ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition ▮▮▮ (▮▮▮▮) ▮▮▮–▮▮▮

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

The method of *N*-grams in large-scale clustering of DNA texts

Z. Volkovich[a,*], V. Kirzhner[b], A. Bolshoy[b], E. Nevo[b], A. Korol[b]

[a] *Department of Software Engineering, ORT Braude College, P.O. Box 78, Karmiel, 20101, Israel*
[b] *Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel*

Received 30 July 2004; received in revised form 2 May 2005; accepted 2 May 2005

Abstract

This paper is devoted to the techniques of clustering of texts based on the comparison of vocabularies of *N*-grams. In contrast to the regular *N*-grams approach, the proposed *N*-grams method is based on calculation of imperfect occurrences of *N*-grams in a text up to a number of mismatched strings. We demonstrated that such an approach essentially improves the resolving capacity of the *N*-grams method for DNA texts. Additionally, we discuss a mutual usage scheme of different clustering technique types to verify the partition quality.

Keywords: *N*-grams; Strings mismatching; Clustering; Genome comparisons; Compositional spectra

1. Introduction

Clustering problems arise in various areas of text mining and information retrieval. Typically, given a language each document is reduced to representation by a vector of frequencies of terms selected in an appropriate way. The next step relates to finding a suitable distance between the vectors, such as Euclidian, Manhattan, Covariance distances, etc. (e.g. Refs. [1,2]), which could provide a reasonable division. Finally, a clustering based on the chosen distance is performed by means of one of the partitioning methods, e.g. resembling the *k*-means method [3] or *k*-medoids procedures [4].

1.1. *N*-grams technique

Probably, Shannon [5] in 1948 was the first to employ *N*-grams for characterizing texts (he also proposed the term *N*-gram). He was speaking about a "discrete source as generating the message, symbol by symbol" [5]. By Shannon, this could be a message written in a natural language, continuous information sources that have been rendered discrete (for example, speech) and, more generally, an abstract stochastic process which generates a sequence of symbols. In particular, Shannon's *N*-grams were defined as formal words (i.e. not related to their real values). Informally speaking, if an object can be represented by a sequence over the taken alphabet A, then one way of performing a feature extraction is to describe it in terms of its subsequences. An *N*-gram is a subsequence of length *N*. Despite its initial narrow usage in the theory of communication, *N*-grams were later applied more widely including such fields as classification of different "texts" like messages in natural and artificial languages, music, images, etc.

Depending on the application field, this approach may use *N*-grams of different length—for example, from 2 to 100 letters. The effect of *N*-grams lengths is discussed in detail in the literature. For instance, the Stores system [6] suggests the value of $N = 3$ because it yields the best selectivity in the search access rate. Other systems have used trigrams in order to conserve memory or disk accesses [7]. Cavnar [8]

[*] Corresponding author. Tel.: +972 4 990 1862; fax: +972 4 990 1852.
E-mail address: zeev@actcom.co.il (Z. Volkovich).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.05.002

employed bigrams and trigrams together in the same system assuming that bigrams provide better matching for individual words while trigrams provide the connections between words to improve phrase matching, thus complementing each other. Cohen [9] and Damashek [10] used 5-grams, while Robertson and Willett [11] used bigrams and trigrams, with no reasons provided for these choices. It is also possible to use initially k-grams and then move to $(k-1)$-grams to improve the results. Thus, Huffman and Damashek [12] and Huffman [13] reached about 20% improvement of garbled text. For Korean text retrieval, bigrams applied in combination with N-grams provided the best 11-point average precision [14].

In general, a comparison of linear symbolic sequences based on the N-grams technique proved to be effective regardless of the origin of the sequence processed including imaging, voice processing, and music. Any such "text" is reduced to a vocabulary of the frequencies of N-grams along the whole sequence or its pieces. Proximity between such vocabularies can then be used for text comparisons.

In this paper, we apply the N-grams technique for classifying DNA sequences considered as text over the four-letter alphabet $\mathfrak{U}=\{\mathrm{A},\ \mathrm{C},\ \mathrm{G},\ \mathrm{T}\}$. One can consider formal words of varying lengths L: word means a string of length L over the alphabet $\mathfrak{U}$. Clearly, these words are exactly Shannon's N-grams. However, in order to retain the standard terminology evolved in the field of bioinformatics, we continue to use the term "word" for discussing concrete biological situations and N-grams in general contexts.

1.2. Clustering technique

Generally speaking, most of the existing clustering methods can be categorized into three groups: partitioning, hierarchical, and density-based approaches. We apply here partitioning and hierarchical methods only. Partitioning methods have the advantage of being able to incorporate knowledge about the size of the clusters by using certain templates and the elements' dissimilarity in the objective function. Such an algorithm is guaranteed to produce clustering for any data although there is currently no generally accepted way to test the null hypothesis of no clustering (e.g. that the data are distributed uniformly).

In addition, the known hierarchical clustering procedures yield a nested sequence of partitions and, as a rule, avoid specifying how many clusters are appropriate. This is achieved by providing a partition from cutting the tree (dendrogram) at some level. Inner statistical tests (see general overview in Ref. [15]) could hardly serve a guide on where to cut the dendrogram. On the other hand, partitioning methods may produce a tighter cluster structure than hierarchical ones and are computationally faster with a larger number of variables in the case of a small number of clusters. Such methods do not usually do well with non-globular clusters, and the difference between various partitioning methods lies in the strategies of making a compromise to find suboptimal solutions. In fact, different methods could yield diverse results. Even with a specific method, the solutions are usually sensitive to initial conditions. Both clustering types may also be used together [16].

