Pergamon

Appl. Math. Lett. Vol. 7, No. 2, pp. 19–23, 1994

0893-9659/94 $6.00 + 0.00

0893-9659(94)E0005-V

Recovering a Tree from the Leaf Colourations It Generates under a Markov Model

M. STEEL
Mathematics Department, University of Canterbury
Christchurch, New Zealand

(Received November 1993; accepted December 1993)

Abstract—We describe a simple transformation that allows for the fast recovery of a tree from the probabilities such a tree induces on the colourations of its leaves under a simple Markov process (with unknown parameters). This generalizes earlier results by not requiring the transition matrices associated with the edges of the tree to be of a particular form, or to be related by some fixed rate matrix, and by not insisting on a particular distribution of colours at the root of the tree. Applications to taxonomy are outlined briefly in three corollaries.

Keywords—Trees, Sequence evolution, Phylogenetic invariants.

1. INTRODUCTION

A fundamental problem in molecular biology is how to use DNA and RNA sequence data to reconstruct evolutionary relationships between the species concerned. Such relationships are generally described by a rooted tree, whose leaves represent the extant species, and are labelled $1, \ldots, n$, and whose remaining vertices (representing ancestral species) are unlabelled and of degree at least 3, except, possibly, for the ancestor of the entire collection, which is regarded as a root vertex of the tree, and may have degree 2. Such trees are called rooted phylogenetic trees [1]. We will let ρ denote the root of T and let $T^{-\rho}$ denote the (phylogenetic) tree obtained from T as follows: if ρ has degree 2, delete ρ and identify its two incident edges, while if ρ has degree at least 3, simply regard ρ as an unlabelled (nondistinguished) vertex.

Suppose there is a set of colours (or states) and that the colour assigned to each vertex v is a random variable, denoted $\chi(v)$. In taxonomic applications, two colours might indicate the two "purine" bases, and the two "pyrimidine" bases; four colours the four nucleotide bases (A, C, G, T), and 20 colours the amino acids. We denote the probability that $\chi(\rho) = x$ by π_x. A simple model of nucleotide mutation assumes, roughly speaking, that starting from ρ these colours change randomly (and independently of changes on other edges) along the edges of T to give the present (observed) leaf colourations. More precisely, direct all the edges of T away from ρ, so that if e has ends v and w, and v lies between w and ρ, we write e as the ordered pair (v, w). Given an event E of the form $[\chi(v_1), \ldots, \chi(v_s)] = [\alpha_1, \ldots, \alpha_s]$, where $V_E := \{v_1, \ldots, v_s\}$ is a set of vertices of T, and an edge $e = (v, w)$, denote by $P_e[x \to y \mid E]$ the conditional probability that $\chi(w) = y$ given that $\chi(v) = x$, and given E. We make the following independence assumption, in which a "descendent" vertex of e is any vertex v for which the path from v to ρ contains e:

ASSUMPTION (A1). *$P_e[x \to y \mid E]$ does not depend on E if V_E does not include any descendent vertex of e.*

The author wishes to thank M.D. Hendy and A. Dress for several helpful comments, and the New Zealand Lotteries Commission for funding this research.

Thus, each edge $e = (v, w)$ of T has an associated transition matrix $M(e)$, with $M(e)_{rs} = P[r \to s \mid \phi]$. Thus each row of $M(e)$ sums to 1. The matrices $M(e)$ and the root values π_x induce (assuming (A1)) a well-defined probability for every possible colouration χ_0 of the leaves. By (A1), this probability is:

$$\sum_{\chi} \prod_{e=(v,w)} \pi_{\chi(\rho)} M(e)_{\chi(v)\chi(w)},$$

where χ ranges over all colouration of the vertices of T, which extend χ_0. Here, we address the taxonomically relevant inverse problem of finding $T^{-\rho}$ given *just* the leaf colouration probabilities (in taxonomic applications, these probabilities can be estimated directly from DNA or RNA sequence data). Note that even with two colours, and $M(e)$ symmetric for all e, it is not always possible to recover $T^{-\rho}$; indeed, if we set the off-diagonal entries of $M(e)$ to a common value x for all edges e, then if $x = 0.5$ or $x = 0$, *every* tree induces the same distribution on the set of leaf bicolourations. Notice that for these two choices of x, the determinant, $\det(M(e))$, of the matrix $M(e)$ equals 0 or 1, respectively. We show here that if, in addition to (A1), we make the following mild (and biologically reasonable) assumption:

