# A Successive Approximations Approach to Character Weighting

James S. Farris

*Systematic Zoology*, Vol. 18, No. 4. (Dec., 1969), pp. 374-385.

Stable URL:

*Systematic Zoology* is currently published by Society of Systematic Biologists.

# A SUCCESSIVE APPROXIMATIONS APPROACH TO CHARACTER WEIGHTING

JAMES S. FARRIS

## Abstract

Farris, J. S. (Dept. Biol. Sci., State Univ., Stony Brook, New York 11790) 1969. A successive approximations approach to character weighting. Syst. Zool., 18:374–385.— Characters that are reliable for cladistic inference are those that are consistent with the true phyletic relationships, that is, those that have little homoplasy. A set of cladistically reliable characters are correlated with each other in a particular non-linear fashion here referred to as hierarchic correlation. Cladistically unreliable characters can be hierarchically correlated only by chance. A technique that infers cladistic relationships by successively weighting characters according to apparent cladistic reliability is suggested, and computer simulation tests of the technique are described. Results indicate that the successive weighting procedure can be highly successful, even when cladistically reliable characters are heavily outnumbered by unreliable ones. [Evolutionary taxonomy. Cladistics. Character weighting.]

Quantitative weighting of characters is of great concern in evolutionary taxonomy. This is so because convergence, parallelism, and evolutionary reversals are possible. Thus, not all characters are equally well correlated with phyletic history. In seeking to infer that phyletic history, we may be able to increase the efficiency of our methods by weighting characters differentially according to their degree of correlation with cladistic relationships.

Classical evolutionary taxonomists have made use of a number of "biological" schemes for character weighting. These rely on such qualities as inferred degree of "functional" or "adaptive importance," degree of "adaptiveness," or "fundamentality of adaptation." Such criteria have been widely criticized (see for example Sokal and Sneath, 1963; Kluge and Farris, 1969) on the grounds that they are subjective and incapable of quantification. A more important property of "biological" criteria may be that they are peripheral to the subject of cladistic inference. Even if we could find some way to measure, for example, the "functional importance" of a character, it would still remain to be demonstrated that the measure of functional importance was correlated with the utility of the character for purposes of cladistic inference. This is not to say that such a demonstration is impossible; some published comments on character weighting are a step in this direction (Farris, 1966; Kluge and Farris, 1969). Nonetheless, the validity of a given weighting criterion should not simply be assumed, and it seems desirable to investigate the connection between weighting methods and the usefulness of characters for cladistic inference.

In this paper I shall present a weighting method developed directly from the concept of *cladistic reliability*: the degree of fit between a character and the phylogeny.

I shall assume throughout that the coding of each character is fixed. References to "poor characters" indicate that the coding of the character implies homoplasy. This need not mean that the homoplasy is "real" in any other sense. Any case of apparent homoplasy can be removed by an appropriate recoding of the character. In that sense, there is no such thing as a "convergent character" or a "character reversal." In this treatment of weighting, therefore, I shall be discussing ways to find a measure of the agreement between the phylogeny and a character coding that has been fixed *a priori* without knowledge of the phylogeny.

A CONSISTENCY MEASURE

We shall need to define a measure of the degree of correlation between a character and a phylogeny. Existing measures of correlation are not appropriate, since we wish to assign a value to the agreement between a character—typically a linear variable—and a branching pattern. We do not wish to restrict the measure of agreement to be large only on "linear," "monotone," or even "single valued" relationships between the character and the pattern—nor do these terms have any clear meaning in this context. We proceed from a set of definitions. A character is in complete agreement with a phyletic branching pattern if that pattern indicates no homoplasy (convergences, parallelisms, or reversals) in the character. The character disagrees with a phyletic pattern to the degree to which that pattern indicates homoplasy in the character.

To measure the amount of homoplasy of a character on a tree, we compute the *patristic unit character length* of the character on the tree, following the definition of *patristic difference* given by Farris (1967). This is done as follows. For the tree pattern in Fig. 1 there are seven nodes, A, B, C, D, E, F, G, with character states 0, 1, 0, 2, 1, 0, 1, respectively. The patristic unit character length of the character on the tree is the amount of change implied by the tree for the character. For the tree and character in Fig. 1 the patristic unit character length is thus calculated by adding the difference in the character between A and B, A and C, B and D, B and E, C and F, and C and G: $1 + 0 + 1 + 0 + 0 + 1 = 3$. For a discretely coded character, as in this example, the patristic unit character length is the "number of steps" of the character for the tree in the sense of Camin and Sokal (1965).

