Simplicity and Informativeness in Systematics and Phylogeny

JAMES S. FARRIS

Abstract

Farris, J. S. (Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11794) 1982. Simplicity and informativeness in systematics and phylogeny. Syst. Zool., 31:413-444.—Advocates of syncretistic classification have generally held that the descriptive and explanatory roles of the biological reference system should be kept separate, and that description and explanation impose conflicting goals on classification. I show that view leads to contradictions, and I summarize earlier demonstrations that the phylogenetic system satisfies both goals at once. Pheneticists have advocated ordinations as an alternative to hierarchic classification, claiming for them superior descriptive power. I show that ordinations are more general than had been thought, as hierarchies can be subsumed under such structures. This viewpoint allows a great many classificatory criteria to be analyzed under a factor or regression model. Present phenetic criteria for clustering amount to evaluating a multiple correlation as the sum of simple correlations. This fallacy can be avoided by regressing data multiply on classificatory factors (groups), but then perfect fit to data can always be achieved, so that the classificatory problem becomes one of choosing among alternative sufficient bases. This problem is approached by analogy with simple structure criteria of factor analysis, and it is found that classificatory simple structure leads to the phylogenetic parsimony criterion. From this new viewpoint the phylogenetic system thus emerges again as a unified descriptive and explanatory structure. Objections to the descriptive power of the phylogenetic system are discussed with emphasis on recent claims by Rohlf and Sokal. It is shown that attempts to defend phenetic classification have comprised nothing other than abandoning or rendering untestable every claimed advantage of phenetic methods. There is thus no possible substantive basis for preference for phenetic grouping. [Phylogenetic systematics; cladistics; information content; simplicity; efficiency; simple structure; factor analysis; phenetics.]

Description and Explanation

The taxon names of the general biological reference system are used as a vocabulary of scientific communication, for describing and discussing biological observations and theories. As such they must be generally agreed on—otherwise attempts at communication would be futile—but further, like any technical terminology, the reference vocabulary must be suited to the nature of its subject. The nature of the subject of any field of science is itself part of the theoretical content of that field, and so scientific terminologies reflect and embody such theory. Scientific vocabularies are then not only a means of description, but also of discussion and explanation in terms of current understanding of the natural processes that give rise to observations. It is the nature of the subject matter that makes it most meaningful and fruitful to use a periodic arrangement—rather than, say, a hierarchy—in the table of the elements, and likewise a hierarchic—and not periodic—arrangement in the biological reference system.

That classifications—or vocabularies—embody theories on the nature of the things categorized is seldom questioned in most fields of science, and biological systematists as well normally recognize the relationship between explanation and scientifically meaningful description. Hennig (1966) certainly understood that connection well, but taxonomists of quite different views, such as Mayr (1969), have also emphasized that biological classifications represent theories. Nonetheless, there is substantial variation in ideas on the consequences of that connection. Hennig realized (as had Darwin) that the natural hierarchy of organisms is produced by the branching pattern of genealogy, so that a description of organic diversity—a classification—meaningful in
terms of underlying process must reflect that genealogy. Mayr also relied on the Darwinian explanation for the existence of the biological hierarchy, but maintained nonetheless that description of diversity requires departure from genealogical grouping. Classifications, according to Mayr (1974:95), are to be formed by "grouping according to a largely phenetic approach." His position in this respect is much like that of Sneath and Sokal (1973:56): "In most cases ... the cladistic and phenetic classifications will largely coincide. However, in the possibly rare cases where this is not so, the general superiority of a phenetic over a cladistic classification becomes even more evident."

Mayr's opinion is certainly much like that of some pheneticists. McNeill (1982), for example, characterized phenetics as "dependent on evolution for its success." This amounts to the claim that evolution justifies or explains the phenetic hierarchy. The success of the periodic table, after all, is due to the ability of the periodic behavior of the elements to justify and explain it. Other pheneticists, however, take quite a different stance. Sneath and Sokal (1973:56) characterized phenetic classification as requiring "only description." They objected to explanation in classification as perpetuating "a system which is hypothetical and unstable by the very nature of its operations." They thus differ from the other authors cited in viewing classification empty of theoretical content as not only possible, but desirable.

It would appear that Sneath and Sokal's attitude on the role of explanation stems from a defective view of how explanation is achieved. They described the theory of evolution as "the most adequate, most unitary, and indeed the simplest hypothesis to which a great variety of phenomena—geographic distribution, physiological adaptation, morphological similarity, or biocenotic complexity—can be related (p. 54). But they then continued, "Phylogeny can thus be seen as the central cause of much biology, yet it cannot be used for an explanatory concept, as phylogenies are not known in the vast majority of instances." The inconsistency of that claim is obvious. Scientific explanations—theories—are seldom literally known. Theories are not taken as explanations because they are known to be true, but rather truth is attributed to them precisely because they provide explanations. Phylogeny in the abstract, furthermore, can hardly be an explanation of any of the phenomena listed. Morphological similarities, for example, might well have turned out some other way for some conceivable phylogeny, hence they can be explained only by particular phylogenies. If different phenomena, such as morphology and biogeography, could not be brought into agreement with a single particular phylogenetic explanation, then it would be nonsense to claim that phylogeny could explain all phenomena at once. Possibly Sneath and Sokal feel (with others) that the general—though not particular—explanatory power of phylogeny is established by just a few well-studied cases, whereas no such achievement is possible for most cases. But it was precisely the ability of phylogeny to explain "a great variety of biological phenomena" that provided the evidence for evolution to begin with. If further cases could not be explained in the same way, then evolution would either have to be abandoned as a theory, or else it would have to be awarded the status of immunity to further evidence, in which case it would not be a scientific theory at all. If Sneath and Sokal's position were based on consistent application of scientific principles, then, they would either have to abandon evolution, or else admit that explanatory phylogenies (as distinguished from those "known" to be true) can generally be found.

There are two further fallacies in the view that a phenetic general reference system can be purely descriptive, totally lacking a hypothetical element. Phenetic classifications cannot be based only on description; some principle must guide the selection of the method used to con-
struct them (cf. Farris, 1977). Nor can a phenetic general reference system entirely lack a hypothetical element, even if description is its only admitted goal. Any classification is necessarily constructed just from particular available characters. If it were deemed to reflect only those characters, however, then it would be (as pheneticists put it) a special purpose, rather than a general purpose, classification. The general reference system must, even on a pheneticist view, be a general purpose arrangement. Proposing a phenetic system of grouping as the general reference system then implies the necessarily hypothetical claim that further characters would conform to the same scheme.

Sneath and Sokal might deny that the hypothesis of generality of phenetic grouping is an evolutionary theory. If so, that denial would constitute at best merely a dispute of definitions, for they do not claim that organic diversity has been produced by something other than evolution. The hypothesis clearly prohibits evolutionary events that are otherwise quite conceivable. Farris (1971), Mickevich (1978, 1980), and even Michener (1978) have all pointed out that mosaic evolution (cross specialization, "discordant" evolution) can cause variation among phenograms for different suites of characters, so that no one phenetic arrangement is generally descriptive.

Of course pheneticists seldom deny that their methods classify the products of evolution. Their position is instead that classifying products ought to be free of suppositions about the process. But process and product can scarcely be entirely independent. Sneath and Sokal did not—and could not—succeed in devising classificatory methods free of the effects of evolutionary phenomena. In attempting to ignore phylogenetic considerations, they merely left the relationship of process to product unrecognized, and so misunderstood.

McNeill's view seems different from that of Sneath and Sokal, but it rests on a similar lack of analysis. He feels that the hierarchy of evolution explains the hierarchy of classification, but also that the same explanation would apply equally well to syncretistic and phylogenetic groupings—that is, to different hierarchical schemes. In the abstract, the idea that chemical properties of elements vary periodically as a function of atomic number might lead to any of a number of periodic tables. But that concept together with the known atomic numbers and chemical properties of elements determines one particular table. Likewise the recognition that evolution produces a hierarchy might by itself be consistent with any classificatory hierarchy, but in the presence of relevant observation it cannot very well explain two different hierarchies at the same time.

Of course it is what observations are relevant that is crucial. Phenetic and phylogenetic methods of grouping differ in the way that they relate character information to the selection of a classification. One might say that they reflect different conceptions of what constitutes evidence for grouping. The phylogenetic procedure of grouping according to putative synapomorphy is based on the connection between apomorphic similarity and kinship, and that connection in turn follows directly from the recognition of evolution as descent with modification. Pheneticists do not usually regard their sort of evidence—raw similarity—as pertinent to evaluating theories of kinship, but rather as a reason for grouping in the abstract. McNeill, in fact, emphasizes that phenetic methods make "no attempt to reflect evolution" (1982:338). But there is a necessary connection between evidence and explanation. An observation can provide evidence for one theory over another only if it is better explained by one of the alternatives. If the hierarchy of evolution is used to justify hierarchies constructed by grouping according to raw similarity, then this entails the claim that raw similarity provides evidence on that evolutionary hierarchy, that is on phylogenetic relationship.

McNeill, to be sure, does not draw that
conclusion, but neither does he argue against it. Rather he never considers the problem, and his thinking is thus seen to differ only in degree from that of Sneath and Sokal: he never bothered to analyze the implications of his claims. Much the same can be said of Mayr. Like McNeill, he resorts to evolution to defend his methods, and like pheneticists generally he justifies his preference for syncretistic arrangements through the assertion that grouping according to raw similarity yields descriptively most effective classifications. For Mayr, as for pheneticists, this amounts to the claim that description and explanation are contradictory goals. That is a most curious notion indeed, inasmuch as no theory can very well be said to explain what it cannot describe. That position, at any rate, was based on nothing, for no advocate of syncretistic grouping had offered any analysis of how phylogenetic classifications explain points of similarity among organisms, or of how any classification describes the distributions of those features.

I have investigated both subjects, and my results show a strong connection indeed between description and explanation: both goals lead to the phylogenetic parsimony criterion. In an earlier paper (Farris, 1979b) I showed that a hierarchy is able to provide an efficient description of characters to the degree that each of the features need occur in the diagnoses of few taxa. Classifications efficient in this sense facilitate description and storage of character information, as they minimize the requirements for redundancy in summarizing the distributions of features: mammary glands need be associated just with the mammals, not listed separately for each mammal. For the same reason, efficient hierarchies are descriptively most informative. They allow the information in a character matrix to be conveyed with a minimum number of group names, so having highest information content on characters per group. In a more recent study (Farris, 1983), I demonstrated that a hypothesis of phylogenetic relationship is able to explain observed similarities among organisms to the extent that it can account for them as identities by virtue of common descent. Features cannot be so explained, of course, just to the degree that the putative tree requires that they originate multiply and independently, that is that it requires hypotheses of homoplasy. Application of the phylogenetic parsimony criterion, choosing the hypothesis of relationships to minimize requirements for hypothesis of homoplasy, then leads to schemes of greatest explanatory power. As I have stressed in both those papers (also Farris, 1980a), the presence of a feature in the diagnoses of several groups corresponds to the interpretation of multiple origins of that trait. When a feature shared by taxa is covered by a single entry in the diagnosis of a more inclusive taxon, likewise, this corresponds to the explanation of that sharing as a consequence of inheritance from the stem species of that inclusive taxon. There is thus a direct correspondence between the descriptive utility of a classification and the explanatory power of that scheme as a phylogenetic hypothesis.

