Vol. 140, No. 3, 1986 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
November 14, 1986 Pages 808-813

A USE FOR PRINCIPAL COORDINATE ANALYSIS IN THE COMPARISON OF PROTEIN SEQUENCES

Kevin J. Woolley* and Medha Athalye

Department of Molecular Biology, University of Edinburgh, Edinburgh, EH49 3JR, Scotland

Received September 30, 1986

Principal Coordinate analysis (PCO) was applied to the comparison of protein sequences. A similarity matrix was derived from a dataset containing 21 *c*-type cytochrome sequences and this was analysed using PCO to produce a plot of the first three principal axes. The relationships indicated from this plot are considered in conjunction with those derived by cluster analysis using the UPGMA method, and the advantages offered by a non-hierarcical method of sequence comparison discussed.

The hypothesis that protein sequence data contains evolutionary information was made soon after homologous protein sequences first became available (1-3). Numerous procedures have since been proposed which attempt to deduce a phylogenetic tree from sequence data (see 4 for an extensive review). However, for datasets containing a high proportion of parallel and back mutations the validity of the trees produced by many of these methods is open to question, especialy for widely divergent members (5).

An alternative approach is to apply a non-hierarchical method, such as ordination analysis. Although these methods do not derive any form of tree-like classification, they do illustrate the relationships between members of the dataset, being particularly useful for distant relationships (6). Ordination methods differ in detail, but the general principal of calculation is the same in all cases. A data matrix (e.g. a NxN similarity matrix) is used to specify coordinates in a multidimensional space.

* Present address: Department of Botany, University of Glasgow, Glasgow, G12 8QQ, Scotland. (To whom all correspondence should be sent)

0006-291X/86 $1.50

This multidimensional plot is then analysed so as to find the best representation (i.e. that containing the maximal amount of the information present in the multidimensional plot) in a lower, usually 2 or 3, dimensional space. The resultant analysis can then be visualy examined, often in conjunction with some form of cluster analysis, for relationships.

Ordination methods have been extensively used in numerical taxonomy (7), but their application to protein sequence data has recieved little attention. Although Ohnishi (8) and Nishikawa (9-11) have both used Ordination analysis for the analysis of protein amino acid composition data.

METHOD

The dataset used consisted of the 21 different *c*-type cytochrome sequences listed in Table 1. Both mitochondrial cytochromes *c* and

TABLE 1
Sequences used in the PCO dataset

Species[1]	Type	Class[2]
Homo sapiens	mitochondrial *c*	IB
Ginko biloba	"	"
Neurospora crassa	"	"
Crithidia oncopelti	"	"
Euglena gracilis	"	"
Tetrahymena pyriformis	"	"
Rhodomicrobium vannielii	prokaryotic c_2	"
Rhodopseudomonas viridis	"	"
Rhodopseudomonas acidophila	"	"
*Rhodopseudomonas globiformis**	"	"
*Rhodopseudomonas salexigens**	"	"
Rhodospirillum rubrum	"	IA
Rhodospirillum photometricum	"	"
Aquaspirillum itersonii+	"	"
Rhodopseudomonas palustris	"	"
Rhodopseudomonas sphaeroides	"	"
Rhodopseudomonas capsulata	"	"
Paracoccus denitrificans‡	"	"
Rhodospirillum fulvum	"	IB
Rhodospirillum molischianum iso-1	"	"
Rhodospirillum molischianum iso-2	"	"

1. All sequences were taken from 15 except those marked (*) which were as given by Ambler (Pers. comm.), (+) Woolley (Pers. comm.), or (I) Ambler (16).
2. As described by Ambler (13, 14).

prokaryotic cytochromes c_2 were included. The sequences were aligned using the scheme described by Dickerson (12). A 21x21 similarity matrix was calculated using the following scoring: all amino acid residue matches were scored as 1, all mismatches as 0, and all null (i.e. deletion/deletion) matches as 0. The total score for each pair of sequences, divided by the nominal length (i.e. including insertions required for sequence alignment) gave the sequence similarity, which was expressed as a percentage. Calculation of the similarity matrix, and Principal Coordinate analysis (PCO) was performed performed using the routines available through the GENSTAT package (Rothemstead version 2.04). The first three coordinates of the PCO were plotted using routines available through the GINO-F graphics package (Cambridge Design Centre). Cluster analysis, using the Unweighted Pair Group Method with Averages (UPGMA) was also applied to the dataset, again using the GENSTAT package.

