Visual TreeCmp: a comprehensive comparison of phylogenetic trees on the web – manual

1. Introduction

A phylogenetic tree represents a historical evolutionary relationship between different species or organisms. There are various methods for reconstructing phylogenetic trees. Applying those techniques usually results in different trees for the same input data. An important problem is to determine how distant from each other two trees reconstructed in such a way are. Comparing phylogenetic trees is also useful in mining phylogenetic information databases. The Visual TreeCmp application was designed to compute distances between arbitrary (not necessary binary) phylogenetic trees. Visual TreeCmp is available in two versions: hosted web-based and stand-alone web-based. Both are Spring framework-based versions of the TreeCmp 2.0 command line application available as part of the same package. The package offers various metrics for rooted and unrooted phylogenies, purely topological as well as weighted metrics (i.e. taking into account numerical lengths/weights of edges in the form of non-negative real numbers). All distances are implemented using polynomial time algorithms and all of them fulfill classic mathematical metric space axioms.

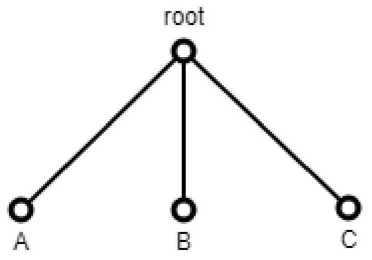
2. Input data format

The TreeCmp software was designed to support BEAST (http://beast.bio.ed.ac.uk/) and MrBayes (http://mrbayes.csit.fsu.edu/) data files, where phylogenetic trees are stored in the NEWICK format. Note that plain text files containing only trees in this format are supported as well. The input file can contain any number of trees separated by a semicolon.
Example:

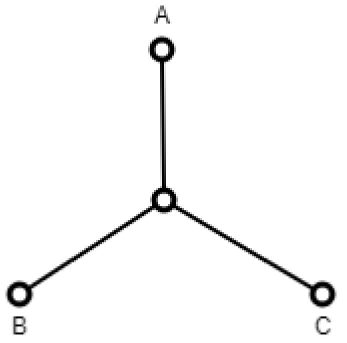
```
((((((((((1,2),3),4),5),6),7),8),9),10);
((((((((2,3),4),5),6),7),8),9),(1,10));
```

Before performing calculations, the validity of entered data is checked. In the case of any incompatibilities with the NEWICK/NEXUS format, the execution of the program will be terminated and an error message will be displayed. If only metrics not including edge weights are used in the calculations (unweighted metrics), the existing weights will not be interpreted. If at least one metric includes weights, positive weights must be assigned to all edges in the trees, otherwise the calculations will be terminated and an error message will be displayed. These restrictions can be relaxed by using the `-W` parameter, which enables applying zero weights, and if there is no weight, zero value will be assigned to the edge as default.

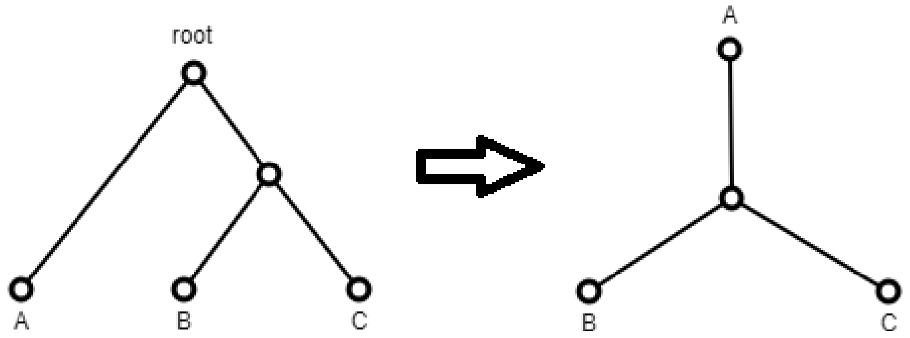
Newick trees are interpreted as rooted, even if there is a multifurcation at the base. The unrooted metrics modify each tree by removing the root indication, and if a node of degree 2 is left behind it is also suppressed. For example, if an unrooted binary tree in the NEWICK format: (A, B, C) is entered in the metric dedicated for a rooted tree, it will be interpreted as a rooted, non-binary tree consisting of a 3-degree root and 3 descendant vertices A, B and C.



