A note on the path interval distance

Jane Ivy Coons^a, Joseph Rusinko^b

^a*North Carolina State University*

^b*Hobart and William Smith Colleges*

Abstract

The path interval distance accounts for global congruence between locally incongruent trees. We show that the path interval distance provides a lower bound for the nearest neighbor interchange distance. In contrast to the Robinson-Foulds distance, random pairs of trees are unlikely to be maximally distant from one another under the path interval distance. These features indicate that the path interval distance should play a role in phylogenomics where the comparison of trees on a fixed set of taxa is becoming increasingly important.

Keywords: Tree Metrics, Phylogenetics, Cophylogenetics

1. Introduction

Biologists display evolutionary relationships between operational taxonomical units, or taxa, as leaves of a tree. In many instances a collection of trees may be constructed from a single set of taxa. In other cases, while the taxa on two trees may be biologically distinct, one may wish to identify taxa on each tree to make a comparison. In this note we assume trees are constructed from a single set of taxa.

Tree metrics have been developed to compare the difference between trees generated using the same set of taxa and to assist when combining a set of trees to form a single tree which reflects the shared aspects of the elements of the set.

When testing reconstruction algorithms one compares a model tree to sample trees reconstructed from simulated data using the algorithm [1, 2, 3]. Alternatively, a step in many reconstruction algorithms involves finding a consensus tree among conflicting trees on the same set of taxa [4, 5, 6]. In the study of cophylogenetics one may identify the taxa on a host tree with taxa on a parasite tree if the parasite is known to live on a unique host. In this case the leaf labels refer to an identified pair of taxa (one host and one parasite) [7, 8, 9]. Trees with the same leaf set are also prevalent in phylogenomics when building a species tree from a set of gene trees [10, 11, 12, 13].

Various factors may cause differences among trees on the same leaf set such as; gene species tree discordance under the coalescent model [14], a parasite's failure to speciate in reaction to a host speciation [15], imperfect reconstruction algorithms, or incongruities among phylogenetic trees. However, given the shared evolutionary history, one expects to find a topological relationship between the trees.

Email addresses: janyivycoons@gmail.com (Jane Ivy Coons), rusinko@hws.edu (Joseph Rusinko)

Preprint submitted to Journal of Theoretical Biology

March 11, 2016

Many tree metrics have been developed to analyze the differences between trees. Perhaps most famous among these methods are the Robinson-Foulds distance and the Nearest Neighbor Interchange (NNI) distance. Recently a new metric, k -interval cospeciation, was proposed in the context of cophylogeny [8]. As k -interval cospeciation has broad applicability in comparing trees with identified taxa sets, we refer to this metric instead as the *path interval distance* as this better reflects the generality of the metric and its relationship to the path-difference metric [16] and the edge-difference distance [17].

The path interval distance illuminates global similarities between trees that classical tree metrics do not uncover. We do not propose that the path interval distance supplant the Robinson-Foulds or NNI distances as their importance and statistical properties have been well documented [18, 19]. However, alternative tree metrics may yield additional insight into the relationship between incongruous trees.

We compare the path interval distance to traditional tree metrics and explore its combinatorial and statistical properties. In particular we show that unlike the Robinson-Foulds metric, the probability that two randomly selected trees are maximally distant under the path interval distance goes to zero as the number of taxa increases. Thus the path interval distance may be useful as a tool for analyzing reconstruction algorithms on simulated data or developing consensus methods in phylogenomics.

2. Relationship Between the Path Interval Distance and Other Discrete Tree Metrics

We restrict our attention to unrooted binary trees where each non-leaf vertex has degree three. At each leaf of the tree is a *taxon* (plural: taxa), such as a gene or species. The number of unrooted binary trees on n taxa is denoted $ub(n)$ and given by,

$$ub(n) = (2n - 5)!! = (2n - 5) \times (2n - 3) \times \dots \times 5 \times 3 \times 1.$$

We define a *cherry* to be a pair of taxa with exactly two edges between them, and a *caterpillar tree* to be a phylogenetic tree with exactly two cherries. The *shortest path* between two taxa on a tree T is the smallest connected subgraph of T that contains both taxa. The *length* of the shortest path is the number of edges in the path. Since path interval distance is a discrete tree metric, we do not consider branch lengths.