The *range* of a character is the difference between its numerically least and greatest states. For the character of Fig. 1, the range is $2 - 0 = 2$. The range of a unit character is the smallest value that the patristic unit
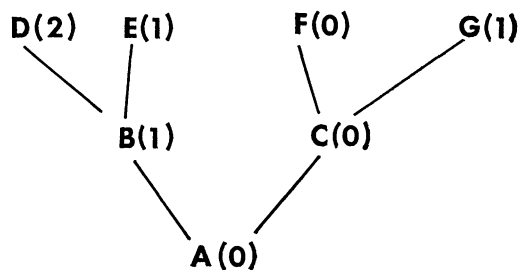


FIG. 1.—A hypothetical tree to illustrate computation of patristic differences.

character length for that character can attain on any tree. I define $c(i)$, the *unit character consistency* of character $i$, as

$$c(i) = \frac{r(i)}{l(i)}, \qquad (1)$$

where $r(i)$ is the range of the character, and $l(i)$ is the patristic unit character length. Necessarily, $c(i)$ varies between 0 and 1 and will increase with the degree of agreement between the character and the tree on which $l(i)$ was calculated. In our example, the $c$ value is ⅔. Unit character consistency is defined analogously to the *consistency* of a data set with a tree (Kluge and Farris, 1969). The prefix "unit character" is intended to distinguish the two uses of "consistency." Note that there is no direct connection between unit character consistencies and consistencies, since the former do not depend on the scales of the unit characters, whereas the latter do. Further, the overall consistency of a set of characters with a tree is not, for instance, in general equal to the mean value of the unit character consistencies of the data set with the tree.

HIERARCHIC CORRELATION

We define as *cladistically reliable* characters that have high unit character consistency with the true cladistic relationships of the group under study. Several char-

TABLE 1.  HYPOTHETICAL DATA SET.
Each OTU has state "1" in the indicated characters
and state "0" otherwise.

| OTU | "1" states in characters |
|-----|--------------------------|
| A | 1,3,7,15 |
| B | 1,3,7,16 |
| C | 1,3,8,17 |
| D | 1,3,8,18 |
| E | 1,4,9,19 |
| F | 1,4,9,20 |
| G | 1,4,10,21 |
| H | 1,4,10,22 |
| I | 2,5,11,23 |
| J | 2,5,11,24 |
| K | 2,5,12,25 |
| L | 2,5,12,26 |
| M | 2,6,13,27 |
| N | 2,6,13,28 |
| O | 2,6,14,29 |
| P | 2,6,14,30 |



FIG. 2.—A tree that fits perfectly all the characters of Table 1.

acters that are all consistent with the cladistic relationships will all fit a common branching pattern, and they are thus in a sense "correlated" with each other. This is not the usual product-moment correlation. To emphasize the distinctness of this "correlation," I shall refer to it as *hierarchic correlation.*

The *defining property* of hierarchic correlation is that a set of variables with high hierarchic correlation will all be highly consistent with a single branching pattern. Characters that are hierarchically correlated may *or may not* be correlated in any other apparent way. To exemplify this point, I have included a hypothetical data set specified by Table 1. All the characters in the data set are completely consistent with the tree in Fig. 2. The characters are all binary, so that the apparent correlations between them can be measured by the *phi* coefficient (the contingency coefficient of Siegel, 1956). The *phi* coefficients for the characters of the hypothetical data set range from 1.0—the maximum possible—for character 1 with character 2, to 0.067—the minimum possible is 0—for character 29 with character 30. The median value of the matrix of the *phi* coefficients is only 0.1.
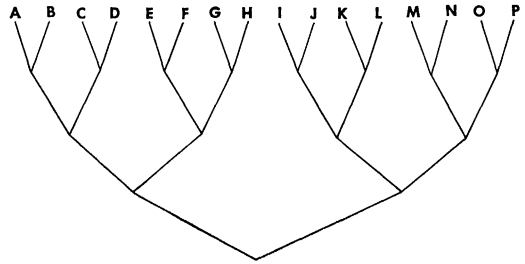
Thus, although the hierarchic correlation of the characters is perfect, the average apparent correlation between characters is not very strong. There is considerable variation in the degree of apparent correlation of pairs of characters, even though all the characters fit the tree perfectly.

Since there is no strong relationship between the apparent correlations of characters and their consistency with cladistic relationships, we should not attempt to estimate the cladistic reliabilities of characters through the apparent correlations of their states.

SUCCESSIVE WEIGHTING

To estimate cladistic reliability, we rely on properties of cladistic systems. By definition cladistically reliable characters are hierarchically correlated with each other. Cladistically unreliable characters are *not* hierarchically correlated with each other. This is so because the phylogeny itself is the only source of a hierarchic pattern for the variation of a character. Characters whose *a priori* codings are poorly related to the phylogeny necessarily vary from the phylogeny each in its own random way with respect to the phylogeny itself. Several unreliable characters will each vary from the phylogeny in its own random way, and chances are very slight that a series of random variables will by accident form a pseudo-hierarchic pattern of variation. If several unreliable characters are developmentally, genetically, or function-

ally related, they may of course vary together much more strongly than independent variables. It is still only by pure chance, however, that a pattern of functional covariation may happen to fit a hierarchic pattern. In general, therefore, we expect hierarchic correlations many times more frequently among cladistically reliable characters than among cladistically unreliable ones.

