SWORDS: A statistical tool for analysing large DNA sequences

PROBAL CHAUDHURI* and SANDIP DAS
Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 BT Road, Kolkata 700 108, India
**Corresponding author (Fax, 91-33-5776680; Email, probal@isical.ac.in).*

In this article, we present some simple yet effective statistical techniques for analysing and comparing large DNA sequences. These techniques are based on frequency distributions of DNA words in a large sequence, and have been packaged into a software called SWORDS. Using sequences available in public domain databases housed in the Internet, we demonstrate how SWORDS can be conveniently used by molecular biologists and geneticists to unmask biologically important features hidden in large sequences and assess their statistical significance.

[Chaudhuri P and Das S 2002 SWORDS: A statistical tool for analysing large DNA sequences; *J. Biosci*. (*Suppl. 1*) **27** 1–6]

1. Introduction

Conventional sequence alignment algorithms and techniques (Doolittle 1990, 1996) for estimating similarities and mismatches among DNA sequences, which are quite popular for analysing and comparing relatively smaller sequences, are not feasible to use when it comes to dealing with sequences that have sizes varying between a few thousand base pairs to a few hundred thousand base pairs. Effective analysis of large DNA sequences requires some form of statistical summarization by reducing the size or the dimension of the data to facilitate numerical computations. At the same time, one is interested in capturing some of the fundamental structural information contained in the sequence data as efficiently as possible. SWORDS, which is an acronym derived from "statistical analysis of words in DNA sequences", is a statistical software developed for exploratory analysis of large DNA sequences based on distributions of DNA words. The development of SWORDS was motivated by the simplicity of the DNA word frequency based approach as compared to other computer intensive and operationally complex techniques (e.g. sequence alignment and homology). Further encouragement came from the empirical experience of several scientists, who have observed that DNA word frequencies are simple yet effective statistical tools to capture information about structural patterns, and these can reveal biologically significant features in a DNA sequence (Nussinov 1980, 1981, 1982, 1984a,b; Karlin and Campbell 1994; Karlin and Cardon 1994; Karlin and Ladunga 1994; Karlin *et al* 1994; Blaisdell *et al* 1996; Pan *et al* 1996; S Basu, D P Burma and P Chaudhuri, unpublished results).

Since a DNA sequence is formed using an alphabet of four letters denoting four DNA bases: adenine (*A*), thymine (*T*), cytosine (*C*) and guanine (*G*), the simplest form of statistical summarization is based on various frequencies of DNA *k*-words, which are *k*-tuples formed using these four letters (Nussinov 1980, 1981, 1982, 1984a,b; Karlin and Campbell 1994; Karlin and Cardon 1994; Karlin and Ladunga 1994; Karlin *et al* 1994; Blaisdell *et al* 1996; Pan *et al* 1996; S Basu, D P Burma and P Chaudhuri, unpublished results) (here $k \geq 1$ is a positive integer). For example, if a sequence runs like *ATTCGGCA . . .*, the first 4-word is *ATTC*, the second one is *TTCG*, the third one is *TCGG* and so on. We will treat relative frequencies of these DNA words as statistical summaries of the given DNA sequence, and a comparison between a pair of DNA sequences to judge their similarities and dissimilarities can be carried out by comparing their associated frequency distributions for DNA words with a given size. The entire DNA sequence can be viewed as a description of its complete molecular structure, and occurrence or non-occurrence of specific

Keywords. DNA sequences; statistical tool; SWORDS

words signifies special features of that molecule. Relative abundance and shortage of certain DNA words are likely to have implications on the molecular structure and stability of genomes, and this may have some connections with the cellular processes like recombination, replication, regulation, repair activities etc.

We will demonstrate in this article how SWORDS can enable molecular biologists to detect the structural signature of a genome and thereby identify phylogenetic relationships among different species reflected in the variation of word distributions in their DNA sequences. There has been an explosive growth of databases consisting of very large DNA sequences. Integrated databases are now globally accessible through the Internet, and this rapid advancement in 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.

