CABIOS

Vol.8. no.1. 1992
Pages 15 – 22

Sequence ordinations: a multivariate analysis approach to analysing large sequence data sets

Desmond G. Higgins

Abstract

Ordination is a powerful method for analysing complex data sets but has been largely ignored in sequence analysis. This paper shows how to use principal coordinates analysis to find low-dimensional representations of distance matrices derived from aligned sets of sequences. The method takes a matrix of Euclidean distances between all pairs of sequence and finds a coordinate space where the distances are exactly preserved. The main problem is to find a measure of distance between aligned sequences that is Euclidean. The simplest distance function is the square root of the percentage difference (as measured by identities) between two sequences, where one ignores any positions in the alignment where there is a gap in any sequence. If one does not ignore positions with a gap, the distances cannot be guaranteed to be Euclidean but the deleterious effects are trivial. Two examples of using the method are shown. A set of 226 aligned globins were analysed and the resulting ordination very successfully represents the known patterns of relationship between the sequences. In the other example, a set of 610 aligned 5S rRNA sequences were analysed. Sequence ordinations complement phylogenetic analyses. They should not be viewed as a complete alternative.

Introduction

In this paper, I describe how to use principal coordinates analysis (Gower, 1966) to examine the pattern of relationships within large data sets of homologous nucleic acid or protein sequences. Numerous methods have been described for performing cluster analysis on groups of sequences. However, these methods all assume an underlying hierarchical structure in the data, this structure being the result of the evolutionary process. In some cases, a hierarchical structure will be difficult to fit. This can be due, for example, to convergent evolution, greatly unequal rates of substitution or simply due to lack of data (short sequences). Further, many phylogenetic methods are computationally extremely expensive for more than a hundred or so sequences. The method that is described here makes few assumptions about the structure of the data and is a very useful, general-purpose technique, even in cases where a hierarchy is easy to fit to a data set. It is also cheap to compute even for huge data sets.

Multivariate analysis is that branch of data analysis which deals with multiple measurements made on one or more samples (Cooley and Lohnes, 1971). One can use multivariate analysis techniques to test hypotheses, but they are most widely used in data exploration or hypothesis generation. The most familiar type of multivariate analysis is ordination, of which principal components analysis or PCA (Hotelling, 1933) is the most widely used. Ordination is used to take a set of objects, initially arranged in a high-dimensional space (defined by the measurements made on the objects) and represent these in a small number (usually two or three) of dimensions, while preserving the inter-object distances as much as possible.

Ordination has been used extremely widely in biology, e.g. in morphometrics (Blackith and Reyment, 1971), taxonomy (Sneath and Sokal, 1973), vegetation analysis (Digby and Kempton, 1987) or psychology (Shepard, 1981) where such methods are generally known as multidimensional scaling techniques. As far as the author is aware, no attempt has been made to analyse alignment or evolutionary distances between sequences using multivariate analysis. PCA is particularly unsuited for analysing sequence distances as it is usually applied to a matrix of correlation coefficients or variances and covariances between normally distributed variables that have been measured on the objects to be ordinated. With sequences, the variables are the different amino acids or bases that occur at different positions. It is not obvious how to analyse such data using PCA.

In this paper an ordination method that can be easily applied to alignment distances is described. This method is principal coordinates analysis (PCOORD), developed by Gower (1966). A computer program for carrying out PCOORD on sequence data sets is offered and some examples of its use are given. PCOORD operates on an $n \times n$ matrix of distances between the n objects to be ordinated. No knowledge of the original variables used to determine the distances is required. The only restriction on the distance matrix is that it should be positive semi-definite, i.e. that all its eigenvalues should be non-negative. This property is guaranteed if the distances used are Euclidean.

PCOORD uses the $n \times n$ distances to embed the n objects (sequences in this case) in a space of up to $n - 1$ dimensions. If the original distances are Euclidean, then by definition there exists an orthogonal space of at most $n - 1$ dimensions where

European Molecular Biology Laboratory, Postfach 10.2209, Meyerhofstrasse 1, 6900 Heidelberg, FRG

the distances between the objects are exactly preserved. In practice, unless the objects are equidistant, as would happen if all objects are randomly related to each other, some of these dimensions account for far more of the original information in the data set than others. By taking the most significant two or three dimensions and plotting the objects along these, the major trends and groupings in the data can be determined by visual inspection. It is a common occurrence to find that the first two dimensions account for >50% of the total information. Each dimension can be interpreted as a 'trend' in the data whose 'importance' can be measured by the proportion of the total variation it accounts for.