If we construct an estimated phyletic tree so that it is most consistent with an initial data set, we may use it as a first—quite possibly, rough—guess at the true cladistic relationships.

The estimated tree can provide a tentative set of unit character consistencies. The unit character consistencies so computed can be taken as estimates of the consistencies of the unit characters with the true phylogeny. The consistencies provide measures of cladistic reliability, and the reliabilities can be used to weight characters. The reweighted character set can then be used to construct a new estimated tree, which will usually differ from the first estimate. The process might be repeated indefinitely. Provided that the methods of tree construction and weighting are fixed, however, the successive trees will all be the same beyond any iteration at which a tree is the same as on the immediately previous iteration. Thus the iterative procedure can be halted as soon as two successive trees have the same form.

The first set of estimated consistencies incorporates information on the hierarchic pattern of the data and so the first reweighted data set contains more easily interpretable information than the initial data set. The second tree is expected to be somewhat closer to the true phylogeny than the initial tree, and the reweighting based on it to be more reliable than the first reweighting. Similarly, we suppose that if methods of weight computation and tree construction are suitably efficient, successive tree estimates and consistency estimates will become progressively better until the process terminates at the correct phyletic tree.

SIMULATION TESTS

Computer programs for estimating trees by successive weighting procedures have already been used to analyze relationships among grebes (DeBenedictis and Farris, in prep), killifish (Farris, in prep), and pygopodid lizards (Kluge, in prep). In each case, the technique has been highly "successful," in the sense that it produced trees in excellent agreement with the preconceptions of the workers. Such "successes" do not, of course, prove anything at all about the validity of successive weighting. I have attempted to gain a better estimate of that validity through a series of computerized simulation tests.

For the tests, a hypothetical phylogenetic tree with 31 nodes and 30 completely consistent characters was specified. Poorly consistent characters in various numbers were assigned to the nodes by a random number generator. For a single run, the data set consisted of 31 OTUs with $30 + k$ characters, where $k$ is the number of randomized characters. A supervisor program submitted a complete data set to a successive weighting program, compared the output of the successive weighting program to the true form of the phylogeny to evaluate the successive weighter, called a random number generator to produce a new complete data set, and repeated the process. In runs so far performed, $k$ varied from 5 to 150.

The successive weighting program formed Prim Networks (see Edwards and Cavalli-Sforza, 1964) of the 31 OTUs. The Prim Network is a shortest network connecting a fixed set of nodes. For these experiments, the length of an internode lying between nodes A and B was computed as

$$\text{dif}(A, B) = \sum_i |X(A, i) - X(B, i)| \, w(i), \quad (2)$$

where $X(A,i)$ is the state of character $i$ for node A and $w(i)$ is the weight assigned to character $i$. Initially all the $w(i)$ were set
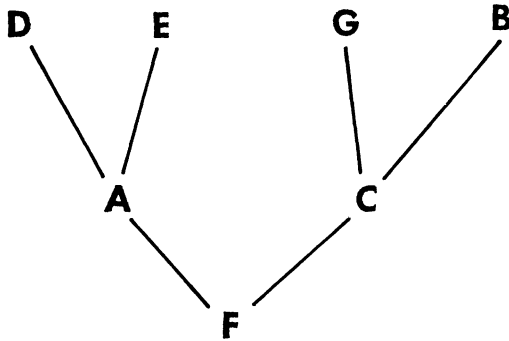
FIG. 3.—A hypothetical tree used to illustrate computation of cladistic difference.
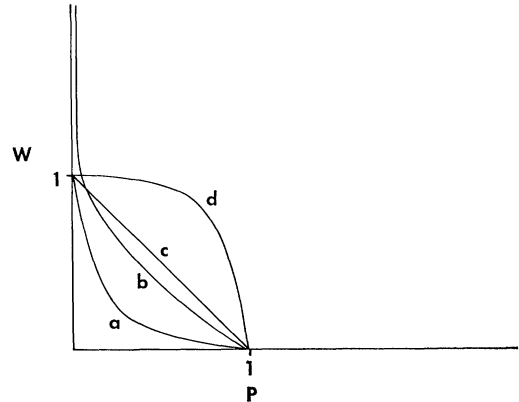


FIG. 4.—Four types of function relating weight ($W$) to probability of change ($P$): a, concave bounded; b, concave unbounded; c, linear; d, convex.

to unity. The true tree for the hypothetical data set could be expressed exactly as a Prim Network of the 31 nodes with the 30 reliable characters. Hence the successive weighting program could—but would not necessarily—retrieve the true tree exactly.

