Bulletin of Mathematical Biology (2020) 82:10
https://doi.org/10.1007/s11538-019-00689-8

Society for Mathematical Biology

ORIGINAL ARTICLE

Most Parsimonious Likelihood Exhibits Multiple Optima for Compatible Characters

Julia Matsieva[1] · **Katherine St. John**[2,3]

Received: 15 November 2018 / Accepted: 12 December 2019 / Published online: 14 January 2020

Abstract
Maximum likelihood estimators are a popular method for scoring phylogenetic trees to best explain the evolutionary histories of biomolecular sequences. In 1994, Steel showed that, given an incompatible set of binary characters and a fixed tree topology, there exist multiple sets of branch lengths that are optima of the maximum average likelihood estimator. Since parsimony techniques—another popular method of scoring evolutionary trees—tend to exhibit favorable behavior on data compatible with the tree, Steel asked if the same is true for likelihood estimators, or if multiple optima can occur for compatible sequences. We show that, despite exhibiting behavior similar to parsimony, multiple local optima can occur for compatible characters for the most parsimonious likelihood estimator. We caution that thorough understanding of likelihood criteria is necessary before they are used to analyze biological data.

Keywords Models of evolution · Phylogenetic trees · Maximum likelihood estimators · Maximum parsimony criteria

1 Introduction

A canonical question in biology is to find the optimal evolutionary tree, or phylogeny, that explains the traits, or characters, of a set of species that are observed in the present moment. These phylogenies provide the basis for future study, ranging from understanding the spread of disease (Janies et al. 2011) to modeling co-evolution of

✉ Julia Matsieva
jmatsieva@ucdavis.edu

Katherine St. John
katherine.stjohn@hunter.cuny.edu

1 Department of Computer Science, University of California, Davis, Davis, CA 95616-8562, USA

2 Department of Computer Science, Hunter College, City University of New York, New York, USA

3 Invertebrate Zoology, American Museum of Natural History, New York, NY, USA

species (Charleston and Perkins 2003), as well as being studied for their own right as evolutionary histories (Bininda-Emonds et al. 2002). The specific order and path that species take to develop their diverse characteristics are not easy to establish by observing these organisms in the present day (Hillis et al. 1996).

There are several popular criteria for determining optimality of a phylogenetic tree (which we will refer to as tree) with respect to a sequence of characters. These optimality criteria give an objective way to evaluate possible evolutionary histories, providing a crucial tool in computational searches for the best tree. Parsimony approaches tend to focus on the combinatorial properties of the problem, such as the shape (or graph-theoretic 'topology') of the tree. More specifically, given a set of characters S and a tree T, the maximum parsimony criterion seeks the tree shape that has the minimum number of character state changes across its edges. In contrast, maximum likelihood methods view the amount of change across each branch, as well as the tree shape, as parameters to be optimized. For every sequence of observed characters S and a tree T with continuous branch lengths $\mathbf{y}$, the maximum likelihood criterion assigns a score to every tree T, representing the likelihood (i.e., $P(S|(T, \mathbf{y}))$) that the sequence of observed characters S is generated by T under a model of evolution. While the parsimony approach seems simpler, both problems are computationally hard (Foulds and Graham 1982; Roch 2006).

At first glance, these differences in the optimality criteria may look like interchangeable technical details; however, the choice of the optimality criterion (and associated parameters) can greatly affect *which* tree is chosen as best. Furthermore, these scores are computed repeatedly during searches for the best tree topology, so improving these computations will greatly speed up the search for the optimal tree. The situation is further complicated by the use of estimation or heuristics, due to the computational complexity of computing the exact answer. Thus, better understanding of these criteria can improve computationally expensive searches for the best tree.

There are several variants of maximum likelihood that are defined differently and thus yield different results. In particular, the states assigned to the internal nodes of the tree are supplementary ("nuisance") parameters that can be handled in different ways. One approach is to take the average of all possible assignments to the internal nodes as the maximum likelihood estimator. Following Barry and Hartigan (1987) and Steel and Penny (2000), we call this implementation of the criterion maximum average likelihood, $M_{\mathrm{av}}L$. Another variant, most parsimonious likelihood, denoted $M_{\mathrm{P}}L$, instead takes the maximum score across all internal state assignments (see Sect. 2 for definitions).

Under the maximum parsimony criterion, there is a unique score for each tree shape, since only the shape of the tree, and not the branch lengths, are considered. Furthermore, for the traditional maximum parsimony criterion, restricting to compatible characters reduces the complexity of the problem of finding the optimal tree from computationally hard to linear time (Gusfield 1991). It was suggested that the same is true for maximum likelihood: that each tree shape has a single local optimum of branch lengths for each character sequence (Fukami and Tateno 1989). Steel (1994) showed that, for a fixed tree, there exist character sequences that yield multiple local optima for $M_{\mathrm{av}}L$ for that tree. That is, given a tree topology, there are multiple ways to assign branch lengths that give a locally optimal $M_{\mathrm{av}}L$ value. His elegant construction used

sequences that were not compatible with the fixed tree. Chor et al. (2000) showed that, for the tree on four leaves, there are character sequences which have one-dimensional *ridges* of optima. More specifically, the branch lengths could be written as a function parameterized by a single variable such that any choice of this variable gives rise to the maximum $M_{\text{av}}L$ value. The sequences demonstrated there were not compatible with the tree topology used in Chor et al. (2000), leading to the still-open question: given *compatible* sequences for a fixed tree, is there a single local optimum for $M_{\text{av}}L$ (Steel 2011)?

This deceptively simple question has significant implications in the computational search for the optimal tree under a maximum likelihood criterion. If there are multiple local optima for data that are compatible with a given tree, then there exist sets of sequences with enough local optima to distort the results of the common heuristic search techniques. On the other hand, if we can show that there is a single set of optimal branch lengths for the simplest situation in which the characters agree with the tree topology, then simple search techniques guaranteed to find the optimum can be used as building blocks for more complex situations. For example, it may be possible to evaluate phylogenetic data in subgroups of compatible sequences and compose the results into a final answer.

