TESTS FOR HIERARCHICAL STRUCTURE IN RANDOM DATA SETS

F. JAMES ROHLF AND DAVID R. FISHER

Abstract

The degree of fit of phenograms (constructed using UPGMA cluster analysis) to similarity matrices based on random data was examined using the cophenetic correlation coefficient. Approximate critical bounds were established for testing for the presence of hierarchic structure in one's data. Sokal's taxonomic distance coefficient was shown to be more sensitive than the product-moment correlation coefficient to differences in the multivariate structure of the sampled distributions.

The purpose of this study was to investigate the degree of fit of phenograms to similarity matrices based on random data (both multivariate normal and uniform distributions were used). The degree of fit was measured using the cophenetic ĉorrelation coefficient (Sokal and Rohlf, 1962). It is of interest to develop a standard for comparsion with actual numerical taxonomic results in order to formulate a test criterion to indicate whether one has sufficient evidence to indicate that the phenetic relationships present in one's data are hierarchic (rather than simply what could be expected from a random sample of a single homogeneous population).

METHODS

Since a mathematical analysis of the statistical properties of the cophenetic correlation coefficient would be rather formidable, two Monte Carlo studies were carried out. In the first, data matrices with 25, 50, 100, 150, and 200 operational taxonomic units (OTU's) and 50, 100, 150, and 200 characters were generated using a pseudo-random number generator (for normal distributions, $\mu = 0$, $\sigma^2 = 1.$) on a digital computer. Hence, each OTU represented an observation from a multivariate normal distribution with parametric means equal to zero, variances equal to unity, and the correlation between all pairs of characters equal to zero. In the second study, data matrices for a few of the above combinations of OTU's and characters were generated using another pseudo-random number generator (uniform distribution over the range 0.0 to 1.0). Both sets of random data matrices were processed using the following procedures. The characters were standardized so that each character would have a sample mean of zero and a sample standard deviation of unity. In the case where the samples were drawn from the normally distributed population this had little effect. However, it was done in order to make the results comparable to standard numerical taxonomic procedures. Correlations and distance coefficients (Sokal, 1961) among the OTU's were then computed and the resulting matrices were clustered using the unweighted pair group method with arithmetic averages (UPGMA) described in Sokal and Sneath (1963). Cophenetic correlations were then computed by correlating elements in the original similarity matrix with the cophenetic values obtained from the phenograms resulting from the cluster analysis (see Sokal and Rohlf, 1962).

RESULTS

Figs. 1 and 2 show the average value of the cophenetic correlations found for each of the combinations of OTU's and characters employed in the study. Fig. 1 shows results of the analyses using correlation coefficients as indices of similarity among the OTU's and Fig. 2 the results for analyses using distance coefficients among the OTU's. In Fig. 1, the average value for the cophenetic correlations is about 0.55 for studies involving only 25 OTU's and this

