RF+ software version 0.2

Description

RF+ is a prototype program for computing RF(+) distances between phylogenetic trees. RF(+) distance is designed to more meaningfully compute the Robinson-Foulds distance between two trees that only have a partially overlapping leaf set. The traditional approach for computing Robinson-Foulds distance between two trees that only have a partially overlapping leaf set is to first restrict the two trees to their shared leaf set and then compute their Robinson-Foulds distance. We refer to distances computed in this way as RF(-) distances. In contrast, the RF(+) distance between two arbitrary trees is computed by first optimally completing each tree on the union of the leaf sets of both trees so as to minimize the Robinson-Foulds distance between them, and then reporting the Robinson-Foulds distance between the two completed trees.

The current prototype implements the algorithms described in the manuscript cited below and can (i) compute the RF(+) distance between a pair of trees where the leaf set of one of the trees is a subset of the leaf set of the other, and (ii) compute the Extraneous-Clade-Free-RF(+) (EF-RF(+)) distance between two trees with arbitrary leaf sets. All trees must be rooted and binary.

We refer the reader to the following paper for further details on RF(+) and EF-RF(+) distance.

"Linear-Time Algorithms for some Phylogenetic Tree Completion Problems under Robinson-Foulds Distance", Mukul S. Bansal, *RECOMB Comparative Genomics Conference (RECOMB-CG) 2018*; LNCS 11183: 209-226.

Implementation details and requirements

RF+ is implemented in Python and requires version 3.0 or greater. The implementation also assumes that ETE 3 toolkit is already installed. ETE toolkit is available freely from etetoolkit.org

We point out that while the algorithms presented in the manuscript cited above have linear, O(n), time complexity, this current implementation of RF+ has O(n log n) time complexity since it implements a slightly suboptimal algorithm for Least Common Ancestor (LCA) computation.

RF+ is freely available open source under GNU GPL.

Usage

RF+ takes as input two or more trees and it compares the first tree with every other tree in the input file. All input trees must be in newick format, must be rooted and binary, and must be in a single input file with each tree appearing on a separate line. By default, the program outputs only the optimal completions. These completions are output in pairs, with each pair consisting of a completion of the first tree and another input tree (considered in order of their appearance in the input file). Note that if the first tree already contains all leaves present in the other tree then only the other tree is completed and the first tree is output as-is. The input file is specified using the "-i" option. An output file (optional) can be specific using the "-o" option. For example,

python3 RF+.py -i input.newick -o output.txt

The “-r” option can be used to output the RF(-) and RF(+) distances between the first tree and each of the other trees. For example,

python3 RF+.py -i input.newick -o output.txt -r

Example dataset

As an example, we provide a small subset of the marsupials dataset used in the paper cited above. This dataset consists of a full tree on 272 species followed by 10 trees that each have a small subset of the leaf set of the full tree. Sample command to run RF+ on this dataset:

python3 RF+.py -i MarsupialsSubset.newick -o outputfile.txt -r