A Probability Measure for Character Compatibility

CHRISTOPHER A. MEACHAM*
Division of Biological Sciences, The University of Michigan, Ann Arbor, Michigan 48109

Received 3 February 1981 revised 18 May 1981

ABSTRACT

It is proved that two undirected binary cladistic characters are compatible iff their smaller states are disjoint or one is a subset of the other. The concept of a cladistic character as an ordered tree of subsets is defined. Cladistic characters that have the same number of elements in their corresponding states are defined to be "nesting equivalent." The equivalence classes of this relation are called "nestings." A certain class of n-tuples is shown to have a biunique correspondence with the $n!$-membered set of all nestings of n binary characters. The model of randomness proposed is that all characters that are nesting equivalent are equally likely. The probability that a pair of undirected binary characters is compatible is derived under this model. This result is extended to collections of undirected binary characters, to collections of directed binary characters, and finally to collections that may include multistate characters. Some proofs are presented which allow a more efficient use of the n-tuple representation of ordered trees of subsets.

INTRODUCTION

Probably the major unifying theme of the biological sciences is the theory of evolution. One aspect of evolution that is important to many biological fields is the concept of evolutionary history. If study shows that some organisms possess a particular feature of anatomy, physiology, or behavior, organisms that are thought to be related to those studied are often, in the absence of conflicting information, inferred to be similar. Conversely, organisms that are thought to be closely related and yet differ in some feature are often studied to obtain critical information about the significance of that feature. In most biological fields, evolutionary histories are obtained intuitively and are usually implicitly derived from taxonomy. One of the challenges of modern systematics is the development of an objective basis for the estimation of evolutionary history.

*Present address: Department of Computer Science,
Memorial University of Newfoundland, St. John's,
Newfoundland, Canada A1C 5S7

©Elsevier North Holland, Inc., 1981
52 Vanderbilt Ave., New York, NY 10017 0025-5564/81/09001+18$02.50

Common to all methods currently recommended as objective bases for inferring evolutionary history is the belief that the features of descendant organisms are derived from the features of their ancestors and that the features of extant organisms thus preserve, in some sense, a record of the evolutionary history of the organisms to which they belong. Each method seeks a pattern in the data with which it is presented and, by a prescribed series of calculations, extracts from the data an estimate of evolutionary history based on the pattern found. Methods of inferring evolutionary history as currently formulated will "succeed" in discovering a pattern even when presented with totally random data. In order to inspire confidence, a method should provide an indication that the pattern discovered differs significantly from a random pattern.

One method that has been proposed for the estimation of evolutionary history is the technique of character compatibility analysis. It is based on facts first noticed by Wilson (1965) [12] and Le Quesne (1969) [8]. Character compatibility analysis has been further developed and given a mathematical foundation by the work of Estabrook and others [1–7, 10]. The results presented here are aimed toward providing a way of calculating the extent to which the observed patterns of compatibility among a set of characters differ from the patterns that would be obtained if the characters bore no relation at all to the evolutionary history of the organisms that possess them.

A *study collection* is a set S of kinds of organisms under investigation. The elements of S are considered *evolutionary units* (EUs) and are assumed to have some historically true evolutionary relationships that can adequately be represented by a tree diagram. A *qualitative character* on S is a basis for assigning the elements of S (the EUs) to mutually exclusive classes called the *states* of the character. For example, the qualitative character "chromosome number" might have the states "$n=7$,""$n=8$," and "$n=9$." The character state "$n=7$" consists of those EUs in S that have a chromosome number $n=7$. A qualitative character is thus a partition of S. A *cladistic character* is a qualitative character together with a hypothesized ordering for the states. In the example above, one would probably hypothesize "$n=8$" to be between "$n=7$" and "$n=9$." The ordered states of a cladistic character constitute a tree of classes of S called a *character-state tree*. If one character state is designated as ancestral, the character-state tree is said to be *directed*; otherwise it is said to be *undirected*. Throughout the discussion that follows, the word *character* should in all cases be understood to mean *cladistic character*.

Le Quesne (1969) [8] introduced the concept of a "uniquely derived" character, so called because the states of such a character have a single evolutionary origin on the historically true evolutionary tree. Estabrook et al. (1975) [2] provide a mathematical formulation of this concept as a "true" character whose evolutionary transformations from state to state occur as single transitions on the true tree in the same order as on the hypothesized character-state tree. Such a character is true because the character-state tree is an incompletely resolved version of the historically true evolutionary tree. For any character there is a set of trees (evolutionary histories) that would make that character true. For a pair of characters, if the corresponding sets of trees have a nonempty intersection, then it is at least possible that both characters are historically true; the two characters are said to be *compatible.* However, if these sets do not intersect, then at least one of the characters is historically incorrect, because there is no possible estimate of evolutionary history that allows both to be true at the same time; the two characters are said to be *incompatible.* Analysis of character compatibility reveals patterns of agreement and disagreement among the hypotheses proposed by the characters in a data set. A *clique* is a set of mutually compatible characters that can, therefore, support the same estimate of evolutionary history. The major result presented here is a method to calculate the probability that an arbitrary collection of random characters is mutually compatible (i.e., consists of members of the same clique) under a reasonable model of randomness.

PAIRS OF UNDIRECTED BINARY CHARACTERS

Consider two undirected binary (two-state) characters on S. Label these two characters K with states k_1, k_2, and label L with states l_1, l_2 so that

$$|l_1| \leqslant |k_1| \leqslant |k_2| \leqslant |l_2|.$$

Define four events:

$$\alpha \equiv \{k_2 \cap l_1 = \varnothing\},$$
$$\beta \equiv \{k_1 \cap l_1 = \varnothing\},$$
$$\gamma \equiv \{k_2 \cap l_2 = \varnothing\},$$
$$\delta \equiv \{k_1 \cap l_2 = \varnothing\}.$$

In addition, define the event $\mathbf{C} \equiv \{K \text{ and } L \text{ are compatible}\}$. Wilson (1965) [12] and Le Quesne (1969) [8] have shown that

$$\mathbf{C} \quad \text{iff} \quad \alpha \text{ or } \beta \text{ or } \gamma \text{ or } \delta. \tag{1}$$

In the trivial case of one-state characters where $|l_1| = 0$, we immediately have **C**. One-state characters are excluded in the discussion that follows because they are always compatible. With this restriction, the relations among the states of K and L are

$$0 < |l_1| \leqslant |k_1| \leqslant |k_2| \leqslant |l_2| < |S|. \tag{2}$$

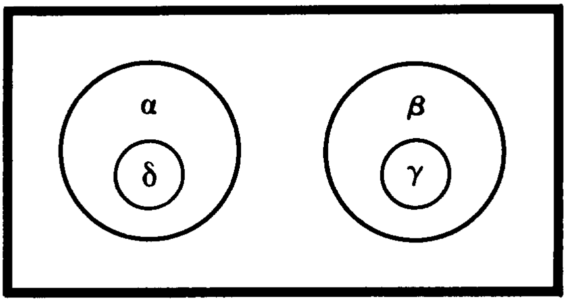

FIG. 1. The Venn diagram for events α, β, γ, and δ.