An alternative explanation for the coordinate axes is as follows. First you find that line which goes through the multidimensional cloud of points so as to minimize the total squared distance of all the points from the line. This is the first dimension or axis and is the dominant one, i.e. the one that accounts for the largest amount of information. The second axis is found similarly but with the added restriction that it must be at right-angles to the first. Each successive axis is found in the same way but must be at right-angles to all the preceding ones. In practice, the axes are found as eigenvectors of the $n \times n$ distance matrix. The 'importance' of each eigenvector is measured by its eigenvalue, the sum of all the eigenvalues being a measure of the total information content of the data set.

I have emphasized the role of multivariate analysis methods in data exploration. However, the techniques may also have a very practical application. Gower (1968) gives a method for superimposing new data points onto a pre-existing ordination when one knows the distance of each new object from each object in the ordination. This leads one naturally to the idea of being able to discriminate automatically and classify new sequences. One can ordinate a set of sequences from known groups and superimpose the positions of test sequences onto the ordination. One can then ask which group each test sequence is closest to. This is a form of discriminant analysis. I do not discuss how to do this in any detail here but multivariate analysis methods appear very promising for use in automatic gene identification and classification.

System and methods

Programs for calculating a distance matrix from a multiple alignment and for carrying out a PCOORD analysis were written in FORTRAN-77 and tested on a Vax 6000 running VMS. The programs are available from the author, free of charge, by sending an e-mail message to the address: higgins@EMBL-HEIDELBERG.DE.

Algorithms

Distances

In order to ordinate a set of sequences one must calculate the distance between every air of sequences (a distance matrix). The underlying assumption of PCOORD is that these distances are Euclidean. The Euclidean distance between two points in a real space is the straight-line distance between them. If the coordinates of the points are known, the distance is usually calculated using Pythagoras's theorem generalized for n dimensions. The distance between points j and k with coordinates x in n dimensions is:

$$d_{jk} = [\Sigma_{i=1,n}(x_{ij} - x_{ik})^2]^{1/2} \quad (1)$$

In the case of sequence distances we do not know the coordinates of the sequences in a space; it is the purpose of the analysis to find these coordinates. Therefore we must use some other means of calculating the distance between every pair of sequences. I first introduce a simple distance measure which is guaranteed to give Euclidean distances between a pair of sequences. I then introduce some variations on the distance measure which are expected to be more sensitive to subtle differences between sequences but which cannot be exactly Euclidean, the ordination will only approximately represent the distances between the sequences. However, the first few axes of the ordination will provide an accurate summary of the main patterns in the data. Since we do not usually use PCOORD to test statistical hypotheses, this is valid. In the next section, where PCOORD is described, I show how to recognize when a distance matrix is not strictly Euclidean and how to measure the degree to which this is so. It should be noted that in the literature on sequence comparison distances, much attention is paid to whether or not a distance measure is 'metric' (e.g. Smith *et al.*, 1981). It is necessary but not sufficient for a distance measure to be metric if it is also to be Euclidean. Unfortunately, the question as to whether or not a given sequence distance measure is Euclidean is rarely addressed.

Gower (1971) described a general measure of similarity which can be converted into a Euclidean distance. In the case of multistate characters, Gower's coefficient is the number of matching characters between two objects, divided by the number of characters considered. This gives a measure between zero and one; zero for complete difference and one for identity. This can be converted to a Euclidean distance by subtracting from one and taking the square root (Gower, 1971). In the case of sequences, the Euclidean distance between two sequences can be calculated as the square root of: the number of sites in the alignment where two sequences differ, ignoring positions where there is a gap in any sequence, divided by the number of sites considered. This is stated formally below. For two sequences i and j with m identical residues from n alignment positions, the Euclidean distance between them is:

$$d_{ij} = [(n - m)/n]^{1/2} \quad (2)$$

The square of this distance measure is widely used in phylogenetic reconstruction where it is simply the percentage divergence between sequences, uncorrected for multiple

substitutions, where one ignores any positions in an alignment where there is a gap in any sequence. Despite the apparent simplicity of this distance measure, in test cases it yields ordinations that are qualitatively as 'good' as ones using supposedly more sensitive measures.

