# A Mathematical Foundation for the Analysis of Cladistic Character Compatibility

G. F. ESTABROOK
*Department of Botany, University of Michigan, Ann Arbor, Michigan 48104*

C. S. JOHNSON, JR.

AND

F. R. McMORRIS
*Department of Mathematics, Bowling Green State University, Bowling Green, Ohio 43403*

Communicated by Stan Ulam

ABSTRACT

Using formal algebraic definitions of "cladistic character" and "character compatibility", the concept of "binary factors of a cladistic character" is formalized and used to describe and justify an algorithm for checking the compatibility of a set of characters. The algorithm lends itself to the selection of maximal compatible subsets when compatibility fails.

## 1. INTRODUCTION

The basic concept of the compatibility of taxonomic characters has been an integral part of the method of phylogenetic systematists ever since Bentham and Hooker published their milestone work ca. 1870. It was certainly inherent in the concepts exposed by Hennig when he constructed a rational methodological basis for phylogenetic systematics in 1956, and preserved in the more modern logical interpretation of this work by Farris, Kluge and Eckardt [6]. Its first appearance as an isolated concept is perhaps with the work of Wilson [9]. The coinage "character compatibility" that we have adopted for use in this exposition has precedent in the work of Camin and Sokal [2], where the concept originated independently of Wilson. An explicit operational procedure for determing the compatibility of a specified pair of taxonomic characters structured to encode determinations of similar and different for a specified sample study of operational evolutionary units

(EU's) was described by LeQuesne [8]. LeQuesne's procedure is applicable as described to pairs of two state and linearly ordered three state qualitative characters. At the Third Annual International Conference on Numerical Taxonomy at Stony Brook, New York, Farris described an ingenious procedure for representing any cladistic character (the character state trees of Camin and Sokal) by several two state characters, thereby apparently extending the applicability of the LeQuesne criterion to the general cladistic character. A formal algebraic concept of true cladistic character was recently offered by Estabrook, Johnson and McMorris [3], wherein a true cladistic character was characterized as a lower semilattice homomorphism from the true evolutionary tree to the character state tree. This concept permitted a formal algebraic definition of character compatibility (Estabrook et al. [4]) that purports to capture the concept employed by working taxonomists.

Various of the works cited above offer more or less formal definitions of cladistic character and character compatibility. The basic concepts are easy to visualize. A cladistic character is the partial estimate of evolutionary relationships suggested in the natural way by the discrete observable expressions (character states) of a single (presumed) homologous correspondence (character) among the EU's under study. If this partial estimate were historically true, such a cladistic character could have evolved its states without parallel evolution and without reversing evolutionary trends. A cladistic character that actually did evolve with no parallelisms or reversals is called divergent by many biologists. (In the terminology which follows, this would be called true on $S'$.) Two cladistic characters are mutually compatible if there exists an estimate (not necessarily true) of evolutionary history that, if it were true, would permit the logical possibility that both characters are divergent.

A natural question is: "Does the pairwise possible divergence of three cladistic characters guarantee, as a logical possibility, the existence of an (estimated) evolutionary tree with respect to which all three could be simultaneously divergent?", or more generally: "Does the pairwise compatibility of an arbitrary finite collection of cladistic characters ensure, as a logical possibility, their simultaneous divergence?". It has been apparent for years that the answer to this question is "yes". In a somewhat more general form, it has recently been proved (Estabrook et al. [4]) as a theorem. We remark here that currently the analogous concept of character state graph applicable to the problem of estimation evolutionary relationships from protein sequence data is under study. At the 1974 Classification Society meeting in Ann Arbor, Michigan, Fitch presented some results indicating that the above theorem is not true for character state graphs. Estabrook and Landrum [5] suggest a simple operational procedure for determining the pairwise compatibility of character state graphs.

In this paper we formalize the concept of the binary factors of a cladistic character and prove that it is mathematically correct to construe them as such. After confirming the obvious result that the binary factors of a cladistic character are mutually compatible, we show that two cladistic characters are compatible if and only if each binary factor of one is compatible with every binary factor of the other. The definition is constructive in the sense that a computer algorithm can easily be transcribed directly from the form of the statement. Having thus represented all cladistic characters by their binary factors, the theorems ensure that the operational procedure of LeQuesne can be confidently implemented to effect the compatibility analysis, suggested by Farris, of an arbitrary collection of cladistic characters.

