Statistical analysis of large DNA sequences using distribution of DNA words

Probal Chaudhuri* and Sandip Das

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 035, India

Conventional sequence alignment techniques for comparing and analysing relatively smaller DNA sequences of nearly equal sizes are not applicable to data consisting of large sequences with widely varying sizes. In this article DNA sequences have been analysed based on distributions of DNA words. DNA word frequencies are simple yet effective statistical tools to capture information about structural patterns, and they can reveal biologically significant features in DNA sequence. Our analysis demonstrates how such simple statistical summaries of large DNA data can enable us to detect the structural signature of a genome as well as to identify phylogenetic relationships among different species reflected in the variation of word distributions in their DNA sequences.

THANKS to the efforts made by scientists located at laboratories spread all over the world, there has been an explosive growth of databases consisting of DNA sequences. Integrated databases are now globally accessible through the Internet. 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 of fruitful scientific results. Conventional sequence alignment algorithms and related techniques[1,2] for determining similarities and dissimilarities among DNA sequences, which are used for relatively smaller sequences are not feasible to use when it comes to dealing with sequences with 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 and at the same time to capture some of the fundamental structural information contained in the sequence data as efficiently as possible.

A DNA sequence is formed using an alphabet of four letters {A, T, C, G} denoting four DNA bases: adenine, thymine, cytosine and guanine, respectively. 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[3–16]. 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 $\#(W_k) = 4^k$, 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 . . . , the first 4-word is ATTC, the second one is TTCG, 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 $\Sigma_{w\in W_k} f_w = 1$. A comparison between a pair of DNA sequences to judge their similarities and dissimilarities can be carried out by comparing their associated frequency vectors $(f_w)_{w\in W_k}$ and $(g_w)_{w\in W_k}$ (say), which is equivalent to comparing the values of f_w and g_w for each $w \in W_k$ in some appropriate way.

Relative abundance and shortage of certain DNA words are likely to have implications on the molecular structures and stability of genomes and this may have some connections with the cellular processes like recombination, replication, regulation, repair activities, etc. An equally important issue is to what extent the relatedness and similarities measured by comparing DNA word frequencies of different sequences are in agreement with known phylogenetic relationships[3,5,7,12,15,16]. Given the simplicity of the word frequency-based approach compared to other computer-intensive and operationally complex techniques (e.g. sequence alignment and homology), and the biologically meaningful and significant results that we have observed it to yield, we have undertaken a statistical study of several genomic sequences based of DNA word frequencies.

Analysis of the complete genome of roundworm (*Caenorhabiditis elegans*)

All six chromosomes of *Caenorhabiditis elegans* have recently been completely sequenced and the data are now available from the Internet. There are five autosomes and one sex chromosome (X-chromosome), and the smallest chromosome (i.e. the 3rd autosome) consists of about 13.26 million base pairs, while the largest chromosome (i.e. the 5th autosome) consists of about 21.6 million base pairs. Clearly, it is not possible to compare and judge the structural similarities of such large sequences with so

*For correspondence. (e-mail: probal@isical.ac.in) much variations in their sizes using any of the standard sequence alignment and homology techniques. Besides, the biological functions of many parts of the genome are not yet well understood and certain parts of the genome might be biologically non-functional. This makes total comparison of these six chromosomes through functional homology virtually impossible. However, it is possible to compare the frequencies of various DNA words in these chromosomes, and that can lead to valuable insights into the extent of their structural similarities.

In Figure 1, we have plotted the frequencies for those fifteen 6-words (i.e. hexanucleotides) that have the largest variation of their frequencies among different chromosomes. We have plotted the frequencies corresponding to different chromosomes using different colours, where the first five are the autosomes and the sixth one is the *X*-chromosome. Frequencies are given along the vertical axis of the graph, while the hexanucleotides are marked along the horizontal axis. Points in the graph have been joined by line segments in order to produce continuous graphs to facilitate visual comparison. In Figure 2, we have presented the dendrogram tree based on average linkage cluster analysis using standard Euclidean distances of these 15-dimensional frequency vectors for the six chromosomes. While there are some striking similarities in the word frequencies for all of these chromosomes reflected in the ups and downs of their frequency plots, it is clearly indicated by the dendrogram that the *X*-chromosome has a very different frequency distribution of hexanucleotides compared to other five autosomes.

