Analytic Solutions for Three Taxon ML_{MC} Trees with Variable Rates Across Sites

Benny Chor* Michael Hendy† David Penny‡

Abstract

We consider the problem of finding the maximum likelihood rooted tree under a molecular clock (ML_{MC}), with three species and 2-state characters under a symmetric model of substitution. For identically distributed rates per site this is probably the simplest phylogenetic estimation problem, and it is readily solved numerically. Analytic solutions, on the other hand, were obtained only recently (Yang, 2000).

In this work we provide analytic solutions for any distribution of rates across sites, provided the moment generating function of the distribution is strictly increasing over the negative real numbers. This class of distributions includes, among others, identical rates across sites, as well as the Gamma, the uniform, and the inverse Gaussian distributions. Therefore, our work generalizes Yang's solution. In addition, our derivation of the analytic solution is substantially simpler. We use the Hadamard conjugation (Hendy and Penny, 1993) to prove a general statement about the edge lengths of any neighboring pair of leaves in any phylogenetic tree (on three or more taxa). We then employ this relation, in conjunction with the convexity of an entropy–like function, to derive the analytic solution.

Key words: Maximum likelihood, phylogenetic trees, molecular clock, Hadamard conjugation, 2-state model, unequal rates across sites.

Corresponding author: David Penny,
Institute of Molecular BioSciences
Massey University
Palmerston North,
New Zealand.

tel +64 6 350 5033, fax +64 6 350 5694,
http://imbs.massey.ac.nz/HTML/penny.html.

*School of Computer Science, Tel-Aviv University, Israel. benny@cs.tau.ac.il. Research supported by ISF grant 418/00. Part of this work was done while visiting Massey University, Palmerston North, New Zealand.

†Institute of Fundamental Sciences and Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand. M.Hendy@massey.ac.nz.

‡Institute of Molecular BioSciences and Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand. D.Penny@massey.ac.nz.

1 Introduction

Maximum likelihood (Felsenstein, 1981) is increasingly used as an optimality criterion for selecting evolutionary trees, but finding the global optimum is difficult computationally, even on a single tree. Because no general analytical solution is available, it is necessary to use numeric techniques, such as hill climbing or expectation maximization (EM), in order to find optimal values. Two recent developments are relevant when considering analytical solutions for simple substitution models with a small number of taxa. Yang (2000) has reported an analytical solution for three taxa with two state characters under a molecular clock. Thus in this special case the tree and the edge lengths that yield maximum likelihood values can now be expressed analytically, allowing the most likely tree to be positively identified. Yang calls this case the "simplest phylogeny estimation problem".

A second development is in Chor *et. al.* (2000), who used the Hadamard conjugation for unrooted trees on four taxa, again with two state characters. As part of that study analytic solutions were found for some families of observed data. It was reported that multiple optima on a single tree occured more frequently with maximum likelihood than has been expected. In one case, the best tree had a local (non global) optimum that was less likely than the optimum value on a different, inferior tree. In such a case, a hill climbing heuristic could misidentify the "optimal" tree. Such examples reinforce the desirability of analytical solutions that guarantee to find the global optima for any tree.

Even though three taxon, two state characters models under a molecular clock is the "simplest phylogeny estimation problem", it is still potentially an important case to solve analytically. It can allow a "rooted triplet" method for inferring larger rooted trees by building them up from the triplets. This would be analogous to the use of unrooted quartets for building up unrooted trees. Trees from quartettree methods are already used extensively in various studies (Bandelt and Dress 1986, Strimmer and von Haeseler 1996, Wilson 1998, Ben-Dor *et. al.* 1998, Erdos *et. al.* 1999). The fact that general analytical solutions are not yet available for unrooted quartets only emphasizes the importance of analytical solutions to the rooted triplets case.

Let ML_{MC} tree denote a maximum likelihood rooted tree under a molecular clock. In this work we provide analytic solutions for three taxon ML_{MC} trees under *any* distribution of variable rates across sites provided the moment generating function of the distribution is strictly increasing over the negative real numbers. This class of distributions includes, as a special case, identical rates across sites. It also includes the Gamma, the uniform, and the inverse Gaussian distributions. Therefore, our work generalizes Yang's solution of identical rates across sites. In addition, our derivation of the analytic solution is substantially simpler. We employ the Hadamard conjugation (Hendy and Penny 1993, Hendy, Penny, and Steel 1994) and convexity of an entropy–like function.

