Joint maximum-likelihood of phylogeny and ancestral states is not consistent

David A. Shaw[1], Vu C. Dinh[2], and Frederick A. Matsen IV*[1]

[1]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
[2]Department of Mathematical Sciences, University of Delaware, Newark, DE, USA

Abstract

Maximum likelihood estimation in phylogenetics requires a means of handling unknown ancestral states. Classical maximum likelihood averages over these unknown intermediate states, leading to consistent estimation of the topology and continuous model parameters. Recently, a computationally-efficient approach has been proposed to jointly maximize over these unknown states and phylogenetic parameters. Although this method of joint maximum likelihood estimation can obtain estimates more quickly, its properties as an estimator are not yet clear. We show that this method of jointly estimating phylogenetic parameters along with ancestral states is not consistent in general. We find a set of parameters that generate data under a four-taxon tree for which this joint method estimates a multifurcating topology in the limit of infinite-length sequences by estimating one or more branches to be zero length. For branch length estimation on the correct topology, we show that this joint method cannot estimate consistent branch lengths except in degenerate cases, and we provide extensive empirical results for outlining the consistent bias in this setting.

*Corresponding author. Email: matsen@fredhutch.org

Introduction

Classical maximum likelihood (ML) estimation in phylogenetics operates by integrating out latent ancestral states at the internal nodes of the tree, obtaining an integrated likelihood [Goldman, 1990]. In a recent paper, Sagulenko et al. [2018] suggest using an approximation to ML inference in which the likelihood is maximized jointly across model parameters and ancestral sequences on a fixed topology. This is attractive from a computational perspective: such joint inference can proceed according to an iterative procedure in which ancestral sequences are first estimated and model parameters are optimized conditional on these estimates. This latter conditional optimization is simpler and more computationally efficient than optimizing the integrated likelihood. But is it statistically consistent?

An estimator is said to be statistically consistent if it converges to the generating model with probability one in the large-data limit; existing consistency proofs for maximum likelihood phylogenetics [Allman et al., 2008, Chai and Housworth, 2011, RoyChoudhury et al., 2015] apply only to estimating model parameters when the ancestral sequences have been integrated out of the likelihood. These proofs do not readily extend to include estimating ancestral states. Moreover, examples of inconsistency arising from problems where the number of parameters increases with the amount of data [Neyman and Scott, 1948] indicate that joint inference of trees and ancestral states may not enjoy good statistical properties. In this case those additional parameters are the states of ancestral sequences. Although Sagulenko et al. [2018] explicitly warn that the approximation is for the case where "branch lengths are short and only a minority of sites change on a given branch," their work motivates understanding the general properties of such joint inference. In particular, one would like to know when this approximate technique breaks down for both topology and branch length inference, even when sequence data is "perfect," i.e., is generated without sampling error according to the exact model used for inference.

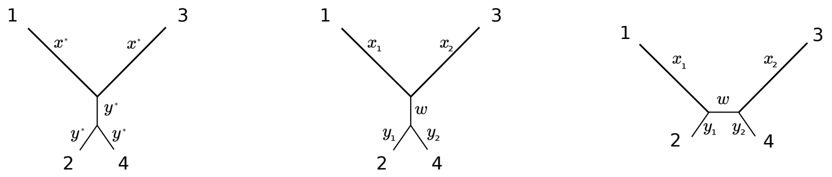
In this paper, we show that jointly inferring trees and ancestral sequences is not consistent in general. To do so, we use a binary symmetric model with data generated on a four-taxon tree: we compute closed form solutions to the joint objective function and demarcate a sizeable area of branch lengths in which joint inference is guaranteed to give a multifurcating tree in the case of perfect sequence data with an infinite number of sites by estimating one or more branch lengths to be zero. We show that, when the topology is known and fixed, joint inference cannot be consistent except in cases of zero or infinite branch length, and we find similar areas through empirical means where joint inference consistently underestimates interior branch lengths.

Phylogenetic maximum likelihood

Assume the binary symmetric model, namely with a character alphabet $\mathcal{A} = \{0, 1\}$ and a uniform stationary distribution [Semple and Steel, 2003]. Let m be the number of tips of the tree, and $p = m-2$ be the number of internal nodes. We observe n independent and identically distributed samples of character data, i.e., an alignment with n columns, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathcal{A}^{m \times n}$ distributed as the random variable Y. The corresponding unobserved ancestral states are $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_n] \in \mathcal{A}^{p \times n}$ and distributed as H with each $\mathbf{h}_i \in \mathcal{A}^p$.

We parameterize branches on the unique unrooted four-tip phylogenetic tree in ways known as the "inverse Felsenstein (InvFels)" tree (Figs. 1a and 1b) and the "Felsenstein" tree (Fig. 1c). The "inverse Felsenstein" terminology comes from Swofford et al. [2001], although it is also called the "Farris" tree [Siddall, 1998, Felsenstein, 2004]. In the standard configuration of this tree, the interior branch parameters are equal to the bottom two parameters as in Fig. 1a. We use this standard configuration as our data generating process, though we do not constrain our branch parameters to be equal when optimizing our objective function.

We parameterize the branches of these trees not with the standard notion of branch length in terms of number of substitutions per site, but with an alternate formulation called "fidelity." The probability of a substitution on a branch with fidelity x is $(1-x)/2$, while the probability of no substitu-



(a) InvFels tree τ^* (b) InvFels tree τ_1 (c) Felsenstein tree τ_2

Figure 1: Three four-taxon trees with fidelities as labeled.

tion is $(1+x)/2$ where $0 \leq x \leq 1$. This parameter quantifies the fidelity of transmission of the ancestral state across an edge [Matsen and Steel, 2007].

Fidelities have useful algebraic properties. As data becomes plentiful, we use the Hadamard transform (see (8) in the Appendix) to compute the exact probabilities that generate each particular configuration of taxa—we call these "generating probabilities"—and these have an especially simple form. For a four-taxon tree, define the general branch fidelity parameter $t = \{x_1, y_1, x_2, y_2, w\}$ where fidelities are ordered in the order of the taxa with the internal branch last (Figs. 1b and 1c). Although we use fidelities exclusively for our theoretical development, we have made our figures in terms of probabilities of substitution $p_x = (1-x)/2$ as they are easier to interpret.

Two paths to maximum likelihood

The standard phylogenetic likelihood approach on unrooted trees under the usual assumption of independence between sites is as follows. For a topology τ and branch fidelities t the likelihood given observed ancestral states $\mathbf{H}$ is

$$L_n(\tau, t; \mathbf{Y}, \mathbf{H}) = \prod_{i=1}^{n} \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t). \tag{1}$$

The probability $\Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t)$ is a product of transition probabilities determined by $\mathbf{Y}$, $\mathbf{H}$, τ, and t [Felsenstein, 2004].

The classical approach is to maximize the likelihood marginalized across ancestral states

$$\tilde{L}_n(\tau, t; \mathbf{Y}) = \prod_{i=1}^{n} \sum_{\mathbf{h}_i \in \mathcal{A}^p} \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) \tag{2}$$

to estimate the tree τ and branch fidelities t.

The alternative approach [Sagulenko et al., 2018] does away with the marginalization and directly estimates the maximum likelihood parameters of the fully-observed likelihood in (1). This is known in statistics as a profile likelihood [Murphy and van der Vaart, 2000] or a relative likelihood [Goldman, 1990], which exists here because $\mathcal{A}$ is a finite set:

$$L'_n(\tau, t; \mathbf{Y}) = \prod_{i=1}^{n} \max_{\mathbf{h}_i \in \mathcal{A}^p} \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) = \max_{\mathbf{H} \in \mathcal{A}^{p \times n}} L_n(\tau, t; \mathbf{Y}, \mathbf{H}). \tag{3}$$

We use $\hat{\mathbf{H}}_n$ to denote an estimate for $\mathbf{H}$ obtained by maximizing (3), and estimate a topology and branch fidelities using this profile likelihood as

$$(\hat{\tau}_n, \hat{t}_n) = \operatorname*{argmax}_{\tau, t} L'_n(\tau, t; \mathbf{Y}). \tag{4}$$

In general, the functional form of (3) is determined by inequalities arising from taking maxima over ancestral states (Table S2) to obtain each conditional likelihood term, these terms depending on the unknown (τ, t). For this reason, in practice, the joint inference strategy estimates $\hat{\mathbf{H}}_n$ for a fixed (τ, t), then $(\hat{\tau}_n, \hat{t}_n)$ given $\hat{\mathbf{H}}_n$, maximizing each of these conditional objectives until convergence [Sagulenko et al., 2018].

Inconsistency of joint inference

We now state our results on the inconsistency of joint inference. All proofs are deferred to the Appendix.

Assume $\mathbf{Y}$ is generated from the InvFels topology τ^* (Fig. 1a) and with true generating branch fidelities $t^* = \{x^*, y^*, x^*, y^*, y^*\}$. Let $\boldsymbol{\xi} = [\xi_j]_{j=1}^{q}$ be the vector of most likely ancestral state splits—the explicit definition for $\boldsymbol{\xi}$ is given in the Appendix. Use $\ell_{\tau^*,t^*}(\tau, t; \boldsymbol{\xi})$ to denote the expected per-site log-likelihood, which can be thought of as the infinite-length sequence case because, as shown in the Appendix,

$$\frac{1}{n} \log L'_n(\tau, t; \mathbf{Y}) \to \ell_{\tau^*,t^*}(\tau, t; \boldsymbol{\xi}). \tag{5}$$

We give ℓ explicitly as (7) in the Appendix. For a fixed τ, let $\hat{t}_n$ maximize the left-hand side of (5) and $\hat{t}$ maximize the right-hand side. We show in the Appendix that $\hat{t}_n \to \hat{t}$, allowing us to focus on only the right-hand side above.

Inconsistent branch length estimation

When the topology is known and fixed and we estimate only branch lengths, we show the following, i.e., that for all x^* and y^* in $(0, 1)$ any branch length estimate is consistently biased.

Theorem 1. *Let $\tau^* = \tau_1$, $t^* = \{x^*, y^*, x^*, y^*, y^*\}$, and $t = \{x_1, y_1, x_2, y_2, w\}$ with $x_1, y_1, x_2, y_2, w > 0$. For all $0 < x^*, y^* < 1$, the solution $\hat{t} := \{\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{w}\}$ given by*

$$\hat{t} = \arg\max_{t} \max_{\boldsymbol{\xi}} \ \ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi})$$

has the property $\hat{t} \neq t^*$.