This example nicely demonstrates how DNA word frequencies can capture structural signatures[15] of these six chromosomes and effectively differentiate the sex chromosome from the five autosomes. It will be appropriate to note here that the frequencies of DNA words consisting of five or fewer letters cannot differentiate these chromosomes very well, and significant differentiation was observed only when hexanucleotide frequencies or higher oligonucleotide frequencies were used.

Analysis of 16S and 18S ribosomal RNA sequences

Phylogenetic relationships among different organisms are of fundamental importance in biology, and one of the prime objectives of DNA sequence analysis is phylogeny

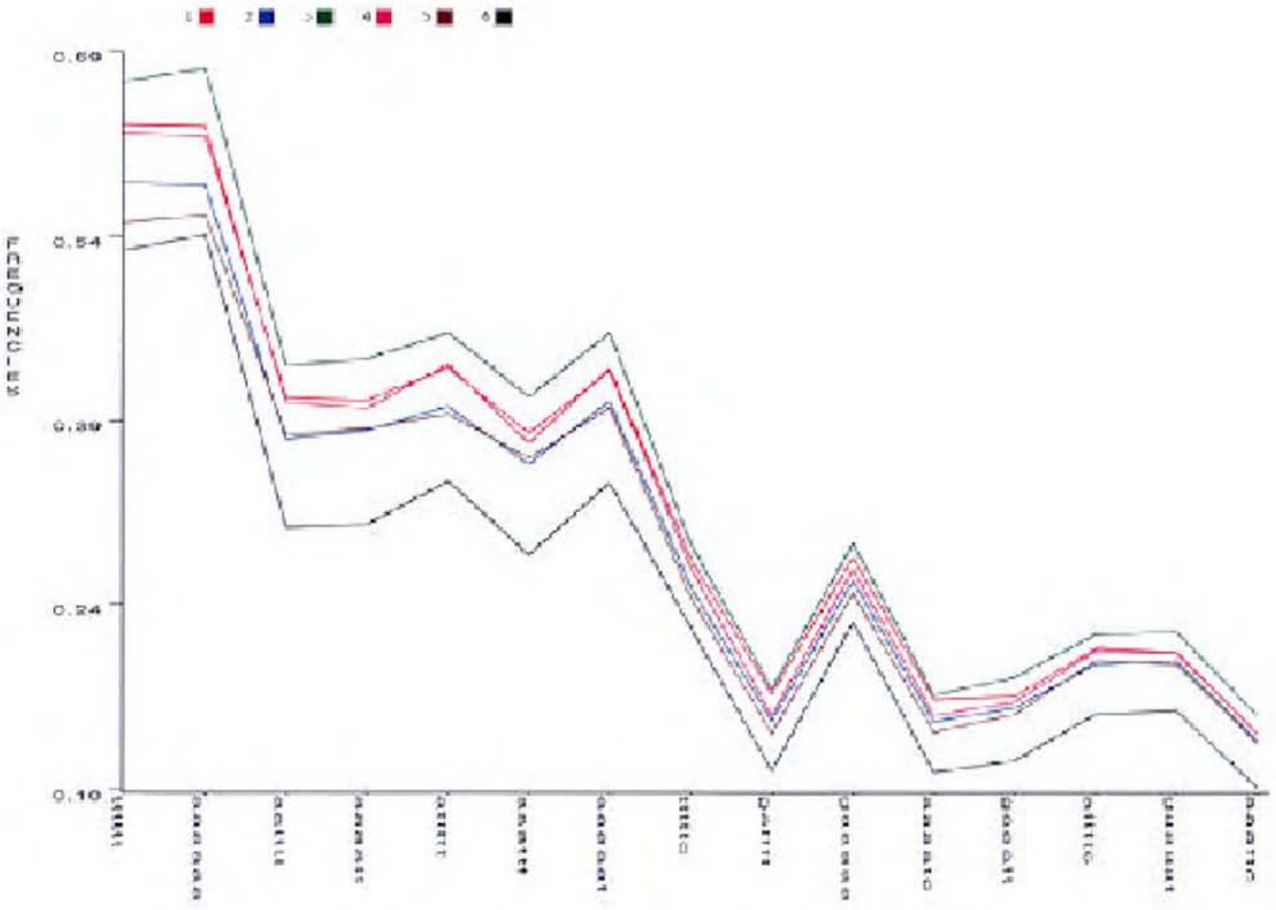

Figure 1. Frequency plots for 15 hexanucleotides with maximum variations.

