Comparing Related Phylogenetic Trees* (Communication)

Tiziana Calamoneri[1], Angelo di Mambro[1], and Blerina Sinaimeri[2,3]**

[1] Computer Science Department, Sapienza University of Rome, Italy
[2] INRIA Grenoble Rhône-Alpes, France
[3] UMR CNRS 5558 - LBBE, Université Lyon 1, France

Abstract. In phylogenetics, several classical distances exist to compare two phylogenetic trees. However, when the evolution in one tree has been influenced by the evolution in the other (*e.g.* two ecologically linked groups of organisms as hosts and their symbionts), other methods are more appropriate to compare the trees. Among the most used ones, there is *phylogenetic tree reconciliation*, *i.e.* mapping of one tree into the other according to certain rules, with a quantification of its quality; we refer to distances based on this concept as *reconciliation distances.* They bring useful information but are unfortunately NP-hard to be computed. It is then interesting to understand whether a polynomial phylogenetic tree distance is correlated to the reconciliation distances.

In this communication we announce a systematic study to compare classical and reconciliation distances and we show that there is not much correlation between them. We then introduce a new distance that is instead correlated with the reconciliation distances and can be computed in polynomial time, hence it represents an efficient alternative to them.

Keywords: phylogenetic trees, tree distances, reconciliation methods.

1 Introduction

The field of phylogenetics has become the central underpinning of research in many areas of biology. Indeed, knowledge of phylogenetic relationships has many important applications such as drug discovery, identifying and tracing the origins of emerging infectious diseases, or guiding genetic improvements in agriculture.

A *phylogenetic tree* is a rooted full binary tree whose leaves represent taxa while the internal nodes the possible ancestors that might have led through evolution to this set of taxa. Comparing phylogenetic trees is a major task in phylogenetic research: comparisons are necessary in different situations as for

** This work has been partially done while the third author was a visiting professor of Sapienza. Partially supported by Sapienza University of Rome, Project "Comparative Analysis of Phylogenies".

example when the solutions provided by different reconstruction methods disagree, when trees derived from different genes are incongruent or when there are phylogenetic trees of ecologically linked groups of organisms. A natural way to compare pairs of phylogenetic trees is to apply a similarity or dissimilarity measure and many distances have been proposed in the literature. Among the most used ones (because they can be computed in polynomial time) there are the Robinson-Foulds distance [12, 7], the distance based on the maximum agreement subtree [13, 5], the triplet distance [3] and the path distance [14]. We will refer to these distances as *classical distances*.

In the cases where we need to compare phylogenetic trees that are derived from different genes [9, 16] or when comparing phylogenetic trees of ecologically linked groups of organisms (like for example host species and their symbionts) [4, 8] some different methods to asses the difference between trees have been introduced. In particular, the amount of dissimilarity between two trees is measured by mapping one tree into the other and quantifying the quality of this mapping according to certain costs. Such a mapping (called *reconciliation*) is a function ρ that maps each internal node of the first tree to a node of the second tree and allows the unique identification of four main macro-evolutionary events: cospeciation, duplication, host switch (or horizontal gene transfer in the case of gene/species), and loss. The amount of incongruences between the trees is found by assigning a cost to each of the four types of events and then seek to find the reconciliation of minimum total cost. We will refer to these distances as *reconciliation distances*.

Nevertheless, the computed reconciliations can sometimes be time-inconsistent, i.e, the inferred host switches may induce contradictory constraints on the dates for the internal nodes of the trees. The problem of finding an optimal time-consistent reconciliation is known to be NP-hard [16, 11] if the first tree is not previously fully dated (that is notoriously difficult); on the contrary, when this constraint is dropped, the problem requires time proportional to the square of the dimension of the trees (*e.g.* [6, 1, 15, 8]).

In the literature there are many comparative studies that analyze the performance of the classical distances (e.g. [14, 2, 10]). These studies compare the distributions of the values of these distances and the possible correlations between them. On the contrary, the reconciliation distances are poorly explored. Indeed, up to our knowledge, the paper [17] is the only one studying reconciliation distances but it considers only a very special case that is very close to classical distances.

