ADAPTIVE HIERARCHICAL CLUSTERING SCHEMES

F. James Rohlf

Abstract
Rohlf, F. J. (Biological Sciences, State Univ., Stony Brook, N. Y. 11790) 1970. Adaptive hierarchical clustering schemes. Syst. Zool., 18:58–82.—Various methods of summarizing phenetic relationships are briefly reviewed (including a comparison of principal components analysis and non-metric scaling). Sequential agglomerative hierarchical clustering schemes are considered in particular detail, and several new methods are proposed. The new algorithms are characterized by their ability to “adapt” to the possible trends of variation found within clusters as they are being formed. A nonlinear version allows the isolation and description of clusters which are parabolic, ring-shaped, etc., by the introduction of appropriate dummy variables. Procedures for computing the best fitting trend line through the cluster are also presented, and problems in measuring the amount of information lost by clustering are discussed. [Phenetics; cluster analysis; numerical taxonomy.]

This paper is concerned with a brief review of some of the techniques of summarizing phenetic similarities that have been proposed for use in numerical taxonomy. One class of methods (sequential agglomerative) is considered in detail and new procedures which allow for elongated and curvilinear clusters are proposed.

The “taxonomy problem” in biology can be described as follows: Given a set of specimens (“operational taxonomic units” or OTU’s, Sokal and Sneath, 1963, which may represent taxa of any rank) known only by a list of their properties or characters, we wish to find the “best” way of describing their often complex patterns of mutual similarities (phenetic relationships). Such relationships do not necessarily imply evolutionary (cladistic) relationships (for a discussion of these approaches, see Sokal and Camin, 1965).

The methods that have been developed appear to have a more general application than just in biological taxonomy, but there are certain facts and assumptions that can be made in biology which influence our choice of methods. As a result, the techniques may or may not be completely valid in other fields. Some of the considerations which influence the development of cluster analyses in biological taxonomy are the following: (1) “All things being equal” we would hope that a system of nested clusters would be found. This is due to the fact that evolution is believed usually to be a divergent process and the distribution of OTU’s in a phenetic space should to some extent reflect this. There are, of course, exceptions to this overall rule which are very important, such as those provided by hybridization and clinal variation. (2) Another consideration is the nature of the character set representing each OTU. We would like to use a “random sampling of characters” or at least a “representative” sampling of characters. But since different sets of characters seem to yield slightly different systems of relationships (Rohlf, 1963; Ehrlich and Ehrlich, 1967; Michener and Sokal, 1966), biologists may have to get used to the idea of using different classifications, based upon different sets of characters, each best for its own special purpose, with overall similarities based on the total character set available at any one time. (3) The selection of OTU’s is also not random. Since we cannot study all organisms, we must select those which are of immediate interest. But even with a specified group of organisms, we usually cannot sample at random. This is so because the distributions of recent (and even fossil) organisms are clumped in a phenetic hyperspace. One needs to pass up many very similar, common specimens to obtain a more interesting sampling of different kinds of organisms. Thus, a preliminary screening of individuals according to their apparent similarities must be made before one can make detailed measurements to analyze their phenetic relationships quan-
titatively. The problem here is mainly the great amount of effort required to record large numbers of characteristics for large numbers of individuals. It is hoped that with the development of automatic data recording devices, it will be possible to gather large amounts of data easily so that random sampling will become practical. As a consequence of points (2) and (3) given above, cluster analytic procedures are only quasi-statistical (see also Sneath, 1967, for further discussion).

A major problem in cluster analysis is that there is no universal agreement on what constitutes a cluster; and, most investigators think that clustering methods depend on the clusters to be found. This may be the reason why almost everyone in this field has his own version of cluster analysis. Moreover, it is important not to cluster one's data blindly, but to develop techniques of clustering which allow one to check the reasonableness of the clusters for the data to which it is applied.

An additional problem which must be faced is the fact that present applications in biological taxonomy require the use of cluster analysis on data with relatively few OTU's. Thus, one does not have a high density of points within clusters (if they exist), but, rather, only vague outlines and suggestions of clusters.

SUMMARY OF NUMERICAL TAXONOMIC METHODS

The standard computational steps performed in a "typical" numerical taxonomic study are outlined below for reference purposes later in the paper.

The following preliminary steps are usually carried out.

A. Data are gathered on $p$ characters (variables) for $t$ individuals (operational taxonomic units or OTU's). The data are usually quantitative measurements of some sort and often are coded into a few class intervals. The result is a $p$ by $t$ matrix $Y$. Frequently there are missing observations in the matrix.

B. The characters are standardized over the OTU's to place them in equivalent units. Usually they are transformed so as to have a mean of zero and a unit standard deviation (or else the ranges may be equalized). It should be pointed out that some procedures are unaffected by such scaling (see below and Friedman and Rubin, 1966, for examples).

C. A matrix of coefficients is computed which measures the relative degree of similarity between all pairs of OTU's. Some coefficients which have been used are: the product-moment correlation coefficient, $r$; average distance, $d$ (Sokal, 1961); generalized distance, $D^2$ and various association coefficients (see Sokal and Sneath, 1963; Rohlf and Sokal, 1965; Huizinga, 1962).

The matrices generated in step C are usually quite large $(t(t-1)/2$ elements). It is therefore necessary to summarize the information on the phenetic relationships implied by them so that the results can be easily comprehended and communicated. Since the results must be simplified, there will almost always be some loss of information and consequent distortion in the final results. The problem is to select a method which gives the proper balance between preservation of information (good fit) and simplicity.

Three classes of methods may be distinguished:

1. Multi-dimensional scaling (ordination). These are techniques in which one obtains the coordinate axes of each OTU in the smallest dimensional space that still preserves sufficient information about the interpoint distances (Kruskal, 1964). This can be accomplished in various ways depending on the restrictions placed upon the system. A special case (in which it is required that the new coordinated axes be linear combinations of the original variables and that the maximum percent of the original variance of the original variables be preserved) is called "principal components analysis" (Rao, 1952; Anderson, 1958; for a taxonomic example see Rohlf, 1967, 1968). Alternatively one has nonmetric scaling (Kruskal, 1964) and parametric mapping
(Shepard and Carroll, 1966). The advantage of these techniques is that they allow one to examine a scatter diagram displaying a summary of the structure of the data without having to first assume that clusters are present. A possible disadvantage for taxonomy is that they do not yield a hierarchical classification. I know of no published applications of the nonmetric scaling techniques to biological taxonomy but the few such analyses I have tried have given very satisfactory results (good [low stress] compression of the data into few dimensions). Fig. 1 shows stereo pairs based on a principal components analysis (Fig. 1a) and nonmetric scaling (Fig. 1b).

