

GeodeMAPS 0.1 Manual

Megan Owen *

October 9, 2008

1 Introduction

GEODEMAPS computes the geodesic distance between two phylogenetic trees, which was described by Billera et al. [1] This algorithm is described in detail in Owen [2].

2 Installation

GEODEMAPS is available for download at www.cam.cornell.edu/~maowen/geodemaps.html. It requires Java version J2SE 5.0 which can be downloaded for free at java.sun.com.

Download and unzip the directory `geodemaps_v0.1`. This directory contains the license file and `geodemaps.jar`, the program that computes the geodesic distance. It also holds the sub-directory `example`, which contains the files to follow the example in Section 4, and the sub-directory `source` code, which contains the source code for `geodemaps.jar`.

3 Running GeodeMAPS

GEODEMAPS is run from the command line, using Java.

```
usage: java -jar geodemaps.jar [options] treefile.
```

treefile is the name of the file containing the list of trees. `geodemaps.jar` computes the geodesic distance between each pair of trees in the file. The trees should be one per line, in Newick format. See <http://evolution.genetics.washington.edu/phylip/newicktree.html> for a description of the Newick format. Although this has not been extensively tested, multifurcating trees should be handled properly.

Options:

*maowen@samsi.info. SAMSI, P.O. Box 14006, Research Triangle Park, NC 277094006, USA. This work was supported in part by NSF grant DMS-0555268 at Cornell University.

-a *algorithm* uses *algorithm* to compute the geodesic distances. Current options are **divide** to run GEODEMAPS-DIVIDE, and **dynamic** to run GEODEMAPS-DYNAMIC. The default is **dynamic**.
-d double check results, by computing each distance with the target tree as the starting tree and vice versa; default is false
-h, --help displays this message
-n normalize (vector of the lengths of all edges has length 1)
-o *outfile* store the output in the file *outfile*
-u unrooted trees (default is rooted trees)
-v, --verbose verbose output

The **-a** option allows users to select which variation of GEODEMAPS they wish to run. Use **-a dynamic** to run the implementation of GEODEMAPS-DYNAMIC, as specified in [2]. Use **-a divide** to run the implementation of GEODEMAPS-DIVIDE, as specified in [2]. GEODEMAPS-DYNAMIC is slower than GEODEMAPS-DIVIDE, but uses less memory.

The **-d** option computes each inter-tree distance twice, and compares the results. Assume the trees are numbered in the order they are listed in the input file. Then the first computation uses the lower-numbered tree as the starting tree and the higher-numbered tree as the target tree, while the second computation used the higher-numbered tree as the starting tree and the lower-numbered tree as the target tree. The two distances should be the same, so an error message is output if they are not.

The **-h** or **--help** option displays a brief usage message, and exits.

The **-n** option normalizes the tree edge lengths before computing the geodesic distance. Specifically, for each tree, we divide the length of each of its edges, both interior and terminal, by the length of the vector of all of these edge lengths. The length of the vector of the new edge lengths is 1.

The **-o *outfile*** option stores the output in the file *outfile*. Otherwise, the default output file is `output.txt`. In the output file, there is one line for each pair of trees. The trees are numbered by their order in the data file, starting at 0. The first column holds the starting tree, the second column holds the target tree, and the final column holds the length of the geodesic distance between the two trees.

The **-u** option is used to indicate that the trees in `treefile` are unrooted. In this case, GEODEMAPS selects one leaf, and roots each of the trees, so that this leaf is the root. Neither the choice of leaf, nor this rooting process, changes the geodesic distance between the trees. The default is rooted trees.

The **-v** or **--verbose** option prints out detailed information about the geodesic.

4 Example

This example corresponds to the trees given in Example 5.2 in [2], where the leaf 3 is replaced by the pair of leaves *3a* and *3b* in each tree. In the first tree, this edge has length 1, while in the second tree this edge has length 05. Thus, the two trees have the common edge $\{3a, 3b\} \cup \{1, 2, 4, 5, 6\}$.

To run this example, type the following at the command line, in the `geodemaps_v0.1` directory:
`java -jar geodemaps.jar -v example/example5.2.common_edge.txt`
 Two files will be generated in the `geodemaps_v0.1` directory: `output.txt`, which contains the geodesic distance between the trees, and `geo_0_1`, which is a copy of the verbose output displayed on the screen and explained below. Compare `output.txt` to `example5.2.common_edge_output.txt` in the `example` subdirectory, and `geo_0_1` to the file of the same name in the `example` subdirectory.