Definition 1. Let T_1 and T_2 be phylogenetic trees on the taxon set X with $|X| = n$. Let $p_i(A, B)$ be the length of the shortest path between taxa A and B on T_i . The path interval distance between T_1 and T_2 , denoted $d(T_1, T_2)$, is the maximum of $|p_1(A, B) - p_2(A, B)|$ over all pairs of taxa $\{A, B\} \in X^2$.

It follows that the path interval distance is a *tree metric* [8] (where this is referred to as k -interval-cospeciation). See [20] for a concise explanation of the relationship between the path interval distance and other tree metrics. In this note we contrast the path interval distance with several of the more popular discrete tree metrics.

Definition 2. A tree metric, d_a , determines d_b if there exists a bijection, $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f(d_a(T_1, T_2)) = d_b(T_1, T_2)$ for all n taxa trees T_1 and T_2 .

We show that several commonly used discrete tree metrics do not determine the path interval distance. In fact, the path interval distance reflects global similarities between trees which other tree metrics do not.

One common tree metric is the nearest neighbor interchange distance. A *nearest neighbor interchange* (NNI) is a tree rearrangement operation in which we switch two subtrees of a tree that are joined by a single edge [21, 22]. The nearest neighbor interchange distance between T_1 and T_2 is the smallest number of NNIs it takes to transform T_1 into T_2 . Finding the NNI distance between two given trees is an NP-complete problem [23], while the path interval distance can be computed in $O(n^2)$ [20].

Huggins et al. ([8]) proved that two trees are path interval distance one apart if and only if they are NNI distance one apart. However, they also introduced a counterexample to the possibility of the path interval distance providing an upper bound for the NNI distance for any $k > 1$.

Theorem 1. *The NNI distance does not determine the path interval distance.*

Proof. Let C_{3x} and D_{3x} be trees built by adding rooted triples on the same leaf set, but with differing cherries to the pendant edges of two $3x$ -taxa caterpillar trees C and D (see Figure 1a and Figure 1b).

Figure 1: Counterexample trees introduced in [8].

Since no NNI operation is able to alter two of these rooted triples at the same time, the NNI distance between C_{3x} and D_{3x} is x . However, the two trees have a path interval distance of two. Moreover by matching a rooted triple from C_{3x} with D_{3x} we maintain path interval distance of two, but decrease the NNI distance by one. Therefore, the NNI distance does not determine the path interval distance. $\square$

Remark 1. Similarly, one can also show that path interval distance does not determine the Robinson-Foulds distance (defined in [24]), the path difference distance (defined in [16]), or the maximum agreement subtree distance (defined in [25]).

The path interval distance between two trees describes global similarities between the trees that other common discrete tree metrics do not. One way of describing this global similarity is through the length of the longest path between two taxa on the trees. Notice that the path interval distance between two trees provides an upper-bound on the difference between the length of the longest path of the two trees. Thus, trees differing by a small path interval distance must share a similar global topology.

While trees C_{3x} and D_{3x} show that the path interval distance and many of the traditional tree metrics can differ greatly, the path interval distance does provide a strict lower bound on the NNI distance denoted $d_{NNI}(T_1, T_2)$.

Theorem 2. For any trees T_1 and T_2 $d(T_1, T_2) \leq d_{NNI}(T_1, T_2)$.

Proof. It suffices to show that one NNI operation changes the number of edges in the path between any two taxa on a tree by at most one. Consider the trees T_i and T_j which differ by a single NNI as shown in Figure 2. An NNI operation can insert or delete at most one edge between any two taxa which implies that $d(T_i, T_j) \leq 1$. Assume that T_1 and T_2 differ by a sequence of NNI moves. Since each NNI move can alter the path interval distance by at most one, it follows that $d_{NNI}(T_1, T_2) \geq d(T_1, T_2)$. $\square$

Figure 2: Trees differing by a single Nearest Neighbor Interchange.

