Evolutionary costs in gene-species reconciliation

Paweł Górecki
gorecki@mimuw.edu.pl
University of Warsaw, PL

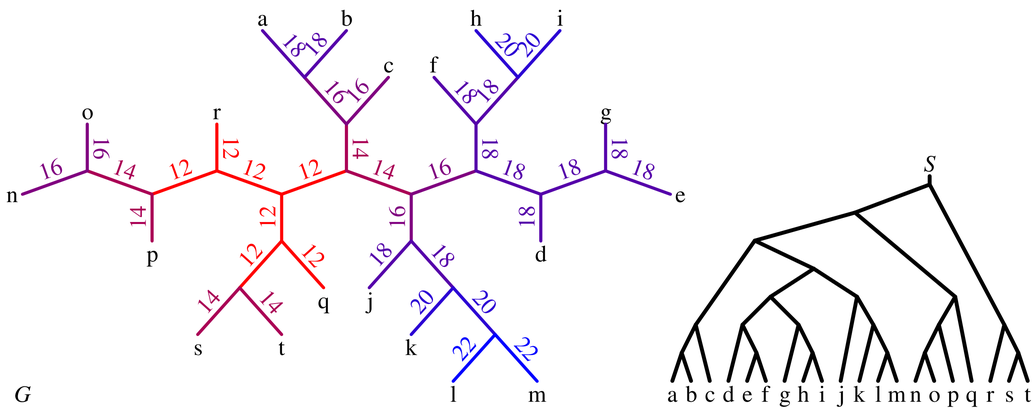
Oliver Eulenstein
oeulenst@iastate.edu
Iowa State University, USA

Jerzy Tiuryn
tiuryn@mimuw.edu.pl
University of Warsaw, PL

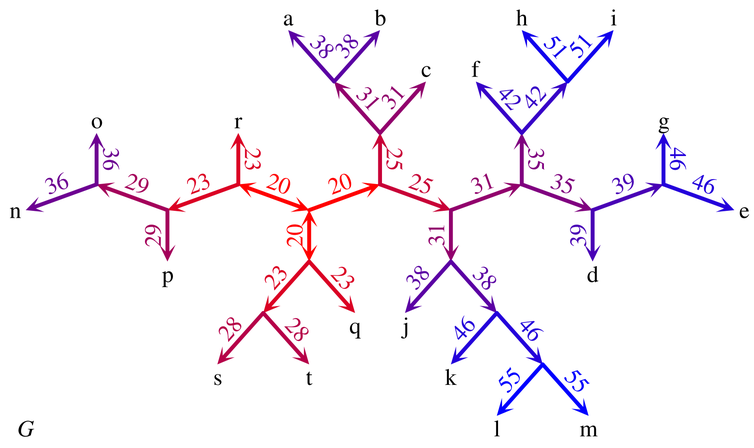
Introduction

Scoring differences between gene and species trees is indispensable for evolutionary studies. While many gene trees are unrooted and species trees are typically rooted, standard scoring tools do not accept such inputs. We present a novel framework that allows to compare rootings of the unrooted gene tree with the species trees by using rooted scoring functions such as Robinson-Foulds (RF), gene duplications (D), gene losses (L), duplication+losses (DL) and deep coalescence (DC).

Cost examples

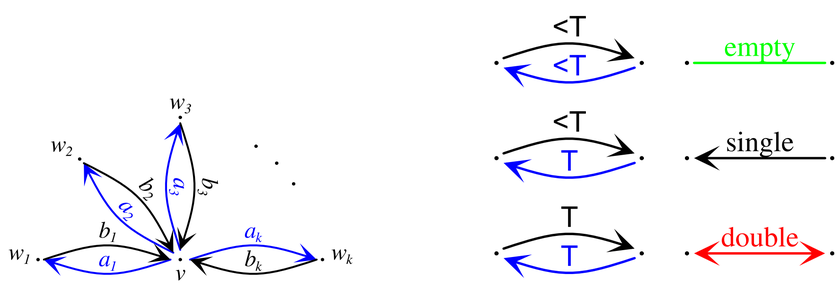


RF cost example. Each edge of the unrooted gene tree G is labeled by the RF score between the species tree S and the rooting of G placed on this edge. There are 7 optimal rootings forming a "valley" with the RF score 12.