reconstruction for understanding evolutionary history of organisms. Many different methods for phylogenetic analysis of DNA sequence data have been proposed and studied in the literature[17–19]. We will now present an analysis of a data set that consists of 16S and 18S ribosomal RNA sequences for twenty-four organisms. These sequences have their sizes varying between 1430 and 1831 nucleotides. The organisms chosen in this case are six bacteria, six archaea, six fungi and six gymnosperm plants and the source of our data is EMBL Data Bank. The result of average linkage cluster analysis based on the frequencies of thirty-five 4-words (tetranucleotides) that have largest between-to-within group variation ratios, where four obvious biological groups are formed by these organisms, is presented in Figure 3. The objective here is to find out to what extent these tetranucleotide frequencies can capture genomic signatures and distinguish organisms that belong to different biological groups[15]. In Figure 3 the bacteria, the archaea, the fungi and the plants form distinct clusters and the eukaryotes clearly separate out from the prokaryotes. It is interesting however that such biologically meaningful clusters could not be formed when frequencies of DNA words consisting of three or fewer letters were considered.

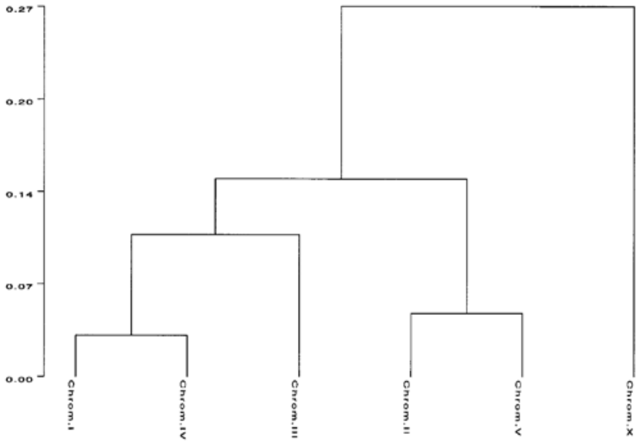

Figure 2. Dendrogram tree based on 15 hexanucleotides.

Coelacanth vs lungfish debate: An analysis of mitochondrial DNA sequences

The transition of life from water to land, leading towards the development of land vertebrates during the Devonian period (approximately 350–400 m.y. ago) is one of the most significant events in the evolutionary history of vertebrates. The origin of terrestrial vertebrates from the aquatic ones involved complex morphological changes as well as physiological innovations and the scarcity of fossil records has created a great deal of controversies among palaeontologists, comparative morphologists and evolutionary biologists for several decades. However, it is

