Version dated: July 12, 2017

CONCATENATION IN THE ANOMALY ZONE

Why concatenation fails near the anomaly zone

Fábio K. Mendes[1*], Matthew W. Hahn[1,2]

[1]*Department of Biology, Indiana University, Bloomington, IN, 47405, USA;*

[2]*Department of Computer Science, Indiana University, Bloomington, IN, 47405, USA*

*E-mail: fkmendes@indiana.edu.

Appendix A

Calculating S_t, the overall support for a topology t

For rooted species tree (((A,B),C),D) (outgroup omitted) and under the infinite sites model, maximum-parsimony methods should recover the topology t that has the largest support (S_t; Eq. A.1 below, but see main text for a more thorough explanation), with support here meaning the total length of gene tree branches that are present as internal branches in topology t. Note that if the infinite sites assumption is violated, the exact relationship between gene tree branch lengths and the support (i.e., the count of site patterns) for different topologies can become less clear due to homoplasy (but see Chifman and Kubatko, 2015, for the case of a four-taxon species tree). Two topologies compete when data is concatenated: the species tree topology (((A,B),C),D), and the anomalous gene tree (AGT) topology ((A,B),(C,D)). Because these two topologies share the internal branch subtending node $\{A, B\}$, one can compare S_4 and S_1 (Table 1, main text) by focusing on the branches these two topologies do *not* share: the branch subtending node $\{A, B, C\}$ (present in the species tree topology) and the branch subtending $\{C, D\}$ (present in the AGT). The species tree topology ($t = 4$; Table 1, main text) will be returned as the most parsimonious (instead of the AGT, $t = 1$) if $S_4 > S_1$.

S_t is defined in the main text as:

$$S_t = \sum_{u;u \in U} \sum_{b;b \in B_{u,t}} P(u)L(b \mid u) \tag{A.1}$$

where U is the set of gene tree topologies that share internal branches with topology t, and $B_{u,t}$ is the set of internal branches that each individual gene tree, u, in U shares with t. $P(u)$ is the probability of gene tree topology u under the species tree (Table 1, main text).

$L(b|u)$ is the expected length in coalescent units (N_e generations) of branch b (in the set $B_{u,t}$) given topology u. For the case where the internal branches of the species tree (x and y; Fig. 1a, main text) have a length of zero (i.e., the species tree is a four-taxon polytomy), finding $L(b|u)$ is straightforward using coalescent theory (Equations 2 and 3, main text).

When the species tree internal branches are not zero, however, a given gene tree topology u can be classified into different coalescent history classes (Degnan and Salter, 2005), the set of which is denoted H. A history class h is defined by the times at which coalescent events take place (Fig. A.1 and Table A.1 and A.2; see below). We can replace the probability of observing each gene tree topology, $P(u)$, with the probability of each history class h in H given u, $G(h \mid u)$. Importantly, we must update the definition of S_t, as the expected branch lengths now depend on h and u:

$$S_t = \sum_{u;u \in U} \sum_{h;h \in H} \sum_{b;b \in B_{u,t}} G(h \mid u) L(b \mid u, h) \tag{A.2}$$

Calculating the probability of a coalescent history class

The probabilities of coalescent history classes given a gene tree topology (defined here as $G(h \mid u)$) have been derived in Pamilo and Nei (1988) and Rosenberg (2002) for the species tree being considered here (for more general cases, see Degnan and Salter 2005). Those calculations make use of the function $g_{ij}(\tau)$ (Tavaré, 1984), defined as:

$$g_{ij}(\tau) = \sum_{k=j}^{i} e^{-k(k-1)\frac{\tau}{2}} \frac{(2k-1)(-1)^{k-j} j_{(k-1)} i_{[k]}}{j!(k-j)! i_{(k)}}. \tag{A.3}$$

where $a_{(k)} = a(a+1)\dots(a+k-1)$ for $k \geq 1$ with $a_{(0)} = 1$; and $a_{[k]} = a(a-1)\dots(a-k+1)$

for $k \geq 1$ with $a_{[0]} = 1$. $g_{ij}(\tau)$ returns the probability that i lineages descend from j lineages τ coalescent units in the past, with $g_{ij}(\tau) = 0$ except when $i \geq j \geq 1$.

