CHAPTER 7

The Construction of a Taxonomic System

In this chapter we discuss how a taxonomic system can be constructed from the resemblances among operational taxonomic units in a study. By this we mean the grouping together of those OTU's to form taxa, employing the affinities found by the methods described in Chapter 6. In order to do this we must discuss the requisites of a taxonomic system, the techniques for creating it, and problems of taxonomic rank.

It must be clearly understood that taxonomic systems are inevitably oversimplified representations of the matrix of affinities among the forms studied. One cannot therefore demand perfection of such systems. The principles of their construction have been little studied, and we feel that this part of the subject would generously repay the attention of taxonomists.

7.1. REQUISITES OF A TAXONOMIC SYSTEM

There are many ways in which we can construct taxonomic systems, according to the purpose in mind. We have assumed throughout this study that we wish to make taxonomic systems which are "natural," and we have discussed in Sections 2.2 and 4.5 the conceptual bases for natural systems. There are, however, a number of requirements essential to any practical taxonomy, whether natural or not.

Every taxonomy is built from units of some sort, and in biological taxonomy these are the organisms and the characters which they possess. We have been considering how one classifies the organisms, from their correlations and resemblances (Q-technique), but one can also consider

the correlations among the characters (R-technique); as an adjunct to the usual systematics this may be valuable, particularly in the making of keys and in the study of causal relations in biology (see Section 7.6). In making taxonomic systems we are impelled by one consideration of overwhelming importance: we can neither list nor remember all the characteristics of various organisms and higher taxa, and we therefore need a system of grouping them into a manageable number of groups whose characters are preponderantly constant. Because of high constancy and mutual intercorrelations of characters, such a grouping will carry a high predictive value. Thus, if we read of a new aphid species we can immediately predict a number of characteristics which this species is expected to possess. Being an aphid, it will with almost complete certainty be a plant feeder, possess a particular type of wing venation, be parthenogenetic in part of its life cycle, produce males by nondisjunction of the sex chromosomes, produce honeydew, secrete wax from cornicles or other glands, and so on. Since an aphid is a homopteran, we can forecast with some accuracy the general construction of its mouth parts, the texture of its wings, and other homopteran characteristics. This type of argument can, of course, be extended to the hexapod and arthropod levels of classification and even higher. It is obviously much easier for us to remember this of the group Aphididae than of each individual aphid. Furthermore, it is impossible to remember or appreciate the innumerable relations between the various OTU's to be classified, but this is easier when they are grouped into fewer inclusive taxa. Work is at present going on to see whether there may be some groups of actinomycetes characterized by the production (at least in a proportion of the strains) of certain kinds and classes of antibiotics (see Silvestri et al., 1962; Arai, Kuroda, and Ito, 1962). If the groups are based on well-correlated characters we may hope that type of antibiotic will also be correlated with the groups.

The prime purpose of a taxonomic system is therefore one of economy of memory. This economy is achieved in one of two ways: (1) either we employ the attributes one at a time in order to cluster our taxonomic entities, which gives us a system such as that used in indexing books by the names of their authors or by their size (monothetic systems), or (2) we attempt to cluster them according to all their attributes considered simultaneously, for which we use measures of affinity between the entities. Intermediates between these two may also be employed. The first, or monothetic, method is "artificial" (though by chance or by selection of the right attributes it may happen to be very nearly "natural"), while the second method is "natural" but incomplete, since the matrix of

affinity values is itself too complex to serve the purpose required and must be analyzed so as to cluster the entities into "natural" taxonomic groups. In either case some form of nomenclature is needed for the resultant groups, but we may postpone discussion of this to a later chapter. In either case, too, the number of resultant groups must not be too large, and their properties must be consistent with the principles on which they were set up. When the groups are monothetic, this is simple; we must only obey our own criteria and put white objects into the pigeonhole labeled "white" and black ones into that labeled "black" (though even this elementary point is often overlooked, as Metcalf, 1954, pointed out in an amusing discussion on artificial keys). When the groups are polythetic ones, we must bear in mind that it is never certain, but only more or less probable, that a member possesses any given feature (see Section 2.2).

The most powerful method of achieving economy of memory is the method of the nested hierarchy. This device allows us to group a large number of taxonomic groups into fewer composite groups of higher rank, and it is only when these groupings are mutually exclusive that it gives the best results; for example, a given genus can belong to only one family, and this family to only one order, and so on. If it is found that some attributes are possessed by all members of one group, the task of remembering the attributes of the group is therefore less. The natural group of mammals has many attributes which are not possessed by the natural group of birds. This property of natural hierarchies is presumably due to the evolution of the natural groups, since it need not be always true—for example, in classifying the members of an interbreeding population, where it may not be possible to establish any satisfactory system of mutually exclusive hierarchies. The advantages of hierarchies are so great that we will generally employ them, even when this means we must distort the system of affinities to some extent.

It is not generally realized that hierarchies, other than purely arbitrary ones, can only be made with certain sorts of distributions of affinities between organisms. Figure 7-1(a) shows a random distribution of organisms envisaged as occupying points in a two-dimensional phenetic space, and Figure 7-1(b) shows a uniform distribution. In neither case can hierarchies be made except by arbitrary lines. In order to make hierarchies, one needs a clumped distribution, as shown in Figure 7-1(c), where the dotted lines indicate a nested hierarchy. The random distribution may give some hierarchical groups, since some parts are clustered by chance, but there are too many "intermediate" types for a satisfactory

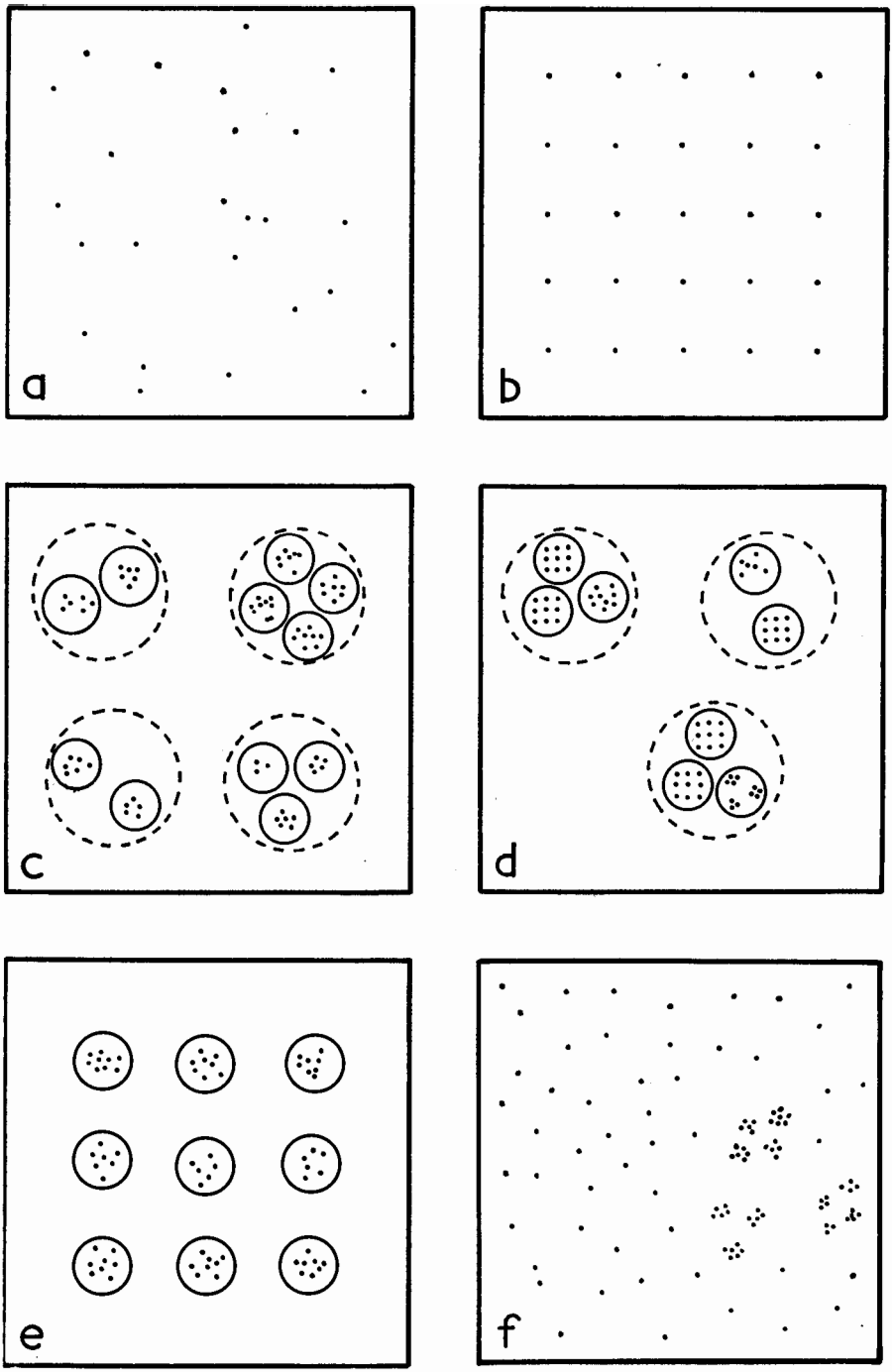

FIGURE 7-1

Different types of clustering (for explanation, see text).

hierarchy. We have no evidence that uniform distributions occur in nature, and random distributions seem likely to occur only at low levels of taxonomic rank, such as among geographical races or panmictic populations (see Section 8.2.3). Note that the clumped distributions must be clumped at each level at which it is wished to make a hierarchical division; whether they are uniform, random, or clumped *within* this level is immaterial for that hierarchical level. Figure 7-1(d) shows a hierarchical arrangement which at the two higher levels is clumped; in Figure 7-1(e) the hierarchy is only possible at the intermediate level, since the circles are regularly spaced. For satisfactory complete hierarchies the taxa must be clustered at every level. Only part of the distribution in Figure 7-1(f) could be arranged hierarchically.

The nature of these distributions should be made clear, for they have received much attention from ecologists; Greig-Smith (1957, pp. 50–84) gives a full discussion of them as applied to two dimensions (see also Thompson, 1956). A random distribution is one in which the probability of occurrence of an entity is independent of the occurrence of other entities in any sampling unit; that is, the presence of an entity at a given point neither raises nor lowers the probability that another entity will occur close to it. A uniform (overdispersed) distribution is one in which there is a decreased probability of another entity occurring close to the first, while in a clumped distribution the probability of this is raised. Although developed for two-dimensional systems, these principles can probably be safely extended to the multidimensional systems of taxonomic affinities.

It is evident that in making clusters of OTU's we place the dividing lines between the groups at places where there are distinct gaps in the combinations of characters we have observed. Above the rank of species (or perhaps of genus) there is no great problem, and gaps are larger and evident on cursory study. So far, there has been no work specifically directed toward measuring such gaps by numerical methods, but there is no reason why they should not be elaborated, if needed, from the numerical techniques described in Chapter 6.

However, when we group a number of quite distinct taxa into a higher taxon, such as grouping genera into families, the presence of the gaps is itself of little help. If there are clear gaps between genera **A**, **B**, **C**, and **D**, this will not tell us whether they should—on grouping them into two families, 1 and 2—be grouped into Family 1 = (**A** + **B** + **C**) and Family 2 = (**D**) or into Family 1 = (**A** + **B**) and Family 2 = (**C** + **D**). The problem is to define the sort of grouping which is to be considered a

natural taxon. This definition may not be easy to state precisely. Clearly the taxon (or taxa) must possess the property of "naturalness" to the highest possible degree. Sneath (1961) has defined natural taxa as "sets composed of all those elements which share x or more features out of y features, where x and y are large numbers, but in which an element may have any combination of features as long as the total number of features shared with every other element of the set is x or more." Note that the qualifying word "all" is inserted in the phrase "sets composed of all those elements." This is necessary, since otherwise one can select a few of the contained entities, and such a selection may not itself be a natural taxon. For example, the selection mice plus whales plus bats is not itself a taxon, but is only an "unnatural" part of the taxon Mammalia, which is composed of all the known kinds of mammals. Also the definition given above yields overlapping taxa, which for convenience are converted into nonoverlapping, hierarchical systems. The study of such systems has not yet been adequately developed, though Woodger (1937, 1951, 1952) and Gregg (1954) have made a start on this.

At a lower level of rank, the problem of division between intergrading groups remains to be discussed. This problem is generally not acute with the higher ranks, except perhaps in microbiology, where there may possibly be large "spectra" of gradually merging forms (see Cowan, 1955). In ecology intergrading groups may be the rule rather than the exception (Goodall, 1953). Discrete, nonoverlapping classes of vegetation or of fauna may be desirable from a practical viewpoint; however, they may have no theoretical validity (Whittaker, 1953). There is no reason why phenetic groups should not be recognized even when all intermediate forms are found between two kinds of creature—for example, where two species hybridize. If the hybrids are in the minority, the division between the two species lies through them, and again this is a problem of cluster analysis. It is analogous to two mountains joined by a saddle: the saddle is the division between them.

One further point requires emphasis: we can only make hierarchical natural groups on the basis of the organisms which are known to us. This is easily seen if one considers a study upon a few forms which appear to fall into two distinct clusters. The two clusters would be two taxa. Yet if it is later found that the forms are connected by a chain of intermediate forms which are more numerous than the extreme forms, we would now have a single cluster which we would have to consider as a single taxon.