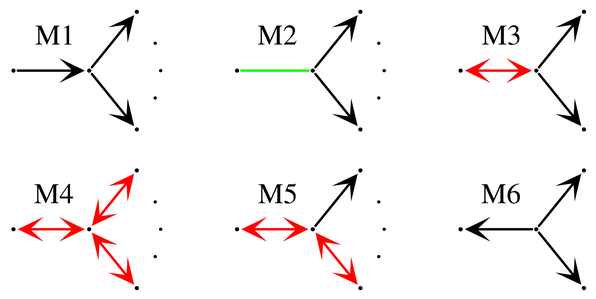


DL cost example with the star-like topology for the trees from the previous figure. There are 3 optimal edges having cost 20.

Gene trees and star topologies



Transforming a gene tree into a star topology. Required: lca-mappings from the rooted subtrees of the input gene tree into a given species tree S.



There are only six types of multifurcated stars that can be present in gene trees. M5 has at least 2 double edges and at least 1 single edge. Star M6 is not present if both trees have multifurcations.

Properties of stars and optimal regions

- Star topology of a given gene tree depends on lca-mappings only.
- Each double and empty edge is optimal for all five costs.
- All edges of star M6 are optimal.

We have the following relationships between optimal regions of gene trees under our cost functions: $D \supseteq DL = L \subseteq DC \subseteq RF$ [1,2,3].

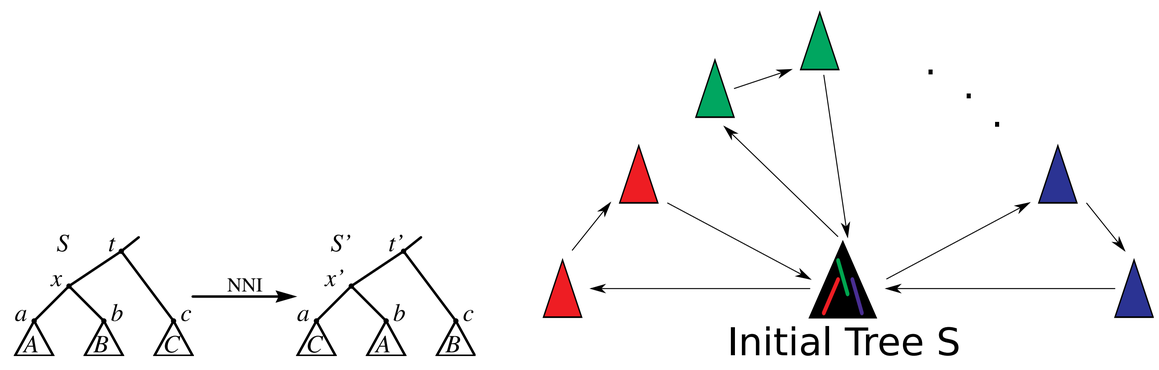
Given unrooted gene tree G and a rooted species tree S, the optimal "unrooted" cost (RF, DC, DL, L and D) as well as the costs of all rootings of G can be computed in $O(n)$ time, where n is the size of input trees [1,2,3].

Applications: improving supertree heuristics

Gene tree parsimony (GTP) problems: given a collection of gene trees and a cost function find an optimal species tree that minimazes the total cost.