THEOREM 1

C *iff* α *or* β.

Proof. It follows from (1) that α or β implies **C**. To show **C** implies α or β, it is sufficient to show that δ implies α and that γ implies β.

Recall δ implies $k_1 \cap l_2 = \varnothing$; l_1, l_2 partition S, so that $(k_1 \cap l_1) \cup (k_1 \cap l_2) = k_1$. Thus, δ implies $k_1 \cap l_1 = k_1$ implies $k_1 \subseteq l_1$, but $|l_1| \leqslant |k_1|$ from (2). Thus, $l_1 = k_1$ and so $k_2 = l_2$. Substituting in δ, we have $l_1 \cap k_2 = \varnothing$, which is α.

A similar argument establishes that γ implies β. ■

Note that $|l_1| \geqslant 1$ implies that α and β are exclusive events, because $(k_1 \cap l_1) \cup (k_2 \cap l_1) = l_1$. Figure 1 is the Venn diagram for the four events. Although Le Quesne (1972) [9] does not provide a proof, this result is implicit in his treatment.

The compatibility of K and L depends only on the occurrence of the exclusive events α and β. This allows the derivation of a simple expression for the probability that two binary characters are compatible. Because (in the nontrivial case)

$$\alpha \quad \text{iff} \quad k_2 \cap l_1 = \varnothing \quad \text{iff} \quad l_1 \subseteq k_1$$

and

$$\beta \quad \text{iff} \quad k_1 \cap l_1 = \varnothing \quad \text{iff} \quad l_1 \subseteq k_2,$$

one can think of events α and β as occurring when l_1 is a subset of k_1 or k_2 respectively.

The basis for modeling the probability of character compatibility assumed here is similar to those proposed by Le Quesne (1972) [9] and Sneath et al.

(1975) [11]. Frequencies of occurrence of EUs within the character states of each character are specified in advance, and then all distributions (with these fixed frequencies) of individual EUs among the character states are considered equally likely. This situation can be described in several algebraically equivalent ways. For example, an urn is filled with $|S|$ balls, $|k_1|$ of which are white and $|k_2| \geqslant |k_1|$ of which are black; $|l_1| \leqslant |k_1|$ balls are drawn. Event α is equivalent to drawing all white balls (l_1 is a subset of k_1), and event β is equivalent to drawing all black balls (l_1 is a subset of k_2). The probability of event α is simply the probability of drawing a white ball on the first draw, $|k_1|/|S|$, times the probability of drawing a white ball on the second draw, $(|k_1|-1)/(|S|-1)$, etc. The probability of event α can be written as

$$\mathrm{P}_\alpha = \prod_{i=0}^{|l_1|-1} \frac{|k_1|-i}{|S|-i} = \frac{|k_1|!(|S|-|l_1|)!}{|S|!(|k_1|-|l_1|)!} = \frac{\binom{|k_1|}{|l_1|}}{\binom{|S|}{|l_1|}}.$$

Likewise,

$$\mathrm{P}_\beta = \prod_{i=0}^{|l_1|-1} \frac{|k_2|-i}{|S|-i} = \frac{|k_2|!(|S|-|l_1|)!}{|S|!(|k_2|-|l_1|)!} = \frac{\binom{|k_2|}{|l_1|}}{\binom{|S|}{|l_1|}}.$$

One may observe that because $|k_1| \leqslant |k_2|$, $\mathrm{P}_\alpha \leqslant \mathrm{P}_\beta$ for any pair of undirected binary characters.

Le Quesne (1972) [9] calculates P, the probability that two undirected binary characters are incompatible, based on the same model of randomness used here. Le Quesne's P is simply related to P_α and P_β: $\mathrm{P} = 1 - \mathrm{P}_\alpha - \mathrm{P}_\beta$. Sneath et al. (1975) [11] present a similar model which obtains the probability EX_k that there are at least k "exceptions" to compatibility. Thus the probability that two undirected binary characters are incompatible is EX_1. Again, $\mathrm{EX}_1 = 1 - \mathrm{P}_\alpha - \mathrm{P}_\beta$. The results to be presented here are generalizations of the theory developed by these workers.

Another powerful consequence of Theorem 1 is that the compatibility of undirected binary characters is determined solely by the relationship between their smaller states, k_1 and l_1:

$$\begin{array}{llll} \alpha & \text{iff} \quad k_2 \cap l_1 = \varnothing & \text{iff} \quad l_1 \subseteq k_1, \\ \beta & \text{iff} \quad k_1 \cap l_1 = \varnothing. \end{array}$$

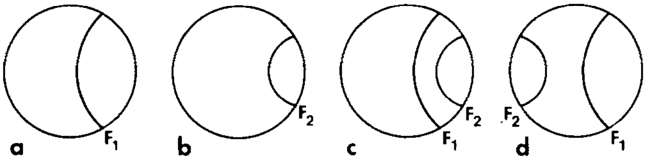

FIG. 2. Diagrams of (a) a binary character F_1; (b) a binary character F_2 ($|F_2| \leq |F_1|$); (c) F_1 and F_2 as α-compatible characters; (d) F_1 and F_2 as β-compatible characters.

Because it is sufficient to consider only the smaller state of a binary character in determining compatibility relationships, it is possible to analyze the conditions necessary for compatibility among sets of characters in terms of these smaller states only. To take fullest advantage of this simplicity in the discussion that follows, I will speak of a character as though it consists only of its smaller state. For example, I will say that character a "contains" character b if the smaller state of character a contains the smaller state of character b. Likewise, I will say the number of EUs in character a "is less than" the number of EUs in character b if the number of EUs in the smaller state of character a is less than the number of elements in the smaller state of character b. In the treatment that follows, the use of set terminology with characters indicates that the smaller state of the character is being discussed. Theorem 1 can now be restated in this fashion:

Two undirected binary characters are compatible iff one is a subset of the other (α*-compatible*) *or they are disjoint* (β*-compatible*).

