Maximum Likelihood on Four Taxa Phylogenetic Trees: Analytic Solutions*

[Extended Abstract]

Benny Chor
School of Computer Science
Tel-Aviv University
Tel-Aviv 39040
Israel
benny@cs.tau.ac.il

Amit Khetan
Dept. of Mathematics
UC Berkeley
Berkeley
California 94720
USA
akhetan@math.berkeley.edu

Sagi Snir†
Computer Science
Department
Technion
Technion City
Haifa 32000
Israel
ssagi@cs.technion.ac.il

ABSTRACT

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. Because no general analytic solution is known, numeric techniques such as hill climbing or expectation maximization (EM), are used in order to find optimal parameters for a given tree. So far, analytic solutions were derived only for the simplest model - *three* taxa, two state characters, under a molecular clock (MC). Quoting Ziheng Yang (2000), who initiated the analytic approach, "*this seems to be the simplest case, but has many of the conceptual and statistical complexities involved in phylogenetic estimation*".

In this work, we give analytic solutions for *four* taxa, two state characters under a molecular clock. 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. We start by presenting the general maximum likelihood problem on phylogenetic trees as a constrained optimization problem, and the resulting system of polynomial equations. In full generality, it is infeasible to solve this system, therefore specialized tools for the MC case are developed.

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). We combine the ultrametric properties of MC trees with the Hadamard conjugation (Hendy and Penny, 1993) to derive a number of topology dependent identities. Employing these identities, we substantially simplify the system of polynomial equations. We finally use tools from algebraic geometry (*e.g.* Gröbner bases, ideal saturation, resultants) and employ symbolic algebra software to obtain *closed form* analytic solutions (expressed parametrically in the input data) for the fork topology, and analytic solutions for the comb. We show that in contrast to the fork, the comb has no closed form solutions (expressed by radicals in the input data). In general, four taxa trees can have multiple ML points (Steel, 1994, Chor *et. al.*, 2001). In contrast, we can now prove that under the MC assumption, both the fork and the comb topologies have a *unique* (local and global) ML point.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences—*biology and genetics*

General Terms

Algorithms

Keywords

Maximum likelihood, phylogenetic trees, molecular clock, Hadamard conjugation, 2-state model, constrained optimization, symbolic algebra, Groebner bases, saturation

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*, the goal is to find the best explanation for the data within the model space. Among tree reconstruction approaches, maximum likelihood (Felsenstein, 1981) is increasingly used as an optimality criterion for inferring trees. In the phylogeny context, this usually means a weighted tree (the weights are parameters of the substitution model for each edge) that maximizes the likelihood of generating the observed sequences.

*Research supported by ISF grant 418/00.

†**Corresponding author.**

RECOMB'03, April 10–13, 2003, Berlin, Germany.

Maximum likelihood (ML) algorithms are computationally intensive, but for tractable cases ML is the method 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 (2000), who worked on three taxa with two state characters under MC [28]. 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, Hendy and Penny [5] using the Hadamard Conjugation of Hendy, Penny, and Steel 1994 [10, 11], together with convexity arguments.

In this work we retain the symmetric two states model of [28, 5] under MC, 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 [4], is to formulate the ML problem as one of constrained optimization, and express it in terms of Lagrange coefficients. We use the Hadamard conjugation [10, 11] 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 [5], and is not solvable by current techniques. (The analytical solutions in [4] were all for special cases where at least 2 out of the 7 input parameters are 0.)

There are two families of rooted topologies for 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. Under MC, the distance from each of the four leaves to the root is the same (Figure 1). The fork topology is somewhat simpler because it has two pairs of equal length edges, while the comb has only one pair of equal length edges. We analyze each case separately, and derive identities that help simplify the general system of equations for each topology. It turns out that the seemingly subtle difference between

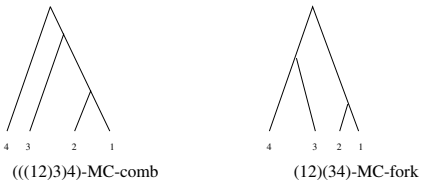

Figure 1: The fork and comb – two rooted topologies on four taxa.

the fork and the comb has significant consequences. By using the "fork identities", the system of equations becomes simple enough it is now solvable by 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. For the comb, the interrelationships between the variables are more complicated, and consequently the resulting system of equations is not solvable by computer algebra tools alone. We employ tools from algebraic geometry (*e.g.* Gröbner bases, ideal saturation, resultants), together with some guess work and, finally, application of `Maple`. We do get analytical solutions to the comb, expressed as roots of a degree 9 univariant polynomial, whose coefficients are polynomials in the input parameter. In this case a closed form solution (expressed as radicals in the input parameters) does not exist.