1.3. Methods of comparison of DNA texts based on N-grams

It is important to realize that the four chemical bases of DNA are exactly the same in all living organisms. Each DNA fragment may be sequenced and its sequence will authentically and adequately represent it. We apply the approach based on the N-grams technique to cluster DNA sequences representing complete genomes. The genome is the total genetic constitution of an organism. Excluding those of viruses, genomes consist of one or more double-stranded molecules of DNA. Certainly DNA fragments, short or long, are molecules (biopolymers) but from an information point of view these molecules may be fully characterized by their sequences and DNA sequences are linear texts over the four letter alphabet: {A, T, G, C}. A complete genome of a relatively simple organism is a pretty long text, hundreds of thousands for the smallest; usually millions of letters. Hereditary fragments of a genome are called genes. The vast majority of genes code for messenger RNAs (mRNAs) that are translated into proteins, with the collection of all protein-coding genes within a genome referred to as the proteome. Some methods of genome classification are based on reduction of a whole genome to its proteome exclusively. The genome also contains a small set of genes coding for structural RNAs. In both prokaryotes and eukaryotes, such RNAs play critical roles in many functions. Collectively, all the protein-coding and structural RNA-coding genes constitute the genic DNA of a genome. For prokaryotes, this is the bulk of the entire genome. In eukaryotes, genic DNA comprises only a fraction (and in some cases a very small fraction) of the total genome.There are methods of genome taxonomy based on comparison of chosen genic elements, protein coding or RNA-coding. However, there are also methods of sequence comparisons unrelated to genomic functional structure. Frequently such methods are called linguistic because they use various N-grams-based techniques.

In one of the recent reviews [17], a few methods of DNA sequence analysis based on counting textual N-grams were presented. As we mentioned above, the most accepted approach to genome comparison is to calculate similarity among one or more pairs of homologous genes of the genomes. This is the strategy of choice in molecular evolution studies. The methods based on counting word occurrences cannot replace the methods based on investigation of homologies. Nevertheless, they serve as a supporting approach in comparative genomics and molecular evolution. One form of statistical summarization is based on analyzing frequencies of oligonucleotides (DNA N-grams). For example, considering N equal to four, there are 256 different 4-grams (256 tetranucleotides over the DNA alphabet). One can treat the set of tetranucleotide frequencies as a statistical summary of the given DNA sequence. By using the frequency distributions of tetranucleotides, one can carry out a comparison between a pair of DNA sequences. It was shown many times that this method could reveal certain biologically significant features in a DNA sequence (see, for example, Refs. [18–20]). This methodology assumes that N (the fixed given size of the N-words) is relatively small allowing for computationally reasonable and statistically justified sequence comparisons.

There are biologically motivated improvements of such a methodology. One direction of development lies in a serious increase of word length. Kirzhner et al. [21,22] proposed a novel natural approach to characterize genomic sequences. According to this approach, a genomic sequence is reduced to a histogram of imperfect occurrences of N-grams. High flexibility characterizes this approach due to an allowance for imperfect matching, so that relatively long words comprising the compositional spectra can be considered. The similarity of spectra obtained on different stretches of the same genome, and, simultaneously, a broad range of dissimilarities between spectral representations of different genomes, justify the usefulness of compositional spectra as an informative genome characteristic.

We analyzed different values of N and possible dimensions of N-grams vectors. It appeared that in the case of a DNA text, a set of 200 N-grams with $N = 10$ may be sufficient to successful clustering of DNA sequences. As it was mentioned above the issue of the number of N-grams (i.e. dimension of a corresponding vector) appears to be very widespread. Usually, the suggestion is to include all N-grams. Their potential number grows exponentially, yet the real number, evidently, is confined by the length of a specific text. It is also important to mention that in contrast to usual N-gram techniques, we have determined the occurrence of a given N-gram in a text with certain inaccuracy [21]. Indeed, small local "mismatching" in a genetic text is usually considered routine in biological systems, as opposed to linguistic texts, when the number of such mistakes should be very limited. In addition, by the allowance of a certain level of mismatching, we obtain better frequency statistics: instead of tens of occurrences of each N-gram we can get hundreds and thousands.

In this paper, we analyze possible distances in the space of N-gram frequencies from the viewpoint of clustering DNA texts. Two clustering approaches are employed. The first one, WPGMA, is agglomerative whereas the second one, Partition Around Medoids (PAM or the k-medoids), is a partition algorithm. The obtained results are compared keeping in mind biological interpretations.

The article is arranged in the following way. In Section 2 we describe, based on our previous work [21,22], the N-grams method as applied to DNA texts. In Section 3, we compare characteristics of several distances from the viewpoint of clustering genomes (DNA sequences). In Section 4, we describe clustering methods and results, which are discussed in Section 5 on the basis of formal and biologically meaningful criteria. The list of species and the clustering results are given in Table 3 in the appendix.

2. Dictionary of N-grams for DNA text

2.1. Calculating compositional spectrum (CS)

In accordance with our previous definitions [21] as a perfect occurrence in a target sequence S means x is a substring of string S. Word y is an imperfect occurrence of x in S means y is a substring of S, and distance (in a given metric space) between x and y is less than a threshold (in the given metrics). Word x is an imperfect occurrence of w_i in S if Hamming distance between word w_i and word x less than given r. This approximate matching can be denoted as "r-mismatching".

Let us consider a set W (dictionary) of n different words (oligonucleotides in biochemical terminology) w_i of length L, $n \ll 4^L$, where 4^L are the total number of maximally possible different words of length L. The quantity n is assumed relatively small. By m_i we denote the number of imperfect occurrences of word w_i of the set W in a target sequence $S : m_i = \text{occ}(w_i|S)$. Now let $M = \Sigma m_i$. The vector of 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.

2.2. N-grams selection and dictionary composition

The pre-selected set W of N-grams effectively distinguishing the texts may (a) not be attached to the texts beforehand, (b) not be unique, and (c) be relatively small. In particular,

(a) For any given parameters L and n we consider a set W as a random sample from the set of all possible N-grams with a sampling procedure that can be represented as a stochastic procedure of word generation. Let the words be produced by sequentially adding new letters (out of the four-letter alphabet) with equal probability of appearance of each letter at the new currently generated position. We call such stochastic procedure uniformly random and any resulted random set W of words will also be referred to as a uniformly random set. Thus, we select a set W independently of the current base of the texts.