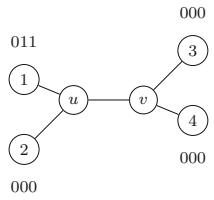
We show that, even for sequences compatible with a tree, there may exist multiple local optima for the branch lengths of that tree for $M_{\text{P}}L$. This surprising result for $M_{\text{P}}L$ suggests that similar behavior might be possible for the $M_{\text{av}}L$ on compatible sequence data, despite being much "smoother" when viewed as a function. We show our results by characterizing the interplay of the individual character functions under different choices of internal node state assignments, in that each assignment yields a smooth function with a single optimum in the interior of the space. We further examine constant characters (those which assign the same value to all the leaves of a tree) and their effect on local optima for most parsimonious likelihood. As the number of constant characters increases, the number of optima decreases and their locations shift toward $\mathbf{1}$. This complements the results of Tuffley and Steel (1997) that the addition of constant characters changes the optimum for maximum average likelihood.

2 Background

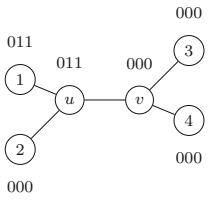
This section includes the definitions and notations that are used for the variants of likelihood examined here [we follow those from Semple and Steel (2003) and Steel and Penny (2000) whenever possible]. An *X-tree* $\mathcal{T}$ is an ordered pair (T, ϕ) where T is a tree with a vertex set V and $\phi : X \to V$ is a map with the property that for each $v \in V$ of degree at most two, $v \in \phi(X)$. A *phylogenetic (X-) tree* $\mathcal{T}$ is an X-tree (T, ϕ) with the property that ϕ is a bijection from X into the leaves of T. We write $V(T)$ to denote the vertex set of T and $E(T)$ to denote the set of *edges* or branches of T. The vertex set $V(T)$ is partitioned into leaves and internal nodes, which we denote $\mathcal{L}(T)$ and $\mathcal{I}(T)$, respectively. Given a leaf v of T, we will write x_v to mean the unique taxon in X that labels v. A *two-state* or *binary character* χ for X is a function $\chi : X \to \{0, 1\}$. For a single character χ, a χ*-labeling* of T is a function $\ell : \mathcal{I}(T) \to \{0, 1\}$ that assigns binary states to the internal labels of $\mathcal{T}$. We say that a

	χ_0	χ_1	χ_2	χ_3	χ_4	χ_{uv}
1	**0**	**1**	**0**	**0**	**0**	**0**
2	**0**	**0**	**1**	**0**	**0**	**0**
3	**0**	**0**	**0**	**1**	**0**	**1**
4	**0**	**0**	**0**	**0**	**1**	**1**

(a) The compatible characters (up to equivalence) on the 4-leaf tree.



(b) The 4-leaf unrooted tree with leaves labeled by the character sequence $S = \chi_0\chi_1\chi_1$.



(c) One of the 8 possible state assignments that extend the characters in (b) to the internal nodes u and v.

Fig. 1 An example of a character sequence on the unrooted tree with four leaves

state assignment $I = \{\ell_1, \ldots, \ell_k\}$ *extends* a set of characters $S = \{\chi_1, \ldots, \chi_k\}$ if ℓ_i is assigns internal states to χ_i. We use χ^ℓ to denote the character χ extended by the internal state assignment ℓ.

We will focus on the 4-leaf unrooted tree, with the leaves $\mathcal{L}(T) = \{1, 2, 3, 4\}$ and internal nodes $\mathcal{I}(T) = \{u, v\}$ shown in Fig. 1. Since there are only two internal nodes, we will denote an extended character function χ^ℓ with $\ell = (b_u, b_v) \in \{0, 1\}^2$ to indicate the extension of character χ to assign $u \mapsto b_u$, $v \mapsto b_v$. In general, there are $2^{|\mathcal{I}(T)|}$ possible ways to extend the two-state character to the internal nodes of T. If a character *sequence* S contains k characters, then there are $2^{k|\mathcal{I}(T)|}$ possible internal state assignments for the sequence.

A *split* on a set of leaves $\mathcal{L}(T)$ is a bipartition of the leaf set into two non-empty sets A and B denoted $A|B$ where $A \subseteq \mathcal{L}(T)$ and $B = \mathcal{L}(T)\backslash A$. Given a tree T and an edge e, the removal of e from T results in two connected components. This bipartition of the leaves is a split. Two splits $A|B$ and $A'|B'$ are *compatible* if at least one of the following intersections is empty: $A \cap B$, $A \cap B'$, $A' \cap B$, or $A' \cap B'$. Each binary character induces a split on the leaves of the tree. For binary characters, this is a bipartition of the leaves—those that have one state of the character versus the other. We will say that the character is *compatible* with a tree if its induced split is compatible with all splits induced by the tree. This technical definition can be interpreted informally as follows: "a collection of characters is compatible if they could all have evolved on some tree without any reverse or convergent transitions" (Semple and Steel 2003). Although a split is defined to be a non-trivial bipartition, we allow inputs containing the *constant* character, which assigns 0 to all $v \in \mathcal{L}(T)$, inducing the trivial split on the leaves of T.

There are many different models of evolution that can be used to evaluate the likelihood of a tree that explains the data (Semple and Steel 2003). We focus on one of the simplest: the 2-state symmetric Markov model of evolution. In this model, for each branch i, we define t_i to be the length of the branch, scaled by a fixed rate of evolution. That is, t_i is proportional to the expected number of state transitions along edge i. Its value is a function of the branch length and the rate of evolution along i, and the range of t_i is $[0, \infty)$. For the branch length t_i, we define the probability $P(t_i)$

$$L(\chi_0^{00}, \mathbf{y}) = \left(\frac{1}{2}\right)^5 (1+y_1)(1+y_2)(1+y_3)(1+y_4)(1+y_{uv})$$
$$L(\chi_1^{00}, \mathbf{y}) = \left(\frac{1}{2}\right)^5 (1-y_1)(1+y_2)(1+y_3)(1+y_4)(1+y_{uv})$$
$$L(\chi_2^{00}, \mathbf{y}) = \left(\frac{1}{2}\right)^5 (1+y_1)(1-y_2)(1+y_3)(1+y_4)(1+y_{uv})$$
$$L(\chi_3^{00}, \mathbf{y}) = \left(\frac{1}{2}\right)^5 (1+y_1)(1+y_2)(1-y_3)(1+y_4)(1+y_{uv})$$
$$L(\chi_4^{00}, \mathbf{y}) = \left(\frac{1}{2}\right)^5 (1+y_1)(1+y_2)(1+y_3)(1-y_4)(1+y_{uv})$$
$$L(\chi_{uv}^{00}, \mathbf{y}) = \left(\frac{1}{2}\right)^5 (1+y_1)(1+y_2)(1-y_3)(1-y_4)(1+y_{uv}).$$

