

The Aims and Principles of Numerical Taxonomy

The contents of this book fall into three main parts. The chapters of the first part furnish an introduction to taxonomic theory in general and to numerical taxonomy in particular. The purpose of this chapter is to outline in summary form the aims and principles of numerical taxonomy. In Chapter 2, we discuss in some detail the conceptual bases of classification, contrasting conventional views with those espoused by a growing number of taxonomists in recent years.

The central part of the book is arranged on a plan that closely reflects the successive steps followed by taxonomists in performing the classificatory process. In Chapter 3 we discuss taxonomic evidence, which comprises the selection of organisms for study, the choice and definition of taxonomic characters, and criteria for homology. The estimation of taxonomic resemblance between organisms follows in Chapter 4. Chapter 5 considers the grouping of organisms into taxa on the basis of these resemblances.

The final part deals with the implications of numerical taxonomy for systematic research. Numerical approaches to the study of phylogeny are detailed in Chapter 6. The next chapter (7) treats the numerical taxonomy of populations, leading to a discussion of phenetic patterns and evolutionary structure. The implications of numerical taxonomy for keys and identification are discussed in Chapter 8,

implications for nomenclature in Chapter 9. A critical examination of numerical taxonomy—its advantages as well as its shortcomings—is featured in Chapter 10. The next chapter (11) is a necessarily brief survey of the application of numerical taxonomy to fields other than biological systematics. The main text concludes in Chapter 12 with a brief outlook on the future of systematics in the computer age.

Two appendixes are provided. A list of applications of numerical taxonomy to different groups of organisms is found in Appendix A; advice on data processing in a numerical taxonomic study is found in Appendix B.

We have made a deliberate attempt to maintain a uniform symbolism for characters, operational taxonomic units (OTU's), and taxa (groups of OTU's), throughout the text. This has meant that in numerous places we have had to report the work of other authors through the use of a symbolism different from that employed in the original publications. We feel that the danger of confusion to the reader who wishes to consult the original sources is more than offset by having a consistent symbolism within the book and we hope that the publication of our book will serve in some small way to retard the development of disparate symbolisms in numerical taxonomy. Even with the best will, it has not been entirely possible to keep symbolism totally consistent, partly because of the limitations on the number of suitable letters in the alphabet. We have attempted to flag all such cases for the reader to further lessen the chances for confusion.

1.1 DEFINITIONS OF SOME TERMS IN TAXONOMY

Adequate definition of taxonomic terms itself would almost require a book. Many terms are used in so many different senses that we have had difficulty in using them in a consistent and exact manner, and no doubt there are still many ambiguities that we have overlooked. The meanings of terms and symbols as they are used in this book can be looked up by way of the Index. There are, however, several that are employed so frequently that we have deemed it desirable to present them here at the outset.

Classification, systematics, and taxonomy are often used interchangeably. In recent years, especially in the United States, there has been a trend toward assigning separate meanings to these terms. In this sense they have been well defined by Simpson (1961), and we follow his usage here.

Systematics. This is defined by Simpson (1961, p. 7) as “*the scientific study of the kinds and diversity of organisms and of any and all relationships among them.*” This definition uses systematics in its widest sense, concerning itself not only with the arrangement of organisms into taxa and with naming them, but also with the causes and origins of these arrangements.

Classification. We have modified Simpson's definition (1961, p. 9), which restricted itself to zoological classification, to more general usage. *Classification is the ordering of organisms into groups (or sets) on the basis of their relationships.* There may be confusion over the term "relationship," which may imply phylogenetic relationship (analyzed in Section 2.3), or which may simply indicate the resemblance or overall similarity as judged by the characters of the organisms without any implication as to their relationship by ancestry. Relationship based on overall similarity may be distinguished from phylogenetic relationship by calling it *phenetic relationship*, employing the convenient term of H. K. Pusey as used by Cain and Harrison (1960b), to indicate that it is judged from the phenotype of the organism and not from its phylogeny. We have restricted the definition of classification to organisms, since this book is primarily intended for the biological taxonomist. However, there are many methods of classification, including numerical taxonomy, which are equally applicable to concepts and entities other than organisms. Classification as defined above is the name of a process; however, it has also been used for the end product of this process. Thus the result of classification is a classification. The term classification has also been employed, mainly in fields outside biology, in the restricted sense of putting entities into distinct classes as opposed to arraying them in a continuous spectrum, cline, or other arrangement (ordination) showing no distinct divisions. We have not restricted the term in this manner. The term classification should also be distinguished from "identification."

