

Maximum Likelihood Jukes-Cantor Triplets: Analytic Solutions

Benny Chor,^{*} Michael D. Hendy,[†] and Sagi Snir[‡]

^{*}School of Computer Science, Tel-Aviv University, Israel; [†]Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; and [‡]Mathematics Department, University of California, Berkeley

Maximum likelihood (ML) is a popular method for inferring a phylogenetic tree of the evolutionary relationship of a set of taxa, from observed homologous aligned genetic sequences of the taxa. Generally, the computation of the ML tree is based on numerical methods, which in a few cases, are known to converge to a local maximum on a tree, which is suboptimal. The extent of this problem is unknown, one approach is to attempt to derive algebraic equations for the likelihood equation and find the maximum points analytically. This approach has so far only been successful in the very simplest cases, of three or four taxa under the Neyman model of evolution of two-state characters. In this paper we extend this approach, for the first time, to four-state characters, the Jukes-Cantor model under a molecular clock, on a tree T on three taxa, a rooted triple. We employ spectral methods (Hadamard conjugation) to express the likelihood function parameterized by the path-length spectrum. Taking partial derivatives, we derive a set of polynomial equations whose simultaneous solution contains all critical points of the likelihood function. Using tools of algebraic geometry (the resultant of two polynomials) in the computer algebra packages (Maple), we are able to find all turning points analytically. We then employ this method on real sequence data and obtain realistic results on the primate-rodents divergence time.

Introduction

Maximum likelihood (ML) is increasingly used as an optimality criterion for selecting evolutionary trees (Felsenstein 1981), but finding the global optimum is a hard computational task, which led to using mostly numeric methods. So far, analytical solutions have been derived only for the simplest models (Yang 2000; Chor, Hendy, and Penny 2001; Chor, Khetan, and Snir 2003)—three and four taxa under a molecular clock, with just two-state characters (Neyman 1971). In this work we present, for the first time, the analytic solutions for a general family of trees with four-state characters, such as DNA or RNA. The model of substitution we use is the Jukes-Cantor model (Jukes and Cantor 1969) under a molecular clock, the simplest four-state model where all substitutions have the same rate. The trees we deal with are on just three taxa, namely, rooted triplets (see fig. 1*b*).

The change from two to four character states incurs a major increase in the complexity of the underlying algebraic system. Like previous works, our starting point is to present the general ML problem on phylogenetic trees as a constrained optimization problem. This gives rise to a complex system of polynomial equations, and the goal is to solve them. The complexity of this system prevents any manual solution, so we relied heavily on Maple, a symbolic mathematical system. However, even with Maple, a simple attempt to solve the system (e.g., just typing solve) fails, and additional tools are required. Spectral analysis (Hendy and Penny 1993; Hendy, Penny, and Steel 1994; Hendy and Snir 2005) uses Hadamard conjugation to express the expected frequencies of site patterns among sequences as a function of an evolutionary tree T and a model of sequence evolution along the edges of T . As in previous works, we use the Hadamard conjugation as a basic building block in our method of solution. However, it turns out that the specific way we represent the system,

which is determined by the choice of variables, plays a crucial role in the ability to solve it. In previous works (Chor, Hendy, and Penny 2001; Chor, Khetan, and Snir 2003), the variables in the polynomials were based on the expected sequence spectrum (Hendy and Penny 1993). This representation leads to a system with two polynomials of degrees 11 and 10. This system is too complex to resolve with the available analytic and computational tools. By employing a different set of variables, based on the path-set spectrum (Hendy and Snir 2005), we were able to arrive at a simpler system of two polynomials whose degrees are 10 and 8. To finesse the construction, we use algebraic geometry combined with Maple. Specifically, we now compute the resultant of the two polynomials, which yields a single, degree 11 polynomial in one variable. The roots of this polynomial yield the desired ML solution. We remark that similar results to those shown here were obtained by Hosten, Khetan, and Sturmfels (2004), using somewhat different techniques. A set of computational algebraic tools for analyzing similar models on small trees based on the methods reported in Catanese et al. (2004); Hosten, Khetan, and Sturmfels (2004); and Sturmfels and Sullivan (2005) is detailed in the web page <http://math.tamu.edu/~lgp/small-trees/>.

It is evident that the currently available heuristic methods can fail to infer the correct tree, even for small number of taxa. This is true not only in the presence of multiple ML points but also in cases where the neighborhood of the (single) ML point is relatively flat. Therefore, a practical consequence of this work is the use of rooted triplets in supertree methods (constructing a big tree from small subtrees). For unrooted trees, the subtrees must have at least four leaves (e.g., “the tree from quartets” problem). For rooted trees, it is sufficient to amalgamate a set of rooted triplets (Aho et al. 1981). Our work enables such approaches to rely on rooted ML triplets based on four characters states rather than two.

Additionally, analytic solutions are able to reveal properties of the ML points that are not obtainable numerically. For small trees, our result can serve as a method for checking the accuracy of the heuristic methods used in practice.

Key words: maximum likelihood, phylogenetic trees, Jukes-Cantor, Hadamard conjugation, analytical solutions, symbolic algebra.

E-mail: m.hendy@massey.ac.nz.