2.—Network Analyses. In these procedures one attempts to find networks (with varying degrees of connectedness) which will reflect the underlying structure of the data. Examples are the use of shortest simply connected networks (Kruskal, 1956; called "Prim" networks by Cavalli-Sforza and Edwards, 1967, who cited the work of Prim, 1957). Jardine and Sibson (1968a and b) and Neely (1969) have suggested the use of networks that are more than simply connected and thus capable of more adequately describing the data structure, but these are more complex to represent. Jardine and Sibson (1968a and b) also discuss methods for constructing a classification from their networks. An example of a shortest connection network is given in Fig. 1.

3.—Cluster Analysis. In these procedures OTU's are allocated to classes. Some of the classes may be nested within other classes if a hierarchical clustering scheme (HCS, Johnson, 1967) is used. If compact clusters are actually present with large gaps between clusters, then these procedures yield very adequate summaries. A large number of clustering algorithms which have been
proposed. Ball (1965), Lance and Williams (1966), Gower (1967), and Bonner (1964) present accounts which compare some of the properties of various methods. Different methods are most suitable for data with certain structure, but relatively little is known about the type of data structure for which each method is most suitable. This point has been considered by Rohlf (1963 and 1967) and Sokal (1966).

Regardless of which of the above types of summarization techniques are used, it is important that they be accompanied by measures of the degree of fit (lack of distortion) between the original matrix of similarity coefficients and the pattern of similarity values implied by the summary. A number of such techniques have been proposed. For example, in principal components analysis one commonly quotes the percentage of the total variance which is contained in the first $k$ dimensions. A useful approach to this problem of fit was proposed by Hartigan (1967) and Jardine et al. (1967). They suggested that one consider the process of constructing a hierarchical clustering scheme as one of fitting a hierarchical tree structure (which has the properties of an ultrametric; Hartigan, 1967) to the original distances (which usually are metric). If the distance between two OTU’s in a phenogram is the level of the lowest cluster which contains both OTU’s, then such distances satisfy the ultrametric inequality $(d_{ij} \leq \max\{d_{ik}, d_{jk}\}$, for all OTU’s in the set). This approach implies the restriction that the level of each cluster ("cophenetic value" of Sokal and Rohlf, 1962) is a monotonically increasing function. Thus, "reversals" found in certain clustering methods (Michener and Sokal, 1957; Farris, 1969) are not permitted. Sokal and Rohlf (1962) proposed the cophenetic correlation (correlation between the elements of the original similarity matrix and the similarity values implied by a scheme of representation such as a cluster analysis, phenogram, principal components analysis, etc., see Fig. 9 below). Hartigan (1967) proposed a weighted sum of squared differences between the original euclidean distances and distances given by the phenogram (the fitted ultrametric). Jardine et al. (1967) proposed a class of measures of distortion, $\Delta\mu$, which involve sums of the absolute values of the differences between the original and implied distances raised to various powers $(1/\mu)$. Jardine and Sibson (1968a) give a normalized version $\tilde{\Delta}\mu$. By varying $\mu$ $(0 \leq \mu \leq 1)$ one places greater or less emphasis on the larger or smaller differences. These are equivalent to the Minkowski metrics used by Kruskal (1964) to measure the average magnitude of the differences. Work such as that of Farris (1969) is needed to study the consequences of using these new coefficients.

As shown by Rubin (1966), Friedman and Rubin (1966), and Hartigan (1967), if one can define a measure of optimality, one can attempt to find the "optimal" classification by trial and error even if a direct algorithm cannot be devised.

It would seem desirable to use measures such as the cophenetic correlation or $\Delta\mu$ which can be applied to the results of many types of analyses, so that one can decide which method gives the best fit to a given set of data. For example, using the cophenetic correlation coefficient as a measure of fit, a principal components analysis using only the first three components gives a better fit to data in which there is little tendency for OTU’s to occur in clusters (e.g., Rohlf, 1967), whereas a cluster analysis can give a better fit if distinct, tight clusters actually exist (e.g., Rohlf, 1968). Since different methods of cluster analysis can yield rather different results when applied to the same data, an objective procedure is to select the method which provides the best fit (if the cophenetic correlation coefficient is used, then one is searching for the least squares best fitting summary of the phenetic relationships).

Principal components analysis is a useful complement to cluster and network analyses. A two- or three-dimensional scatter diagram shows the longer distances fairly well but it distorts the smaller
scaling tends to evenly distribute the distortions among both the large and small distances (see Fig. 9 below).

**Hierarchical Clustering Schemes (HCS)**

In the present study, we shall consider only the class of clustering algorithms which can be characterized by being sequential and agglomerative.

A **sequential** clustering process forms clusters in a regular stepwise manner rather than “simultaneously” as in procedures such as those proposed by Rohlf (1966) and Hartigan (1967) which attempt to optimize the entire tree. Sequential procedures are usually much faster than simultaneous ones. This is important if large numbers of OTU’s are to be considered.

*Agglomerative* clustering procedures begin with pairs of similar OTU’s and build up clusters by contrast with divisive methods which start with the entire set of OTU’s and divide it into subsets and these further into sub-subsets, etc. (see for example the method of Williams and Lambert, 1959). Gower (1967) points out some problems which arise in divisive-type cluster analyses.

Examples of cluster analytic techniques that satisfy the above restrictions are single linkage, complete linkage, weighted and unweighted pair group methods (see Sokal and Sneath, 1963; and Sokal, 1966).

The computational algorithm for all of these methods can be outlined as follows (the only differences between the methods being in the details of step 3 below):

1. Read in the matrix to be analyzed [in the account given below we will assume, for notational simplicity, that an average distance matrix (Sokal, 1961) is used].
2. Find all reciprocally closest pairs of OTU’s (i.e., OTU’s $I$ and $J$ are mutually closest if the smallest $d$ in column $I$ is in row $J$ and the smallest $d$ in column $J$ is in row $I$). If the matrix is symmetric at least one such pair must exist. In the case of ties an arbitrary tie-breaking rule must be used.
3. For each of these pairs of OTU’s add a new row and column to the matrix. The
coefficients, which give the distance between each of these new clusters and the previous OTU’s are then computed. For example, if column \( t + 1 \) corresponds to the cluster of OTU’s \( I \) and \( J \), then

\[
d_{k, t+1} = d_{k, I + J} = f(d_{kI}, d_{kJ}; d_{IJ})
\]

for \( k = 1, t \) (\( \neq I, J \)). Also distances of the type

\[
d_{t+1, t+2} = d_{I + J, L + M} = f(d_{IL}, d_{IM}, d_{JL}, d_{JM}; d_{LM}).
\]

Column \( t + 2 \) corresponds to a cluster consisting of the pair of OTU’s \( L \) and \( M \). A variety of functions, \( f(\ldots; \ldots) \), have been used (such as the minimum, maximum, or average of the distances listed before the first semicolon and correspond to what are called single linkage, complete linkage, and average linkage cluster analysis, respectively, by Sokal and Sneath, 1963). Proctor (1966) and Gower (1967) have proposed functions (see below) which result in a “centroid” clustering. Lance and Williams (1966) also give several such functions.