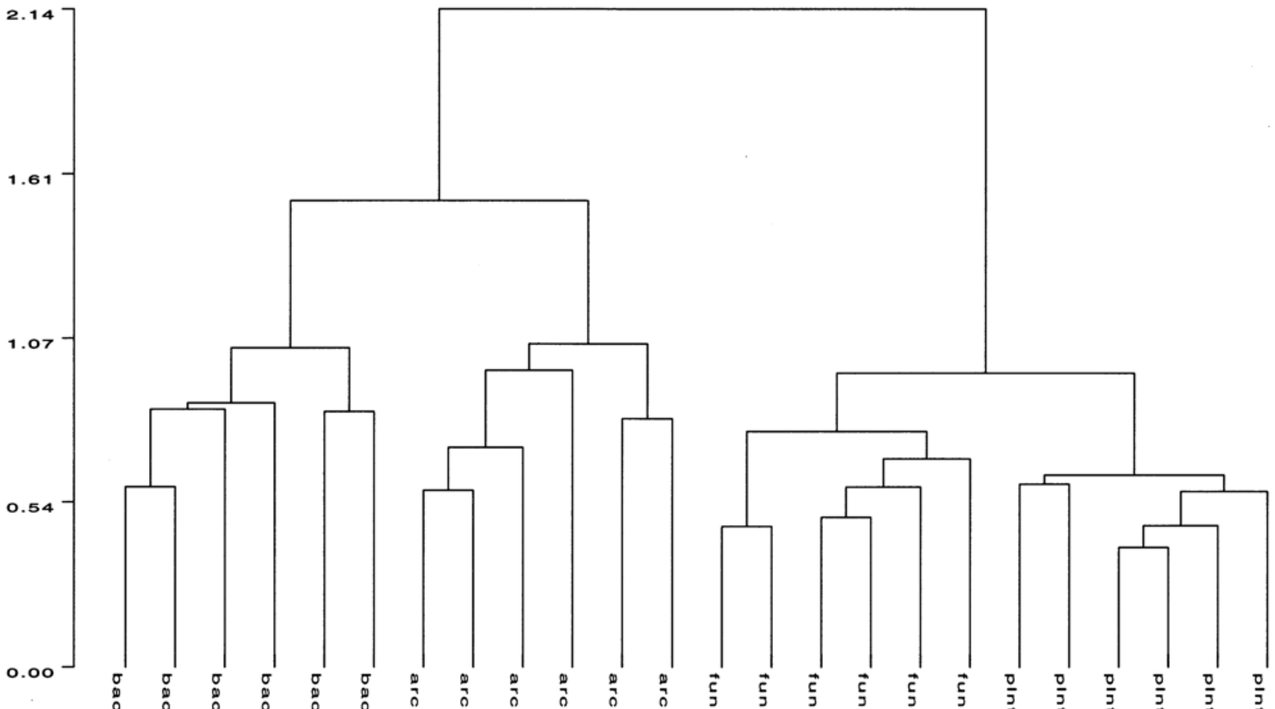

Figure 3. Dendrogram tree based on 35 tetranucleotides.

generally agreed upon 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[20–22]. In the 19th and early 20th century, coelacanths were known only from fossil records, and they vanished from the fossil records approximately 65–70 m.y. ago. As a result they were believed to be history and palaeontologists have credited them as the ancestor of all tetrapods, and it is still one of the predominant textbook dogmas. The first living coelacanth was caught off the Comoro Islands near the east coast of South Africa at the mouth of Chalumna River a few days before Christmas in 1938. Sixty years later in July 1998, another coelacanth population was discovered by American and Indonesian scientists off Sulawesi, Indonesia. This second location is about 10,000 km east of the first location where the fish was found alive.

The other member of the class of lobe-finned fish is the lungfish. In the lower Devonian period, there were a variety of different species of lungfish, which lived in both marine and freshwater environment. However, there are only a very small number of species that are alive today and these are the Australian lungfish, the South American lungfish and the African lungfish. They are obligate air-breathers with external nasal openings that are important for any animal that needs to breathe and chew at the same time.

An extensive amount of research work has been carried out and published in recent issues of leading science journals by geneticists based on the mitochondrial DNA of the lungfish, the coelacanth and various other vertebrates such as mammals, birds, fish and amphibians to determine the relative phylogenetic positions of the coelacanth and the lungfish in the evolutionary tree of vertebrates[20–22]. Their analysis seems to suggest some statistical evidence for the mitochondrial genome of the lungfish being closer than that of the coelacanth to mitochondrial genome of land-living animals and amphibians. However, their findings are not unambiguous, have a lot of subjective elements in terms of the choice of the parts of mitochondrial DNA data as well as the method of comparison, and different methods applied to different parts of the mitochondrial DNA data pointed towards different possibilities. Since the sizes of vertebrate mitochondrial DNA sequences vary from 15,000 to 18,000 bases, conventional alignment and sequence matching techniques are not directly applicable to the entire mitochondrial genome and one has to work with parts of it separately.

