Discrete Mathematics 16 (1976) 141-147
© North-Holland Publishing Company

AN ALGEBRAIC ANALYSIS OF CLADISTIC CHARACTERS

G.F. ESTABROOK
Department of Botany, University of Michigan, Ann Arbor, Michigan 48104, U.S.A.

C.S. JOHNSON, Jr. and F.R. McMORRIS
Department of Mathematics, Bowling Green State University, Bowling Green, Ohio 43403, U.S.A.

Received 23 September 1974

Cladistic characters are used by many numerical taxonomists in the estimation of evolutionary history. We make use of semilattice theory to give an algebraic formulation of the ideas involved in this process and to give rigorous proofs of theorems which justify certain operational procedures in current use. In particular, we discuss certain compatibility tests for a collection of characters.

1. Introduction

The problem of estimating the evolutionary history of a set S of evolutionary units has challenged biologists since the time of Darwin. If one views S as a partially ordered set by taking $x \leq y$ to mean "x is an ancestor of y", then the evolutionary history of S may be viewed as a partially ordered set S' containing S. In the large majority of cases it is considered reasonable to assume that S' is a finite tree lower semilattice in which the greatest lower bound $x \wedge y$ is the most recent common ancestor of x and y. In practice S' is, of course, unknown and in attempting to estimate it, working comparative biologists make use of "cladistic characters" on S. A cladistic character is, in mathematical terms, an equivalence relation on S together with a partial ordering of the equivalence classes intended to represent evolutionary relationships among the "character states". The structuring of cladistic characters on a given study collection S is by no means an exact procedure, being subject to many possible errors in judgment by the biologist. It may happen that several characters on the same S turn out to be "incompatible", either intuitively or in subtler ways, in which case one or more of the characters may be restructured or thrown out. This structuring process is highly intuitive, it having been noticed years ago and accepted as reasonable, for example, that several characters are compatible if they are pairwise compatible. In the present paper we give algebraic formulations of all the above notions and rigorous proofs of theorems which give mathematical justification for (1) the exclusion of certain characters (the non-isotone ones), (2) a relatively simple compatibility test, and (3) the practice of inferring compatibility from pairwise compatibility. It is our feeling that these definitions and results are an accurate reflection of current practice in this field. For further biological background and motivation we refer the reader to references [1-4].

2. Definitions and results

We suppose throughout that all sets are finite. EU's are evolutionary units.

Definition 2.1. A *tree poset* is a partially ordered set having the property that $a \leq c$ and $b \leq c$ together imply $a \leq b$ or $b \leq a$. 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 and S^* will represent an estimate of S', the (unknown) true evolutionary history of S. By taking $x \leq y$ to mean "x is an ancestor of y" we view S as a tree poset, S' and S^* as tree semilattices in which $x \wedge y$ is the most recent common ancestor of x and y.

Definition 2.2. A *cladistic character on* S is a map $K: S \rightarrow P$ and a *cladistic character on* S^* is an onto map $K: S^* \rightarrow P$ where P is a tree semilattice (the *character state tree*).

Definition 2.3. Let S^* be a tree semilattice. A cladistic character $K: S^* \rightarrow P$ is *true* if and only if K satisfies the following three conditions for $a, b \in S^*$:

(i) $\bar{a} \in K^{-1}(K(a))$ where $\bar{a} = \wedge K^{-1}(K(a))$

(ii) $a \leq b$ implies $K(a) \leq K(b)$

(iii) $K(a) \leq K(b)$ implies $\bar{a} \leq \bar{b}$.

Definitions 2 and 3 are discussed in some detail in [3] and we will simply "translate" Definition 3 here. Condition (i) asserts that a character state must contain the most recent common ancestor of the EU's belonging to it. Part (ii) requires that if one EU x is an ancestor to another EU y, then the state to which x belongs is ancestral, in the character state tree of K, to the state to which y belongs. Finally (iii) says that if one character state is ancestral to another in the character state tree, then the most recent common ancestor in the one state is ancestral to the most recent common ancestor in the other state.

The proof of the following theorem can be found in [3].

Theorem 2.1. *A cladistic character is true if and only if it is a semilattice homomorphism.*

Theorem 2.2 provides a quick check to determine whether a cladistic character could possibly be true on the historically correct S'. We find that cladistic characters which reverse the evolutionary directions evidenced in S may be excluded from consideration.

