8

Identification and Discrimination

As we began Chapters 4 and 5, so we shall also begin this chapter by presenting, in Section 8.1, the form in which the data are given—in this chapter, for purposes of identification. We must, of course, already have groups of individuals or OTU's against which to identify an unknown, and these groups will normally be taxa. General considerations for identification and discrimination follow in Section 8.2, and sequential and simultaneous keys are considered in Sections 8.3 and 8.4, respectively. We conclude the chapter with a discussion of discriminant functions in Section 8.5.

Work on identification has not been as intensive in recent years as has been the work on classification. Thus much of the discussion that follows must be tentative and programmatic rather than definite. However, we hope that just as our earlier outline of procedures for classification (Sokal and Sneath, 1963) led to an increased development and improvement of such methods, so the sections that follow will stimulate biologists, mathematicians, and computer scientists to produce a theory and technology of taxonomic keys compatible with our present knowledge and capabilities. There are already signs that the field will advance swiftly.

8.1 THE IDENTIFICATION MATRIX

A data matrix arranged for purposes of identification may be called an identification matrix $\mathscr{I}$. It is shown in Table 8-1. It consists of a number of submatrices, that is, it is partitioned vertically into q blocks, each block representing a taxon $\mathbf{J} = (\mathbf{A}, \mathbf{B}, \ldots, \mathbf{J}, \ldots, \mathbf{Q})$. Within any block $\mathbf{J}$ are the individuals or OTU's that provide the information on the taxon, i.e., they are the sample of organisms that represent the taxon. These OTU's are numbered $\mathbf{a_J}, \ldots, \mathbf{j_J}, \ldots, \mathbf{t_J}$ within each block $\mathbf{J}$. The rows of the matrix represent the n characters $(1, 2, \ldots, i, \ldots, n)$ for the OTU's. A character state value in this matrix has three subscripts; thus $X_{i\mathbf{jK}}$ is the value for the ith character of the $\mathbf{j}$th OTU of the $\mathbf{K}$th taxon. One or more subscripts will be omitted when the meaning is clear. The matrix may, of course. be partitioned differently if the taxonomic rank of the taxa to be considered is changed; thus in studying a family, for example, one partition might be into tribes. another into genera. The $\mathscr{I}$ matrix may often be the same as the original data matrix (Section 4.1) except that the OTU's are reordered and grouped into taxa.

TABLE 8.1.
The identification matrix $\mathscr{I}$ and vector $\mathbf{u}$.

Characters	Taxa			
	OTU's in taxon **A**	OTU's in taxon **J**	OTU's in taxon **Q**	Unknown OTU
	$\mathbf{a_A}, \ldots, \mathbf{j_A}, \ldots, \mathbf{t_A}$	$\mathbf{a_J}, \ldots, \mathbf{j_J}, \ldots, \mathbf{t_J}$	$\mathbf{a_Q}, \ldots, \mathbf{j_Q}, \ldots, \mathbf{t_Q}$	$\mathbf{u}$
1	$X_{1\mathbf{aA}}, \ldots, X_{1\mathbf{jA}}, \ldots, X_{1\mathbf{tA}}$	$X_{1\mathbf{aJ}}, \ldots, X_{1\mathbf{jJ}}, \ldots$	$X_{1\mathbf{aQ}}, \ldots$	$X_{1\mathbf{u}}$
2	$X_{2\mathbf{aA}}, \ldots, X_{2\mathbf{jA}}, \ldots, X_{2\mathbf{tA}}$	$X_{2\mathbf{aJ}}, \ldots, X_{2\mathbf{jJ}}, \ldots$	$X_{2\mathbf{aQ}}, \ldots$	$X_{2\mathbf{u}}$
$\vdots$				
n	$X_{n\mathbf{aA}}, \ldots, X_{n\mathbf{jA}}, \ldots, X_{n\mathbf{tA}}$	$X_{n\mathbf{aJ}}, \ldots$	$X_{n\mathbf{aQ}}, \ldots$	$X_{n\mathbf{u}}$

On the right of Table 8-1 we have a column vector for the unknown OTU (an individual) to be identified, symbolized by $\mathbf{u}$. Its elements are character state values symbolized as $X_{i\mathbf{u}}$, where $i = 1, 2, \ldots, i, \ldots, n$, as before. Capital letters symbolize taxa to emphasize their resemblance to matrices rather than to vectors. Thus. though we may replace the taxon by character averages (for example), these are obtained by operating on an $n \times t_\mathbf{J}$ block of the $\mathscr{I}$ matrix; the vectors representing the averages are collected into a new matrix, in which column vectors represent taxa as averages of character values. Clearly, too, we may have many unknowns to identify, but at any one instant we normally have only one. Successive identifications thus formally entail replacing $\mathbf{u}$ by other unknowns in turn.

In most applications the matrix will not be partitioned horizontally; the same n characters will normally be recorded for all taxa. However, though n may be the number of characters studied in the whole numerical taxonomic study, we may discard some of them as being of little value in identification and to reduce unnecessary computation. Where it is pertinent we will use $m < n$ to show that n has been reduced to a smaller character set (this is a use of m different from that in Section 4.4, where it means the number of matches in an association coefficient, and is different from m used for the number of states of a character in various sections).

Since characters are quite properly weighted for identification, we require a symbol for this, and use $w_{i\mathbf{J}}$ for the weight of the ith character when testing $\mathbf{u}$ as a member of taxon **J**, or $w_{i\mathbf{JK}}$ when deciding between taxa **J** and **K**. This is principally used in discriminant analysis (Section 8.5). Over all characters the weights constitute a vector $\mathbf{w}$. Moreover, certain characters should be preferred over others because they are readily and constantly observable, so that an additional weight, $e_{i\mathbf{J}}$ (expressed as a vector $\mathbf{e}_{\mathbf{J}}$), may be given to symbolize ease of observation of character i in taxon **J**.

Missing values in the identification matrix may be coded NC, but they are of two types that may sometimes require separate symbols. These may simply be unrecorded values (e.g., petal color blue, but unrecorded) or they may be inapplicable (e.g., petal color when there are no petals). They need distinguishing, because a specimen with blue petals can be excluded as a member of a species without petals, but it could belong to a species whose petal color had not been recorded.

The identification matrix is often transformed into some other matrix before constructing a scheme for identification. There are two main types of transformed matrix. The first replaces the $t_{\mathbf{J}}$ columns of a taxon **J** by one or two columns representing some simplified summary of the character values, such as their means, ranges, standard deviations, and for 0,1 characters in particular, the proportion of OTU's with a given state. The second main form is a variance-covariance matrix (or a correlation matrix) between characters together with vectors of means, the starting point for discriminant analyses.

8.2 GENERAL CONSIDERATIONS

The objects of any identification scheme are ease and certainty of identification (Davis and Heywood, 1963). All other considerations are secondary. If one identifies an unknown specimen, this presupposes that one already has taxa with which to identify it. The form of the identification matrix shows this clearly. We distinguish therefore between classification in the sense of making classes, clusters, or taxa, and identification. The use of the word classification by many statisticians to mean identification is particularly confusing, and this is why we emphasize the point. There are some strategies that combine the two procedures, usually by successively "identifying" new individuals, but these also require criteria for deciding when identification with an existing class is unacceptable, so that new classes may be started (e.g., Ornstein, 1965; Rosen, 1967). These methods then become effectively cluster analyses, and we believe the distinction is a useful one.

The main methods used in identification are keys and discriminant functions. By far the commonest and most versatile are the former. Two differences between identification keys and classifications may be noted. Keys are not necessarily natural classifications in any of the usual senses. The divisions of the key may be quite arbitrary, as long as they are convenient for identifying specimens. Also, the same taxon can key out many times in different parts of the key; it need not have a unique position. Discriminant functions are more restricted in scope and much less often used. The various subdivisions of these methods are described later in this section, so we digress now to some general points about discriminatory characters.

We noted in Section 8.1 that characters are quite properly weighted for purposes of identification. There are two main approaches to calculating the appropriate weights, $w_{i\mathbf{J}}$. The most usual is based on the frequencies of various character states in different taxa but ignores correlations between characters. A detailed discussion is given by Ledley and Lusted (1959a). Since highly correlated characters tend to behave as a single character, this approach is likely to give overestimates of the probability that a given identification is correct (a good study of this is that of Mosteller and Wallace, 1964). The other approach considers the correlations between characters and is employed particularly in discriminant analysis. It is theoretically more powerful and precise.

Another, different form of weighting noted in the previous section is weighting according to ease of observation of different characters, $e_{i\mathbf{J}}$. Characters that are prominent, unlikely to be confused, and found in all specimens and during much of the life cycle (or in plants, throughout the year) are to be preferred. These weights. though unavoidably subjective in part, should also take account of the chance of loss of organs through damage or the cost of obtaining a given measurement.

In practice it is usual to reduce the original list of n characters to a smaller list of m characters, (the smallest effective number). The choice of characters for discrimination may be carried out in many ways. Inspection of the original tables of data after rearranging the columns to give the $\mathscr{I}$ matrix is the most usual (e.g. Steel, 1965; Moss, 1968a). A character that is invariant throughout is clearly useless, but unless the data being used are part of a larger study, such characters will have already been deleted. For two-state characters one can use the algebraic difference between the frequencies in two taxa of the 1 state symbolized as G by Sneath (1962); the most discriminatory characters have the highest values of G (positive or negative). This is a simple method that Hall (1965b) found useful in a botanical study, but nonadditive scoring (Section 4.8) causes difficulty.

Gyllenberg (1963) obtains the 0,1 characters most useful as discriminators (on the average) as follows. The proportion of the 0 or 1 values for character i, whichever is the greater, is noted for each taxon, and Gyllenberg calls this C. The sum of C over all q taxa, ΣC_i for character i, is then a measure of the value of i for separating groups. Characters that are least variable within taxa score highest, and $\frac{1}{2}q \leq \Sigma C \leq q$.