In words, the joint estimation procedure never recovers the true generating t^* except in cases of zero or infinite branch length. This is apparent given Table S1, as the solution $\hat{t}$ is a linear combination of $p_{\tilde{y}_j}$ values, and no generating probability contains an x^* or y^* term.

Convergence to degenerate topology

Given data generated on τ_1 there exist true nonzero branch lengths such that the estimator $\hat{t}$ maximizing the right-hand side of (5) has an internal branch of length zero.

Theorem 2. *Let $\tau^* = \tau_1$, $t^* = \{x^*, y^*, x^*, y^*, y^*\}$, and $t = \{x_1, y_1, x_2, y_2, w\}$ with $x_1, y_1, x_2, y_2, w > 0$. There exists an open set of $0 < x^*, y^* < 1$ such that the solution $\hat{t} := \{\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{w}\}$ given by*

$$\hat{t} = \arg\max_{t} \max_{\boldsymbol{\xi}} \ \ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi})$$

has the property $\hat{w} \equiv 1$.

This result implies an inconsistency as we estimate the interior branch length to be zero (i.e., interior branch fidelity is one) in an open set of values for x^* and y^* (Fig. S2). As we consider different topologies τ_1 and τ_2 for $\hat{t}$, the incorrect topology τ_2 attains a likelihood value at its maximum equal to that of the true topology τ_1 in the limit. In other words, if $w = 1$ the objective functions $\ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi})$ and $\ell_{\tau^*,t^*}(\tau_2, t; \boldsymbol{\xi})$ are equivalent. We elaborate on this point in the Appendix. The proof is through analytically reducing the general case to 81 separate cases (Table S3) to obtain a closed form maximal value for each.

We provide the following as an intuition for the theoretical development. For a particular site pattern, to obtain the joint maximum likelihood function we maximize over ancestral states. For the internal branch—the branch between the two internal nodes—we have a choice of $(1 + w)$ or $(1 - w)$ in each of our likelihood terms depending on which ancestral state corresponds to the highest conditional log-likelihood. As $(1+w) > (1-w)$, a maximization procedure tends to prefer the $(1 + w)$ term, though this is not guaranteed because the maximum depends on the values of the unknown branch parameters t. Nevertheless, this tendency to include $(1+w)$ terms in the likelihood results in a positive bias of branch fidelities, i.e., estimating branch lengths to be shorter than truth. This is apparent in the "long x^*, short y^*" scenario as these are the cases in which the most likely ancestral states are the same for each internal node letting $x_1 = x_2 = x^*$ and $y_1 = y_2 = y^*$ ($\xi_j = \emptyset$ for all j in Table S3). If we allow multifurcating trees in our inference, then we can think of this as an instance of converging to the wrong topology, as the true $y^* \neq 1$.

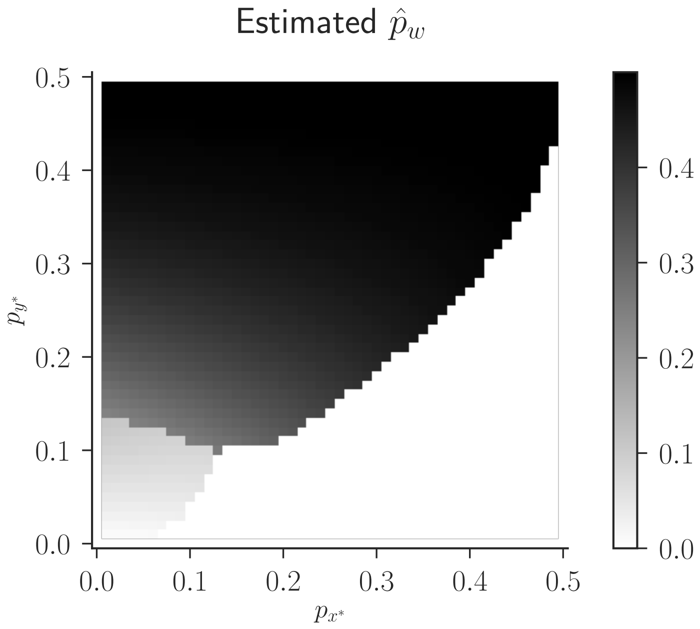


Figure 2: Estimates for $\hat{p}_w = (1 - \hat{w})/2$ when optimizing (3), where the true value for p_w is p_{y^*}. Data generated as in Fig. S2. The white region in the lower right highlights which values of x^* and y^* result in an interior branch being estimated as length zero, resulting in an inconsistency.

Empirical validation

Direct numerical optimization confirms our theoretically-derived bounds and provides a more detailed picture compared to the analytically-derived region (Fig. S2). To verify the regions of inconsistency and obtain a clearer picture of the closed form parameter estimates, we plot the optimal $\hat{w}$ via joint estimation (Fig. 2). As before, the region of inconsistency encompasses almost half of the branch fidelity space; given the correct topology, there are many situations where we estimate the interior branch length to be zero.

In our optimization procedure, we again consider the 81 separate cases (Table S3) and, for each function, we compute the closed form solution for $\hat{t}$. We compute these maxima over a lattice in steps of 10^{-2} for $x^*, y^* \in (0, 1)$. Our optimization code can be found at https://github.com/

matsengrp/joint-inf/.

In estimating the interior branch length w, we find a systematic bias in the joint inference procedure even when the true branches are short (Fig. 3). As data are generated with parameters $\{x^*, y^*, x^*, y^*, y^*\}$, the true value for w is y^*. There are discontinuities in the fit (Fig. 2) due to the choice of which ancestral state splits are maximal, so we investigate the bias in the region where p_{x^*} and p_{y^*} are both small, i.e., $p_{x^*}, p_{y^*} \leq .1$, as these short-branch cases should be the best settings for joint optimization [Sagulenko et al., 2018]. Although the estimates for $\hat{p}_w$ are better than the estimates when p_{y^*} is small and p_{x^*} is large (Fig. 2), joint inference still predictably underestimates the interior branch length. Additionally, the bias estimates $\hat{p}_w - p_{y^*}$ given $p_{x^*}, p_{y^*} \leq .1$ range from $[-4 \times 10^{-2}, 3 \times 10^{-3}]$.

Inference on the integrated likelihood performs as expected where $\hat{w}$ is equal to y^* regardless of the value of x^* (Fig. S3). We use L-BFGS-B when optimizing (2). The errors in this case are lower than machine tolerance showing that, even in cases where joint inference is supposed to do well, it still fails to achieve a low error from truth.

Discussion

We have shown that jointly inferring ancestral states and phylogenetic parameters [Sagulenko et al., 2018] is not consistent in general. Specifically, in the case of four-taxon trees with infinite data, we have obtained nontrivial regions of generating parameters that result in a type of topological inconsistency: the joint inference procedure estimates zero-length branches, which can be considered as a multifurcating topology. Also, the incorrect topology attains the same likelihood as the topology that generated the data by fixing this branch to have zero length. Since the parameters with the highest likelihood given the generating topology include a zero-length branch, we cannot exclude the possibility that the incorrect topology with this branch having nonzero length is more likely to be observed, though we have not found regions where this is the case. The regions of inconsistency we found arise when the top two branches of the generating trees are

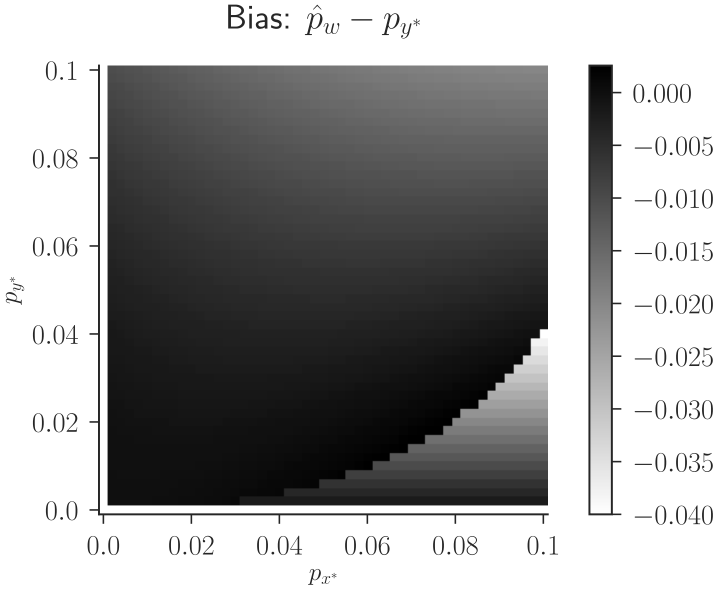


Figure 3: Bias in branch length estimation. Even in regions with short branch length ($p_{x^*}, p_{y^*} \leq .1$) where joint optimization should perform well, there is systematic bias toward shorter branch lengths.

"long," that is, when the top branch fidelities tend to be small, and when the lower branches are "short," i.e., have large fidelities. We see that this inconsistency occurs even if some branches are short. This expands on the empirical findings of poor estimation given long branches in Sagulenko et al. [2018] (their Figures 2 and 3). However, the problems are not just for long branches as Sagulenko et al. [2018] imply: even when all branches are short there is a consistent bias, and the bias is on the same order as the magnitude of the parameters (Fig. 3). In addition, we have shown there are no nontrivial generating parameters that yield consistent branch length estimates.

Joint inference of tree parameters and ancestral sequences is a type of profile likelihood, a well-studied subject in statistics [Murphy and van der Vaart, 2000]. Many properties regarding the performance of maximum likelihood estimates obtained using this approach are known, and many methods exist to overcome their undesirable properties, e.g., the method of sieves [Geman and Hwang, 1982]. A potential solution in this case using the method of sieves could be to project the column-wise ancestral states into a lower-dimensional space, allowing the degrees of freedom in the ancestral state columns to grow with n, albeit more slowly than $O(n)$. Elsewhere in statistics literature, the failure of maximum likelihood estimates to obtain consistent estimates as the number of parameters goes to infinity have been shown by the Neyman-Scott paradox [Neyman and Scott, 1948], though parameters tending to infinity is not a necessary condition for inconsistency [Le Cam, 1990]. Consistency proofs of standard maximum likelihood estimates of phylogeny (2) are recent [Allman et al., 2008, Chai and Housworth, 2011, RoyChoudhury et al., 2015], and no results have been obtained for profile likelihood. We have furthered progress in understanding the limitations of this joint optimization procedure.

