Journal of Theoretical Biology 374 (2015) 54–59

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Do tree split probabilities determine the branch lengths?

Benny Chor [a], Mike Steel [b,*]

[a] School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel
[b] Biomathematics Research Centre, School of Mathematics and Statistics, University of Canterbury, Christchurch 8140, New Zealand

HIGHLIGHTS

- When are phylogenetic tree branch lengths determined by tree split probabilities?
- We prove that this holds for any tree when the branch lengths are sufficiently small.
- We prove that it also holds for trees with up to four leaves, without further assumptions.
- Our results extend to certain models with more than 2 states.

ARTICLE INFO

Article history:
Received 24 October 2014
Received in revised form
27 January 2015
Accepted 19 March 2015
Available online 3 April 2015

Keywords:
Phylogenetic tree reconstruction
Evolutionary model
Markov process
Hadamard transform
Systems of polynomial equations
Inverse function theorem

ABSTRACT

The evolution of aligned DNA sequence sites is generally modeled by a Markov process operating along the edges of a phylogenetic tree. It is well known that the probability distribution on the site patterns at the tips of the tree determines the tree topology, and its branch lengths. However, the number of patterns is typically much larger than the number of edges, suggesting considerable redundancy in the branch length estimation. In this paper we ask whether the probabilities of just the 'edge-specific' patterns (the ones that correspond to a change of state on a single edge) suffice to recover the branch lengths of the tree, under a symmetric 2-state Markov process. We first show that this holds provided the branch lengths are sufficiently short, by applying the inverse function theorem. We then consider whether this restriction to short branch lengths is necessary. We show that for trees with up to four leaves it can be lifted. This leaves open the interesting question of whether this holds in general. Our results also extend to certain Markov processes on more than 2-states, such as the Jukes–Cantor model.

1. Background

When a discrete character evolves on a tree under a Markov process, the probability distribution on site patterns at the leaves of the tree is determined by the tree and its branch lengths (Felsenstein, 2004; Semple and Steel, 2003). What is less obvious is that this process is invertible for many models – that is, the probability distribution on site patterns at the leaves uniquely identifies both the tree and its branch lengths.

This fundamental property underlies all statistical approaches for inferring evolutionary relationships from aligned genetic sequence data. In this setting, the 'discrete character' refers to the pattern of nucleotides across the species at each site, and the frequency of this pattern across the sequences provides some estimate of the probability of that pattern. In this paper we are interested in what the probability distribution says about the branch lengths of the underlying tree (we will assume this topology is known). Notice that the number of site patterns grows exponentially with the number n of leaves, yet the number of branches of the tree (for which the branch lengths are being estimated) grows linearly with n. For example, in the case of a symmetric 2-state model, there are effectively 2^{n-1} site patterns, while the number of edges is between n (for the star tree) to $2n-3$ (for a completely resolved binary tree).

This suggests a basic question – do we need all the site pattern probabilities to infer the branch lengths? More precisely, if a tree has k edges (branches), are there k site patterns whose probabilities under the model might identify the lengths of these branches?

One motivation for this question is that in practice, many site patterns will simply never occur (indeed most will not, if our sequence length grows at most polynomially with n, since the number of site patterns grows exponentially with n). This is a problem if we try to estimate pattern probabilities from their relative frequency.

There is an natural candidate for a particular choice of k site patterns – for each edge we take the site pattern in which all the leaves on one side of the edge are in one state, and all the leaves on the other side of the edge are in a different state – we refer to

* Corresponding author. Fax: +64 33642587.
E-mail addresses: benny@cs.tau.ac.il (B. Chor), mike.steel@canterbury.ac.nz (M. Steel).

http://dx.doi.org/10.1016/j.jtbi.2015.03.024
0022-5193/

such a site pattern as a *tree split* for this edge. From a practical perspective, the tree splits are patterns that are likely to be observed in the data, since they require just one change of state in the tree. They also correspond to the primary divisions of the species into two groups (e.g. vertebrates vs. invertebrates) and so have a clear phylogenetic meaning.

The question of whether the tree split probabilities determine the branch lengths is a delicate one – we prove that for the 2-state symmetric model, the answer is yes for 4-leaf resolved (binary) trees and for 4-leaf star trees, and we conjecture that it holds true for arbitrary phylogenetic trees. This conjecture is supported by our proof that the branch lengths are determined by the tree split probabilities for any tree (on any number of leaves) when these branch lengths are close to zero.

Our approach exploits the Hadamard representation for the 2-state model (Hendy and Penny, 1989, 1993), as well as computational (symbolic) algebraic analysis tools. In the concluding comments we point out how our results also extend to certain 4-state model, including the Jukes–Cantor and Kimura 2ST models, or more generally to certain models with an even number of states. Our results also complement other recent algebraic analysis of models on trees with a small number of leaves, including Klaere and Liebscher (2012) and Sumner and Jarvis (2009).

