



Analytic solutions of maximum likelihood on forks of four taxa [☆]

Benny Chor ^{a,*}, Sagi Snir ^{b,1}

^a School of Computer Science, Tel-Aviv University, Tel-Aviv 39040, Israel

^b Department of Mathematics, University of California, Berkeley, CA 94720, USA

Received 19 July 2004; received in revised form 26 May 2005; accepted 4 April 2006

Abstract

This work deals with symbolic mathematical solutions to maximum likelihood on small phylogenetic trees. Maximum likelihood (ML) is increasingly used as an optimality criterion for selecting evolutionary trees, but finding the global optimum is a hard computational task. In this work, we give general analytic solutions for a family of trees with *four* taxa, two state characters, under a molecular clock. Previously, analytical solutions were known only for *three* taxa trees. The change from three to four taxa incurs a major increase in the complexity of the underlying algebraic system, and requires novel techniques and approaches. Despite the simplicity of our model, solving ML analytically in it is close to the limit of today's tractability.

Four taxa rooted trees have two topologies – the *fork* (two subtrees with two leaves each) and the *comb* (one subtree with three leaves, the other with a single leaf). Combining the properties of molecular clock fork trees with the Hadamard conjugation, and employing the symbolic algebra software Maple, we derive a number of topology dependent identities. Using these identities, we substantially simplify the system of polynomial equations for the fork. We finally employ the symbolic algebra software to obtain *closed form* analytic solutions (expressed parametrically in the input data).

© 2006 Elsevier Inc. All rights reserved.

[☆] Research supported by ISF grant 418/00. Part of these results were presented at the RECOMB 2003 conference in Berlin.

* Corresponding author.

E-mail addresses: benny@cs.tau.ac.il (B. Chor), ssagi@math.berkeley.edu (S. Snir).

¹ The research of S. Snir done while at the department of Computer Science, Technion, Haifa 32000, Israel.

Keywords: Maximum likelihood; Phylogenetic trees; Molecular clock; Analytic solutions; Hadamard conjugation; Symbolic manipulation

1. Introduction

The study of evolution and the construction of phylogenetic (evolutionary) trees are classical subjects in biology. DNA sequences from a variety of organisms are rapidly accumulating, providing the data to a number of sequence based approaches for phylogenetic trees reconstruction. Given a set of n aligned *sequences*, one per species, the goal is to find the best explanation for the data within the model space. Among tree reconstruction approaches, maximum likelihood [5] is increasingly used as an optimality criterion for inferring trees. The goal is to produce a weighted tree (the weights are parameters of the substitution model for each edge) that maximizes the likelihood of generating the observed sequences. The *molecular clock* assumption is independent of maximum likelihood. It postulates that phylogenetic trees are rooted, and the length of the path from each leaf to the root is the same (Fig. 1).

Algorithms for finding maximum likelihood (ML) trees are computationally intensive, but for tractable cases ML is the criteria of choice. Because no general analytical solution is available, numeric techniques (such as hill climbing or expectation maximization) are used in order to find optimal likelihood values for any given tree. The first to consider *analytical solutions* for simple substitution models with a small number of taxa was Yang, who worked on three taxa with two state characters under molecular clock [16]. Yang calls this ‘the simplest phylogeny estimation problem’, but adds that it ‘has many of the conceptual and statistical complexities involved in phylogenetic estimation’. The solution of Yang was generalized and its derivation was simplified by Chor et al. [2] using the Hadamard conjugation of Hendy et al. [8,9], together with convexity arguments.

In this work we retain the symmetric two states model of Neyman [13], as used in [16,2] under molecular clock, but increase the number of taxa to four. The change from three to four taxa incurs a major increase in the complexity of the underlying system of polynomial equations, and requires novel techniques and approaches. Our starting point, like [1], is to formulate the ML problem as one of constrained optimization, and express it in terms of Lagrange coefficients. We use the Hadamard conjugation [8,9], together with the symbolic algebra software Maple,

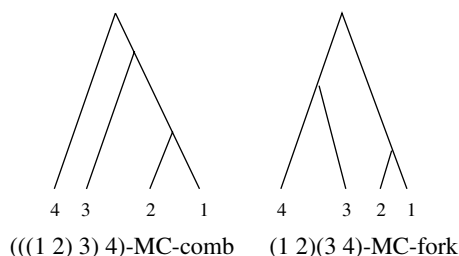


Fig. 1. The fork and comb – two rooted topologies on four taxa.

to simplify the resulting system of polynomial equations. This yields a system of nine degree 5 polynomials in nine variables. This system is substantially more complex than the three taxa system [2], and is not solvable by current techniques. (We note that the analytical solutions in [1] were all for special cases where at least 2 out of the 7 input parameters are 0.)

