J. Math. Biol. 46, 479–503 (2003)
Digital Object Identifier (DOI):
10.1007/s00285-002-0185-3

Mathematical Biology

Srabashi Basu · Debi Prosad Burma · Probal Chaudhuri

Words in DNA sequences: some case studies based on their frequency statistics

Received: 15 October 2000 / Revised version: 8 October 2002
Published online: 28 February 2003 –

Abstract. One of the critical requirements of data analysis involving large DNA sequences is an effective statistical summarization of those sequences. In this article DNA sequences have been analyzed based on word frequencies. Our analysis focuses on the detection of structural signature of a genome reflected in word frequencies and identification of phylogenetic relationships among different species reflected in the variation of word distributions in their DNA sequences. We have carried out a statistical study of the complete genome of baker's yeast, of various ribosomal RNA sequences from different prokaryotic and eukaryotic organisms and of the full genomes of some bacteriophages. Our exploratory analysis amply demonstrates the usefulness of DNA word frequencies in reducing the dimensionality of large sequences while retaining some of the structural information there that can have biological significance. Some conceptual issues that arise in course of our investigation have been addressed. A few interesting problems related to the statistics of DNA words have been pointed out with some indication of their possible solutions. The work has been partially motivated by the fact that sequence alignment and homology techniques that are quite popular for comparing and analyzing relatively smaller DNA sequences of nearly equal sizes are not applicable to data consisting of large sequences with widely varying sizes, which may contain segments with unknown or no biological functions, and consequently their comparison through functional homology is either impossible or extremely difficult.

1. Introduction: statistical summarization of DNA sequence data using word frequencies

Rapid advancement in automated DNA sequencing technology has created the need for summarization of large volumes of sequence data so that effective statistical analysis can be carried out leading to fruitful scientific results. Various known sequence alignment algorithms and techniques for estimating the homologies and mis-matches among DNA sequences [see e.g., Doolittle (1990, 1996)

S. Basu, P. Chaudhuri: Theoretical Statistics and Mathematics unit, Indian Statistical Institute, 203 B.T. Road, Calcutta 700108, India. e-mail: srabashi@isical.ac.in; probal@isical.ac.in

D.P. Burma: Molecular Biology Unit, Institute of Medical Sciences, Banaras Hindu University, Varanasi, 221005, India

Current address: CF186, Salt lake, Calcutta 700064, India

Research presented here was supported in part by a grant from Indian Statistical Institute.

Key words or phrases: Average linkage clustering – Chernoff's faces – Dendrograms – DNA words – F-ranks of words – F-ratios of words – l_1-distance – Phylogenetic relationships – Rank correlation – Single linkage clustering

and Waterman (1995)] that are used for comparing sequences of relatively smaller sizes are not feasible to use when it comes to dealing with sequences having sizes varying between a few thousand base pairs to a few hundred thousand base pairs. Even for comparing sequences of smaller sizes, the standard alignment and matching algorithms are known to be time consuming and laborious, and the utility of more rapid and parsimonious procedures that may be somewhat rough in nature yet useful in producing quick and significant results is well appreciated. Effective analysis of large DNA sequences requires some form of statistical summarization by reducing the dimension of the data to facilitate numerical computations and at the same time to capture some of the fundamental structural information contained in the sequence data as efficiently as possible.

Since a DNA sequence is formed using an alphabet of four letters (i.e. A, T, C and G) denoting four DNA bases, the simplest form of statistical summarization that one can think of is based on various frequencies of DNA k-words, which are k-tuples formed using these four letters. For an integer $k \geq 1$, let W_k denote the set of all possible k-words formed using the alphabet $\{A, T, C, G\}$. Clearly there are 4^k possible k-words, and for a given DNA sequence and a given word $w \in W_k$, we will denote by f_w the relative frequency of the word w in the sequence, where the words in the sequence may have one or more overlapping letters. For example, if a sequence runs like $ATTCGGCA\ldots$, the first 4-word is $ATTC$, the second one is $TTCG$, the third one is $TCGG$ and so on. We will view the 4^k-dimensional frequency vector $(f_w)_{w\in W_k}$ as a form of statistical summary of the given DNA sequence, and we trivially have $f_w \geq 0$ for all $w \in W_k$ and $\sum_{w\in W_k} f_w = 1$. It is easy to see that a given DNA sequence is in general not uniquely determined by its k-word frequecies unless of course k is as large as the size of the given sequence [see Arratia, Martin, Reinert and Waterman (1996) for issues and results related to recovery of sequences from k-word frequencies]. Nevertheless, a comparison between a pair of DNA sequences to judge their structural similarities and dissimilarities can be carried out by comparing their associated word frequency vectors $(f_w)_{w\in W_k}$ and $(g_w)_{w\in W_k}$ (say), and it involves comparing the values of f_w and g_w for each $w \in W_k$ in some appropriate way.

Markov chain models are quite popular in the analysis of sequence data based on word frequencies, and perhaps they are one of the most widely used stochastic models for this purpose [see e.g., Waterman (1995) and Reinert, Schbath and Waterman (2000) for some excellent reviews]. Consider a $(k-1)$-step ($k \geq 2$) homogenous Markov chain with the state space same as the DNA alphabet $\{A, T, C, G\}$, and for $w = (u_1, \ldots, u_k) \in W_k = \{A, T, C, G\}^k$, let $p_w = Pr(u_k|u_1, \ldots, u_{k-1})$ denote its $(k-1)$-step transition probability. Then for this Markov model, it is easy to see that the following result holds [see e.g. Section 2.2 in Reinert, Schbath and Waterman (2000)].

Result 1.1. *For an observed DNA sequence S, the conditional likelihood given the first $k-1$ letters of S will be*

$$L(S) = \prod_{w\in W_k} (p_w)^{(N-k+1)f_w},$$

where $(f_w)_{w\in W_k}$ is the 4^k-dimensional k-word frequency vector associated with S, and N is the length of S (i.e., the number of nucleotides in S). The form of the conditional likelihood implies that $(f_w)_{w\in W_k}$ is a sufficient statistic for the transition probability parameters p_w's, and the maximum likelihood estimates of these parameters can be derived in terms of f_w's.

Result 1.1. provides one of the simplest but rather interesting formal statistical motivation for the use of k-word frequencies to summarize DNA sequence data. Under the assumption of homogenous Markov property of a DNA sequence, DNA word frequencies can be viewed as a *sufficient summarization* that does not lose any of the relevant statistical information contained in the original sequence. Distributions of k-words in a DNA sequence under Markov models and many related statistical and probabilistic results can be found in Waterman (1995) and Reinert, Schbath and Waterman (2000).

Relative abundance and shortage of certain DNA words are likely to have implications on molecular structures and stability of genomes, and this may have some connections with cellular processes like recombination, replication, regulation, repair activities etc. Another important issue is to what extent the relatedness and similarities measured by comparing DNA word frequencies of different sequences are in conformity with known phylogenetic relationships. Our primary objective in this paper is to make a critical evaluation of the potentials of DNA word frequencies as a useful statistical summary of sequence data. We intend to investigate the extent to which the *structural signature* in the genome of a species is reflected in those frequencies [see Pevzner, Borodovsky and Mironov (1989a, 1989b) for some related work]. Further, we will try to develop an understanding of how far phylogenetic relationships among different species can be captured by distances based on the word frequencies obtained from their DNA sequences [see also Chaudhuri and Das (2001, 2002)]. It is expected that a careful analysis of word frequencies may also reveal valuable insights and interesting facts concerning the evolutionary process of different parts of a genome. In addition to performing case studies with specific genomic sequences using careful exploratory analysis, we have tried to focus on some of the conceptual issues underlying the statistical analysis of DNA word frequencies. Some theoretical questions that arise in course of our data analysis have been addressed. Our statistical analysis is largely exploratory in nature, and it is not explicitly based on any specific probability model unlike what has been sometimes done in the literature in the past.

2. Structural signature of a genome reflected in word frequencies

Usually a prokaryotic organism like a bacteria or an archaea has its DNA material organised in the form of a circular or a linear genome inside its cell while a eukaryotic cell has the DNA material distributed in several chromosomes. As the genome grew in size in course of evolution leading to complex eukaryotic cells from more primitive and simpler prokaryotic cells, one plausible hypothesis is that the chromosomes were formed by fission of large genomes. In such a case, one would expect certain basic structural similarities in different chromosomes in the

cell of an eukaryotic species. Recently Chaudhuri and Das (2001, 2002) analyzed all six chromosomes (five autosomes and one sex chromosome) of *Caenorhabditis elegans* i.e., the round-worm using word frequencies and reported some interesting observations. A related question is to what extent different parts of the same genome have similar structure. In case a genome is not yet fully sequenced, one would still like to carry out statistical analysis of partially sequenced genomes as has been done by several researchers in the past [e.g., Blaisdell, Campbell and Karlin (1996) extensively analyzed partially sequenced bacteriophage genomes using word frequencies]. If different parts of the same genome have some basic structural similarities, such analysis will be justified, and one can make valid statistical inference based on incomplete genome data.

2.1. An analysis of the complete genome of Baker's yeast

We will now present an analysis of the full genome of *Saccharomyces cerevisiae* i.e., baker's yeast. Yeast genome consists of sixteen chromosomes, and the smallest one among them is a sequence of 230,209 nucleotides while the largest one is a sequence of 1,583,176 nucleotides. Clearly, it is not possible to compare and judge the structural similarities of such large sequences with so much variations in their sizes using any of the standard sequence alignment and homology techniques (see Section 3 for some details). Besides, the biological functions of many parts of the genome are not well understood, and certain parts of the genome might be biologically non-functional (i.e., they might consist of the so called *junk DNA*). This makes total comparison of these sixteen chromosomes through functional homology virtually impossible. However, it is possible to compare the frequencies of various DNA words in these chromosomes, and this can lead to an idea of the extent of their structural similarities.