From Equation (A.2), comparing S_1 and S_4 requires computing $G(h \mid u)$. Note, however, that because some of the history classes contribute the same support to S_t, we do not have to calculate $G(h \mid u)$ for all values of h. For example, history classes 2, 4 and 5 given $u = 4$ all contribute 1 to S_4, and so their probabilities $(\delta_1 + \delta_2 + \delta_3)$ can be evaluated to $(1 - (g_{21}(y)g_{21}(x) + g_{22}(y)g_{31}(x)\frac{1}{3}))$ (Table A.1).

Calculating expected branch lengths

After calculating the probabilities of the different coalescent history classes, $G(h \mid u)$, we now must calculate the expected gene tree branch lengths for each t contributed by each h. For our purposes in comparing the species tree and the AGT, the only branches that matter are those supporting node $\{A, B, C\}$ and node $\{C, D\}$. Evaluating S_4, for example, would entail summing the expected branch lengths in all coalescent histories from all three gene tree topologies that have node $\{A, B, C\}$ (Fig. A.1; this is equivalent to summing all branches highlighted in red).

Again, expected branch lengths can be obtained with coalescent theory (Tables A.1 and A.2) if we assume clock-like evolution. Some of the expected branch lengths (such as those from history classes 2, 4 and 5, given $u = 4$; Table A.1) are simply the expected time until coalescence of two lineages (N_e generations = 1 coalescent unit). For the remaining history classes, however, we must find the expected times of coalescence of either two lineages, or three lineages into their MRCA *conditioning* on finding the MRCA within a branch of length τ. The former is used when finding the support for the species tree ($t = 4$) coming from history class 1 of the congruent topology ($h = 1$ and $u = 4$; Fig. A.1): here, two lineages must coalesce in x, so we must subtract the expected time of coalescence (conditioning on it happening in x) from $1 + x$.

(Note that branch lengths measured in coalescent units as derived here are informative of the support they provide to competing topologies only if we make the assumption that N_e is the same across species and along the species tree. This assumption is necessary because coalescent units conflate time and effective population sizes. A “wide and long” [large internode distance and N_e] and a “thin and short” [small internode distance and N_e] can have the same length in coalescent units and be equivalent in the distributions of discordant topologies they allow for – but may have different distributions of site patterns, which can then influence the support they provide to competing topologies.)