The remainder of this paper is organized as follows: In subsection 2 we explain the Hadamard conjugation and its relation to maximum likelihood. In Section 3 we state and prove our main technical theorem. Section 4 applies the

theorem to solve ML_{MC} analytically on three species trees. Finally, Section 5 presents some implications of this work and directions for further research.

2 Hadamard Conjugation and ML

The Hadamard conjugation (Hendy and Penny 1993, Hendy, Penny, and Steel 1994) is an invertible transformation linking the probabilities of site substitutions on edges of an evolutionary tree T with edge set $E(T)$ to the probabilities of obtaining each possible combination of characters. It is applicable to a number of simple models of site substitution: Neyman 2 state model (Neyman 1971), Jukes–Cantor model (Jukes and Cantor 1969), and Kimura 2ST and 3ST models (Kimura 1983). For these models, the transformation yields a powerful tool which greatly simplifies and unifies the analysis of phylogenetic data. In this section we explain the Hadamard conjugate and its relationships to ML.

We now introduce a notation that we will use for labeling the edges of unrooted binary trees. (For simplicity we use four taxa, but the definitions extend to any n.) Suppose the four species, $1, 2, 3$ and 4, are represented by the leaves of the tree T'. 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 $\{1, 2\}$. Each edge e of T induces a split of the taxa, namely the two sets of leaves on the two components of T resulting from the deletion of e. Hence the central edge of the tree $T' = (12)(34)$ in the brackets notation induces the split identified by the subset $\{1, 2\}$. For brevity we will label this edge by e_{12} as a shorthand for $e_{\{1,2\}}$. Thus $E(T') = \{e_1, e_2, e_{12}, e_3, e_{123}\}$ (see Figure 1).

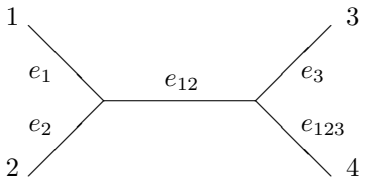


Figure 1: The tree $T' = (12)(34)$ and its edges

We use a similar indexing scheme for splits at a site in the sequences: For a subset $\alpha \subseteq \{1, ..., n-1\}$, we say that a given site i is an α-split pattern if α is the set of sequences whose character state at position i differs from the i-th position in the n-th sequence. Given a tree T with n leaves and edge lengths $\mathbf{q} = [q_e]_{e \in E(T)}$ $(0 \leq q_e < \infty)$ (where q_e is the expected number of substitutions per site, across the edge e), the expected probability (averaged over all sites) of generating an α-split pattern ($\alpha \subseteq \{1, \ldots, n-1\}$) is well defined (this probability may vary across sites, depending on the distribution of rates). Denote this

expected probability by $s_\alpha = Pr(\alpha\text{-split}|T, \mathbf{q})$. We define the *expected sequence spectrum* $\mathbf{s} = [s_\alpha]_{\alpha \subseteq \{1,\ldots,n-1\}}$. Having this spectrum at hand greatly facilitates the calculation and analysis of the likelihood, since the likelihood of observing a sequence with splits described by the vector $\hat{\mathbf{s}}$ given the sequence spectrum $\mathbf{s}$ equals

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

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 $H_n^{-1} = \frac{1}{2^n} H_n$.

The length of an edge q_e, $e \in E(T)$ in the tree T was defined as the expected number of substitutions (changes) per site along that edge. The *edge length* spectrum of a tree T be 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

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

The Hadamard conjugation specifies a relation between the expected sequence spectrum $\mathbf{s}$ and the edge lengths spectrum $\mathbf{q}$ of the tree.

Proposition 1 *(Hendy and Penny 1993) Let T be a phylogenetic tree on n leaves with finite edge lengths ($0 \leq q_e < \infty$ 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 for $H = H_{n-1}$ we have:*

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

where the exponentiation function exp 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*.