2. An analysis of mitochondrial genomes of some vertebrates

Phylogenetic relationships among different organisms are of fundamental importance in biology. One of the prime objectives of DNA sequence analysis is phylogeny reconstruction for understanding evolutionary history of different species. Many different methods for phylogenetic analysis of DNA sequence data have been proposed and studied in the biology literature (Felsenstein 1983, 1988; Nei 1996; Strimmer and van Haelsler 1996). On the other hand, the cluster analysis (Everitt 1993) is a well known and widely used statistical technique which is useful in investigating the relative closeness of different organisms based on a given similarity or distance measure. In many ways, cluster analysis is different from phylogenetic analysis though they have some intrinsic similarities. Consequently, a dendrogram tree (Everitt 1993) produced by a cluster analysis is not the same as a phylogenetic tree. SWORDS performs cluster analysis using distance measures that are computed by comparing DNA word frequencies for different sequences corresponding to different organisms.

We will now explore to what extent the similarities and dissimilarities measured by comparing DNA word frequencies of different sequences obtained from different organisms are in agreement with known phylogenetic relationships (Nussinov 1984a; Karlin and Campbell 1994; Karlin and Ladunga 1994; S Basu, D P Burma and P Chaudhuri, unpublished results). For this purpose, we present an analysis of a data set that consists of DNA sequences of complete mitochondrial genomes of fourteen vertebrates. These vertebrates include six mammals (human, gorilla, blue whale, finback whale, kangaroo and

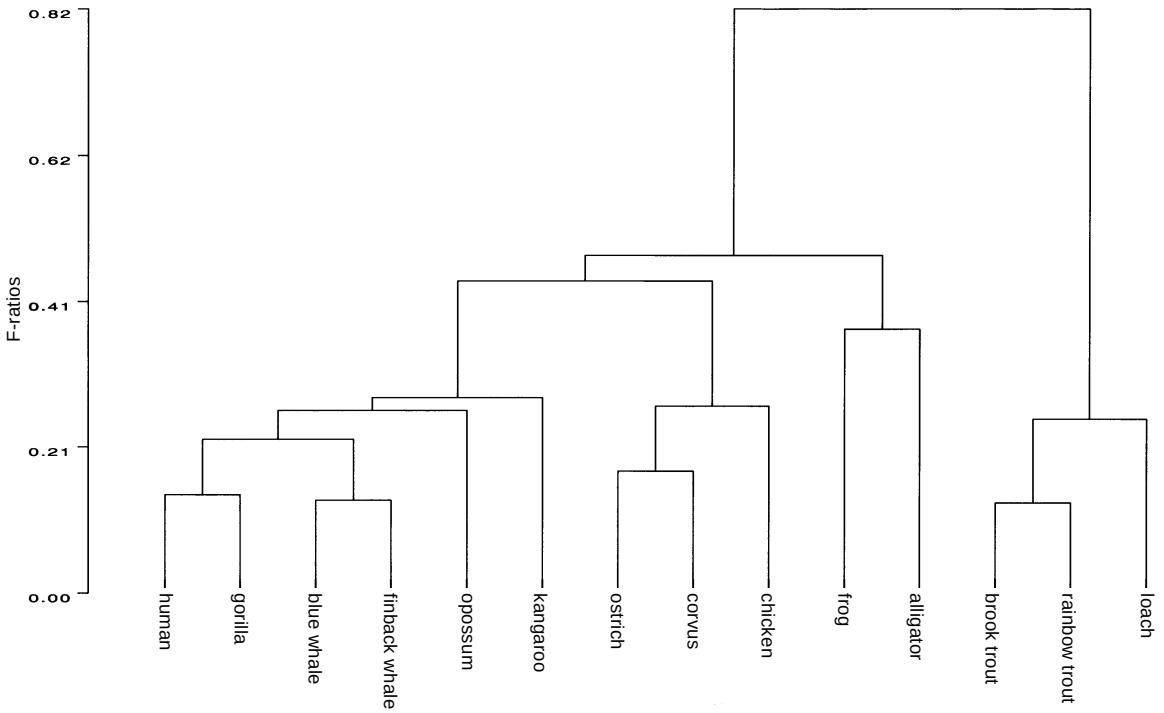

Figure 1. Dendrogram based on 30 tetranucleotides with largest F-ratios.