To summarize, a taxonomic system should be "natural" in an empirical sense, and thus should be of high predictive value. The system is best

arranged in the form of a nested hierarchy, and dividing lines are to be placed at gaps in the combinations of characters observed.

7.2. FREQUENCY DISTRIBUTIONS OF SIMILARITY COEFFICIENTS

Illuminating insights into the phenetic relationships among the OTU's in a numerical taxonomic study can be obtained by an inspection of the frequency distributions of the similarity coefficients. A frequency distribution of correlation coefficients between pairs of species was published by Michener and Sokal (1957) in their study of the bees of the *Hoplitis* complex. A primary mode among these correlation coefficients was shown at $r = 0.38$. This represented the most frequent class of correlation coefficients found between the species in this study (mostly intergeneric relationships). A secondary mode at $r = 0.78$ indicated relationships among closely related species. Bimodality or multimodality of this sort substantiates the nested arrangement of clusters of the OTU's in a given study. Rohlf and Sokal (1963) observed similar multimodal structure in distributions of correlation and distance coefficients based on standardized and not standardized characters. Standardization of characters generally accentuated the multimodality of the distributions. Even when multimodality was not clearly indicated, observed variances of similarity coefficients were greatly in excess of expectations. A pronounced bimodal distribution (each peak approximating to a normal curve) is apparent for the similarity values for strains of two species of bacteria (Sneath, 1957b); one peak indicates the interspecific similarity values, the other the intraspecific values. These findings and others lend some initial justification to the procedures for clustering the matrix, by demonstrating that the arrangement of the OTU's in the general study is not entirely at random.

The variation of similarity coefficients may be useful as an indication of the homogeneity of the OTU's. If to a homogeneous nucleus of OTU's we add other OTU's of the same homogeneous group, the variance will in general remain about the same. As we add OTU's of a markedly dissimilar group, the variance will rise steeply, and bimodal or multimodal curves of the affinity value distribution may appear.

7.3. TECHNIQUES FOR DESCRIBING TAXONOMIC STRUCTURE

Several techniques have been employed in numerical taxonomy to ascertain and describe structure in matrices of similarity coefficients.

Such techniques have been used in various fields such as ecology and psychology to group related items into ecological associations or personality types. Since the requirements for clustering methods in taxonomy differ sufficiently from those in other fields, we shall restrict our account to taxonomic techniques. These methods are still somewhat in a state of flux, and modifications are to be expected in the next few years, especially as they relate to the development of ever faster and more sophisticated computational equipment.

The clustering techniques described below are generally applicable to all three types of coefficients of similarity—coefficients of association, correlation, and distance. Distance coefficients are conventionally coded in such a way that the larger the coefficient the greater the distance between OTU's. Hence there is a negative relationship between coefficients of association and correlation, on the one hand, and of distance, on the other. Since most cluster methods are designed to recognize the greatest similarity first and lesser similarities later, distance coefficients are for purposes of computation conventionally changed in magnitude in such a way that the greatest distances appear small and the smallest distances great. A convenient scheme for such a conversion is to calculate the complement of the distance from a convenient number, usually ten. We therefore define the tens-complement of the distance as $10 - d_{jk}$. When the similarity coefficients in a clustering method are limited in range from zero to one, the tens-complement of the distance is divided by 10. This results in a similarity value of one for a distance value of zero.

Three techniques for describing taxonomic structure in similarity matrices are discussed below. The simplest of these is differential shading of a similarity matrix (Section 7.3.1); more complex, but more usefully descriptive, are the various types of cluster analysis, described in Section 7.3.2; the most complex of the methods is factor analysis, described in Section 7.3.3.

7.3.1. Differential shading of the similarity matrix

This is the most obvious technique for recognizing in one overall glance the groupings among the OTU's of a similarity coefficient matrix. The method consists of adopting a system of grouping the similarity coefficients into from five to ten evenly spaced classes arrayed by order of magnitude and representing each of these classes by a different degree of shading in the squares of a half matrix. Generally, the highest value is shown darkest and the lowest value lightest. The shading patterns need

to be chosen with care, so that visually they present an even progression. It has been found empirically that for seven shades, ranging from white to black, the following densities of shading (expressed as the percentage of the area of the square that is black) give good results: 0, 12.5, 25, 50, 75, 87.5, 100. Barring and crosshatching are easier to use than stippling. One can then see the half matrix represented as a pattern of different shades, generally limited by a diagonal of squares of the darkest value, representing the similarity of the individual OTU's with themselves (see Figure 7-2). Alongside these can be found densely shaded triangular submatrices, representing groups of related OTU's. Such groups are frequently found on first inspection, unless the array of OTU's has been deliberately randomized, when the squares containing the darkest shade will be scattered throughout the matrix. A shaded similarity coefficient matrix exhibiting reasonably good initial clustering is shown in Figure 7-2. By skillful rearrangement of the sequence of the OTU's, these clusters can be more sharply defined and areas of light and dark can be separated with greater precision. An example of such a "cleaned up" diagram is shown in Figure 7-3. Although this has not been generally done, it might be of advantage to represent such figures as symmetrical

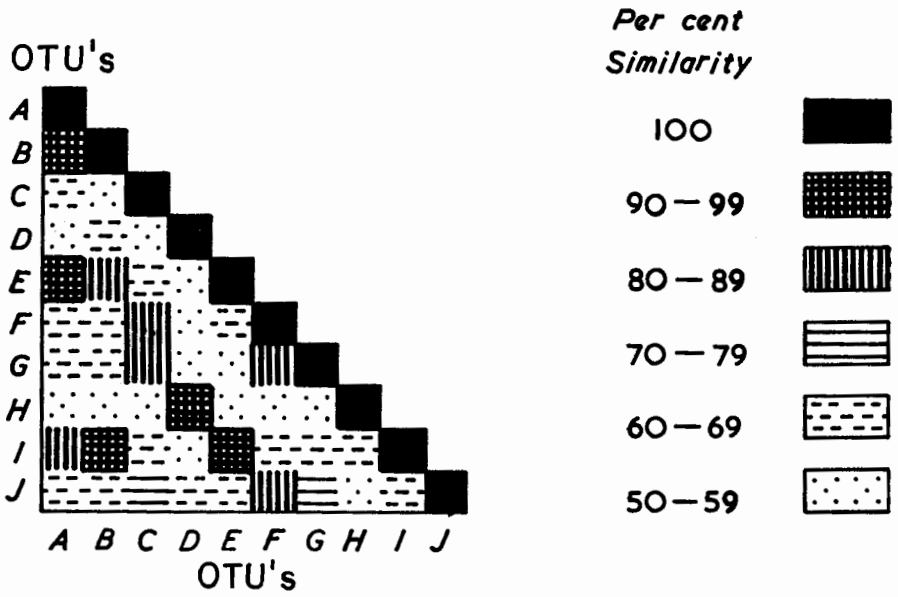

FIGURE 7-2

A shaded similarity matrix with the OTU's arranged in a haphazard order (for explanation, see text).

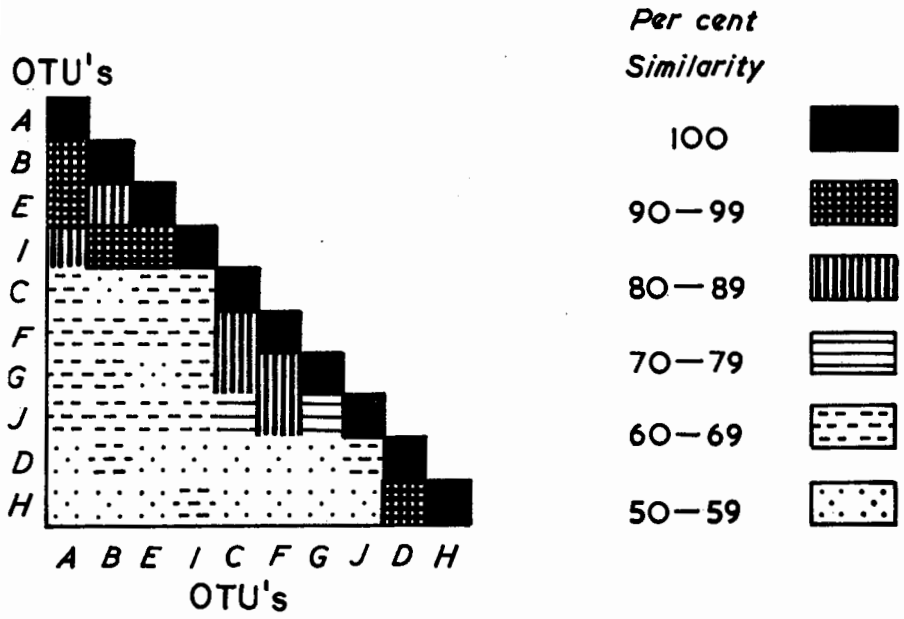

FIGURE 7-3

The similarity matrix of Figure 7-2 after rearrangement by clustering, or the bringing together of OTU's that are very similar to one another.

matrices. In such cases the clusters will be represented as dark squares, as shown in Figure 7-4, which is the symmetrical representation of Figure 7-3 above. One can then visualize the search for group structure as a rearranging of the rows or columns of this matrix in such a way as to obtain the optimum structure in the system. Such a procedure has been suggested by Robinson (1951). But this rearrangement of the rows of a matrix is not as simple as it sounds, because every time the position of a row changes the position of the corresponding column will change. Criteria for optimum structure in a shaded matrix, which must be a function of both the size of the groups as well as the depth of shade, have not yet been developed. When many taxa are used, such diagrams become quite unwieldy. All the individual squares have to be kept very small, making a clear definition of taxa difficult.

7.3.2. Cluster analysis

We mean by this general term a large class of numerical techniques for defining groups of related OTU's based on high similarity coefficients.

7.3.2.1. Elementary cluster analysis

This is the simplest of the various methods of clustering. It consists of arbitrarily selecting a level on the scale of similarity coefficients. All coefficients above this level are written down and the relationships expressed by these coefficients are indicated by lines or links connecting the OTU's, which are represented as points. Selection of a very high coefficient of similarity as criterion for clustering would yield only a few small clusters, just as only the higher peaks of a mountain range would appear as islands on a topographic map if all the area below some high contour line were obliterated. When the criterion for admission of similarity coefficients is lowered, more OTU's join the established clusters, new clusters form, and old ones coalesce. This is analogous to the joining of mountain peaks on a map by successively lowering contour lines. Sooner or later clusters will overlap by this method, some OTU being a member of two clusters simultaneously. Because of the possibility of overlapping clusters, elementary cluster analysis is generally an unsatisfactory procedure. When carried out in large matrices, it must be done by means of a systematic procedure. This is, of course, also necessary if the compu-

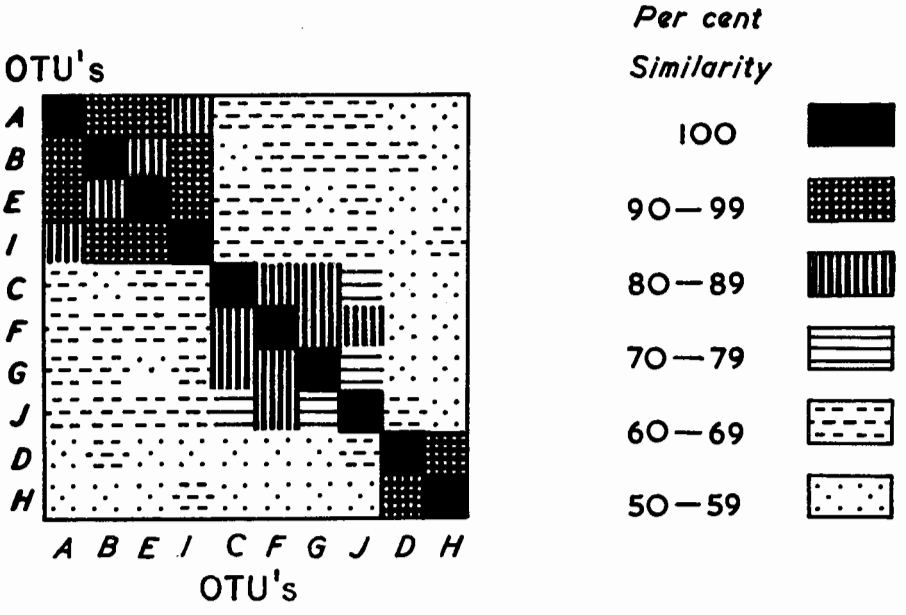

FIGURE 7-4

The symmetric matrix derived from Figure 7-3, as if by reflection in a mirror set along the principal diagonal.

tation is to be carried out by a computer program. Diagrammatic methods of this type, using heavy lines (or multiple lines) for high similarity links, have been used by several authors, for example, Boeke (1942).

