Appl. Math. Lett. Vol. 5, No. 6, pp. 63–67, 1992
Printed in Great Britain. All rights reserved

0893-9659/92 $5.00 + 0.00

SPECTRAL ANALYSIS AND A CLOSEST TREE METHOD FOR GENETIC SEQUENCES

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

M. D. HENDY
Massey University, Mathematics Department
Palmerston North, New Zealand

L. A. SZÉKELY
Eötvös University, Department of Computer Science, Budapest, Hungary
and
Institut für Diskrete Mathematik, Bonn, Germany

P. L. ERDŐS
University of Twente, Faculty of Applied Mathematics
Enschede, The Netherlands

(Received June 1992)

Abstract—We describe a new method for estimating the evolutionary tree linking a collection of species from their aligned four-state genetic sequences. This method, which can be adapted to provide a branch-and-bound algorithm, is statistically consistent provided the sequences have evolved according to a standard stochastic model of nucleotide mutation. Our approach exploits a recent group-theoretic description of this model.

1. INTRODUCTION

Recently there has been a desire to apply stochastic models to the reconstruction of evolutionary trees from aligned genetic sequences. This has led to statistically consistent methods, such as maximum likelihood, being favoured over alternative, though more popular approaches, such as the method of maximum parsimony. One problem with maximum likelihood, however, is that it is computationally very difficult, and branch-and-bound algorithms are unknown. A further, more fundamental, problem arises in deciding whether one tree is "significantly better" than another on the basis of their likelihood ratios alone [1].

An alternative approach relies on the natural equivalence between trees whose leaves are labelled by the species, and collections of pairwise compatible splits (bipartitions) of the set of species (see [2]). In this setting, one applies a suitable, tree-independent transformation to the data to obtain a measure of how well the data supports each possible split of the species set as being an actual split of the underlying tree. This philosophy is implicit in the recent "split decomposition" approach of Bandelt and Dress [3], which deals with distance data rather than sequences, and in the "spectral analysis" method of Hendy and Penny [4].

This latter procedure is confined, however, to sequences with two-character states. Here, we provide an extension to deal with the four-character states A, C, G, T which occur with nucleotide (DNA) sequences. Furthermore, if the sequences have evolved according to Kimura's 3-parameter model (see [5,6]), the transformation will be statistically consistent by singling out just the splits that exist in the underlying tree, as the length of the sequences becomes sufficiently large [7].

Research of L. A. Székely supported by the A.v. Humboldt-Stiftung and U.S. Office of Naval Research under the contract N-0014-91-J-1385.

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

A simple (and statistically consistent) method for choosing an optimal tree from the spectrum—a "closest tree"—is also described, generalizing the closest tree method of Hendy [8]. Unlike maximum likelihood, this method shares with parsimony the property of selecting the tree(s) which minimizes an explicitly-representable function of the data. This, in turn, allows the development of a branch-and-bound algorithm.

Our results rely on a recent application of discrete Fourier transformations to analyse probability distributions induced by randomly colouring the edges of a tree by elements of the Kleinian group $\mathbf{Z}_2 \times \mathbf{Z}_2$ [6,9,10]. Applications to real data will appear elsewhere (Penny, Hendy and Steel, in preparation).

DEFINITIONS.

(1) *Denote a set of n species by the label set $L = \{R, 1, \dots, n-1\}$, where R is an arbitrarily-selected "root" species. A colouration χ of L by the Kleinian group $G = \mathbf{Z}_2 \times \mathbf{Z}_2$ defines a pair of subsets σ_1, σ_2 of $L \setminus \{R\}$ as follows: For $i = 1, 2$, let π_i denote the projection of G onto its i^{th} component, and let*

$$\sigma_i = \{k \in L, k \neq R : \pi_i(\chi(k) + \chi(R)) = 1\}.$$

(2) *An aligned collection of genetic sequences of length m can be regarded as an m-tuple of colourations of L by G, by representing the four nucleotides A, G, C, T, by the elements $(0,0), (0,1), (1,0), (1,1)$ respectively, as in [6]. The sets of corresponding components of these sequences are called sites. For such an m-tuple, let $x_\Theta(\Theta = (\sigma_1, \sigma_2))$ denote the proportion of the colourations which correspond to the induced pair σ_1, σ_2 by the bijection in (1). By ordering the pairs Θ, the x_Θ's form a vector $\mathbf{x}$ having 4^{n-1} non-negative components which sum to 1.*

