MCdist 0.5 – an application for computing the Matching Cluster distance between phylogenetic trees - manual

Damian Bogdanowicz

1. Introduction

A phylogenetic tree represents 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 two trees reconstructed in such a way are from each other. Comparing phylogenetic trees is also useful in mining phylogenetic information databases. The MCdist application was designed to compute the Matching Cluster (MC) distance between arbitrary (not necessary binary) **rooted** phylogenetic trees.

2. Input data format

The MCdist software was designed to support BEAST (http://beast.bio.ed.ac.uk/) and MrBayes (http://mrbayes.csit.fsu.edu/) date 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.

3. Output data format

All output files created by the application regardless of chosen mode have similar structure. Output files are tab separated text files (TSV), which means that they can be easily read by various data analysis software (e.g. MS Excel, OpenOffice.org). An output file consists of two sections. The first section contains formatted in rows values of distances in the MC metric (or aggregate values in case of widow mode). The second section contains summary data computed based on all rows that appears in the first section.

4. Running MCdist

The MCdist application is distributed as a zip archive. In order to unpack the file any software supporting zip compression, for example free software 7-zip (http://www.7-zip.org/), can be used. In order to run the MCdist application Java VM in version at least 1.5 is required.

4.1. Directory structure

Directory			Description
bin			contains main jar file: **MCdist.jar** and lib folder with necessary open source libraries: pal-1.5.1.jar (http://www.cebl.auckland.ac.nz/pal-project/) and commons-cli-1.2.jar (http://commons.apache.org/cli/)
config			contains xml configuration file
doc			contains this manual
	examples		contains subdirectories with examples
		beast	contains an example input file created using BEAST
		mr_bayes	contains an example input file created using MrBayes
		plain	contains an example input file with plain trees
src			contains source code of this application

4.2. Command line syntax

Usage:

```
java -jar MCdist.jar -w <size>|-s|-m -i <inputfile> -o <outputfile>
```

Note that options order is important.

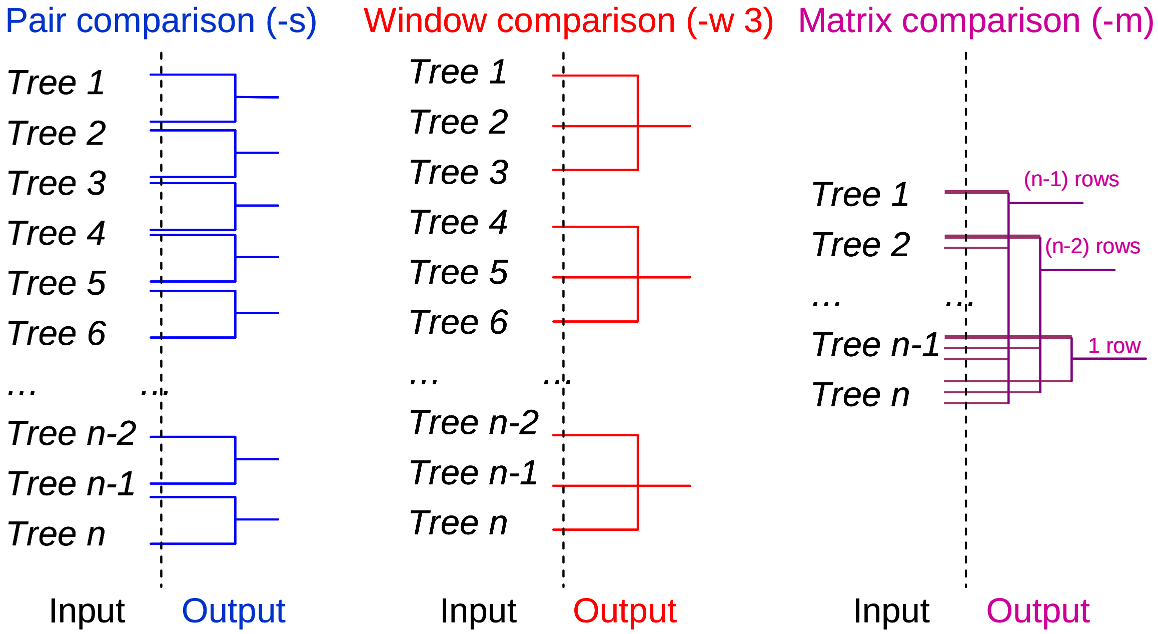
- The comparison mode options (only one option should be specified):
  - `-s` – pair comparison mode; every two neighboring trees in the input file are compared,
  - `-w <size>` – window comparison mode; every two trees within a window with a specified size are compared – the average distance and the standard deviation go to the output file,
  - `-m` – matrix comparison mode; every two trees in the input file are compared.

- IO options (both options should be specified):
  - `-i <inputfile>` - input data file with trees in the Newick format,
  - `-o <outputfile>` - output data file with the results of computations.

4.3. Types of analysis

There are three different types of available reports:
- Overlapping pair comparison,
- Window comparison,
- Matrix comparison.

Details of the computation of these reports are explained in the picture below.



5. Example

Input file: `\doc\examples\plain\plain.trees`
Invocation: `java -jar MCdist.jar -w 2 -i plain.trees -o plain.trees.w_2.out`
Console output:

```
MCdist version 0.5-b18
-----
Active options:
Type of the analysis: window comparison mode (-w) with window size: 2
Input file: plain.trees
Output file: plain.trees.w_2.out
-----
2011-04-13 22:46:32: Start of scanning input file: plain.trees
2011-04-13 22:46:32: End of scanning input file: plain.trees
2011-04-13 22:46:32: 4 valid trees found in file: plain.trees
2011-04-13 22:46:32: Start of calculation...please wait...
2011-04-13 22:46:32: 0.00% completed...
2011-04-13 22:46:32: 50.00% completed...
2011-04-13 22:46:32: 100.00% completed.
2011-04-13 22:46:32: End of calculation.
2011-04-13 22:46:32: Total calculation time: 16 ms.
```

Output file: plain.trees.w_2.out:

```
state  MatchingCluster (avg)     MatchingCluster (stddev)
1      7.0000 0.0000
2      3.0000 0.0000
---------
Summary:
Name   Avg    Std    Min    Max    Count
MatchingCluster    5.0    2.0    3.0    7.0    2
```

6. License