The use of some simple diagrams can aid in the visualization of relationships among compatible characters. Figure 2(a) is a diagram of a single binary character, F_1. The circle represents the entire study set, S. The character, F_1, is the curved line that divides S into two states. The smaller state of F_1 is the biconvex subset of S on the right. The larger state of F_1 is the concave-convex subset on the left. F_2, diagrammed in Figure 2(b), is a smaller character than F_1; $|F_1| \geq |F_2|$. Figure 2(c) shows F_1 and F_2 as α-compatible characters; F_2 is a subset of F_1. Figure 2(d) shows F_1 and F_2 as β-compatible characters; F_1 and F_2 are disjoint.

COLLECTIONS OF UNDIRECTED BINARY CHARACTERS

This model for the probability of compatibility of pairs of undirected binary characters can be extended to collections of such characters. I wish to derive the probability that, given the frequencies of EUs in the states of a collection of undirected binary characters, this collection is compatible assuming all distributions (with these fixed frequencies) of EUs among the character states to be equally likely. McMorris (1977) [10] has shown that a

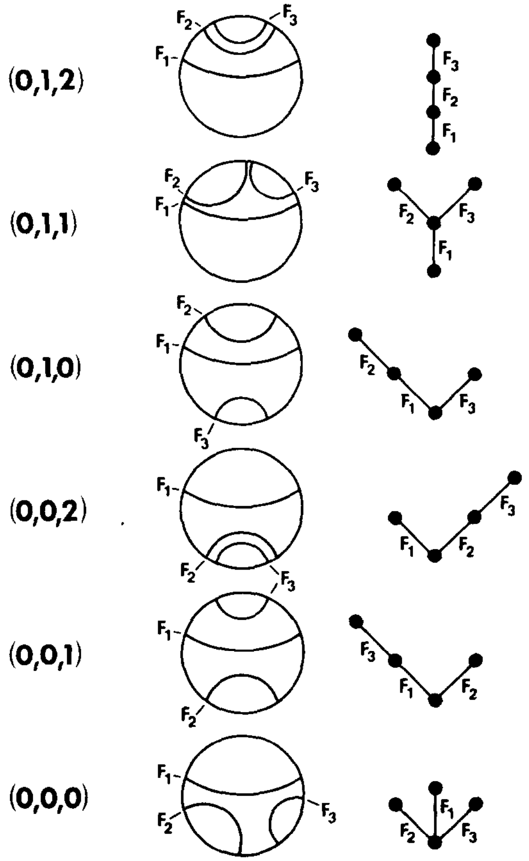

FIG. 3. All possible nestings of three binary characters. In the column on the left are the 3-tuple representations; in the central column are the diagrams of the nestings; in the column on the right are the corresponding trees. The trees are drawn as they would appear if all characters were directed CEP.

collection of undirected binary characters is compatible iff all pairs are compatible. Theorem 1 together with McMorris's result indicates that such a collection will be compatible iff each pair of characters is disjoint or one is a subset of the other. As Figure 2 shows, a pair of such characters can be compatible in two different ways that preserve the frequencies of EUs in the states of the binary characters. Figure 3 shows the diagrams for the six ways that three undirected binary characters can nest to form a tree. Also shown is the tree that corresponds to each of the six diagrams. The diagrams are equivalent to the trees in the sense that they are representations of trees of subsets.

DEFINITION 1 (Estabrook and McMorris, 1980 [6])

A *tree of subsets* of S is a collection $\mathfrak{T}$ of nonempty subsets of S such that

(1) $S \in \mathfrak{T}$,
(2) if $A, B \in \mathfrak{T}$ and $A \cap B \neq \varnothing$, then $A \subseteq B$ or $B \subseteq A$.

For reasons that will become clear, I adopt the convention that $F_0 = S$.

THEOREM 2

A collection $F_1, F_2, \ldots, F_n$ of undirected binary characters on S is compatible iff $\{F_0, F_1, F_2, \ldots, F_n\}$ is a tree of subsets of S.

Proof. This is a direct consequence of Estabrook and McMorris's (1980) [6] Theorem 1 and Definition 6. ■

The following theorem will be helpful:

THEOREM 3

If $F_1, F_2, \ldots, F_{n-1}$ is a compatible collection of undirected binary characters and $|F_n| \leqslant \min\{|F_1|, |F_2|, \ldots, |F_{n-1}|\}$, then $F_1, F_2, \ldots, F_n$ is a compatible collection iff $F_n \cap F_i \neq \varnothing \Rightarrow F_n \subseteq F_i$, $0 \leqslant i \leqslant n-1$.

Proof. $F_1, F_2, \ldots, F_n$ a compatible collection implies $\{F_0, F_1, F_2, \ldots, F_n\}$ a tree of subsets of S implies ($F_n \cap F_i \neq \varnothing \Rightarrow F_n \subseteq F_i$ or $F_i \subseteq F_n$); but $|F_n| \leqslant |F_i|$; thus $F_n \subseteq F_i$, $0 \leqslant i \leqslant n-1$.

$\{F_0, F_1, F_2, \ldots, F_{n-1}\}$ a tree subsets of S and ($F_n \cap F_i \neq \varnothing \Rightarrow F_n \subseteq F_i$) implies $\{F_0, F_1, F_2, \ldots, F_n\}$ a tree of subsets of S implies $F_1, F_2, \ldots, F_n$ a compatible collection. ■

From Theorem 3, it is clear that a set of n characters labeled so that $|F_1| \geqslant |F_2| \geqslant \cdots \geqslant |F_n|$ forms a compatible collection iff the ith character is a subset of one of the classes of the partition of S formed by the first $i-1$ characters. There are two possible three-class partitions of S produced by two compatible undirected binary characters (Figure 2). By Theorem 3, there are three ways a third character can be added to each of these two partitions to form a compatible collection. There are thus six possible four-class partitions of S formed by three binary characters (Figure 3). In short, n undirected binary characters can potentially be compatible in $n!$ ways. Each way in which a collection of undirected binary characters can be compatible corresponds to a way of nesting these characters. In order to calculate the probability that a collection of characters is compatible, it is necessary to sum over all possible nestings the probability of each nesting.

