

A CONSTRAINT OPTIMIZATION PROBLEM IN PHYLOGENETICS

RAPHAEL HAUSER AND BHALCHANDRA D. THATTE

ABSTRACT. We consider a rooted phylogenetic tree under molecular clock, a two-state character, and a two-state symmetric substitution model (Neyman model). We examine the problem of inferring the ancestral root state by the maximum likelihood and the maximum parsimony methods. In particular, we would like to investigate if there are trees and characters for which the two methods can give different ancestral states. This project is based on a problem suggested by Mike Steel.

A *phylogenetic X -tree* is a tree $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$ on a leaf set $X = \{1, \dots, m\} \subset V(\mathcal{T})$ with no vertices of degree 2 (except the root vertex if we consider a rooted phylogenetic tree). We consider only binary phylogenetic trees.

A *character f* is a function $f : X \rightarrow \mathcal{C}$ for some set $\mathcal{C} := \{c_1, c_2, c_3, \dots, c_r\}$ of r *character states*. For example, one is often interested in 4 states A,T,G,C.

Next we describe the so called r -state symmetric model, also known as the N_r -model. The N_r -model assumes a uniform distribution of states at the root, and assumes equal rates of transitions between any two distinct character states [8]. For any edge $e \in E(\mathcal{T})$, let P_e denote the probability that a substitution from a character state c_i to another character state c_j occurs on edge e for $c_i \neq c_j$. Furthermore, let Q_e denote the probability that no substitution occurs on edge e . In the N_r -model we have $0 \leq P_e \leq 1/r$ for all $e \in E(\mathcal{T})$. We have $(r-1)P_e + Q_e = 1$. The N_r -model assumes that substitutions occur independently on different edges. Note that for $r = 4$, the N_r -model coincides with the Jukes-Cantor model [7]. For $r = 2$, it is known as the Neyman model [8]. See also references [9], [12] for background on phylogenetic methods and models of molecular evolution.

In this project we will only consider two states, say α and β , evolving on a rooted binary phylogenetic tree under the Neyman model.

We consider phylogenetic trees constrained by a *molecular clock*, that is, we suppose that the probability of a change of state from the root ρ to any leaf is a fixed number P independent of the leaf. We sometimes say that P is the height of the tree. In the N_r model, we have $0 \leq P \leq 1/r$ (and in fact $0 \leq P_e \leq 1/r$ for every edge e of the tree).

Suppose we observed a two-state character f on a rooted phylogenetic X -tree \mathcal{T} . We ask the question: what was the ancestral root state? Two of the most popular methods in phylogenetics are the maximum likelihood method [5] and the maximum parsimony method [6]. In the maximum parsimony method (which we call MP), one assigns states to internal nodes of the tree so that the observed data f can be explained with a minimum number of substitutions. Thus it is a purely combinatorial (and model-free) method. In the maximum likelihood method, the root state is the state that maximizes the likelihood of observing the character f .

Problem 1. (Computationally) search for examples of trees \mathcal{T} and characters f for which MP and ML disagree on the root state or verify that they agree on at least small trees.

Equivalence between MP and ML has been examined in a related context in [11].

We illustrate the two methods with a simple example. Figure 1 shows a tree \mathcal{T} on 5 leaves and a character f . The diagram on the left shows the most parsimonious assignment of internal states. The diagram on the right shows substitution probabilities on the edges.

Key words and phrases. phylogenetics, ancestral state reconstruction, Neyman model, maximum parsimony, maximum likelihood.

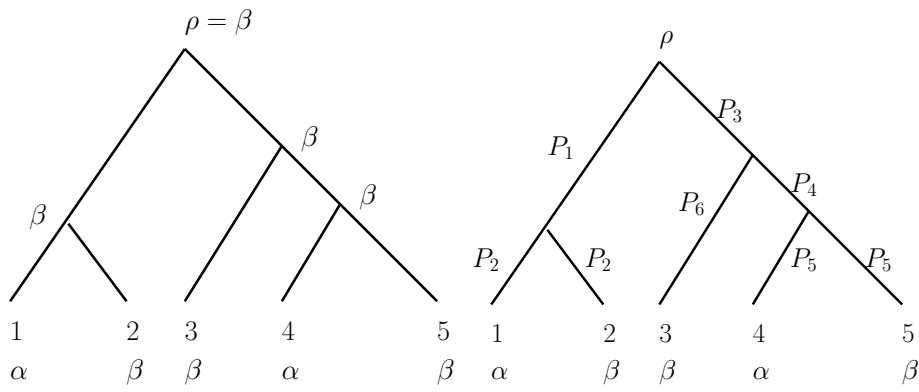


FIGURE 1. Assignment of internal states according to MP (left); the same phylogenetic tree showing substitution probabilities (right).

