Phylogenetic mixtures on a single tree can mimic a tree of another topology

Frederick A. Matsen and Mike Steel
Biomathematics Research Centre
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Corresponding Author:
Frederick A. Matsen
phone: +64 3 364 2987 x7431
fax: +64 3 364 2587
email: ematsen@gmail.com

Keywords: Phylogenetics; Mixture Model; Sequence Evolution; Model Identifiability

Abstract

Phylogenetic mixtures model the inhomogeneous molecular evolution commonly observed in data. The performance of phylogenetic reconstruction methods where the underlying data is generated by a mixture model has stimulated considerable recent debate. Much of the controversy stems from simulations of mixture model data on a given tree topology for which reconstruction algorithms output a tree of a different topology; these findings were held up to show the shortcomings of particular tree reconstruction methods. In so doing, the underlying assumption was that mixture model data on one topology can be distinguished from data evolved on an unmixed tree of another topology given enough data and the "correct" method. Here we show that this assumption can be false. For biologists our results imply that, for example, the combined data from two genes whose phylogenetic trees differ only in terms of branch lengths can perfectly fit a tree of a different topology.

It is now well known that molecular evolution is heterogenous across time and position (Simon et al., 1996). A classic example is that of stems and loops of ribosomal RNA: the evolution of one side of a stem is strongly constrained to match the complimentary side, whereas for loops different constraints exist (Springer and Douzery, 1996). Heterogenous evolution between genes is also widespread, where even the general features of evolutionary history for neighboring genes may differ wildly (Ochman et al., 2000). Presently it is not uncommon to use concatenated sequence data from many genes for phylogenetic inference (Phillips et al., 2004), which can lead to very high levels of apparent heterogeneity (Baldauf et al., 2000). Furthermore, empirical evidence using the covarion model shows that sometimes more subtle partitions of the data can exist, for which separate analysis is difficult (H-C Wang and Roger, 2007).

This heterogeneity is typically formulated as a mixture model (Pagel and Meade, 2004). Mathematically, a phylogenetic mixture model is simply a weighted average of site pattern frequencies derived from a number of phylogenetic trees, which may be of the same or different topologies. Even though many phylogenetics programs accept aligned sequences as input, the only data actually used in the vast majority of phylogenetic algorithms is the derived site pattern frequencies. Thus, in these algorithms, any record of position is lost and inhomogeneous evolution appears identical to homogeneous evolution under an appropriate phylogenetic mixture model. For simplicity, we call a mixture of site pattern frequencies from two trees (which

may be of the same or different topologies) a *mixture of two trees*; when the two trees have the same underlying topology, the mixture will be called a *mixture of branch length sets on a tree*.

Mixture models have proven difficult for phylogenetic reconstruction methods, which have historically sought to find a single process explaining the data. For example, it has been shown that mixtures of two different tree topologies can mislead MCMC-based tree reconstruction (Mossel and Vigoda, 2005). It is also known that there exist mixtures of branch lengths on one tree which are indistinguishable from mixtures of branch lengths on a tree of a different topology (Steel et al., 1994; Štefankovič and Vigoda, 2007a,b). Recently, simulations of mixture models from "heterotachous" (changing through time) evolution have been shown to cause reconstruction methods to fail (Ruano-Rubio and Fares, 2007).

The motivation for our work is the observation that both theory and simulations have shown that in certain parameter regimes, phylogenetic reconstruction methods return a tree topology different from the one used to generate the mixture data. The parameter regime in this class of examples is similar to that shown in Figure 1, with two neighboring pendant edges which alternate being long and short. After mixing and reconstruction, these edges may no longer be adjacent on the reconstructed tree. We call this "mixed branch repulsion." This phenomenon has been observed extensively in simulation (Kolaczkowski and Thornton, 2004; Spencer et al., 2005; Philippe et al., 2005; Gadagkar and Kumar, 2005) and it has been proved that certain

distance and maximum likelihood methods are susceptible to this effect (Chang, 1996; Štefankovič and Vigoda, 2007a,b). Up to this point such results have been interpreted as pathological behavior of the reconstruction algorithms, which has led to a heated debate about which reconstruction methods perform best in this situation (Steel, 2005; Thornton and Kolaczkowski, 2005). Implicit in this debate is the assumption that a mixture of trees on one topology gives different site pattern frequencies than that of an unmixed tree of a different topology. This leads to the natural question of how similar these two site pattern frequencies can be.

Here we demonstrate that mixtures of two sets of branch lengths on a tree of one topology can exactly mimic the site pattern frequencies of a tree of a different topology under the two-state symmetric model. In fact, there is a precisely characterizable region of parameter space (of codimension two) where such mixtures exist. Consider two quartet trees of topology 12|34, as shown in Figure 1. Label the pendant branches 1 through 4 according to the taxon labels, and label the internal edge with 5. The first branch length set will be written $t_1, \ldots, t_5$ and the second $s_1, \ldots, s_5$. Now, let $k_1, \ldots, k_4$ satisfy the following system of inequalities:

$$k_1 > k_3 > k_4 > 1 > k_2,$$

$$\tfrac{1-k_1^2}{k_1}\tfrac{1-k_4^2}{k_4} + \tfrac{1-k_2^2}{k_2}\tfrac{1-k_3^2}{k_3} > 0,$$

$$\tfrac{k_1+k_4}{1+k_1k_4} \cdot \tfrac{k_2+k_3}{1+k_2k_3} > 1.$$

Then there exist nonzero internal branch lengths t_5 and s_5, mixing weights, and positive numbers $\ell_1, \ldots, \ell_4$ such that if for $i = 1, \ldots, 4$, $k_i = \exp(-2(t_i - s_i))$ and $t_i \geq \ell_i$, the corresponding mixture of two 12|34 trees will have the same site pattern frequencies as a single tree of the 13|24 topology. A more precise statement of this proposition along with proof is provided in the Appendix. We have illustrated two examples of branch length sets satisfying these criteria in Figure 1. Although the parameter regime including these examples is of dimension two less than the ambient space, examples which are close to satisfying the above system of inequalities will have pattern probabilities which are effectively indistinguishable to those for a tree of a different topology.