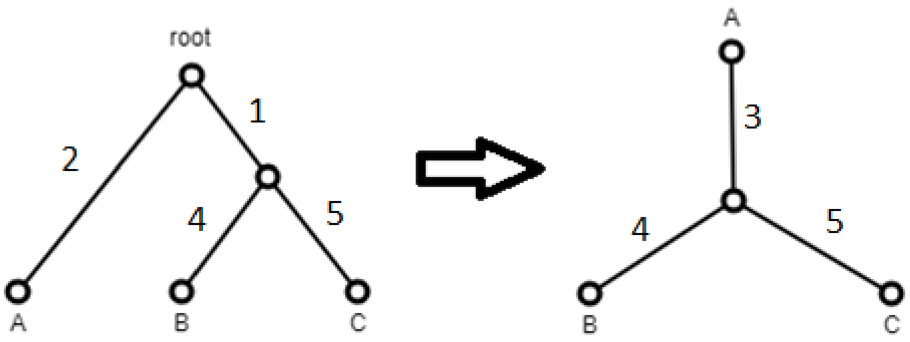
In the case when a rooted binary tree is entered to the metric dedicated for an unrooted tree, the root will be treated as an internal vertex or will be automatically suppressed if its degree equals 2. For example, if a rooted binary tree in the NEWICK format (A, B, C) is entered in the metric for a unrooted tree, then the root will be treated as an internal vertex and the tree will be interpreted as an unrooted, binary tree.



However, after entering the rooted tree (A, (B, C)) to the metric dedicated for unrooted trees, the root will be suppressed as in the figure below.



In the case with a weighted tree (a tree with weights on the edge) the sum of edge weights incident to the removed root vertex has been assigned to the newly created edge as in the figure below.



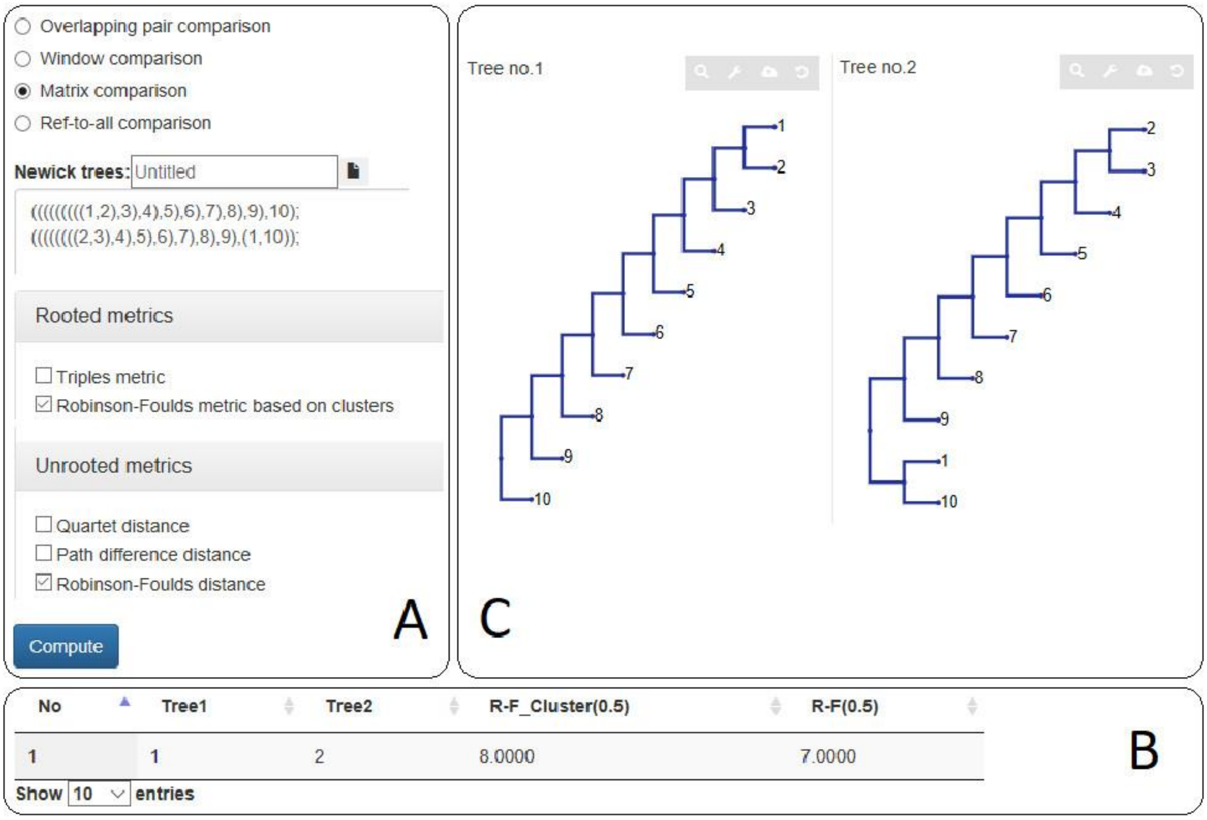
Summing up, Newick trees are interpreted as rooted, even if there is a multifurcation at the base. The unrooted metrics modify each tree by removing the root indication, and if a node of degree 2 is left behind, it is also suppressed.