For the case of unequal rates across sites, the following generalization applies:

Proposition 2 *(Waddell, Penny, and Moore 1997) Let T be a phylogenetic tree on n leaves with finite edge lengths ($0 \leq q_e < \infty$ for all $e \in E(T)$). Assume that sites mutate according to a symmetric substitution model, with unequal rates across sites, so that $M : R \to R$ be the moment generating function of the rate distribution. Let $\mathbf{s}$ be the expected sequence spectrum. Then for $H = H_{n-1}$,*

$$\mathbf{s} = \mathbf{s}(\mathbf{q}) = H^{-1}(M(H\mathbf{q})) \ ,$$

where the function M is applied element wise to the vector $\rho = H\mathbf{q}$.

This transformation is called the *Hadamard conjugation of M*. Specific examples of the moment generating function include

- For equal rates across sites, $M(\rho) = e^{\rho}$.
- For the uniform distribution in the interval $[1-b, 1+b]$ with parameter b $(1 \geq b > 0)$, $M(\rho) = \frac{1}{2b\rho}\left(e^{(1+b)\rho} - e^{(1-b)\rho}\right)$.
- For the Γ distribution with parameter k $(k > 0)$, $M(\rho) = (1 - \rho/k)^{-k}$.
- For the inverse Gaussian distribution with parameter d $(d > 0)$, $M(\rho) = e^{d(1-\sqrt{1-2\rho/d})}$.

Notice that for $k \to \infty$, the Γ distribution converges to the equal rates distribution.

3 Technical Results

Under a molecular clock, a model tree on $n \geq 2$ taxa has at least two sister taxa i and j whose pendant edges q_i and q_j are of equal length ($q_i = q_j$). Our first result states that if $q_i = q_j$, then the corresponding split probabilities are equal ($s_i = s_j$). Knowing that a pair of these variables attains the same value simplifies the analysis of the maximum likelihood tree in general, and in particular makes it possible for the case of $n = 3$ taxa. Furthermore, if $q_i > q_j$ and the moment generating function M is strictly increasing in the range $(-\infty, 0]$, then the corresponding split probabilities satisfy $s_i > s_j$.

3.1 Main Technical Theorem

Theorem 1 *Let i and j be sister taxa in a phylogenetic tree T on n leaves, with edge weights $\mathbf{q}$. Let $\mathbf{s}$ be the expected sequence spectrum, let $H = H_{n-1}$, and let M be a real valued function such that*

$$\mathbf{s} = H^{-1}M(H\mathbf{q}),$$

then:

$$q_i = q_j \Longrightarrow s_i = s_j;$$

and if the function M is strictly monotonic ascending in the range $(-\infty, 0]$ then:

$$q_i > q_j \Longrightarrow s_i > s_j.$$

Proof: Let $X = \{1, 2, \ldots, n\}$ be the taxa set with reference element n, and let $X' = X - \{n\}$. Without loss of generality $i, j \neq n$. For $\alpha \subseteq X'$, let $\alpha' = \alpha \Delta \{i, j\}$ (where $\alpha \Delta \beta = (\alpha \cup \beta) - (\alpha \cap \beta)$ is the symmetric difference of α and β). The mapping $\alpha \to \alpha'$ is a bijection between

$$X_i = \{\alpha \subseteq X' | i \notin \alpha, j \in \alpha\}$$

and

$$X_j = \{\alpha \subseteq X' | i \in \alpha, j \notin \alpha\}.$$

Note that the two sets X_i and X_j are disjoint. Writing $h_{\alpha,i}$ for $h_{\alpha,\{i\}}$ we have

$$\alpha \in X_i \Longrightarrow h_{\alpha,i} = 1, h_{\alpha,j} = -1, h_{\alpha',i} = -1, h_{\alpha',j} = 1.$$

On the other hand, if $\alpha \notin X_i \cup X_j$ then $h_{\alpha,i} = h_{\alpha,j}$. Hence

$$\begin{aligned}
s_i - s_j &= 2^{-(n-1)} \textstyle\sum_{\alpha \subseteq X'} (h_{\alpha,i} - h_{\alpha,j}) M(\rho_\alpha) \\
&= 2^{-(n-1)} \left(\textstyle\sum_{\alpha \in X_i} (h_{\alpha,i} - h_{\alpha,j}) M(\rho_\alpha) + \sum_{\alpha \in X_j} (h_{\alpha,i} - h_{\alpha,j}) M(\rho_\alpha) \right) \\
&= 2^{-(n-1)} \left(\textstyle\sum_{\alpha \in X_i} (h_{\alpha,i} - h_{\alpha,j}) M(\rho_\alpha) + \sum_{\alpha \in X_i} (h_{\alpha',i} - h_{\alpha',j}) M(\rho_{\alpha'}) \right) \\
&= 2^{-(n-1)} \left(\textstyle\sum_{\alpha \in X_i} 2M(\rho_\alpha) - \sum_{\alpha \in X_i} 2M(\rho_{\alpha'}) \right) \\
&= 2^{-(n-2)} \textstyle\sum_{\alpha \in X_i} (M(\rho_\alpha) - M(\rho_{\alpha'})) \ .
\end{aligned}$$