7.3.2.2. Clustering by single linkage (Sneath's method)

The clustering methods which follow differ from elementary cluster analysis in that they do not group taxa related above a certain fixed criterion (level of similarity coefficient), but instead first cluster those OTU's most related, gradually admitting more members into the cluster by lowering the criteria of admission. The method practiced by Sneath (1957b) and followed by a number of authors is directly related to elementary linkage analysis. First it clusters together those strains mutually related with the highest possible similarity coefficient; then it successively lowers the level of admission by steps of equal magnitude. Thus if the first criterion is a similarity coefficient of 0.99, the next ones may be 0.98, 0.97, and so on. The computer program designed to process data according to Sneath's technique lists in successive stages the code numbers of the OTU's in the groups as they coalesce. One might, for instance, obtain a record as follows:

S	OTU's
0.99	1, 2
0.98	1, 2, 3 4, 5
0.97	1, 2, 3 4, 5
.	.
.	.
.	.
0.80	1, 2, 3, 4, 5

This indicates that OTU's 1 and 2 join at a similarity value of 0.99; OTU 3 joins them at 0.98, while OTU's 4 and 5 join at the same level; not until $S = 0.80$ do 4 and 5 join 1, 2, and 3.

The admission of an OTU or a cluster into another cluster is by what we may call the criterion of single linkage. By this we mean that if a similarity level of 0.88 would admit an OTU into a cluster, a single linkage at that level with any member of that cluster would suffice to warrant admission. Similarly, any pair of OTU's (one in each of two clusters) related at the critical level will make their clusters join. Thus, while two clusters may be linked by this technique on the basis of a single

bond, many of the members of the two clusters may be quite far removed from each other.

In order to overcome this difficulty, Sneath suggests recalculating the mean similarity values both within groups and between groups. This can, of course, be done at any of several hierarchic levels. The average similarity value is calculated by computing one of two quantities. The first quantity ΓS is the so-called square mean—the mean of all $t \times t$ similarity coefficients including the self-comparison in the principal diagonal. The other quantity, ΔS, the triangle mean, is based on the so-called "strictly triangular matrix," which includes only the triangular portion of the similarity coefficients, excluding self-comparisons. These quantities are related. The ΔS is lower than ΓS or approaches it when t, the number of taxa in a cluster, becomes large:

$$\Delta S = \frac{\Gamma S t - 1}{(t - 1)},$$

$$\Gamma S = \frac{\Delta S(t - 1) + 1}{t}.$$

On the whole it would appear that the ΔS method of averaging would be preferable, since self-correlations are a function of the number of taxa included and would distort the estimates of the relationships among OTU's.

7.3.2.3. Clustering by complete linkage (Sørensen's method)

This method, described by Sørensen (1948) for ecological studies, has not been used in numerical taxonomy. It corresponds in most details to Sneath's method, except that admission of an OTU into a cluster is by what we might call the complete linkage criterion. A given OTU joining a cluster at a certain similarity coefficient S_i must have relations at that level or above with every member of the cluster. Thus single bonds with just one member of the cluster would not be sufficient to effect the juncture. Where the possible groups overlap (where there is a choice of two attachments which a given OTU can make), Sørensen prefers fusion to yield the larger group, or the cluster with the greater number of OTU's. If these are equally large, he would choose the juncture so as to have as few as possible residual groups; and if this criterion turns out to be indecisive, he recommends choosing the combination with the highest average similarity coefficient.

After all junctions permitted at a given criterion have been made, he computes a new similarity matrix, using means of similarity coefficients.

It is obvious that with different initial levels of similarity coefficients the resulting clustering is likely to vary. Sørensen does not feel that these differences in appearance of the final dendrogram, based on differences in the limits by which he admits groups to a cluster, matter very much. However, we feel that the computational procedure leading to dendrograms should be an unequivocal one; therefore we prefer the recalculation of the similarity coefficient matrix at regular and short intervals, as practiced in the following clustering method.

7.3.2.4. Clustering by average linkage (the group methods of Sokal and Michener)

These group methods are a class of clustering techniques proposed by Sokal and Michener (1958). These authors suggested their techniques for the analysis of correlation coefficient matrices, but with some minor exceptions, to be mentioned below, the group methods can be applied to all types of similarity coefficient matrices. We shall discuss them here in this general context. They base admission of any individual into a cluster on the average of the similarities of that individual with the members of the cluster. This average similarity was called $\overline{L}_n$ in the original paper, but now is called $\overline{S}_n$. By members of the cluster we mean either the original OTU's or the smaller clusters composing a higher ranking cluster (see below on how these are formed). As the cluster grows and more remote relatives are considered as prospective members, the value of $\overline{S}_n$ of necessity becomes lowered. In their original study Sokal and Michener suggested that when any one prospective member would lower the $\overline{S}_n$ value of a cluster by 0.03, the prospective member should not be included, and that similarity coefficients should be recalculated among all clusters already formed at that level as well as between all clusters and those OTU's that have remained single. The value of 0.03 was empirically arrived at and referred to correlation coefficients employed by Sokal and Michener (1958) in their study. With different studies and different similarity coefficients this value may need to be adjusted.

By such a method the size of the clusters at any level is likely to vary, and the number of OTU's joining a new cluster is also liable to vary. Perhaps no OTU would join a given cluster in any one computational cycle; this would be because all prospective members had higher relationships to other clusters than to the one under consideration or because prospective members would cause too large a drop in $\overline{S}_n$ values. In other situations an appreciable number of OTU's might join the groups before the $\overline{S}_n$ level is depressed by 0.03. Since the number of OTU's joining a

cluster varies, the above procedure has been called the variable group method, contrasting with the so-called pair-group method. This latter method permits only one OTU to join a cluster during any one computational cycle. This OTU is always the one having the highest average similarity value with the cluster. As soon as all prospective members have joined their clusters, a new similarity matrix of all clusters with each other and with single stems is recalculated. Thus by the pair-group method more recomputation of similarity coefficient matrices is necessary, because during any one clustering cycle only one OTU or cluster can join with another OTU or cluster.

In their original study Sokal and Michener (1958) tried both the pair-group and the variable group method and found little difference between them. They preferred the variable group method at the time, but more recently the pair-group method has been recommended because it is considerably easier to program, particularly for computers of medium capacity and speed. The relative merits of the two methods will be further discussed in Section 7.3.2.6.

The recomputation of the correlation coefficients analyzed by Sokal and Michener (1958), after each clustering cycle had been completed, was originally carried out by Spearman's sums of variables method (Spearman, 1913; Holzinger and Harman, 1941). The general formula for this computation is

$$r_{qQ} = \frac{\square qQ}{\sqrt{q + 2\,\Delta q}\,\sqrt{Q + 2\,\Delta Q}},$$

where $\square qQ$ is the sum of all correlations between members of one group with the other group, Δq is the sum of all correlations between members of the first group, ΔQ is a similar sum between members of the second group, q is the number of OTU's in group 1, and Q is the number of OTU's in group 2. The details of this computation are shown in Appendix A.3. Spearman's sums of variables method correlates the sums of the variables making up any one cluster with sums of variables in any other cluster. In the special case where we would wish to calculate the correlation between a single OTU (x) and a new group (q), the formula is amended to

$$r_{xq} = \frac{\sum r_{xq}}{\sqrt{q + 2\,\Delta q}}.$$

A peculiarity of Spearman's method is that so-called reversals in correlation level are possible; that is, the sums of variables may be correlated at a slightly higher level than the variables composing them. For exam-

ple, if **A** and **B** have formed the nucleus of a group at $r_{AB} = 0.9$ and **C** is about to join them, by the rules of the variable group method both r_{AC} and r_{BC} must be $< r_{AB}$. It can then be shown that $r_{(A+B)C}$ (the correlation between cluster **A** + **B** and OTU **C**) must be < 0.925. Thus $r_{(A+B)C}$, while it will usually be $< r_{AB}$, could be slightly greater. Similar situations can be shown to exist with larger groups. The increases found by Sokal and Michener were well below the mathematically possible limits. In all such cases the relations were represented as multifid furcations of all the stems involved in the reversal and at the highest of the several $\overline{S}_n$ levels considered. When Δq and/or ΔQ are very small or negative, large reversals are possible. When $\square qQ$ is negative, reversals will also be negative. Therefore, when Rohlf (1962) employed correlation coefficients based on standardized characters, he found that the reversals for the coefficients near zero occurred in a negative direction, depressing the mean correlations appreciably. This would indicate that Spearman's method is not suitable for such correlations.

When the group methods are applied to distances or coefficients of association it would not be proper to recalculate correlation coefficients at the end of each clustering cycle. For this reason the new relationships among the clusters are calculated as arithmetic averages of all the coefficients involved in the correlations of any two clusters. This method has been used by Rohlf (1962) and Sneath (1962); an example is given in Appendix A.3. Using this simple method of averages, reversals of similarity coefficients during clustering are, of course, impossible, and clusters

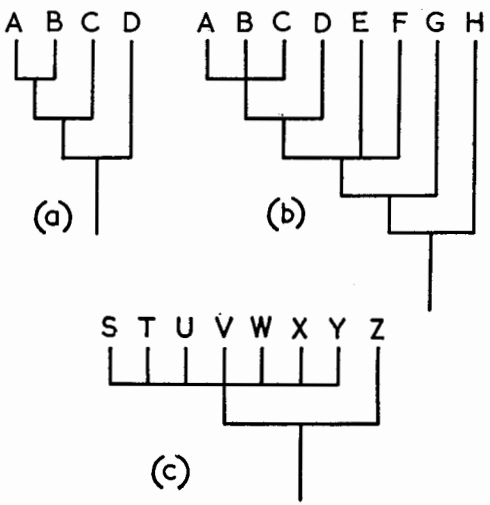

FIGURE 7-5

The weighting of stems in building clusters, modified from Sokal and Michener (1958) (for explanation, see text).

will then decrease steadily or at worst retain the same $\bar{S}_n$ value through several clustering cycles.

A consideration which we have not so far discussed is the problem of weighting stems. The simple diagram of Figure 7-5(a) will clarify this issue. The **A** and **B** represent two OTU's with the highest similarity coefficient. The $\bar{S}_n$ for **C** against **A** and **B** is significantly below S_{AB}, so that **A** and **B** are represented as being closer to each other than they are to **C**. When studying the relation of a fourth OTU, **D**, with group **ABC**, we face the following problem: should we calculate the correlation of **ABC** against **D** with **A**, **B**, and **C** equally weighted or should we weight **A** = **B** and **AB** = **C**? Rephrased biologically, the problem is whether to relate **D** with the homogeneous group **ABC**, or with the stem **AB-C**, where **C** carries as much weight in determining the relation with **D** as do **A** and **B** together. In such a simple case the two alternatives may not produce very different results, but in a situation such as that depicted in Figure 7-5(b), OTU **H** might be weighted as $\frac{1}{8}$ of the group **A-H** or $\frac{1}{2}$, depending on the system adopted. Similarly, **B** would be weighted $\frac{1}{8}$ in the former case but only $\frac{1}{7\frac{1}{2}}$ in the latter. When dealing with fairly large groups the second method would therefore reduce the weight of the members admitted early and increase the weight of those OTU's admitted later. The same problem is found in a situation such as that shown in Figure 7-5(c). By the first method **T** is weighted $\frac{1}{8}$; by the second method it is weighted only $\frac{1}{14}$. Neither of the two methods is entirely satisfactory. By the first method we are reducing the importance of OTU's **H** and **Z** in representing groups **A-H** and **S-Z**, respectively. The clustering methods discussed in Sections 7.3.2.1 to 7.3.2.3 have all been of the unweighted kind. Most of the applications of the technique of Sokal and Michener have employed weighted methods. The relative merits of the two approaches are discussed in Section 7.3.2.6.

7.3.2.5. Central or nodal clustering (the method of Rogers and Tanimoto)

The method proposed by Rogers and Tanimoto (1960) is a clustering method, but its actual procedures are markedly different from the preceding ones. Rogers and Tanimoto started with an association coefficient ranging from zero (absence of association) to one (complete association).

The coefficient which they used was their own similarity coefficient, S_{RT}, but their methods can be applied to any such coefficient, as was shown by Silvestri et al. (1962), who applied Rogers and Tanimoto's method to the simple matching coefficient S_{SM}.

Rogers and Tanimoto's clustering technique can be summarized as follows. After a matrix of similarity coefficients has been obtained, a value R_i is obtained, which symbolizes the number of nonzero similarity coefficients that a given OTU i has with other OTU's. It is obvious that in a matrix of t taxa the maximum value of R_i is $t - 1$. Rogers and Tanimoto originally developed the criterion for R_i as the number of OTU's with which OTU i has at least one attribute in common; hence $S_{RT} > 0$. It might seem questionable whether such small similarity values should be considered in computing R_i. Alternatively one might consider only those similarity coefficients greater than the expected value or greater than a given criterion which had been shown empirically to give useful clusters. This latter approach was adopted by Silvestri et al. (1962), who in setting up their R_i values considered only similarity coefficients larger than 0.65. These R_i values are indices of typicality. The larger the R_i, the more "typical" a given OTU is of the group under study.

Next to be computed is the quantity H_i, which is defined by

$$H_i = \prod_{j=1}^{j=t} S_{ij}, \qquad j \neq i,\ S_{ij} > 0.$$

This is the product of all the similarity coefficients of OTU i with those of the other OTU's, except with itself, and with those coefficients equal to zero (or not considered for evaluating R_i). The higher the value of H_i, the more "typical" will be OTU i. All the OTU's are then grouped in a table in order of descending value of R_i; in those cases where R_i is equal for two or more OTU's the order is by descending value of H_i.

