Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems 81 (2005) 208–222
BioSystems
www.elsevier.com/locate/biosystems

Large-scale genome clustering across life based on a linguistic approach

Valery Kirzhner [a,*], Alexander Bolshoy [a], Zeev Volkovich [b], Abraham Korol [a], Eviatar Nevo [a]

[a] *Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel*
[b] *ORT Braude College, Software Engineering Department, Carmiel 20101, Israel*

Received 10 April 2005; accepted 13 April 2005

Abstract

With the availability of genome sequences, the possibility of new phylogenetic reconstructions arises in order to reveal genomic relationships among organisms. According to the compositional-spectra (CS) approach proposed in our previous studies, any genomic sequence can be characterized by a distribution of frequencies of imperfect matching of words (oligonucleotides). In the current application of CS-analysis, we attempted to analyze the cluster structure of genomes across life. It appeared that compositional spectra show a clear three-group clustering of the compared prokaryotic and eukaryotic genomes. Unexpectedly, this grouping seriously differs from the classical Universal Tree of Life structure represented by common kingdoms known as Eubacteria, Archaebacteria, and Eukarya. The revealed CS-clustering displays high stability, putatively reflecting its objective nature, and still enigmatic biological significance that may result from convergent evolution driven by ecological selection. We believe that our approach provides a new and wider (compared to traditional methods) perspective of extracting genomic information of high evolutionary relevance.

Keywords: Comparative genomics; Sequence comparisons; Oligonucleotide; Occurrences; Clustering; Ecological convergence

1. Introduction

Evolutionary, structural, and functional perspectives comprise the major foci of current activity in molecular sequence analysis. The central problem of sequence analysis is gene organization. Another fundamental problem, the total genome organization, has also attracted attention and become important in recent years when long stretches of DNA and entire genomes were sequenced for many species, largely prokaryotic, but also eukaryotic organisms. Structural heterogeneity of DNA sequences at the supragenic level using the ever-increasing amount of sequences in databases can be analyzed opening new perspectives for relating structure to function. This is an especially challenging problem in light of the fact that gene-coding material comprises

* Corresponding author. Tel.: +972 4 8288040; fax: +972 4 8246554.
E-mail address: valery@esti.haifa.ac.il (V. Kirzhner).

0303-2647/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2005.04.003

a relatively small portion of the genome in eukaryotes. Clearly, the most straightforward approach to extract historical (evolutionary) information from sequence organization is to compare the coding part of the genome. This was the strategy of choice for the vast majority of molecular taxonomists. Woese attempted to systematize 16S rRNA data and to build the current "Universal Tree of Life" on that data (Woese, 1987; Woese et al., 1990; see also Doolittle, 1999). Many other efforts, mainly based on various protein sequences, failed to generate a consistent phylogeny. More recently, joint analysis of multiple protein groups was employed (e.g., Feng et al., 1997; Tekaia et al., 1999). The availability of genome sequences opened a new possibility of phylogenetic reconstructions of the relationships between organisms in the most objective way (Snel et al., 1999). Another way to integrate the genomic information is whole proteome analysis (Tekaia et al., 1999; Wolf et al., 2002; Lin and Gerstein, 2000), albeit the resulting dendrograms cannot be considered as phylogenetic trees.

Phylogenetic reconstruction from sequence information using only similarity assessments of aligned homologous genes or regions was critically discussed by Karlin and Cardon (1994), Karlin et al. (1997), Brocchieri (2001) and Gribaldo and Philippe (2002).

There are different methods based on counting oligomers in nucleotide and amino acid sequences. Such methods are analogous to the formal linguistic analysis of human texts. Naturally, one of the main applications of the methods based on counting words is the comparison of long genetic sequences, in particular, complete genomes. Such comparisons do not require any preliminary or posteriori alignments to reveal homologous fragments in both sequences. Therefore, the alternative name for the linguistic methods of genomic comparisons might be "alignment-free sequence comparisons". A linguistic tool for sequence analysis assumes some form of statistical summarization to facilitate numerical computations. An obvious motivation is in the simplicity of the DNA word frequency-based approach as compared to other computer intensive and operationally complex techniques (e.g., sequence alignment and homology). One form of statistical summarization is based on various frequencies of DNA k-words, which are k-tuples over the DNA alphabet $\{A, C, G, T\}$ (here $k \geq 1$ is a positive integer). We proposed a novel natural approach to characterize genomic sequences (Kirzhner et al., 2002, 2003). This attitude like the former linguistic measures is based on the counting occurrences of fixed words from a predefined sufficiently large set of words of a fixed predefined length. A measure based on this approach is called compositional spectrum (CS) and actually means a histogram of imperfect word occurrences.

According to the CS-approach, any genomic sequence can be characterized by a distribution of frequencies of imperfect matching of words from a set W that is called compositional spectrum (Kirzhner et al., 2002, 2003). In the current application of the compositional spectra, we have attempted to analyze the widely accepted taxonomic scheme that clusters all genomes across life into three kingdoms: Eukarya, Eubacteria and Archaea. It appeared that the large-scale clustering of genomes based on compositional spectra disagrees with the foregoing three-component system. The revealed CS-clustering displays high stability, putatively reflecting its objective nature, which may be affected by ecological stresses. In contrast, the "three-kingdom scheme" does not pass the test of cluster stability when the CS-distance is employed for genome comparisons.

2. Methods and sequences

2.1. Compositional spectrum of a DNA sequence

In our previous work (Kirzhner et al., 2002, 2003) we introduced the notion of *compositional spectrum*. Below is the essence of the CS-method.

Consider some word w of length L in the alphabet $\{A, T, C, G\}$ and sequence S in the same alphabet. Let sequence S contain a word x, which differs from word w not more than in r positions. Let us suppose that word x is an imperfect occurrence of w in S and this approximate matching can be denoted as "r-mismatching".

Let us consider now some set W of n words w_i of length L. Let each word w_i has m_i imperfect occurrences in a sequence S. Now let $M = \Sigma m_i$. The frequency distribution $F(W, S)$ of $f_i = m_i/M$ will be referred to as a "compositional spectrum" of the sequence S relative to the set W. According to (Kirzhner et al., 2002, 2003), let $L = 10$, $n = 200$, $r = 2$ and, also, let sequences

S represent sufficiently long fragments of a genome (≈103–105 bp). To produce a word set W, we employed a random generator assuming equal probabilities of appearance of each of the four nucleotides at any current position of any word.

2.2. Distance between DNA sequences based on compositional spectra

According to Kirzhner et al. (2002, 2003) we define d between two sequences S и S' as the distance between their spectra $F(W, S)$ and $F(W, S')$. Let us, in turn, define this distance between spectra as $d = 1 - \rho$ ($0 \leq d \leq 2$), where ρ is the Spearman rank correlation (Kendall, 1970). Note that the proposed measure will be proportional to the number of pair-wise 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.

2.3. Cluster analysis

In order to build the clusters two similar agglomerative methods have been employed. *WPGMA* (Sneath and Sokal, 1973) is the most common, weighted pair-group method average. The average distance is calculated from the distance between each point in a cluster and all other points in another cluster. The two clusters with the lowest average distance are joined together to form a new cluster. *Complete-linkage clustering* (see, among others, Sneath and Sokal, 1973) is also known as the maximum or furthest-neighbor method. The distance between two clusters is calculated as the greatest distance between members of the relevant clusters.

In order to verify the clustering results, we employed an additional method of clustering based on a different methodology, compared to the hierarchical one used in our major analysis. Namely, the partitioning method partition around matoid (PAM), or the *k*-medoids (Kaufman and Rousseeuw, 1990), was chosen. PAM procedure is known to find the best partition for a priory defined number of clusters.

2.4. Sequences

Genomic sequences of 49 genomes of Eukaryota, Eubacteria, and Archaea were used in this study. These included large stretches sampled from the data on 27 full genomes, and summed-up contigs of 13 partly sequenced genomes served to produce for each species two different target sequences, each of 200–500 kb (referred to as A and B). In a few cases, the available material was only sufficient to build one such target (A), so to maintain the same structure of the algorithms the B sequence was taken equal to A (see Appendix A).