There are two families of rooted topologies on four taxa: Topologies with two taxa in each subtree of the root, which we call *fork* topologies, and topologies where one subtree of the root has three taxa, which we call *comb* topologies. In this work we focus on the fork topology. We first derive identities that help simplify the general system of equations for the fork topology. By using these identities, we obtained a simplified system of three polynomials in three variables with degree four. At this stage, the system became simple enough to enable its solution – not manually, but by using computer algebra tools (e.g., Maple). This leads to the derivation of *closed form* analytical solutions, expressed as rational functions in the input parameters. This solution is unique, implying a single local and global ML point.

Analytic solutions do reveal properties of the maximum likelihood points that are not obtainable numerically. Using them we can show here that every molecular clock fork tree has a unique (global and local) ML point. It stands in contrast to the situation *without* the molecular clock hypothesis, where uniqueness does not always hold, as proved in [14,1]. This is the main contribution of this work as this property is of great importance in validating quartet likelihood methods, where the ML point is estimated using hill climbing.

The application of sophisticated mathematical and algebraic tools to derive optimal solutions in the area of mathematical phylogenetics has attracted considerable interest lately. This work is one of the first steps in this direction. Here, the emphasis is on the mathematical background and techniques used to derive the results, whereas in a companion paper submission [4], phylogenetic aspects are emphasized.

The remainder of this work is organized as follows: Section 2 introduces definitions, notations, and briefly explains the Hadamard conjugation [8,9] and its relation to maximum likelihood. In Section 3, technical properties of general trees, and particular properties of the molecular clock fork, are developed and proved. In Section 4 we derive the closed form solution to the ML molecular clock forks, while in Section 5 we give some concluding remarks and directions for further research.

2. Definitions, notations, and the Hadamard conjugation

In this section we define the model of substitution we use, introduce useful notations, and briefly describe the Hadamard conjugation.

2.1. Definitions and notations

We start with a tree labeling notation that will be useful for the rest of the work. For simplicity we use four taxa, but the definitions extend to any n . A *split* of the species is any partition of $\{1, 2, 3, 4\}$ into two disjoint subsets. We will identify each split by the subset which does not contain 4 (in general n), so that for example the split $\{\{1, 2\}, \{3, 4\}\}$ is identified by the subset $\{1, 2\}$. For brevity, to label objects subscribed by a split, we concatenate the members of the split. Each edge e

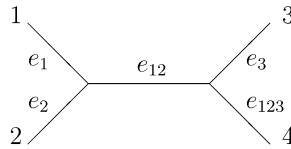


Fig. 2. The tree $T = (12)(34)$ and its edges.

of a phylogenetic tree T induces a split of the taxa, i.e., the cut induced by removing e . We denote the edge e by the cut it induces. For instance, the central edge of the tree $T = (12)(34)$ induces the split $\{\{1, 2\}, \{3, 4\}\}$, that is identified by the subset $\{1, 2\}$ and therefore this edge is denoted e_{12} . Thus, $E(T) = \{e_1, e_2, e_{12}, e_3, e_{123}\}$ (see Fig. 2).

In Neyman model [13] of 2-state evolution with symmetric probabilities of substitution, each character at a species admits one out of two states, without loss of generality $\{x, y\}$. Hence, a character evolving along an evolutionary tree T with n leaves, induces a split pattern between the leaves admitting the state x and y .

In this 2-state model, the length of an edge q_e , $e \in E(T)$ in the tree T is defined as the expected number of substitutions (changes) per site along that edge. Given the edge lengths of T : $\mathbf{q} = [q_e]_{e \in E(T)}$ ($0 \leq q_e < \infty$), the probability of generating an α -split pattern ($\alpha \subseteq \{1, \dots, n-1\}$) is well defined. Denote this probability by $s_\alpha = Pr(\alpha\text{-split} | T, \mathbf{q})$. Using the same indexing scheme as above, we define the *expected sequence spectrum* (expected spec) $\mathbf{s} = [s_\alpha]_{\alpha \subseteq \{1, \dots, n-1\}}$.

The *edges length spectrum* (edges spec) of a tree T with n leaves is the 2^{n-1} dimensional vector $\mathbf{q} = [q_\alpha]_{\alpha \subseteq \{1, \dots, n-1\}}$, defined for any subset $\alpha \subseteq \{1, \dots, n-1\}$ by

$$q_\alpha = \begin{cases} q_e & \text{if } e \in E(T) \text{ induces the split } \alpha, \\ - \sum_{e \in E(T)} q_e & \text{if } \alpha = \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

2.2. Hadamard conjugation

The Hadamard conjugation [8,9] is an invertible transformation that specifies a relation between the expected sequence spectrum \mathbf{s} and the edge lengths spectrum \mathbf{q} of the tree. In other words, the transformation links the probabilities of site substitutions on edges of an evolutionary tree T , to the probabilities of obtaining each possible combination of characters. The Hadamard conjugation is applicable to a number of site substitution models: Neyman 2 state model, Jukes–Cantor model [10], and Kimura 2ST and 3ST models [11] (the last three models are applicable to four states characters, such as DNA or RNA). For these models, the transformation yields a powerful tool which greatly simplifies and unifies the analysis of phylogenetic data, and in particular the analytical approach to ML.