Interestingly, each of these sixteen DNA sequences contains about 30% A's, 30% T's, 20% C's and 20% G's, and there is only negligible variation in mononucleotide (1-word) frequencies among these sixteen chromosomes. Another noteworthy fact is that each of the $4^6 (= 4,096)$ 6-words occurs with positive frequency in each of the chromosomes, and in each chromosome, some of the $4^8 (= 65,536)$ 8-words do not occur. In each of the seven larger chromosomes, all of the $4^7 (= 16,384)$ 7-words occur with positive frequency, while in each of the nine smaller chromosomes, some of the 7-words turn out to be missing. If k-word frequencies are to be used to summarize a given DNA sequence, a question that naturally arises is what value(s) of k one should use. There does not seem to be a complete and general mathematical solution to this problem that is known in the literature though there are various suggestions put forward by different authors based on specific probability models for sequences that are applicable to special situations. For instance, in the case of very large words, many of those words are expected to be missing in a given DNA string. Rahmann and Rivals (2000) investigated expected number of missing words of a specified length in a random text. Let us now consider the following result.

Result 2.1. *If the total length of a sequence is N, for any $k > \log_4(N - k + 1)$ (i.e. $4^k > N - k + 1$), some of the k-words will be missing in that sequence.*

Proof. Clearly, there are 4^k possible distinct k-words that can be formed using the alphabet $\{A, T, C, G\}$. On the other hand, the total number of distinct k-words in a sequence of length N must be less than $(N - k + 1)$. □

Since $N \gg k$, we will have $\log_4(N - k + 1) \approx \log_4 N$, and it appears that the frequencies of words having size larger than $\log_4 N$ may not be statistically very informative due to the "sparsity" of such words in the sequence. For the sixteen chromosomes in yeast, the values of $\log_4 N$ vary between 8.91 and 10.30. We feel that $\log_4 N$ can be used as a preliminary empirical guideline for choosing appropriate value(s) of k for carrying out meaningful statistical analysis based on k-word frequencies.

It will be appropriate to note here that a word with length of the order of $\log N$ or bigger in an i.i.d or a Markov sequence of length N is typically considered a *rare word* and Poisson or compound Poisson approximation for its frequency count has been studied in detail in the literature [see e.g. Arratia, Goldstein and Gordon (1990), Godbole and Schaffner (1993), Geske et al. (1995), Arratia et al (1996), Reinert and Schbath (1998, 1999), Reinert, Schbath and Waterman (2000)]. However, Result 2.1 above is completely deterministic in nature and is not derived from any specific probability model.

In Figures 2.1 and 2.2 we have plotted the frequencies for 2-, 3-, 4- and 5-words in all sixteen chromosomes. In each plot, different DNA words have been plotted along the horizontal axis, and their frequencies have been plotted along the vertical axis. In the horizontal axis, words having the same size have been arranged in a specific manner, where the letter A comes first and then the letters T, G and C come in that order. For example, in the case of 2-words, the words $AA, AT, AG, AC, TA, TT, \cdots, CG, CC$ come in that order and are plotted along the horizontal axis of the first graph. Similarly, for 3-words, we have $AAA, AAT, AAG, AAC, \cdots, CCA, CCT, CCG, CCC$ appearing along the horizontal axis in that order. In each of the plots, consecutive points have been joined by line segments to produce a continuous line graph for each chromosome in order to facilitate visual comparison of relatively high and low word frequencies in different chromosomes through ups and downs in the graphs. Graphs for 16 chromosomes are overlaid in each plot to enable a visual comparison of word frequency patterns for different chromosomes. There is a high degree of similarity in the word frequencies for all of these chromosomes, and it is clearly visible in the plots. The frequency plots for different chromosomes are virtually indistinguishable. The striking similarities in the word frequencies of these sixteen chromosomes are further highlighted in figure 2.3, where for word sizes 2 through 5, we have plotted the maximum and the minimum frequencies of each of those ten words of a specific size that have the *largest ranges*. Since the *range of a word* is the difference between the maximum and the minimum frequencies of the word across sixteen chromosomes, these are the words with largest variabilities in their frequencies among different chromosomes. Very small ranges for all these words are clearly noticeable in all cases re-confirming a very similar word frequency patterns in all sixteen chromosomes.

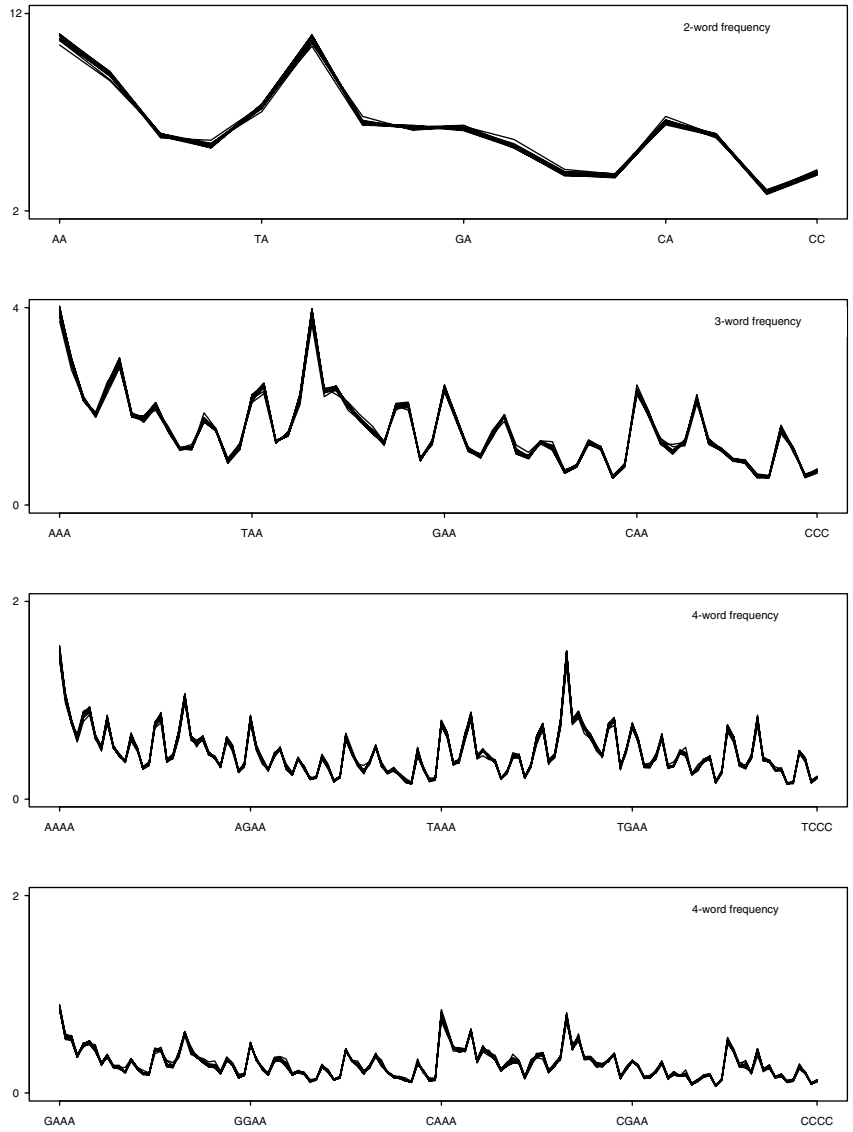


Fig. 2.1. Frequency (in %) plots for 2-, 3- and 4-words for sixteen chromosomes of baker's yeast.

In order to make a further comparison of all sixteen chromosomes in terms of their structural similarities reflected in the frequency distributions of their DNA words, one may also consider relative ranks of different k-words in the chromosomes in terms of their frequencies. Ranks of k-words quantify their relative abundance or scarcity in a sequence, and comparison of ranks of each k-word for different chromosomes can reveal important structural features and similarities of the chromosomes. Let $(r_w)_{w\in W_k}$ be the vector of ranks of the k-words computed by

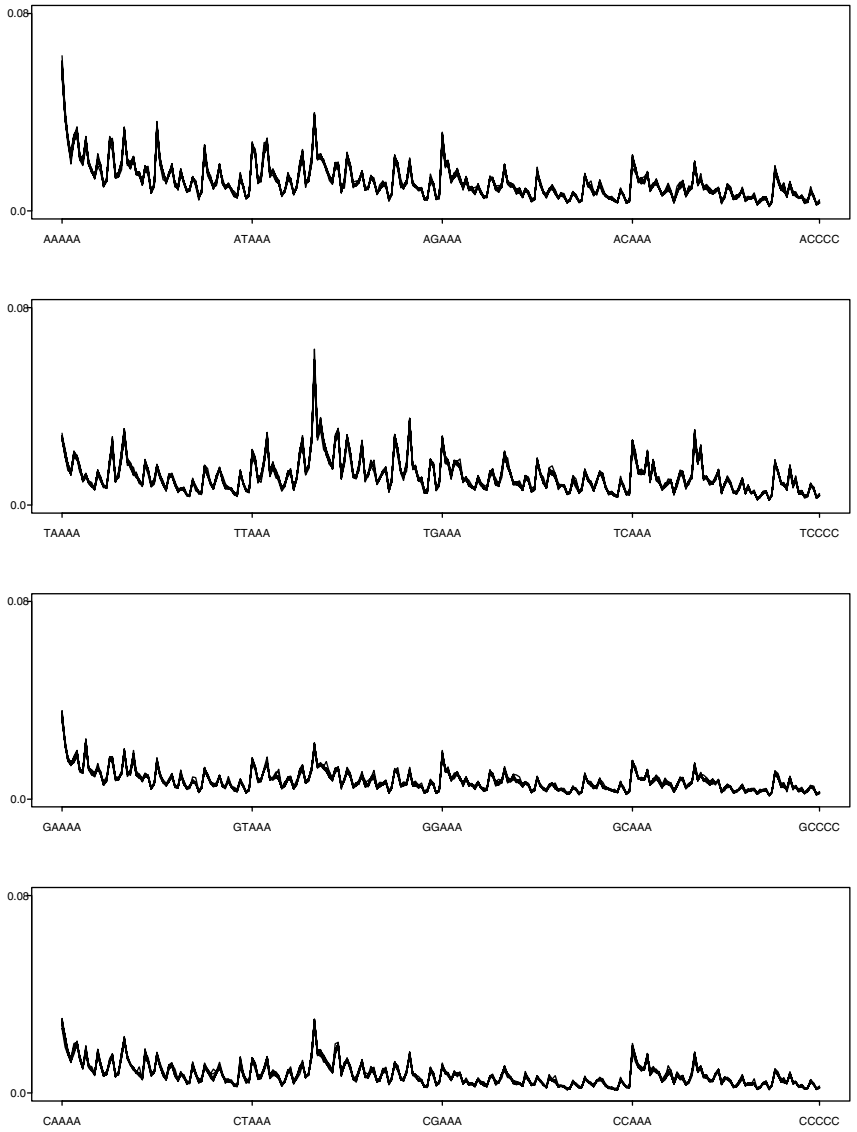


Fig. 2.2. Frequency (in %) plots for 5-words for sixteen chromosomes of baker's yeast.

