APPENDIX **B**

Some Hints on Techniques, Sources, and References

The rapid advance of computer techniques has made it inappropriate for us to present an appendix of worked examples such as was featured in our *Principles of Numerical Taxonomy*. There are, however, still practical points that may be conveniently brought together here. These come under the following headings: preparing material for numerical taxonomic studies; computational methods and strategy; and sources of computer programs, and bibliographies and reviews of special areas of numerical taxonomy. A full treatment of practical and computational aspects will appear in a forthcoming volume by F. J. Rohlf and P. M. Neely.

In selecting material for numerical taxonomic studies the investigator should have a clear idea of the kind of variation he wishes to explore, among both characters and OTU's. We believe that in most organisms it will be possible to find the necessary numbers of characters for analysis. In some difficult groups there may be few available characters, which in itself would show that previous taxonomies must have been based on inadequate data; any improvement in the classification, therefore, must first require new methods of study. This difficulty may also arise with studies at very high taxonomic ranks, as it may not be easy to know what characters can be selected for comparing, for example, an insect with an echinoderm; in such cases chemical data, especially protein sequences, are useful.

The selection of OTU's also requires attention, particularly if space-distorting clustering methods are employed, because of the sensitivity of these methods to the sample of OTU's chosen to represent the taxa under study. Ideally one would study all the species of a genus, all the genera of a family, and so on, but although this is often not feasible, efforts should be made to make the set of OTU's as complete as possible. A more difficult problem is that of incomplete data for the characters of the OTU's. Again the worker may have to undertake

extensive studies to obtain the missing data (which again implies the inadequate bases of previous taxonomies). Our present advice on when to exclude OTU's or characters because of incompleteness of the data has been given in Section 4.12.

Substantial saving of time and effort can come from a planned approach to collecting the data. In some applications (e.g., microbiology) it is inconvenient to add OTU's to a partly completed study; in others (e.g., when using highly specialized techniques in chemistry or electron microscopy) it is inconvenient to reexamine OTU's for further characters. It is therefore useful to make preliminary lists of OTU's and characters early in the study, and this will also ensure that obvious kinds of information are looked for and recorded. After all the data have been recorded, unnecessary and laborious copying should be avoided. This is facilitated by first inspecting all the OTU's and characters and clearing up any ambiguities. A checklist of OTU's should then be made. Next, the specification of the computer program should be studied to see in what format the data should be presented. It should now be possible to code the character state values in the form required for punching. Some characters may be rejected at this stage for various reasons, and it is best to place the characters in some logical order to avoid accidental repetitions and inconsistences. It usually does not matter what order the OTU's are in, because ties in resemblance values are uncommon when large numbers of characters are employed. Further details can be found in Sneath (1967d).

One aspect of computational strategy is the writing of convenient computer programs for carrying out the computations. There are technical aspects that, though extremely important for the success of numerical taxonomic work, go beyond the scope of this book and are discussed in detail in a forthcoming volume by Rohlf and Neely. Among other problems, Rohlf and Neely concern themselves with questions such as how data matrices should be read in and stored, which particular algorithms should be used for various matrix computations, and what particular combination of central processor and peripheral equipment such as tape drives, disks, and drums permits an optimal analysis of a data matrix of given dimensions. The successful solution of these problems by several programming groups (e.g., the team headed by F. J. Rohlf, first at the University of Kansas and subsequently at the State University of New York at Stony Brook, or that of G. N. Lance and W. T. Williams at Canberra, Australia, among others) has greatly aided the advance of numerical taxonomic work all around the world. Several taxonomic program packages have been developed (e.g., those of Rohlf, Kishpaugh, and Kirk, 1971; of Wishart 1969e, 1970; of Gower and Ross of Rothamsted Experimental Station; of Sackin at Leicester University; and of Lance and Williams at Canberra); these permit a wide choice of resemblance coefficients and methods for displaying taxonomic structure.

