From Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, E. M. Keramidas, Ed., 254–257, (1991).

Reconstruction of Evolutionary Trees from Pairwise Distributions on Current Species

Joseph T. Chang and John A. Hartigan
Department of Statistics
Yale University

Abstract

Suppose that the evolution of a character possessed by a number of current species is modelled as a Markov random field on an evolutionary tree. Suppose that for each pair of current species we know the joint probability distribution of the pair of characters possessed by that pair of species. We give conditions under which the evolutionary tree can be reconstructed from knowledge of these pairwise joint distributions, that is, conditions under which there is only one evolutionary tree topology consistent with the given pairwise distributions. In this way we establish consistency of a method for reconstructing evolutionary trees using pairwise distributions estimated from observed homologous DNA sequences.

1 Introduction

Evolutionary relationships among species are commonly conceptualized in terms of an "evolutionary tree." A tree consists of nodes and arcs. The *degree* of a node is the number of arcs incident to the node. Nodes of degree one are *terminal nodes*, and nodes of higher degree are *internal nodes.* In an evolutionary tree, the terminal nodes are labelled by current species observable today, and the internal nodes correspond to ancestral species. We assume speciation events occur at internal nodes, so that we do not allow nodes of degree two. The scientific problem of interest to us is to infer the evolutionary tree relating a given set of current species. This inference is to be based on data, which might typically be a set of observed DNA sequences, one from each of the given current species. For most of the paper, we will restrict our attention to the topology of the tree together with the labels of the terminal nodes. In particular, we are not interested in length of time, direction of time, or the root of the tree.

Let T denote a finite set of current species, let C denote a finite set of characters, and for each $t \in T$ let X_t denote the character possessed by species t. For example, C might be the set of four nucleotides, and X_t might identify the nucleotide occupying a particular site in the DNA of a representative of species t. We consider $\{X_t : t \in T\}$ to be random variables generated by a Markov random field model on an evolutionary tree. To describe the model, we begin with the tree $\mathcal{T} = (S, A)$, characterized by its set of nodes (or species) S and its set of arcs A. S may be decomposed into the union $S = T \cup N$ of the set T of terminal nodes and the set N of nonterminal nodes; since current species correspond to terminal nodes, there is no conflict with the notation T introduced above. Each arc $a \in A$ is undirected and may be represented as a subset $\{r, s\}$ containing two distinct nodes $r, s \in S$. For each $s \in S$ let X_s be a random variable taking values in C. We assume that $\{X_s : s \in S\}$ is a Markov random field on $\mathcal{T}$, which means that for each $s \in S$ the conditional distribution of X_s given all of the other values $\{X_r : r \neq s\}$ is the same as the conditional distribution of X_s given just the values $\{X_r : \{r, s\} \in A\}$ at the "neighbors" of s. This completes the description of the probabilistic model for the evolution of a single character. In general we observe n characters for each species. In this case, we make the standard but undoubtedly unrealistic assumption that distinct characters are independent and identically distributed (*iid*), that is, we imagine that $X^1, \ldots, X^n$ are *iid*, where each $X^i = \{X^i_s : s \in S\}$ is a Markov random field on $\mathcal{T}$.

The following brief remarks about methods of evolutionary tree reconstruction are intended to provide some context for this work; the excellent survey of Felsenstein (1988) should be consulted for more background and references. The most popular method is probably the *parsimony* method of Camin and Sokal (1965), which chooses a tree in which characters can be assigned to nodes so that the number of changes of character across arcs of the tree is minimal over all trees. This method has the very considerable virtue of ease of implementation. However, the unfortunate truth observed by Felsenstein (1978) is that parsimony is *inconsistent*; in fact, Felsenstein exhibited an example in which the probability that parsimony would choose an incorrect tree approached one as the number n of observed characters per species approached infinity. A maximum likelihood method for the present model was considered by Barry and Hartigan (1987b). This method overcomes parsimony's defect of inconsistency, at the cost of a great increase in computational difficulty. The distance method of Barry and Hartigan (1987a), which will be described more fully below, is intermediate between parsimony and maximum likelihood in terms of computational difficulty. The question that originated the present investigations was whether the distance method is consistent. This question will be addressed in section 3.