Identification. This is the *allocation or assignment of additional unidentified objects to the correct class once a classification has been established.* Thus, a person using a key to the known wild flowers of Yellowstone National Park "identifies" a given specimen as a goldenrod. Some mathematicians and philosophers would call this process classification also, but we follow conventional usage of biologists in distinguishing strictly between the two.

Taxonomy. Simpson used this term to mean "*the theoretical study of classification, including its bases, principles, procedures and rules*" (Simpson, 1961, p. 11). By this definition the bulk of the subject matter of our book is concerned with taxonomy. We would include the theoretical aspects of identification under taxonomy as well. Taxonomy, like classification, has also been used to designate the end products of the taxonomic process. Since this is a generally accepted usage, we will occasionally employ it in this sense. Taxonomy is frequently made synonymous with classification and we have not attempted to distinguish rigorously between the terms "numerical taxonomy" for theory and "numerical classification" for practice. By contrast, Blackwelder (1967a) understands taxonomy to mean "... the day-to-day practice of dealing with the kinds of organisms. This includes the handling and identification of specimens, the publication of the data, the study of the literature, and the analysis of the variation shown by the specimens." Most of these activities

are procedural, related to taxonomy but of no theoretical significance. We shall have little to say regarding them.

We use the term *taxon* (plural *taxa*) as an abbreviation for taxonomic group of any nature or rank, as suggested by H. J. Lam (in Lanjouw, 1950) and Rickett (1958).

1.2 DEFINITION OF NUMERICAL TAXONOMY

Before proceeding, it is necessary that we clearly define our use of the term “numerical taxonomy.” We mean by it *the grouping by numerical methods of taxonomic units into taxa on the basis of their character states*. The term includes the drawing of phylogenetic inferences from the data by statistical or other mathematical methods to the extent to which this is possible. These methods require the conversion of information about taxonomic entities into numerical quantities. We have preferred “numerical taxonomy” to “quantitative taxonomy,” since the latter would include other methods, such as serology or paper chromatography.

In view of the development of numerical methods for estimating cladistic relationships subsequent to the publication of *Principles of Numerical Taxonomy*, it has been suggested (by, among others, W. H. Wagner, personal communication) that the entire field be renamed “numerical systematics.” Because of the generally accepted breadth of the meaning of the term systematics (see previous section) and the numerous aspects of systematics that are not suited to the approaches discussed in this book, we prefer to retain the established definition. Jardine and Sibson (1971) suggest the name “mathematical taxonomy,” and Blackith and Reyment (1971) employ “multivariate morphometrics.” A synonym, “taxometrics,” has been suggested for numerical taxonomy (by Mayr, 1966, in lieu of “taximetrics,” proposed by Rogers, 1963), and we shall use the adjective “taxometric” occasionally in place of the more cumbersome “numerical taxonomic.” However, the term numerical taxonomy seems well established and we retain it here.

We do not wish to widen the concept of numerical taxonomy to include every application of statistical or other numerical methods in systematic research. Our approach consists of a variety of numerical techniques; but when such techniques are not applied to problems of classification they are not included in numerical taxonomy.

The practice of numerical taxonomy embraces a number of fundamental assumptions and philosophical attitudes toward taxonomic work, which we shall discuss and defend in detail in the sections that follow. None of the attitudes and assumptions is new. They, as well as isolated attempts at a numerical treatment of taxonomic relationships, date back more than 200 years. However, we would prefer to restrict the appellation numerical taxonomy to the integrated approach of recent years, which is bringing about a revision of the theory and practice of taxonomy.

1.3 THE FUNDAMENTAL POSITION OF NUMERICAL TAXONOMY

The fundamental position of numerical taxonomy may be summarized in the following principles (modified from Sneath, 1958), which embody concepts that can be traced to Michel Adanson (1727–1806), a French botanist whose views are discussed in some detail in Section 2.2. The views represented by these principles are therefore frequently called neo-Adansonian.