Mol. Biol. Evol. 23(3):626–632, 2006

doi:10.1093/molbev/msj069

Advance Access publication November 30, 2005

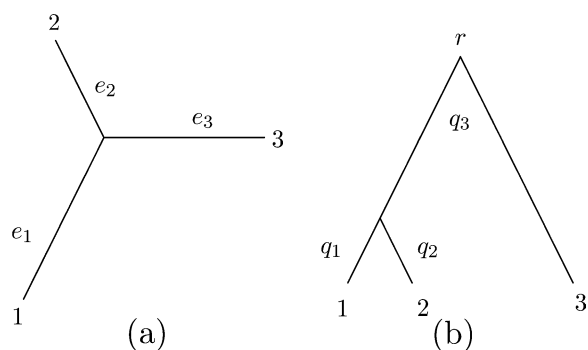


FIG. 1.—(a) A tree T on three leaves, 1, 2, and 3, with edges labeled. (b) A rooted tree on three leaves, with root r , with edge weights q_1, q_2 , and q_3 . If $q_1 = q_2 \leq q_3$, then this tree satisfies the molecular clock as the path lengths from r to each leaf are the same.

The remainder of this work is organized as following. In the next section we describe the materials and methods we used in this work. Specifically, in *Definitions, Notations, and the Hadamard Conjugation* we provide definitions and notations used throughout the rest of this work. In *Definitions, Notations, and the Hadamard Conjugation* we develop the identities and structures induced by the Jukes-Cantor model, while in *Obtaining the ML Solution* we develop ML on phylogenetic trees and subsequently solves the system for the special case of three species under Jukes-Cantor and molecular clock. In *Results and Discussion* we give results and discuss future research directions: *Results on Genomic Sequences* reports on experimental results of applying our method on real genomic sequences, while in *Directions for Future Research* we conclude with three open problems.

Materials and Methods

Here we detail on the methods and tools we employed in order to obtain our results. These are developed specifically for the tree T on three taxa referred to as 1, 2, and 3. We label the leaves of T as 1, 2, and 3 and the edges (branches) as e_1, e_2 , and e_3 , as illustrated in figure 1a. Taxon 3 is chosen to be the reference taxon. Our analysis is focused on the site substitutions required to transform the reference sequence to those of 1 and 2 under Kimura’s 3-substitution model (K3ST) (Kimura 1981). We will subsequently impose the constraints of the Jukes-Cantor model (Jukes and Cantor 1969), and under a molecular clock, on the rooted tree of figure 1b.

Definitions, Notations, and the Hadamard Conjugation

Given aligned nucleotide sequences for the three taxa 1, 2, and 3, there are 4^3 nucleotide combinations possible at each site. We refer to each of these as a “character pattern.”

A character pattern $\begin{bmatrix} \chi_1 \\ \chi_2 \\ \chi_3 \end{bmatrix}$ can be described by the state χ_3 at taxon 3, together with the substitutions $\begin{bmatrix} \chi_3 \rightsquigarrow \chi_1 \\ \chi_3 \rightsquigarrow \chi_2 \end{bmatrix}$.

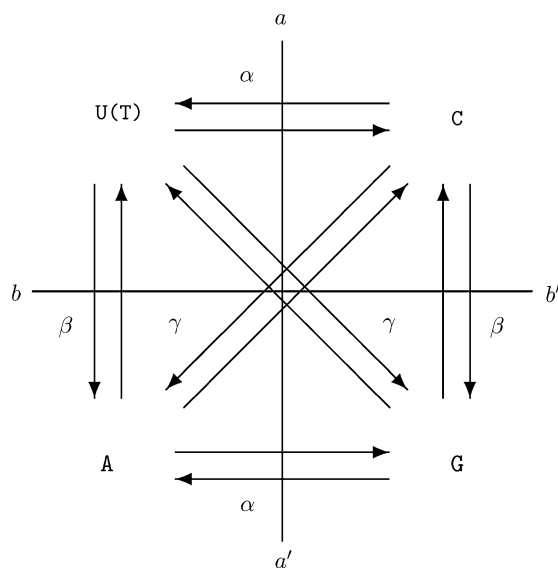


FIG. 2.—K3ST model. Substitution types α and γ cross the line aa' , and β and γ cross the line bb' .

Kimura (1981) in his K3ST model, describes three classes of substitution: the transitions; and two types of transversions. Following Hendy and Snir (2005), we use α to denote a transition, β and γ to denote the two transversion types, together with ϵ to indicate no substitution, as described in figure 2. For example, the character pattern

$\begin{bmatrix} T \\ A \\ C \end{bmatrix}$ is obtained with the assignment of C to taxon 3, and

the substitution pattern $\begin{bmatrix} C \rightsquigarrow T \\ C \rightsquigarrow A \end{bmatrix} = \begin{bmatrix} \alpha \\ \gamma \end{bmatrix}$ for taxa 1 and 2.

In Hendy, Penny, and Steel (1994) it is shown that under the K3ST model, the probability of a substitution pattern is independent of the character state at taxon 3. We index each site pattern by a pair (X, Y) of subsets of $\{1, 2\}$, where X is the set of taxa with substitutions which cross the line aa' in (fig. 2), and Y is the set of taxa with substitutions which cross the line bb' . We denote the probability of site pattern (X, Y) as $s_{X,Y}$, which when not ambiguous, we index by the lists of elements of X and of Y , for brevity.

Thus, for example, the substitution pattern $\begin{bmatrix} \alpha \\ \gamma \end{bmatrix}$ is indexed by the pair of subsets $(\{1, 2\}, \{2\})$, and the probability of obtaining this substitution pattern is written as $s_{12,2}$.