comparing their frequencies in a DNA sequence. For a pair of chromosomes α and β, a simple way to compare their associated rank vectors $(r_w)_{w\in W_k}$ and $(s_w)_{w\in W_k}$ (say) for k-words will be to compute some form of rank correlation coefficient based on the two rank vectors. For instance, one can compute Spearman's rank correlation coefficient

$$R_S^{(k)}(\alpha,\beta) = \frac{\sum_{w\in W_k} r_w s_w - 4^{k-1}(4^k+1)^2}{4^k(4^k+1)(4^k 2+1)/6 - 4^{k-1}(4^k+1)^2},$$

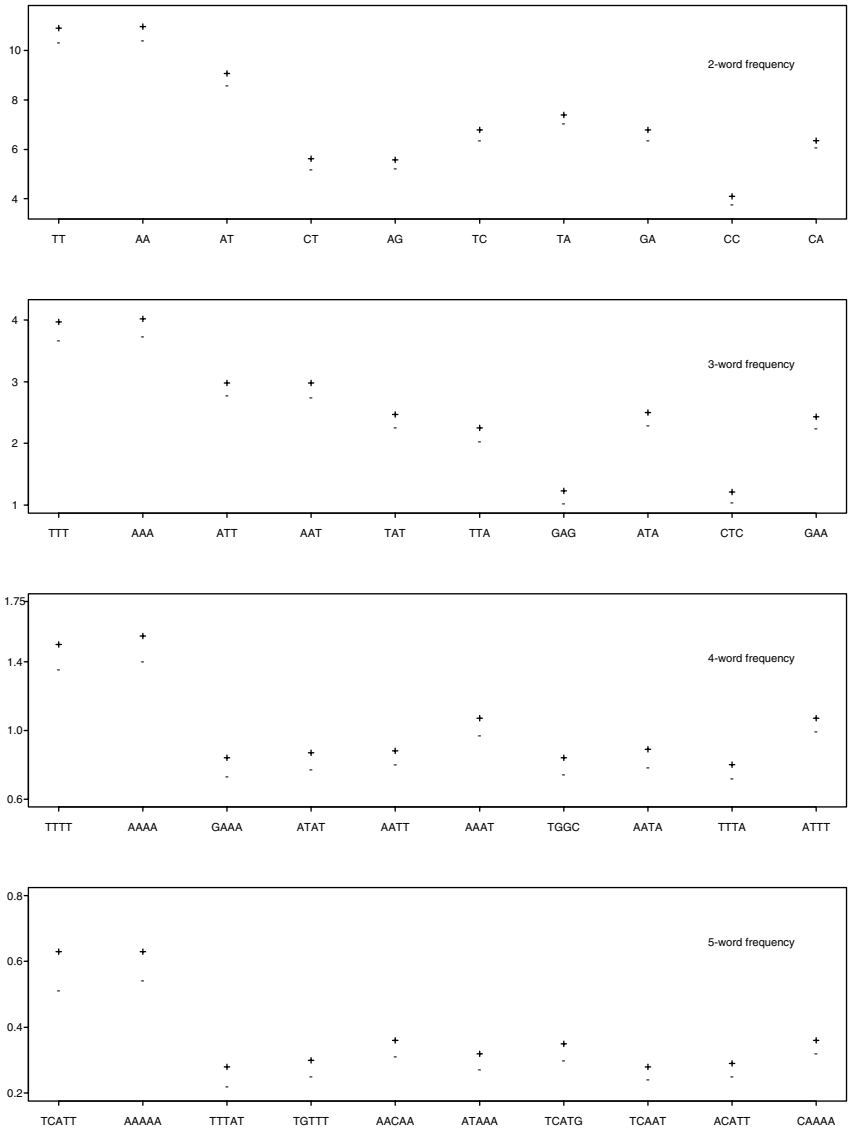


Fig. 2.3. Maximum (+) and minumum (−) frequencies of words of yeast chromosomes.

the value of which lies between -1 and $+1$. Strong structural similarity between α and β, which is reflected in similar patterns of relative abundance and scarcity of different DNA k-words in the two sequences, will make the value of $R_S^{(k)}(\alpha, \beta)$ very close to $+1$. On the other hand small and negative values (if any) of $R_S^{(k)}(\alpha, \beta)$ will be an indication of structural dissimilarity between α and β. In Figure 2.4 we have plotted along the vertical axes the values of pairwise rank correlation coefficients for all sixteen chromosomes in the cases of 3-, 4-, 5- and 6-words. The

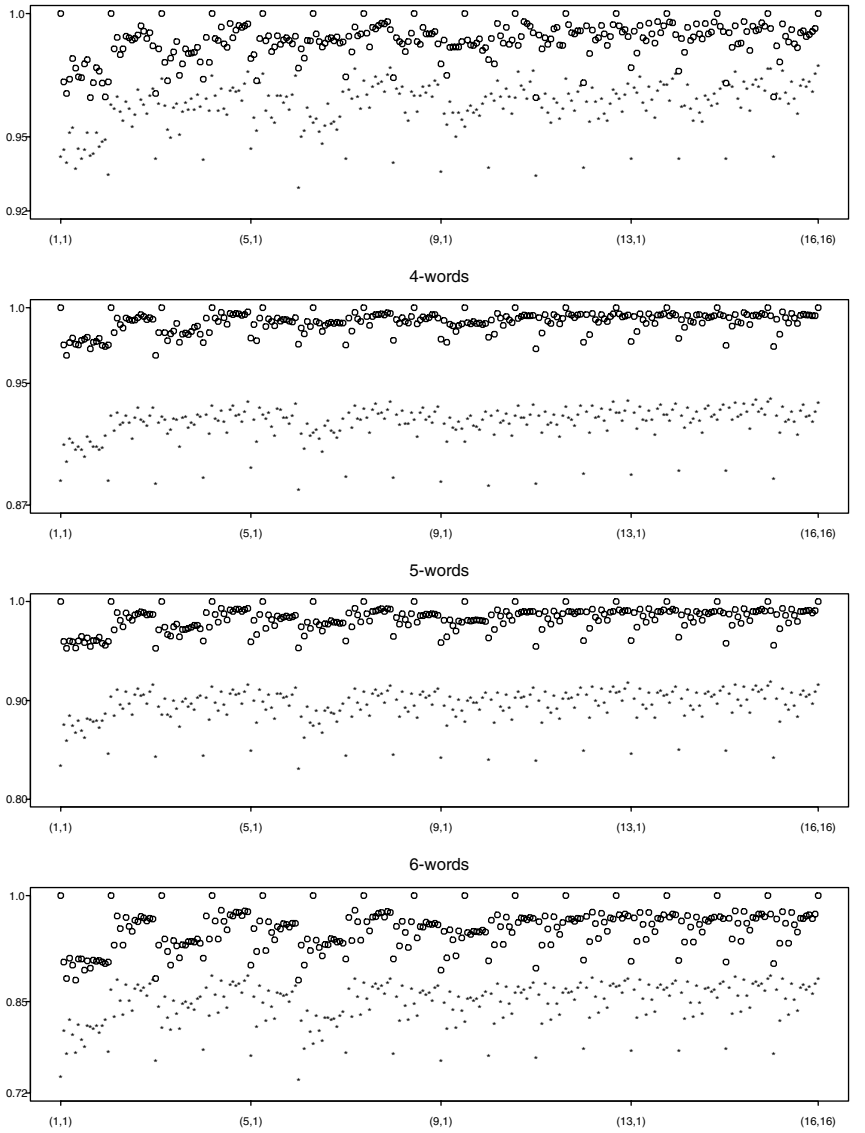


Fig. 2.4. Spearman's rank correlations for chromosomes of yeast (○) and their cross rank correlations with random sequences (⋆).

cases of ties among word frequencies were handled in usual ways for this rank based analysis. The pairs arranged along the horizontal axes correspond to different chromosome pairs, and in each case they have been arranged in the order (1, 1), . . . , (1, 16), (2, 1), . . . , (2, 16), (3, 1), . . . , (15, 16), (16, 1), . . . , (16, 16). In all cases, the values of rank correlation coefficients are quite high as is evident in the figures indicating a very strong structural similarity of all sixteen chromosomes.

Let us observe here that the plots for dinucleotide frequencies for these chromosomes do not just reflect their mono-nucleotide frequencies. Since the frequencies of A and T are essentially same (i.e., $f_A \approx f_T$) and so are the frequencies of C and G (i.e., $f_C \approx f_G$) for each of the chromosomes, in the case of a straight-forward reflection of mono-nucleotide frequencies in di-nucleotide frequencies, one would have observed

$$f_{AA} \approx f_{TT} \approx f_{AT} \approx f_{TA},$$

$$f_{CC} \approx f_{GG} \approx f_{CG} \approx f_{GC} \text{ and}$$

$$f_{AC} \approx f_{TC} \approx f_{AG} \approx f_{TG} \approx f_{CA} \approx f_{CT} \approx f_{GA} \approx f_{GT},$$

which does not happen to be true here. This implies that the "doublets" have their own frequency patterns that are not derivable from "singlet" frequencies, and the strong similarity of the di-nucleotide frequencies of all sixteen chromosomes is not a mere consequence of their similar mono-nucleotide composition.

