

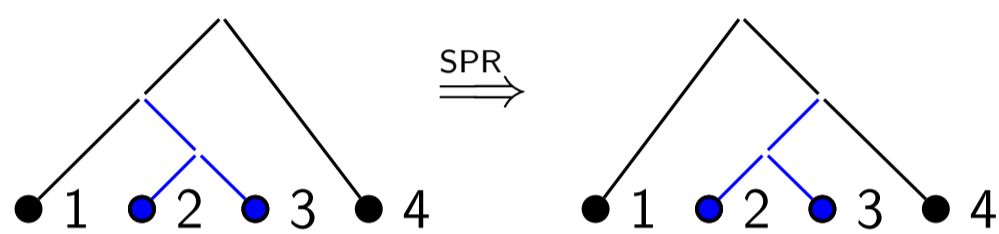
Abstract

In evolutionary biology one is interested in computing various distance-measures between certain trees. One of these measures is induced by subtree-prune-and-regraft (SPR) operations. The SPR-distance between phylogenetic trees can be expressed in terms of the number of trees in a maximum agreement forest. However, this does not hold for fully-labeled trees. An approach to computing the SPR-distance between fully-labeled trees is to run a guided breadth-first-search in the corresponding network-graph.

The SPR-Distance between Phylogenetic Trees

A **phylogenetic tree**, also called evolutionary tree, is a rooted binary leaf-labeled tree. It is used in biology to model evolutionary relationships between different species, i.e the leaves represent the extant species and the structure of the tree corresponds to their ancestral relationship.

To measure the dissimilarity of two phylogenetic trees the **SPR-distance** counts the minimum number of **subtree-prune-and-regraft (SPR) operations** that are necessary to transform one tree into another. Such an SPR operation consists of pruning a subtree and regrafting it at a preexisting edge in the remaining tree.



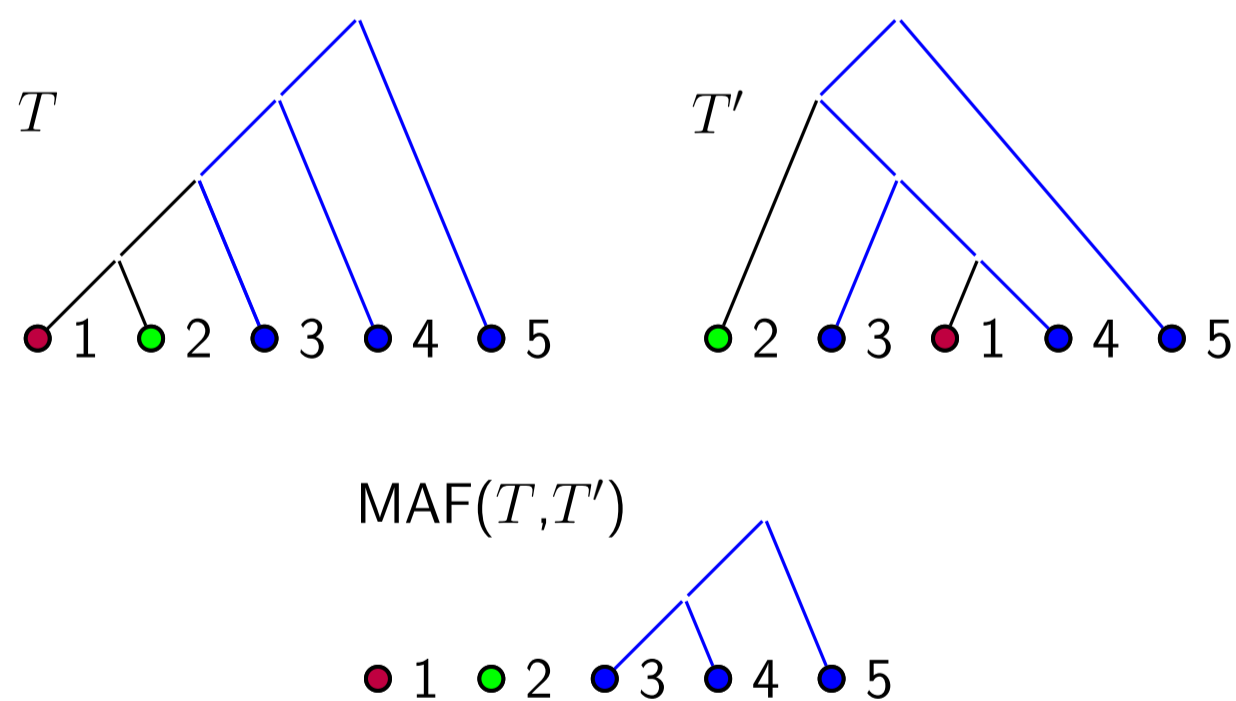
The figure above shows an example of an SPR operation where the subtree containing the leaves labeled 2 and 3 is pruned and reattached at the edge above the leaf labeled 4.