Even in cases where it is feasible to derive them, analytical solutions will most probably *not* replace numeric approaches in ML based tree reconstruction packages. But the analytic solutions do reveal properties of the maximum likelihood points that are not obtainable numerically. For example, we show that every MC-tree has a unique (global and local) ML point. Without the molecular clock hypothesis, this uniqueness does not hold, as proved in [23, 4]. Another consequence of our work is that methods for constructing large trees from smaller subtrees are inherently inconsistent. We give examples of phylogenetically reasonable input data on four taxa under MC, where the 4 ML subtrees on three taxa are compatible, but the four taxa tree they imply differs from the true four taxa ML tree.

The remaining of this work is organized as follows: Section 2 introduces definitions, notations, and briefly explains the Hadamard conjugation [10, 11] and its relation to maximum likelihood. In Section 3 we show some technical properties on general trees, and particular properties of the MC fork and comb. In Section 4 we derive the closed form solution to the MC fork. In Section 5 we derive the analytical solution to the MC comb. Section 6 exhibits the inconsistency of constructing large trees from smaller subtrees. In Section 7 we demonstrate two cases of applying our results to genomic DNA sequences, while in Section 8 we give some concluding remarks and directions for further research. The appendix contains several technical details and proofs that substantiate and clarify the claims in Section 5.

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 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 Figure 2).

In Neyman 2 states model [17], each character at a species admits one out of two states, without loss of generality $\{\mathtt{x},\mathtt{y}\}$. Hence, a character evolving along an evolutionary

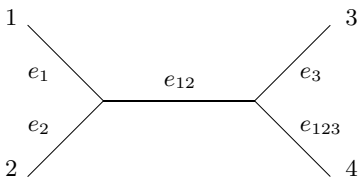

Figure 2: The tree $T' = (12)(34)$ and its edges

tree T with n leaves, induces a split pattern between the leaves admitting the state $\mathbf{x}$ and $\mathbf{y}$.

In the 2 states 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 ($\alpha \subseteq \{1, \ldots, n-1\}$) 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) $\mathbf{s} = [s_\alpha]_{\alpha \subseteq \{1,\ldots,n-1\}}$.

The *edges lengths* spectrum (edges spec) of a tree T with n leaves is the 2^{n-1} dimensional vector $\mathbf{q} = [q_\alpha]_{\alpha \subseteq \{1,\ldots,n-1\}}$, defined for any subset $\alpha \subseteq \{1, \ldots, n-1\}$ 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 [10, 11] 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 [14], and Kimura 2ST and 3ST models [16] (the last two are applicable to "normal", four states DNA). 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 (1977, p. 45), defined inductively for $n \geq 0$ by $H_0 = [1]$ and $H_{n+1} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}$. For example,

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

It is convenient to index the rows and columns of H_n by lexicographically ordered subsets of $\{1, \ldots, n\}$. 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.: *(Hendy and Penny 1993) 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*

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

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

This transformation is called the *Hadamard conjugation*.

DEFINITION 2.: *A vector $\hat{\mathbf{s}} \in \mathcal{R}^{2^{n-1}}$ satisfying*

$$\sum_{\alpha \subseteq \{1,\ldots,n-1\}} \hat{\mathbf{s}}_\alpha = 1$$

and $H\hat{\mathbf{s}} > \mathbf{0}$ is called conservative.

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

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

where the ln function is applied element-wise to the vector $H\hat{\mathbf{s}}$. 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 MC 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 structure together with the Hadamard conjugation to derive a number of identities on each of the two topologies (fork and comb). Using the derived identities, the system is substantially simplified in both cases, to the point where analytic solutions can be derived.

3.1 General ML System

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

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

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

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

Without loss of generality, we describe the systems for the unrooted trees corresponding to (12)(34) MC-fork and the (((12)3)4) MC-comb (Figure 3). The unrooted versions of

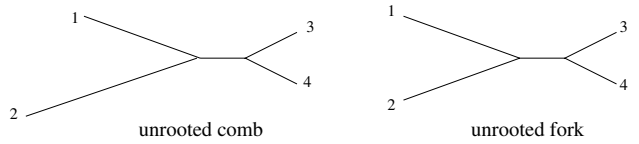

Figure 3: Unrooted layout of the MC fork and comb

the two trees both have e_{12} as their "central" edge. The expected spec $\mathbf{s}$ 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 $\alpha \notin E(T)$.

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 two constraints $q_{13}(\mathbf{s}) = 0$ and $q_{23}(\mathbf{s}) = 0$. The approach taken in [4] 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 $\mathbf{s}$, so we seek the point or points where

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

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 $\hat{s}_\alpha$ are not variables - they are parameters detemined by the four *input* sequences. (For bvrevity, we will use q_α and not $q_\alpha(\mathbf{s})$ in the sequel.)