The time required for computation depends greatly on the numerical method employed. For most methods, in which the full similarity matrix is computed, the time is roughly proportional to nt^2, so that increasing the number of OTU's has more effect than increasing the number of characters. For association analysis the reverse holds, as the time is proportional to n^2t. Any method for which the time rises too steeply with increasing t or n is impracticable for realistic taxonomic problems. Examples are the methods of Edwards and Cavalli-Sforza (1965), and complete searches for seriation (Kendall, 1963) or rooted trees (Fitch and Margoliash, 1968), for which the times are roughly proportional to 2^{t-1}, $t!$, and $(2t)!/(t+1)!2^{t-1}$ respectively. Some methods require only the computation of part of the S matrix (e.g., that of Rose, 1964, and use of a graph-theoretical approach to single linkage clustering by Gower and Ross, 1969). Special combinatorial algorithms (e.g., Wishart, 1969a, 1970) can reduce the time appreciably. Češka (1968) gives a method whereby the mean values of association coefficients within and between sets of OTU's may be computed directly from the $n \times t$ matrix without calculating all similarity values (provided the sets are already known). Further information on computing times is given by W. T. Williams (1964) and Lance and Williams (1966b).

For most present computer installations, with more conventional methods, capacity is limited by t with an upper limit of usually about 400. In some applications very large numbers of OTU's, perhaps many thousands, must be processed, and special methods are then needed. Although experience with such methods is still limited, they have been described by a number of authors. These include methods of Lockhart and Hartman (1963), Crawford and Wishart (1967; 1968), Ross (1969d), Kaminuma, Takekawa, and Watanabe (1969), and Switzer (1970), as well as methods of Rose (1964) and of Gower and Ross (1969).

These fast methods may also be used to divide very large data sets into manageable subsets, but because many of them are monothetic there is the risk of misplacement of a few OTU's unless special reallocation facilities are provided (e.g., Crawford and Wishart, 1968). Other ways of handling more OTU's than can be accommodated at once have been mentioned in Section 3.1. With ordination methods it is usually possible to obtain a desired end result by algebraic manipulation of either Q- or R-type matrices, so one can choose the technique involving least computation according to whether n or t is the greater (see Gower, 1966a,b, 1967b; Orloci, 1967a). Some steps in orthodox taxonomy are analogous to using R analyses to break down very large sets of OTU's into smaller ones, and this strategy is also available.

An allied problem is to add a new point to an ordination, and a method has been described by Gower (1968) for which an example is given by Wilkinson (1970b). A general solution for n dimensions to the related problem of matching diagrams (see Section 3.3) has been derived by Gower (1971b), as follows. The n-dimensional coordinates of the h points are first referred to the centroids of their respective diagrams A and B, giving two $h \times n$ matrices ($h > n$), **A** and **B**. If **A** and **B** have originally different numbers of columns, the smaller is filled out with zeros on the right to ensure that **A** and **B** are both $h \times n$ matrices. One then computes $\mathbf{R} = \mathbf{A}'\mathbf{B} = \mathbf{USV}'$ where **U** and **V** are orthogonal (their elements scaled so the sum of squares of columns is unity) and **S** is diagonal. **U**, **V**, **S** are obtained as eigenvectors of the equations $\mathbf{RR}'\mathbf{U} = \mathbf{US}^2$ and $\mathbf{R}'\mathbf{RV} = \mathbf{VS}^2$ or directly by a singular value decomposition algorithm (Golub and Reinsch, 1970). The required orthogonal rotation matrix **H** that minimizes least square distances between corresponding h-points of A and B is $\mathbf{H} = \mathbf{VU}'$. This is not unique if **R** has rank less than $h - 1$, but this simply means that several possible rotations exist. The columns of **U** and **V** may each be changed in sign independently, without affecting the validity of the above solution, giving 2^n different results, each one associated with a different set of reflections of B relative to A. The reflection that gives the best fit is obtained by calculating $\mathbf{U}'\mathbf{RV}$ for some solution **U** and **V** and then multiplying the ith column of **V** by the sign of the ith diagonal element of $\mathbf{U}'\mathbf{RV}$. The new **V** is then used in subsequent calculations. These considerations are conveniently covered by calculating the matrix $\mathbf{G} = \mathbf{JH}$, where **J** is the reflection matrix, a diagonal matrix of $+1$ and -1 elements whose signs are derived as explained above. **G** gives the best fit after allowing for reflection. The coordinates of B referred to A are given by $\delta\mathbf{BG}$, where δ is a scaling factor of B. The value of δ that gives the minimum sum of squared distances between corresponding pairs of points, $\Sigma\Delta^2$, is obtained as $\delta = \text{trace}\,(\mathbf{BGA}')/\text{trace}\,(\mathbf{BB}')$. After scaling **B** by multiplying by δ, $\Sigma\Delta^2$ is then trace $(\mathbf{AA}')$, which equals δ^2 trace $(\mathbf{BB}')$. However, because of the nonreciprocal scaling this gives when fitting B to A compared with A to B, other methods of scaling may be preferred. A simple method is to scale **A** and **B** to have the same total sums of squares, say unity; this amounts to dividing **A** and **B** by the square roots of the traces of $\mathbf{A}'\mathbf{A}$ and $\mathbf{B}'\mathbf{B}$ respectively. Sometimes it is clear that both A and B are on the same scale so no additional scaling is required.