opossum), three birds (chicken, ostrich and corvus), three fish (rainbow trout, brook trout and loach), one amphibia (frog) and one reptile (alligator). These DNA sequences are available in the public domain GenBank and EMBL databases that are accessible through the Internet. The dendrogram tree presented in figure 1 is produced by SWORDS as a result of average linkage cluster analysis using 30 tetra-nucleotide frequencies. The tree shows a clear clustering of mammals, birds and fish in three distinct clusters. Among different mammals, human and gorilla form a cluster, two whales form another cluster and the two marsupials are clearly separated from other mammals. Among the fish, two trouts form a cluster, and the cluster of three aquatic vertebrates is distinctly separated from terrestrial and amphibian vertebrates. The clustering and the relative positions of the mammals, birds, amphibia and reptiles conform with the standard evolutionary tree derived using fossil records and other molecular as well as morphological evidence.

In figure 2 we have presented star plots (Pan *et al* 1996) produced by SWORDS based on the same frequency data. In each star, there are 30 rays emerging out of the central point. Each of these rays points to a particular direction and corresponds to one specific tetra-nucleotide frequency (i.e. one among those 30 tetra-nucleotide frequencies used in the cluster analysis carried out by SWORDS). These stars have been drawn in such a way that the length of each ray in a star is proportional to the value of the particular tetra-nucleotide frequency to which it corresponds. Visual comparison of the shapes of the stars clearly reveal similarities of different animals within each of the homogenous biological groups of mammals, birds and fish. Also noteworthy is the striking dissimilarity between the shapes of the stars corresponding to aquatic vertebrates and the shapes of those corresponding to terrestrial ones.

Among 256 possible tetra-nucleotides, 30 tetranucleotides used in this analysis have been chosen in such a way that they have the highest between to within group variation ratios. These ratios were calculated using the mammals, the birds and the fish as three distinct biological groups. We call them F-ratios following the terminology used in the standard statistical decomposition and analysis of variation, when there are different groups present in the data. Figure 3 shows a graphical display of these tetra-nucleotides together with their F-ratios. Very high F-ratios for *GGGA* and *ATCA* compared to other tetra-nucleotides are quite striking in this diagram. It implies that the frequencies for these two 4-words vary a lot across different homogenous biological groups, and this might have some evolutionary implications. It will be appropriate to note here that di-nucleotide and tri-nucleotide frequencies could not adequately discriminate among mammals, birds and fish when we tried similar analysis with 2-words and 3-words using SWORDS. Four-words are the words of the smallest size that could produce biologically meaningful clusters.

3. Another look at "the coelacanth vs lungfish controversy"

The origin of terrestrial vertebrates from the aquatic ones in course of evolution during the Devonian period (approximately 350–400 million years ago) involved complex morphological and physiological changes, and is considered to be one of the most significant events in the evolutionary history of vertebrates. It is generally

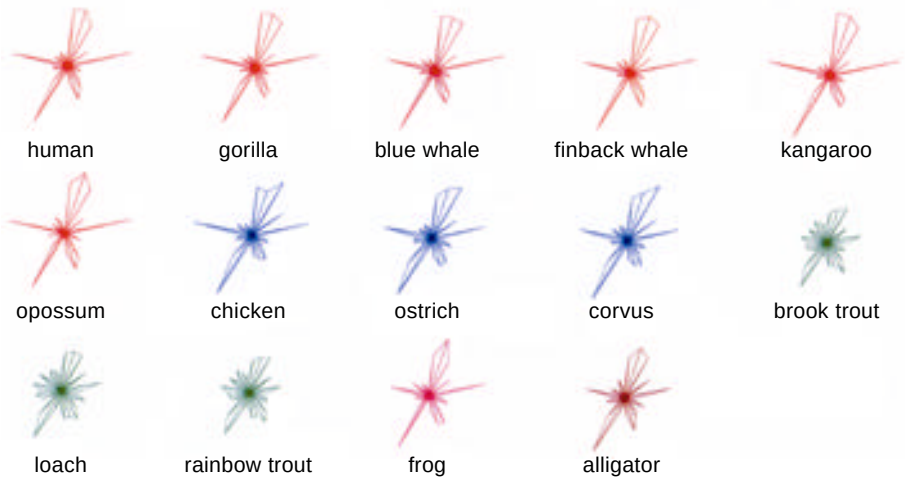

Figure 2. Star plots for 30 tetranucleotide frequencies with largest F-ratios.