REGRESSION

The equivalence of efficiency of explanation with efficiency of description is precisely what one would want in a scientific vocabulary (cf. Sober, 1975, 1983; Farris, 1983). For evolutionary taxonomists, then, that should settle the matter of choosing a basis for the general reference system. Not so for pheneticists, who have never been much interested in theoretical content. They might well attempt to salvage phenetics by proposing some other relationship between characters and arrangements. They would seem to have a wide choice of procedures to defend, as there are a great number of phenetic clustering methods. Beyond this, some pheneticists write of such ordination methods as principal components analysis and factor analysis as methods for representing "taxonomic structure," considering these descriptively superior to hierarchies. As has been pointed out before (Schuh and Farris, 1981; Farris, 1981), or-
Ordinations as they are now employed are hardly distinctively phenetic, nor are they even taxonomic, for they define no groups. For much the same reason, they offer just about the worst possible choice of a reference system, providing neither a vocabulary nor an explanatory structure in any obvious way. None of that, of course, would deter pheneticists from advocating ordinations, if by doing so they might hope to maintain their stance that explanation and scientific description can be kept separate. At any rate, ordinations might be made taxonomically more useful in some way.

The diversity of nonphylogenetic methods would seem to complicate the task of analyzing them in a coherent framework. I shall show, however, that these seemingly different procedures can be so treated, and that doing so yields new insight on the general weakness of phenetic clustering.

Ordinations as they are now used typically appear as drawings showing the relative positions of points in the space of possible combinations of attributes. As such they seem utterly different from classifications, which, if drawn at all, appear as tree diagrams. The formal structures underlying the drawings, however, are much more alike.

An ordination is calculated by choosing a set of axes (factors) for the character space, so that the original characters are represented as functions (usually linear combinations) of these axes. Applications of ordinations usually employ just a few of those axes, these being chosen to reflect the data as closely as possible. The reason for this is that no more than three independent axes can be drawn as a spatial representation. The limitation is of the diagram, not of the ordination as such, and no one seems to have argued that it would be undesirable to represent data still more accurately through a greater number of axes, were it feasible to visualize such a representation.

The descriptive adequacy of a (low-dimensional) ordination is typically judged by the degree to which variation in the new axes accounts for variation in the original characters, thus in effect by the (multiple) correlation obtained by (multiply) regressing original characters on the new axes. Evaluation may also be based on the correlation between distances (compliments of raw similarities) between pairs of terminal taxa computed from the original characters and from the ordination. As Sneath and Sokal (1973: 278) put it, “cophenetic values based on ordinations are the actual distances (in any specified metric) in the ordination.” They term the correlation between original and cophenetic values cophenetic correlation.

As is well known (cf. Farris, 1979a, b) pheneticists also employ the cophenetic correlation between original and tree-implied distances to evaluate hierarchies. Classifications and ordinations then both provide summaries of the distances among taxa, and the two representations are alike in the way in which the conformity of summary to data is judged, although in phenetic practice the way in which the summary is made differs between the methods.

That observation applies just to distances, but the same connection can be seen between ordinations and phenetic methods for assessing conformity of classifications to characters. In the k-means method (MacQueen, 1967) a partition (single-level classification, suite of mutually exclusive classes) of objects into k classes is chosen to minimize the variation (in the sense of sum of squares) within groups. This can be viewed as a kind of regression analysis. The independent variables (factors) are group memberships, the centroids of the groups provide the predicted values from the regression, and the variation within groups represents residual variation, departure of observation from prediction. To the extent that the selected grouping succeeds in accounting for variation among objects, that variation is “explained” by the grouping, just as with a regression. (Although the explanation is only formal, and need not correspond to a scientific expla-
nation.) To the extent that the observed traits of members of groups correspond to the predicted values of those groups, the scheme is predictive or informative concerning those traits. the predicted values for groups provide the means of retrieving information on traits from the grouping. Likewise in the method of Gower (1974), each group of a partition is assigned a predictor state for each character, and membership of groups (for a given number of groups) is chosen to minimize the total number of departures of observed from predicted character states. In the method of Duncan and Estabrook (1976) a most “informative” partition is selected to minimize entropy (maximize negentropy, i.e., Shannon information) of qualitative characters conditional on the partition. As entropy measures variation in qualitative traits, this procedure is analogous to the others. It has the shortcoming that it provides no predicted states for the groups, so that there is no means of actually retrieving the information that the arrangement supposedly contains. I presume, however, that some predictor function could be devised.

These various proposals offer different formulae for measuring departure of observed from predicted character states. While a single principle underlies them all, they might well in practice, nonetheless, lead to preference for different partitions. Justification of a phenetic procedure would thus require defending one measure of conformity against other possibilities. But that issue will be seen to be inconsequential, for its solution will be obtained incidentally in the course of addressing a more pressing problem. These methods all apply just to single level classifications, whereas systematic practice requires a means of evaluating alternative hierarchies.

Gower proposed a simple way of extending his measure to hierarchies. He regarded a hierarchic classification as a series of partitions—one for each hierarchic level—and assessed it by summing his single-level statistic over levels. No such generalization has been suggested for the other measures mentioned here, but the same idea would apply just as well to the sum-of-squares and entropy statistics. While phenetic clustering methods have not been based on any explicit concept of conformity of characters to hierarchies, they do seem to correspond to Gower’s sort of extension. Operating by repeatedly grouping mutually most similar taxa, phenetic clustering in effect tries to maximize the homogeneity (averaged over characters) of groups at every level. For the sum-of-squares measure, UPGMA performed on squared Euclidean distances would have that effect. Clustering according to the entropy statistic itself (as in Williams et al., 1966) would accomplish the same for the entropy measure. Divisive clustering algorithms fit the same characterization as well. That of Edwards and Cavalli-Sforza (1965) proceeds by finding a sum-of-squares optimal binary partition for the entire collection of objects, then likewise dividing each of those groups, repeating this until a resolved hierarchy is formed. It simply applies k-means (k = 2) to each selection, and so attempts to optimize, if anything, the sum of sum-of-squares criteria. Again, the same would hold for divisive clustering according to an entropy measure.

SUMMING

While measures of conformity vary, then, phenetic methods all embody the same general view of how hierarchies conform to characters. The relationship between a character and a single level classification is like a correlation, and that between characters and hierarchies is effectively a sum of correlations, one for each hierarchic level, or, perhaps, one for each group. This sort of evaluation of hierarchies seems an obvious extension from the single level case, but in fact it encounters some difficulties. One general problem with this approach results from a characteristic common to the single-level statistics. For a single character
with two states, any of these measures can be optimized by the binary partition each of whose groups is distinguished by one of the states. But any of them can also always be optimized by a trivial partition, one with a separate group for each object. This occurs because the statistics are directly sensitive only to the homogeneity within groups, rather than the conformity of the groups to the pattern of distribution of the states. There seems no dispute that the latter grouping is generally undesirable. Sneath and Sokal, for example, comment that “a single presence-absence character would obviously only permit the recognition of two taxa, the ‘haves’ and the ‘have-nots.’” Similarly, Sokal (1977:5) observed that for natural classifications, “states should be homogeneous within taxa and heterogeneous among them.” I showed before (Farris, 1979b) that considerations of descriptive efficiency (informativeness, “naturalness” in a purely descriptive sense) would lead to preference for partitions whose groups are distinguished by states, not just homogeneous.

Gower did address this problem for partitions; he proposed a second measure that can be used to inhibit formation of undistinguished groups. Duncan and Estabrook attempted the same, and in application of k-means this principle is also typically taken into account if only by inspection. Since Gower’s second measure is not included in his evaluation of hierarchies, however, it is not immediately germane to that problem. We shall see later, however, that the problem of groups without distinguishing states does arise in hierarchies in another way.

Assessing a hierarchy by summing measures such as those discussed over levels has at least two drawbacks, seemingly different, but actually related. In Gower’s procedure, each departure of the predicted state of an object from the observed state counts as a single error, with the result that choice of a partition is sensitive to abundance of objects. For the simple data,

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

the best partition into two groups would be AB/CD. Suppose however that A and B are higher taxa, the members of each all alike for the characters given, while C and D are single species. One might wish to evaluate a partition of species, rather than of predefined groups. If A and B each included five species, and if the fit measure were evaluated on the set of 12 species, either A/BCD or B/ACD would be a better division into two groups than is AB/CD according to Gower’s statistic. The same property carries over if a hierarchy is evaluated by summing values for levels. If each of A–D is regarded as a single object, the hierarchy ((A, B), (C, D)) is optimal by Gower’s measure, whereas if A and B each included five identical objects, (A, (B, (C, D))) would be preferred by the same method. In practice such a criterion might lead to separation of rodents and bats at a higher level than one of those were separated from monotremes, simply as a result of the great number of species in the first two groups.

The second problem is more directly concerned with summation of single-level measures. A character that serves to distinguish some particular group, rodents, say, will be well correlated with that group of the hierarchy, but it can hardly be correlated with more inclusive groups—mammals, amniotes, . . . , vertebrates, . . . , metazoans. If fit of the character to the hierarchy is measured by summing single-level fits over levels, the heterogeneity of rodent characters within those more inclusive groups is effectively taken as a weakness of the hierarchy. But that is illogical, as no hierarchy could avoid that “fault” unless it made rodents coordinate with all other life—and even
then the same problem would arise for the distinguishing characters of other groups. The difficulty in this case seems to be that the interpretability of the measure on partitions does not carry over when summing over levels. For a partition, heterogeneity within groups implies that some observed states do not match group predictors, so that there are errors of prediction. So far as Gower's measure is concerned, the same interpretation applies to the heterogeneity of rodent characters among vertebrates, but that interpretation is no longer reasonable, for no one could maintain that the present reference system makes incorrect predictions concerning which organisms have the distinguishing characters of rodents.

The other side of the same coin is that characters homogeneous within a group are counted as successful predictions repeatedly, at every level of the hierarchy at which the group is distinct. In the classification (A, (B, (C, D))) Gower's method counts a state distinguishing A in each of three successive partitions, A/BCD, A/B/CD, and A/B/C/D. The features distinguishing D, however, are counted as correct predictions just once, in the last of those three. That would appear to imply that the distinctive traits of D are less well predicted than those of A, but that is also illogical, as those features are predictively just alike: each sort simply distinguishes a terminal taxon. If heterogeneity of D's distinctive traits within BCD were taken to imply a fault of prediction of those features, that would again amount to the view that the present system is defective because not all vertebrates are rodents. The descriptive utility of the taxon A itself, moreover, is effectively judged according to how the other taxa are subdivided, but again, there is no apparent logical connection. Here, too, the interpretability of the single-level statistic does not carry over. In the case of a partition, each state of each object provides a single source of evidence on the choice of grouping. Each group has a predictor for each character, this does or does not match the state of an included object, and the preferred partition minimizes errors. When the same measure is summed over levels, however, "separate" predictions on each object are made at each level, so that each entry in the character matrix is treated as several lines of evidence. In fact those multiple uses of the same observation cannot be independent, nor can the predictions made at different levels. The groups of a hierarchy are required to be well-nested, so that lower- and higher-level subdivisions must be related.

Similar difficulties arise in other suggestions for assessing hierarchies. Archie (1980) has proposed a whole suite of statistics which, while differing in detail, share the characteristic of summing a fit measure over several levels. Interestingly, Archie strongly criticizes Gower's method. These procedures do not merit much discussion, as Archie himself was unable to find any basis for deciding among the many possible statistics that he had devised. This reflects a simple lack of any coherent theoretic framework in his approach. It seems worth noting in passing, however, that his thinking is based on a quite astounding premise: that only one state of each character can be predicted. On that supposition, once it has been determined to predict that a vertebrate will have a dorsal tubular nerve cord, it is then not permissible to predict whether a crab will lack that structure. It is rather difficult to see how a measure based on such a principle could be of much use in assessing the practical utility of classifications.

