RF+ software version 1.0

Description

RF+ is a 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 manuscripts cited below and can (i) compute the RF(+) distance between two trees with arbitrary leaf sets output the corresponding optimal completions, and (ii) compute the Extraneous-Clade-Free-RF(+) (EF-RF(+)) distance between two trees with arbitrary leaf sets and output the corresponding optimal completions. Input trees can be rooted or unrooted but must be binary. If the two given trees share no leaves in common, then their optimal completions are trivial and the software will return a tree which merely joins the two trees together with a new root node.

We refer the reader to the following two papers for further details on RF(+) and EF-RF(+) distances.

Optimal Completion and Comparison of Incomplete Phylogenetic Trees Under Robinson-Foulds Distance
Keegan Yao and Mukul S. Bansal
Under review.

Linear-Time Algorithms for Phylogenetic Tree Completion Under Robinson-Foulds Distance
Mukul S. Bansal.
Algorithms for Molecular Biology, 15:6, 2020.

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

This version of RF+ was implemented by Keegan Yao and Ashim Ranjeet under the supervision of Mukul Bansal and 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 with only leaf node labels, no edge lengths, and must be in a single input file with each tree appearing on a separate line. For each pair of compared trees, the

program outputs: (i) basic statistics such as leaf set sizes and sizes of the union and intersection of the two leaf sets, (ii) RF(-), RF(+), and EF-RF(+) distances between the two trees, and (iii) optimal completions of both trees. 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
```

By default, optimal completions are computed under RF(+) distance. To compute optimal completions under the EF-RF(+) distance instead, the "-ext" command line option can be used. For example,

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

Finally, the "-u" option can be used to compute unrooted EF-RF(+) and/or RF(+) distances and completions. For example,

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

Example dataset

As an example, we provide a sample input file consisting of 10 rooted trees. RF+ can be executed on this input file as follows:

```
python3 RF+.py -i MarsupialSubset.newick -o outputfile.txt
```