Fig. 2 The labeled likelihood functions for the compatible characters for the tree in Fig. 1 extended with the 00 state assignment

that the character stays the same across the edge and probability $Q(t_i)$ that it changes across the edge as:

$$P(t_i) = \frac{1}{2}(1+e^{-2t_i}), \qquad Q(t_i) = \frac{1}{2}(1-e^{-2t_i}). \tag{1}$$

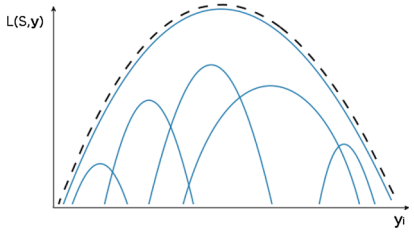
These transitions are symmetric: the probability of that a character changes from 0 to 1 across a branch equals the probability of changing from 1 to 0. We change variables to simplify the notation by setting $y_i = e^{-2t_i}$. The range of each y_i is [0, 1]. We use $\mathbf{y}$ to refer to the vector $(y_i)_{i \in E(T)}$. Then, the likelihood that an observed character was derived from T is given by:

$$L(\chi^\ell, \mathbf{y}) = \left(\frac{1}{2}\right)^{|E(T)|} \prod_{i \in E(T)} (1 + (-1)^{\delta_i} y_i) \tag{2}$$

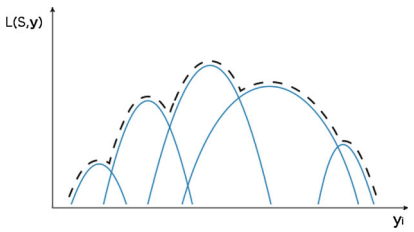
where $\delta_i = 1$ if $\chi^\ell(u) \neq \chi^\ell(v)$ and 0 otherwise for $i = (u, v) \in E(T)$, modeling the notion that state transitions across edges are unlikely under this model of evolution.

For the unrooted tree on four leaves given in Fig. 1, the vector $\mathbf{y} = (y_1, y_2, y_3, y_4, y_{uv})$ corresponds to the edge lengths after the change of variables. Figure 1a lists the six binary characters, up to equivalence, that are compatible with the tree. Let χ_0 be the character that assigns the same state to all of the leaves, i.e., the constant character. We note that the constant character fits the definition above of binary characters (i.e., it is a function from the leaves of T to {0,1}), though it is often not included in the character set since it does not have two distinct states. We name the remaining characters by the leaf on which they differ (for χ_1, χ_2, χ_3, and χ_4). The remaining compatible character that has different labels for the leaves on opposite sides of the edge (u, v) we call χ_{uv}. The labeled character likelihood functions for the compatible characters of the tree in Fig. 1 with the internal nodes u, v assigned states (0, 0) are shown in Fig. 2.

When we fix an extension of the set of characters to all the internal nodes, we can view this as inducing a single labeled likelihood function (see Fig. 3):



(a) The case in which there is a single state assignment I for all $\chi \in S$ such that $L_{MP}(S, \mathbf{y}) = L(S, I, \mathbf{y})$ for all $\mathbf{y}$. This would imply that L_{MP} has a single optimum on $[0, 1]^{|E(T)|}$.



(b) The case in which there is no single individual likelihood function that "covers" all of the other optima for all $\mathbf{y}$, implying that L_{MP} exhibits multiple optima, even though each $L(S, I, \mathbf{y})$ is shown to have a single optimum in Lemma 6.

Fig. 3 A drawing showing several individual labeled likelihood functions, projected down to a single component, and their interactions. The dashed line denotes the value of L_{MP}, the maximum over all of the individual functions at each point

Definition 1 Given a tree T and a sequence of characters, $S = \{\chi_1, \ldots, \chi_k\}$, on the leaves of the tree, and an extension $I = \{\ell_1, \ldots, \ell_k\}$ of the character sequence to the internal nodes, we define a *labeled likelihood function* for I to be the function

$$L(S, I, \mathbf{y}) = \prod_{\ell \in I} L(\chi^{\ell}, \mathbf{y}), \tag{3}$$

where $\mathbf{y}$ with components $y_i = e^{-2t_i}$ and t_i the length of the branch i of the tree and $t_i \geq 0$.

With this notation in mind, we can now formally define the variations of likelihood: maximum average likelihood ($M_{\text{av}}L$) and most parsimonious likelihood ($M_{\text{P}}L$). To compute the maximum *average* likelihood, we average over all of the possible internal states. Given a tree T and a sequence of characters $S = \{\chi_1, \ldots, \chi_k\}$, the average likelihood function can be written as

$$L_{\text{av}}(S, \mathbf{y}) = \left(\frac{1}{2}\right)^k \prod_{\chi \in S} \sum_{\ell \text{ extends } \chi} L(\chi^{\ell}, \mathbf{y}), \tag{4}$$

where S is the sequence (multiset) of characters and each ℓ extends the character to assign states to the internal nodes of T. For example, the data set $S = \{\chi_0, \chi_1, \chi_1\}$ from Fig. 1 has average likelihood function:

$$L_{\text{av}}(S, \mathbf{y}) = \left(\frac{1}{2}\right)^3 \left(L(\chi_0^{00}, \mathbf{y}) + L(\chi_0^{01}, \mathbf{y}) + L(\chi_0^{10}, \mathbf{y}) + L(\chi_0^{11}, \mathbf{y})\right) \cdot \left(L(\chi_1^{00}, \mathbf{y}) + L(\chi_1^{01}, \mathbf{y}) + L(\chi_1^{10}, \mathbf{y}) + L(\chi_1^{11}, \mathbf{y})\right)^2.$$