We believe that this similarity between site pattern frequencies generated by mixtures of branch lengths on one tree and corresponding frequencies on a different tree is what is leading to the mixed branch repulsion observed in theory and simulation. Furthermore, it is possible that even the simple case presented here is directly relevant to reconstructions from data. First, it is not uncommon to simplify the genetic code from the four standard bases to two (pyrimidines versus purines) in order to reduce the effect of compositional bias when working with genome-scale data on deep phylogenetic relationships (Phillips et al., 2004). Second, when working on such relationships concatenation of genes is common (Baldauf et al., 2000), for which a phylogenetic mixture is the expected result. Finally, the region of parameter space bringing about mixed branch repulsion may become more

extensive as the number of concatenated genes increases. Therefore in concatenated gene analysis it may be worthwhile considering incongruence in terms of branch lengths and not just in terms of topology (Rokas et al., 2003; Jeffroy et al., 2006), as highly incongruent branch lengths may produce artifactual results upon concatenation.

Mixed branch repulsion may be more difficult to detect than the usual model mis-specification issues; in the cases presented here the mis-specified single tree model fits the data perfectly. In contrast, although using the wrong mutation model for reconstruction using maximum likelihood can lead to incorrect tree topologies (Goremykin et al., 2005), the resulting model mis-specification can be seen in the resulting poor likelihood score. In the mixtures presented here, there is no way of telling when one is in the mixed regime on one topology or an unmixed regime on another topology. Furthermore, model selection techniques such as the Akaike Information Criterion (Posada and Buckley, 2004) which penalize parameter-rich models would, in this case, choose a simple unmixed model and thereby select a tree that is different from the historically correct tree if the true process was generated by a mixture model.

The derivation of the zone resulting in mixed branch repulsion is a conceptually simple application of the two pillars of theoretical phylogenetics: the Hadamard transform and phylogenetic invariants (Semple and Steel, 2003; Felsenstein, 2004). The Hadamard transform is a closed form invertible transformation (expressed in terms of the discrete Fourier transform) for

gaining the expected site pattern frequencies from the branch lengths and topology of a tree or vice versa. Phylogenetic invariants characterize when a set of site pattern frequencies could be the expected site pattern frequencies for a tree of a given topology. They are identities in terms of the discrete Fourier transform of the site pattern frequencies. Therefore, to derive the above equations, we simply insert the Hadamard formulae for the Fourier transform of pattern probabilities into the phylogenetic invariants, then check to make sure the resulting branch lengths are positive.

Similar considerations lead to an understanding of when it is possible to mix two branch length sets on a tree to reproduce the site pattern frequencies of a tree of the same topology (see Appendix for details). For a quartet, either two neighboring pendant branch lengths must be equal between the two branch length sets of the mixture, or the sum of one pair of neighboring pendant branch lengths and the difference of the other pair must be equal. For trees larger than quartets, the allowable mixtures are determined by these restrictions on the quartets (results to appear elsewhere). For pairs of branch lengths satisfying these criteria, any choice of mixing weights will produce site pattern frequencies satisfying the phylogenetic invariants.

Intuitively, one might expect that when two sets of branch lengths mix to mimic a tree of the same topology, some sort of averaging property would hold for the branch lengths. However, this need not be the case, as demonstrated by Figure 2. In fact, it is possible to mix two sets of branch lengths on a tree to mimic a tree of the same topology such that a resulting

pendant branch length is arbitrarily small while the corresponding branch length in either of the branch length sets being mixed stays above some arbitrarily large fixed value.

The results in this paper shed some light on the geometry of phylogenetic mixtures (Kim, 2000). As is well known, the set of phylogenetic trees of a given topology forms a compact subvariety of the space of site pattern frequencies (Sturmfels and Sullivant, 2005). The first part of our work demonstrates that there are pairs of points in one such subvariety such that a line between those two points intersects a distinct subvariety (see Figure 3). Therefore the convex hull of one subvariety has a region of intersection with distinct subvarieties. This is stronger than the recently derived result by Štefankovič and Vigoda (2007a,b) that the convex hulls of the varieties intersect. The second part of our work shows that there exist pairs of points in a subvariety such that the line between those points intersects the subvariety. Furthermore, it demonstrates that when such a line between two points intersects the subvariety in a third point, then a subinterval of the line is contained in the subvariety.

This geometric perspective can aid in understanding practical problems of phylogenetic estimation under maximum likelihood. The question of when maximum-likelihood selects the "wrong" topology given mixture data was initiated by Chang (1996) who found a one-parameter space of such examples and recently continued by Štefankovič and Vigoda (2007a) who found a two-parameter space of such examples. Our results show that an

eleven-parameter space of such examples exist, which is the most possible for a mixture of quartet trees.