The same tree of subsets can be obtained in more than one way. If for example, $S = \{a, b, c, d, e\}$, $F_1 = \{a, b\}$, and $F_2 = \{d, e\}$, then we have the set $\{\{a, b, c, d, e\}, \{a, b\}, \{d, e\}\}$ of subsets of S, which by Definition 1 is a tree of subsets of S. If $F_1=\{\mathrm{d,e}\}$ and $F_2=\{\mathrm{a,b}\}$, we have $\{\{\mathrm{a,b,c,d,e}\},\{\mathrm{d,e}\},\{\mathrm{a,b}\}\}$ which is the same tree of subsets of S. In the treatment that follows, it is necessary to distinguish among different ways of obtaining the same tree. I introduce the concept of an ordered set of subsets of S.

DEFINITION 2

An *ordered set of subsets* of S is an ordered collection $\mathcal{F}=\{F_0, F_1, F_2,\dots, F_n\}$ of $n+1$ nonempty subsets of S such that $|F_0|\geqslant|F_1|\geqslant\cdots\geqslant|F_n|$.

DEFINITION 3

Two ordered sets of subsets $\mathcal{F},\mathcal{G}$ are *identical*, written $\mathcal{F}=\mathcal{G}$, iff $|\mathcal{F}|=|\mathcal{G}|=n+1$ and $F_i=G_i$ for all $0\leqslant i\leqslant n$.

An ordered set of subsets of S that satisfies Definition 1 is called an *ordered tree of subsets* of S.

DEFINITION 4

Two ordered trees of subsets $\mathcal{F},\mathcal{G}$ are *nesting equivalent* if $|\mathcal{F}|=|\mathcal{G}|$ and for all i, j such that $0\leqslant i\leqslant n$, $0\leqslant j\leqslant n$, $|F_i|=|G_i|$, and $F_j\subseteq F_i \Leftrightarrow G_j\subseteq G_i$.

The classes of the relation "nesting equivalent" are called *nestings*. A nesting is thus a set of ordered trees of subsets of S.

Consider n-tuples of the form $(f_1, f_2,\dots, f_n)$ where each f_i is an integer $0\leqslant f_i\leqslant i-1$. I wish to prove that there is a biunique correspondence between the $n!$ n-tuples of this form and the $n!$ nestings of n binary characters.

DEFINITION 5

An ordered n-tuple $(f_1, f_2,\dots, f_n)$, $0\leqslant f_i\leqslant i-1$, is a *representation* of the ordered tree $\mathcal{F}$ of subsets of S if $F_i\subseteq F_{f_i}$, $1\leqslant i\leqslant n$, and $F_i\cap F_j=\varnothing$, $f_i<j<i$.

Because every subset is contained in $F_0=S$, it is clear that every ordered tree of subsets has some representation.

THEOREM 4

There is a one-to-one correspondence between nestings and representations.

The following lemma is useful:

LEMMA

If an ordered tree of subsets of S has the representation $(f_1, f_2,\dots, f_n)$, *then* $F_q\subseteq F_p$, $q>p$, *iff there is some i such that* $F_q\subseteq F_i$ *and* $f_i=p$.

Proof of lemma

If $F_q\dot{\subseteq} F_p$, $q>p$, then there is some F_i such that $i=\min\{j\,|\,F_q\subseteq F_j$ and $q\geqslant j>p\}$. By Definition 5, $f_i=p$.

That $f_i=p$ implies that $F_i \subseteq F_p$ and $i>p$. $F_q \subseteq F_i$ and $F_i \subseteq F_p$ implies $F_q \subseteq F_p$. $F_q \subseteq F_p$ and $F_i \subseteq F_p$ and $f_i=p$ implies, by Definition 5, that $i \leqslant q$ implies $q>p$. ■

Proof of theorem

Assume $(f_1, f_2, \ldots, f_n)$ is a representation of $\mathcal{F}$ and $(g_1, g_2, \ldots, g_n)$ a representation of $\mathcal{G}$. It is sufficient to show that $\mathcal{F}$ is not nesting equivalent to $\mathcal{G}$ iff $(f_1, f_2, \ldots, f_n) \neq (g_1, g_2, \ldots, g_n)$.

Proving the forward implication, assume $\mathcal{F}$ is not nesting equivalent to $\mathcal{G}$. Without loss of generality we assume that there is some $F_q \subseteq F_p$, $q>p$, but that $G_q \not\subseteq G_p$. If $f_q=p$, then by Definition 5, $G_q \not\subseteq G_p$ implies $g_q \neq p$ and we are done. If $f_q=g_q \neq p$, then by the lemma there is some $i>p$ such that $F_q \subseteq F_i$ and $f_i=p$. But $g_i=p$ and $G_q \not\subseteq G_p$ implies $G_q \not\subseteq G_i$. If $f_q=i$, then $G_q \not\subseteq G_i$ implies $g_q \neq i$. If $f_q=g_q \neq i$, reapplication of the lemma will eventually give us some new i such that $f_q=i$ but $G_q \not\subseteq G_i$ implies $g_q \neq i$, which implies $(f_1, f_2, \ldots, f_n) \neq (g_1, g_2, \ldots, g_n)$.

Conversely, assume there is some i such that $f_i \neq g_i$. Without loss of generality assume $f_i > g_i$. This implies by Definition 5 that $F_i \subseteq F_{f_i}$ but $G_i \not\subseteq G_{f_i}$, which implies $\mathcal{F}$ is not nesting equivalent to $\mathcal{G}$. ■

I use the notation $\mathrm{P}[(f_1, f_2, \ldots, f_n)]$ to denote the probability of the nesting whose representation is $(f_1, f_2, \ldots, f_n)$. Given that $F_1, F_2, \ldots, F_{i-1}$ are compatible and form the nesting represented by $(f_1, f_2, \ldots, f_{i-1})$, the probability that F_i is compatible with $F_1, F_2, \ldots, F_{i-1}$ in such a way as to form the nesting represented by $(f_1, f_2, \ldots, f_i)$ is, by Theorem 3, the probability that $F_i \subseteq F_{f_i}$ and $F_i \cap F_j = \varnothing$ for $f_i < j < i$, that is, the probability that F_i is contained in the "residuum," r_i, of EUs in character F_{f_i} that are not also contained in some F_j, $f_i < j < i$:

$$r_i = F_{f_i} \cap \left(\bigcap_{f_i < j < i} \tilde{F}_j \right)$$

and

$$|r_i| = |F_{f_i}| - \sum_{f_i < j < i,\, f_j = f_i} |F_j|. \tag{3}$$

Thus,

$$\mathrm{P}\big[(f_1, f_2, \ldots, f_i)\big] = \mathrm{P}\big[(f_1, f_2, \ldots, f_{i-1})\big] \cdot \mathrm{P}\big[F_i \subseteq r_i \big| (f_1, f_2, \ldots, f_{i-1})\big],$$

where

$$\mathrm{P}[F_i \subseteq r_i | (f_1, f_2, \ldots, f_{i-1})] = \prod_{j=0}^{|F_i|-1} \frac{|r_i|-j}{|S|-j} = \frac{|r_i|!(|S|-|F_i|)!}{|S|!(|r_i|-|F_i|)!} = \frac{\binom{|r_i|}{|F_i|}}{\binom{|S|}{|F_i|}}.$$