Next a separation figure, S_i, is calculated for the character, which is the product of the number of taxa, q_1, in which the character is predominantly 1 and of the number of taxa, q_0, in which it is predominantly 0 (using chosen cutoff levels such as 0.9 and 0.1). The value of S_i is greatest for characters that divide the taxa as nearly as possible into equal halves. The general usefulness of a character as a discriminator is indicated by the rank figure, R_i, which is $\Sigma C_i \times S_i$. Characters with the highest R_i are preferred for constructing identification schemes. It is often sufficient to calculate the values of S_i and an example of its use to select new tests in bacteriology is given by Lapage and Bascomb (1968).

Maccacaro (1958), Möller (1962a,b,c) and Jičín, Pilous, and Vašíček (1969) use rather similar methods but employ information statistics. This can be generalized to multistate characters, and the expression $\Sigma_{\mathbf{J}=1}^{q}[-(\Sigma_{g=1}^{m} p_{g\mathbf{J}} \log_2 p_{g\mathbf{J}})]$, where $p_{g\mathbf{J}}$ is the proportion of the gth of the m states of the character in the Jth taxon, may be useful. Niemalä, Hopkins, and Quadling (1968) give two methods for 0,1 characters. The first is to compute for each character the quantity $\log(q_1 + q_0)! - (\log q_1! + \log q_0!)$. The highest values are given by the characters that are the best separators. By an extension of this last formula they also obtain the m characters that are jointly the best (which are not necessarily those with the highest values when considered singly). The second method is to operate on the $\mathscr{I}$ matrix and to delete characters in turn, providing the deletion does not make any pair of taxa indistinguishable. The character states for the taxa are recoded as 1 and -1 (it is implied that the commoner state is used), and $A_i = |\Sigma_q X_{i\mathbf{J}}|$ is calculated for each character. The characters are then discarded in diminishing order of A_i (this deletes the least useful characters first) until the chosen number, m, remain.

Another way to rank characters is the method developed by Bonham-Carter (1967a), who does so by the magnitude of chi-square values. His null hypothesis is the independence of the taxa from marginal totals of the characters summed over all taxa. Several of the information-theoretic methods can also be adapted for this purpose (e.g., Estabrook, 1967; Bisby, 1970b).

The classificatory method of Lockhart and Hartman (1963) and association analysis (Williams and Lambert, 1959; Section 5.4) extract discriminatory characters in the course of constructing monothetic groups. Although most taxa are at least partly polythetic, one may find some character states that sharply distinguish any two taxa; that is, they are present in all members of one taxon and absent in all members of the other. There may be no single states of this kind, but it may be possible to distinguish the taxa by using several character states that occur with different frequencies in the two taxa. This latter situation, phenetic overlapping, is found in taxa that are fully polythetic (see Section 2.2). It is here that discriminant analysis (Section 8.5) is particularly valuable.

The minimum number of characters for discrimination is easy to calculate. No more groups can be distinguished than the product of the number of character states. Thus three characters, two of three states and one of four states, allow at the most the distinction of $3 \times 3 \times 4 = 36$ groups. In general: log (number of distinguishable groups) $\leq$ Σ(log number of character states). Actually, because of character correlations only rarely will the number of distinguishable groups be as large as the product of the number of character states. In practice many more characters are required than the theoretical minimum: examination of various dichotomous keys in the literature shows that a given character seldom serves as a convenient separator in more than one part of the key, so that one needs about as many characters as there are branches, even if one character (and not more) is used at each branch point. Since for a dichotomous key $q - 1$ branches are needed to separate q taxa, the ratio of characters to taxa, m/q, is usually over 1 and may be as high as 2 or 3 if the taxa are difficult to separate, or if the author wishes to make a very reliable key. The contrast between theory and practice is seen by the fact that the theoretically minimum number of characters required to separate an estimated ten million species of living organisms is only 24 two-state characters, whereas Munroe (1964) believes that about 500 characters would be needed in practice. Munroe's figure is probably an underestimate, but since it is not likely that ten million characters would be needed, this suggests that the relation of m to q is still poorly understood (see Ledley and Lusted, 1959a,b; Osborne, 1963a,b) and needs further study.

It should be noted that there are two possible errors in identification. First, an unknown may be identified as a member of taxon **J** when it should be identified with another taxon in the scheme, **K**. Second, an unknown is identified with **J** but belongs to a taxon outside the study entirely. Some schemes use a criterion for successful identification and include a provision for recording "no identification made" to guard against the second danger. But this possibility is still a serious danger, because such tests will not always work. If, for example, some very similar taxa were inadvertently omitted it is likely the characters that discriminate between them and the included taxa might not have been chosen. Identification schemes should therefore be comprehensive with regard to taxa, and limitations of age, sex, life stage, etc., must be clear. A third type of error is of course the exclusion of an unknown from any of the known taxa when in fact it is a member of one of these.

Identification methods overlap a good deal, but we divide them into two main types, the sequential and the simultaneous. The *sequential methods* are the usual diagnostic keys and certain related schemes like multiple entry keys. Sequential methods can be divided into monothetic and polythetic ones. The *simultaneous methods* include discriminant functions and also others where some measure of agreement over all characters is employed, so that the identification can be made at one step. The synoptic table is an informal device of this kind. In many of these methods the unknown is in effect placed in a phenetic space and its closeness to (or inclusion within) known clusters is determined; they can therefore be considered as phenetic distance models. But not all involve distances, so the broader term of simultaneous methods is preferred here. Simultaneous methods are almost always polythetic.

Discriminant functions are probabilistic by design, and any of the others can be made so. By probabilistic we mean that some measure is given of the likelihood that the identification of a given specimen is correct. The need for this is least at high ranks, where taxa are sharply distinct.

This aspect should be given more attention and identification schemes should wherever possible be thoroughly tested with specimens that were not used in their construction. Probabilistic considerations also enter in another way. Schemes can be devised that identify members of some taxa with higher probability than others. In some work the Bayesian approach, in which the probabilities of correct identification are highest for the commonest taxa, might be desirable on the grounds that occasional misidentification of a rarity was less serious than misidentification of common forms; but in other work a converse approach might be desirable. It may be noted that these probabilities are not necessarily related to the weights given to the characters, for in monothetic keys the effective weight at a given couplet is infinite, because all other characters are ignored at this division.

Yet another consideration of probability that will affect the identification procedures, especially in large scale screening, is the Bayesian consideration of the likelihood of a given taxon being found in nature. We are less likely to identify an unknown OTU **u** as belonging to taxon **J** if we know that only very few individuals belonging to **J** have ever been collected. These considerations, which have not so far been extensively applied to identification schemes, are relevant to both sequential and simultaneous procedures.

Identification schemes, like other algorithms, can handle only a certain limited number of taxa conveniently. If there are too many taxa they must be divided into several schemes, and a sequential strategy is then superimposed on that of the schemes themselves. In this the schemes are much affected by practical considerations—length or complexity, the number of characters demanded, etc.—all of which must be balanced against their success rate.

Computers will increasingly be used in this field as electronic data processing comes into use in systematics. The only practicable way of originally calculating discriminant functions is by computer. Some taxometric programs now provide lists of characters of high discriminatory value. We need, however, to distinguish two different uses of computers in this connection. First, the computer may make a key or discriminant function that can be printed and used independently. This may be their major use for some time to come. Second, the identification scheme may be stored in the computer so that it is used "on line". This is most promising for simultaneous methods with large collections of data (e.g., Goodall, 1968a; Lapage et al., 1970). We have pointed out elsewhere (Sokal and Sneath, 1966) that this will make acute the problem of standardizing descriptive terms throughout large taxa.

8.3 SEQUENTIAL KEYS

Sequential keys can be constructed according to any desired sequence of divisions of the set of taxa into successively smaller subsets. The first decision is whether the sequence is to follow the established taxonomic hierarchy, or whether the most efficient but probably quite artificial system is to be used. Since the objectives of identification are different from those of classification, there is no strong reason why the taxonomic hierarchy should be embodied in the key, although in large studies it may be convenient to set up successive keys based on selected rank categories. Thus in a family one may have a key to genera, and for each genus a separate key to species, but the key to genera need not show the subfamilies and tribes.

As noted earlier, a given taxon may occur many times at the tips of the key (though this may lead to the suspicion that it is not a very natural taxon). This is better than to construct the key so that the taxon occurs only once with many cross-references from other parts of the key (Davis and Heywood, 1963). Osborne (1963a,b) believes that such repetition (he calls keys of this type reticulated) is likely to be inefficient on mathematical grounds.

Although keys can have more than two alternatives at each step, the clarity of dichotomous keys is a considerable advantage, and we therefore restrict our discussion to them. Osborne discusses several aspects of the branching structure and how one may make the key as short and efficient as possible. A dichotomous key for q taxa has $q - 1$ branch points (unless taxa occur more than once at the tips). The number of furcations is thus the same however it is arranged (Figure 8-1), but if the OTU's are split off one at a time, using distinctive characters, then the key requires $q - 1$ different characters (Figure 8-1,*b*). Furthermore the average number of characters that must be examined to identify an unknown specimen is higher (and very much higher for large q), than if the paths bifurcate repeatedly (Figure 8-1,*a*). Osborne notes that the latter type of key is generally easiest to use, most rapid and most reliable. There are occasional exceptions (because it may sometimes be convenient first to dispose of a few highly distinctive taxa using characteristic features), but in general each division should be made on a character that as nearly as possible divides the taxa under consideration at that point into equal halves. This conclusion is also reached on grounds of information theory (Maccacaro, 1958; Rescigno and Maccacaro, 1961; Möller, 1962a,b,c) and underlies several of the methods for choosing diagnostic characters mentioned in the last section. With repeated branching one can key out 2^m taxa in m levels of the key, theoretically using only m characters. Osborne points out that if the chance of making a mistake in answering a question is the same for each character, this key will give fewest errors. These considerations may not be important in practice, however. The procedure of finding characters that give division into equal numbers of taxa could lead to unreliability at later branches. The distinctive characters may be less