Previous work in phylogenetics has developed consistency counterexamples using similar four-taxon topologies to the one used here [Felsenstein, 1978]. In this previous work, when simulating data under the Felsenstein topology τ_2, as the number of observations increases, the InvFels topology τ_1 becomes more likely when performing a particular estimation procedure. We have shown cases in which, when generating from the InvFels topology, we converge to a multifurcating topology, with one or more branch lengths estimated to be zero. Moreover, the inconsistency demonstrated by Felsenstein [1978] is attributed to long branch attraction, i.e., the fact that there may be multiple long branches where parallel changes are more likely than a single change along a short branch. This is not the case here; while analytically the inconsistency occurs when the top two branches are long and the bottom three are short, we see empirically that this inconsistency is present in roughly half of the entire parameter space, and occurs when the true branches generate data that more likely has no change along the interior branch. Additionally, we generate data on the InvFels tree τ_1 while Felsenstein [1978] generates data on the Felsenstein tree τ_2. Difficulties in phylogenetic estimation when generating data on the InvFels tree have been found by Siddall [1998], though Swofford et al. [2001] show that sequence length plays a major role in these issues.

The case of joint inference of a phylogenetic likelihood is discussed in Goldman [1990]. There, Goldman provides a worked example in which estimating a topology with fixed branch lengths is equivalent to parsimony and thus not guaranteed to be consistent, though he does not discuss the inconsistency of joint inference in general. We show cases where the incorrect topology attains an equal likelihood value at the maximum as the correct topology, and, moreover, if we know the correct topology, we show cases where branch lengths are severely biased and cannot be consistent. Finally, just prior to his conclusion, he discusses when parsimony gives the same answer as maximum likelihood, concluding that the question is ill-posed since parsimony estimates different parameters than maximum likelihood, i.e., it assumes equal branch lengths. We render the question well-posed: the joint inference procedure outlined here estimates the same parameters as classical maximum likelihood—topology and branch lengths—albeit implicitly estimating ancestral states as well. We are able to provide much more detail on how large branch lengths must be for general joint inference to fail to be consistent.

We have shown an inconsistency when performing joint inference on branch lengths given an InvFels topology and investigated the performance of branch parameter estimation. There is substantial scope for future work to make these results more precise and more general. All of these results hold only for a simple binary symmetric model on four-taxon trees, and extensive simulation is necessary to understand how these results extend to more complicated general cases, such as applied examples with larger trees or more realistic mutation models that are of interest to practitioners. Also, given that many of the bounds presented here are in the form of level sets of multivariate polynomials, a more formal approach using algebraic geometric techniques may reveal more stable or interesting patterns of inconsistency; see Sturmfels [2002] for a thorough treatment of solving systems of polynomial equations. Finally, all of the material presented here concerns joint estimation under maximum likelihood, and does not pose any problem for other settings, such as joint sampling of trees and ancestral sequences in a Bayesian framework.

Acknowledgements

We thank Richard Neher, Vladimir Minin, and Joe Felsenstein for helpful discussions. Insight from the reviewers and the editors greatly improved this manuscript.

This work was supported by National Institutes of Health grants R01-GM113246, R01-AI120961, U19-AI117891, and U54-GM111274 as well as National Science Foundation grants CISE-1561334 and CISE-1564137. The research of Frederick Matsen was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation.

References

Elizabeth S Allman, Cécile Ané, and John A Rhodes. Identifiability of a markovian model of molecular evolution with Gamma-Distributed rates.

Adv. Appl. Probab., 40(1):229–249, 2008. ISSN 0001-8678. URL http://www.jstor.org/stable/20443578.

Juanjuan Chai and Elizabeth A Housworth. On rogers' proof of identifiability for the GTR + Γ + I model. *Syst. Biol.*, 60(5):713–718, October 2011. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syr023. URL http://dx.doi.org/10.1093/sysbio/syr023.

Joseph Felsenstein. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410, 1 December 1978. ISSN 0039-7989. doi: 10.2307/2412923. URL http://www.jstor.org/stable/2412923.

Joseph Felsenstein. *Inferring phylogenies*. Sinauer Associates, Inc., Sunderland, Massachusetts, 2004.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, 10(2):401–414, 1982.

Nick Goldman. Maximum likelihood inference of phylogenetic trees, with special reference to a poisson process model of DNA substitution and to parsimony analyses. *Syst. Biol.*, 39(4):345–361, December 1990. ISSN 1063-5157. doi: 10.2307/2992355. URL https://academic.oup.com/sysbio/article-abstract/39/4/345/1646997?redirectedFrom=fulltext.

Lucien Le Cam. Maximum likelihood: An introduction. *International Statistical Review*, 58(2):153–171, Aug 1990.

Frederick A. Matsen and Mike Steel. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic Biology*, 56(5):767–775, 1 October 2007. ISSN 1063-5157. doi: 10.1080/10635150701627304. URL http://sysbio.oxfordjournals.org/content/56/5/767.abstract.

Susan A. Murphy and Aad W. van der Vaart. On profile likelihood. *Journal of the American Statistical Association*, 95(450):449–465, 2000. ISSN 0162-1459. doi: 10.2307/2669386. URL http://www.jstor.org/stable/2669386.

Jerzy Neyman and Elizabeth L. Scott. Consistent estimates based on partially consistent observations. *Econometrica*, 16(1):1–32, 1948. ISSN 0012-9682, 1468-0262. doi: 10.2307/1914288. URL http://www.jstor.org/stable/1914288.

Arindam RoyChoudhury, Amy Willis, and John Bunge. Consistency of a phylogenetic tree maximum likelihood estimator. *Journal of Statistical Planning and Inference*, 161:73–80, June 2015. ISSN 0378-3758. doi: 10.1016/j.jspi.2015.01.001. URL http://www.sciencedirect.com/science/article/pii/S0378375815000038.

Pavel Sagulenko, Vadim Puller, and Richard A Neher. TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol*, 4(1):vex042, January 2018. ISSN 2057-1577. doi: 10.1093/ve/vex042. URL http://dx.doi.org/10.1093/ve/vex042.

Charles Semple and Mike Steel. *Phylogenetics*. Oxford University Press, New York, NY, 2003.

Mark E Siddall. Success of parsimony in the Four-Taxon case: Long-Branch repulsion by likelihood in the farris zone. *Cladistics*, 14(3):209–220, 1 September 1998. ISSN 0748-3007, 1096-0031. doi: 10.1111/j.1096-0031.1998.tb00334.x. URL http://dx.doi.org/10.1111/j.1096-0031.1998.tb00334.x.

Bernd Sturmfels. Solving systems of polynomial equations. In *American Mathematical Society, CBMS Regional Conferences Series, No. 97*, 2002.

David L. Swofford, Peter J. Waddell, John P. Huelsenbeck, Peter G. Foster, Paul O. Lewis, and James S. Rogers. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods.

Systematic Biology, 50(4):525–539, August 2001. ISSN 1063-5157. URL http://www.ncbi.nlm.nih.gov/pubmed/12116651.

A.W. van Der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, 3. Cambridge University Press, 1998. ISBN 9780521496032. URL https://books.google.com/books?id=udhfQgAACAAJ.

Appendix

Site split formulation

We begin by introducing "site splits." We use site splits to formalize the notion that a given site pattern is equally probable to its complement under the binary symmetric model. This is a standard step in the description of the Hadamard transform (Section 8.6 of Semple and Steel [2003]), although our approach is complicated slightly by the inclusion of ancestral states.

Since we have a finite character alphabet, for a given column i there are a finite number of possible assignments of characters to tips $\mathbf{y}_i$ or internal nodes $\mathbf{h}_i$. For the binary symmetric model, the alphabet $\mathcal{A}$ is $\{0, 1\}$. Take the tip labels of τ to be $\{1, \ldots, m\}$. For likelihood calculation under the binary symmetric model, we describe a given $\mathbf{y}_i$ as a subset of indices $\tilde{y} \subseteq \mathcal{Y} := \{1, \ldots, m-1\}$, commonly called a "site split." Define the complement of $\mathbf{y}$ as $\overline{\mathbf{y}}$, and let $\mathbf{y}_{i,k}$ be the label of the kth tip in the ith alignment column. We define the site split $\tilde{y}$ for a $\mathbf{y}_i$ as the set of tips labeled with 1 in $\mathbf{y}_i$ if the mth tip is not labeled with 1, and as the set of tips labeled with 1 in $\overline{\mathbf{y}}_i$ if the mth tip is labeled with 1. Taking such a complement simplifies but does not change the result of likelihood computation because the probability of observing a particular collection of binary characters is equivalent to the probability of its complement under the binary symmetric model.

For a fixed topology τ, we define an ordered set of internal node labels $\{1, \ldots, p\}$ for $\mathbf{h}_i$ and similarly use a subset of characters $\tilde{h} \subseteq \mathcal{H} := \{1, \ldots, p\}$ to describe a realization $\mathbf{h}_i$. In this case we cannot use the same complement trick as before: the probability of observing an ancestral state split conditional on a site split is not invariant to taking its complement. We thus define an "ancestral state split" $\tilde{h}$ for an internal node $\mathbf{h}_i$ to be the set of internal nodes labeled with 1 if the mth tip is not labeled with 1, and as the set of internal nodes labeled with 1 in $\overline{\mathbf{h}}_i$ if the mth tip is labeled with 1. We emphasize that the ancestral state split complementing procedure depends on tip states, not ancestral states: both site splits and ancestral state splits are defined by whether the mth element of $\mathbf{y}_i$ is labeled as 1.

We enumerate the site splits $\tilde{y}_j$ of which there are $q = |\mathcal{P}(\mathcal{Y})|$ in total where $\mathcal{P}$ denotes the power set. Similarly we enumerate ancestral state splits $\tilde{h}_k$ of which there are $r = |\mathcal{P}(\mathcal{H})|$ in total.

We first fix notation.