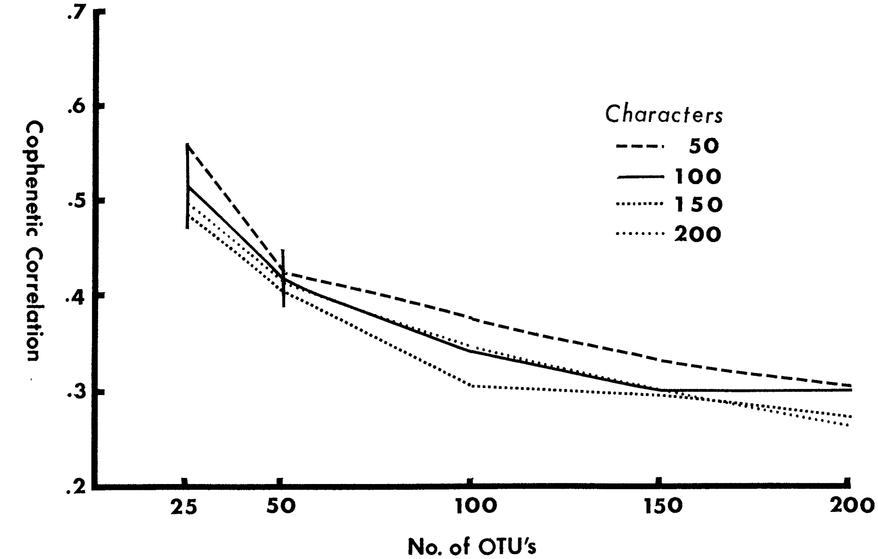

Fig. 1.—Plot of average magnitude of cophenetic correlation versus number of OTU's for various numbers of characters. The cophenetic correlations were computed using UPGMA cluster analysis based on a matrix of correlation among OTU's computed from standardized normally distributed random numbers. Vertical lines for 100 characters and 25 and 50 OTU's indicate the limits of $\bar{Y}\pm2s$ (mean ± 2 standard deviations) for these two cases. Sample sizes for the points on which the graph is based are shown in Table 1.

magnitude drops as one increases the number of OTU's in the study. There appears to be a tendency for studies based on fewer characters to have a higher cophenetic correlation than studies based upon larger numbers of characters, although the difference is not great. This indicates that UPGMA cluster analysis can by chance find apparent clusters when there are relatively few OTU's and characters in the study.

Fig. 2, using a distance coefficient as an index of taxonomic similarity, shows a somewhat different pattern. The magnitude of the cophenetic correlation appears to drop only slightly as one increases the number of OTU's from 25 to 50. There is also only a slight indication that the average value of cophenetic correlation might drop as one increases the number of characters in the study.

Two of the combinations of characters and OTU's received additional replications so that it would be possible to estimate the reliability of the average cophenetic correlations found above (see Table 1). This was done for 100 characters and for 25 and 50 OTU's. It was found that the standard deviations were much higher for cophenetic correlations based upon distance coefficients than those based on correlation coefficients. The standard deviations were 0.01850 and 0.00615, respectively. On this basis, it is possible to make crude estimates of 5% critical bounds for testing the null hypothesis that one has a single homo-

Table 1. Sample sizes for the average cophenetic correlations graphed in Figs. 1 and 2.

		Number of OTU's				
		25	50	100	150	200
	50	1	1	1	1	1
Number of	100	13	13	2	2	2
characters	150	1	1	1	1	1
	200	2	3	2	1	1

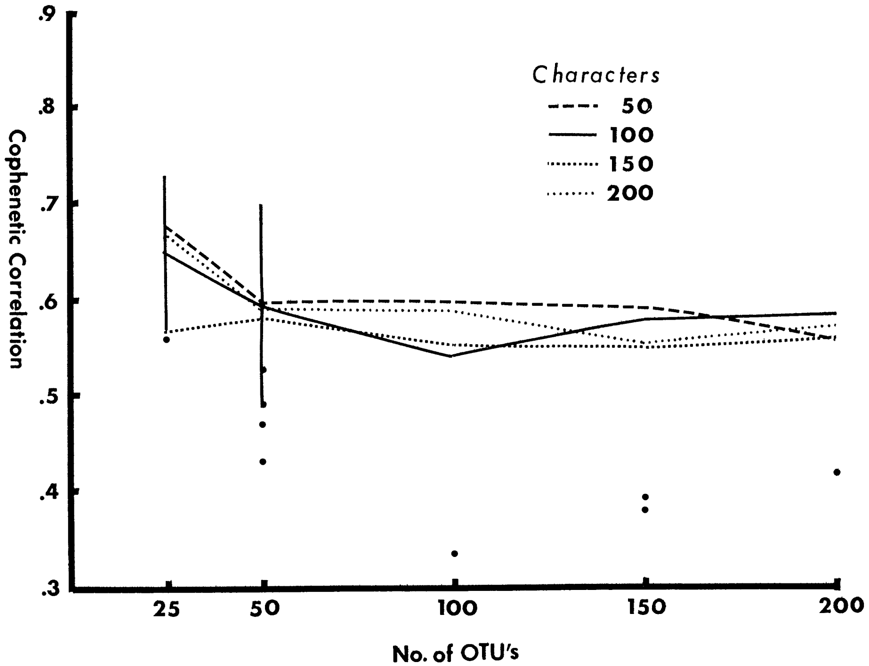

FIG. 2.—Graph of average magnitude of cophenetic correlation versus number of OTU's for various numbers of characters. The cophenetic correlations were computed using UPGMA cluster analysis based upon matrices of taxonomic distances among OTU's computed from standardized normally distributed random numbers. Vertical lines for 100 characters and 25 and 50 OTU's indicate the limits of $\bar{Y} \pm 2s$ for these two samples. Sample sizes for the points on which the graph is based are shown in Table 1. In addition, cophenetic correlations for uniformly distributed random numbers are shown as points.