Computing the SPR-distance is NP-hard!

Maximum Agreement Forests

An **agreement forest (AF)** for two phylogenetic trees is a forest of phylogenetic trees that appears in both of them; a **maximum agreement forest (MAF)** is an agreement forest with a minimal number of trees.

A maximum agreement forest can be obtained by deleting a minimum number of edges in either one of the given trees such that every tree in the forest appears in both of them.



The above figure shows two phylogenetic trees and a maximum agreement forest for them which consists of three trees. The colors illustrate how the three trees in this maximum agreement forest appear in T and T' . Also T' can be obtained from T by two SPR operations.

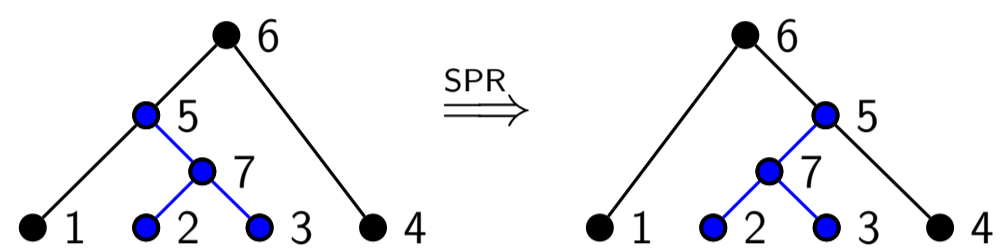
If maf denotes the number of trees in a maximum agreement forest, then for any two phylogenetic trees T and T' on the same label set:

$$\mathit{dist}_{SPR}(T, T') = \mathit{maf}(T, T') - 1$$

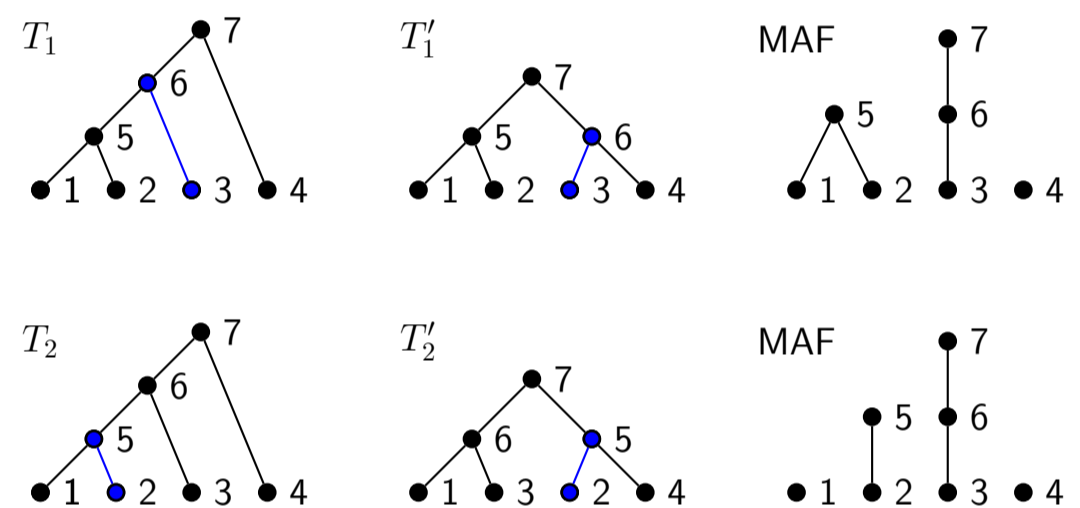
This beautiful correspondence is the main tool used in many proofs concerning the SPR-distance. It is the key idea to express the SPR-distance problem as an **integer linear programming problem (ILP)** and to develop an **approximation algorithm** for computing the SPR-distance with approximation-ratio 3.