Definition. *Let the mapping from site patterns to site splits*

$$\psi : \mathcal{A}^m \to \mathcal{P}(\mathcal{Y})$$

be

$$\psi(\mathbf{y}) = \begin{cases} \{i' \in \{1, \ldots, m-1\} : \mathbf{y}_{i,i'} = 1\} & \textit{if}\ \mathbf{y}_{i,m} = 0, \\ \{i' \in \{1, \ldots, m-1\} : \overline{\mathbf{y}}_{i,i'} = 1\} & \textit{if}\ \mathbf{y}_{i,m} = 1, \end{cases}$$

and the mapping from ancestral states and tip states to ancestral state splits

$$\xi : \mathcal{A}^m \times \mathcal{A}^p \to \mathcal{P}(\mathcal{H})$$

be

$$\xi(\mathbf{y}, \mathbf{h}) = \begin{cases} \{i' \in \{1, \ldots, p\} : \mathbf{h}_{i,i'} = 1\} & \textit{if}\ \mathbf{y}_{i,m} = 0, \\ \{i' \in \{1, \ldots, p\} : \overline{\mathbf{h}}_{i,i'} = 1\} & \textit{if}\ \mathbf{y}_{i,m} = 1. \end{cases}$$

Then, given a site pattern–valued random variable Y and an ancestral state–valued random variable H, define the random variables

$$\Psi := \psi(Y)$$

and

$$\Xi := \xi(Y, H).$$

The mapping ψ operates by returning the tips labeled as 1 in a site pattern to obtain a site split in $\mathcal{P}(\mathcal{Y})$ if the set of tips labeled 1 is not in $\mathcal{P}(\mathcal{Y})$. The mapping ξ is defined by whether the tip states have their complements taken or not: if the set of tips labeled 1 in $\mathbf{y}$ is in $\mathcal{P}(\mathcal{Y})$, $\xi(\mathbf{y}, \mathbf{h})$ is the set of tips labeled 1 in $\mathbf{h}$; otherwise, the set of tips labeled 1 in $\overline{\mathbf{y}}$ necessarily is in $\mathcal{P}(\mathcal{Y})$ and so $\xi(\mathbf{y}, \mathbf{h})$ is $\overline{\mathbf{h}}$.

We now consider the ith factor of (1). As a consequence of assuming a binary symmetric model, for some $\tilde{y}_j \in \mathcal{P}(\mathcal{Y})$ the mapping $\psi(\mathbf{y}_i)$ has the property

$$\begin{aligned}
\Pr(\Psi = \tilde{y}_j, \Xi = \tilde{h}_k \mid \tau, t) &= \Pr(\Psi = \psi(\mathbf{y}_i), \Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \tau, t) \\
&= \Pr((Y = \mathbf{y}_i, H = \mathbf{h}_i) \cup (\overline{Y} = \mathbf{y}_i, \overline{H} = \mathbf{h}_i) \mid \tau, t) \\
&= \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) + \Pr(\overline{Y} = \mathbf{y}_i, \overline{H} = \mathbf{h}_i \mid \tau, t) \\
&= 2 \cdot \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t)
\end{aligned}$$

where $\overline{Y}$ is the complement of the site pattern–valued random variable Y and has the same distribution as Y (similarly for H). Since

$$2 \cdot \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) = \Pr(\Psi = \psi(\mathbf{y}_i), \Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \tau, t),$$

given (τ, t), there exist sets $\eta_1(\tau, t), \ldots, \eta_q(\tau, t)$ such that $\xi_j \in \eta_j(\tau, t)$ satisfies

$$\max_{\tilde{h}_k \in \mathcal{P}(\mathcal{H})} \Pr(\Psi = \tilde{y}_j, \Xi = \tilde{h}_k \mid \tau, t) = \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t).$$

In other words, for the jth site split, $\eta_j(\tau, t) \subseteq \mathcal{P}(\mathcal{H})$ is the set of most likely ancestral state splits for that particular site split, topology and set of branch lengths, i.e., $\eta_j(\tau, t)$ is a set of sets of most likely internal node labels. Here, ξ_j is one of possibly many equiprobable ancestral state splits in $\eta_j(\tau, t)$. For each $\mathbf{y}_i$, $\xi(\mathbf{y}_i, \cdot)$ is surjective as it can map values from $\mathcal{A}^p$ to all elements in $\mathcal{P}(\mathcal{H})$. This can be seen by using the definition of $\xi(\mathbf{y}_i, \cdot)$ and assuming $\mathbf{y}_{i,m} = 0$, where in this case each of the 2^p values of $\mathbf{h}$ correspond to each of the 2^p elements of $\mathcal{P}(\{1, \ldots, p\})$. The same can be done for the case of $\mathbf{y}_{i,m} = 1$, implying $\xi(\mathbf{y}_i, \cdot)$ is surjective. From this we have

$$\begin{aligned}
\max_{\mathbf{h}_i} 2 \cdot \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) &= \max_{\mathbf{h}_i} \Pr(\Psi = \psi(\mathbf{y}_i), \Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \tau, t) \\
&= \max_{\tilde{h}_k \in \mathcal{P}(\mathcal{H})} \Pr(\Psi = \tilde{y}_j, \Xi = \tilde{h}_k \mid \tau, t) \\
&= \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t)
\end{aligned}$$

for some j. Thus, each term in the likelihood can be collapsed into terms relating only to site splits and ancestral state splits, indexed by j, as opposed to individual observations, indexed by i.

Example

We follow with an example computing these probabilities and likelihoods. Consider the fixed, binary four-taxon tree τ_1 in Fig. 1a. The set of all possible character assignments is

$$\begin{aligned}\mathcal{P}(\{1,2,3,4\}) = \{&\emptyset, \{1,2,3,4\}, \{1\}, \{2,3,4\}, \{2\}, \{1,3,4\}, \{3\}, \{1,2,4\},\\ &\{1,2\}, \{3,4\}, \{1,3\}, \{2,4\}, \{2,3\}, \{1,4\}, \{1,2,3\}, \{1,4\}\}\end{aligned}$$

where each set indicates the tips assigned the character 1. For example, $\emptyset$ is the labeling 0000 and $\{1,3,4\}$ is the labeling 1011. Symmetry allows us to group adjacent pairs in $\mathcal{P}(\{1,2,3,4\})$ into equiprobable splits, letting $\mathcal{Y} = \{1,2,3\}$. The unique site splits, collapsing complements, are

$$\begin{aligned}\mathcal{P}(\mathcal{Y}) &= \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}\\ &=: \{\tilde{y}_1, \ldots, \tilde{y}_8\}.\end{aligned}$$

Since we identify character complements, we do not consider the additional splits

$$\begin{aligned}&\mathcal{P}(\{1,2,3,4\}) \setminus \mathcal{P}(\mathcal{Y}) =\\ &\quad \{\{1,2,3,4\}, \{2,3,4\}, \{1,3,4\}, \{1,2,4\}, \{3,4\}, \{2,4\}, \{1,4\}, \{4\}\},\end{aligned}$$

the symmetry of the binary character model allowing us to focus only on the elements of $\mathcal{P}(\mathcal{Y})$. This tree has two internal nodes with $\mathcal{H} = \{1,2\}$ and unique ancestral state splits

$$\mathcal{P}(\mathcal{H}) = \{\emptyset, \{1\}, \{2\}, \{1,2\}\}.$$

Internal node 1 is the node connected to leaves 1 and 3 while internal node 2 is connected to leaves 2 and 4. The mapping from characters to splits in this case depends on the characters at the tips and the ancestral states. For example, we take both $\psi(0000) = \emptyset$ and $\psi(1111) = \emptyset$. Similarly, we have $\xi(0000, 00) = \emptyset$ and $\xi(1111, 11) = \emptyset$, needing to take the complement of all the characters present on the tree to identify splits. We cannot identify complements for ancestral states in the same way as tip states since, for $\tilde{y} \in \mathcal{P}(\mathcal{Y})$,

$$\Pr(\Psi = \tilde{y}, \Xi = \emptyset \mid \tau, t) \neq \Pr(\Psi = \tilde{y}, \Xi = \{1, 2\} \mid \tau, t)$$

in general.

For each site split $\tilde{y} \in \mathcal{P}(\mathcal{Y})$, we maximize the likelihood over all $\tilde{h} \in \mathcal{P}(\mathcal{H})$. A maximum occurs at one of possibly several ancestral state splits in $\mathcal{P}(\mathcal{H})$, defined via $\eta_j(\tau, t)$ for the jth site split. As a simple example, say all branch lengths correspond to a probability p $(< 1/2)$ of changing character along that branch, with $t = \{p, p, p, p, p\}$. The probabilities of observing ancestral state splits for $\tilde{y}_1 = \emptyset$ are

$$\Pr(\Psi = \emptyset, \Xi = \emptyset \mid \tau, t) = (1 - p)^5,$$

$$\Pr(\Psi = \emptyset, \Xi = \{1\} \mid \tau, t) = \Pr(\Psi = \emptyset, \Xi = \{2\} \mid \tau, t) = p^3(1 - p)^2,$$

$$\Pr(\Psi = \emptyset, \Xi = \{1, 2\} \mid \tau, t) = p^4(1 - p).$$

The set of most likely ancestral states contains a single element, here $\eta_1(\tau, t) = \{\emptyset\}$. Then, taking $\xi_1 \in \eta_1(\tau, t)$ we have

$$\Pr(\Psi = \emptyset, \Xi = \xi_1 \mid \tau, t) = \Pr(\Psi = \emptyset, \Xi = \emptyset \mid \tau, t) = (1 - p)^5.$$

For $\tilde{y}_5 = \{1, 2\}$ we have

$$\Pr(\Psi = \{1, 2\}, \Xi = \emptyset \mid \tau, t) = \Pr(\Psi = \{1, 2\}, \Xi = \{1, 2\} \mid \tau, t) = p^2(1 - p)^3,$$

$$\Pr(\Psi = \{1, 2\}, \Xi = \{1\} \mid \tau, t) = \Pr(\Psi = \{1, 2\}, \Xi = \{2\} \mid \tau, t) = p^3(1 - p)^2.$$

Here, the set of most likely ancestral states is $\eta_5(\tau, t) = \{\emptyset, \{1,2\}\}$, and, for $\xi_5 \in \eta_5(\tau, t)$,

$$\Pr(\Psi = \{1,2\}, \Xi = \xi_5 \mid \tau, t) = p^2(1-p)^3.$$

Site split likelihood

The likelihood in (3) can be written as

$$\begin{aligned} L'_n(\tau, t; \mathbf{Y}) &= \max_{\mathbf{H}} \ L_n(\tau, t; \mathbf{Y}, \mathbf{H}) \\ &= \prod_{i=1}^{n} \max_{\mathbf{h}_i} \ \Pr(Y = \mathbf{y}_i, H = \mathbf{h}_i \mid \tau, t) \\ &\propto \prod_{i=1}^{n} \max_{\mathbf{h}_i} \ \Pr(\Psi = \psi(\mathbf{y}_i), \Xi = \xi(\mathbf{y}_i, \mathbf{h}_i) \mid \tau, t) \\ &= \prod_{i=1}^{n} \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t) \\ &= \prod_{j=1}^{q} \left[\Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t)\right]^{n_j(\mathbf{Y})} \end{aligned} \tag{6}$$

for $\tilde{y}_j \in \mathcal{P}(\mathcal{Y})$ and some $\xi_j \in \eta_j(\tau, t)$ with $1 \leq j \leq q$ where $n_j(\mathbf{Y})$ is the number of columns in $\mathbf{Y}$ that project to site split $\tilde{y}_j$.