geneous normally distributed cluster versus the alternative hypothesis that one has a system of nested clusters. These critical bounds (not confidence limits) are indicated by the vertical lines in Figs. 1 and 2 ($\bar{Y}\pm2s$; mean ± twice the standard deviation). Most studies of actual taxonomic data have had cophentic correlations well above these limits (e.g., values of 0.8 are typical and values of 0.9 are not too unusual).

It would, of course, be desirable to replicate each combination of characters and OTU's a large number of times, but owing to the large amount of computer time involved, this was not practical (each "observation" represents a single conventional numerical taxonomic study).

In the second half of this study, a few data matrices were constructed using uniformly distributed random numbers as values in the data matrices. When correlation coefficients were used, the uniform data gave results similar to that obtained before, and the results are not shown. However, the distance coefficients gave very different results. The cophenetic correlations obtained for distances were much lower for the uniform data than for the normally distributed data. This can be seen in Fig. 2 where these cophenetic correlations are shown as points. These results indicate that the distance coefficient is more sensitive to different patterns of the distribution of points than is the correlation coefficient. The distance coefficient also appears to have the advantage that it is less affected by the number of OTU's included in the study (especially when the number of OTU's is greater than 50).

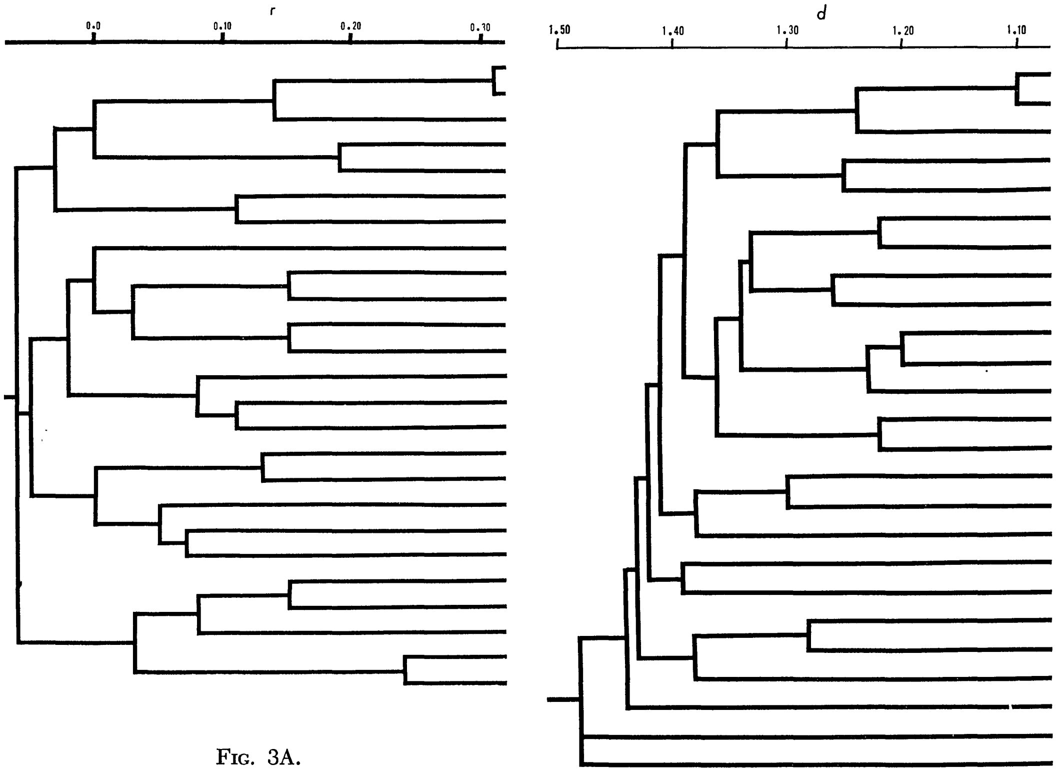
r
0.0
0.10
0.20
0.30
d
1.50
1.40
1.30
1.20
1.10

FIG. 3A.