The resulting system, in its full generality, is beyond the reach of current computer algebra techniques. The key to deriving analytical solutions is to combine the Hadamard conjugation with the molecular clock structure in order to derive a number of identities on each of the two topologies (fork and comb). Using the derived identities, the system is substantially simplified in both cases (though in a different manner). The molecular clock assumption yields different edge length relations in the two cases. For the fork, we have $q_1 = q_2$ and $q_3 = q_{123}$, while the comb satisfies $q_1 = q_2$ and $q_1 + q_{12} = q_3$. We emphasise that the system of equations does not take explicitly into account *inequalities* like $q_{12} \geq 0$ or $q_{123} \geq q_{12} + q_1$. The system is hard enough to solve as it is. Of course, the final ML point (solution) should satisfy the relevant inequlities, as otherwise it would not correspond to a "real" phylogenetic tree.

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 MC, 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 [5].

THEOREM 1.**:** *[5] 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 MC-fork has two pairs of sister taxa, each with equal edge lengths on the tree. The next theorem is a generalization of the previous one, yielding one additional identity for the MC-fork.

THEOREM 2.**:** *Consider a tree T on n leaves, with two sister taxa i and j such that $q_i = q_j$ (see figure 4). Let $\mathbf{s}$ be the expected spec, such that $\mathbf{s} = H^{-1}\ln(H\mathbf{q})$. Then for every $\alpha \subseteq \{1, 2, ..., n-1\} \setminus \{i, j\}$, $s_{\alpha\cup\{i\}} = s_{\alpha\cup\{j\}}$.*

PROOF. The argument is essentially 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 $\beta \subseteq \{1, 2, ..., n-1\} \setminus \{i, j\}$, $q_{\beta\cup\{i\}} = q_{\beta\cup\{j\}}$. This means that any function of the edge spec, $\mathbf{q}$, is invariant under interchange of i and j. In particular, $s_{\alpha\cup\{i\}} = s_{\alpha\cup\{j\}}$. □

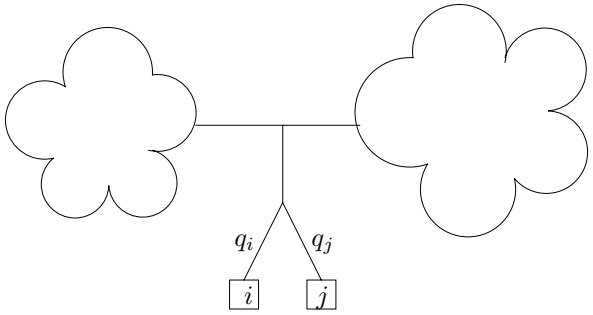

Figure 4: A general tree with two sister taxa i and j s.t. $q_i = q_j$

CLAIM 1.**:** *Let T be a $(1,2)(3,4)$-MC-fork. Then $s_3 = s_{13}(1 - 2s_1 - 2s_{13})/(2s_1 + 2s_{13})$*

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

$$\begin{aligned} 8q_{13} &= -\ln(1 - 4s_1 - 4s_{13}) \\ &\quad -\ln(1 - 4s_3 - 4s_{13}) + \ln(1 - 4s_1 - 4s_3) \ . \end{aligned}$$

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. □

The next technical claim deals with conservative points $\mathbf{s} \in \mathbb{R}^8$ (namely $H\mathbf{s} > 0$) 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 2.**:** *Let $\mathbf{s} = (s_\emptyset, s_1, s_2, s_{12}, s_3, s_{13}, s_{23}, s_{123}) \in \mathbb{R}^8$ be a conservative point satisfying $s_1 = s_2$ and $s_{13} = s_{23}$, and let $\mathbf{q} = H^{-1}\ln H\mathbf{s}$. Then $\mathbf{s}$ satisfies $q_{13}(\mathbf{s}) = q_{23}(\mathbf{s})$.*

PROOF. By the Hadamard conjugation we get:

$$\begin{aligned} 4(q_{13} - q_{23}) &= \ln(1 - 2\,s_1 - 2\,s_{12} - 2\,s_{123} - 2\,s_{13}) - \\ &\quad \ln(1 - 2\,s_2 - 2\,s_{12} - 2\,s_{23} - 2\,s_{123}) - \\ &\quad \ln(1 - 2\,s_1 - 2\,s_{12} - 2\,s_3 - 2\,s_{23}) + \\ &\quad \ln(1 - 2\,s_2 - 2\,s_{12} - 2\,s_3 - 2\,s_{13}) \end{aligned}$$

But by the assumption $s_1 = s_2$ and $s_{13} = s_{23}$,
so $(1 - 2\,s_1 - 2\,s_{12} - 2\,s_{123} - 2\,s_{13})$ (the first term) equals

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

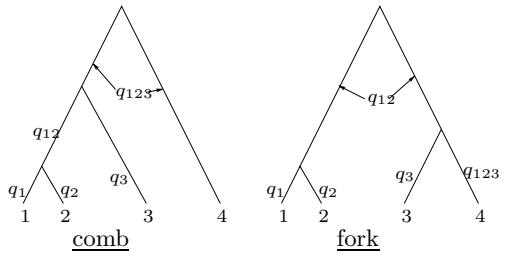

Figure 5: trees with edges length example

The following two statements refer to the MC-comb.