The 16 site pattern probabilities are arranged as a 4×4 matrix

$$S = \begin{bmatrix} s_{\emptyset,\emptyset} & s_{\emptyset,1} & s_{\emptyset,2} & s_{\emptyset,12} \\ s_{1,\emptyset} & s_{1,1} & s_{1,2} & s_{1,12} \\ s_{2,\emptyset} & s_{2,1} & s_{2,2} & s_{2,12} \\ s_{12,\emptyset} & s_{12,1} & s_{12,2} & s_{12,12} \end{bmatrix},$$

called the “sequence spectrum.” These probabilities can be parameterized in several ways, in Hendy and Penny (1993) and Hendy, Penny, and Steel (1994) for the K3ST model,

they are given as functions of the expected numbers of substitutions of each type on each edge.

In the Jukes-Cantor model, the expected numbers of the three substitution types on an edge e_i are the same, that is,

$$q_\alpha(e_i) = q_\beta(e_i) = q_\gamma(e_i).$$

We will write this common value as q_i . Thus, q_1 , q_2 , and q_3 are the expected numbers of substitutions of each type, on edges e_1 , e_2 , and e_3 , respectively. Following the results of Hendy, Penny, and Steel (1994); Steel, Hendy, and Penny (1998); and Hendy and Snir (2005) specialized to the Jukes-Cantor model on T for three taxa, we express these expected numbers in the 4×4 matrix

$$Q = \begin{bmatrix} -3(q_1 + q_2 + q_3) & q_1 & q_2 & q_3 \\ q_1 & q_1 & 0 & 0 \\ q_2 & 0 & q_2 & 0 \\ q_3 & 0 & 0 & q_3 \end{bmatrix},$$

called the “edge-length spectrum.” The major result of Hendy, Penny, and Steel (1994) then becomes the following.

Theorem 1

$$S = H^{-1} \exp(HQH)H^{-1}, \tag{1}$$

where

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \tag{2}$$

$H^{-1} = (1/4)H$, and the log function \ln is applied to each component of the matrix HQH .

Equation (2) in Theorem 1 is equivalent to earlier expressions (Steel et al. 1992; Székely et al. 1993) of Hadamard conjugations for the K3ST model, which used Hadamard matrices of 2^n rows and columns applied to vectors of 2^n entries, with $q_i = q_\alpha(e_i) = q_\beta(e_i) = q_\gamma(e_i)$. The current expression has recently been developed in Hendy and Snir (2005). A full proof of this result follows directly from applying Theorem 7 of Hendy and Snir (2005) on the tree T . This approach has been followed in order to clarify the role of path sets, which explain the intermediate terms in the conjugations, enabling us to derive and interpret the values in equation (8) directly.

We now define several auxiliary matrices that will be useful in the sequel:

$$J = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad A_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix},$$

$$A_4 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.$$

The following identities relating H and these six matrices hold:

$$A_0 + A_1 + A_2 + A_3 + A_4 = J, \tag{3}$$

$$HJH = 16A_0, \tag{3}$$

$$HA_0H = J, \tag{4}$$

$$HA_1H = 4(A_0 + A_2) - J, \tag{5}$$

$$HA_2H = 4(A_0 + A_1) - J, \tag{6}$$

$$HA_3H = 4(A_0 + A_3) - J. \tag{7}$$

The edge-length spectrum of an arbitrary 3-tree can be expressed as the 4×4 matrix,

$$Q = \begin{bmatrix} -3(q_1 + q_2 + q_3) & q_1 & q_2 & q_3 \\ q_1 & q_1 & 0 & 0 \\ q_2 & 0 & q_2 & 0 \\ q_3 & 0 & 0 & q_3 \end{bmatrix} = q_1A_1 + q_2A_2 + q_3A_3 - 3(q_1 + q_2 + q_3)A_0.$$

Now, from equations (3–7), we see

$$HQH = -4[(q_1 + q_3)A_1 + (q_2 + q_3)A_2 + (q_1 + q_2)A_3 + (q_1 + q_2 + q_3)A_4]$$

$$= -4 \begin{bmatrix} 0 & q_1 + q_3 & q_2 + q_3 & q_1 + q_2 \\ q_1 + q_3 & q_1 + q_3 & q_1 + q_2 + q_3 & q_1 + q_2 + q_3 \\ q_2 + q_3 & q_1 + q_2 + q_3 & q_2 + q_3 & q_1 + q_2 + q_3 \\ q_1 + q_2 & q_1 + q_2 + q_3 & q_1 + q_2 + q_3 & q_1 + q_2 \end{bmatrix},$$

so applying the exponential function to each element of the matrix HQH , we obtain the so called path-set spectrum, R :

$$R = \exp(HQH) = \begin{bmatrix} 1 & x_1x_3 & x_2x_3 & x_1x_2 \\ x_1x_3 & x_1x_3 & x_1x_2x_3 & x_1x_2x_3 \\ x_2x_3 & x_1x_2x_3 & x_2x_3 & x_1x_2x_3 \\ x_1x_2 & x_1x_2x_3 & x_1x_2x_3 & x_1x_2 \end{bmatrix} \tag{8}$$

$$= A_0 + x_1x_3A_1 + x_2x_3A_2 + x_1x_2A_3 + x_1x_2x_3A_4,$$

where

$$x_i = e^{-4q_i}. \tag{9}$$

The x_i values can replace the q_i values as the defining parameters and are called the “path-set variables.” The entries of R relate to the probabilities of differences between the end points of paths in T (Hendy and Snir 2005).