Definition 1. A *Hadamard matrix* of order ℓ is an $\ell \times \ell$ matrix A with ± 1 entries such that $A^t A = \ell I_\ell$.

We will use a special family of Hadamard matrices, called Sylvester matrices in MacWilliams and Sloan [12], defined inductively for $n \geq 0$ by

$$H_0 = [1] \quad \text{and} \quad H_{n+1} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}.$$

For example,

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad \text{and} \quad H_2 = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

It is convenient to index the rows and columns of H_n by lexicographically ordered subsets of $\{1, \dots, n\}$. Denote by $h_{\alpha,\gamma}$ the (α, γ) entry of H_n , then $h_{\alpha,\gamma} = (-1)^{|\alpha \cap \gamma|}$. This implies that H_n is symmetric, namely $H_n^t = H_n$, and thus by the definition of Hadamard matrices $H_n^{-1} = \frac{1}{2^n} H_n$.

Proposition 1 [8]. *Let T be a phylogenetic tree on n leaves with finite edge lengths ($q_e < \infty$ for all $e \in E(T)$). Assume that sites mutate according to a symmetric substitution model, with equal rates across sites. Let \mathbf{s} be the expected sequence spectrum. Then*

$$\mathbf{s} = \mathbf{s}(\mathbf{q}) = H_{n-1}^{-1} \exp(H_{n-1}\mathbf{q}), \tag{1}$$

where the exponentiation function $\exp(x) = e^x$ is applied element-wise to the vector $\rho = H_{n-1}\mathbf{q}$. That is, for $\alpha \subseteq \{1, \dots, n-1\}$, $s_\alpha = 2^{-(n-1)} \sum_\gamma h_{\alpha,\gamma} (\exp(\sum_\delta h_{\gamma,\delta} q_\delta))$.

This transformation is called the Hadamard conjugation.

Definition 2. A vector $\hat{\mathbf{s}} \in \mathbb{R}^{2^{n-1}}$ satisfying $\sum_{\alpha \subseteq \{1, \dots, n-1\}} \hat{s}_\alpha = 1$, $\hat{\mathbf{s}} \geq 0$, and $H_{n-1}\hat{\mathbf{s}} > \mathbf{0}$ is called conservative.

For conservative data $\hat{\mathbf{s}}$, the Hadamard conjugation is invertible, yielding:

$$\gamma = \gamma(\hat{\mathbf{s}}) = H_{n-1}^{-1} \ln(H_{n-1}\hat{\mathbf{s}}),$$

where the \ln function is applied element-wise to the vector $H_{n-1}\hat{\mathbf{s}}$. We note that γ is not necessarily the edge length spectrum of any tree. On the other hand, the expected sequence spectrum of any tree T is always conservative.

3. Maximum likelihood on four taxa trees

In this section, we describe how the system of equations is set up, and how molecular clock is used to simplify it. We begin with the formulation of the general maximum likelihood problem as a constrained optimization problem, and the resulting system of polynomial equations. Then we use the molecular clock model properties together with the Hadamard conjugation to derive a number of identities relevant to the fork. Using the derived identities, the system is substantially simplified in both cases, to the point where analytic closed form solutions can be derived.

3.1. General ML system

Given an input data ψ of n aligned, two-states sequences as rows, every column (site) in ψ induces a split. Let \hat{s}_α be the number of columns in ψ inducing the split α ($\alpha \subseteq \{1, \dots, n-1\}$). The vector $\hat{s} = [\hat{s}_\alpha]_{\alpha \subseteq \{1, \dots, n-1\}}$, indexed analogously to the expected sequence spectrum, is called the *observed sequence spectrum* (observed spec). The likelihood of producing the observed spec \hat{s} , given the expected spec s equals

$$L(\hat{s}|\mathbf{s}) = \prod_{\alpha \subseteq \{1, \dots, n-1\}} Pr(\alpha\text{-split}|\mathbf{s})^{\hat{s}_\alpha} = \prod_{\hat{s}_\alpha > 0} s_\alpha^{\hat{s}_\alpha}.$$

In the specific case of a four taxa *unrooted* tree:

$$L(\hat{s}|\mathbf{s}) = s_\emptyset^{\hat{s}_\emptyset} \cdot s_1^{\hat{s}_1} \cdot s_2^{\hat{s}_2} \cdot s_{12}^{\hat{s}_{12}} \cdot s_3^{\hat{s}_3} \cdot s_{13}^{\hat{s}_{13}} \cdot s_{23}^{\hat{s}_{23}} \cdot s_{123}^{\hat{s}_{123}}.$$