3. Results: compositional-spectra distances and major separation (classification) of organisms

We tested the “three-kingdom” classification and some other species’ subdivision that we found by using the CS-distance (Fig. 1). According to this last structure, cluster I contains all the considered sequences from human genome, mouse, *Arabidopsis thaliana* mitochondrial genome, and sequences from several Eukarya and thermophilic Archaea, whereas cluster II contains some number of Eukaryotes and Prokaryotes, and cluster III contains prokaryotes (both Eubacteria and Archaea) and a free-living Eukaryote *Leishmania*. Therefore, in the obtained clustering, Eukarya sequences appeared in two out of three clusters, whereas Archaea are spread in all three clusters. Our aim is to demonstrate that this major tripartite subdivision indeed displays the main features of clustering, if the CS-distance is employed, and vice versa: the traditional “three-kingdom” scheme does not fit the clustering criteria if the CS-distance employed.

3.1. General procedure

In order to conduct large-scale comparisons of different species, we first generate 100 sets W_i ($i = 1, \ldots, 100$) of words using the method described in Section 2.1. For each set W_i the spectra of all the DNA sequences under consideration have been calculated, resulting in a matrix D_i of pair-wise distances between these sequences. Furthermore, by applying the described methods of clustering for each matrix D_i, we have obtained the hierarchic structures of embedded clusters. Such structures appear automatically when the agglomerative technique is applied: “start with the points as individual clusters and, at each step, merge the most similar or closest pair of clusters”.

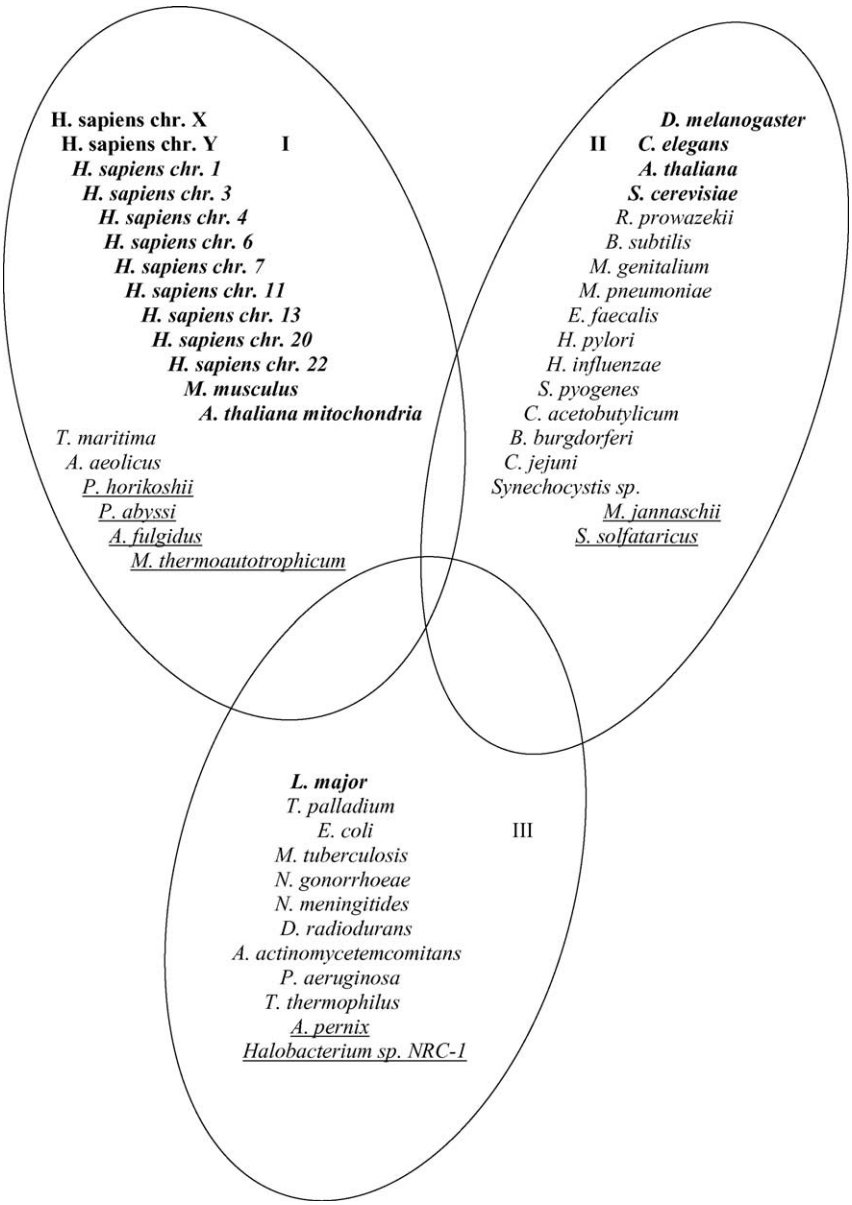

Fig. 1. Mixed-structure resulting from compositional-spectra distances. Cluster I contains all the considered sequences from the human genome, mouse, *A. thaliana* mitochondrial genome and sequences from several Eukarya and thermophilic Archaea, whereas cluster II contains some number of eukaryotes and prokaryotes, and cluster III contains prokaryotes (both Eubacteria and Archaea) and the free-living *Leishmania* (Eukarya: bold, Archaea: underline).

Let us first mark a certain pre-selected group of species, the proximity between which we are going to test. Then, by applying the *WPGMA* method, let us build step-by-step the system of clusters. At the initial state, each species is (by definition) considered an elementary cluster. At each stage, the algorithm joins the two nearest clusters (i.e., groups of species). This is a regular way of agglomerative algorithms' function-

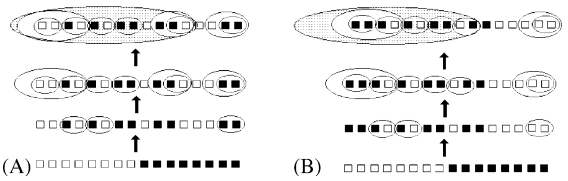

Fig. 2. Schematic presentation of *U*-cluster building. Marked elements are in black and non-marked in white. Marked elements are (A) poorly clustered (many non-marked elements entered *U*-cluster); (B) well clustered (only two non-marked elements entered *U*-cluster). In both cases, the process of cluster merging is shown that is terminated upon first appearance of *U* (shaded) cluster (harboring ≥75% of marked elements).

ing. However, we should not wait until the process is completed, i.e., when all species are united into one common cluster. Instead, we will stop the process on the stage when no less than 75% of the selected group of species can be united in the same cluster (let us denote it as *U*-cluster).

It should be noted that the *U*-cluster is the minimal (for its volume) set in the *WPGMA* integration process that includes no less than 75% of elements of the selected group of species. Clearly, the cluster structure obviously shows that if the selected species join directly during the clustering process, we might expect that the *U*-cluster will contain: (i) more than the required 75% of the marked species; but (ii) comparatively few species which do not belong to this group. Moreover, if in the course of cluster merging the marked species are going to join with the non-marked ones, the *U*-cluster will contain each of the marked species with some number of the non-marked ones. In Fig. 2 we demonstrate two such possible variants. Thus, the goal of the test is to evaluate in the *U*-cluster, the number of such non-marked species that will be absorbed by the *U*-cluster during the merging process required for binding 'almost all' marked species. The described test has been applied for the evaluation of the tested clusters stability.

Note that the chosen threshold level (75%) for stopping the clustering procedure was, to some extent, an arbitrary one. But it should be higher than 50% enabling the *U*-cluster to absorb the majority of marked elements; likewise, it should be less than 100% allowing for filtering out the elements randomly associated with the marked one. In all tests, the same threshold level (75%) was employed allowing for comparisons of the results across different experiments.