3. Using Visual TreeCmp

The most convenient way to use the application is through GUI available at: https://eti.pg.edu.pl/treecmp/WEB. To determine any metric we need at least two phylogenetic trees, e.g.:

```
(((((((((1,2),3),4),5),6),7),8),9),10);
((((((((2,3),4),5),6),7),8),9),(1,10));
```

In addition, it is necessary to select at least one metric and approve the calculations. The completed form needed to determine two metrics (Robinson-Foulds and Robinson-Foulds cluster) is presented in Fig.1A. The Fig.1B shows a report with calculated values of the given metrics. After clicking on the corresponding row, the visualization of the compared trees can be seen – Fig.1C. Please note that in the case of the Robinson-Foulds cluster metric, the root was automatically removed from the given trees. The only vertex of degree 2 has been shrunk. The result between RF and RFC differs by 1.



○ Overlapping pair comparison
○ Window comparison
● Matrix comparison
○ Ref-to-all comparison

Newick trees: Untitled

```
((((((((((1,2),3),4),5),6),7),8),9),10);
(((((((((2,3),4),5),6),7),8),9),(1,10));
```

Rooted metrics

□ Triples metric
☑ Robinson-Foulds metric based on clusters

Unrooted metrics

□ Quartet distance
□ Path difference distance
☑ Robinson-Foulds distance

No	Tree1	Tree2	R-F_Cluster(0.5)	R-F(0.5)
1	1	2	8.0000	7.0000

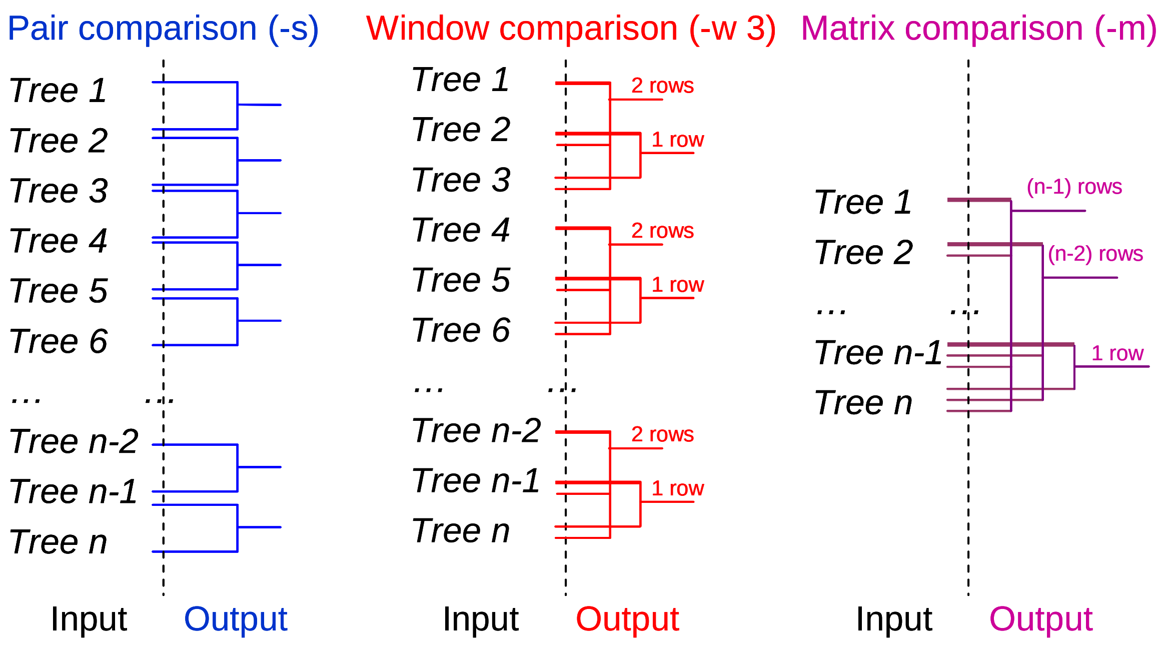
Show 10 entries

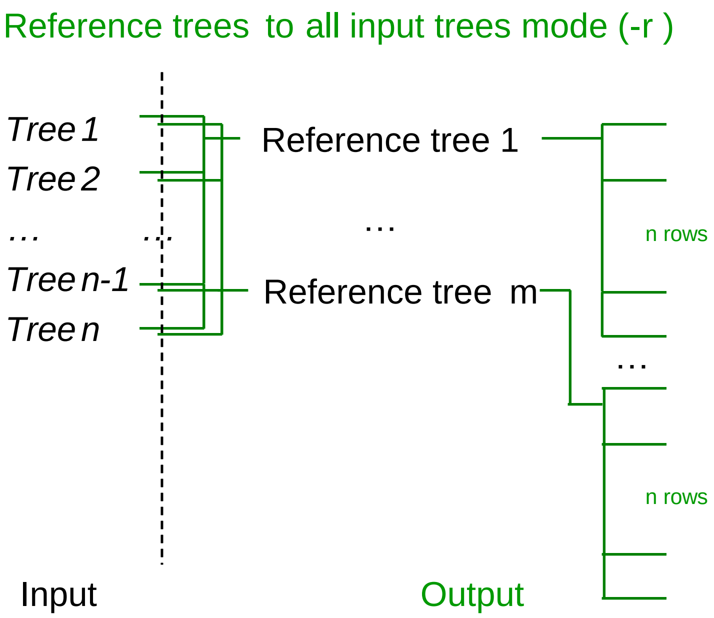
Figure 1- (A) An example of the completed form. (B) Calculated values of RF and RFC metrics. (C) The visualization of compared trees with the applied Phylo.io library.

Mandatory options:

- The comparison mode options (only one option should be specified):
  - `Overlapping pair comparison` – an overlapping pair comparison mode; every two neighboring trees in the input file are compared,
  - `Window comparison` – a window comparison mode; every two trees within a window with a specified size are compared – the average distance and the standard deviation go to the output file,
  - `Matrix comparison` – a matrix comparison mode; every two trees in the input file are compared.