Tree estimates were compared to the true tree through the corresponding matrices of cladistic differences between OTUs. The matrices of cladistic differences were constructed using a modification of the definition of Farris (1967). The *cladistic difference* between two nodes on a tree diagram was taken to be the number of internodes lying on the path between the nodes. In Fig. 1, D and E are each directly connected to B; hence each has a cladistic difference of 1 from B according to the network of Fig. 1. Similarly, the cladistic difference in Fig. 1 between F and E is 4.

The index of difference between two cladistic difference matrices—hence two networks—was the mean squared difference between corresponding elements of the matrices, the constant zero values of the main diagonal being ignored. The equation,

$$d(Q, P) = 2 \sum_{i, j < i} \frac{[q(i, j) - p(i, j)]^2}{t(t-1)}, (3)$$

expresses this relationship. $Q$ and $P$ are the two matrices of cladistic difference being compared, $i$ and $j$ index rows and columns of $Q$ and $P$, $q(.,.)$ and $p(.,.)$ are indexed

elements of $Q$ and $P$, and $t$ is the number of OTUs. The value of $d(Q, P)$ is zero if $Q$ and $P$ are identical.

For the network with seven OTUs in Fig. 1, the matrix of cladistic differences is:

|   | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 0 | 1 | 1 | 2 | 2 | 2 | 2 |
| B | 1 | 0 | 2 | 1 | 1 | 3 | 3 |
| C | 1 | 2 | 0 | 3 | 3 | 1 | 1 |
| D | 2 | 1 | 3 | 0 | 2 | 4 | 4 |
| E | 2 | 1 | 3 | 2 | 0 | 4 | 4 |
| F | 2 | 3 | 1 | 4 | 4 | 0 | 2 |
| G | 2 | 3 | 1 | 4 | 4 | 2 | 0. |

Fig. 3 can be taken as a different network for the same OTUs, and produces the cladistic difference matrix,

|   | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| A | 0 | 3 | 2 | 1 | 1 | 1 | 3 |
| B | 3 | 0 | 1 | 4 | 4 | 2 | 2 |
| C | 2 | 1 | 0 | 3 | 3 | 1 | 1 |
| D | 1 | 4 | 3 | 0 | 2 | 2 | 4 |
| E | 1 | 4 | 3 | 2 | 0 | 2 | 4 |
| F | 1 | 2 | 1 | 2 | 2 | 0 | 2 |
| G | 3 | 2 | 1 | 4 | 4 | 2 | 0. |

The comparison $d$ (Fig. 1, Fig. 3) has the value 1.8095. No optimality properties are claimed for $d$ as a measure for comparing networks. It is intended simply as a descriptive index of amount of difference.

Both the original characters of the hypothetical OTUs and the characters sup-

TABLE 2.   SUCCESS OF SUCCESSIVE WEIGHTING FOR VARIOUS WEIGHT FUNCTIONS.  EACH ENTRY IS THE MEAN
OF 10 SIMULATION RUNS.

| Number of random characters | Average difference before iteration | Average difference after iteration (average number of iterations) | | | |
|---|---|---|---|---|---|
| | | $w(i) = 1 - p(i)$ | $w(i) = 1 - (p(i))^2$ | $w(i) = (1 - p(i))^2$ | $w(i) = (p(i))^{-3} - 1$ |
| 5 | 2.6 | 1.9(2.5) | 1.9(2.5) | 0.4(4.5) | 0.0(2.0) |
| 10 | 4.4 | 4.0(2.2) | 4.2(2.3) | 1.3(3.9) | 0.0(2.3) |
| 15 | 5.8 | 5.0(2.3) | 5.5(2.4) | 2.5(4.0) | 0.0(2.6) |
| 20 | 6.0 | 4.9(2.5) | 5.6(2.7) | 3.4(4.2) | 0.0(2.9) |
| 25 | 7.1 | 5.5(2.5) | 7.5(2.5) | 3.2(4.0) | 0.0(3.1) |
| 30 | 6.0 | | 5.8(2.0) | 4.9(3.6) | 0.0(3.0) |
| 35 | 9.1 | | 8.8(2.2) | | 0.0(3.5) |
| 40 | 6.7 | | 8.3(2.3) | | 0.0(3.0) |
| 45 | 9.0 | | | | 0.0(3.2) |
| 50 | 8.5 | | | | 0.0(3.5) |
| 55 | 10.2 | | | | 0.0(3.5) |
| 60 | 9.8 | | | | 0.0(3.1) |
| 90 | 13.0 | | | | 0.0(3.3) |
| 120 | 12.0 | | | | 0.0(2.8) |
| 150 | 10.2 | | | | 0.0(3.6) |

plied by the random number generator were binary valued. Consequently, the unit character consistency of any character could be computed from the number of times the character changed on a given network. A character is said to change in a particular internode of the tree if the two end-nodes of that internode have different states for the character. If character $i$ changes $l(i)$ times, the proportion

$$p(i) = \frac{l(i)}{t-1}, \qquad (4)$$

where $t$ is the number of nodes on the tree, varies between 0 and 1 for binary characters and is related to the consistency:

$$p(i) = ((t-1)c(i))^{-1}, \qquad (5)$$

since the range, $r(i)$, of a binary character is 1. High values of $p(i)$ are associated with low reliability of characters.