4. Rows and columns corresponding to the OTU’s which have clustered are now deleted. The size of the matrix is reduced by the number of pairs found in step 2.

5. Steps 2 through 4 are repeated until there is only a single row and column left in the matrix.

After steps 3 and 4 have been taken for the first time, some of the columns in the matrix correspond to single OTU’s while other columns may correspond to a cluster containing a variable number of OTU’s. This makes it possible to make the new distances computed in the second and subsequent passes through step 3 a function of the number of OTU’s within each cluster as well. For example, in the unweighted pair group method using arithmetic averages (UPGMA) one calculates a weighted average distance:

\[
d_{k; I + J} = (n_I d_{kI} + n_J d_{kJ})/(n_I + n_J).
\]

In centroid clustering the following functions are used:

\[
d_{I, J, K} = f(d_{KI}, d_{KJ}; d_{IJ}) = n_I d_{KI} + n_J d_{KJ} - n_I n_J d_{IJ}/(n_I + n_J)
\]

\[
d_{I, J, L + M} = f(d_{IL}, d_{IJ}, d_{IM}, d_{JM}; d_{IJ}; d_{LM}) = (n_I n_L d_{IL} + n_J n_L d_{IJ} + n_I n_M d_{IM} + n_J n_M d_{JM})/(n_I + n_J)(n_L + n_M)
\]

\[
- n_I n_J d_{IJ}/(n_I + n_J)^2
\]

\[
- n_L n_M d_{LM}/(n_L + n_M)^2,
\]

where \( n_I, n_J, \) etc., are the number of OTU’s in cluster \( I, J, \) etc.

It is possible to diagram the results of a cluster analysis in the form of a tree-like structure called a “phenogram” by Camin and Sokal (1965). An example is given in Fig. 2. A computer program “TAXON” (available from the author) has been prepared which performs the analysis described above with built-in options for a variety of functions and special provisions for large matrices which cannot fit in the main core storage. The program also depicts a phenogram.

A simpler program is available in Bonham-Carter (1967) which also has provision for drawing a phenogram on an automatic plotter.

**Analysis of the Results of a HCS**

The statistical reliability of a phenogram is not known. However, the effectiveness (as measured by the cophenetic correlation coefficient) of the phenogram in summarizing the original distance matrix indicates that, of the various functions (to be used in step 3 of the HCS) which have been proposed so far, a weighted average has always yielded the best fit (Sokal and Rohlf, 1962, Farris, 1969). This procedure has been called the “unweighted pair group method using arithmetic averages” or UPGMA (Sokal and Sneath, 1963). The question arises, however, whether a simultaneous clustering procedure might produce a phenogram with less distortion. Rohlf (1966) and Hartigan (1967) have shown this to be possible. It would not be computationally practical to evaluate the criterion
for all possible phenograms, nor does it seem possible to develop an algorithm for finding the "optimum" least squares best-fitting phenogram. Therefore a trial-and-error approach must be taken in which an initial phenogram is adjusted by moving OTU's from one branch to another branch of the phenogram in an attempt to decrease the measure of distortion. The existing algorithms require large amounts of computer time. Thus this approach is not yet practical for large scale use. Farris (1969) gives an algorithm which yields phenograms with higher cophenetic correlations than those given by UPGMA but his technique permits reversals.

When comparing different sets of data using the same clustering procedure, one may also use these measures of distortion as an index to measure the extent to which the phenetic relationships are hierarchic (Rohlf and Fisher, 1968). If, for example, the cophenetic correlation is very high, say above 0.95, one may feel satisfied that imposing a system of nested clusters has not caused undue distortion. If the correlation is much lower, say about 0.6 or 0.7, one may well question the assumption that one has a system of nested clusters. In such cases it has been found very useful to investigate whether the lack of fit pertains to the phenogram as a whole, or if, perhaps, there is distortion only in certain regions.

A helpful technique is to plot frequency distributions of the distance coefficients on each of the branches of the phenogram (Fig. 3). Since each branch unites two clusters, the coefficients in the frequency distribution are the distances between all of the OTU's in one branch and the OTU's in the other.

A detailed study of these "residuals" indicated that the deviations are often not "random." Typically most of these extreme residual distances involved only one or two of the OTU's in one cluster and most of the OTU's in another cluster. Further study revealed that this was caused by an OTU, say I (a member of a fairly tight cluster) being relatively close to an OTU, J. OTU J, however, was not particularly close to the OTU's to which OTU I was close. This indicated, geometrically, that OTU J was on the "outer edge" of a cluster. If the cluster
were elongated into a hyper-ellipsoid, the effect would be even stronger. For example, compare the placement of OTU's 8 and 14 in Fig. 4 (a bivariate scatter diagram of artificial data) with their placement in the phenogram in Fig. 2 constructed from the same data (Fig. 3 is also of this same data). R-type principal components analysis with computation of the projections of OTU's onto the principal axes and the construction of three-dimensional models enables a visual confirmation of this type of relationship in actual multivariate data (Rohlf, 1967, 1968).

**Fig. 4.**—Example of effect of varying the value of the artificial variance in the linear AHCS discussed in text. a. Two-way scatter diagram of artificial data. b, c, and d. Results of clustering with variance equal to 100, 1.0, and 0.01, respectively. Lines indicate points included in clusters at each successive step in the clustering process. The exact shape of the lines has no meaning.

**ADAPTIVE HIERARCHICAL CLUSTERING SCHEMES (AHCS)**

One can envision many different ways in which the basic HCS could be extended to allow for clusters which are ball shaped, elongated, or, perhaps, even curved in various ways. One could try reclustering the same data using a variety of techniques in order to see which method gives the best results. In practice one would probably find, however, that one method worked well for a few of the clusters but another method worked better for some of the other clusters in the same set of data. It would
seem desirable to develop a method sensitive to trends in the shape of the clusters as they are formed, automatically adjusting the algorithm to be more efficient for the types of clusters which seem to be appearing. It would also seem to be important that the clustering procedure not only indicate the cluster to which each OTU belongs but also yield information about the structure of the cluster itself.