Theorem 2.2. *Let S be a tree poset, P a tree semilattice, and K a map $K: S \rightarrow P$. Then there exists a tree semilattice S^* extending S ($S \subseteq S^*$ and $x \leq y$ in S implies $x \leq y$ in S^*) and an extension of K to a true cladistic character on S^* if and only if $a \leq b$ implies $K(a) \leq K(b)$ for all $a, b \in S$ (i.e., K is isotone).*

Proof. Necessity is obvious from Theorem 2.1.

To prove sufficiency, we first enlarge the relation $\leq$ partially ordering S so that $K^{-1}(K(z))$ is a chain (any two elements are comparable) for each $z \in S$. We may do this because of the fact (see [5]) that every poset can be embedded in a chain. Letting $\leqslant$ denote this enlarged relation, we note that $\leqslant$ is reflexive and antisymmetric on S, transitive on each $K^{-1}(K(z))$, but not necessarily transitive on all of S.

Now let S^* be the disjoint union of S and P, and extend K to S^* by defining $K(x) = x$ for all $x \in P$. We now have K an onto map $K: S^* \rightarrow P$. For each $x \in P$ define $x \leqslant y$ for all $y \in K^{-1}(K(x))$, so $\leqslant$ is now a reflexive relation on S^* extending the original partial order on S and having the property that each $K^{-1}(K(z))$ is a chain. Notice also that $x \leqslant y$ in S^* implies $K(x) \leq K(y)$, since either $K(x) = K(y)$ or $x \leq y$ in the original partial order on S giving $K(x) \leq K(y)$ by hypothesis. Define a relation $\preceq$ on S^* by $x \preceq y$ if and only if either $K(x) = K(y)$ and $x \leqslant y$, or $K(x) < K(y)$. We claim that $(S^*, \preceq)$ is a tree semilattice extending S on which the extension of K is a true cladistic character. It is clear that $x \leq y$ implies $x \preceq y$, that $x \preceq y$ implies $K(x) \leq K(y)$, and that $\preceq$ is reflexive and antisymmetric. To see transitivity, suppose $x \preceq y \preceq z$. If $K(x) = K(y) = K(z)$, we have $x \preceq z$ by transitivity of $\leqslant$ in $K^{-1}(K(x))$. If $K(x) = K(y) < K(z)$, $K(x) < K(y) = K(z)$, or $K(x) < K(y) < K(z)$, we have $x \preceq z$ since $K(x) < K(z)$. To show that $\preceq$ satisfies the tree condition, assume $x, y \preceq z$. This implies that $K(x), K(y) \leq K(z)$. Since P is a tree we have $K(x) \leq K(y)$ or $K(y) \leq K(x)$. If $K(x) < K(y)$ or $K(y) < K(x)$, we have $x \preceq y$ or $y \preceq x$ by definition, and if $K(x) = K(y)$ we have $x \leqslant y$ or $y \leqslant x$ in $K^{-1}(K(x))$ since $K^{-1}(K(x))$ is a chain. Thus $\preceq$ is a tree partial order on S^*. It is easy to show that any two elements $x, y \in S^*$ have a lower bound, namely, the smallest element in $K^{-1}(K(x) \wedge K(y))$. Hence $(S^*, \preceq)$ is a tree semilattice. The fact that $K: S^* \rightarrow P$ satisfies Definition 2.3 is immediate and the proof is complete.

Lemma 2.1. *Let $f: A \rightarrow B$ be a homomorphism from the tree semilattice A into the semilattice B. Then $\mathrm{Im}(f)$ is a tree subsemilattice of B.*

Proof. Suppose $f(x), f(y) \leq f(z)$ for $x, y, z \in A$. Since $x \wedge z$, $y \wedge z \leq z$, we have $x \wedge z \leq y \wedge z$ or $y \wedge z \leq x \wedge z$. Thus $f(x \wedge z) \leq f(y \wedge z)$ or $f(y \wedge z) \leq f(x \wedge z)$, from which it follows that $f(x) \leq f(y)$ or $f(y) \leq f(x)$.

Notation. Let $P_1, \ldots, P_n$ be tree semilattices and

$$\mathscr{P} = \prod_{i=1}^{n} P_i = \{(p_1, \ldots, p_n) \colon p_i \in P_i\}.$$

$\mathscr{P}$ is a semilattice (but in general not a tree) with respect to the partial order $(p_1, \ldots, p_n) \leq (q_1, \ldots, q_n)$ iff $p_i \leq q_i$ for $i = 1, \ldots, n$. The meet operation in $\mathscr{P}$ is $(p_1, \ldots, p_n) \wedge (q_1, \ldots, q_n) = (p_1 \wedge q_1, \ldots, p_n \wedge q_n)$. We let $\rho_i \colon \mathscr{P} \to P_i$, $i = 1, \ldots, n$ be the projection onto P_i. That is, $\rho_i(p_1, \ldots, p_n) = p_i$. Each ρ_i is clearly a semilattice homomorphism. We will use this notation for the remainder of this paper.

