J. Math. Biology 4, 195—200 (1977)

Journal of Mathematical Biology

© by Springer-Verlag 1977

When Are Two Qualitative Taxonomic Characters Compatible?

G. F. Estabrook, Ann Arbor, Michigan, and **F. R. McMorris,** Bowling Green, Ohio

Received October 25, 1976

Summary

A proof is given of a procedure that has previously appeared claiming to determine when two amino acid positions on a protein could both possibly be divergent taxonomic characters. An algorithm for executing this procedure is described.

Introduction

Proteins such as cytochrome *c* and myoglobin as well as nucleotide sequences have been used by many biologists to construct estimates of the evolutionary history of various collections of taxa (Boulter et al., 1972; Fitch and Margoliash, 1967; Fitch, 1975; Moore et al., 1973). Each position on a sequence is considered a taxonomic character. If it were known for certain which characters were divergent (i.e., underwent no reversals or parallel evolution) within the particular collection of taxa under study, then these characters could be used to indicate various branches of the true evolutionary history of the collection, assuming of course that this evolutionary history is in the form of a tree.

Obviously it is impossible to know positively if a character did indeed undergo divergent evolution of its states, but it nevertheless is an important problem to determine if it is at least logically possible for a set of characters to all be divergent (Estabrook, 1972; Sneath et al., 1975). Recently a procedure has been offered by Estabrook and Landrum (1975) and independently by Fitch (1975) that purportedly tells when two characters could both be divergent. These authors did not have mathematical verification that their procedure does indeed do what they claim. It is the purpose of this note to place this procedure on firm mathematical ground and present an algorithm that allows its implementation on the computer.

Rather than have most of the space in this short note devoted to definitions, we will assume some knowledge of graph theory terminology as found in (Roberts, 1976) and the algebraic terms and results as found in (Estabrook et al., 1975, 1976 a, b; McMorris, 1975).

Results

Throughout, S will denote a (finite) study collection of evolutionary units (EU's) and S' will denote the true evolutionary history of S. We need some necessary definitions and comments.

A *tree* is a connected graph having no circuits. Note that for any two distinct vertices of a tree, there exists a unique chain joining them. If a vertex r of a tree is picked and all edges oriented toward r (or all oriented away from r), this gives a digraph which is often referred to as a *rooted tree*. In most methods for estimating evolution, including the present one, the assumption is made that S consists of EU's for which S' is a rooted tree containing S. A rooted tree may also be viewed as a tree (lower) semilattice, so that the most recent common ancestor of two EU's is their greatest lower bound in the semilattice S'.

A *qualitative taxonomic character* on S is a mapping from S onto another set, called the *character state set* of the character. We will simply use the word "character" for the above notion. Examples of characters are the various amino acid positions on homologous proteins of the EU's. In particular (Boulter et al., 1972), if $S=\{$sunflower, cotton, tomato$\}$, then the fourth amino acid position K_4 of cytochrome c is given by K_4(sunflower)=Ala, K_4(cotton)=Gln, and K_4(tomato)=Asp. Notice that if $K: S \to P$ is a character on S, then K induces a partition of S with cells $K_\alpha = \{x \in S: K(x) = p_\alpha\}$ for each $p_\alpha \in P$.

Now let T be a tree containing S as a subset of its vertex set. A map (character) $K: T \to P$ is said to be *ideally related to* T if each K_α is a connected subgraph of T. For the record, Fitch would say that T is a "most parsimonious tree" for K (Fitch, 1975). If K is ideally related to T, then each K_α is convex, i.e., if $x, y \in K_\alpha$ and $x, x_1, \ldots, x_n, y$ is a chain joining x and y, then $x_i \in K_\alpha$ for all $i=1, \ldots, n$. Note also that the definition of ideally related makes sense when T is rooted if we replace "connected" with "weakly connected". Now if K is ideally related to T and $T=S'$, then K is what we have called *divergent*, i.e., $\wedge K_\alpha \leq \wedge K_\beta \leq x$, with $x \in K_\alpha$ implies that $K_\alpha = K_\beta$ (Estabrook et al., 1975). Divergent characters are important since they are able to distinguish branches of S'.

