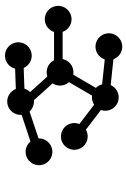
MulRF

A platform-independent software package for phylogenetic analysis
using multi-copy gene trees

[Version 1.2]

User's Manual



Contents

List of Figures

Chapter 1

Introduction

MulRF is a platform-independent software package for inferring species trees from incongruent multi-copy gene trees using a generalization of the Robinson-Foulds (RF) distance to multi-labeled trees (i.e., trees with multiple leaves having the same label). The underlying generic tree distance measure and fast running time makes MulRF useful for analyzing large genomic data sets, in which multiple evolutionary processes as well as phylogenetic error may contribute to gene tree discord. MulRF implements fast local search based heuristics which greatly speed up species tree inference [1, 2]. It provides the following features:

- The input gene trees can be **weighted** for the species tree inference.
- **Topological constraints** can be imposed on the inferred species tree.
- Running **multiple replicates** of the tree search heuristic is automated.
- Weighted or un-weighted **MulRF distances** of a given species tree from a collection of input trees, as well as the MulRF distance to each gene tree, can be computed.
- The input gene trees and the resulting species tree can be **illustrated**.

NOTE: Throughout this manual we describe the input trees as multi-copy gene trees. However, **MulRF is not limited to gene tree input**. Indeed, it can be used to build a supertree from **any collection of single and/or multi-labeled trees, including gene trees, species trees, or taxonomy trees**.

Chapter 2

Using MulRF

This chapter gives an overview of how to use MulRF. Topics include downloading steps, file formats, and user interfaces.

2.1 Download

MulRF is an **open source program**, and the source code is available for download at `https://github.com/ruchiherself/MulRFRepo`. If you do not want to compile the source code, the **platform-independent binary executable** is available at `http://genome.cs.iastate.edu/CBL/MulRF/`. This executable has been successfully tested on Windows 7, Mac OS X 10.9, and Ubuntu 12.04. To install the binary executable, simply download and unzip `MulRF1.2.zip` (or `MulRF1.2.tar.gz` for Linux) to a computer.

For Mac and Windows: move inside the directory `MulRF1.2` and click the executable jar file `MulRF` to begin.

For Linux and others: open the terminal shell and move to the directory `MulRF1.2`. Then type "`java -jar MulRF.jar`".

NOTE: 1) MulRF requires **Java Runtime Environment (JRE) 5.0 or later version** to be installed on the computer to run. 2) **Linux users** should set the **execute permissions** of executables in `MulRF1.2 -> executables` before running MulRF.

2.2 Input/Output File Format

2.2.1 Gene and Species Tree Files

A gene tree file is an ASCII format text file containing a collection of input gene trees. The gene trees must be in Newick format terminated by a semicolon, and they can be non-binary (unresolved). In the gene trees, the leaf names must correspond to a species name (e.g., the name of the species from which the gene was sequenced). MulRF allows multi-copy gene trees as input, i.e., gene trees that have multiple leaves (gene copies) with the same label or name, corresponding to a single species. Here is an example

of a multi-copy gene tree in Newick format: `(((speciesA,speciesB),speciesC),speciesA);`

The following rules must be followed when creating a gene tree or species tree file:

- The species names in the species tree must be unique (occur only once in the tree).
- Labels with non-alphabetic characters (e.g., spaces) must be encapsulated in apostrophes or quotation marks, e.g., `speciesA`, `"species A"` or `'species A'`.
- A **comment** in the input file must be encapsulated in a square brackets. A comment can appear anywhere, including inside the Newick format tree, e.g., `( (speciesA,speciesB), [example comment text] speciesC );`.
- A gene tree can span multiple lines.
- Gene trees can be **unrooted or rooted**, which is specified by the prefix `[&U]` or `[&R]` for unrooted or rooted cases, respectively. A gene tree without any prefix is considered rooted. If an input tree is rooted, the root will not be considered during the species tree search; instead, the gene tree will be treated as an unrooted topology. Also, if an input gene tree has a root, then this root may get lost in the tree search; that means if the gene tree is rewritten in the output file then its default Newick root may differ from its initial root.
- The input trees can have **branch lengths**, but MulRF does not use branch length data to calculate the MulRF distance or infer the species tree.
- If a gene tree is **weighted**, the MulRF distance for that weighted gene tree is multiplied by its weight. Thus, gene trees with a higher weight should have a larger impact in the tree search. The weight can be any positive real number from 0 to 1. A weight of 0 means that the gene tree has no effect on the MulRF analysis (i.e., it is equivalent to not including the gene tree in the analysis). A weight can be specified by the prefix `[&WEIGHT=<value>]`. If a weight is not specified, the input tree is given a weight of 1. Here is an example of assigning weight to an input tree: `[&WEIGHT=0.5]((speciesA,speciesB),speciesC);`