(3) *Let H' denote the $2^{n-1} \times 2^{n-1}$ symmetric matrix $[(-1)^{|\sigma \cap \sigma'|}]$, ($\sigma, \sigma'$ subsets of $\{1, \dots, n-1\}$), and let H denote the Kronecker product of H' with itself. Since H' is a symmetric Hadamard matrix (see [4]), H is also, so that:*

$$H^{-1} = 4^{1-n} H. \tag{1.1}$$

(4) *For vector $\mathbf{y} \in (\mathbf{R}^+)^k$, (resp. $\mathbf{y} \in \mathbf{R}^k$) define $\log(\mathbf{y})$ (resp. $\exp(\mathbf{y})$) to be the vector whose i^{th} component is $\log_e(y_i)$ (resp. $\exp(y_i)$) for $i = 1, \dots, k$.*

(5) *For $\mathbf{x}$ as in (2), if $(H\mathbf{x})_i > 0$ for all i, define*

$$\begin{aligned} \gamma(\mathbf{x}) &= H^{-1}(\log(H\mathbf{x})), \\ &= 4^{1-n} H(\log(H\mathbf{x})), \qquad \textit{by equation (1.1).} \end{aligned} \tag{1.2}$$

The vector $\gamma(\mathbf{x})$ with components $\gamma_\Theta(\mathbf{x})$, ($\Theta = (\sigma_1, \sigma_2)$) is called the *conjugate spectrum* of the sequence data. Note that, since $\sum_\Theta x_\Theta = 1$,

$$\sum_\Theta \gamma_\Theta(\mathbf{x}) = 0. \tag{1.3}$$

2. USING CONJUGATE SPECTRA TO INVERT KIMURA'S 3-PARAMETER MODEL

The transformation defined by (1.2) is independent of any considerations involving trees, or of how the aligned sequences may have evolved. We now show that if the sequences have indeed evolved on a tree according to a standard stochastic model, then the conjugate spectrum converges to a vector which identifies that tree, and provides the other parameters in the model.

DEFINITION. *A phylogenetic tree on L, is a collection T of nonempty subsets of $L' = \{1, \dots, n-1\}$ with the properties:*

(1) *$L' \in T$ and $\{i\} \in T$, for all $i \in L'$.*

(2) *For ρ, $\rho' \in T$, $\rho \cap \rho' \in \{\rho, \rho', \emptyset\}$.*

There is a well-known bijection ϕ between the sets of phylogenetic trees on L and the (graph-theoretic) trees which have $|L|$ leaves labelled with distinct elements of L, and whose non-leaf vertices are unlabelled and of degree at least three (see [2]). Under ϕ, $\rho \in T$ corresponds to the edge of $\phi(T)$ which seperates the leaves whose labels are in ρ from the leaves whose labels are in $L - \rho$. The pair $\{\rho, L-\rho\}$ is said to be a *split* of $\phi(T)$.

Suppose that T is a phylogenetic tree on L. We assume that the nucleotides at different sites on the aligned sequence have evolved identically and independently from (unknown) ancestral states at some "ancestral point" in $\phi(T)$—such a "point" could be a non-leaf vertex, or the midpoint of an edge. Then, with probability 1, x_Θ converges (as the length of the sequences tends to infinity) to its expected value, denoted f_Θ. Note that f_Θ is the probability of observing, at any given site, a leaf-colouring that induces the pair $\Theta = (\sigma_1, \sigma_2)$.

To calculate f_Θ we need a model to describe the generation of leaf colourings. Kimura's 3-parameter model [5] is a continuous Markov model of nucleotide changes prescribed by three independent probabilities on each edge of $\phi(T)$. The changes are categorised into three types, the *transitions*, A interchanging with G ($A \leftrightarrow G$) and $C \leftrightarrow T$, and two types of *transversions*, $A \leftrightarrow C$, $G \leftrightarrow T$ and $A \leftrightarrow T$, $C \leftrightarrow G$. For the edge of $\phi(T)$ corresponding to ρ the expected numbers of these changes can be shown to be E_ρ^2, E_ρ^1 and E_ρ^3 as defined below, when a continuous time Markov process is assumed.

Evans and Speed showed [6] that under this model, if the nucleotide characters are identified with the elements of the Kleinian group G, then the change of x to y is a function of $y - x \in G$. They identified the three types of change with $(0,1), (1,0)$ and $(1,1)$ respectively. We generalise this model with a model M in which we randomly edge colour $\phi(T)$, with $p_\rho(g)$ being the probability the edge corresponding to ρ is coloured g. $P(\rho, g) = p_\rho(g)$ maps $T \times G \rightarrow [0,1]$. This induces a leaf colouring Θ, when we colour R by $(0,0)$, and colour leaf $\ell \neq R$ by the sum of the edge colours in the path in $\phi(T)$ from ℓ to R. Let $f_\Theta(T,P)$ be the probability of inducing the pair of subsets $\Theta = (\sigma_1, \sigma_2)$ as in (1). Let

$$\mathbf{P}_\rho = \begin{bmatrix} p_\rho(1,0) \\ p_\rho(0,1) \\ p_\rho(1,1) \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix},$$

and let

$$\mathbf{q}_\rho = \begin{bmatrix} q_\rho^1 \\ q_\rho^2 \\ q_\rho^3 \end{bmatrix} = -\frac{1}{2}\log(\mathbf{j} - 2A\mathbf{P}_\rho), \quad \mathbf{E}_\rho = A^{-1}\mathbf{q}_\rho.$$

Note that $A^{-1} = A - \frac{1}{2}J$ (where $J = \mathbf{j}\mathbf{j}^{\mathrm{T}}$) and $\mathbf{P}_\rho = \frac{1}{4}\mathbf{j} - \frac{1}{2}A^{-1}\exp(-2A\,\mathbf{E}_\rho)$.