In this communication we announce a systematic study to compare a set of most used reconciliation distances among them and to classical distances. We show that there is not much correlation between reconciliation and classical distances. Hence, we introduce a new distance that is inspired by reconciliation distances but does not compute reconciliations, so guaranteeing on the one hand polynomial time to be computed, and on the other hand correlation with reconciliation distances.

2 Classical and reconciliation distances

A phylogenetic tree is a rooted binary tree whose leaves correspond to the considered taxa and whose internal nodes have degree exactly three (except the root, having degree two). For a tree T, we call $V(T)$, $L(T)$ and $V_I(T)$, the set of its nodes, its leaves and its internal nodes, respectively. All the distances below are considered for pairs of trees defined on the same set of leaves L. Furthermore, they are all normalized in order to obtain a value in $[0, 1]$.

2.1 Classical distances

We will focus on the following distances that are defined for rooted and binary trees, that are most used in biology and that can be computed in polynomial time.

Robinson-Foulds (RF) distance [12] : For every $u \in V_I(T)$, we define the *cluster* of u as $C(u) = \{l \,|\, l \text{ is a leaf of } (T(u))\}$. Let $\mathcal{C}(T) = \{C(u) | u \in V_I(T)\}$.

$$d_{rf}(T_1, T_2) = \frac{|\mathcal{C}(T_1)/\mathcal{C}(T_2)| + |\mathcal{C}(T_2)/\mathcal{C}(T_1)|}{2}.$$

Maximum Agreement Subtree (MAST) distance [13, 5]: Let $L' \subseteq L(T_1))$. The subtree of T_1 induced by L', $S_{T_1}(L')$ is an *agreement subtree of* T_1 *and* T_2 if and only if it is isomorphic to the subtree $S_{T_2}(L')$. The *maximum agreement subtree problem* asks for the largest agreement subtree of T_1 and T_2, and the number of its leaves is denoted by $MAST(T_1, T_2)$.

$$d_{mast}(T_1, T_2) = n - MAST(T_1, T_2).$$

Triplet distance [3]: A *triplet* is a set $\{i, j, k\} \subseteq L$. The triplet distance $d_t(T_1, T_2)$ between T_1 and T_2 is defined as the number of triplets whose topology differ in the two trees.

Path distance: [14]: Let $l_T(v, w)$ denote the distance between v and w in T. The path distance d_p between two rooted phylogenetic trees T_1 and T_2 is equal to:

$$d_p(T_1, T_2) = \sum_{i,j \in L} |l_{T_1}(i, j) - l_{T_2}(i, j)|.$$

2.2 Reconciliation distances

To present the reconciliation method we will take as an example the host-symbiont context. Specifically, in this case, two phylogenetic trees T_1 and T_2 representing the evolution of symbionts and their host species respectively, and a function σ mapping the leaves of T_1 to the leaves of T_2, (σ presents the nowadays infections) are given in input. In our special case in which the trees to be compared have the same leaf set, σ is a bijection. We recall here (sometimes in an informal way, in order not to overburden the exposition) some basic notions on reconciliations while the formal and complete definitions can be found *e.g.* in [16]. Given T_1, T_2, σ, a *reconciliation* ρ is a function $\rho : V(T_1) \to V(T_2)$ that extends σ (*i.e.* $\rho(v) = \sigma(v)$ for all $v \in L(T_1)$) and satisfies some biologically motivated constrained. A reconciliation ρ associates each internal node v of T_1 to an event among: *cospeciation* (when the children of v are mapped by ρ one in the left subtree of $\rho(v)$ and one in the right subtree of $\rho(v)$), *duplication* (when the children of v are mapped by ρ either both in the left subtree of $\rho(v)$ or both in the right subtree of $\rho(v)$) and *host-switch* (when a child of v is mapped by ρ in the subtree rooted at $\rho(v)$ while the other child is mapped in a node that is neither a descendent nor an ancestor of $\rho(v)$), while each arc (u, v) of T_1 is associated to a certain number of loss events $l_{(u,v)} \geq 0$ that is equal to $l_{T_2}(\rho(u), \rho(v))$ if $\rho(u)$ is an ancestor of $\rho(v)$. Given a vector $C = \langle c_c, c_d, c_s, c_l \rangle$ of real values that correspond to the costs of each type of event, the most parsimonious (or optimal) reconciliations are the ones that minimize the total cost, *i.e.* $\min_\rho cost(\rho) = \sum_{i \in \{c,d,s,l\}} e_i \, c_i$, where e_i is the number of events of type i in ρ. We use the value of the optimal cost as a measure of similarity between the trees. We consider 7 different reconciliation distances by varying the cost of the events.