## 2. DEFINITIONS AND RESULTS

We begin by recalling some of the relevant definitions and results from our earlier papers [3, 4]. We suppose throughout that all sets are finite.

*DEFINITION 1*

A *tree poset* is a partially ordered set having the property that $a \leqslant c$ and $b \leqslant c$ together imply $a \leqslant b$ or $b \leqslant a$. A *tree semilattice* is a tree poset in which any two elements $a$ and $b$ have a greatest lower bound, denoted $a \wedge b$.

In what follows, $S$ will denote a fixed set of EU's under study, $S'$ will represent the (unknown) true evolutionary history of $S$, and $S^*$ will represent an estimate of $S'$. By taking $x \leqslant y$ to mean "$x$ is an ancestor of $y$", we view $S$ as a tree poset, and $S'$ and $S^*$ as tree semilattices in which $x \wedge y$ is the most recent common ancestor of $x$ and $y$. In most cases $S$ would consist of the maximal elements of $S'$ (contemporary EU's) and have no comparabilities.

*DEFINITION 2*

A *cladistic character on* $S$ is a map $K: S \to P$, and a *cladistic character on* $S^*$ is an onto map $K: S^* \to P$, where $P$ is a tree semilattice (the *character state tree*). $0 \equiv \wedge P$.

By Theorem 2 of [4] we may restrict our attention to characters which are order-preserving [$a \leqslant b$ implies $K(a) \leqslant K(b)$]. The following definition is discussed in [3]. It corresponds loosely to what many biologists would call divergent.

*DEFINITION 3*

A cladistic character $K: S' \to P$ is *true* if and only if it satisfies the following three conditions for all $a, b \in S'$.

(i) $\bar{a} \in K^{-1}(K(a))$, where $\bar{a} = \bigwedge K^{-1}(K(a))$.
(ii) $a \leqslant b$ implies $K(a) \leqslant K(b)$.
(iii) $K(a) \leqslant K(b)$ implies $\bar{a} \leqslant \bar{b}$.

*LEMMA 1 (SEE [3])*

*A cladistic character is true if and only if it is a semilattice homomorphism* $(K(a \wedge b) = K(a) \wedge K(b)$ *for all* $a, b \in S')$.

*DEFINITION 4*

A set of cladistic characters $K_i : S \to P_i$, $i = 1, \ldots, n$, is *compatible* if there exists a tree semilattice $S^*$ extending $S$ ($S \subseteq S^*$ and $x \leqslant y$ in $S$ implies $x \leqslant y$ in $S^*$) such that each $K_i$ can be extended to a cladistic character $K_i^* : S^* \to P_i$ that would be true if $S^* = S'$.

Notice that compatibility guarantees nothing except the existence of a logically possible estimate $S^*$, whereas incompatibility guarantees that the characters, as presently structured, cannot all be true on $S'$. An "internal" test for compatibility is given by Theorem 3 of [4]. Its application is made easier, computationally, by the pairwise criterion (Theorem 4 of [4]) mentioned in the introduction, and easier still by use of the pairwise criterion in conjunction with binary factorization, which we describe next.

*DEFINITION 5*

A *binary character* is a cladistic character whose character state tree has two elements.

Recall that if $K_1 : S \to P_1$ and $K_2 : S \to P_2$ are two characters, we define $K_1 \times K_2 : S \to P_1 \times P_2$ by $(K_1 \times K_2)(x) = (K_1(x), K_2(x))$. Notice that $K_1 \times K_2$ is not in general a character, since $P_1 \times P_2$ is not a tree unless either $P_1$ or $P_2$ has only one element. In particular, when $K_1$ and $K_2$ are binary, $P_1 \times P_2$ looks like

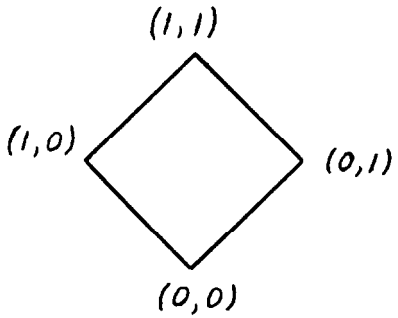


FIG. 1.

Assuming, as we do, that all characters on $S$ are order-preserving, the special case of Theorem 3 of [4] of interest to us here is as follows.