2. Model and notations

In the Neyman 2-state model (Neyman, 1971), each character admits one out of two states, for example, purines and pyrimidines. Without loss of generality, we denote these states by 0 and 1. We use the symmetric Poisson model, where for each edge e of the tree T, there is a corresponding probability p_e $(0 \le p_e < 1/2)$ that the character states at the two incident vertices of e differ, and this probability is independent of the state at the initial vertex. For a 2-state character, this probability p_e that the endpoints of e at a site are in different states is the same as the probability of having an odd number of substitutions per site across the edge e. The expected number of substitutions per site across the edge e equals $q_e = -\frac{1}{2}\ln(1-2p_e)$. The value q_e is referred to the *(branch) length* of edge e. Measuring the tree edges by q_e $(0 \le q_e < \infty)$, we get an additive measure on the tree, namely the expected number of substitutions between each pair of leaves (because expected values are additive). Such a phylogenetic tree with branch lengths is a probabilistic model that emits any given pattern of states at its leaves with a well defined probability. Notice that the limits $q_e \to 0$ and $q_e \to \infty$ correspond to the limits $p_e \to 0$ and $p_e \to \frac{1}{2}$, respectively.

The observed sequences at the leaves can be represented by a matrix, ψ, where the number of rows equals the number of species ($n=4$ in our case), and the number of columns equals the common length of the sequences. In biological terms, this matrix is just a data alignment – that is, each column consists of an aligned site of (binary) character states across the n species. For 2-state characters, it is convenient to 'summarize' the observed data ψ by a vector of observed frequencies of splits, $\hat{\mathbf{s}}$. This vector simply counts how many sites share any specific pattern. Under a fully symmetric 2-state model, the probability of a pattern is equal to that of its complement (where all 0 and 1 are interchanged). We make the following convention about indexing the patterns obtained in the sequences over $n=4$ species, labeled 1, 2, 3, and 4, with the sequences $x_1, x_2, x_3, x_4 \in \{0,1\}^n$: We identify a site pattern by the subset of species 1, 2, 3 whose character at that site is different from that of species 4. More generally (i.e. for any value of n) for every $\alpha \subseteq \{1,\ldots,n-1\}$, an *$\alpha$-split pattern* is a pattern where all taxa in the subset α have one character (0 or 1), and the taxa in the complement subset have the second character (there are two such patterns). The value $\hat{s}_\alpha$ equals the number of times that α-split patterns appear in the data. For $n=4$ there are $2^3 = 8$ possible patterns, and the vector of observed sequence frequencies is $\hat{\mathbf{s}} = [\hat{s}_\varnothing, \hat{s}_1, \hat{s}_2, \hat{s}_{12}, \hat{s}_3, \hat{s}_{13}, \hat{s}_{23}, \hat{s}_{123}]$.

3. The tree split probabilities determine the branch lengths locally

In this section, we show that the (multivariate) inverse function theorem implies that branch lengths can be recovered from tree split probabilities provided the branch lengths are not too large. Recall that the inverse function theorem provides a sufficient condition for a function f from an N-dimensional space A to another N-dimensional space B to be invertible in the neighborhood of some point $a \in A$. This condition is that the function f be continuously differentiable in a neighborhood of a, and its Jacobian matrix (of first derivatives) be non-singular at a. In this paper, a *phylogenetic tree* refers to an unrooted tree with labeled leaves, and with every internal vertex having degree strictly greater than 2 (Semple and Steel, 2003).

Theorem 3.1. *Let T be any phylogenetic tree, on any number of leaves. Under the 2-state symmetric model, the probabilities of the tree splits determine the branch lengths of T in some neighborhood of the origin. That is, provided all the branch lengths are sufficiently small then they can be uniquely recovered from the tree split probabilities they induce.*

Proof. To simplify notation in this section, given a phylogenetic tree T with k edges, label the edges $e_1, e_2, \ldots, e_k$. For each $i \in \{1,\ldots,k\}$, let α_i denote the tree split corresponding to e_i; let s_i be the probability of generating the pattern α_i on T under the symmetric 2-state model; let q_i be the branch length of edge e_i, and let $p_i = \frac{1}{2}(1-e^{-2q_i})$, which is the probability of a change of state on edge $e_i = (v_1^i, v_2^i)$. Consider the two subtrees of T which result from removing the edge e_i (but not the nodes v_1^i, v_2^i). Let T_1, T_2 denote the resulting subtrees, rooted at v_1^i, v_2^i, respectively. Let Q_i^1 be the probability of the event 'all leaves of T_1 are in the same state as v_1^i', and Q_i^2 be the probability of the event 'all leaves of T_2 are in the same state as v_2^i'. Let R_i^1 denote the probability of the event 'all leaves of T_1 are in the same state and they differ from the state of v_1^i', and R_i^2 denote the probability of the event 'all leaves of T_2 are in the same state and they differ from the state of v_2^i'. We note that under the 2-state symmetric model, changes of state on different edges are independent events. By considering whether or not there is a change of state on edge e_i, the following identity holds for all i:

$$s_i = p_i Q_i^1 Q_i^2 + (1-p_i)(Q_i^1 R_i^2 + R_i^1 Q_i^2). \tag{1}$$

Note that $Q_i^1, Q_i^2, R_i^1, R_i^2$ involves only the terms p_j for $j \ne i$, and that when all the p_j terms are zero we have

$$R_i^1 \mid_{\mathbf{p}=\mathbf{0}} = R_i^2 \mid_{\mathbf{p}=\mathbf{0}} = 0 \quad \text{and} \quad Q_i^1 \mid_{\mathbf{p}=\mathbf{0}} = Q_i^2 \mid_{\mathbf{p}=\mathbf{0}} = 1. \tag{2}$$