The lower bound is strict since for any n there exist two n -taxa trees with path interval distance k which can be rearranged to match in a series of k NNI moves. For example, in Figure 3, we have with two n -taxa caterpillar trees where $d_{NNI}(T_1, T_2) = d(T_1, T_2)$.

Figure 3: Two caterpillar trees which are maximally distant under NNI and the path interval distance.

3. The $(n - 3)$ -Interval Neighborhood

With over 1200 citations, the Robinson-Foulds Distance is one of the most commonly used tree metrics. However, under this distance it is possible to make a minor change to a tree and end up with a tree which is maximally distant from the original [26]. This is in part possible because such a large portion of trees are maximally distant from one another under the Robinson-Foulds distance. Motivated by this, we consider the neighborhood of trees maximally distant from a tree under the k -interval neighborhood.

Definition 3. *The k -interval neighborhood of a tree T , denoted $IN_k(T)$, is the set of all trees which are path interval distance k from T .*

Recall that the 1-interval neighborhood was the set of trees which differed by a single nearest neighbor interchange. It follows that $2(n - 3)$ trees lie in the 1-interval neighborhood regardless of the underlying tree topology. We now examine the alternate extreme, the $(n - 3)$ -interval neighborhood of a tree.

Lemma 1. *Given n -taxa trees T_1 and T_2 , the path interval distance $d(T_1, T_2) \leq n - 3$. Moreover, if $d(T_1, T_2) = n - 3$, then one of the trees must be a caterpillar.*

Proof. Consider two n -taxa trees T_1 and T_2 . Then T_1 and T_2 each have exactly $n - 3$ internal edges. Therefore, any two taxa must have between 2 and $n - 1$ edges separating them. So the maximum path interval distance between T_1 and T_2 is $|(n - 1) - 2|$. The second statement follows from the fact that only a caterpillar tree has a path of length $(n - 1)$. $\square$

As a result, the $(n - 3)$ -interval neighborhood is dependent on the tree topology.

Theorem 3. *For tree T with $n \geq 6$ taxa and c cherries,*

$$|IN_{(n-3)}(T)| = \begin{cases} (8n - 30)(2n - 9)!! + (2n^2 - 18n + 40)(n - 4)! & \text{if } c = 2, \\ c(n^2 - 5n + 7 - c)(n - 4)! & \text{if } c \geq 3. \end{cases}$$

Proof. Case 1: When $c = 2$, T is a caterpillar tree with cherries (a, b) and (c, d) . We can construct T' which is maximally distant from T either by:

- placing two taxa on opposite cherries of T into the same cherry on T' , or
- placing two taxa from the same cherry of T into opposite cherries of T' , when T' is a caterpillar.

First, we count the trees obtained by placing two taxa on opposite cherries of T into the same cherry on T' . There are four such pairs of taxa: (a, c) , (a, d) , (b, c) and (b, d) . Since prescribing a cherry reduces the number of available taxa by one we can construct $4ub(n - 1)$ trees with one of the four pairs as a cherry. However, it is possible for cherries (a, c) and (b, d) to appear on the same tree and for cherries (a, d) and (b, c) to appear on the same tree so $2ub(n - 2)$ trees will have been double-counted. So $4ub(n - 1) - 2ub(n - 2) = 4(2n - 7)!! - 2(2n - 9)!!$ is the number of trees path interval distance $n - 3$ from T yielded by putting taxa from opposite cherries in T into the same cherry in T' .