If, because of the nature of a particular data set, one does wish to score gaps, there are two variations on the above distance measure. Firstly, one can simply treat a gap character as an extra residue type and continue as before. This will increase the distance between two sequences if one has a gap in a position where the second does not. Distances calculated using this treatment are still guaranteed to be Euclidean. Unfortunately, if very long gaps are present, as frequently happens in DNA sequence alignments, this may unduly weight the effect of the gaps. The second method is only to ignore any positions in the alignment where there is a gap in either of the two sequences being considered, i.e. one should not ignore positions where there is a gap in a third sequence that is not present in either of the two sequences being scored. Distances calculated using this variation will not be exactly Euclidean. Gaps in an alignment are the equivalent of missing data points in numerical taxonomy (Sneath and Sokal, 1973). Provided the number of gaps is small relative to the total amount of data, the distances may still be very close approximations to Euclidean distances.

The identity-based distance measure and the variations given above are perfectly adequate for most DNA or RNA sequence data sets. However, with proteins, one frequently wishes to weight differentially the distance between different pairs of amino acids. The most commonly used way of doing this is to use an amino acid similarity matrix such as a PAM 250 matrix (Dayhoff, 1978), converted to distances. Unfortunately, the distances between the 20 amino acids in a Dayhoff matrix are not Euclidean, i.e. one cannot exactly represent the distances in a space of 19 or fewer dimensions. This is obvious because in order for a distance to be metric, the distance between an object and itself must be zero. This is not the case with a Dayhoff matrix converted to distances, which is therefore not metric and hence not Euclidean. This means that distances calculated between entire sequences using a Dayhoff matrix will also not be Euclidean. This is unfortunate as the PAM 250 matrix has been shown to be a sensitive indicator of relatedness between highly diverged proteins. One solution is to use a less sophisticated amino acid distance matrix that is Euclidean. Such a matrix is described by Smith and Smith (1990). It arranges the 20 amino acids into a three-level hierarchy of similarity. Converted to distances it arranges the amino acids as follows:

Distance	Amino acids
3	any two not scored below
2	[DEKRHNQST] [ILVFWYCM]
1	[DE] [KRH] [NQ] [ST] [ILV] [FWY] [AG]
1	DEKRHNQSTILVFWYCMAGP

This scoring system can be used as an alternative to the identity-based distance measure, described earlier.

The last point to be considered is how to deal with data sets where the sequences are not already aligned. Multiple alignments are to be preferred in general because it means that 'like' is being compared with 'like' in all of the sequences. This is also true of phylogenetic reconstruction. The two examples used in this paper came from published alignments. If an alignment is not available from the literature, one can use an automatic multiple alignment program such as Clustal by Higgins and Sharp (1988, 1989) or align the sequences manually. However, in some cases this will not be possible or desirable, usually because the sequences are highly divergent and too difficult to align consistently. One must then use pairwise alignments and calculate a distance in each case. The problem then is how to find a distance meausre that will be Euclidean. Unfortunately, as far as the author is aware, no measures of distance that can be used in this way are known that are Euclidean. The best that can be done is to use one that is known to be metric. The metric nature of alignment distances is discussed by Smith *et al.* (1981). The main restriction is that terminal gaps in each alignment must be scored with a gap penalty function. There are two disadvantages to using separately calculated distances. Firstly, it will be prohibitively slow for long (>1000 residues) sequences or for many sequences (>50). Secondly, there is no guarantee that the distances will even closely approximate Euclidean distances. However, in some cases, there will be no alternative.

Principal coordinates analysis

Starting with an $n \times n$ distance matrix D, derived from n sequences, PCOORD is carried out using the following procedure described by Gower (1966).

The $n \times n$ matrix D has elements d_{jk}. Compute the $n \times n$ matrix E with elements e_{jk}.

$$e_{jk} = -1/2d_{jk}^2 \text{ for } i \text{ and } j = 1,n \qquad (3)$$

Compute the $n \times n$ matrix F with elements f_{jk}.

$$f_{jk} = e_{jk} - e_{j\cdot} - e_{\cdot k} + e_{\cdot\cdot} \qquad (4)$$

where $e_{j\cdot}$ is the mean of row j; $e_{\cdot k}$ the mean of column k; $e_{\cdot\cdot}$ is the grand mean of the matrix E. This has the effect of centring the data so that the multidimensional cloud of points will have the origin of a set of coordinate axes as its centroid. The diagonal elements of matrix F are the squared distances of each point from the centroid. Therefore the trace of F gives a measure of the total variation (sums of squares) in the data set.