Let

$$L''_n(\tau, t; \mathbf{Y}) = \prod_{j=1}^{q} \left[\Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t)\right]^{n_j(\mathbf{Y})}$$

be the final product in (6). Assume n observations are generated from a model with parameters (τ^*, t^*). We have

$$\frac{1}{n} \log L''_n(\tau, t; \mathbf{Y}) = \sum_{j=1}^{q} \frac{n_j(\mathbf{Y})}{n} \cdot \log \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t)$$

so that, in the $n \to \infty$ limit,

$$\begin{aligned}&\frac{1}{n}\log L_n''(\tau, t; \mathbf{Y})\\ &\qquad \to \sum_{j=1}^{q} \Pr(\Psi = \tilde{y}_j \mid \tau^*, t^*) \cdot \log \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t).\end{aligned} \tag{7}$$

Hadamard representation

We state the Hadamard representation of site split generating probabilities—that is, probabilities of obtaining particular site splits given a tree—following Section 8.6 of Semple and Steel [2003]. For each edge e define the edge "fidelity" for that edge as

$$\theta(e) = 1 - 2p(e)$$

where $p(e)$ is the probability of a character change along edge e. For an even-sized subset of $Y \subseteq \mathcal{S}$, let the path set $P(Y)$ be the set of edges in the path connecting both elements of Y. For n taxa, the probability of observing site split $A \in \mathcal{P}(\mathcal{Y})$ is

$$p_A = \frac{1}{2^{n-1}} \sum_{Y \subseteq \mathcal{S} : |Y| \equiv 0 (\mathrm{mod}\ 2)} \left[(-1)^{|Y \cap A|} \prod_{e \in P(Y)} \theta(e) \right]. \tag{8}$$

By convention, we set $P(\emptyset) = \emptyset$ and $\prod_{e \in \emptyset} \theta(e) = 1$. For notational convenience, let

$$p_{\tilde{y}_j} := \Pr(\Psi = \tilde{y}_j \mid \tau_1, t),$$

for any site split $\tilde{y}_j$. Table S1 contains calculations of site split probabilities for the trees in Fig. 1.

Likelihood computations

To compute the likelihood of observing a set of data, we need $\Pr(\Psi = \tilde{y}_j, \Xi = \tilde{h}_k \mid \tau, t)$ for each $\tilde{h}_k$ and $\tilde{y}_j$. Using branch fidelities, the probability of a character change along a branch with fidelity parameter x is $(1 - x)/2$, while the probability of a character remaining the same is $(1 + x)/2$. See

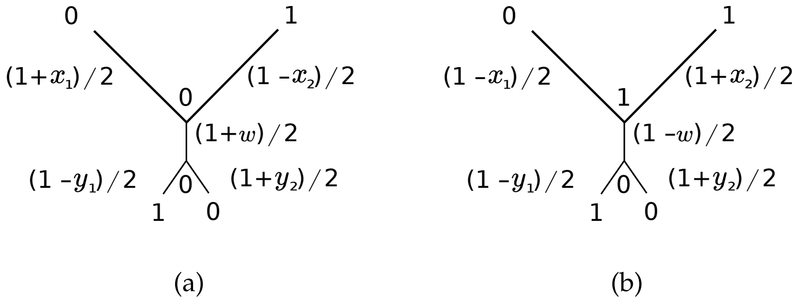


Figure S1: Example likelihood computations on the InvFels tree τ_1 for fidelities $t = \{x_1, y_1, x_2, y_2, w\}$. Edges labeled by the probability of substitution along that edge. In (a), we compute the product to obtain $\Pr(\Psi = \{2,3\}, \Xi = \emptyset \mid \tau_1, t) = (1+x_1)(1-x_2)(1+y_1)(1-y_2)(1+w)/32$. In (b), the same process yields $\Pr(\Psi = \{2,3\}, \Xi = \{1\} \mid \tau_1, t) = (1+x_1)(1-x_2)(1+y_1)(1-y_2)(1-w)/32$.

Fig. S1 for the parameters on an example site pattern on the InvFels tree. Likelihood computations for all site splits and ancestral state splits are in Table S2 for the InvFels tree.

Convergence of branch parameters

For a fixed τ, we show that $\hat{t}_n \to \hat{t}$ for

$$\hat{t}_n = \arg\max_{t \in \mathcal{T}} \frac{1}{n} \log L'_n(\tau, t; \mathbf{Y})$$

and

$$\hat{t} = \arg\max_{t \in \mathcal{T}} \ell_{\tau^*, t^*}(\tau, t; \boldsymbol{\xi}).$$

Using the notation in Section 5.2.1 in van Der Vaart [1998], we let

$$m_t(\mathbf{y}) = \sum_{j=1}^{q} 1\{\psi(\mathbf{y}) = \tilde{y}_j\} \cdot \log \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, t)$$

so that

$$\frac{1}{n} \log L'_n(\tau, t; \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^{n} m_t(\mathbf{y}_i)$$

and

$$\ell_{\mathcal{T}^*,t^*}(\tau, t; \boldsymbol{\xi}) = E[m_t].$$

To show $\hat{t}_n \to \hat{t}$, we use Wald's consistency proof [p. 48, Theorem 5.14 of van Der Vaart, 1998], which requires four conditions. The first is that $\mathcal{T}$ is compact, which is obviously true. The second is that

$$E\left[\sup_{t \in \mathcal{T}} m_t\right] < \infty,$$

and, since $m_t(\mathbf{y})$ is nonpositive for all t and $\mathbf{y}$, this property holds. The remaining conditions are on the maps

$$\mathbf{y} \mapsto \sup_t m_t(\mathbf{y})$$

and

$$t \mapsto m_t(\mathbf{y}).$$

We need the first map to be measurable, which is evident since the domain $\mathcal{A}^m$ of the mapping is a finite set, and so all subsets of the domain are also finite and thus measurable. Finally, we must have the the second mapping be upper-semicontinuous for almost all $\mathbf{y}$. For a fixed ancestral state split $t \mapsto m_t(\mathbf{y})$ is continuous for all $\mathbf{y}$. If we move about in $\mathcal{T}$, a different ancestral state split becomes more likely, though when we maximize over ancestral state splits we obtain a continuous function since the maximum over continuous functions is also continuous. This ensures the upper-semicontinuous property of this mapping, and shows $\hat{t}_n \to \hat{t}$, allowing our consistency results to be proved using $\ell_{\mathcal{T}^*,t^*}(\tau, t; \boldsymbol{\xi})$.

Properties of the joint objective function

Consider the InvFels tree τ_1 with arbitrary fidelities, i.e., $t = \{x_1, y_1, x_2, y_2, w\}$. Next we show that the likelihood $\ell_{\tau_1,t}(\tau_1, t; \boldsymbol{\xi})$ remains unchanged if x_1 and x_2 are exchanged or if y_1 and y_2 are. Although this property should not be surprising due to symmetry, we write it out for completeness. This holds for a general t, and thus holds setting $t = t^*$. Using the Hadamard transform, we calculate the generating probabilities on the InvFels tree. For site split $\emptyset$,

$$\begin{aligned}\Pr(\Psi = \emptyset \mid \tau_1, t) &= \frac{1}{8}(1 + x_1x_2 + y_1y_2 + x_1y_1w + x_1y_2w + y_1x_2w + x_2y_2w + x_1y_1x_2y_2)\\ &= \frac{1}{8}(1 + x_1x_2 + y_1y_2 + w[x_1y_1 + x_1y_2 + y_1x_2 + x_2y_2] + x_1y_1x_2y_2)\\ &= \frac{1}{8}(1 + x_1x_2 + y_1y_2 + w[x_1 + x_2][y_1 + y_2] + x_1y_1x_2y_2),\end{aligned}$$

and this probability is unchanged when x_1 is exchanged with x_2 and y_1 is exchanged with y_2. Similarly, for site split $\{1, 3\}$,

$$\Pr(\Psi = \{1, 3\} \mid \tau_1, t) = \frac{1}{8}(1 + x_1x_2 + y_1y_2 - w[x_1 + x_2][y_1 + y_2] + x_1y_1x_2y_2),$$

which also is invariant to exchanging x_1 with x_2 and y_1 with y_2.

All other generating probabilities differ only in the signs of each term (see Table S1). For example, for site split $\{1\}$ we have

$$\Pr(\Psi = \{1\} \mid \tau_1, t) = \frac{1}{8}(1 - x_1x_2 + y_1y_2 + w[-x_1 + x_2][y_1 + y_2] - x_1y_1x_2y_2)$$

and for site split $\{3\}$ we have

$$\Pr(\Psi = \{3\} \mid \tau_1, t) = \frac{1}{8}(1 - x_1x_2 + y_1y_2 + w[x_1 - x_2][y_1 + y_2] - x_1y_1x_2y_2)$$

meaning if we exchange the values of x_1 and x_2 then these probabilities swap values, regardless of what we do with y_1 and y_2. We show that for site splits $\{1\}$ and $\{3\}$, exchanging x_1 and x_2 also swaps the values of the likelihood terms, again independent of what happens to y_1 and y_2 (Table S2). Indeed, the corresponding possibilities for the likelihood values are

$$\begin{aligned}\Pr(\Psi = \{1\}, \Xi = \emptyset \mid \tau_1, t) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 + w)(1 + y_1)(1 + y_2);\\ \Pr(\Psi = \{1\}, \Xi = \{1\} \mid \tau_1, t) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 - w)(1 + y_1)(1 + y_2);\end{aligned}$$

$$\begin{aligned}
\Pr(\Psi = \{1\}, \Xi = \{2\} \mid \tau_1, t) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 - w)(1 - y_1)(1 - y_2);\\
\Pr(\Psi = \{1\}, \Xi = \{1, 2\} \mid \tau_1, t) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 + w)(1 - y_1)(1 - y_2);
\end{aligned}$$

for site split $\{1\}$ and

$$\begin{aligned}
\Pr(\Psi = \{3\}, \Xi = \emptyset \mid \tau_1, t) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 + w)(1 + y_1)(1 + y_2);\\
\Pr(\Psi = \{3\}, \Xi = \{1\} \mid \tau_1, t) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 - w)(1 + y_1)(1 + y_2);\\
\Pr(\Psi = \{3\}, \Xi = \{2\} \mid \tau_1, t) &= \frac{1}{32}(1 + x_1)(1 - x_2)(1 - w)(1 - y_1)(1 - y_2);\\
\Pr(\Psi = \{3\}, \Xi = \{1, 2\} \mid \tau_1, t) &= \frac{1}{32}(1 - x_1)(1 + x_2)(1 + w)(1 - y_1)(1 - y_2);
\end{aligned}$$

for site split $\{3\}$, which shows the likelihood remains unchanged if x_1 and x_2 are swapped.