1. The greater the content of information in the taxa of a classification and the more characters on which it is based, the better a given classification will be.
2. A priori, every character is of equal weight in creating natural taxa.
3. Overall similarity between any two entities is a function of their individual similarities in each of the many characters in which they are being compared.
4. Distinct taxa can be recognized because correlations of characters differ in the groups of organisms under study.
5. Phylogenetic inferences can be made from the taxonomic structures of a group and from character correlations, given certain assumptions about evolutionary pathways and mechanisms.
6. Taxonomy is viewed and practiced as an empirical science.
7. Classifications are based on phenetic similarity.

Principles 1 through 3 are elaborated in Section 1.4, “The Estimation of Resemblance”; principle 4 is treated in Section 1.5, “The Construction of Taxa”; principle 5 is discussed in Section 1.6, “The Recognition of Phyletic Relationships”; principles 6 and 7 are dealt with in Section 1.7, “Phenetic and Phylogenetic Taxonomy.”

In practice the operations of numerical taxonomy are carried out in the following sequence: organisms and characters are chosen and recorded; the resemblances between organisms are calculated; taxa are based upon these resemblances; and last, generalizations are made about the taxa (such as inferences about their phylogeny, choice of discriminatory characters, etc.).

It should be clear that generalizations about the taxa cannot be made before one has recognized the taxa; that taxa cannot be recognized before resemblances between organisms are known; and that these resemblances cannot be estimated before organisms and their characters have been examined. Therefore, although some of these steps may be in effect combined in certain computational methods, or the whole procedure may be repeated a second time for some special reason, the order of the steps within the procedure cannot be changed without destroying the rationale of the classificatory process.

1.4 THE ESTIMATION OF RESEMBLANCE

Estimation of resemblance is the most important and fundamental step in numerical taxonomy. It commences with the collection of information about characters in the

taxonomic group to be studied. This information may already exist and merely require extraction from the literature, or it may have to be discovered entirely or partly *de novo*. In most cases both of these procedures will need to be applied. For the method to be reliable, many characters are needed. All kinds of characters are equally desirable: morphological, physiological, ethological, and sometimes even distributional ones. One must guard only against introducing bias into the choice of characters and against meaningless characters (Section 3.7).

From our assertion of the equal taxonomic value of every character (see Sections 3.2 and 3.3 for our definition of a character) it is only a small step to the practice of assigning equal weight for every character when using it to evaluate taxonomic relationships. This is a point in direct conflict with traditional taxonomic practice, over which much controversy has raged. We discuss this issue in some detail in Section 3.9. The impossibility of developing criteria for a priori weighting of taxonomic characters has been quite generally conceded.

The actual computation of a measure of resemblance can be done in a variety of ways (Chapter 4). Resemblance is expressed by coefficients of similarity usually ranging between unity and zero, the former for perfect agreement, the latter for none whatever, or by coefficients of dissimilarity (distance) usually ranging between zero and an undefined positive value, the former for identity, the latter for maximal distance or disparity. The calculations are likely to be rather tedious and computer processing is needed for any but very minor studies.

Resemblance coefficients are tabulated in matrix form with one coefficient for every pair of taxonomic entities. If a symmetrical (mirror image) matrix is to be tabulated for t entities, a $t \times t$ matrix will result. Similarities among taxonomic entities can be represented geometrically by points in a space. A maximum of $t - 1$ dimensions is needed for a correct representation of the t points (taxonomic entities) in the space. The distances between the points can be regarded as taxonomic distances.

1.5 THE CONSTRUCTION OF TAXA

Classification in numerical taxonomy is generally based on a matrix of resemblances, in which taxa are constructed through various techniques designed to disclose and summarize the structure of the matrix. By some methods classification can be carried out on the original data matrix. A rough, graphical representation of the structure of the matrix can be obtained by shading the various elements of the matrix differentially (see Figure 5-16) if they have previously been roughly grouped so that supposedly similar forms are near each other. If, as is methodologically preferable, the entities are placed in the matrix without predetermined order, visual grouping is not easily accomplished without rearrangement. The various computational methods for clustering will process the data equally efficiently whether they are ordered or not. Since computational methods simul-

taneously provide some numerical evaluation of the taxonomic relationships, they are to be preferred.