The OTU having the highest R_i and the highest H_i value is considered the most typical one in the whole study, as will be shown below. It is in fact the centroid of the system, when the matching coefficients are expressed in logarithmic form. The most typical OTU is called the prime node of the study and around it is formed a cluster of OTU's having high similarity coefficients with it. The problem now is to find a criterion to determine the number of OTU's that are to go in the cluster. In order to do this a second node has to be found, which is generally the second highest member in the typicality array, although not always. The radius of the cluster of OTU's around the prime node must be such that it does not include the OTU forming the secondary node.

At this point we should introduce the concept of distance as used by Rogers and Tanimoto (1960). They convert the similarity coefficients ranging from zero to one into distances defined as $d_{ij} = -\log_2 S_{ij}$. These

distances will range from zero (when $S_{ij} = 1$) to infinity (when $S_{ij} = 0$); therefore, similarity values of zero are not so converted and, as has been seen above, are omitted from consideration. These distances permit the visualization of taxonomic similarity as taxonomic distance similar to the distances previously discussed in Section 6.2.3. However, they define a so-called semimetric space in which Euclidean properties of distance are not necessarily obeyed—the sum of two sides of a triangle is not necessarily greater than the third. Thus similarity between OTU's **A** and **B** and also between **B** and **C** need not necessarily imply similarity between **A** and **C**. By using the logarithmic formulation we can now calculate

$$-\log_2 H_i = \sum_{j=1}^{j=t} d_{ij} = \sum_{j=1}^{j=t} (-\log_2 S_{ij}), \qquad i \neq i, S_{ij} > 0,$$

expressed as distances. The most typical OTU—the one presumably forming the primary node—will have the lowest $-\log_2 H_i$ value expressed as the sum of the distances.

The formulation of similarities as distances and the expression of them as the negative logarithm of the similarity coefficient permit the relation of these procedures in numerical taxonomy to information theory. These are interesting relationships, even though not particularly useful at this stage in the development of the concept.

If a second node has been determined, the distance to this node can be computed. If we call this distance $d_{1,2}$, then all OTU's related to the primary node OTU at distances less than $d_{1,2}$ are to be included. The resulting cluster now has to be tested for homogeneity. Rogers and Tanimoto (1960) do this by computing a quantity $u_n[(d_{ij})]$, which is a measure of the inhomogeneity of the cluster. The exact definition of $u_n[(d_{ij})]$ is given by the equations

$$u_n[(d_{ij})] = \frac{\varepsilon_n(g, h) - E_n[(d_{ij})]}{\varepsilon_n(g, h)} = 1 - \frac{E_n[(d_{ij})]}{\varepsilon_n(g, h)},$$

$$\varepsilon_n(g, h) = \log_2 \left[\frac{n - g}{2} (n - g - 1) - h \right],$$

where n is the number of OTU's in the cluster in addition to the primary node OTU. (Please note that n is used here with a different connotation than throughout the rest of this book.) The matrix (d_{ij}) is the symmetric distance matrix of the OTU's in the cluster, g is the number of zeros in (d_{ij}) which lie above the principal diagonal, and h is the number of infinite elements above the principal diagonal which are not also in the

same rows and columns as the g zeros. The term $E_n[(d_{ij})]$ is defined as

$$E_n[(d_{ij})] = -\tfrac{1}{2} \sum_{ij}{}' \frac{d_{ij}}{T_n[(d_{ij})]} \log_2 \frac{d_{ij}}{T_n[(d_{ij})]},$$

where d_{ij} represents the elements of the matrix (d_{ij}), Σ' indicates summation of finite terms only after repeated rows and columns are deleted, and

$$T_n[(d_{ij})] = \tfrac{1}{2}(\sum_{ij}{}' d_{ij}).$$

This measure of inhomogeneity is successively computed as OTU after OTU is added to the primary node cluster in the order of their distances from it. When this measure suddenly takes a large jump in value, the "natural" limits of the cluster have been exceeded and, again, the last OTU to be added is removed. Sometimes it is necessary to remove also the OTU closest to the periphery of the cluster, exchanging it for another possible contender in order to see whether the cluster would be more homogeneous with a different composition. After the primary clump has been determined, it is removed from the study and a secondary clump is found among the remaining OTU's. This procedure is repeated with the residual number of OTU's until such a time as all the OTU's have joined clusters or until only a few residual ones remain. These are then attached to those clusters to which they seem to fit best.

The measure of inhomogeneity devised by Rogers and Tanimoto necessitates fairly complex computational facilities, and for this reason Silvestri et al. (1962) simplified it by studying the R_i values of the OTU's in a cluster and rejecting those OTU's with an R_i value much inferior to the possible maximum of $c - 1$ (if c is the number of strains in the cluster, $R_i \max = c - 1$).

A new clustering method is proposed by Lockhart and Hartman (1963) in a recent publication. The groups, although polythetic, are made monothetic by discarding all characters which vary within them. This method avoids the main disadvantage of monothetic methods—the moving of an aberrant OTU to a place removed from its "natural" polythetic place. It employs the number, d_c, of characters in which it differs from a given index OTU (after discarding those characters that are not constant as the group is built up). For details of the method the reader is referred to the original source. The authors report that monothetic groupings obtained for 50 representative microorganisms were found to be essentially similar to the polythetic groups obtained by other methods.

7.3.2.6. Comparison and evaluation of clustering techniques

We have not as yet a systematic study comparing the effects of the different methods of clustering discussed above. For this reason the discussion which follows must be largely based upon theoretical considerations whose significance remains to be empirically validated.

A technique for evaluating different clustering methods by means of cophenetic correlation coefficients has been developed by Sokal and Rohlf (1962). These coefficients are described in greater detail below (Section 7.4). But for the purpose of such a comparison we would compute a correlation between the original similarity coefficients on which a dendrogram is based and the so-called cophenetic values, which are a matrix of coded similarity values extracted from the dendrogram. We may postulate that in an ideal case the dendrogram should reproduce exactly the amount of information on similarity available in the similarity coefficient matrix. Thus a perfect correlation between original similarity coefficients and the cophenetic values in a given dendrogram would show that there has been no distortion whatsoever on converting the data into a dendrogram. It is most unlikely that this would happen in any actual set of data. The extent to which the cophenetic correlation departs from 1.0 will be a measure of distortion which the clustering technique and the dendrogram impose on the taxonomic relationships. We may evaluate different methods of clustering by comparing the magnitude of cophenetic correlation coefficients between the original similarity coefficients and cophenetic values based on a variety of different dendrograms prepared for the same OTU's. When such a study was made, using 23 selected species from the 97 species of the *Hoplitis* complex analyzed by Michener and Sokal (1957), Sokal and Rohlf (1962) found that among the four clustering methods tested—weighted pair-group method, weighted variable-group method, unweighted variable-group method, and weighted pair-group method, using averages rather than Spearman's sums of variables method—the last of the methods gave the highest correlation with the original correlation coefficients ($r = 0.86$). The differences were not very impressive since the greatest distortion, which occurred in the weighted pair-group method, still gave a cophenetic correlation of 0.80 with the original correlation coefficient. Comparing an average linkage method and a single linkage method with original similarity coefficients for the data by Hamann (1961) as processed by Sneath (unpublished), we found cophenetic correlations of 0.59 and 0.34, respectively, showing that the average linkage method represented the data considerably better, although not too well.

Each of the clustering methods described in the previous section is valid in its own right if consistently applied. It is to be expected that the single linkage method will provide for the very rapid coalescing of groups, and hence may not provide a sufficient amount of taxonomic detail. But the complete linkage method may require the lowering of the similarity criterion by a considerable amount in order to establish groups by that method. In both cases the admission of a new member depends upon a single S value, the highest or lowest, which may be unrepresentative for many reasons. It seems to us, therefore, that methods based on averages, such as the various group methods of Sokal and Michener, have an advantage in bestriding a middle path between the extreme positions. Certain problems remain, particularly when a given OTU **A** has much similarity to another OTU **B** but not to **C**, while **B** has higher relations with **C** than it has with **A**. In such a case, average clustering is necessarily unrepresentative of the true relations. In clustering work in fields other than biological taxonomy this problem is not so serious because overlapping clusters are permitted.

A comparison by Sneath (unpublished) of Hamann's (1961) similarity coefficients among families of Monocotyledons shows that clustering by the average method draws out the differences in similarity coefficients considerably, so that the lowest stems unite near a coefficient of zero, while with single linkage methods the lowest stems unite at a coefficient of 60 (out of a maximum of 100). Furthermore, in addition to this condensation of detail produced by the single linkage method, some rearrangements of the families of plants are suggested. In particular, one family (the Cyanastraceae), which shows uniformly low similarity to all other OTU's except one (the Liliaceae), was grouped with the Liliaceae by the single linkage method but was widely separated from all OTU's by the average linkage method. The cophenetic correlation coefficient (see Section 7.4) between the two dendrograms was only 0.28.

If we adopt the method of average linkage, the problem of weighting OTU's within joining clusters assumes importance. The nature of the problem has already been discussed. Sokal and Michener (1958), in discussing the merits of the two methods, felt that the optimum system of weighting would be one between the two extremes—weighting late arrivals more than earlier members of the group, yet not weighting them equally to the entire early stem. Since this was not possible without the renewed introduction of a subjective element into the procedure, they adopted the method of weighting each new member as equal to the sum total of all old group members. They thought such a procedure to be the

less objectionable method of the two, in view of the underlying assumed phylogenetic causes of the phenetic relationships under study. Thus it was felt that clusters of several OTU's or stems bearing a single OTU equally represent independent evolutionary lines. Weighting of stems in this manner does not provide the best classification, yet we lack a better criterion for the moment. A computer could be programmed to apply a series of different weights, where the weight could be some function of the size of the group. Thus, if a cluster representing ten OTU's has another OTU joining them, there are two extreme choices: to weight the single joiner equal to all ten, or to weight the newcomer only one-eleventh. The weight furnished by the computer could be in between these values as a function of the number of OTU's in each stem. But it may legitimately be argued that weighting of stems distorts the affinities among the OTU's. Computation of similarity coefficients among clusters on an unweighted basis is the most faithful method of condensation of the original coefficients. Rohlf (1962) has found this to be true in his data, when tested by cophenetic correlations. However, to follow a consistently unweighted policy may lead to absurdities in certain situations. For example, if we wish to know the affinity of a reptile with the mammals and represent the latter by 100 rodents but only 5 representatives of other orders, an unweighted reptile-mammal affinity would represent mainly a reptile-rodent affinity. A further advantage of the unweighted scheme is that a true average similarity is computed for which confidence limits can be estimated (Rohlf, 1962). Each system of weighting is defensible on several grounds, and we do not as yet have sufficient experience to decide the relative merits of the various systems. In light of our knowledge to date it may well be that the differences between the two extreme choices are so slight that a compromise solution such as suggested above would be generally acceptable.

In deciding between pair-group and variable-group procedures, we note that the former will show less distortion of the original similarity coefficient matrix and be devoid of an arbitrary criterion of group formation. But the variable-group method does not differentiate between fine differences in order of clustering, which may be nonsignificant. Again, we need more experience to decide on the preferred procedure; however, in this instance it is well known that the two alternatives produce very similar results. Also, limitations of computational equipment may determine the choice of method.

In devising our method in such a way as to avoid overlapping clusters, we are in fact biasing the data to yield discrete definable clusters. Such

biases are deliberately introduced as a regular function of the system because of the nature of the classification we are attempting to construct. Biological classification should be constructed by nested nonoverlapping categories, if only for purposes of convenience. The underlying phylogeny makes such an arrangement the only reasonable one, at least among the higher categories. This may well be an aspect of the clustering process in which numerical taxonomy differs from classificatory procedure in related sciences such as psychology, ecology, or language classification. Overlapping language forms are permissible, as are overlapping psychological types and overlapping ecological associations. Overlapping taxa of the same rank are, however, not permissible. (For a contrary view see the paper by Michener, referred to in Section 10.7.)

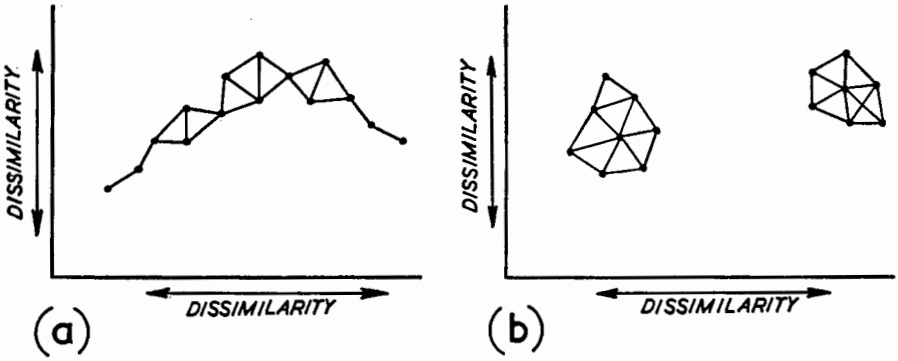

FIGURE 7-6

The effects of different methods of linkage in forming clusters. (a) Single linkage; the clusters may be long and narrow. (b) Complete linkage; the clusters are compact.