RESULTS AND DISCUSSION

The plot of the first three coordinates from the PCO analysis is shown in Figs. 1-3. The first, second, and third dimensions respectively contain

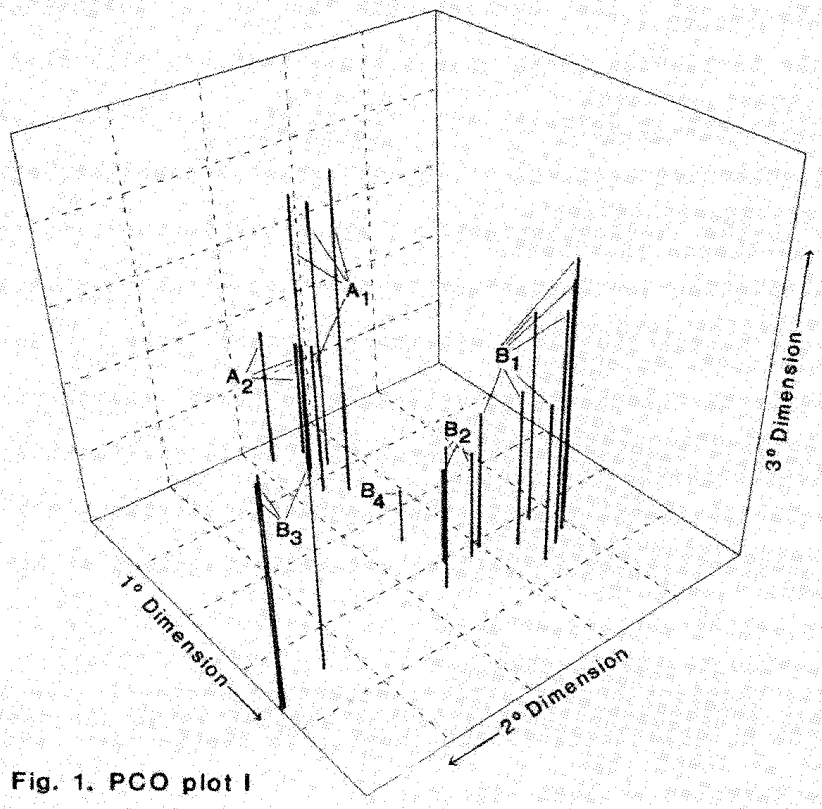

Fig. 1. PCO plot I

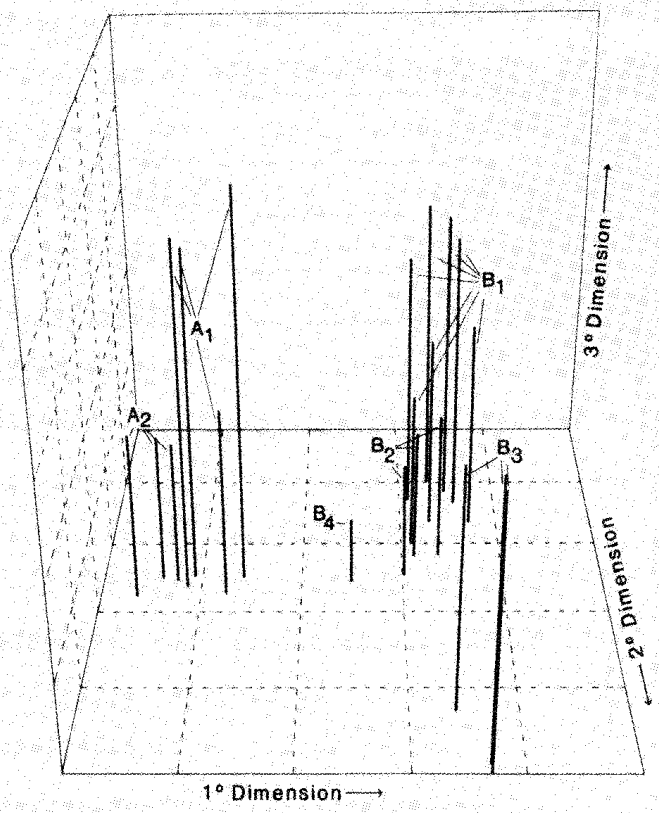

Fig. 2. PCO plot II

12.8%, 10.4% and 7.3% of the total information present in the similarity matrix, a total of 30.5%. This total is enough to be reasonably confident that the figure does not contain any gross distortions of the sequence relationships. For the purposes of this discussion the PCO plot will be examined in conjunction with the UPGMA analysis shown in Fig. 3.