The rest of the trees in $\text{IN}_{(n-3)}(T)$ are constructed by placing two taxa in the same cherry of T into opposite cherries of a caterpillar tree T' . There are $(n-2)!$ caterpillar trees with a and b in opposite cherries. However, $2(n-3)! - (n-4)!$ of those contain cherry (a, c) or (b, d) , which we counted in the previous step. Similarly, there are $2(n-3)! - (n-4)!$ trees with cherry (a, d) or (b, c) which have already been counted. So by fixing a and b in opposite cherries of a caterpillar tree we yield

$$(n-2)! - 4(n-3)! + 2(n-4)!$$

new trees in $\text{IN}_{(n-3)}(T)$.

By the same argument, we obtain another $(n-2)! - 4(n-3)! + 2(n-4)!$ new trees by fixing c and d in opposite cherries without double counting. Therefore, for a caterpillar tree,

$$\begin{aligned} |\text{IN}_{(n-3)}(T)| &= 4(2n-7)!! - 2(2n-9)!! + 2(n-2)! - 8(n-3)! + 4(n-4)! \\ &= (8n-30)(2n-9)!! + (2n^2 - 18n + 40)(n-4)!. \end{aligned}$$

Case 2: Let T be a tree with $c \geq 3$ cherries. If $d(T, T') = n-3$ holds, then T' must be a caterpillar. Moreover, there must be two taxa that were in a cherry on T which are in opposite cherries on T' .

For each cherry (i_l, j_l) on T , $1 \leq l \leq c$, there exist $(n-2)!$ caterpillar trees with i_l and j_l in opposite cherries. For i_1 and j_1 , this amounts to $(n-2)!$ trees counted in $\text{IN}_{(n-3)}(T)$. For (i_2, j_2) , there are $(n-2)!$ caterpillar trees with i_2 and j_2 in opposite cherries. However, $(n-4)!$ of these have cherries (i_1, i_2) and (j_1, j_2) and another $(n-4)!$ have cherries (i_1, j_2) and (j_1, i_2) . So of the $(n-2)!$ caterpillar trees with i_2 and j_2 in opposite cherries, $2(n-4)!$ were already counted. Similarly for each (i_l, j_l) , $2(l-1)(n-4)!$ have already been counted for the previous cherries of T . So

$$\begin{aligned} |\text{IN}_{(n-3)}(T)| &= \sum_{l=1}^c (n-2)! - 2(l-1)(n-4)! \\ &= c(n-2)! - 2(n-4)! \sum_{l=2}^c (l-1) \\ &= c(n-2)! - \frac{2(n-4)!c(c-1)}{2} \\ &= c(n^2 - 5n + 7 - c)(n-4)!. \end{aligned} \tag{1}$$

□

Theorem 4. *For all n -taxa trees with three or more cherries, a tree with three cherries has the smallest $(n-3)$ -interval neighborhood, and a tree with $\lfloor \frac{n}{2} \rfloor$ cherries has the largest. The size of the $(n-3)$ -interval neighborhood increases as a quadratic in the number of cherries c .*

Proof. For an n taxa tree with $c \geq 3$ cherries, by Equation 1 the size of the maximal interval is quadratic function in c with a vertex at $c = \frac{n^2-5n+7}{2}$. So, if c could exceed that

value, the size of $IN_{(n-3)}(T)$ would start to decrease. The greatest number of cherries a binary tree can have is $\lfloor \frac{n}{2} \rfloor$. So we need to show that $\lfloor \frac{n}{2} \rfloor \leq \frac{n^2-5n+7}{2}$ for all $n \geq 6$. It suffices to find where $\frac{1}{2}(n^2 - 5n + 7) \leq \frac{n}{2}$, which is where

$$\frac{1}{2}n^2 - 3n + \frac{7}{2} \leq 0.$$

The roots of this quadratic are located at $n = 3 \pm \sqrt{2}$. So the inequality is only true when $n \leq 3 + \sqrt{2}$. The theorem follows by noting that any tree with at least three cherries must have at least six taxa. $\square$

We use the preceding two theorems to compare the size of the $(n - 3)$ -interval neighborhood for caterpillar trees with that of trees with more cherries.