MULTIPLE REGRESSION

The seemingly obvious extension of single-level statistics to hierarchies just by summing, then, leads to a host of difficulties. The idea that fit of classifications to data can be treated as a regression of characters on group memberships provides a simple way to encapsulate these problems and also offers a solution to them. Assessing a hierarchy by summing fit measures over levels is like evaluating a multiple correlation by summing
single correlations, and the drawbacks of that obviously faulty procedure can be avoided by instead regressing characters multiply on group membership.

Adopting a multiple regression view obviates the problems of Gower's sort of summing. As a multiple regression would provide just a single prediction equation for each character, each state of a taxon would be compared to its prediction just once, so that there would be no redundant use of evidence. Dependence between levels would automatically be taken into account by a multiple regression, so that the nesting of a hierarchy would no longer offer a difficulty. Specious errors resulting from variation of rodent characters among vertebrates would not arise. In a multiple regression, a dependent variable is well predicted if it is well correlated with any of the predictor variables, so that the correlation of distinguishing rodent characters with the Rodentia would be correctly assessed as implying that those characters are well predicted. The imperfect correlation of those same characters with other groups would now suggest no spurious errors. Finally, as will be seen, abundance of taxa would have no effect on the evaluation, although the reasons for this will turn out to be related to a further issue in assessing hierarchies.

To show how to apply a multiple regression approach to classification, I shall first consider the usual least-squares procedure. Each character of a data matrix $C$ is to be represented as a linear combination of group memberships. Denoting the matrix of coefficients of those combinations as $L$, and the representation of the hierarchy itself as $H$, the model can be written $C = HL$. For a given hierarchy $H$, the least-squares choice of coefficients $L$ is found by $L = (H'H)^{-1}H'C$. Here $C$ is taken as having one row for each terminal taxon and one column for each character. $H$ has a row for each terminal, and its columns correspond to variables that describe the hierarchy. The rows of $L$ also correspond to those variables, its columns to characters. A suitable coding of the classification can be obtained by associating a variable with each split of a bifurcating tree, all the terminals on one side being assigned a code of 1, those on the other side a $-1$, and those outside the group a 0. One further variable, corresponding to the all-inclusive group is coded as a constant 1 for all terminals.

As an example, the classification $((A, B), (C, D))$ has the $H$ matrix

<table>
<thead>
<tr>
<th>A/B C/D</th>
<th>AB/CD</th>
<th>ABCD</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

With this $H$, the $L$ matrix for Data Set 1, above, is found to be

$$
\begin{array}{cccc}
\frac{1}{2} & -\frac{1}{2} & 0 & 0 \\
0 & 0 & \frac{1}{2} & -\frac{1}{2} \\
\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\end{array}
$$

The predicted character matrix $HL$ is identical to that of Data Set 1, so that the multiple correlation of the characters on the hierarchy is 1.0. Prediction is free of error.

If the same calculations are carried out with A and B of Data Set 1 taken as each comprising five species, the resulting multiple correlation is again 1.0 for the same classification. This method then lacks the peculiar sensitivity to taxon size found in Gower's method.

The accuracy of this procedure raises a new problem, though. It results from the fact that the matrix $(H'H)$ has rank equal to the number of distinct objects in the classification. But with the coding method outlined, this will be true of any clas-
sification whatever, so that any classification will yield perfect predictions for any data. To take just one example, consider the classification \(((A, C), (B, D))\). For Data Set 1 this gives an L matrix

\[
\begin{array}{cccc}
\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\
0 & \frac{1}{2} & 0 & -\frac{1}{2} \\
\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2}
\end{array}
\]

and predicted states HL identical to Data Set 1 again. While the problem of making accurate predictions has thus been solved completely, the procedure now seems to offer no way to choose among hierarchies on the basis of accuracy of prediction.

**SIMPLE STRUCTURE**

One might suspect that the disappearance of accuracy of prediction as a criterion is a peculiarity of multiple regressions, but in fact the same result arises in other approaches. Smith and Kirsch (reviewed by Mitter, 1980) proposed a way of extending the entropy statistic to apply to hierarchies so as to avoid summing redundant "information" at different levels. As Mitter pointed out, their statistic evaluates to 0 information loss for any hierarchy whatever. Alternatively, it might be objected that the degeneracy results from a common pitfall of regressions: that using enough parameters will always allow a perfect fit. That is true in a sense, but the observation offers no escape from the problem. The number of parameters fitted in these examples, for one thing, is no greater than that in Gower's method. They are simply better chosen. No one, I presume, would suggest that one ought to choose parameters poorly simply in order to avoid the problem raised by a better fit. Moreover, any suite of variables adequate to place t separate taxa in a hierarchy (even an unresolved one) would necessarily furnish at least t possible combinations of values. Prediction from those values to the states of the terminal taxa could then always be made fully accurate. While in regressions in general, perfect fit can be avoided, if desired, by reducing the number of predictor variables, this is not possible for hierarchies.

While this problem seems surprising, that is so mostly by contrast with the less accurate prediction procedures discussed earlier. It is actually a characteristic feature of hierarchies. I showed before (Farris, 1979b) that any classification could be provided with diagnoses adequate to recover a character matrix accurately. But that does not mean that there is no way of choosing among classifications on the basis of character data. While any classification will recover the data accurately, not all will do so with equal efficiency or informativeness. Of more immediate interest here is that the same problem is also a familiar one in the context of ordination methods. In any multidimensional ordination, the new axes may be subjected to any nonsingular transformation without altering the fraction of variation in the data explained by the ordination. In the present case, the variables used to describe the classification act as axes of a space, and, as it develops, the axes of any classification provide a sufficient basis to account for all the variation in the data. In ordination in general, then, the issue arises of how to choose among possible axes for the space. Two sorts of considerations are used in approaching this subject: convenience or simplicity of prediction, and scientific interpretability of axes.

Convenience of prediction can be appreciated by inspecting the L matrix for classification \(((A, B), (C, D))\). The first character of Data Set 1 is accurately recovered by the matrix multiplication \(C = HL\), but accomplishing this involves representing that character as \((\frac{1}{2})(A/B) + (\frac{1}{4})(AB/CD) + (\frac{1}{4})(ABCD)\). It would be considerably simpler or more efficient to use a different sort of rule, such as: character 1 takes on state 1 just in A. Doing so would also be more informative in the sense of Sober (1975, 1983): finding the value of the state requires specification of fewer parameters. To find the state of A from HL one must know in effect that A is grouped with B, that A and B together are
separated from C and D, and that A belongs to the all-inclusive group. To retrieve the same information by the suggested rule, one need know only that A is the terminal in question. The two classifications used as examples differ in informativeness in this sense. Clumsy as the L matrix for ((A, B), (C, D)) is, that for ((A, C), (B, D)) is worse still. The former hierarchy recovers the last character as \((\frac{1}{2})(AB/CD) + (\frac{1}{2})(ABCD)\), while the latter must resort to \(-(\frac{1}{2})(A/B) - (\frac{1}{2})(C/D) + (\frac{1}{2})(ABCD)\). A similar observation holds for character 5.

Simplicity of description (or prediction) then favors the hierarchy ((A, B), (C, D)) for Data Set 1, and inspection of those data readily shows that this arrangement is indeed preferable for that purpose. The main groups A+B and C+D correspond to the last two characters of those data, while the alternatives A+C and B+D are unrelated to any of the characters. That is hardly to say, however, that either of these representations of data is as simple as it might be, as we shall see later.

The problem of choosing among alternative representations of data in terms of new axes—factors—has received much attention in the psychometrics literature, where it is generally considered under the subject of factor analysis (for a review see Harman, 1960). There the criterion of choice usually applied is some version of the idea of simple structure. This means, in present terms, selecting the hierarchy \(H\) so that the representation of the characters summarized by \(C = HL\) is simple in the sense that \(L\) has many zero coefficients, or, somewhat more generally, so that, say, the sum of unsigned entries of \(L\) is as small as possible. In the example just considered, classification ((A, B), (C, D)) is preferable by either measure, having two fewer nonzero \(L\) entries and a sum of unsigned entries smaller by unity than does ((A, C), (B, D)). There is a considerable literature devoted to techniques for choosing factors so as to optimize such criteria, but those methods are largely inapplicable to biological systematics, inasmuch as they are not aimed at producing a hierarchic classification of objects ("cases," i.e., taxa). The ideas used by psychometricians to justify preference for simple structure solutions nonetheless offer interesting parallels to systematic principles.

In the preceding discussion the variables used to describe the hierarchy were chosen primarily for convenience in calculating the inverse of \(H'H\), but for purposes of achieving simplicity in the description \(C = HL\) it will be useful to encode the classification in a different way. Take each column of \(H\) to be the characteristic variable of one of the groups of the classification. Such a variable takes on value unity for each terminal in the group, and value zero for every other terminal. On this convention the \(H\) matrix for classification ((A, B), (C, D)) would be the same as Data Set 1, except that a seventh variable, having value 1 for every terminal, would be needed to describe the all-inclusive group. With this convention, it is seen that \(C = HL\) corresponds to a very simple system of equations. Take \(L\) to be a \(6 \times 6\) identity matrix with a seventh all-0 row appended. Each variable in the data is now described as a function of just one of the groups of the classification, evidently the minimum possible. The number of zeros in \(L\) is clearly maximized. The sum of entries in this new \(L\) is the same as that for the old \(L\) for the same grouping. This is to be expected in this case, as the old and new \(H\) are related by a transformation that conserves the sum of unsigned entries. The two sums are not really comparable, of course, as the total number of entries (zero and otherwise) in the new \(L\) is nearly twice as great as in the old; the average value of entries is smaller by the same ratio. The sum as such aside, it is clear that the new coding method allows a simpler description of the data in terms of the classification in a more basic sense. The coefficients of the new \(L\) are readily interpreted in terms of the grouping. Character 5, for example, is retrieved just as \((1)(A+B)\), that is, it is characterized as taking on state 1 just in group A+B. Likewise for the other
characters. As few groups appear in the description of any one character, this scheme is simpler and more informative than the old in Sober's sense, as already mentioned.

This extreme simplicity is possible in this case because the characters themselves are all 0/1 valued, but in most systematic applications, characters would take on a wider variety of values. Then it would be too much to expect that the entries of L would be either 0 or 1, and further, their magnitude would be of interest. I shall thus concentrate on minimizing the sum of unsigned coefficients of L.

In psychometric application, it is usually desired that coefficients (factor loadings) be interpretable in terms of regression coefficients in the usual, least-squares sense, for the reason that psychometric ordinations are generally computed from matrices of correlations between variables. Quite a different meaning can be applied to the entries in L in the present problem. Observe that the equation \( C = HL \) represents the state \( c(i, j) \) of taxon \( i \) for character \( j \) as the inner product of the \( i \)th row of \( H \) with the \( j \)th column of \( L \). As the columns of \( H \) are the characteristic variables of the clusters of the classification, this product will have nonzero entries just for those columns of \( H \) that denote clusters to which taxon \( i \) belongs. As the entries of \( H \) are all 0 or 1, \( c(i, j) \) is represented just as the sum of entries \( 1(k, j) \), \( k \) indexing just the clusters containing taxon \( i \). Consider those clusters as ordered from most- to least-inclusive, so that \( 1(l, j) \), the entry for the all-inclusive group, represents an initial value for the \( j \)th character (the initial values are all conveniently 0 for the earlier example), and each subsequent value represents an increment or decrement in the value of the character. As \( C = HL \) can always be assured, it is known that the sum of the initial value, the increments, and the decrements will be precisely \( c(i, j) \). If the clusters are interpreted (perhaps only formally) as monophyletic groups, they correspond to stem species, and an increment or decrement \( 1(k, j) \) can be interpreted as the change in character \( j \) in evolution between the stem species of the \( k \)th group and the next most inclusive group. The initial value on this interpretation is the plesiomorphic state for the all-inclusive group, that is, the state of the stem species of that group. The observed state \( c(i, j) \) is then represented as the combination of a plesiomorphic condition together with a series of evolutionary modifications culminating in the feature of terminal taxon \( i \).