To demonstrate this last fact, let $\delta_{\mathrm{KL}}(p, q)$ be the Kullback-Leibler divergence of q from p:

$$\delta_{\mathrm{KL}}(p, q) = \sum_i p_i \log \frac{p_i}{q_i}.$$

Note that δ_{KL} is a continuous function when probability distributions p and q have no components equal to zero, i.e. sit in the interior of the probability simplex $\mathring{\Delta}$. Let $V_{12|34}$ be the space of trees of topology 12|34 considered as a compact subvariety of the space of site pattern frequencies. For $p \in \mathring{\Delta}$ and compact subvariety V, let $\delta_{\mathrm{KL}}(p, V)$ be the minimum of $\delta_{\mathrm{KL}}(p, v)$ where v ranges over V.

The above results show that mixtures of data from a tree of topology 12|34 can equal a point $m \in \mathring{\Delta}$ on the variety $V_{13|24}$. Because the parameter regime for mixtures mimicking a tree of the same topology is disjoint from that for mixtures mimicking the same topology, $\delta_{\mathrm{KL}}(V_{12|34}, m) > 0$. Now, because $m \in \mathring{\Delta}$, there exists an open ball $B_m \subset \mathring{\Delta}$ containing m. Thus δ_{KL} is continuous on B_m and because $\delta_{\mathrm{KL}}(m, V_{13|24}) = 0 < \delta_{\mathrm{KL}}(m, V_{12|34})$ there exists an open ball $B'_m \subset B_m$ such that $\delta_{\mathrm{KL}}(m', V_{13|24}) < \delta_{\mathrm{KL}}(m', V_{12|34})$ for all $m' \in B'_m$. By definition, maximum likelihood will thus choose a tree of topology 13|24 for any site pattern data $m' \in B'_m$. The pre-image of B'_m under the continuous map sending two sets of quartet branch lengths and a mixing

parameter to the resulting mixture model data is the required eleven-parameter space.

In general, we are surprised by the results in this paper given the level of development of the tools useful for solving this type of problem. Many interesting questions remain: for what other site substitution models does the phenomenon shown here present itself? What are the criteria for mixed branch repulsion for trees larger than quartets? What is the zone of branch length parameter space for which the resulting mixture is closer (in some meaningful way) to the expected site pattern frequencies of a tree of different topology than to those for a tree of the original topology?

Acknowledgments

The authors would like to thank Dennis Wong for advice on the figures. Funding for this work was provided by the Allan Wilson Centre for Molecular Ecology and Evolution, New Zealand.

References

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977.

Chang, J. T. 1996. Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134:189–215.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Press, Sunderland, MA.

Gadagkar, S. R. and S. Kumar. 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol 22:2139–2141.

Goremykin, V. V., B. Holland, K. I. Hirsch-Ernst, and F. H. Hellwig. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. Mol Biol Evol 22:1813–1822.

H-C Wang, E. S., M. Spencer and A. Roger. 2007. Testing for covarion-like evolution in protein sequences. Mol. Biol. Evol. 24:294–305.

Jeffroy, O., H. Brinkmann, F. Delsuc, and H. Philippe. 2006. Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231.

Kim, J. 2000. Slicing hyperdimensional oranges: the geometry of phylogenetic estimation. Mol Phylogenet Evol 17:58–75.

Kolaczkowski, B. and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Mossel, E. and E. Vigoda. 2005. Phylogenetic MCMC algorithms are misleading on mixtures of trees. Science 309:2207–2209.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Pagel, M. and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571–581.

Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50.

Phillips, M. J., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455–1458.

Posada, D. and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst Biol 53:793–808.

Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Ruano-Rubio, V. and M. Fares. 2007. Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. Syst. Biol. 56:68–82.

Semple, C. and M. Steel. 2003. Phylogenetics vol. 24 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford.

Simon, C., L. Nigro, J. Sullivan, K. Holsinger, A. Martin, A. Grapputo, A. Franke, and C. McIntosh. 1996. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes. Mol Biol Evol 13:923–932.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol 22:1161–1164.

Springer, M. S. and E. Douzery. 1996. Secondary structure and patterns of evolution among mammalian mitochondrial 12S rRNA molecules. J Mol Evol 43:357–373.

Steel, M. 2005. Should phylogenetic models be trying to "fit an elephant"? Trends Genet 21:307–309.

Steel, M. A., L. A. Szekely, and M. D. Hendy. 1994. Reconstructing trees when sequence sites evolve at variable rates. J Comput Biol 1:153–163.

Sturmfels, B. and S. Sullivant. 2005. Toric ideals of phylogenetic invariants. J Comput Biol 12:204–228.

Thornton, J. W. and B. Kolaczkowski. 2005. No magic pill for phylogenetic error. Trends Genet 21:310–311.

Štefankovič, D. and E. Vigoda. 2007a. Phylogeny of mixture models: Robustness of maximum likelihood and non-identifiable distributions http://arxiv.org/abs/q-bio.PE/0609038.

Štefankovič, D. and E. Vigoda. 2007b. Pitfalls of heterogeneous processes for phylogenetic reconstruction. Syst. Biol. 56:113–124.

Appendix

In this section we provide more precise statements and proofs of the propositions in the text. The proofs will be presented in the reverse order than they were stated in the main text— first the fact that it is possible to mix two branch lengths on a tree to mimic a tree of the same topology, then that it is possible to mix branch lengths to mimic a tree of a distinct topology.