2.3 A new distance

We introduce a new distance that we call *related subtree (RS)* which tries to keep into account the relation between T_1 and T_2 but without performing the time consuming computation of a reconciliation. Let v be an internal node of T_1 and L_v the set of leaves in $T_1(v)$ (*i.e.* the subtree of T_1 rooted in v); consider now the same leaf set L_v in T_2 and let $T_2(L(v))$ be the smallest subtree of T_2 that connects all the leaves in $L(v)$, notice that $T_2(L(v))$ is rooted at the *lca* in T_2 of the leaf set $L(v)$ and can have nodes of degree 2, so it does not coincide with the subtree induced by $L(v)$ in T_2 and it is neither a phylogenetic tree. We hence define the following measure:

$$d_{rs}(T_1, T_2) = \sum_{v \in V_I(T_1)} |diam(T_1(v)) - diam(T_2(L(v)))|.$$

3 Computational results

To evaluate and compare the performance of these distances, we carry out two different types of experiments. For each of the experiments $N = 1000$ pairs of randomly chosen phylogenetic trees on $n = 20, 25,$ and 30 leaves are considered.

Distributions and Correlation The first set considers the distribution of the values of these distances. For each one of the 1000 pairs of trees, the normalized values of all the 12 distances previously defined, were computed. In general, the results indicate that the reconciliation distances have similar distributions that are near to a normal distribution. On the other side, the classical distances behave differently. In particular the distributions of the RF and the triplet distances appear shifted to the right, implying that most of the pairs of trees are far apart in these distances. This confirms what is known about the behavior of RF for phylogenetic trees [14]. The path distance and the RS distance seem to have a distribution similar to the normal distribution. The distributions suggest that the path, RS and the reconciliation distances are better at discriminating between trees since they attain a larger spread of values.

We also studied the correlation among the distances. It is not surprising that a correlation among many of the reconciliation distances rd_C exists as these distances differ only by the choice of the costs, *e.g.* there is a strong correlation among rd_{0231} and rd_{0121}. The classical distances do not appear correlated among them. This is also expected from what is known in the literature. The results show that the RS distance seems the one that is better correlated to the reconciliation distances. This again suggests that the RS distance could be a good alternative to the reconciliation distances.

Comparison between Classical and Reconciliation Distances A second set of experiments arises from the observation that requiring a correlation could be too much in order to state that two measures behave similarly; namely, observing what is needed in contexts like gene-species reconciliations, fixing a (species) tree T_1, and given many (gene) trees $T_2^1, \ldots, T_2^k$, we say that two distances d' and d'' *agree* if every time that $d'(T_1, T_2^i) \leq d'(T_1, T_2^j)$ then it also holds that $d''(T_1, T_2^i) \leq d''(T_1, T_2^j)$ and every time that $d'(T_1, T_2^i) \geq d'(T_1, T_2^j)$ then it also holds that $d''(T_1, T_2^i) \geq d''(T_1, T_2^j)$. Even if d' and d'' have a low correlation index, if they agree on all the pairs of gene trees T_2^i, T_2^j, then they will both determine the same (gene) tree as the closest to T_1. It is hence worth to compute the percentage of agreement between pairs of measures. We fixed a tree T_1 and generated 1000 pairs of trees (T_1, T_2). For each pair of trees the values of the 12 distances were computed. From the results, it came out that –among the polynomially computable distances– the RS distance is the one that agrees more with the reconciliation distances.