Without loss of generality, we describe the systems for the unrooted trees corresponding to (12)(34) molecular clock fork. Fig. 3 shows the molecular clock fork $T = (12)(34)$ on the left, and on the right its equivalent unrooted tree. Both versions have e_{12} as their ‘central’ edge. The expected spec s of these trees can be represented as a point in \mathbb{R}^8 whose edge lengths satisfy

- $q_\alpha(s) \geq 0$ for all $\alpha \in E(T)$.
- $q_\alpha(s) = 0$ for all $\alpha \notin E(T)$, $\alpha \neq \emptyset$.
- $q_\emptyset(s) = -\sum_{\alpha \neq \emptyset} q_\alpha(s) < 0$.

Thus, q_{13} and q_{23} must equal zero, and we can formulate the problem of maximizing the likelihood function as a constrained maximization problem: Find the maximum value of L under the constraints $q_{13}(\mathbf{s}) = 0$ and $q_{23}(\mathbf{s}) = 0$. For $\alpha \neq \emptyset$, we say that q_α is out of the boundary if $q_\alpha < 0$. We can ignore the ‘out of boundary’ requirements when maximizing the likelihood, provided we eventually verify that the resulting ML tree is indeed reasonable, namely all its edges are non-negative (inside the boundary). The approach taken in [1] is to initially express the set of critical points using Lagrange multipliers. By Proposition 1, every q_α is a function of the expected spec s , so we seek the point or points where

$$\nabla L = \lambda_1 \nabla q_{13}(\mathbf{s}) + \lambda_2 \nabla q_{23}(\mathbf{s}).$$

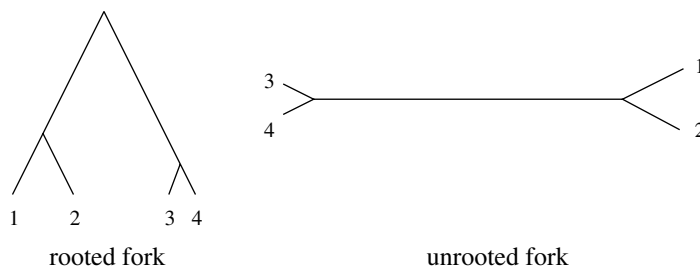


Fig. 3. Rooted layout of the molecular clock fork $T = (12)(34)$ (left), and its unrooted version (right).

This gives rise to a system of ten degree 5 polynomial equations in ten variables: The eight s_α variables, and two additional ‘Lagrange’ variables (λ_1 and λ_2). We emphasize that the eight \hat{s}_α are not variables – they are parameters determined by the four *input* sequences. (For brevity, we will use q_α and not $q_\alpha(\mathbf{s})$ in the sequel.)

Even with the simple model of $n = 4$, 2-state characters, extensive computer algebra is required to solve the equations describing the turning points of L . As the complexity of these equations will grow exponentially with the number of taxa n , it is unlikely that that current computer algebra tools can solve more general systems, beyond $n = 5$ taxa or under more general models of 4-state substitution with $n > 4$. Therefore, the key for deriving analytical solutions for the specific fork topology, is to combine the Hadamard conjugation with properties of the molecular clock structure. This in turn, provides a number of identities on the fork topology (also known as *invariants*). Using the derived identities, the system is substantially simplified. For the fork, we have $q_1 = q_2$ and $q_3 = q_{123}$ (see Fig. 4). We emphasise that the system of equations does not take explicitly into account *inequalities* like $q_{12} \geq 0$ or $q_{12} + \min(q_{123}, q_1) > \max(q_1, q_{123})$. The system is hard enough to solve as it is. These inequalities need to be checked on any putative solution.

3.2. Simplifying identities

The key to our simplifications is the use of lengths relations among the edges (the \mathbf{q} variables), which follow from molecular clock, in order to derive identities on the expected spec variables (the \mathbf{s} variables). The following relation on the expected spec variables is proved in [2].

Theorem 1 [2]. *Let i and j be sister taxa in a phylogenetic tree T with n leaves and edge weights \mathbf{q} . Let \mathbf{s} be the expected spec, such that $\mathbf{s} = H^{-1} \ln(H\mathbf{q})$; then $q_i = q_j$ implies $s_i = s_j$ and $q_i > q_j$ implies $s_i > s_j$.*

Under a molecular clock, the four taxa molecular clock fork has two pairs of sister taxa i, j and k, ℓ , such that $q_i = q_j$ and $q_k = q_\ell$. The next theorem is a generalization of the previous one, yielding one additional identity for the molecular clock fork.