The basic goal of simple structure approaches is to obtain a factor solution capable of scientifically meaningful interpretation. In many applications no theory is available to guide formulation of a criterion of simplicity. Even when an accepted theory can be so applied, it may be difficult to derive from this an explicit criterion. And even when a criterion can be justified, it may suggest no ready means of effective calculation. Harman (1960:291) comments of this subject that parsimony is “one of the fundamental standards in selecting a preferred solution,” but that this idea does “not always carry an explicit meaning” in factor analysis. In the systematic problem, however, the interpretation just suggested provides an immediate benefit in all three respects. That the phylogenetic parsimony solution is interpretable in terms of real processes is obvious, but a means of computation is supplied as well. The rows of \( L \) correspond to groups, whence to putative stem species. The total of the unsigned entries in the \( k \)th row of \( L \) represents the total change in all characters in evolution to the \( k \)th stem from the stem of the next most inclusive group. The total of \( L \) entries exclusive of those for the all-inclusive group is just the total length of the tree corresponding to the hierarchy described by \( H \). Given \( H \), entries of \( L \) can be found to minimize their unsigned total conditional on the initial values by the Wagner optimization method of Far-
ris' (1970). Alternative classifications, of course, will generally admit of different total amounts of change (total unsigned entries of L), and the classification admitting of the simplest (most informative) description of data—again conditional on the initial values—can be found, at least to a good approximation, by applying the Wagner method (described in its earliest version by Kluge and Farris, 1969; for improved methods see Mickevich, 1978; Schuh and Farris, 1981).

In factor analysis in general, a criterion like the sum of loadings can determine a set of axes only up to an arbitrary rotation of reflection. In the systematic problem, certain transformations of classifications correspond to other hierarchies that differ, say, just in rooting. In the phylogenetic interpretation in particular the problem of choosing a "rotation"—a rooting—is solved by the simple structure criterion itself, that is by the outgroup criterion (cf. Farris, 1982). Among possible alternative trees with the same length, that one is selected that allows outside taxa to be connected to it most parsimoniously. In the discussion before, the L entries for the all-inclusive group were considered just as initial values, as hypotheses of plesiomorphy. When the outgroup criterion is used to root the tree, and so provide inferences of plesiomorphy, the entries for the all-inclusive group are no longer initial values as such, but now become increments and decrements in the context of a more inclusive problem. In this respect, the role of parsimony in determining hierarchies is just as I have explained before (Farris, 1979b:511), and there seems to be no need for further discussion.

Most parsimonious trees have the property that their grouping does not depend on the relative frequencies of congruent characters in data. If in Data Set 1 there were a large number of characters like character 1, then a phenogram for those data would have clusters (A, (B, C, D)), whereas for the data as presented the clustering would be ((A, B), (C, D)). Supposing the hypothesis of plesiomorphy to be fixed, the most parsimonious tree would be the same in both cases. Farris (1979b, 1980b) has shown in detail why this property of phenograms leads to loss of descriptive and explanatory power, but here it is of interest to concentrate on the invariance itself. Mickevich (1978) has pointed out that variation in relative frequencies of autapomorphies of different groups (in this case, state 1 of character 1 distinguishes just taxon A) is responsible for the incongruence of phenograms based on different suites of characters. This instability of phenetic clustering precludes use of phenograms as general purpose arrangements, but further, and more to the point, obviates as well interpretation of phenetic groups in terms of any general process. Interestingly, a similar idea has been used in psychometric factor analysis. Thurstone, the originator of the simple structure concept, considered it (1947:361) "a fundamental requirement of a factorial method." Kaiser (1958) argued for his own simple structure method (varimax) precisely on the grounds that its "solution is invariant un-
der changes in the composition of the test battery” (p. 195). One might dispute, of course, whether psychometricians have succeeded in identifying repeatable psychological factors, but that is an empirical issue of little relevance to systematics. It is heartening, however, that others have recognized a connection in principle between informative description and repeatability as guides to the recognition of natural processes.

**CONSEQUENCES AND CONTROVERSIES**

From a purely phylogenetic standpoint, this factor-analytic treatment seems to offer little that is new. It was already known that parsimonious phylogenetic classifications combine explanatory power with efficiency of description. The value of this derivation lies instead in what it shows of the defects of other approaches to classification. It is now seen that ordinations provide neither an alternative to hierarchic classification nor a legitimate reason for separating description from explanation. Systematists thus no longer need be concerned with those aspects of the pheneticist stance. The multiple regression viewpoint, moreover, brings out the basic fallacy of phenetic classificatory criteria, those that sum fit measures over levels as though either the membership or the characteristics of inclusive and included taxa might be independent.

Another benefit of the same sort is provided by Sneath and Sokal’s suggestion, already quoted, for recovering distance information from ordinations. The present treatment has been concerned primarily with the relationship between classifications and characters. This seems entirely appropriate, as characters are in fact the source of evidence for selecting groupings. Pheneticists, however, have almost universally concentrated instead on the conformity of classifications to matrices of distances or raw similarities, generally (although inaccurately) claiming that their methods are superior for achieving such conformity. On viewing classifications as ordinations, however, it is seen at once that any classification that describes character information accurately will likewise allow any sort of distance computed from character data to be recovered with perfect accuracy from the classification. Fit to matrices of raw similarity thus becomes immaterial as a classificatory criterion, being subsumed under descriptive information content on characters. In fact, this observation offers a great advantage for recovery of distance information. It is seldom the case that a phenogram will fit a distance matrix perfectly, and even if one distance statistic showed such a fit, it is unlikely that some other measure of raw similarity would be so fortunate for the same character data. Recovering distances or similarities from a classification by way of character data not only allows perfect recovery of distances, it allows perfect recovery of every kind of distance measure simultaneously.

It seems then that there never was any real need to resort to syncretistic arrangements for purposes of description. But, of course, that conclusion has been reached before, and it has proved controversial, at least in the sense that pheneticists have offered a great variety of nominal reasons for denying it. It seems suitable then to extend the present discussion by reviewing the faults of the more recent of those arguments.

The writings of opponents of phylogenetic systematics have generally revealed a curious reluctance to address substantial issues. Both pheneticists and eclecticists, to be sure, long insisted that syncretistic arrangements were necessary for effective description of the characteristics of organisms. But as late as Sneath and Sokal (1973), this claim was put forward seemingly just as an article of faith. It was based on no comparisons of the descriptive power of phylogenetic and syncretistic classifications, nor had any measure been proposed that might have been used to make such comparisons. Rather than attempting to produce evidence for the descriptive superiority of their methods, these workers relied in-
stead on a variety of auxiliary objections to phylogenetic systematics. Phylogenetic classification was criticized as impossible, requiring knowledge of the true phylogeny (Gingerich, 1979, and many others), or as being unable to place unknown, extinct organisms (Sokal, 1975). It was dismissed as pointless, inasmuch as it might instead be represented as a tree diagram (McNeill, 1979). Its criteria were objected to as overly stringent, as they would determine the classification (Estabrook, 1981). None of these critics ever bothered to consider, apparently, that phylogeneticists do not claim to use the true phylogeny, that no method can place anything unknown, that any classification might be drawn as a tree (or phenogram), or that any criteria whatever that led to a classification would thereby have determined it. Many other "arguments" of this general type clutter the literature (for further discussion see Farris, 1977, 1979b). They are entirely irrelevant to choice of a reference system, either because they depend on portraying phylogenetic systematics as something other than what phylogeneticists in fact advocate, or because they claim as weaknesses of phylogenetics in particular properties that instead are necessarily common to all methods of classification.

It was phylogeneticists—not pheneticists or eclecticists—who finally began to compare methods of classification with respect to the properties in which the superiority of syncretistic schemes had been claimed to reside. The confrontation of results with preconceptions provoked controversy, of course, but it did have the immediate benefit of obviating reliance on criticisms of the sort just mentioned. At least, none of the cited authors has shown any inclination to attempt to defend his point as pertinent or well-reasoned. The old reluctance to come to grips with concrete issues nonetheless remained, now simply taking on new forms. Much of the discussion written in attempts to reply to the new findings tends more to obscure questions than to resolve them.

**EQUIVOCATIONS**

The simplest way to obscure is by shifting definitions—that is, by equivocation. The issue at hand seems perfectly clear. Pheneticists have always advocated classification by clustering according to raw ("over-all") similarity, and this on grounds of superior descriptive power. It is precisely this opinion that distinguishes pheneticists from phylogeneticists. Phylogeneticists object to the pheneticist position by contending that clustering by raw similarity does not yield descriptive superior schemes. Pheneticists might reply to that charge directly, but instead often proceed by denying that clustering by raw similarity is phenetics. Thus, Colless (1981) maintained that phenetic classifications are not phenograms (diagrams of raw similarity clusterings), but are instead related to the latter by "imperatives that still remain unanalyzed." Rohlf and Sokal (1981:463) go a step further: "If a classification is intended to represent a similarity scheme it is phenetic, if it is intended to show evolutionary branching sequences it is cladistic."2

To the extent that such disclaimers were taken seriously, their effect would be to make impossible any rational criticism of phenetic classification, simply by rendering it likewise impossible to determine what objective properties—as distinguished from intents—a phenetic classification might possess. It seems pertinent, then, to point out that the latter authors, at least, do not themselves adhere to their subjectivist definition: "We would predict that the phenetic methods will be more congruent . . . since they are

---

2 The purpose of such obfuscations, of course, is to provide a simple means for escaping criticism. Past errors of pheneticists can be dismissed on the grounds that they were not aspects of "phenetics." McNeill (1982) has his own distinctive way of making this excuse. He introduces the concept of "true phenetics," which he leaves undefined, although he does provide examples. Amusingly, McNeill lists just two of these exemplars, himself and Jardine and Sibson (1971). This puts him in a position to defend "phenetics" without the bother of addressing the defects in the views of other (one presumes "false") pheneticists, such as Colless or Rohlf and Sokal.
sampled from the entire phenome rather than a subset representing shared derived similarity (synapomorphy)” (Rohlf and Sokal, 1981:466). And, “data sets with little homoplasy could yield quite different phenetic and cladistic classifications because of inherent differences in the structure of the approaches” (Rohlf and Sokal, 1981:469). But plainly if the distinction between phenetics and cladistics were only a matter of intent, then it would be nonsense to write of structural differences in the approaches, as would Sneath and Sokal’s claim, quoted earlier of superiority for phenetic classification. If the distinction between phenetics and cladistics were something other than that the former groups by raw similarity, whereas the latter groups by putative synapomorphy, then the quoted prediction would be nonsense as well.