We will next make an attempt to investigate to what extent 2-word frequencies are reflected in k-word frequencies for $k \geq 3$. This can be done as follows. We can generate sixteen random sequences each using the alphabet $\{A, T, C, G\}$ and a 1-step homogeneous Markov process whose 1-step transition probabilities are estimated from the 2-word frequencies of a chromosome (so that different chromosomes are used for different sequences). Then we can compare the k-word frequencies for sixteen random sequences with those for sixteen actual chromosomes for $k \geq 3$. Let α^* denote the random Markov sequence generated using the 2-word frequencies of the chromosome α. For each of the sixteen chromosomes, we will make the sequence α^* to have the same length as α, and both will begin with the same first letter. We have computed the cross-correlation coefficients of ranks of k-words given by $R_S^{(k)}(\alpha^*, \beta)$ for $k = 3, 4, 5$ and 6 and all 256 pairs α^*, β of sixteen chromosomes. The values have been plotted in Figure 2.4 together with the pairwise rank correlation coefficients of actual chromosomes. It is clearly noticeable in all the plots that the values of cross-correlations of ranks for k-words of actual chromosomes and random Markov sequences are consistently smaller than the corresponding values of pairwise rank correlations among actual chromosomes. This is an evidence for the fact that k-words for $k \geq 3$ have their intrinsic frequency patterns in each of these chromosomes that are not derivable from their 2-word frequencies. This analysis on the one hand proves the existence of a very strong structural similarity among the sixteen chromosomes, which is reflected in their different k-word frequencies, and on the other hand, it reveals that the frequencies of higher order words in this case have their own patterns that are not completely derived from lower order word frequencies.

Let us clarify here that we are using rank correlations merely as descriptive statistics to measure the similarity between any pair of DNA sequences as reflected in their oligonucleotide frequencies. We were motivated to use rank correlations after observing how nicely the peaks and the valleys tend to match for all sixteen chromosomes in the frequency plots presented in Figures 2.1 and 2.2. Rank correlations are not being used here to test any statistical hypothesis concerning the order of a fitted Markov model in a formal manner – their use in this paper is limited to

developing descriptive tools for making visual comparisons as demonstrated in the plots in Figure 2.4. Reinert, Schbath and Waterman (2000, section 2.3) discussed the use of likelihood ratio tests for determining the appropriate order of the Markov chain that would fit a given sequence data. It is possible to carry out such tests to establish the inadequacy of first order homogenous Markov models for these sixteen chromosomes. However, our rank correlation based analysis is technically much simpler than more formal procedures like likelihood ratio tests, and the results can be presented using easily comprehensible graphs with nice visual evidence. The fact that the cross-correlation between a real chromosome and a simulated chromosome is always smaller than that between two real chromosomes for all possible 240 pairwise comparisons makes a fairly convincing argument for the inappropriateness of the first order homogenous Markov model for the Yeast chromosomes.

2.2. Analysis of some bacteriophage genomes

Statistical analysis of many virus genomes based on DNA word frequencies has been reported in the literature by several authors [Purm, Rodolphe and de Turckheim (1995), Blaisdell, Campbell and Karlin (1996), Leung, Marsh and Speed (1996), etc.]. The neutral mutation and random drift hypothesis, which is popularly known as *the neutral theory of evolution* [see Kimura (1983)], states that evolutionary changes at the molecular level and variability within species is mostly caused by mutation and random genetic drift, and it is not so much due to *Darwinian selection*. The proponents of neutral theory advocate that in the case of molecular evolution the intensity of Darwinian selection is so weak that mutation pressure and random drift become the prevalent factors there. An important characteristic of virus genomes is the high speed of their evolution due to faster point mutations and genomic re-arrangements compared to nuclear genes of higher organisms. In a way this has made virus genomes convenient materials for the establishment of the neutral theory of evolution, and perhaps this can be a good justification for their probabilistic modeling and statistical analysis too. Due to limited size of a virus genome, it can be very useful in studying the evolution of DNA through structural comparison of different fragments of the genome.

For our analysis, we have chosen four bacteriophage genomes: $\Phi X174$ (5,386 nucleotides), $G4$ (5,577 nucleotides), $F1$ (6,407 nucleotides) and $PF3$ (5,833 nucleotides), and our source for their sequences is the EMBL Data Bank. All of them are single stranded lytic phages, and *Escherichia coli* is the host for the first three while the host for the last one is *Pseudomonas aeruginosa*. We have considered five fragments of each of these bacteriophage genomes. The first two fragments are formed by taking two equal (or almost equal) and disjoint halves of the entire genome, and the other three fragments are formed by taking three equal (or almost equal) and disjoint parts of the full genome. We have labelled the five fragments corresponding to each of the four viruses as follows: A1–A5 (for $\Phi X174$), B1–B5 (for $G4$), C1–C5 (for $F1$) and D1–D5 (for $PF3$). The results of average linkage cluster analysis based on simple l_1-distances of 3- and 4-word frequencies of twenty fragments are presented in the dendrograms in Figure 2.5. Here the l_1-distance between a pair of frequency vectors is given by $\sum_w |f_w - g_w|$, where f_w's and

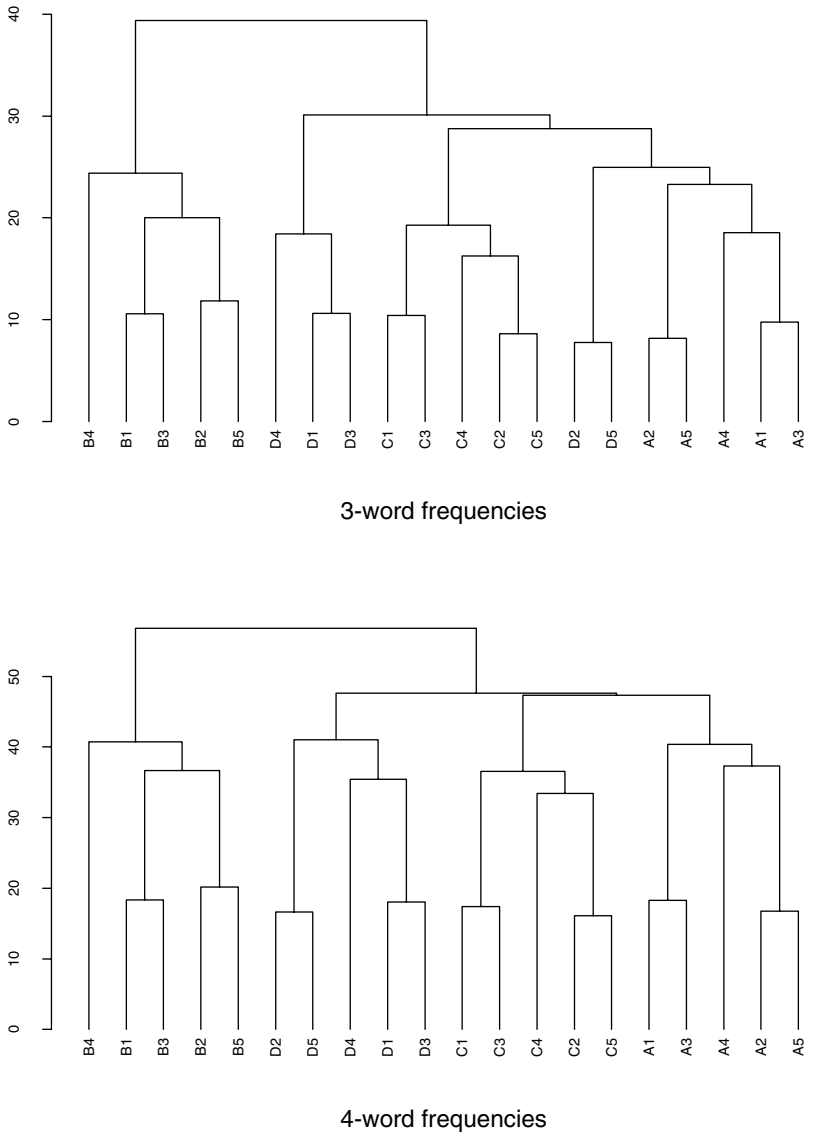


Fig. 2.5. Dendrograms for average linkage cluster analysis of fragments of phage genomes.

g_w's are the frequencies corresponding to 3-words (or 4-words) in two fragments of the same (or different) phage genome(s). In both the dendrograms, fragments of the same bacteriophage genome exhibit an overall tendency of forming their own separate clusters, and this is indicative of the presence of some form of structural signature in different fragments of the same phage that is amply captured by word frequencies. The four bacteriophages chosen here are known to have certain biological similarities among themselves, and there is not much variations in the sizes of their genomes. In spite of that there appears to be some intrinsic patterns

in the word frequencies of different fragments of the same phage genome that can adequately distinguish them from the fragments of a different phage genome. Note that here we have used only the l_1-distances among different frequency vectors. This distance was chosen because of its simplicity and its interpretability in terms of direct comparison of different word frequencies in a pair of sequences. However, we expect that many other reasonable distances will lead to similar results.

We have observed in course of our data analysis that k-word frequencies are not very informative when the value of k is too small *nor if k is too large*. For instance, if we carry out average linkage cluster analysis of the twenty fragments of four bacteriophage genomes considered above using the l_1-distances of 6-word and 7-word frequencies, we obtain the dendrograms presented in Figure 2.6. It is clearly visible in the two dendrograms that the results become worse as we move from hexanucleotides to heptanucleotides, and in the latter case heterogenous clusters are formed, where fragments of different phage genomes occur in the same cluster. We now state the following interesting result.