(b) Every uniformly random set defined in (a) may be used as an N-gram dictionary. Theoretically, different dictionaries may result in a different correlation between text distances. However, the remarkable fact is that this is not the case for DNA texts [21]. In particular, a very high correlation is characteristic for sequences taken randomly from the same genomes.

(c) Concerning the size of the dictionary W and a fixed length of a word in the dictionary, the following considerations based on our previous massive comparisons may be helpful [21]. Remarkably, selection of appropriate dictionary parameters depends on a text size. As a first step of reduction, we randomly chose a continuous genome fragment of the size $|S| = 500,000$ bp. Except a few small parasites, all organisms have genomes larger than 1 Mbp, so a random extraction of such a fragment is possible leaving the rest of a genome for the verification of results. A next step is to choose a size N for a N-gram dictionary. There are a few different considerations in a choice of this parameter. From the biological point of view, the length should be in the range of 8–20 bases, typical for protein binding sites. A next parameter, a number of mismatches should correspond to about 20–30% of a site length. Alternatively, it is possible to get a theoretical estimate of a character word length. For this purpose we would use the theory used in DNA sequencing and technology, which was named sequencing by hybridization (SBH). According to this approach the whole sequence of the length N is reconstructed from a complete set of its subsequences of the predefined length (N-grams). This theory also considers a possibly ambiguous reconstruction resulting in a few variants of a whole sequence, so the theory provides "the probability of unique reconstruction". Theoretical estimations for the chosen length of the text (500,000 bp) give the following values: size L from 10 (words as patterns with gaps of un-sampled positions—imperfect matching with a number of mismatches close to 20% of the L value), up to 20 letters (perfect matching)[23–25].

It is obvious, that a problem of reconstruction, especially partial, is similar to a problem of classification. Thus, basing both on empirical and theoretical considerations we took sets of words of a fixed length from 10 up to 20. In a text of 500,000 letters a word of the length 10 has in average about 10 perfect occurrences. Allowing two mismatches numbers of imperfect occurrences of words of length 10 vary in significant range—from hundreds up to thousand. It creates stretched enough scale. At significant increase in length of a word, say $L = 15$, allows the number of mismatches r to grow from 7 to 8 if we consider the same number of words (for e.g. 8) in sequences of the same length, i.e. the number of identical letters remains constant—about 8. Therefore we chose parameters $L = 10, r = 2$ though, it seems that, some variation of these parameters is insignificant for the further considerations.

The size n of the dictionary was established empirically minding a problem of further clustering. For this procedure we took a set of approximately 50 genomes. In one cycle of the procedure for the fixed dictionary of the size n, all pairwise distances between any pair of those genomes were calculated. Repeating this step for 100 dictionaries of the length n we obtained an average dispersion as a function of n. The saturation happened around $n = 200$. Thus, in this study we deal with CS for parameters $L=10, n=200, r=2$.

Testing of various dictionaries with such parameters showed, that CS various parts of genomes of the size of 500,00 bp appeared to be practically identical.

2.3. Material

The described approach was applied to the analysis of cluster structures in a set of genomes including DNA sequences of 37 species of Eukaryota, Eubacteria, and Archaea (see Appendix, Table 3, column A).

3. Distances in the compositional spectra space

By definition, a compositional spectrum of the sequence S relative to the set W is a vector of frequencies $F(W, S)$. There are various methods to measure dissimilarity between two distribution-vectors $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n)$ and $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$. In this paper we compare several possible distances.

It is associated with one of the famous problems of the Cluster Analysis concerning the identification of the optimum number of clusters. As the synthesis process continues, increasingly dissimilar clusters must be fused, i.e. the classification becomes increasingly artificial. Usage of the histogram of distances appears to be the most appropriate way to handle this problem (see Refs. [26,27]). Namely, local minima of the histogram of all pairwise distances make a cluster structure of the data such that the masses of all local peaks indicate the density of the clusters' concentration. (A number of local minima could provide a lower bound for the number of possible clusters.) Note that this methodology makes it possible to assess the discriminating capabilities of a distance measure without running any clustering procedure. For example, detection of two essential local histogram minimum points leads to an assumption that there are at least three clusters.

We examined the next distances:

1. The Euclidian distance that is defined for two distributions

$\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n)$ and $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n)$ by the conventional way as

$$d_1(\mathbf{t}, \mathbf{u}) = \sum_{i=1}^{n} (\mathbf{t}_i - \mathbf{u}_i)^2.$$

2. The Manhattan distance

$$d_2(\mathbf{t}, \mathbf{u}) = \sum_{i=1}^{n} \mathrm{abs}(\mathbf{t}_i - \mathbf{u}_i).$$

3. The Max-distance

$$d_3(\mathbf{t}, \mathbf{u}) = \max_i \, \mathrm{abs}(\mathbf{t}_i - \mathbf{u}_i).$$

4. *KS*-like distance, in the case when the coordinates of the vectors $\mathbf{t}$ and $\mathbf{u}$ can be interpreted as frequencies

$$d_4(\mathbf{t},\mathbf{u}) = \max_i |F_i - U_i|,$$

where

$$\mathbf{F}_i = \frac{\sum_{j=1}^{i}\mathbf{t}_j}{\sum_{j=1}^{n}\mathbf{t}_j}, \quad \mathbf{U}_i = \frac{\sum_{j=1}^{i}\mathbf{u}_j}{\sum_{j=1}^{n}\mathbf{u}_j}$$

are the relative cumulative distributions. This distance is associated with the well-known Kolmogorov–Smirnov statistics.

5. The correlation dissimilarity

$$d_5(\mathbf{t},\mathbf{u}) = (1 - R(\mathbf{t},\mathbf{u}))/2,$$

where $R(\mathbf{t},\mathbf{u})$ is the Pearson correlation coefficient.

6. The Spearman dissimilarity

$$d_6(\mathbf{t},\mathbf{u}) = (1 - \boldsymbol{\rho}(\mathbf{t},\mathbf{u}))/2,$$

where $\rho(\mathbf{t},\mathbf{u})$ is the Spearman rank correlation coefficient.