Computer programs for average linkage clustering which take into account the variance of the members of the clusters that are formed have so far not been developed. If single linkages are permitted, complicated serpentine clusters may be formed; see Figure 7-6(a). The actual affinity between the "end links" of the serpentine cluster may be very remote indeed. On the other hand, if complete linkage is practiced, the clusters are likely to be very compact and well defined; see Figure 7-6(b). Such clusters could also be conceived as clouds of OTU's in a hyperspace, and they could be defined as hyperspheroids in that space. No attempt at such a definition has yet been made, although the method of Rogers and Tanimoto does in fact delimit clusters by defining hyperspheres around central points. It would seem that straggly, serpentine types of clusters

are not very desirable in biological taxonomy; for that reason also the use of average cluster methods is recommended.

The method of Rogers and Tanimoto has the advantage that when carried out in its original formulation—by computing a coefficient of inhomogeneity—one is able to obtain a measure of the homogeneity of clusters which, as discussed above, would be a desirable feature of a clustering procedure. At the end of the clustering procedure the groups themselves are re-examined, to test whether the most natural groupings have been obtained. Through the mathematical relations to information theory concepts we may be nearing an understanding of the information content of natural taxa in biology.

However, on the debit side we find that the method of Rogers and Tanimoto defines only a series of primary nodes without connecting these into a dendrogram. It centers these nodes around the most typical OTU in each case and, instead of studying the mutual relations among the OTU's simultaneously in a similarity matrix, disperses these as it were to several centers of attraction, the nodal points. Furthermore there is much indeterminacy in the clustering solutions obtained by this method. A number of decisions during clustering are left to the discretion of the operator. This feature was deliberately built into the routine by its authors and has been made a part of the computer classification program (The IBM 704 Taxonomy Application). Such flexibility goes counter to our aims to make the clustering procedure repeatable. Furthermore, viewing the prospects of numerical taxonomy for the immediate future, it is unlikely that an appreciable number of taxonomists will become familiar enough with computer techniques that the kind of close computer-man association necessary for running this program is going to be realized. It would be preferable to have a program which, once the data are fed into the machine, will yield an unequivocal similarity scheme. In our limited experience with this particular technique, we have found that the results obtained by the Rogers and Tanimoto clustering method are usually reflected in various other types of clustering methods yielding more conventional dendrograms. These other methods have also the additional features of joining clusters at precisely calculated levels. Hill et al. (1961) and Silvestri et al. (1962) compared the Rogers and Tanimoto method and the single linkage method on the same affinity coefficients. The dendrograms were on the whole very similar. We cannot at this stage wholeheartedly recommend the procedure of Rogers and Tanimoto, although it does have interesting features which may on further development lead to fruitful results.

In summary we might recommend, at this stage in the development of numerical taxonomy, the average linkage method of clustering employed as a variable-group method or as a pair-group method. The decision which to employ will often depend on available computer facilities. The weighted method is to be preferred until such time as functions for intermediate weighting are developed. There may, however, be legitimate reasons for preferring unweighted clustering methods in a given study. Recomputation of similarity matrices at the end of each clustering cycle is to be done by Spearman's method for correlation coefficients based on unstandardized characters and by ordinary averages for correlation coefficients based on standardized characters and for all other similarity coefficients.

7.3.3. Factor analysis

A method of representing taxonomic structure which is related to but considerably more involved than the clustering methods is factor analysis. Sokal (1958) appears to have been the first to employ factor analysis to indicate taxonomic relationships from a similarity matrix expressed in the form of correlation coefficients. Morishima and Oka (1960) followed Sokal's proposal, applying factor analysis to a taxonomic correlation coefficient matrix but without rotation to simple structure. Factor analysis of Q-type matrices, originated by Stevenson (1936), has been used repeatedly in psychology.

Factor analysis when applied to correlations among taxa may be interpreted as a statistical method for describing the complex interrelationships among taxa in terms of the smallest number of factors. An OTU is placed in the taxon corresponding to the factor to which it is most closely related. The degree of similarity between an OTU and the average aspect of the taxon which a factor represents is given by the factor loadings. The higher the factor loading, the more typical is the OTU of the taxon. In a sense each factor represents the "type" of a taxon. Hence there is a superficial similarity between factor analysis and the type concept, now generally in disrepute (see Simpson, 1961). The distinction between such essentially idealistic concepts and the empirically and statistically based typology of numerical taxonomy is made by Sokal (1962b).

Multiple factor analysis is a branch of multivariate statistics which, in examining a complex set of phenomena stated in terms of correlations among the variables under consideration (in our case the correlations among OTU's), attempts to express these phenomena as functions of a

small number of new variables. These new variables (called factors) should contain the maximum amount of information for describing these relationships. Two different methods of factor analysis are customarily practiced: the principal components method is largely employed by British factor analysts, while multiple factor analysis with rotation to simple structure is widely accepted in the United States. Factor analysis is sufficiently complex that a detailed discussion and explanation of the subject would require a book in itself. Computational details for the method can therefore not be given here, but readers are referred to Cattell (1952), Fruchter (1954), Harman (1960), and Thurstone (1947) as suitable introductory texts.

In the most extensive application of factor analysis to classificatory work, Rohlf and Sokal (1962) employed centroid factor extraction with subsequent rotation to simple structure. This is the most commonly employed form of multiple factor analysis. Centroid factor extraction describes the interrelationships among OTU's in terms of an arbitrary orthogonal (uncorrelated) system; that is, the new variables (factors) are uncorrelated. The first centroid factor accounts for most of the covariation among the OTU's, the second centroid factor for somewhat less, the third still less, and so on. The relative amounts of information contained by each factor can be determined and are usually expressed as a percentage of the total amount of information. The problem of how many factors should be extracted is a complicated subject, and readers are referred to the books mentioned in the previous paragraph for a discussion of this issue.

After the factors have been extracted they are transformed by rotation of coordinate axes to another coordinate system, no longer restricted to orthogonality, which reveals the interrelationships among the OTU's in their simplest form. The criteria for this constellation, known as "simple structure," require that any one factor influence only some of the variables (OTU's) in each study and affect other variables little or not at all. In a simple structure solution, therefore, we do not find a general factor but instead find group factors. Furthermore, any one variable (OTU) should not be affected by all factors. Correlation between the factors is permitted and is frequently necessary if simple structure is to be obtained.

Rotation to simple structure was an immensely tedious and partly subjective method only a few years ago; however, the recent development of so-called analytical (computational) procedures has simplified the process considerably. The concept of simple structure is somewhat controversial, but factors which emerge in a simple structure solution have

been shown to correspond to meaningful entities which reappear in different related studies (Cattell, 1952).

Rohlf and Sokal (1962) applied these methods to sets of 40 species from among the 97 of the *Hoplitis* complex of Michener and Sokal (1957) and obtained the results illustrated in Figure 7-7 for one of these sets. In general they found very good agreement between results of the weighted pair-group method and factor analysis. In Section 8.1.3 we will discuss in some detail the so-called pregroup-exgroup problem. The apparent isolation of exgroup OTU's by the weighted pair-group method, compared to their position by orthodox classification, is probably due to the computational procedures implied in these group methods. The group methods, as well as most other forms of cluster analysis, consider only the highest correlations in a matrix at each clustering cycle. Since relatively isolated OTU's do not have very high correlations with most of the OTU's in the matrix, they are in effect left out until $\bar{S}_n$ is decreased to the average level of correlation of these relatively isolated OTU's with the other OTU's, which exaggerates their degree of isolation. Factor analysis, on the other hand, simultaneously employs all the correlations of each OTU with every other OTU. Therefore the isolated OTU's are considered from the start with the group of OTU's with which they have the highest relationships. All of the species designated by Michener and Sokal (1957) as being exgroup species (see Section 8.1.3) were placed by factor analysis (Rohlf and Sokal, 1962) in the same general groups of species in which they had been placed by Michener and Sokal after the pregroup-exgroup corrections had been made. It would therefore seem that at least for the pregroup-exgroup problem in the *Hoplitis* complex the phenetic classification based on factor analysis is close to the one which Michener and Sokal presumed to be the evolutionary one.

Factor analysis should also predict (by low communalities; see Rohlf and Sokal, 1962) which OTU's might be incorrectly placed by the weighted variable-group method. One difference between factor analysis and the weighted variable-group method is that the latter gives more detail of taxonomic structure, especially at the lower levels, whereas factor analysis only indicates the cluster to which an OTU belongs and the degree to which each of the OTU's resembles an "average" representative of the cluster. This limitation of factor analysis might be an advantage in that it prevents one from attempting to interpret differences which are probably not reliable (see Michener and Sokal, 1957, p. 161). If more detail is desired within a cluster, one has to limit the scope of the study.

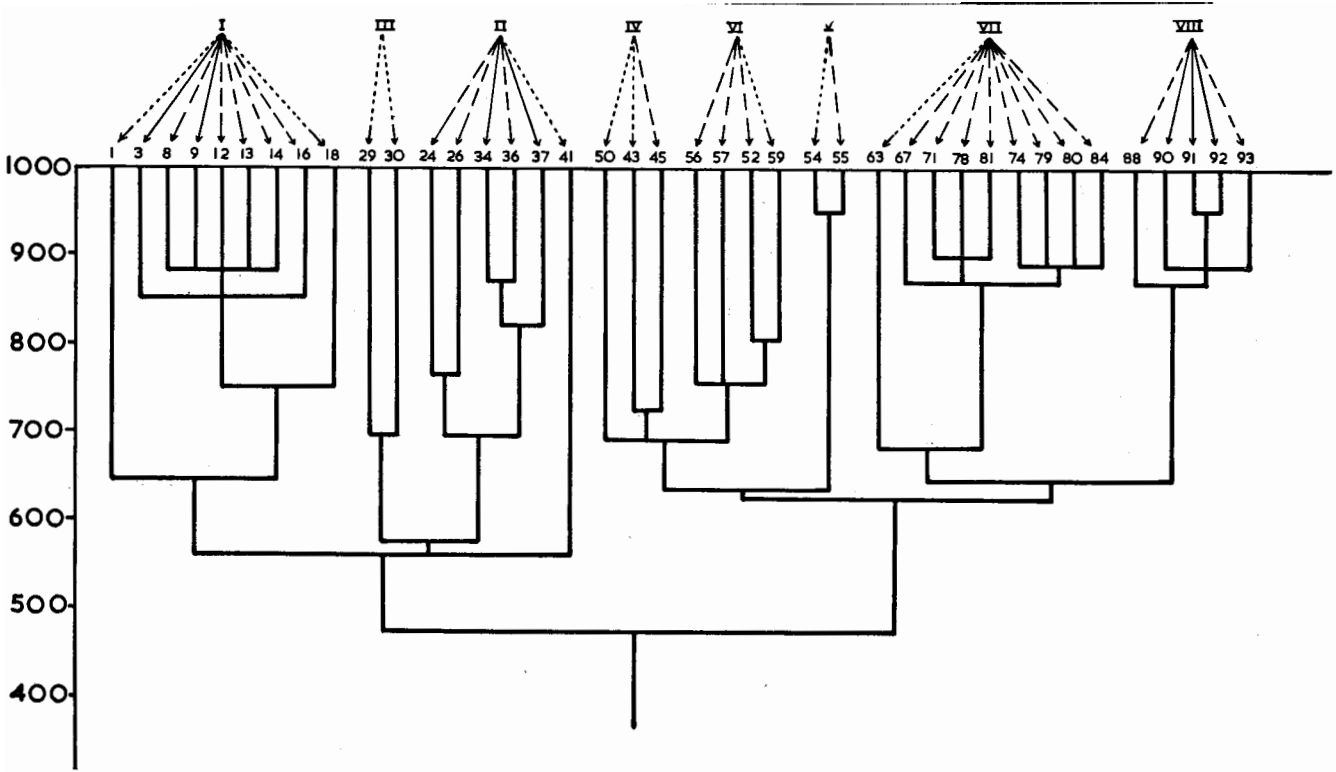

FIGURE 7-7

Dendrogram of 40 species of bees of the Hoplitis *complex (Michener and Sokal, 1957) obtained by the weighted pair-group method. The results of centroid factor analysis with rotation to simple structure are shown across the top in the form of arrow diagrams from the factors to the species. The numbers at the top of the dendrogram are code numbers for the species. Ordinate of dendrogram: magnitude of correlation coefficient between joining stems multiplied by 1000. In the arrow diagrams solid lines correspond to standard partial regression coefficients (primary patterns) >0.9, dashed lines to coefficients between 0.70 and 0.89, and dotted lines indicate the relatively highest coefficient between a factor and a species without very high coefficients with any factor. [Redrawn from Rohlf and Sokal, 1962, Figure 1.]*

On comparing the results of factor analysis with the dendrogram given by the weighted pair-group method, Rohlf and Sokal (1962) found that no straight line could be drawn across a dendrogram to yield exactly the groups given by the corresponding factor analysis. This means that the groups formed by the factor analysis are approximately, but not exactly, at the same hierarchic level. Morishima and Oka (1960) also found the weighted pair-group method yielding results quite similar to those of the factor analysis which they employed, although these authors did not rotate their factor matrix to simple structure.

The main limitation of the usefulness of factor analysis in finding taxonomic structure is the amount of computation necessary with even a small number of OTU's. However, with increased availability of digital computers this limitation is of less importance, although cluster analysis will always be much more rapid than factor analysis.