2 Identifiability of the Tree

A principal ingredient in the consistency proof is an "identifiability" result that says that under the assumptions of our Markov model and certain other conditions, distributions of pairs of the form (X_t, X_u), where t and u are terminal nodes, determine the evolutionary tree. This result in turn follows from Lemma 1 below, which says that knowing the values of an "additive function" on pairs of terminal nodes of a tree is enough to determine the tree.

The statement of Lemma 1 requires some definitions. Let $\mathcal{T}_1 = (S_1, A_1)$ and $\mathcal{T}_2 = (S_2, A_2)$ be two trees with $S_1 = T \cup N_1$ and $S_2 = T \cup N_2$, so that the terminal nodes of $\mathcal{T}_1$ and $\mathcal{T}_2$ are the same. We say that $\mathcal{T}_1$ and $\mathcal{T}_2$ are *equivalent* if there is a bijective "relabelling" function $\rho : S_1 \to S_2$ such that $\rho(t) = t$ for all $t \in T$ and $A_2 = \{\{\rho(r), \rho(s)\} : \{r, s\} \in A_1\}$. In this case we write $\mathcal{T}_1 \sim \mathcal{T}_2$. Thus, $\mathcal{T}_1 \sim \mathcal{T}_2$ means that $\mathcal{T}_1$ and $\mathcal{T}_2$ are the same up to a possible relabelling of nonterminal nodes.

Next, for a given tree $\mathcal{T} = (S, A)$, let $\vec{A} = \{(r, s) : \{r, s\} \in A\}$ be the set of *directed arcs* of $\mathcal{T}$. Then for all distinct $r, s \in S$ either $\pi(r, s) := \{(r, s)\} \subset \vec{A}$ or there is a unique $n \geq 1$ and a unique sequence $s_1, \ldots, s_n$ of distinct nodes such that $\pi(r, s) := \{(r, s_1), (s_1, s_2), \ldots, (s_n, s)\} \subset \vec{A}$. This defines $\pi(r, s)$, the *path from r to s*. We say that a function $f : S \times S \to \mathbb{R}$ is *additive* on the tree $\mathcal{T}$ if for all $r, s \in S$ we have

$$f(r, s) = \sum_{a \in \pi(r,s)} f(a),$$

with the sum being defined to be 0 if $r = s$.

Lemma 1 *Let $\mathcal{T}_1 = (S_1, A_1)$ and $\mathcal{T}_2 = (S_2, A_2)$ be two evolutionary trees with the same set of terminal nodes T. Suppose there exist functions $f_1 : S_1 \times S_1 \to \mathbb{R}$ and $f_2 : S_2 \times S_2 \to \mathbb{R}$ such that*

1. *$f_i(r, s) + f_i(s, r) \neq 0$ for all $\{r, s\} \in A$ and $i = 1, 2$*
2. *f_i is additive on $\mathcal{T}_i$ for $i = 1, 2$*
3. *$f_1(t, u) = f_2(t, u)$ for all $t, u \in T$.*

Then $\mathcal{T}_1 \sim \mathcal{T}_2$.

Results appearing in the papers of Dobson (1974) and Sattath and Tversky (1977) are clearly closely allied but apparently not the same. They focus on existence of trees satisfying certain conditions, while we are interested in uniqueness. We also find the fact that our additive function need not be nonnegative to be interesting.

The key to the identifiability result mentioned above is the notion of distance introduced by Barry and Hartigan (1987a), which takes the form

$$d(r, s) = -(1/4) \log[\det(P^{rs})]$$

for four-valued characters, where P^{rs} is the Markov transition matrix whose $(i, j)th$ entry is $P\{X_s = j | X_r = i\}$. For the Markov model we have assumed, d is an additive function. The identifiability result is then just the statement obtained by taking the additive function in Lemma 1 to be Barry and Hartigan's distance. The conditions required are

$$\det(P^{rs}) > 0 \quad \text{for} \quad \{r, s\} \in A \qquad (2.1)$$

and

$$\det(P^{rs})\det(P^{sr}) < 1 \quad \text{for} \quad \{r,s\} \in A. \qquad (2.2)$$