As stated in the main text, the general strategy of the proofs is simple: use the Hadamard transform to calculate Fourier transforms of site pattern probabilities and then insert these formulas into the phylogenetic invariants. These steps would become very messy except for a number of simplifications: First, because the discrete Fourier transform is linear, a transform of a mixture is simply a mixture of the corresponding transforms. Second, the fact that the original trees satisfy a set of phylogenetic invariants reduces the complexity of the mixed invariants. Finally, the product of the exponentials of the branch lengths appear in all formulas, and division leads to a substantial simplification.

First we remind the reader of the main tools and fix notation. Note that for the entire paper we will be working with the two-state symmetric (also known as Cavender-Ferris-Neyman) model.

The Hadamard transform and phylogenetic invariants

For a given edge e of branch length $\gamma(e)$ we will denote

$$\theta(e) = \exp(-2\gamma(e)) \tag{1}$$

which ranges between zero and one for positive branch lengths. We call this number the "fidelity" of the edge, as it quantifies the quality of transmission of the ancestral state across the edge. For $A \subset \{1, \ldots, n\}$ of even order, let $q_A = (H_{n-1}\bar{p})_A$ be the Fourier transform of the split probabilities, where H_n is the n by n Hadamard matrix. (Semple and Steel, 2003)

Quartet trees will be designated by their splits, i.e. 13|24 refers to a quartet with taxa labeled 1 and 3 on one side of the quartet and taxa 2 and 4 on the other.

By the first identity in the proof of Theorem 8.6.3 of (Semple and Steel, 2003) one can express the Fourier transform of the split probabilities in terms of products of fidelities. That is, for any subset $A \subset \{1, \ldots, n\}$ of even order,

$$q_A = \prod_{e \in \mathcal{P}(T,A)} \theta(e) \tag{2}$$

where $\mathcal{P}(T, A)$ is the set of edges which lie in the paths connecting the taxa in A to each other. This set is uniquely defined (again, see (Semple and Steel, 2003)).

From this equation, we can derive values for the fidelities from the Fourier transforms of the split probabilities. In particular, it is simple to write out the fidelity of a pendant edge on a quartet. For example,

$$\theta_1 = \sqrt{\frac{\theta_1\theta_5\theta_4 \cdot \theta_1\theta_2}{\theta_2\theta_5\theta_4}} = \sqrt{\frac{q_{14}\ q_{12}}{q_{24}}}.$$

In general, we have the following lemma:

Lemma 1. *If a, b, and c are distinct pendant edge labels on a quartet such that a and b are adjacent, then the fidelity of a pendant edge a is*

$$\sqrt{\frac{q_{ab}\ q_{ac}}{q_{bc}}}.$$

A similar calculation leads to an analogous lemma for the internal edge:

Lemma 2. *The fidelity of the internal edge of an $ab|cd$ quartet tree is*

$$\sqrt{\frac{q_{ac}\ q_{bd}}{q_{ab}\ q_{cd}}}.$$

This paper will also make extensive use of the method of phylogenetic invariants. These are polynomial identities in the Fourier transform of the split probabilities which are satisfied for a given tree topology. Invariants are understood in a very general setting (see (Sturmfels and Sullivant, 2005)), however here we only require invariants for the simplest case: a quartet tree with the two-state symmetric model. In particular, for the quartet tree $ab|cd$, the two phylogenetic invariants are

$$q_{abcd} - q_{ab}\ q_{cd} = 0 \tag{3}$$

$$q_{ac}\ q_{bd} - q_{ad}\ q_{bc} = 0. \tag{4}$$

A set of $q.$ mimic the Fourier transforms of site pattern frequencies of a nontrivial tree exactly when they satisfy the phylogenetic invariants and have corresponding edge fidelities (given by Lemmas 1 and 2) between zero and one.

This paper is primarily concerned with the following situation: a mixture of two sets of branch lengths on a quartet tree which mimics the site pattern frequencies of an unmixed tree. We fix the following notation: the two branch length sets will be called t_i and s_i, the corresponding fidelities will be called θ_i and ψ_i, and the Fourier transforms of the site pattern frequencies will be called $q.$ and $r.$, respectively. The internal edge of the quartet will carry the label $i = 5$, and the pendant edges are labeled according to their terminal taxa (e.g. $i = 2$ is the edge terminating in the second taxon). The mixing weight will be written α, and we make the convention that the mixture will take the t_i branch length set with probability α time and s_i with probability $1 - \alpha$.

Mixtures mimicking a tree of the same topology

In this section we describe conditions on mixtures such that a nontrivial mixture of two branch lengths on 12|34 can give the same probability distribution as a single tree of the same topology.

Mixing two branch length sets on a 12|34 quartet tree with the above notation leads to the following form of invariant (3) for a resulting tree also of topology 12|34:

$$\begin{aligned}(\alpha + 1 - \alpha)(\alpha\, q_{1234} + (1-\alpha)\, r_{1234}) - \\ (\alpha\, q_{12} + (1-\alpha)\, r_{12})(\alpha\, q_{34} + (1-\alpha)\, r_{34}) = 0.\end{aligned} \tag{5}$$

Multiplying out terms then collecting, there will be a $\alpha^2(q_{1234} - q_{12}q_{34})$ term which is zero by the phylogenetic invariants for the 12|34 topology. Similarly, the terms with $(1-\alpha)^2$ vanish. Dividing by $\alpha\,(1-\alpha)$ which we assume to be nonzero, equation (5) becomes

$$q_{1234} + r_{1234} - (q_{12}r_{34} + r_{12}q_{34}) = 0.$$

Applying invariant (3) for the 12|34 topology and simplifying leads to the following equivalent form of (5):

$$(q_{12} - r_{12})(q_{34} - r_{34}) = 0. \tag{6}$$

The same sorts of moves lead to the second invariant of the mixed tree:

$$q_{13}r_{24} + r_{13}q_{24} - (q_{14}r_{23} + r_{14}q_{23}) = 0. \tag{7}$$

The fact that α doesn't appear in these equations already delivers an interesting fact: if a mixture of two branch lengths in this setting satisfy the phylogenetic invariants for a single α, then they do so for all α. Geometrically, this means if the line between two points on the subvariety cut out by the phylogenetic invariants intersects the subvariety non trivially then it sits entirely in the subvariety.