The characters K and L on S are said to be *compatible* if and only if there exists a tree T containing S such that K and L can be extended to maps that are both ideally related to T. Hence if K and L are compatible, then it is possible that both are divergent, while if K and L are not compatible, then at least one is not divergent (i.e., misrepresents evolutionary relationships).

Now let $K: S \to P = \{p_1, \ldots, p_m\}$ and $L: S \to Q = \{q_1, \ldots, q_n\}$ be characters on S. The *matrix* of K and L is the $m \times n$ matrix (a_{ij}) with $a_{ij} \in \{0, 1\}$ and $a_{ij}=1$ if and only if there exists $x \in S$ such that $K(x)=p_i$ and $L(x)=q_j$. As an example, let $K: S \to \{A, C, G\}$ and $L: S \to \{A, U\}$ with $S=\{r, x, y, z, w\}$ be given by $K(x)=$ $=K(r)=K(y)=A$, $K(z)=C$, $K(w)=G$, $L(r)=L(x)=L(z)=A$, $L(y)=L(w)=U$. Then setting $p_1=A$, $p_2=C$, $p_3=G$ and $q_1=A$, $q_2=U$, the matrix, for this indexing, is

$$\begin{array}{c} \\ A \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cc} A & U \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{array}$$

Clearly different indexing of P and Q will result in different data matrices, but the indices are usually fixed at the start of an investigation so that the non-uniqueness problem does not occur in practice.

The *data graph* is formed by considering the non-zero entries of the data graph as vertices and defining an edge between two vertices if they are adjacent non-zero entries in the same row or column. Labeling the data graph by placing those EU's that map onto each vertex we have for above example the following labeled data graph.

We can now state the theorem which is a translation of the unproven procedure announced in (Estabrook and Landrum, 1975) and (Fitch, 1975).

Theorem: *The characters K and L on S are compatible if and only if their data graph has no circuits.*

Proof: First it is important to notice that the initial indexing of the states of K and L does not effect the existence of circuits. Assume that the data graph has no circuits, fix an indexing of the states, $K: S \to \{p_1, \ldots, p_m\}$, $L: S \to \{q_1, \ldots, q_n\}$, and label the graph with the EU's at the vertices. This labeled graph is easily seen to be a forest (a disjoint union of trees).

From the construction of the labeled data graph, the set of all EU's represented in row i is precisely $K_i = \{x \in S: K(x) = p_i\}$, and the set of all EU's in column j is precisely L_j. If $x_1 \ldots x_l$ is a vertex with $l > 1$ relabel it as $(x_1 \ldots x_l)^*$. For each such vertex $(x_1 \ldots x_l)^*$, add the vertices $x_1, x_2, \ldots, x_l$ and define a new edge from each x_i to $(x_1 \ldots x_l)^*$. We now have a labeled forest F containing S as a subset of its vertex set. Define $K^* : F \to \{p_1, \ldots, p_m\}$ by $K^*(z) = K(z)$ for all $z \in S$ and $K^*((x_1 \ldots x_l)^*) = K(x_1)$ $(= K(x_2) = \ldots = K(x_l))$. Joining the components of F together in the obvious way to form a tree T, it is clear that K^* extends K and is ideally related to T. A similar argument holds for L and thus we have shown that K and L are compatible.

For the converse, we must use some results from our earlier work. Recall (Estabrook et al., 1975) that a rooted tree can be considered as a tree (lower) semilattice, and that a *cladistic character* on S is a map $K: S \to P$ where P is a tree semilattice. We have shown in (Estabrook et al., 1976 b) that two cladistic characters K and L on S are compatible if and only if $\langle \mathrm{Im}(K \times L) \rangle$ is a tree subsemilattice of the semilattice $P \times Q$, where $\mathrm{Im}(K \times L)$ denotes the image of $K \times L$ in $P \times Q$ and $\langle \mathrm{Im}(K \times L) \rangle$ is the subsemilattice generated by $\mathrm{Im}(K \times L)$. It was noted in (McMorris, 1975) that the qualitative characters K and L are compatible if and

only if there exist three semilattice orderings of P and Q making K and L compatible cladistic characters.