  - `Ref-to-all comparison` – a reference trees to all trees mode. Each tree in the input file is compared to all reference trees.

Details of the computation flow in each of these cases are explained in the pictures below.





- The metric option. At least one and at most 18 metrics can be specified (numbers in square brackets correspond to the reference list. Metrics should be separated by spaces.

  Metrics for rooted trees:
  - `Triples` – the Triples metric (Crichlow et al. 1996),
  - `RFCluster(0.5)` – the Robinson-Foulds metric based on clusters (Robinson and Foulds 1981),
  - `MatchingPair` – the Matching Pair metric (Bogdanowicz and Giaro 2014),
  - `NodalSplittedWeighted` – the Nodal Splitted metric with L^2 norm (Cardona et al. 2010),
  - `GeoRooted` – the Geodesic (BHV) rooted distance for weighted trees (Owen, Provan 2009),
  - `NodalSplitted` – the Nodal Splitted weighted metric with L^2 norm (Cardona et al. 2010),

  - `MatchingCluster` – the Matching Cluster metric (Bogdanowicz et al. 2012),
  - `MAST` – the Rooted Maximum Agreement Subtree distance (Farach and Thorup 1994),
  - `RFClusterWeighted(0.5)` – the Robinson-Foulds weighted metric based on clusters (Robinson and Foulds 1979),
  - `CopheneticL2Weighted` – the Cophenetic weighted metric with L^2 norm (Cardona, Mir, Rosselló, Rotger and Sánchez 2013),
  - `CopheneticL2` – the Cophenetic metric with L^2 norm (Cardona, Mir, Rosselló, Rotger and Sánchez 2013).

  Metrics for unrooted trees:

  - `RFWeighted(0.5)` – the Robinson-Foulds weighted distance (Robinson and Foulds 1979),
  - `Quartet` – the Quartet distance (Estabrook 1985),
  - `PathDifferne ce` – the Path difference distance (Steel and Penny 1993),
  - `RF(0.5)` – the Robinson-Foulds distance (Robinson and Foulds 1981),
  - `MatchingSplit` – the Matching Split distance (Bogdanowicz and Giaro 2012),
  - `UMAST` – the Unrooted Maximum Agreement Subtree distance (Farach and Thorup 1994),
  - `GeoUnrooted` – the Geodesic (BHV) unrooted distance for weighted trees (Owen, Provan 2009).

- Newick trees – an input data file with trees in the NEWICK/NEXUS format. Using `Ref-to-all` comparison mode, both input files should be specified.