Rohlf and Sokal (1981:463) also write that, “we would interpret a synapomorphy scheme . . . as a special kind of phenetic relationship.” This makes the motivation for their inconsistency of usage clear. Workers known as “pheneticists” aim to defend “phenetics” as superior by twisting words so that whatever turns out to be superior can be called “phenetics.” But those same workers still hope that what is in fact (and what in fact they regard as) phenetics—clustering by raw similarity—will still prove superior, and other manipulations of meaning plays roles in their arguments to this point as well. Mickevich (1978, 1980) showed that phylogenetic classification is more stable than are phenetic groupings when different sources of evidence, such as morphological and biochemical characters, are considered. Rohlf and Sokal (1980) countered that she had not investigated “stability.” They argued just by listing four types of experiment that they were willing to term assessments of stability, carefully leaving out the type that she had done. The sincerity of their position on this point can perhaps best be appreciated by comparing it with that of Sneath and Sokal (1973:31), who list instability (incongruence) in Mickevich’s sense as one of the four major problems of phenetic classification. Farris (1979b) found that phylogenetic methods for representing distance information are superior to phenetic methods for that purpose. Rohlf and Sokal (1981), just as before, present a list (p. 470) of what they are willing to allow as “cladistic” methods, then suggest (p. 479) that Farris’ method is not cladistic because it is not in their list. The method in question is like ones proposed by Cavalli-Sforza and Edwards (1967) and by Fitch and Margoliash (1967) for phylogenetic inference. Sneath and Sokal discuss both under “The Study of Phylogeny.” Farris, at any rate, most certainly intended that his trees be interpretable as “evolutionary branching sequences.” Those schemes are thus cladistic according to Rohlf and Sokal’s own usage of intent. Perhaps Rohlf and Sokal intended that only their own intentions would count in determining what is or is not cladistic. That would seem to correspond to their position in arrogating to themselves the authority to decree that the word “stability” may apply to invariance of classification under any sort of perturbation whatever, with the sole exception of the effect that Mickevich studied.

RETREATS

That pheneticists find it necessary to resort to terminological trickery in order to defend their position surely raises the suspicion that that position has grave defects that must be concealed, but of course this by itself says little of what those flaws might be. There is nonetheless a readily recognized correspondence between the equivocations just mentioned and the faults of phenetics at a more substantive level. Just as the aim of Rohlf and Sokal’s pretended usage of “phenetics” is to make it impossible to criticize phenetics, much other discussion is aimed at abandoning any pheneticist claim that might prove subject to refutation. Just as Rohlf and Sokal’s lists were devised to rule out unfavorable findings by definition, so other arguments aim to show that any unfavorable results must be irrelevant. But this
tactic of defense leads ultimately to a contradiction. Phenetics was originally advocated on the grounds that it possessed certain concrete advantages. To the extent that every test of those advantages is rationalized away as irrelevant, those claims of advantage become empirically empty. To the extent that those advantages are no longer claimed, there is no reason for adopting a phenetic approach. Pheneticists are driven at last to maintaining that classification ought to be phenetic for reasons that can be neither analyzed nor corroborated.

That is a curious position indeed for workers who began by insisting on empiricism, objectivity, and repeatability in systematics (cf. Sokal and Sneath, 1963), and that point will serve as well as a first example of abandoning positions. Contrast that stance with Colless' classification according to "imperatives that still remain unanalysed." That of course amounts just to simple refusal to admit what phenetic classification is, but other arguments that are seemingly more definitive have much the same effect. Rohlf and Sokal (1980) complained of Mickevich's (1978) comparison of phenetic and phylogenetic methods that she had used the Manhattan distance for phenetic clustering. Pheneticists, they wrote, would not do such a thing. Colless (1980) decried as "eccentric" Mickevich and Johnson's (1976) phenetic analysis of one of the same data sets that Mickevich (1978) employed. He insisted that the Manhattan distance should be used. Mickevich (1980) repeated her earlier comparisons with a different distance statistic, finding that phenetic clusterings still have inferior stability. Rohlf and Sokal (1981) now complain that she should have used the correlation coefficient instead, as would have "most pheneticists." Jardine and Sibson (1971:33)—one of McNeill's examples of "true phenetics"—comment of this application of the correlation coefficient: "This is absurd." This is far from exhausting the list of such examples; several others of the same type are given by Mickevich (1980), Schuh and Farris (1981), and Mickevich and Farris (1981). With the exception of the unusually forthright Jardine and Sibson, none of these pheneticists has ever been known to criticize others' choice of method. It rather seems that "poor" phenetic techniques can be discovered to be such only when used in comparative studies by phylogeneticists. While these complaints seem to differ from Colless' obscure imperatives, their effect is just the same. Phenetics cannot be criticized if no one can discover exactly what "phenetics" is.

Sneath and Sokal (1973) stressed the desirability of (descriptive) naturalness in classification, characterizing this quality in terms of effectiveness of representing character information. Farris (1979b) analyzed the concept of naturalness and found that phylogenetic classifications possess this property to a greater degree than do phenograms. Rohlf and Sokal (1981) leave naturalness out of their list(!) of desirable properties of classifications, and while they do discuss something called "naturalness," this now pertains only to describing raw similarities, not characters. Sneath and Sokal also held that natural classifications are also predictive. Farris (1979b) showed that both prediction and informativeness could be assessed through the same quantitative measure. Schuh and Polhemus (1980) illustrated that concept in a concrete case, attributing both properties to a phylogenetic classification. Rohlf and Sokal now make no mention of Farris' derivation, but they do object to Schuh and Polhemus' observation by denying that there are any known grounds for relating these criteria. They do not discuss, of course, the implications of their position for Sneath and Sokal's claim.

In listing types of stability, Rohlf and Sokal (1980) had among their aims to object that Mickevich (1978) had not investigated all of those, implying that each type would provide useful information on evaluating methods. One of these types was stability under change in clustering method. Mickevich (1980) pointed out that this can have no possible bearing on
choice of method. Rohlf and Sokal (1981) still list stability types, but do not mention this one. Another of the original types was stability on addition of terminal taxa. Schuh and Polhemus (1980) and Schuh and Farris (1981) found higher stability of this type for phylogenetic than for phenetic methods. Rohlf and Sokal (1981) make no mention of those results (although they cite both papers), but they do suggest (p. 467) that “complete stability upon the addition of [terminal taxa] is an inherently unattainable (and undesirable) goal.” Still another of the original types was stability of classification under change in character coding. Again, both Schuh and Polhemus (1980) and Schuh and Farris (1981) found phylogenetic classification to be more stable in this respect. Again, Rohlf and Sokal (1981) offer no comment on these results, but this type of stability is no longer listed. Instead, one now finds stability of similarity matrices under changes in coding listed as a possible criterion—and Rohlf and Sokal emphasize that similarity matrices are not classifications. As to the application of this criterion, “one might wish to compare phenetic and phylogenetic results, “if a comparable scale can be found.” The original criteria of Rohlf and Sokal seem generally to have become irrelevant as soon as the possibility arose that phylogeneticists might investigate their consequences. Evidently, they had no real purpose other than to provide an excuse for complaining about Mickevich’s results.

The last of the original types of stability concerned changing the set of characters used. Mickevich (1978) did investigate the effect of using different character sets. Rohlf and Sokal (1980:98) contended that her results “were thus not suitable tests [of this] aspect of taxonomic stability,” but were instead tests of the hypothesis of nonspecificity of Sokal and Sneath (1963). On the surface, this seems just another play on words, but there is a deeper and more important issue. Sneath and Sokal (1973), as already noted, had reckoned differences between classifications based on different kinds of characters among the “major problems” of phenetic classification. The reason that it is such a problem, as Mickevich (1978, 1980), Farris (1979b), Mickevich and Farris (1981), and even Ehrlich and Ehrlich (1967), have all pointed out, is that if different sorts of characters require different arrangements, then no one arrangement can be suitable for all kinds of characters. The pheneticist aim of a general purpose descriptive classification would then be an impossibility.

While objecting to Mickevich’s results on a variety of pretexts, Rohlf and Sokal (1980) agreed to her conclusion that phenetic methods would produce incongruent classifications from different kinds of characters. Rohlf and Sokal (1981:470), indeed, insist that “reliable phenetic methods should [emphasis original] yield different classifications based on these different character classes.” But what, then, of the “major problem”? Rohlf and Sokal (1980) imply that a satisfactory general purpose classification can still be obtained by basing it on a suite of characters obtained by stratified sampling of the population of all kinds of characters. Mickevich (1978) had already pointed out that this is infeasible, and Mickevich (1980) observed that stratified sampling is impossible in consequence of Rohlf and Sokal’s own admission that characters cannot be randomly selected. Rohlf and Sokal (1981) make no effort to rebut any of these points, and they now make no mention of stratified sampling, either. But neither do they offer any explanation of how Sneath and Sokal’s “major problem” of phenetic classification is to be solved.

INVENTIONS

In each of the cases just discussed, criteria and ideas that had been used by pheneticists in attempts to bolster their position are retracted—or simply left unmentioned—as soon as those concepts appear instead in criticisms of phenetics. The aim of those discussions is clear enough: to dismiss any consideration that might lead to discovery of a weakness of
phenetics. Those retreats from earlier positions are tacit, but other arguments attempt much the same on ostensibly more explicit grounds.

Some "explicit grounds" comprise nothing other than misrepresenting pertinent information. Sokal and Rohlf (1981) offered a wide variety of complaints on the calculations of Schuh and Polhemus (1980), but just two of those points survived in Rohlf and Sokal (1981). These are that in comparing stability of phenetic and phylogenetic methods in random character selection experiments, Schuh and Polhemus used just one random partition of characters, and that the character coding of that study was unsuitable for phenetic analysis. Schuh and Farris (1981) had replied to Sokal and Rohlf by repeating Schuh and Polhemus' investigation with the same coding recommended by Sokal and Rohlf for that application and using 100 random partitions, as Sokal and Rohlf had also suggested. Rohlf and Sokal make no mention of this, but they do claim that their "major criticisms are not addressed" in the paper by Schuh and Farris. In fact, Schuh and Farris not only addressed those points, but proceeded exactly as Sokal and Rohlf had demanded in both respects. Their new results showed higher stability for phylogenetic classification under random character selection, refuting the prediction of Rohlf and Sokal on this subject, already quoted. Seemingly unable in this case to concoct some further excuse for denying the relevance of stability comparisons, Rohlf and Sokal instead pretend that Schuh and Farris' new calculations were never performed.

A different sort of deception underlies another aspect of Rohlf and Sokal's position on the significance of congruence studies for phylogenetic methods. They maintain that incongruence of phylogenetic classifications derived from different types of characters reveals a weakness of phylogenetic systematics (although, of course, such incongruence would not be a "problem" for phenetics!), and in particular that Mickevich's (1978) results demonstrate that defect. Their basis for that claim is that she reported consensus information indices for the Wagner method considerably less than "the perfect consensus of 1.0" (Rohlf and Sokal, 1980:100). But as Mickevich (1980) pointed out, two fully congruent—identical—trees can have a value much less than 1.0 for that index. Rohlf and Sokal are presumably aware of this, as Rohlf (1981) has noted as much himself. Rohlf and Sokal (1981:466) nonetheless persist: "Mickevich (1978) found in practice that cladograms constructed from different sets of characters differed appreciably." Their criticism of cladistics is based in this instance on what can only be viewed as an intentional misrepresentation of the properties of Mickevich's consensus information statistic. Along the same lines, Janowitz (1980) claimed to have found examples in which clustering by special similarity yielded lower cophenetic correlation than did clustering by raw similarity, so refuting Farris' (1977, 1979b) findings. Farris (1980a) discovered that Janowitz had used some other, undescribed, measure as the "cophenetic correlation" for special similarity.

Fairness

But those are extreme cases, and most of the arguments advanced by pheneticians can be more charitably characterized as reasoning that does not bear close inspection. The claim, already mentioned, that Mickevich's congruence study was a test of the hypothesis of nonspecificity rather than a "proper" test of stability has been made repeatedly (Rohlf and Sokal, 1980, 1981; Sokal and Rohlf, 1981). Equally often (Mickevich, 1978, 1980; Schuh and Farris, 1981; and even Farris, 1971) it has been pointed out that the hypothesis of nonspecificity—"that there are no large classes of genes affecting exclusively one class of characters" (Sokal and Sneath, 1963:85)—cannot logically be related to the differences in congruence properties of phenetic and phylogenetic methods. The truth of a genetic mechanism can hardly depend on what
clustering method is used. Rohlf and Sokal (1981) continue to maintain their old position on this point, but make no effort to justify a connection between the hypothesis and the presumed test. Their argument, to the extent that they have one, seems to consist of no more than endless repetition.