Lemma 2.2. *Let T be a tree subsemilattice of $\mathscr{P}$. Then there exists a tree subsemilattice T^* of $\mathscr{P}$ such that $T \subseteq T^*$ and $\rho_i(T^*) = P_i$ for $i = 1, \ldots, n$.*

Proof. We prove this lemma by showing that the elements necessary to make each $\rho_i \colon T \to P_i$ onto can be adjoined to T one by one, resulting in a tree subsemilattice at each step. Since all sets are finite, a finite number of such steps will suffice. In the proof we let T represent the tree subsemilattice resulting from step k $(k \geq 0)$ and show how to carry out step $k+1$.

Without loss of generality we assume $\rho_1(T)$ is a proper subset of P_1. Let $e_1 \in P_1 \setminus \rho_1(T)$.

Case 1. Assume that there exists $a_1 \in \rho_1(T)$ such that $e_1 \leq a_1$. We may assume that e_1 is covered by a_1 in P_1, choosing a larger e_1 and smaller a_1 if necessary.

Since $a_1 \in \rho_1(T)$, there exists $x = (a_1, x_2, \ldots, x_n) \in T$. Let $a = \wedge\{x \in T \colon \rho_1(x) = a_1\}$. Thus $a = (a_1, a_2, \ldots, a_n) \in T$ and $(a_1, x_2, \ldots, x_n) \in T$ implies $a_i \leq x_i$ for $i = 2, \ldots, n$. Now let $T^* = T \cup \{e\}$ where $e = (e_1, a_2, \ldots, a_n)$.

We first show that T^* is a tree poset. There are two cases that must be considered. Suppose $c, d \leq e$ where $c, d \in T$. Since $e \leq a$ and T is a tree, we have $c \leq d$ or $d \leq c$. The other case is when $e, d \leq c$ for $d, c \in T$. Now $e \leq c$ implies $(e_1, a_2, \ldots, a_n) \leq (c_1, \ldots, c_n)$ and thus in P_1 we have $e_1 \leq a_1 \wedge c_1 \leq a_1$. Since a_1 covers e_1 we must have $e_1 = a_1 \wedge c_1$, contradicting the fact that $e_1 \notin \rho_1(T)$, or $a_1 \wedge c_1 = a_1$. This yields $a \leq c$ so that $a, d < c$, and thus $a \leq d$ or $d \leq a$. If $a \leq d$ we have $e \leq d$. If $d \leq a$ we have $d_1 \leq a_1$ in P_1. But $e_1 < a_1$ in P_1 also, so that $d_1 \leq e_1$, giving $d \leq e$, or $e_1 < d_1$. If $e_1 < d_1$ then $e_1 \leq a_1 \wedge d_1 \leq a_1$, and since a_1 covers e_1 we have $e_1 = a_1 \wedge d_1$, a contradiction to $e_1 \notin \rho_1(T)$, or $a_1 \wedge d_1 = a_1$. Now $a_1 \wedge d_1 = a_1$ gives $a_1 \leq d_1$ and hence $a_1 = d_1$. Since $d \in T$ and $\rho_1(d) = a_1$, we have $a \leq d$ and hence $e \leq a = d$.

We now must show that T^* is a subsemilattice of $\mathscr{P}$. Let $b = (b_1, \ldots, b_n) \in T$. Then we claim that $b \wedge e$ is equal to e or $b \wedge a$, both elements of T^*. Since $e \leq a$, we have $b \wedge e \leq b \wedge a$. Now $e \leq a$ and $b \wedge a \leq a$ imply $e \leq b \wedge a$ or $b \wedge a \leq e$, since T^* is a tree poset. From $e \leq b \wedge a \leq b$ we have $e \wedge b = e$ and from $b \wedge a \leq e$ we have $b \wedge a \leq e \wedge b$, giving $e \wedge b = b \wedge a$.