WARNING: Due to rounding errors, **fractional weights** can sometimes cause MulRF to malfunction. However, weights like 0.5 or 0.25 work fine.

2.2.2 Constraints File

A constraints file allows the user to impose constraints on the topology of the inferred species tree. As the name indicates, the file contains a set of topological constraints for the species tree. Each constraint is a comma-separated list of species names terminated with a semicolon. In the resulting species tree, the species in a constraint will be closer to each other than to the rest of the species. Multiple constraints can be specified, but any single species can be part of at most one constraint. For an example, a constraint file may contain the following two constraints: `cat, dog, turtle; rice, arabidopsis;`

NOTE: The constraints option can only be used if the **leaf adding** option is selected to generate the **initial species tree**.

2.2.3 Scoring File

A scoring file is used to compute the MulRF distance from a given set of gene trees to a species tree. A scoring file contains a species tree followed by a list of gene trees. A scoring file has the same format as a gene tree and species tree file (Section 2.2.1), except that the first tree in a scoring file must be a species tree.

2.2.4 Output File

The first tree in the output file of MulRF search is the species tree with lowest MulRF score found during the heuristic search. The total reconciliation cost of this species tree appears in a comment right before this tree. The species tree is followed by a list of all the input gene trees, with the individual MulRF score of each gene tree (against the species tree) appearing in a comment with that tree.

All the input files should be stored inside the `inputData` folder under the project home directory. After execution, output files are stored in a unique folder inside the `outputData` folder under the project home directory. The default output file folder name is a system-generated hexadecimal number followed by the input file name.

2.3 User Interface

MulRF's user interface includes a log window to record results of analysis, gene and species tree display windows, and various gene tree, species tree, and output folder table windows.

2.3.1 Log window

The log window displays all the selections made in the drop-down menu options and updates the user on the species tree search in the real-time. It appears at any first action in MulRF's main window and can be found in the top-left corner throughout the analyses. The log window scrolls down to display the latest inserted line. The contents of the log window are also saved to a text file `logFile` in the project home directory.

2.3.2 Tree Display Windows

MulRF displays gene and species tree by converting their Newick file into PNG format tree image using ATV [3].

NOTE: MulRF returns an **unrooted species tree**, and the root in the tree display window is **arbitrary**.

2.3.3 Table Windows

- **Input File Statistics Table:** Upon selection of an input gene tree file, MulRF creates an input file statistics table. It lists all the input gene trees, with their leaf counts, weights, and `view` buttons to display them. Only one gene tree file can be selected at a time.

- **Output Folder Statistics Table:** After a species tree search, MulRF creates an output folder statistics table. It lists all the generated output files as a button with the MulRF distance of the species tree in each file from the collection of input gene trees.
- **Output File Statistics Table:** Clicking the output file button in the output folder statistics table creates an output file statistics table. It lists the inferred species trees, followed by a list of input gene trees used in the analysis. This table also includes the score, weight, and a `view` button for each tree. The score of a species tree is its MulRF distance from all the input gene trees. For an input gene tree, the score denotes the MulRF distance of it from the species tree listed above it.

Chapter 3

Sample Run

Here we present an example of how to run MulRF to infer a species tree from a sample data set. Our sample input file - `vertebrates02.newick`, has ten multi-copy gene trees (see Figure 3.1). This input gene tree file is located in `inputData` folder inside MulRF's home directory.