Theorem 5. *Of all n -taxa trees, the caterpillar tree has the largest $(n - 3)$ -interval neighborhood.*

Proof. Let T be an n taxa caterpillar tree, and T' by an n taxa tree with $\frac{n}{2}$ cherries. Then by Theorem 3 we have

$$|IN_{n-3}(T)| - |IN_{n-3}(T')| = (8n-30)(2n-9)!! + (2n^2-18n+40)(n-4)! - \frac{n}{2}(n^2-5n+7-\frac{n}{2})(n-4)!.$$

Combining terms yields a difference of

$$(8n - 30)(2n - 9)!! + \left(\frac{-n^3}{2} + \frac{19n^2}{4} - \frac{43n}{2} + 40\right)(n - 4)!.$$

For all values of $n \geq 4$ this exceeds,

$$(2n - 9)!! + \frac{-n^3}{2}(n - 4)!.$$

Using induction one can show this value is positive for all $n \geq 20$. A direct computation yields that $|IN_{n-3}(T)| > |IN_{n-3}(T')|$ for all n from six to nineteen. $\square$

We have shown that the caterpillar tree has the largest $(n - 3)$ -interval neighborhood. We will show that a very small portion of all trees lie in this largest neighborhood, which is a direct contrast to the distribution under the Robinson-Foulds distance which we review here.

Definition 4. [22, 24] *A split is a bipartition of the set of taxa induced by an edge of a tree T . The Robinson-Foulds distance between T_1 and T_2 is defined to be the normalized number of splits induced by T_1 and not by T_2 , and vice versa.*

Bryant and Steel ([18]) showed that asymptotically the proportion of trees that share s splits follows a Poisson distribution with mean $\lambda = \frac{c}{2n}$, where n is the number of taxa and c is the number of cherries. For a given tree T , the expected proportion of trees that share s splits with T is given by $\frac{\lambda^s e^{-\lambda}}{s!}$.

When two trees, T_1 and T_2 have $s = 0$, this means that they share no splits, and that they are the farthest possible Robinson-Foulds distance apart. The expected proportion of trees which share 0 splits with an n taxa tree with exactly c cherries is given by $e^{-c/2n}$.

For a fixed c , this expected value goes to one as n approaches infinity. So if we fix a number of cherries on a tree and increase the number of taxa, it becomes increasingly likely that we will choose a tree that is the maximum Robinson-Foulds distance away from T .

However, if we let c be the maximum number of cherries possible, we have

$$\lim_{n \rightarrow \infty} e^{-\lfloor n/2 \rfloor / 2n} = e^{-1/4} \approx 0.7788.$$

These results demonstrate that for large taxa, a high proportion of trees will be maximally distant from one another under the Robinson-Foulds metric. In a stark contrast, two random trees are unlikely to be maximally distant from one another under the path interval distance.

Theorem 6. *The proportion of trees in the $(n - 3)$ -interval neighborhood of a given tree approaches 0 as n approaches infinity.*

Proof. Let T be an n taxa tree. By Theorem 3 we have

$$\begin{aligned} |IN_{n-3}(T)| &\leq 2((n-2)! + 2((n-4)! - 2(n-3)!) + 2(2n-7)!! - (2n-9)!!) \\ &\leq 2((n-2)! + 2(n-4)! + 2(2n-7)!!) \\ &\leq 2(n-2)! + 4(n-4)! + 4(2n-7)!! \\ &\leq 6(n-2)! + 4(2n-7)!! \end{aligned}$$

Therefore the proportion of trees maximally distant from a given tree is less than:

$$\begin{aligned} \frac{6(n-2)! + 4(2n-7)!!}{ub(n)} &= \frac{6(n-2)! + 4(2n-7)!!}{(2n-5)!!} \\ &= \frac{6(n-2)!}{(2n-5)!!} + \frac{4(2n-7)!!}{(2n-5)!!} \\ &= \frac{6(n-2)!}{(2(n-2)-1)!!} + \frac{4}{2n-5}. \end{aligned}$$

The results follows from taking the limit of this expression as n approaches infinity. $\square$

4. Conclusion

As the compilation of a variety of trees with identifiable leaf sets becomes an ever more common tool in phylogenomics, we suggest that scientists consider the path interval distance as an alternative or supplement to Robinson-Foulds or NNI distances both in the design and testing of consensus methods.