We have analysed DNA sequences for complete genomes of coelacanth, lungfish, frog, alligator, six mammals (human, gorilla, blue whale, finback whale, kangaroo

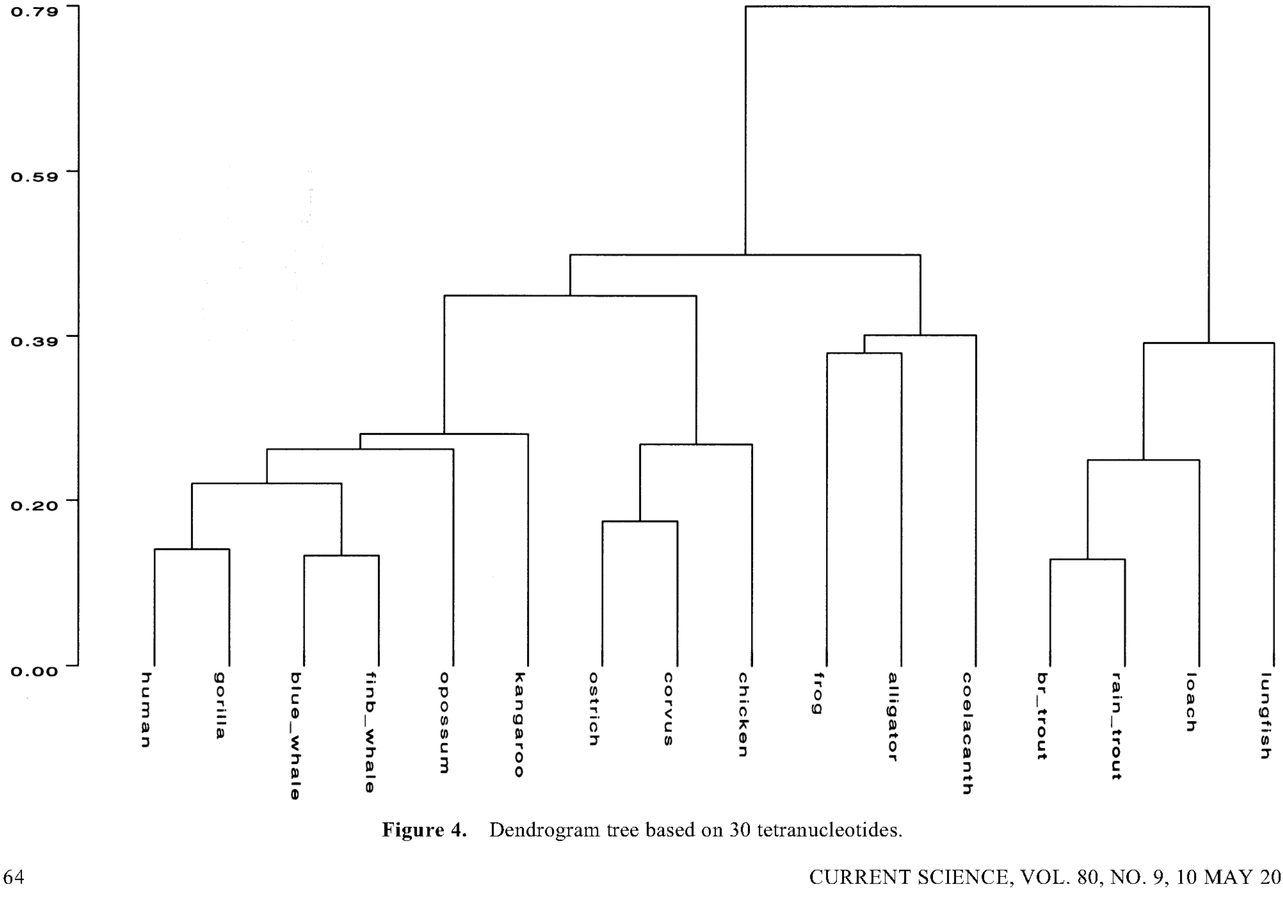

Figure 4. Dendrogram tree based on 30 tetranucleotides.