Result 2.2. *For any two distinct DNA sequences S and S^*, let $l(S, S^*)$ denote the largest possible value of k for which S and S^* contain a common k-word. In other words, for any $k > l(S, S^*)$, the set of k-words that occur in S will be completely disjoint from the set of k-words in S^*. Hence, if $(f_w)_{w\in W_k}$ and $(g_w)_{w\in W_k}$ are the frequency vectors for k-words in the sequences S and S^* respectively, we must have*

$$\sum_{w\in W_k} f_w g_w = 0 \quad \text{and} \quad \sum_{w\in W_k} |f_w - g_w| = 2$$

for all $k > l(S, S^*)$. In general, if we have a sample $\mathcal{S} = \{S_1, \ldots, S_n\}$ of n DNA sequences, and

$$l(\mathcal{S}) = \max_{1\le i<j\le n} l(S_i, S_j),$$

the l_1-distance between S_i and S_j based on k-word frequencies will be exactly equal to 2 for all $i \neq j$ and any $k > l(\mathcal{S})$.

Proof. We begin by observing that for any $k > l(S, S^*)$, the set of k-words that occur in S will be completely disjoint from the set of k-words in S^*. Hence, if $(f_w)_{w\in W_k}$ and $(g_w)_{w\in W_k}$ are the frequency vectors for k-words in the sequences S and S^* those frequency vectors will have the property that whenever the frequency of a k-word in one frequency vector is positive, the frequency of the same k-word in the other frequency vector must be zero. Since the frequencies of any vector are non-negative quantities adding up to 1, the result follows. □

The main implication of Result 2.2 is that cluster analysis of the sequences in $\mathcal{S}$ using l_1-distances of the frequencies of words larger than $l(\mathcal{S})$ will produce a dendrogram with a non-informative structure, where one after another different sequences will be linked in a chain, and no subgroups or clusters among these sequences will be visible irrespective of whether such subgroups really exist there or not. The value of $l(\mathcal{S})$ may sometimes turn out to be too large. However, it is our empirical experience that even for moderately large values of k, due to sparsity

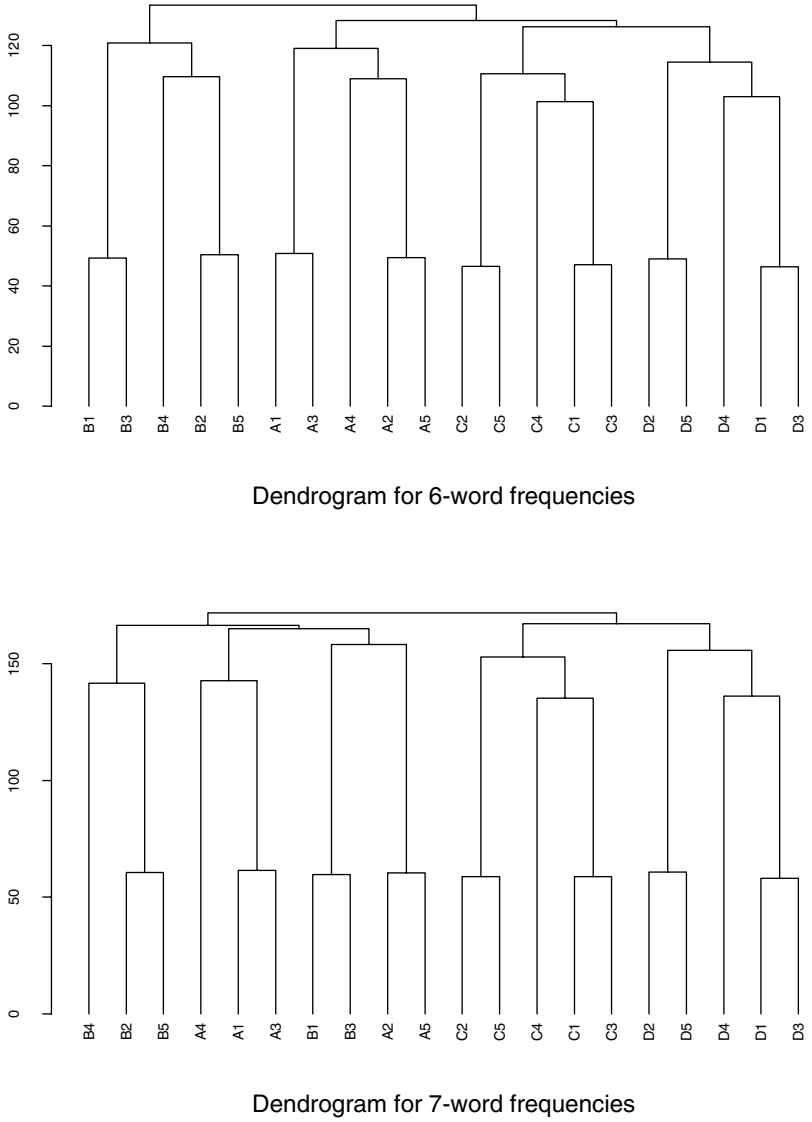


Fig. 2.6. Average linkage cluster analysis of fragments of bacteriophage genomes.

of k-words in the sequences, there is usually very little variation among the pairwise l_1-distances computed from k-word frequencies of different sequence pairs. As a result, cluster analysis using those k-word frequencies will produce a chain structure in the dendrogram and cannot reveal any of the subgroups or clusters that might be visible if analysis is carried out with the frequencies of smaller words.

An important issue that arises at this point is how to choose those k-words that are more informative compared to other k-words in determining the structural

signature of any specific virus genome. This is equivalent to identifying those k-words whose frequencies do not vary much among different fragments of the same bacteriophage but vary substantially among the fragments coming from different phages. Given n different phage genomes and m fragments of each genome, let $f_w^{a,\alpha}$ denote the frequency of the k-word w in the fragment α of the phage a. Define now the ratio

$$F(w) = \frac{m^{-1}(n-1)^{-1} \sum_{a \neq b} \sum_{\alpha,\beta} |f_w^{a,\alpha} - f_w^{b,\beta}|}{(m-1)^{-1} \sum_{a} \sum_{\alpha \neq \beta} |f_w^{a,\alpha} - f_w^{a,\beta}|},$$

which may be interpreted as the l_1-distance version of the traditional F-ratio of between and within group variations used in statistical analysis of variance. Since all along we have used the l_1-distance so far, we have defined these F-ratios also using the same distance instead of the more conventional version based on squared Euclidean distance. However, we do not expect the main findings here to change very much if we use the more conventional version of F-ratio or any other version of it based on some reasonable distance measure. Note that a large value of $F(w)$ corresponds to a word w for which the dispersion among the frequencies $f_w^{a,\alpha}$ across different phages (i.e., a's) is larger than the dispersion among these frequencies across different fragments (i.e., α's) of the same phage. One can then rank the k-words according to their F-values, and we will call it the F-ranks of the k-words. It seems meaningful to choose k-words with high F-ranks in order to capture structural signatures of different virus genomes.

In the case of twenty fragments of the four phage genomes that we have formed, we selected fifteen 3-words and fifteen 4-words that have highest F-ranks among all 3-words and 4-words respectively. Then these word frequencies were utilized to carry out Chernoff's face analysis [see Chernoff (1973)], where fifteen frequencies were allowed to determine the following fifteen feature parameters for the faces: 1 for area of face, 2 for shape of face, 3 for length of nose, 4 for location of mouth, 5 for curve of smile, 6 for width of mouth, 7, 8, 9, 10, 11 for location, separation, angle, shape, width of eyes respectively, 12 for location of pupil, and 13, 14, 15 for location, angle and width of eyebrows. Results of the analysis are presented in Figures 2.7 and 2.8. In both the figures, the similarity among the faces in each column corresponding to different fragments of the same virus genome is quite noticeable, and so are the dissimilarities of the faces in different columns that correspond to different virus genomes.

3. Word frequencies and phylogenetic relationships

Phylogenetic relationships among different organisms are of fundamental importance in biology, and one of the prime objectives of DNA sequence analysis is the construction of phylogenetic trees for understanding the evolutionary history of organisms. Many different methods for phylogenetic analysis of DNA sequence data have been proposed and studied in the literature. Readers are referred to review articles by Felsenstein (1983, 1988) and Nei (1996) and the books by Weir (1996)

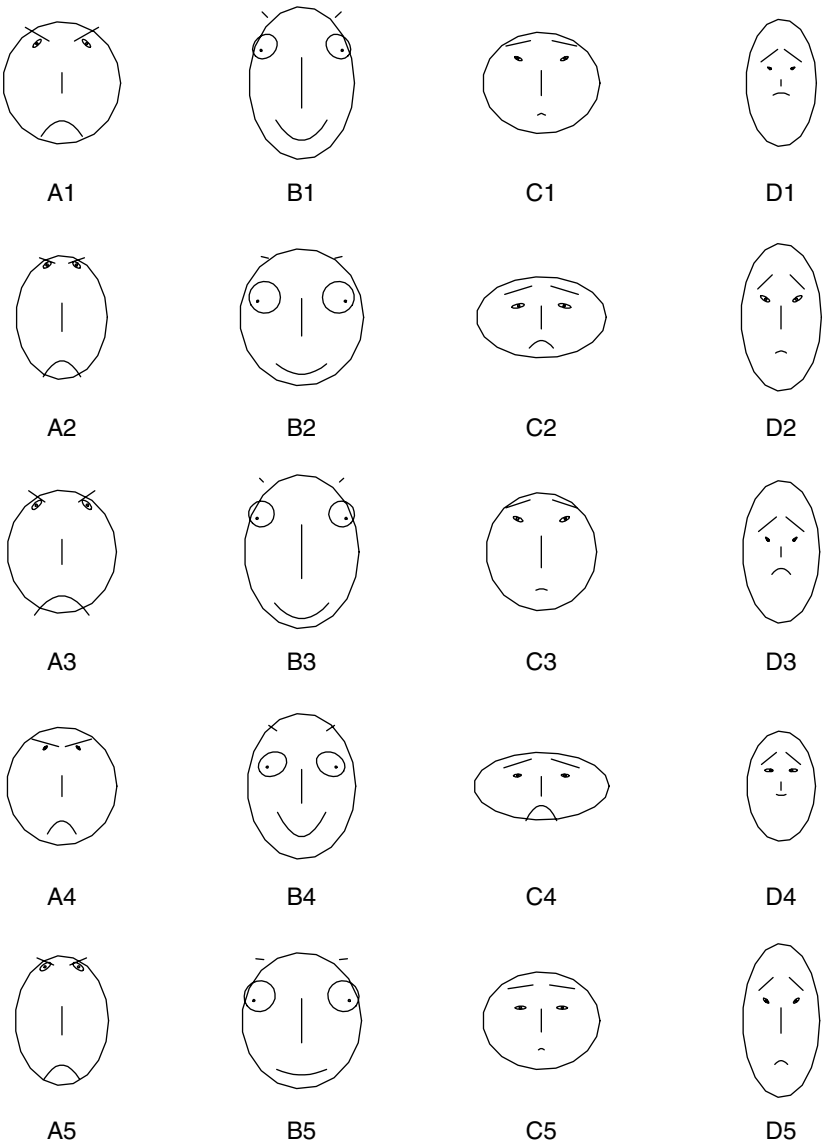


Fig. 2.7. Chernoff's faces drawn with fifteen highest F-rank 3-word frequencies for fragments of phage genomes.

and Lange (1997) for extensive discussion of such methods and related statistical problems. A number of standard programs [e.g., GCG by Devereux, Haeberli and Smithies (1984), PHYLIP by Felsenstein (1989), PUZZLE by Strimmer and von Haesler (1996), PAUP by Swofford (1997)] are available to construct phylogenetic trees from *evolutionary distances* among different species based on DNA sequences extracted from those species.