believed that the coelacanth together with different types of lungfish and the extinct rhipidistians formed the class of *lobe-finned fish* (Sarcopterygii) from which the tetrapods originated, and different varieties of *ray-finned fish* (Actinopterygii) are only distantly related to tetrapods (Zardoya and Meyer 1996a,b, 1997). Geneticists and molecular biologists have done extensive research based on the mitochondrial DNA of the lungfish, the coelacanth and various other vertebrates such as mammals, birds, fish and amphibia in their attempt to determine the relative phylogenetic positions of the coelacanth and the lungfish in the evolutionary tree of vertebrates (Zardoya and Meyer 1996a,b, 1997). Their work suggests that the mitochondrial genome of the lungfish is closer to that of the coelacanth than to the mitochondrial genomes of land living and amphibian animals. However, these findings are not free from ambiguity and are based on subjective choice of parts of mitochondrial DNA data as well as the method of comparison. It is known that different methods applied to different parts of the mitochondrial DNA data can present different phylogenetic relationships among the coelacanth, the lungfish and other aquatic and terrestrial vertebrates. As the sizes of vertebrate mitochondrial DNA sequences vary from 15000 bases to 18000 bases, direct application of conventional alignment and sequence matching techniques to the entire mitochondrial genome is not computationally feasible. One is forced to work with parts of the genome separately in order to use those methods for comparing sequences.

We have analysed DNA sequences for the complete mitochondrial genome of the coelacanth, the lungfish and the fourteen vertebrates studied in the preceding section using SWORDS. The dendrogram tree presented in figure 4 is the result of the average linkage cluster analysis using the same 30 tetra-nucleotide frequencies that were used in the analysis presented in the preceding section. The tree shows a clear clustering of coelacanth with frog, alligator and other terrestrial vertebrates, while lungfish clusters itself with the three different ray-finned fish forming a separate cluster of aquatic vertebrates.

An important question that arises at this point is how strong is the statistical evidence in favour of a *coelacanth–tetrapod link* and a *lungfish–fish link* that is present in the dendrogram produced by SWORDS. In other words, one would like to get an idea about the statistical significance of such a link in a dendrogram tree. One way to address this question is to consider segments or parts of the complete sequence for each of the mitochondrial genomes and then compare them with the DNA word frequencies for only those segments instead of complete sequences. The dendrogram tree can be