Although we refer to S as a sequence of characters, all of our computations on S are commutative, so we use multiset notation throughout the manuscript. [1]

Barry and Hartigan (1987) suggest another approach, *most parsimonious likelihood*, where, instead of averaging the values over all possible state assignments at each point $\mathbf{y}$, the state assignment that gives the best score is chosen. This approach echoes that of maximum parsimony, in that the best possible value under all state assignments is chosen. Barry and Hartigan write "we call this technique most parsimonious because the values of internal nodes are usually assigned to agree as much as possible with neighboring nodes" (Barry and Hartigan 1987), and suggest that they expect this technique to be easier to apply than maximum average likelihood. We find that this does not seem to be the case (see Sect. 6). For a fixed underlying tree, T, we interpret the description of most parsimonious likelihood as the function:

$$L_{\text{MP}}(S, \mathbf{y}) = \max_{I \text{ extends } S} L(S, I, \mathbf{y}) = \max_{I \text{ extends } S} \prod_{\ell \in I} L(S^{\ell}, \mathbf{y}) \tag{5}$$

where S is the observed sequence of characters on the leaves of the tree. We note that there is not a bijection between the character sequences and the corresponding labeled likelihood functions, since different internal state assignments can yield the same labeled likelihood function. (We detail when labeled likelihood functions are the same in Sect. 4.)

While maximum average likelihood averages the $L(S^{\ell}, \mathbf{y})$ values over all internal state assignments ℓ that extend S, most parsimonious likelihood chooses the best internal state assignments *for each character copy*. Thus, it is consistent with Eq. 5 for two *copies* of the same *character* to receive different state assignments for the internal nodes of the tree. As we show in Sect. 4, each of these individual labeled functions has at most one local optimum on $[0, 1]^{|E(T)|}$. If its local optimum is not "covered"—that is, exceeded in value—by the value of any other labeled function at that point, we call it a "lump".

Definition 2 Let T be a tree and S be a sequence of characters on the leaves of T. If $\mathbf{y}^*$ is a local optimum of labeled likelihood function $L(S, I, \mathbf{y})$ on $[0, 1]^{|E(T)|}$ for some state assignment I of S, then we call $\mathbf{y}^*$ a *lump* of $L_{\text{MP}}(S, \mathbf{y})$ if $L(S, I', \mathbf{y}^*) \leq L(S, I, \mathbf{y}^*)$ for all other internal state assignment $I' \neq I$, of S.

We show that each labeled likelihood function has at most one local maximum and characterize the lumps in Sect. 4.

3 Multiple Optima

We include a running example to show the difference between the likelihood variants. We note that unlike the maximum parsimony criterion, $M_{\text{P}}L$ takes the best internal state

[1] Our goal with the use of the term "sequence" is to call back to the scientific process of obtaining observations in a continuous setting. We expect likelihood estimates to change as more samples are discovered and characterized. We do not use the term to refer to unaligned DNA sequences for a single taxon.

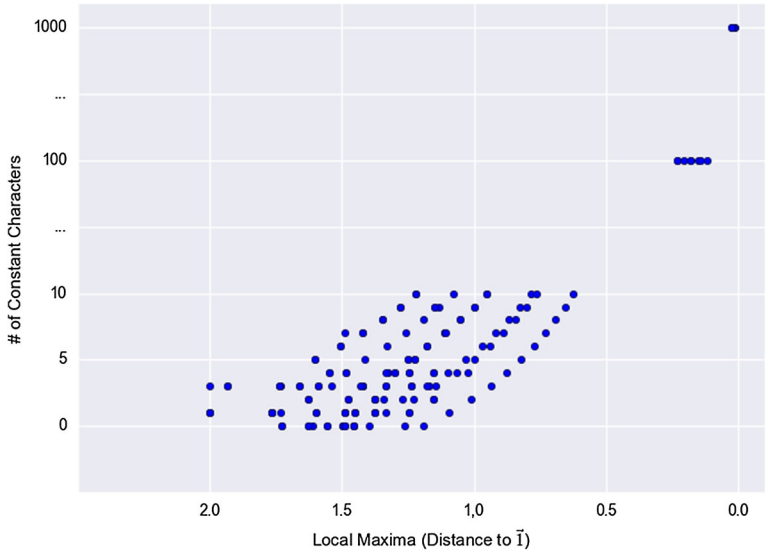


Fig. 4 The local maxima of the running example in Fact 4 ordered by their distance to the vector **1** and increasing number of additional constant characters (along the y-axis)

assignments for each set of branch lengths, not a single best internal state assignment for each tree shape. More specifically, given a multiset S of characters, the best internal state assignment for each character is chosen to maximize $L_{\mathrm{MP}}(\mathbf{y})$ for *each* $\mathbf{y}$. We show:

Theorem 3 *For a phylogenetic tree T, there exists a sequence of compatible characters S such that most parsimonious likelihood $L_{\mathrm{MP}}(S, \mathbf{y})$ has multiple optima.*

We demonstrate one such sequence in Fact 4; however, the character sequences that exhibit this behavior are not difficult to locate. Most of the datasets we explore (using the code described in Sect. 5) exhibit this behavior. In fact, according to Lemma 6, it is possible to engineer datasets whose most parsimonious state assignment has a lump on the interior of $(0, 1)^{|E(T)|}$. Finally, we can generalize Fact 4 to larger trees, since every likelihood function, as given by Eq. 2, can be rewritten as

$$L(\chi^{\ell}, \mathbf{y}) = \left(\frac{1}{2}\right)^{|E(T)|} \prod_{i \in E_1} (1 + (-1)^{\delta_i} y_i) \prod_{i \in E_2} (1 + (-1)^{\delta_i} y_i)$$

for any partitions E_1, E_2 of $E(T)$. Therefore, any extension to a larger tree that preserves the labels on both sides of the five edges of the 4-leaf tree from Fig. 1 will retain all of the lumps found in Table 1.

The example below, with constant characters, three copies of the characters, χ_1 and χ_2 and five copies for the for the character, χ_{uv}, corresponding to the middle edge, has multiple optima for their compatible tree (summarized in Table 1).

