

MSdist 0.5 – an application for computing the Matching Split 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 MSdist application was designed to compute the Matching Split (MS) distance between arbitrary (not necessary binary) **unrooted** phylogenetic trees. MSdist is a part of TreeCmp application.

2. Input data format

The MSdist 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 MS 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 MSdist

The MSdist 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 MSdist application Java VM in version at least 1.5 is required.

4.1. Directory structure

Directory		Description
bin		contains main jar file: MSdist.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 MSdist.jar -w <size>|-s|-m -i <inputfile> -o <outputfile>
```

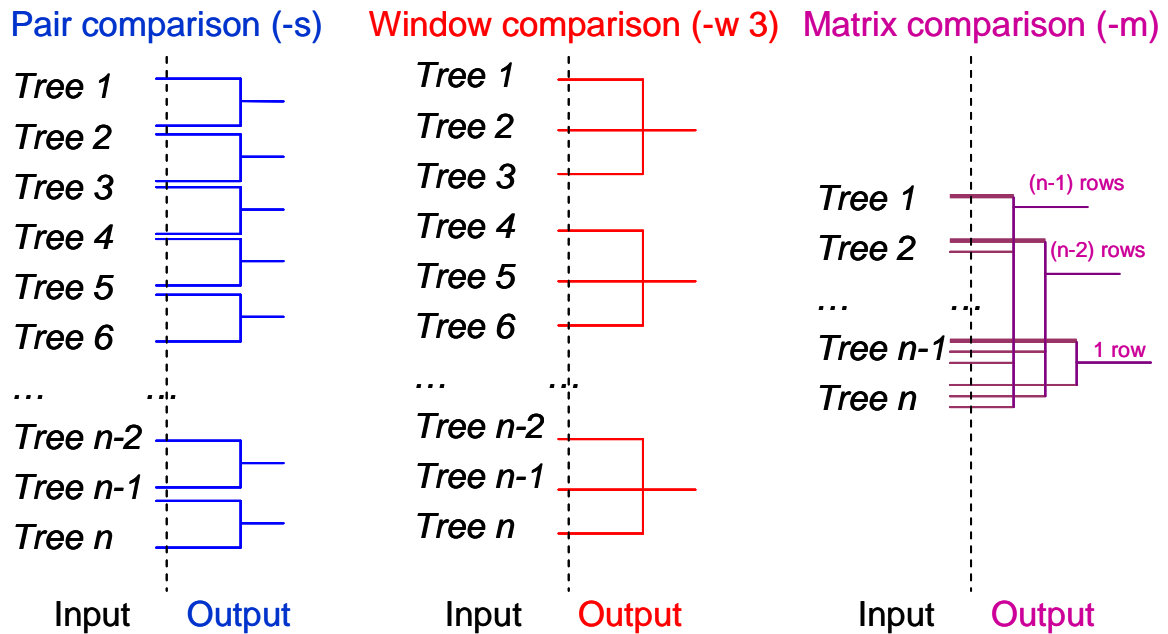
- The comparison mode options (only one option should be specified):
 - `-s` – pair comparison mode; every two neighboring trees in the input file are compared,
 - `-w <size>` - window comparison mode; every two trees within a window with a specified size are compared – the average distance and the standard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

- 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\beast\testBSP.newick

Invocation: java -jar MSdist.jar -w 3 -i testBSP.newick -o testBSP.newick_w_3.out

Console output:

```

2011-04-22 23-27-29: Start of scanning input file: testBSP.newick
2011-04-22 23-27-29: End of scanning input file: testBSP.newick
2011-04-22 23-27-29: 11 valid trees found in file: testBSP.newick
2011-04-22 23-27-29: Start of calculation...please wait...
2011-04-22 23-27-29: 0.00% completed...
2011-04-22 23-27-29: 10.00% completed...
2011-04-22 23-27-29: 20.00% completed...
2011-04-22 23-27-29: 30.00% completed...
2011-04-22 23-27-29: 40.00% completed...
2011-04-22 23-27-29: 50.00% completed...
2011-04-22 23-27-29: 60.00% completed...
2011-04-22 23-27-29: 70.00% completed...
2011-04-22 23-27-29: 80.00% completed...
2011-04-22 23-27-29: 90.00% completed...
2011-04-22 23-27-29: 100.00% completed.
2011-04-22 23-27-29: End of calculation.

```

Output file testBSP.newick_w_3.out:

```
state MatchingSplit (avg)      MatchingSplit (stddev)
1      60.0000      5.8878
2      10.6667      2.4944
3      14.0000      1.4142
4      9.0000 0.0000
-----
Summary:
Name   Avg   Std   Min   Max   Count
MatchingSplit 23.416666666666668 21.1979755322688 9.0 60.0 4
```

6. License

Copyright (C) 2010, Damian Bogdanowicz

This program is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but **WITHOUT ANY WARRANTY**; without even the implied warranty of **MERCHANTABILITY** or **FITNESS FOR A PARTICULAR PURPOSE**. See the GNU General Public License for more details.

You should have received a copy of the GNU General Public License along with this program. If not, see <http://www.gnu.org/licenses/>.