CLAIM 3.: *Let T be a $(((1,2)3)4)$-MC-comb. Then*

$$\begin{aligned}-1+4s_1+2s_3+2s_{123} &= (-1+4s_1+4s_{12}-2s_3+2s_{123})\\ &\quad (-1+8s_1+4s_{12}-4s_3).\end{aligned}$$

The proof of Claim 3 is very similar to the proof of Claim 1.

CLAIM 4.: *Consider the $(((12)3)4)$-MC-comb. Then $s_{13}=s_1+s_{12}-s_3$.*

PROOF. By the Hadamard conjugation we get:

$$\begin{aligned}&s_1+s_{12}-s_3-s_{13}=\\ &-\frac{1}{4}\,e^{(-2\,q_2-2\,q_{12}-2\,q_{23}-2\,q_{123})}-\frac{1}{4}\,e^{(-2\,q_1-2\,q_{12}-2\,q_{13}-2\,q_{123})}\\ &+\frac{1}{2}\,e^{(-2\,q_3-2\,q_{13}-2\,q_{23}-2\,q_{123})}+\frac{1}{4}\,e^{(-2\,q_2-2\,q_{12}-2\,q_3-2\,q_{13})}\\ &-\frac{1}{4}\,e^{(-2\,q_1-2\,q_{12}-2\,q_3-2\,q_{23})}\,.\end{aligned}$$

Substituting $q_{13}=q_{23}=0$ and the $(((12)3)4)$-MC-comb identities $q_1=q_2$ and $q_3=q_1+q_{12}$ completes the proof. □

4. SOLVING THE MC-FORK

In this section we develop the analytic solutions for the ML MC-fork.

THEOREM 3.: *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 $\sum_\alpha \hat{s}_\alpha = c$. Then the expected spec of the ML $(12)(34)$-MC-fork equals:*

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

PROOF. The $(12)(34)$-MC-fork satisfies $q_1=q_2$ and $q_3=q_{123}$, so by Theorem 1, $s_1=s_2$ and $s_3=s_{123}$. By Theorem 2, $q_1=q_2$ implies $s_{13}=s_{23}$. We substitute $s_2=s_1$, $s_{23}=s_{13}$ and $s_{123}=s_3$. Using Claim 1,

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

so after substituting the resulting value for s_3, q_{13} becomes identically zero. Now $s_1=s_2$, $s_{13}=s_{23}$, and by definition $\mathbf{s}$ is conservative. Therefore $\mathbf{s}$ satisfies the conditions of Claim 2, and thus $q_{13}=q_{23}$, so $q_{23}=0$ as well. This means that after making this series of substitutions the two constraints $(q_{13}=0,\,q_{23}=0)$ are satisfied, so in order to look for critical points of L it suffices to solve $\nabla L=\mathbf{0}$. 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 can now be expressed as a function of three free variables s_1,s_{12},s_{13} and eight given parameters (the $\hat{s}_\alpha$). The three partial derivatives in ∇L yield a system of three polynomial equations that is too involved to solve manually. We applied `Maple` to solve it. The solutions for the three variables s_1,s_{12},s_{13} appear in the statement of the theorem. The other variables are obtained by back substitutions. □

The question of uniqueness of the ML point for phylogenetic analysis has raised considerable interest in the past [7, 27, 23, 4] . It is now known that even four taxa ML trees exhibit datasets giving rise to multiple ML points [23, 4]. In contrast, our result implies uniqueness for the ML MC-fork.

COROLLARY 4.: *Each MC-fork topology has a unique local and global maximum likelihood point.*

5. SOLVING THE MC-COMB

In this section we explain very briefly the process leading to the analytical solution of the MC-comb. Additional details and explanations appear in the Appendix.