Note that we have no assurance $|F_i| > |r_i|$, in which case $\mathrm{P}[F_i \subseteq r_i | (f_1, f_2, \ldots, f_{i-1})] = 0$. By iteration,

$$\mathrm{P}[(f_1, f_2, \ldots, f_n)] = \prod_{i=1}^{n} \frac{\binom{|r_i|}{|F_i|}}{\binom{|S|}{|F_i|}}. \tag{4}$$

Because $r_1 = F_0 = S$, $\mathrm{P}(F_1 \subseteq r_1) = 1$.

Equations (3) and (4) can be used to develop a straightforward algorithm to calculate the probability that a set of undirected binary characters form a compatible collection by calculating the probability of the nestings represented by each n-tuple and summing these probabilities over all n-tuples.

Many n-tuples may be representations of impossible nestings for a given set of characters because of "packing" considerations. It is often the case that for some F_i, $|F_i| > |r_i|$. Because many n-tuples correspond to impossible nestings, an algorithm to calculate the probability of compatible collections can be made more efficient by eliminating such n-tuples from consideration. I begin by noting that because $|F_1| \geqslant |F_2| \geqslant \cdots \geqslant |F_n|$, the n-tuple $(0, 1, 2, \ldots, n-1)$ represents a possible nesting for any arbitrary collection of undirected binary characters.

THEOREM 5

If the ordered n-tuple $(f_1, f_2, \ldots, f_{i-1}, f_i, i, i+1, \ldots, n-2, n-1)$, $0 \leqslant f_j \leqslant j - 1$, *is a representation of an impossible nesting, then all n-tuples* $(f_1, f_2, \ldots, f_{i-1}, f_i, a_{i+1}, \ldots, a_{n-1}, a_n)$ *where the elements* a_j, $i+1 \leqslant j \leqslant n$, *range over all legitimate values*, $0 \leqslant a_j \leqslant j-1$, *are representations of impossible nestings.*

Proof. If $(f_1, f_2, \ldots, f_{i-1}, f_i, i, i+1, \ldots, n-2, n-1)$ is a representation of an impossible nesting, then because for all j $(i+1 \leqslant j \leqslant n)$ we have $r_j = F_{j-1}$ and $|F_j| \leqslant |F_{j-1}|$, it follows that there must be some f_j $(1 \leqslant j \leqslant i)$ such that $|F_{f_j}| > |r_{f_j}|$, which implies that all $(f_1, f_2, \ldots, f_{i-1}, f_i, a_{i+1}, \ldots, a_{n-1}, a_n)$ are representations of impossible nestings. ■

For example, knowing that (0,0,1,1,2,5,6,7,8,9) represents an impossible nesting tells us that the $6 \times 7 \times 8 \times 9 \times 10 = 30240$ 10-tuples

$(0,0,1,1,2,\cdot,\cdot,\cdot,\cdot,\cdot)$ represent impossible nestings. In the worst case, $n!$ calculations are necessary, but Theorem 5 can permit the number of calculations to be reduced to a lower limit of 2^{n-1}.

COLLECTIONS OF DIRECTED BINARY CHARACTERS

A further extension of this model can be made to directed binary characters. Two kinds of directed binary characters are distinguished.

DEFINITION 6

A directed binary character is said to be *common-equals-primitive* (CEP) if the number of EUs in the primitive state is greater than or equal to the number of EUs in the advanced state, else it is said to be *rare-equals-primitive* (REP).

I introduce the notation $\vec{F}_1 = \uparrow$, read "the direction of F_1 is CEP," and $\vec{F}_1 = \downarrow$, "the direction of F_1 is REP."

THEOREM 6

A set of directed binary characters $F_1, F_2, \ldots, F_n$ is compatible iff the set is compatible as undirected characters and there exist no F_p, F_q such that ($\vec{F}_p = \uparrow$ and $\vec{F}_q = \downarrow$ and $F_q \subsetneq F_p$) or ($\vec{F}_p = \vec{F}_q = \downarrow$ and $F_p \cap F_q = \varnothing$).

Proof. Estabrook et al. (1976) [3] have shown that a set of directed binary characters is compatible iff each pair is compatible. Thus, I need only demonstrate here that Theorem 6 is true for $n=2$ (a pair of directed binary characters).

That a pair of directed binary characters is compatible implies that the undirected characters are compatible implies by Theorem 1 that they are disjoint or one is contained in the other. There are four combinations for directing two characters which, in combination with these two nestings, produce eight exhaustive possibilities for two directed characters F_1, F_2 such that $|F_1| \geqslant |F_2|$:

(a) $\vec{F}_1 = \vec{F}_2 = \uparrow,\ F_2 \subseteq F_1$, (e) $\vec{F}_1 = \vec{F}_2 = \uparrow,\ F_1 \cap F_2 = \varnothing$,
(b) $\vec{F}_1 = \uparrow,\ \vec{F}_2 = \downarrow,\ F_2 \subset F_1$, (f) $\vec{F}_1 = \uparrow,\ \vec{F}_2 = \downarrow,\ F_1 \cap F_2 = \varnothing$,
(c) $\vec{F}_1 = \downarrow,\ \vec{F}_2 = \uparrow,\ F_2 \overset{\to}{\subseteq} F_1$, (g) $\vec{F}_1 = \downarrow,\ \vec{F}_2 = \uparrow,\ F_1 \cap F_2 = \varnothing$,
(d) $\vec{F}_1 = \vec{F}_2 = \downarrow,\ F_2 \subseteq F_1$, (h) $\vec{F}_1 = \vec{F}_2 = \downarrow,\ F_1 \cap F_2 = \varnothing$.