Of course Rohlf and Sokal are not seriously interested in viewing Mickevich's results as a test of the nonspecificity hypothesis. They make no effort to interpret those findings in terms of the truth of that theory—which in any case would be impossible. Rather their aim in making that claim is to provide support for another contention, that stability tests ought to be performed instead with randomly selected sets of characters. That this position amounts to abandoning the earlier stance of Sneath and Sokal has already been seen. The implied exclusiveness of types of tests is in any event illogical, as one might easily perform tests in both ways. Further, as Rohlf and Sokal (1980, 1981) have already admitted, characters are not—and cannot be—randomly selected in systematic practice. As Farris (1971) and Mickevich (1980) have pointed out, this means that random selection tests have little bearing on the practical utility of methods. Rohlf and Sokal make no attempt to rebut that observation. But why, then, their insistence on random selection? The answer is all too readily apparent. Rohlf and Sokal had already conceded that phenetic methods would prove unstable when classifications from different types of characters are compared, but they believed (as already quoted) that clustering by raw similarity would make a better showing in the type of test they propose.

That belief was ill-founded, as Schuh and Farris (1980) have demonstrated, but Rohlf and Sokal's argument is of a type that is of some further interest. Their stand is simply that tests that produce results unfavorable to phenetics must ipso facto not be the right tests. Any such test, to put it bluntly, must be unfair. Most of what has been put forward as explicit objections to comparisons of phenetic and phylogenetic classification—as distinguished from equivocation, evasions, and retreats—corresponds to this general pattern.

McGinley and Michener (1980:92) are perhaps most forthright in this sort of argument: "Replication [of groups, i.e., stability] is likely to result from cladistic...relationships. Therefore Nelson's (1979) methodological dice are loaded in the direction of recognizing clades." But many others have done much the same. Farris (1979a) found that clustering by special similarity allowed more accurate retrieval of phenetic distance information than did clustering by raw similarity. Colless (1981) noticed that Farris' method was particularly good at recovering distances between the reference point (outgroup) and other terminal taxa. He suggested, accordingly, that fit of observed to retrieved distance should be computed with those particular parts of the data left out. Rohlf and Sokal (1981) make the same suggestion, but offer a further argument, to which I shall return below. Sokal and Rohlf (1981) objected to use of an outgroup to root cladograms in Schuh and Polhemus' (1980) stability comparisons. They had noticed that this provided cladograms with a source of stability not available to phenograms, and insisted that to make stability measures "comparable," cladograms should be rooted in another way. Rohlf and Sokal (1981) also object to Mickevich's (1980) method of rooting. Observing that it produces higher stability for cladograms, they conclude from this alone that its use introduces "a serious new bias." Bottjer (1980; refuted in detail by Farris, 1980b) criticized Farris' (1979b) measure of information content of classifications on the grounds that it would tautologically favor phylogenetic methods. McNeill (1982) took a similar position. Rohlf and Sokal (1981:477) offer much the same criticism of Farris' (1979a, b) methods for retrieving distance information from phylogenetic classifications: "the effectiveness of the special similarity transformation is a mathematical prop-
erty of the transformation and is not based on any 'naturalness' of the data in an evolutionary or phylogenetic sense." They again contend that phenetic and phylogenetic results are not "comparable," and offer new comparisons of phenetic with "cladistic" methods entirely of their own creation. These yield, of course, results inferior to those of phenograms.

As if pheneticists had ever before used "naturalness" to refer to anything other than purely descriptive efficacy—or had ever shown any concern for evolutionary considerations! And indeed, any of these arguments go back to the by now familiar theme of saving "phenetics" just by discarding all of its old premises. Sneath and Sokal (1973:57) cite with approval Colless (1969) to the effect that any maximally predictive scheme is by definition—tautologically—phenetic. For a further example: Sokal and Rohlf (1962) advocated UPGMA clustering on the grounds that it gives higher cophenetic correlation than other phenetic methods. Farris (1969) showed that this is an algebraic consequence of the method, but this has not led Rohlf and Sokal to abandon their favorite. Tautologous (or mathematical) results, it seems, are entirely legitimate, so long as they seem to favor grouping by raw similarity. Pheneticists had likewise never been known before to claim that evaluations of classifications ought properly to leave out part of the data. Indeed, if that might be done, why not just use the traditional taxonomists' "good characters" instead of bothering with overall similarity? Nor can Sneath and Sokal's claim, already quoted, be sensibly interpreted to mean anything other than that phenetics is superior to cladistics, not, "phenetics is superior so long as the outgroup is not counted." Having long argued (tautologically, or so they believed) that clustering by raw similarity is superior because it uses more information—symplesiomorphy as well as synapomorphy—pheneticists now complain that the use of outgroups (whose purpose, of course, is to distinguish apomorphy from plesiomorphy) provides a kind of information that phenograms cannot utilize. Nor could Sneath and Sokal's contention be taken as intending, "phenetics is superior to cladistics, provided that by 'cladistics' is meant something other than what cladists advocate." "Superior," furthermore, is a comparative, and the claim implies a comparison. If, as they now attempt, pheneticists contend that their techniques are not comparable to legitimate cladistic methods, then they tacitly admit that their previous claims on behalf of their approach were meaningless.

Application of any sort of criterion to selection of methods might involve a tautology, and any criterion that favors one method over another might be accused of bias. But such charges by themselves do not provide a rational argument. If they did, then an accurate algorithm for simple addition might be dismissed as tautologically so in favor of some defective procedure, and advocates of the theory that elephants weigh less than mice could dismiss balances as biased. No comparison can be legitimately discarded simply on the observation that it provides a certain result. Conformity of a method to a criterion cannot legitimately be criticized just as a mathematical property. Instead the conformity of methods to criteria and the results of comparisons can only be evaluated by consideration of the criteria themselves. Criteria and comparisons can honestly be dismissed only when it can be demonstrated that they possess undesirable properties on grounds extrinsic to their behavior in the disputed application.

COMPARABILITY

Pheneticists have made no effort to defend (as distinguished from assert) the proposition that stability is undesirable in a classificatory system. Rohlf and Sokal (1981) do note that stability achieved by abandoning conformity to data is of no benefit. Farris (1980a) and Mickevich and Farris (1981) had pointed out the same, and drew the conclusion that the inferior descriptive power of phenetic methods
implies that what stability those techniques might have is of no value. All this serves just to concentrate attention on the descriptive efficacy of methods. Rohlf and Sokal’s discussion, indeed, is aimed mostly at providing a justification for their stance that phenetic and phylogenetic methods of description cannot be compared.

Rohlf and Sokal offer several different objections to Farris’ (1977, 1979a, b, 1980a, b) method of clustering by special similarity. They first follow Janowitz (1979; see also Farris, 1979a) in contending that the high cophenetic correlation for special similarity might have been achieved simply by making that measure of resemblance independent of data. But Farris (1979a, b) had already demonstrated that raw similarities can be recovered from special similarities, in consequence of which the latter convey just as much information on data as do the former. After only a page of text, Rohlf and Sokal take note of this and proceed to more novel discussion. Farris (1979a) had recovered raw similarities from special similarity clusterings using just the first taxon of each data set as the reference point. Farris (1980a) had noted that choice of reference point could affect accuracy of recovery of similarities and had recommended trying each of the taxa as a reference point, retaining the scheme with highest cophenetic correlation, and rerooting this, if necessary. Rohlf and Sokal carefully avoid mention of that second paper. They object to the first on the grounds that one must use each of the taxa as a reference point, then evaluate the procedure by summarizing those results. Their argument to this extent comprises denying Farris’ conclusions on the properties of his method simply by substituting some other “cladistic” method. It is informative to contrast their stance on this regard with their repeated insistence, already cited, that comparisons of phenetics and cladistics must properly utilize the particular phenetic methods that Rohlf and Sokal advocate.

Rohlf and Sokal also follow Colless (1981) in deciding that cophenetic correlations must be evaluated with distances from the reference point to other taxa omitted from consideration. They give more of a supporting argument than did Colless (less would be impossible), but what they write does not hold up well to scrutiny. They first note that the retrieved special similarities between the reference point and other taxa are all zero, concluding from this (1981:474) that, “it would seem reasonable to remove the reference [taxon] from the comparison,” since the zero values, “do not depend on the observed data.” Of course those values were not intended to depend on the data, but on the contrary, serve to provide a mechanism for retrieving data effectively. This part of their argument is much like objecting to the sample mean as a description of a sample on the grounds that the sum of signed deviations of observations from that mean is necessarily zero. In fact, of course, the sample mean is prized for exactly that property, among others. But those zero values are rather beside the point, for they pertain only to retrieved special similarities, and of course Rohlf and Sokal are much more interested in denying that phylogenetic methods recover information on raw similarities. For this purpose, those authors also wish to exclude from consideration the retrieved raw similarities between the reference point and other taxa. These do depend on the data, and further, as Rohlf and Sokal (1981:476) themselves observe, those distances “will be perfectly encoded and then recovered by the special similarity transformation and its inverse . . . . Thus, again, it does not seem appropriate to include [these] values . . . .” Their reason for excluding these data is nothing other than that the phylogenetic method fits them perfectly! Having resorted to such a premise, it is surprising that Rohlf and Sokal did not save themselves effort by simply dismissing any data set that gave better fit with phylogenetic than with phenetic methods. Or, to be more precise, that they did not do so in so many words.
It seems unnecessary to comment on the character of Rohlf and Sokal's argument as it is revealed by these quotations, but it is worth noting that this rationalization illustrates another weakness of their position. Farris had used the cophenetic correlation as a measure of conformity of a classification to a distance matrix just because that statistic had been advocated for that purpose by Rohlf and Sokal themselves. Their original use for that measure had been as a way of selecting among possible phenetic arrangements, that is, possible placements of taxa. In Farris' method, the distance information for the reference point is perfectly retrieved, as the placement of the reference point on the tree is chosen to have that property. But in classifications in general that taxon need not be so placed, and if it were not, then its distances might be quite inaccurately retrieved. The role of the cophenetic correlation as a means of choosing among classification includes selecting a placement of the reference point. If the cophenetic correlation were calculated with the reference point excluded, then any placement of that taxon would be allowed by that criterion. The original cophenetic correlation was a potentially meaningful basis for comparing methods precisely because it was a criterion applicable to choosing arrangements. In attempting to evade the consequences of their own criterion, Rohlf and Sokal would replace it with another measure that is not useful for that original purpose. It did not occur to them, apparently, that if a criterion cannot be used to select arrangements, it cannot very well be used to choose among methods of classification, either.

“SIGNIFICANCE”

Rohlf and Sokal have other criticisms of special similarity clustering, but these are related to their comments on Farris' (1979b) second method for retrieving phenetic distance information from cladograms, and so I shall not discuss them separately. As Farris (1979b) noted, this second method, additive branch-length fitting, renders criticisms of special similarity clustering obsolete. The second phylogenetic method is a generalization and an improvement of the first, and provides still more accurate recovery of distance information. Disputing the efficacy of special similarity clustering would thus do pheneticists little good, unless they could also justify dismissing additive fitting.

As Farris (1979b) pointed out, both special similarity clustering and additive fitting are able to retrieve distance information more accurately than can phenograms because the first two methods can fit a length separately to each branch of a tree, while the last can fit only clustering levels. To Rohlf and Sokal this suggests an analogy with stepwise regression analysis, and they contend that a significance test is needed to determine whether the additional “parameters” of the phylogenetic methods improve fit more than would be expected from sampling error alone. This of course requires a distribution, a problem that they solve simply by postulating that the entries of a distance matrix are independent and multivariate normally distributed. Working from that assumption, they employ an F-ratio to judge the significance in the difference in fit between level-fitting and branch-fitting.