7.4. THE REPRESENTATION OF THE RESULTS OF CLUSTERING

Similarity matrices which have been clustered by simple shading methods are not generally represented in any other form except as the dark triangles in the original shaded matrix (see Figure 7-3). However, if the clusters are grouped again, one or more secondary matrices may be shown in which each new OTU represents a cluster of former OTU's. One could conceive of such shaded diagrams as cross-sectional transects through a dendrogram.

The most common and convenient representation of the results of numerical taxonomy is by dendrograms. Except for their rectangular nature they appear very much like the customary phylogenetic trees, but they are strictly based on phenetic evidence and should not imply descent. The abscissa of such a dendrogram has no special meaning, serving only to separate the OTU's, while the ordinate is in some similarity coefficient scale usually from zero to one and frequently multiplied by 100 or by 1000 in order to avoid decimal places. Points of junction between stems along such a scale mean that the resemblance between the two stems is at the similarity coefficient value shown on the ordinate. When using correlation coefficients and Spearman's sums of variables method, the highest level of correlation is customarily chosen in the case of reversals, and the level of juncture indicates a correlation no higher than the given value. Michener and Sokal (1957) drew their dendrograms with the tips pointing upward and the final joined stem pointing down-

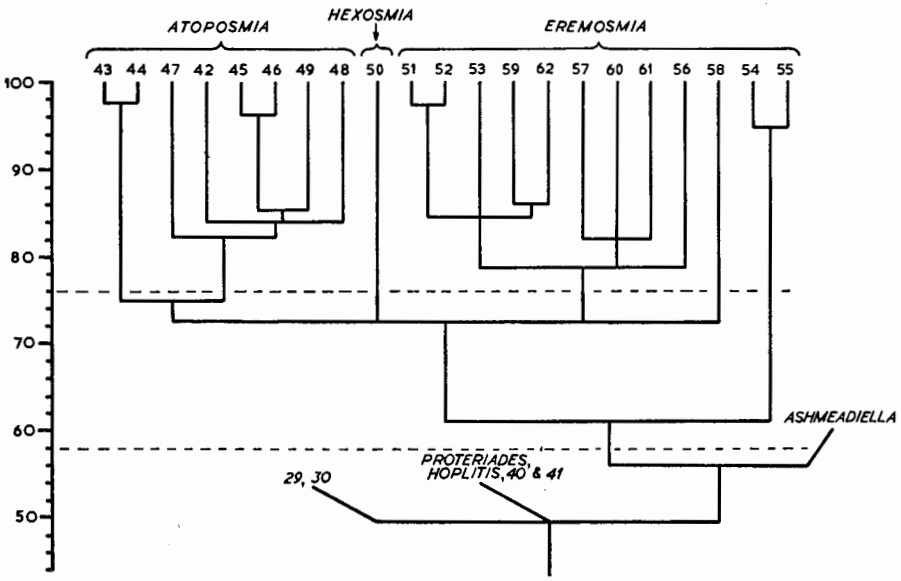

FIGURE 7-8

Dendrogram for the bee genus Anthocopa, *a portion of the* Hoplitis *complex* (*Michener and Sokal, 1957*). *This dendrogram was obtained by the weighted variable-group method.* Ordinate: *magnitude of correlation coefficient multiplied by one hundred. Correlations between any two joining stems can be found by reading the value on the ordinate corresponding to the horizontal line connecting the stems. This value becomes approximate and maximal in cases of multifid furcations. "Roofs" over the species numbers at the tips of the lines delineate subgenera. These subgenera are based on Michener's formal classification established by conventional methods before the numerical taxonomic study was carried out. "Stubs" at the base of the dendrogram indicate connections with related taxa. Broken horizontal lines were drawn across the dendrogram by Michener and Sokal in an attempt to delimit genera and subgenera* (*at* r *levels of 76 and 58, respectively*). *More recent practice would tend to label these lines phenon lines yielding 76-phenons and 58-phenons, respectively. Using this terminology, we arrive at six 76-phenons in Anthocopa, which are* (*43 . . . 44*), (*47 . . . 48*), (*50*), (*51 . . . 56*), (*58*), *and* (*54 . . . 55*). [*Redrawn from Figure 7 in Michener and Sokal* (*1957*).]

ward on the page, as shown in Figure 7-8. Most authors have followed their lead. Another version is to present the dendrogram lying on its side with the tips of the OTU's pointing to the right. Such an arrangement has been chosen by Sneath (1962), for example (see Figure 7-9). It has the advantage that if the nomenclature of the classification is to be listed in the same diagram this can be done more conveniently with the diagram lying on its side.

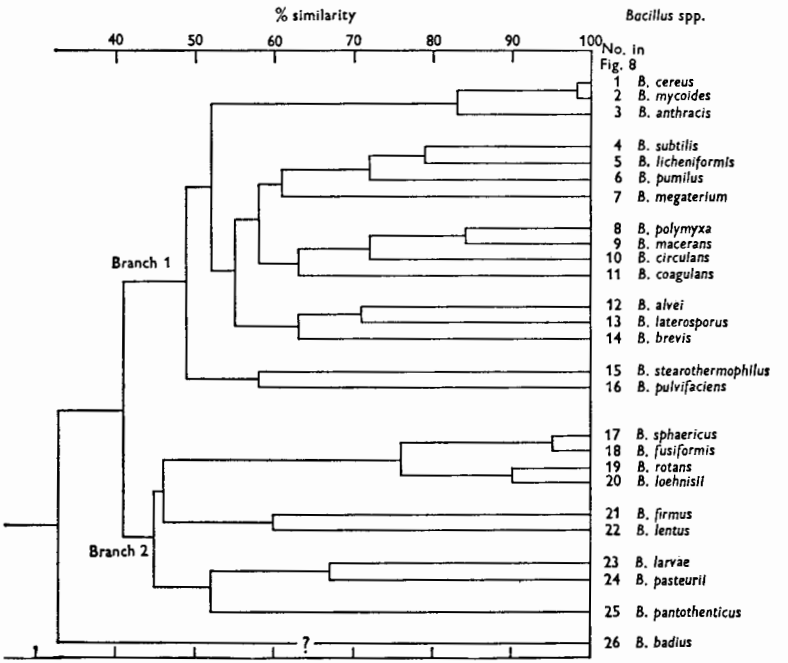

FIGURE 7-9

Dendrogram of 26 species of bacteria of the genus Bacillus. [*Sneath (1962), Figure 9. Reproduced by permission from the 12th Symposium of the Society for General Microbiology.*]

Taxonomic rank can be assigned to OTU's by drawing horizontal lines across (vertically oriented) dendrograms at given levels (see next section for details). While the scale alongside any one dendrogram is a reliable indicator of the level at which two stems come together, it is not to be construed as representing relationships among the tips of the dendrograms, which are impossible to represent adequately in a two-dimensional graph. As we have seen in Chapter 2, representation of phenetic relationships among the OTU's in a study is not possible by a dendritic scheme in two or even three dimensions.

Comparison between the methods of representation by different clustering techniques can be carried out by the technique of cophenetic correlations devised by Sokal and Rohlf (1962). These authors divided the similarity values along the ordinate into a suitable number of equal class intervals by drawing *phenon lines* as class limits across the dendrogram (see Figure 7-10, where the range of similarity values has been divided into eight classes). The number of classes into which the variation should be divided will depend upon the number of OTU's being classified. As a

very rough guide, dendrograms of less than ten OTU's need not be divided into more than four classes, while dendrograms involving as many as 100 OTU's should be divided into at least ten classes. A further consideration should be that the class intervals should be fine enough to reveal a reasonable amount of structural detail in the dendrogram to be analyzed. Persons planning to do such computations on a desk calculator should employ the minimum number of classes necessary. But increasing the number of classes never does any harm from a statistical point of view. A computer program developed by Rohlf divides the range of similarity values into 50 classes. Schemes could also be developed which would handle the actual similarity value at which two stems join. The statistical consequences of using actual juncture levels rather than co-

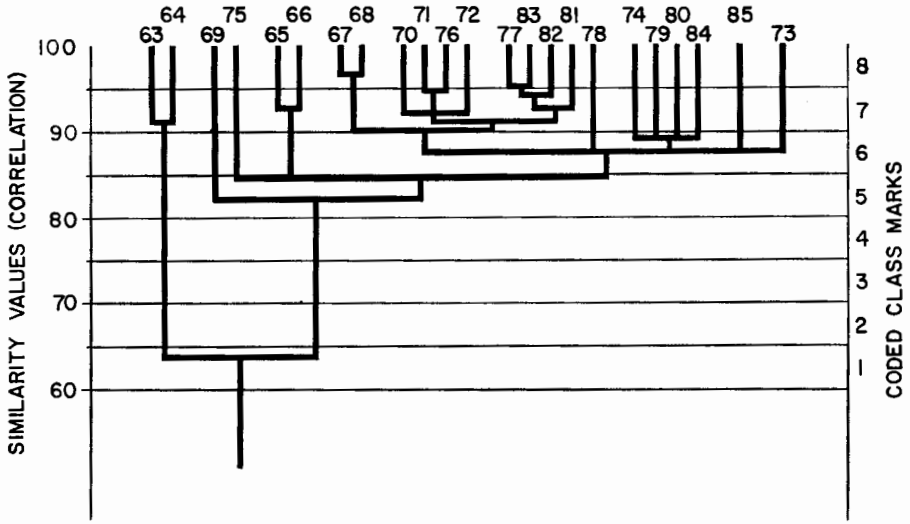

FIGURE 7-10

Dendrogram or diagram of relationships among 23 species of the bee subgenera Chilosmia *and* Ashmeadiella s. str., *taken from Rohlf and Sokal (1962) and based on data by Sokal and Michener (1958). The relationships were obtained by the weighted pair-group method (WPGM). The ordinate is graduated in a Pearson product-moment correlation coefficient scale (coded by multiplying by 100). Numbers across the top of the figure are species code numbers which are identified in Rohlf and Sokal (1962) or Michener and Sokal (1957). Horizontal lines across the dendrogram are phenon lines defining taxa at the minimum level of similarity at which the phenon line cuts the ordinate. Class intervals delimited by phenon lines along the similarity scale have had their class marks coded (on the right side of the dendrogram). [From Figure 1, Sokal and Rohlf (1962). Permission of the editors of* Taxon.]

phenetic values are probably slight, based on the well-known effects of grouping of frequency distributions.

Once the class intervals along the ordinate have been established, each class mark should be coded on a scale starting with unity at the low end (the end having the lowest similarity value) and increasing by unit steps. Thus with ten classes the highest class should be coded 10. These values will then be proportional to the similarity values, except in the case of distances, where they will be complementary and where inverse coding —starting with unity at the highest level—might be advised. The coding is a computational convenience for desk calculator operation; actual class marks can be used in digital computer programs.

The cophenetic value of two OTU's was defined by Sokal and Rohlf (1962) as the class mark of the class (between phenon lines) in which their stems are connected. For example, in Figure 7-10 we can see that species 65 and 66 are connected in class interval 7. Hence their cophenetic value is 7. Similarly, the cophenetic value of species 69 with species 74 is 5, since that is the level at which these OTU's are connected. The closer the relationship between the two OTU's, the higher their cophenetic values. It is convenient to record cophenetic values in matrix form, resembling a matrix of similarity values (see Appendix Table A-17).

A comparison of dendrograms is made simply by calculating an ordinary product-moment correlation coefficient between the corresponding elements of the two matrices of cophenetic values to be compared. These coefficients have been called cophenetic correlation coefficients. For this procedure, each half matrix can be imagined as strung out in single file, column by column. For t OTU's there will be $t(t-1)/2$ elements in a half-matrix of similarity coefficients. The magnitude of the cophenetic correlation coefficient will describe the amount of agreement between the two dendrograms being compared. In their original study Sokal and Rohlf (1962) found that the weighted variable-group method was closest to the weighted pair-group method; the dendrograms prepared by the weighted variable-group method resembled those prepared by the weighted pair-group method at a cophenetic correlation coefficient of 0.95. Comparison between average cluster methods and single linkage cluster methods on Hamann's (1961) similarity matrix of monocotyledonous plants gave a value of only 0.28.

In view of the difficulty of representing the phenetic relations among all the OTU's by means of a dendrogram, other graphic methods of representation have been attempted. Models or projections of models which represent the taxa as points or little spheres in a character space

have frequently been used. It is of course generally impossible to represent all the relations among t taxa in a three-dimensional space except in unusual cases, as where the rank of the matrix does not exceed 3. However, in a surprising number of instances one can represent a relatively small number of taxa in a three-dimensional space without doing too much violence to their distances. It is, of course, distance matrices particularly which can be so represented (see Figure 7-11). These representations are not particularly useful for purposes of publication (however, see Lysenko and Sneath, 1959) but are often of great interest for private study by the investigator who can get a different "feeling" for the quantified relations among the taxa by seeing them in this particular

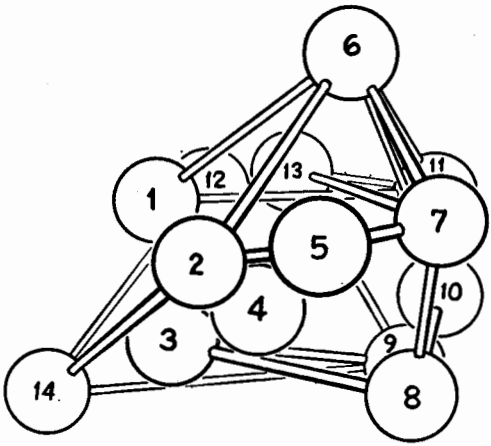

FIGURE 7-11

Taxonomic model of the Enterobacteriaceae. The taxa are represented by spheres connected by rods indicating taxonomic distance. [*After Lysenko and Sneath (1959).*]

form. Looking at these models and photographing them from different angles and points of view also helps in this connection.