These numerical methods are collectively called cluster analysis (see Sections 5.4 and 5.5). They are methods for establishing and defining clusters of mutually similar entities from the $t \times t$ resemblance matrix. These clusters may be likened to hills and peaks on a topographic chart, and the criteria for establishing the clusters are analogous to the contour lines of such a map. Rigid criteria correspond to high elevation lines that surround isolated high peaks—for example, species groups in a matrix of resemblances between species. As the criteria become more relaxed the clusters grow and become interrelated in the same way that isolated peaks acquire broader bases and become connected to form mountain complexes and eventually chains, with progress from higher- to lower-level contour lines. The clusters are generally based on phenetic resemblances only and have no necessary phyletic connotations. Differences in methods of clustering refer mainly to rules for forming clusters and for partitioning the organisms in taxonomic (character) space.

The important common aspect of all these methods is that they permit the delimitation of taxonomic groups in an objective manner, given a matrix of coefficients of relationship. Boundaries for taxonomic groups can be visualized as the contour lines already discussed or they can be represented as the intersections of horizontal transects with the branches of the tree-like diagrams of relationship commonly employed in numerical taxonomy. Comparable limits can be drawn for all taxonomic groups within a particular study. Boundaries or transects at progressively lower levels of resemblance would create taxa of increasingly higher taxonomic rank.

The number and position of boundaries or transects should follow some pre-arranged system. Clearly they will depend on the size of the similarity matrix: too many boundaries would provide too fine a classification; too few would leave much structure unrevealed. The aims of the investigator, conventions in the particular group, and questions of convenience and esthetics would all affect the placing of boundaries. We consider the number of taxa to be established at any rank a relatively unimportant and arbitrary detail. But once boundaries have been established, the structure within a taxon (and hence the number of lower ranked taxa contained therein) depends entirely on the resemblance values of entities and is not subject to the manipulations of the investigator. In other words, *the position and number of the boundaries or transects is arbitrary, but they must be based on comparable criteria in all regions of the taxonomic space under consideration.*

We view monotypic taxa or very sizeable ones with equanimity. Their occurrence does not lead us, respectively, to lumping or splitting. In taking such a position we are motivated by an effort to supply taxa with some objective and definable criteria for intragroup cohesion. This will largely depend on the method of cluster formation (Sections 5.3 and 5.4).

Biologists who use the results of taxonomic research are much concerned with the stability of a classification. The stability of a scheme based on numerical taxonomy may be affected in two ways.

1. More information (in the form of new characters) may accumulate. If the initial evaluation of resemblances has been based on a large sample of characters it is our contention that the relative similarities would change little on the addition of further characters (Section 3.8).

2. New taxonomic entities may be included in subsequent studies. Application of the previous criteria for levels and number of transects may result in new and different taxa. Agreement will have to be reached by practitioners of numerical taxonomy on whether the established system should be rearranged to suit the new results or whether the new data should be judged by standards of relationship already established—that is, whether the boundaries should be continued at the level at which they were drawn in the first study (see Section 5.10).

What level in the hierarchy are we to call a subgenus, a genus, a family? Have these terms any significance of their own other than as indications of the relative levels of the boundaries? It is generally accepted in conventional systematics that genera (and other categories) in such diverse groups as insects, birds, and flowering plants do not represent taxa of equivalent relationship. Is it possible for numerical taxonomy to set up such equivalent categories, although these would be based on entirely different groups of characters (discussion in Section 5.11)? It would appear preferable to employ a new series of terms for the hierarchic system established by numerical taxonomy, which would include in the terms a quantitative estimate of the similarity among members of the group. The phenon terminology (see Section 5.11) is an attempt at such a new system. Other, more sophisticated methods remain to be developed.

An important aspect of the construction of natural taxa is their representation. This can be done by means of tree-like diagrams indicating phenetic relationship, graphs, or three-dimensional models of taxa in phenetic hyperspace, and two-dimensional projections or stereoscopic images of such models (see Section 5.9).

The representation of taxonomic relationships in a space of two or three dimensions (ordination; see Section 5.6) leads to investigations of patterns of taxonomic structure that are of considerable ecological and evolutionary interest (see Sections 5.14 and 7.3).