Fact 4 *The most parsimonious likelihood function $L_{\mathrm{MP}}(S, \mathbf{y})$ on dataset $S = \{3\chi_1, 3\chi_2, 5\chi_{uv}\}$ has 17 distinct lumps in $[0, 1]^5$, four of which occur strictly in $(0, 1)^5$.*

Table 1 The 17 individual lumps of $L_{\text{MP}}(S, \mathbf{y})$ on dataset $S = \{3\chi_1, 3\chi_2, 5\chi_{uv}\}$ with their scaled L_{MP} and L_{av} scores

	$\mathbf{y}$	$\log 2^{5.11} L_{\text{MP}}(S, \mathbf{y})$	$\log 2^{5.11} L_{\text{av}}(S, \mathbf{y})$	$\nabla \log 2^{5.11} L_{\text{av}}(S, \mathbf{y})$
1	$\left(\frac{5}{11}, \frac{5}{11}, 1, 1, \frac{1}{11}\right)$	17.65	14.44	$(-1.82, -1.82, 5.5, 5.5, -4.04)$
2	$\left(\frac{5}{11}, \frac{5}{11}, 1, 1, 0\right)$	17.61	14.8	$(-1.55, -1.55, 5.5, 5.5, -3.77)$
3	$\left(\frac{1}{11}, \frac{1}{11}, 1, 1, \frac{5}{11}\right)$	16.52	14.81	$(-2.51, -2.51, 5.5, 5.5, -0.98)$
4	$\left(\frac{1}{11}, 0, 1, 1, \frac{5}{11}\right)$	16.47	15.03	$(-2.48, -2.33, 5.5, 5.5, -0.5)$
5	$\left(0, \frac{1}{11}, 1, 1, \frac{5}{11}\right)$	16.47	15.03	$(-2.33, -2.48, 5.5, 5.5, -0.5)$
6	$\left(\frac{5}{11}, \frac{5}{11}, \frac{1}{11}, \frac{1}{11}, 1\right)$	10.07	6.54	$(-2.12, -2.12, -3.26, -3.26, -0.79)$
7	$\left(\frac{1}{11}, \frac{1}{11}, \frac{5}{11}, \frac{5}{11}, 1\right)$	10.07	8.95	$(-4.35, -4.35, 3.58, 3.58, -0.79)$
8	$\left(\frac{1}{11}, 0, \frac{5}{11}, \frac{5}{11}, 1\right)$	10.03	9.32	$(-4.36, -3.79, 3.86, 3.86, -0.4)$
9	$\left(0, \frac{1}{11}, \frac{5}{11}, \frac{5}{11}, 1\right)$	10.03	9.32	$(-3.79, -4.36, 3.86, 3.86, -0.4)$
10	$\left(\frac{5}{11}, \frac{5}{11}, 0, 0, 1\right)$	9.98	7.18	$(-1.55, -1.55, -3.77, -3.77, 0.0)$
11	$\left(\frac{3}{11}, \frac{3}{11}, \frac{3}{11}, \frac{3}{11}, 1\right)$	9.28	6.82	$(-3.24, -3.24, 0.05, 0.05, -1.74)$
12	$\left(\frac{9}{11}, \frac{1}{11}, \frac{1}{11}, \frac{1}{11}, \frac{7}{11}\right)$	6.82	7.07	$(-0.78, -1.65, -2.18, -2.18, -0.92)$
13	$\left(\frac{1}{11}, \frac{9}{11}, \frac{1}{11}, \frac{1}{11}, \frac{7}{11}\right)$	6.82	7.07	$(-1.65, -0.78, -2.18, -2.18, -0.92)$
14	$\left(\frac{9}{11}, \frac{1}{11}, 0, 0, \frac{7}{11}\right)$	6.73	7.52	$(-0.17, -1.5, -2.69, -2.69, 0.0)$
15	$\left(\frac{1}{11}, \frac{9}{11}, 0, 0, \frac{7}{11}\right)$	6.73	7.52	$(-1.5, -0.17, -2.69, -2.69, 0.0)$
16	$\left(\frac{9}{11}, \frac{1}{11}, \frac{3}{11}, \frac{3}{11}, \frac{5}{11}\right)$	6.33	7.12	$(-1.7, -1.81, 0.55, 0.55, -3.12)$
17	$\left(\frac{1}{11}, \frac{9}{11}, \frac{3}{11}, \frac{3}{11}, \frac{5}{11}\right)$	6.33	7.12	$(-1.81, -1.7, 0.55, 0.55, -3.12)$

Note that entries 12, 13, 16, 17 occur strictly on the interior of the domain of L_{MP}. The third column gives the scaled average likelihood value at each lump. The last column shows the maxima of L_{MP} are not maxima of $L_{\text{av}}(S, \mathbf{y})$, which has nonzero gradient

In Fig. 4, the lowest row ($y = 0$) values of the right figure show the 17 local maxima for our running example. Since it is difficult to visualize 5-dimensional space, we have mapped the values to a line, ordered by their distance to the point $\mathbf{1} = [1, 1, 1, 1, 1]$. The y-axis is indexed by the number of additional constant characters added. Adding constant characters reduces the number of local maxima and moves them toward $\mathbf{1}$. (We state this formally in Corollary 7.)

Lemma 5 *Let $\chi_0 : X \to \{0, 1\}$ be the constant character, defined $\chi_0(x) = 0, \forall x \in X$, and let S be a sequence of characters that contains $k_0 > 0$ copies of the constant character, and for every $\mathbf{y} \in [0, 1]^{|E(T)|}$, let I be a state assignment that extends S such that:*

$$L_{\text{MP}}(S, \mathbf{y}) = L(S, I, \mathbf{y}).$$

Then, I has at least k_0 copies of the zero state assignment defined $\ell_0(v) = 0$ for all $v \in \mathcal{I}(T)$.