The successive weighting was performed by assigning weights as functions of the $p(i)$ values. Several different functions $w(i) = f(p(i))$ were investigated. They fall into four main categories (Fig. 4): *linear*, $w(i) = 1 - p(i)$; *convex*, $w(i) = 1 - (p(i))^k$, $k > 1$; *concave and bounded*, $w(i) = (1 - p(i))^k$, $k > 1$; and *concave and unbounded*, $w(i) = (p(i))^{-k} - 1$, $k \geqslant 1$.

RESULTS AND INTERPRETATIONS

The results of the simulation tests are summarized in Table 2. The first column of the Table gives the number of randomized characters, the second column gives the difference between the true tree and the unweighted Prim Network, and the third through sixth columns give, for various weighting functions, the difference between the true and estimated trees and the number of iterations required to achieve a stable solution by successive weighting. All tabulated values are averages of ten runs.

The application of the successive weighting algorithm almost always improves the estimate of the true tree. The efficiency of the estimate, however, varies sharply between weighting functions. Among bounded functions, the concave is most effective, the convex is least effective, and the linear is intermediate. By far the most effective is the unbounded concave weight function, which achieves a perfect estimate of the true tree even when unreliable characters outnumber reliable characters five to one.

The differences before iteration may become quite large (the largest in the study was 21.2), so that the first estimate of the phyletic tree is in general quite inaccurate.

Nonetheless, the best weighting function invariably produced an accurate answer. This fact suggests that initial approximations do not need to be very good in order for the successive weighting procedure to be effective, and hence that the validity of the successive weighting process is not strongly tied to the assumption that an unweighted, estimated tree is a reasonable first guess at the true tree.

Since the concave weight functions performed better than the convex one, it appears it is more effective to weight strongly for very reliable characters than it is to weight strongly against very unreliable ones. The difference between the two possibilities actually lies in the treatment of mediocre characters. With a concave weight function, only very reliable characters receive high weight, and intermediately poor characters are devaluated almost as much as are the characters that are definitely unreliable. With a convex weight function, only very unreliable characters are heavily weighted against, and mediocre characters receive almost as much weight as very reliable characters. The results of the simulation runs seem to indicate that it is most expedient to treat suspicious characters as if they were unreliable.

The most successful weighting function weighted very strongly. The ratio between the weight for a character that changed once and the weight for a character that changed 29 times is about $2.5 \times 10^5$. For the linear weight function, the weight ratio is 29; for the convex weight function it is 15.2; and for the bounded concave weight function, it is 841. Thus the degree of success of the successive weighting technique is positively related to the strength of weighting. As far as available data show, the stronger the weighting, the more effective the procedure.

INITIAL WEIGHT ESTIMATES

In the simulation tests, all the characters were binary with initial weight unity.

There was no initial influence of weighting coefficients, and the successive weighter was able immediately to detect the hierarchic structure of the data. In real data applications, however, the characters may not all have such conveniently equivalent codings. The possibility exists that an "unweighted" data set may imply weighting coefficients quite uncorrelated with the cladistic reliabilities of the unit characters. Such a case might lead to reduced efficiency of the successive weighter. It is desirable to develop objective means of controlling the weight structure of the data supplied to the successive weighter.

Some weighting methods rely only on the distribution of a unit character in the data. Farris (1966) and Kluge and Farris (1969) point out that the conservatism of a character should be inversely related to $s_W$, the standard deviation of the character within OTUs. Further, the relative tendency of characters to undergo convergence should be reflected by the ratio $s_W{}^2/s_B{}^2$ where $s_B{}^2$ is the variance of a character between OTUs. Initial weights based on these variability criteria were employed in the applied studies mentioned above.

Of greater theoretical interest from the standpoint of this paper are initial weighting criteria that—like the successive weighter—depend on the covariation of the unit characters. Several measures of covariation between characters have been employed in taxonomy. It is illuminating to separate those measures into two categories. Measures of *phenetic correlation* depend only on the apparent covariation of characters. Measures of *hierarchic correlation*, on the other hand, depend on the correspondence between characters and dendritic patterns of some kind. In the first group fall the *phi* coefficient, the product-moment correlation coefficient applied to between-OTU covariation of characters, various forms of factor analysis, and the predictivity of Throckmorton (1968; see also Farris, Kluge, and Eckardt, in press, *a*). In addition to the consistency measure used

by the successive weighter, measures of hierarchic correlation include the *compatibility matrix* of Camin and Sokal (1965) and the *character-pair matrix* of Le Quesne (1969).