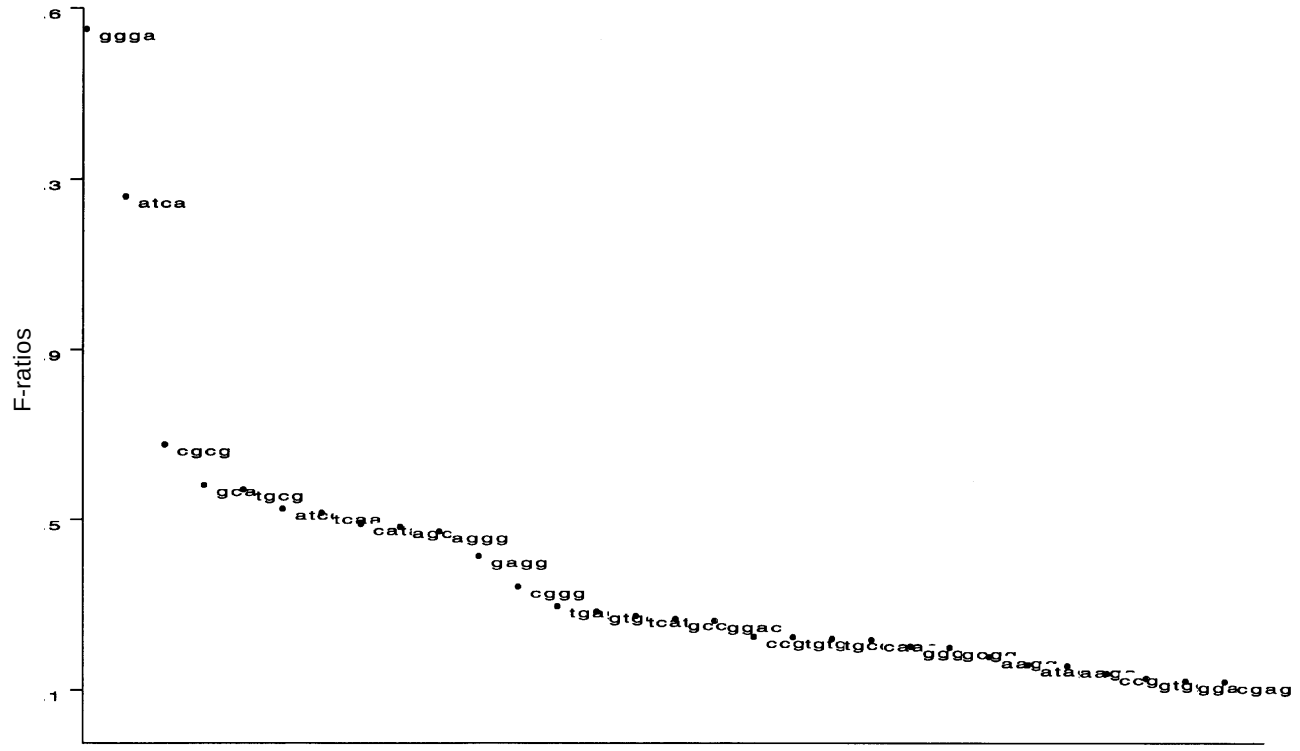

Figure 3. F-ratios for 30 tetranucleotides.