7. The Kendall dissimilarity

$$d_7(\mathbf{t},\mathbf{u}) = (1 - \tau(\mathbf{t},\mathbf{u}))/2,$$

where $\tau(\mathbf{t},\mathbf{u})$ is the Kendall rank correlation coefficient.

We investigated all distances according to their potential ability to produce a "meaningful" partition using the same set of 37 chosen species (Appendix). We have also taken two sequences 300–400 kb long from each genome, excluding the Human genome (where 11 sequences were taken). Thus, the initial set (database) included 85 sequences. Such a structure of the database (more than one sequence from every genome) will allow a natural control of the grouping quality in data. Then, the compositional spectrum (*N*-gram vector) had been calculated for every sequence as described in Section 2.

According to the obtained results, we divided all distance measures into two groups. The first group included: Euclidian, Manhattan, Max, and *KS*-like distances. The second group contained the dissimilarities based on correlation: Person, Spearman, and Kendall. A typical example of the distance histogram in the first group is given by the Euclidian distance (Fig. 1).

This histogram points to a possibility of two or three clusters being very different in their sizes. However, generally speaking, it looks like one big cluster. This fact could be indirectly demonstrated by an attempt to group the data into two clusters using the K-means algorithm. The final partition becomes unstable from the point of view of the Multivariate MANOVA procedure. It is very hard to expect a reasonable clustering to be obtained by means of such inappropriate measure.

An example of a histogram for the second group is provided in Fig. 2, using the distance based on Kendall rank correlation coefficient.

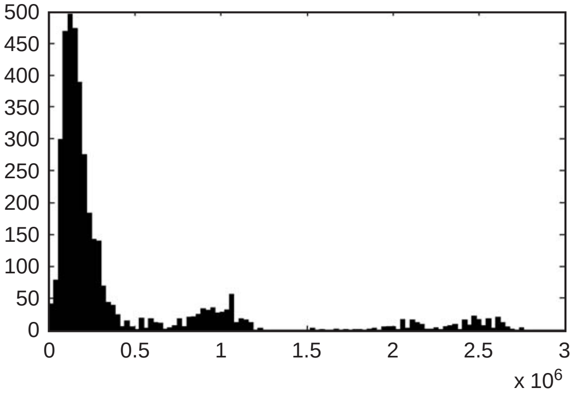

Fig. 1. Histogram of the Euclidian distance.

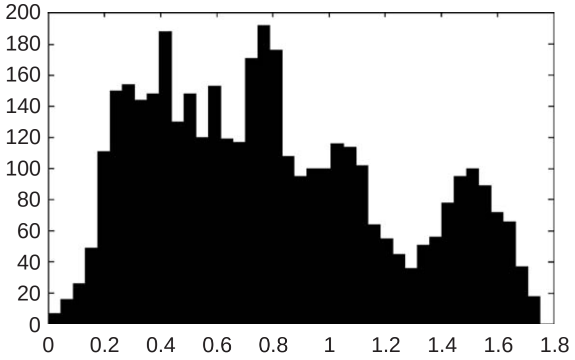

Fig. 2. Fig. 2. Histogram of the Kendall distance.

One can recognize here some three local minima of the histogram, but disparities between masses of the local maximum areas are not so considerable compared to the previous group of distances (see Fig. 1). This fact allows us to assume three or more clusters in the data.

We can observe three major local minima of the histogram, but disparities between the masses of areas corresponding to the local maxima are not so considerable comparing those in the previous group of distances. Therefore, we can assume that there are three or more clusters in the data. Consequently, the second group seems to be preferable for the effectiveness of the subsequent clustering. The second group also includes a measure based on the *N*-grams vector correlation that is often used for building distance matrices. However, we used the distance based on the Kendall coefficient that seems to give the best estimation of the cluster structure.

4. Clustering data

Using the distance based on the Kendall coefficient we apply two clustering methods corresponding to two different clustering techniques—hierarchical and partitioning.

4.1. Weighted pair group method with arithmetic mean (WPGMA)

The WPGMA is the simplest method of tree construction [28]. It was originally developed for constructing taxonomic phenograms, i.e. trees that reflect the similarities between elements of finite sets. In evolutionary studies these elements are usually referred to as operational taxonomic units (OTUs). In order to get an idea about cluster stability, we complement the WPGMA method by the following procedure that will be clarified through the treatment of our major data set on 37 species.

The procedure is as the follows. As it was noted in the previous paragraph the number of clusters in the considered species collection appears to be not less than three. We used the WPGMA method to divide the set into three clusters and considered the validation of each cluster by means of the following approach. First of all, we mark the cluster elements and repeat N times the hierarchical WPGMA procedure based on various distance matrixes (where N is a pre-selected number). These matrices are built on some randomly chosen dictionaries and corresponding compositional spectra. The clusterization procedure is stopped if at least 75% of the marked items are concentrated in a one of the formed clusters. Reiteration of the described procedure for another distances template can lead, generally speaking, to another cluster. As a result, each data set element gains an array of the relative frequencies of the membership. Thus, a data set together with an attached occurrence table is called U-cluster; i.e. each element is presented by its estimated membership probability. A cluster is considered "stable" if the majority of the marked elements possess significantly higher probability values in the U-cluster than the non-marked ones do.

We conducted this test 100 times with 100 different (but uniformly distributed) dictionaries (accordingly, using the produced distance matrixes) with the same parameters $L = 10$, $n = 200$, and $r = 2$. Then, we calculated the percentage of occurrences of every element (OUT) in this or that cluster within all 100 tests. For example, the result obtained for the first cluster is reproduced in Fig. 3.

Averaging-out was done with 100 runs using 100 different sets W. X-axis represents all 11 sampled human genome sequences (*Homo sapiens* chromosomes X, Y and 1,3,4,6,7,11,13,20,22) and (for simplicity) one of the two sequences per genome for all remainder species (if both sequences for each genome were represented they would appear as adjacent neighbors). The vertical axis shows the percentages of occurrences of each species in the given cluster. The species identified as the cluster members are marked by squares on the lower curve. We see that this cluster has a kernel appearing in no less than 75% of the cases and random components appearing in not more than 20% of the cases. This kernel can be considered as a cluster.