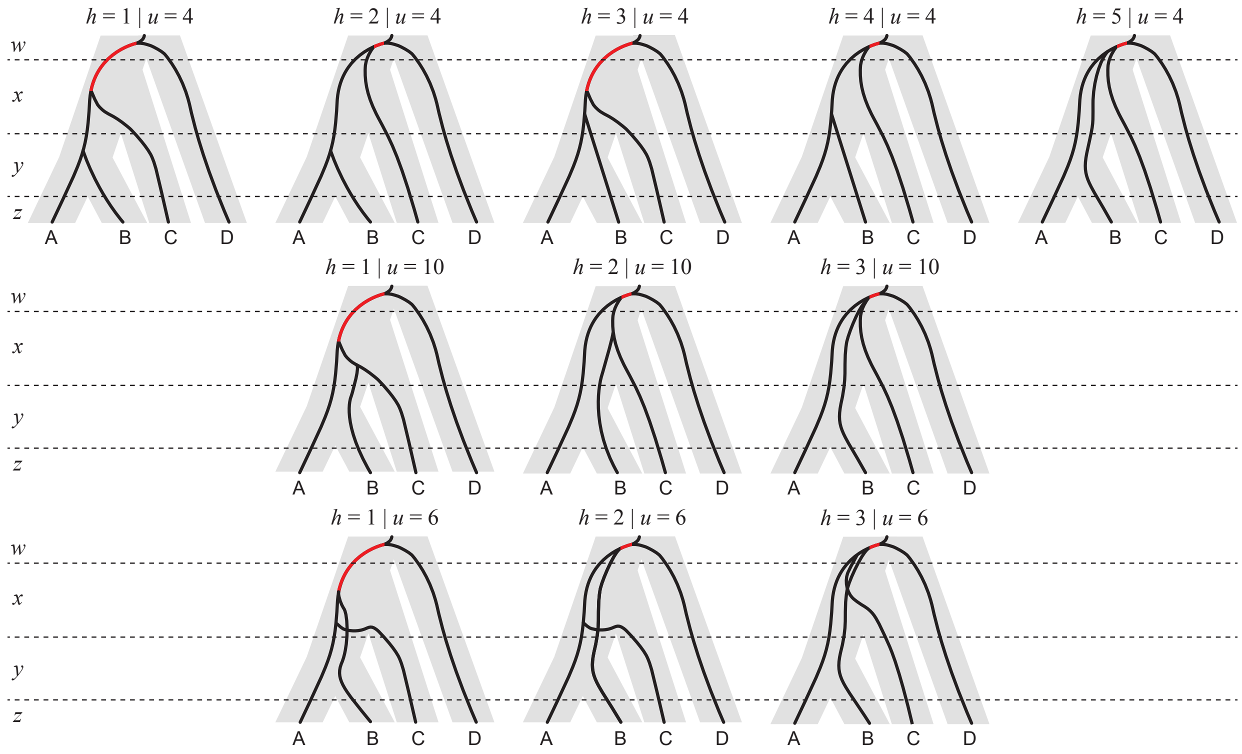

Figure A.1: All history classes from all gene tree topologies that share node {A,B,C} with the species tree topology. Branches in red represent the contributed support of each history class to the species tree topology.

In order to derive the expected time of coalescence of two lineages conditioning on a coalescent event happening within a branch of length τ, we use the fact that the expected time of coalescence of two lineages, v, is exponentially distributed (with $\lambda = 1$), with *pdf*:

$$f(v_2; 1) = \begin{cases} e^{-v_2} & x \geq 0, \\ 0 & \text{otherwise}, \end{cases} \tag{A.4}$$

and *cdf*:

$$F(v_2 = \tau; 1) = \begin{cases} 1 - e^{-\tau} & x \geq 0, \\ 0 & \text{otherwise}. \end{cases} \tag{A.5}$$

Note that in the *cdf* above, we equate $v_2 = \tau$ because we are interested in the probability of coalescence before time τ.

We can then define the *pdf* of v_2 given that a coalescent event happens within a branch of length τ, by dividing Equation (A.4) by Equation (A.5):

$$f_\tau(v_2 \mid \text{Coalescence}) = \begin{cases} \frac{e^{-v_2}}{1-e^{-\tau}} & 0 \leq v_2 < \tau, \\ 0 & \text{otherwise}, \end{cases} \tag{A.6}$$

and then finally calculate the *pdf* for the expected time for two lineages to coalesce in a branch of length τ, conditioning on a coalescence event happening, $q(\tau)$:

$$q(\tau) = E[f_\tau(v_2 \mid \text{Coalescence})] = \int_0^\tau v_2 \frac{e^{-v_2}}{1 - e^{-\tau}} dv_2 = 1 - \frac{\tau}{e^\tau - 1}. \tag{A.7}$$

Importantly, $q(\tau)$ converges on 1 coalescent unit, as expected (Fig. A.2).

The same logic outlined above can be used to derive the expected time of coalescence of three lineages into their MRCA within a branch of length τ, conditioning on their coalescence taking place in that branch. In this case, the expected time of coalescence

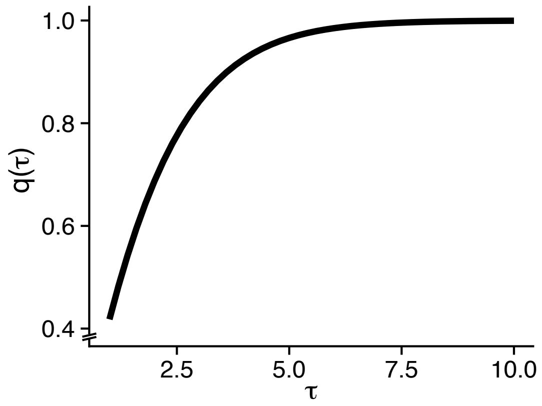

Figure A.2: Expected time of coalescence of two lineages within a branch of length τ, conditioning on a coalescence event happening.

of three lineages into their MRCA, v_3, can be seen as a variable resulting from the convolution of two exponentially distributed random variables (with $\lambda = 1$ and $\lambda = 3$, respectively). If we name the *pdf*s of these two exponential variables $k(v_3)$ and $l(v_3)$, we can define the *pdf* of the convolved variable:

$$f_{k+l}(\alpha) = \int_{-\infty}^{\infty} k(v_3) l(\alpha - v_3) dv_3 = -\frac{(e^{\alpha\lambda_1} - e^{-\alpha\lambda_2})\lambda_1\lambda_2}{\lambda_1 - \lambda_2}, \tag{A.8}$$

for $\alpha > 0$. Replacing $\lambda_1 = 1$ and $\lambda_2 = 3$, we obtain *pdf*:

$$f_{k+l}(\alpha) = \begin{cases} \frac{3}{2}(-e^{-3v_3} + e^{-v_3}) & v_3 > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{A.9}$$

and *cdf* (similarly to what was done above, we equate $v_3 = \tau$):

$$F_{k+l}(\alpha) = \begin{cases} \frac{1}{2}(2 + e^{-3\tau} - 3e^{-\tau}) & x > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{A.10}$$