We can gain more insight by considering these equations in terms of

fidelities. Direct substitution using (2) into (6) gives

$$(\theta_1\theta_2 - \psi_1\psi_2)(\theta_3\theta_4 - \psi_3\psi_4) = 0.$$

This equation will be satisfied exactly when the branch lengths satisfy

$$t_1 + t_2 = s_1 + s_2 \quad \text{or} \quad t_3 + t_4 = s_3 + s_4. \tag{8}$$

The corresponding substitution into (7) and then division by $\theta_2\theta_5\theta_4\psi_2\psi_5\psi_4$ gives after simplification

$$\left(\frac{\theta_1}{\theta_2} - \frac{\psi_1}{\psi_2}\right)\left(\frac{\theta_3}{\theta_4} - \frac{\psi_3}{\psi_4}\right) = 0$$

This equation will be satisfied exactly when the branch lengths satisfy

$$t_1 - t_2 = s_1 - s_2 \quad \text{or} \quad t_3 - t_4 = s_3 - s_4. \tag{9}$$

To summarize,

Proposition 3. *The mixture of two* $12|34$ *quartet trees with pendant branch lengths* t_i *and* s_i *satisfies the* $12|34$ *phylogenetic invariants for the binary symmetric model exactly (up to renumbering) when either* $t_1 = s_1$ *and* $t_2 = s_2$*, or* $t_1 + t_2 = s_1 + s_2$ *and* $t_3 - t_4 = s_3 - s_4$*.*

As described above this proposition makes no reference to the mixing weight α.

In quartets where $t_1 = s_1$ and $t_2 = s_2$, the resulting tree will also have pendant branch lengths t_1 and t_2:

Proposition 4. *A mixture of two* $12|34$ *quartet trees with branch lengths* t_i *and* s_i *which satisfies* $t_1 = s_1$ *and* $t_2 = s_2$ *will have resulting branch lengths for the first and second taxa equal to* t_1 *and* t_2*, respectively.*

Proof. Let the fidelity of the edges leading to taxon one and two be denoted μ_1 and μ_2. We have by Lemma 1 with $a = 1$, $b = 2$ and $c = 3$,

$$\mu_1 = \sqrt{\frac{(\alpha\theta_1\theta_2 + (1-\alpha)\psi_1\psi_2) \cdot (\alpha\theta_1\theta_5\theta_3 + (1-\alpha)\psi_1\psi_5\psi_3)}{\alpha\theta_2\theta_5\theta_3 + (1-\alpha)\psi_2\psi_5\psi_3}}$$

This fraction is equal to θ_1 after substituting $\psi_1 = \theta_1$ and $\psi_2 = \theta_2$, which are implied by the hypothesis. The same calculation implies that $\mu_2 = \theta_2$. □

In the rest of this section we note that anomalous branch lengths can emerge from mixtures of trees mimicking a tree of the same topology. In particular, it is possible to mix two sets of branch lengths on a tree to mimic a tree of the same topology such that a resulting pendant branch length is arbitrarily small while the corresponding branch length in either of the branch length sets being mixed stays above some arbitrarily large fixed value.

Proposition 5. *There exist branch length sets on the quartet with the same arbitrarily long branch lengths for a given pendant edge which mix to mimic a tree with an arbitrarily short branch length under the binary symmetric model.*

Proof. To get such an anomalous mixture, set $\theta_1 = \psi_1$, $\theta_3 = \psi_3$, $\theta_4 = \psi_4$,

$\theta_2 = \psi_5$, $\theta_5 = \psi_2$, and $\alpha = .5$. The equations (8) and (9) are satisfied because $\theta_3 = \psi_3$ and $\theta_4 = \psi_4$, and therefore $t_3 = s_3$ and $t_4 = s_4$. This implies that the mixture will indeed satisfy the phylogenetic invariants.

Now, because again the Fourier transform of a mixture is the mixture of the Fourier transform, using Lemma 1 we have

$$\mu_1 = \frac{\theta_1|\theta_2 + \theta_5|}{\sqrt{\theta_2\theta_5}} \tag{10}$$

Now note that by making the ratio θ_2/θ_5 small, it is possible to have μ_1 be close to one although θ_1 can be small. This setting corresponds (via (1)) to the case of the first branch length of the resulting tree to be going to zero although the trees used to make the mixture may have long first branch lengths. It can be checked by calculations analogous to (10) that the other fidelities of the tree resulting from mixing will be, in order, $\sqrt{\theta_2\theta_5}$, θ_3, θ_4, $\sqrt{\theta_2\theta_5}$. These are clearly strictly between zero and one, so the resulting tree will have positive branch lengths. □

Mixtures mimicking a tree of a different topology

In this section we answer the question of what branch lengths on a quartet can mix to mimic a quartet of a different topology.