Similar calculations were done for the other two clusters, the result being represented in Table 3, column 2. We refer to the obtained clustering as U-clustering.

4.2. Partitioning around medoids

The partitioning around medoids (PAM) method was introduced by Kaufman and Rousseeuw [4]. The PAM procedure gets as its arguments a dissimilarity matrix of the elements and suggests the number of clusters k. The routine is built on searching for k representative objects, or medoids, around the points to be clustered. The clusters are created

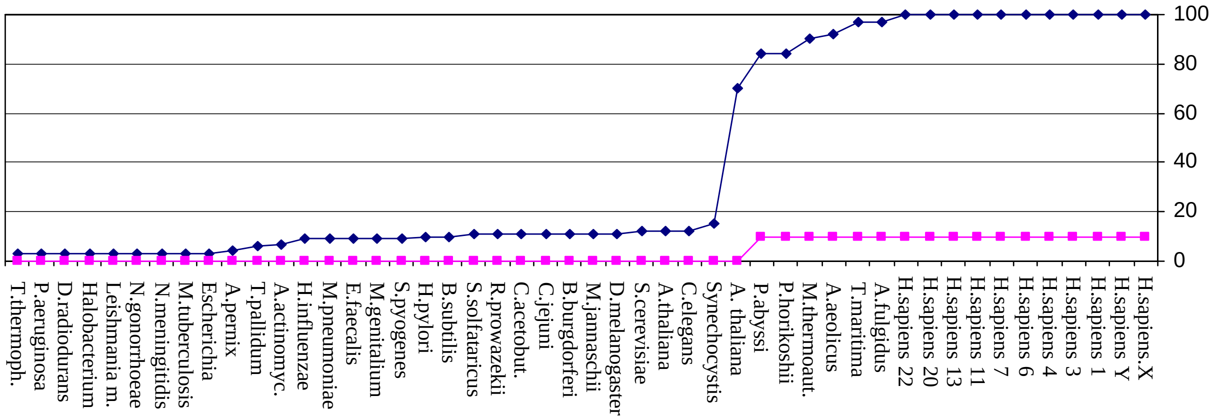

Fig. 3. Fig. 3. A cluster construction example by means the WPGMA method.

Table 1
Different correlation coefficients of U- and P-partitions

	3	4	5	6	7	8	9
Cramer’s_V	0.6165	0.8138	0.8891	0.9537	0.9825	0.9825	0.9825
Rand	0.679	0.8364	0.8654	0.899	0.9127	0.9099	0.9055
Jain & Dubes	0.4511	0.589	0.625	0.6858	0.7146	0.7049	0.6925
Fowlkes & Mallows	0.6351	0.7403	0.7745	0.8325	0.8613	0.8563	0.8474

Table 2
Two-way joint distribution of 3 clusters of U-partition and P_7-partition

Clusters	1	2	3	4	5	6	7
1	0	0	0	11	6	8	0
2	34	0	0	0	0	0	2
3	0	14	8	0	0	0	2

by passing on each element to the most similar medoid. The algorithm seeks for an optimal value of the sum of the dissimilarities of the observations to their closest medoid. A minimum partition for the objective function is a final medoids-set, such that no single movement of an observation relative to the medoids can decrease the objective function value. The PAM approach looks more robust and efficient than the known k-means algorithm and is implemented in clustering packages R and S-Plus.

4.3. Verifying the results of clustering

When using the hierarchical method, we rested upon an estimate of a number of clusters obtained in Section 3. According to this estimate, the number of clusters when using the Kendall coefficient for compositional spectra of 37 species is expected to be no less than three. However, the actual number of clusters may be greater. Disagreement between the amounts of clusters determined by different procedures may derive from the fusion of several smaller clusters into one big cluster.

The results of clustering with PAM method (P-clusters) substantially depend on the chosen number of clusters. Therefore, we decided to use the PAM method for calculation of the cluster structure starting from three as the minimum number of clusters. The intention is to get a “correct fragmentation” of the 3-cluster structure found earlier with the hierarchical method. Namely, we refer here to such a fragmentation, with PAM method, when every P-cluster of this fragmentation intersects only one U-cluster. We call such a kind of mutual cluster structure as a coordinated structure. We believe that such a coordinated structure makes possible the restoration of genuine clustering. Note that the coordinated structure always exists, for example, in the trivial situation when the number of P-clusters is equal to the number of elements. If this case is the only example of a coordinated structure, then clustering fragmentation must be considered inadequate. However, even if a non-trivial coordinated structure is achieved, cluster fragmentation might be close to the trivial one, e.g. with the number of P-clusters close to the total number of elements.

In order to characterize the quality of the coordinated structure, we employed several correlation coefficients. Keeping in mind reaching a coordinated structure with a minimal number of P-clusters, we subsequently calculate the partitions $P_3, P_4, \ldots,$ etc. until the achievement of a coordinated cluster structure. It is interesting to note that coordination is not a “heritable” feature. In particular, if the PAM cluster structure with some given number of clusters T is coordinated with a given hierarchical structure, then it is not necessary that for a value greater than T the structure will also be coordinated.