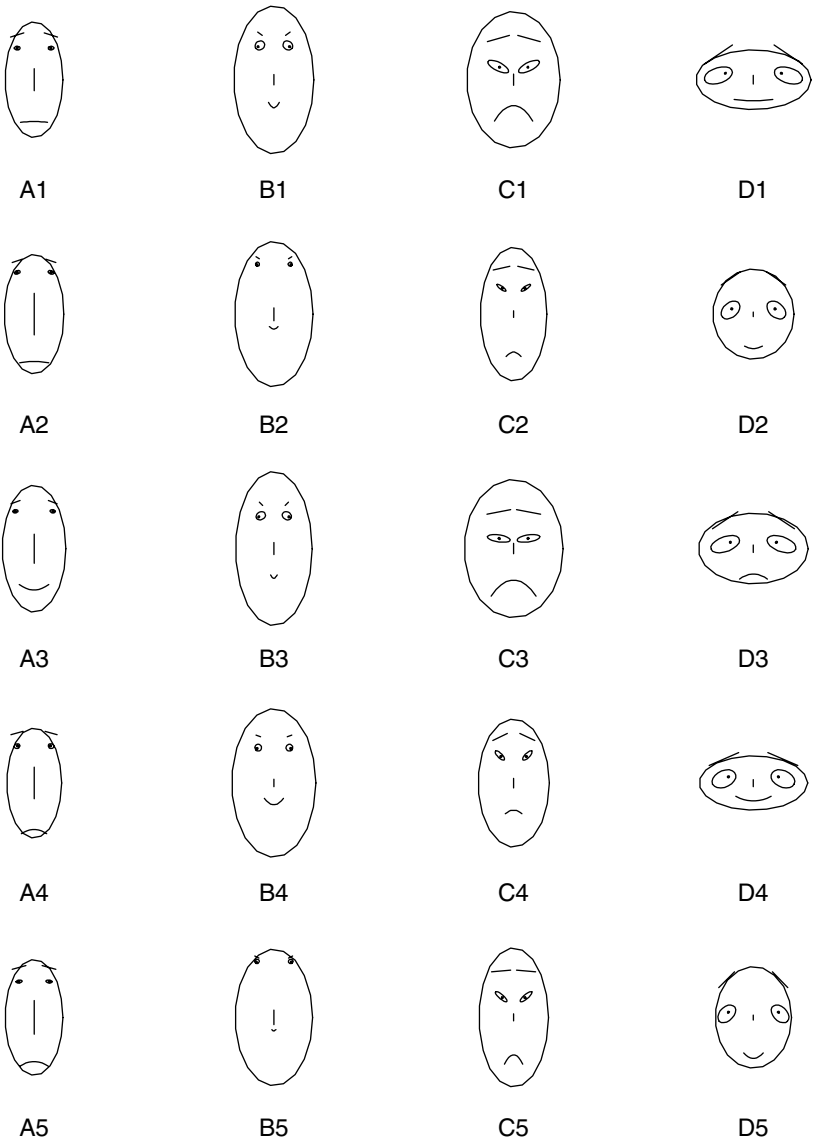


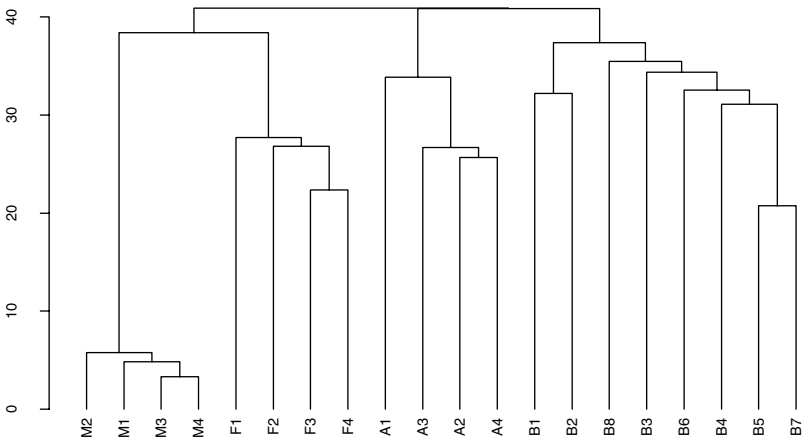
Fig. 2.8. Chernoff's faces drawn with fifteen highest F-rank 4-word frequencies for fragments of phage genomes.

Suppose that we have n DNA sequences denoted by $S_1, S_2, \ldots, S_n$, and we want to compute an evolutionary distance $d(S_i, S_j)$ for each pair (S_i, S_j) $1 \leq i < j \leq n$, which can be used to infer about phylogenetic relationships among the sequences S_i's. For such purpose, stretches of DNA sequences having some common function in different species are aligned and evolutionary distances are computed by counting the nucleotide replacements at different locations (or sites) [see e.g. Waterman (1995), Needleman and Wunsch (1970)]. It is usually necessary to create gaps for

such alignment of sequences in order to take care of possible insertion and deletion of certain parts of a DNA sequence in course of evolution, and one needs to impose proper penalty values for such gaps while computing evolutionary distances. Further, often it is required to assign different weights to different types of nucleotide replacements at a specific site in order to get a biologically meanigful distance. Choice of those weights as well as the creation of gaps and the setting of penalty values for those gaps in aligned sequences are based on very subjective considerations, and on many occasions they are known to lead to controversial phylogenetic trees. Besides, such methods though fairly popular in the study of molecular evolution, are extremely computer intensive and feasible only if neither the size of each of the sequences nor their number n is too large. Alternatively, one can use the l_1-distance of frequency vectors associated with a pair of sequences as in Section 2.2, and in that case we have $d(S_i, S_j) = \sum_{w \in W_k} |f_w - g_w|$, where $(f_w)_{w \in W_k}$ and $(g_w)_{w \in W_k}$ are frequency vectors corresponding to k-words in two sequences S_i and S_j respectively. Clearly, the l_1-distance for k-word frequencies are conceptually and computationally much simpler than the evolutionary distance computed from nucleotide substitutions at different sites and occurrence of inserted and deleted segments in aligned sequences.

We now present an analysis of a data set that consists of 16S and 18S ribosomal RNA sequences for twenty organisms. These sequences have their sizes varying between 1471 and 1869 nucleotides, and the source of the data is EMBL Data Bank. The organisms chosen in this case are as follows. Eight bacteria: (B1) *Acholeplasma laidawii*, (B2) *Mycoplasma gallisepticum*, (B3) *Heliobacterium chlorum*, (B4) *Deinococcus radiodurans*, (B5) *Rochalimaea quintana*, (B6) *Anaplasma marginale*, (B7) *Brucella abortus* and (B8) *Bacteroids fragilis*; four archaea: (A1) *Methanothrix soehngenii*, (A2) *Halobacterium marismotui*, (A3) *Halobacterium volcanii* and (A4) *Halococcus morrhuae*; four fungi: (F1) *Phytophthora megasperma*, (F2) *Mucor racemosus*, (F3) *Coccidioides immitis* and (F4) *Candida albicans*; and four mammals: (M1) *Rattus norvegicus* (rat), (M2) *Oryctolagus cuniculus* (rabbit), (M3) *Mus musculus* (mouse) and (M4) *Homo sapiens* (human). The results of single and average linkage cluster analysis based on 4-word frequencies are presented in Figure 3.1. Here clustering has been carried out in a hierarchical agglomerative manner, and as usual in each stage the distance between a pair of clusters C and C' is defined as the minimum of the distances $d(S, S')$ for single linkage procedure and as the average of the distances $d(S, S')$ in the case of average linkage procedure for all possible pairs (S, S') with $S \in C$ and $S' \in C'$ [see e.g., Everitt (1993)]. Let us observe that in both the dendrograms in Figure 3.1 the bacteria, the archaea, the fungi and the mammals form distinct clusters, and the eukaryotes clearly separate out from the prokaryotes. It will be appropriate to note here that 4-words are the words of smallest size for which the dendrogram tree based on the l_1-distances among the frequencey vetors consists of such biologically homogenous and meaningful clusters. If the same analysis is repeated with smaller words (i.e. 2-words or 3-words), we do not get a dendrogram with such a nice clustering of bacteria, archaea, fungi and mammmals in four distinct groups.