The UPGMA dendrogram may be broadly divided into two groups of sequences, labelled A and B. These groups also correspond to classes IA and IB proposed by Ambler (13,14) in his scheme for the sequence classification of c-type cytochromes. From the examination of the PCO plot it may be seen that the separation of sequences along the first dimension also corresponds to this scheme. The only sequence that does not appear to conclusively cluster with either of the two large groups is that belonging to *Rhodopseudomonas salexigenes* (labelled B_4). However this is also in agreement with the UPGMA diagram, where it clusters at a low similarity level with the 'B' group.

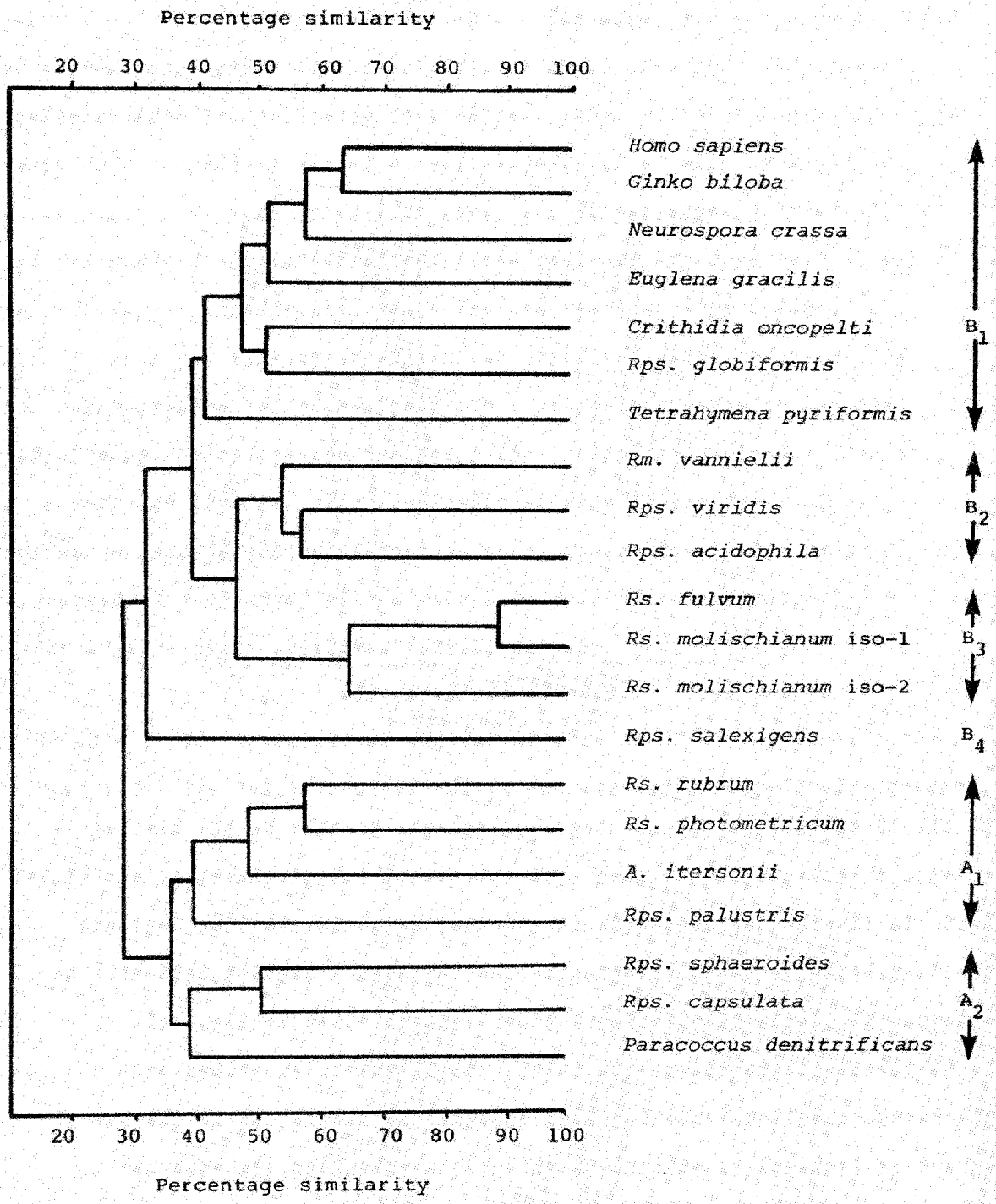

Fig. 3. UPGMA dendrogram of *c*-type cytochrome dataset

Further examination of the UPGMA dendrogram shows that the two large groups can be subdivided into several smaller groups, labelled A_1, A_2 and B_1 - B_4. When all three dimensions shown in the PCO plot are examined most of these subgroups can be clearly seen. For instance subgroups B_3 and B4 may be seen in the large 'B' group. The separation of the B_1 and B_2

subgroups is not as apparent as might be expected from the UPGMA dendrogram but this may be explained by variations in similarity values for individual sequences in the groups. For instance, *Rhodopseudomonas viridis* cytochrome c_2 (clustered by UPGMA into group B2) has a higher average similarity for the B_1 group (42.9%) than does *Tetrahymena pyriformis* mitochondrial cytochrome *c* (40.0% clustered by UPGMA in the B1 group). The A1 and A2 subgroups also show the same effect. This shows how the use of Ordination analysis can illustrate factors not easily apparent from an hierarcical treatment of data.

CONCLUSION