Kimura's model implies the following two conditions, while model M only requires Condition 1.

CONDITION 1. *q_ρ^i exists for $i = 1,2,3, \rho \in T$.*

CONDITION 2. *$E_\rho^i \geq 0$ for $i = 1,2,3, \rho \in T$.*

Notice that Condition 1 can be re-phrased as:

CONDITION 1′. *For $\rho \in T$, and distinct nonzero g, $g' \in G$, $p_\rho(g) + p_\rho(g') < \frac{1}{2}$.*

It is easily checked that model M together with Conditions 1 and 2 characterise the leaf-colouration probability distributions which can arise under Kimura's 3-parameter model (with its implicit assumption of a continuous-time process). If we drop Condition 2, we have a slightly more general model, which we will refer to as the *generalized (Kimura 3-parameter) model.*

Let $C(T) = \{(\emptyset,\emptyset)\} \cup \{\rho^{(i)} : \rho \in T,\ i = 1,2,3\}$ where $\rho^{(1)} = (\rho, \emptyset)$, $\rho^{(2)} = (\emptyset, \rho)$, $\rho^{(3)} = (\rho, \rho)$.

THEOREM 1. *(Inversion formulae) For a leaf-colouration distribution $\mathbf{f} = \mathbf{f}(T,P)$ arising under the generalized model, we have:*

$$\gamma_\Theta(\mathbf{f}) = \begin{cases} 0, & \text{if } \Theta \notin C(T), \\ E_\rho^i, & \text{if } \Theta = \rho^{(i)}, \quad \rho \in T, \\ -\sum\limits_{\rho \in T} \mathbf{j} \cdot \mathbf{E}_\rho, & \text{if } \Theta = (\emptyset, \emptyset). \end{cases}$$

In particular, the conjugate spectra of aligned sequences provide consistent estimators (under the generalized model) of the underlying parameters (T and P).

PROOF. For $i = 1,2,3$, let $\mathbf{q}^i$ be the vector obtained from $\{q^i_\rho\}$ by indexing the subsets ρ of $L' = \{1,\ldots,n-1\}$. For $\Theta = (\sigma_1,\sigma_2)$ let $v^i_{\Theta,\rho} = (-1)^{|\sigma_i \cap \rho|}$ for $i = 1,2$, and let $v^3_{\Theta,\rho} = v^1_{\Theta,\rho} . v^2_{\Theta,\rho}$. For $i = 1,2,3$, let K^i be the 4^{n-1} by $|T|, 0-1$ matrix with $K^i_{\Theta,\rho}$ $(\Theta = (\sigma,\sigma'))$, $\rho \in T$, defined by the equations:

$$\begin{bmatrix} K^1_{\Theta,\rho} \\ K^2_{\Theta,\rho} \\ K^3_{\Theta,\rho} \end{bmatrix} = \frac{1}{4}(\mathbf{j} - 2A^{-1}\mathbf{v}_{\Theta,\rho}), \qquad \text{where } \mathbf{v}_{\Theta,\rho} = \begin{bmatrix} v^1_{\Theta,\rho} \\ v^2_{\Theta,\rho} \\ v^3_{\Theta,\rho} \end{bmatrix}.$$

Applying the main result from [10], it can be shown that:

$$\mathbf{f} = H^{-1}\exp(\mathbf{r}), \qquad \text{where } \mathbf{r} = \sum_{i=1}^{3} K^i(-2\mathbf{q}^i).$$

(Specifically, with $P(T,X_i)$ defined as in [10], let $P_3(X_1,X_2) = P(T,X_1) \cap P(T,X_2)$, $P_i(X_1,X_2) = P(T,X_i) \setminus P_3(X_1,X_2)$, for $i = 1,2$; then the edge corresponding to ρ lies in $P_i(X_1,X_2)$ iff $K^i_{\Theta,\rho} = 1$, where $\Theta = (X_1 \cap L', X_2 \cap L')$.)