Correlations among taxa can also be represented as angles (angle = arc cos r) between vectors, the tips of which represent the OTU's (Rohlf and Sokal, 1962). Various taxa can then be represented as clusters of vectors, more or less like knitting needles sticking out of an orange. However, we again run into the problem of showing relations which need more than three dimensions for a true representation.

7.5. CRITERIA OF RANK

We believe that rank should be based on phenetic criteria, as indeed we believe is the usual practice of taxonomists. These criteria have usually been based on two points: (1) that the internal phenetic diversity of taxa of equal rank should be as nearly equal as possible; (2) that gaps

between taxa of equal rank should be as nearly equal as possible. The second point is to cover the cases of monotypic genera, families, and so on, as with *Gingko biloba*, which is the only species of its order. These two rules may sometimes be inconsistent, and we must then decide which to choose. This depends on the kind of cluster analysis employed. Simpson (1961, pp. 133–134) would, for instance, retain as a genus the very diverse taxon *Rattus*, although the divergence among its species is greater than the average distance between *Rattus* and the closely related genus *Thallomys*.

A case in point is the birds, a class which has much suffered from "splitters" and from the naming of trivial variants as subspecies. There is evidence that phenetically and genetically they are, for vertebrates, a group of small variability. The most aberrant are the penguins and some ratites, but these are not as different from "typical" birds as whales and bats are from "typical" mammals. Birds are remarkably similar to one another serologically (DeFalco, 1942), and interspecific and intergeneric hybrids are relatively highly fertile (Huxley, 1943, p. 146; Sandnes, 1957). The difficulty of making satisfactory classifications (either phyletic or phenetic classifications based on comparative morphology as discussed by Stresemann, 1959) and the large number of monotypic genera also suggest that the category of genus in birds is of lower rank than the genus in other vertebrates. Possibly a more uniform scheme would treat the birds as a class containing only one subclass, with the latter containing only a few orders, and would reduce most avian orders to families and most families to genera.

Those who have devised techniques for numerical taxonomy have suggested that they can be used to decide the rank of the taxa which they yield, and some suggestions have been made that agreement might be reached by biologists on the similarity levels which should define the categories of rank. To say that this is premature is to state the obvious, but the likely developments deserve some discussion. Sneath (1961) has pointed out that there is a lower limit to the groupings which can be fitted into a nonarbitrary hierarchy. For example, different mutants of one species cannot be so arranged; it is impossible to decide whether white cats are of higher rank than long-haired cats. Such groups are not phenetic taxa; they are "rankless taxa" and cannot be satisfactorily handled by hierarchic subspecific nomenclature. Many so-called subspecies are of this nature.

The criteria which have been proposed are all very similar in essence: the same criterion for a given rank must be applied to all parts of an

analysis. That is, where a hierarchical tree has been made, the line defining a given rank must be a straight line drawn across it at some one affinity level. The line must not bend up and down according to personal and preconceived whims about the rank of the taxa. We believe that in the foreseeable future each major group will have to be standardized separately. No useful standard can yet be applied to both bees and jellyfish, but within the megachilid bees, or perhaps within all the bees, some worthwhile standardization might obtain. To make this practicable there would have to be agreement on the rank of the whole group under study; we also would have to decide on the rank of the OTU's employed, which will frequently be a category considered to be a species. The other ranks could then be intercalated evenly.

The particular body of data to which this agreed primary scale was to be fitted would also need to be agreed upon. Any subsequent studies should always include these characters (if at all possible), even if they were augmented by new characters. There is something to be said for choosing certain type specimens and defining these as points of reference, for in the event of disagreement on whether the original data were representative it would be possible to make further analyses using additional features of these types or of acceptable replacement types. Nevertheless, it is our conviction that if adequate and representative samples of the known taxa and of the characters are chosen, subsequent analyses will seldom show marked discrepancies in rank. Such a reference system should obviously make the least alteration of well-established systems of taxonomy. Another useful convention would be to use a parallel system based on affinities, such as 80-phenons, which would on each occasion be equated with that rank which the worker accorded to it (see Section 9.1.1). This would be preferable to any attempt to give fixed taxonomic ranks to the phenon scale, since it would lose its point and its flexibility. This is not to say that ranks might not by agreement be equated with the phenon scale in a particular named taxonomic study.

The consequences to taxonomic rank of adding or removing taxa from a study would, we believe, not prove a serious problem, provided that only a small proportion of the OTU's were involved. The coefficients of affinity themselves should be little affected if the recommendations in Chapter 6 are followed. The methods of cluster analysis employed (see Section 7.3.2) will themselves have some influence on the ranks, since these methods summarize the affinity matrix in slightly different ways. In general we expect this effect will be as great as that produced by omitting a small proportion of OTU's.

As work proceeds on the higher ranks there are sure to be changes in the relative ranks of some taxa. For example, it might become apparent when a thorough comparative taxonomy of all insect orders was made that the Blattaria, for example, which had been treated initially as an order, were only of familial rank. For this reason some new ranks might have to be intercalated to express this, or possibly a revision of the system might be forced. This should be a pressing reason for comparative work at various hierarchic levels to be attempted as soon as possible. In this connection, such techniques as comparative serology may have a valuable part to play. Whatever the growth of numerical taxonomy may be, there should be great efforts to build it up by coordinated work between different specialities, calling for much more cooperation than has occurred in the past. To make an imperfect but legitimate analogy with map making, a number of reference points over wide areas will be of as great value in systematics as surveyor's bench marks are in cartography.

Do we have enough ranks to handle the number of taxa which exist among living organisms? The use of numerical taxonomy may lead to a need for finer gradations of rank, with appropriate names for the new ranks. It is therefore of some interest to ask whether the existing ranks will be adequate for this task. Two separate questions arise. (1) Is the present hierarchy adequate to handle the taxa of living creatures if it only aims at giving a schematic and convenient system? (2) Is it adequate if it tries to reflect accurately the values for rank which numerical taxonomy may provide?

The first question is easy to answer if we confine ourselves to species and higher taxa. The number of species of living creatures is probably between 1.5×10^6 and 2.0×10^6, a majority of these being insects. Allowing for fossil species, it is unlikely that systematists will study more than 10^7 species, even though the number of such fossil species over geological time may have been in total several powers of ten greater than this. The seven conventional ranks of kingdom, phylum (in zoology) or division (in botany), class, order, family, genus, and species would then be adequate if, on the average, each rank contained ten taxa of the rank below. In some instances the taxon would contain more than ten of the next subordinate taxa, but if they did not exceed one hundred the system would still be workable, especially with the help of the intermediate ranks such as subfamily, tribe, and others. The addition of subspecies and variety would also allow for ample scope at infraspecific levels.

The second question is not so easy to answer, for we cannot foresee sufficiently the scope of numerical taxonomy. In theory, at least, the

present system of ranks would be quite inadequate. Consider those organisms with 10^9 bits of genetic information (see Section 5.3.1). This would allow $2^{1,000,000,000}$ possible combinations of features, and if all combinations gave viable organisms this would allow the existence of about $10^{300,000,000}$ different forms of creatures. This is by no means the upper limit of the genetic potential of living creatures. In theory we might need to measure affinity coefficients sometimes between creatures which were identical, and sometimes between those which differed in $10^{300,000,000}$ respects. The few rank categories would be far too few to allow the flexibility needed. Several million ranks (each taxon of which contained some hundreds of the next lower taxa) would be barely adequate to express the numerical affinity coefficients in terms of rank. Yet in practice nothing as extreme could arise even if we were able to make numerical analyses of the size implied above. For if we consider only species, we have only 10^7 of these. Even considering individuals will not lead to such fantastically large numbers, for if the number of protons and electrons in the observable universe is about 10^{80}, and geological time is about 10^{17} seconds, it is clear that there could never have been on earth $10^{300,000,000}$ individual creatures even if each had lived for only one second. In addition, most of the possible combinations would have resulted in inviable organisms, though we cannot yet guess what percentage would be viable.

It is clear from the above example that most of the character hyperspace (which we can use to represent the relations of possible organisms) is empty. It seems likely that the present-day organisms are scattered somewhat unevenly through it because there are large tracts of it which appear to be unoccupied—for example, the enormous tracts which would hold the organisms intermediate between higher animals and higher plants, if such creatures could exist. Indeed, if we could visualize the phenetic position of all organisms past and present, we would expect them to represent a dendrogram in hyperspace outlining the phyletic tree of living creatures.

7.6. THE RELATION BETWEEN Q AND R TECHNIQUES IN NUMERICAL TAXONOMY

We have already briefly described differences between Q and R correlation matrices. Based on the same original data, the former represent correlations between subjects (in the case of numerical taxonomy between OTU's), while the latter represent correlations between char-

acters. Since both sets of correlation matrices are obtained from the same $n \times t$ data matrix, it is obvious that the two are related to each other, and in fact they can in certain cases be transformed, one into the other (Thomson, 1951). Three aspects of R correlations are considered below.

The R correlations are important for their intrinsic interest. Little work on R correlations in taxonomic groups with OTU's as high or higher than the species level has been done. There are a number of studies in which OTU's represent individuals within populations or among populations (Hammond, 1957; Jolicoeur, 1959; Kraus and Choi, 1958; Sokal, 1959, 1962a; Sokal and Hunter, 1955; Sokal and Rinkel, 1963). However, we know only of the study by Stroud (1953) of R correlations using species as OTU's. Much of interest remains to be learned here. We would like to know how large correlations among characters are when different-sized groups are studied and whether different types of characters such as length and weight will correlate differently from meristic or qualitative characters. Correlations within larger taxa and among larger taxa are equally feasible and could be done by partitioning covariances, similar to the technique suggested by Sokal (1962a).

Since characters are correlated, the Q matrix of correlations between OTU's does not have a sampling distribution expected of ordinary correlation coefficients. This is because each reading for an OTU is not an independent sample from a common population but represents different characters. We have referred to this problem previously as that of the heterogeneity of column vectors. As we have seen, the problem of the reliability of correlation coefficients is somewhat alleviated by standardizing rows (standardizing the variates for each character). However, the problem of redundancy of information because of character correlations remains. We need to obtain more information on the magnitude of the effect of correlations among characters on the sampling distribution of correlations among taxa.

Work at present underway in Sokal's laboratory will investigate the relations between R matrices and Q matrices. One reason why no more has been done in this connection in numerical taxonomy is that in work done so far usually fewer OTU's than characters have been measured. It has been simpler, because of limited capacity of computers, to calculate correlations among OTU's than correlation among characters. As computational equipment gets better and faster, we shall be able to attack these problems more efficiently.

Factor analyses of R correlation matrices should be of great interest. Such analyses would presumably reduce the R correlation matrices to

primary components, possibly orthogonal or deliberately orthogonal, by use of principal axes methods. This would mean that the correlations among n characters could be analyzed into k factors, where $n > k$. By reducing a character correlation matrix through factor analysis, we are isolating independent dimensions of variation of the R matrix, hoping thereby to remove redundancy from our studies. Studies of this sort have recently been carried out by Sokal (1962a), Sokal and Rinkel (1963), and Defayolle and Colobert (1962). This has interesting implications for numerical taxonomy. If there is redundancy in character information it might be possible to isolate the factors from an R correlation matrix and to calculate factor scores for each OTU on the factors obtained. We could then employ only those characters which provide independent information on the taxa concerned. This would reduce the number of character scores on the basis of which OTU's could be classified. It would, however, not reduce computational effort, owing to the tedium of factor analysis. It might be argued that after such an analysis characters would in fact be weighted in terms of the independent amount of information which they contain. This might open the door to a numerical taxonomy among OTU's based on weighted characters. These would, of course, not be weighted by their presumed phylogenetic importance but by the criteria mentioned above. Our thoughts in this connection should therefore not be misconstrued. Studies of this sort have not advanced beyond the programmatic stage. We have at the moment no idea how many common factors would be found in an R correlation matrix involving higher ranking OTU's. Studies of R matrices at the lowest taxonomic levels have generally not produced many factors (for a discussion of this point see Sokal and Rinkel, 1963).

7.7. THE PUBLICATION OF RESULTS

The publication of the results of numerical taxonomy may raise a number of problems. In many instances it would be desirable to publish the data on the characters together with the method of scoring them, and also the full affinity matrix. This information would be needed by workers who wish to re-examine the group by similar techniques. However, these data would take up a great deal of space, and it may be more convenient to arrange for microfilm records of the manuscript data to be made available to other workers. The original data matrix and the affinity matrix may be deposited in a library and the fact noted in the published work; for example, the Science Library in London has pro-

vision for the deposition of certain material of this kind. The data may also be stored in a form suitable for data processing, in punched card or tape format. Where it is possible to publish the lists of characters which were employed and their definitions, scaling, and coding, together with records of the material employed, this should be done; if the full details are too lengthy, they should be deposited with the data matrix. It would be valuable to publish, in addition to any diagnostic keys, a list of the most constant characters of the taxa which have been recognized, for these are of much practical value to subsequent workers.