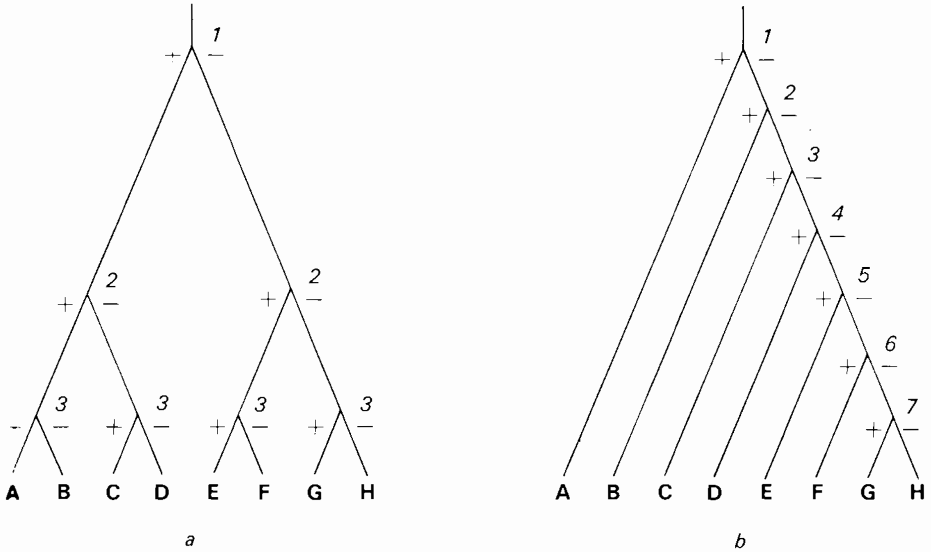


FIGURE 8-1
Two arrangements of a dichotomous key for eight taxa, **A** to **H**. The characters used are symbolized as *1* to *7*, each with two states, present (+) or absent (−). *a*, The paths branch repeatedly, and only the theoretical minimum of three characters is required. The number of characters to be examined to identify an unknown specimen is three. *b*, One taxon is split off at a time on the basis of a distinctive character. Now seven characters are required and the average number that must be examined is 28/8 = 3.5.

liable to errors than the others. Also, as noted in the last section, it is rare that anything like the theoretical minimum will be sufficient.

The "bracket" and the "indented" keys are the most common forms, though the terminology is confusing since either can be indented; they differ mainly in typographic layout and are illustrated by Mayr (1969a, p. 278). The bracket key is the most generally useful, because it can be worked in reverse so that one can retrace a false lead. We do not describe the details of making keys and refer the reader to the articles of Ainsworth (1941), Voss (1952), Metcalf (1954), Stearn (1956), and Mayr (1969a). It may be quite difficult to make a good key that is simple, short, and efficient. Blackwelder (1967a) notes that at high taxonomic ranks the choice of characters may be very difficult, because some exceptions are likely to occur with most characters (for example, there are arthropods without legs).

Monothetic Sequential Keys

Monothetic sequential keys have a single contrasting statement in each couplet, referring to only one character, to be answered (in principle) by a single yes or no. They are, of course, vulnerable to exceptions, as are all monothetic schemes.

Osborne (1963a) suggests that the characters should be scaled on an integer scale of 1 to 4 so chosen that the compiler can be fairly sure the user will recognize the correct integer from his examination of the specimens to be identified. With this scheme the key will be reliable if every taxon differs from every other by at least a score of 2 upon one or more of the characters.

Several on-line computer schemes are now being developed. Rypka et al. (1967) and Rypka and Babb (1970) have incorporated Gyllenberg's scheme (Section 8.2) into a computer program with some additional modifications. They first compute Gyllenberg's S for each character and select the character with the highest value. This is the best initial separator; they then choose as the next character the one giving the highest joint S with the first. The third character chosen as divisor is that with highest joint S with the previous two, and so on. The joint S is calculated as follows

$$S_{\text{joint}} = \tfrac{1}{2}[q^2 - (q_a^2 + q_b^2 + \ldots + q_z^2)]$$

where q is the total number of taxa, and $q_a + \ldots + q_z$ are the numbers of taxa possessing the z various unique combinations of 0,1 states in the n characters considered. With binary characters as here, $z = 2^n$, but many of the combinations will probably not occur. Rypka and his coworkers note that one can compute directly the best pairs, triples, quadruples, etc., of characters, using all possible combinations of the characters, but this makes heavy demands on computing time. Although intended to identify bacteria, this method is likely to make insufficient provision for exceptional isolates, and may be more suited to higher organisms.

Multiple entry keys are another group of keying schemes that are generally monothetic and sequential. A good description of one form is given by Leenhouts (1966). Each taxon is listed against each character arranged under the two leads, for example:

Leaflets	Taxa
(a) Entire	**A B D (F) (G*)**
(b) Not entire	**C E (F) (G*)**

Taxa that can possess either state are in parentheses, and those whose state is unknown are given also asterisks. To use the key one chooses any character and excludes taxa that do not agree, then chooses another character, and continues until only one taxon is left. The principle is readily applicable to superimposed punched cards ("peek-a-boo" systems), of which a good example is a key to the families of flowering plants produced by Hansen and Rahn (1969). Each card represents a character with a fixed position on it for each taxon; selections of cards are superimposed until only one perforation remains, indicating the required taxon.

For example, overlapping cards No. 8 (tendrils present), No. 53 (flowers zygomorphic), and No. 133 (carpel 1) leaves only family—No. 29 (Papilionaceae). The principle is also readily applied to on-line computing, and descriptions of systems that follow a similar strategy have been given by Boughey, Bridges, and Ikeda (1968), Goodall (1968a), and Morse (1971). An advantage of multiple entry keys is that the user may employ any of the key characters that are available on the specimen; with the usual keys he must have the very characters required for each couplet in turn.

Although monothetic groups are seldom required in taxonomy, monothetic cluster methods (e.g., Maccacaro, 1958; Williams and Lambert, 1959; Lockhart and Hartman, 1963) may be useful for constructing keys because they yield characters that are likely to be near optimal for key making. For this purpose Gower (1967a) has suggested subdividing on the character that, at each dichotomy, maximizes the multiple correlation between it and all previously unused characters.

Polythetic Sequential Keys

Polythetic sequential keys are keys in which at least some couplets consist of several statements about different characters. These are the commonest form of taxonomic key. The reasons for using several characters are threefold: (1) one or more characters may be unobservable on some specimens (for example, damaged specimens or plants not in flower); (2) there are a few taxa (or individuals) exceptional in the most readily observed characters; and (3) the user may make a mistake in deciding about a character. In each case the other characters help the user to decide which branch to take. The basic idea is that of the majority vote; unless the key says otherwise, the user is best advised to follow the majority verdict (we note, however, that this is rarely stated explicitly, and some workers intend the first character to be more important than the others). In other words, the user gives preference to the alternative most similar to his specimen, but no one character is essential. The strategy is thus basically polythetic, consisting of a comparison of the specimen with the statements in the couplet, followed by choice of the best match.

This procedure affords many advantages, not the least of which is a better prospect of accurate probabilistic estimates. It does have the disadvantage of being somewhat less clear-cut. Also, the procedural rules we have just mentioned are not self-evident. Polythetic keys, therefore, require some formalizing. With monothetic keys, at any branch point the single character has decisive value (all others have zero weight). In polythetic keys the characters require weights, either differential weights or else a specific statement that they are equal. Hall (1965b) gives an illustration of a key in which such weights are attached to each character in the couplets. These weights are not only the statistical discriminatory weights, w_i, but also the ease of observation values, e_i. The latter cannot be so readily estimated as the former, for they depend a good deal on the experience of the user, so the practical problems of using keys are clearly relevant here. For example, are the difficulties due more to poor character descriptions (perhaps they are described in highly technical language without a diagram, one of the biggest stumbling blocks for the inexperienced) or to difficulty of observation, as when special microscopic preparations are essential?

Several methods have been developed for producing keys automatically by computer (Morse, 1968, 1971; Pankhurst, 1970a,b; Hall, 1970). Pankhurst gives extensive details of the technique he uses. The key is constructed from a table of character state values for the taxa. These values can be qualitative or quantitative, and provision is made for missing data, although if there are many missing entries this greatly increases the difficulty of making a key. The number of characters at each branch point can be controlled. Either an "indented" or "bracketed" key can be produced (either of them with or without typographic indentation) in a form ready for use (Figure 8-2). When a taxon keys out, all remaining distinctive characters for that taxon are furnished by the computer program; they are useful as a check on the identification. The user can allocate weights, w_i or e_i, to each character, or he can weight the taxa to obtain short identification routes to taxa of his choice (such as commoner ones). Taxa are allowed to key out several times if this is the only way to get a key, and highly distinctive taxa key out early (but this is allowed to happen only rarely because it interferes with the attempt to optimize the key). The basic procedure is to find characters that divide the taxa into equal halves, with preference given to dichotomies over polychotomies. Pankhurst uses a separation function $F = F_1 + F_2$, where

$$F_1 = (k - 2)^2 \text{ and } F_2 = \sum_{\mathrm{b}=1}^{k} |1 - (q_{\mathrm{b}}\, k/q_{\mathrm{a}})|$$

and where, at a given branch point, a, with q_{a} taxa under consideration, there are k subgroups each containing q_{b} taxa. The divisions on different possible characters are tested, and that with minimum F is preferred subject to certain accessory conditions, such as that characters with high w_i are considered first. The methods of Morse and Hall use rather similar principles. Morse (1971) makes special provision for characters that are variable within taxa. The character for the initial couplet is the one giving the highest value of $DV \times \exp\{CV\}$ provided the character is not unknown or inapplicable for any of the taxa under consideration. DV is calculated as $2q_Tq_F + \frac{1}{2}q_V(q_T + q_F)$ where q_T, q_F, and q_V are the numbers of taxa for which the answer to the first lead is true, false, or variable, respectively. The value CV is a "convenience value" given to the character by the user. This procedure is repeated for subsequent branches of the key.