The paper by Carmichael, George, and Julius (1968) is one attempt in this direction: their algorithm employs different criteria for deciding when one should stop adding OTU’s to a cluster. Their program points out various statistics including flags indicating the reason why clustering was stopped at a specified level for a particular cluster. This yields some indication as to the nature of the cluster.

I propose to use a measure of distance between clusters that takes into account current knowledge of the size and shape trends found within each cluster. By this method, as successive OTU’s added to the cluster slightly change its shape, the equation used to compute the distance between a cluster and all new candidates for inclusion will be adjusted successively so that new OTU’s consistent with the extrapolated trends in the cluster are considered to be closer to the cluster than OTU’s which depart appreciably from this trend. The ultimate success of this procedure is obviously limited by the extent to which OTU’s are actually grouped into distinct clusters which follow relatively simple and consistent shape trends. A cluster with a very complex shape (for example, a helix) represented by only a few OTU’s and with a large amount of random noise superimposed will be very difficult to resolve and recognize for what it is. If such points are relatively distant from other clusters, then clusters can be recognized but their underlying shapes will not be apparent.

The procedure developed below has the limitation that one must specify beforehand the class of cluster shapes considered permissible. Within that class the algorithm can automatically select the proper shape. I shall first describe the logic of the algorithm in terms of clustering hyperellipsoidal clusters and extend the procedure in the following section to clusters of more general shapes.

**LINEAR AHCS**

This allows for the possibility that at least some of the clusters may have the form of hyperellipsoids, rather than be limited to hyperspheroid.

The basic modification required is that the functions used in step 3 of the algorithm given above must be changed so that if (at a given state in the clustering procedure) there is evidence that the cluster is elongated, OTU’s close to the ends of an elongated cluster should be considered “closer” than OTU’s at the same euclidian distance from the mean of the cluster but along the sides of the cluster. Single linkage cluster analysis tends to do this but it does not take into consideration the fact that the density of points at the ends of the clusters is usually quite low, so that the inter-point distances are greater there.

Let us consider a simple example for the case in which one has only two variables. If the two variables are uncorrelated within the cluster (Fig. 5a), then two points such as a₁ and a₂ will be considered equally distant from the mean of the cluster. However, if the two variables are correlated within a cluster (see Figs. 5b and 5c), then points such as b₁ and c₁ should be considered to be closer to the centroid of their clusters than points b₂ and c₂. The concentric ellipses represent the loci of points which are equally distant from the centroids of the clusters. The higher the correlation between the variables, the more marked will be the discrepancy between ordinary euclidian distances between points and the centroid of their cluster and the distance function employed in this algorithm.

The equation for the distance function employed in this study is as follows:

\[ D_{i,j} = \sqrt{(\bar{Y}_j - \bar{Y}_i)^T S^{-1}_T (\bar{Y}_j - \bar{Y}_i) / S_T} \]
and $J$, for the $p$ variables (the superscript "$t$" indicates matrix transportation).

This distance function is very similar to Mahalanobis’s generalized distance, $D$ (see Rao, 1952; Seal, 1964; or Anderson, 1958, for accounts), but it differs from it in two important ways. First, the within group V-CV matrix of just one group is used rather than a pooled V-CV matrix (the reason for not pooling is that we have no reason to expect the different V-CV matrices to be homogeneous). Secondly, we have included the generalized variance $|S_I|$ as a factor. This was found necessary since we wished to take the shape of a cluster into consideration and not its size or hypervolume. If this is not done then one finds that all points are close to large clusters and distant from small ones.

It is important that the properties of $D_{i\rightarrow j}$ be understood so that one will know what to expect of this type of cluster analysis. The data plotted in Fig. 6a have the following V-CV matrix:

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} .0012388 & .0003788 \\ .0003788 & .0002936 \end{bmatrix}.$$  

The eigenvalues are $\lambda_1 = .0013718$ and $\lambda_2 = .0001606$, which are the diagonal elements of the diagonal matrix $\Lambda$. The matrix of column eigenvectors (normalized) is

$$F = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} = \begin{bmatrix} 0.9435 & -0.3314 \\ 0.3314 & 0.9435 \end{bmatrix}.$$  

The first eigenvector (corresponding to largest eigenvalue) points in the direction of the principal axis of the equal frequency ellipsoid which could be fitted to the data (see Sokal and Rohlf, 1969, for an example of the computation for the bivariate case). The equation of the line representing the major trend of variation is given by the equation

$$C_{12}Y_{1i} + C_{22}Y_{2i} = K_i$$  

constructed using the eigenvector corresponding to the smallest eigenvalue. Note that if there is relatively little scatter of points (deviations) from the principal axis,
the values of $K_t$ (which would be obtained if the $t$ data points were substituted into this equation) will almost be constant. If we take $K$ as the mean value which we would obtain, this gives us the equation of the best fitting trend line, i.e., the line with a minimum squared deviations from it (Fig. 6b):

$$0.3314 Y_{1t} + (-0.9435) Y_{2t} = -0.0330.$$  

Analogously with $p$ greater than 2, the vector corresponding to the smallest root gives us the equation of the best fitting line. If the last two eigenvectors are very small, relative to the rest, then the last two vectors yield the equation of the best fitting plane, etc.

This approach appears contrary to that usually taken in a principal components analysis where one usually computes only those eigenvectors corresponding to the largest eigenvalues. In the present case, one is interested in just the opposite, since the inverse of the within group variance-covariance matrix is employed in the equation for $D_{1-t}$. As pointed out by Sebestyén (1962) the eigenvectors corresponding to the smallest eigenvalues expresses the "property" most invariant within the cluster.
Since \[ S = FA^{-1}F^t, \]
\[ S^{-1} = FA^{-1}F^t. \]

Thus it is possible to write:

\[ D_{L \rightarrow J}^2 = (Y_J - Y_J)^t FA^{-1} F^t (Y_J - Y_J) v, \]

where \( v \) equals the generalized variance (a constant in the present context). If \( P = F^t Y \), the projection of the OTU's onto the eigenvectors, then

\[ D_{L \rightarrow J}^2 = v \sum_{i=1}^p \lambda_i^{-1} (P_{ij} - P_{ii})^2, \]

where the index \( i \) refers to the \( p \) eigenvectors. Thus, the squared differences in projections of each OTU onto each axis is weighted according to the reciprocal of the eigenvalue corresponding to that axis. The largest weight is given to the eigenvector with the smallest \( \lambda \).

This weighting procedure has some similarities to that used by Pearson (1926) for the coefficient of racial likeness (which ignores the correlations among the characters) and to the weighting scheme of Farris (1966) (which ignores the correlations among the characters and also implies that the within group variation is the same in all groups).

These equations can, perhaps, be better understood if we re-express them in terms of their corresponding geometric operations.