By the definition of the Hadamard conjugate,

$$\rho_\alpha = \sum_{\beta \subseteq X'} h_{\alpha,\beta} q_\beta \quad \text{, so} \quad \rho_\alpha - \rho_{\alpha'} = \sum_{\beta \subseteq X'} (h_{\alpha,\beta} - h_{\alpha',\beta}) q_\beta \ .$$

Now for $\beta = \emptyset$ we have $h_{\alpha,\beta} = h_{\alpha',\beta} = 1$ so the contribution of $\beta = \emptyset$ to $\rho_\alpha - \rho_{\alpha'}$ is zero. Likewise, for any split $\beta \subseteq X'$ ($\beta \neq \emptyset$), which does not correspond to an edge $e \in E(T)$, $q_\beta = 0$. So the only contributions to $\rho_\alpha - \rho_{\alpha'}$ may come from splits β corresponding to edges in T. Now since i and j are sister taxa in T, every edge $e \in E(T)$ that is not pendant upon i or j does not separate i from j. Thus the split β corresponding to such edge e satisfies $\beta \notin X_i \cup X_j$, and the parities of $|\alpha \cap \beta|$ and $|\alpha' \cap \beta|$ are the same, so

$$h_{\alpha,\beta} = (-1)^{|\alpha \cap \beta|} = (-1)^{|\alpha' \cap \beta|} = h_{\alpha',\beta} \ .$$

Thus the only contributions to $\rho_\alpha - \rho_{\alpha'}$ comes from the two edges that connect the pair of neighboring leaves i and j to their parent. That is,

$$\rho_\alpha - \rho_{\alpha'} = (h_{\alpha,i} - h_{\alpha',i}) q_i + (h_{\alpha,j} - h_{\alpha',j}) q_j \ ,$$

and for $\alpha \in X_i$ we get $\rho_\alpha - \rho_{\alpha'} = 2(q_i - q_j)$.

Thus if $q_i = q_j$ then for every $\alpha \in X_i$ we have $\rho_\alpha = \rho_{\alpha'}$, so $M(\rho_\alpha) = M(\rho_{\alpha'})$ and $s_i - s_j = 2^{-(n-2)} \sum_{\alpha \in X_i} (M(\rho_\alpha) - M(\rho_{\alpha'})) = 0$, hence $s_i = s_j$.

If $q_i > q_j$ then for every $\alpha \in X_i$ we have $\rho_\alpha > \rho_{\alpha'}$. Now $q_\emptyset = -\sum_{e \in E(T)} q_e$, and for every $e \in E(T)$, $q_e \geq 0$. Since $\rho_\alpha = \sum_{\beta \subseteq X'} h_{\alpha,\beta} q_\beta$ and $h_{\alpha,\emptyset} = 1$ we conclude that $\rho_\alpha \leq 0$ for all $\alpha \subseteq X'$. Therefore, if M is strictly monotone ascending in $(-\infty, 0]$ then $M(\rho_\alpha) > M(\rho_{\alpha'}), \forall \alpha \in X_i$. Since $s_i - s_j = 2^{-(n-2)} \sum_{\alpha \in X_i} (M(\rho_\alpha) - M(\rho_{\alpha'}))$, we have $s_i > s_j$. □

We remark that the moment generating functions M in the four examples of Section 2 (equal rates across sites, uniform distribution with parameter b, $0 < b \leq 1$, Gamma distribution with parameter k, $0 < k$, and inverse Gaussian distribution with parameter d, $0 < d$) are strictly increasing in the range $\rho \in (-\infty, 0]$. We reiterate that the statement in Theorem 1 holds for trees in general, not just on 3 leaves.

