Documentation for Wag2020 Program

Ward C. Wheeler
Division of Invertebrate Zoology,
American Museum of Natural History,
200 Central Park West, New York, NY, 10024, USA;
wheeler@amnh.org

February 3, 2021

Running Title: Wag2020

0.1 Introduction

This is the first version of documentation for the program Wag2020. This program is designed to produce a variety of phylogenetic distance-based trees and output the result in multiple formats as well. All source code, precompiled binaries, test data, and documentation are available from `https://githib.com/wardwheeler/wag2020`.

This first version is brief.

0.2 Input Distance Matrix Format

Input distance matrix should be in cvs format (can be exported from spreadsheet programs), with first line of taxon (terminal) names and square matrix following. First column (from second row) contains distances, not repeat of names.

Matrices must be symmetrical and non-negative.

Beware of the MS-Excel non-conforming CVS output. A final line feed needs to be added to Excel created CVS files.

0.3 Output Formats

Graph outputs can be in either Graphviz 'dot' or Newick formats. Dot files can be visualized in a variety of ways using Graphviz (e.g. dot, neanto, twopi) into pdf, jpg and a large variety of other formats. Newick outputs of trees can be visualized in a large number of programs (e.g. FigTree; `http://tree.bio.ed.ac.uk/software/figtree/`)

Example dot command line:
dot -Tpdf myDotFile.dot > myDotFile.pdf

For some reason on OSX the 'pdf' option does not seem to work. You can use '-Tps2' and that will generate a postscript file that preview can read and convert to pdf.

0.4 Command options

There are multiple program options that require specification. There are defaults for all except the input matrix. Parameters are given with options in a range 'a to b' (a-b) with any value in the interval, or alternates 'a or b' (a—b). File options require a valid filename. All commands are followed by a colon ':' before the option with no spaces on the command-line, but spaces are allowed in an input parameter file. Capitalization (for commands, but not filenames) is ignored. Commands can be in any order (or entered from a file). Commands in a parameter file can include comment lines (preceded by two dashes '- -'), with each command on a single line.

If a single argument is passed to the program, that argument is assumed to be a parameter file. Command-lines must contain at least two arguments. Parameter files must be text only–created by a text editor or exported as text only Otherwise, unpredictable weirdness may occur.

- Input:“distanceMatrixFile”
  Default:none
  This commands specifies the input distance matrix. The filename must be in double quotes. The matrix must be square and all values non-negative.

- stub:“stubFileName”
  Default:none
  This commands specifies the stub of output files if specified, otherwise the input sistance matrix filename is used. The stub string must be in double quotes. This command will overwrite an existing file.

- Outgroup:“outputGroupName”
  Default:first taxon in distance matrix
  This commands specifies the outgroup for rooting the tree. The taxon name must be in double quotes.

- Output:“outputFile”
  Default:none
  This commands specifies the output tree filename. The filename must be in double quotes. This command will overwrite an existing file.

- additionsequence:best|random:n|NJ|WPGMA
  Default:best
  Specifies the distance tree construction method. “Best” denotes Farris (1972) method (dWag), “random” is a random addition sequence with number of replicates specified by “n”using Farris (1972) distance calculations, NJ is to generate Neighbor-Joining (Saitou and Nei, 1987), and WPGMA for, well WPGMA (Sokal and Michener, 1958).

- excludedTaxa:“taxonNametFile”
  Default:none
  This commands specifies the the filename containing the names of taxa to be excluded from analysis. The filename must be in double quotes.

- buildSet:best|random:n|all
  Default:best
  Specifies the trees retained after the initial build to be (potentially) retained on for refinement. “Best” denotes retention of those of lowest cost, “all” retains all trees found, “best” can take an optional integer argument to limit the number of trees of lowest cost retained.

- keepSet:best|random:n|all
  Default:best
  Specifies the trees retained after refinement. “Best” denotes retention of those of lowest cost, “all” retains all trees found, “best” can take an optional integer argument to limit the number of trees of lowest cost retained.

- firstPairChoice:closest|furthest
  Default:closest
  Specifies the initial pair of taxa to start tree building with the “best” addition sequence

(Farris, 1972). “Closest” is the pair with smallest distance and “furthest” that pair with largest distance. If “random” is specified for addition sequence, this option is ignored.

- refinement:none|OTU|SPR|TBR
  Default:none
  Specifies the method to refine trees generated by the initial build. “OTU” removes and readds each terminal taxon, “SPR” and “TBR” perform those methods of tree refinement. NJ an WPGMA tree can be refined, but the tree cost determination method is that of dWag (Farris, 1972). Refinement methods are those of Wheeler (in prep.).

- outputSet:best:n|random:n
  Default:best
  Specifies the method of retaining trees for both build and refinement.. “Best” denotes retaining trees of lowest cost limited to ‘n’, “random” retains ‘n’ trees uniformly chosen at random.

The program requires at least one input graph file and at least two input graphs (they could be in the same file).

0.5 Program Use

The program is invoked from the command-line as in:

```
wag2020 input:testData.csv” additionSequence:best ...
```

Execution in Parallel

By default the program will execute using a single process core. By specifying the options ‘+RTS -NX -RTS’ where ‘X’ is the number of processors offered to the program. These are specified after the program as in (for 4 parallel threads):

```
wag2020 +RTS -N4 -RTS other options...
```

Acknowledgments

The author would like to thank NSF REU DBI-1358465, DARPA SIMPLEX N66001-15-C-4039, and Robert J. Kleberg Jr. and Helen C. Kleberg foundation grant “Mechanistic Analyses of Pancreatic Cancer Evolution” for financial support.

Bibliography

Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. *American Naturalist* 106:645–668.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4:406–425.

Sokal, R. R. and Michener, C. D. 1958. A statistical method for evaluating systematic relationships. 38:1409–1438.