Proof From the definitions given in Sect. 2, we can write

$$L(\chi_0^{\ell_0}, \mathbf{y}) = \left(\frac{1}{2}\right)^{|E(T)|} \prod_{i \in E(T)} (1 + y_i).$$

Since log preserves order, we can also write

$$\log L(\chi_0^{\ell_0}, \mathbf{y}) = |E(T)| \log \frac{1}{2} + \sum_{i \in E(T)} \log(1 + y_i).$$

We can see from Fig. 1 that if we were to change any one of the internal state assignments to 1, we would create a state change on three edges of T. If we were to set *all* internal state assignments to 1, we would still have disagreements on all of the edges incident to the leaves of T. Therefore, ℓ_0 is the only state assignment that does not give rise to any disagreements on T. Therefore, since $\mathbf{y} \in [0, 1]^{|E(T)|}$, then $\log(1 + y_i) \geq \log(1 - y_i)$, and since log preserves order,

$$L(\chi_0^{\ell_0}, \mathbf{y}) \geq L(\chi_0^{\ell}, \mathbf{y}) \tag{6}$$

for all $\ell \neq \ell_0$, $\mathbf{y} \in [0, 1]^{|E(T)|}$. Then, it follows that, since ℓ_0 is the best state assignment for a single constant character, then

$$L(\{k_0\chi_0\}, \{k_0\ell_0\}, \mathbf{y}) = \prod_{k=1}^{k_0} L(\chi_0^{\ell_0}, \mathbf{y}) \geq \prod_{k=1}^{k_0} L(\chi_0^{\ell \neq \ell_0}, \mathbf{y}), \tag{7}$$

where we use multiset notation for k_0 copies of the constant character, and k_0 copies of the all-zero state assignment for those characters. Then, if S contains k_0 copies of the constant character, $L_{\text{MP}}(S, \mathbf{y})$ can always be rewritten as

$$L_{\text{MP}}(S, \mathbf{y}) = \max_{I \text{ extends } S} \left(\left(\prod_{\ell \text{ extends } \chi_0} L(\chi_0^{\ell}, \mathbf{y}) \right) \cdot L(S \backslash \{k_0\chi_0\}, I, \mathbf{y}) \right)$$

and then, by Eq. 7,

$$L_{\text{MP}}(S, \mathbf{y}) = \prod_{k=1}^{k_0} L(\chi_0^{\ell_0}, \mathbf{y}) \max_{I \text{ extends } S \backslash \{k_0\chi_0\}} L(S \backslash \{k_0\chi_0\}, I, \mathbf{y}).$$

□

Therefore, for constant characters, if suffices to only check the single internal state assignment when computing the most parsimonious likelihood.

4 Characterizing Labeled Likelihood Functions

In this section, we analyze the individual labeled likelihood functions that are the building blocks of the overall most parsimonious likelihood function L_{MP}. While this paper focuses on compatible characters, we note that we do not assume compatible character sets for the results in this section.

In order to locate the extrema of an MP likelihood function given in Eq. 5, we compute the common roots of the system of its first partial derivatives. We note that a labeled likelihood function $L(S, I, \mathbf{y})$ simplifies to a product of monomials in each of the tree edge variables. This means that the likelihood function for character set S and internal state assignment I that extends S is given by

$$L(S, I, \mathbf{y}) = \frac{1}{2}^{|E(T)||S|} \prod_{i \in E(T)} (1 + y_i)^{p_i} (1 - y_i)^{n_i}, \tag{8}$$

where $p_i + n_i = |S|$ for all $i \in E(T)$. These values depend on internal state assignment I and are obtained by resolving the δ_i functions in Eq. 2, which add one to p_i if the two endpoints of edge i agree and add one to n_i otherwise. As such, p_i and n_i are uniquely determined for each S and I. Further, p_i is the total number of state agreements over edge $i \in E(T)$ and n_i is the total number of state disagreements. We note that $\sum_{i \in E(T)} n_i$ is the parsimony score for T with leaves labeled by S and internal nodes by I. We let $\mathbf{p}$ be the vector of the values of p_i and similarly $\mathbf{n}$ be the vector of the values of n_i.

Lemma 6 *Every labeled likelihood function $L(S, I, \mathbf{y})$ has a global maximum at $\mathbf{y}^*$ with coordinates*

$$y_i^* = \frac{p_i - n_i}{p_i + n_i}$$

where p_i and n_i satisfy Eq. 2 for I and for $i \in E(T)$ for $p_i, n_i \neq 0$.

Before we prove this claim, we point out that these optimal branch lengths are intuitive and similar to the ranking obtained from the traditional maximum parsimony criterion. Since p_i gives the number of state agreements over an edge and n_i gives the number of disagreements, the value above is reminiscent of taking a consensus vote along each edge. We expect edges with more state agreements to be considered more likely under the evolutionary assumption that state transitions are rare.

Proof of Lemma 6 Let I be an internal state assignment with associated p_i and n_i for $i \in E(T)$ and let $|S| = k$. We examine the following cases: (1) where $p_i, n_i \neq 0$, (2) where $p_i = 0$, and (3) where $n_i = 0$.

In Case (1), the first partial derivative of $L(S, I, \mathbf{y})$:

$$\begin{aligned}\frac{\partial}{\partial y_i} L(S, I, \mathbf{y}) &= \Big(p_i(1+y_i)^{p_i-1}(1-y_i)^{n_i} - n_i(1+y_i)^{p_i}(1-y_i)^{n_i-1}\Big)\frac{1}{2^k} \\ &\quad \prod_{j\neq i\in E(T)} (1+y_j)^{p_j}(1-y_j)^{n_j} \\ &= \Big((p_i-n_i)-(p_i+n_i)y_i\Big)\frac{L(S, I, \mathbf{y})}{(1+y_i)(1-y_i)}\end{aligned} \tag{9}$$

which means that $\frac{\partial}{\partial y_i} L(S, I, \mathbf{y}) = 0$ lies on the hyperplane defined by:

$$y_i = \frac{p_i - n_i}{p_i + n_i}$$

regardless of the other values of $y_{j\neq i}$. In order to characterize this critical point, we inspect the value of the Hessian matrix of second partial derivatives evaluated at this point (Stewart 2005). This is a generalization of the second derivative test to the multivariable case; if the Hessian can be shown to be negative definite at $\mathbf{y}^*$, then $\mathbf{y}^*$ is a maximum of $L(S, I, \mathbf{y})$. Let $a_i = p_i - n_i$ and $b_i = p_i + n_i$; then let $y_i^* = \frac{a_i}{b_i}$. We can rewrite the first partial derivative, using a_i and b_i as:

$$\frac{\partial}{\partial y_i} L(S, I, \mathbf{y}) = \frac{a_i - b_i y_i}{1 - y_i^2} L(S, I, \mathbf{y}) \tag{10}$$