Extract the eigenvectors and eigenvalues of the matrix F. Normalize each eigenvector so that its sum of squares equals the corresponding eigenvalue. Rank the eigenvectors in order of decreasing eigenvalue. Each eigenvector has n elements. The

*r*th element of the *p*th eigenvector is the coordinate of the *r*th sequence on the *p*th axis of the ordination. The variation (sum of squares) accounted for by the *p*th eigenvector is the ratio of the *p*th eigenvalue to the sum of all eigenvalues (the sum of all eigenvalues is more conveniently calculated as the trace of the matrix F above). The above procedure is guaranteed to preserve exactly the distance between two points as measured by their coordinates in $n - 1$ dimensions, providing the original distances are Euclidean. If the original distances are not Euclidean, then some of the eigenvalues will be negative. The first few eigenvectors will still provide a useful summary of the variation in the data set, but the overall fit will only be approximate. The absolute values of the negative eigenvalues will be a measure of the degree to which the ordination approximates the distances in the original data. The ordination is visualized by plotting the positions of the points along the first two or three axes.

Implementation

In this section, the methods described above are applied to two published data sets of sequences—one protein, the other RNA. These data sets were chosen because they are large and well understood. In both cases, the alignments used were based on secondary or tertiary structure information.

Globins

Bashford *et al.* (1987) published a study of all the globins in the PIR database (Sidman *et al.*, 1988). At that time this consisted of 226 sequences. Bashford *et al.* aligned the entire set of sequences using a combination of structural superposition of those sequences with known crystal structures and automatic alignment of the remainder. A PCOORD analysis of these 226 globins is shown in Figures 1 and 2. This was derived from a 226 × 226 distance matrix using the amino acid weight matrix of Smith *et al.* (1990) where gaps were only ignored for each two-sequence comparison. Figure 1 shows axis 1 versus axis 2 of the ordination, while Figure 2 shows axis 1 versus axis 3. For purposes of illustration, the sequences are divided into four groups: (i) alpha globins, including zeta globins; (ii) beta globins including epsilon, delta and gamma globins; (iii) myoglobins; (iv) plant, invertebrate and lamprey globins. The percentage of the total variation in the data set accounted for by variation along each of the first 12 axes of the ordination is given in Table I.

The first three axes of the ordination account for a total of 42% of the total variation in the data set. The fourth axis accounts for only 3%. Therefore, all of the major trends in the data can be seen by examining Figures 1 and 2. Any other trends are minor in comparison. Between the three first axes of the ordination, the four provisional groups are arranged in a tetrahedral manner. Axis 1 serves mainly to separate the alpha and beta globins from the myoglobins; axis 2 serves to separate the alpha from the beta globins, while axis 3 is devoted to separating the plant and invertebrate globins from the rest. The plant globins acutally form a neat group at the top of Figure 2 at one end of axis 3.

It is not surprising that the four provisional groups separate well. This information can also be gleaned from a phylogenetic tree of the same sequences. However, there are several features

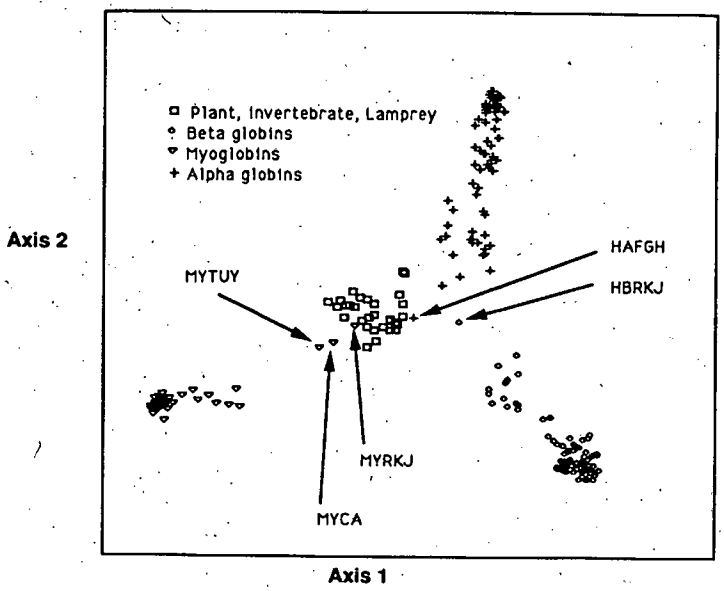

Fig. 1. Plot of axis 1 versus axis 2 of a principal coordinates analysis of 226 globin sequences. Four symbols are used to split the sequences into four broad groupings. The anomalous positions of five sequences are indicated with arrows and their PIR database names are given.