In many ways, cluster analysis is different from conventional phylogenetic analysis though they have some intrinsic similarities. For instance, the well known



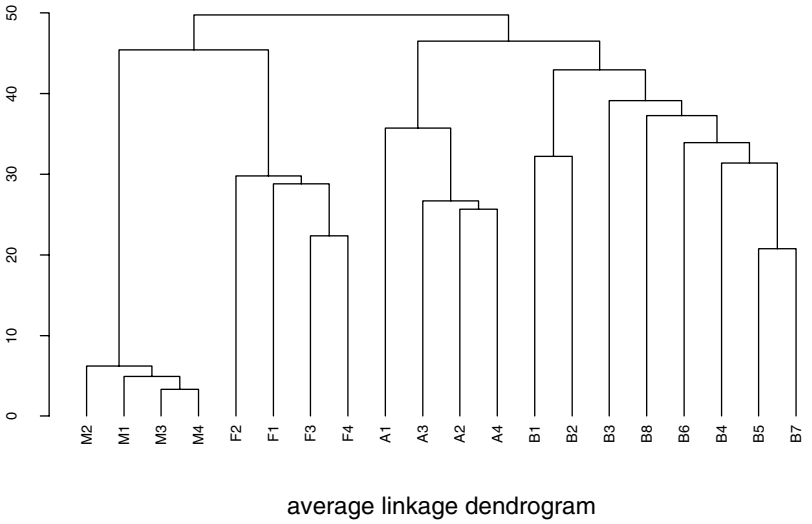


Fig. 3.1. Cluster analysis of 16S- and 18S-RNA data using l-1 distance of tetranucleotide frequencies.

UPGMA phylogenetic tree is virtually same as the average linkage dendrogram produced from a given distance matrix. However, there are many other methods for constructing phylogenetic trees, and they usually contain more information about the evolutionary history than a dendrogarm tree such as the time needed for evolutionary divergence, differential rates of evolution across different edges of the tree

and possible locations of the unobserved ancestors of the observed species in the evolutionary tree. Nevertheless, the two types of trees are related in some fundamental ways, and the species that are close neighbours in a phylogenetic tree are expected to be close neighbours in a dendrogram tree also and vice versa if the same evolutionary distance is used in the phylogenetic analysis and the cluster analysis. Further, the merging evolutionary distances of different clusters in a dendrogram tree are expected to be related to evolutionary time needed for divergence. Consequently, a dendrogram tree, which is usually computationally much simpler to construct than standard phylogenetic trees, is a very useful and convenient tool for studying phylogenetic relationships among a given set of species.

It is quite clear from our cluster analysis presented in this section that different biological species exhibit different distributions of DNA words in their ribosomal RNA sequences, and this can be amply utilized to cluster different organisms into homogenous biological groups by a straight-forward comparison of their DNA word frequencies. Obviously, the technique used here is much simpler than procedures that are commonly used in the construction of phylogenetic trees based on some evolutionary models (e.g. maximum likelihood or maximum parsimony phylogenetic analysis using alligned sequences). It appears that one can use distances based on DNA word frequencies as an alternative and effective statistical tool for quick and convenient determination of phylogenetic relations among different biological species. Recently Chaudhuri and Das (2001, 2002) carried out cluster analysis of mitochondrial genomes of several fish, amphibia, reptiles, birds and mammals using word frequencies and demonstrated how such analysis can reveal facts that are of great significance in the evolutionary biology of vertebrates.

4. Concluding remarks and discussion

An extensive analysis of dinucleotide frequencies (i.e., frequencies of doublets or 2-words) in various sequences from eukaryotic and prokaryotic genomes was carried out by Nussinov (1980, 1981, 1982, 1984a, 1984b) in an attempt to explore and understand compositional heterogeneity in different DNA sequences reflected in the preference for the usage of certain doublets, which can be measured by the bias in doublet frequencies. In a series of papers, Karlin and his co-authors [see e.g., Karlin and Cardon (1994), Karlin and Campbell (1994), Karlin and Ladung (1994), Karlin, Ladunga and Blaisdell (1994), Blaisdell, Campbell and Karlin (1996)] made a detailed study of frequencies of DNA words of different sizes (mainly doublets, triplets and quadruplets) for sequences from various bacteriophage genomes as well as bacterial (i.e., prokaryotic) and eukaryotic genomes. Their analyses were based on *relative abundance distance* computed from word frequency data, and their emphasis was on general relatedness, similarities and contrasts among genomic sequences, which they distinguished from the problem of construction of phylogenetic trees.

Godbole and Schaffner (1993), Geske et al. (1995), Martindale and Konopka (1996), Reinart and Schbath (1998, 1999) and Reinert, Schbath and Waterman (2000) investigated the possibility of modeling the frequency distribution of oligonucleotides in DNA sequences using different types of probability laws. Unusual

frequencies of certain DNA words in the *Escherichia coli* genome and possible statistical and bigological implications of such over- and under-representation of those words have been investigated by Phillips, Arnold and Ivarie (1987a, 1987b) whose analysis was based on certain Markov chain models for DNA sequences. Purm, Rodolphe and de Turckheim (1995) and Leung, Marsh and Speed (1996) carried out a similar analysis for some virus genomes [see also Schbath, S., Prum, B. and de Turckheim (1995)]. While Markov and Hidden Markov Models [see e.g., Churchill (1989) and Muri (1998)] are two very popular and widely studied models for DNA sequences, it has been pointed out by several authors [see e.g., Pevzner (1992), Karlin and Brendel (1993) and Leung, Marsh and Speed (1996)] that homogeneous Markov models quite often do not fit observed DNA sequence data very well. On the other hand, it is also well known that [see e.g. Waterman (1995), Reinert, Schbath and Waterman (2000)] word frequencies are $N^{1/2}$-consistent estimates of corresponding word probabilities in a random sequence generated from a finite alphabet, and they are approximately normally distributed when the sequence size N tends to infinity so long as the sequence satisfies stationarity and the ρ-mixing property, both of which are much weaker conditions than the assumption of homogeneous Markov property. It was noted in Reinert, Schbath and Waterman (2000) that virtually nothing is available in the literature on the statistical behavior of word frequencies obtained from a sequence satisfying Hidden Markov Models. These are some of the reasons why in this article we have decided to depend on a detailed exploratory and model-free analysis instead of depending on the assumption of Markov or any other related property for DNA sequences. Our empirical investigation amply demonstrates the usefulness of word frequencies as statistical summaries containing some biologically significant structural information from the sequence data.

Our next result is about how the frequencies of larger words are related to the frequencies of smaller words in a DNA sequence.

Result 4.1. *For $N \gg k > k^* \geq 1$, different k^*-word frequencies can be well approximated by different orthogonal weighted sums of k-word frequencies.*

Proof. For $k \geq 2$, let v be a $(k-1)$-word in $\{A, T, C, G\}^{(k-1)}$, and denote the k-words (v, A), (v, T), (v, C) and (v, G) by w_1, w_2, w_3 and w_4 respectively all of which are elements of $\{A, T, C, G\}^k$. Also, denote the k-words (A, v), (T, v), (C, v) and (G, v) by w_5, w_6, w_7 and w_8 respectively. Then it is straight-forward to verify that the difference

$$g_v - \left(\frac{N-k+1}{N-k+2}\right)(f_{w_1} + f_{w_2} + f_{w_3} + f_{w_4})$$

is equal to either zero or $1/(N-k+2)$, where N is the total length of the DNA sequence, g_v is the frequency of v, and f_{w_1}, f_{w_2}, f_{w_3} and f_{w_4} are the frequencies of w_1, w_2, w_3 and w_4 respectively. Similarly, the difference

$$g_v - \left(\frac{N-k+1}{N-k+2}\right)(f_{w_5} + f_{w_6} + f_{w_7} + f_{w_8})$$

will also be equal to either zero or $1/(N-k+2)$, where f_{w_5}, f_{w_6}, f_{w_7} and f_{w_8} are the frequencies of w_5, w_6, w_7 and w_8 respectively. In other words, we have

$$|g_v - \{(N-k+1)/(N-k+2)\}(f_{w_1} + f_{w_2} + f_{w_3} + f_{w_4})| \leq 1/(N-k+2)$$

and

$$|g_v - \{(N-k+1)/(N-k+2)\}(f_{w_5} + f_{w_6} + f_{w_7} + f_{w_8})| \leq 1/(N-k+2).$$

Since typically $N \gg k$, these immediately imply that the frequency of any $(k-1)$-word is approximately equal to a weighted sum of the k-word frequencies, where the weights are all non-negative (equal to zero or $(N-k+1)/(N-k+2)$). This can be further generalized to the fact that for $1 \leq k^* < k$ and $N >> k$, the frequency of a k^*-word will be approximately equal to a weighted sum of k-word frequencies, where the weights are all non-negative (equal to zero or $(N-k+1)/(N-k^*+1)$), and the error in approximation is not larger than $(k-k^*)/(N-k^*+1)$. Next, we observe that for any two distinct k^*-words v and z, if $g_v \approx \sum_{w \in W_k} \lambda_w f_w$ and $g_z \approx \sum_{w \in W_k} \mu_w f_w$ – the non-negative weights λ_w's and μ_w's can be so chosen that for a k-word w, λ_w and μ_w cannot be simultaneously positive (i.e., if one them is positive, the other one must be zero), and consequently the two weight vectors will be orthogonal to each other in the sense that $\sum_{w \in W_k} \lambda_w \mu_w = 0$. This is a simple consequence of the fact that for any two distinct k^*-words v and z, the two sets of (k^*+1)-words $\{(v, A), (v, T), (v, C), (v, G)\}$ and $\{(z, A), (z, T), (z, C), (z, G)\}$ are completely disjoint. □

Recall at this point the F-ratios and the F-ranks of k-words introduced in Section 2.2. One can nicely generalize the main ideas there as follows. If there are altogether n sequences in the sample that form r biologically homogeneous disjoint clusters $C_1, \ldots, C_r$ (which are known *a priori*), the F-ratio associated with the weighted sums of k-word frequencies based on a given weight vector $(\lambda_w)_{w \in W_k}$ such that $\lambda_w \geq 0$ can be defined as

$$F = \frac{\left\{ \sum_{1 \leq i < j \leq r} \sum_{S \in C_i, S^* \in C_j} \left| \sum_{w \in W_k} \lambda_w (f_w^{(S)} - f_w^{(S^*)}) \right| \right\} \left(\sum_{1 \leq i < j \leq r} n_i n_j \right)^{-1}}{\left\{ \sum_{i=1}^{r} \sum_{S, S^* \in C_i, S \neq S^*} \left| \sum_{w \in W_k} \lambda_w (f_w^{(S)} - f_w^{(S^*)}) \right| \right\} \left(\sum_{i=1}^{r} n_i (n_i - 1) \right)^{-1}}.$$