Using a_i and b_i to simplify notation, the second partial derivative is:

$$\frac{\partial^2}{\partial y_i^2} L(S, I, \mathbf{y}) = \left(\frac{-b_i}{1-y_i^2} + \frac{2y_i(a_i - b_i y_i)}{(1-y_i^2)^2} + \frac{(a_i - b_i y_i)^2}{(1-y_i^2)^2}\right) L(S, I, \mathbf{y})$$

When $\mathbf{y} = \mathbf{y^*}$, where $y_i = \frac{a_i}{b_i}$, we have:

$$\frac{\partial^2}{\partial y_i^2} L(S, I, \mathbf{y})|_{\mathbf{y}=\mathbf{y}^*} = \frac{-b_i}{1 - (\frac{a_i}{b_i})^2} L(S, I, \mathbf{y}^*)$$

Therefore, the second partial derivative with respect to y_i is negative at y_i^*. Finally, we observe that the cross partial derivatives of $L(S, I, \mathbf{y})$ can be written generally as

$$\frac{\partial^2}{\partial y_i y_j} L(S, I, \mathbf{y}) = (a_i - b_i y_i) \frac{\partial}{\partial y_j}\left(\frac{L(S, I, \mathbf{y})}{1 - y_i^2}\right),$$

demonstrating that the cross partial derivative also has a root at $y_i^* = \frac{p_i - n_i}{p_i + n_i}$. Thus, we can summarize that in Case (1) when $p_i, n_i \neq 0$, the ith row of the Hessian matrix of second partial derivatives has a negative number on the diagonal and zero everywhere else, when evaluated at $y_i^* = \frac{a_i}{b_i} = \frac{p_i - n_i}{p_i + n_i}$.

Next, we examine Case (2) where $p_i = 0$. In this case, the partial derivative of $L(S, I, \mathbf{y})$ is

$$\frac{\partial}{\partial y_i} L(S, I, \mathbf{y}) = -n_i (1 - y_i)^{n_i - 1} \frac{L(S, I, \mathbf{y})}{(1 - y_i)^{n_i}} = -n_i \frac{L(S, I, \mathbf{y})}{(1 - y_i)},$$

which only has roots at $y_i = 1$, where $L(S, I, \mathbf{y})$ has roots. The roots still exist for the cross partial derivative, and the second partial derivative with respect to y_i, given below:

$$\frac{\partial^2}{\partial y_i^2} L(S, I, \mathbf{y}) = n_i (n_i - 1) \frac{L(S, I, \mathbf{y})}{(1 - y_i)^2}.$$

Since the ith partial derivative has no roots on the interior, then $L(S, I, \mathbf{y})$ has no critical point on the interior.

Finally, Case (3) is similar to Case (2). If $n_i = 0$, then

$$\frac{\partial}{\partial y_i} L(S, I, \mathbf{y}) = p_i (1 + y_i)^{p_i - 1} \frac{L(S, I, \mathbf{y})}{(1 + y_i)^{p_i}} = p_i \frac{L(S, I, \mathbf{y})}{(1 + y_i)},$$

which puts the root at $y_i = -1$, which is not in the domain of $L(S, I, \mathbf{y})$. That root is still there for all the different partial derivatives, so for the matching partial derivative, we get

$$\frac{\partial^2}{\partial y_i^2} L(S, I, \mathbf{y}) = p_i (p_i - 1) \frac{L(S, I, \mathbf{y})}{(1 + y_i)^2}.$$

Otherwise, there are no critical points on the interior, because the ith component of $\mathbf{y}$ lands on the boundary. □

Therefore, in the case where $p_i, n_i \neq 0$, for all $i \in E(T)$, the function $L(S, I, \mathbf{y})$ has a single maximum in the interior. We know that it is a maximum, because each row of the Hessian has a negative value on the diagonal and zero everywhere else; thus, it is negative definite at

$$y_i = \frac{p_i - n_i}{p_i + n_i}.$$

The occurrence of a maximum on the interior of L is entirely determined by the exponents on the monomials corresponding to the edges. We also know that the choice of the internal state assignment changes the values of p_i and n_i.

Since the MP likelihood function, given in Eq. 5, is defined as the maximum over all choices of internal state assignments, we can demonstrate sets of characters S such that L_{MP} has multiple local maxima on the interior. We note that the polynomials can have negative roots, but these are not maxima of L_{MP}, which is only defined on $\mathbf{y} \in [0, 1]^{|E(T)|}$ as given in Eq. 1.

We showed in Lemma 5 that the best state assignment for the internal nodes for the constant character is the all-zero assignment, which creates agreements across all edges. Therefore, a consequence of Lemma 5 is that adding k_0 constant characters to a given dataset increases p_i by k_0 for all $i \in E(T)$. This change shifts the location of $\mathbf{y}^*$, the local maximum of $L(S, I, \mathbf{y})$ for all I, regardless of the original values of S. In the limit, the local maximum of $L(S, I, \mathbf{y})$ shifts toward $\mathbf{y} = \mathbf{1}$ as more constant characters are added. As a corollary to Lemmas 5 and 6, we note that:

Corollary 7 *Let S be a sequence of characters and let $S_k = S \cup \{k\chi_0\}$ be S with an additional k constant characters. Then, let M_k be the set of lumps of $L_{\mathrm{MP}}(S_k, \mathbf{y})$ for a fixed tree T, then*

$$\lim_{k\to\infty} \sum_{\mathbf{y}^* \in M_k} |\mathbf{1} - \mathbf{y}^*| = 0.$$

This behavior is illustrated in Fig. 4. The optima move toward $\mathbf{1}$ as constant characters are added.

5 Implementation

As shown in Eq. 5, there are two components to computing most parsimonious likelihood. First, it is necessary to compute optima of the individual labeled likelihood functions $L(S, I, \mathbf{y})$; then, it is necessary to maximize $L(S, I, \mathbf{y})$ at every $\mathbf{y}$ over all state assignments I that extend the dataset S. For the first step, we leverage the result in Sect. 4 to locate the optima of the individual labeled functions $L(S, I, \mathbf{y})$. By Lemma 6, the location of the maximum $\mathbf{y}^*$ of $L(S, I, \mathbf{y})$ depends entirely on the resulting $\mathbf{p}, \mathbf{n}$ derived from the internal state assignment I. This requires no numerical root-finding since the result is derived already.