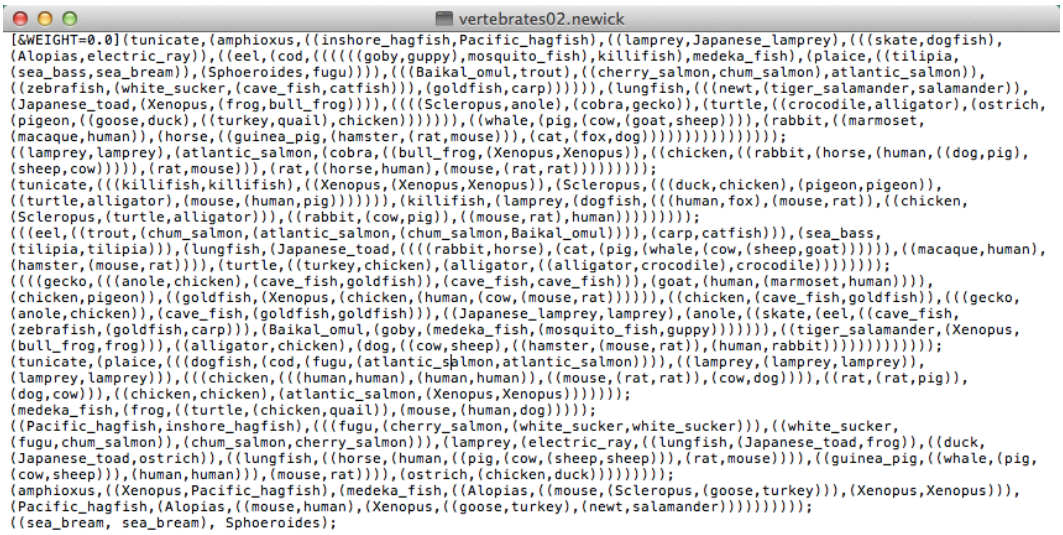


Figure 3.1: Input gene tree file

To estimate an optimal species tree for this collection of gene trees, the user must launch MulRF and perform the following steps:

1. Click `File->Open Gene Tree File`, and a file chooser window will appear. Select `vertebrates02.newick` and click `Open`. This will set a log window at the top-left corner, and a statistics table of the selected

gene tree file. The `Input File Statistics Table` lists gene trees with their number of leaves, weights, and `View` buttons (Figure 3.2). The log window shows the selected input file name.