Now, consider the Jacobian matrix of partial derivatives

$$\mathbf{J} = \left[\frac{\partial s_i}{\partial p_j}\right].$$

From Eq. (1) and the fact that p_i does not appear in Q_i^1, Q_i^2 and R_i^1, R_i^2 we have

$$\frac{\partial s_i}{\partial p_i} = Q_i^1 Q_i^2 - (Q_i^1 R_i^2 + R_i^1 Q_i^2)$$

and from Eq. (2) this equals 1 when $\mathbf{p} = \mathbf{0}$.

Similarly, for $j \neq i$, Eq. (1) gives

$$\frac{\partial s_i}{\partial p_j} = p_i \frac{\partial}{\partial p_j} Q_i^1 Q_i^2 + (1-p_i)\frac{\partial}{\partial p_j}(Q_i^1 R_i^2 + R_i^1 Q_i^2),$$

and so, when $\mathbf{p}=\mathbf{0}$ the first term on the right vanishes (since $p_i = 0$) and we have

$$\frac{\partial s_i}{\partial p_j}|_{\mathbf{p}=\mathbf{0}} = \frac{\partial}{\partial p_j}(Q_i^1 R_i^2 + R_i^1 Q_i^2)|_{\mathbf{p}=\mathbf{0}}.$$

Now, notice that $Q_i^1 R_i^2 + R_i^1 Q_i^2$ is a multinomial polynomial of the variables p_k, $k \neq i$. We argue that this polynomial has no term of the form cp_j for a constant $c \neq 0$. Suppose otherwise, then setting $p_j = \frac{1}{4}$ and $p_k = 0$ for all $k \neq j$, Eq. (1) shows that the pattern α_i occurs with probability $\frac{1}{4}c \neq 0$ under such an edge probability assignment. However, since there is no vertex of degree 2 in T, edge e_i is the only edge for which α_i has positive probability when all but one p-term is set to zero. This contradicts the assumption that we are in the setting where $j \neq i$. Since $Q_i^1 R_i^2 + R_i^1 Q_i^2$ has no term of the form cp_j ($c \neq 0$) it follows that:

$$\frac{\partial}{\partial p_j}(Q_i^1 R_i^2 + R_i^1 Q_i^2)|_{\mathbf{p}=\mathbf{0}} = 0.$$

Summarizing, we have $\partial s_i / \partial p_i|_{\mathbf{p}=\mathbf{0}} = \delta_{ij}$ (Kronecker delta), and so at $\mathbf{p}=\mathbf{0}$, $\mathbf{J}$ is the $k \times k$ identity matrix. Since the map $\mathbf{p} \mapsto (s_1, \ldots, s_k)$ is a polynomial map, and therefore continuously differentiable, and since the Jacobian of this map is invertible at $\mathbf{p}=\mathbf{0}$, the conditions for the multivariate inverse function theorem apply at this point. Thus, for some neighborhood of $\mathbf{p}=\mathbf{0}$, the function $(p_1, \ldots, p_k) \mapsto (s_1, \ldots, s_k)$ is invertible. Finally, observe that the map from $[0, \infty)^k$ onto $[0, 1/2)^k$ defined by $(q_1, \ldots q_k) \mapsto (p_1, \ldots p_k)$ is invertible, and so the theorem now follows. □

Theorem 3.1 begs the question: how small do the branch lengths need to be in order for invertibility to hold? In the next section we show that we can obtain invertibility holds for all finite branch lengths for trees with at most four leaves.

4. Exact analysis for small trees

For phylogenetic trees with two and three leaves it is easily verified that the tree splits determine the branch lengths with no restriction required on the size of these branch lengths. Thus we will consider trees with four leaves, of which there are two types – the resolved binary tree and the star tree. We will show that the probability distribution on tree splits determines the branch lengths of the binary tree for all non-negative branch lengths, and then establish that the same holds for the star tree.

A useful tool in this analysis is the Hadamard representation of the 2-state symmetric model, which we first recall.

4.1. Hadamard representation

Given a tree T with n leaves and edge probabilities $\mathbf{p} = [p_e]_{e \in E(T)} (0 \leq p_e < \frac{1}{2})$, the probability of generating an α-split pattern ($\alpha \subseteq \{1, \ldots, n-1\}$) is determined (and is equal for all sites). Denote this probability by $\mathbf{s}_\alpha = \Pr(\alpha\text{-split}|T, p)$.