Theorem 2. *Consider a tree T on n leaves, with two sister taxa i and j such that $q_i = q_j$ (see Fig. 5). Let \mathbf{s} be the expected spec, such that $\mathbf{s} = H^{-1} \ln(H\mathbf{q})$: Then for every $\alpha \subseteq \{1, 2, \dots, n - 1\} \setminus \{i, j\}$, $s_{\alpha \cup \{i\}} = s_{\alpha \cup \{j\}}$.*

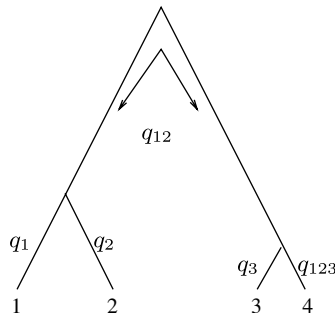


Fig. 4. In the (12)(34) molecular clock fork, $q_1 = q_2$ and $q_3 = q_{123}$.

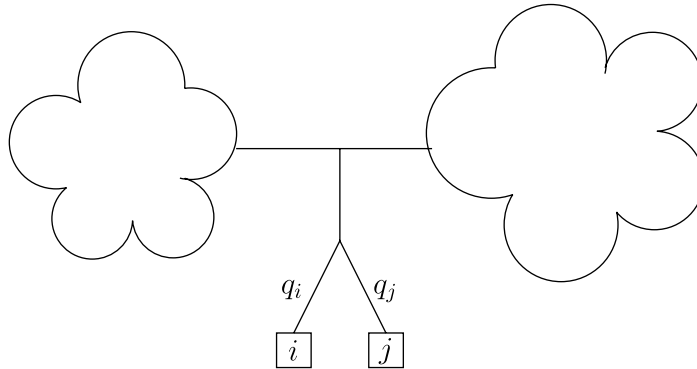


Fig. 5. A general tree with two sister taxa i and j s.t. $q_i = q_j$.

Proof. The proof is a symmetry argument, saying that i and j are interchangeable. More formally, since i and j are sister taxa in T and $q_i = q_j$, then for every other taxa k in T ($k \neq i, j$) the length of the tree paths from i to k and from j to k is the same. By the definition of the edge spec, this implies that for every $\beta \subseteq \{1, 2, \dots, n - 1\} \setminus \{i, j\}$, $q_{\beta \cup \{i\}} = q_{\beta \cup \{j\}}$. This means that any function of the edge spec, \mathbf{q} , is invariant under interchange of i and j . In particular, $s_{\alpha \cup \{i\}} = s_{\alpha \cup \{j\}}$. \square

Claim 3. Let T be a (1, 2)(3, 4) molecular clock fork. Then the expected sequence spectrum of T , satisfies the relation

$$s_3 = s_{13}(1 - 2s_1 - 2s_{13}) / (2s_1 + 2s_{13}).$$

Proof. Substituting $s_1 = s_2, s_{123} = s_3$ and $s_{13} = s_{23}$ in Proposition 1, we get

$$q_{13} = -\frac{1}{8}(\ln(1 - 4s_1 - 4s_{13}) + \ln(1 - 4s_3 - 4s_{13}) - \ln(1 - 4s_1 - 4s_3)).$$

By equating q_{13} to zero, taking exponent, and multiplying through by the denominator, we get $(1 - 4s_1 - 4s_3) = (-1 + 4s_1 + 4s_{13})(-1 + 4s_3 + 4s_{13})$. Arithmetic manipulation yields the claimed equality. \square

The next technical claim deals with conservative points $\mathbf{s} \in \mathbb{R}^{\otimes}$ (namely $H_{n-1}\mathbf{s} > 0$) satisfying $s_1 = s_2, s_{13} = s_{23}$, and $\sum_{\alpha \subseteq \{1,2,3\}} s_\alpha = 1$. (These points need not be the expected spec of a tree.) This technical claim will be useful in simplifying the system of polynomial equations that we solve in Section 4.

Claim 4. Let $\mathbf{s} = (s_\emptyset, s_1, s_2, s_{12}, s_3, s_{13}, s_{23}, s_{123}) \in \mathbb{R}^{\otimes}$ be a conservative point satisfying $s_1 = s_2$ and $s_{13} = s_{23}$, and let $\mathbf{q} = H^{-1} \ln H_{n-1}\mathbf{s}$. Then \mathbf{s} satisfies $q_{13}(\mathbf{s}) = q_{23}(\mathbf{s})$.