3.2. Large-scale clustering structures across life

3.2.1. Mixed-structure and its stability

Using the foregoing method we tested the cluster structure of the considered species shown in Fig. 1 (referred to as "mixed-structure", or *m*-structure). Then, this result was compared with the standard "three-kingdom-structure". It appeared that *m*-structure is highly stable with respect to the selected set of words W_i. We can demonstrate this by means of the following test. Let us mark all elements of the cluster I (except the mouse genome and *A. thaliana* mitochondrial genome) that, in our analysis, mainly included humans and thermophilic bacteria. For each set of words W_i (i=1, . . ., 100), the *U*-cluster was obtained, denoted as U_i, in order to indicate its dependence on the set W_i. Subsequently, the proportion of cases of participation of each species' across all U_i-clusters was derived (Fig. 3A). According to the results presented in Fig. 3A, all human sequences proved to belong to the U_i-clusters for each set W_i. The same can be stated about mouse, although the 100% presence of this genome was a nice surprise, because it was not marked as a member of the 'selected group'. Similarly, *A. thaliana* mitochondrial genome was not marked, but proved to belong to nearly 70% of U_i-clusters. Remarkably, thermobacterial sequences from cluster I appeared in more than 80% of U_i-clusters. A certain number of 'other' species (that did not belong to cluster I by *m*-structure) could also, in principle, join U_i-clusters. However, we were satisfied to find that the level of their "participation" in U_i-cluster did not exceed 20%. The same test was also conducted for two other clusters, II and III, of our *m*-structure. As one can see from Figs. 4A and 5A, the stability of these clusters manifested by U_i-cluster pattern, is similar to that found for cluster I. These results allow us to assume that the proposed *m*-structure indeed reflects some robust (objective) relationships among the compared species across life. If this is the case, one would expect that the foregoing stability reflected by the *U*-clusters will disappear if instead of the genuine clusters I, II, or III, we will take for similar considerations some mixtures of these clusters. That was exactly our next test.

We also employed another hierarchical clustering method, complete-linkage clustering. Within the framework of our procedure this method resulted in the same *m*-structure. For verification of the result,

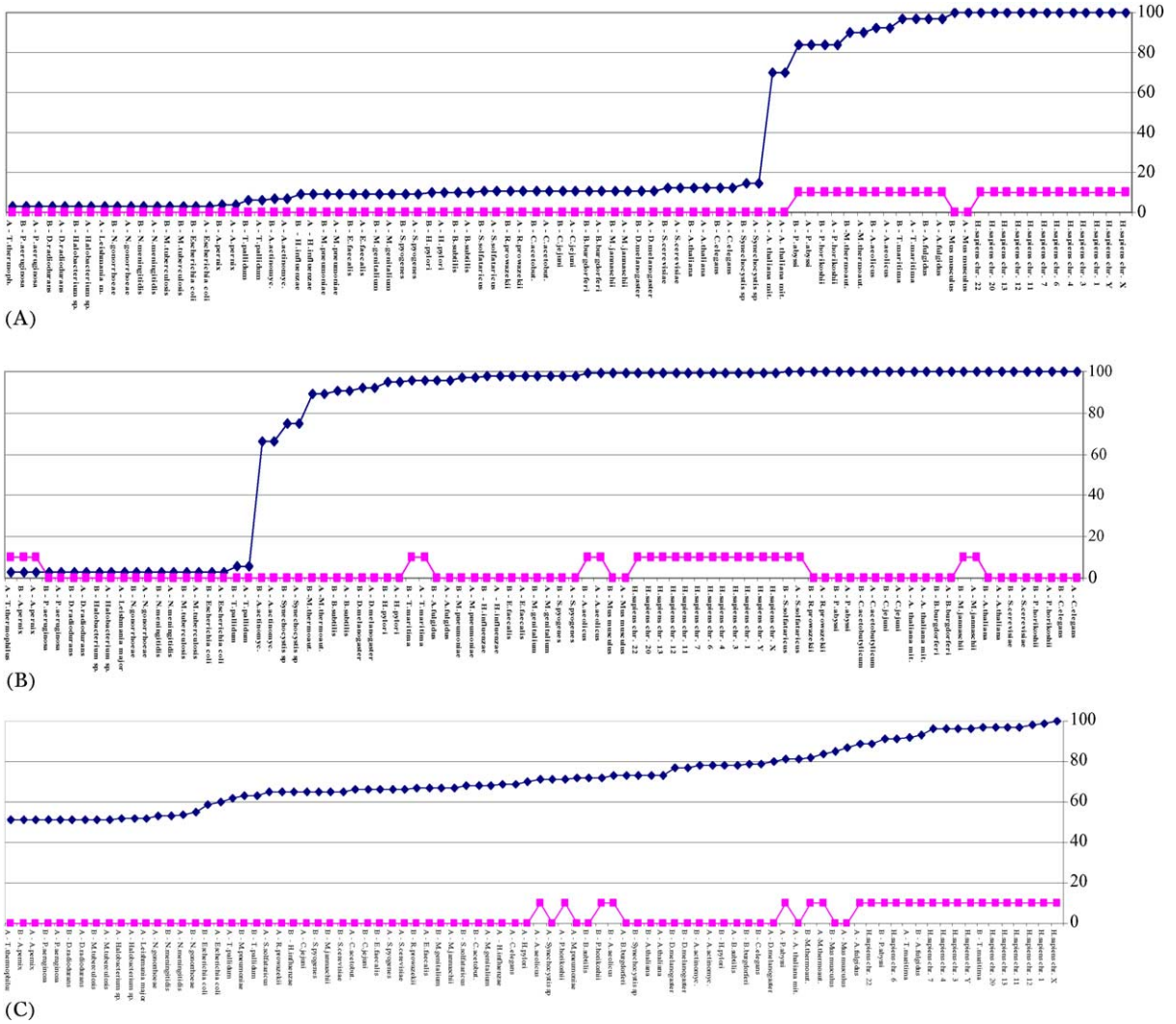

Fig. 3. Frequencies of species entry to cluster I. Axis *Y* denotes the percentage of species' participation in clusters I. Value 100 apparently means that the corresponding species belongs to each cluster U_i, whereas value 0 demonstrates that the species does not belong to any cluster. On axis *X* there are species names. For the sake of convenience of analysis they are organized according to the participation in clusters percentage decrease. The upper curve corresponds to the participation percentage, while the lower curve indicates 'selected species group'. For these elements, $y = 10$, for the rest, $y = 0$. (A) Evaluating the stability of species entries to cluster I. The results of averaging on 100 realizations (clusters U_i, see Section 3.1) are shown. It can be seen that the marked elements of cluster I enter in nearly all clusters U_i, whereas the remainder (no-marked) species appear randomly in not more than 20% of clusters U_i. (B) Reduced stability of species entries to cluster I, after replacing a part of marked species by species from other clusters (four marked thermophilic Archaea, *P. horikoshii*, *P. abyssi*, *A. fulgidus*, and *M. thermoautotrophicum* were replaced with three Archaea from clusters II and III, *A. pernix*, *M. jannaschii*, *S. solfataricus* and *T. thermophilus*). (C) Reduced stability of species entries to cluster I, after reshuffling one of two sequences representing each genome (sequence *A* of each species was reshuffled for *C* and *G* letters *A* without altering *GC*-content and *GC*-positions; sequence *B* of each species remained unchanged).

the partition method (PAM) was applied that is more robust and efficient than the known *k*-means algorithm. A series of partitions were obtained for different number of clusters (3–9), and the best correspondence to *m*-structure was found for seven-cluster partition (Fig. 6A). In this case, Cramer correlation coefficient between *U*-clusters and PAM-clusters was 0.9537. Note, that only cluster *P*7 overlaps with two