These eight possibilities are shown in Figure 4. Note that only in diagrams (b) and (h) are the directions of the characters inconsistent. Thus, two directed binary characters are compatible iff they are compatible as undirected characters and not [($\vec{F}_1 = \uparrow$, $\vec{F}_2 = \downarrow$ and $F_2 \subsetneq F_1$) or ($\vec{F}_1 = \vec{F}_2 = \downarrow$ and $F_1 \cap F_2 = \varnothing$)]. ■

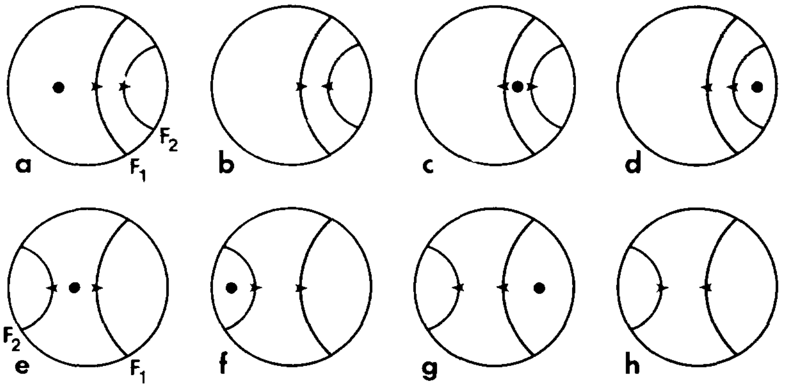

FIG. 4. All possible relationships between two directed binary characters, F_1 and F_2 ($|F_1| \geq |F_2|$). The arrow points from the primitive state to the advanced state for each character. In (a), (b), (e), and (f), F_1 is CEP; in (c), (d), (g), and (h), F_1 is REP. In (a), (c), (e), and (g), F_2 is CEP; in (b), (d), (f), and (h), F_2 is REP. The dot in each diagram indicates the root. In diagrams (b) and (h), the directions of the two characters are inconsistent. Hence in these two cases, the characters are incompatible as directed characters even though they would be compatible if undirected.

Theorem 6 immediately indicates that no REP character can be contained in a CEP character and that of any two REP characters, one must contain the other. These conditions place rigid restrictions on n-tuples that can represent compatible nestings of directed binary characters. The largest REP character in a collection must be contained in F_0 only, the second largest REP character must be contained in the largest, the third in the second, etc. The element of an n-tuple that corresponds to an REP character is always the index of the next larger REP character or 0 if no such character exists. For example, if in a collection of ten characters the 3rd, 5th, 6th, 8th, and 10th characters are REP, only 10-tuples of the form $(\cdot,\cdot,0,\cdot,3,5,\cdot,6,\cdot,8)$ can represent possible nestings.

Because an n-tuple representation can only indicate that a character with a larger index is contained in a character with a smaller index and because CEP characters can only be contained in REP characters and not vice versa, for any two characters of the same size but different directions, the REP character must be given a lower index than the CEP character. For directed characters, an added condition must be made to the definition of an ordered set of subsets (Definition 2), i.e., $(|F_p| = |F_q| \text{ and } \vec{F}_p \neq \vec{F}_q \text{ and } p < q) \Rightarrow (\vec{F}_p = \downarrow$ and $\vec{F}_q = \uparrow)$.

COLLECTIONS OF MULTISTATE CHARACTERS

The final generalization of this model is to arbitrary collections that may include multistate characters, directed or not. As before, I assume that, given the observed frequencies of EUs in the character states, all distributions of EUs among the character states are equally likely. For multistate characters, I additionally require that the ordering of character states be fixed. Using the terms defined above, I assume all ordered trees of subsets of S that are nesting equivalent to the observed character state trees are equally likely.

Let $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$ be a collection of m multistate characters on S, and $n_i = |\mathcal{K}_i| - 1$. Each $\mathcal{K}_i$ is an ordered tree of subsets of S, $\{K_{i0}, K_{i1},, \ldots, K_{in_i}\}$ $(K_{i0} = S)$, whose representation is $(k_{i1}, k_{i2}, \ldots, k_{in_i})$. The subsets K_{ij}, $1 \leqslant j \leqslant n_i$, of S are the smaller states of the binary cladistic characters called the factors of $\mathcal{K}_i$. Estabrook et al. (1976) [4] and, in a different context, Estabrook and McMorris (1980) [6] have proved that a collection of multistate characters is compatible iff their binary factors are compatible.

Define an ordered tree of subsets of S, called $\mathcal{F}$, whose elements correspond to the factors of the multistate characters. Let $n = \Sigma_{i=1}^{m} n_i$. For every element F_p, $1 \leqslant p \leqslant n$, of $\mathcal{F}$, there is a corresponding factor K_{uv}, $1 \leqslant u \leqslant m$, $1 \leqslant v \leqslant n_u$, written $F_p \simeq K_{uv}$.

DEFINITION 7

The elements of an ordered set $\mathcal{F}$ of subsets of S are said to *correspond* to the factors of a set of multistate characters on S, $\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_m$, if

(1) $|\mathcal{F}| = (\Sigma_{i=1}^{m} |\mathcal{K}_i| - 1) + 1 = n + 1$,
(2) $F_0 = S$,
(3) $F_p \simeq K_{uv} \Rightarrow |F_p| = |K_{uv}|$,
(4) $|F_0| \geqslant |F_1| \geqslant \cdots \geqslant |F_n|$,
(5) $(K_{uv} \simeq F_p \text{ and } K_{uw} \simeq F_q \text{ and } v < w) \Rightarrow p < q$,
(6) $(|F_p| = |F_q| \text{ and } \vec{F}_p \neq \vec{F}_q \text{ and } p < q) \Rightarrow (\vec{F}_p = \downarrow \text{ and } \vec{F}_q = \uparrow)$.

Essentially, these conditions mean (1) that $\mathcal{F}$ has as many factors as the sum of the factors in the characters, $\mathcal{K}_i$, (2) that S is contained in $\mathcal{F}$, (3) that each factor in $\mathcal{F}$ is the same size as its corresponding factor, (4) that the elements of $\mathcal{F}$ are ordered properly for an ordered set of subsets, (5) that any two factors corresponding to factors of the same character are ordered the same way in $\mathcal{F}$ as in the original character, and (6) that for any two factors of the same size but different directions, the REP character has the lower index.

If the factors F_i were independent binary characters, it would be possible to calculate the probability of compatibility by summing over all n-tuple representations the probability of each representation. However, because these elements are factors of multistate characters, the model of randomness requires that we only consider as possibilities those n-tuples that represent nestings where the ordered set of elements corresponding to the factors of each original multistate character is nesting equivalent to the original multistate character. Algorithmically, we need a way of verifying that the representations of the original character state trees and the ordered tree of subsets of the corresponding elements are the same.