The starting point is similar to the fork. We seek the point or points satisfying $\nabla L=\lambda_1\nabla q_{13}(\mathbf{s})+\lambda_2\nabla q_{23}(\mathbf{s})$. We apply the "comb identities" derived in subsection 3.2. A change of the variables and of the input parameters leads to a somewhat simplified system, where the two constraints $q_{12}=0,q_{13}=0$ are already satisfied. Having no constraints to satisfy, we now look for point(s) satisfying $\nabla L=0$. This later system has three free variables and three partial derivatives. Unfortunately, this system involves rational functions, and not just polynomials. The obvious route to tackle the problem is to equate the numerators of these rational functions to zero. There are two drawbacks to this approach: It yields high degree polynomials (total degree 5), and may introduce many "spurious zeroes" of ∇L. We now apply a mixture of algebraic geometry techniques (Gröbner bases, saturation of an ideal), a bit of guesswork, some luck, and heavy application of computer algebra software (`Maple`). With all these we discover three simpler polynomials (total degree 3) in three variables, whose zeroes are exactly the zeroes of ∇L, without any spurious zeroes.

One of these polynomials is linear in one of the variables, so we can express this variable as a polynomial in the two remaining variables (of total degree 2). Thus, we are left with two polynomials in two variables, t_1 and t_2. Viewed as polynomials in t_1, the degree of both polynomials is just 3. To find the common zeroes of these two polynomials, we

compute their resultant with respect to t_1. After simplification, we get a single univariate polynomial in t_2, of degree 9, whose coefficients (there are 3891 of those) are polynomials of total degree 5 in the (given) input parameters. The Galois group of this polynomial is the whole symmetric group S_9, so it is a "generic" degree 9 polynomial which cannot be solved by radicals. This means that we cannot get closed form solutions (expressed with radicals of the input parameters) like we got for the fork. To solve any particular instance, a univariate polynomial root finder (such as **fsolve** in Maple), is employed. All of the s_α are then readily obtainable from the solutions to this one equation. In principle, the degree 9 equation could yield up to 9 solutions to the ML problem. However, the following theorem states that the ML point in the valid region is unique (proof omitted):

THEOREM 5.: *If all the $\hat{\mathbf{s}}_\alpha$ are positive integers then L has a unique critical point where all s_α are between 0 and 1.*

6. CONSISTENCY OF A TREE RECONSTRUCTION METHOD

The algorithmic paradigm of *divide and conquer* is used under a variety of contexts in phylogeny. These include quartets methods [2, 25, 12, 3, 15, 13, 19], super tree methods [21, 18, 24], and others. In general, the input is a set of n aligned sequences over the taxa set $\{1, 2, \ldots, n\}$, and ℓ subsets $S_1, S_2, \ldots, S_\ell \subset \{1, 2, \ldots, n\}$. Phylogenetic trees are constructed for each subset S_i of taxa. These ℓ subtrees are then joined to construct a single tree on all n taxa. Variants include rooted or unrooted trees, as well as the question of what to do in case the constructed subtrees are incompatible (either due to sampling errors, or to shortcomings of the algorithm).

Perhaps the simplest variant deals with rooted trees (and subtrees), under the requirement that the "large tree" be constructed only if all subtrees are compatible. For this rooted case, Aho *et. al.* [1] devised an efficient algorithm. (In contrast, for unrooted trees the analogous problem is NP-hard [22]). A natural question to ask is the *consistency* of the divide and conquer paradigm. Suppose all subtrees are consistent, and each is correctly constructed (according to the optimization criteria employed). Would the "large tree" also optimize the same criteria?

In this section we consider a simple, specific instance of this general consistency problem. We consider $n = 4$ taxa, all $\binom{4}{3} = 4$ rooted triplets (subtrees on 3 species), and ML under MC optimization criteria. In general, given an observed spec $\hat{s}$ for a taxa set N, the construction of the observed spec for a subset $K \subseteq N$ is easily done by "projecting" $\hat{s}$ on the input space of K. In particular, to infer the ML MC-triplet for an observed spec $\hat{s} = [\hat{s}_0, \hat{s}_1, \hat{s}_2, \hat{s}_3]$ on the taxa set $X = \{1, 2, 3\}$, we use the following theorem:

THEOREM 6.: *(Chor, Hendy and Penny 2001) Assume $\hat{s}_3 < \hat{s}_2 < \hat{s}_1$. Then the ML MC-triplet equals T_1, where T_1 is the ML tree for the family of trees where $q_2 = q_3$.*

We can now show that for ML MC criteria, divide and conquer is *inconsistent*.

THEOREM 7.: *There is an observed spec*

$$\hat{s} = [\hat{s}_0, \hat{s}_1, \hat{s}_2, \hat{s}_{12}, \hat{s}_3, \hat{s}_{13}, \hat{s}_{23}, \hat{s}_4]$$

over four taxa such that the set of ML MC-triplets for the four observed specs induced by $\hat{s}$ infer the wrong tree.