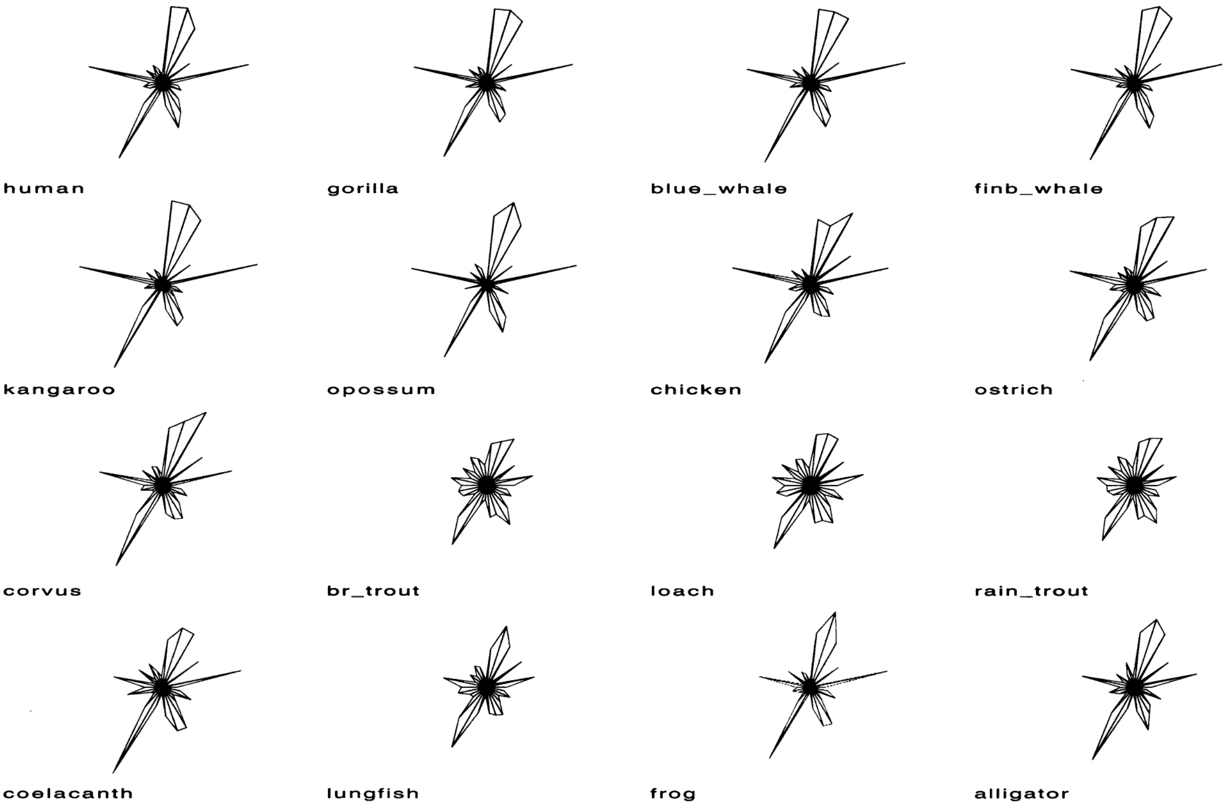

Figure 5. Star plots for 30 tetranucleotide frequencies.

and opossum), three birds (chicken, ostrich and corvus) and three ray-finned fish (rainbow trout, brook trout and loach). These DNA sequences are available in the public domain Genbank databases that are accessible through the Internet. The dendrogram tree presented in Figure 4 is the result of average linkage cluster analysis using 30 tetranucleotide frequencies. There is a clear clustering of all the mammals, the birds and the fish in this dendrogram tree and these biological groups as well as the reptiles and the amphibians appear in this tree in an order which is in agreement with the present knowledge of vertebrate evolution. The tree also shows the clustering of coelacanth with frog, alligator and other terrestrial vertebrates, while lungfish clusters itself with three different ray-finned fish, forming a separate cluster of aquatic vertebrates. Thirty tetranucleotides used here have been chosen in such a way that they have the highest between-to-within group variation ratios among 256 possible tetranucleotides. These ratios were calculated using the mammals, the birds and the ray-finned fish as three distinct biological groups. Star plots for those 30 tetranucleotide frequencies for different sequences presented in Figure 5 also indicate relative closeness of coelacanth to land-living vertebrates and relative closeness of lungfish to different varieties of ray-finned fish. However, one should be cautious while interpreting these statistical results and more detailed and careful analysis of oligonucleotide frequencies is necessary before one can confidently assert the relative positions of the coelacanth and the lungfish in the vertebrate evolutionary tree. In particular, it was remarked by the referee that high evolutionary rates of the mitochondrial genomes of some of the vertebrates may sometimes produce lots of unique derived characters in their genomic sequences, and it is possible that this may lead to enhanced similarities among different species and bias results obtained from oligonucleotide frequencies. We intend to pursue these issues and a more extensive analysis of DNA word frequencies in a future paper.