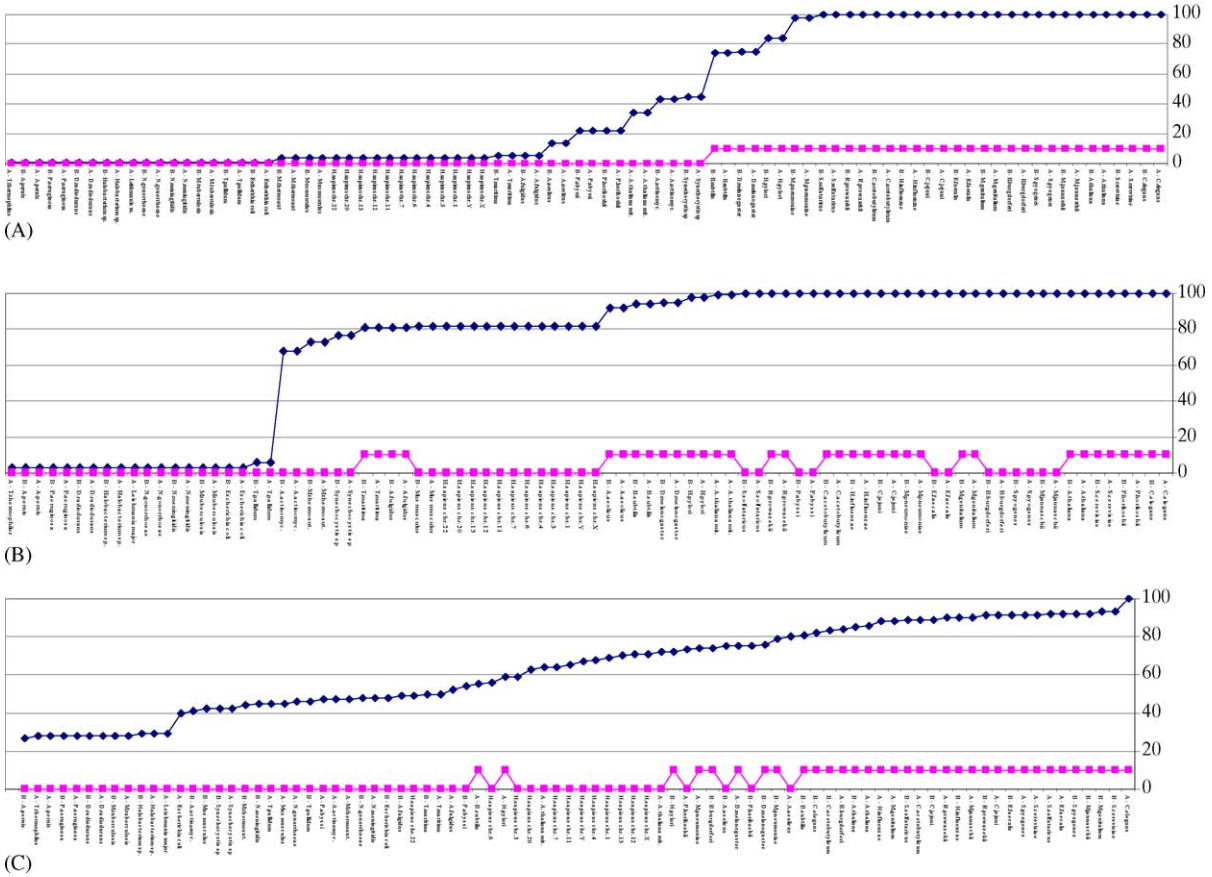

Fig. 4. Frequencies of species entry to cluster II. (A) Evaluating the stability of species entries to cluster I. The results of averaging on 100 realizations (clusters U_i, see Section 3.1) are shown. It can be seen that the marked elements of cluster I enter in nearly all clusters U_i, whereas the remainder (no-marked) species appear randomly in not more than 20% of clusters U_i. (B) Reduced stability of species entries to cluster II, after replacing a part of marked species by species from other clusters (five prokaryotes *M. jannaschii*, *S. solfataricus*, *S. pyogenes*, *B. burgdorferi*, and *E. faecalis* were replaced with five species from cluster I: *P. horikoshii*, *A. fulgidus*, *A. aeolicus*, *T. maritima*, and *A. thaliana* mit.). It is visible, that the new cluster now includes practically all considered species. (C) Reduced stability of species entries to cluster I, after reshuffling one of two sequences representing each genome (sequence A of each species was reshuffled for *C* and *G* letters *A* without altering *GC*-content and *GC*-positions; sequence *B* of each species remained unchanged).

U-clusters, whereas each one of the other clusters *P*1, . . ., *P*6 unequivocally correspond to only one of the *U*-clusters. For comparison, in Fig. 6B the correspondence between PAM partition on seven clusters and three-kingdom clustering is shown. Clearly, there is no one-to-one correspondence between the two types of clustering.

3.2.2. Effect of miss-anchoring

Let us again consider cluster I, in which the four thermophilic Archaea will be substituted with other four Archaea (from clusters II and III) that did not belong to this cluster. In other words, in addition to human sequences, we "marked" these four new Archaea genomes that actually should not belong to cluster I. Application of the proposed test to this mixed group gives a result shown in Fig. 3B. In contrast to the original clear clustering pattern (Fig. 3A), virtually no merging, leading to formation of a "non-trivial" *U*-cluster, was possible in the second case unless nearly all 100% of species were united for each set W_i. As before, the same test was conducted for the remaining two clusters of the *m*-structure with similar results (Figs. 4B and 5B).