The application of Ordination methods to the comparison of protein sequence data may be used to give a graphical representation of the similarity relationships between sequences. As such, these methods offer a useful adjunct to tree-based procedures, especially where the relationships between subgroups of sequences is being considered.

ACKNOWLEDGMENTS

We would like to thank Dr. R. P. Ambler for his advice on all aspects of this work, which was supported by the Science and Engineering Research Council (UK).

REFERENCES

1. Crick, F. H. C. (1958) Symp. Soc. Exp. Biol. 12, 138-163
2. Anfinsen, C. B. (1958) in The Molecular Basis of Evolution, John Wiley & Sons inc, New York.
3. Zuckerkandl, E. & Pauling, L. (1965) J. Theor. Biol. 8, 357-366
4. Felsenstein, J. (1982) Quart. Rev. Biol. 57, 379-404
5. Meyer, T. E., Cusanovich, M. A. & Kamen, M. D. (1986) Proc. Natl. Acad. Sci. U.S.A. 83, 217-220
6. Rohlf. F. J. (1968) Syst. Zool. 17, 246-255
7. Sneath, P. H. A. & Sokal, R. R. (1973) in Numerical Taxonomy, W. H. Freeman & Company, San Fransisco, U.S.A.
8. Ohnishi. K. (1978) in Evolution of Protein Molecules (Matsubara, H & Yamanaka, T. Eds.), pp75-87, Japan Scientific Societies Press, Tokyo
9. Nishikawa, K. & Ooi, T. (1982) J. Biochem. 91, 1821-1824
10. Nishikawa, K., Kubota, Y. & Ooi, T. (1983a) J. Biochem. 94, 981-995
11. Nishikawa, K., Kubota, Y. & Ooi, T. (1983b) J. Biochem. 94, 997-1007
12. Dickerson, R. E. (1980) Scientific American 242, 98-111
13. Ambler, R. P. (1977) in The Evolution of Metalloenzymes, Metalloproteins & Related Materials (Leigh. G. J., Eds.), pp 100-118 Symposium Press
14. Ambler, R. P. (1982) in From Cyclotrons to Cytochromes (Kaplan, N. O. & Robinson, A., Eds.), pp263-280, Academic Press, London
15. Dayhoff, M. O. (1972) in Atlas of Protein Sequence and Structure, Volume 5 (Supplements 1973, 1976 & 1978) National Biomedical Research Foundation, Washington.
16. Ambler, R. P., Meyer, T. E., Kamen, M. D., Schichman, S. A. & Sawyer, L. (1981) J. Mol. Biol. 147, 351-356