Fig. 6a shows some artificial data representing a single cluster for which two variables \( Y_1 \) and \( Y_2 \) are correlated within the cluster (they may be uncorrelated in other clusters). In Fig. 6b are shown the major and minor axes for the 90% equal frequency ellipsoid fitted to this data. In Fig. 6c, the data have been rotated and translated so the origin of the graph is now at the mean and the major and minor axes are used as the coordinate axes (a result of the operation \( P = F^t Y \)). In Fig. 6d, the coordinate axes have been rescaled by multiplying them by \( \lambda^{-1} \). In this new space, the distance of a point at coordinates \((A, B)\) to the mean \((\bar{Y}_1, \bar{Y}_2)\) can be computed using the standard Pythagorean equation. Fig. 6e shows contours of points which have an equal distance from the mean of the cluster. In Fig. 6f both axes have been drawn to the same scale, causing the contours to become circular in outline. Consequently, in Fig. 6f, ordinary euclidian distances are equal to the generalized distances of Fig. 6e. Thus, OTU's near the ends of an elongated cluster would in fact be nearer to the center of the cluster in the conventional sense and would be added to a cluster in preference to OTU's located along the sides of the cluster.

If this distance function is used in step (3) of the algorithm, step (2) must also be modified, since there may not be any reciprocally close pairs of OTU's. This is due to the fact that in general \( D_{L \rightarrow J} \neq D_{J \rightarrow L} \). When such cases arise, we can simply find the smallest \( D_{L \rightarrow J} \) value and consider the two OTU's involved as a mutually close pair.

Even when reciprocally close pairs are found, other OTU's may "interfere" in the sense that if OTU's \( I \) and \( J \) are mutually closest \( D_{K \rightarrow L} \) may be less than either \( D_{L \rightarrow J} \) or \( D_{J \rightarrow L} \). In such cases, OTU's \( I \) and \( J \) are not treated as a mutually closest pair of OTU's.

Fig. 7 shows the results of this algorithm applied to a set of simple data. The data represent seven species of limpets (marine gastropods) plotted with respect to two variables describing shell shape (ratio of the height of the shell to its length and the ratio of the distance between the apex of the shell, see diagram in Fig. 7a). With only two variables used one should not place much taxonomic importance on the clusters obtained, but the clusters do seem reasonable for the given set of data.

It should be pointed out that this type of clustering is not affected by linear transformations of the original data. Thus, standardization of the characters has no effect and need not be performed (as is also the case in the techniques proposed by Friedman and Rubin, 1966). The effect of the size of the organisms is also minimized if one has samples of both large and small specimens within each OTU (thus, the
be most efficient to carry out a principal components analysis on the among OTU’s V-CV matrix and use only the first few, $k$, components as variables in order to lighten the computational load. The V-CV matrices for each OTU must also be transformed so that they are compatible with these principal components (i.e., they are transformed using the same eigenvectors, new $S = F^T S F$). If there is no information about the covariation of characters within the OTU’s, as when individual specimens are used as OTU’s, one can do one of two things: (1) One can enter the OTU’s into the analysis and cluster on the basis of ordinary euclidean distances until a cluster has enough OTU’s in it so that its variance-covariance matrix is of full rank ($n > p$). A problem with this is that one might run out of OTU’s before this happens. This could be avoided by first performing a principal components analysis, using only the first few components as variables. (2) Another procedure would be to generate artificial V-CV matrices for each OTU. In the experiments carried out so far, we have used V-CV matrices which imply that the concentration ellipsoids around each OTU are actually hyperspheres. The hypervolume of these hyperspheres was left as a parameter which could be adjusted. The most desirable value is not known. If the volume is relatively large, then the clustering will differ little from centroid clustering described before and hyperspherical clusters will tend to be formed. If, however, the value is very small, then, once a cluster composed of two or more OTU’s is formed, the method will be very sensitive to the shapes of the clusters. The precise value to be used must depend upon several factors: 1, The scale of measurements used for each variable (thus standardization is recommended); 2, The amount by which the OTU’s deviate from tightly packed spheroidal or ellipsoidal clusters (since this need not be the same for all clusters no one value may be optimal for all of the clusters); and 3, The true dimensionality of the phenetic space.

Fig. 7.—Two-way scatter diagram showing the means of 7 species of limpets and 90% equal frequency ellipses fitted to the individual observations for each group. Insert (left side view of a limpet) defines measurements used in the ratios for the abscissa and ordinate of the figure. b. Results of clustering using the linear AHCS. Details are discussed in text.

technique proposed by Burnaby, 1966, is not necessary).

For most numerical taxonomic studies, this procedure would be prohibitive since it would require at least $n + 1$ individuals to be measured on all $p$ characters for each of the $t$ OTU’s, since $2t - 1$ principal components analyses would need to be performed. Computer time would also be a problem if $p$ (the number of characters) were large.

Several shortcuts are possible. It would
Experience so far has not yet enabled the development of optimal values for the variances in the artificial variance-covariance matrices. All one can do at present is to try several values and observe the resulting phenograms and cophenetic correlations. However, it is now less obvious how the cophenetic values should be computed. If the method of Sokal and Rohlf (1962) is used, the cophenetic correlations steadily decrease as the variances decrease even though the phenograms seem subjectively to improve. For example, the phenograms in Figs. 3a, 4b, 4c, and 4d have cophenetic correlations equal to 0.781, 0.733, 0.723, and 0.643, respectively.

There are two ways in which the cophenetic correlation can be raised without changing the branching patterns of the phenogram. First, one can adjust the level of each branch so that it is at a level corresponding to the average of the distances between the OTU's in the two clusters being united (rather than the logarithm of the hypervolume of the resulting cluster). When this is done to yield Fig. 3b we note that such an adjustment can yield reversals which must be eliminated by changing the two bifurcations to a trifurcation at a level corresponding to their weighted average. This adjustment raised the cophenetic correlation for the phenogram in Fig. 3b from 0.656 to 0.689. The small change reflects the fact that the logarithm of the hypervolume is not a bad scale to start with. Secondly, the cophenetic correlation may be raised by adding additional existing information to the phenogram. For example, one could somehow show that three of the clusters were elongated and oriented in particular directions. If this were done the cophenetic values would have to be computed taking this into consideration and the resulting cophenetic correlation for Fig. 3b would undoubtedly be higher. Algorithms and computer programs are being developed to perform such computations.