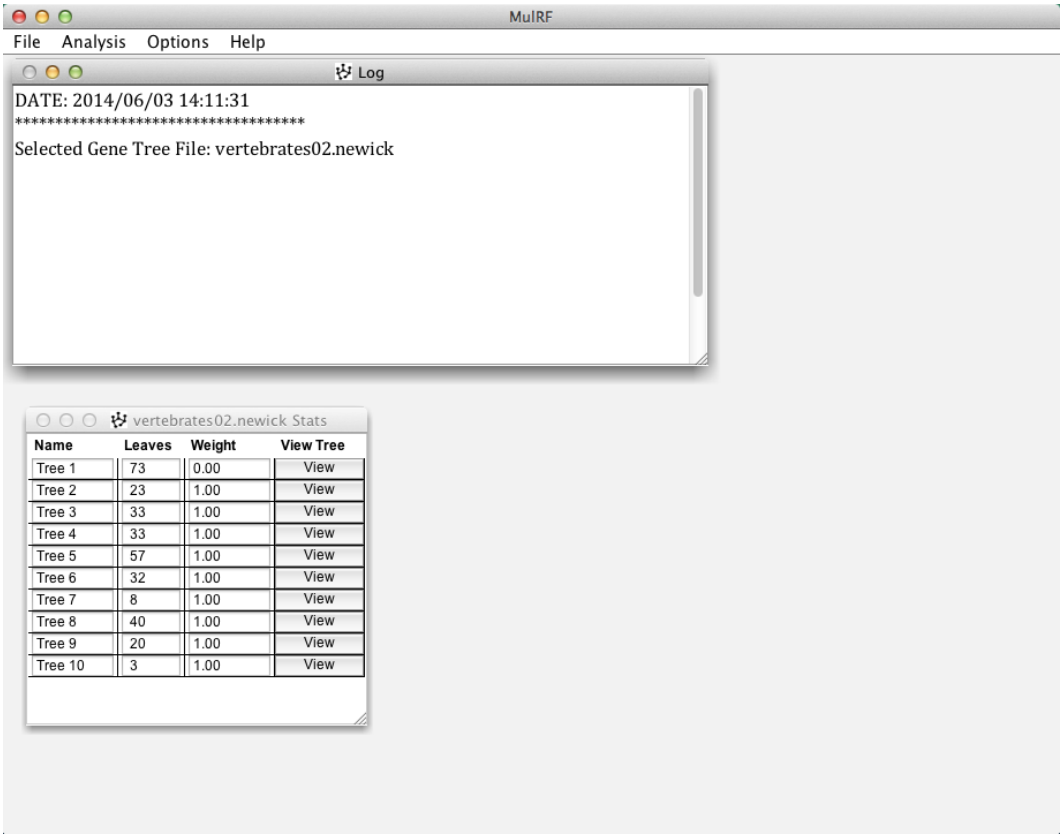


Figure 3.2: After selecting the input gene tree file

2. Now click `Analysis->Tree search`. This step is typically the most time-consuming, depending on the input size and details of the heuristic search. Observe the log window for continuous updates (Figure 3.3).

3. Once the tree search is over, a statistics table for the output folder appears. The rows in this table correspond to the generated output files inside the new folder under `outputData` folder (Figure 3.4). The first column in the output folder statistics table has buttons to display the statistics table of the corresponding files. The second column shows the MulRF score of the resulting species tree in each file. Clicking the file name button shows the statistics table of the file (Figure 3.4). It lists species tree followed by input gene trees with their score, weight, and `View` buttons to display them. Observe that the weight of the first gene tree is 0; thus, this gene tree does not have any impact on the species tree search.

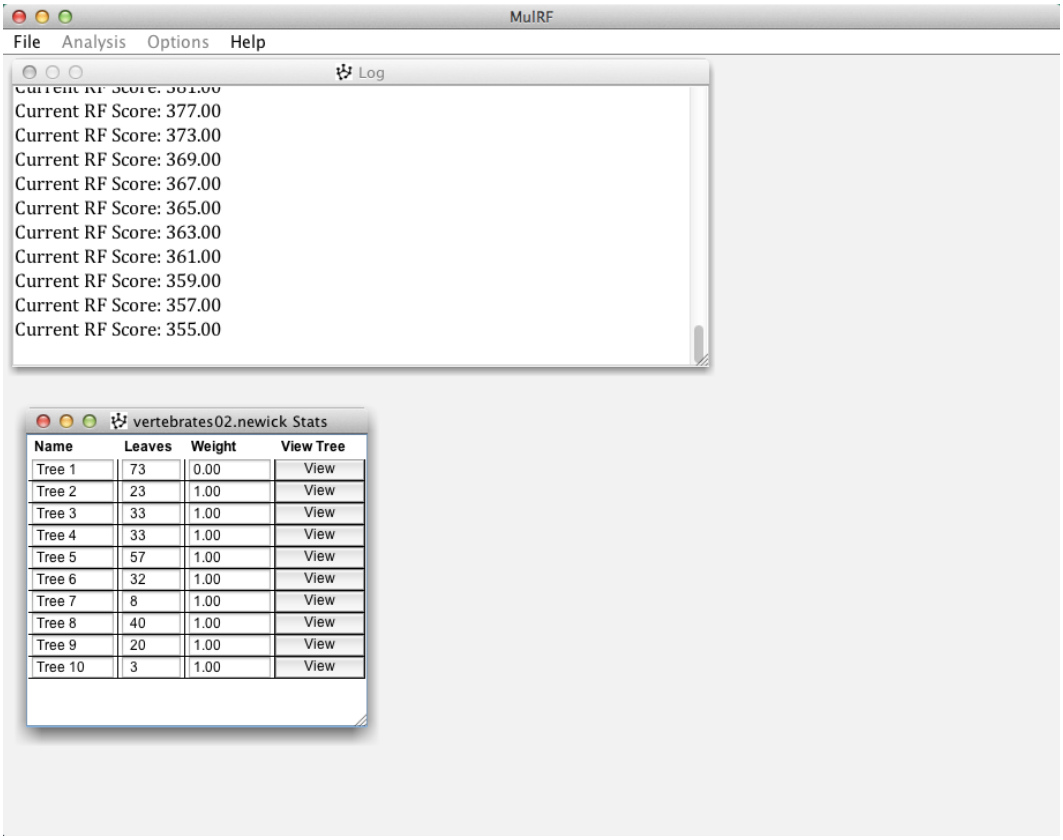


Figure 3.3: Running tree search

The generated output files are now available inside a new folder under `outputData`. Observe the name of the newly generated folder in the log window.