Case 2. Assume that there does not exist an element $a_1 \in \rho_1(T)$ such that $e_1 \leq a_1$. Since we may assume $O \in \rho_1(T)$ (if it is not, add $(O, \ldots, O)$ to T where O denotes the least element of P_i), there exists $a_1 \in \rho_1(T)$ such that $a_1 < e_1$. We may assume that a_1 is covered by e_1. Let $a = (a_1, a_2, \ldots, a_n)$ be an element of T such that $\rho_1(a) = a_1$, and let $T^* = T \cup \{e\}$ where $e = (e_1, a_2, \ldots, a_n)$.

We assert that T^* is a tree. As before we have two cases. If $d, e \leq c$ where $d, c \in T$, then $e_1 \leq c_1$ with $c_1 \in \rho_1(T)$ which is impossible. Therefore we must check only the case when $c, d \in T$ and $c, d \leq e$. This gives $c_1 < e_1$, and since $a_1 < e_1$ we have $c_1 \leq a_1$ or $a_1 \leq c_1$. If $a_1 < c_1$, then $a_1 < c_1 < e_1$, contradicting the fact that e_1 covers a_1. Hence $c_1 \leq a_1$ thereby giving $c \leq a$. Similarly one can show $d \leq a$, and since T is a tree we have $d \leq c$ or $c \leq d$.

To show that T^* is a subsemilattice let $b = (b_1, \ldots, b_n) \in T$. We claim that $b \wedge e = b \wedge a$. Now $a \leq e$ and $b \wedge e \leq e$ give $a_1 \leq e_1$ and $b_1 \wedge e_1 \leq e_1$, implying that $a_1 \leq b_1 \wedge e_1$ or $b_1 \wedge e_1 \leq a_1$. If $b_1 \wedge e_1 \leq a_1$, then $b_1 \wedge e_1 \leq a_1 \wedge b_1$, from which it follows that $b \wedge e \leq b \wedge a$. This gives $b \wedge e = b \wedge a$. If $a_1 \leq b_1 \wedge e_1 \leq e_1$, then $a_1 = b_1 \wedge e_1$ or $b_1 \wedge e_1 = e_1$ since a_1 is covered by e_1. Now $a_1 = b_1 \wedge e_1 \leq b_1$ implies that $b \wedge a = b \wedge e$, and $b_1 \wedge e_1 = e_1$ implies that $e_1 \leq b_1$ where $b_1 \in \rho_1(T)$, which is impossible. The proof is now complete.

In Definition 2.4 we generalize the concept of compatible characters suggested by Camin and Sokal in [1]. These authors define two characters to be compatible if there exists an estimate of evolutionary history with respect to which both are true.

Definition 2.4. Let S be a tree poset and P_i a tree semilattice for $i = 1, \ldots, n$. A set of isotone maps $K_i : S \to P_i$, $i = 1, \ldots, n$, is *compatible* if there is a tree semilattice S^* extending S such that each K_i can be extended to a true cladistic character $K_i^* : S^* \to P_i$.

The following theorem gives a useful compatibility test.

Theorem 2.3. *Let S be a tree poset and P_i a tree semilattice for $i = 1, \ldots, n$. The isotone maps $K_i : S \to P_i$, $i = 1, \ldots, n$, are compatible if and only if $\langle \mathrm{Im}(K) \rangle$ is a tree subsemilattice of $\mathscr{P}$, where $K : S \to \mathscr{P}$ is defined by $K(x) = (K_1(x), \ldots, K_n(x))$ and $\langle \mathrm{Im}(K) \rangle$ denotes the subsemilattice of $\mathscr{P}$ generated by $\mathrm{Im}(K)$.*

Proof. Assume that the K_i's are compatible. Then there exists a tree semilattice S^* and semilattice homomorphisms $K_i^* : S^* \to P_i$, $i = 1, \ldots, n$. Now $K^* : S^* \to \mathscr{P}$ given by $K^*(x) = (K_1^*(x), \ldots, K_n^*(x))$ is a homomorphism and by Lemma 2.1, $\mathrm{Im}(K^*)$ is a tree semilattice. Since $\mathrm{Im}(K) \subseteq \mathrm{Im}(K^*)$, we are done.

For the converse, assume $\langle \mathrm{Im}(K) \rangle$ is a tree. Then by Lemma 2.2 there exists a tree semilattice P_0 extending $\langle \mathrm{Im}(K) \rangle$ such that $\rho_i(P_0) = P_i$ for $i = 1, \ldots, n$. From Theorem 2.2 there exists a tree semilattice S^* extending S and a homomorphism K^* from S^* onto P_0 extending K. For each i, let $K_i^* = \rho_i \circ K^*$. Thus K_i^* is an onto homomorphism from S^* to P_i extending K_i, which is what we wanted to show.