Proposition 6. *Let $k_1, \ldots, k_4$ satisfy the following inequalities:*

$$k_1 > k_3 > k_4 > 1 > k_2 > 0, \tag{11}$$

$$\tfrac{1-k_1^2}{k_1}\tfrac{1-k_4^2}{k_4} + \tfrac{1-k_2^2}{k_2}\tfrac{1-k_3^2}{k_3} > 0, \tag{12}$$

$$\tfrac{k_1+k_4}{1+k_1k_4} \cdot \tfrac{k_2+k_3}{1+k_2k_3} > 1. \tag{13}$$

Then there exists π_5 such that for any $\pi_5 < k_5 < \pi_5^{-1}$ sufficiently close to either π_5 or π_5^{-1} there exists a mixing weight such that for any $t_1, \ldots, t_5$ and $s_1, \ldots, s_5$ satisfying $\pi_5 = \exp(-2(t_5 + s_5))$ and $k_i = \exp(-2(t_i - s_i))$ for $i = 1, \ldots, 5$, the corresponding mixture of two $12|34$ trees will satisfy the phylogenetic invariants for a single tree of the $13|24$ topology. The resulting internal branch length is guaranteed to be positive, and the pendant branch lengths will be positive as long as the pendant branch lengths being mixed are sufficiently large.

Proof. Let $m_\cdot$ denote the Fourier transform of the site pattern frequencies of the mixture. The invariants for a tree of topology $13|24$ are (by (3) and (4))

$$m_{1234} - m_{13}m_{24} = 0 \tag{14}$$

$$m_{12}m_{34} - m_{14}m_{23} = 0 \quad . \tag{15}$$

As before, we insert the mixture of the Fourier transforms of the

pattern frequencies into the invariants. For the first invariant,

$$\begin{aligned}(\alpha + 1 - \alpha)(\alpha\, q_{1234} + (1-\alpha)\, r_{1234}) \\ -(\alpha\, q_{13} + (1-\alpha)\, r_{13})(\alpha\, q_{24} + (1-\alpha)\, r_{24}) = 0.\end{aligned}$$

Multiplying, this is equivalent to

$$\begin{aligned}\alpha^2(q_{1234} - q_{13}q_{24}) \\ +\alpha(1-\alpha)\,(q_{1234} + r_{1234} - (q_{13}r_{24} + r_{13}q_{24})) \\ +(1-\alpha)^2(r_{1234} - r_{13}r_{24}) = 0.\end{aligned} \tag{16}$$

A similar calculation with the second invariant leads to

$$\begin{aligned}\alpha^2(q_{12}q_{34} - q_{14}q_{23}) \\ +\alpha(1-\alpha)\,(q_{12}r_{34} + r_{12}q_{34} - (q_{14}r_{23} + r_{14}q_{23})) \\ +(1-\alpha)^2(r_{12}r_{34} - r_{14}r_{23}) = 0\end{aligned} \tag{17}$$

Rather than (16) and (17) themselves, we can take (16) and the difference of (16) and (17). Because the $q_{\cdot}$ and $r_{\cdot}$ come from a tree with topology $12|34$, they satisfy $q_{1234} = q_{12}q_{34}$ and $q_{13}q_{24} = q_{14}q_{23}$ and the same equations for r. Thus the difference of (16) and (17) can be simplified to (assuming $\alpha(1-\alpha) \neq 0$)

$$\begin{aligned}q_{1234} + r_{1234} - (q_{12}r_{34} + r_{12}q_{34}) \\ = q_{13}r_{24} + r_{13}q_{24} - (q_{14}r_{23} + r_{14}q_{23}).\end{aligned} \tag{18}$$

We would like to ensure that the tree coming from the mixture has nonzero internal branch length. By Lemma 2 this is equivalent to showing that

$$m_{13}\ m_{24} > m_{14}\ m_{23} \tag{19}$$

Substituting in for the mixture fidelities and simplifying results in

$$\begin{aligned}
&\alpha^2(q_{13}q_{24} - q_{14}q_{23}) \\
&\qquad +\alpha(1-\alpha)\left(q_{13}r_{24} + r_{13}q_{24} - (q_{14}r_{23} + r_{14}q_{23})\right) \\
&\qquad +(1-\alpha)^2(r_{13}r_{24} - r_{14}q_{23}) > 0
\end{aligned}$$

The first and last terms of this expression vanish because the $q_{\cdot}$ and $r_{\cdot}$ satisfy the 12|34 phylogenetic invariants coming from (3) and (4). Simplifying leads to

$$q_{13}r_{24} + r_{13}q_{24} > q_{14}r_{23} + r_{14}q_{23}. \tag{20}$$

Define $k_i = \psi_i/\theta_i$ for $i = 1, \ldots, 5$ and $\rho = \alpha/(1-\alpha)$. Note that

$$0 < k_i < \infty \text{ and } \theta_i < \min(k_i^{-1}, 1) \tag{21}$$

is equivalent to $0 < \theta_i < 1$ and $0 < \psi_i < 1$. Define

$$\begin{aligned}
\chi_{12} &= k_1k_2 + k_3k_4 \qquad & \chi_{13} &= k_1k_3 + k_2k_4 \\
\chi_{14} &= k_1k_4 + k_2k_3 \qquad & \chi_{1234} &= 1 + k_1k_2k_3k_4.
\end{aligned}$$

Later we will make use of the fact that the $\chi_{\cdot}$ are invariant under the action of the Klein four group.