Under conditions (2.1) and (2.2), the pairwise distributions of characters at terminal nodes determine the evolutionary tree. Condition (2.2) corresponds to condition 1 of the lemma. To get an idea of its significance, note that an example of a situation it rules out is $P^{rs} = P^{sr} = I$ for two internal nodes r and s joined by an arc. This is reasonable, since in such a case we could eliminate the arc $\{r,s\}$ and combine nodes r and s into one node without changing any probability distributions at the terminal nodes of the tree. Condition (2.1) ensures that the logarithm in Barry and Hartigan's distance is defined. This would presumably hold in biologically realistic models, for example, models in which characters evolve as a Markov chain in continuous time. In any case, both conditions (2.1) and (2.2) may be relaxed by the device of using the distance $d^*(r,s) = -(1/4)\log|\det(P^{rs})|$ in place of d. The resulting conditions would be $\det(P^{rs}) \neq 0$ and $|\det(P^{rs})\det(P^{sr})| < 1$ for $\{r,s\} \in A$.

3 Consistency

The method Barry and Hartigan (1987a) propose for choosing an evolutionary tree from given data applies the least squares idea of Cavalli-Sforza and Edwards (1967) in the following manner. For each ordered pair (t,u) of terminal nodes, form the estimated distances

$$\hat{d}(t,u) = -(1/4)\log[\det(\hat{P}^{tu})], \qquad (3.1)$$

where $\hat{P}^{tu}$ is the usual empirical estimate of P^{tu}. For a candidate tree $\mathcal{T} = (S,A)$ under consideration, for each $(r,s) \in \vec{A}$ introduce a variable x_{rs}. Define the "departure from additivity" of the tree $\mathcal{T}$ to be the minimum of the quantity

$$\sum_{t,u\in T}\left(\hat{d}(t,u) - \sum_{(r,s)\in\pi(t,u)} x_{rs}\right)^2$$

over all possible values of the variables x_{rs}. Choose the tree having the smallest departure from additivity.

To state the consistency result, let $\mathcal{T}$ denote the true evolutionary tree, and as usual assume that $X^1, X^2, \ldots$ are *iid* Markov random fields on $\mathcal{T}$. Suppose that $\mathcal{T}^n$ denotes the estimate given by a tree reconstruction method when applied to the data $X^1, \ldots, X^n$. We say that the method is *strongly consistent* if with probability one there is a finite N such that $\mathcal{T}^n = \mathcal{T}$ for all $n \geq N$.

Theorem 2 *Suppose the true evolutionary tree $\mathcal{T} = (S,A)$ is bifurcating, that is, all internal nodes have degree 3. Then under conditions (2.1) and (2.2), the method of Barry and Hartigan (1987a) is strongly consistent.*

The assumption that the true tree is bifurcating rules out nodes of degree higher than 3. This restriction is necessary for the following reason. Lemma 1 states that different trees cannot have exactly the same distances between pairs of terminal nodes. However, if the true tree has nodes of degree higher than 3, then there are different trees that may have distances arbitrarily close to the true distances. If the true tree is bifurcating and the true model satisfies the conditions (2.1) and (2.2), then there are no such different trees.

4 No Information in Asymmetry

The distance d is asymmetric in general: $d(r,s) \neq d(s,r)$. Since the symmetrized distance function

$$\tilde{d}(r,s) = [d(r,s) + d(s,r)]/2$$

is also additive, one could work with $\tilde{d}$ rather than d, and effectively cut in half the number of equations and unknowns in each least squares calculation. However, replacing d by $\tilde{d}$ involves ignoring some of the information in the data, so that one might suspect that we would pay for the gain in computational simplicity by sacrificing efficiency. It turns out that as far as the method of Barry and Hartigan is concerned, no efficiency at all is lost by symmetrizing. The reason is contained in the following result.