Now, $r_\Theta = \sum_{\rho\in T}(-\mathbf{j}\cdot\mathbf{E}_\rho + \mathbf{v}_{\Theta,\rho}\cdot\mathbf{E}_\rho)$, that is, $\mathbf{r} = \sum_{\rho\in T}\sum_{i=1}^{3}((-E^i_\rho)\mathbf{j} + E^i_\rho\,\mathbf{v}^i_\rho)$, where $\mathbf{v}^i_\rho$ is the vector obtained from $v^i_{\Theta,\rho}$ by ordering the Θ. Now,

$$(H^{-1}\mathbf{j})_\Theta = \begin{cases} 1, & \text{if } \Theta = (\emptyset,\emptyset), \\ 0, & \text{otherwise,} \end{cases} \qquad (H^{-1}\mathbf{v}^i_\rho)_\Theta = \begin{cases} 1, & \text{if } \Theta = \rho^{(i)}, \\ 0, & \text{otherwise,} \end{cases}$$

and since $\gamma(\mathbf{f}) = H^{-1}\mathbf{r}$, the result follows. ∎

If x_Θ is the relative frequency of colouring Θ in some observed sequences we can take $\mathbf{x} = (x_\Theta)$ as an estimate of $f(T', P')$ where T', P' are unknown. For each T we can find P to minimize the Euclidean distance $D(T,\mathbf{x}) = \|\gamma(\mathbf{x}) - \gamma(\mathbf{f}(T,P))\|$. A *closest tree* for $\mathbf{x}$ is a tree T which minimizes $D(T,\mathbf{x})$. This can be found using a branch and bound search over all trees as in [11], to minimize $\delta(T,\mathbf{x})$ (below).

Write $\gamma_\Theta = \gamma_\Theta(\mathbf{x})$, and let $\gamma_T = \sum_{\Theta\in C(T)} \gamma_\Theta$.

COROLLARY. *(Approximation formulae)*

(1) *For any tree T the values E^i_ρ which minimize $D(T,\mathbf{x})$ are given by:*

$$E^i_\rho = \gamma^{(i)}_\rho - \frac{1}{c}\gamma_T, \qquad c = |C(T)| = 3|T| + 1.$$

(2) *A closest tree for $\mathbf{x}$ is a tree T which minimizes*

$$\delta(T,\mathbf{x}) = -\sum_{\Theta\in C(T)} \gamma^2_\Theta + \frac{\gamma^2_T}{c}.$$

For this tree,

$$D^2(T,\mathbf{x}) = \sum_{\Theta} \gamma^2_\Theta + \delta(T,\mathbf{x}).$$

REMARKS.

(1) Without condition 1, P is not uniquely determined by $f(T,P)$.

(2) If, in part (1) of the Corollary the optimal values for E^i_ρ are negative, this suggests that (under the Kimura 3-parameter model) ρ does not correspond to an edge in the tree which generated the data.

(3) Unlike the conjugate spectra considered in [4], here there exists $4^{n-1} - c$ pairs Θ such that $\gamma_\Theta(\mathbf{f}(T,P)) = 0$ for all (T,P). This provides a useful check on the applicability of the model as a method for reconstructing trees.

References

1. W.-H. Li and M. Gouy, Statistical methods for testing molecular phylogenies, In *Phylogenetic Analysis of DNA Sequences*, (Edited by M.M. Miyamoto and J. Cracraft), Oxford University Press, Oxford, 249–277, (1991).
2. H.-J. Bandelt and A. Dress, Reconstructing the shape of a tree from observed dissimilarity data, *Advances in Applied Mathematics* 7, 309–343 (1986).
3. H.-J. Bandelt and A. Dress, A canonical decomposition theory for metrics on a finite set, *Advances in Mathematics* (to appear).
4. M.D. Hendy and D. Penny, Spectral analysis of phylogenetic data, University of Bielefield, ZiF-Nr 91/23, preprint, (1991).
5. M. Kimura, Estimation of evolutionary sequences between homologous nucleotide sequences, In *Proceedings of the National Academy of Sciences USA*, Vol. 78, pp. 454–458, (1981).
6. S.N. Evans and T.P. Speed, Invariants of some probability models used in phylogenetic inference, *Annals of Statistics* (to appear).
7. M.D. Hendy and D. Penny, A framework for the quantitative study of evolutionary trees, *Systematic Zoology* **38** (4), 297–309 (1989).
8. M.D. Hendy, A combinatorial description of the closest tree algorithm for finding evolutionary trees, *Discrete Mathematics* **96** (1), 51–58 (1991).
9. L.A. Székely, M.A. Steel and P.L. Erdős, Fourier calculus on finite sets and evolutionary trees (submitted).
10. L.A. Székely, P.L. Erdős, M.A. Steel and D. Penny, A Fourier inversion formula for evolutionary trees, *Applied Mathematics Letters* (to appear).
11. M.D. Hendy and D. Penny, Branch and bound algorithms to determine minimal evolutionary trees, *Mathematical Biosciences* **59**, 277–290 (1982).