We can then define the *pdf* of v_3 given a coalescent event happens within a branch of length τ, by dividing Equation (A.9) by Equation (A.10):

$$f_\tau(v_3 \mid \text{Coalescence}) = \begin{cases} \frac{3(-e^{-3v_3}+e^{-v_3})}{2+e^{-3\tau}-3e^{-\tau}} & 0 \leq v_3 < \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{A.11}$$

The last step is to calculate the *pdf* for the expected time for two lineages to coalesce in a branch of length τ, conditioning on a coalescence event happening, $r(\tau)$:

$$\begin{aligned} r(\tau) = E[f_\tau(v_3 \mid \text{Coalescence})] &= \int_0^\tau v_3 \frac{3(-e^{-3v_3}+e^{-v_3})}{2+e^{-3\tau}-3e^{-\tau}} dv_3 = \\ &= \frac{1+8e^{3\tau}+3b-9e^{2\tau}(1+\tau)}{3(-1+e^\tau)^2(1+2e^\tau)}. \end{aligned} \tag{A.12}$$

Finally, we must again verify the convergence of $r(\tau)$, except in this case the expectation is $1+\frac{1}{3}$ coalescent units (Fig. A.3).

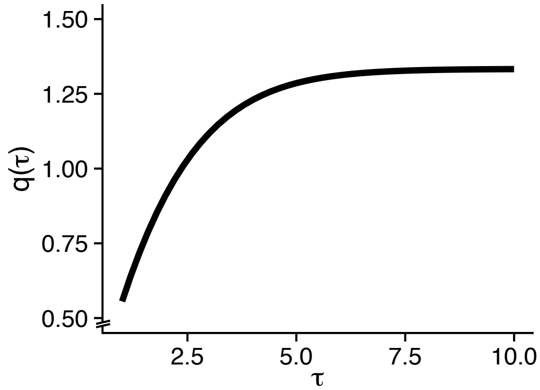

Figure A.3: Expected time of coalescence of three lineages within a branch of length τ, conditioning on a coalescence event happening.

Table A.1: Gene trees supporting the species tree topology through the branch subtending node {A,B,C} (branch lengths in N_e generations).

Topology	u	History class, h	Branches containing 1^{st} and 2^{nd} coalescences	Probability of history class, $G(h \mid u)$	Expected branch length, $L(b \mid u, h)$
((AB)C)D)	4	1	y, x	$g_{21}(y)g_{21}(x)$	$1 + x - q(x)$
		2	y, w	δ_1	1
		3	x, x	$g_{22}(y)g_{31}(x)\frac{1}{3}$	$1 + x - r(x)$
		4	x, w	δ_2	1
		5	w, w	δ_3	1
((BC)A)D)	10	1	x, x	$g_{22}(y)g_{31}(x)\frac{1}{3}$	$1 + x - r(x)$
		2	x, w	κ_1	1
		3	w, w	κ_2	1
((AC)B)D)	6	1	x, x	$g_{22}(y)g_{31}(x)\frac{1}{3}$	$1 + x - r(x)$
		2	x, w	ζ_1	1
		3	w, w	ζ_2	1

Table A.2: Gene trees supporting the species tree topology through the branch subtending node {C,D} (branch lengths in N_e generations).

Topology	u	History class, h	Branches containing 1^{st} and 2^{nd} coalescences	Probability of history class, $G(h \mid u)$	Expected branch length, $L(b \mid u, h)$
((AB)(CD))	1	1	y, w	$g_{22}(y)g_{33}(x)\frac{1}{3}\frac{1}{3}$	$1 + \frac{1}{6}$
		2	x, w	β_1	1
		3	w, w	β_2	1
((CD)A)B)	14	1	w, w	1	$\frac{1}{3}$
((CD)B)A)	15	1	w, w	1	$\frac{1}{3}$

Appendix B

Simulations across the phylogenetic space of a four-taxon species tree

In order to understand the behavior of different tree estimation methods across phylogenetic space, we used the coalescent model to simulate gene trees from an asymmetric species tree with four species in its ingroup, ((((A:z,B:z):y,C):x,D):w,E), where z, y, x and w are the lengths of terminal branches A and B, and the internal branches subtending (A,B), ((A,B),C) and (((A,B),C),D), respectively. Branch E leads to the outgroup, so the internal branch length w was always large enough so no ILS happened between E and any of the remaining taxa.