4 Three Taxa $\mathbf{ML}_{MC}$ Trees

We first note that for three taxa, the problem of finding analytically the ML trees without the constraint of a molecular clock is trivial. This is a special case of unconstrained likelihood for the multinomial distribution. On the other hand, adding a molecular clock makes the problem interesting even for $n = 3$ taxa, which is the case we treat in this section.

For $n = 3$, let s_0 be the probability of observing the constant site pattern (`xxx` or `yyy`). Let s_1 be the probability of observing the site pattern which splits 1 from 2 and 3 (`xyy` or `yxx`). Similarly, let s_2 be the probability of observing the site pattern which splits 2 from 1 and 3 (`yxy` or `xyx`), and let s_3 be the probability of observing the site pattern which splits 3 from 1 and 2 (`xxy` or `yyx`).

Consider unrooted trees on the taxa set $X = \{1, 2, 3\}$ that have two edges of the same length. Let $\mathcal{T}_1$ denote the family of such trees with edges 2 and 3 of the same length ($q_2 = q_3$), $\mathcal{T}_2$ denote the family of such trees with edges 1 and 3 of the same length ($q_1 = q_3$), and $\mathcal{T}_3$ denote the family of such trees with edges 2 and 1 of the same length ($q_2 = q_1$). Finally, let $\mathcal{T}_0$ denotes the family of trees with $q_1 = q_2 = q_3$. We first see how to determine the ML tree for each family.

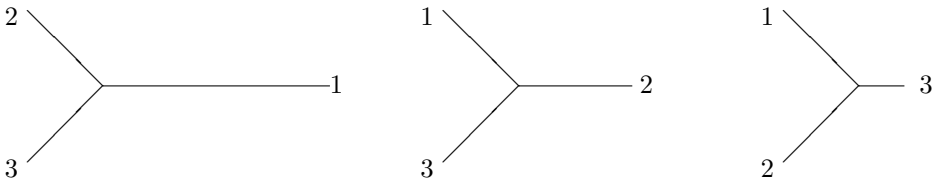


Figure 2: Three trees in the families $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3$, respectively

Given an observed sequence of m sites, let m_0 be the number of sites where

all three nucleotides are equal, and let m_i $(i = 1, 2, 3)$ be the number of sites where the character in sequence i differs from the state of the other sequences. Then $m = m_0 + m_1 + m_2 + m_3$, and $f_i = m_i/m$ is the frequency of sites with the corresponding character state pattern.

Theorem 2 *Let (m_0, m_1, m_2, m_3) be the observed data. The ML tree in each family is obtained at the following point:*

- *For the family $\mathcal{T}_0$, the likelihood is maximized at T_0 with $s_0 = f_0$, $s_1 = s_2 = s_3 = (1 - f_0)/3$.*
- *For the family $\mathcal{T}_1$, the likelihood is maximized at T_1 with $s_0 = f_0$, $s_1 = f_1, s_2 = s_3 = (f_2 + f_3)/2$.*
- *For the family $\mathcal{T}_2$, the likelihood is maximized at T_2 with $s_0 = f_0$, $s_2 = f_2, s_1 = s_3 = (f_1 + f_3)/2$.*
- *For the family $\mathcal{T}_3$, the likelihood is maximized at T_3 with $s_0 = f_0$, $s_3 = f_3, s_1 = s_2 = (f_1 + f_2)/2$.*

Proof: The log likelihood function equals

$$l(m_0, m_1, m_2, m_3|\mathbf{s}) = \sum_{i=0}^{3} m_i \log s_i,$$

and for the normalized function $\ell = l/m$ we have

$$\ell(m_0, m_1, m_2, m_3|\mathbf{s}) = \sum_{i=0}^{3} f_i \log s_i \ .$$

Consider, without loss of generality, the case of the $\mathcal{T}_1$ family. We are interested in maximizing ℓ under the constraint $q_2 - q_3 = 0$. Since 2 and 3 are a pair of sister taxa in the family of trees $\mathcal{T}_1$, by Theorem 1 the equality $q_2 = q_3$ implies $s_2 = s_3$. Substituting $s_0 = (1 - s_1 - 2s_2)$, a maximum point of the likelihood must satisfy