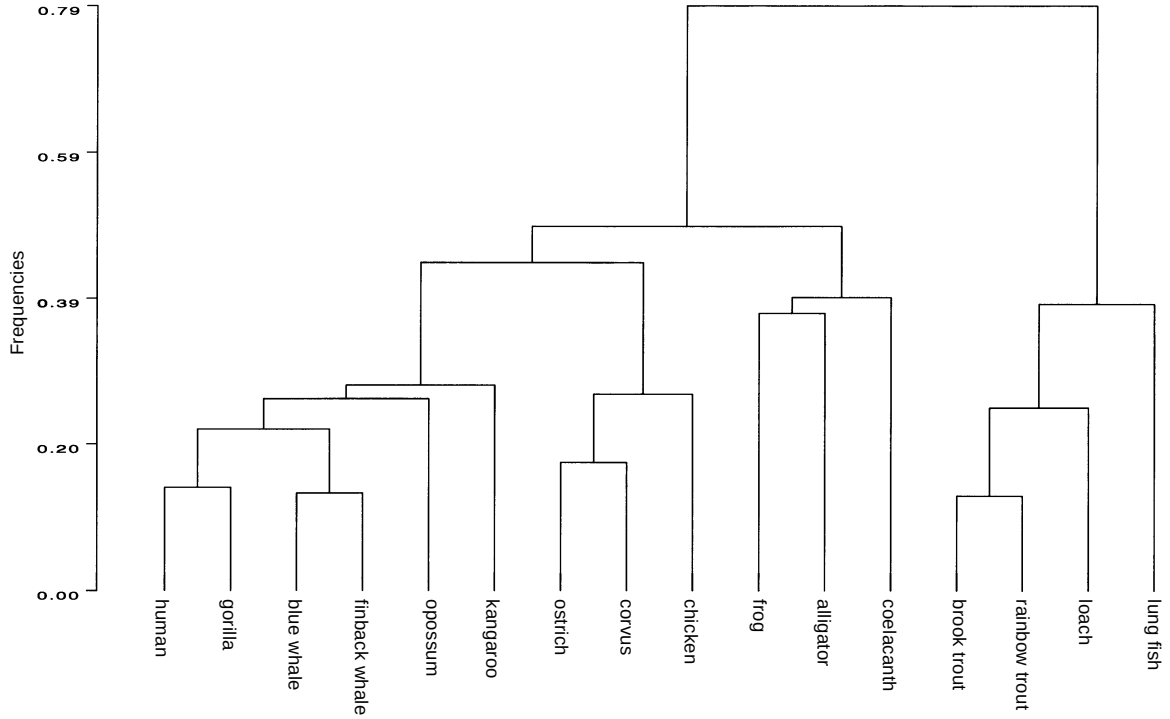

Figure 4. Dendrogram based on 30 tetranucleotides frequencies.

recomputed based on distances obtained from the word frequencies for randomly selected segments from complete genomes. The topologies of different dendrograms obtained from repeated random sampling of segments can give an idea about the amount of statistical variation present in the similarities and dissimilarities among different segments of these DNA sequences. The basic idea here is closely related to the bootstrap and the jack-knife techniques (Felsenstein 1985; Zharkikh and Li 1992a,b, 1995; Felsenstein and Kishino 1993; Hills and Bull 1993) used in statistical validation of phylogenetic trees constructed using aligned DNA sequences, where one randomly samples different aligned sites. SWORDS can perform repeated random sampling of such segments of mitochondrial genomes in order to measure the statistical significance of the *coelacanth–tetrapod link* as well as that of the *lungfish–fish link* in the dendrogram constructed using DNA word frequencies. With the current set of sixteen vertebrates and 30 tetranucleotides mentioned above, these two links were observed to occur 95%–98% of the times, when SWORDS was made to repeatedly compute the dendrogram tree by comparing the word frequency distributions of randomly selected segments of complete mitochondrial genomes. The percentages reported here are based on several thousand random trials, and they vary slightly depending on the sizes of the selected segments.

4. Concluding remarks

Earlier authors have investigated the possibility of modelling the frequency distribution of oligonucleotides in DNA sequences using certain specific probability laws (Pevzner *et al* 1989a,b; Pevzner 1992; Schbath *et al* 1995; Martindale and Konopka 1996; Reinert and Schbath 1999). Unusual frequencies of certain DNA words in *Escherichia coli* and virus genomes and possible statistical and biological implications of such over- and under-representation of those words have been studied in the literature based on Markov chain models for DNA sequences (Phillips *et al* 1987a,b; Prum *et al* 1995; Leung *et al* 1996). However, it is also well known that frequently standard and popular probability models do not fit observed DNA sequence data very well, and this is one of the reasons why SWORDS has been developed as *an exploratory and model-free data analytic tool* instead of depending on specific probability laws or models for DNA sequences.

Word frequencies are very convenient and natural statistical summaries for large DNA sequences. In view of the ever growing sizes of DNA data bases as well as the sizes of the sequences that are available in these data bases – we feel that SWORDS will be a useful statistical tool for molecular biologists and geneticists. We have tried to demonstrate in this article how SWORDS can be

used for exploratory analysis of DNA data for discovering biologically significant features hidden in large sequences using simple yet effective statistical techniques. This free software can run on WINDOWS 95/98/NT as well as LINUX platforms, and it is available with us (probal@isical.ac.in; http://www.isical.ac.in/~probal).

Acknowledgements

We thank Professor Debi Prosad Burma for introducing us to this topic and his helpful discussions. Thanks are also due to the referee who contributed a number of helpful comments and suggestions.