ASSUMPTION (A2). $\det(M(e)) \neq 0, \pm 1$ *for all edges* e; $\pi_x \neq 0$, *for all colours* x,

then $T^{-\rho}$ can be uniquely recovered. Note that (A2) does not require $M(e)$ to be diagonalizable, nor to have all its eigenvalues real. For 2×2 symmetric transition matrices, the model described by (A1) and a stronger form of (A2), in which $1 > \det(M(e)) > 0$, is the model described by Cavender [2] (see also [3,4]) and, for this model, $T^{-\rho}$ can be uniquely recovered, for instance by spectral analysis relative to the group $\mathbf{Z}_2$ (see [5]). However, even for this special model, the rooted tree T (as opposed to $T^{-\rho}$) cannot be found without a further assumption, namely that $\pi_1 \neq \pi_2$. We now describe an analytical result which allows $T^{-\rho}$ to be found easily and quickly (i.e., in polynomial time) in the general (nonsymmetric) case, with any number of colours, under Assumptions (A1) and (A2).

Note that we do not make any assumption about the actual process occurring on an edge which produces net random transitions of states between its ends and, in particular, we do not assume any sort of fixed continuous-time process, let alone a "rate" matrix constant across edges of the tree (as in [6]). Also, we do not make any further assumption about the root distribution π or the structure on the family of transition matrices, apart from those properties prescribed by (A2). Our treatment is, therefore, valid for a much wider class of models than is usually considered in molecular taxonomy (see [6]).

2. THE MAIN RESULT

Given the above model, let $f_{ij}(x, y)$ denote the probability that leaf i is coloured x and leaf j is coloured y (this is a sum of the probabilities of a subset of leaf colourations). Hence, for example, $\sum_{x \neq y} f_{ij}(x, y)$ is the probability that leaves i and j are differently coloured. Also, note that:

$$\sum_{x,y} f_{ij}(x, y) = 1. \tag{1}$$

By indexing the colours, $f_{ij}(x, y)$ forms a square matrix, $F_{ij} = [f_{ij}(x, y)]$, thus, we can define

$$\psi_{ij} := -\ln \left[\left| \det(F_{ij}) \right| \right]. \tag{2}$$

THEOREM. *There is a unique (unrooted) phylogenetic tree, namely* $T^{-\rho}$, *and a unique strictly positive valued function* λ^* *defined on the edges of* $T^{-\rho}$ *such that, for all* i, j, ψ_{ij} *is the sum of* $\lambda^*(e)$ *over all edges* e *on the path in* $T^{-\rho}$ *joining* i *and* j. *Both* $T^{-\rho}$ *and* λ^* *can be reconstructed from the* ψ_{ij} *values in polynomial time.*

PROOF. For leaves i and j of T, let $v(i,j)$ denote the vertex of T, which is the last vertex common to the paths from ρ to i and from ρ to j (i.e., the "most recent" common ancestor of i and j). For a vertex v of T let $\pi_k(v)$ denote the probability that v is in state k (by (A1), this will be a function of π and the transition matrices on the path from the root to v). An inductive argument based on (A2) shows that

$$\pi_k(v) \neq 0 \qquad \text{for all } v \text{ and all } k, \tag{3}$$

for otherwise some column of one of the matrices $M(e)$ would consist entirely of zeros, and this would imply $\det(M(e)) = 0$. For economy, we will write $\pi_k(i,j)$ to denote $\pi_k(v(i,j))$. Let

$$\Delta(i,j) := \prod_k \pi_k(i,j),$$

and let $\phi(e)$ denote the absolute value of the determinant of $M(e)$:

$$\phi(e) := \left|\det(M(e))\right|.$$

Since the eigenvalues of a transition matrix have modulus at most 1, (A2) gives:

$$0 < \phi(e) < 1. \tag{4}$$

We claim that

$$\psi_{ij} = -\ln\left[\Delta(i,j)\right] - \sum_{e \in P(T;i,j)} \ln\left[\phi(e)\right], \tag{5}$$

where, for vertices v, v' in T, $P(T;v,v')$ denotes the path in T connecting v and v'. Note that the right hand side of (5) is real, and positive, by (3) and (4). Now, by (A1),

$$f_{ij}(x,y) = \sum_k \pi_k(v)\, P_{k,x}\, Q_{k,y}, \tag{6}$$

where $P = \left[P_{x,k}\right] := \prod_{e \in P(T;v,i)} M(e)$; $Q = \left[Q_{x,k}\right] := \prod_{e \in P(T;v,j)} M(e)$, and where $v := v(i,j)$. Note that (6) can be rewritten as the matrix equation:

$$F_{ij} = P^t \Pi Q,$$

where P^t is the transpose of P, and Π is the diagonal matrix with $\pi_k(v)$ as its kk^{th} entry. Thus, $\det(F_{ij}) = \prod_k \pi_k(i,j) \times \det(P) \times \det(Q)$.

Also, we have:

$$\left|\det(P)\right| = \prod_{e \in P(T;v,i)} \phi(e); \qquad \left|\det(Q)\right| = \prod_{e \in P(T;v,j)} \phi(e),$$

so that

$$\left|\det(F_{ij})\right| = \prod_k \pi_k(i,j) \times \prod_{e \in P(T;i,j)} \phi(e),$$

which establishes the claim (5).

We now define a function λ on the edges of T and show that it is real, strictly positive, and "realises" ψ_{ij} on T. Given such a λ, we obtain a function λ^* on the edges of $T^{-\rho}$ which also has these properties by setting:

$$\lambda^*(e) = \begin{cases} \lambda(e), & \text{if } \rho \text{ has degree} > 2, \text{ or if } e \text{ is not incident with } \rho \\ \lambda(e_1) + \lambda(e_2), & \text{if } \rho \text{ has degree 2, with incident edges } e_1, e_2, \text{ and } e \\ & \text{is the edge of } T^{-\rho} \text{ obtained by identifying } e_1 \text{ and } e_2. \end{cases}$$

The rest of the theorem will then follow by the well-known existence and uniqueness results involving the four point condition and additive realisations of dissimilarity measures on unrooted phylogenetic trees (see, for instance, [1]). To define λ, we proceed as follows: for any edge $e = (v, w)$ of T, where w is a leaf, set

$$\lambda(e) = -\ln\left[\phi(e)\right] - 0.5 \ln\left[\prod_k \pi_k(v)\right],$$

and for any edge $e = (v, w)$ of T for which neither of v, w are leaves, set

$$\lambda(e) = -\ln\left[\phi(e)\right] - 0.5 \ln\left[\prod_k \pi_k(v)\right] + 0.5 \ln\left[\prod_k \pi_k(w)\right]. \tag{7}$$

Then, from (3) and (4), $\lambda(e)$ is real, and for edges incident with a leaf $\lambda(e) > 0$ by (3). Furthermore, it is easily checked from (5) that $\psi_{ij} = \sum_{e \in P(T;i,j)} \lambda(e)$. Thus, it remains to check that $\lambda(e) > 0$ in (7). Let us first suppose $M = \left[M_{\mu\nu}\right]$ is any $r \times r$ matrix with nonnegative entries and x is row vector of length r with nonnegative entries. We claim that

$$\prod_\mu (xM)_\mu \geq \prod_\mu x_\mu \times \left|\det(M)\right|. \tag{8}$$

To obtain this, note that the left hand side of (8) is just:

$$\prod_\mu \left(\sum_\nu x_\nu M_{\nu\mu}\right) \geq \prod_\mu x_\mu \times \left(\sum_\sigma M_{\sigma(1)1} M_{\sigma(2)2} \ldots M_{\sigma(r)r}\right), \tag{9}$$