1.6 THE RECOGNITION OF PHYLETIC RELATIONSHIPS

Although the major emphasis of numerical taxonomy to date has been the creation of stable phenetic groups, the methodology also offers constructive approaches for the study of phyletic lineages and evolution (Chapter 6). Measures of phenetic similarity between organisms of different geological periods will provide objective

information regarding rate and direction of evolution, or may assist in the solution of some stratigraphic problems. Much useful information may also be obtained for studies of speciation by comparing phenetic differences with genetic or geographical data.

Recent years have seen renewed emphasis on operational approaches for studies that are phyletic or phylogenetic. As we explain in Section 2.3, these two words have meanings that sometimes overlap. Such analyses aim at estimating the branching sequences of the evolutionary lines represented by the taxa in a study. The basic data comprise conventional morphological characters as well as amino acid sequences in proteins. The methods employed so far attempt to find the most parsimonious (minimum length) phylogenetic trees needed to establish the phenetic pattern of the taxa. This type of reasoning has been employed by phylogeneticists for many years. What is new is the systematization of the procedure and the simultaneous consideration of many characters, assigning them equal a priori weights but sometimes weighting them during clustering on the basis of some operational criteria. By these methods, and others under development, evolutionary lineages are being estimated by means of computers.

1.7 PHENETIC AND PHYLOGENETIC TAXONOMY

Sokal and Camin (1965) have distinguished operational, empirical, and numerical approaches to taxonomy, whose importance is considered to be in that order. In *operational taxonomy* statements and hypotheses about nature must be subject to meaningful questions, i.e., those that can be tested by observation or experiment (see Section 2.1). *Empirical taxonomy* is based on many observed and recorded characters, taxa being grouped according to a majority of shared characters (Simpson, 1961; Sokal, 1962b). The method by which the number of shared characters is determined varies with the empirical school. It may, or may not, be quantitative. Numerical taxonomy as generally practiced is both operational and empirical.

The taxonomic procedures outlined in the earlier sections of this chapter are of a strictly empirical nature. As such they are related to the procedures of some typologists in the past. This is not, however, an automatic disqualification of our views, as Simpson (1961) has implied (for a defense of "statistical typology" see Sokal, 1962b). The fundamental test of the validity of empiricism in taxonomy must be whether it can be used as a consequential and consistent method for arranging organized nature. We believe that it can be so used and that it is the only reasonable approach.

A basic attitude of numerical taxonomists is the strict separation of phylogenetic speculation from taxonomic procedure. Taxonomic relationships are evaluated purely on the basis of the resemblances existing *now* in the material at hand. These phenetic relationships do not take into account the origin of the resemblance found nor the rate at which resemblances may have increased or decreased in the past.

This attitude is taken because, as is discussed in detail in Section 2.6, we do not at the moment possess a classificatory scheme, graphic or otherwise, able simultaneously to yield information on the degree of resemblance, descent, and rate of evolutionary progress. Any scheme attempting to combine these approaches would be excessively complicated. The separation of overall similarity (phenetics) from evolutionary branching sequences (cladistics) is an important advance in taxonomic thinking (see Section 2.3 for a detailed discussion). It is often difficult to make this separation, in view of long established methods of thought acquired with the early training of most practicing taxonomists of the present generation. Yet the most recent generation of taxonomists just out of graduate school has had little difficulty with this concept and recognizes the value of the separate ways of regarding taxonomic relationships.

Not only do we insist on the separation of phenetic from phylogenetic considerations in taxonomic procedure, but we also feel that only phenetic evidence can be used to establish a satisfactory classification. We hold this belief for several reasons.

1. The available fossil record is so fragmentary that the phylogeny of the vast majority of taxa is unknown. Evolutionary branching sequences must be inferred largely from phenetic relationships among existing organisms.
2. Phenetic classification is possible for all groups. By contrast, cladistic classification, based on branching sequences, requires historical inferences about the direction of evolution in a group of organisms.
3. Even when fossil evidence is available, this evidence itself must first be interpreted in a strictly phenetic manner—with the exception that a time scale is given in addition, which may restrict certain interpretations of the phylogeny—since the criteria for choosing the ancestral forms in a phylogeny are phenetic and are based on the phenetic relationship between putative ancestor and descendant.
4. From the point of view of biology in general, it is probably of more interest to describe the overall similarity of organisms than their branching sequences. If the classifications are to have predictive value, it is evident that those based on overall similarity will be most predictive.