Computer-made keys are generally as short or shorter than manual ones, and if appropriate values of w_i are chosen they appear comparable to manual keys in quality. It is likely that in the near future they will become superior to those

```
1    Stem 0–10 cm                                                                          2
  2  Sterile rosettes absent, capitula more than 3 cm                        17 J. fontqueri
  2  Sterile rosettes present, capitula up to 3 cm                                         3
    3  Capitula obconical, involucral bracts lax, patent or recurved          15 J. humilis
    3  Capitula subglobose, involucral bracts appressed                      16 J. taygetea
1  Stem more than 10 cm                                                                    4
  4  Pappus shorter than achene                                             11 J. polyclonos
  4  Pappus longer than achene                                                             5
    5  Involucral bracts lax, patent or recurved                                           6
      6  Capitula more than 3 cm                                                           7
        7  Stem leafy throughout                             10b J. mollis. ssp. moschata
        7  Stem leafy at base                                                              8
          8  Involucral bracts lanceolate                                       10 J. mollis
          8  Involucral bracts linear                                      14 J. glycacantha
      6  Capitula up to 3 cm                                                               9
        9  Stem woody at base                                                6 J. albicaulis
        9  Stem herbaceous                                                                10
        10  Basal leaves subglabrous above, tomentose beneath, achene more than
            5 mm                                                             9 J. eversmanii
        10  Basal leaves puberulent above, tomentose beneath, achene 2–5 mm  12 J. ledebouri
    5  Involucral bracts appressed                                                        11
    11  Basal leaves subglabrous above, tomentose beneath                                 12
      12  Distal crown of achene inconspicuous                                            13
        13  Capitula subglobose                                               8 J. cyanoides
        13  Capitula hemispherical                                        13 J. consanguinea
      12  Distal crown of achene conspicuous                                              14
        14  Rhizome absent                                               2 J. stoechadifolia
        14  Rhizome present                                             3 J. tzar-ferdinandi
    11  Basal leaves arachnoid tomentose                                                  15
      15  Sterile rosettes present                                                        16
        16  Basal leaves pinnatifid, capitula obconical                         4 J. pinnata
        16  Basal leaves entire, capitula hemispherical                     7 J. kirghisorum
      15  Sterile rosettes absent                                                         17
        17  Stem woody at base, basal leaves entire                        1 J. linearifolia
        17  Stem herbaceous, basal leaves pinnatifid                          5 J. tanaitica
```

FIGURE 8-2
A computer generated key of an "indented" type for European species of *Jurinea* (Compositae). The figure has been arranged to reflect, for the most part, the format a computer line-printer would adhere to, though in some details (typeface and line width), line-printer output would differ. [From Pankhurst (1970b).]

generally made by hand, and can be made even when the number of taxa is otherwise discouragingly large. It is, however, necessary to provide considerable amounts of accurate data in the form of the matrix of taxa and characters, though this would generally be easy after a numerical taxonomic study. It is especially important for the range of within-taxon variation to be known.

Estimates of the probability of correct identification are likely to be more accurate with several characters to a couplet than one. The methods of estimating these are discussed in the next section, for they are basically the same as for simultaneous keys (but on a restricted character set). If the characters are quite few, it may be feasible to do direct counts of OTU's with different character combinations and use these as rough estimates of the underlying natural phenetic distributions. Because the manual testing of keys is a laborious business, it would be useful to have a computer program that would generate hypothetical specimens by a Monte Carlo process from plausible frequency distributions of the character states. It could then test computer-made keys for their success rating and also pick out taxa that are not readily separable, which need further attention. Polythetic sequential keys should not list more characters than are necessary in practice, or much of their convenience will be lost.

Any sequential key can be made probabilistic, though this is rarely done. The main attempts have been made by Möller (1962a,b,c) and Hill and Silvestri (1962). Each tip of the key has an associated probability that a specimen that keys out to this position will be identified correctly. These probabilities should be as close to 1.0 as possible. A major problem is the accurate estimation of the probabilities: with monothetic keys one would expect that errors of estimate would accumulate rather readily. It is likely that Monte Carlo methods would have to be used in the way mentioned above.

8.4 SIMULTANEOUS KEYS

Simultaneous keys are those in which the unknown is compared in turn with all the taxa in the hope of obtaining an unambiguous identification with one of them in a single step. Their form is most often a table of m characters against the q taxa in which entries are the typical or commonest values for the taxa. The vector $\mathbf{u}$ of the unknown is compared in turn with each column, and the taxon with which it shows closest agreement is taken as the correct identification. The underlying concept is thus polythetic, and a simultaneous key is formally the same as a multiple branch point in a sequential polythetic key. Indeed, in any large study it becomes necessary to adopt a sequential strategy by breaking the full table into sections and to make the identification first to the major groups of taxa, and then to individual taxa. This is because too large a table is inefficient, as many of the characters are redundant for any one attempted identification; it is therefore best to use successive small tables. A good example is the work of Cowan (1965) and Cowan and Steel (1965), where a table is used to identify two major groups of genera, and for each such group a separate table is provided to carry identification to the genus. Such a scheme is, of course, virtually a multiple choice sequential polythetic key, but with numerous characters.

Any resemblance measure can be used to assess the best match. The usual one (as in Cowan and Steel's work) is the number of agreements for 0,1 characters. Cowan and Steel note, however, that there are difficulties with this simple method (which is analogous to using S_{SM} with equal weighted characters). Among these is the problem that for some characters the 0 states may have dubious significance; they may indicate a clear-cut negative or that chemical tests may not have been performed properly. And of course the characters are not explicitly weighted. Corlett, Lee and Sinnhuber (1965) have used this method in a computer-based scheme where a punched card with 19 test results is fed in to afford identification of a bacterium. A similar computer method is described by Walker et al. (1968) for pollen grains. Elimination of taxa as possible answers must be made on some chosen low value of the resemblance measure. Increased power would come from giving each character value X_{iu} a weight w_{iJ} and an ease of observation value e_{iJ}, so that during computation the contribution to the resemblance due to X_{iu} was multiplied by $w_{iJ} \times e_{iJ}$; what is effectively weighting of this kind is used in some of the methods described below and in the next section.

A related concept is that of giving each taxon a limiting envelope and treating it as if it occupied a definite volume in A-space. An unknown is then identified with the taxon closest to it. Furthermore, one can tell if the unknown is outside any taxon or is intermediate between two taxa. This concept is mainly associated with discriminant analysis but need not be restricted to it. Any phenetic space can be used, and angular measures of resemblance can be treated as great circle distances on a unit hypersphere (see Firschein and Fischler, 1963). Ordinary Euclidean distances are easier to handle, however. This model therefore extends the idea of a central position of a taxon to include also a measure of its size, most readily as the measure of the radius of a hypersphere. This is satisfactory if the taxa are roughly hyperspherical, but if they are markedly elongated because of pronounced correlations between characters, then discriminant analysis is better (discriminant analysis effectively makes the taxa as nearly hyperspherical as possible in a transformed phenetic space). The model will break down if intermediate forms are very numerous, so it is assumed that they are relatively uncommon.

Problems for which this model is suited occur in bacterial taxonomy, and Gyllenberg (1964, 1965b) has proposed a detailed scheme. Examples of its use are given by Gyllenberg and Rauramaa (1966). Gyllenberg actually used correlation coefficients, and also reduced the A-space to three dimensions by principal component analysis, but here we describe a more general form.

A taxon is defined by the coordinates of its centroid ($\bar{\mathbf{x}}_{\mathbf{J}}$) and by a radius $r_{\mathbf{J}}$. Different measures for $r_{\mathbf{J}}$ have been mentioned in Section 5.2, including that suggested by Gyllenberg, which is twice the root mean square of the distances of the OTU's of the cluster from the centroid, and which is likely to overestimate the effective radius. It is preferable to determine a radius empirically as described in Section 5.2, such that it encloses a chosen percentage of OTU's. The identification matrix is thus converted into a new matrix, $\mathscr{L}$. This has m rows and q columns, recording the centroids as the mean value of the characters within each taxon, together with an additional row vector giving the radii of the taxa.

The distance between the unknown **u** and the centroid of each taxon is calculated. If the distance of **u** from the centroid of a taxon **J**, $d_{\bar{\mathbf{x}}_{\mathbf{J}},\mathbf{u}}$, is less than $r_{\mathbf{J}}$, the unknown lies within that taxon. If the unknown lies outside any taxon it is recorded as unidentified (or possibly as an intermediate if it lies between two taxa). If **u** lies within only one taxon it is identified as belonging to it. Some taxa may overlap, and **u** may then lie within the hyperspheres of two or more taxa. Gyllenberg suggests that the unknown is then best identified with the taxon **J** for which the ratio $r_{\mathbf{J}}/d_{\bar{\mathbf{x}}_{\mathbf{J}},\mathbf{u}}$ is greatest. This is not necessarily the same as the taxon whose centroid is nearest to

u, as can be seen in Figure 8-3 for the unknowns marked $\mathbf{u}_2$ and $\mathbf{u}_3$. This figure also illustrates the other points of the scheme.

Clearly the system will break down if there is considerable overlap between hyperspheres, perhaps necessitating reclassification. Overlap is readily found by testing if any distance between centroids is less than the sum of the appropriate radii. The system is likely to be satisfactory if the taxa are approximately hyperspherical, that is, if character correlations are not, on the average, great.

The center of a taxon is best represented by the centroid, although other central measures can be used (Section 5.2). The hypothetical median organism of Liston, Weibe, and Colwell (1963) has been used as the center of taxa by Bogdanescu and Racotta (1967), and identifications were made by calculating distances from these. Hutchinson, Johnstone, and White (1965) used as the cluster center the OTU with minimal variance of distances to other members of the cluster. We expect that in spaces of high dimensionality other central measures like the centrotype would also be suitable.

In models not explicitly conceived as distance models, but closely analogous, there have been several attempts at calculating probabilities of correct identification.

Macnaughton-Smith (1965) suggests for 0,1 data a criterion for identification that appears to have several advantages. For each taxon **J**, the constant $C_{\mathbf{J}}$ is calculated:

$$C_{\mathbf{J}} = (n - 1)\log t_{\mathbf{J}} - \sum_{i=1}^{n} \log(t_{\mathbf{J}} - t_{\mathbf{J},1i})$$

where

$$t_{\mathbf{J},1i} = \sum_{j_{\mathbf{J}}=1}^{t_{\mathbf{J}}} X_{i\mathbf{j}\mathbf{J}}$$

Also in each taxon one calculates for each character the quantity $A_{i\mathbf{j}}$

$$A_{i\mathbf{j}} = \log(t_{\mathbf{J}} - t_{\mathbf{J},1i}) - \log t_{\mathbf{J},1i}$$

For an unknown, **u**, one calculates for each taxon in turn the sum of C and all the A_i's that refer to those characters scored 1 in the unknown. Identification is with the taxon for which this sum is least. The logarithm of the probability of misclassification is proportional to this sum. The arbitrary choice of zero for log 0 may be needed to avoid indeterminacy.

