Proof for Theorem 2 in "First steps toward the geometry of cophylogeny"

Peter Huggins
Lane Center for Computational Biology
Carnegie Mellon University

Megan Owen
North Carolina State University
Raleigh, NC, USA

Ruriko Yoshida
Department of Statistics
University of Kentucky

Theorem 1 (Theorem 2 in the main text)**.** *Suppose T_H, T_P are unrooted trees on the same set of leaves. Then T_H and T_P satisfy 1-max path difference cospeciation if and only if T_H and T_P differ by at most 1 NNI operation.*

Proof. It is clear that if T_P is at most 1 NNI from T_H, then T_P and T_H satisfy 1-max path difference cospeciation.

To show that if there is 1-max path difference cospeciation, then T_H and T_P differ by at most one NNI, we use induction on the number of leaves. The base case occurs when there are 4 leaves. Then, there can be most one NNI, and thus the number of edges between leaves changes by at most 1.

If some cherry (a, b) in T_H that is also a cherry in T_P, then replace these cherries with the same leaf and we are done by induction. It remains to consider the case when no cherry (a, b) in T_H is also a cherry in T_P.

1-max path difference cospeciation implies a and b are at most 3 edges apart, and thus exactly 3 edges apart, since they do not form a cherry. Then without loss of generality, a forms a cherry with some subtree S_c, and b is attached just above this cherry, as shown in Fig. 1. Futhermore, we can assume, without loss of generality, that in a sequence of NNIs transforming T_H into T_P, the NNI moving S_c to between a and b occurs last. Let T_c be the second last tree in this sequence of NNIs. Then by our hypothesis, T_H and T_c differ by at least 1 NNI. If T_H and T_P differ by exactly 1 NNI, then there are two cases. If this NNI is done about an edge in S_c, then some leaf l in S_c moves one edge closer to the root of S_c. This implies the number of edges between a and l decreases by 2 between T_H and T_P, which is a contradiction. Otherwise, a, b, and S_c are all contained in one of the subtrees involved in the interchange, say A (using the notation from Fig. 2). Then this NNI moves a one edge closer to each leaf in B, and the second NNI moves a one edge closer to the root of A, and hence to each leaf in B. Thus, the number of edges between a and any edge in B decreases by 2 between T_H and T_P, which is a contradiction.

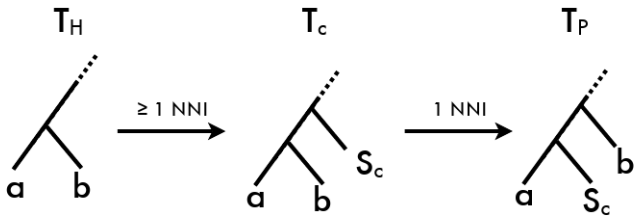

Figure 1: The trees used in the proof of Theorem 2.

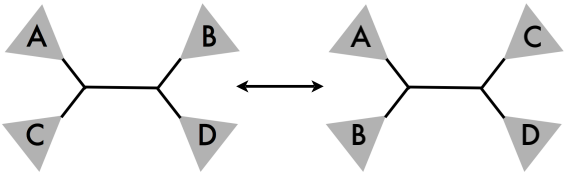

Figure 2: An NNI operation.

Thus at least two NNIs are needed to transform T_H into T_c. Both T_H and T_c contain the cherry (a, b), so replace this cherry with the leaf ab in T_H to get the tree T'_H, and in T_c to get the tree T'_c.

By the induction hypothesis on T'_H and T'_c, there are distinct leaves i, j such that

$$|d_{T'_c}(i,j) - d_{T'_H}(i,j)| > 1. \tag{1}$$

Case ($i, j \notin S_c$ and $i, j \neq ab$) or ($i, j \in S_c$) : Then $d_{T'_c}(i,j) = d_{T_c}(i,j) = d_{T_P}(i,j)$ and $d_{T'_H}(i,j) = d_{T_H}(i,j)$. Plugging into (1) gives $|d_{T_P}(i,j) - d_{T_H}(i,j)| > 1$.

Case $i \notin S_c$, and $j = ab$: Then $d_{T'_c}(i, ab) = d_{T_c}(i,a) - 1 = d_{T_P}(i,a) - 1$, and $d_{T'_H}(i, ab) = d_{T_H}(i,a) - 1$. Plugging into (1) gives $|d_{T_P}(i,a) - d_{T_H}(i,a)| > 1$.

Case $i \in S_c$, $j \notin S_c$, and $j \neq ab$: Consider the subtree of T'_c containing ab, S_c, j, and the paths between them. Let x be the number of edges between j and the common ancestor of ab and i. Let y be the number of edges from the root of subtree S_c to i. Let V be the interior vertex where the paths from ab, i, and j meet in T'_H. Let u be the number of edges between ab and V. Let v be the number of edges between j and V. Let w be the number of edges between i and V. Then $d_{T_c}(i,j) = 1 + x + y$. Since $d_{T'_c}(i,j) = d_{T_c}(i,j)$ and $d_{T'_H}(i,j) = d_{T_H}(i,j)$, then $|d_{T_H}(i,j) - d_{T_c}(i,j)| > 1$, which implies $d_{T_H}(i,j) \leq x+y-1$ or $d_{T_H}(i,j) \geq x+y+3$. Now $d_{T_P}(i,j) = 2+x+y$, so 1-max path difference cospeciation implies $d_{T_H}(i,j) = x+y+3$. By definition of v and w, $v + w = d_{T_H}(i,j) = x+y+3$.

We have $d_{T_P}(a,i) = 2+y$, $d_{T_P}(b,i) = 3+y$, and $d_{T_H}(a,i) = d_{T_H}(b,i)$. So 1-max path difference cospeciation implies $d_{T_H}(a,i) = d_{T_H}(b,i) = 2+y$ or $3+y$. This implies $u+w = 1+y$ or $2+y$. We also have $d_{T_P}(a,j) = 2+x$, $d_{T_P}(b,j) = 1+x$, and $d_{T_H}(a,j) = d_{T_H}(b,j)$.

Then 1-max path difference cospeciation implies $d_{T_H}(a,j) = d_{T_H}(b,j) = 1 + x$ or $2 + x$. This implies $u + v = x$ or $1 + x$.

Then either $(v+w)-(u+w) = v-u = x+2$ or $x+1$. If $v-u = x+2$ and $u+v = x$, then $2v = 2x+2$ or $v = x+1$. This implies $u = -1$, which is impossible. If $v-u = x+2$ and $u+v = 1+x$, then $2v = 2x+3$, which is impossible, because all variables are numbers of edges, and hence integers. If $v - u = x + 1$ and $u + v = x$, then $2v = 2x + 1$, which is also impossible because all variables are integers. Finally, if $v - u = x + 1$ and $u + v = x$, then $2v = 2x + 2$, which implies $v = x + 1$ and $u = -1$, which is impossible. Therefore, $v + w = d_{T_H}(i,j) \neq x + y + 3$, which is a contradiction.

Case $i = ab$ and $j \in S_c$: Then $d_{T_c'}(ab,j) = d_{T_c}(b,j) - 1 = d_{T_P}(b,j) - 1$ and $d_{T_H'}(ab,j) = d_{T_H}(b,j) - 1$. Plugging into (1) gives $|d_{T_P}(b,j) - d_{T_H}(b,j)| > 1$.

Therefore, there exist at least two leaves, such that the number of edges between them changes by more than 1 from T_H to T_P, which is a contradiction. □