Fig. 8 shows an example of this method applied to the mosquito pupae data of Rohlf (1967) (see also Fig. 1). a. Results of non-metric multidimensional scaling for two dimensions. Stress = 0.170, cophenetic correlation equals 0.967. Species code numbers and generic groupings are also indicated: 1–20 Aedes; 21–25 Culex; 26, 27, and 33–38 Culiceta; 28 and 29 Orthopodomyia; 30 Mansonia; 31 Uranotaenia; 32 Wyeomyia; and 39–45 Anopheles. Results of UPGMA cluster analysis. Cophenetic correlation is 0.913. Results of AHCS (presented in text) with artificial variance equal to 0.001 and using projections of the 45 species onto a 9-dimensional space as input data. Cophenetic correlation is 0.643.

Rohlf (1967). The projections of the 45 species onto the first nine principal components were used as input. The spacing of
Fig. 9.—Comparisons of amount and patterns of distortion introduced by various methods of summarization of phenetic relationships. On all figures the ordinates are the average taxonomic distance between all pairs of species and the abscissas are the estimated distances or cophenetic values from various methods. The following correlation coefficients refer to cophenetic correlation coefficients. a. UPGMA, $r = 0.913$. b. AHCS as is Fig. 8c, $r = 0.643$. c. Principal components analysis with projections based on only two dimensions, $r = 0.938$. d. Non-metric multidimensional scaling analysis in two dimensions, $r = 0.967$.

points in Fig. 8 was determined by a multidimensional scaling analysis (cf. Fig. 5 in Rohlf, 1967). Fig. 8a gives the species code numbers and the generic groupings. Fig. 8b indicates the results of a UPGMA cluster analysis applied to the original distance matrix. Similar results were obtained using AHCS with a very large variance. Note the tendency for “round” clusters to be formed. Fig. 8c shows the results of AHCS with a variance of 0.001. The results are very good by the standards of conventional taxonomy. All of the species of *Aedes*, *Culex*, and *Orthopodomyia* form their own clusters. There are only two “misplaced” species. Species 30 joins 45 before it clusters with the rest of *Anopheles* and species 26 is placed with the *Culex*.

Fig. 9 compares the amounts and patterns of distortion introduced to the pupae data by these various techniques. The original distances are plotted along the ordinates of
the figures and estimated distances (co-
phenetic values) are plotted along the ab-
scissas.

**NONLINEAR ADAPTIVE HCS**

If one suspects that the clusters to be
found in a given set of data are neither
hyperspheroidal nor elongated into hyper-
ellipsoid but, rather, that they may take the
form of clouds of points around some cur-
vilinear function (see Fig. 10 for examples),
then the techniques in the preceding sec-
tion will not adequately isolate and de-
scribe the clusters. There are (at least) two
main approaches that we may take. First,
we could in some way determine the re-
gression function (or at least a numerical
approximation; Sneath, 1966) which fits
the shape of the cluster and then compute
the distance between an OTU and the curve
representing the cluster. Sebestyen (1962,
section 5.1) points out the problems in-
volved with this method. The second pro-
cEDURE would be a simple extension of the
method given in the preceding section. We
would wish to compute the distance be-
tween an OTU and a cluster using a dis-
tance function which takes into considera-
tion the apparent shape of the cluster, as
well as the extent to which the OTU's fit the
underlying curve (if it exists).

This latter procedure can be easily ac-
complished by simply adding new dummy
variables to the original data matrices. The
precise nature of the new dummy variables
will depend upon the class of underlying
curves that one believes might be appro-
priate. For example, if we wish to allow
the possibility of points being clustered
around points (hyperspheroidal clusters),
lines (hyperellipsoidal clusters), parabolas,
circles (ring-shaped clusters), ellipses, or
hyperbolas, then we would simply add the
squares of the original variables as well as
their cross-products of the variables taken

<-

**Fig. 10.**—Examples of AHCS applied to non-
linear clusters. a. Artificial data with three clusters
in the shape of a parabola, an ellipse, and a circle.
b and c. Results of clustering, using artificial var-
iances equalling 0.001 and 0.0001, respectively.
two at a time. Thus, if we have only two variables, say \( Y_1 \) and \( Y_2 \) we would now have five variables (in general the total number of variables would be \( p + p(p+1)/2 \)), \( Y_1, Y_2, Y_1^2, Y_2^2, Y_1Y_2 \). One could then input into the computer program, described above, the means for each OTU on these variables and the variance-covariance matrices for each group for these variables. In order to avoid very large numbers it is desirable to standardize the original variables before computing their powers.

If, at a given stage in the clustering process, the data points describe a ring, the generalized distance between the OTU’s and this cluster would be small for OTU’s close to this ring and would be relatively large for OTU’s away from the ring. It should be pointed out that OTU’s near the mean of the cluster (in the center of the ring) would be considered relatively far away from the cluster, as one would wish for such a procedure. In general, OTU’s along the direction of the main trend of the cluster would be considered relatively close and OTU’s not in the direction of the general trend of the cluster would be considered distant. The amount of within-group variation, i.e., the extent to which the points closely follow a trend, affects the degree to which the distance function reflects the trends. If there is much scatter and the data only vaguely follow a parabola, for example, then the distance will differ little from what would be obtained if the nonlinear dummy variables had not been included. Since the properties of the nonlinear generalized distances employed in this method are not well known, an account is given below which describes their more important properties.

The mathematical basis for this technique is a simple generalization of the procedure described in the preceding section (Sebestyen, 1962). Let us consider an example such as the parabolic shaped cluster extracted from Fig. 10 and displayed in Figure 11a. For purposes of illustration, let us add only the one dummy variable, \( Y_3 = Y_1^2 \), which we know we will need (Figure 11b). The same results would be obtained if we included \( Y_2^2 \) and \( Y_1Y_2 \), but that would require one to visualize a five-dimensional space, which is undesirable for illustrative purposes.

Note that since the ordinate in Figure 11b is \( Y_1^2 \) the data points must all lie on the surface of a parabolic cylinder as shown in the figure. The mean of the sample of points will lie at the centroid of the distribution and not on the surface of the parabolic cylinder.