Goodall's deviant index (Goodall, 1966b) can also be used to estimate the probability of correct identification, as can his probabilistic similarity index (Goodall, 1964, 1966c).

Considerable success has been achieved in the difficult field of bacterial identification by using a method based on conditional probabilities of 0,1 characters (Dybowski, Franklin, and Payne, 1963; Dybowski and Franklin, 1968; Lapage

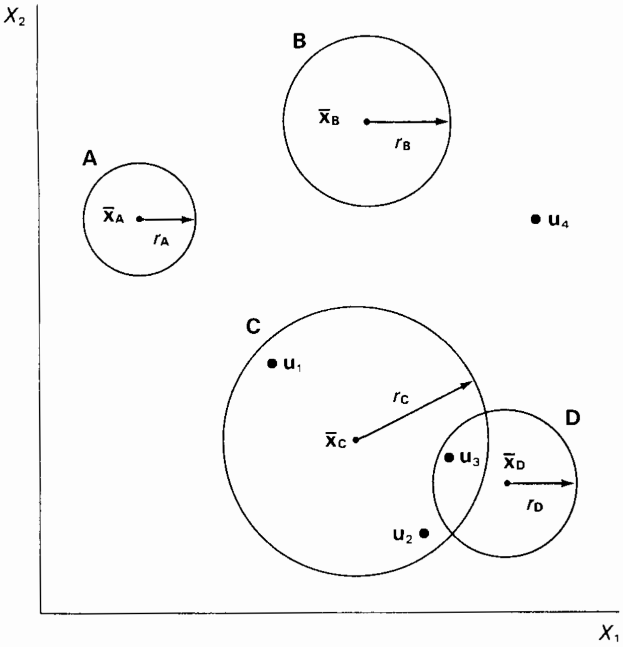


FIGURE 8-3

Identification as a process in A-space. The four taxa, **A**, **B**, **C**, and **D**, are represented by circles. They have centroids on the two character axes X_1 and X_2 shown by the central dots, $\bar{\mathbf{x}}_\mathbf{A}$, $\bar{\mathbf{x}}_\mathbf{B}$, $\bar{\mathbf{x}}_\mathbf{C}$, and $\bar{\mathbf{x}}_\mathbf{D}$ and dimensions shown by the radii of the circles $r_\mathbf{A}$, $r_\mathbf{B}$, $r_\mathbf{C}$, and $r_\mathbf{D}$.

An unknown $\mathbf{u}_1$ lies within circle **C** and no other, and is identified with **C**. The unknown $\mathbf{u}_2$ would also be allocated to **C**, although it is closer to the center of **D** than the center of **C**. The unknown $\mathbf{u}_3$, which is within both circles **C** and **D** would be regarded as an intermediate form, or else allocated to **C** by the ratio rule given in the text. This is because the ratio of $r_\mathbf{C}$ to the distance $\mathbf{u}_3$ to $\bar{\mathbf{x}}_\mathbf{C}$ is about 1.4, greater than the ratio of $r_\mathbf{D}$ to the distance of $\mathbf{u}_3$ to $\bar{\mathbf{x}}_\mathbf{D}$ (about 1.1). The unknown $\mathbf{u}_4$ is outside any circle and remains unidentified.

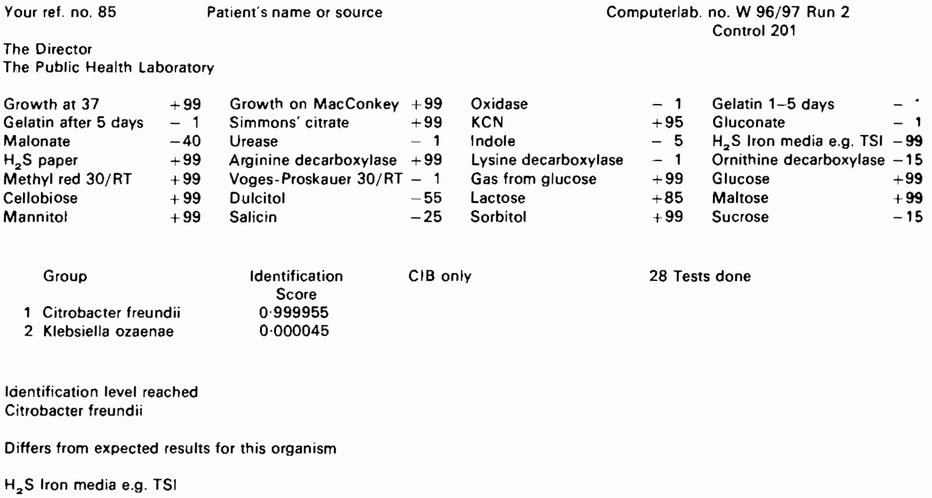
et al., 1970). A matrix $\mathscr{P}$ is stored in the computer that contains the proportion of state 1 for each of the m characters (mostly biochemical tests) in the q taxa. The entries $p_{i\mathbf{J}}$ lie between 0 and 1, but they are never set exactly to 0 or 1 for two reasons: (a) some exceptional bacterial strains must always be expected, as well as occasional mistakes in performing tests; and (b) values of 0 or 1 will rule out a possible identification completely if an atypical result occurs because this leads to multiplication by zero in the process described below. In practice limiting values of 0.01 and 0.99 are suitable. The basic logic is as follows: if in taxon **J** the proportion of state 1 of a given character h is, say, 0.2, then for that character the probability that an unknown that scores 1 belongs to taxon **J** is taken as $p_{h\mathbf{J}}$, which here equals 0.2. If the unknown scores 0, it is taken as $1 - p_{h\mathbf{J}}$, here 0.8. Similarly if a second character, i, is considered, with $p_{i\mathbf{J}}$ of 0.7, then the probabilities associated with state 1 and 0 are taken as 0.7 and 0.3 respectively. On considering both characters the probabilities are multiplied, so that in this example the probability for an unknown with the character states 1 and 1 is 0.2 × 0.7 = 0.14. The highest joint probability is given by an unknown possessing the majority states (in this example 0 and 1 respectively for characters *h* and *i*, giving 0.8 × 0.7 = 0.56).

The unknown is therefore compared with each taxon in turn and the individual probabilities are multiplied together for as many characters as are available, to obtain *L*, the joint likelihood values:

$$L_{\mathbf{J}} = \prod_{i=1}^{m} |X_{i\mathbf{u}} + p_{i\mathbf{J}} - 1|$$

The above formula assumes independence of characters.

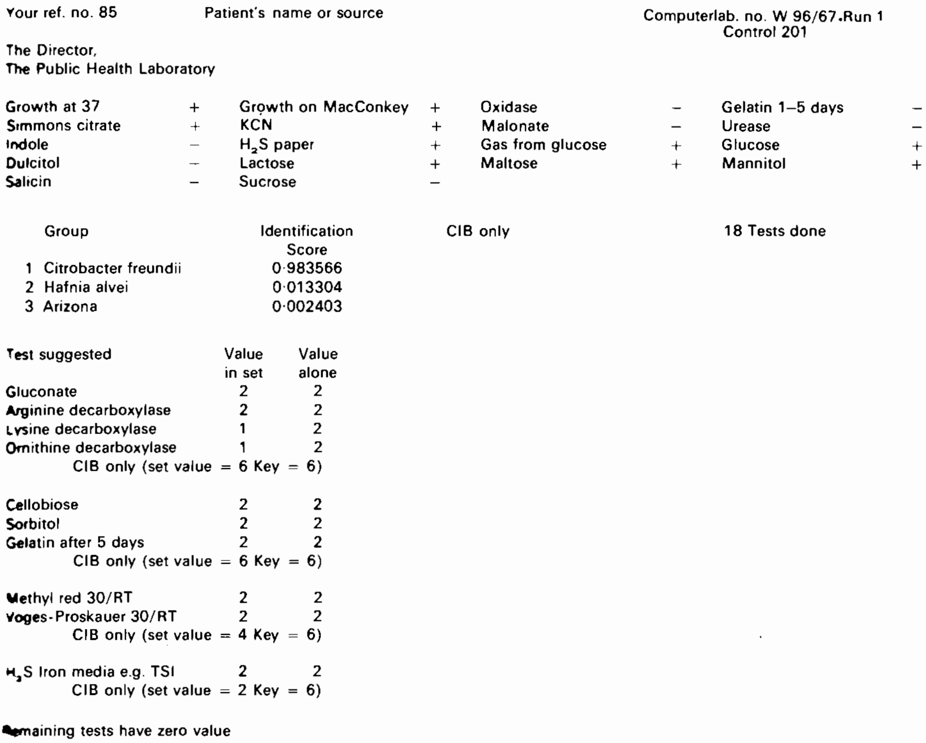
As the number of characters is increased the joint likelihood becomes vanishingly small for a misidentification. For a correct identification it also falls, but more slowly. To compensate for this Lapage et al. (1970) calculate $L_{\mathbf{J}}/\Sigma_{\mathbf{J}=1}^{q} L_{\mathbf{J}}$, and call this the *identification score*, which seems superior to earlier proposals by Dybowski and Franklin (1968). If the score reaches a sufficiently high level, such as 0.999. for the comparison of **u** with one taxon, this is accepted as a successful identification. If this level is not reached the program prints out the most likely candidates, and also valuable information on what tests should be made next in order to clinch the identification if possible. Figures 8-4 and 8-5 show examples of computer output.



```
Your ref. no. 85             Patient's name or source                     Computerlab. no. W 96/97 Run 2
                                                                                       Control 201
The Director
The Public Health Laboratory

Growth at 37          +99   Growth on MacConkey +99   Oxidase               – 1   Gelatin 1–5 days       – 1
Gelatin after 5 days  – 1   Simmons' citrate    +99   KCN                   +95   Gluconate              – 1
Malonate              –40   Urease              – 1   Indole                – 5   H2S Iron media e.g. TSI –99
H2S paper             +99   Arginine decarboxylase +99 Lysine decarboxylase – 1   Ornithine decarboxylase –15
Methyl red 30/RT      +99   Voges-Proskauer 30/RT – 1 Gas from glucose      +99   Glucose                +99
Cellobiose            +99   Dulcitol            –55   Lactose               +85   Maltose                +99
Mannitol              +99   Salicin             –25   Sorbitol              +99   Sucrose                –15


    Group                         Identification       CIB only                  28 Tests done
                                      Score
  1 Citrobacter freundii            0·999955
  2 Klebsiella ozaenae              0·000045


Identification level reached
Citrobacter freundii

Differs from expected results for this organism

H2S Iron media e.g. TSI
```