Using the verbose (-v) option, gives the following output, which we will now explain.

```
Starting tree: (((((4:1,5:1):0.88,(3a:1,3b:1):1):0.47,2:1):0.73,1:1):0.83,6:1);
Target tree: (((((3a:0.2,3b:1):0.5,4:1):0.15,2:1):0.87,(5:1,6:1):0.42):0.7,1:1);
```

Starting tree edges:

Length	Edge
0.88	4,5
1.0	3a,3b
0.47	3a,3b,4,5
0.73	2,3a,3b,4,5
0.83	1,2,3a,3b,4,5

Target tree edges:

Length	Edge
0.5	3a,3b
0.15	3a,3b,4
0.87	2,3a,3b,4
0.42	5,6
0.7	2,3a,3b,4,5,6

The first section of the output file gives the two trees, in the Newick format from the input file. Next, the edges of each tree are listed. The first column is the length of the edge, and the second column is the leaves below that edge in the tree. These leaves uniquely specify the edge, as they are simply the elements in one of the two blocks forming the partition of the leaves, made by the edge.

Leaf contribution squared 0.64

(Leaf contribution squared = square of the length of the vector whose i-th element is the absolute value of the difference between the length of the edge ending in leaf i in the first tree and the length of the edge ending in leaf i in the second tree.)

The second section specifies the contribution the edges ending in leaves make to the geodesic distance. In this example, only the edge to leaf 3a changes length, from 1 to 0.2. The leaf contribution squared is calculated by finding the absolute value of the change in length for each edge, and summing the squares of these values. In this example, the leaf contribution squared is $|1 - 0.2|^2 + \sum_{i=1}^6 (|1 - 1|^2) = 0.64$. See [2, Section 2.3] for more details.

Common edges are: (Length = abs. value of difference in length between the two trees)

Length	Edge
0.5	3a,3b

Common edges contribution squared: 0.25

(sum of squares of above differences in length)

=====

The third section of the output specifies the contribution of the common edges to the geodesic

distance. The squared contribution is calculated by taking the absolute value in the change of length for each common edge, and summing the squares of these values. In this example, there is one common edge, so the common edges contribution squared is $|1 - 0.5|^2 = 0.25$. See [2, Section 2.3] for more details.

The following section repeats for each pair of subtrees that are found by removing the common edges from each tree. All leaves below a common edge are represented as one leaf, using the notation of one of the original leaves with a * beside it. In this example, there is one common edge, dividing the leaves 3a and 3b from the rest of the leaves. Since the two leaves 3a and 3b form a trivial subtree, that contributes 0 to the geodesic, we only consider the other subtree. In this subtree, the leaf 3a* represents the two original leaves 3a and 3b. This subtree problem is equivalent to the problem solved in [2, Example 5.2].

Now finding the geodesic between the following subtrees, which have no edges in common:

Leaves or subtree representatives in subtrees:

1
2
3a*
4
5
6

Starting subtree edges:

Length	Edge
0.88	4,5
0.47	3a*,4,5
0.73	2,3a*,4,5
0.83	1,2,3a*,4,5

Target subtree edges:

Length	Edge
0.15	3a*,4
0.87	2,3a*,4
0.42	5,6
0.7	2,3a*,4,5,6

Geodesic distance between above subtrees, ignoring edges ending in leaves: 2.68321

Ratio sequence corresponding to the geodesic:

Ratio 0: 1.130047

Total length and corresponding edges dropped:

0.997647	4,5
	3a*,4

Total length and corresponding edges added:

0.882836	3a*,4
	2,3a*

Ratio 1: 1.185714

Total length and corresponding edges dropped:

0.83 1,2,3a*

Total length and corresponding edges added:

0.7 2,3a*,4

Ratio 2: 1.738095

Total length and corresponding edges dropped:

0.73 1,2

Total length and corresponding edges added:

0.42 2,3a*

For each pair of subtrees, the above information describes the geodesic between them. It does this by giving the ratio sequence associated with the geodesic. Each ratio corresponds to a transition between orthants in the tree space. The edges added and dropped to get to the new orthant are specified (by the set of leaves below the edge), along with their lengths. If the lengths of the edges dropped or added are l_1, l_2, \dots, l_k , then the total length is $\sqrt{\sum_{i=1}^k l_i}$. The ratio is the total length of the edges dropped over the total length of the edges added. The ratios should be non-descending. Notice that the computed subtree geodesic distance differs from the geodesic distance given in [2, Example 5.2] due to rounding in computing the latter distance.

Geodesic distance between start and target tree is 2.844225

Finally, the leaf and common edge contributions are combined with the geodesic distances between the subtrees to give the geodesic distance between the original trees. The formula for the square of the geodesic distance is

$$(\text{leaf contrib.})^2 + (\text{common edge contrib.})^2 + \sum_{i=1}^{\# \text{subtrees}} (\text{geo. dist. between } i^{\text{th}} \text{ subtree pair})^2$$

5 License

Copyright (C) 2008 Megan Anne Owen

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/>.

References

- [1] L. Billera, S. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27:733–767, 2001.

- [2] M. Owen. *Distance Computation in the Space of Phylogenetic Trees*. PhD thesis, Cornell University, 2008.