The mean, \( \hat{Y} \), and the V-CV matrix for the data plotted in Figure 11b are (the \( p \) variables are \( Y_1, Y_2, Y_3 = Y_1^2 \)):

\[
\hat{Y}' = (0.0, 0.0, 0.885314)
\]

\[
S = \begin{bmatrix}
1.000000 & 0.008470 & 0.012959 \\
0.008470 & 1.000000 & 0.774072 \\
0.012959 & 0.774072 & -0.614745 \\
\end{bmatrix}
\]

The eigenvalues of \( S \) are \( \lambda_1 = 1.60541, \lambda_2 = 0.9967, \) and \( \lambda_3 = 0.00967 \). The matrix of column eigenvectors (normalized) is:

\[
F = \begin{bmatrix}
-0.024199 & 0.995462 & -0.005044 \\
-0.787594 & -0.022171 & -0.615802 \\
-0.615718 & -0.019029 & 0.787900 \\
\end{bmatrix}
\]

<FIG. 11.—Diagram showing geometric interpretation of generalized distances computed using non-linear data. a. Artificial data from Fig. 10a. b. Data of a with a new dummy variable \( Y_3^2 \) added so that the points are now plotted in a three-dimensional space. Since the ordinate is \( Y_3^2 \) all points must lie in a parabolic cylinder. The cylinder is cut by the plane determined by the first two principal components of the variance-covariance matrix among the three variables. c. Parabolic cylinder as in b cut by a series of parallel planes. The intersections of these planes with the cylinder are shown as a family of parabolas at the top of the figure. These parabolas represent one set of curvilinear coordinate axes as explained in the text. d. Artificial data in a with the curvilinear coordinate axes generated by principal components I, II, and III. e. Curvilinear coordinate axes as in Fig. d with scales weighted according to the reciprocal of the eigenvalues of the principal components. f. Loci of points of distance \( .5, 1, 2, \) and \( 3 \) from the mean of the cluster. This figure is analogous to Fig. e.>
As before, the eigenvectors point in directions of maximum variance. In a cartesian space one can interpret the coordinate of a point \((a, b)\) on the \(Y_1\) axis as being the value of \(K\) which makes the line \(Y_1 = K\) intersect with the point \((a, b)\). \(K\) obviously equals \(a\) in this simple case. Analogously, the coordinate of a point \((Y_1, Y_2, Y_3)\) on axis \(I\), for example, is the value of \(P_I\) which satisfies the equation: 
\[-0.024199 Y_1 - 0.787594 Y_2 - 0.615718 Y_3 = P_I\]  
(\(P_I\) the coefficients are taken from the first eigenvector). The locus of all points which yield a particular value of \(P_I\), say zero, is a plane cutting through the 3-dimensional space as shown in Fig. 11c. If we take into consideration the fact that \(Y_3 = Y_1^2\) and consider only points on the parabolic cylinder we note that the projection of the intersection of this plane and the cylinder onto the \(Y_1, Y_2\) plane is a parabola. By varying the value of \(P_I\) we get a family of parabolas. All points on the same parabolar have equal projections onto axis \(I\). In a similar manner (using the second and third eigenvectors) we can have planes for axes \(II\) and \(III\) as shown in Fig. 11c. The projections of these intersections are a series of curves almost parallel to the \(Y_2\) axis for axis \(II\) and a series of parabolas for axis \(III\).

Fig. 11d shows these three families of curves simultaneously. This figure is analogous to Fig. 6c in that we can now express the coordinate of each point on the cylinder in terms of its projection (coordinate) on the three curvilinear coordinate axes corresponding to the three eigenvectors. For example, the curvilinear coordinates of the second point \((Y_1 = -1.07557, Y_2 = 0.28687)\) is \((P_I = -0.9122, P_{II} = -1.0942, P_{III} = 0.7402)\) computed by multiplication by the eigenvectors as before \((P = FY)\). These coordinates can be confirmed by an examination of Fig. 11d.

A general discussion of curvilinear coordinate systems is given in Hildebrand (1952).

The smaller eigenvalue and its associated eigenvector indicate the general trend of the data and the closeness of the points to this curve. In the present example we see that the best fitting curve is the parabola 
\[ -0.005 Y_1 - 0.616 Y_2 + 0.788 Y_1^2 = 0.700\]  
which can be rewritten as
\[0.616 \left( Y_2 + \frac{0.700}{0.616} + \frac{(0.005)^2}{4(0.788)(0.616)} \right) = 0.788 \left( Y_1 - \frac{0.005}{0.788} \right)^2,\]

or simply \(Y_2^2 = (0.788/0.616)(Y_1^2)\) after a simple translation of the axes.

If we had included the additional dummy variables \(Y_2^2\) and \(Y_1 Y_2\), the trend equation would be of the general form
\[AY_1^2 + BY_1 Y_2 + CY_2^2 + DY_1 + EY_2 = \bar{P}\]

By a rotation of the axes one can eliminate \(Y_1 Y_2\) term. By completing the square and translating the rotated axes one can arrive at a simplified equation which makes the form of the curve more obvious (see standard analytical geometry texts for details of the methods for carrying out these operations). If a miscellaneous collection of dummy variables are employed simplification of the trend equation may be more difficult. The magnitude of the smallest eigenvalue measures the fit of the data to this trend curve. It is the variance of the deviations of the \(n\) data points measured perpendicularly from the curve.

Analogously to Figure 6d, we can construct Figure 11e in which the coordinate axes have been weighted by \(\lambda^4\). The
shapes of the axes are unaffected by this operation, only the scales are changed. We cannot measure distance directly in this space. As above, we must compute distances using the standard formula:

\[ D_{k,l}^2 = \sum_{i=1}^{p} \lambda_i^{-1} (P_{ki} - P_{li})^2, \]

where \( P_{ki} \) are the means projections of the sample points onto the curvilinear axes, \( P_{li} \) are the curvilinear coordinates of a point in the space depicted in Fig. 11b, and the \( \lambda_i \) are the eigenvalues.

Contours of points on the parabolic cylinder, which are equidistant from the sample mean, are shown in Fig. 11f. One will note that, as in Fig. 6e, points which deviate from the mean in the direction of the principal trend of the data (in the present case, parabolic) are indicated to be relatively closer to the mean than are points which deviate in other directions.

It should be pointed out that, unless the points have identical values of \( Y_1 \), the mean of the sample will not lie in the surface of the parabolic cylinder shown in Fig. 11b. Thus the mean cannot be plotted in Figs. 11d or 11e and a distance equal to zero cannot be shown in Fig. 11f (of course, the mean of another sample could be identical to the present sample mean, the distance between these two means would be correctly indicated to be zero). It is difficult (but not impossible) to construct a plot analogous to Figure 6f. Considering only data points we can study \( D_{k,l}^2 \) as a function of only two variables \( Y_1 \) and \( Y_2 \).

Fig. 12 shows the distance functions we have considered as surfaces in three dimensions. Figure 12a represents the ordinary cartesian distance, the height (\( \bar{Z} \)) of the surface at each point (\( Y_1, Y_2 \)) is equal to the distance of the point (\( Y_1, Y_2 \)) from the mean of the sample at (\( \bar{Y}_1, \bar{Y}_2 \)). Figs. 12b, c, and d are similar plots where the cluster is an ellipsoid, parabola, or a ring, respectively. Figs. 6e and 11f represent some of the contours of Figs. 12b and c respectively. Fig. 12d clearly shows that points at the centroid of a ring shaped cluster are further away from the cluster than are points along the ring itself.