4 Conclusions

In this communication we announce a systematic study to compare a set of most used classical distances with a set of distances based on the reconciliations. We show that there is not much correlation in between. Hence, we introduce a new distance that turns out to be correlated with the reconciliation distances and hence can provide an alternative to classical distances. As a future direction it is interesting to extend this studies to mul-labeled trees (*i.e.* trees where more than one leaf may be labeled with the same label). This will allow to test these studies to real datasets where mul-labeled trees are common (*e.g.* it is common that one symbiont species is associated to more than one host species).

References

1. Bansal, M.S., Alm, E., Kellis, M.: Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. Bioinformatics **28**(12), i283–i291 (2012)
2. Bogdanowicz, D., Giaro, K., B, W.: Treecmp: comparison of trees in polynomial time. Evolutionary Bioinformatics **8**, 475–487 (2012)
3. Brodal, G., Fagerberg, R., Mailund, T., Pedersen, C., Sand, A.: Computing the triplet and quartet distance between trees of arbitrary degree. In Proceedings of the annual ACM-SIAM Symposium on Discrete Algorithms (SODA) pp. 1814–1832 (2013)
4. Charleston, M.A.: Jungles: a new solution to the host/parasite phylogeny reconciliation problem. Mathematical Biosciences **149**(2), 191–223 (May 1998)
5. Cole, R., Farach-Colton, M., Hariharan, R., Przytycka, T., Thorup, M.: An $O(nlogn)$ algorithm for the maximum agreement subtree problem for binary trees. SIAM J. Comput. **30**(5), 1385–1404 (2000)
6. Conow, C., Fielder, D., Ovadia, Y., Libeskind-Hadas, R.: Jane: a new tool for the cophylogeny reconstruction problem. Algorithms for Molecular Biology **5**(16), 10 pages (February 2010)
7. Day, W.: Optimal algorithms for comparing trees with labeled leaves. Journal of Classification **2**(1), 7–28 (1985)
8. Donati, B., Baudet, C., Sinaimeri, B., Crescenzi, P., Sagot, M.: Eucalypt: efficient tree reconciliation enumerator. Algorithms for Molecular Biology **10**(1), 3 (2015)
9. Doyon, J.P., Hamel, S., Chauve, C.: An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. IEEE/ACM Transactions on Computational Biology and Bioinformatics **9**(1), 26–39 (2011)
10. Kuhner, M.K., Yamato, J.: Practical performance of tree comparison metrics. Systematic Biology **64**(2), 205–214 (2015)
11. Libeskind-Hadas, R., Charleston, M.A.: On the computational complexity of the reticulate cophylogeny reconstruction problem. Journal of Computational Biology **16**(1), 105–117 (2009)
12. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Math. Biosci. **55**, 131–147 (1981)
13. Steel, M., Warnow, T.: Kaikoura tree theorems: Computing the maximum agreement subtree. Inform. Process. Lett. **48**, 77–82 (1993)
14. Steel, M.A., Penny, D.: Distributions of tree comparison metrics: Some new results. Systematic Biology **42**(2), 126–141 (1993)
15. Stolzer, M.L., Lai, H., Xu, M., Sathaye, D., Vernot, B., Durand, D.: Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics **28**(18), i409–i415 (2012)
16. Tofigh, A., Hallett, M., Lagergren, J.: Simultaneous identification of duplications and lateral gene transfers. Journal of IEEE/ACM Transactions on Computational Biology and Bioinformatics **8**(2), 517–535 (2011)
17. Zheng, Y., Zhang, L.: Are the duplication cost and Robinson-Foulds distance equivalent? Journal of computational biology : a journal of computational molecular cell biology **21**(8), 578–90 (2014)