Proof. By the Hadamard conjugation we get:

$$4(q_{23} - q_{13}) = \ln(1 - 2s_1 - 2s_{12} - 2s_{123} - 2s_{13}) - \ln(1 - 2s_2 - 2s_{12} - 2s_{23} - 2s_{123}) - \ln(1 - 2s_1 - 2s_{12} - 2s_3 - 2s_{23}) + \ln(1 - 2s_2 - 2s_{12} - 2s_3 - 2s_{13}).$$

By the assumption $s_1 = s_2$ and $s_{13} = s_{23}$, so $(1 - 2s_1 - 2s_{12} - 2s_{123} - 2s_{13})$ (the first term) equals $(1 - 2s_2 - 2s_{12} - 2s_{23} - 2s_{123})$ (the second term) and $(1 - 2s_1 - 2s_{12} - 2s_3 - 2s_{23})$ (third term) equals $(1 - 2s_2 - 2s_{12} - 2s_3 - 2s_{13})$ (fourth term). This implies $q_{13}(\mathbf{s}) - q_{23}(\mathbf{s}) = 0$. \square

4. Solving the molecular clock fork

In this section we develop the analytic solutions for the ML molecular clock fork.

Theorem 5. *Let $\hat{\mathbf{s}} = (\hat{s}_\emptyset, \hat{s}_1, \hat{s}_2, \hat{s}_{12}, \hat{s}_3, \hat{s}_{13}, \hat{s}_{23}, \hat{s}_{123})$ be the observed spec, and let $c = \sum_\alpha \hat{s}_\alpha$. Then the expected spec of the ML (12)(34) molecular clock fork equals:*

$$\begin{aligned}
 s_{13} = s_{23} &= (\hat{s}_{23}^2 + \hat{s}_1\hat{s}_{23} + \hat{s}_2\hat{s}_{23} + \hat{s}_{123}\hat{s}_{23} + 2\hat{s}_{13}\hat{s}_{23} + \hat{s}_3\hat{s}_{23} + \hat{s}_{13}\hat{s}_3 \\
 &\quad + \hat{s}_{13}^2 + \hat{s}_{13}\hat{s}_{123} + \hat{s}_2\hat{s}_3 + \hat{s}_2\hat{s}_{123} + \hat{s}_2\hat{s}_{13} + \hat{s}_1\hat{s}_3 + \hat{s}_1\hat{s}_{123} + \hat{s}_1\hat{s}_{13})/2c^2, \\
 s_{12} &= \frac{(c - \hat{s}_2 - \hat{s}_1 - \hat{s}_{23} - \hat{s}_{13})(c - \hat{s}_3 - \hat{s}_{23} - \hat{s}_{123} - \hat{s}_{13})\hat{s}_{12}}{c^2(c - \hat{s}_1 - \hat{s}_2 - \hat{s}_3 - \hat{s}_{123} - \hat{s}_{13} - \hat{s}_{23})}, \\
 s_1 = s_2 &= -(\hat{s}_{23}^2 - c\hat{s}_{23} + \hat{s}_2\hat{s}_{23} + \hat{s}_{123}\hat{s}_{23} + \hat{s}_3\hat{s}_{23} + \hat{s}_1\hat{s}_{23} + 2\hat{s}_{13}\hat{s}_{23} - c\hat{s}_{13} + \hat{s}_{13}\hat{s}_3 \\
 &\quad + \hat{s}_{13}^2 + \hat{s}_{13}\hat{s}_{123} + \hat{s}_2\hat{s}_3 + \hat{s}_2\hat{s}_{123} - \hat{s}_2c + \hat{s}_2\hat{s}_{13} + \hat{s}_1\hat{s}_3 + \hat{s}_1\hat{s}_{123} - \hat{s}_1c + \hat{s}_1\hat{s}_{13})/2c^2 \\
 s_3 = s_{123} &= -(\hat{s}_{13}\hat{s}_1 + \hat{s}_{23}\hat{s}_1 + \hat{s}_3\hat{s}_1 + \hat{s}_{123}\hat{s}_1 - \hat{s}_{23}c - \hat{s}_{123}c - \hat{s}_3c - \hat{s}_{13}c + \hat{s}_{123}\hat{s}_{23} + \hat{s}_{123}\hat{s}_{13} + \hat{s}_3\hat{s}_{23} \\
 &\quad + \hat{s}_3\hat{s}_{13} + \hat{s}_{123}\hat{s}_2 + \hat{s}_{23}\hat{s}_2 + \hat{s}_{23}^2 + \hat{s}_{13}\hat{s}_2 + 2\hat{s}_{13}\hat{s}_{23} + \hat{s}_{13}^2 + \hat{s}_3\hat{s}_2)/2c^2.
 \end{aligned}$$