5. Results

5.1. Coordinated cluster structure

As it was mentioned above, clustering aims at extracting inner structure in a data. A fundamental question referred to as the problem of cluster validation, i.e. the problem of judgment the “correct” number of clusters. In general, clustering algorithms make different assumptions about the set structure. Since for the cases of interest one does not know whether these assumptions are satisfied by the data, different clustering methods lead to a variety of answers. For that reason a universal method of finding the true number of clusters for given data does not exist. However, in the spirit of Roth et al. [29], we can consider the notion of cluster stability as a criterion for this purpose. In our case, this concept suggests good matching between PAM partitions

Table 3
Summary table of the used species and their U- and P-partitions

A	B	3	4	5	6	7	8	9
Operational taxonomic units	U-structure	PAM	PAM	PAM	PAM	PAM	PAM	PAM
Homo sapiens chr. X	1	0	3	3	3	3	3	3
Homo sapiens chr. Y	1	0	3	3	3	3	3	3
Homo sapiens chr. 1	1	0	3	3	3	3	3	3
Homo sapiens chr. 3	1	0	3	3	3	3	3	3
Homo sapiens chr. 4	1	0	3	3	3	3	3	3
Homo sapiens chr. 6	1	0	3	3	3	3	3	3
Homo sapiens chr. 7	1	0	3	3	3	3	3	3
Homo sapiens chr. 11	1	0	3	3	3	3	3	3
Homo sapiens chr. 13	1	0	3	3	3	3	3	3
Homo sapiens chr. 20	1	0	3	3	3	3	3	3
Homo sapiens chr. 22	1	0	3	3	3	3	3	3
Pyrococcus hor	1	0	0	0	5	5	5	5
Pyroccocus hor	1	0	0	0	5	5	5	5
Methanobacterium thermoautotroph	1	0	3	4	4	4	4	8
Methanobacterium thermoautotroph	1	1	3	4	4	4	4	8
Archaeoglobus fungulus	1	0	3	4	4	4	4	4
Archaeoglobus fungulus	1	2	2	4	4	4	4	4
Aquifex aeolicus	1	0	0	4	5	5	5	5
Aquifex aeolicus	1	0	0	4	5	5	5	5
A. thaliana mitochondrial genome	1	0	3	3	5	5	5	5
A. thaliana mitochondrial genome	1	0	3	3	5	5	5	5
Thermotoga maritima	1	0	0	4	4	4	4	4
Thermotoga maritima	1	0	0	4	4	4	4	4
Pyrococcus abyssi	1	0	0	0	5	5	5	5
Pyrococcus abyssi	1	0	0	0	5	5	5	5
Caenorhabditis elegans	2	0	0	0	0	0	0	0
Caenorhabditis elegans	2	0	0	0	0	0	0	0
Synechocystis sp.	2	2	2	2	2	6	6	6
Synechocystis sp.	2	2	2	2	2	6	6	6
Saccharomyces cerevisiae	2	0	0	0	0	0	0	0
Saccharomyces cerevisiae	2	0	0	0	0	0	0	0
Arabidopsis thaliana	2	0	0	0	0	0	0	0
Arabidopsis thaliana	2	0	0	0	0	0	0	0
Drosophila melanogaster	2	0	0	0	0	0	0	0
Drosophila melanogaster	2	0	0	0	0	0	0	0
Methanococcus jan	2	0	0	0	0	0	0	0
Methanococcus jan	2	0	0	0	0	0	0	0
Streptococcus pyogenes	2	0	0	0	0	0	0	0
Streptococcus pyogenes	2	0	0	0	0	0	0	0
Borrelia burgdorferi	2	0	0	0	0	0	0	0
Borrelia burgdorferi	2	0	0	0	0	0	0	0
Bacillus subtilis	2	0	0	0	0	0	0	0
Bacillus subtilis	2	0	0	0	0	0	0	0
Helicobacter pylori	2	0	0	0	0	0	0	0
Helicobacter pylori	2	0	0	0	0	0	0	0
Mycoplasma genitalium	2	0	0	0	0	0	0	0
Mycoplasma genitalium	2	0	0	0	0	0	0	0
Enterococcus faecalis	2	0	0	0	0	0	0	0
Enterococcus faecalis	2	0	0	0	0	0	0	0
Mycoplasma pneumoniae	2	0	0	0	0	0	0	0
Mycoplasma pneumoniae	2	0	0	0	0	0	0	0
Campylobacter jejuni	2	0	0	0	0	0	0	0
Campylobacter jejuni	2	0	0	0	0	0	0	0
Haemophilus influenzae	2	0	0	0	0	0	0	0
Haemophilus influenzae	2	0	0	0	0	0	0	0

Table 3 (*continued*)

A	B	3	4	5	6	7	8	9
Operational taxonomic units	U-structure	PAM	PAM	PAM	PAM	PAM	PAM	PAM
Sulfolobs solfat	2	0	0	0	0	0	0	0
Sulfolobs solfat	2	0	0	0	0	0	0	0
Clostridium acetobutylicum	2	0	0	0	0	0	0	0
Clostridium acetobutylicum	2	0	0	0	0	0	0	0
Rickettsia prowazekii	2	0	0	0	0	0	0	0
Rickettsia prowazekii	2	0	0	0	0	0	0	0
Escherichia coli	3	2	2	2	2	2	2	2
Escherichia coli	3	2	2	2	2	2	2	2
Treponema pallidum	3	2	2	2	2	2	7	7
Treponema pallidum	3	2	2	2	2	2	7	7
Thermus thermophilus	3	1	1	1	1	1	1	1
Thermus thermophilus	3	1	1	1	1	1	1	1
Mycobacterium tuberculosis	3	1	1	1	1	1	1	1
Mycobacterium tuberculosis	3	1	1	1	1	1	1	1
Neisseria meningitides	3	2	2	2	2	2	2	2
Neisseria meningitides	3	2	2	2	2	2	2	2
Neisseria gonorrhoeae	3	2	2	2	2	2	2	2
Neisseria gonorrhoeae	3	2	2	2	2	2	2	2
Leishmania major	3	1	1	1	1	1	1	1
Leishmania major	3	1	1	1	1	1	1	1
halobacterium SRc-1	3	1	1	1	1	1	1	1
halobacterium SRc-1	3	1	1	1	1	1	1	1
Deinococcus radiodurans	3	1	1	1	1	1	1	1
Deinococcus radiodurans	3	1	1	1	1	1	1	1
Pseudomonas aeruginosa	3	1	1	1	1	1	1	1
Pseudomonas aeruginosa	3	1	1	1	1	1	1	1
Actinobacillus actinomycetemcomitans	3	0	0	0	0	6	6	6
Actinobacillus actinomycetemcomitans	3	2	2	2	2	6	6	6
Aeropyrum pernix	3	1	1	1	1	1	1	1
Aeropyrum pernix	3	1	1	1	1	1	1	1

A—list of species; B—U-partitions (different clusters are marked by different numbers); (3–9)—P-partitions with corresponding numbers of clusters.

and U-structure. Moreover, we are interested to reach a partition having suitable biological interpretation.