We explored the phylogenetic space of this species tree by simulating 20,000 gene trees at different x- and y- value combinations (measured in coalescent units, where 1 unit = N_e generations), with x varying from 0.015 to 0.285 in 0.015 increments, and y varying from 0.05 to 0.95 in 0.05 increments – for a total of 361 combinations comprising a square xy-grid (w and z were fixed for this initial set of simulations to 12 and 1 coalescent units, respectively). In addition, we further explored phylogenetic space by simulating along the xy-grid four more times: (i) with $z = 0.1$ and $z = 10$ (one each; w was fixed at 12 coalescent units), and (ii) with $w = 8$ and $w = 20$ (one each; z was fixed at 1 coalescent unit). Simulated gene trees were used in conjunction with the Jukes-Cantor nucleotide evolution model (Jukes and Cantor, 1969) and $\theta = 0.04$ to simulate one 1-kb locus alignment per tree. All 20,000 simulated alignments from each xy-grid point were concatenated and used in downstream analyses. Coalescent simulations were done with ms (Hudson, 2002) and sequences were simulated with Seq-Gen (Rambaut and Grassly, 1997).

Comparing empirical and expected support for the species tree and the anomalous tree

We summarized the difference in phylogenetic signal favoring the species tree (SP) versus the anomalous gene tree (AGT) by computing the SP:AGT ratio of the sums of branch lengths supporting each tree. Branch length support for both trees was calculated at 19 grid points along the diagonal of the xy-grid (from $x = 0.015$ and $y = 0.05$, to $x = 0.285$ and $y = 0.95$, and for $x = y = 0$), with 100 replicates for every point, each replicate consisting of 20,000 gene trees.

For each replicate in each grid point, we computed the support for the species tree by adding the lengths of all internal branches subtending ((A,B),C); these branches were present in 3 of the 15 possible topologies: (((A,B),C),D), (((A,C),B),D), and (((B,C),A),D) (outgroup omitted). Similarly, we added the lengths of all internal branches subtending (C,D) in order to obtain the branch length support for the anomalous tree; these branches are found in topologies ((A,B),(C,D)), (((C,D),A),B), and (((C,D),B),A). Finally, we compared the SP:AGT ratios of branch length support at each grid point to the expected theoretical ratios (see Appendix A).

Evaluating tree inference methods on concatenated alignments across phylogenetic space

Phylogenies were estimated from the concatenated alignments across the xy-grid using neighbor-joining, parsimony, and maximum-likelihood as implemented in PAUP* v4.0a150 (Swofford, 2002). Maximum-likelihood estimation was done exhaustively, as in Kubatko and Degnan (2007): all 15 possible rooted topologies had their likelihoods evaluated and the top one was reported. We also estimated the maximum-likelihood tree with heuristic search; in this case PAUP* reported one single best tree in all but one point on the grid.

Inferring site pattern likelihoods under the maximum-likelihood tree

The 20 million sites in each concatenated alignment were first classified into one of

44 unique site pattern bins, after coding the ancestral state (the base present in the outgroup E) as “0”, and the derived states as “1”, “2” or “3” depending on how many different states were present at a given site. This procedure is possible because the Jukes-Cantor model does not incorporate transition-transversion bias, and so site pattern ((((AA)G)G)A), for example, is equivalent to ((((AA)C)C)A); both would be coded as “00110”.

The likelihood of all site patterns was computed for the maximum-likelihood tree at the grid point closest to the origin ($x = 0.015$ and $y = 0.05$). Likelihood computations were done with PAUP*.

References

Chifman, J. and L. S. Kubatko. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. Journal of Theoretical Biology 374:35–47.

Degnan, J. H. and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Hudson, R. R. 2002. Generating samples under a wright-fisher neutral model. Bioinformatics 18:337–338.

Jukes, T. H. and C. R. Cantor. 1969. Evolution of protein molecules. Academic press, New York.

Kubatko, L. S. and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under the coalescence. Systematic Biology 56:17–24.

Pamilo, P. and M. Nei. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5:568–583.

Rambaut, A. and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in the Biosciences 13:235–238.

Rosenberg, N. A. 2002. The probability of topological concordance of gene trees and species trees. Theoretical Population Biology 61:225–247.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their application in population genetics models. Theoretical Population Biology 26:119–164.