We now use the compatibility test in Theorem 2.3 to prove a fact which has been suspected for years — that one need only test pairs to determine the compatibility of an arbitrary set of cladistic characters.

Theorem 2.4. *Let S be a tree poset. The isotone maps $K_i : S \to P_i$, $i = 1, \ldots, n$, are compatible if and only if they are pairwise compatible.*

Proof. It is clear that compatible maps are pairwise compatible. Assume the K_i's are pairwise compatible. By Theorem 2.3 we must show that $\langle \mathrm{Im}(K) \rangle$ is a tree in $\mathscr{P}$, where K is as before. Suppose

$$K(x_1) \wedge \ldots \wedge K(x_m), K(y_1) \wedge \ldots \wedge K(y_n) \leq K(z_1) \wedge \ldots \wedge K(z_l).$$

Then

$$K_i(x_1) \wedge \ldots \wedge K_i(x_m), K_i(y_1) \wedge \ldots \wedge K_i(y_n) \leq K_i(z_1) \wedge \ldots \wedge K_i(z_l) \text{ for all } i.$$

Since each P_i is a tree we have that either

$$K_i(x_1) \wedge \ldots \wedge K_i(x_m) \leq K_i(y_1) \wedge \ldots \wedge K_i(y_n) \tag{1}$$

or

$$K_i(y_1) \wedge \ldots \wedge K_i(y_n) \leq K_i(x_1) \wedge \ldots \wedge K_i(x_m). \tag{2}$$

If (1) holds for all i or if (2) holds for all i, we are done. Otherwise there exist i and j such that

$$K_i(x_1) \wedge \ldots \wedge K_i(x_m) < K_i(y_1) \wedge \ldots \wedge K_i(y_n) \tag{3}$$

and

$$K_j(y_1) \wedge \ldots \wedge K_j(y_n) < K_j(x_1) \wedge \ldots \wedge K_j(x_m).$$

Since K_i and K_j are compatible, we have $\langle \mathrm{Im}(K_i \times K_j) \rangle$ a tree in $P_i \times P_j$ where $(K_i \times K_j)(x) = (K_i(x), K_j(x))$. Now

$$(K_i \times K_j)(x_1) \wedge \ldots \wedge (K_i \times K_j)(x_m), (K_i \times K_j)(y_1) \wedge \ldots \wedge (K_i \times K_j)(y_n)$$
$$\leq (K_i \times K_j)(z_1) \wedge \ldots \wedge (K_i \times K_j)(z_l)$$

implies that

$$(K_i \times K_j)(x_1) \wedge \ldots \wedge (K_i \times K_j)(x_m) \leq (K_i \times K_j)(y_1) \wedge \ldots \wedge (K_i \times K_j)(y_n)$$

or

$$(K_i \times K_j)(y_1) \wedge \ldots \wedge (K_i \times K_j)(y_n) \leq (K_i \times K_j)(x_1) \wedge \ldots \wedge (K_i \times K_j)(x_m).$$

In other words, $K_i(x_1) \wedge \ldots \wedge K_i(x_m) \leq K_i(y_1) \wedge \ldots \wedge K_i(y_n)$ and $K_j(x_1) \wedge \ldots \wedge K_j(x_m) \leq K_j(y_1) \wedge \ldots \wedge K_j(y_n)$, or $K_i(y_1) \wedge \ldots \wedge K_i(y_n) \leq K_i(x_1) \wedge \ldots \wedge K_i(x_m)$ and $K_j(y_1) \wedge \ldots \wedge K_j(y_n) \leq K_j(x_1) \wedge \ldots \wedge K_j(x_m)$. Either of these contradicts (3). Hence we must have (1) or (2) and the proof is complete.

References

[1] J.H. Camin and R.R. Sokal, A method for deducing branching sequences in phylogeny, Evolution 19 (1965) 311–326.

[2] G.F. Estabrook, Cladistic methodology: A discussion of the thepretical basis for the induction of evolutionary history, Ann. Rev. Ecol. Syst. 3 (1972) 427–456.

[3] G.F. Estabrook, C.S. Johnson, Jr. and F.R. McMorris, An idealized concept of the true cladistic character, Math. Biosciences 23 (1975) 263–272.

[4] W.J. LeQuesne, A method of selection of characters in numerical taxonomy, Syst. Zool. 18 (1969) 201–205.

[5] E. Szpilrajn, Sur l'extension de l'ordre partiel, Fund. Math. 16 (1930) 386–389.