PROOF. By example. Let $\hat{s} = [400, 10, 10, 30, 20, 30, 30, 21]$. The observed sequence specs obtained by projection on the taxa set $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{1, 3, 4\}$ and $\{2, 3, 4\}$ are $[421, 40, 40, 50]$, $[420, 40, 40, 51]$, $[410, 40, 50, 51]$ and $[410, 40, 50, 51]$ respectively. The ML MC-triplets inferred by Theorem 6 are $\{(1, 2), 3\}$, $\{(1, 2), 4\}$, $\{(1, 3), 4\}$ and $\{(2, 3), 4\}$ respectively. Applying Aho *et. al.* algorithm on this set of triplets trees yields the $(((12), 3), 4)$-MC-comb. Using our techniques we can find that the likelihood of ML $(((12), 3), 4)$-MC-comb is -683.85. However, the ML (12)(34)-MC-fork achieves a bigger likelihood, -683.42.

□

7. RESULTS ON GENOMIC SEQUENCES

In order to evaluate our method we tested them on two genomic sequences. Since our methods are designed for four species, we took four homolog sequence from species where it is believed that MC approximately holds. We aligned the sequences using CLUSTALW [26]. We then converted the sequences to two states (purines – A,G and pyrimidines – C,T) and applied our formulae to get the best ML-MC phylogenetic tree out of the 3 forks and the 12 combs. We then compared this tree to the one obtained by running Phylip [6] under MC for the "regular" four states DNA sequences.

Our first test involved four primates: Human, Chimpanzee (pan), Sumatran orangutan (pongo) and Gorilla. The gene examined was the D-loop mitochondrial DNA (accession numbers X90314, AF176766, X97708, and AF089820, respectively), with sequence length 320 bp (after gaps removal) . The observed spec for this gene is

$$\hat{s} = [211, 4, 3, 13, 26, 10, 3, 50].$$

The MC tree produced by our method is the comb (((human,pan),pongo),gorilla). It is commonly believed that these four primates evolved along a different tree – the (((human,pan),gorilla),pongo)-MC-comb [9]. However, as indicated in [20], primate mitochondrial sequences often give rise to a tree different than the consensus one. Phylip ML-MC program (dnamlk), working in the *four states* model, inferred the same tree for this gene, with small differences in the proportions between branches' lengths. Figure 7 depicts the two inferred trees. Furthermore, the Homo-Pan clade was still inferred.

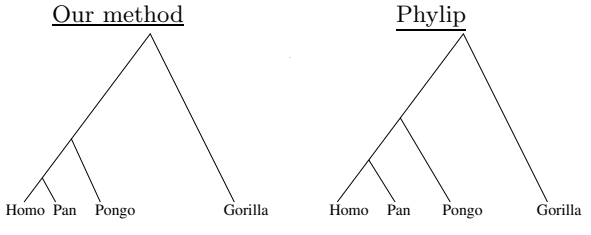

Figure 6: Inferred ML-MC trees for the D-loop mitochondrial DNA

Our second computational test was on the four species Human, Chimpanzee (pan), Rat (rattus), and Mouse (mus). The gene examined was the NK cell receptor D gene (accession numbers AF260135, AF259063, AF009511, and AF030313,

respectively). The observed spec is

$$\hat{s} = [539, 0, 0, 79, 9, 0, 0, 15].$$

The tree inferred by our method and was almost identical to the one inferred by Phylip. It is shown in Figure 7.

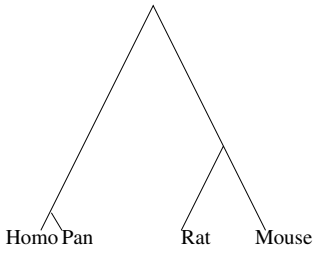

Figure 7: ML MC tree for NK cell receptor D. gene

8. CONCLUDING REMARKS

This work introduced novel applications of algebraic techniques to get analytical solutions for 4 taxa ML trees under MC in the 2-states model. We believe this is a significant step in this area of research. It would be of interest to extend these techniques to 4 taxa trees *without MC*, and possibly to four state characters under the Jukes-Cantor model of substitution. One major promise of this line of research is to rigorously prove global properties of ML trees (for example, the uniqueness of ML point under the MC model). Numeric approaches can only suggest conjectured properties. A specific open problem is the number of local ML points for non MC, four taxa trees, where currently the only known cases are 1, 2, or infinitely many local ML points. Even though we resort to a numeric tool (root finder) at the final step of the MC-comb, the situation is very different than the numeric methods used in ML in general. The uniqueness of local maxima for the MC-comb implies that hill climbing methods need not employ multiple starting points.

An interesting observation is that in the few tests we conducted on genomic data, there was no significant difference in the outcome when using our approach (2 states) vs. the "regular" 4 states ML algorithms. Of course, this may change when a larger number of taxa is considered. Finally, we remark that the use of Hadamard conjugation seems unavoidable. It is possible to formulate the MC-fork, for example, as a system with three probabilities, 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 Figure 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 5 and up to 63 monomials that we get for the MC-fork.