Using these definitions, direct substitution using (2) into (16), (18), and (20) and some simplification shows that the set of equations

$$\begin{aligned}\rho^2(1-\theta_5^2)+\rho(\chi_{1234}-\theta_5\psi_5\chi_{13}) \\ +(1-\psi_5^2)(\chi_{1234}-1)>0\end{aligned} \tag{22}$$

$$\chi_{1234}-\chi_{12}=\theta_5\psi_5(\chi_{13}-\chi_{14}) \tag{23}$$

$$\chi_{13}>\chi_{14} \tag{24}$$

is equivalent to equations (14), (15) and (19).

Assign variables A, B, and C in the standard way such that (22) can be written $A\rho^2+B\rho+C$. The A and C terms are strictly positive, thus the existence of a $0<\rho<\infty$ satisfying this equation implies

$$B<0 \text{ and } B^2-4AC>0. \tag{25}$$

On the other hand, (25) implies the existence of a $0<\rho<\infty$ satisfying (22).

Equation (23) is simply satisfied by setting

$$\theta_5\psi_5=\frac{\chi_{1234}-\chi_{12}}{\chi_{13}-\chi_{14}}. \tag{26}$$

However, in doing so, we must require that this ratio is strictly between zero

and one. The fact that it must be less than one can be written

$$\chi_{14} + \chi_{1234} < \chi_{12} + \chi_{13} \tag{27}$$

which by a short calculation is equivalent to (13). In the next paragraph it will be shown that this ratio being greater than zero follows from other equations.

It is also necessary to check that the variable $B < 0$ after substituting for $\theta_5\psi_5$, namely that

$$\chi_{1234} - \frac{\chi_{1234} - \chi_{12}}{\chi_{13} - \chi_{14}}\chi_{13} < 0.$$

Multiplying by $\chi_{13} - \chi_{14}$ which is positive by (24) this equation is equivalent to

$$\chi_{12}\chi_{13} < \chi_{1234}\chi_{14} \tag{28}$$

which by a short calculation is equivalent to (12). The conclusion then is that the existence of a $\rho \geq 0$ satisfying (22) is equivalent to (12) and $B^2 - 4AC > 0$ given the rest of the invariants.

Now, (24) and (28) imply that $\chi_{12} < \chi_{1234}$. Therefore, according to (26) the product $\theta_5\psi_5$ is greater than zero given (24). For convenience, set $\pi_5 = \theta_5\psi_5$, which as described is determined by $k_1, \ldots, k_4$. Now, θ_5 being less than one and ψ_5 being less than one are equivalent to

$$\pi_5 < k_5 < \pi_5^{-1}. \tag{29}$$

In summary, the problem of finding branch lengths and a mixing parameter such that the derived variables satisfy (14), (15) and (19) is equivalent to finding k_i and θ_i satisfying (12), (13), (21), (24), (26), (29) and $B^2 - 4AC > 0$, which can be written

$$(\chi_{1234} - \pi_5\chi_{13})^2 - 4(1 - \pi_5/k_5)(1 - \pi_5 k_5)(\chi_{1234} - 1) > 0. \tag{30}$$

However, this last equation can be satisfied while fixing the other variables by taking k_5 close to π_5 or π_5^{-1} while satisfying (29).

Now we show that (possibly after relabeling) equation (11) is equivalent to (24) in the presence of the other inequalities. Recall that the $\chi_\cdot$ are invariant under the action of the Klein group acting on the indices of k_i. Because the invariants are equivalent to equations which can be expressed in terms of the $\chi_\cdot$ with θ_5 and ψ_5, we can assume that $k_1 \geq k_2$ and $k_1 \geq k_3$ by renumbering via an element of the Klein group.

Now, subtract $\chi_{12}\chi_{14}$ from (28) to find

$$\chi_{12}(\chi_{13} - \chi_{14}) < (\chi_{1234} - \chi_{12})\chi_{14}.$$

Rearranging (27), it is clear that this implies that

$$\chi_{12} < \chi_{14}. \tag{31}$$

Inserting the definition of the $\chi_\cdot$ into (24) and (31) shows that these equations

are equivalent to

$$0 < (k_1 - k_2)(k_3 - k_4) \text{ and } 0 < (k_1 - k_3)(k_4 - k_2). \tag{32}$$

We have assumed by symmetry that $k_1 \geq k_2$ and $k_1 \geq k_3$; now (32) shows that k_1 can't be equal to either k_2 or k_3. Also, (32) shows that $k_3 > k_4$ and $k_4 > k_2$. All of these inequalities put together imply that $k_1 > k_3 > k_4 > k_2$, which directly implies (24).

Furthermore, another rearrangement of (27) using the inequality (31) leads to $\chi_{1234} < \chi_{13}$. This after substitution gives $(1 - k_1k_3)(1 - k_2k_4) < 0$, which implies that it is impossible for all of the k_i to be either less than or greater than one.

Note that (12) excludes the case $k_1 > k_3 > 1 > k_4 > k_2$; this leaves $k_1 > 1 > k_3 > k_4 > k_2$ and $k_1 > k_3 > k_4 > 1 > k_2$. We can assume the latter without loss of generality by exchanging the θ_i and the ψ_i (which corresponds to replacing k_i with k_i^{-1}) and renumbering.

So far we have described how to find values for the branch lengths so that the invariants (3) and (4) and the internal branch length inequality (19) are satisfied. However, we also need to check that the resulting pendant branch lengths for the tree are positive. Here we describe how this can be achieved by taking a lower bound on the values of t_i.