Many computer programs now have facilities for graphic output, ranging from diagrams constructed on the line printer to cathode ray displays, but most commonly on a graph plotter (for example the GRAFPAC subroutines of Rohlf, 1969). Subroutines for producing phenograms or cladograms include those of Bartcher (1966), Bonham-Carter (1967b), and McCammon and Wenniger (1970). Rearranged similarity matrices are also often provided

(sometimes with only the first significant digit and no spaces between digits, giving much the same impression as shaded similarity matrices). Reordered $n \times t$ matrices (e.g., Bonham-Carter, 1967b) are useful for selecting diagnostic characters. Ordination plots can also be provided, though problems occur if plotted points overlap. Stereograms are assuming increasing importance, and here exact positioning is very important, as pointed out by Rohlf (1968). Formulae for stereograms are given by Fraser and Kovats (1966) and by Rohlf (1968). Rohlf's formulae can be calculated on desk calculators, and are given below.

For any given OTU with coordinates X_{I}, X_{II}, and X_{III} on three ordination axes I, II, and III, one calculates its position for the stereoimages on axes X (horizontal) and Y (vertical). First the coordinates are suitably scaled by subtracting the minimum value for any OTU in the study, and dividing by a constant M that is conveniently taken as rather larger than the greatest of the ranges of the values on I, II, and III. The scaled values are indicated by primes

$$X'_{\mathrm{I}} = (X_{\mathrm{I}} - X_{\mathrm{I,min}})/M$$
$$X'_{\mathrm{II}} = (X_{\mathrm{II}} - X_{\mathrm{II,min}})/M$$
$$X'_{\mathrm{III}} = (X_{\mathrm{III}} - X_{\mathrm{III,min}})/M$$

It is most convenient to choose the axis with the greatest range as I, and that with the least as III.

Next, viewing points are chosen, where the left viewing point has the coordinates L_{I}, L_{II}, and H, and for the right they are R_{I}, R_{II}, and H. Rohlf recommends for general use $L_{\mathrm{I}} = L_{\mathrm{II}} = R_{\mathrm{II}} = 1/2$, $R_{\mathrm{I}} = 2/3$, and $H = 3$. The position for the left stereoimage is then

$$X_L = (HX'_{\mathrm{I}} - L_{\mathrm{I}}X'_{\mathrm{III}})/(H - X'_{\mathrm{III}})$$
$$Y_L = (HX'_{\mathrm{II}} - L_{\mathrm{II}}X'_{\mathrm{III}})/(H - X'_{\mathrm{III}})$$

and for the right stereoimage

$$X_R = (HX'_{\mathrm{I}} - R_{\mathrm{I}}X'_{\mathrm{III}})/(H - X'_{\mathrm{III}})$$
$$Y_R = (HX'_{\mathrm{II}} - R_{\mathrm{II}}X'_{\mathrm{III}})/(H - X'_{\mathrm{III}})$$

The X, Y coordinates are calculated for each OTU and may also be calculated for the corners of a rectangular box that acts as a viewing frame, which can then be drawn in with straight lines. Because of the need for accuracy, enlarged drawings should be made and reduced photographically to the size appropriate to the viewing device to be used. If an oblique view is preferred, convenient viewing points are given by $L_{\mathrm{I}} = L_{\mathrm{II}} = R_{\mathrm{II}} = 1.5$, $R_{\mathrm{I}} = 1.7$. $H = 3$.