$$\frac{\partial \ell}{\partial s_i} = 0 \quad (i = 0, 1) \ ,$$

implying

$$\frac{f_1}{s_1} = \frac{f_0}{s_0} \ ,$$

$$\frac{f_2 + f_3}{2s_2} = \frac{f_0}{s_0} \ .$$

Denote $d = f_0/s_0$, then we have $f_2 + f_3 = 2ds_2$. Adding the right hand sides and left hand sides of this equality to these of $f_1 = ds_1$ and $f_0 = ds_0$, we get

$$f_0 + f_1 + f_2 + f_3 = d(s_0 + s_1 + 2s_2) \ .$$

Since both $f_0 + f_1 + f_2 + f_3 = 1$ and $s_0 + s_1 + 2s_2 = 1$, we get $d = 1$. So the ML point for the family $\mathcal{T}_1$ is attained at the tree T_1 with parameters

$$s_0 = f_0, s_1 = f_1, s_2 = s_3 = (f_2 + f_3)/2 \ .$$

We denote by T_2, T_3, T_0 the three corresponding trees that maximize the function ℓ for the families $\mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_0$. The weights of these three trees can be obtained in a similar fashion to T_1. □

Theorem 3 *Assume $m_3 \leq m_2 \leq m_1$. Then the ML_{MC} tree equals T_1.*

Proof: By Theorem 2, the maximum likelihood tree under the condition that two edges have the same length is one of the trees T_1, T_2, or T_3. Let

$$G(p) = f_0 \log f_0 + p \log p + (1 - f_0 - p) \log \frac{(1 - f_0 - p)}{2} \ .$$

Substituting the values s_0, s_1, s_2, s_3 for each tree in the expression

$$\ell(m_0, m_1, m_2, m_3 | \mathbf{s}) = \sum_{i=0}^{3} f_i \log s_i \ ,$$

and somewhat abusing the notation, we get the following values for the function ℓ on the three trees

$$\begin{aligned} \ell(T_1) &= G(f_1) \ , \\ \ell(T_2) &= G(f_2) \ , \\ \ell(T_3) &= G(f_3) \ . \end{aligned}$$

The function $G(p)$ behaves similarly to minus the binary entropy function (Gallager, 1968)

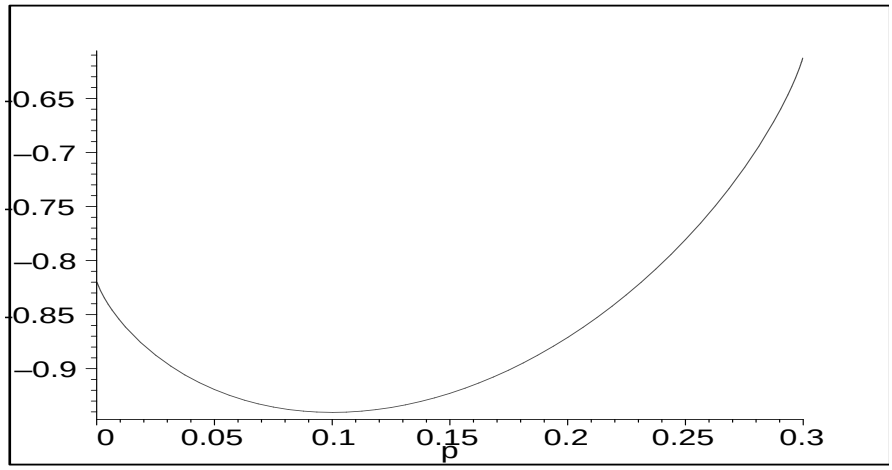
$$-H(p) = p \log p + (1 - p) \log(1 - p) \ .$$

The range where $G(p)$ is defined is $0 \leq p \leq 1 - f_0$. In this interval, $G(p)$ (like $-H(p)$) is negative and $\cup$-convex. So G has a single minimum at the point p_0 where its derivative is zero, $dG(p)/dp = 0$. Solving for p we get $p_0 = (1 - f_0)/3$.

Now $f_3 \leq f_2 \leq f_1$ and $G(p)$ is $\cup$-convex. Therefore, out of the three values $G(f_1), G(f_2), G(f_3)$, the maximum is attained at either $G(f_3)$ or at $G(f_1)$, but not at $G(f_2)$ (unless $f_2 = f_1$ or $f_2 = f_3$).