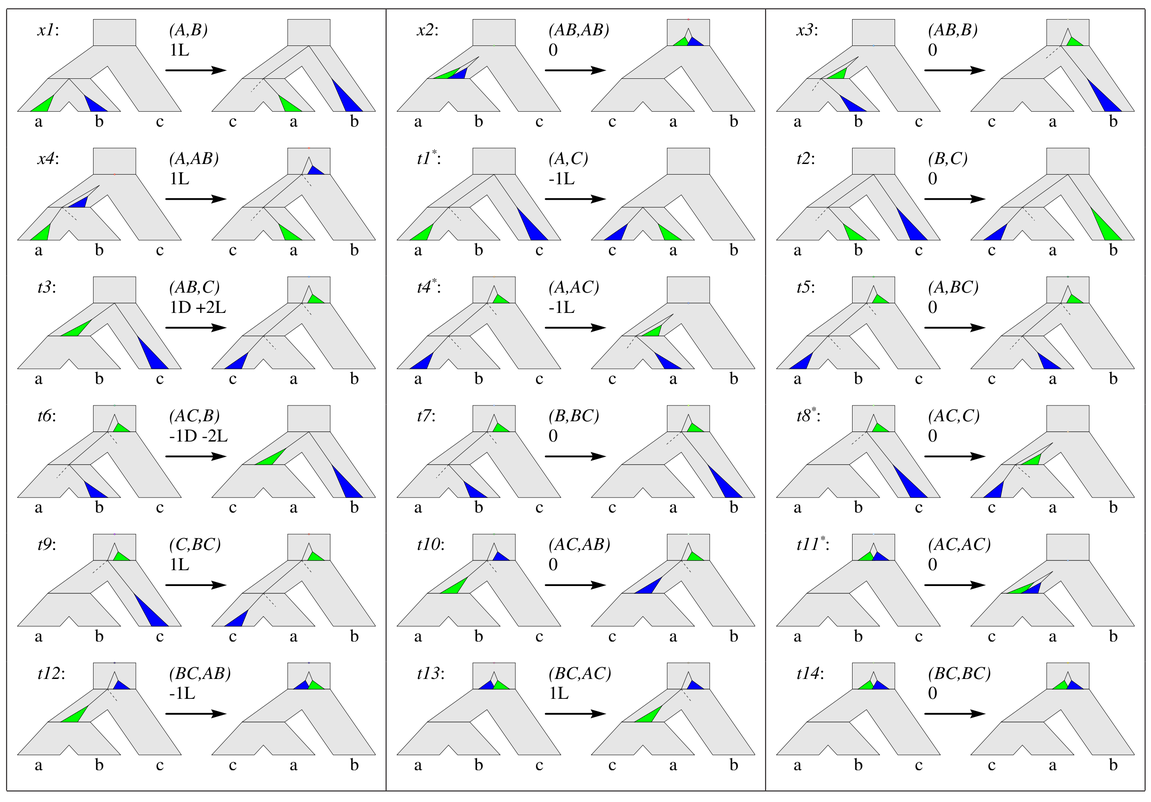
GTP problems are usually NP-complete. They often are addressed by local search heuristics that perform a stepwise search of the tree space, where each step is guided by an exact solution to an instance of *a local search problem*.

Local search problem (NNI): given an unrooted gene tree G and a rooted species tree S find an optimal species tree S' in the 1-NNI neighborhood of S.



Fast local search algorithm

Reconciliation algorithm based on star topologies can be adopted to design fast local search algorithms.



The NNI operation changes locally the embedding of a rooting of G into S. This property allows the efficient detection of the new positions of gene duplications and gene losses in the species tree. In consequence, the local search problem for the DL cost (and other costs) can be solved in $O(n)$ time [4]. GTP software is available at `http://bioputer.mimuw.edu.pl/gorecki/fasturec`.

References

[1] P. Górecki and J. Tiuryn. Inferring phylogeny from whole genomes. Bioinformatics, 23(2):e116-22, 2007. [2] P. Górecki and O. Eulenstein. A Robinson-Foulds measure to compare unrooted trees with rooted trees. LNCS, 7292:102-114, 2012. [3] P. Górecki and O. Eulenstein. Deep coalescence reconciliation with unrooted gene trees: Linear time algorithms. LNCS 7434, accepted to COCOON 2012. [4] P. Górecki and O. Eulenstein. GTP supertrees from unrooted gene trees: linear time algorithms for NNI based local searches. LNCS, 7292:83-105, 2012.

This work was supported by MNiSW (#N N301 065236), National Science Foundation (#0830012 and #106029), NCN (#2011/01/B/ST6/02777) and the NIMBioS Working Group "Gene Tree Reconciliation".