Using the same indexing scheme as above,when $n=4$, we define the vector of pattern generation probabilities $\mathbf{s} = [s_\emptyset, s_1, s_2, s_{12}, s_3, s_{13}, s_{23}, s_{123}]$. This vector is termed the expected sequence spectrum in Hendy and Penny (1993), where the indexing scheme is explained as well.

Even though in principle, the edge lengths $q_e = [q_e]_{e \in E(T)}$ determine the vector $\mathbf{s}$ of pattern generation probabilities, it is not obvious how to actually compute this vector, given the edge lengths. This is where the Hadamard conjugation (Hendy and Penny, 1993; Hendy et al., 1994) comes in. This transformation yields a powerful tool, which greatly simplifies and unifies the analysis of phylogenetic data.

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

We will use a special family of Hadamard matrices, whose sizes, ℓ, are powers of 2 (MacWilliams and Sloane, 1977), defined inductively for $n \geq 0$ by $H_0 = [1]$ and $H_{n+1} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}$. 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 $H_n^{-1} = 2^{-n} H_n$.

Proposition 4.1 (*Hendy and Penny, 1993*). *If* $\mathbf{p} < \mathbf{1/2}$ *then* $\mathbf{s} = \mathbf{s}(\mathbf{q}) = H^{-1} \exp(H\mathbf{q})$ *where* $H = H_{n-1}$, *namely for* $\alpha \subseteq \{1, \ldots, n-1\}$,

$$\mathbf{s}_\alpha = 2^{-(n-1)} \sum_\gamma h_{\alpha\gamma} \left(\exp\left(\sum_\delta h_{\gamma\delta} q_\delta \right) \right).$$

We note that the transformation is reversible, so if $H\mathbf{s} > \mathbf{0}$ (all entries are positive) then $\mathbf{q} = \mathbf{q}(\mathbf{s}) = H^{-1} \ln(H\mathbf{s})$. Thus, when $n=4$, the eight components in the expected sequence spectrum uniquely determine the five edge lengths of the corresponding four taxa tree.

The question we explore is if the five components that correspond to splits in the tree, namely $s_1, s_2, s_{12}, s_3, s_{123}$ also determine the five edge lengths $q_1, q_2, q_{12}, q_3, q_{123}$. In other words, is the mapping $(q_1, q_2, q_{12}, q_3, q_{123}) \mapsto (s_1, s_2, s_{12}, s_3, s_{123})$ one to one?

4.2. Resolved tree on four leaves

In this case, the tree has five edges with non-negative lengths, $0 \leq q_\alpha < \infty$. To describe the mapping $(q_1, q_2, q_{12}, q_3, q_{123}) \mapsto (s_1, s_2, s_{12}, s_3, s_{123})$, let us first denote

$$a_1 = e^{-2q_1}, \quad a_2 = e^{-2q_2}, \quad a_{12} = e^{-2q_{12}}, \quad a_3 = e^{-2q_3}, \quad a_{123} = e^{-2q_{123}},$$

where the a_α are in the interval $(0, 1]$ (corresponding to edge lengths $0 \leq q_\alpha < \infty$). Then by the Hadamard transform (with the 8-by-8 matrix, H_3)

$$\begin{aligned}
8s_1 &= 1 + a_{123}a_3 - a_1a_2 + a_{12}a_2a_3 + a_{12}a_{123}a_2 - a_1a_{12}a_3 - a_1a_{12}a_{123} \\
&\quad - a_1a_{123}a_2a_3 \\
8s_2 &= 1 + a_{123}a_3 - a_1a_2 - a_{12}a_2a_3 - a_{12}a_{123}a_2 + a_1a_{12}a_3 \\
&\quad + a_1a_{12}a_{123} - a_1a_{123}a_2a_3 \\
8s_{12} &= 1 + a_{123}a_3 + a_1a_2 - a_{12}a_2a_3 - a_{12}a_{123}a_2 - a_1a_{12}a_3 \\
&\quad - a_1a_{12}a_{123} + a_1a_{123}a_2a_3 \\
8s_3 &= 1 - a_{123}a_3 + a_1a_2 - a_{12}a_2a_3 + a_{12}a_{123}a_2 - a_1a_{12}a_3 \\
&\quad + a_1a_{12}a_{123} - a_1a_{123}a_2a_3 \\
8s_{123} &= 1 - a_{123}a_3 + a_1a_2 + a_{12}a_2a_3 - a_{12}a_{123}a_2 \\
&\quad + a_1a_{12}a_3 - a_1a_{12}a_{123} - a_1a_{123}a_2a_3
\end{aligned}$$

while the pattern probabilities of the three non-tree splits are

$$\begin{aligned}
8s_{13} &= 1 - a_{123}a_3 - a_1a_2 - a_{12}a_2a_3 + a_{12}a_{123}a_2 + a_1a_{12}a_3 - a_1a_{12}a_{123} \\
&\quad + a_1a_{123}a_2a_3 \\
8s_{23} &= 1 - a_{123}a_3 - a_1a_2 + a_{12}a_2a_3 - a_{12}a_{123}a_2 - a_1a_{12}a_3 \\
&\quad + a_1a_{12}a_{123} + a_1a_{123}a_2a_3 \\
8s_\emptyset &= 1 + a_{123}a_3 + a_1a_2 + a_{12}a_2a_3 + a_{12}a_{123}a_2 + a_1a_{12}a_3 \\
&\quad + a_1a_{12}a_{123} + a_1a_{123}a_2a_3
\end{aligned}$$