From Theorem 1, we find the expected sequence spectrum is

$$S = H^{-1}RH^{-1} \tag{10}$$

$$= \begin{bmatrix} a_0 & a_1 & a_2 & a_3 \\ a_1 & a_1 & a_4 & a_4 \\ a_2 & a_4 & a_2 & a_4 \\ a_3 & a_4 & a_4 & a_3 \end{bmatrix}, \tag{11}$$

where

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \mathbf{a} = \frac{1}{16} \begin{bmatrix} 1 & 3 & 3 & 3 & 6 \\ 1 & 3 & -1 & -1 & -2 \\ 1 & -1 & 3 & -1 & -2 \\ 1 & -1 & -1 & 3 & -2 \\ 1 & -1 & -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ x_2x_3 \\ x_1x_3 \\ x_1x_2 \\ x_1x_2x_3 \end{bmatrix}. \tag{12}$$

Thus, we see that the expected frequency of each site pattern takes one of the five a_i values, each of which is a function of the three parameters $x_1, x_2,$ and x_3 .

In particular, when we impose the molecular clock condition $q_1 = q_2$, and hence $x_1 = x_2$, equation (12) reduces to

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \mathbf{a} = \frac{1}{16} \begin{bmatrix} 1 & 6 & 3 & 6 \\ 1 & 2 & -1 & -2 \\ 1 & 2 & -1 & -2 \\ 1 & -2 & 3 & -2 \\ 1 & -2 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ x_1x_3 \\ x_1^2 \\ x_1^2x_3 \end{bmatrix}, \tag{13}$$

and we observe $a_1 = a_2$.

Obtaining the ML Solution

When we have an aligned sequence of length c for each of the three taxa 1, 2, and 3, we can count the relative frequencies (normalized to sum to 1) of the substitution patterns among the sites. For $(X, Y) | X, Y \subseteq \{1, 2\}$, we set $f_{X,Y}$ to be the relative frequency of observing the site pattern (X, Y) , and let F be the 4×4 matrix with entries $f(X, Y)$ indexed as in the sequence spectrum S . If the sequences were generated under the Jukes-Cantor model on the tree T with edge weights q_1, q_2, q_2 , then the expected number of substitution pattern (X, Y) is $f_{X,Y} = c s_{X,Y}$, where c is the number of sites.

However, in c independent samples, the observed frequencies will include sampling error, so we cannot directly conclude $S = F$. ML provides a method of estimating the parameters $q_1, q_2,$ and q_3 from the observed frequencies F .

Given a set of edge lengths $q_1, q_2,$ and q_3 , the “likelihood” of observing F under the Jukes-Cantor model on T is

$$L(T) = \prod_{X,Y \in \{1,2\}} s_{X,Y}^{f_{X,Y}}, \tag{14}$$

where the entries in S are obtained from equation (11). This gives identities among the pattern probabilities $s_{X,Y}$, so grouping the common factors in equation (12) gives

$$L(T) = \prod_{j=0}^4 a_j^{f_j}, \tag{15}$$

where

$$\begin{aligned} f_0 &= F_{\emptyset,\emptyset}, \\ f_1 &= F_{\emptyset,1} + F_{1,\emptyset} + F_{1,1}, \\ f_2 &= F_{\emptyset,2} + F_{2,\emptyset} + F_{2,2}, \\ f_3 &= F_{\emptyset,12} + F_{12,\emptyset} + F_{12,12}, \\ f_4 &= F_{1,2} + F_{1,12} + F_{2,1} + F_{2,12} + F_{12,1} + F_{12,2}. \end{aligned}$$

The expected sequence spectrum S can be expressed as a function of the three variables $x_1, x_2,$ and x_3 , so the values which maximize the likelihood L are obtained when each of the three partial derivatives, $\partial L / \partial x_j = 0$ ($j = 1, 2, 3$). In contrast to previous works (Chor et al. 2000; Chor, Hendy, and Penny 2001; Chor, Khetan, and Snir 2003; Chor and Snir 2004), that operated in the space of the expected sequence variables, $S_{D,E}$, here we are operating in the space of the path-set variables $\mathbf{x} = [x_1 \ x_2 \ x_3]$. This eliminates the need to introduce the constraints of the ML points being on a “tree surface.” By the chain rule, we get:

$$\frac{\partial L}{\partial \mathbf{x}} = L \cdot \sum_{i=0}^4 \frac{f_i \partial a_i}{a_i \partial \mathbf{x}}, \tag{16}$$

which must be $\mathbf{0}$ at each ML point. From equation (12), we can determine the matrix of partial derivatives,

$$\begin{aligned} \frac{\partial \mathbf{a}}{\partial \mathbf{x}} &= \begin{bmatrix} \partial a_i \\ \partial x_j \end{bmatrix} \\ &= \frac{1}{16} \begin{bmatrix} 3 & 3 & 3 & 6 \\ 3 & -1 & -1 & -2 \\ -1 & 3 & -1 & -2 \\ -1 & -1 & 3 & -2 \\ -1 & -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 0 & x_3 & x_2 \\ x_3 & 0 & x_1 \\ x_2 & x_1 & 0 \\ x_2x_3 & x_1x_3 & x_1x_2 \end{bmatrix}, \end{aligned} \tag{17}$$

hence, at each turning point (where $\partial L / \partial \mathbf{x} = \mathbf{0}$) we have from equation (14)

$$\begin{aligned} &\begin{bmatrix} f_0 & f_1 & f_2 & f_3 & f_4 \\ a_0 & a_1 & a_2 & a_3 & a_4 \end{bmatrix} \begin{bmatrix} 3 & 3 & 3 & 6 \\ 3 & -1 & -1 & -2 \\ -1 & 3 & -1 & -2 \\ -1 & -1 & 3 & -2 \\ -1 & -1 & -1 & 2 \end{bmatrix} \\ &\times \begin{bmatrix} 0 & x_3 & x_2 \\ x_3 & 0 & x_1 \\ x_2 & x_1 & 0 \\ x_2x_3 & x_1x_3 & x_1x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{18}$$