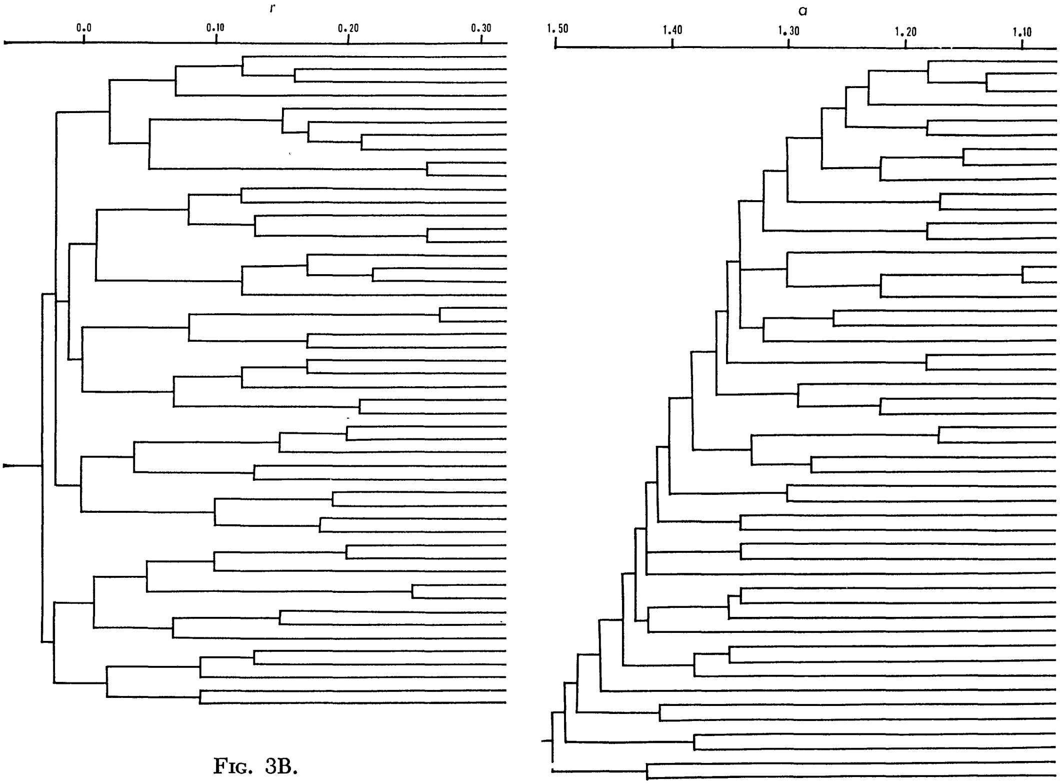
r
0.0
0.10
0.20
0.30
a
1.50
1.40
1.30
1.20
1.10

FIG. 3B.