For unusually difficult jobs, consultation with computer experts in processing multivariate data is essential. In a study of the suborder Blattaria by Huber (1968) involving 177 OTU's representing 37 species scored for 446 characters, a major problem was storing the original data matrix on tape for subsequent computer processing. In spite of competent assistance it took the better part of a month before the data were successfully converted from cards to tape. The subsequent handling of the data was relatively routine by the NT-SYS system at The University of Kansas, although on some computers such large numbers of characters would also present a problem.

Another aspect of computational strategy is how much time a given investigator should devote to learning how to program by himself, whether he should attempt to implement numerical taxonomy programs at his own computation center or use larger, remote facilities where these programs are already implemented if he has the opportunity to do so. There are still many problems in transferring programs to another machine, or even to the same machine at another installation. Recent tendency has been to develop sophisticated systems at several large computation centers and have these used by clients outside the institution, because such

a strategy would generally be considerably less expensive and less time-consuming for a potential user than to attempt to implement even a simple numerical taxonomy program at his own computation center. The availability of long distance, time-shared computing makes this approach even more attractive. Our present advice therefore would be that for "one shot jobs" it is simplest and most economical to send the data to a center that is equipped to handle them. However, persons who wish to do numerical taxonomy on a routine basis should establish at their centers a series of programs that would carry out these computations for them. The recent development, by various teams of workers, of libraries of basic numerical taxonomy routines not interrelated as a computing system, but standing independent of each other and written in a simple, widely compatible style, has made this strategy more feasible than before.

Programs that are written for numerical taxonomy should have detailed write-ups to enable machine operators unfamiliar with the computations and taxonomists unfamiliar with computers to do the work with maximum facility. Information on operating instructions should include how many OTU's and characters can be processed, the exact format of input and output, estimates for execution, and restrictions on the kinds of characters permitted. It is particularly important that the write-up should not only describe the general idea of what the program will do, but should also give in detail the actual algebra used (unless it is a very standard procedure for which reference to a publication would be adequate). This information is essential for the user to enable him to be certain that the program carries out the kind of analysis that he requires. It is also extremely useful for the write-up to contain a small worked example with input data and results for checking the program if it is implemented on a new machine.

In *Principles of Numerical Taxonomy* a grouping of numerical taxonomic programs was outlined, based on suggestions by Sneath and Rohlf in *Taxometrics*, **2**, December 1962. Many programs incorporate several groups, so that these are not always convenient in practice, but they illustrate the logical arrangement of subprograms in a computer program package and may therefore still be of use in planning program layout. They are briefly described below with minor modifications.

GROUP 1

Control programs control the subsequent programs of Groups 2–9. Control programs call up different subprograms as required and direct the flow of operations.

GROUP 2

Translation programs take in descriptions of OTU's in words or diagrams and convert them into appropriate numerical codes. These programs may eventually be able to remove much of the tedium of coding characters from the shoulders of the taxonomist, and are now being used particularly in key-making programs.

GROUP 3

Character conversion programs convert data to the form necessary for computing resemblance coefficients, such as standardization (or other transformations), transposition of rows and columns, and augmentation (or deletion) of OTU's (or characters).

GROUP 4

Resemblance coefficient programs use the output of Group 3 programs and calculate matrices of resemblance between OTU's. Some special procedures in cladistic analysis (e.g., character compatibility) may be conveniently included here.

GROUP 5

Programs for analysis of taxonomic structure, a group originally restricted to programs for cluster analysis, can now conveniently be extended to include (a) cluster analysis programs,

(b) ordination programs, (c) programs for cladograms, and (d) cophenetic programs. This group takes resemblance matrices or their equivalent and yield, as output, dendrograms, cluster parameters, ordination plots, distortion measures, etc.

GROUP 6
Data extraction programs extract data from earlier steps, answering specific questions addressed to the study. They include (a) programs that compute average resemblances within and between specified clusters of OTU's and (b) identification programs. They require additional input to indicate specified phenons (sometimes single OTU's).

GROUP 7
Interstudy coordination programs store and sort out previous studies, establish reference taxa and their characters, and correlate different studies (e.g., by computing distortion measures).

GROUP 8
Publication programs convert results into forms that are legible and publishable, such as diagnostic keys and graphic outputs of quality suitable for use as phenograms and ordination plots.

GROUP 9
Miscellaneous programs; some programs do not readily fit into the other classes.