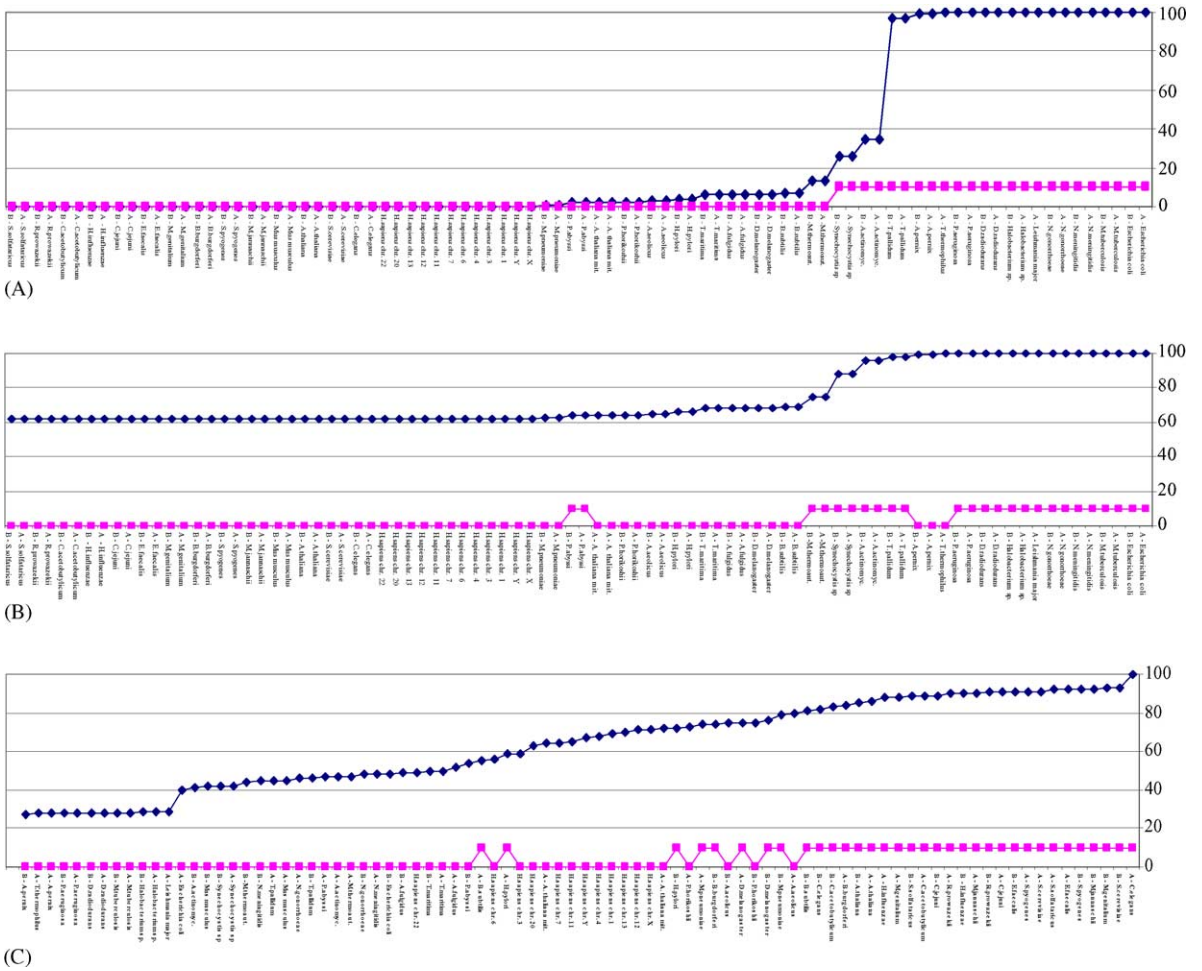

Fig. 5. Frequencies of species entry to cluster III. (A) Evaluating the stability of species entries to cluster I. The results of averaging on 100 realizations (clusters U_i, see Section 3.1) are shown. (B) Reduced stability of species entries to cluster II, after replacing a part of the marked species by species from other clusters (two prokaryotes *T. thermophilus* and *A. pernix* were replaced with two species from clusters I and II: *P. abyssi* and *T. pallidum*). (C) Reduced stability of species entries to cluster I, after reshuffling one of two sequences representing each genome (sequence *A* of each species was reshuffled for *C* and *G* letters without altering *GC*-content and *GC*-positions; sequence *B* of each species remained unchanged).

3.2.3. GC-permutation test

An important question is whether or not the revealed *m*-structure and the patterns manifested in *U*-clusters are derived from the similarity of the sequences with respect to their *GC*-content. This is a relevant question when compositional spectra are used as a basis for sequence comparisons. Nevertheless, in our previous work (Kirzhner et al., 2002, 2003) we have investigated this problem and demonstrated that *GC*-content affects the distance significantly only when the $(G + C)$-values vary greatly. However, even upon large variations in *GC*-content, the CS-distance remains sensitive to the origin of the sequence, i.e., for two sequences with some difference between their *GC*-contents, CS-distance is much smaller if the sequences belong to the same genome.

In the current study, the role of *GC*-content in formation of the analyzed clusters was tested directly.

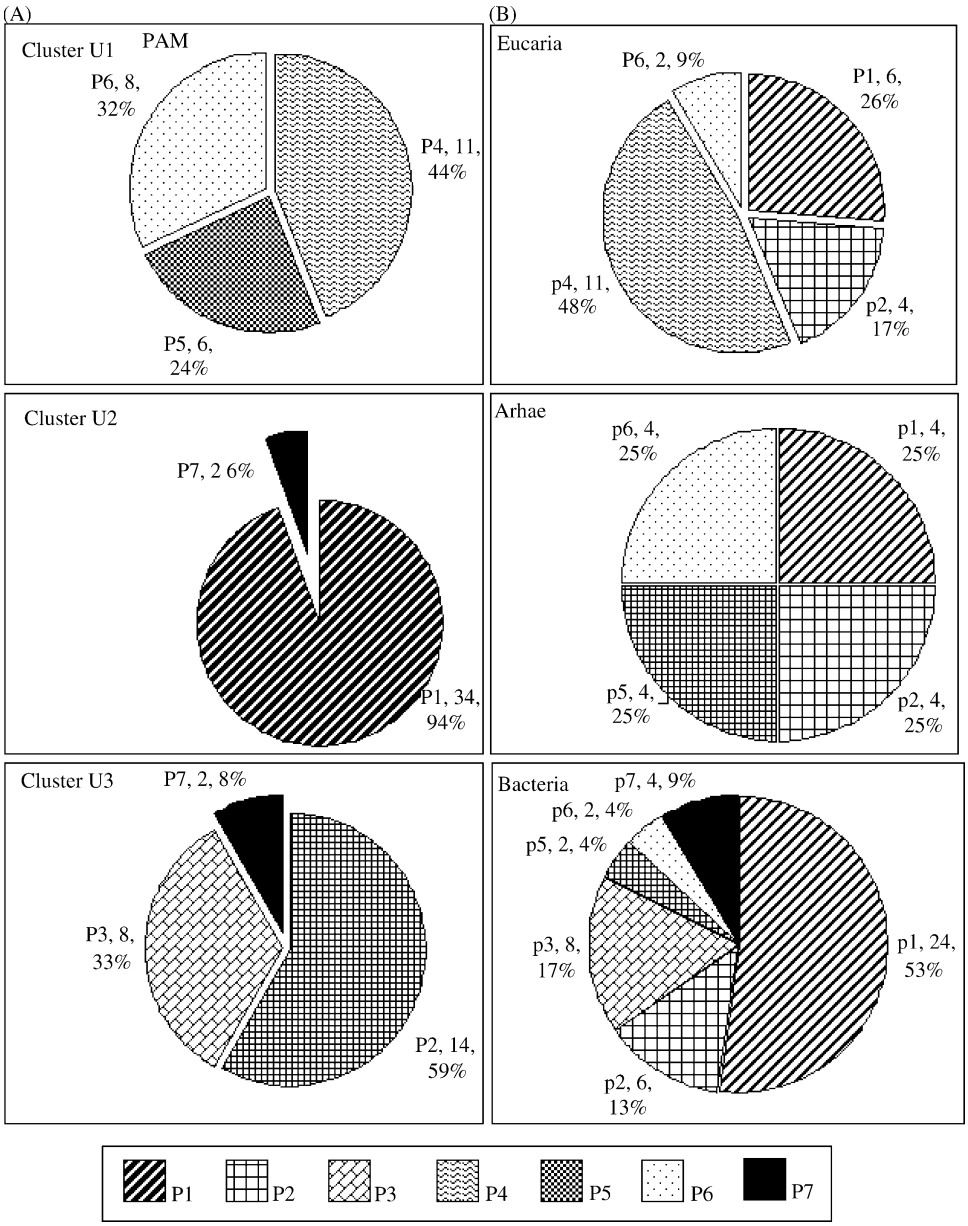

Fig. 6. Correspondence between PAM-clustering and each of *m*-structure or three-kingdom structure. *P*1–*P*7 denotes the clusters of PAM-partitioning. Each fragment is annotated by cluster number, its number of elements, and its proportion (%) in the cluster it belongs to. (A) Correspondence between *m*-structure and PAM; (B) correspondence between three-kingdom partition and PAM.

Namely, for almost all of the considered species, we employed two different sequences *A* and *B*. This allowed the use of a simple procedure: let us leave sequence *B* as it is, whereas sequence *A* can be transformed as follows. Consider '*GC*-positions' in sequence *A*, where there are letters *G* or *C*. Let the letters' frequencies relatively to their total number $(G+C)$ in sequence *A* equal p_G and $p_C = 1 - p_G$, re-

spectively. Now, let us interchange the letters *G* and *C* as follows: into each *GC*-position we put letters *G* or *C* with probabilities p_G and p_C, respectively. Thus, we randomly change the distribution of *G* and *C* letters in sequence *A* without altering *GC*-content and *GC*-positions. Similarly, let us interchange letters *A* and *T* in the same sequence *A*, leaving sequence *B* without changes. (For human sequences, 6 out of 12 contigs were considered as *A*, and 6 as *B*.) Our previous analysis was applied to the entire set of sequences *A* and *B*. As 'selected group of species' we have marked the same elements of cluster I, as was done before (see Fig. 3A). Remarkably, $G \leftrightarrow C$ and $A \leftrightarrow T$ permutations totally destroyed the previously revealed clear pattern (compare Fig. 3A and C). Analogous results were obtained for clusters II and III (compare Fig. 4A versus C, and Fig. 5A versus C). Therefore, the revealed *m*-structure and the respective patterns manifested in *U*-clusters, that presumably reflect some important aspects of species-genome similarities, cannot be explained by *GC*-content alone.

