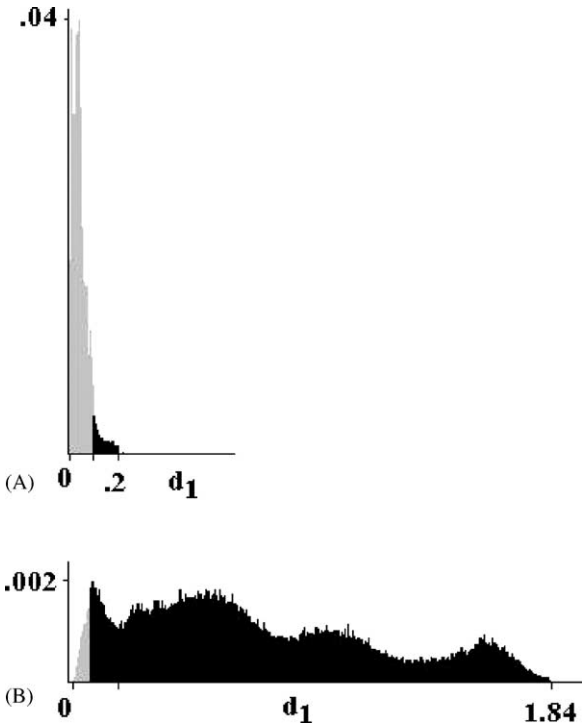

Fig. 2. Distribution of intra- (A) and intergenomic (B) distances d_1 for 48 species. Calculations were performed for 100 realizations of uniformly random sets of words of length 10, $|W| = 200$ (r-mismatching with $r = 2$ was used in calculations of the spectra).

4.3. Intra- and intergenomic relationships of compositional spectra

Compositional spectra may be used for characterization of intra- and intergenomic relationships. We found that, as a rule, intragenomic distances are smaller than intergenomic distances (a few exceptions exist and are partially discussed below). In other words, for every chosen set of words, distance between two fragments from the same genome tends to be smaller than distance between a fragment from the selected genome and any fragment belonging to another genome. In Fig. 2, we show histograms of intra- and intergenomic distances.

Let us consider an application of the method to the study of intragenomic heterogeneity. The analysis was performed on 22 complete prokaryotic genomes. Each genome was broken down into overlapped fragments of length of 300–400 kb with overlapping of 100 kb. Compositional spectra were computed for all these fragments, and two opposite cases are presented in Fig. 3. The case of small smooth changes in spectra (homogeneous spectral composition) is typical for a broad majority of the studied genomes and are presented in Fig. 3A. However, some species have non-homogeneous genomic composition, and our technique is sensitive to this feature. In Fig. 3B, we present spectral distribution of *Borrelia burgdorferi*. Some of the tested segments have

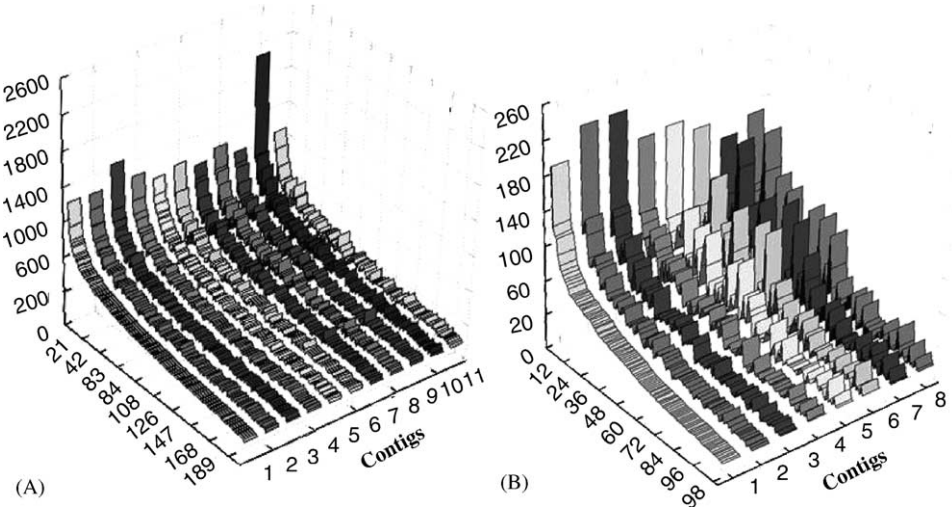

Fig. 3. Compositional spectra of contiguous fragments of complete genomes of *Mycobacterium tuberculosis* (A) and *B. burgdorferi* (B).