It should be rare for a pair trees to be maximally distant from one another under the path interval distance. This in itself is not sufficient to suggest that the path interval distance be an alternative to the Robinson-Foulds distance when comparing trees in simulation studies. However, Xi et al. ([20]) have simulated the distribution of the path interval distance among random twelve taxa trees and noted that they resemble a Gaussian distribution. Further, they have constructed both lower and upper bounds

on the expected value of this distribution for pairs of trees drawn from the uniform distribution on the space of phylogenetic trees. We refer the reader to this sequel which also examines the distribution of path interval distances between a species tree and gene trees generated under the coalescent process.

When biologists seek to reconstruct the deep history of life on earth it is important that relationships between larger clades be accurate, and that the tree have a high degree of global accuracy. Consider C_{3x} and D_{3x} from Figure 1. For large values of x the Robinson-Foulds distance will be quite large, while C_{3x} and D_{3x} will still have a path interval distance of two. These large trees would be assigned a worse "score" based on the Robinson-Foulds distance. The fact that $d(C_{3x}, D_{3x}) = 2$ however, tells us that these trees are actually similar to one another. While they may not be a perfect match on a local level, we would likely want their scores to be very good since their topologies match and they exhibit large degree of global congruence. The Robinson-Foulds distance does not facilitate this as well as the path interval distance. On the other hand the path interval distance does not help us quantify the number of errors on the tree, a task perhaps better suited to the Robinson-Foulds distance.

Refrégier et. al. [27] constructed phylogenetic trees for Caryophyllaceae and Microbotryum (hosts and parasites, respectively) using a Bayesian 50% majority rule consensus tree approach. They noted "some broad-scale congruence between the host and parasite phylogenies appears by visual inspection... while fine-scale congruence between the host and parasite phylogenies is weak, in particular among the parasites of *Silene*." They were able to determine this through visual inspection, as their trees were relatively small. The path interval distance provides a mechanism for quantifying their observation.

We believe that path interval distance also has applications to research in the coalescent model of evolution. A phylogenomic reconstruction method was recently proposed by [28] in which a variation of the Robinson-Foulds distance was used as a guiding principle in reconstructing species trees from gene trees. Given the seemingly Gaussian nature of the path interval distance noted in [20], we suggest in future work that phylogenomic reconstruction algorithms consider leveraging the path interval distance.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1358534. Research reported in this publication was supported by an Institutional Development Award (IDeA) from the National Center for Research Resources (5 P20 RR016461) and the National Institute of General Medical Sciences (8 P20 GM103499) from the National Institutes of Health.

References

- [1] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyML 3.0, *Systematic Biology* 59 (3) (2010) 307–321.
- [2] Y. Lin, V. Rajan, B. M. Moret, Fast and accurate phylogenetic reconstruction from high-resolution whole-genome data and a novel robustness estimator, *Journal of Computational Biology* 18 (9) (2011) 1131–1139.
- [3] W. Salzburger, G. B. Ewing, A. Von Haeseler, The performance of phylogenetic algorithms in estimating haplotype genealogies with migration, *Molecular Ecology* 20 (9) (2011) 1952–1963.