Ideally the input to the successive weighter should contain as much information on cladistic reliability as possible. Hence we choose among measures of covariation on the basis of their power to detect hierarchic correlations. The product-moment correlation coefficient and the closely related *phi* coefficient measure directly the between-OTU covariation of character states. The example above suffices to show that such covariation is not well correlated with cladistic reliability. The predictivity of Throckmorton (1968) measures the covariation of similarity coefficients derived from selected character sets. While Throckmorton apparently intended predictivity as a measure of cladistic reliability, that statistic is also only loosely connected with hierarchic correlation. Farris, Kluge, and Eckardt (in press, *a*) showed that the covariation of similarities derived from hierarchically perfectly correlated sets of characters could be positive, negative, or intermediate. Measures of phenetic correlation, then, offer a poor basis for an initial weighting procedure.

The compatibility matrix of Camin and Sokal (1965) is a square matrix, $C$, in which the rows and columns correspond to characters. The entry $C(i, j)$ is the minimum number of extra steps for character $i$ on a tree which fits character $j$ exactly. $C(i, j)$ is clearly related to the hierarchic correlation of $i$ and $j$. If $C(i, j)$ is zero, $i$ and $j$ are mutually consistent. A nonzero $C(i, j)$ indicates some inconsistency between $i$ and $j$. The exact magnitude of a nonzero entry, $C(i, j)$, cannot, however, be readily interpreted. The estimated reliability of character $i$ or $j$ is a function of its lack of consistency with a tree based on all characters. While there is some correlation between entries $C(i, j)$ and the consistency of character $i$ with the tree for all characters, the

relationship is not a simple one, and we may not safely assume that the cladistic reliability of $i$ is computable from, say, $\sum_j C(i, j)$.

The approach of Le Quesne (1969) avoids the perils of attempting interpretation of the exact value of $C(i, j)$. The character-pair matrix, $H$, is a binary-valued array in which $H(i, j)$ is unity if $C(i, j)$ is nonzero, and zero otherwise. Le Quesne's statistic, $N_i = \sum_j H(i, j)$, is the number of characters with which character $i$ is incompatable. Le Quesne described a simple method for computing $H(i, j)$ in the case of binary-valued characters. A contingency table, $T(i, j)$:

Character $i$

|   | 0 | 1 |
|---|---|---|
| 0 |   |   |
| 1 |   |   |

Character $j$    ,

whose entries are the numbers of OTUs in each of the four possible categories, is constructed. The number of nonzero boxes of $T(i, j)$ determines $H(i, j)$. If that number is four, $H(i, j)$ is unity. Otherwise, $H(i, j) = 0$.

The counts, $N_i$, seem to offer a good first approximation to cladistic reliabilities. In a study with $p$ reliable and $q$ unreliable characters, we would expect the unreliable characters to be inconsistent both with each other and with the reliable characters. The $N_j$ for an unreliable character, $j$, should be nearly $p + q$. Such high $N$ values are closely approached by 3 of the 4 apparently unreliable characters of the *Argodrepana* data analyzed by Le Quesne. A reliable character, $k$, will also be inconsistent with most of the unreliable characters, but will be consistent with virtually all of the reliable characters. Hence $N_k$ will be about $q < p + q$.

Le Quesne's method can be readily extended to multistate characters by using *additive binary coding* (Sokal and Sneath, 1963; an example in an evolutionary ap-

plication is given by Farris, Kluge, and Eckardt, in press, $b$) of the characters. Each binary coding variable is then treated as a character. This approach has the advantage that each step of a multistate character may be individually evaluated with respect to cladistic reliability. Use of additive binary coding and separate evaluation of steps is also one way to extend the successive weighting technique to multistate characters. Preliminary results of a study to be published elsewhere suggest that this may be the most effective such extension.

Le Quesne suggested simply removing characters with high $N$ until all the $H$ entries remaining are zero. This would be a statisfactory technique for data such as Le Quesne's for *Argodrepana*, in which there are many characters of perfect cladistic reliability and only a few very unreliable characters. In general, however, we need to be more cautious. It may happen that many of the most reliable characters of a large data set show one or two cases of convergence, so that elimination of non-zero $H$ entries would not be possible. Characters that are in general quite unreliable may nonetheless contain useful information on some restricted region of the evolutionary tree, and a character deletion approach would cause loss of that information. Further, although we may regard $N_i$ as a good indicator of reliability, we do not necessarily expect that it will be completely accurate. We rely on the successive weighter to check our conclusions based on $N_i$, but this can be done only if all the characters are present in the input to the successive weighter. Finally, it is impossible simply to delete poor binary characters when an additive binary coding is used for multistate characters: we cannot allow a single step to be deleted from the middle of a multistate character, nor would we wish to delete an entire multistate character that had only one or two unreliable steps.