For site splits $\{2\}$ and $\{1, 2, 3\}$, exchanging y_1 and y_2 swaps the values of the generating probabilities, independent of what happens to x_1 and x_2. In the case of the likelihood values, we see that the values for these site splits swap as well, though, we look at the complement of the most likely ancestral state split. In other words, the function value for the likelihood also swaps between site splits $\{2\}$ and $\{1, 2, 3\}$, though the most likely ancestral state split is different. Indeed,

$$\begin{aligned}
\Pr(\Psi = \{2\}, \Xi = \emptyset \mid \tau_1, t) &= \frac{1}{32}(1 + x_1)(1 - y_1)(1 + x_2)(1 + y_2)(1 + w);\\
\Pr(\Psi = \{2\}, \Xi = \{1\} \mid \tau_1, t) &= \frac{1}{32}(1 - x_1)(1 - y_1)(1 - x_2)(1 + y_2)(1 - w);\\
\Pr(\Psi = \{2\}, \Xi = \{2\} \mid \tau_1, t) &= \frac{1}{32}(1 + x_1)(1 + y_1)(1 + x_2)(1 - y_2)(1 - w);\\
\Pr(\Psi = \{2\}, \Xi = \{1, 2\} \mid \tau_1, t) &= \frac{1}{32}(1 - x_1)(1 + y_1)(1 - x_2)(1 - y_2)(1 + w);
\end{aligned}$$

for site split $\{2\}$ and

$$\Pr(\Psi = \{1, 2, 3\}, \Xi = \emptyset \mid \tau_1, t) = \frac{1}{32}(1 - x_1)(1 - y_1)(1 - x_2)(1 + y_2)(1 + w);$$

$$\Pr(\Psi = \{1,2,3\}, \Xi = \{1\} \mid \tau_1, t) = \frac{1}{32}(1+x_1)(1-y_1)(1+x_2)(1+y_2)(1-w);$$
$$\Pr(\Psi = \{1,2,3\}, \Xi = \{2\} \mid \tau_1, t) = \frac{1}{32}(1-x_1)(1+y_1)(1-x_2)(1-y_2)(1-w);$$
$$\Pr(\Psi = \{1,2,3\}, \Xi = \{1,2\} \mid \tau_1, t) = \frac{1}{32}(1+x_1)(1+y_1)(1+x_2)(1-y_2)(1+w);$$

for site split $\{1,2,3\}$, which shows the likelihood remains unchanged if y_1 and y_2 are swapped.

For site splits $\{1,2\}$ and $\{2,3\}$ we see the following. By exchanging only x_1 with x_2, the generating probabilities and likelihood values swap between these two site splits. The same is true of the generating probabilities if we exchange only y_1 and y_2, except, for the case of the likelihood values, we again look at the complement of the most likely ancestral state split as in the case of splits $\{2\}$ and $\{1,2,3\}$. Now, if we exchange both x_1 with x_2 and y_1 with y_2, we see these generating probabilities remain unchanged, and, for the likelihood values, we look at the complement of the most likely ancestral state split and see these values also remain unchanged.

Thus exchanging x_1 with x_2 and y_1 with y_2 does not change the value of the log-likelihood $\ell_{\tau_1,t}(\tau_1, t; \boldsymbol{\xi})$. Therefore we can reduce the number of candidate likelihoods we need to search by, without loss of generality, assuming $x_2 \geq x_1$ and $y_2 \geq y_1$, with these likelihoods given in Table S3 after maximizing over ancestral state splits.

Theorems and proofs

We begin by showing an inconsistency in branch length estimation on the InvFels tree.

Theorem 1. *Let $\tau^* = \tau_1$, $t^* = \{x^*, y^*, x^*, y^*, y^*\}$, and $t = \{x_1, y_1, x_2, y_2, w\}$ with $x_1, y_1, x_2, y_2, w > 0$. For all $0 < x^*, y^* < 1$, the solution $\hat{t} := \{\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{w}\}$ given by*

$$\hat{t} = \arg\max_{t} \max_{\boldsymbol{\xi}} \ \ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi})$$

has the property $\hat{t} \neq t^*$.

Proof. For a fixed, known $\boldsymbol{\xi}$, there exists a closed form solution to $\hat{t} := \{\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{w}\}$ solving

$$\hat{t}_{\boldsymbol{\xi}} = \arg\max_t \ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi}).$$

We show in this case that the log-likelihood ℓ attains a unique maximum at $\hat{t}_{\boldsymbol{\xi}}$. For fixed $\boldsymbol{\xi}$, the log-likelihood can be decomposed into a sum of functions of each variable, i.e.,

$$\ell_{\tau^*,t^*}(\tau^*, t, \boldsymbol{\xi}) = \sum_{j=1}^{q} c_j \cdot \log h_{j,x_1}(x_1) + \sum_{j=1}^{q} c_j \cdot \log h_{j,y_1}(y_1) + \sum_{j=1}^{q} c_j \cdot \log h_{j,x_2}(x_2) + \sum_{j=1}^{q} c_j \cdot \log h_{j,y_2}(y_2) + \sum_{j=1}^{q} c_j \cdot \log h_{j,w}(w).$$

Due to this additive form, all off-diagonal terms of the Hessian for this function are zero, so we show that the diagonal terms are nonpositive. Without loss of generality we focus on the variable x_1 and the log-likelihood proportional to

$$\ell(x_1) = \sum_{j=1}^{q} c_j \cdot \log h_{j,x_1}(x_1).$$

Doing calculation as in Figure S1, each functional form, suppressing constants with respect to x_1 and the initial $1/32$ constant, is

$$h_{j,x_1}(x_1) \propto (1 + x_1)^{e_j}(1 - x_1)^{1-e_j}$$

for $e_j \in \{0, 1\}$, which, simplifying, results in

$$\ell(x_1) \propto \left(\sum_{j=1}^{q} c_j e_j\right) \log(1 + x_1) + \left(\sum_{j=1}^{q} c_j (1 - e_j)\right) \log(1 - x_1) \tag{9}$$

$$= \left(\sum_{j=1}^{q} c_j e_j\right) \log(1 + x_1) + \left(1 - \sum_{j=1}^{q} c_j e_j\right) \log(1 - x_1), \tag{10}$$

which has second derivative

$$\ell''(x_1) = -\left(\frac{\sum_j c_j e_j}{(1+x_1)^2} + \frac{1-\sum_j c_j e_j}{(1-x_1)^2}\right).$$

As $x_1 \in (0,1]$, we need only $0 \leq \sum_j c_j e_j \leq 1$ to imply the diagonal terms of the Hessian are nonpositive. Since $\sum_j c_j = 1$ and $e_j \in \{0,1\}$, then $0 \leq \sum_j c_j e_j \leq 1$ and $\ell''(x_1) \leq 0$. Applying similar arguments to the other variables, the Hessian for the log-likelihood has nonpositive diagonal terms and off-diagonal terms equal to zero, and $\hat{t}$ uniquely maximizes ℓ.

Now, by straightforward calculus, we solve for the unique maximum $\hat{x}_1$ by setting the first derivative of (10) to zero to obtain

$$\hat{x}_1 = 2 \cdot \left(\sum_{j=1}^{q} c_j e_j\right) - 1$$

where

$$\sum_{j=1}^{q} c_j e_j = \sum_{j=1}^{q} \mathbf{1}\{\text{site split } j \text{ has term } (1+x_1)\} \cdot p_{\tilde{y}_j}.$$

As an example, Table S4 shows the maximal ancestral state splits and corresponding likelihood values for $\boldsymbol{\xi}_0 = [\emptyset]_{j=1}^{q}$. In this case,

$$\sum_{j=1}^{q} c_j e_j = p_\emptyset + p_2 + p_3 + p_{23} = \frac{1}{2} + \frac{1}{2}x^*(y^*)^2$$

and $\hat{x}_1 = x^*(y^*)^2$.

We show that solutions of this form never obtain $\hat{t} = t^*$ except in cases of zero or infinite branch length. Given Table S1, all solutions to $\hat{x}_1$ have the form

$$\hat{x}_1 = a_{x_1,0} + a_{x_1,1}(x^*)^2 + a_{x_1,2}(y^*)^2 + a_{x_1,3}x^*(y^*)^2 + a_{x_1,4}(x^*)^2(y^*)^2.$$

where $a_{x_1,k}$ are constants independent of x^* and y^*—in fact, $a_{x_1,k}$ takes values in the set $\{i/8 : i = -4, -3, \ldots, 7, 8\}$. The true branch fidelity for

x_1 is x^*, and the only cases to possibly obtain $\hat{x}_1 = x^*$ are when $y^* = 1$ or when $(x^*)^2 = x^*$, i.e., one of the generating branch parameters is zero or infinite length; the same is true for x_2. A similar argument for y_1, y_2, and w shows that estimates can only be consistent when $(y^*)^2 = y^*$, i.e., $y^* = 0$ or $y^* = 1$. □

We now proceed to show there exist x^* and y^* such that the interior branch parameter w is estimated as exactly one, indicating convergence to a multifurcating topology.