FIGURE 8-4

Computer identification: complete identification. The unknown has been identified as *Citrobacter freundii* with probability of over 99.99 percent. The next alternative, *Klebsiella ozaenae*, has a probability of less than 0.01 percent. The percent values of $p_{i\mathbf{J}}$ for *Citrobacter freundii* on the 28 tests done, the results (+ or −) found with this unknown, and an aberrant test result, have also been shown. The figure has been arranged to reflect the format of a report such as a computer line-printer generates Actual line-printer output would differ in some details. [From Lapage et al. (1970).]



```
Your ref. no. 85          Patient's name or source                       Computerlab. no. W 96/67.Run 1
                                                                                     Control 201
The Director,
The Public Health Laboratory

Growth at 37        +    Growth on MacConkey  +    Oxidase              -    Gelatin 1-5 days    -
Simmons citrate     +    KCN                  +    Malonate             -    Urease              -
Indole              -    H2S paper            +    Gas from glucose     +    Glucose             +
Dulcitol            -    Lactose              +    Maltose              +    Mannitol            +
Salicin             -    Sucrose              -

    Group                     Identification         CIB only                      18 Tests done
                                  Score
  1 Citrobacter freundii        0·983566
  2 Hafnia alvei                0·013304
  3 Arizona                     0·002403

Test suggested              Value     Value
                            in set    alone
Gluconate                     2         2
Arginine decarboxylase        2         2
Lysine decarboxylase          1         2
Ornithine decarboxylase       1         2
        CIB only (set value = 6 Key = 6)

Cellobiose                    2         2
Sorbitol                      2         2
Gelatin after 5 days          2         2
        CIB only (set value = 6 Key = 6)

Methyl red 30/RT              2         2
Voges-Proskauer 30/RT         2         2
        CIB only (set value = 4 Key = 6)

H2S Iron media e.g. TSI       2         2
        CIB only (set value = 2 Key = 6)

Remaining tests have zero value
```

FIGURE 8-5

Computer identification: the unknown has not given an identification score that is high enough, although the most probable answer is *Citrobacter freundii*. Four sets of new tests are suggested, and the user may then perform any or all of the four. Commonly the first set is sufficient. The relative value of the new tests is also indicated (for details see the original article). The figure has been arranged to reflect the format of a report such as is generated by a computer line-printer. Actual line-printer details would differ. [From Lapage et al. (1970).]

The scheme implemented by Lapage and his colleagues is now receiving extensive testing and has shown itself to be extremely powerful for identifying bacteria of medical importance. This is despite the fact it does not take character correlations into account, does not use a criterion to exclude misidentification of strains of taxa not represented in the matrix (i.e., there is in effect no critical radius of the taxa if the system is viewed as analogous to a distance model), and may be sensitive to vigor and pattern differences. The number of characters required for a high percentage of identifications is about 30, but this is quite economical since it represents a ratio of m/q of about 0.5, whereas with conventional methods the ratio is about 1. Very few misidentifications occur, and the identification rate appears satisfactory in view of imperfections in the present classification of bacteria (and consequently in the $\mathscr{P}$ matrix) and the frequency of aberrant bacterial strains in nature (for further discussion on these points see Sneath, 1969a, 1972).

Simultaneous keys are well suited for use on-line with a computer, as the data tables can be easily stored and the computations swiftly made. Hall (1969a) describes a modification that makes provision for excluding numerous very unlikely possibilities, thus increasing the speed of identification. Simultaneous keys are less useful in printed form, because matching of the unknown on the columns is troublesome. Several mechanical devices have been suggested for use with them (Cowan and Steel, 1960, 1965; Olds, 1970), and "peek-a-boo" punched cards can also be adapted to them (Yourassowsky et al., 1965); such modifications overlap with sequential techniques described in the previous section.

Other possibilities in simultaneous identification methods are the use of automatic scanning devices, the output of which is discussed in Sections 3.3 and 3.4 These may one day allow identification directly from the specimen. There is also rapid advance in automated methods of biochemical analysis, which could be coupled to an on-line identification scheme.

Related to simultaneous identification techniques are programs developed by Lance, Milne, and Williams (1968), which take hierarchical classifications or ordinations (and the data matrices underlying these) and output mean differences in desired characters for any specified groups. Although much information could be obtained as a by-product of the classificatory procedures, Lance, Milne, and Williams recommend that it be done as a separate run inasmuch as obtaining all the possible comparisons would be far too time-consuming and produce excessive printed output. Also it is impossible to know which particular comparisons will be of interest until the classifications have initially been obtained and examined by the investigator.

8.5 DISCRIMINANT ANALYSIS

In previous sections of this chapter we have mentioned the idea of weighting characters for identification. When the weighting is done in a manner that maximizes the probability of correctly identifying unknown specimens from a few close or overlapping taxa it leads to the branch of multivariate statistics that we discuss in this section. Most of these methods can be viewed as extensions of taxon distance models, such as the model illustrated in the last section, where the character axes have been transformed.

Discriminant Functions

A linear discriminant function is a linear function $\mathbf{z}$ of characters describing OTU's that weight the characters in such a way that as many as possible of the OTU's in one taxon have high values for $\mathbf{z}$ and as many as possible of another have low values. so that $\mathbf{z}$ serves as a much better discriminant of the two taxa than does any one character taken singly. The n characters are almost always reduced to a smaller set.

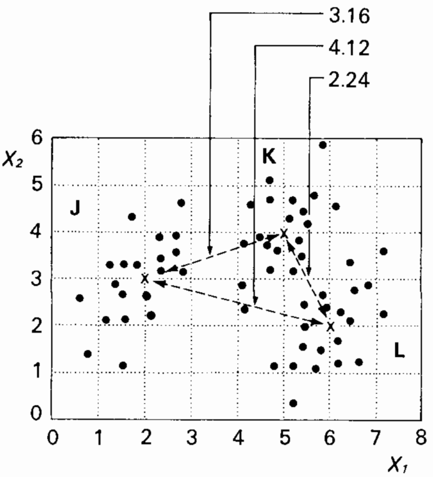


FIGURE 8-6
Three clusters of OTU's representing three taxa, **J**, **K**, and **L**, for two character axes X_1 and X_2. The crosses represent the centroids, which are $\bar{\mathbf{x}}_{\mathbf{J}} = 2.0, 3.0$; $\bar{\mathbf{x}}_{\mathbf{K}} = 5.0, 4.0$; and $\bar{\mathbf{x}}_{\mathbf{L}} = 6.0, 2.0$. The pooled variance-covariance matrix **W** is:

		Characters	
		1	*2*
Characters	*1*	0.5	0.4
	2	0.4	1.0

and the inverse matrix $\mathbf{W}^{-1}$ is:

$$\begin{array}{rr} 2.941 & -1.176 \\ -1.176 & 1.471 \end{array}$$

The Euclidean distances between the centroids are shown.

m. The function is also such that it has maximal variance between groups relative to the pooled variance within groups.

As originally described (Fisher, 1936), the discriminant function was applied to two taxa, and was later generalized to many taxa. It is calculated from the pooled variances and covariances between the *m* characters within each taxon. In its original form this is a weighted average of the variances and covariances of the characters in taxa **J** and **K**. When generalized, the variances and covariances for all the taxa being considered are pooled. In addition, the means of each character for each taxon are required, representing the centroids of the taxa.

The method is illustrated by Figures 8-6 to 8-8. Figure 8-6 shows three taxa, **J**, **K**, and **L** for two character axes, X_1 and X_2. The taxa are shown as clusters of individuals (OTU's), but we are not now concerned with the values for the individuals, but only with descriptive parameters of the three clusters. These are their centroids and their dispersions.

The variances and covariances between the characters are first calculated, yielding the three $m \times m$ matrices (here $m = 2$); these are then averaged to give a pooled within-groups variance-covariance matrix **W**. This is shown in the legend to Figure 8-6 together with the values of the centroids of the taxa.