References

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

Doolittle R F 1990 Molecular evolution: computer analysis of protein and nucleic acid sequences; *Methods Enzymol.* **183** 1–735

Doolittle R F 1996 Molecular evolution: computer methods for macromolecular sequence analysis; *Methods Enzymol.* **266** 1–711

Everitt B S 1993 *Cluster Analysis* (London: Edward Arnold)

Felsenstein J 1983 Statistical inference of phylogenies (with Discussion); *J. R. Stat. Soc.* (*Ser. A*) **146** 246–272

Felsenstein J 1985 Confidence limits on phylogenies: an approach using the bootstrap; *Evolution* **39** 783–791

Felsenstein J 1988 Phylogenies from molecular sequences: inference and reliability; *Annu. Rev. Genet.* **22** 521–565

Felsenstein J and Kishino H 1993 Is there something wrong with the bootstrap? A reply to Hillis and Bull; *Syst. Biol.* **42** 193–200

Hillis D M and Bull J J 1993 An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis; *Syst. Biol.* **42** 182–192

Karlin S and Campbell A M 1994 Which bacterium is the ancestor of the animal mitochondrial genome?; *Proc. Natl. Acad. Sci. USA* **91** 12842–12846

Karlin S and Cardon L R 1994 Computational DNA sequence analysis; *Annu. Rev. Microbiol.* **44** 619–654

Karlin S and Ladunga I 1994 Comparisons of eukaryotic genomic sequences; *Proc. Natl. Acad. Sci. USA* **91** 12832–12836

Karlin S, Ladunga I and Blaisdell B E 1994 Heterogeneity of genomes: measures and values; *Proc. Natl. Acad. Sci. USA* **91** 12837–12841

Leung M-Y, Marsh G M and Speed T P 1996 Over- and under-representation of short DNA words in herpesvirus genomes; *J. Comput. Biol.* **3** 345–360

Martindale C and Konopka A K 1996 Oligonucleotide frequencies in DNA follow a Yule distribution; *Comput. Chem.* **20** 35–38

Nei M 1996 Phylogenetic analysis in molecular evolutionary genetics; *Annu. Rev. Genet.* **30** 371–403

Nussinov R 1980 Some rules in the ordering of nucleotides in the DNA; *Nucleic Acids Res.* **8** 4545–4562

Nussinov R 1981 Nearest neighbor nucleotide patterns: structural and biological implications; *J. Biol. Chem.* **256** 8458–8462

Nussinov R 1982 Some indications for inverse DNA duplication; *J. Theor. Biol.* **95** 783–793

Nussinov R 1984a Doublet frequencies in evolutionary distinct groups; *Nucleic Acids Res.* **12** 1749–1763

Nussinov R 1984b Strong doublet preferences in nucleotide sequences and DNA geometry; *J. Mol. Evol.* **20** 111–119

Pan A, Basu S, Dutta C, Burma D P and Mukherjee R 1996 Nucleotide frequency map: a new technique for pictorial representation of dinucleotide frequencies; *Curr. Sci.* **71** 50–53

Pevzner P A 1992 Nucleotide sequences versus Markov models; *Comput. Chem.* **16** 103–106

Pevzner P A, Borodovsky M Y and Mironov A A 1989a 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** 1013–1026

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

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

Phillips G, Arnold J and Ivarie R 1987b 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

Prum B, Rodolphe F and de Turckheim E 1995 Finding words with unexpected frequencies in deoxyribonucleic acid sequences; *J. R. Statist. Soc.* **B57** 205–220

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

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

Strimmer K and van Haelsler A 1996 Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies; *Mol. Biol. Evol.* **13** 964–969

Zardoya R and Meyer A 1996a Evolutionary relationships of the coelacanth lungfishes and tetrapods based on the 28S ribosomal RNA sequences; *Proc. Natl. Acad. Sci. USA* **93** 5449–5454

Zardoya R and Meyer A 1996b The complete nucleotide sequence of the mitochondrial genome of the lungfish (*Protopterus dolloi*) supports its phylogenetic position as a close relative of land vertebrates; *Genetics* **142** 1249–1263

Zardoya R and Meyer A 1997 The complete DNA sequence of the mitochondrial genome of a "living fossil" the coelacanth (*Latimeria chalumnae*); *Genetics* **146** 995–1010

Zharkikh A and Li W H 1992a Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock; *Mol. Biol. Evol.* **9** 1119–1147

Zharkikh A and Li W H 1992b Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock; *J. Mol. Evol.* **35** 356–366

Zharkikh A and Li W H 1995 Estimation of confidence in phylogeny: the complete and partial bootstrap technique; *Mol. Phylogenet. Evol.* **4** 44–63