Next we write the likelihood of observing f on \mathcal{T} . Suppose f_l is the restriction of f on the left subtree (on the leaf set $\{1, 2\}$), and f_r is the restriction of f on the right subtree (on the leaf set $\{3, 4, 5\}$). The likelihood of observing f maybe written as

$$\begin{aligned} P(f|\mathcal{T}, \rho = \alpha) &= P(f_l|\mathcal{T}, \rho = \alpha)P(f_r|\mathcal{T}, \rho = \alpha) \\ &= P_2(1 - P_2) ((1 - P_3)P_6 + P_3(1 - P_6)) P_5(1 - P_5), \end{aligned}$$

which is a polynomial in the edge substitution probabilities. Similarly, one can write $P(f|\mathcal{T}, \rho = \beta)$ as a polynomial in the edge substitution probabilities.

In the maximum likelihood method we first calculate $A = \sup P(f|\mathcal{T}, \rho = \alpha)$, $B = \sup P(f|\mathcal{T}, \rho = \beta)$, where the supremum is obtained over the parameters P_i . Then we assign the root ρ the state α if $A > B$, the state β if $B > A$, and the state $\{\alpha, \beta\}$ if $A = B$. Note that even the maximum parsimony method may assign the state $\{\alpha, \beta\}$ to the root.

Obtaining a global maximum of the likelihood polynomial will be the biggest challenge in this project, and we might be able to consider only trees on at most 5 or 6 leaves. We refer to [10, 3, 2, 1, 4] for some analytical results on the maximum likelihood calculations.

The constraints for the optimization problem are imposed by the molecular clock. That is, the probability of substitution from the root to any leaf is P . So the constraints in the above example are

$$\begin{aligned} P &= P_1(1 - P_2) + (1 - P_1)P_2 \\ &= P_3(1 - P_6) + (1 - P_3)P_6 \\ &= P_3P_4P_5 + P_3(1 - P_4)(1 - P_5) + (1 - P_3)P_4(1 - P_5) + (1 - P_3)(1 - P_4)P_5. \end{aligned}$$

We can verify that the constraints can be simplified by the substitution $P_i = (1 - X_i)/2$. With this substitution the constraints become $X = X_1X_2 = X_3X_6 = X_3X_4X_5$, which can therefore be linearized by taking logarithms.

REFERENCES

- [1] B. Chor, M. Hendy, and S. Snir. Maximum likelihood jukes-cantor triplets: Analytic solutions. *Molecular Biology and Evolution*, 23(3):626–632, March 2006.
- [2] B. Chor, A. Khetan, and S. Snir. Maximum likelihood on molecular clock comb: Analytic solutions. *Journal of Computational Biology*, 13(3):819–837, April 2006.
- [3] Benny Chor and Sagi Snir. Analytic solutions of maximum likelihood on forks of four taxa. *Mathematical Biosciences*, 208(2):347 – 358, 2007.
- [4] Benny Chor and Sahi Snir. Molecular clock fork phylogenies: Closed form analytic maximum likelihood solutions. *Systematic Biology*, 53:963–967(5), December 2004.
- [5] J. Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

- [6] W. M. Fitch. Toward defining the course of evolution: minimum changes for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.
- [7] T. Jukes and C. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. New York Academic Press, 1969.
- [8] J. Neyman. Molecular studies of evolution: A source of novel statistical problems. In *Statistical Decision Theory and Related Topics*, pages 1–27. New York Academic Press, 1971.
- [9] Charles Semple and Mike Steel. *Phylogenetics*, volume 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2003.
- [10] Mike A. Steel. The maximum likelihood point for a phylogenetic tree is not unique. *Systematic Zoology*, 43:560–564, 1994.
- [11] C. Tuffley and M. A. Steel. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. of Math. Biol.*, 59:581–607, 1997.
- [12] Ziheng Yang. *Computational Molecular Evolution*. Oxford University Press, 2006.

E-mail address, Raphael Hauser: `hauser@comlab.ox.ac.uk`

OXFORD UNIVERSITY COMPUTING LABORATORY, PARKS ROAD, OXFORD OX1 3QD, UNITED KINGDOM.

E-mail address, Bhalchandra D. Thatte: `thatte@stats.ox.ac.uk`

DEPARTMENT OF STATISTICS, UNIVERSITY OF OXFORD, 1 SOUTH PARKS ROAD, OXFORD OX1 3TG, UNITED KINGDOM.