To accept taxonomic relationships based only on phenetic criteria is difficult for many present-day biologists. For almost a century there has been an intimate conceptual association between taxonomic and phylogenetic reasoning, so that terms such as “specialized,” “primitive,” “analogous,” and many others have assumed double meanings whose differences are rarely made distinct. Yet the effort to do so must be made to achieve clarity and to progress in the understanding of taxonomic principles.

The issues summarized in this section are discussed in greater detail in Sections 2.2 to 2.6.

1.8 THE ADVANTAGES OF NUMERICAL TAXONOMY

The methods discussed in this book have a number of advantages, which will be listed below. First, however, we should refer to their principal aims, *repeatability* and *objectivity*. If observations are repeatable within an acceptable error and if taxonomic procedures are clearly circumscribed, it is hoped that numerical methods will lead different scientists employing the same data base and working independently to obtain comparable estimates of the resemblance among any group of organisms.

Closely tied up with repeatability is the notion of objectivity. The *Random House Dictionary of the English Language* defines "objective" as "free from personal feelings or prejudice; based on facts; unbiased." This is a relative concept, seldom fully realized. By including many characters without previous arbitrary selection or elimination, and by providing explicit methods of processing the data and evaluating the results, we reduce subjective bias and increase objectivity.

Of the manifold advantages of numerical taxonomy we may briefly mention the following, which will be discussed in greater detail in appropriate sections of the text.

1. Numerical taxonomy has the power to integrate data from a variety of sources, such as morphology, physiology, chemistry, affinities between DNA strands, amino acid sequences of proteins, and more. This is very difficult to do by conventional taxonomy.

2. Through the automation of large portions of the taxonomic process, greater efficiency is promoted (Sokal and Sneath, 1966). Thus, much taxonomic work can be done by less highly skilled workers or automata.

3. The data coded in numerical form can be integrated with existing electronic data processing systems in taxonomic institutions and used for the creation of descriptions, keys, catalogs, maps, and other documents.

4. Being quantitative, the methods provide greater discrimination along the spectrum of taxonomic differences and are more sensitive in delimiting taxa. Thus they should give better classifications and keys than can be obtained by the conventional methods.

5. The creation of explicit data tables for numerical taxonomy has already forced workers in this field to use more and better-described characters. This necessarily will improve the quality of conventional taxonomy as well.

6. A fundamental advantage of numerical taxonomy has been the reexamination of the principles of taxonomy and of the purposes of classification. This has benefited taxonomy in general, and has led to the posing of some fundamental questions discussed in numerous subsequent sections of this book and in many publications in the recent systematic literature.

7. Numerical taxonomy has led to the reinterpretation of a number of biological concepts and to the posing of new biological and evolutionary questions. Thus, the method is coming into its own as a heuristic tool in biological research. In Chapters 6 and 7 and Section 10.3 we take up these novel aspects of numerical taxonomy.

1.9 IDENTIFICATION OF SPECIMENS

Once a classification has been established by some operational method, the construction by computer of appropriate keys for identifying specimens is an obvious next step. Some of the errors that are occasionally committed in setting up dichotomous keys would be avoided by logical and consistent programs. For purposes of taxonomic keys and identification of specimens, weighting of characters must be introduced to emphasize those characters that are most effective in distinguishing between previously established taxa. Weight must not only be statistical in the sense of providing efficient criteria for differentiation, but must also take into account the ease of observing or measuring a given character, the probability of its being recordable in a study specimen, the chance of structures bearing the character being damaged or confused, etc. Alternative structures must be provided for identification in case the preferred structures are not available.

Although the logic of key making may change little, the physical form of keys can be adapted to electronic data processing and a variety of forms may prove useful in different circumstances. The more sophisticated keys will usually be probabilistic; that is, they will give, on assumptions reasonable in the light of current knowledge, the likelihood that the identification of a given specimen is correct. On-line computing permits a new interaction between taxonomists and computers for the identification of specimens. Taxonomic keys can be stored in the computer and can be presented in the form of a dialogue between the taxonomist and the computer. The subject of keys and identification is discussed in some detail in Chapter 8.