Non-linear AHCS is not limited to clusters which take on the shapes of conic sections (as is the case in the simple example described above). By introducing the appropriate dummy variables a variety of cluster shapes can be handled. The principal limitations of this approach is that one must decide beforehand which class of functions are appropriate for a given set of data. But note that the amount of computation necessary increases very rapidly as additional dummy variables are added.

Figs. 10b and c show the results of the proposed clustering procedure (with dummy variables \( Y_1^2, Y_2^2 \) and \( Y_1Y_2 \) added) for two values of the diagonal elements of the artificial within group V-CV matrices (see above). A small value clearly results in isolating the clusters properly. An examination of the eigenvectors and eigenvalues at each step in the clustering also correctly indicate the shape of each cluster. In contrast to earlier examples, very small variances still resulted in essentially the same clusters.

As in the case of other clustering procedures, AHCS's success is limited by the closeness and compactness of clusters, but this technique has the ability to follow clear nonlinear trends. AHCS also has the advantage that it yields the equation of the best fitting curve through the cluster as well as its degree of fit (in contrast to the techniques described by Sneath, 1966; Carmichael et al., 1968; and Ball, 1965).

As was the case for the linear HCS the cophenetic correlations (using the conventional cophenetic values) were rather low (0.691 and 0.573 for Figs. 10b and c). A general technique for computing cophenetic values for nonlinear clusters is more complex and is being developed.

It is also quite clear that the concept of weighted curvilinear coordinates can be applied to other related statistical methods. For example, Sebestyen (1962) developed nonlinear generalized discriminant func-
tions (his examples were in terms of discrimination between classes made up of disjoint subclasses). Techniques of nonlinear factor analysis are given by MacDonald (1962, 1967). Gnanadesikan and Wilk (1969) discuss and give an example of nonlinear principal components analysis.

DISCUSSION

In order to be worth interpreting, a phenogram must impose a relatively small amount of distortion of the pattern of phenetic relationships to be found in the original distance matrix. Thus, it is very important that the distortion between the phenogram and the original distance matrix be measured by some stated criterion. Without such a measure one has no basis for interpreting a phenogram seriously. Many indices are possible (see above). If (in an actual study) the cophenetic correlation coefficient is quite high one should be encouraged to try to interpret the clusters which emerge. However, this only sets a limit to when one should start interpreting. As can be seen from the examples presented above, even a phenogram that fits the phenetic relationships in a way that intuitively appears quite poor will yield cophenetic correlation coefficients well above the range of values expected, due to chance alone (Rohlf and Fisher, 1968). Thus, even a coefficient near 0.9 does not guarantee that the phenogram serves as a sufficiently good summary of the phenetic relationships. The situation is somewhat analogous to problems in regression where one cannot tell from the percentage of variation explained by linear regression whether or not the deviations from regression represent simple random scatter or are systematic (indicating that, perhaps, a quadratic regression would have been more appropriate). The only way of determining this is to make some sort of plot where one can visually attempt to see patterns in the residuals; or curvilinear regression may be done to see whether a significantly better fit is obtained.

Likewise in cluster analyses, one could either examine the "residuals" or deviations between the estimated and actual distances (as in Figs. 3 and 9) or try a variety of cluster analyses to see which one gives the highest cophenetic correlation. As Sokal and Rohlf (1962) suggested, it has been found that, of the sequential agglomerative HCS's which use distances in a euclidean space and place similar OTU's together, the UPGMA procedure always yields the highest cophenetic correlation regardless of the structure of the data. Thus, for this class of techniques, the UPGMA procedure is clearly to be preferred. However, we have seen above how poorly UPGMA clusters certain types of data structures. Thus we should not be satisfied with results given by UPGMA, but should investigate whether it is possible to get better degree of fit. It is also important that techniques be developed which enable more information to be included in a phenogram. The frequency distributions on the branches of the phenogram as shown in Figure 3 show information about the relative degree of fit of different parts of the phenogram to the data. In addition, it would be desirable to annotate the phenogram with information concerning the nature of the structure of the clusters. For example, it is important whether the points in a cluster are distributed in the form of an ellipsoid, paraboloid, ring, etc. In the case of clusters with a linear order, it would also seem appropriate that the OTU's be listed in the phenogram in some natural sequence reflecting their position in the cluster (rather than simply in the order in which they are added to the cluster, which means that OTU's toward the extreme ends of an elongated cluster will often be placed side by side in the phenogram).

These additional considerations add much complexity to the results of a cluster analysis, but I believe they are necessary in order to achieve an acceptable level of summarization of one's data, so that one may safely use the results in biology. Some of
the criticisms by conventional taxonomists of the results of numerical taxonomic studies have been justifiable in that some of the taxonomic conclusions which they questioned were artifacts of the clustering procedure—more direct (but more laborious) examination of the original distance matrix indicated relationships more consistent with conventional classifications than with the phenograms (see for example Barr and Chapman, 1964, and Rohlf, 1964).

Comprehension of multivariate relationships is difficult. The best advice that can be given to someone who needs practical results from numerical taxonomy is to try a variety of techniques for summarizing his data. Different methods will expose somewhat different aspects of the phenetic relationships. If in doubt about conflicting results one should not hesitate to directly examine the original data. If a single unique "result," rather than a general understanding, is required then that summary with the lowest measure of distortion should be used.

ACKNOWLEDGMENTS

Contribution number 1427 from the Department of Entomology, The University of Kansas and number 7 from the program in Ecology and Evolution, State University of New York, Stony Brook. This work was supported, in part, by a grant (GB-4927) from the National Science Foundation. Computer time was made available on the GE 635 by the University of Kansas Computation Center.

The paper has benefitted from my discussions with a number of people. Particularly helpful were discussions at the Second Numerical Taxonomy Symposium (Sokal, 1969), the Numerical Taxonomy Discussion Group at The University of Kansas, and with Dr. Joseph B. Kruskal and the Mathematics Group at Bell Telephone Laboratories at Murray Hill, N.J. E. J. Carmone furnished me a copy of Dr. Kruskal's MDSCAL-IV program. Drs. R. R. Sokal and J. S. Farris read the manuscript and made many useful suggestions.

Patricia R. Rohlf assisted in the preparation of the illustrations and in the typing of the manuscript. Dr. A. J. Rowell generously prepared Figs. 11b and 11c.

REFERENCES


Goodall, David W. 1968. Affinity between an
Jackson, R. A. And T. Crovello. 1969. (In preparation.)

Department of Biology, State University of New York, Stony Brook, New York 11790.