This seems like straightforward cookbook statistics, and indeed that is its fault. The procedure is simply carried over from more mundane applications, and nothing about it is suitable for the problem at hand. Elements of a distance matrix can hardly be independent of one another. Rohlf and Sokal admit this for distances calculated for character data, but claim that the postulate may nonetheless be reasonable for such quantities as immunological distance. Even those distances, however, usually (although not always, see Farris, 1981; Farris, Kluge, and Mickevich, 1979, 1982) obey the triangle inequality, and that is enough to show non-independence (Hartigan, 1975; interestingly, Rohlf and Sokal cite this work). In fact, Rohlf and Sokal have made
the simple mistake of misidentifying the premise needed for their test. The distances as such need not be independent; it is only required that their departures from a tree model (“errors”) be so. But even that less restrictive assumption can scarcely hold. If an instance of homoplasy is so placed in the tree that it produces convergence between the members of two separate clusters, the pairwise distances between terminals of those two groups will obviously depart from a tree model dependently.

Problems of distribution aside, the stepwise regression analogy is itself rather strained. In usual applications, the significance test is performed just to determine whether there is any benefit to be gained by adding another fitted parameter, with the idea that in the interests of simplicity, the number of parameters should be kept as few as is consistent with data. Rohlf and Sokal use the same test to assess improvement of branch-fitting over level-fitting on fully-resolved trees. As the tree is in either case fully resolved, there is no real difference in simplicity between the two arrangements compared. The advantage of simplicity in ordinary regressions is that few parameters need be taken into account in framing descriptions or in making extrapolations (predictions). In Farris’ method, description of distance information by a classification is achieved by placing a value in the diagnosis of each taxon. In the phenetic method as well, each taxon has an associated value, its clustering level, with the difference that all the terminal taxa are required to have the same value. As each taxon will have a diagnosis in any event, there is no practical advantage to be gained by restricting the entries. In usual regression analysis, the decision made is whether to include a further parameter, but if one is included, it will be just another term in the equation, of the same algebraic sort as the others. Branch-lengths, however, are not the same kind of quantity as clustering levels. They are related to retrieved distance information by a different equation. Further, they are not quite comparable to clustering levels in terms of number, since in general several branch-lengths are required to convey the same information as a single clustering level. Finally, in a properly performed stepwise regression, the denominator of the F-ratio is an estimate of sampling error. In the ratio used by Rohlf and Sokal, the denominator is taken as the departure from the tree model. As it is quite probable (homoplasy, again) that the data depart from perfect fit to a tree model by more than sampling error alone, this practice has the effect of rendering the calculated F-ratio smaller than the proper value. That, of course, fits Rohlf and Sokal’s purposes nicely, as it decreases the chance that a phylogenetic scheme will have to be judged significantly better than a phenogram.

Finding an excuse to deny cases of better fit for the phylogenetic method is in fact the principal goal of Rohlf and Sokal’s discussion. Farris (1979a) had used a significance test as well. His was a non-parametric paired comparison, performed simply by tallying the number of better fits for each method (0 for phenograms, 9 for special similarity clustering) then evaluating the binomial probability (low, to say the least) of so extreme a departure from the hypothesis (Sneath and Sokal’s) that phenetics is superior whenever the two approaches give different results. Rohlf and Sokal offer no objection to this test as such; their denial of Farris’ conclusion comprises mostly doing away with the data themselves. Their discussion of testing, however, rests on the tacit presumption that only differences “significant” by their standard can be properly regarded as differences in the paired comparisons. That is a rather curious premise. In usual applications of such tests, one does not try to judge whether individual observations differ only by sampling error in counting the differences in either direction. On the hypothesis, for example, that the difference between two paired quantities is as likely to be positive as negative, observing that all of 100 differences come out
in the same direction is very suspicious indeed, and such observation is normally taken as good evidence against the hypothesis. If in such a case the hypothesis were then defended on the grounds that the differences were rather small—which is in fact irrelevant—one would be justified in regarding the defense as at best ad hoc. Differences need not be large to be real, and the aim of the paired comparison is to judge whether the differences, whatever their magnitude on some chosen scale, conform to a consistent pattern. If one were to begin by declaring that only large differences are to be counted, the effect might very well be to prevent the pattern from being discovered. That, of course, would conform to Rohlf and Sokal's aims very well indeed.

Rohlf and Sokal might, one presumes, insist on some grounds that only differences “significant” by their test provide evidence, although what those grounds might be is not apparent from their paper. It is, however, informative in this connection to consider another part of their argument. On finding that branch-length fitting is superior in most cases even by their own test, they attempt (as I shall discuss below) to dismiss this result by declaring that the phylogenetic procedure is not a method of classification at all. They then proceed to evaluate a “cladistic” method entirely of their own devising and having the convenient property of generally yielding lower cophenetic correlation than phenograms. In their table 4 they summarize paired comparisons of their “cladistic” method with phenetic clustering, noting in the last column of that table which method is “best.” “Best” is established just by comparing cophenetic correlations, and is determined entirely by the direction of the difference; there is no provision for values being not “significantly” different. In fact, on the same assumptions that Rohlf and Sokal had made for their own tests, none of those differences is “significant,” but Rohlf and Sokal make no mention of this, and present their conclusions on “best” anyway. Their attitude, evidently, is that differences must be “significant” in order to support phylogenetic classification, whereas no such requirement applies to differences that can be represented as favoring phenetics.

Those comments apply to Rohlf and Sokal’s tests of both phylogenetic methods, but in evaluating special similarity clustering in particular, those authors manage to make a further, particularly basic, mistake. All of the above aside, the F-ratio can only be applied if the parameters are fitted to minimize the residual sum of squares. Special similarity clustering, however, does not produce branch lengths for a least-squares fit to a distance matrix. Farris (1980a) had pointed this out but again, Rohlf and Sokal avoid mention of that paper.

“CLASSIFICATION”

In evaluating special similarity clustering with their F-ratio, Rohlf and Sokal leave out the distances from the reference point and consider results for each reference point separately, rather than seeking an optimal choice. Their results are statistically meaningless for the reasons just given, and scientifically meaningless as well—unless one supposes that it is legitimate to leave out just those parts of the data in which the phylogenetic method gives optimal fit. Their analysis is in any event meaningless as an evaluation of phylogenetic methods, except on the premise that Rohlf and Sokal may legitimately redefine “cladistic” to mean anything they please, then attribute the faults of their “cladistics” to what phylogenetic systematics is in fact. On the premise, that is, that equivocation is a valid mode of argument.

For branch-length fitting, Rohlf and Sokal seem unable to identify any part of the data that can be rationalized away.

Readers who find it surprising that the authors of a widely-used biometry text could have made so many elementary statistical mistakes might profitably recall at this point that until Mickevich (1980) pointed out their blunder, Rohlf and Sokal were quite prepared to use “stratified” sampling without being able to make random selections.
Neither does their statistical reasoning raise much of an issue in practice, since even by their own tests the phylogenetic method gives higher cophenetic correlation in every case, and only three of these differences fail to meet their criterion of "significance." To avoid the obvious implication of these results, Rohlf and Sokal fall back on their last argument: that the products of the phylogenetic method are not classifications at all, and so cannot be compared to phenetic clusterings.

The grounds that they advance for this idea are spread over several parts of their discussion. They first introduce a set of conventions for devising "flow charts" of taxonomic methods. Here they formalize a number of routine ideas, as that analysis proceeds from observation to coding, to calculation (of similarities, say), ... to classification. One might well wonder at the need to present such material, but Rohlf and Sokal's purpose in doing so is clear from the way in which end results show up in their charts. Ordinations, dendrograms, and (undirected) "trees" are awarded distinct symbols in the diagrams, and so are portrayed as separate—and mutually exclusive—sorts of results, with the implication that the different constructs are not "comparable." Now of course it is true that some ordinations are not trees, and that some (undirected) trees are not dendrograms (i.e., directed tree diagrams), but this does not imply exclusivity. Rohlf and Sokal's terminology obscures the relationships between these quantities: that every dendrogram is also a "tree" in their sense, and that either sort of tree is also, among other things, a coordinate system, and so a kind of ordination. Their aim in all this is just to argue that "trees cannot be directly transformed into classifications" (Rohlf and Sokal, 1981:462). That there is no coherency to their view is shown by their subsequent admission (p. 462), that there are procedures "for directly converting an unrooted tree into a classification."

Rohlf and Sokal's Appendix A is also devoted to recounting elementary material. Here they characterize ultrametrics (ultrametric distances are the retrieved distances of phenograms), four-point metrics (branch-length fitted trees with positive branch-lengths yield four-point metric retrieved distances), and the relationship between classification and ultrametrics. Here, too, their argument proceeds primarily by taking note only of those connections that they find it convenient to admit. They do not mention that every ultrametric distance is also a four-point metric distance, which is the reason why properly performed branch-length fitting can never yield a worse fit to observed distances than does phenetic clustering (Farris, 1979b). They discuss a non-invertible transformation for obtaining ultrametric distances from four-point distances, and conclude from this that there is no unique correspondence between classifications (which they define to be ultrametrics) and four-point distances. But this is irrelevant, since as they later admit, the transformation used by Farris (1979b) is invertible. It is irrelevant for other reasons as well, and these reasons involve the crux of Rohlf and Sokal's argument. In their main text they admit that the retrieved distances of Farris' method fit observed distances more closely than do those of phenograms, but contend (p. 476) that the former "are not ultrametrics and therefore are not equivalent to a classification." They conclude from this (p. 479) that "It is not appropriate to compare [a measure of fit of similarities to a classification] with [a measure of fit of similarities to trees] when one wishes to determine 'a classification with maximal content of distance information' [emphasis ours] (Farris, 1979b). The unrooted Wagner tree is simply not a classification." Farris (1979b) had in fact used rooted Wagner trees, but this simple falsehood is immaterial, as Rohlf and Sokal aim to apply the same criticism to distances retrieved from directed phylogenetic trees as well.

Rohlf and Sokal's claim rests on several subsidiary premises, the first of which is that a classification must be an ultrametric distance. In their Appendix A, they
argue to this point by noting that any hierarchy (well-nested suite of groups) implies a distance that satisfies the ultrametric conditions: take the derived distance between a pair of terminals to be the categorical rank of the smallest inclusive group containing both. That is true so far as it goes, but the claimed equivalence of classifications to ultrametric distances does not follow. One could just as well say that the same hierarchy defines an ultrametric similarity: the similarity between two terminals is the number of groups to which both belong.

Obvious as this may seem, ignoring that idea is nonetheless crucial to Rohlf and Sokal’s argument, but other points need to be discussed before this can be made clear. The sort of rank distances (or similarities) just described are not in fact what either pheneticists or cladists use to retrieve distance information from a classification. For that purpose, pheneticists associate a numerical clustering level with each group, then take the retrieved distance between two terminals to be the clustering level of the smallest group including the pair. It is at once seen from this that the claimed equivalence of classifications with ultrametric retrieved distances rests merely on equivocation, tacitly redefining “classification” to mean “phenogram.” Phenograms—and only phenograms—associate clusterings with derived distances in this way. It is hardly true of classifications in general that their levels of nesting are intended to represent levels of phenetic difference. Those nesting levels might well be intended to correspond with antiquity of groups, for example. And a value associated with a group need not represent a degree of difference. It might just as well correspond to a degree of similarity. Even if a similarity, furthermore, it need not be raw similarity; it might just as well correspond (special) similarity.