Let us compare PAM-structures P_3–P_9 (see appendix, Table 3, columns 3–9) with U-structure. A formal way to do it is to exploit external indices of partition agreement. Usually, the calculation of these scores is built on so-called cross-tabulation, or contingency table. Entries n_{ij} of this table denote the number of objects that are in both clusters i and j, $i=1,\dots,N$, $j=1,\dots,M$ for two different partitions P_N and P_M. We use the most known coefficients [15,30,31]. In addition, the regular Cramer correlation coefficient was employed as well. The coefficients together with U-structure are presented in Table 1.

U-partition is fixed and contains 3 clusters whereas P-partitions are represented sequentially with P varying from 3 to 9 clusters.

As we can see now, the best correlation is reached in the case when the number of PAM clusters is 7. Two-way joint distribution, representing U-clustering and P_7-partition, is represented in Table 2.

All clusters of P_7-partition, except one, are fully included into 3 clusters of U-partition. We can see here almost perfect matching, that is, the first U-cluster consists of three P-clusters (4, 5, 6) only; the second U-cluster almost coincides with the first one in P-partition; and the third U-cluster consists of two P-clusters. Thus, we have unconditionally found a coordinated cluster structure. It turns out that three clusters of the U-structure are coordinated with the clusters of P_7-structure. It is natural to ask which of the two structures is more suitable for biological interpretation. In our case, we are interested in determining large similarity groups; therefore, U-structures appear to be an appropriate tool, however P_7-partition gives more thinly subordinated clustering.

5.2. Biological interpretation of the partition

A main requirement should be that both representatives of every genome occur in the same cluster. This is a formal indication of a robust clustering. Detailed discussion of different biological notions is obviously not in the scope of this paper. However, we can see some logic in the obtained U-structures. Our speculation is that the presented U-structure largely follows a pattern of ecological convergence. Indeed, the closest cluster to humans is the set of thermophilic prokaryotes (see Table 3). Some of the thermophilic prokaryotes appeared to be far from the main thermophilic cluster, but even in this case they joined other clusters by pairs (e.g. *Methanococcus jannaschii*, *Sulfolobus solfataricus*, *Aeropyrum pernix*, *Thermus thermophilus*. Note that both *S. solfataricus* and *Aeropyrum pernix* inhabit highly extreme environments). One can speculate that high temperature was the common denominator, largely converging the genome structures of mammals and thermophilic prokaryotes. This idea of ecological convergence is also supported by the fact that the cluster of thermophilic prokaryotes is represented by species that belong to both Eubacteria (*Thermotoga maritima*, *Aquifex aeolicus*) and Archaea (*Pyrococcus horikoshii*, *Pyrococcus abyssi*, *Archaeoglobus fulgidus*, and *Methanobacterium thermoautotrophicum*). Another possible complementary explanatory factor in the revealed Archaea–Eubacteria relationships may derive from differentiation of prokaryotes with respect to aerobic–anaerobic metabolism. Indeed, almost all of the prokaryotes that manifested an ecological "affinity" to eukaryotes are anaerobic. An exception is *Aquifex aeolicus*, which is micro-aerobic using oxygen at very low concentrations [32]. There is one thermophilic and anaerobic organism (*Methanococcus jannaschii*) that is located distantly from the foregoing cluster. It is, however, a strict anaerobe that dies when exposed to oxygen. In other words, two parameters, temperature and oxygen, can almost perfectly explain the foregoing clustering revealed by compositional spectra. The second U-cluster brings together Eucaria, Eubacteria, and Archaea and we cannot connect this association with any simple uniform pattern. The third cluster unites CG-rich genomes.

6. Conclusion

The approach of text reduction to a relatively small dictionary of the proposed *N*-grams in the paper gives an opportunity to successfully classify DNA-texts. Doing so, we have implemented two basic clustering techniques and their coordination principle. The proposed *N*-grams approach stems from the desirability to address the uncertainty associated with DNA-texts. We assume that such an approach may be useful in the analysis of phonetic and image "texts" as well.

Acknowledgements

We are grateful to referees for constructive suggestions which have materially improved the presentation of the paper.

Appendix

The list of species and the dustering results, i.e. PAM structures $P_3 - P_9$ with U-structure is presented in Table 3.

References

[1] M.W. Berry, M. Browne, Understanding Search Engines: Mathematical Modeling and Text Retrieval, SIAM Book Series: Software, Environments and Tools, 1999.

[2] J. Kogan, C. Nicholas, V. Volkovich, Text mining with information-theoretical clustering, Comput. Sci. Eng. 5 (6) (2003) 52–59.

[3] E. Forgy, Cluster analysis of multivariate data: efficiency vs. interpretability of classifications, Biometrics 21 (3) (1965) 768–769.

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

[5] C.E. Shannon, A mathematical theory of communication, Bell System Tech. J. 27 (1948) 379–423, 623–656.

[6] T. De Heer, Experiments with syntactic traces in information retrieval, Inform. Storage Retrieval 10 (1974).

[7] E.S. Adams, A.C. Meltzer, Trigrams as index elements in full text retrieval, observations and experimental results, ACM Composite Science Conference, 1993.

[8] W.B. Cavnar, Using an N-gram-based document representation with a vector processing retrieval model, The Fourth Text Retrieval Conference (TREC-3), 1995.

[9] J.D. Cohen, Highlights: language- and domain-independent automatic indexing terms for abstracting, J. Am. Soc. Inform. Sci. 46 (3) (1995).

[10] M. Damashek, Gauging Similarity with N-grams: language-independent categorization of text, Science 267 (1995).