The SPR-Distance between Fully-Labeled Trees

A **fully-labeled tree** is a rooted binary tree where all vertices are labeled. Analogously to the case of phylogenetic trees, SPR operations and an induced SPR-distance can be defined for fully-labeled trees.



Unfortunately, the correspondence between the SPR-distance and the size of a maximum agreement forest does not hold for fully-labeled trees:

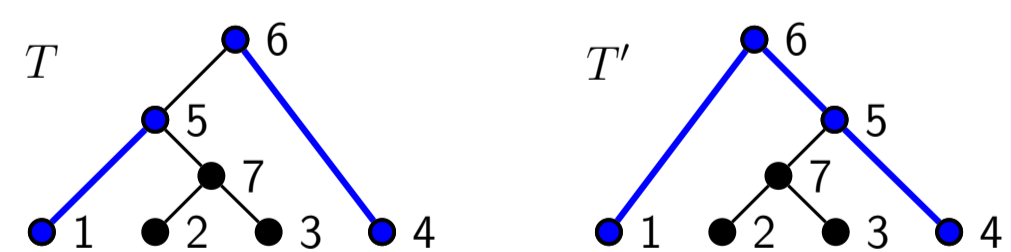


Tree T'_1 as well as tree T'_2 can be obtained by pruning and regrafting the edge colored in blue. But the MAF for T_1 and T'_1 consists of only 3 trees whereas the MAF for T_2 and T'_2 consists of 4 trees.

Computing the SPR-Distance by BFS

The SPR-distance between two fully-labeled trees T and T' can be computed by **breadth-first-search (BFS)** in the corresponding **network graph**, i.e. the graph with all fully-labeled trees on the same label sets as T and T' as vertices and an edge between two vertices if and only if their corresponding trees can be transformed into each other by exactly one SPR operation.

Such a computation can be made much faster, if it is not necessary to search all neighbors. Therefore one is interested in a parameter that indicates which directions are good. The **difference-path** of T and T' is defined as the union of all edges in the paths of T' induced by the edges in $E(T) \setminus E(T')$.



Conjecture:

For any two fully-labeled trees on the same set of leaf-labels and interior-labels there is at least one sequence of SPR operations that transforms one tree into the other on which the number of edges in the diff-path decreases.

References

1. B. L. Allen and M. Steel, *Subtree Transfer Operations and Their Induced Metrics on Evolutionary Trees*, Annals of Combinatorics 5 (2001) 1-15
2. M. Bordewich, C. McCartin and C. Semple, *A 3-Approximation Algorithm for the Subtree Distance between Phylogenies*, Journal of Discrete Algorithms Vol. 6 no. 3 (2008), pp. 458-471
3. M. Bordewich and C. Semple, *On the Computational Complexity of the Rooted Subtree Prune and Regraft Distance*, Annals of Combinatorics 8 (2004), pp. 409-423
4. J. Hein, T. Jiang, L. Wang and K. Zhang, *On the complexity of comparing evolutionary trees*, Discrete Applied Mathematics 71 (1996), pp. 153-169
5. Y. Wu, *A practical method for exact computation of subtree prune and regraft distance*, Bioinformatics Vol. 25 no. 2 (2009), pp. 190-196