- [4] M. S. Bansal, J. Dong, D. Fernández-Baca, Comparing and aggregating partially resolved trees, *Theoretical Computer Science* 412 (48) (2011) 6634–6652.
- [5] D. Bryant, A classification of consensus methods for phylogenetics, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 61 (2003) 163–184.
- [6] R. Libeskind-Hadas, Y.-C. Wu, M. S. Bansal, M. Kellis, Pareto-optimal phylogenetic tree reconciliation, *Bioinformatics* 30 (12) (2014) 87–95.
- [7] C. Conow, D. Fielder, Y. Ovadia, R. Libeskind-Hadas, Jane: a new tool for the cophylogeny reconstruction problem, *Algorithms for Molecular Biology* 5 (1) (2010) 16.
- [8] P. Huggins, M. Owen, R. Yoshida, First steps toward the geometry of cophylogeny, in: T. Hibi (Ed.), *Harmony of Gröbner Bases and the Modern Industrial Society*, Singapore, 2012, pp. 99–116.
- [9] Y. Ovadia, D. Fielder, C. Conow, R. Libeskind-Hadas, The cophylogeny reconstruction problem is NP-complete, *Journal of Computational Biology* 18 (1) (2011) 59–65.
- [10] J. Heled, A. J. Drummond, Bayesian inference of species trees from multilocus data, *Molecular Biology and Evolution* 27 (3) (2010) 570–580.
- [11] J. Hughes, M. Kennedy, K. P. Johnson, R. L. Palma, R. D. Page, Multiple cophylogenetic analyses reveal frequent cospeciation between pelecyaniform birds and pectinopygus lice, *Systematic Biology* 56 (2) (2007) 232–251.
- [12] L. Liu, L. Yu, Estimating species trees from unrooted gene trees, *Systematic Biology* 60 (5) (2011) 661–667.
- [13] S. Mirarab, R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, T. Warnow, Astral: Genome-scale coalescent-based species tree estimation, *Bioinformatics* 30 (17) (2014) i541–i548.
- [14] J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent, *Trends in Ecology & Evolution* 24 (6) (2009) 332–340.
- [15] R. D. Page, *Tangled trees: phylogeny, cospeciation, and coevolution*, University of Chicago Press, 2003.
- [16] M. A. Steel, D. Penny, Distributions of tree comparison metrics some new results, *Systematic Biology* 42 (2) (1993) 126–141.
- [17] W. Williams, H. Clifford, On the comparison of two classifications of the same set of elements, *Taxon* (1971) 519–522.
- [18] D. Bryant, M. Steel, Computing the distribution of a tree metric, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* 6 (3) (2009) 420–426.
- [19] N. D. Pattengale, E. J. Gottlieb, B. M. Moret, Efficiently computing the robinson-foulds metric, *Journal of Computational Biology* 14 (6) (2007) 724–735.
- [20] J. Xi, J. Xie, R. Yoshida, Distributions of topological tree metrics between a species tree and a gene tree, *arXiv preprint arXiv:1501.02374*.
- [21] G. W. Moore, M. Goodman, J. Barnabas, An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets, *Journal of Theoretical Biology* 38 (3) (1973) 423–457.
- [22] D. F. Robinson, Comparison of labeled trees with valency three, *Journal of Combinatorial Theory, Series B* 11 (2) (1971) 105–119.
- [23] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, L. Zhang, On distances between phylogenetic trees, in: *SODA*, Vol. 97, Citeseer, 1997, pp. 427–436.
- [24] D. F. Robinson, L. R. Foulds, Comparison of phylogenetic trees, *Mathematical Biosciences* 53 (1) (1981) 131–147.
- [25] C. Finden, A. Gordon, Obtaining common pruned trees, *Journal of Classification* 2 (1) (1985) 255–276.
- [26] S. Böcker, S. Canzar, G. W. Klau, The generalized Robinson-Foulds metric, in: *Algorithms in Bioinformatics*, Springer, 2013, pp. 156–169.
- [27] G. Refrégier, M. Le Gac, F. Jabbour, A. Widmer, J. A. Shykoff, R. Yockteng, M. E. Hood, T. Giraud, Cophylogeny of the anther smut fungi and their Caryophyllaceae hosts: prevalence of host shifts and importance of delimiting parasite species for inferring cospeciation, *BMC Evolutionary Biology* 8 (1) (2008) 100.
- [28] R. Chaudhary, J. G. Burleigh, D. Fernández-Baca, Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance., *Algorithms for Molecular Biology* 8 (1) (2013) 28.