1.10 NOMENCLATURE

Any change in taxonomic procedure will inevitably entail a change in nomenclature. Although in a formal way it can be, in practice the subject of nomenclature cannot be dissociated from taxonomy per se. As long as we are going to study and group organisms, by whatever criteria, we will want names or labels to refer to them. We shall see in Chapter 9 that the present system of nomenclature does not adequately fulfill the various functions expected of it. Some changes in the system will become necessary simply due to the inevitable adoption of electronic data processing for taxonomic information. Proposals for revision of the nomenclature are only in their early formative stages. They will be reviewed in Chapter 9.

If numerical taxonomy has a special effect on nomenclature, it will be mostly through revision of existing classifications. If a numerical study shows considerable changes from the relationships implied by the orthodox nomenclature, the author has no choice but to alter the classification and to make such changes of name as are then required. One must do this if one is to provide other biologists with the benefits of improved classifications.

The present overdependence of the system of nomenclature upon types has been changed by the advent of numerical taxonomy, for it is now in principle possible to

determine the limit of taxonomic groups, which formerly had been a matter of individual opinion. Numerical nomenclature for intermediate forms may also prove of some use.

1.11 THE DEVELOPMENT OF NUMERICAL METHODS IN TAXONOMY

The earliest attempts to apply numerical methods to taxonomy date from the rise of biometrics in the last century. As early as 1898 Heincke used a measure of phenetic distance to distinguish between races of the herring, while in 1909 Czekanowski employed a distance coefficient in physical anthropology. It was early realized that biometrics could be applied to systematics, but the only important development was that of discriminant functions (Fisher, 1936), which is useful in only a limited area of taxonomy.

One of the earliest statistics of interest to systematists was the "Coefficient of Racial Likeness" (Pearson, 1926). It was extensively applied in physical anthropology but does not seem to have been taken up by taxonomists. The *C.R.L.* was a type of taxonomic similarity coefficient, and was subsequently developed by Mahalanobis in the form of the "Generalized Distance" statistic, which is also formally a coefficient of this kind (see Rao, 1948). Anderson and Whitaker (1934) and Anderson and Abbe (1934) employed a similar statistic, which they called the "General Index." These statistics, though mathematically adequate, did not lead to notable advances for two reasons: (1) they were developed mainly as discriminant functions to aid in the allocation of individuals to existing taxa and not as methods for creating taxa; and (2) as a consequence of (1) these workers selected principally those characters which gave the best discrimination between the taxa, and—since it is usually only necessary to employ a small number of such characters once they have been found—these methods were in practice based on few rather than many characters. Some of the characters were selected on a priori grounds, and their small number led to instability on repeating the work with other characters. The lack of adequate computing facilities in the early part of this century also limited multivariate work to few characters. These techniques, with others developed later, have been very widely and successfully used for the study of certain limited taxonomic problems. We may, for example, cite the elegant work of Blackith (1957) on sexual and phase variation in locusts, Reymont (1963) on fossils, and that of Jeffers and Black (1963).

Some studies with aims similar to those of numerical taxonomy today should be mentioned. Smirnov (1925) established types on a quantitative basis. His work has been discussed and evaluated from different points of view by Hennig (1950) and Sokal (1962b). Haltenorth (1937) in a study of similarities among eight species of large cats developed a coefficient similar to that of Cain and Harrison (1958). At about the same time Zarapkin (1939) developed a rather elaborate technique called

Divergenzanalyse, arriving at a quantity analogous to taxonomic distance. The *Affinitätsrechnung* of Schilder and Schilder (1951) is also a computation of taxonomic distance. We believe that these methods did not succeed at the time they were developed because of the entrenched nature of phylogenetic systematics and since for any substantial number of characters or taxa the methods advocated by these authors presented computational difficulties insurmountable at the time.