Acknowledgments

We would like to thank Mike Hendy and Mike Steel for very helpful discussions and for comments on an earlier version of this work. We also thank Dan Geiger and Shmuel Onn for fruitful discussions. Finally, we would like to thank the Singular team from the Centre for Computer Algebra, University of Kaiserslautern, for their assistance in the prelimary computational steps of this work.

9. REFERENCES

[1] A. Aho, Y. Sagiv, T. Szymanski, and J. Ullman. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. *SIAM Journal of Computing*, 10(3):405–421, 1981.

[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] A. Ben-Dor, B. Chor, D. Graur, R. Ophir, and D. Pelleg. Constructing phylogenies from quarbcgoptets: Elucidation of eutherian superordinal relationships. *Jour. of Comput. Biology*, 5(3):377–390, 1998.

[4] B. Chor, M. Hendy, B. Holland, and D. Penny. Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *MBE*, 17(10):1529–1541, 2000. Earlier version appeared in RECOMB 2000.

[5] B. Chor, M. Hendy, and D. Penny. Analytic solutions for three taxon mlmc trees with variable rates across sites. In *WABI 2001*, 2001.

[6] J. Felsenstein. PHYLIP - phylogenetic inference package, (version 3.2). *Cladistics*, 5:164–166, 1989.

[7] K. Fukami and Y. Tateno. On the uniqueness of the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. *J. Mol. Evol.*, 28:460–464, 1989.

[8] I. Gelfand, M. Kapranov, and A. Zelevinsky. *Discriminants, Resultants and Multidimensional Determinants*. Birkhauser, Boston, 1994.

[9] D. Graur and W. Li. *Fundamentals of Molecular Evolution*. Sinauer Assoc., 2000.

[10] M. D. Hendy and D. Penny. Spectral analysis of phylogenetic data. *J. Classif.*, 10:5–24, 1993.

[11] M. D. Hendy, D. Penny, and M. Steel. Discrete fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci. USA.*, 91:3339–3343, 1994.

[12] T. Jiang, P. Kearney, and M. Li. Orchestrating quartets: approximation and data correction. In *IEEE Symp. Foundation of Computer Science (FOCS)*, pages 416–425, Palo Alto, California, November 1998.

[13] K. S. John, T. Warnow, B. M. E. Moret, and L. Vawter. Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining. In *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2001.

[14] T. Jukes and C. Cantor. Evolution of protein molecules. In H. Munro, editor, *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[15] P. Kearney. The ordinal quartet method. In *RECOMB98, the Second Annual International Conference on Computational Molecular Biology*, New York, 1998. ACM.

[16] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, 1983.

[17] J. Neymann. Molecular studies of evolution: A source of novel statistical problems. In S. Gupta and Y. Jackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York, 1971.

[18] R. Page. Modified mincut supertrees. In *R. Guigo and D Gusfield (Eds): WABI 2002, LNCS 2452*, 2002.

[19] V. Ranwez and O. Gascuel. Quartet-based phylogenetic inference: Improvements and limits. *Molecular Biology and Evolution*, 18:1103–1116, 2001.

[20] M. Ruvolo. Molecular phylogeny of the hominoids: Inferences from multiple independent dna sequence data sets. *Molecular Biology and Evolution*, 14(7):248–265, 1997.

[21] C. Semple and M. Steel. A supertree method for rooted trees. *Discrete Applied Mathematics*, 103:147–158, 2000.

[22] M. Steel. The complexity of reconstructing trees from qualitative characters and subtress. *Journal of Classification*, 9(1):91–116, 1992.

[23] M. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Syst. Biology*, 43(4):560–564, 1994.

[24] M. Steel, A. Dress, and S. Boker. Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology*, 49:363–368, 2000.

[25] K. Strimmer and A. von Haeseler. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13(7):964–969, 1996. Software available at `ftp://ftp.ebi.ac.uk/pub/software/unix/puzzle/`.

[26] J. Thompson, D. Higgins, and T. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalty and weight matrix choice. *Nucleic Acids Research*, 22:4673–4780, 1994.

[27] E. Tillier. Maximum likelihood with multiparameter models of substitution. *J. Mol. Evol.*, 39:409–417, 1994.

[28] Z. Yang. Complexity of the simplest phylogenetic estimation problem. *Proc. R. Soc. Lond. B*, 267:109–119, 2000.