Proof. The (12)(34) molecular clock fork satisfies $q_1 = q_2$ and $q_3 = q_{123}$, therefore by [Theorem 1](#), $s_1 = s_2$ and $s_3 = s_{123}$. By [Theorem 2](#), $q_1 = q_2$ implies $s_{13} = s_{23}$. Substituting $s_2 = s_1$, $s_{23} = s_{13}$ and $s_{123} = s_3$, our likelihood function becomes:

$$L(\mathbf{s}|\hat{\mathbf{s}}) = s_\emptyset^{\hat{s}_\emptyset} \cdot s_1^{\hat{s}_1 + \hat{s}_2} \cdot s_{12}^{\hat{s}_{12}} \cdot s_3^{\hat{s}_3 + \hat{s}_{123}} \cdot s_{13}^{\hat{s}_{13} + \hat{s}_{23}}.$$

By [Claim 3](#), $q_{13} = 0$ implies $s_3 = s_{13}(1 - 2s_1 - 2s_{13})/(2s_1 + 2s_{13})$, which we substitute for s_3 in order to satisfy the constraint of a (12)(34) tree. This also eliminates the need to use the Lagrange multiplier corresponding to this constraint. Now since $s_1 = s_2$ and $s_{13} = s_{23}$ and by definition \mathbf{s} is conservative, so \mathbf{s} satisfies the conditions of [Claim 4](#), and thus $q_{13} = q_{23}$, so $q_{23} = 0$ as well. By this way the two constraints are satisfied, and we need not use any of the Lagrange multipliers. Eventually, we use the former identities $s_1 = s_2$ and $s_{13} = s_{23}$ to substitute for s_2 and s_{23} and we remain with a likelihood function L with only three variables (s_1, s_{12}, s_{13}).

Therefore, in order to look for critical points of L we should just equate the partial derivatives of L with respect to these remaining variables to zero.

The final step is to use the fact that the s_α variables add up to 1 and substitute $s_\emptyset = 1 - \sum_{\alpha \subseteq \{1,2,3\} \setminus \emptyset} s_\alpha$. The likelihood function becomes

$$L(\mathbf{s}|\hat{\mathbf{s}}) = \left(\frac{s_1 - 2s_1^2 - 2s_1s_{13} - s_{12}s_1 - s_{12}s_{13}}{s_1 + s_{13}} \right)^{(c-\hat{s}_1-\hat{s}_2-\hat{s}_3-\hat{s}_{123}-\hat{s}_{12}-\hat{s}_{13}-\hat{s}_{23})} \cdot s_{12}^{\hat{s}_{12}} s_{13}^{(\hat{s}_{13}+\hat{s}_{23})} s_1^{(\hat{s}_1+\hat{s}_2)} \left(\frac{s_{13}(1 - 2s_1 - 2s_{13})}{2(s_1 + s_{13})} \right)^{(\hat{s}_3+\hat{s}_{123})},$$

where $c = \sum_{\alpha \subseteq \{1, \dots, 3\}} \hat{s}_\alpha$.

This is a polynomial in three free variables s_1, s_{12}, s_{13} and eight (given) parameters. As we argued before, each ML point will be a critical point if $\nabla L = \mathbf{0}$, namely will satisfy the three polynomial equations:

$$\frac{\partial L}{\partial s_1} = 0, \quad \frac{\partial L}{\partial s_{12}} = 0, \quad \frac{\partial L}{\partial s_{13}} = 0.$$

The partial derivative with respect to s_{12} is

$$\begin{aligned} \frac{\partial L}{\partial s_{12}} = & -\hat{s}_{12}s_1 + 2\hat{s}_{12}s_1^2 + 2\hat{s}_{12}s_1s_{13} + cs_{12}s_1 + cs_{12}s_{13} - \hat{s}_1s_{12}s_1 - \hat{s}_1s_{12}s_{13} - \hat{s}_2s_{12}s_1 - \hat{s}_2s_{12}s_{13} \\ & - \hat{s}_3s_{12}s_1 - \hat{s}_3s_{12}s_{13} - \hat{s}_{123}s_{12}s_1 - \hat{s}_{123}s_{12}s_{13} - \hat{s}_{13}s_{12}s_1 - \hat{s}_{13}s_{12}s_{13} - \hat{s}_{23}s_{12}s_1 - \hat{s}_{23}s_{12}s_{13}. \end{aligned}$$

The other two derivatives are rather lengthy, and are easily reproducible with Maple, so we omit them here. In order to solve the system, so we applied Maple ‘solve’ to express three free variables s_1, s_{12}, s_{13} as rational functions in the input parameters $\hat{\mathbf{s}}$. The solutions appear in the statement of the theorem. The other variables are obtained by back substitutions. \square

The question of uniqueness of the ML point for phylogenetic analysis has raised considerable interest in the past [7,15,14,1]. It is now known that even four taxa ML trees exhibit datasets giving rise to multiple ML points [14,1]. In contrast, our result implies uniqueness for the ML molecular clock fork.