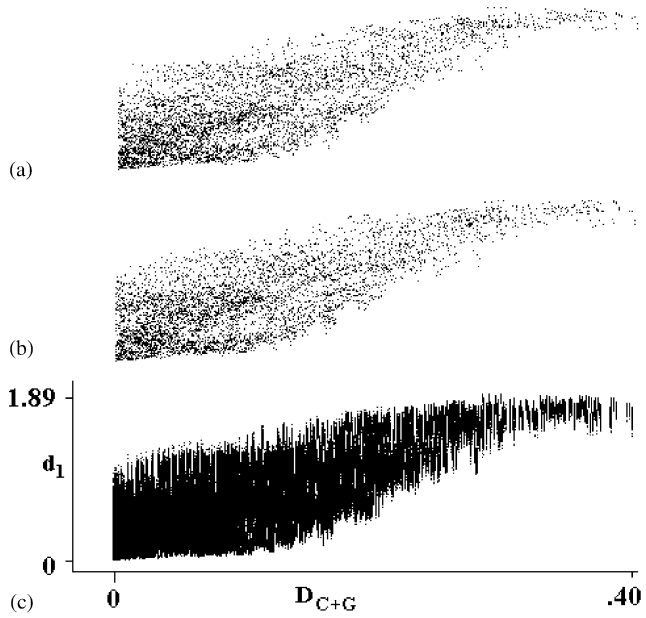

Fig. 4. Effect of $C+G$ content on d_1-distance between genomic sequences. Pairwise comparison of 48 genomic sequences (see description in Section 4.1) was conducted using random uniformly word sets W_i ($i=1,\ldots,100$), with $|W|=200$ and word length 10, as elsewhere. The derived d_1 values (Y-axis) are displayed against the absolute values of the differences in $G+C$ content (D_{C+G}—represented by X-axis). (A) Combined results of all pairwise ($48\times 47/2$) comparisons, each conducted over 100 sets W; (B) and (C) combined results of all pairwise comparisons for two different sets W taken randomly from the foregoing 100 sets.

different spectra and appeared to be far away from one another. This inconsistency can easily be explained by the recent finding of McInerney [22], who revealed that the genome of *B. burgdorferi* has two different patterns of codon usage. Thus, our method can identify heterogeneous genomes. Noteworthy differences among compositional spectra of compared sequences could not be explained exclusively by variation in $G+C$ content. For example, differences within 10%-range in $G+C$ composition of two fragments from the same genome do not have any considerable influence on their distance, whereas this may not be the case for intergenomic distances for certain pairs of species (Fig. 4). The figure shows that when D_{C+G}-distance is < 20–25% there is virtually no relationship between D_{C+G}- and d_1-distance, whereas big differences in $G+C$ content have a dominating influence on d_1.

5. Conclusions

Analyzing structural heterogeneity of DNA sequences at the supragenic level is considered one of the main targets of the current stage of genomic studies. This is an especially challenging problem because gene-coding material constitutes a relatively small part of the genome for most of the eukaryotes. The natural and simple approach proposed in this study seems to be especially useful for examining inter- 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 obtained pattern proved 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 spectra are multivariate descriptions of the genomes and genome fractions, hence invite a distance that could allow for quantitative comparisons between genomes.

References

[1] V. Brendel, H.G. Busse, Nucl. Acids Res. 12 (1984) 2561.
[2] V. Brendel, J.S. Beckmann, E.N. Trifonov, J. Biomol. Struct. Dyn. 4 (1986) 11.
[3] P. Pevzner, M. Borodovsky, A. Mironov, J. Biomol. Struct. Dyn. 6 (1989) 1013.
[4] S. Pietrokovski, J. Hirshon, E.N. Trifonov, J. Biomol. Struct. Dyn. 7 (1990) 1251.
[5] M.S. Gelfand, Biosystems 30 (1993) 277.
[6] D.B. Searls, Comput. Appl. Biosci. 13 (1997) 333.
[7] P.R. Sibbald, S. Banerjee, J. Maze, J. Theoret. Biol. 136 (1989) 475.
[8] A.O. Schmitt, H. Herzel, J. Theoret. Biol. 188 (1997) 369.
[9] A.K. Konopka, D. Chatterjee, Gene. Anal. Tech. 5 (1988) 87.
[10] E.N. Trifonov, in: R.H. Sarma, M.H. Sarma (Eds.), Making in Structure and Methods, Adenine Press, Albany, 1990, p. 69.
[11] A.E. Gabrielian, A. Bolshoy, Comput. Chem. 23 (1999) 263.
[12] A. Kalogeropoulos, Yeast 9 (1993) 889.
[13] R. Nussinov, Nucl. Acids Res. 10 (1980) 4545.
[14] S. Karlin, V. Brendel, Science 257 (1992) 39.
[15] A. Apostolico, M.E. Bock, S. Lonardi, X. Xu, J. Comput. Biol. 7 (2000) 71.

[16] G. Reinert, S. Schbath, M.S. Waterman, J. Comput. Biol. 7 (2000) 1.
[17] A.K. Konopka, in: D. Smith (Ed.), Biocomputing: Informatics and Genome Projects, Academic Press, San-Diego, 1994.
[18] M.S. Waterman, Introduction to Computational Biology, Chapman & Hall, London, 1995.
[19] C. Scapoli, A. Rodriguez-Larralde, S. Volinia, M. Beretta, I. Barrai, Comput. Appl. Biosci. 10 (1994) 465–470.
[20] S. Volinia, C. Scapoli, R. Gambari, R. Barale, I. Barrai, Nucl. Acids Res. 19 (1991) 3733.
[21] M.G. Kendall, Rank Correlation Methods, Charles Griffin & Co. Ltd., London, 1970.
[22] J.O. McInerney, Proc. Natl. Acad. Sci. USA 95 (1998) 10 698.