of the ordination that would not be at all apparent on a tree. Firstly, there are a number of sequences that are outliers from their respective provisional groupings. Five of these outliers are indicated by arrows on Figures 1 and 2. These are a frog alpha globin (PIR name: HAFGH); a shark beta globin (HBRKJ); and three fish myoglobins (MYTUY, MYCA and MYRKJ). Maeda and Fitch (1982) state that the frog alpha globin is a monomeric protein that may function as a myoglobin. They state that it shares many sequence features of both alpha and beta globins and that it is therefore close to the ancestral sequence of both groups. The position of this sequence on the ordination is exactly consistent with this finding. The sequence of MYTUY is from a yellowfin tuna and was reported by Watts *et al.* (1980). They reported the sequence to be very different from those of other vertebrate myoglobins. Again, its position on the ordination is consistent with this. The author has been unable to find any information in the literature as to the affinities of the other three outlier sequences. In a phylogenetic tree, these intermediate sequences would be arbitrarily placed into one of the major globin groupings. Their intermediate position, which is so obvious on the ordination, would not be apparent.

The second feature of the ordination plots that would not be obvious on a phylogenetic tree is the overall pattern of the globin radiation. Starting in the centre of the plots near the invertebrate, plant and lamprey globins, the three other groups of myoglobins, alpha globins and beta globins can be seen radiating out in different directions. The most diverged sequences are at the tips of the clusters and the less diverged sequences lie towards the centre.

The distance measure used here is not guaranteed to be Euclidean because positions with a gap in any sequence were not excluded from all comparisons. However, the sum of all negative eigenvalues (−0.013) is minute compared to the sum of all eigenvalues (21 885.298) and the approximation to a Euclidean data set is therefore almost perfect. Indeed only one eigenvalue is negative. By expectation, this should be equal to zero. However, rounding errors during the calculation will produce small fluctuations from zero. Qualitatively identical ordinations were obtained using the simple square root of the percentage difference distance or the amino acid distance matrix of Smith and Smith (1990), whether or not positions in the alignment with a gap in any sequence were excluded. These result show the method to be very robust. A broadly similar ordination was obtained from a multiple alignment that was automatically generated using the Clustal package of Higgins and Sharp (1988, 1989).

5S ribosomal RNA sequences

A compilation of >600 5S rRNA sequences from eukaryotes, prokaryotes mitochondria and chloroplasts has been published by Specht *et al.* (1990). These are short RNA molecules of ~120 nucleotides, essential for ribosome function. They have a highly conserved secondary structure composed of five helices (stems) connected by loops. Specht *et al.* provide an alignment of all known 5S rRNA sequences based on these conserved secondary structure regions. This alignment was analysed to give the ordination in Figures 3 and 4 of 610 sequences.

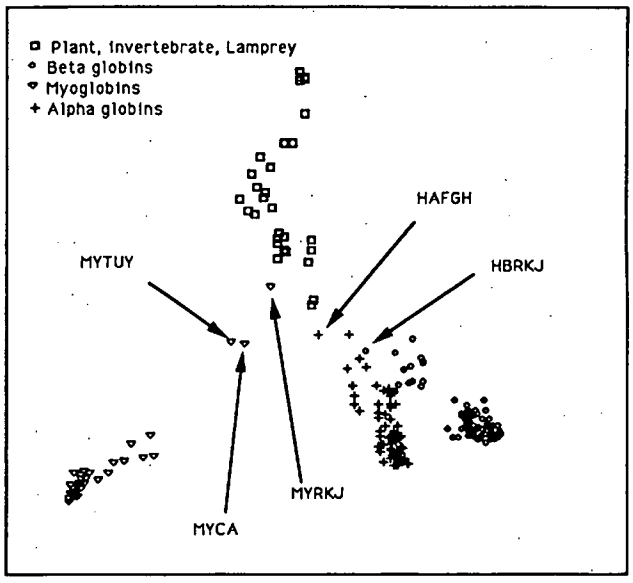

Fig. 2. Plot of axis 1 versus axis 3 of a principal coordinates analysis of 226 globin sequences. Four symbols are used to split the sequences into four broad groupings. The anomalous positions of five sequences are indicated with arrows and their PIR database names are given.