Since the affinity matrix is of such importance (for the usual hierarchical tree is a gross oversimplification of it), it may be useful to publish it as a shaded diagram (Renkonen, 1938; Sneath, 1957b; Sneath and Cowan, 1958); see Figure 7-3. Not only does this take up less space, but the shading gives a clear visual impression of the affinities, provided that the OTU's have been arranged so as to bring together (as far as possible) the members of each taxon. The shaded diagram has some disadvantages: it is not easy to handle more than 8 or 10 different shades, so that the affinity values can only be represented to the nearest 10% or thereabouts, and this may be too coarse a spacing to allow for the best cluster analysis; it also must be converted back into numerical form if further analysis is attempted. Nevertheless, in many instances differences of much less than 10% may not be significant, so the loss of information may not in reality be very great.

The use of models (such as Anderson and Abbe, 1934; Lysenko and Sneath, 1959) may prove useful, but they are more helpful for obtaining a better idea of the relations while one is studying a group, or for teaching purposes, than they are for publication. For most purposes a dendrogram will be essential in any publication. It is important to specify what measures of affinity and clustering were employed. It may, in addition, be helpful to note the electronic computer and the program used, since other workers may greatly benefit if they can use programs which have already been prepared for these techniques.

7.8. THE INCORPORATION OF ADDITIONAL DATA INTO A CLASSIFICATION (INTERSTUDY COORDINATION)

After a numerical taxonomic study has been completed, two kinds of additional data are likely to be forthcoming. First, new study of the organisms may reveal characters other than those employed in the earlier

work. Second, information may become available on OTU's which are related to those previously studied but which for one reason or another had not been included in the study before. The second problem may also arise when persons are studying similarities among a large number of organisms and are unable to process all the data simultaneously because of limitations of the computer. We shall take up these problems in turn.

Adding new characters will be warranted only if the new characters are quite numerous or if the earlier classification has been based on relatively few characters. If the hypothesis of the matches asymptote holds and a sufficient number of characters have been employed previously, the new characters should not change the arrangement of the taxa appreciably. We have so far no experience with situations of this sort and shall have to await work in this field before coming to definite conclusions on the number of characters which must be added before a revision of a group need be made. In such circumstances records of new characters should probably be deposited in some central agency (see in this connection our thoughts on the future of systematics, Section 10.5) until a sufficient number have accumulated in order to warrant revision of the group. It may also be that in future years there will be electronic files of taxonomic information available at certain central locations. This would permit the new characters to be added to the old. When taxonomic reprocessing of the data is indicated, all characters new and old would be considered. Consideration should also be given to studies of correlations of characters. It may be that newly studied characters will have to be correlated with characters that have already been established, and only new information in the sense of character variation not represented by the previously studied characters will be added to the eventual data matrix.

Related to the above problems are situations where two studies are made on the same material but using different samples of characters (even if partially overlapping). We expect that the similarity values would be very close in the two studies (provided the conditions mentioned in Section 5.4 are fulfilled, notably that the samples of characters are both large ones). As has been pointed out (Sneath, 1957b), the correspondence cannot always be expected to be close in actual affinity indices, but the *relative* values for given pairs will probably be more constant. If, for example, the similarity of sparrow to sparrow is 90% in one survey, but sparrow to duck is 50%, then a second study might yield the following results: sparrow to sparrow 95% and sparrow to duck 75%. We do not yet know very much about the mathematical properties

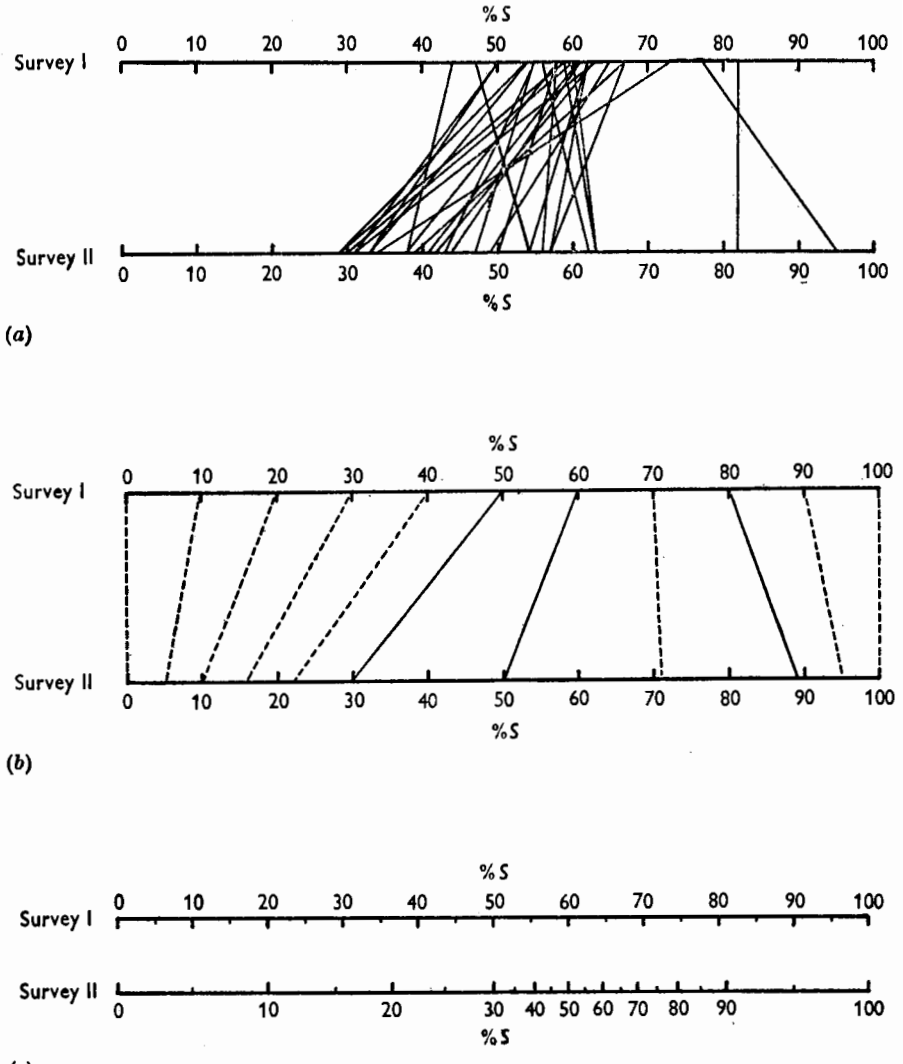

FIGURE 7-12

Calibration of one S *value scale with another.* [*Sneath and Cowan (1958). Reproduced by permission of the* Journal of General Microbiology.] (*a*) *Method of calibration. A line is drawn from the* S *value found in one survey for the comparison of a particular pair of strains to the* S *value found in the other survey for the comparison of the same pair of strains. All the comparisons which were made in both surveys are similarly entered.* (*b*) *The approximate mean correspondence between the two scales of* S *is obtained by drawing lines from one scale to the other based on the mean slopes of the lines in figure* (*a*). *The* S *values of 0 and 100% are the same on both scales. Dashed lines are only presumptive, as they are not based on information in* (*a*). (*c*) *From* (*b*) *the two scales are calibrated against each other by distorting one of them.*

of these relative values in different surveys. In the example given above the ratios of $1 - S$ are constant, but it would be hazardous to predict what relations will in fact be found to be the most usual.

It is possible to compare one study with another, so as to calibrate the two scales of rank. For a very simple example (Sneath and Cowan, 1958), see Figure 7-12. The mathematical treatment has not yet been developed, but it is likely that if the affinity values in the two studies are plotted as a scatter diagram, using a suitable transformation, the best-fitting line as given by a least squares statistic would be sufficient to allow us to calibrate the affinity value scale of one study in terms of the affinity value scale of the other study. If we employ affinity coefficients scaled to lie between 0 and 1 and plot the affinity values for two studies on the same pair of organisms, we can make a scatter diagram which illustrates the relation between the two scales. The points, if numerous enough and spaced throughout the whole range of affinity values, will lie with some scatter along a curved line, except in the unlikely event that the two scales are virtually identical. The ends of the curve will approach 0% at the lower ends of the two scales and 100% at the upper ends, and the intervening part of the curve may well be sigmoid. With correlation coefficients, the upper and lower limits will be -1 and $+1$, but similar relations will hold. This will mean that the curve will be of cubic or higher order, which will be difficult to fit to the scatter diagram without the use of an electronic computer. With the increased availability of electronic computing facilities it will be commonly practicable to fit the best-fitting higher-order curve to the scatter diagram, and this should prove to be sufficiently accurate for any ordinary work.

Even without curve fitting, we find two-way frequency distributions of similarity coefficients of great interest. Figure 7-13 shows such a scattergram for similarities between pairs of species (in males and females separately) of the 97 species of bees in the *Hoplitis* complex (Michener and Sokal, 1963). The similarity coefficients are correlation coefficients based on standardized characters. There is a clear positive correlation between the two variables ($r = 0.71$) although the scatter at the upper end of the distributions is quite wide. Thus when expressed in correlation coefficients there is considerable congruence between male and female similarity values.

Adding new OTU's to a study presents more serious problems. If many OTU's are added, it is obvious that a revision of the entire group is necessary. On the other hand, if only a single one or a few OTU's are added, the computation of the correlation coefficients between these new

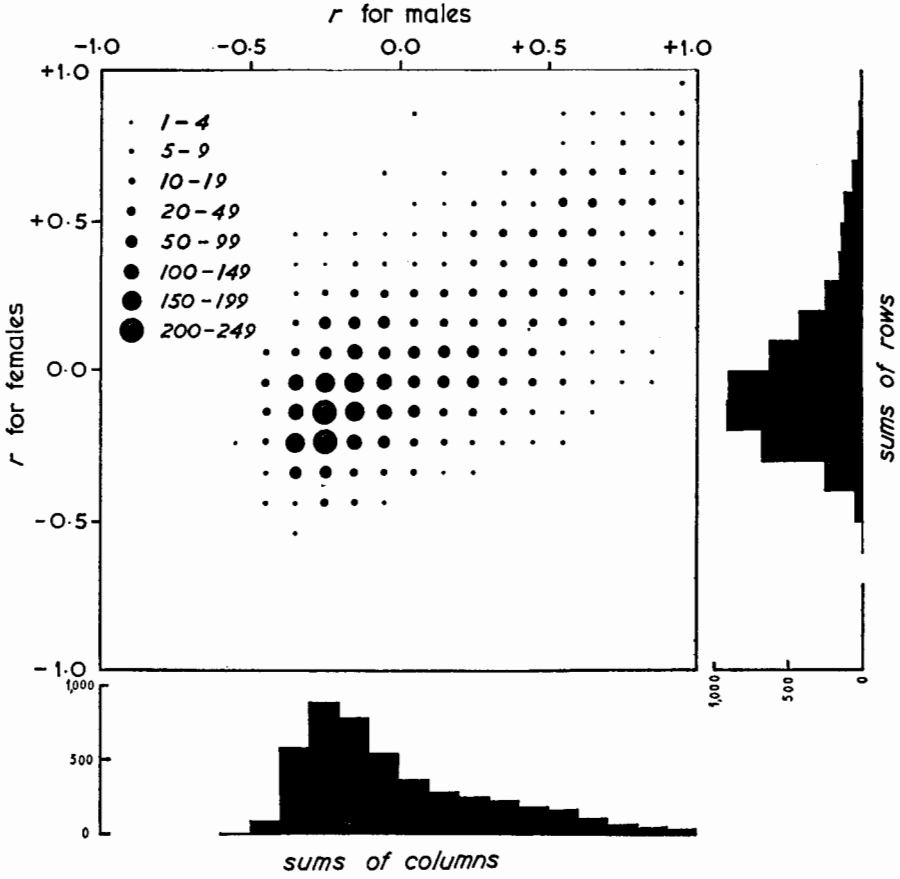

FIGURE 7-13

Two-way frequency distribution of similarity coefficients (in this case r*) for the comparison of similarity matrices based, respectively, on male and female characters in the* Hoplitis *complex. [Unpublished data from Michener and Sokal.] The abscissa and the ordinate indicate the magnitude of the correlation coefficient from males and females, respectively. The frequencies in each cell of the two-way frequency distribution are represented by solid circles. The size of each circle indicates the magnitude of the frequency according to the key shown in the upper left corner of the graph. The graph is based on 4,656 correlation coefficients. The correlation coefficient between the two variables is 0.71.*

OTU's and the others is relatively simply carried out. However, we are then faced with preparing a new dendrogram from the augmented similarity matrix. Such a procedure has two drawbacks. Having to recompute a cluster analysis of the similarity matrix is time consuming; second, since relationships would inevitably be changed to some degree,

the advantage of the stability of an analysis is to some degree negated. It is therefore quite important to make efforts to obtain reasonably complete and representative taxonomic groups before undertaking a revisional study by numerical taxonomy.

Another approach may be to set up a series of standard OTU's, against which newcomers may be compared. Such a method would locate the newly added OTU's in hyperspace with respect to the standards but would not necessarily locate them with respect to each other nor with respect to other OTU's previously studied. This is therefore only a stopgap measure, which cannot take the place of a complete analysis.