There are now many computer programs for numerical taxonomy. Most of these are unpublished, and workers must contact those who have written them; however, there are several sources through which these may be traced, in addition to standard sources of papers on applications to numerical taxonomy. The following periodicals among others contain descriptions and sometimes full program listings: the *Computer Journal*, *Applied Statistics*, *Behavioral Science*, and the *Kansas Geological Survey Computer Contributions*. Two newsletters, *Taxometrics* (issued by the National Collection of Type Cultures, Colindale, London N.W. 9) and the *Classification Programs Newsletter* (issued by the M.R.C. Microbial Systematics Unit, University of Leicester) contain lists of programs and their sources.

The series of computer contributions of the Kansas Geological Survey is especially valuable because these contain full descriptions and examples of input and output as well as the programs themselves. Numbers of special interest include the following: Bartcher (1966) for cladistic relationships by the Camin–Sokal method; Bonham-Carter (1967b) for Q cluster analysis of binary WPGM or UPGM; Wahlstedt and Davis (1968) for principal components (and a form equivalent to principal coordinates); Wishart (1969e) for numerous resemblance coefficients and cluster methods; Ondrick and Srivastava (1970) for correlations and R and Q factor analysis with varimax rotation; Demirmen (1969) for iterative reallocation in cluster analysis; Reyment, Ramden, and Wahlstedt (1969) for Mahalanobis distance; Reyment and Ramden (1970) for canonical variates; and McCammon and Wenniger (1970) for a special form of phenogram called the dendrograph. An extensive compilation of programs for environmental sciences (Tarrant, 1972) lists many taxometric ones. Many useful multivariate statistical programs are given in Cooley and Lohnes (1971). Sokal and Rohlf (1969) include programs for basic statistical procedures. Certain algorithms of use in graphs and trees are given by Gower and Ross (1969), Ross (1969a,b,c), and Farris (1970). Wishart (1970) gives many useful combinatorial formulae for clustering methods that can considerably reduce time and programming. Other papers that consider improved algorithms for various methods are Proctor (1966), Jensen (1969), Vinod (1969), Wishart (1969a), and Cole and Wishart (1970)). Estabrook and Brill (1969) describe a program for taxonomic data processing.

There are now numerous publications reviewing special areas in numerical taxonomy. Our earlier volume (Sokal and Sneath, 1963) is available in French translation from Laboratoire Central, Compagnie Française de Petrole, Bordeaux, and a summary is available in Russian (Sokal, 1968). The proceedings of a numerical taxonomy symposium have been edited by Cole (1969). Publications mainly on methodology in special groups of organisms include Lockhart and Liston (1970) and Sneath (1972) in microbiology, and J. McNeill (in preparation) in botany. Reviews of numerical taxonomy in special groups include the following: in zoology Funk (1963, acarology), Johnston (1964, acarology), Moss and Webster (1970, nematology); in botany Williams (1967b), Gilmartin (1967b), and Sneath (1969d); in microbiology Sneath (1962, 1964a, 1968c), Lysenko (1963a), Véron (1969), and Colwell (1971). A more general review is that of Rogers, Fleming, and Estabrook (1967). Similar reviews on applications in paleontology and subjects outside systematics have been listed in Section 6.5 and in Chapter 11.

Certain methodological fields are also covered by recent or forthcoming works: in mathematics, Fernandez de la Véga (1965), Lerman (1970), and Jardine and Sibson (1971); on computational aspects, Rohlf and Neely (in preparation); on cluster analysis, Spence and Taylor (1970) and Wishart (1969b, 1970); on phyletics, Kluge and Farris (1973).

Useful bibliographies may be found in various issues of *Taxometrics* (a KWIC index to this has been issued) and the *Classification Society Bulletin*. Journals of special interest to numerical taxonomists include *Systematic Zoology*, *Taxon*, the *Classification Society Bulletin*, the *Journal of General Microbiology*, *Biometrics*, and *Computers and the Humanities*.

Introductory texts for statistics and mathematics for numerical taxonomy include Sokal and Rohlf (1969), Simpson, Roe, and Lewontin (1960), Schwartz (1961), Searle (1966), Graybill (1969), and Cooley and Lohnes (1971); more advanced treatments of multivariate methods are found in Rao (1952), Anderson (1958), Seal (1964), Morrison (1967), Harman (1967), and Van de Geer (1971).