*THEOREM 1*

*Two binary characters $K_1: S \to P_1$ and $K_2: S \to P_2$ are compatible if and only if the image of $K_1 \times K_2$ in $P_1 \times P_2$ is a tree poset.*

We caution the reader that this theorem may fail when $K_1$ and $K_2$ are not binary.

*DEFINITION 6*

Let $K: S \to P$ be a cladistic character, and let $p_1, \ldots, p_n$ be the nonzero elements of $P$. For each $i = 1, \ldots, n$, let $T_i = \{0, 1\}$, and define $K_i: S \to T_i$ by $K_i(x) = 1$ if $K(x) \geqslant p_i$ and $K_i(x) = 0$ otherwise. The binary characters $K_i$ are called the *binary factors* of $K$, and we write $K = K_1 \otimes \cdots \otimes K_n$.

The following fact is not of interest computationally, but it describes the sense in which the $K_i$ constitute a factorization of $K$.

*LEMMA 2*

*Let $K: S \to P$ be a character with binary factors $K_1, \ldots, K_n$, and let $K_1 \times \cdots \times K_n: S \to T_1 \times \cdots \times T_n$ be defined as in the $n = 1$ case above. Then there is a (semilattice) embedding $\phi$ of $P$ in $T_1 \times \cdots \times T_n$ such that $K_1 \times \cdots \times K_n = \phi \circ K$.*

*Proof.* Define $\phi: P \to T_1 \times \cdots \times T_n$ by $\phi(p) = (a_1, \ldots, a_n)$ with $a_i = 1$ if and only if $p \geqslant p_i$. It is clear that $K_1 \times \cdots \times K_n = \phi \circ K$ and that $\phi$ is one-to-one, so we must show that $\phi$ is a semilattice homomorphism. For $p, q \in P$ set $\phi(p \wedge q) = (a_1, \ldots, a_n)$, $\phi(p) = (b_1, \ldots, b_n)$, and $\phi(q) = (c_1, \ldots, c_n)$. Then $a_i = 1$ is equivalent to $p \wedge q \geqslant p_i$, which is equivalent to $p \geqslant p_i$ and $q \geqslant p_i$, which is equivalent to $b_i = 1$ and $c_i = 1$, which is equivalent to $b_i \wedge c_i = 1$. Hence $\phi(p \wedge q) = \phi(p) \wedge \phi(q)$.

The next fact is intuitively clear, but we will need to make use of it.

*LEMMA 3*

*Let $K = K_1 \otimes \cdots \otimes K_n$ be a character on $S$. Then $K_i$ is compatible with $K_j$ for each $i$ and $j$.*

*Proof.* Assume that, for some $i$ and $j$, $K_i$ is not compatible with $K_j$. Using Theorem 1 and the diagram for $P_1 \times P_2$ which preceeds it, we may suppose that we have $x, y, z \in S$ with $(K_i(x), K_j(x)) = (1, 1)$, $(K_i(y), K_j(y)) = (0, 1)$, and $(K_i(z), K_j(z)) = (1, 0)$. Since $p_i, p_j \leqslant K(x)$ and $P$ is a tree, we must have $p_i \leqslant p_j$ or $p_j \leqslant p_i$, but $p_i \leqslant p_j \leqslant K(y)$ contradicts $K_i(y) = 0$ and $p_j \leqslant p_i \leqslant K(z)$ contradicts $K_j(z) = 0$.

We now show that compatibility can be checked working only with binary factors, to which Theorem 1 can be applied.

*THEOREM 2*

*Let* $K = K_1 \otimes \cdots \otimes K_n$ *and* $L = L_1 \otimes \cdots \otimes L_m$ *be characters on* $S$. *Then* $K$ *and* $L$ *are compatible if and only if* $K_i$ *is compatible with* $L_j$ *for* $i = 1, \ldots, n$ *and* $j = 1, \ldots, m$.