The difficulty of computing L_{MP} stems from the second step that requires finding the internal state assignment with the maximum L_{MP} value at each point. Even the small dataset in Fig. 1 has $|S| = 11$ character copies, which means there are a total of $2^{2\cdot 11} = 4096$ internal state assignments I that extend S on the tree in Fig. 1. We explore this space by starting at a *most parsimonious* state assignment I_{MP} that extends S and changing one assignment at a time in a breadth-first order.[2] For each subsequent labeling I, we check if the value of $L(S, I, \mathbf{y}^*)$ at its maximum is bounded above by $L(S, I', \mathbf{y}^*)$ for all state assignments I' that have been examined already. By definition, a most parsimonious state assignment I_{MP} has the *minimum* number of label disagreements for that dataset, that is, the lowest value of $\sum_{i\in E(T)} n_i$ corresponding to I_{MP}. Therefore, we expect it to have "competitive" values of L_{MP} for $y_i > 0$, allowing us to filter out the maxima of individual labeled functions that are *not* lumps of L_{MP} in an efficient order. Note that none of the $M_{\mathrm{P}}L$ extrema are optimal for $M_{\mathrm{av}}L$.

These techniques are implemented as a proof of concept in the mathematics software system SageMath (Stein et al. 2015) and supplemented with scripts in Python 2.7 (Foundation 2010). Despite several optimization strategies, such as leveraging the

[2] Note that I_{MP} yields the first lump given in Table 1.

result from Lemma 5 and choosing I_{MP} as the starting point, the exhaustive search performed by this methods exhibits poor runtime performance, as it takes around 5–10 min to process the small example given in this paper. However, we chose this trade-off to ensure that the method does not mis-identify any lumps. A more sophisticated treatment of the combinatorics involved in converting state assignments into $\mathbf{p}, \mathbf{n}$ would be a necessary step in developing more efficient algorithms for locating the optima for L_{MP}. We encourage researchers interested in experimenting with our code to contact us via email.

6 Conclusion and Future Work

Our analysis of most parsimonious likelihood illuminates surprising behavior. Even for character sequences that are compatible with the tree, multiple sets of branch lengths give local optima for the estimator. Symmetry in the character sequences is a natural way to have multiple optima. The number of lumps (including those on the boundary) can be quite high, bounded by the number of possible internal state assignments of the nodes. When given a single compatible character, the optima occur on the boundary of the tree lengths. We show that the addition of constant characters to the dataset moves the optima of the most parsimonious likelihood criterion in predictable ways, similar to the work of Tuffley and Steel for maximum average likelihood (Tuffley and Steel 1997).

While we characterize the behavior of most parsimonious likelihood for compatible characters, the original question of Steel (2011) of the behavior of maximum average likelihood for compatible characters is still open. The difficulties in characterizing the behavior of maximum average likelihood arise from the number of variables and the high degree of the polynomials involved; however, if one is interested in the value of L_{av} at a certain point, the computation is simple. We observe that most parsimonious likelihood has the opposite problem: the behavior of the individual labeled functions is known and their optima are simple to compute; however, we do not currently have a method for computing L_{MP} at a point that is not known to be an optimum of any labeled function. This leads to the interesting computational problem of efficiently finding and ranking the labeled likelihood functions with maxima that are closest to a given point, which is necessary to evaluate L_{MP} at that point.

Acknowledgements We would like to thank Dan Gusfield, Rob Gysel, Mike Steel, and Ward Wheeler for helpful comments and conversations. This work was partially supported by a grant from the Simons Foundation to KS.

References

Barry D, Hartigan J (1987) Statistical analysis of hominoid molecular evolution. Stat Sci 2:191–207

Bininda-Emonds ORP, Gittleman JL, Steel MA (2002) The (super) tree of life. Ann Rev Ecol Syst 33:265–89

Charleston MA, Perkins SL (2003) Lizards, malaria, and jungles in the caribbean. In: Page RD (ed) Tangled trees: phylogeny, cospeciation, and coevolution. University of Chicago Press, Chicago, pp 65–92

Chor B, Hendy MD, Holland BR, Penny D (2000) Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol Biol Evol 17(10):1529–1541

Foulds LR, Graham RL (1982) The Steiner problem in phylogeny is NP-complete. Adv Appl Math 3(1):43–49

Foundation PS (2010) Python language reference, version 2.7. http://www.python.org. Accessed 29 Apr 2019

Fukami K, Tateno Y (1989) On the maximum likelihood method for estimating molecular trees: uniqueness of the likelihood point. J Mol Evol 28(5):460–464

Gusfield D (1991) Efficient algorithms for inferring evolutionary trees. Networks 21(1):19–28

Hillis DM, Mable BK, Moritz C (1996) Molecular systematics. Sinauer Associates, Sunderland

Janies DA, Treseder T, Alexandrov B, Habibb F, Chen JJ, Ferreira R, Catalyurek U, Varon A, Wheeler WC (2011) The supramap project: linking pathogen genomes with geography to fight emergent infectious diseases. Cladistics 27:61–66

Roch S (2006) A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. IEEE/ACM Trans Comput Biol Bioinform 3(1):92–94

Semple C, Steel M (2003) Phylogenetics. Oxford lecture series in mathematics and its applications, vol 24. Oxford University Press, Oxford

Steel M (2011) The penny ante challenge problems: open problems from the New Zealand phylogenetics meetings. www.math.canterbury.ac.nz/bio/events/south2012/files/penny_ante_problems.pdf. Accessed 8 Aug 2019

Steel MA (1994) The maximum likelihood point for a phylogenetic tree is not unique. Syst Biol 43(4):560–564

Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol Biol Evol 17(6):839–850

Stein W et al (2015) Sage mathematics software (version 6.6). The Sage Development Team. http://www.sagemath.org. Accessed 8 Aug 2019

Stewart J (2005) Multivariable calculus: concepts and contexts. Brooks/Cole, Pacific Grove ISBN 0-534-41004-9

Tuffley C, Steel M (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull Math Biol 59(3):581–607