4. The output species tree (or gene trees) can be displayed by clicking the `View` button in the output file statistics table (Figure 3.5).

   The contents of the log window are stored inside `logFile` under the project home directory.

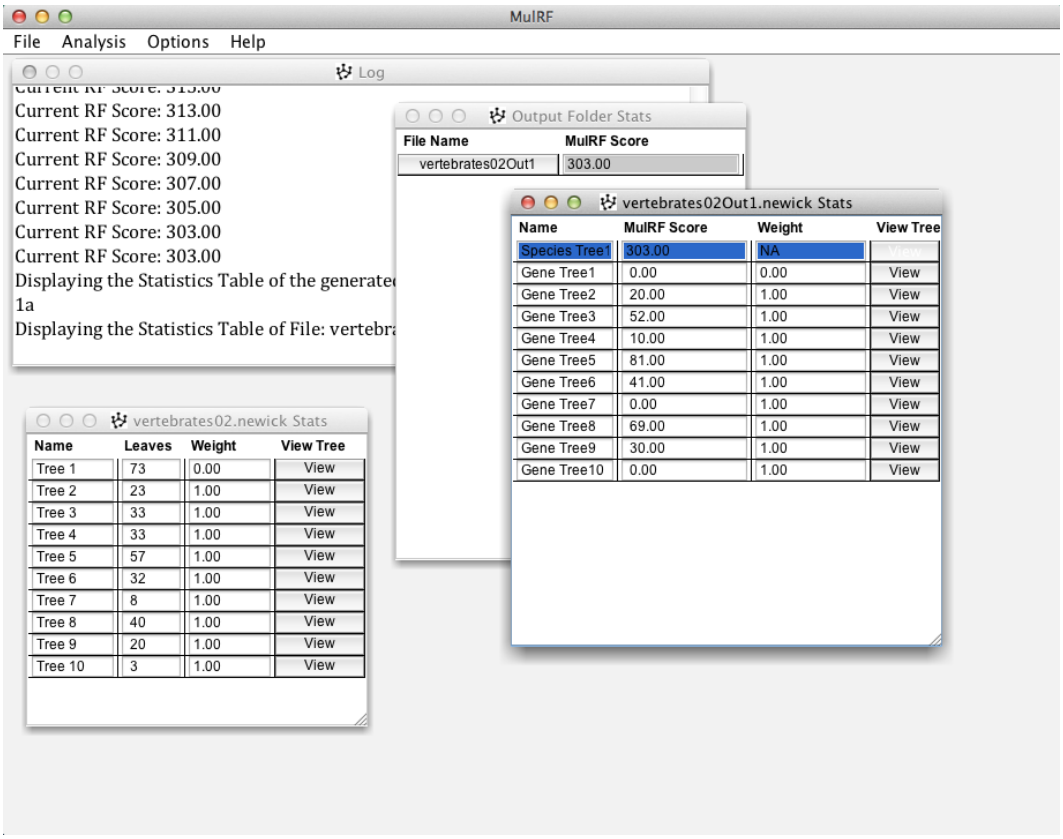


Figure 3.4: Statistics tables for the newly generated output folder and file

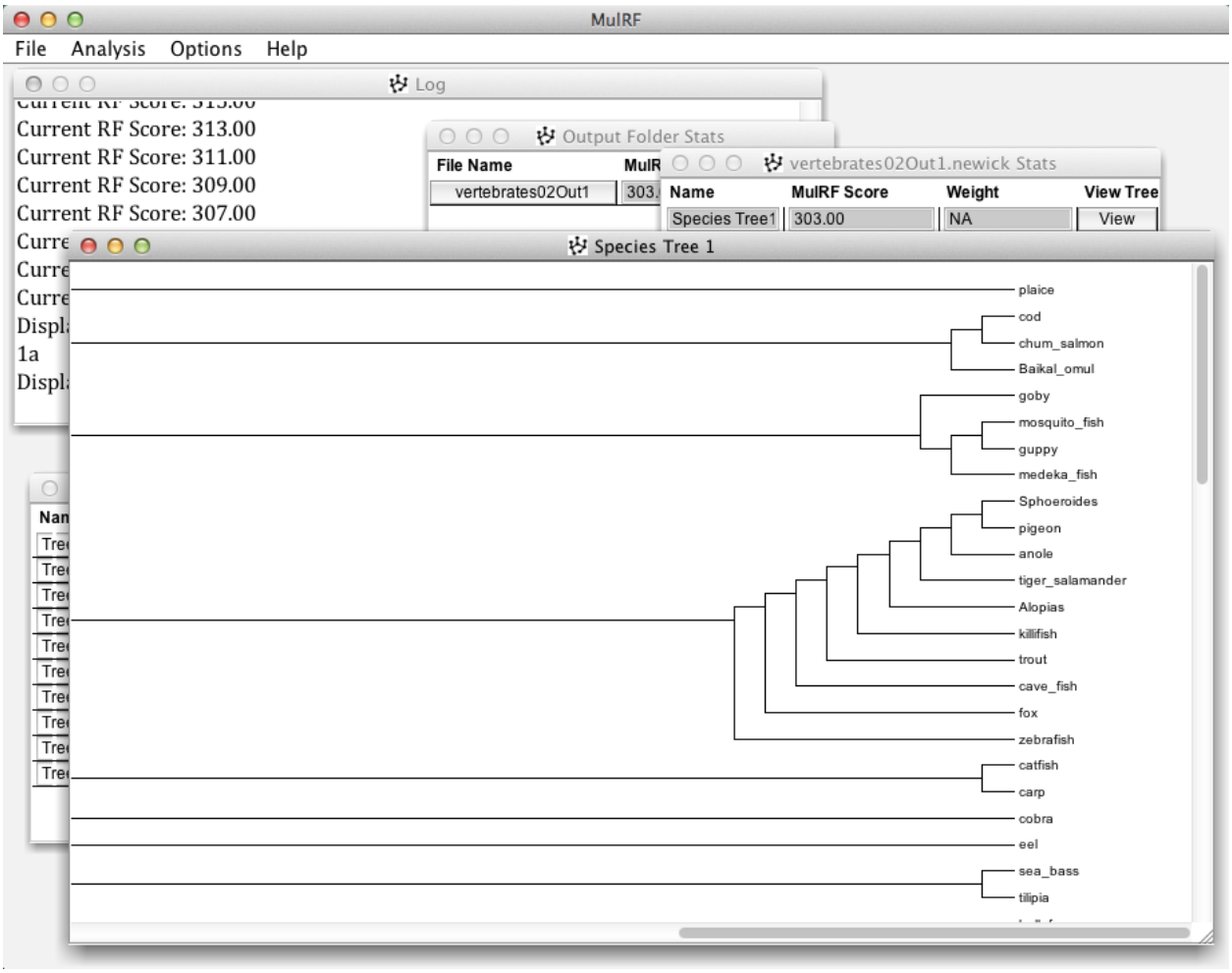


Figure 3.5: Displaying the output species tree

Chapter 4

Available Features

This chapter gives a description of available features in MulRF's user interface. The menu bar contains four item: `File`, `Analysis`, `Options`, and `Help`.

4.1 File

The `File` menu item assists in selecting the input files to run MulRF analyses (Figure 4.1). Its drop-down menu contains following items:

Figure 4.1: File menu item in MulRF

4.1.1 Open Gene Tree File

This allows the user to select an input gene tree file for the species tree search from folder `inputData` in the project home directory. The menu item is disabled as soon as file is selected, and it remains disabled until the input file is closed. The same command can also be executed by typing `CTRL-G`.

4.1.2 Close Gene Tree File

This enables the user to close a previously selected gene tree file. Closing a gene tree file enables the user to select another input gene tree file using the `Open Gene Tree File`.

4.1.3 Open Scoring File

This menu item allows the user to select a scoring file from the folder `inputData` in the project home directory. The menu item is disabled as soon as a file is selected, and it remains disabled until the scoring file is closed. The same command can also be executed by typing `CTRL-S`.

4.1.4 Close Scoring File

This menu item closes the previously selected scoring file. After closing the scoring file, the user can use `Open Scoring File` to select another scoring file.

4.1.5 Exit

Clicking the exit submenu item closes (quits) the application. This can also be done by typing `CTRL-Q` in Windows and `'Command-key'-Q` in Mac OS.

4.2 Analysis

This menu item assists in initiating the heuristic tree search and computing the MulRF score (Figure 4.2). Its drop-down menu contains following items:

4.2.1 Tree search

This submenu allows to start the heuristic search. Note that the gene tree file should be opened and options (if required) should be set before clicking it and beginning the tree search.

