Geodesic Treepath Problem (GTP) algorithm Manual

Megan Owen, J. Scott Provan *†

March 17, 2017

1 Introduction

The Geodesic Treepath Problem (GTP) algorithm is a polynomial algorithm to compute the geodesic distance between two phylogenetic trees, which was introduced by Billera et al. [1] This algorithm is described in detail in [3].

2 Installation

GTP is available for download. It requires Java version SE 6 which can be downloaded for free at java.sun.com.

The following files are available at
http://comet.lehman.cuny.edu/owen/code.html

- the executable file gtp.jar
- the GNU general public license under which this code is released
- the source code for gtp.jar
- a directory containing the input and output files to follow the example in Section 4

3 Running GTP

GTP is run from the command line, using Java. The following command assumes you are in the directory containing `gtp.jar`.

`usage: java -jar gtp.jar [options]` *`treefile`*.

*M. Owen is with the Department of Mathematics and Computer Science at Lehman College, City University of New York (CUNY), New York, NY 10468. E-mail: megan.owen@lehman.cuny.edu

†J.S. Provan is with the Department of Statistics and Operations Research at the University of North Carolina, Chapel Hill, NC, 27599. E-mail: provan@email.unc.edu.

treefile is the name of the file containing the list of trees. `gtp.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. The input trees can also be multifurcating (non-binary).

Options:

`-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 `-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. This option should be chosen when the outermost bracket in the Newick representations of the trees contains three branches or subtrees. In this case, GTP 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, which correspond to there being just two branches or subtrees in the outermost bracket in the Newick representations of the trees.

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

4 Example

This example corresponds to the trees given in Example 5.4 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 $\{3a, 3b\}|\{1, 2, 4, 5, 6\}$.

To run this example, put `gtp.jar` and the directory `example` in the same directory. Type the following at the command line, in that directory:

```
java -jar gtp.jar -v example/example5.4_common_edge.txt
```

The file `output.txt`, which contains the geodesic distance between the trees, will be generated

in the current directory and details about the geodesic (available in the file `example/geo_0_1`) are output to the screen. This output can be piped to a file to store it. Compare `output.txt` to `example5.4_common_edge_output.txt` in the `example` subdirectory, and the screen output to `geo_0_1` 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:
 Edge ID  Length  Edge
 0        0.88    4,5
 1        1       3a,3b
 2        0.47    3a,3b,4,5
 3        0.73    2,3a,3b,4,5
 4        0.83    1,2,3a,3b,4,5
Target tree edges:
 Edge ID  Length  Edge
 0        0.5     3a,3b
 1        0.15    3a,3b,4
 2        0.87    2,3a,3b,4
 3        0.42    5,6
 4        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 [3, Section 4] for more details.

```
Common edges are:  (Length = abs.  value of difference in length between the two
trees)
 Edge ID  Length  Edge
 1        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 [3, Section 4] 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 orginal 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 5.4].

```
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:
 Edge ID  Length  Edge
 0        0.88    4,5
 2        0.47    3a*,4,5
 3        0.73    2,3a*,4,5
 4        0.83    1,2,3a*,4,5
Target subtree edges:
 Edge ID  Length  Edge
 1        0.15    3a*,4
 2        0.87    2,3a*,4
 3        0.42    5,6
 4        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:
Combinatorial type:  {0,2}/{1,2};{4}/{4};{3}/{3};

Ratio 0:  1.13004777
Total length and corresponding edges dropped:
 0.997647  4,5
           3a*,4,5
```

```
Total length and corresponding edges added:
 0.882836  3a*,4
           2,3a*,4

Ratio 1:  1.18571429
Total length and corresponding edges dropped:
 0.83  1,2,3a*,4,5

Total length and corresponding edges added:
 0.7  2,3a*,4,5,6

Ratio 2:  1.73809524
Total length and corresponding edges dropped:
 0.73  2,3a*,4,5

Total length and corresponding edges added:
 0.42  5,6
--------------------------------------------------------------------------------
```

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, ..., 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 5.4] due to rounding in computing the latter distance. The combinatorial type of the geodesic summarizes which edges (given by their IDs) are added and dropped at each step. This representation is unique for each combinatorial type of geodesic between two tree topologies.

```
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

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

5 License

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] Megan Owen. Computing geodesic distances in tree space. *SIAM Journal on Discrete Mathematics*, 25(4):1506–1529, 2011.

[3] Megan Owen and J Scott Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(1):2–13, 2011.