To calculate the discriminant function between **J** and **K** the inverted **W** matrix is multiplied by the vector $\boldsymbol{\delta}_{\mathbf{JK}} = [(\bar{X}_{1\mathbf{J}} - \bar{X}_{1\mathbf{K}}), (\bar{X}_{2\mathbf{J}} - \bar{X}_{2\mathbf{K}}), \ldots, (\bar{X}_{m\mathbf{J}} - \bar{X}_{m\mathbf{K}})]$. This gives the discriminant function as a vector **z**.

$$\mathbf{z}_{\mathbf{JK}} = \mathbf{W}^{-1}\boldsymbol{\delta}_{\mathbf{JK}}$$

This vector consists of a series of weights, w_i, which we here symbolize as $z_1, z_2, \ldots, z_m$ for characters *1*, *2*, ..., *m*. In the example $\mathbf{z}_{\mathbf{JK}} = -7.647, 2.057$. These weights are then multiplied by the observed character values of the unknown individual, and summed to give a discriminant score, *DS*.

$$DS_{\mathbf{u}} = z_1 X_{1\mathbf{u}} + z_2 X_{2\mathbf{u}} + \ldots + z_m X_{m\mathbf{u}}$$

This score is used for discriminating between members of **J** and **K** by calculating three reference scores for the centroid of **J**, for the centroid of **K**, and for the point midway between them—$DS_{\mathbf{J}}$, $DS_{\mathbf{K}}$, and $DS_{0.5}$ respectively. These are obtained from the equations

$$\begin{aligned} DS_{\mathbf{J}} &= \bar{\mathbf{x}}_{\mathbf{J}}\mathbf{z}' \\ DS_{\mathbf{K}} &= \bar{\mathbf{x}}_{\mathbf{K}}\mathbf{z}' \\ DS_{0.5} &= \tfrac{1}{2}(\bar{\mathbf{x}}_{\mathbf{J}} + \bar{\mathbf{x}}_{\mathbf{K}})\mathbf{z}' \\ &= \tfrac{1}{2}(DS_{\mathbf{J}} + DS_{\mathbf{K}}) \end{aligned}$$

The midway point assumes that the frequency of members of **J** and **K** in the population are equal. The values for the example are shown in Figure 8-7, which also shows the geometric effect the transformation has on the original character

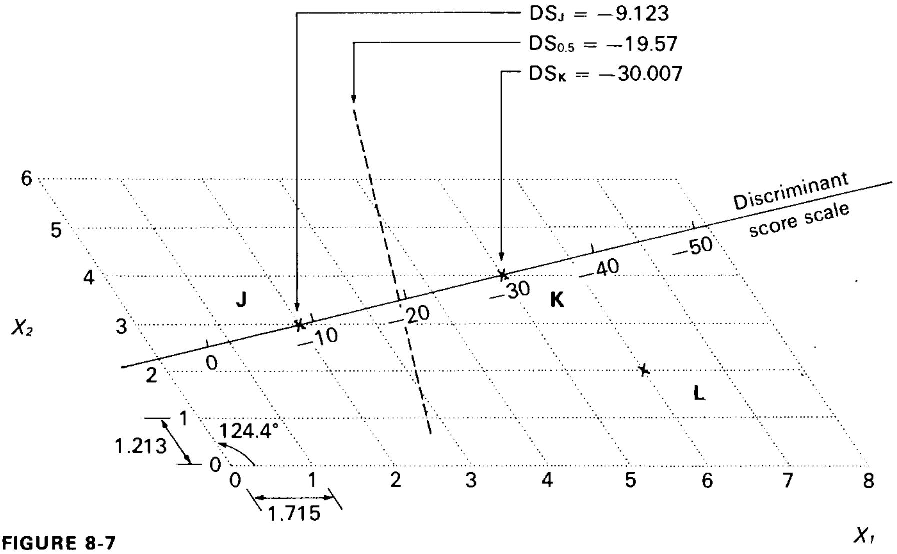


FIGURE 8-7
The effect of the discriminant analysis transformation upon the geometry of Figure 8-6, and the discriminant score scale for taxa **J** and **K**.

The transformation has two effects: the mean intrataxon variance is equalized for each character axis, and the axes are skewed (according to functions of the covariances), so as to make the clusters as nearly hyperspherical as possible. The angle of skewing is shown together with the new scales of the transformed axes. The discriminant score scale for separating **J** and **K** is also shown. The scores are calculated as $(-7.647 \times X_1) + (2.057 \times X_2)$ as explained in the text. For example, an unknown at $X_1 = 3$, $X_2 = 4$ has a score of -14.713. The points representing individual OTU's have been omitted for clarity. The discriminant function vector $\mathbf{z} = -7.647, 2.057$ is obtained by multiplying $\mathbf{W}^{-1}$ by $\boldsymbol{\delta}_{\mathbf{JK}}$ which is $(2.0 - 5.0), (3.0 - 4.0) = -3, -1$. See Figure 8-6.

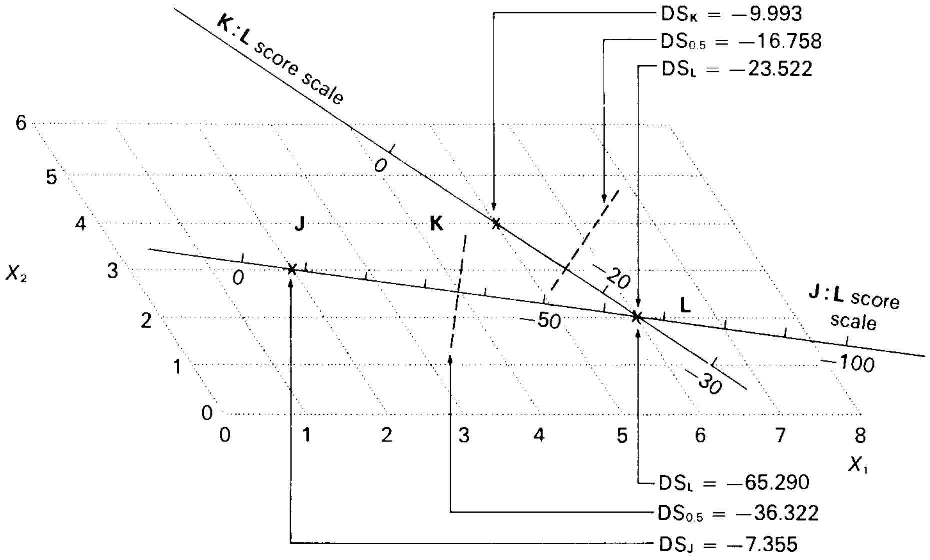


FIGURE 8-8
The discriminant score scales for separating taxa **J** and **L**, and **K** and **L**.

scales. The value of $DS_{0.5}$ is halfway between the other two scores and defines a plane midway between the centroids and perpendicular to the line joining them, shown by the dashed line in Figure 8-7. If the observed score for an unknown, DS_{u}, lies on the DS_{J} side, the unknown is allocated to taxon **J**, and if on the DS_{K} side, to taxon **K**. The length of the line between the centroids of **J** and **K** measured in discriminant function units is the square root of Mahalanobis' D^2 and the absolute difference between the scores DS_{J} and DS_{K} is also equal to D^2.

It should be noted that a different discriminant function, and discriminant scores, are calculated for each pair of taxa. Figure 8-8 shows the other two discriminant lines; it will be seen that the scales are different in size and orientation. It will also be evident from Figures 8-6 to 8-8 how discriminant functions are valuable when character values overlap, particularly when many characters are involved and one cannot draw scattergrams of the clusters.

Blackith (1965) has pointed out that the vector angle between two discriminant function vectors measures contrasts of form of the taxa. If the angle is small the functions measure similar contrasts of form, and large angles represent distinct contrasts.

There are two important uses of the distance D^2_{JK} between the scores DS_{J} and DS_{K}. First, one can test whether this indicates that the centroids are significantly different. For this, one uses an F test with m and $(t_{\mathrm{J}} + t_{\mathrm{K}} - m - 1)$ degrees of freedom and tests the ratio

$$\frac{D^2_{\mathrm{JK}}(t_{\mathrm{J}}t_{\mathrm{K}})(t_{\mathrm{J}} + t_{\mathrm{K}} - m - 1)}{(t_{\mathrm{J}} + t_{\mathrm{K}})(t_{\mathrm{J}} + t_{\mathrm{K}} - 2)m}$$

This ratio is related to Hotellings' T^2, as follows (Rao, 1952, p. 74):

$$T^2 = \frac{t_{\mathbf{J}} t_{\mathbf{K}}}{t_{\mathbf{J}} + t_{\mathbf{K}}} D^2_{\mathbf{JK}}$$

Second, one can determine the contribution that each character makes to D^2. and hence, see if any of them have such little discriminatory power that they are unlikely to be worth using. Alternatively, one can choose the best few characters from the set, and ascertain by the F test whether enough have been selected. The percent contribution of character i is $100 \times (z_i\delta_i/D^2)$ where z_i and δ_i are the ith elements of vectors $\mathbf{z}_{\mathbf{JK}}$ and $\boldsymbol{\delta}_{\mathbf{JK}}$. This criterion does not consider correlations between characters; if two or more characters are correlated they contribute to D^2 to a greater extent than this test suggests.

Although it is usual to take the midpoint between centroids as the criterion for identifying an unknown, there is no reason why one need do this. If it were very important to be sure of identifying all members of taxon **J** even at the price of misidentifying some members of **K** by allocating them to **J** in error, one can choose a criterion lying closer to the center of **K** than of **J**. $DS_{0.5}$ gives equal probability of misclassification of unknowns from either taxon.

The probability of misclassification can be calculated on the assumption of a multivariate normal distribution and also that the unknown does belong to **J** or **K** (and not to some distant cluster). A primary purpose of a discriminant function is to minimize the probability of wrong assignment of unknown individuals. If an unknown lies upon the $DS_{\mathbf{J}}$ side of $DS_{0.5}$, we can ask how many standard errors it is from $DS_{\mathbf{K}}$ and consult tables of the normal distribution. If it is many standard errors, then it is very unlikely to be a member of **K** misclassified as a member of **J**. The standard deviation of a taxon in D-space is taken as 1.0 in every direction. The square root of the difference between $DS_{\mathbf{J}}$ and $DS_{\mathbf{K}}$, i.e., D itself, gives the number of standard deviations the centroids are apart, and half of this corresponds to the $DS_{0.5}$ plane. An unknown lying on the $DS_{\mathbf{J}}$ side therefore is over $\frac{1}{2}D_{\mathbf{JK}}$ standard deviations from **K**.

Figures 8-6 to 8-8 illustrate several important points. The use of a given discriminant function implies that the unknown does belong to one or the other of the two taxa being considered. If instead it belongs to a quite different cluster. located far off in the space, it may have almost any discriminant score and may thus appear to belong to one or the other of the two taxa under consideration, when it really belongs to a third. Also, when there are many taxa one has to test against a large number of discriminant functions. These two problems are largely overcome by the use of D^2 as described below. Harder to overcome is the fact that all the usual methods of discriminant analysis assume that the dispersion matrices of the taxa are homogeneous (that is, the clusters all have much the same size, shape, and orientation in phenetic space) and that the clusters have multivariate normal distributions.