3.3. Poor clustering structure of the "three-kingdom" scheme

The aforementioned technique of *U*-cluster comparisons was also applied for testing the clustering quality of the "three-kingdom" scheme. The results are represented in Fig. 7. Unlike the *m*-structure that displayed a rather high robustness, each cluster of the three-kingdom scheme displayed a rather diffused organization. Indeed, if one marks the elements of one of the kingdoms, application of the described procedure generates corresponding *U*-cluster that includes practically all elements of the three clusters. Clearly, this indicates a much poorer organization of the "three-kingdom" scheme compared to that of the *m*-structure. We realize that this result derives from the method of sequence comparisons based on compositional spectra. However, the very fact that the pairwise CS-distances proved robust with respect to the sampled parts of the compared genomes, the selected set of words, words length, and other parameters, allows us to hypothesize that the *m*-structure does reflect some important relationships among species not represented by the more conventional "three-kingdom" scheme.

4. Discussion

4.1. Major schemes of molecular phylogenetic classification

Various sources of sequence information have been employed to analyze the phylogenetic relationships across life. The first phylogenetic trees based on DNA, RNA, and protein data employed relatively small and highly conserved sequences (Woese, 1987; Woese et al., 1990) and have lead to the foundation of the *Universal Tree of Life* concept. Despite its general attractiveness and biological tractability, this idea encountered various conceptual problems (putting aside numerous technical obstacles). Therefore, starting from comparisons of single genes (like rDNA) or proteins, it was extended to groups of genes (Wolf et al., 2002; Brown et al., 2001; Snel et al., 1999), the non-coding DNA (Rogozin et al., 2002), and even whole-genome structures (Lin and Gerstein, 2000; Fitz-Gibbon and House, 1999; Snel et al., 1999; Wolf et al., 2002). Consequently, it becomes obvious that different sets of reference genes (sequences) result in a wide variation of genomes clustering (Wolf et al., 2002; Brocchieri, 2001; Lin and Gerstein, 2000; Brown and Doolittle, 1997). In other words, the derived schemes based on different parts of the genome proved to result in different phylogenies. These inconsistencies became especially clear during the last few years when huge bulks of sequence information were generated by genomic projects. It is noteworthy that these results justify, to some extent, the assumptions that horizontal transfer of genetic material had a major effect on a large-scale evolutionary time (Campbell, 2000; Wolf et al., 2002). In light of such possibilities, it is highly desirable to reach reasonable robustness of phylogenetic reconstructions based on relative positioning of genes in the compared genomes (reviewed Wolf et al., 2002).

4.2. Compositional spectra

Our study is based on compositional-spectra analysis that can be considered an alternative to the class of more prevalent methods based on sequence alignment (see discussion in Brocchieri, 2001). The method and its possible applications were described earlier (Kirzhner et al., 2002, 2003). It is noteworthy that given the chosen parameters of our algorithm (word length

L, word number n in the set W, and the level of allowed mismatching r), the total length of the tested sequence covered by the set W (taking into account the frequency of words occurrence) covers, as a rule, much less than 1% of the sequence. Usually, we employ 200 such W sets to ensure statistical stability of the results, but in fact, these sets give a very good correspondence in the pair-wise distances between compared sequences (Kirzhner et al., 2002, 2003). Clearly, CS accumulates some information from both the coding and non-coding parts of the genome.

The contribution of the coding part of the genome to the "cladistic information" in the CS is presumably close to the one obtained by more standard approaches of species clustering based on sequence comparisons of all genes. However, the interconnection between CS (i.e., oligonucleotide abundances) and the genome sequence can be discussed also in the following formal sense. Various methods are now being proposed for recovering a total DNA sequence by measuring the occurrences of all words of some length (say, $L = 10–15$) in the sequence (e.g., Pevzner, 1989; Preparata et al., 1999; Reinert et al., 2000; Shamir and Tsur, 2002). From such a formal point of view, the CS-approach can be represented as a three-step treatment: (i) using Hamming distance, we define the neighborhoods of each word of length L, (ii) select an arbitrary subset W of words, and (iii) for each word $w_i \in W$ calculate the combined frequency of occurrence of w_i and all words of its neighborhood in the target sequence. Clearly, it would be impossible to restore the target sequence based on CS using set W of a small size (e.g., 200). It is also clear that the set of all sequences of a given length sharing the same CS should be relatively small. Nevertheless, in the limiting case of using the frequencies of all possible words all of length L together with their neighborhoods is equivalent to the basic situation when the frequency of occurrence is known for each word separately. The latter allows one to restore the target sequence effectively (Pevzner, 1989; Preparata et al., 1999). In other words, CS contains in a holographic-like (nutshell) form the information about the structure of the targeted sequence. We can conclude that, by accumulating various kinds of signals, CS characterizes the sequence (genome) "in general", allowing quantifying differences and similarities between species. The latter may derive from a joint effect of phylogenetic divergence, ecological (convergent and divergent) adaptations, horizontal transfer, etc.

This consideration clarifies, to some extent, the foregoing m-structure obtained in our species clustering based on CS. Our results undoubtedly demonstrate that the classical "three-kingdom" scheme is not reflected in the information based on word occurrences. On the contrary, the m-structure proved to be robust with respect to the arbitrary chosen "vocabularies" W indicating the existence of some objective information absorbed by CS that normally remains unexploited by more standard taxonomic methodologies that result in the "three-kingdom" structure.

4.3. M-structure: bridging to other classifications

The concept of the Universal Tree of Life suggests that life is divided into three primary groupings, commonly known as Eubacteria, Archaebacteria, and Eukaryotes. This simple and clear message underwent some erosion due to subsequent studies based on wider molecular-genetic material (e.g., Brocchieri, 2001). Thus, using whole proteome comparisons Tekaia et al. (1999) derived phenograms where Archaebacteria shared one cluster with Eubacteria, but not with Eukaryotes. A tree built on the overall occurrence of orthologs also proved different from the traditional ribosomal phylogeny, although a traditional tree could be derived based on a certain group of proteins (Lin and Gerstein, 2000). It is interesting that different phenograms may be obtained by using slow-evolving and fast-evolving molecular positions (Brinkmann and Philippe, 1999). The dependence of phenogram variability on the chosen sets of targeted sequences was observed in other studies as well (Golding and Gupta, 1995; Brown and Doolittle, 1997). Brown et al. (2001) also showed that a "proper" choice of molecular data may result in the three-kingdom scheme, but in fact, Archaebacteria and Eukaryotes showed a higher proximity to each other than to Eubacteria cluster. In discussing the taxonomical relationships between these major groups of organisms, Mayr indicates "that the transcription, translation, and splicing machineries of the Archaebacteria resemble those of the Eukaryotes, while the majority of the functional genes, coding primarily for metabolic enzymes, transport systems and enzymes of cell wall biogenesis resemble the Eubacterial ones" (Mayr, 1998).

Let us compare our *m*-structure with the phenogram from Brown et al. (2001) that fits the three-kingdom scheme. Three major features of *m*-structure deserve attention here: (1) *Methanococcus jannaschii* belongs to another cluster other than Archaea; (2) Eubacteria *Thermotoga maritima* and *Aquifex aeolicus* lie in the same cluster as mammalia and Archaea (Fig. 1I); and Eukarya appeared in two different clusters (Fig. 1, I and II). In connection with the features (1), it is worth mentioning that 44% of the *M. jannaschii* (Archaea) gene products showed significantly higher similarity to bacteria than to eukaryotic proteins as their closest homologues (Koonin et al., 1997). According to our tests, the first cluster was very robust (Fig. 3A). Besides mammalia, it includes some thermophilic bacteria that proved, with no exceptions, to be also anaerobic. According to Forterre et al. (2002) the archaeal ancestor was probably a hyperthermophilic anaerobe. Replacing a part of these Archaea by aerobic Archaea results in trivialization of cluster I (Fig. 3B). Concerning feature (2), we believe that the proximity of the mentioned thermophilic Eubacteria to Archaea is more adequate than their assignment to Eubacteria (as done by Brown et al., 2001). Note that based on genome-wise comparisons, these two Eubacteria are closer to Archaea (Nelson et al., 1999).