Concluding remarks

Word frequencies are very convenient and natural statistical summaries for large DNA sequences. In view of the massive sizes of DNA sequences that are available these days, thanks to automated biotechnology, there is a genuine need for effective summarization of such data before any meaningful scientific analysis is carried out. Our analysis has amply demonstrated that distribution of DNA words can capture biologically significant structural patterns in large DNA sequences, and it can be used as a tool for phylogenetic analysis. We have developed a software called SWORDS[16] (an acronym for Statistical analysis of WORDS in DNA Sequences) that is capable of extensive exploratory statistical analysis of large DNA sequences using distributions of DNA words. This software can run on WINDOWS 95/98/NT platforms and is available for free from us (probal@isical.ac.in).

1. Doolittle, R. F., *Methods Enzymol.*, 1990, **183**, 1–735.
2. Doolittle, R. F., *Methods Enzymol.*, 1996, **266**, 1–711.
3. Basu, S., Burma, D. P. and Chaudhuri, P. , 2000 (unpublished manuscript submitted for publication).
4. Blaisdell, B. E., Campbell, A. M. and Karlin, S., *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 5854–5859.
5. Karlin, S. and Campbell, A. M., *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 12842–12846.
6. Karlin, S. and Cardon, L. R., *Annu. Rev. Microbiol.*, 1994, **44**, 619–654.
7. Karlin, S. and Ladunga, I., *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 12832–12836.
8. Karlin, S., Ladunga, I. and Blaisdell, B. E., *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 12837–12841.
9. Nussinov, R., *Nucleic Acids Res.*, 1980, **8**, 4545–4562.
10. Nussinov, R., *J. Biol. Chem.*, 1981, **256**, 8458–8462.
11. Nussinov, R., *J. Theor. Biol.*, 1982, **95**, 783–793.
12. Nussinov, R., *Nucleic Acids Res.*, 1984, **12**, 1749–1763.
13. Nussinov, R., *J. Mol. Evol.*, 1984, **20**, 111–119.
14. Pan, A., Basu, S., Dutta, C., Burma, D. P. and Mukherjee, R., *Curr. Sci.*, 1996, **71**, 50–53.
15. Deschavanne, P. J., Giron, A., Vilain, J., Fagot, G. and Fertil, B., *Mol. Biol. Evol.*, 1999, **16**, 1391–1399.
16. Chaudhuri, P. and Das, S., *J. Biosci.*, 2001 (to appear).
17. Felsenstein, J., *J. R. Stat. Soc. Ser. A*, 1983, **146**, 246–272.
18. Felsenstein, J., *Annu. Rev. Genet.*, 1988, **22**, 521–565.
19. Nei, M., *Annu. Rev. Genet.*, 1996, **30**, 371–403.
20. Zardoya, R. and Meyer, A., *Proc. Natl. Acad. Sci. USA*, 1996, **93**, 5449–5454.
21. Zardoya, R. and Meyer, A., *Genetics*, 1996, **142**, 1249–1263.
22. Zardoya, R. and Meyer, A., *Genetics*, 1997, **146**, 995–1010.

ACKNOWLEDGEMENTS. We thank Prof. Devi Prosad Burma for introducing us to this topic and many helpful discussions. We also thank the referee who read an earlier version of the paper with great interest and provided several helpful comments and suggestions.