*Proof.* Suppose that $K: S \to P$ and $L: S \to Q$ are compatible. Let $K^*: S^* \to P$ and $L^*: S^* \to Q$ be the extensions of Definition 4, let $\phi_K: P \to T_1 \times \cdots \times T_n$ and $\phi_L: Q \to T_1 \times \cdots \times T_m$ be the maps of Lemma 2, and let $\rho_i: T_1 \times \cdots \times T_k \to T_i$ be the projection map $\rho_i(x_1, \ldots, x_k) = x_i$. All maps involved are homomorphisms, and we define homomorphisms $K_i^*: S^* \to T_i$ and $L_j^*: S^* \to T_j$ by $K_i^* = \rho_i \circ \phi_K \circ K^*$ and $L_j^* = \rho_j \circ \phi_L \circ L^*$. These extend $K_i$ and $L_j$, since $K_i = \rho_i \circ \phi_K \circ K$ and $L_j = \rho_j \circ \phi_L \circ L$, and are onto, since $K^*$ and $L^*$ are. Thus $K_i$ and $L_j$ are compatible.

Suppose, conversely, that each $K_i$ is compatible with each $L_j$. The set of characters $K_1, \ldots, K_n, L_1, \ldots, L_m$ is pairwise compatible by Lemma 3 and hence compatible by Theorem 4 of [4]. Thus there is an extension $S^*$ of $S$ and extensions $K_i^*, L_j^*$ of the $K_i, L_j$ to $S^*$. Now it is an easy exercise to show that if $B$ is a subsemilattice of a tree semilattice $A$, then $f: A \to B$ given by $f(a) = \max\{b \in B \mid b \leqslant a\}$ is a semilattice homomorphism which leaves elements of $B$ fixed. Let $f_K$ be the homomorphism obtained in this manner when $A = \mathrm{Im}(K_1^* \times \cdots \times K_n^*)$, which is a tree semilattice by Lemma 1 of [4], and $B = \langle \mathrm{Im}(K_1 \times \cdots \times K_n) \rangle$, the semilattice generated by $\mathrm{Im}(K_1 \times \cdots \times K_n)$ in $T_1 \times \cdots \times T_n$. Since $\langle \mathrm{Im}(K_1 \times \cdots \times K_n) \rangle \subseteq \phi_K(P)$, we may define a semilattice homomorphism $K^*: S^* \to P$ by $K^*(x) = \phi_K^{-1}(f_K(K_1^* \times \cdots \times K_n^*(x)))$. We define $L^*: S^* \to Q$ in a similar manner, and it is clear that $K^*$ and $L^*$ extend $K$ and $L$. If $K^*$ and $L^*$ are onto, we are done. If not, we observe that $\mathrm{Im}(K^* \times L^*)$ is a tree subsemilattice of $P \times Q$ (Lemma 1 of [4] again), and hence $K^*$ and $L^*$ are compatible by Theorem 3 of [4], making $K$ and $L$ compatible as well.

## CONCLUSION

To test a set of characters for compatibility, one may take them two at a time, compute their binary factors using Definition 6, and apply Theorem 1 to pairs of binary factors. This procedure is justified by Theorem 4 of [4] and by Theorem 2. If a set of characters proves to be incompatible and it is felt that maximal compatible subsets are most likely to be historically correct, then the data from the pairwise checking can be compiled in a "compatibility table" which can be used to select maximal compatible subsets.

*Johnson and McMorris thank the Faculty Research Committee of Bowling Green State University for grants supporting this work.*

## REFERENCES

1 G. Bentham and J. D. Hooker, *Genera Plantarum* (1862–1883), London.
2 J. H. Camin and R. R. Sokal, A method for deducing branching sequences in phylogeny, *Evolution* **19**, 311–326 (1965).
3 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).
4 G. F. Estabrook, C. S. Johnson, Jr., and F. R. McMorris, An algebraic analysis of cladistic characters, to be published.
5 G. F. Estabrook and L. Landrum, A simple test for the possible simultaneous evolutionary divergence of two amino acid positions, *Taxon*, **24**, 53–57 (1975).
6 J. S. Farris, A. G. Kluge and M. J. Eckardt, A numerical approach to phylogenetic systematics, *Syst. Zool.* **19**, 172–189 (1970).
7 W. Hennig, Systematik und Phylogenese, *Ber. Hundertj. dtsch. ent. Ges.*, 1956.
8 W. J. LeQuesne, A method of selection of characters in numerical taxonomy, *Syst. Zool.* **18**, 201–205 (1969).
9 E. O. Wilson, A consistency test for phylogenies based on contemporaneous species, *Syst. Zool.* **14**, 214–220 (1965).