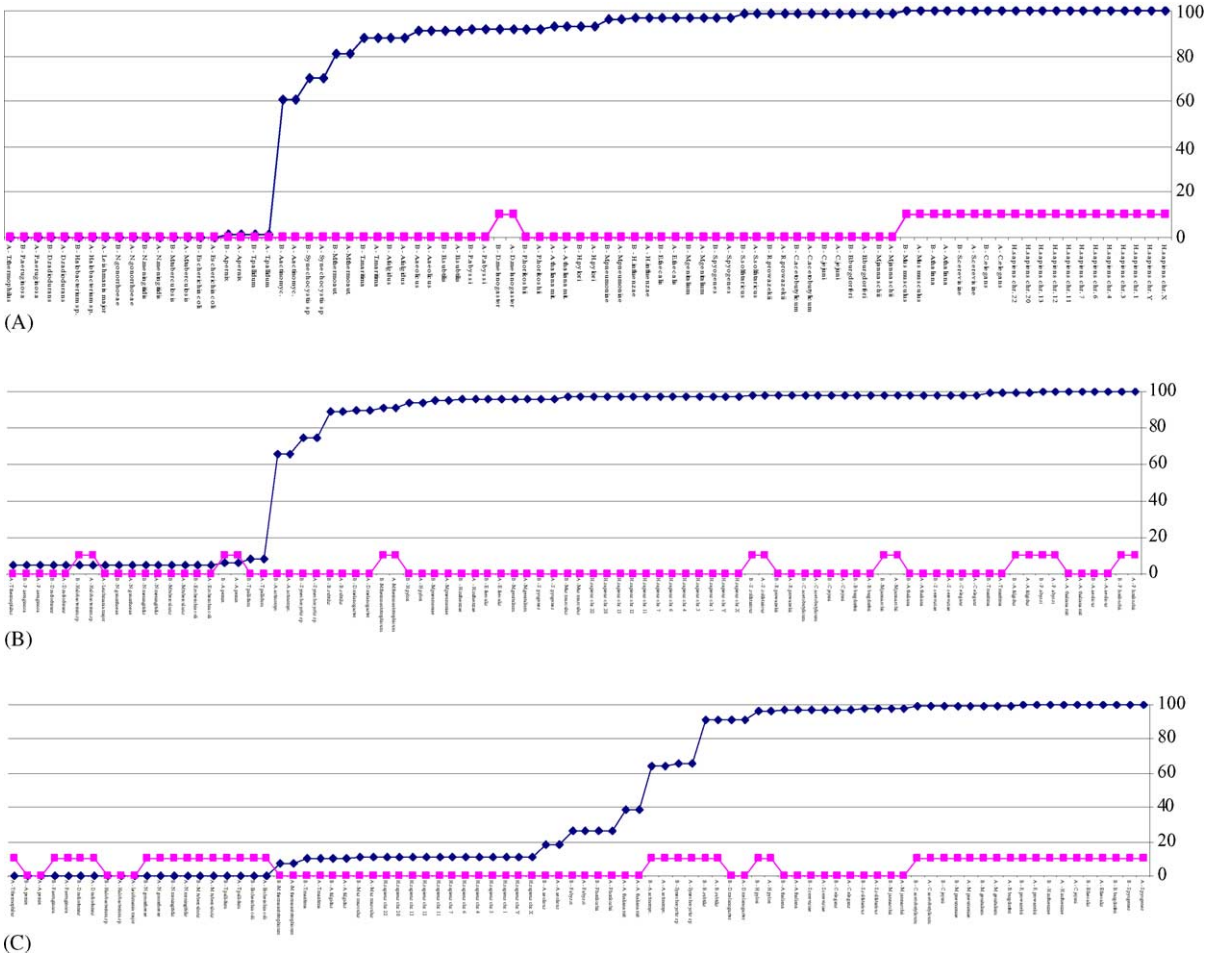

Fig. 7. Instability of species entries to the clusters of the three-kingdom scheme. The results of averaging on 100 realizations (clusters U_i, see Section 3.1) are shown. It can be seen that nearly all species (independent on "marked" or "non-marked" identifier) appear in nearly all clusters U_i: (A) Eukarya; (B) Archeabacteria; and (C) Eubacteria.

Concerning the feature (3) of *m*-structure, we would like to recall a statement by Golding and Gupta (1995) that "the eukaryotic cell nucleus formed is from the fusion of an archaebacteria and a gram-negative bacteria." If so, the appearance of eukaryotic species in a shared cluster with Eubacteria (Fig. 1II) may derive from their some "relatedness" to gram-negative bacteria. Note that gram-negative bacteria also showed up in the third cluster free of eukaryotes, but all gram-negatives from cluster II are anaerobic whereas from cluster III they are all aerobic. As a contrast to the foregoing stability of the *m*-structure, we would like to again discuss our results of the application of the proposed approach to the known "three-kingdom scheme". The results showed a very poor clustering (Fig. 7). Indeed, in order to include no less than 75% of "marked" (anchored) species for any of the three kingdoms, we also had to include nearly all (>90%) of the remainder ("non-marked") species from the other two kingdoms.

5. Conclusions and prospects

The analysis presented in this paper indicated that the classical "three-kingdom scheme" does not pass the test of cluster stability when the CS-distance is employed for genome comparisons. This result is not surprising because different characteristics of the genome (e.g., specific genes, rRNA, intergenic spacers, and even gene number in the genome) taken as a reference for taxonomic reconstruction, have resulted in quite different phenograms (Golding and Gupta, 1995; Tekaia et al., 1999; Brinkmann and Philippe, 1999; Lin and Gerstein, 2000; Brocchieri, 2001; Wolf et al., 2002). It becomes evident that using any one genomic parameter for such reconstructions may be questioned. Moreover, it was suggested that more objective synthetic classifications are possible only if genomic sequences are complemented by other evolutionarily valuable information (Mayr, 1998; Gupta, 1998; Karlin et al., 2002). Our results also suggest that there are diverse interrelationships between genomes that only partially can be accounted for by the "three-kingdom scheme". In this sense, the CS-analysis can be considered as one more step toward the synthetic classification. Of major interest for future analysis, would be employing the discriminating ability of CS-analysis to unravel the biological meaning of the detected robust patterns, like *m*-structure. In particular, we are interested to test whether ecological determinants play a role in these patterns, e.g., in the form of convergent ecological selection.

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 name of 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 "*Mycobacterium tuberculosis* (A: 1199341 (358717), 0.65; B: 2579341 (358913), 0.65)" means that *M. tuberculosis* 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.

A.1. Eukaryota

Homo sapiens chr. X (NT 011528) (539188, 0.40); *Homo sapiens* chr. Y (NT011864) (539595, 0.40); *Homo sapiens* chr. 1 (NT 004302) (539495, 0.36); *Homo sapiens* chr. 3 (NT 002444) (543554, 0.46); *Homo sapiens* chr. 4 (NT 006051) (535847, 0.44); *Homo sapiens* chr. 6 (NT 007122) (599072, 0.43); *Homo sapiens* chr. 7 (NT 007643) (447791, 0.36); *Homo sapiens* chr. 11 (NT 008933) (506578, 0.37); *Homo sapiens* chr. 13 (NT 009796) (537882, 0.38); *Homo sapiens* chr. 20 (NT 011328) (540270, 0.44); *Homo sapiens* chr. 22 (NT 001454) (701877, 0.50); *Mus musculus* (A: chr7, AC012382, 276523, 0.44; B: chr.11, AL603707, 234182, 0.49); *Caenorhabditis elegans* (A: chr1, 1-438825, 0.36; B: chr2, 1-350000, 0.36); *Drosophila melanogaster* (A: chr.2 AE003641, 299556, 0.42; B: chr. X, AE003506, 300000, 0.43); *Arabidopsis thaliana* (A: chr.1, NC 003071.1 100000-531319, 0.36; B: chr.1, NC 003075.1, 400000–727859, 0.36); *A. thaliana* mitochondrial genome (A: NC

001284.1, 366923, 0.45); *Saccharomyces cerevisiae* (A: chrii, 1-800000, 0.38; B: chrxv, 1-800000, 0.39); *Leishmania major* (A: AE001274, 1-171770, 0.62).