The three relations of equation (18) are rational functions in the three variables x_i . If we multiplied by the common denominators, we obtain three polynomials, each of degree 14 in the variables. Solving these simultaneously might be

feasible, however, we can simplify the system by imposing the additional constraints

$$q_1 = q_2 \leq q_3,$$

which constrain the edge lengths to satisfy the molecular clock, as illustrated in figure 1b. These constraints imply $x_1 = x_2 \geq x_3$. In our analysis below, we will explicitly impose the equality $x_1 = x_2$ to find the turning points. The inequality will need to be tested on any potential solution and, if it were not satisfied, a maximum could be sought on the boundary of the valid tree domain, where $x_1 = x_2 = x_3$.

The constraint $x_1 = x_2$ implies $a_1 = a_2$, so by replacing x_1 and a_1 , equation (18) reduces to

$$\begin{bmatrix} \frac{f_0}{a_0} & \frac{f_1 + f_2}{a_2} & \frac{f_3}{a_3} & \frac{f_4}{a_4} \end{bmatrix} \begin{bmatrix} 6 & 3 & 6 \\ 4 & -2 & -4 \\ -2 & 3 & -2 \\ -2 & -1 & 2 \end{bmatrix} \times \begin{bmatrix} x_3 & x_2 \\ 2x_2 & 0 \\ 2x_2x_3 & x_2^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{19}$$

From equation (12), each a_i is a polynomial in x_2, x_3 , so multiplying equation (18) by $a_0a_2a_3a_4$ gives two polynomial equations in the two variables x_2, x_3 and the observed frequencies f_j . We refer to these polynomial equations as E_1 and E_2 .

We now show that the system of two resulting polynomials $\{E_1, E_2\}$ has only finitely many solutions, all of which we can find. The major tool used here is the “resultant” of two polynomials. Let $f(x) = \sum_{i=0}^d a_i x^i$ and $g(x) = \sum_{j=0}^d b_j x^j$ be two polynomials in one variable, x . The resultant of f and g , denoted $Res(f, g, x)$, is a polynomial in the coefficients a_i and b_j of f and g , which is 0 whenever f and g have a common zero. The coefficients can themselves be unknowns or functions of other variables, in which case, the resultant replaces the two polynomials f and g by a single polynomial in one fewer variable.

Computing the resultant is a classical technique for eliminating one variable from two equations. There is an elegant formula for computing it due to Sylvester and another due to Bezout, which have been implemented in the computer algebra package Maple.

We can compute the resultant $ER = Res(E_1, E_2, x_3)$ of E_1 and E_2 with respect to x_3 . This eliminates x_3 from the equations and yields a single polynomial ER , in just x_2 and the parameters. The polynomial ER has the form:

$$ER = kf_0f_{12}f_3f_4x_2^{13}(3x_2 + 1)(2x_2^2 + x_2 + 1)(3x_2^2 + 1) \times (3x_2^2 + 3x_2 + 2)(x_2 - 1)^2(x_2 + 1)^3 \cdot P_0,$$

where P_0 is the degree 11 polynomial with 288 monomials displayed in the appendix and k is some large constant.

Theorem 2 The turning points of L (eq. 14) corresponding to realistic trees (namely, trees with positive edge lengths) are exactly the roots of P_0 .

Proof. The only factors in ER (eq. 20) that admit positive real roots are P_0 and $(x_2 - 1)$. However, $x_2 = 1$ implies $q_2 = 0$, which is not realistic. \square

Corollary 3 The Jukes-Cantor triplet has a finite number of ML points.

Proof. P_0 has at most 11 different solutions and, for each such solution, we can back substitute to obtain all the values of x_3 . \square

Maxima on the Boundaries of the Parameter Space

The realistic parameter space R is bounded by the constraints

$$0 \leq q_1 = q_2 \leq q_3 \leq \infty,$$

which implies

$$0 \leq x_3 \leq x_1 = x_2 \leq 1.$$

The analysis above finds maxima at turning points in the interior of R , it is also possible that the maximum value of L can be on the boundary, which is not a turning point. To find these we must consider the turning points of L when the parameters are constrained to the boundary.

We identify three boundary subspaces

$$\text{I : } x_3 = 0 \leq x_1 = x_2 \leq 1; \quad \text{II : } 0 \leq x_3 = x_1 = x_2 \leq 1; \\ \text{III : } 0 \leq x_3 \leq x_1 = x_2 = 1.$$

We must independently examine maxima on each boundary subspace. In each case, the likelihood function on the boundary can be described by a single parameter x_i , then solving $dL/dx_i=0$ gives a single polynomial constraint.