APPENDIX

A. COMB DETAILS

We start out with a linear invertible change of variables:

$$t_0 = s_0 = 1 - 4s_1 + s_3 - 3s_{12} - s_{123}, t_1 = 1 - 8s_1 - 4s_{12} + 4s_3,$$

$$t_2 = 1 - 4s_1 - 4s_{12} + 2s_3 - 2s_{123}, t_3 = 1 - 4s_1 - 2s_3 - 2s_{123}.$$

The quadratic constraint of Claim 3 can now simply be written as $t_3 = t_1 t_2$.

Under this change of variables we can remove all of the constraints and reduce the problem to maximizing a function of three free variables. To simplify presentation we rewrite the parameter space a bit, setting $g_1 = \hat{s}_1 + \hat{s}_2$, $g_{12} = \hat{s}_{12}$, $g_{13} = \hat{s}_{13} + \hat{s}_{23}$, $g_3 = \hat{s}_3$, and $g_{123} = \hat{s}_{123}$.

Substituting, and applying Gröbner bases, saturation of ideals by polynomials, some guesses and a good deal of luck and `Maple` hacking, we can now show (proof omitted)

PROPOSITION 2.: *The ML points of the MC-comb are among the zeros of the following polynomial system.*

$$\begin{aligned}
m_0 &= c(t_1^2 - 2t_2t_1 + 4t_2^2 + 8t_0 - 6t_2 - 5) + \\
&\quad g_1(-8t_0 + 4t_1 + 2t_2t_1 + 2t_2 + 8) + \\
&\quad g_3(2 + 4t_2 + 2t_1) + 2g_{12}(1 + t_1)(1 + t_2) + \\
&\quad g_{13}(-8t_0 + 2t_2t_1 + 10t_2 + 4t_1 + 8) + 2g_{123}t_2(3 + t_1) \\
m_1 &= c(-2t_2^2 - 3t_2t_1 - 3t - 2 + 4t_2t_0 - 1 + 4t_1t_0 - t_1 + 2t_2^3) + \\
&\quad g_1(t_2^2 + t_2^2t_1 + 4t_0 + 2t_2t_1 - 4t_2t_0 + 2t_2 + t_1 + 1) + \\
&\quad 2g_3(2t_2^2 + 3t_2t_1 + t_2 + 1 + t_1) + g_{12}(1 + t_2)^2(1 + t_1) + \\
&\quad g_{13}(4t_0 + t_2^2t_1 + 5t_2^2 - 4t_2t_0 + 4t_2 + 4t_2t_1 + t_1 + 1) + \\
&\quad g_{123}t_2(3t_2 + t_2t_1 + 3t_1 + 1) \\
m_2 &= 2c(1 + t_2)(-t_2t_1 - t_2 + 2t_2t_0 - t_0 + t_1t_0) + \\
&\quad g_1(2t_2 + 4t_0 - 4t_2t_0 + 2t_2^2t_1 + 2t_2^2 + 2t_2t_1) + \\
&\quad 2g_3t_2(1 + t_1)(1 + t_2) + 2g_{12}t_2(1 + t_1)(1 + t_2) + \\
&\quad 2g_{13}(1 + t_2)(t_2 + t_2t_1 + 2t_0) + 2g_{123}t_2(1 + t_1)(1 + t_2)
\end{aligned}$$

The equations m_0, m_1, m_2 are of low degree, as m_0 is of total degree 2, while m_1 and m_2 are of total degree 3. Furthermore, t_0 appears *linearly* in s_0, so it is possible to eliminate it. This leaves us with just two polynomials in t_1 and t_2. We shall see that this system 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. In particular the coefficients can themselves be unknowns, or functions of other variables, in which case the resultant replaces the two polynomials f and g with a single polynomial in one fewer variable.

Computing the resultant is a classical technique for eliminating one variable from two equations. There is a very nice formula due to Sylvester and another due to Bezout which have been implemented in most computer algebra packages, such as `Maple`. We use the following property in order to obtain the common root of $f(x)$ and $g(x)$ as a function of the coefficients a_i and b_j [8, pp. 109].

THEOREM 8.: *[8] Let $R(a, b) = Res(f, g, x)$. If a, b are such that f and g have a common zero, therefore $R(a, b) = 0$, and moreover this common zero is unique without multiplicity, then we can actually find the common root as a ratio of two polynomials in a and b. Namely,*

$$x = \frac{\partial R}{\partial a_1} \Big/ \frac{\partial R}{\partial a_0} \quad .$$

Applying this theorem, we find a single univariate polynomial $E(t_2)$ in just one variable. This turns out to be a polynomial of degree 9 whose 3891 coefficients are themselves polynomials of total degree up to 5 in the input parameters. The Galois group of this polynomial is the whole symmetric group S_9, so it is a "generic" degree 9 polynomial which cannot be solved by radicals. For any particular choice of the parameters we use a root finder (**fsolve** in `Maple`) to find the real roots of $E(t_2)$.