[11] A.M. Robertson, P. Willett, Searching for historical word-forms in a database of 17th century english text using spelling-correction methods, 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1992.

[12] S. Huffman, M. Damashek, Acquaintance: a novel vector-space N-gram technique for document categorization, The Third Text REtrieval Conference (TREC-3), 1995.

[13] S. Huffman, Acquaintance: language-independent document categorization by N-grams, The Fourth Text REtrieval Conference (TREC-4), 1996.

[14] J.H. Lee, J.S. Ahn, Using N-grams for Korean text retrieval, 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996.

[15] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, New York, 1988.

[16] R. Douglass, D.R. Cutting, J. Karger, O. Pedersen, J.W. Tukey, Scatter/gather: a cluster-based approach to browsing large document collections, SIGIR '92, 1992, pp. 318–329.

[17] A. Bolshoy, DNA sequence analysis linguistic tools, Rev. Appl. Bioinform. 2 (2003) 103–112.

[18] S. Pietrokovski, J. Hirshon, E.N. Trifonov, Linguistic measure of taxonomic and functional relatedness of nucleotide sequences, J. Biomol. Struct. Dynamics 7 (1990) 1251–1268.

[19] S. Karlin, I. Ladunga, BE. Blaisdell, Heterogeneity of genomes—measures and values, Proc. Natl. Acad. Sci. USA 91 (1994) 12837–12841.

[20] S. Karlin, C. Burge, Dinucleotide relative abundance extremes: a genomic signature, Trends Genet. 11 (1995) 283–290.

[21] V. Kirzhner, A. Korol, A. Bolshoy, E. Nevo, Compositional spectrum—revealing patterns for genomic sequence characterization and comparison, Physics A 312 (2002) 447–457.

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

[23] S.A. Heath, F.P. Preparata, J. Young, Sequencing by hybridization using direct and reverse cooperating spectra, Proceedings of the Sixth Annual International Conference on Computational Biology, 2002, pp. 186–193.

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

[25] F.P. Preparata, Sequencing-by-hybridization revisited: the analog-spectrum proposal, Comput. Biol. Bioinform. 1 (2004) 47–52.

[26] N.N. Aprausheva, A new approach for the cluster detection, Computer Center of the Russian Academy of Science, Moscow, 1993 (in Russian).

[27] L.J. Latecki, R. Venugopal, M. Sobel, S. Horvath, Tree-structured partitioning based on splitting histograms of distances, IEEE International Conference on Data Mining (ICDM'03), Melbourne, FL, USA, 2003.

[28] R.R. Sokal, C.D. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas Sci. Bull. 38 (1958) 1409–1438.

[29] V. Roth, V. Lange, M. Braun, J. Buhmann, A resampling approach to cluster validation, COMPSTAT 2002, available at http://www.cs.unibonn.De/~braunm.

[30] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (1971) 846–850.

[31] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clustering, J. Am. Stat. Assoc. 78 (1983) 553–584.

[32] G. Deckert, P.V. Warren, T. Gaasterland, W.G. Young, A.L. Lenox, D.E. Graham, R. Overbeek, M.A. Snead, M. Keller, M. Aujay, R. Huber, R.A. Feldman, J.M. Short, G.J. Olsen, R.V. Swanson, The complete genome of the hyperthermophilic bacterium Aquifex aeolicus, Nature 392 (1998) 353–358.

About the Author—ZEEV (VLADIMIR) VOLKOVICH is an Associate Professor in the Department of Software Engineering at College ORT Braude Karmiel, Israel. He received a Ph.D. in Probability Theory and Mathematical Statistics in 1982. Dr. Volkovich is doing research in statistical pattern recognition, feature selection, text mining, stochastic models, and generalized convolutions. His recent research includes classification with unsupervized learning and clustering initialization.

About the Author—VALERY KIRZHNER is a Professor in the Faculty of Science, University of Haifa. He holds a Ph.D. in Mathematics from Leningrad Department of Steklov Institute of Mathematics (Academy of Science former USSR). He is an author more then 100 papers in fields of mathematics, mathematical and computer modeling in biology, population dynamics and bioinformatics.

About the Author—ALEXANDER BOLSHOY is an Associate Professor of Computational Biology in the Institute of Evolution at Haifa University. Dr. Bolshoy's research is aimed to reveal the biological role of certain evolutionary preserved genomic patterns using statistical techniques and different methods of sequence and structure comparison. He graduated from the faculty of mathematics at the Moscow State University, USSR and received his Ph.D. in Structural Biology at the Weizmann Institute of Science. Since then he has worked at the Institute for Biological Sciences, NRC of Canada; at the Center for Biological Sequence Analysis at the Technical University of Denmark; at NCBI/NLM/NIH and in the School of Informatics at Indiana University.

About the Author—Professor EVIATAR NEVO is a world-leading scientist in evolutionary biology. Nevo contributed to biodiversity evolution in nature involving genes, genomes, phenomes, populations, species, and ecosystems of bacteria, fungi, plants, animals, and humans focusing on adaptation and speciation. Nevo founded the Institute of Evolution at the University of Haifa in 1976 and has been the Director since its inception. The Institute has become a world center of excellence conducting active integrative research in biodiversity, molecular biology, genomic, and organismal evolutions linking *field*, *laboratory*, and *theoretical* research programs across life. Nevo is a foreign member of the National Academy of Science, USA, the Linnean Society of London, and the Ukraine Academy of Sciences. He was awarded numerous prizes. He is also the author and/or co-author of 19 books and hundreds papers in diverse fields of evolutionary biology.

About the Author—ABRAHAM KOROL is a Professor of Genetics at Faculty of Science, University of Haifa. He holds a Ph.D. in Biology from the Institute of General Genetics (Moscow, Academy of Science former USSR), Doctor of Science in Biology from Leningrad University, and M.Sc. in Computer Science from Leningrad Technical University. He is an author of four books and about 200 papers in fields of genetics, evolutionary and population biology, mathematical and computer modeling in biology and bioinformatics. He is head of laboratory of mathematical and population genetics and laboratory of *Drosophila* genetics.