Define an n-tuple $(c_1, c_2, \dots, c_n)$ such that $c_i = u \Rightarrow F_i \simeq K_{uv}$. That is, the ith element of this n-tuple is the index of the character to whose factor the ith element of $\mathcal{F}$ corresponds. Define an n-tuple $(s_1, s_2, \dots, s_n)$ such that $s_i = p \Rightarrow (F_i \simeq K_{c_i w}$ and $k_{c_i w} = v$ and $K_{c_i v} \simeq F_p)$. The ith element of this n-tuple is the index of the next larger factor F_{s_i} that contains F_i and corresponds to a factor of the same multistate character, $\mathcal{K}_{c_i}$. I also define a set $\mathcal{L}_i$, $1 \leqslant i \leqslant m$, of subsets of $\mathcal{F}$, such that $\mathcal{L}_i = \{F_0\} \cup \{F_j | c_j = i\}$ and the elements of each $\mathcal{L}_i$ are ordered by their indices as elements of $\mathcal{F}$. Thus each $\mathcal{L}_i$ is the set of elements of $\mathcal{F}$ that correspond to the factors of $\mathcal{K}_i$. The notation for the elements and for the representation of $\mathcal{L}_i$ is parallel to that of $\mathcal{K}_i$. The model of randomness requires that only representations of $\mathcal{F}$ where $\mathcal{L}_i$ is nesting equivalent to $\mathcal{K}_i$, $1 \leqslant i \leqslant m$, be allowed.

THEOREM 7

Given an ordered tree of subsets of S, $\mathcal{F}$, whose elements F_i, $1 \leqslant i \leqslant n$, correspond to the factors of the characters $\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_m$; then $\mathcal{L}_u$ is nesting equivalent to $\mathcal{K}_u$ for $1 \leqslant u \leqslant m$ iff for all p, q such that $1 \leqslant p \leqslant n$, $s_p < q < p$, we have $F_p \subseteq F_{s_p}$ and $c_q = c_p \Rightarrow F_q \cap F_p = \varnothing$.

Proof. Because of condition (5) of Definition 7 and the definition of $\mathcal{L}_u$, we have $L_{uv} \simeq K_{uv}$, and by condition (3), $|L_{uv}| = |K_{uv}|$, $1 \leqslant u \leqslant m$, $0 \leqslant v \leqslant n_u$.

Proving the forward implication, assume $\mathcal{L}_u$ is nesting equivalent to $\mathcal{K}_u$, $1 \leqslant u \leqslant m$. Let $k_{uw} = v$; then by Definition 5, $K_{uw} \subseteq K_{uv}$ and $K_{uw} \cap K_{ui} = \varnothing$, $v < i < w$. By the definition of s_p, $k_{uw} = v$ implies $K_{uv} \simeq F_{s_p}$ and $K_{uw} \simeq F_p$. As above, $L_{uv} \simeq K_{uv}$, but $L_{uv} \simeq K_{uv} \sim F_{s_p}$ implies $L_{uv} = F_{s_p}$. Similarly, $L_{uw} = F_p$. That $\mathcal{L}_u$ is nesting equivalent to $\mathcal{K}_u$ implies $L_{uw} \subseteq L_{uv}$ and $L_{uw} \cap L_{ui} = \varnothing$, $v < i < w$, which implies $F_p \subseteq F_{s_p}$ and $c_q = c_p \Rightarrow F_q \cap F_p = \varnothing$, $s_p < q < p$.

Conversely, assume

(1) $c_p = u$,
(2) for $s_p < q < p$, $c_q = u \Rightarrow F_q \cap F_p = \varnothing$ and $F_p \subseteq F_{s_p}$,
(3) $L_{uw} = F_p$ and $L_{uv} = F_{s_p}$.

That $F_p \subseteq F_{s_p}$ implies $L_{uw} \subseteq L_{uv}$. That for $s_p < q < p$, $c_q = u \Rightarrow F_q \cap F_p = \varnothing$ implies that for $v < i < w$, $L_{uw} \cap L_{ui} = \varnothing$. By Definition 5, $l_{uw} = v$, but by the definition of s_p, $k_{uw} = v$. Thus for all u, w such that $1 \leqslant u \leqslant m$, $1 \leqslant w \leqslant n_u$, we have $l_{uw} = k_{uw}$. By Theorem 4, $\mathcal{L}_u$ is nesting equivalent to $\mathcal{K}_u$, $1 \leqslant u \leqslant m$. ■

The expression for the probability of each representation of $\mathcal{F}$ is similar to Equation (4), but in the case of multistate characters we are constrained to draw each subset F_i of S from F_{s_i}. So

$$P[(f_1,f_2,\ldots,f_n)]=\prod_{i=1}^{n}\frac{\binom{|r_i|}{|F_i|}}{\binom{|F_{s_i}|}{|F_i|}}. \tag{5}$$

From the previously defined n-tuples and Equations (3) and (5), it is possible to construct an algorithm that calculates the probability that an arbitrary collection of characters is mutually compatible under the proposed model of randomness.

EXAMPLES

I now present some examples of calculated probability of compatibility. The notation $(n_1|n_2)$ is used for a binary character that has n_1 EUs in one state and n_2 EUs in the other. The notation $(n_1|n_2|n_3)$ is used for a three-state character. For directed characters, the number of EUs in the primitive state is in italics.

Two undirected binary characters $F_1=(7|3)$ and $F_2=(8|2)$ can be compatible in two ways, represented by (0,0) and (0,1):

$$\mathrm{P}_\alpha=\mathrm{P}[(0,1)]=\frac{\binom{3}{2}}{\binom{10}{2}}=\frac{1}{15},\qquad \mathrm{P}_\beta=\mathrm{P}[(0,0)]=\frac{\binom{7}{2}}{\binom{10}{2}}=\frac{7}{15}.$$

The total probability is $\mathrm{P}_T=\frac{8}{15}$.

If the same two characters are directed in this way: $F_1=(\mathit{7}|3)$ and $F_2=(8|\mathit{2})$, only representations of the form $(\cdot,0)$ are possible because F_2 is REP. The total probability is $\mathrm{P}_T=\mathrm{P}[(0,0)]=\frac{7}{15}$.