Theorem 2. *Let $\tau^* = \tau_1$, $t^* = \{x^*, y^*, x^*, y^*, y^*\}$, and $t = \{x_1, y_1, x_2, y_2, w\}$ with $x_1, y_1, x_2, y_2, w > 0$. There exists an open set of $0 < x^*, y^* < 1$ such that the solution $\hat{t} := \{\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{w}\}$ given by*

$$\hat{t} = \arg\max_{t} \max_{\boldsymbol{\xi}} \ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi})$$

has the property $\hat{w} \equiv 1$.

Proof. As we have a closed form solution to our likelihood problem, we compute the optimal solution given Table S2. Let

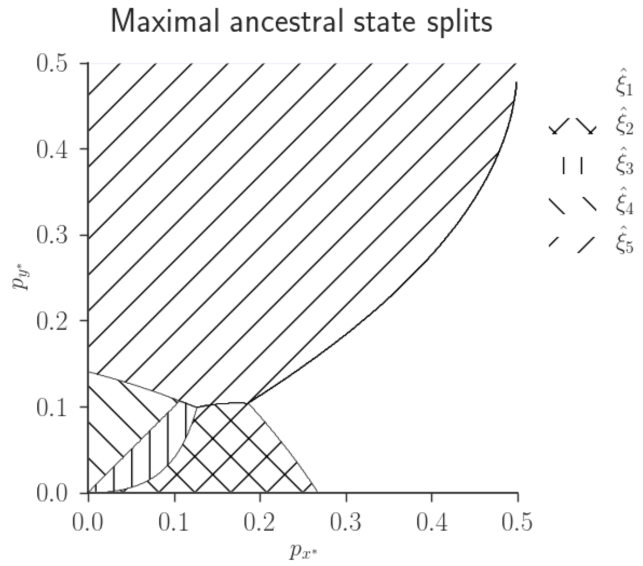
$$\hat{t}_{\boldsymbol{\xi}} = \operatorname*{argmax}_{t} \ \ell_{\tau^*,t^*}(\tau, t; \boldsymbol{\xi}).$$

be the closed form solution for t for a fixed maximal ancestral state split $\boldsymbol{\xi}$. We need only consider the possibilities for choices of ancestral state splits in Table S3 as opposed to Table S2. Upon excluding cases of infinite branch lengths (i.e., any of x_1, y_1, x_2, y_2, w equal to zero) and the redundant cases of $x_1 > x_2$ and $y_1 > y_2$, we obtain

$$\hat{\boldsymbol{\xi}} = \operatorname*{argmax}_{\boldsymbol{\xi}} \ \ell_{\tau^*,t^*}(\tau_1, \hat{t}_{\boldsymbol{\xi}}; \boldsymbol{\xi}).$$

We show the maximal ancestral states in Fig. S2.

Mapping each maximal ancestral state split to each likelihood value, we see that $\hat{w} \equiv 1$ if $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_1$ or $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_2$, which encompasses the bottom-right region of Figure S2. □



$\hat{\boldsymbol{\xi}}_1$	$\{\emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset, \emptyset\}$
$\hat{\boldsymbol{\xi}}_2$	$\{\emptyset, \emptyset, \emptyset, \emptyset, \{1,2\}, \emptyset, \emptyset, \emptyset\}$
$\hat{\boldsymbol{\xi}}_3$	$\{\emptyset, \emptyset, \emptyset, \emptyset, \{1,2\}, \emptyset, \emptyset, \{1\}\}$
$\hat{\boldsymbol{\xi}}_4$	$\{\emptyset, \emptyset, \emptyset, \emptyset, \{1,2\}, \emptyset, \{1,2\}, \{1\}\}$
$\hat{\boldsymbol{\xi}}_5$	$\{\emptyset, \emptyset, \emptyset, \{1\}, \{1\}, \emptyset, \{1\}, \{1\}\}$

Figure S2: Regions of maximal ancestral state splits on the InvFels tree τ_1.

The regions in Fig. S2 are analytically-derived regions of inconsistency in terms of probabilities of a character change along a branch for “perfect” data generated on the InvFels topology (Fig. 1) with $p_{w^*} = p_{y^*}$ (in terms of fidelities, $w^* = y^*$). As the region of degeneracy in Fig. S2 gives the values of x^* and y^* where $\hat{w}$ is guaranteed to be one, we converge on a multifurcating topology in these cases. It is easy to see that when $\emptyset$ is the maximal ancestral state split, we have the same log-likelihood for τ_1 and τ_2. Moreover, if $w = 1$, the internal branch becomes zero-length and the two topologies are indistinguishable. Let $\mathcal{T}_0$ be such that, for $t^* = \{x^*, y^*, x^*, y^*, y^*\}$, $t^* \in \mathcal{T}_0$ corresponds to x^* and y^* falling in the region in Fig. S2 where $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\xi}}_1$. We can see this results in the likelihood of both topologies being equal, i.e.,

$$\begin{aligned}
&\max_{t:t^*\in\mathcal{T}_0} \ell_{\tau^*,t^*}(\tau_1, t; \boldsymbol{\xi}) \\
&\qquad = \max_{t:\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_1, w=1, \tau=\tau_1} \Pr(\Psi = \tilde{y}_j \mid \tau^*, t^*) \cdot \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, \{x_1, y_1, x_2, y_2, w\}) \\
&\qquad = \max_{t:\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_1, w=1, \tau=\tau_2} \Pr(\Psi = \tilde{y}_j \mid \tau^*, t^*) \cdot \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau, \{x_1, y_1, x_2, y_2, w\}).
\end{aligned}$$

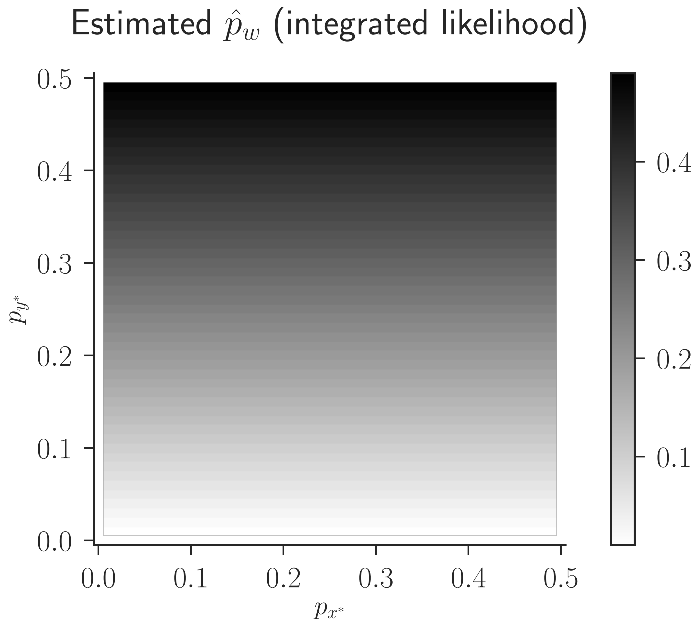


Figure S3: Estimates for $\hat{p}_w$ when computing $(\hat{x}_1, \hat{y}_1, \hat{x}_2, \hat{y}_2, \hat{w})$ using L-BFGS-B optimizing the classical integrated likelihood (2) rather than a joint optimization procedure.

InvFels tree $\tau = \tau^*$, $t^* = \{x^*, y^*, x^*, y^*, y^*\}$

$\tilde{y}_j$	$p_{\tilde{y}_j}$	$8 \cdot \Pr(\Psi = \tilde{y}_j \mid \tau, t)$
$\emptyset$	$p_\emptyset$	$1 + (x^*)^2 + (y^*)^2 + 4x^*(y^*)^2 + (x^*)^2(y^*)^2$
$\{1\}$	p_1	$1 - (x^*)^2 + (y^*)^2 - (x^*)^2(y^*)^2$
$\{2\}$	p_2	$1 + (x^*)^2 - (y^*)^2 - (x^*)^2(y^*)^2$
$\{3\}$	p_3	$1 - (x^*)^2 + (y^*)^2 - (x^*)^2(y^*)^2$
$\{1, 2, 3\}$	p_{123}	$1 + (x^*)^2 - (y^*)^2 - (x^*)^2(y^*)^2$
$\{1, 2\}$	p_{12}	$1 - (x^*)^2 - (y^*)^2 + (x^*)^2(y^*)^2$
$\{2, 3\}$	p_{23}	$1 - (x^*)^2 - (y^*)^2 + (x^*)^2(y^*)^2$
$\{1, 3\}$	p_{13}	$1 + (x^*)^2 + (y^*)^2 - 4x^*(y^*)^2 + (x^*)^2(y^*)^2$

InvFels tree $\tau = \tau_1$, $t = \{x_1, y_1, x_2, y_2, w\}$

$\tilde{y}_j$	$p_{\tilde{y}_j}$	$8 \cdot \Pr(\Psi = \tilde{y}_j \mid \tau, t)$
$\emptyset$	$p_\emptyset$	$1 + x_1x_2 + y_1y_2 + w[x_1 + x_2][y_1 + y_2] + x_1y_1x_2y_2$
$\{1\}$	p_1	$1 - x_1x_2 + y_1y_2 + w[-x_1 + x_2][y_1 + y_2] - x_1y_1x_2y_2$
$\{2\}$	p_2	$1 + x_1x_2 - y_1y_2 + w[x_1 + x_2][-y_1 + y_2] - x_1y_1x_2y_2$
$\{3\}$	p_3	$1 - x_1x_2 + y_1y_2 + w[x_1 - x_2][y_1 + y_2] - x_1y_1x_2y_2$
$\{1, 2, 3\}$	p_{123}	$1 + x_1x_2 - y_1y_2 + w[x_1 + x_2][y_1 - y_2] - x_1y_1x_2y_2$
$\{1, 2\}$	p_{12}	$1 - x_1x_2 - y_1y_2 + w[-x_1 + x_2][-y_1 + y_2] + x_1y_1x_2y_2$
$\{2, 3\}$	p_{23}	$1 - x_1x_2 - y_1y_2 + w[x_1 - x_2][-y_1 + y_2] + x_1y_1x_2y_2$
$\{1, 3\}$	p_{13}	$1 + x_1x_2 + y_1y_2 + w[-x_1 - x_2][y_1 + y_2] + x_1y_1x_2y_2$