Corollary 6. *Each molecular clock fork topology has a unique local and global maximum likelihood point.*

It is known that ML is *consistent*, namely with high probability, for long enough input sequences, the correct tree is the tree maximizing the likelihood [6, Chapter 16]. Therefore, if the species under study evolved according to a molecular clock fork tree and the number of sites is sufficiently large, then we would expect the ML point to be a close approximation of the true tree point. The true tree is conservative (i.e., $\mathbf{q} > 0$), so in particular, we expect the ML point to be conservative. In the rare cases where we find that the ML point is not conservative, then this is not a realistic solution, and because the likelihood function is convex, the maximum point will be on a boundary.

Our results were stated for the ML (12)(34) molecular clock fork. By permuting the labels, they apply to the other two molecular clock forks as well.

5. Concluding remarks

By applying novel algebraic techniques, we extended the state of the art in analytic solutions of ML to four taxa under molecular clock in the 2-states model. We believe this is a significant step

in this area of research. We remark that the use of Hadamard conjugation in this context seems necessary. Not only does it allow variable elimination, but the resulting system is substantially simplified. Alternatively, it is possible to formulate the (12)(34)-molecular clock–fork as a system with only three variables, p_1 for the two pendant edges of taxa 1 and 2, p_3 for 3 and 4, and p_{12} for the central edge (see Fig. 3). However, this formulation yields a polynomial system of total degree 12, with up to 550 monomials per equation. This should be contrasted with the system of total degree 4 and up to 75 monomials that we get for the molecular clock–fork. It is this difference that enabled us to bring the problem to that degree of complexity that is solvable, when employing symbolic algebra tools.

The next step in this line of research is to extend this result to the other molecular clock four taxa topology – the comb (see Fig. 1). It has one fewer of the simple equalities $q_i = q_j$, and has a more complicated equality instead. This small change makes the resulting system substantially harder to solve. An analytic solution for this more involved case was obtained recently [3]. Like the fork, the comb also has a unique ML solution. But unlike the fork, the comb does not have a closed form solution (rational functions of the input parameter). Instead, the solution is expressed as the root of a degree nine polynomial.

Acknowledgements

We would like to thank Mike Hendy, Mike Steel, Shmuel Onn and Ziheng Yang for very helpful discussions and for comments on earlier versions of this work. We also wish to thank the anonymous referee for very helpful comments and corrections.

References

- [1] B. Chor, M. Hendy, B. Holland, D. Penny, Multiple maxima of likelihood in phylogenetic trees: an analytic approach, *MBE* 17 (10) (2000) 1529.
- [2] B. Chor, M. Hendy, D. Penny, Analytic solutions for three taxon mlmc trees with variable rates across sites, in: *WABI*, 2001.
- [3] B. Chor, A. Khetan, S. Snir, Maximum likelihood on molecular clock comb: analytic solutions, *J. Comput. Biol.* 13 (3) (2006) 819 (Earlier version in *RECOMB* 2003, Berlin).
- [4] B. Chor, S. Snir, Maximum likelihood molecular clock forks: closed form analytic solutions, *Syst. Biol.* 53 (6) (2004) 963.
- [5] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, *J. Mol. Evol.* 17 (1981) 368.
- [6] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, 2003.
- [7] K. Fukami, Y. Tateno, On the uniqueness of the maximum likelihood method for estimating molecular trees: uniqueness of the likelihood point, *J. Mol. Evol.* 28 (1989) 460.
- [8] M.D. Hendy, D. Penny, Spectral analysis of phylogenetic data, *J. Classif.* 10 (1993) 5.
- [9] M.D. Hendy, D. Penny, M. Steel, Discrete Fourier analysis for evolutionary trees, *Proc. Natl. Acad. Sci. USA* 91 (1994) 3339.
- [10] T. Jukes, C. Cantor, Evolution of protein molecules, in: H. Munro (Ed.), *Mammalian Protein Metabolism*, Academic Press, New York, 1969, p. 21.
- [11] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University, Cambridge, 1983.
- [12] F. MacWilliams, N. Sloan, *The Theory of Error-Correcting Codes*, Elsevier Science Publishers, NorthHolland, 1977.

- [13] J. Neyman, Molecular studies of evolution: a source of novel statistical problems, in: S. Gupta, Y. Jackel (Eds.), *Statistical Decision Theory and Related Topics*, Academic Press, New York, 1971, p. 1.
- [14] M. Steel, The maximum likelihood point for a phylogenetic tree is not unique, *Syst. Biol.* 43 (4) (1994) 560.
- [15] E. Tillier, Maximum likelihood with multiparameter models of substitution, *J. Mol. Evol.* 39 (1994) 409.
- [16] Z. Yang, Complexity of the simplest phylogenetic estimation problem, *Proc. R. Soc. Lond. B* 267 (2000) 109.