For example, for the tree in Fig. 1, in the special case where all five edges have the same branch length q let $a=e^{-2q}$. Then $s_1=s_2=s_3=s_{123}=\frac{1}{8}(1-a^4)$, and $s_{12}=\frac{1}{8}(1+2a^2-4a^3+a^4)$. For the two non-tree splits we have $s_{13}=s_{23}=\frac{1}{8}(1-2a^2+a^4)$, and, in addition, $s_\varnothing=\frac{1}{8}(1+2a^2+4a^3+a^4)$.

Returning to the general setting, notice that

$$1-2s_1-2s_2-2s_3-2s_{123}=a_1a_{123}a_2a_3\in(0,1]. \tag{3}$$

Moreover, it can be checked, using the above mentioned identities, that

$$s_{13}+s_{23}=\tfrac{1}{4}(1-a_1a_2-a_3a_{123}+a_1a_2a_3a_{123}),$$

and

$$s_\varnothing+s_{12}=\tfrac{1}{4}(1+a_1a_2+a_3a_{123}+a_1a_2a_3a_{123}),$$

and so the following linear inequality holds:

$$s_\varnothing+s_{12}-s_{13}-s_{23}=\tfrac{1}{2}(a_1a_2+a_3a_{123})\geq 0. \tag{4}$$

This inequality will be used shortly to identify which one of two solutions to a quadratic equation is valid. Moreover, the following two inequalities (used in the next section) can also be readily verified from the eight s_α identities, mentioned above:

$$s_\varnothing+s_{23}-s_{12}-s_{13}=\tfrac{1}{2}(a_1a_{12}a_{123}+a_{12}a_2a_3)\geq 0. \tag{5}$$

$$s_\varnothing+s_{13}-s_{12}-s_{23}=\tfrac{1}{2}(a_1a_{12}a_3+a_{12}a_{123}a_2)\geq 0. \tag{6}$$

Since the Hadamard transformation $\mathbf{s}=\mathbf{s}(\mathbf{q})=H^{-1}\exp(H\mathbf{q})$ is one-to-one, to show that the mapping $(q_1,q_2,q_{12},q_3,q_{123})\mapsto(s_1,s_2,s_{12},s_3,s_{123})$ is one to one for all choices of non-negative branch lengths, it suffices to show that s_{12}, s_{23} and $s_\varnothing$ can be determined by the five remaining s-values, corresponding to tree splits.

The Hadamard transformation allows us to express each q_α in terms of the seven values $(s_1,s_2,s_{12},s_3,s_{13},s_{23},s_{123})$ ($s_\varnothing$ equals 1 minus the sum of these seven s_α). For the tree T of Fig. 1, two splits are not realized by any edge of T and this corresponds to the identities $q_{13}=q_{23}=0$. This gives rise to two invariants involving the seven s_α (Cavender and Felsenstein, 1987). After some manipulation and simplification, we derive the following two invariants (Chor et al., 2000):

$$\begin{aligned}0=&\,1-2s_1-2s_2-2s_3-2s_{123}\\&-(-1+2s_1+2s_2+2s_{13}+2s_{23})(-1+2s_3+2s_{13}+2s_{23}+2s_{123})\end{aligned}$$

$$\begin{aligned}0=&\,(-1+2s_1+2s_{12}+2s_{13}+2s_{123})(-1+2s_2+2s_{12}+2s_3+2s_{13})\\&-(-1+2s_2+2s_{12}+2s_{23}+2s_{123})(-1+2s_1+2s_{12}+2s_3+2s_{23})\end{aligned}$$

From the first equation, we get a quadratic equation for $x=s_{13}+s_{23}$:

$$\begin{aligned}0=&\,1-2s_1-2s_2-2s_3-2s_{123}\\&-(-1+2s_1+2s_2+2x)(-1+2s_3+2s_{123}+2x)\end{aligned}$$