where the second summation is over all permutations σ of $(1, 2, \ldots, r)$, and so this sum is at least $\left|\det(M)\right|$, since the permanent of a nonnegative matrix is never smaller than the absolute value of its determinant. Now, by (A1), $\left[\pi_1(w), \ldots, \pi_r(w)\right] = \left[\pi_1(v), \ldots, \pi_r(v)\right] M(e)$, and so, applying (8) to the case $M = M(e)$ and $x = \left[\pi_1(v), \ldots, \pi_r(v)\right]$, and noting from (4), that $\det(M)^2 < \left|\det(M)\right| = \phi(e)$, we obtain

$$\prod_k \pi_k(w) > \prod_k \pi_k(v) \times \phi(e)^2.$$

Taking the natural logarithm of this inequality and multiplying by $-\frac{1}{2}$ shows that the expression in (7) is positive, as required. This completes the proof.

COROLLARY 1. *Each phylogenetic tree T, up to the placement of its root, is uniquely defined by the collection of probabilities of the leaf colourations it induces under Assumptions* (A1) *and* (A2).

In taxonomic applications, the probability of each leaf colouration is often estimated simply as the observed proportion of sites in a collection of aligned sequences which correspond to this colouration. Provided the sites in the sequence have evolved identically and independently (the i.i.d. model), these estimates will tend, with probability 1, to the true probability value as the length of the sequences increases. More generally, this statement holds if the sites evolve identically and with limits on the degree of pairwise correlation between states at different sites, as allowed by Bernstein's Theorem (see [7]). In either situation, we have the following result.

COROLLARY 2. *A computationally efficient and statistically consistent algorithm to reconstruct unrooted phylogenetic trees from aligned sequence data satisfying the i.i.d. (or weaker) assumption described above (as well as* (A1), (A2)*) is the following procedure.*

Step 1. *For each pair i, j, and each x, y, estimate $f_{ij}(x, y)$ by setting it equal to the proportion of sites in which i and j are in state x and y, respectively.*
Step 2. *Using (2), calculate ψ_{ij} for each pair i, j.*
Step 3. *Use a suitable dissimilarity-based tree reconstruction method (e.g., Bandelt and Dress's split decomposition method [8]), taking the ψ_{ij} as the dissimilarity values.*

In the final corollary, we demonstrate the existence of "phylogenetic invariants" for the semigroup of transition matrices satisfying (A2). (Phylogenetic invariants for a semigroup G of transition matrices are polynomial functions of the leaf colouration probabilities which, for at least one rooted phylogenetic tree, take the same value for any choice of the matrices $M(e)$ from G).

COROLLARY 3. *Phylogenetic invariants exist for the semigroup $G_1(\det M(e) \neq 0, \pm 1)$ with arbitrary root distribution ($\pi_x \neq 0$). Specifically, for any subset of four leaves i, j, k, l from T, we have $P(T; i, j) \cap P(T; k, l) = \phi$ if and only if:*

$$\det(F_{ik}F_{jl}) - \det(F_{il}F_{jk}) = 0.$$

REFERENCES

1. H.-J. Bandelt and A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Adv. Appl. Math.* **7**, 309–343 (1986).
2. J.A. Cavender, Taxonomy with confidence, *Math. Biosci.* **40**, 271–280 (1978).
3. J.S. Farris, A probability model for inferring evolutionary trees, *Syst. Zool.* **22**, 250–256 (1973).
4. J. Pearl and M. Tarsi, Structuring causal trees, *J. Complexity* **2**, 60–77 (1986).
5. M.D. Hendy, The relationship between simple evolutionary tree models and observable sequence data, *Syst. Zool.* **38**, 310–321 (1989).
6. F. Rodriguez, J.L. Oliver, A. Marin and J.R. Medina, The general stochastic model of nucleotide substitution, *J. Theor. Biol.* **142**, 485–501 (1990).
7. A. Rényi, *Probability Theory*, North-Holland Publishing, Amsterdam, (1970).
8. H.-J. Bandelt and A. Dress, A canonical decomposition theory for metrics on a finite set, *Adv. Math.* **92** (1), 47–105 (1992).
9. J.A. Cavender and J. Felsenstein, Invariants of phylogenies: Simple cases with discrete states, *J. Classification* **4**, 57–71 (1987).