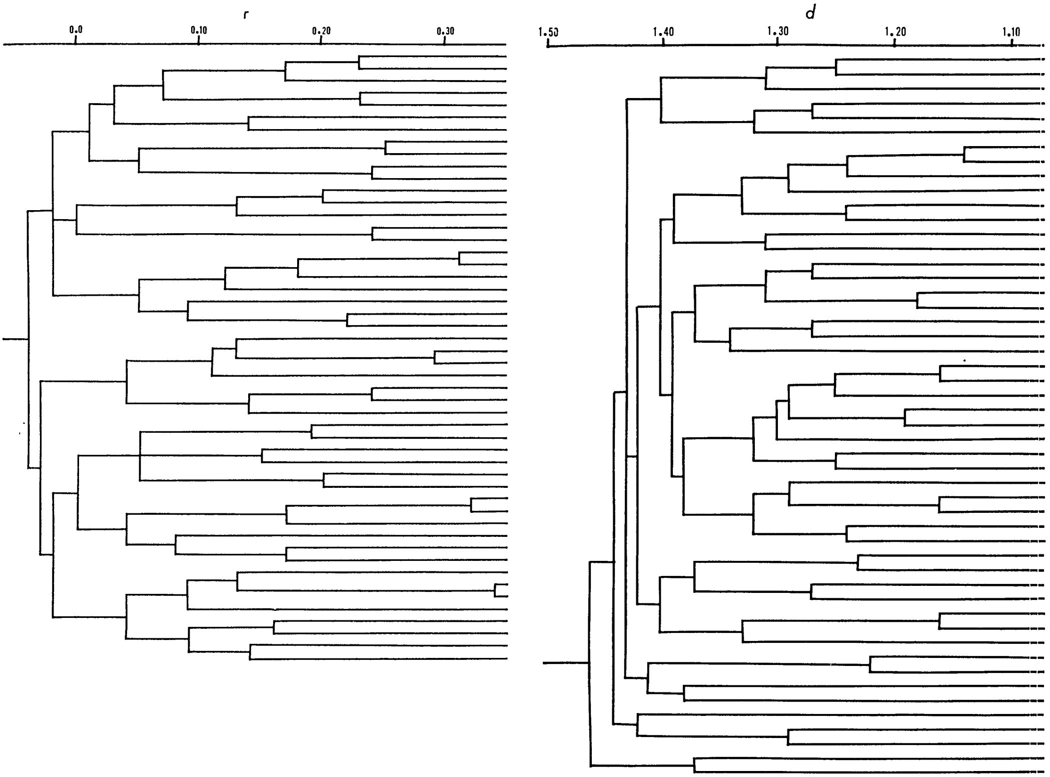

FIG. 4.—Sample phenograms based upon numerical taxonomic studies of uniformly distributed random data. UPGMA cluster analysis applied to correlation (r) and distance (d) matrices for 50 OTU's and 100 characters.

We also examined the phenograms which resulted from the various matrices, in order to allow a qualitative comparison of these "random" phenograms and those based upon actual data. No quantitative indices have been developed previously which completely describe the form and shape of a phenogram. A few samples of phenograms are shown in Figs. 3 and 4. The form of the resulting phenograms appears to be, in some respects, distinctively different from that found in most actual numerical taxonomic studies. The most noticeable characteristics of these "random" phenograms (in contrast to typical phenograms with which we are familiar) is that all the branching in the phenogram is restricted to a rather narrow region. For example, note that the correlation phenogram in Fig. 3A has a scale ranging from only about -0.05 to 0.31 and the corresponding distance phenogram has a scale ranging from only about 1.5 to 1.1. Thus, there are not very many extremely close OTU's or OTU's that are quite distant from one another. Another impression one gains is that at any one phenon level there tend to be many clusters of relatively few OTU's rather than relatively few distinctive clusters with large numbers of OTU's, the latter being more common in our experience with

←

FIG. 3.—Sample phenograms based upon numerical taxonomic studies of normally distributed random data. **A** shows phenograms resulting from UPGMA cluster analysis applied to correlation (r) and distance (d) matrices for 25 OTU's and 100 characters. **B** shows similar phenograms for a study based on 50 OTU's and 100 characters.

actual data. This indicates that there is rather less diversity in the structure of the random data than one usually finds in actual biological data. Distance phenograms for uniform data (see Fig. 4) are particularly distinctive, being much more symmetrical than the "typical" distance phenogram. The distance phenograms in Fig. 3 show the typical degree of skewness.

ACKNOWLEDGMENTS

Contribution No. 1388 from the Department of Entomology, The University of Kansas. This investigation was supported by a grant (GB-4927) from the National Science Foundation. All computations were carried out using the NT-SYS system of computer programs (developed by the senior author with the assistance of John Kishpaugh and Ron Bartcher) on the GE-625 computer at The University of Kansas Computation Center. The authors appreciate the assistance of Patricia Rohlf who prepared the illustrations and Drs. P. H. A. Sneath and R. R. Sokal who read the manuscript and made many useful suggestions.

REFERENCES

Sokal, R. R. 1961. Distance as a measure of taxonomic similarity. Syst. Zool., 10:70–79.

Sokal, R. R., and F. J. Rohlf. 1962. The comparison of dendrograms by objective methods. Taxon, 11:33–40.

Sokal, R. R., and P. H. A. Sneath. 1963. The principles of numerical taxonomy. Freeman, San Francisco, 359 p.

Department of Entomology, The University of Kansas, Lawrence, Kansas 66044; and Department of Zoology, The University of Kansas, Lawrence, Kansas 66044.