In Farris’ (1979b) method, the levels associated with clusters of a branch-length fitted tree are precisely levels of derived similarity: the level is the path-length distance between the stem of the cluster and the root of the tree. Part of Rohlf and Sokal’s position is that this method involves just unrooted trees, but it is seen from this description that such is not the case. The derived similarity values depend on the location of the root, and if the root were changed, then so would be the levels. Rohlf and Sokal do not dispute this; they simply do not mention it. So far as these levels are concerned, then, the claim that Farris’ constructs do not correspond to classifications (i.e., rooted trees) is only misrepresentation.

The levels in Farris’ method provide hierarchies with a way of retrieving information on derived similarities, and if that were all that they provided, it is unlikely that Rohlf and Sokal would have bothered to contend that a hierarchy ceases to be a classification upon incorporating such levels. But the same levels offer a way of retrieving information on raw similarities—or phenetic distances—too, and the finding that this recovery is more accurate than that of phenograms is of course the principal motive for contention. Farris’ method retrieves phenetic distance information from derived similarity levels by the same equation that is used for that purpose in special similarity clustering, the inverse of the special similarity transformation. While Rohlf and Sokal simply ignore this part of Farris’ procedure in connection with branch-length fitting, they do dispute distance retrieval by special similarity clustering on grounds whose faults are pertinent here as well. As I have already noted, Farris’ method places a similarity level in the diagnosis of each taxon. Pheneticists also associate a level with each cluster, although as to the location of those levels in a practical classification they are rather vague. The levels cannot be, as Rohlf and Sokal seem to imply, the ranks of taxa. Ranks need not correspond to distance information at all, and, even if they did, could convey only rough information, as there are so few categories in use. Be that as it may, Rohlf and Sokal are perfectly willing to use levels associated with taxa in their own way of recovering distance
information, but comment of the similar use by special similarity clustering that it involves the “additional” information contained in the levels of the terminal taxa. Their underlying reasoning, to the extent that there is any, seems to be that terminal taxa do not have diagnoses. They offer no explicit defense of that premise, however, perhaps hesitant to invite ridicule. At the best this is another case of the familiar story. Having long claimed that their methods utilize more information (overall similarities) than phylogenetic procedure, pheneticists now complain when a phylogenetic procedure utilizes information that phenograms cannot.

The same plaintive tone is to be found as well in Rohlf and Sokal’s remaining objection, already quoted: that retrieved distances from a phylogenetic tree do not correspond to a classification because they are not ultrametric, thus that fits of the phenetic and phylogenetic methods cannot be compared, “when one wishes to determine ‘a classification [emphasis original] with maximal content of information.’” It has already been seen that this reasoning depends on the false presumption that a classification must be an ultrametric distance. Phenograms are “demonstrated” to be the best means of classification simply by stating the desired conclusion as a premise—“classification” means phenogram. But it is useful to pursue the implications of that peculiar usage. If classification meant ultrametric distance, then Rohlf and Sokal’s last claim would amount to: in order to say that a quantity retrieved from a classification is included in the information content of the classification, that quantity must be a classification. By that reasoning, “I learn from this book that France is in Europe” would have to be assigned the implication that the location of France is a bit of printed paper.

Returning to rationality, if a phylogenetic scheme allows retrieval of distances, then those distances are among the information contained in that classification, whether those distances are themselves a “classification” in any sense or not. If a phylogenetic scheme retrieves distances more accurately than does a phenogram, then it is more informative concerning those distances than is the phenetic arrangement. The core of Rohlf and Sokal’s argument is that phenograms cannot be compared to methods that store information in some way other than does phenetic clustering. On that supposition, to be sure, it would be quite impossible to show that a non-phenetic method is superior to phenetics in information content. But the reason for that impossibility has nothing to do with the merits of phenetics; it is comparison itself that is ruled out, by definition. But “superior,” again, is a comparative, and so the stance that Rohlf and Sokal have adopted amounts to no more than the tacit confession that their claim of superiority for phenetic clustering never had any substantive basis.

**CHARACTER INFORMATION**

Of course, as I have already observed, methods for storing and retrieving distance information as such are of little real importance, as in most applications distances are computed from character data, and so distance information can be recovered from a classification by calculation from retrieved character information. That approach has the advantage that it avoids influence of arbitrary choice of a distance measure on construction of the classification itself. It has already been seen that when this approach is taken, accuracy of retrieval of distance information necessarily becomes irrelevant to choice of classification, as character states and so distances will always be perfectly recovered, provided only that the classification is equipped with adequate diagnoses. Only the efficiency, or simplicity, of representation of character information then remains pertinent. Farris (1979b) had already demonstrated that phylogenetic classifications provide more efficient, or informative, summaries of character information—are more “natural” in the sense of Mill (1874) and Gilmour (1961)—than clusterings by raw similarity. Rohlf
and Sokal write nothing of Sneath and Sokal's (1973) strong emphasis on the importance of Gilmour's ideas, mentioning the latter author only in connection with Farris, and using "naturalness" (futilely) to refer only to retrieval of distance information. Where Sneath and Sokal (1973: 188) had characterized "economy of reliance on memory" as "a paramount taxonomic goal," Rohlf and Sokal write nothing of the economy—that is, parsimony, i.e., efficiency—of representation of character information. This is not entirely to their detriment, to be sure. At least they make no effort to rearrange definitions so that parsimonious arrangements are not classifications, or to fabricate reasons why efficiency or informativeness is undesirable in classification.

Nonetheless in the process of avoiding Farris' result, Rohlf and Sokal manage to make some mistakes worthy of note. Farris (1979b, 1980b) had observed that a taxon has descriptive utility to the degree that it has efficiently diagnostic character states, that is states distinguishing it from its relatives. Of this Rohlf and Sokal (1981: 568) write, "none of the statements defining a Gilmour natural taxon restrict the occurrence of identical states outside a taxon. Pheneticists have never claimed that each taxon is distinguished by a unique character state or even by predominantly unique character states." They go on to accuse Farris (1977, 1979b) of misinterpreting Sneath (1961) to the contrary. Interestingly, in quoting Sneath, they omit the part of the passage (Sneath, 1961:122) that mentions character states. But even on the supposition that Rohlf and Sokal's revision of Sneath were accurate, that would not be enough to establish the claim that "pheneticists have never" associated distinguishing states with natural taxa. Sneath and Sokal (1973), and Sokal (1977:5), both already quoted, had declared, respectively, that a single presence/absence character would permit only two classes, and that natural taxa predict that characters will be homogeneous within and heterogeneous between them. Indeed, Rohlf and Sokal (1981:468) themselves observe, "For character predictivity to be high, characters must be relatively homogeneous within taxa but differ among taxa." At the least, then, Rohlf and Sokal would need to explain how the two classes of the first case could fail to have a distinguishing state, and how states that are homogeneous within and heterogeneous between taxa can fail to distinguish those taxa. Interestingly, too, Archie (1980), whom Rohlf and Sokal cite with approval, had also taken note of the passage from Sokal (1977), and had placed on it the same interpretation as that suggested here (which is also that of Farris, 1979b). Rohlf and Sokal make no mention of any of this, although they discuss Archie's methods, which were based—if not very logically—on this interpretation.

But even if Rohlf and Sokal's version of what pheneticists have "never" done were truthful, this would comprise neither a defense of phenetics nor a criticism of phylogenetic systematics. It would mean only that pheneticists had been pursuing their "objective" approach without ever bothering to specify the relationship between characters and classifications.

This smacks strongly of the familiar story; a criterion that had been used to support phenetics is abandoned, to be replaced by nothing substantial, and earlier discussion of it denied, as soon as it is discovered to be a weakness of phenetics instead. The rest of Rohlf and Sokal's discussion lends support to this interpretation. "General utility," they (1981:467) write, "albeit hard to define, lies at the base of the phenetic concept of natural classification. Pheneticists argue that classification based on high within-group similarity in as many features as possible ... will be widely useful ... since similar [terminal taxa] and taxa are placed together." And, to be sure, if one groups by similarity, similar taxa will indeed be placed together, but this says nothing of how those groups might be "useful." If utility is now "hard to define," pheneticists did not always view it so. The new
stance conveys no information beyond that Rohlf and Sokal are no longer willing to take any concrete position.

This leaves only prediction, and here Rohlf and Sokal again decline to advance any substantive argument. They make no mention of Farris’ (1979b) discussion of this subject, nor do they offer any concept of their own. They defer to Archie’s (1980) views, observing that his assortment of “predictivity” measures gives complex results “that do not support any one school.” They provide no defense of—or any comment on—the faults of Archie’s approach, already noted above. But this can no longer be surprising, for Rohlf and Sokal had shown no concern for either consistency or rational justification in any of their other discussion, adopting instead the strategy of obscuring whatever ideas or criteria might prove a danger to phe

CONCLUSION

Rohlf and Sokal list just five main types of “desirable properties of classification.” These are: fit to a similarity matrix, stability, general utility, fit to known cladistic relationship, and “phylogenetic” optimality criteria. I do not imagine that they intend to compare phenetic and phylogenetic classification with respect to the last two.

For the first, they offer a variety of illogical objections to my earlier demonstrations of better fit for phylogenetic methods, deleting data as they find it convenient, and resorting at last to the premise that classifications are phenograms by definition. But this hardly matters, for they finally conclude (1981:483) that phenetic and phylogenetic procedures cannot be compared in this respect.

For the second, they admit that phenograms yield incongruent arrangements for different sorts of characters, but never attempt to explain how those different schemes can be combined into a general-purpose classification. Beyond this, they list several subsidiary types, most of which come to some such fate as being “an inherently unattainable (and undesirable) goal.” And in their conclusions we find (pp. 482–483), “congruence in phenograms and cladograms need not be measured in the same way.”

The third “lies at the base” of phenetic classification, but it is “hard to define,” and in their conclusions (p. 481) we read, “The goals of phenetic taxonomy have been stated in loose terms which need to be refined.” But they decline to divulge what the nature of the refinement might be, and in discussing this subject—or failing to do so—they contrive not to admit how characters might be related to a phenetic classification. They make no attempt to rebut my demonstration that phylogenetic classification offers a more informative, natural, economical, and efficient representation of characters than do phenetic arrangements. Indeed, in contrast to earlier advocacy of phenetics, they make no substantive claims whatever of informativeness, naturalness, economy, or efficiency for their approach.

In the beginning of phenetics, its methods were championed as objective and empirical, its classifications as repeatable, natural (“general-purpose”), and superior. At the end, there is no stated objective, and so nothing meaningful to repeat. There is no admitted relationship between characters—evidence—and classifications, and so no possibility of either empiricism or informativeness. Repeatability, so far as it concerns congruence, is no longer sought, nor, in consequence, can naturalness be. Superiority is reduced to the curious kind that exists only in the absence of comparison, for the very possibility of comparison—or any meaningful evaluation—is disavowed.

The phylogenetic system truly provides all that phenetics once claimed, and far more beyond this. It is a unified system that explains the similarities among organisms while describing them effec-
tively at the same time. In attempting to evade the consequences of the information content of the phylogenetic system, pheneticists have proceeded only by abandoning or rendering meaningless every claim of advantage for their methods that they had ever advanced. They have themselves shown that there never was any logical basis for preference for syncretistic systems. I can think of no better epitaph for a view that was never better epitaph for a view that was never. In attempting to evade the consequences of the information content of the phylogenetic system, pheneticists have proceeded only by abandoning or rendering meaningless every claim of advantage for their methods that they had ever advanced. They have themselves shown that there never was any logical basis for preference for syncretistic systems. I can think of no better epitaph for a view that was never better epitaph for a view that was never.

REFERENCES


Manuscript received July 1982
Revised October 1982