Let $u=1-2s_1-2s_2$, $v=1-2s_3-2s_{123}$ and $w=1-2s_1-2s_2-2s_3-2s_{123}$. The quadratic equation for x can now be written as

$$0=w-(-u+2x)(-v+2x),$$

or

$$4x^2-2x(u+v)-(w-uv)=0.$$

Solving this equation gives

$$x=\tfrac{1}{4}\left[(u+v)\pm\sqrt{(u+v)^2+4(w-uv)}\right]. \tag{7}$$

Although there are two possible real solutions for x, only the negative square root provides a valid solution. To see this notice that

$$(u+v)/4=(1-s_1-s_2-s_3-s_{123})/2=(s_\varnothing+s_{12}+s_{13}+s_{23})/2,$$

and since, by inequality (4), $x=s_{13}+s_{23}\leq s_\varnothing+s_{12}$ it follows that $x\leq(u+v)/4$, and thus only the negative square root term can apply in (7). For this solution of $x=s_{13}+s_{23}$, the second invariant can be rewritten as a quadratic equation in $y=s_{13}-s_{23}$, where all coefficients are known

$$\begin{aligned}0=&\,(-1+2s_1+2s_{12}+2s_{123}+x+y)\\&\times(-1+2s_2+2s_{12}+2s_3+x+y)\\&-(-1+2s_2+2s_{12}+2s_{123}+x-y)\\&\times(-1+2s_1+2s_{12}+2s_3+x-y)\end{aligned}$$

In fact, this is really just a linear equation in y, since the y^2 terms cancel, to give

$$y=(AB-CD)/(A+B+C+D),$$

where

$$A=-1+2s_1+2s_{12}+2s_{123}+x,$$

$$B=-1+2s_2+2s_{12}+2s_3+x,$$

$$C=-1+2s_2+2s_{12}+2s_{123}+x,$$

$$D=-1+2s_1+2s_{12}+2s_3+x.$$

In summary, we have a unique solution for $x=s_{13}+s_{23}$, and for this we have a unique solution for $y=s_{13}-s_{23}$. These two uniquely determine $s_{13}=\frac{1}{2}(x+y)$ and $s_{23}=\frac{1}{2}(x-y)$. Once s_{13},s_{23} are determined, the final value $s_\varnothing$ is determined by the linear invariant

$$s_\varnothing+s_1+s_2+s_{12}+s_3+s_{13}+s_{23}+s_{123}=1.$$

Returning again to the special case of the tree in Fig. 1 with all five branch lengths equal to q, and $a=e^{-2q}$, the above analysis yields $u=v=\frac{1}{2}(1+a^4)$ and $w=a^4$, so that the negative root of Eq. (7) is

$$x=\tfrac{1}{4}(1+a^4)=s_{13}+s_{23}.$$

Then $A=B=C=D$ and so $y=0$, which gives $s_{13}=s_{23}=\frac{1}{2}x=\frac{1}{8}(1+a^4)$.

4.3. The star tree on four leaves

The star tree is a special case of the resolved tree on four nodes, where the internal edge q_{12} is of length 0. However, the uniqueness result we have for the resolved tree does *not* directly imply a similar one for the star, as we have one fewer observed pattern probability. We want to show that in this case, the mapping $(q_1,q_2,q_3,q_{123})\mapsto(s_1,s_2,s_3,s_{123})$ is one to one. While this does not directly follow from the previous result, a similar algebraic approach does work here as well:

$$\begin{aligned}8s_1&=1+a_{123}a_3-a_1a_2+a_2a_3+a_{123}a_2-a_1a_3-a_1a_{123}-a_1a_{123}a_2a_3\\8s_2&=1+a_{123}a_3-a_1a_2-a_2a_3-a_{123}a_2+a_1a_3+a_1a_{123}-a_1a_{123}a_2a_3\\8s_3&=1-a_{123}a_3+a_1a_2-a_2a_3+a_{123}a_2-a_1a_3+a_1a_{123}-a_1a_{123}a_2a_3\\8s_{123}&=1-a_{123}a_3+a_1a_2+a_2a_3-a_{123}a_2+a_1a_3-a_1a_{123}-a_1a_{123}a_2a_3\end{aligned}$$

In case of the star tree on $n=4$ leaves, only the four edges with pendant leaves, q_1,q_2,q_3,q_{123}, could have non-zero lengths. The splits corresponding to internal edges are not present in this tree and this leads to the identities: $q_{12},q_{13},q_{23}=0$. Expressing these edges in terms of $\mathbf{s}$, slightly manipulating and simplifying the expressions, we get the following three quadratic invariants:

$$\begin{aligned}(\mathrm{I}_1)\quad&(-1+2s_1+2s_2+2s_3+2s_{123})\\&=(-1+2s_1+2s_{12}+2s_{13}+2s_{123})\cdot(-1+2s_2+2s_{12}+2s_3+2s_{13}).\end{aligned}$$

$$\begin{aligned}(\mathrm{I}_2)\quad&(-1+2s_1+2s_2+2s_3+2s_{123})\\&=(-1+2s_2+2s_{12}+2s_{23}+2s_{123})\cdot(-1+2s_1+2s_{12}+2s_3+2s_{23}).\end{aligned}$$

$$(I_3)\quad (-1+2s_1+2s_2+2s_3+2s_{123}) = (-1+2s_1+2s_2+2s_{13}+2s_{23})\cdot(-1+2s_3+2s_{13}+2s_{23}+2s_{123}).$$