Assume edges a and b are adjacent on the 12|34 trees being mixed, and a and c are adjacent on the resulting 13|24 tree. Then, by Lemma 1 and

(2), the fidelity of the pendant a edge is

$$\sqrt{\frac{(\alpha\theta_a\theta_b + (1-\alpha)\psi_a\psi_b)(\alpha\theta_a\theta_5\theta_c + (1-\alpha)\psi_a\psi_5\psi_c)}{\alpha\theta_b\theta_5\theta_c + (1-\alpha)\psi_b\psi_5\psi_c}}.$$

In order to assure that the resulting pendant branch length for edge a is positive, we must show that the above fidelity is less than one. This is equivalent to showing that θ_a must satisfy

$$\theta_a < \sqrt{\frac{\alpha + (1-\alpha)k_bk_5k_c}{(\alpha + (1-\alpha)k_ak_b)(\alpha + (1-\alpha)k_ak_5k_c)}} \tag{33}$$

for all such a, b, c triples. Thus this equation along with (21) imply upper bounds for θ_a; by the definition of fidelities these translate to lower bounds for t_a. This concludes the proof. □

Note that the proof actually completely characterizes (up to relabeling) the set of branch lengths and mixing weights such that the resulting mixture mimics a tree of different topology.

Proposition 7. *If two sets of branch lengths on the* 12|34 *tree mix to mimic a tree of the topology* 13|24 *then up to relabeling the associated k_i must satisfy the inequalities (11), (12), (13), and (29); the θ_i must satisfy the inequalities (21) and (33). The two required equalities are that the product $\theta_5\psi_5$ must satisfy (26), and the associated ρ must satisfy (22).*

Figure 1: Mixtures of two sets of branch lengths on a tree of a given topology can have exactly the same site pattern frequencies as a tree of a different topology under the two-state symmetric model. The notation in the diagram showing $x * T_1 + (1 - x) * T_1^{'} = T_2$ means that the indicated mixture of the two branch lengths sets shown in the diagram gives the same expected site pattern frequencies as the tree T_2. The diagrams show two examples of this "mixed branch repulsion"; the general criteria for such mixtures is explained in the text. The branch length scale in the diagrams is given by the line segment indicating the length of a branch with 0.5 substitutions per site. Note that the mixing weights in this example have been rounded.

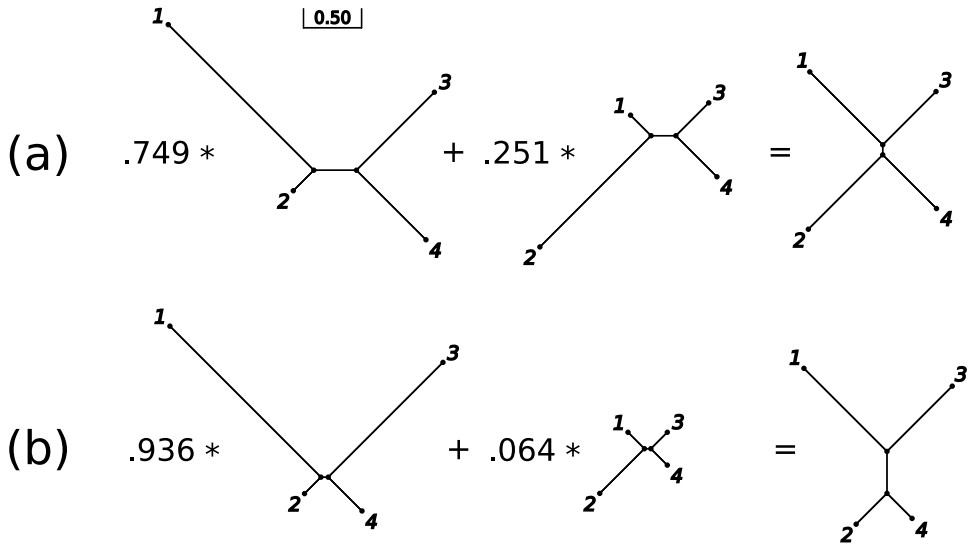

Figure 2: Mixtures of two sets of branch lengths on a tree of a given topology can have exactly the same site pattern frequencies as a tree of the same topology under the two-state symmetric model. The criterion for the occurrence of this phenomenon is explained in the text and an example is shown in the figure. Note in particular that the branch lengths need not average: for example, the branch length for the pendant edge leading to taxon 1 virtually disappears after mixing.

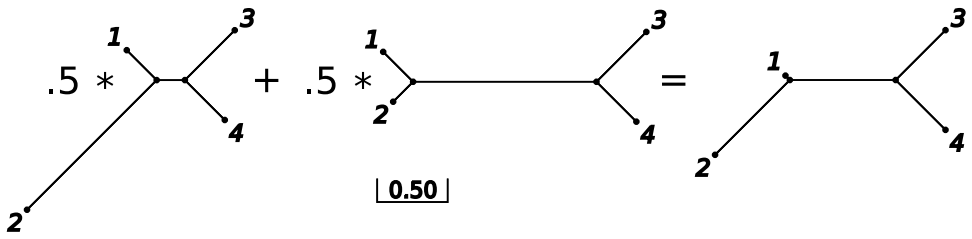

Figure 3: A geometric depiction of the main result. The ambient space is a projection of the seven-dimensional probability simplex of site pattern frequencies for trees on four leaves. The gray sheet is a subset of the two-dimensional subvariety of the site pattern frequencies for trees of the 12|34 topology, while the black sheet is an analogous subset for the 13|24 topology. The bold line represents the possible mixtures for the two sets of branch lengths for the 12|34 topology in Figure 1a. The fact that these two sets of branch lengths can mix to make a tree of topology 13|24 is shown here by the fact that the bold line intersects the black sheet.