4.2.2 MulRF score

This submenu allows to compute the MulRF distance from a given species tree to a collection of gene trees.

Figure 4.2: Analysis menu item in MulRF

4.3 Options

This menu item assists in customizing the heuristic tree search (Figure 4.3). These options should be set (if required) before initiating the heuristic species tree search. The drop-down menu contains following items:

Figure 4.3: Options menu item in MulRF

4.3.1 Starting Species Tree Generation

This allows the user to choose the method to be used to obtain the initial starting species tree for the heuristic search. There are two available choices:

1. **Leaf Adding:** This is a **greedy step-wise method** to build an initial starting species tree; it is the default choice. This method considers the input trees as rooted, by using the default Newick format tree roots, and uses the rooted MulRF distance. The rooted MulRF distance is computed based on the the rooted RF distance, instead of unrooted RF distance, between the extended species tree and the multi-copy gene tree. The greedy step-wise method randomly selects three leaves from the species leaf set and considers all the triplets that these three leaves can form. It randomly selects a triplet with the lowest rooted MulRF score based on the input gene trees as the backbone tree. Next, it randomly selects a new leaf (species) to add to the initial 3-taxon backbone tree. The algorithm tries to add this leaf in all possible positions in the current tree, and it stores all those positions that give the best score (lowest rooted MulRF distance to the input gene trees). It then randomly selects a position from among these best locations and adds the leaf at that location. This process of adding a new leaf is repeated until the starting tree includes all species. Once all the leaves are added, we have our initial species tree. Note that the root of this initial species tree will not be considered further in the heuristic search.

2. **User Specified:** This allows the user to choose a file containing a species tree in Newick format. This user-supplied species tree is then used as the starting species tree for the heuristic search.

4.3.2 Output Folder Name

This submenu item allows the user to specify a name for the folder in which the output files (with the results of the analysis) will be stored. The default name for the output folder is composed of input gene tree file name, followed by a unique system-generated hexadecimal number.

4.3.3 No. of Replicates

The tree search heuristic is a **randomized algorithm**, and the search may get caught in a local minima. To explore the tree space more thoroughly, we recommend executing the tree search heuristic multiple times on the same data set. MulRF automates this by running multiple replicates, each using a different random seed to build the starting tree and to order the edges of the species tree for subtree prune and regraft (SPR) operations. MulRF allows the user to enter the number of replicate runs to be performed. The `No. of Replicates` submenu provides the following two choices:

1. **1 (Default):** This sets the number of replicates to the default value 1.
2. **Customize:** This opens an input dialog box, where the user can change the default to any input number (less than 100).

4.3.4 Constraints File

This menu item allows the user to select constraints file to impose constraints on the topology of the inferred species tree. Clicking of `Constraints File` submenu opens the file chooser window to select the constraints file from `inputData` folder under the project home directory.

4.3.5 Random Seed

The random seed allows the user to specify the seed for the random number generator used by MulRF. There are two available choices:

1. **System Generated (Default):** The default random seed is the system wall clock time.
2. **Customized:** This allows the user to enter their own random seed. This is helpful when the user wishes to repeat a particular analysis. Analyses using the same random seed, the same data set, and the same search options will produce the same output.

4.4 Help

This menu item provide help to the MulRF user (Figure 4.4).Its drop-down menu contains following two items:

4.4.1 About

This provides a brief information about MulRF.

4.4.2 Help Contents

This is designed to guide the user through the available features in MulRF's user interface. The items in the left can be clicked to see the corresponding details in the right section of the help content window. A search feature is also provided to search and see the details of different features. The same command can also be executed by typing `CTRL-H`.

Figure 4.4: Help menu items in MulRF

Bibliography

[1] R. Chaudhary, B. Boussau, J. G. Burleigh, and D. Fernández-Baca. Assessing approaches for inferring species trees from multi-copy genes. (Under review).

[2] R. Chaudhary, J. G. Burleigh, and D. Fernández-Baca. Inferring species trees from incongruent multi-copy gene trees using the robinson-foulds distance. *Algorithms for Molecular Biology*, 28:8, 2013.

[3] C. M. Zmasek and S. R. Eddy. ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, 17:383–384, 2001.