Let $x=s_{12}+s_{13}$, $y=s_{12}+s_{23}$, and $z=s_{13}+s_{23}$. Note that $x+y+z=2s_{12}+2s_{13}+2s_{23}$. The three equations (I_1)-(I_3) become three quadratic equations, each in one of these variables, which can therefore be solved separately. Recall that s_1, s_2, s_3, s_{123} are known.

$$(I_1)\quad (-1+2s_1+2s_2+2s_3+2s_{123}) = (-1+2s_1+2s_{123}+2x)\cdot(-1+2s_2+2s_3+2x).$$

$$(I_2)\quad (-1+2s_1+2s_2+2s_3+2s_{123}) = (-1+2s_2+2s_{123}+2y)\cdot(-1+2s_1+2s_3+2y).$$

$$(I_3)\quad (-1+2s_1+2s_2+2s_3+2s_{123}) = (-1+2s_1+2s_2+2z)\cdot(-1+2s_3+2s_{123}+2z).$$

Now all of these equations are of the same quadratic form as those used to determine x in the resolved tree setting. Indeed, Eq. (I_3) is exactly the same (with z playing the role of the earlier x) and the inequality (4) required to exclude one of the two roots of the quadratic equation applies just as validly for the star tree; as do inequalities (5) and (6) These latter two inequalities show that the quadratic equations for x and y in Eqs. (I_1) and (I_2) (respectively) have unique solutions. Having uniquely determined x, y and z, notice that

$$s_{12}=\tfrac{1}{2}(x+y-z),\quad s_{13}=\tfrac{1}{2}(x+z-y)\quad\text{and}\quad s_{23}=\tfrac{1}{2}(y+z-x).$$

Finally, as in the resolved tree case, $s_\varnothing$ is determined by the linear constraint:

$$s_\varnothing+s_1+s_2+s_{12}+s_3+s_{13}+s_{23}+s_{123}=1.$$

4.3.1. The star tree and the inverse function theorem

Let T_n denote the star tree on n leaves. If p_i denotes the probability of a state change (0 to 1 or visa versa) on the edge incident with leaf i, $i=1,2,\ldots,n$, then the probability s_i of obtaining the tree split pattern that separates leaf i from the other leaves (i.e. the probability that leaf i has a different state to all the other leaves) is given, in terms of $p_1, p_2, \ldots, p_n$ $(0\le p_i<1/2)$ by

$$s_i = p_i\prod_{j\neq i}(1-p_j)+(1-p_i)\prod_{j\neq i}p_j. \tag{8}$$

This formula is easily verified by observing that there are two ways to obtain the tree split that separates leaf i from the other leaves – either there is a change on the edge incident with leaf i, and no changes on any of the other edges (the first term), or there is no change on the leaf incident with leaf i but there are changes on all the other edges (the second term).

In the special case with $n=4$, consider the 4-by-4 Jacobian matrix:

$$\mathbf{J}=\left[\frac{\partial s_i}{\partial p_j}\right]_{1\le i,j\le 4}.$$

Suppose $\mathbf{J}$ is invertible (the determinant is non zero) in a certain domain, and the underlying functions (s_1, s_2, s_3, s_4 in our case) are differentiable. Then by the (multivariate) inverse function theorem, these functions are locally invertible.

We employed the symbolic mathematical software Maple (version 17) to compute the Jacobian and its determinant. It turns out that the determinant can be expressed as a product of three simple multivariate polynomials:

$$\begin{aligned}\det(\mathbf{J}) &= (2p_1p_2-p_1+1-p_2-p_3+2p_3p_4-p_4)\\ &\times(2p_1p_3-p_1+1-p_2-p_3+2p_2p_4-p_4)\\ &\times(2p_1p_4-p_1+1-p_2+p_2p_3-p_3-p_4)\end{aligned}$$

Each factor has a similar structure, and can be represented as follows:

$$(2p_1p_2-p_1+1-p_2-p_3+2p_3p_4-p_4) = \tfrac{1}{2}(2p_1-1)\cdot(2p_2-1)+\tfrac{1}{2}(2p_3-1)\cdot(2p_4-1).$$

It is clear that if all p_i are in the range $0\le p_i<1/2$, then each factor is strictly positive, so the determinant is positive, as desired. So by the (multivariate) inverse function theorem, under these conditions, the functions are *locally* invertible for any neighborhood in the domain $0\le p_1, p_2, p_3, p_4<1/2$. By contrast, the invertibility results in the previous section are *global*.

5. Concluding comments

We have made the first steps towards settling the question of whether the probability distribution of tree splits suffices to determine the branch lengths in the 2-state symmetric model. We have shown that this holds in two cases:

- for any tree if the branch lengths are sufficiently short, and
- the branch lengths are strictly positive and the tree has at most four leaves.

We conjecture that invertibility holds for any phylogenetic tree (with any number of leaves) and across the space of all non-negative branch lengths, however a proof will require a different or modified approach. The algebraic approach may be extended to slightly larger numbers of taxa, but the underlying complexity of solving such systems of polynomial equations is quite prohibitive as the number of taxa grows. For the approach employing the inverse function theorem to be applicable, it will be required to show global, rather than just local, invertibility.