Instead of deleting poor characters, we shall assign characters initial weights as

functions of $N_i$. Letting $N_T$ stand for the total number of characters in the study, we can base a weight function conveniently on $N_i/N_T$. The weight, $w(i)$, needs to decrease rapidly with $N_i/N_T$. In studies with many unreliable characters and few reliable ones, the difference between $N_j/N_T$ and $N_k/N_T$ may not be great, even if characters $j$ and $k$ differ substantially in reliability. If the number of unreliable characters is large, a rapidly increasing weight function is needed to provide large differentials in weight between reliable and unreliable characters. By analogy with the simulation test, we might expect an initial weight function such as $w(i) = (N_T/N_i)^3 - 1$ to be reasonably effective. Again, for application to multistate variables, $i$ may index either a binary character or a binary coded step of a multistate character.

We may wonder how likely it is that the successive weighter will extensively modify the conclusions of the initial weighting procedure. There will undoubtedly be data sets for which it is possible to reach valid conclusions on the basis of the initial matrix, $H$, alone. In general, however, we would expect the successive weighter to provide more detailed information on cladistic reliability than is available in the initial estimate. This is so for two reasons. First, the $H$ matrix reflects only pairwise covariation of characters, while the consistencies, $c(i)$ of the successive weighter provide a measure of higher-order interactions between characters $i$ and the rest of the entire set of characters. Second, the $H$ matrix is symmetric, so that it is quite possible for reliable characters to have moderately high $N_i$, since they are inconsistent with several unreliable characters. For data with many unreliable characters, therefore, initial weight estimates based on $N_i/N_T$ may be unable to make a strong distinction between reliable and unreliable characters. The successive weighter, on the other hand, utilizes a non-symmetric relationship by measuring the consistency between character $i$ and the entire tree. As

is amply shown by the simulation runs, the successive weighter can develop large weighting differentials between reliable and unreliable characters, even when the proportion of unreliable characters in the data is quite large.

The several weighting criteria, $s_W$, $s_W{}^2/s_B{}^2$, $N_i/N_T$, and $c(i)$, treated in this paper are not mutually exclusive. They can all be used in a single procedure. Programs to handle these many kinds of weighting information are now being developed.

### DISCUSSION

A number of possible difficulties need to be discussed. Some of my more orthodoxly pheneticist colleagues have suggested that the results are mathematically trivial: that any successive weighting program is bound to appear to work, and hence the simulation tests have no meaning. This objection does not appear to be valid, since the data indicate that the successive weighting technique works well only if a proper weight function is used.

Is the operation of the program on a hypothetical data set a reasonable test of the program's efficiency for analyzing the real data? Within limits, I believe that it is. The hypothetical data sets presented to the successive weighting program were simply matrices of character states. Some of the characters were consistent with a common branching pattern and some were not—the program had no *a priori* information on which were which. Surely a real data set is no more than that, at least in the context of computer analysis. From the standpoint of the operation of the successive weighting algorithm, real and hypothetical data sets seem quite directly comparable.

The unreliable characters in the simulation runs were produced by a random number generator, while unreliable characters in real data are produced causally when distinct phyletic lines respond similarly to similar selective situations. Does this dichotomy between causality and randomness imply that the hypothetical data set

was too unrealistic? Probably not: the feature of unreliable characters that is of interest in this context is their random distribution with respect to the phylogeny. Unreliable characters necessarily have this random property despite the fact that their distribution is caused by selection. It is only the randomness that concerns us, and that randomness, it would seem, can be effectively mimicked by a random number generator.

To what extent would the effectiveness of the successive weighting technique be reduced by correlations of functional or adaptive origin between characters? There are two answers to this. The first is that the effects of functional correlations are still being studied, and programs to successively weight, taking seemingly correlated complexes of characters into account, are being developed. The second—and perhaps more interesting—is that the characters used were correlated. Since there were only 31 OTUs in the hypothetical data sets, the dimensionality of the character space could not exceed 30. Up to 180 total characters were employed; hence there must have been partial correlations between most of the randomized characters. From the standpoint of computer analysis, such partial correlations should have the same effect as partial functional correlations between characters. The most successful weighting algorithm functioned perfectly despite this handicap. We do not need to fear that small amounts of functional correlation in our data will render our procedures ineffectual.

One unpleasant possibility is that some type of data set might cause a successive weighting program to interate indefinitely without achieving a stable solution. All the existing successive weighting programs have safeguards against that possibility, but they have never been needed. In numerous simulation tests and analyses of real data, the number of iterations has never exceeded 20, and the overall average number of iterations is between 3 and 5. Non-ter-

minating analyses do not seem to be a serious practical danger of this technique.