Other early methods, specifically intended for taxonomy, are those of Forbes (1933), Anderson and Owenbey (1939), Sturtevant (1939, 1942), Boeke (1942), James (1953), Stallings and Turner (1957), Hudson, Lanzillotti, and Edwards (1959), and Chillcot (1960), all based on variations of matching coefficients. Terentjev (1931, 1959) introduced a method of defining clusters of characters called "pleiades." These authors did not develop their methods sufficiently to meet the main needs of numerical taxonomy and hesitated to give equal weight to every feature or to employ large numbers of characters. Similar trends can also be seen in the history of bacterial classification, where the earlier reliance on a few morphological or physiological characters has given place to attempts at classification in the Adansonian tradition (see Sneath, 1962). There developed in Poland during the 1950's the Wroclaw school of taxonomy based on the coefficient of Czekanowski, whose procedures for classification were similar to numerical taxonomy (Florek et al., 1951a,b; Perkal, 1951). Instead of a formalized and exact clustering, the Polish workers use graphic approaches involving linkage diagrams and multiple contour lines. The technique apparently remained undiscovered by English-speaking numerical taxonomists whose approach relied more upon numerical and computer techniques. However, a review by Hubac (1964) drew attention to the work of the Wroclaw group and caused at least one worker to adapt the Wroclaw techniques for the graphic representation of conventional taxometric results (Moss, 1967). (See also Section 5.9.)

One of the main conceptual difficulties that retarded progress in numerical taxonomy was the problem of weighting and the liberating effect of accepting equal weighting can scarcely be overemphasized. The use of many characters is also a prerequisite of progress. In addition, the use of methods of cluster analysis in building the taxonomic hierarchy has been a major advance. It is in these three developments that numerical taxonomy chiefly differs from the earlier ideas and methods.

The modern work in numerical taxonomy was initiated by the almost simultaneous publications of Sneath (1957a,b) at the National Institute for Medical Research, London, on bacteria, and of Michener and Sokal (1957) and Sokal and Michener (1958) at the University of Kansas, on bees. These authors developed their ideas independently and, upon learning of each other's work prior to the publication of these papers, initiated a correspondence and collaboration which resulted in the development of principles and techniques (Sneath and Sokal, 1962) and culminated in the publication of a book on this subject (Sokal and Sneath, 1963). At almost the same time two other groups independently developed numerical approaches to the taxonomy problem. Cain and Harrison (1958) at Oxford

proposed a mean character difference for taxonomic purposes, and Rogers and Tanimoto (1960) at the New York Botanical Garden developed a taxonomic technique based on a coefficient of association.

Since those early days, and especially since the publication of our *Principles of Numerical Taxonomy*, there has been a very rapid increase in the development of methods for numerical taxonomy and in the application of these techniques. The subject is still in a period of active growth and change and we lack the historical perspective to single out every significant landmark in its development, but the following chronological chart may give an idea of the development of the field. The number of papers published in the field has risen from 60 that appeared in the five years from 1957 to 1961 to over 200 annually in recent years.

- | | |
|-----------|---|
| 1957–1961 | Development of first methods and of theory of numerical taxonomy. |
| 1962 | First test of the nonspecificity hypothesis (Rohlf); first criterion of goodness of a classification (Sokal and Rohlf); first comprehensive publication of theory and methods (Sneath and Sokal). |
| 1963 | Publication of <i>Principles of Numerical Taxonomy</i> (Sokal and Sneath). |
| 1964 | First cladistic technique for continuous characters (Edwards and Cavalli-Sforza). |
| 1965 | First cladistic technique for discrete characters (Camin and Sokal); first critical analysis of the methodology of numerical taxonomy (Williams and Dale). |
| 1966 | First paper on taxonomy of scanned images (Sokal and Rohlf); electronic data processing linked to taxonomy (Sokal and Sneath); first generalization of hierarchical clustering methods (Lance and Williams). |
| 1967 | Numerical cladistics of proteins (Fitch and Margoliash); first attempts at definitions of operational homology (Jardine, Key, Sneath); congruence of phenetics with genome message (Heberlein, De Ley, Tijtgat); first computer-oriented identification methods (Goodall, Morse, Dybowski, Franklin). |
| 1968 | Investigation of intra-OTU variation (Crovello). |
| 1969 | First generalized methods for numerical cladistics (Kluge and Farris); reexamination of the logic of biological classifications (Jardine); attempts at statistical validation of clustering procedures (Mountford, Switzer). |
| 1970 | Reexamination of the biological species concept (Sokal and Crovello). |
| 1971 | Publication of <i>Mathematical Taxonomy</i> , stressing logically consistent criteria for taxonomy and overlapping clusters (Jardine and Sibson). |