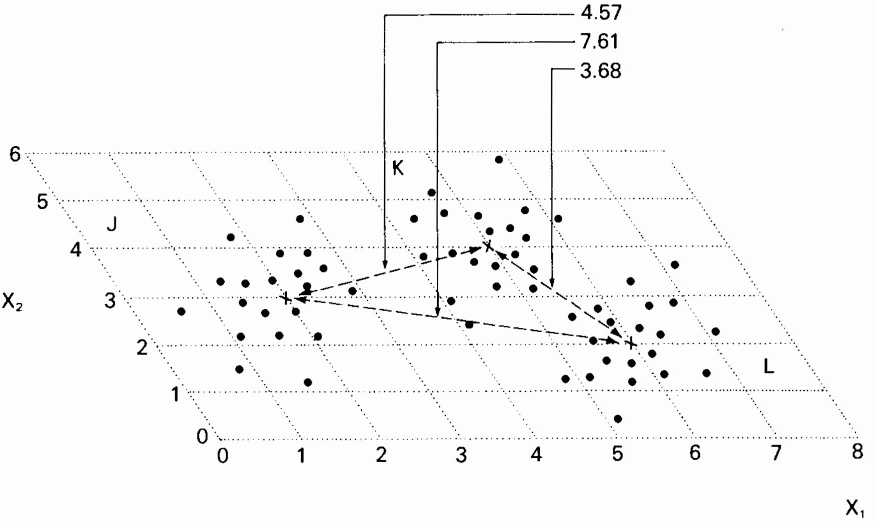


FIGURE 8-9
The transformation into discriminant analysis space from Figure 8-6 shown in more detail. The points representing the OTU's are indicated, together with the Euclidean distances in the new space between the centroids. These distances are the values of D (i.e., the square roots of D^2).

Mahalanobis' D-space and Multiple Discriminant Analysis

Figure 8-9 shows the original clusters in the transformed D-space, but without the discriminant score scales. It was noted earlier that D^2 could be obtained between the centroids from the discriminant scores. However following Mahalanobis (1936), one can calculate it between any pair of points f and g by the equation

$$D_{fg}^2 = \boldsymbol{\delta}_{fg}' \mathbf{W}^{-1} \boldsymbol{\delta}_{fg}$$

Then the square root of D^2 is simply the Euclidean distance in the D-space. This can be seen in Figure 8-9, where the distances between the centroids are marked. The method therefore transforms the original space into a new space, in which the original axes are stretched and also skewed so that they are no longer at right angles. The length of a unit in dimension i is p times the original units, where p is the square root of the a_{ii} element of the matrix $\mathbf{W}^{-1}$. The direction cosine between the dimensions h and i is equal to $a_{hi}/\sqrt{a_{hh}a_{ii}}$. Gower (1967b) gives a representation with correlation coefficients.

In the special case where none of the characters are correlated (so that all covariances are zero) the transformation simply stretches the axes but leaves them at right angles; the length of a unit in dimension i is then $1/s_i$ times the original units, where s_i is the mean within-cluster standard deviation of character i. That is, the axes are stretched in inverse proportion to the standard deviation. If in addition all variances are equal to s^2, then D is $1/s$ times the ordinary Euclidean distances. The D units are of a kind that can be called "ease of discrimination units." Confidence limits in D-space can be found from the sampling variance of D^2, which is $D^2/(t_1 + t_2 + \ldots + t_q - q)$.

Canonical Variates

Because D can be represented in an orthogonal system of axes (though the orientation is arbitrary), one can perform analyses on distances between taxa or individuals, in particular by principal coordinate analysis (Gower, 1966a, 1967b). Canonical variates and multiple discriminant analysis are equivalent except in minor particulars. The coordinates of the points in an orthogonal system can be obtained, and the orthogonal axes are the canonical variates, which can be used also for discrimination. In Figure 8-10 the orthogonal D axes are shown, obtained by principal coordinate analysis of D distances between thę centroids of **J**, **K**, and **L**. It is clear that one can then readily identify unknowns by seeing whether they fall within critical distances of taxon centroids. Gower (1966a) points out that though there is rarely need to do so, one can transform into D-space even if one has only a single taxon: one considers the whole set of OTU's as one large group. The positions of OTU's will then be made such that the entire set forms a hyperspherical cluster, within which, of course, there may be subclusters. Note that the relations between vector lengths are preserved in D-space. An OTU twice as big as another but of the same shape will lie on the same line from the origin in Figure 8-9, but twice as far away. Discriminant functions and D^2 are less sensitive to general size factors than taxonomic distance; but an unknown of the same shape as a member of a taxon may appear outside that taxon if it differs much in size.

There are certain difficulties with discriminant analysis. If any character is invariant in each of the taxa, the matrix **W** cannot be inverted, unless "generalized inverses" are used. Yet the character might have a unique state for each OTU, and by itself be a perfect discriminator. There are also difficulties in choosing the limited set of best characters from the large number that should be employed in numerical taxonomy. Characters with means that are well separated in relation to the variances and that are not highly correlated with other characters are in general the best, but optimal methods of choosing them pose statistical and computational problems (see Feldman, Klein, and Honingfeld, 1969). We believe that for most taxonomic work it will be possible to choose a nearly optimal set by inspection. It has been shown by Dunn and Varady (1966) that rather large numbers of individuals are required in each taxon for reliable discriminant functions. The gain in discriminatory power over simpler methods may not be very great (Sokal, 1965) particularly with 0,1 characters (Gilbert, 1968; Kurczynski, 1970) and simple discriminants based on equal weight for each character can be quite effective (e.g., Kim, Brown, and Cook, 1963). Discriminant functions have most value for very close clusters

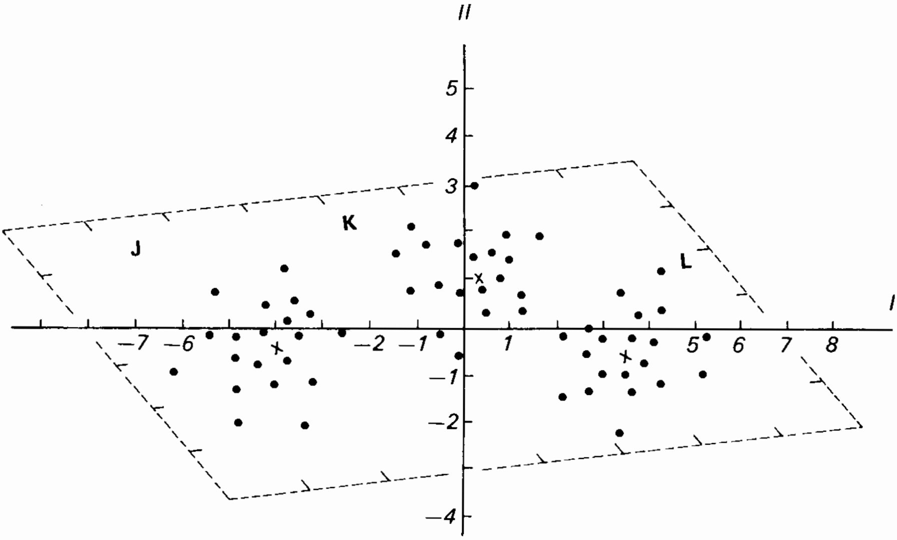


FIGURE 8-10
Canonical variates I and II superimposed on the positions of OTU's in Figure 8-9. The axes are principal axes, and the new origin is at the centroid of the three taxon centroids, 4.33, 3.0. The original character axes are shown by the frame of broken lines. The new axes are scaled in D units, which are effectively within-taxon average standard deviations in the transformed space.

that partly overlap; good examples are those of Giles and Elliot (1962, 1963) on human skulls of different sexes and racial groups. Sokal (1965) lists examples of their use in taxonomy. Hill et al. (1965) obtain a type of discriminant function from their gradient factor analysis. DuPraw (1965a) achieved excellent discrimination of wings of honeybees of different geographical origin by multiple discriminant analysis. Blackith and Reyment (1971) describe numerous applications of D^2, discriminant functions, and canonical variates. New discriminant methods have been suggested by Hall (1968) and Saila and Flowers (1969).

Among references dealing with more complex methods than discriminant functions and D^2, and reviewing various parts of the field of discrimination, are Rao (1952), Sebestyen (1962), Reyment (1963), Kossack (1963), Sokal (1965), and Chaddha and Marcus (1968). Related work is that of Birnbaum and Maxwell (1961). Cavalli (1949) discusses the "mean correlation coefficient," defined as the mean of the $n(n - 1)/2$ correlation coefficients between n pairs of characters, and discusses its relation to Gini's synthetic coefficient and the work of Zarapkin. Penrose (1954) found that it is possible to obtain, in a simple manner, good approximations to D^2 by using an average measure of the correlations between characters.

Sebestyen (1962) points out that measures that reduce the general size factor may make discrimination more difficult, but the point at issue is whether the difference in size in the particular case is indeed a reliable discriminant or an artifact of sampling, for example. Sebestyen considers that the most powerful methods are those giving equiprobable envelopes of clusters, but they have been little developed and require very large numbers of individuals. He also discusses some nonlinear methods, as does Rohlf (1970). Williams and Lance (1968) also discuss this general problem under the heading of extrinsic criteria of patterns; they conclude that nonlinear multiple regression may sometimes be a suitable technique but that the area is in need of deeper study.

It may often be useful to employ the simple method of Lubischew (1962) for testing single characters as discriminators. He calculates his coefficient of discrimination $K = (\bar{X}_{i\mathbf{A}} - \bar{X}_{i\mathbf{B}})^2/2s_i^2$ where s_i^2 is the pooled variance for character i from taxa **A** and **B** (that is an average weighted by the numbers of individuals). The greater K is, the better i is as a discriminator. This takes no account of correlation between characters. The probability of misclassification is approximately the probability that a normal deviate exceeds $\sqrt{K/2}$, so that, for example, with $K = 7.68$, 95 percent of identifications will be correct. This requires the distributions of X in **A** and **B** to be approximately normal and of equal variance.

General Conclusions on Identification and Discrimination

Recommendations in this area are somewhat tentative because of the small experience with alternative numerical methods. We suggest that for large studies with well-separated taxa the sequential methods are best. Simultaneous keys are useful for highly polythetic groups without much overlap, but discriminant analysis is indicated where there are a few close groups in which identification must be as certain as possible. The addition of simple probabilistic values to the traditional methods should prove rewarding.