Proposition 3 *Let a tree $\mathcal{T}$ having terminal nodes T be given. Suppose we are also given the estimated distances $\hat{d}(t,u)$ of (3.1) for all $t,u \in T$. For additive functions f on $\mathcal{T}$ define $S(f)$ and $\tilde{S}(f)$ to be*

$$\sum_{t,u\in T}[f(t,u) - \hat{d}(t,u)]^2$$

and

$$\sum_{t,u \in T} \left(\frac{f(t,u) + f(u,t)}{2} - \frac{\hat{d}(t,u) + \hat{d}(u,t)}{2} \right)^2,$$

respectively. Then we have

$$\inf_f S(f) = \inf_f \tilde{S}(f),$$

where the infima are taken over all functions f additive on $\mathcal{T}$.

5 Identifiability of the Full Model

Although section 2 showed that pairwise distributions over terminal nodes determine a tree, it is interesting that such pairwise distributions do not determine the full model, that is, they do not determine the Markov transition matrices P^{rs} for $\{r,s\} \in A$. This can be seen in the smallest nontrivial case: a tree having 3 terminal nodes $T = \{a,b,c\}$ and one nonterminal node $N = \{m\}$, say. Begin with an "original" model having marginal probability vector π^m at node m and Markov transition matrices P^{ma}, P^{mb}, and P^{mc}. These specifications determine the complete joint distribution of the Markov random field $\{X_a, X_b, X_c, X_m\}$, and in particular the Markov transition matrices P^{am}, P^{bm}, and P^{cm}. Let $\mathbf{1}$ denote a vector of ones and let a prime "$(')$" denote transpose. Then it turns out that if R is an invertible matrix satisfying the conditions

1. $R\mathbf{1} = \mathbf{1}$
2. $(R^{-1})'\pi^m R^{-1} = \mathrm{diag}(\tilde{\pi}^m)$ for some probability vector $\tilde{\pi}^m$
3. RP^{mi} and $P^{im}R^{-1}$ have nonnegative entries for $i = a,b,c$,

then the model having marginal probability vector $\tilde{\pi}^m$ at node m and Markov transition matrices $\tilde{P}^{mi} = RP^{mi}$ for $i = a,b,c$ has the same pairwise distributions over the terminal nodes as the original model. It is not difficult to find such examples; in fact, two characters are enough.

On the other hand, under conditions, the joint distribution of the values $\{X_t : t \in T\}$ at all of the terminal nodes is enough to determine the full model, at least for two characters, as the following result shows.

Proposition 4 *Let $\mathcal{T}_1 = (S_1, A_1)$ and $\mathcal{T}_2 = (S_2, A_2)$ be two evolutionary trees with the same set of terminal nodes T. For i=1 and 2, let $X_i = \{X_i(s) : s \in S_i\}$ be a Markov random field on $\mathcal{T}_i$ taking on two values. Suppose that conditions (2.1) and (2.2) hold with P^{rs}, P^{sr}, and A replaced by P_i^{rs}, P_i^{sr}, and A_i, respectively, for i=1,2. Suppose also that the joint distributions of $\{X^1(t) : t \in T\}$ and $\{X^2(t) : t \in T\}$ are the same. Then $\mathcal{T}_1 \sim \mathcal{T}_2$, and we have equality of the full joint distributions of $\{X_1(s) : s \in S_1\}$ and $\{X_2(\rho(s)) : s \in S_1\}$, where ρ is the "relabelling" function in the definition of the equivalence $\mathcal{T}_1 \sim \mathcal{T}_2$.*

We conjecture that a similar result holds for the case of more than two characters.

6 References

1. Barry, D. and Hartigan, J. A. (1987a). Asynchronous distance between homologous DNA sequences. *Biometrics* **43** 261–276.
2. Barry, D. and Hartigan, J. A. (1987b). Statistical analysis of hominoid molecular evolution. *Statistical Science* **2** 191–210.
3. Camin, J. H. and Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19** 311–326.
4. Cavalli-Sforza, L. L. and Edwards, A. W. (1967). Phylogenetic analysis: models and estimation procedures. *Evolution* **32** 550-570.
5. Dobson, A. J. (1974). Unrooted trees for numerical taxonomy. *Journal of Applied Probability* **11** 32–42.
6. Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27** 401–410.
7. Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.* **22** 521–565.
8. Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika* **42** 319–345.