Now assume that the characters K and L are compatible and that the data graph contains a circuit. It is not hard to see that the states may be relabeled in such a way that $a_{11}, a_{12}, a_{22}, a_{23}, a_{33}, \ldots, a_{k1}$ $(k \geq 2)$ are vertices on the circuit. Note that, after relabeling, vertex a_{ij} is not necessarily in the i-th row and j-th column. Now $a_{11}, a_{12}, \ldots, a_{k1}$ on the circuit implies that the elements

$$(p_1, q_1), (p_1, q_2), (p_2, q_2), (p_2, q_3), \ldots, (p_k, q_1)$$

are members of $\mathrm{Im}\,(K \times L)$ in $P \times Q$ and since K and L are compatible, there are tree semilattice orderings of P and Q so that $\langle \mathrm{Im}\,(K \times L)\rangle$ is a tree semilattice.

We have $(p_1 \wedge p_2, q_2)$, $(p_1, q_1 \wedge q_2) \leq (p_1, q_2)$ in $\langle \mathrm{Im}\,(K \times L)\rangle$ so that either $(p_1, q_1 \wedge q_2) \leq (p_1 \wedge p_2, q_2)$ or $(p_1 \wedge p_2, q_2) \leq (p_1, q_1 \wedge q_2)$. Assume the first case. A similar argument will work for the second case. Now $(p_1, q_1 \wedge q_2) \leq (p_1 \wedge p_2, q_2)$ implies that $p_1 = p_1 \wedge p_2$ and hence $p_1 < p_2$.

Since $(p_1 \wedge p_2, q_2), (p_2, q_2 \wedge q_3) \leq (p_2, q_2)$ and $p_1 < p_2$, we have $q_2 < q_3$.

Since $(p_2, q_2 \wedge q_3), (p_2 \wedge p_3, q_3) \leq (p_2, q_3)$ and $q_2 < q_3$, we have $p_2 < p_3$.

Since $(p_2 \wedge p_3, q_3), (p_3, q_3 \wedge q_4) \leq (p_3, q_3)$ and $p_2 < p_3$, we have $q_3 < q_4$.

Continuing this way we have $p_1 < p_2 < \ldots < p_k$ and $q_2 < q_3 < \ldots < q_k$ immediately before checking

$$(p_{k-1} \wedge p_k, q_k), (p_k, q_k \wedge q_1) \leq (p_k, q_k)$$

which yields $q_k < q_1$.

Finally, $(p_1 \wedge p_k, q_1), (p_k, q_k \wedge q_1) \leq (p_k, q_1)$ implies $q_1 < q_k$ which is the desired contradiction. Q.E.D.

Implementation of Results

In an actual taxonomic study many characters are used. Although it is not true that a pairwise compatible collection of characters is always compatible (McMorris, 1975), the taxonomist might still want to know those characters that are pairwise compatible. We now give a method by which the result proved in the previous section can be implemented on the computer and used for large data sets.

We say that the unordered pair of characters $K : S \to P = \{p_1, \ldots, p_m\}$ and $L : S \to Q$ is *simply compatibly extended* to the unordered pair of characters $K^* : S^* \to P^* = \{p_1, \ldots, p_m, p_{m+1}\}$, $L^* : S^* \to Q^*$ if $Q^* = Q$, $S \subseteq S^*$, K agrees with K^* on S, L agrees with L^* on S, $K^*(x) = p_{m+1}$ for all $x \in S^* \setminus S$, and $L^*(x) = L^*(y)$ for all $x, y \in S^* \setminus S$.

