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 $\mathbf{p} = \mathbf{0}$.