For three undirected binary characters $F_1=(6|4)$, $F_2=(7|3)$, and $F_3=(8|2)$, there are six nestings:

$$\mathrm{P}[(0,1,2)]=\frac{\binom{4}{3}}{\binom{10}{3}}\cdot\frac{\binom{3}{2}}{\binom{10}{2}}=\frac{1}{450},\qquad \mathrm{P}[(0,1,1)]=\frac{\binom{4}{3}}{\binom{10}{3}}\cdot 0=0,$$

$$\mathrm{P}[(0,1,0)]=\frac{\binom{4}{3}}{\binom{10}{3}}\cdot\frac{\binom{6}{2}}{\binom{10}{2}}=\frac{1}{90},\qquad \mathrm{P}[(0,0,2)]=\frac{\binom{6}{3}}{\binom{10}{3}}\cdot\frac{\binom{3}{2}}{\binom{10}{2}}=\frac{1}{90},$$

$$\mathrm{P}[(0,0,1)] = \frac{\binom{6}{3}}{\binom{10}{3}} \cdot \frac{\binom{4}{2}}{\binom{10}{2}} = \frac{1}{45}, \qquad \mathrm{P}[(0,0,0)] = \frac{\binom{6}{3}}{\binom{10}{3}} \cdot \frac{\binom{3}{2}}{\binom{10}{2}} = \frac{1}{90}.$$

The 3-tuple (0, 1, 1) represents an impossible nesting because $|r_3| = 1 < |F_3| = 2$. Then $\mathrm{P}_T = \frac{13}{225}$.

If these three characters are directed, with $F_1 = (6|\mathit{4})$, $F_2 = (\mathit{7}|3)$, $F_3 = (8|\mathit{2})$, then F_1 and F_3 are REP; therefore only representations of the form $(0, \cdot, 1)$ are possible, but (0, 1, 1) is eliminated as above because of packing, which leaves only $\mathrm{P}_T = \mathrm{P}[(0,0,1)] = \frac{1}{45}$.

Given an undirected three-state character (6|1|3) and a binary character (8|2), the three-state character has two binary factors: (6|4) and (7|3). This gives a total of three factors: $F_1 = (6|4)$, $F_2 = (7|3)$, and $F_3 = (8|2)$, as above, but now, because F_1 and F_2 correspond to factors of the same multistate character, we have the added requirement that $F_2 \subseteq F_1$. So possible 3-tuples must have the form $(\cdot, 1, \cdot)$. This gives three possibilities (0, 1, 2), (0, 1, 1), and (0, 1, 0). Again (0, 1, 1) represents an impossible packing. So

$$\mathrm{P}[(0,1,2)] = \frac{\binom{4}{3}}{\binom{4}{3}} \cdot \frac{\binom{3}{2}}{\binom{10}{2}} = \frac{1}{15} \quad \text{and} \quad \mathrm{P}[(0,1,0)] = \frac{\binom{4}{3}}{\binom{4}{3}} \cdot \frac{\binom{6}{2}}{\binom{10}{2}} = \frac{1}{3}$$

for a total $\mathrm{P}_T = \frac{6}{15}$. If these two characters are directed, with $(6|\mathit{1}|3)$ and $(8|\mathit{2})$, we again have the directed factors $F_1 = (6|\mathit{4})$, $F_2 = (\mathit{7}|3)$, and $F_3 = (8|\mathit{2})$. As above, because of the multistate character, only 3-tuples $(\cdot, 1, \cdot)$ are possible, but because F_1 and F_3 are REP only 3-tuples of the form $(0, \cdot, 1)$ are possible. Thus, only the 3-tuple (0, 1, 1) is potentially possible, but as above, $|r_3| = 1 < |F_3| = 2$ for (0, 1, 1) and this 3-tuple is impossible because of packing. Consequently, the characters $(6|\mathit{1}|3)$ and $(8|\mathit{2})$ cannot possibly be compatible; $\mathrm{P}_T = 0$.

This paper was initiated by a series of conversations with Leslie R. Landrum. I thank George F. Estabrook for his encouragement and valuable comments. Other persons who contributed materially to the development of my ideas deserve recognition: T. O. Duncan, K. L. Fiala, T. R. Getty, P. C. Harley, F. A. Hartman, G. D. McPherson, N. L. Oden, S. G. Poss, D. L. Schultz, and J. G. Strauch, Jr. I gratefully acknowledge the support of a National Science Foundation Graduate Fellowship. This paper is a portion of a doctoral dissertation to be submitted to the Horace H. Rackham School of Graduate Studies of The University of Michigan.

REFERENCES

1 G. F. Estabrook, Cladistic methodology: a discussion of the theoretical basis for the induction of evolutionary history, *Ann. Rev. Ecol. Syst.* 3:427–456 (1972).

2 G. F. Estabrook, C. S. Johnson, Jr., and F. R. McMorris, An idealized concept of the true cladistic character, *Math. Biosci.* 23:263–272 (1975).

3 G. F. Estabrook, C. S. Johnson, Jr, and F. R. McMorris, An algebraic analysis of cladistic characters, *Discrete Math.* 16:141–147 (1976).

4 G. F. Estabrook, C. S. Johnson, Jr, and F. R. McMorris, A mathematical foundation for the analysis of cladistic character compatibility, *Math. Biosci.* 29:181–187 (1976).

5 G. F. Estabrook and L. R. Landrum, A simple test for the possible simultaneous evolutionary divergence of two amino acid positions, *Taxon* 24:609–613 (1975).

6 G. F. Estabrook and F. R. McMorris, When is one estimate of evolutionary relationships a refinement of another? *J. Math. Biol.* 10:367–373. (1980).

7 G. F. Estabrook and C. A. Meacham, How to determine the compatibility of undirected character state trees, *Math. Biosci.* 46:251–256 (1979).

8 W. J. Le Quesne, A method of selection of characters in numerical taxonomy, *Syst. Zool.* 18:201–205 (1969).

9 W. J. Le Quesne, Further studies based on the uniquely derived character concept, *Syst. Zool.* 21:281–288 (1972).

10 F. R. McMorris, On the compatibility of binary qualitative taxonomic characters, *Bull. Math. Biology* 39:133–138 (1977).

11 P. H. A. Sneath, M. J. Sackin, and R. P. Ambler, Detecting evolutionary incompatibilities from protein sequences, *Syst. Zool.* 24:311–332 (1975).

12 E. O. Wilson, A consistency test for phylogenies based on contemporaneous species, *Syst. Zool.* 14:214–220 (1965).