Additional options (optional):

- `Normalized distances` – reports normalized distances δ_m for a particular metric (works only for unweighted metrics) m (Bogdanowicz et al. 2012; based on an average value from pre-computed data). This functionality is available for trees with the number of leaves between 4 and 1000. Note that a normalized tree similarity for a particular metric m (NTS_m) can be expressed by a normalized distance as follows: $NTS_m. = 1 - \delta_m$ (Bogdanowicz et al. 2012).
- `Prune trees` – prunes compared trees if needed. This option is designed to allow the comparison of trees with different (partially overlapping) sets of taxa. After using this option three additional columns appear in the output file (see section 4 for details).
- `Include summary` – includes a summary section in the output file.
- `Zero weights allowed` – weights of zero value are allowed. If there is no weight, its default value will be set to zero.
- `Bifurcation trees only` – allow only bifurcating trees.

Note that if a rooted tree (with bifurcation in the root) is compared using metrics for unrooted trees, the tree will be automatically transformed into an unrooted one, i.e., the bifurcation will be replaced with an arbitrary trifurcation.

All elements of the GUI including report output data along with the use of more extensive test cases are described in detail in the WEB user's manual available at:
https://eti.pg.edu.pl/TreeCmp/manual_tree_cmp.html
or in the application menu under the button ?.

4. Running and useful Java VM parameters

Visual TreeCmp application is available in two forms:

1. Web application – Java Web application server is required to run `VisualTreeCmp.web` file. This application has been tested with the Apache Tomcat server. Tomcat Web Application Deployment is described with details here: https://tomcat.apache.org/tomcat-8.0-doc/deployer-howto.html#Deployment_on_Tomcat_startup.

2. Standalone application with WEB GUI – can be run locally without the server installation needed. Running `VisualTreeCmp.jar` file:

   ```
   java -jar VisualTreeCmp.jar
   ```

   The application will be visible at: http://localhost:8080/TreeCmp

   We can change the default port from 8080 to another one, e.g. 8083:

   ```
   java -jar VisualTreeCmp.jar --server.port=8083
   ```

   In the case of an analysis of large trees the following exceptions might occur:

   - Exception in thread "main" java.lang.OutOfMemoryError: Java heap space

     To solve the problem increase Java heap space memory limit using JVM option –Xmx

     Example:

     ```
     java -Xmx4096m -jar VisualTreeCmp.jar
     ```

   - Exception in thread "main" java.lang.StackOverflowError at pal.io.FormattedInput.skipWhiteSpace(FormattedInput.java:111) at pal.io.FormattedInput.readNextChar(FormattedInput.java:131) at pal.tree.ReadTree.readNH(ReadTree.java:81)
     …..
     at pal.tree.ReadTree.readNH(ReadTree.java:89)

     To solve the problem increase Java thread stack size limit using JVM option –Xss
     Example:

     ```
     java -Xss1m -jar VisualTreeCmp.jar
     ```

     These options can be used in conjunction.

License

References

1. Bogdanowicz D, Giaro K: **Matching Split Distance for Unrooted Binary Phylogenetic Trees**. *IEEE/ACM Trans Comput Biol Bioinform* 2012, **9**: 150-160.
2. Bogdanowicz D, Giaro K: **Comparing Phylogenetic Trees by Matching Nodes Using the Transfer Distance between Partitions**. Submitted, 2014.
3. Bogdanowicz D, Giaro K., Wróbel B. **TreeCmp: comparison of trees in polynomial time.** *Evol. Bioinform.* 2012, in press.
4. Cardona G, Llabrés M, Rosselló F, Valiente G: **Nodal distances for rooted phylogenetic trees**, *J Math Biol* 2010 **61**:253-276.
5. Critchlow DE, Pearl DK, Qian C**: The Triples Distance for Rooted Bifurcating Phylogenetic Trees,** *Syst Biol* 1996, **45**: 323-334.
6. Estabrook GF, McMorris FR, Meacham CA: **Comparison of Undirected Phylogenetic Trees Based on Subtrees of Four Evolutionary Units**. *Syst Biol* 1985, **34**:193-200.
7. McKenzie A, Steel M, **Distributions of cherries for two models of trees**. *Math Biosci* 2000, **164**:81-92.
8. Owen M, Provan J. **A Fast Algorithm for Computing Geodesic Distances in Tree Space**. *IEEE/ACM Trans Comput Biol Bioinform* 2009. **8**: 2-13.
9. Robinson D, Foulds LR **Comparison of weighted labelled trees**. *Combinatorial Mathematics VI* 1979, **748**:119-126.
10. Robinson DF, Foulds LR: **Comparison of phylogenetic trees**. *Math Biosci* 1981, **53**:131-147.
11. Steel MA, Penny D: **Distributions of Tree Comparison Metrics – Some New Results**. *Syst Biol* 1993, **42**:126-141.
12. Semple C, Steel M: **Phylogenetics**, Oxford University Press 2003.