In case I, with $x_3 = 0, x = x_1 = x_2$, we find

$$\frac{dL}{dx} = 0 \Rightarrow Lx(3(f_0 + f_3)(1 - x^2) - (f_1 + f_2 + f_4)(1 + 3x^2)) = 0,$$

which has a single positive root

$$x = \left(1 - \frac{4}{3}(f_1 + f_2 + f_4)\right)^{\frac{1}{2}},$$

and a zero at $x = 0$.

In case II, with $x = x_1 = x_2 = x_3$, we find

$$\frac{dL}{dx} = 0 \Rightarrow Lx(1 - x)^2(9f_0(1 + x)(1 - x)(1 + x + 2x^2) + (f_1 + f_2 + f_3)(1 + 9x^2 + 6x^3)(1 - 3x)(1 + 2x) - 3f_4(1 + 9x^2 + 6x^3)(1 + x + 2x^2)) = 0,$$

which factorizes into a polynomial of degree 5, and $x(1 - x)^2$ which has zeros at $x = 0, 1$.

In case III, with $x = x_3, x_1 = x_2 = 1, L = 0$ (unless $f_1 = f_2 = f_4 = 0$, whence L is undefined).

Results and Discussion

Here we report experimental results obtained by employing our method on genomic sequences. We conclude by discussion and pointing out future research directions.

Results on Genomic Sequences

In order to evaluate our method, we tested it on three sets of real genomic sequences. In each case, we found a single solution to $P_0 = 0$ in the realistic parameter space,

for each of the three rooted triples. By evaluating the likelihood at each of these turning points, we found each was a maximum.

We looked at the natural killer cell receptor D gene on human, mouse, and rat (accession numbers AF260135, AF030313 and AF009511, respectively). We aligned the sequences using ClustalW (Thompson, Higgins, and Gibson 1994). Next, we computed the observed sequence spectrum F (eq. 21, as explained in *Obtaining the Maximum Likelihood Solution*).

$$F = \begin{bmatrix} 424 & 18 & 18 & 80 \\ 1 & 7 & 2 & 2 \\ 7 & 4 & 4 & 4 \\ 27 & 1 & 2 & 40 \end{bmatrix}. \quad (21)$$

We calculated the ML value for each of the three rooted trees under the model for the three species. The (rat, mouse, and human) tree was maximal, with edge lengths $q_1 = q_2 = 0.0197$ to rat and mouse and $q_3 = 0.1061$ to human, giving the log likelihood $\ln L = -870.2$. This result suggests the human-rodent split at 70 MYA and the mouse-rat split at 20 MYA, a result consistent with commonly accepted dates.

We also calculated the ML value for each of the three rooted trees for the beta actin gene, for the three species guinea pig, goose, and *Caenorhabditis elegans* (accession numbers AF508792, M26111, and NM_076440, respectively), finding the ((guinea pig, goose), *C. elegans*) tree maximal, with $q_1 = q_2 = 0.021819$ and $q_3 = 0.050188$ giving $\ln L = -1241.5$. Finally, we calculated the ML value for each of the three rooted trees for the histone gene of *Drosophila melanogaster*, *Hydra vulgaris*, and human (accession numbers AY383571, AY383572, and NM_002107, respectively), finding the ((*D. melanogaster*, *H. vulgaris*), Human) tree maximal, with $q_1 = q_2 = 0.001555$ and $q_3 = 0.012740$ with $\ln L = -86.835133$.

Each of the results above agree closely with the numerical values obtained using the popular phylogenetic reconstruction packages PHYLIP (Felsenstein 1995) and PAUP* (Swofford 1998), which use iterative methods to estimate the maxima. In each case, the likelihood function had a unique maximum in the parameter range.

Directions for Future Research

The progress made here brings up a number of open problems:

Our ML solutions are derived from the roots of a univariate, degree 11 polynomial. This implies that the number of ML solutions is finite. It would be interesting to explore the question of “uniqueness” of the solution. If this is the case, it will most likely follow from the existence of a single solution corresponding to a realistic tree, as in Chor, Khetan, and Snir (2003).

The Jukes-Cantor substitution model is a special case of the family of Kimura substitution models. It would be interesting to further extend the result in this paper for the other models (two and three parameters) of the Kimura family. It would be interesting to extend these results to rooted trees with “four leaves” under Jukes-Cantor model and a molecular clock. Here we have two different

topologies—the fork and the comb (Chor, Khetan, and Snir 2003). It is expected that such extension will face substantial technical difficulties.

Acknowledgments

Thanks to Joseph Felsenstein for his fruitful discussions, to Bernd Sturmfels for enlightening comments on this manuscript and informing us about his manuscript (Hosten, Khetan, and Sturmfels 2004), and to the two other reviewers for their constructive comments. This research was supported by the Israel Science Foundation grant 418/00.

Appendix

The polynomial P_0 is presented where the coefficients are functions of the observed pattern frequencies, f_i with $f_{12} = f_1 + f_2$. Using Maple we find:

$$\begin{aligned} P_0 = & (432f_4^3 + 216f_{12}^3 + 1080f_4^2f_{12} + 1728f_4f_{12}f_0 + 1296f_4^2f_0^2 \\ & + 432f_3^3 + 1080f_3^2f_0 + 216f_0^3 + 1080f_4^2f_0 + 1080f_3^2f_{12} \\ & + 864f_4f_0^2 + 864f_4f_{12}^2 + 648f_{12}f_0^2 + 1296f_4^2f_3 \\ & + 2160f_4f_3f_0 + 2160f_4f_3f_{12} + 1728f_0f_{12}f_3 + 648f_{12}^2f_0 \\ & + 864f_3f_{12}^2 + 864f_3f_0^2)x_2^{11} + (4212f_4^2f_3 + 2844f_3f_{12}^2 \\ & + 900f_3^3 + 3330f_4f_{12}^2 + 4788f_4f_3f_0 + 1656f_4^3 \\ & + 4878f_4f_{12}f_0 + 216f_0^3 + 1350f_{12}f_0^2 + 4068f_4^2f_{12} \\ & + 6894f_4f_3f_{12} + 1800f_3^2f_0 + 1116f_3f_0^2 + 1548f_4f_0^2 \\ & + 3456f_4f_3^2 + 3960f_0f_{12}f_3 + 2988f_4^2f_0 + 2826f_3^2f_{12} \\ & + 918f_{12}^3 + 2052f_{12}^2f_0)x_2^{10} + (4446f_4f_{12}^2 + 1566f_4f_3^2 \\ & + 2655f_3f_{12}^2 - 540f_4f_0^2 - 315f_{12}f_0^2 + 6435f_4f_3f_{12} \\ & - 162f_3^3 + 1449f_{12}^2f_0 - 486f_0^3 + 306f_4f_3f_0 - 1638f_3f_{12}^2 \\ & + 1323f_3^2f_{12} + 3123f_4f_{12}f_0 + 1890f_4^2f_0 + 5436f_4^2f_{12} \\ & + 2268f_4^3 + 288f_0f_{12}f_3 + 3996f_4^2f_3 + 1278f_{12}^3 \\ & - 1314f_3^2f_0)x_2^9 + (1866f_4^3 + 327f_4^2f_3 - 1203f_4f_{12}f_0 \\ & + 909f_4f_3f_{12} - 2346f_3^2f_{12} - 2088f_{12}f_0^2 - 3135f_4f_3^2 \\ & + 3558f_4f_{12}^2 - 3702f_3^2f_0 + 4542f_4^2f_{12} - 594f_0^3 \\ & - 2646f_3f_0^2 - 2313f_4f_0^2 + 882f_{12}^3 - 1650f_3^3 - 774f_{12}^2f_0 \\ & + 276f_3f_{12}^2 - 4746f_0f_{12}f_3 - 735f_{12}^2f_0 - 5760f_4f_3f_0)x_2^8 \\ & + (-1134f_3^3 + 3546f_4^2f_{12} + 2754f_4f_{12}^2 + 1326f_4^3 \\ & - 2868f_4f_{12}f_0 - 1062f_3f_{12}^2 - 954f_3^2f_0 - 3096f_0f_{12}f_3 \\ & - 1446f_{12}^2f_0 - 1950f_4^2f_3 + 534f_{12}^3 - 2580f_4f_3f_{12} \\ & - 3000f_3^2f_{12} - 1854f_4^2f_0 + 162f_0^3 - 3996f_4f_3^2 \\ & - 1152f_{12}f_0^2 - 4272f_4f_3f_0 - 1332f_4f_0^2 + 342f_3f_0^2)x_2^7 \\ & + (108f_{12}f_0^2 + 150f_{12}^3 - 900f_3^2f_{12} + 978f_4f_{12}^2 \\ & + 1842f_3^2f_0 + 486f_0^3 - 456f_4f_3f_0 - 3192f_4f_{12}f_0 \\ & + 36f_4f_0^2 - 1974f_4^2f_0 + 510f_3^3 + 450f_4^3 - 1308f_4f_3^2 \\ & - 3192f_4f_3f_{12} - 894f_3f_{12}^2 + 1818f_3f_0^2 - 2166f_4^2f_3 \end{aligned}$$