A.2. Eubacteria

Bacillus subtilis (A: 1199941 (579647), 0.43; B: 2219941 (399002), 0.41); *Streptococcus pyogenes* (A: 239941 (690238), 0.38; B: 1079941 (696345), 0.39); *Mycoplasma genitalium* (A: 1 (287593), 0.33; B: 278581 (288000), 0.30); *Mycoplasma pneumoniae* (A: 239941 (199523), 0.40; B: 539941 (199523), 0.40); *Mycobacterium tuberculosis* (A: 1199341 (358717), 0.65; B: 2579341 (358913), 0.65); *Synechocystis* sp. (A: 719941 (349960), 0.48; B: 2699941 (350000), 0.47); *Helicobacter pylori* (A: 599941 (320335), 0.39; B: 1439941 (320387), 0.39); *Escherichia coli* (A:599941 (519942), 0.51; B: 2999941 (542976), 0.51); *Deinococcus radiodurans* (A: 599941 (399971), 0.67; B: 1799941 (399983), 0.66); *Thermotoga maritima* (A:59941 (370054), 0.46; B:1259941 (366191), 0.46); *Aquifex aeolicus* (A: 599941 (399976), 0.43; B: 1199941 (400002), 0.44); *Neisseria meningitides* (A: 599941 (361259), 0.51; B: 1199941 (373905), 0.52); *Neisseria gonorrhoeae* (A: 350020, 0.53; B: 355192, 0.54); *Campylobacter jejuni* (A: 59341 (399984), 0.31; B: 1079341 (400002), 0.30); *Haemophilus influenzae* (A: 119941 (399863), 0.38; B: 1139941 (399981), 0.38); *Clostridium acetobutylicum* (A: 315781 (347567), 0.32; B: 3000001 (340105), 0.31); *Treponema pallidum* (A: 59941 (275933), 0.52; B: 719941 (275984), 0.53); *Pseudomonas aeruginosa* (A: 720001 (345206), 0.65; B: 1920001 (355163), 0.67); *Actinobacillus actinomycetemcomitans* strain HK1651 (A: 6000 (338681), 0.45; B: 800000 (344947), 0.45); *Rickettsia prowazekii* (A: 239941 (276000), 0.29; B:719941 (276000), 0.29); *Borrelia burgdorferi* (A: (1) 399967, 0.29; B: 400000 (399979), 0.29).

A.3. Archaea

Halobacterium sp. NRC-1 (A: 240001 (191652), 0.62; B: 64001 (211652), 0.64); *Pyrococcus horikoshii* (A: 240001 (399992), 0.42; B: 840001 (400002), 0.42); *Pyrococcus abyssi* (A: 360000 (360000), 0.45; B: 1200001 (360000), 0.45); *Archaeoglobus fulgidus* (A: 12061 (399986), 0.48; B: 1200061 (400002), 0.48); *Methanococcus jannaschii* (A: 120001 (399868), 0.32; B: 840001 (399977), 0.31); *Methanobacterium thermoautotrophicum* (A:599941 (344374), 0.49; B: 1199941 (344455), 0.50); *Aeropyrum pernix* (A: 360061 (400002), 0.58; B: 960061 (400002), 0.57); *Sulfolobus solfataricus* AE006641 (A: 400001 (584947), 0.36; B: 1220002 (308779), 0.36).

References

Brinkmann, H., Philippe, H., 1999. Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16 (6), 817–825.

Brocchieri, L., 2001. Phylogenetic inference from molecular sequences: review and critique. Theor. Popul. Biol. 59 (1), 27–40.

Brown, J.R., Doolittle, W.F., 1997. Archaea and the prokaryote-to-eukaryote transition. Microbiol. Mol. Biol. Rev. 61, 456–502.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E., Stanhope, M.J., 2001. Universal trees based on large combined protein sequence data sets. Nat. Genet. 28, 281–285.

Campbell, A.M., 2000. Lateral gene transfer in prokaryotes. Theor. Popul. Biol. 57 (2), 71–77.

Doolittle, W.F., 1999. Phylogenetic classification and the universal tree. Science 284, 2124–2128.

Feng, D., Cho, G., Doolittle, R.F., 1997. Determining divergence times with a protein clock: update and reevaluation. Proc. Natl. Acad. Sci. U.S.A. 94, 13028–13033.

Fitz-Gibbon, S.T., House, C.H., 1999. Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. 27 (21), 4218–4222.

Forterre, P., Brochier, C., Philippe, H., 2002. Evolution of the Archaea. Theor. Popul. Biol. 61 (4), 409–422.

Golding, G.B., Gupta, R.S., 1995. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol. Biol. Evol. 12 (1), 1–6.

Gribaldo, S., Philippe, H., 2002. Ancient phylogenetic relationships. Theor. Popul. Biol. 61 (4), 391–408.

Gupta, R.S., 1998. Life's third domain (Archaea): an established fact or an endangered paradigm? Theor. Popul. Biol. 54 (2), 91–104.

Karlin, S., Brocchieri, L., Trent, J., Blaisdell, B.E., Mrazek, J., 2002. Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. Theor. Popul. Biol. 61 (4), 367–390.

Karlin, S., Cardon, L.R., 1994. Computational DNA sequence analysis. Annu. Rev. Microbiol. 48, 619–654.

Karlin, S., Mrazek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179 (12), 3899–3913.

Kaufman, L., Rousseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York.

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

Kirzhner, V., Korol, A., Bolshoy, A., Nevo, E., 2002. Compositional spectrum—revealing patterns for genomic sequence characterization and comparison. Physica A 312, 447–457.

Kirzhner, V., Nevo, E., Korol, A., Bolshoy, A., 2003. One promising approach to a large-scale comparison of genomic sequences. Acta Biotheor. 51 (2), 73–89.

Koonin, E.V., Mushegian, A.R., Galperin, M.Y., Walker, D.R., 1997. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. Mol. Microbiol. 25 (4), 619–637.

Lin, J., Gerstein, M., 2000. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. Genome Res. 10 (6), 808–818.

Mayr, E., 1998. Two empires or tree? Proc. Natl. Acad. Sci. U.S.A. 95, 9720–9723.

Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., Fraser, C.M., et al., 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature 399 (6734), 323–329.

Pevzner, P.A., 1989. 1-Tuple DNA sequencing: computer analysis. J. Biomol. Struct. Dyn. 7 (1), 63–73.

Preparata, F., Frieze, A., Upfal, E., 1999. Optimal reconstruction of a sequence from its probes. J. Comput. Biol. 6 (3–4), 361–368.

Reinert, G., Schbath, S., Waterman, M.S., 2000. Probabilistic and statistical properties of words: an overview. J. Comput. Biol. 7 (1–2), 1–46.

Rogozin, I.B., Makarova, K.S., Natale, D.A., Spiridonov, A.N., Tatusov, R.L., Wolf, Y.I., Koonin, E.V., 2002. Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. Nucleic Acids Res. 30 (19), 4264–4271.

Shamir, R., Tsur, D., 2002. Large scale sequencing by hybridization. J. Comput. Biol. 9 (2), 413–428.

Sneath, P.H.A., Sokal, R.R., 1973. Numerical Taxonomy, the Principles and Practice of Numerical Classification. W.H. Freeman, San Francisco.

Snel, B., Bork, P., Huynen, M.A., 1999. Genome phylogeny based on gene content. Nat. Genet. 21 (1), 108–110.

Tekaia, F., Lazcano, A., Dujon, B., 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. 9, 550–557.

Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51, 221–271.

Woese, C.R., Kandler, O., Wheelis, M.L., 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukarya. Proc. Natl. Acad. Sci. U.S.A. 87, 4576–4579.

Wolf, Yu.I., Rogozin, I.B., Grishin, N.V., Koonin, E.V., 2002. Genome trees and the tree of life. Trends Genet. 18 (9), 472–479.