Table I. Percentage of total variation accounted for by variation along the first 12 axes of the globin and 5S rRNA ordinations

Axis	Globin	5S rRNA
1	20.8	17.0
2	14.5	5.8
3	7.2	5.3
4	3.1	4.4
5	2.7	3.7
6	2.3	3.1
7	1.8	2.8
8	1.5	2.3
9	1.4	2.1
10	1.3	2.0
11	1.2	1.8
12	1.2	1.8

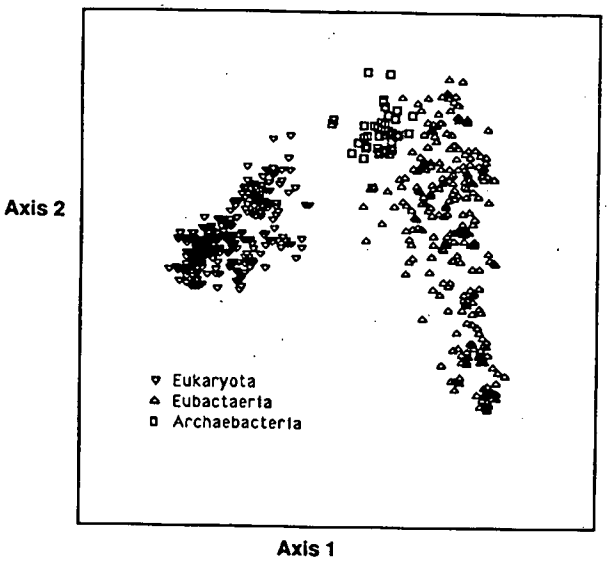

Fig. 3. Plot of axis 1 versus axis 2 of a principal coordinates analysis of 610 5S rRNA sequences.

The ordination was calculated from a 610 × 610 distance matrix using the distance formula in equation (2). Gaps were ignored only for each two sequence comparisons. The percentage variation accounted for by each of the first 12 axes of the ordination is given in Table I. The first three axes account for a total of 28% of the total variation. This figure is lower than for the previous alignment of the globins for two reasons. Firstly, this data set is three times larger, meaning that here will be more general variation. Secondly, the 5S sequences are from a much broader range of organisms. This means that the number of major patterns in the data is greater. The sequences are divided into three major groups on the ordination plots: Eukaryota, Eubacteria (including plastids and mitochondria) and Archaebacteria. Many groupings of the sequences are possible but with 610 points to plot on each figure, it is self-defeating to use more than three symbols. Therefore, some of the features

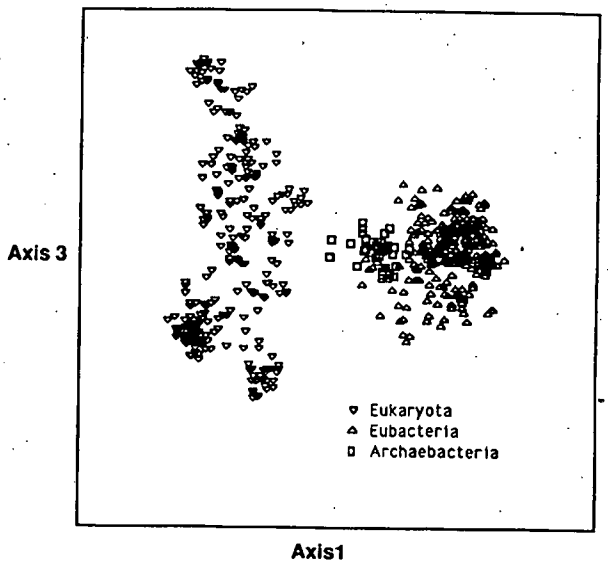

Fig. 4. Plot of axis 1 versus axis 3 of a principal coordinates analysis of 610 5S rRNA sequences.

of the ordination must be described in words rather than symbols.

Axis 1 of the ordination splits the sequences into Eukaryotes and Prokaryotes with the Archaebacteria in the centre. Axis 2 stretches the eubacterial grouping into two overlapping groups with the gram-negative bacteria at the bottom and the gram-positive bacteria at the top of the cluster, nearest the Archaebacteria. The mitochondria (just four sequences) and the 20 or so chloroplast sequences are clustered with gram positives. Axis 3 stretches out the eukaryotic sequences. Four main groupings of eukaryote separate, with the fungi forming a large group at the top of the cluster; the higher plants and green algae cluster tightly at the bottom right of the cluster; the metazoa form a tight cluster at the bottom left; and the rest of the eukaryotes form a diffuse group in the centre. This is a complicated ordination with many important patterns. In order to examine some of the patterns in more detail one would need to take subsets of the sequences and reordinate them.