Since $f_3 + f_2 + f_1 = 1 - f_0$ and $f_3 \leq f_2 \leq f_1$, we have $f_3 \leq (1 - f_0)/3 \leq f_1$, namely the two "candidates" for ML points are on different sides of the minimum point. The point f_3 is strictly to the left and the point f_1 is strictly to the right (except the case where $f_3 = f_1$ and the two points coincide). If $G(f_1) \geq G(f_3)$, then the tree T_1 is the obvious candidate for ML_{MC} tree. Indeed, T_1 satisfies $s_3 = s_2 < s_1$, so by Theorem 1, $q_3 = q_2 < q_1$. Thus, a root can be placed on the edge e_1 so that the molecular clock assumption is satisfied.

As a specific example, consider the case where $f_0 = 0.7$. The function $G(p)$ for this case is depicted in the next figure. For $f_1 = 0.21, f_2 = 0.05$ and $f_3 = 0.04$, we have $G(f_1) = -.8565 > G(f_3) = -.9088$



The function $G(p)$ for $f_0 = 0.7$ $(0 \leq p \leq 0.3)$.

We certainly could have a case where $G(f_3) > G(f_1)$. For example, if $f_0 = 0.7$ is the same then the function $G(p)$ is the same, and for $f_1 = 0.15, f_2 = 0.14, f_3 = 0.01$ we get $G(f_3) = -.8557 > G(f_1) = -.9227$. However, in this case the tree T_3 has $s_3 < s_1 = s_2$, implying (by Theorem 1) $q_3 < q_1 = q_2$. Therefore there is no way to place a root on an edge of T_3 so as to satisfy a molecular clock. In fact, any tree with edge lengths $q_3 < q_1 = q_2$ does not satisfy a molecular clock. So the remaining possibilities could be either the tree T_0 (where $s_1 = s_2 = s_3 = (1 - f_0)/3$) or the tree T_1. As T_0 attains the minimum over the function G, the tree T_1 will always give the greater likelihood (except in the redundant case $f_1 = f_3$, where all these trees collapse to T_0). This completes the proof of Theorem 3. □

The case $m_2 < m_3 < m_1$ and its other permutations can clearly be handled similarly.

5 Discussion and Open Problems

In the case where $G(f_3) > G(f_1)$, T_1 is still the ML_{MC} tree. However, if the difference between the two values is significant, it may give a strong support for rejecting a molecular clock assumption for the given data m_0, m_1, m_2, m_3. This would be the case, for example, when $0 \approx m_3 \ll m_1 \approx m_2$.

Two natural directions for extending this work are to consider four state characters, and to extend the number of taxa to $n = 4$, either with or without the molecular clock assumption.

The question of constructing rooted trees from rooted triplets is an interesting algorithmic problem, analogous to that of constructing unrooted trees from unrooted quartets. The biological relevance of triplet based reconstruction methods is also of interest.

Acknowledgements: Thanks to Sagi Snir for helpful comments on earlier versions of this manuscript.

References

Bandelt, H.-J., and A. Dress, 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7:309–343.

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

Chor, B., M. D. Hendy, B. R. Holland , and D. Penny, 2000. Multiple Maxima of Likelihood in Phylogenetic Trees: An Analytic Approach. *Mol. Biol. Evol.*, Vol. 17, No.10, September 2000, pp. 1529–1541.

Erdos, P., M. Steel, L. Szekely, and T. Warnow, 1999. A few logs suffice to build (almost) all trees (i). *Random Structures and Algorithms*, 14:153–184.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.*, 17:368–376.

Gallager, R.G. Information Theory and Reliable Communication, Wiley, New York (1968).

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.

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

MacWilliams, F., and N. Sloan, 1977. *The Theory of Error-Correcting Codes.* North-Holland, Elsevier Science Publishers.

Neyman, J., 1971. 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.

Strimmer, K., and A. von Haeseler, 1996. Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, 13(7):964–969.

Waddell, P., D. Penny, and T. Moore, 1997. Hadamard conjugations and Modeling Sequence Evolution with Unequal Rates across Sites. *Molecular Phylogenetics and Evolution*, 8(1):33–50.

Wilson, S.J., 1998. Measuring inconsistency in phylogenetic trees. *Journal of Theoretical Biology*, 190:15–36.

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