Felsenstein tree $\tau = \tau_2$, $t = \{x_1, y_1, x_2, y_2, w\}$

$\tilde{y}_j$	$p_{\tilde{y}_j}$	$8 \cdot \Pr(\Psi = \tilde{y}_j \mid \tau, t)$
$\emptyset$	$p_\emptyset$	$1 + x_1y_1 + x_2y_2 + w[x_1 + y_1][x_2 + y_2] + x_1y_1x_2y_2$
$\{1\}$	p_1	$1 - x_1y_1 + x_2y_2 + w[-x_1 + y_1][x_2 + y_2] - x_1y_1x_2y_2$
$\{2\}$	p_2	$1 - x_1y_1 + x_2y_2 + w[x_1 - y_1][x_2 + y_2] - x_1y_1x_2y_2$
$\{3\}$	p_3	$1 + x_1y_1 - x_2y_2 + w[x_1 + y_1][-x_2 + y_2] - x_1y_1x_2y_2$
$\{1, 2, 3\}$	p_{123}	$1 + x_1y_1 - x_2y_2 + w[-x_1 - y_1][-x_2 + y_2] - x_1y_1x_2y_2$
$\{1, 2\}$	p_{12}	$1 + x_1y_1 + x_2y_2 + w[-x_1 - y_1][x_2 + y_2] + x_1y_1x_2y_2$
$\{2, 3\}$	p_{23}	$1 - x_1y_1 - x_2y_2 + w[x_1 - y_1][-x_2 + y_2] + x_1y_1x_2y_2$
$\{1, 3\}$	p_{13}	$1 - x_1y_1 - x_2y_2 + w[-x_1 + y_1][-x_2 + y_2] + x_1y_1x_2y_2$

Table S1: 8 times the site split probabilities $p_{\tilde{y}_j}$ on the true InvFels tree τ^* with $t^* = \{x^*, y^*, x^*, y^*, y^*\}$, and on the InvFels tree τ_1 and Felsenstein tree τ_2 with $t = \{x_1, y_1, x_2, y_2, w\}$ obtained using the Hadamard transform.

$\tilde{y}_j$	$\tilde{h}_k$	$32 \cdot \Pr(\Psi = \tilde{y}_j, \Xi = \tilde{h}_k \mid \tau_1, t)$
$\emptyset$	$\emptyset$	$(1+x_1)(1+y_1)(1+x_2)(1+y_2)(1+w)$
	$\{1\}^*$	$(1-x_1)(1+y_1)(1-x_2)(1+y_2)(1-w)$
	$\{2\}^*$	$(1+x_1)(1-y_1)(1+x_2)(1-y_2)(1-w)$
	$\{1,2\}^*$	$(1-x_1)(1-y_1)(1-x_2)(1-y_2)(1+w)$
$\{1\}$	$\emptyset$	$(1-x_1)(1+y_1)(1+x_2)(1+y_2)(1+w)$
	$\{1\}$	$(1+x_1)(1+y_1)(1-x_2)(1+y_2)(1-w)$
	$\{2\}^*$	$(1-x_1)(1-y_1)(1+x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1+x_1)(1-y_1)(1-x_2)(1-y_2)(1+w)$
$\{2\}$	$\emptyset$	$(1+x_1)(1-y_1)(1+x_2)(1+y_2)(1+w)$
	$\{1\}^*$	$(1-x_1)(1-y_1)(1-x_2)(1+y_2)(1-w)$
	$\{2\}$	$(1+x_1)(1+y_1)(1+x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1-x_1)(1+y_1)(1-x_2)(1-y_2)(1+w)$
$\{3\}$	$\emptyset$	$(1+x_1)(1+y_1)(1-x_2)(1+y_2)(1+w)$
	$\{1\}$	$(1-x_1)(1+y_1)(1+x_2)(1+y_2)(1-w)$
	$\{2\}^*$	$(1+x_1)(1-y_1)(1-x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1-x_1)(1-y_1)(1+x_2)(1-y_2)(1+w)$
$\{1,2,3\}$	$\emptyset$	$(1-x_1)(1-y_1)(1-x_2)(1+y_2)(1+w)$
	$\{1\}$	$(1+x_1)(1-y_1)(1+x_2)(1+y_2)(1-w)$
	$\{2\}^*$	$(1-x_1)(1+y_1)(1-x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1+x_1)(1+y_1)(1+x_2)(1-y_2)(1+w)$
$\{1,2\}$	$\emptyset$	$(1-x_1)(1-y_1)(1+x_2)(1+y_2)(1+w)$
	$\{1\}$	$(1+x_1)(1-y_1)(1-x_2)(1+y_2)(1-w)$
	$\{2\}$	$(1-x_1)(1+y_1)(1+x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1+x_1)(1+y_1)(1-x_2)(1-y_2)(1+w)$
$\{2,3\}$	$\emptyset$	$(1+x_1)(1-y_1)(1-x_2)(1+y_2)(1+w)$
	$\{1\}$	$(1-x_1)(1-y_1)(1+x_2)(1+y_2)(1-w)$
	$\{2\}$	$(1+x_1)(1+y_1)(1-x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1-x_1)(1+y_1)(1+x_2)(1-y_2)(1+w)$
$\{1,3\}$	$\emptyset$	$(1-x_1)(1+y_1)(1-x_2)(1+y_2)(1+w)$
	$\{1\}$	$(1+x_1)(1+y_1)(1+x_2)(1+y_2)(1-w)$
	$\{2\}^*$	$(1-x_1)(1-y_1)(1-x_2)(1-y_2)(1-w)$
	$\{1,2\}$	$(1+x_1)(1-y_1)(1+x_2)(1-y_2)(1+w)$

Table S2: 32 times likelihood values for all site splits $\tilde{y}_j$ and ancestral state splits $\tilde{h}_k$ of the InvFels tree τ_1. Ancestral states with * are never maximal provided parameters are in $(0,1]$. By combinations of $\tilde{h}_k$, there are $3^5 \cdot 4^2 = 3,888$ possible forms for the likelihood.

$\tilde{y}_j$	$\eta_j(\tau_1, t)$	ξ_j	$32 \cdot \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau_1, t)$
$\emptyset$	$\{\emptyset\}$	$\emptyset$	$(1+x_1)(1+y_1)(1+x_2)(1+y_2)(1+w)$
$\{1\}$	$\{\emptyset\}$	$\emptyset$	$(1-x_1)(1+y_1)(1+x_2)(1+y_2)(1+w)$
$\{2\}$	$\{\emptyset\}$	$\emptyset$	$(1+x_1)(1-y_1)(1+x_2)(1+y_2)(1+w)$
$\{3\}$	$\{\emptyset, \{1\}, \{1,2\}\}$	$\emptyset$	$(1+x_1)(1+y_1)(1-x_2)(1+y_2)(1+w)$
		$\{1\}$	$(1-x_1)(1+y_1)(1+x_2)(1+y_2)(1-w)$
		$\{1,2\}$	$(1-x_1)(1-y_1)(1+x_2)(1-y_2)(1+w)$
$\{1,2,3\}$	$\{\emptyset, \{1\}, \{1,2\}\}$	$\emptyset$	$(1-x_1)(1-y_1)(1-x_2)(1+y_2)(1+w)$
		$\{1\}$	$(1+x_1)(1-y_1)(1+x_2)(1+y_2)(1-w)$
		$\{1,2\}$	$(1+x_1)(1+y_1)(1+x_2)(1-y_2)(1+w)$
$\{1,2\}$	$\{\emptyset\}$	$\emptyset$	$(1-x_1)(1-y_1)(1+x_2)(1+y_2)(1+w)$
$\{2,3\}$	$\{\emptyset, \{1\}, \{1,2\}\}$	$\emptyset$	$(1+x_1)(1-y_1)(1-x_2)(1+y_2)(1+w)$
		$\{1\}$	$(1-x_1)(1-y_1)(1+x_2)(1+y_2)(1-w)$
		$\{1,2\}$	$(1-x_1)(1+y_1)(1+x_2)(1-y_2)(1+w)$
$\{1,3\}$	$\{\emptyset, \{1\}, \{1,2\}\}$	$\emptyset$	$(1-x_1)(1+y_1)(1-x_2)(1+y_2)(1+w)$
		$\{1\}$	$(1+x_1)(1+y_1)(1+x_2)(1+y_2)(1-w)$
		$\{1,2\}$	$(1+x_1)(1-y_1)(1+x_2)(1-y_2)(1+w)$

Table S3: 32 times likelihood values on the InvFels tree τ_1. Due to the symmetry of the likelihood, WLOG we let $x_2 \geq x_1$ and $y_2 \geq y_1$ and maximize over ancestral state splits to reduce the number of possible functional forms to consider. Likelihoods with multiple entries have maxima determined by unknown branch length parameters. Because in 4 cases there are 3 possibilities for ξ_j, there are $3^4 = 81$ possible forms for the likelihood.

$\tilde{y}_j$	$\eta_j(\tau_1, t)$	ξ_j	$32 \cdot \Pr(\Psi = \tilde{y}_j, \Xi = \xi_j \mid \tau_1, t)$
$\emptyset$	$\{\emptyset\}$	$\emptyset$	$(1 + x_1)(1 + y_1)(1 + x_2)(1 + y_2)(1 + w)$
$\{1\}$	$\{\emptyset\}$	$\emptyset$	$(1 - x_1)(1 + y_1)(1 + x_2)(1 + y_2)(1 + w)$
$\{2\}$	$\{\emptyset\}$	$\emptyset$	$(1 + x_1)(1 - y_1)(1 + x_2)(1 + y_2)(1 + w)$
$\{3\}$	$\{\emptyset, \{1\}, \{1, 2\}\}$	$\emptyset$	$(1 + x_1)(1 + y_1)(1 - x_2)(1 + y_2)(1 + w)$
$\{1, 2, 3\}$	$\{\emptyset, \{1\}, \{1, 2\}\}$	$\emptyset$	$(1 - x_1)(1 - y_1)(1 - x_2)(1 + y_2)(1 + w)$
$\{1, 2\}$	$\{\emptyset\}$	$\emptyset$	$(1 - x_1)(1 - y_1)(1 + x_2)(1 + y_2)(1 + w)$
$\{2, 3\}$	$\{\emptyset, \{1\}, \{1, 2\}\}$	$\emptyset$	$(1 + x_1)(1 - y_1)(1 - x_2)(1 + y_2)(1 + w)$
$\{1, 3\}$	$\{\emptyset, \{1\}, \{1, 2\}\}$	$\emptyset$	$(1 - x_1)(1 + y_1)(1 - x_2)(1 + y_2)(1 + w)$

Table S4: 32 times the maximal likelihood values on the InvFels tree τ_1 where $\emptyset$ is the most likely ancestral state split for each site split.