The hypothetical data sets employed all contained a branching pattern, and once that was found, no further interpretation was required. Real data sets, however, may consist entirely of unreliable characters and thus not contain a true branching pattern at all. Since a network-forming program is certain to produce a tree, whether it makes sense or not, some care is required in interpreting the results of computerized tree-forming techniques with real data. Again, this topic is still under investigation, but I can offer a few preliminary remarks.

If the real data set contains a substantial number of cladistically reliable characters, we may reasonably expect the successive weighting program to produce a good estimate of the true tree. The reliable characters will then be highly consistent with the output tree, and if most of the characters in the data set are in good agreement with the output tree, we are justified in believing our answer to be correct. Indeed, if the output tree is consistent with a "large" number of characters, we should accept the output tree, even if the consistent characters were not a majority of the character set.

If the data set does not contain enough good characters to specify a good estimate of the true tree, the output tree will be inconsistent with most of the characters, and it may happen that none of the characters will be highly consistent with the output tree. If this occurs, we may be sure that we have bad data.

Future research into this problem will be concerned with developing precise criteria along the lines of the arguments above. Theory involved with incorporating the number of characters and the number of OTUs in the data set into the evaluation will probably be relevant, and studies of the bearing of character coding on the interpretation may well prove fruitful.

From these considerations, it seems reasonable to conclude that the tests performed may be considered indicative of the inferential power of the successive weighting procedure. While development of techniques to apply the procedure to real data remains to be done, the considerable efficiency of the successive weighting algorithm seems to promise that methods of cladistic inference may soon become highly reliable.

## SUMMARY

The problem of weighting characters is approached through the concepts of unit character consistency and cladistic reliability. Unit characters are highly consistent with a tree if they show little homoplasy on that tree. Unit characters are cladistically reliable if they are consistent with the true phyletic tree.

A set of cladistically reliable characters is consistent with a single tree—the true one—and hence the characters are correlated with each other in a particular way here referred to as hierarchic correlation. A set of cladistically unreliable characters can be hierarchically correlated only by accident.

A good estimate of the true cladistic relationships for a group can be achieved by forming an estimated tree, weighting characters according to their consistency with the tree, and repeating until the tree no longer changes between iterations. This procedure is workable because it reinforces the influence of hierarchically correlated characters on the tree estimate, while eliminating the effects of characters that are random with respect to a branching pattern.

Computer simulation was employed to test the efficiency of the successive weighting procedure. The results varied with the form of the function relating consistency to weight. The most effective weight function gave flawless results even when only 30 reliable characters and 150 unreliable ones were included in a hypothetical data set.

A modification of the character selection technique of Le Quesne (1969) is suggested as a means to obtain an initial set of weights for the input data to the successive weighter.

The use of successive weighting algo-

rithms may soon make it possible to construct reliable techniques for cladistic inference.

### REFERENCES

CAMIN, J. H., AND R. R. SOKAL. 1965. A method for deducing branching sequences in phylogeny. Evolution, 19:311–326.

DEBENEDICTIS, P., AND J. S. FARRIS. In prep. A quantitative study of the phyletic relationships among the grebes.

EDWARDS, A. W. F., AND L. L. CAVALLI-SFORZA. 1964. Reconstruction of evolutionary trees. *In* Phenetic and phylogenetic classification, Publication 6, Systematics Association. London.

FARRIS, J. S. 1966. Estimation of conservatism of characters by constancy within biological populations. Evolution, 20:587–591.

FARRIS, J. S. 1967. The meaning of the relationship and taxonomic procedure. Syst. Zool., 16:44–51.

FARRIS, J. S. In prep. A quantitative study of the phyletic relationships within *Fundulus* and *Profundulus* (Teleostei: Cyprinodontidae).

FARRIS, J. S., A. G. KLUGE, AND M. J. ECKARDT. In press, *a*. On predictivity and efficiency. Syst. Zool.

FARRIS, J. S., A. G. KLUGE, AND M. J. ECKARDT. In press, *b*. A numerical approach to phylogenetic systematics. Syst. Zool.

KLUGE, A. G., AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool., 18:1–32.

KLUGE, A. G. In prep. The evolution of the lizard family Pygopodidae.

LE QUESNE, W. J. 1969. A method of selection of characters in numerical taxonomy. Syst. Zool., 18:201–205.

SIEGEL, S. 1956. Nonparametric statistics for the behavioral sciences. McGraw-Hill, New York. 312 pp.

SOKAL, R. R., AND P. H. A. SNEATH. 1963. Principles of numerical taxonomy. Freeman, San Francisco. 359 pp.

THROCKMORTON, L. 1968. Concordance and discordance of taxonomic characters in *Drosophila* classification. Syst. Zool., 17:355–387.

*Department of Biological Sciences, State University of New York at Stony Brook, Stony Brook, L.I., N. Y. 11790.*