$$\begin{aligned}
& -1086f_{12}^2f_0 + 1278f_4^2f_{12} + 24f_0f_{12}f_3)x_2^6 + (570f_4^2f_{12} \\
& + 270f_0^3 + 696f_0f_{12}f_3 + 206f_4^3 + 738f_3f_0^2 + 1236f_4f_3f_0 \\
& - 1334f_4^2f_3 - 1002f_4^2f_0 - 684f_{12}^2f_0 - 1878f_4f_{12}f_0 \\
& + 122f_{12}^3 + 1014f_3^3 + 486f_4^2f_{12} - 320f_3f_{12}^2 + 828f_4f_0^2 \\
& + 920f_4f_3^2 + 542f_3^2f_{12} + 666f_{12}f_0^2 - 1846f_4f_3f_{12} \\
& + 1482f_3^2f_0)x_2^5 + (-6f_4^2f_{12} - 562f_4f_3f_{12} - 162f_3f_0^2 \\
& + 660f_0f_{12}f_3 + 348f_3^2f_{12} - 206f_3f_{12}^2 - 162f_{12}^2f_0 \\
& - 498f_4f_{12}f_0 - 344f_4^2f_3 + 648f_{12}f_0^2 + 1344f_4f_3f_0 \\
& + 666f_4f_0^2 - 26f_{12}^3 - 324f_4^2f_0 + 330f_3^3 + 1014f_4f_3^2 \\
& - 30f_4f_{12}^2 - 2f_4^3 + 222f_3^2f_0 - 54f_0^3)x_2^4 + (-138f_3^3 \\
& + 408f_0f_{12}f_3 + 26f_{12}^3 + 38f_4^2f_3 - 74f_3f_{12}^2 - 246f_3^2f_0 \\
& - 6f_4^3 - 132f_4f_{12}f_0 + 624f_4f_3f_0 + 180f_4f_0^2 + 22f_4f_{12}^2 \\
& + 40f_3^2f_{12} - 84f_4f_3f_{12} + 144f_{12}f_0^2 - 10f_4^2f_{12} + 30f_{12}^2f_0 \\
& - 270f_3f_0^2 - 114f_4^2f_0 + 220f_4f_3^2 - 162f_0^3)x_2^3 \\
& + (50f_4f_3f_{12} - 28f_4f_{12}^2 - 4f_{12}^3 + 74f_3^2f_{12} + 42f_4^2f_0 \\
& + 18f_4f_{12}f_0 - 2f_4^3 - 162f_3^2f_0 - 126f_3f_0^2 + 84f_4f_3f_0 \\
& - 30f_{12}^2f_0 + 26f_3f_{12}^2 - 54f_0^3 + 72f_4^2f_0 + 18f_4^2f_3 - 90f_3^3 \\
& - 18f_{12}f_0^2 - 26f_4^2f_{12} + 96f_0f_{12}f_3 - 28f_4f_3^2)x_2^2 \\
& + (-24f_0f_{12}f_3 + 9f_{12}f_0^2 - 15f_3f_{12}^2 - 54f_4f_3f_0 - 2f_4^3 \\
& + 4f_4f_{12}^2 + 2f_4^2f_{12} - 36f_3f_0^2 + 11f_4f_3f_{12} + 3f_{12}^2f_0 + 2f_4^2f_3 \\
& - 6f_4f_3^2 + 15f_3^2f_{12} - 12f_3^3 + 27f_4f_{12}f_0 - 48f_3^2f_0)x_2 \\
& + (f_4^2f_3 - 9f_4f_0^2 + 3f_4^2f_0 + 2f_3f_{12}^2 - 2f_3^2f_{12} + f_4f_3^2 \\
& + 6f_0f_{12}f_3 - 3f_4f_3f_{12} - 3f_4f_{12}f_0).
\end{aligned}$$

Literature Cited

- Aho, A. V., Y. Sagiv, T. G. Szymanski, and J. D. Ullman. 1981. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM J. Comput.* **10**:405–421.
- Catanese, F., S. Hosten, A. Khetan, and B. Sturmfels. 2004. The maximum likelihood degree. (<http://front.math.ucdavis.edu/math.AG/0406533>).
- Chor, B., M. Hendy, B. Holland, and D. Penny. 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. *Mol. Biol. Evol.* **17**:1529–1541. (Earlier version appeared in RECOMB 2000).
- Chor, B., M. Hendy, and D. Penny. 2001. Analytic solutions for three taxon mlmc trees with variable rates across sites. Pp. 204–213 in B. Moret and O. Gascuel, eds. *Lecture notes in Computer Science 2149 in Workshop on Algorithms in Bioinformatics 2001*. Springer-Verlag, Heidelberg, Germany.
- Chor, B., A. Khetan, and S. Snir. 2003. Maximum likelihood on four taxa phylogenetic trees: analytic solutions. Pp. 76–83 in W. Miller, M. Vingron, S. Istrail, P. Pevzner, and M. Waterman, eds. *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology (RECOMB)*. AMC Press, Berlin.
- Chor, B., and S. Snir. 2004. Molecular clock fork phylogenies: closed form analytic maximum likelihood solutions. *Syst. Biol.* **53**(6):963–967.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- . 1995. PHYLIP (phylogeny inference package). Distributed by the author, Department of Genetics, University of Washington, Seattle.
- Hendy, M. D., and D. Penny. 1993. Spectral analysis of phylogenetic data. *J. Classif.* **10**:5–24.
- Hendy, M. D., D. Penny, and M. A. Steel. 1994. Discrete Fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci. USA* **91**:3339–3343.
- Hendy, M. D., and S. Snir. 2005. Hadamard conjugation for the Kimura 3st model: combinatorial proof using pathsets. (<http://arxiv.org/abs/q-bio.PE/0505055>).
- Hosten, S., A. Khetan, and B. Sturmfels. 2004. Solving the likelihood equations. (<http://front.math.ucdavis.edu/math.ST/0408270>).
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–132 in H. N. Munro, ed. *Mammalian protein metabolism*. Academic Press, New York.
- Kimura, M. 1981. Estimation of evolutionary sequences between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* **78**:454–458.
- Neyman, J. 1971. Molecular studies of evolution: a source of novel statistical problems. Pp. 1–27 in S. Gupta and Y. Jackel, eds. *Statistical decision theory and related topics*. Academic Press, New York.
- Steel, M., M. D. Hendy, and D. Penny. 1998. Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results. *Discrete Appl. Math.* **88**:367–396.
- Steel, M. A., M. D. Hendy, L. A. Székely, and P. L. Erdős. 1992. Spectral analysis and a closest tree method for genetic sequences. *Appl. Math. Lett.* **5**:63–67.
- Sturmfels, B., and S. Sullivant. 2005. Toric ideals of phylogenetic invariants. *J. Comput. Biol.* **12**:204–228.
- Swofford, D. L. 1998. PAUP*beta. Sinauer Associates, Sunderland, Mass.
- Székely, L., P. L. Erdős, M. A. Steel, and D. Penny. 1993. A Fourier inversion formula for evolutionary trees. *Appl. Math. Lett.* **6**:13–17.
- Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalty and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4780.
- Yang, Z. 2000. Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Lond. B* **267**:109–116.

William Martin, Associate Editor

Accepted November 21, 2005