Our results pertain to the symmetric 2-state model, however, they are also directly relevant to certain 4-state models, such as the Jukes–Cantor model, or the Kimura 2ST model. To see why, notice that under the Jukes–Cantor model, if we partition the four bases into two sets of size two (e.g. $\{A,T\},\{G,C\}$ or one of the other two possible pairings) and regard these two sets as (hyper)-states, then the corresponding process on a tree is precisely described by a symmetric 2-state model. Thus, from just the probabilities of the resulting tree splits, we have shown how in certain settings it is possible to identify the branch lengths of the tree. From these, the corresponding branch lengths for the original Jukes–Cantor model (as measured by the expected number of substitutions on each edge) are then obtained by multiplying the 2-state branch lengths by the factor $\frac{3}{2}$. For the Kimura 2ST model, we need to use the following partition $\{A,G\},\{C,T\}$ corresponding to purines and pyrimidines. In that case, once again we obtain an induced symmetric 2-state model on these (hyper)-states, and once again the branch lengths under the Kimura 2ST model will be obtained from those for the 2-state process by multiplying by a factor that depends on the transition-to-transversion rate. There is nothing special about four states here either; any number of even states allows for models for which a 'lumped' 2-state process follows the 2-state symmetric model.

Finally, we discuss briefly a related but simpler question, namely whether there is any collection of k linear combinations of the site pattern probabilities that identifies the branch lengths. In this case, the answer is 'yes' for any number of leaves (and any tree). This can be seen by combining three observations that apply for a wide range of substitution models (not just on two states): (i) the expected probability that any pair of leaves i, j differ in state is a sum of certain pattern probabilities, (ii) this expected

probability can be transformed to give the sum of the branch lengths between i and j in the tree (Felsenstein, 2004), and (iii) k such (carefully selected) pairs of path distances suffice to recover the length of all the k edges (Dress et al., 2012).

For example, for the quartet tree $T = 12|34$, and the 2-state symmetric model, consider the five linear combinations $L_{12} = s_1+s_2+s_{13}+s_{23}$, $L_{13} = s_1+s_3+s_{12}+s_{23}$, $L_{14} = s_1+s_{12}+s_{13}+s_{123}$, $L_{23} = s_2+s_3+s_{12}+s_{13}$, and $L_{34} = s_3+s_{13}+s_{23}+s_{123}$. Then L_{xy} is the probability that leaf x and leaf y are in different states, and so $\kappa_{xy} = -\frac{1}{2}\log(1-2L_{xy})$ is the length of the path between leaf x and y in T under the edge lengths determined by $\mathbf{q}$. Those five κ values now uniquely determine the five branch lengths of T (this may be verified directly, or by observing that the five pairs of leaves described form a 'pointed x-cover' of T for the leaf choice $x=1$ and so, by Theorem 7 of Dress et al. (2012), the associated κ values determine the branch lengths).

Acknowledgments

We thank the two anonymous reviewers for several helpful suggestions. This work was initiated while B.C. was on a sabbatical visit at the University of Canterbury, New Zealand. M.S. thanks the Allan Wilson Centre for funding support. The work of B.C. was not supported by the ISF (Israeli Science Foundation) or any other funding agency.

References

Cavender, J.A., Felsenstein, J., 1987. Invariants of phylogenies in a simple case with discrete states. J. Classif. 4, 57–71.
Chor, B., Hendy, M.D., Holland, B.R., Penny, D., 2000. Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol. Biol. Evol. 17 (10), 1529–1541.
Dress, A.W.M., Huber, K.T., Steel, M., 2012. 'Lassoing' a phylogenetic tree. I.: Basic properties, shellings, and covers. J. Math. Biol. 65 (1), 77–105.
Felsenstein, J., 2004. Inferring Phylogenies, Sinauer Associates, Sunderland, MA.
Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Biol. 38 (4), 297–309.
Hendy, M.D., Penny, D., 1993. Spectral analysis of phylogenetic data. J. Classif. 10, 5–24.
Hendy, M.D., Penny, D., Steel, M.A., 1994. A discrete Fourier analysis for evolutionary trees. Proc. Natl. Acad. Sci. USA 91, 3339–3343.
Klaere, S., Liebscher, V., 2012. An algebraic analysis of the two-state Markov model on tripod trees. Math. Biosci. 237 (1–2), 38–48.
MacWilliams, F., Sloane, N., 1977. The Theory of Error Correcting Codes. Elsevier Science Publishers, Amsterdam, North-Holland.
Neyman, J., 1971. Molecular studies of evolution: a source of novel statistical problems. In: Gupta, S.S., Yackel, J. (Eds.), Statistical Decision Theory and Related Topics. Academic Press, New York, NY, pp. 1–27.
Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford.
Sumner, J.G., Jarvis, P.D., 2009. Markov invariants and the isotopy subgroup of a quartet tree. J. Theor. Biol. 258 (2), 302–310.