The above arrangement of sequences does not help us to choose between the conflicting phylogenies advocated by Lake (1988) on the one hand and Woese (1987), Fox *et al.* (1980) and Gouy and Li (1989) on the other, for the major groups of organisms. Lake favours the monophyly of the so called Eocytes, a group of the Archaebacteria, as used here, and Eukaryotes. There are only five Eocyte sequences used in this study and they simply form a cluster within the rest of the Archaebacteria. The alternative hypothesis is that all the Archaebacteria are monophyletic. Although it is tempting to regard the grouping of the Eocytes with the rest of the Archaebacteria as evidence of the latter theory, there is insufficient data to draw definite conclusions. Further, the

arrangement of sequences that we observe could actually be consistent with either theory. All that can be safely concluded is that the Archaebacteria occupy an intermediate position between the other two major groups.

Just as with the globin ordination, the treatment of gaps in the 5S rRNA distance matrix theoretically will give distances that are not strictly Euclidean. As was stated earlier, the deviation will not be noticeable if the number of gaps is small relative to the total amount of data. In this case, the 5S rRNA alignment does have much more gaps than the globin alignment and there is a correspondingly higher number of negative eigenvalues. However, the sum of all the negative eigenvalues (−136.55) is still trivial compared to the sum of all eigenvalues (15 278.15) and the fit to a Euclidean data set is again almost perfect. It is desirable not to throw away all sites with a gap in any sequence in order to guarantee an absolutely perfect Euclidean fit because in this case one throws away half of the data.

Discussion

One obvious question to ask about using ordination to examine sequence data sets is: why bother; why not simply draw a tree using one of the standard phylogenetic estimation methods and leave it at that? Ordination should not be seen as a complete alternative to methods that give a phylogeny directly. Rather, the techniques described here should be considered to be complementary to phylogenetic analyses.

The main reason for using ordination is that it provides views of a data set that are normally invisible or unreliable on a tree. For most tree estimation methods, the easiest features to detect are the grouping together of closely related sequences, while the most difficult features are the very deep branches, reflecting ancient evolutionary events. By contrast, ordinations tend to show the major trends in a data set very clearly along the first few axes, while the fine details of affinity between closely related sequences are often hidden on minor axes. Therefore the two approaches are complementary.

Parts of a phylogeny that are unreliable may be arbitrarily represented on a tree or different methods may simply give different answers. An ordination will give a more realistic and robust representation of the affinities of difficult to place sequences. If a sequence is difficult to place into either of two groups, one should prefer to know that it is intermediate, rather than have it arbitrarily placed into one of the two. In extreme cases, such as data sets consisting of short protein motifs of say 20 amio acids (e.g. helix−turn−helix DNA binding domains), most of the phylogenetic information may be lost due to convergence, extreme selection and general lack of data (sequences are too short). Phylogenetic methods will find it dificult to detect all but the most obvious groupings in these cases. Only strictly hierarchical patterns will be seen. Ordinations, however, will still succeed in revealing trends or patterns of a non-hierarchical nature.

Providing the distances used are Euclidean, then PCOORD will be extremely robust to small variations in data sets, i.e. one will find the same patterns if one ordinates data sets consisting of slightly different numbers of sequences or different lengths of sequence. This is because PCOORD does not transform, rescale or filter the data except in deciding which trends to show on which axes. The analysis is mainly a display tool for helping to explore the distances. If the distances are not Euclidean, then this does not hold so well. However, in the examples given in this paper and in numerous test cases run by the author, the departures from a perfect fit to a Euclidean model are always trivial with real sequence data using the distance functions described in the methods section.

One further advantage of PCOORD is that it is extremely fast, even for enormous data sets. The ordination of 610 5S rRNA sequences, described earlier, took 13 min of CPU time to compute on a VAX 6000-420. The shorter ordination of 226 globin sequences took 30 s CPU time. Ordinations of <100 sequences can be carried out in just a few seconds. Depending on the methods used, a phylogenetic analysis of 610 5S rRNA sequences could take literally years to compute.