Here $f_w^{(S)}$ and $f_w^{(S^*)}$ are the frequencies of the k-word w in the sequences S and S^* respectively, and $n_1, \ldots, n_r$ are the numbers of sequences in $C_1, \ldots C_r$ respectively such that $n_1 + \ldots + n_r = n$. Further, given any family of 4^k-dimensional weight vectors, one can sort them according to the F-ratios associated with their corresponding weighted sums of k-word frequencies, and this will yield a definition of F-ranks for these weight vectors. Weights that will lead to larger values of the F-ratio will be biologically more significant. If the *a priori* known biological clusters $C_1, \ldots, C_r$ represent different species groups, such weights may provide valuable

insights into the process of speciation in course of evolution through variation in the distributions of DNA words across different species in the sample. In particular, one may try to construct an appropriate orthogonal family of 4^k-dimensional weight vectors so that the corresponding weighted sums of the k-word frequencies will have appropriately large F-ratios, and cluster analysis based on the l_1-distances computed from those weighted sums will yield biologically meaningful homogenous clusters (i.e., clusters that reflect *a priori* known biological relationships and hierarchies). This leads to a distribution free and data driven procedure for choosing appropriate DNA words and their weights for a given set of DNA sequences, which we hope will be a useful alternative as well as an effective supplement to the approach of choosing the appropriate word size under homogeneous Markov models using likelihood ratio tests as has been done by earlier authors [see Section 2.3 in Reinert, Schbath and Waterman (2000)].

Word frequencies are very naive yet quite natural and useful statistical summaries of DNA sequences. In view of the massive size of DNA sequence data that has been made available these days by automated biotechnology, there is a real need for statistical summarization to gain information from this kind of data. It is fairly transparent from our data analysis and discussion presented in this article that one can conveniently use word frequencies to capture structural patterns present in the DNA sequences of the same organism or organisms forming homogenous biological groups. We have also observed that frequency distributions for DNA words tend to be different in the sequences obtained from different biological groups. All these can make such relatively simple statistical summary measures very useful tools for analysis of large DNA sequences, which is currently one of the most important and interesting problems that lie in the interface between statistics and molecular biology.

Acknowledgements. Authors are thankful to the Distributed Information Centres of the Department of Biotechnology (Government of India) located at University of Puna (Pune), Bose Institute (Calcutta) and Indian Institute of Science (Bangalore) for their help in acquiring some of the DNA sequences analyzed in this paper. Thanks are also due to two anonymous referees, who carefully read earlier versions of the paper. They pointed out some very relevant recent references and made several useful comments.

References

Arratia, R., Goldstein, L., Gordon, L.: Poisson approximation and the Chen-Stein method. Stat. Sci. **5**, 403–434 (1990)

Arratia, R., Martin, D., Reinert, G., Waterman, M.S.: Poisson approximation for long repeats in a random sequence with application to sequencing by hybridization. J. Comput. Biol. **3**, 425–463 (1996)

Blaisdell, B.E., Campbell, A.M., Karlin, S.: Similarities and Dissimilarities of phage genomes. Proc. Natl. Acad. Sci. USA **93**, 5854–5859 (1996)

Chaudhuri, P., Das, S.: Statistical analysis of large DNA sequences using distribution of DNA words. Curr. Sci. **80**, 1161–1166 (2001)

Chaudhuri, P., Das, S.: SWORDS: a statistical tool for analyzing large DNA sequences. J. Biosci. **27**, 1–6 (2002)

Chernoff, H.: The use of faces to represent points in k-dimensional space graphically. J. Amer. Statist. Assoc. **68**, 361–368 (1973)

Churchill, G.A.: Stochastic models for heterogeneous DNA sequences. Bull. Math. Biol. **51**, 79–94 (1989)
Devereux, J.P., Haeberli, P., Smithies, O.: A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**, 387–395 (1984)
Doolittle, R.F.: Molecular evolution: computer analysis of protein and nucleic acid sequences. Meth. Enzymol. **183**, 1–735 (1990)
Doolittle, R.F.: Molecular evolution: computer methods for macromolecular sequence analysis. Meth. Enzymol. **266**, 1–711 (1996)
Everitt, B.S.: *Cluster Analysis*. Edward Arnold: London (1993)
Felsenstein, J.: Statistical inference of phylogenies. J. R. Statist. Soc. (A) **146**, 246–272 (1983)
Felsenstein, J.: Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**, 521–565 (1988)
Felsenstein, J.: PHYLIP – phylogeny inference package (Version 3.4). Cladistics **5**, 164–166 (1989)
Godbole, A.P., Schaffner, A.A.: Improved Poisson approximations for word patterns. Adv. Appl. Prob. **25**, 334–347 (1993)
Geske, M.X, Godbole, A.P., Schaffner, A.A., Skolnick, A.M., Wallstrom, G.L.: Compound Poisson approximations for word patterns under Markovian hypotheses. J. Appl. Prob. **32**, 877–892 (1995)
Karlin, S., Brendel, V.: Patchiness and correlations in DNA sequences. Science **259**(5095), 677–680 (1993)
Karlin, S., Campbell, A.M.: Which bacterium is the ancestor of the animal mitochondrial genome? Proc. Natl. Acad. Sci. USA **91**, 12842–12846 (1994)
Karlin, S., Cardon, L.R.: Computational DNA sequence analysis. Annu. Rev. Microbiol. **44**, 619–654 (1994)
Karlin, S., Ladunga, I.: Comparisons of eukaryotic genomic sequences. Proc. Natl. Acad. Sci. USA **91**, 12832–12836 (1994)
Karlin, S., Ladunga, I., Blaisdell, B.E.: Heterogeneity of genomes: measures and values. Proc. Natl. Acad. Sci. USA **91**, 12837–12841 (1994)
Kimura, M.: *The Neutral Theory of Molecular Evolution*. Cambridge: University Press, Cambridge: 1983
Lange, K.: *Mathematical and Statistical Methods for Genetic Analysis*. New York: Springer Verlag, 1997
Leung, M.-Y., Marsh, G.M., Speed, T.P.: Over- and underrepresentation of short DNA words in herpesvirus genomes. J. Comput. Biol. **3**, 345–360 (1996)
Martindale, C., Konopka, A.K.: Oligonucleotide freqnecies in DNA follow a Yule distribution. Comp. Chem. **20**, 35–38 (1996)
Muri, F.: Modeling bacaterial genomes using Hidden Marokov Models. In: *Compstat'98 Proceedings in Computational Statistics*, R. Payne, P.J. Green (eds), pp. 89–100. Physica-Verlag, Heidelberg. 1998
Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence for two proteins. J. Mol. Biol. **48**, 443–453 (1970)
Nei, M.: Phylogenetic analysis in molecular evolutionary genetics. Annu. Rev. Genet. **30**, 371–403 (1996)
Nussinov, R.: Some rules in the ordering of nucleotides in the DNA. Nucleic. Acids Res. **8**, 4545–4562 (1980)
Nussinov, R.: Nearest neighbor nucleotide patterns: structural and biological implications. J. Biol. Chem. **256**, 8458–8462 (1981)
Nussinov, R.: Some indications for inverse DNA duplication. J. Theor. Biol. **95**, 783–793 (1982)
Nussinov, R.: Doublet frequencies in evolutionary distinct groups. Nucleic. Acids Res. **12**, 1749–1763 (1984a)
Nussinov, R.: Strong doublet preferences in nucleotide sequences and DNA geometry. J. Mol. Evol. **20**, 111–119 (1984b)

Pevzner, P.A.: Nucleotide sequences versus Markov models. Comput. Chem. **16**, 103–106 (1992)

Pevzner, P.A., Borodovsky, M.Y., Mironov, A.A.: Linguistics of nucleotide sequences I: the significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. J. Biomol. Struct. Dyn. **6**(5), 1013–1026 (1989a)

Pevzner, P.A., Borodovsky, M.Y., Mironov, A.A.: Linguistics of nucleotide sequences II: stationary words in genetic texts and the zonal structure of DNA. J. Biomol. Struct. Dyn. **6**(5), 1027–1038 (1989b)

Phillips, G., Arnold, J., Ivarie, R.: Mono- through hexa-nucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic. Acids Res. **15**, 2611–2626 (1987a)

Phillips, G., Arnold, J., Ivarie, R.: The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. Nucleic. Acids Res. **15**, 2627–2638 (1987b)

Prum, B., Rodolphe, F., de Turckheim, E.: Finding words with unexpected frequencies in deoxyribonucleic acid sequences. J. R. Statist. Soc. (B) **57**, 205–220 (1995)

Rahmann, S., Rivals, E.: Exact and efficient computation of the expected number of missing and common words in random texts. In: *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching, Lecure Notes in Computer Science*, **1848**, pp. 375–387, New York: Springer–Verlag, 2000

Reinert, G., Schbath, S.: Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. J. Comput. Biol. **5**, 223–253 (1998)

Reinert, G., Schbath, S.: Large compound Poisson approximations for occurrences of multiple words. In: *Statistics in Molecular Biology and Genetics*, F. Seillier-Moiseiwitsch (ed.), IMS Lecture Notes and Monograph Series, vol. 33, pp. 257–275. IMS, Hayward, California, 1999

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

Schbath, S., Prum, B., de Turckheim, E.: Exceptional motifs in different Markov chain models for statistical analysis of DNA sequences. J. Comput. Biol. **2**, 417–437 (1995)

Strimmer, K., von Haesler, A.: Quartet puzzling: a quartet maximum likelihood method for reconstructing tree toplogies. Mol. Biol. Evol. **13**, 964–969 (1996)

Swofford, D.L.: *PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods), Version 4.0*, Sinauer Associates: Sunderland, 1997

Waterman, M.S.: *Introduction to Computational Biology*, Chapman and Hall: New York, 1995

Weir, B.S.: *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Assoc.: Sunderland, 1996