Notice that if K and L can be simply compatibly extended to K^* and L^*, then the matrix for K^* and L^* is identical to that of K and L except for containing an additional row or column that has all entries 0 except for a single 1. Thus we see that if K^* and L^* simply compatibly extend K and L, then K and L are compatible

if and only if K^* and L^* are compatible. This is true since the existence of circuits in a graph is not effected by adjoining one vertex and one edge.

The characters K^* and L^* *compatibly extend* K and L if there exists characters $K^{(i)}, L^{(i)}$ $1 \leq i \leq s$ such that $K^{(1)} = K$, $L^{(1)} = L$, $K^{(s)} = K^*$, $L^{(s)} = L^*$ with $K^{(i+1)}$ and $L^{(i+1)}$ a simple compatible extension of $K^{(i)}$ and $L^{(i)}$ for $1 \leq i \leq s-1$. Hence if K^* and L^* compatibly extend K and L, then K^* and L^* are compatible if and only if K and L are compatible.

If the matrix of two characters contains a row or column with a single nonzero entry, removal of the EU's that correspond to this entry produces a pair of characters of which the original pair is a simple compatible extension. If this process of removing the EU's corresponding to single nonzero entries in rows or columns is continued, one of two matrix types must eventually be encountered:

1. A matrix with only one nonzero entry, or
2. A matrix with more than one nonzero entry in every row and every column.

We will argue that encountering type 1 is sufficient for the original two characters to be compatible, and that encountering type 2 is sufficient for the original characters to be incompatible. Thus, the simple procedure of successively removing rows or columns with single nonzero entries from the original matrix of two characters constitutes an algorithm for checking their compatibility, since matrix types 1 and 2 exhaust all cases and are exclusive.

If a matrix has only one nonzero entry, then clearly the data graph of those characters has no circuits. Hence any pair of characters that is a compatible extension of a pair with such a matrix, is compatible.

Suppose that the matrix is of type 2. Then, starting at any entry that is a 1 proceed first to an adjacent 1 in the same row, then staying in this new column proceed to an adjacent 1. Now, staying in this new row, proceed to an adjacent 1. This process may continue because every row and column has at least two 1's. Since there are only a finite number of 1's, this process will ultimately revist a 1 thus forming a circuit in the data graph. Hence the original characters, being a compatible extension of incompatible characters, are not compatible.

A computer program in FORTRAN implementing this algorithm is available from the authors.

References

Boulter, D., Ramshaw, J. A. M., Thompson, E. W., Richardson, M.: A Phylogeny of higher plants based on the amino acid sequences of cytochrome *c* and its biological implications. Proc. R. Soc. Lond. *B 181*, 441—455 (1972).

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

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

Estabrook, G. F., Johnson, C. S., jr., McMorris, F. R.: An algebraic analysis of cladistic characters. Discrete Math., in press (1976 a).

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

Estabrook, G. F., Landrum, L.: A simple test for the possible simultaneous evolutionary divergence of two amino acid positions. Taxon *25*, 609—613 (1975).

Fitch, W. M.: Toward finding the tree of maximum parsimony. In: The Eighth International Conference on Numerical Taxonomy (Estabrook, G. F., ed.), pp. 189—220. San Francisco: W. H. Freeman and Company 1975.

Fitch, W. M., Margoliash, E.: Construction of phylogenetic trees. Science *155*, 279—284 (1967).

McMorris, F. R.: Compatibility criteria for cladistic and qualitative taxonomic character. In: The Eighth International Conference on Numerical Taxonomy (Estabrook, G. F., ed.), pp. 399—415. San Francisco: W. H. Freeman and Company 1975.

Moore, G. W., Barnabas, J., Goodman, M.: A method for constructing maximum parsimony ancestral amino acid sequences on a given network. J. Theor. Biol. *38*, 459—485 (1973).

Roberts, F.: Discrete Mathematical Models. Englewood Cliffs, N. J.: Prentice-Hall 1976.

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

Prof. G. F. Estabrook
Department of Botany
University of Michigan
Ann Arbor, MI 48109, U.S.A.

Dr. F. R. McMorris
Department of Mathematics
Bowling Green State University
Bowling Green, OH 43403, U.S.A.