Finally, it is becoming increasingly common to use sets of aligned sequences or profiles (Gribskov *et al.*, 1987) in database similarity searches and alignments. These sets of aligned sequences are used to represent groups of sequences and are found to be more sensitive in finding similarity than when single sequences are used. Normally, the degree of similarity of a sequence to a profile is measured as some function of the residues in common between the sequence and the profile, at aligned positions. I propose that the profile be considered as a cloud of points in a space and that the similarity of a sequence to it be measured as a distance from the centroid of the cloud. In the case of the globin ordination shown earlier, consider the case of a new sequence of unknown affinity. One can calculate the distance between the new sequence and each sequence in the aligned set of globins. Then, using the method of Gower (1968), one can place the new sequence on the ordination and ask which group it is closest to. If the groups are optimally spread in space then this will be an efficient method of discrimination. In practice, if the groups are not well separated in space then the discrimination will not be optimal. One needs a method that will arrange the reference groups in space so as to maximize the distances between the centroids and minimize the overlaps. Such methods are very well known when the variables are normally distributed (Fisher, 1936) but cannot be easily applied to sequence data. The development of equivalent methods for sequence data would seem to be a very useful and imporant area of research.

Acknowledgements

The author is particularly grateful to Martin Vingron for mathematical advice. Thanks go to Thomas Specht and Jörn Wolters for sending the 5S rRNA alignment and Arthur Lesk for the globins. I thank Peter Sibbald, Graham

Cameron, Toby Gibson and Rob Kempton for very helpful discussions. Finally I thank Robert Blackith who first taught me about principal coordinates analysis in 1978.

References

Bashford,D., Chothia,C. and Lesk,A.M. (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.*, **196**, 199–216.

Blackith,R.E. and Reyment,R.A. (1971) *Multivariate Morphometrics.* Academic Press, London.

Cooley,W.W. and Lohnes,P.R. (1971) *Multivariate Data Analysis.* Wiley, New York.

Dayhoff,M.O. (1978) *Atlas of Protein Structure and Function.* National Biomedical Research Foundation, Silver Spring, MD, Vol. 5, Suppl. 3.

Digby,P.G.N. and Kempton,R.A. (1987) *Multivariate Analysis of Ecological Communities.* Chapman & Hall, London.

Fisher,R.A. (1936) The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, **7**, 179–188.

Fox,G.E., Stackebrandt,E., Hespell,R.B., Gibson,J., Maniloff,J., Dyer,T.A., Wolfe,R.S., Balch,W.E., Tanner,R.S., Magrum,L.J., Zablen,L.B., Blakemore,R., Gupta,R., Bonen,L., Lewis,B.J., Stahl,D.A., Luehrsen,K.R., Chen,K.N. and Woese,C.R. (1980) The phylogeny of prokaryotes. *Science*, **209**, 457–463.

Gouy,M. and Li,W.-H. (1989) Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. *Nature*, **339**, 145–147.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–328.

Gower,J.C. (1968) Adding a point to vector diagrams in multivariate analysis. *Biometrika*, **55**, 582–585.

Gower,J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857–871.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.

Higgins,D.G. and Sharp,P.M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*, **73**, 237–244.

Higgins,D.G. and Sharp,P.M. (1989) Fast and sensitive multiple sequence alignments on a microcomputer. *Comput. Applic. Biosci.*, **5**, 151–153.

Hotelling,H. (1933) Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, **24**, 417–441.

Lake,J.A. (1988) Origin of the eukaryotic nucleus determined by rate-invariant analysis of rRNA sequence. *Nature*, **331**, 184–186.

Maeda,N. and Fitch,W.M. (1982) Isolation and amino acid sequence of a monomeric hemoglobin in heart muscle of the bullfrog, *Rana catesbeiana*. *J. Biol. Chem.*, **257**, 2806–2815.

Myers,E.W. and Miller,W. (1988) Optimal alignments in linear space. *Comput. Applic. Biosci.*, **4**, 11–17.

Shepard,R.N. (1981) Multidimensional scaling, tree fitting, and clustering. *Science*, **210**, 390–398.

Sidman,K.E., George,D.G., Barker,W.C. and Hunt,L.T. (1988) The protein identification resource (PIR). *Nucleic Acids Res.*, **16**, 1869–1871.

Smith,R.F. and Smith,T.F. (1990) Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA*, **87**, 118–122.

Smith,T.F., Waterman,M.S. and Fitch,W.M. (1981) Comparative biosequence metrics. *J. Mol. Evol.*, **18**, 38–46.

Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy.* Freeman, New York.

Specht,T., Wolters,J. and Erdmann,V.A. (1990) Compilation of 5S rRNA and 5S rRNA gene sequences. *Nucleic Acids Res.*, **18**, 2215–2230.

Watts,D.A., Rice,R.H. and Brown,W.D. (1980) The primary structure of myoglobin from yellowfin tuna (*Thunnus albacares*). *J. Biol. Chem.*, **255**, 10916–10924.

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

Received on February 14, 1991; accepted on June 5, 1991