Documentation for EUNCon 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

March 24, 2021

Running Title: EUNCon

0.1 Introduction

This is the first version of documentation for the program EUNCon. This program is designed to produce a variety of phylogenetic supergraphs from input graphs in a variety of formats 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/EUNCon`.

This first version is brief.

0.2 Input Graph Formats

Graphs may be input in the graphviz "dot" format `https://graphviz.org/`, Newick (as interpreted by Gary Olsen; `https://evolution.genetics.washington.edu/phylip/newick_doc.html`), Enhanced Newick Cardona et al. (2008), and Forest Enhanced Newick (defined by Wheeler, 2020) formats.

Quickly, Forest Enhanced Newick (FEN) is a format based on Enhanced Newick (ENewick) for forests of components, each of which is represented by an ENewick string. The ENewick components are surrounded by '<' and '>'. As in <(A, (B,C)); (D,(E,F));>. Groups may be shared among ENewick components.

0.3 Output Formats

Graph outputs can be in either Graphviz 'dot' or FEN 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. FEN outputs of single trees (ie forest with a single component) are rendered as enewick. Newick files can be visualized in a large number of programs (e.g. FigTree; `http://tree.bio.ed.ac.uk/software/figtree/`, Dendroscope; `https://uni-tuebingen.de/fakultaeten/mathematisch-naturwiss fachbereiche/informatik/lehrstuehle/algorithms-in-bioinformatics/software/`). When FEN/Enewick files are output, leaf vertices are modified if they have indegree > 1, creating a new node as parent to that leaf and redirecting the leaf's in-edges to that new node with a single edge connecting the new node to the leaf.

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 (> blah.ps) that Preview can read and convert to pdf.

0.4 Command options

There are only a few program options that require specification. There are defaults for all but input graphs. 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. For input graphs, wildcards are allowed (ie '*' and '?'). All commands are followed by a colon ':' before the option with no spaces. Capitalization (for commands, but not filenames) is ignored. Commands can be in any order (or entered from a file as stdin '< filename).

- Reconcile:eun|cun|majority|strict|Adams
  Default:eun
  This commands specifies the type of output graph. EUN is the Edge-Union-Network Miyagi and Wheeler (2019), CUN the Cluster Union Network (Baroni et al., 2005), majority (with fraction specified by 'threshold') specifies that a values between 0 and 100 of either vertices or edges will be retained. If all inputs are trees with the same leaf set this will be the Majority-Rule Consensus (Margush and McMorris, 1981). Strict requires all vertices be present to be included in the final graph. If all inputs are trees with the same leaf set this will be the Strict Consensus (Schuh and Polhemus, 1980). Adams denotes the Adams II consensus (Adams, 1972).

- Compare:Combinable|identity
  Default:combinable
  Species how group comparisons are to be made. Either by identical match [(A, (B,C))≠(A,B,C)], combinable sensu Nelson (1979) [(A, (B,C)) consistent with (A,B,C)]. This option can be used to specify "semi-strict" consensus (Bremer, 1990).

- Threshold:(0-100)
  Default:0
  Threshold must be an integer between 0 and 100 and specifies the frequency of vertex or edge occurrence in input graphs to be included in the output graph. Affects the behavior of 'eun' and' majority.'

- Connect:True|False
  Default:False
  Specifies the output graph be connected (single component), potentially creating a root node and new edges labeled with "0.0".

- EdgeLabel:True|False
  Default:True
  Specifies the output graph have edges labeled with their frequency in input graphs.

- VertexLabel:True|False
  Default:False
  Specifies the output graph have vertices labeled with their subtree leaf set.

- OutFormat:Dot|FENewick
  Default:Dot
  Specifies the output graph format as either Graphviz 'dot' or FEN.

- OutFile:filename
  Default:euncon.out
  Specifies the output graph file name. No conventions are enforced.

- Any string that does not contain a colon, ':', is assumed to be an input graph file.

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:
euncon reconcile:eun threshold:0 outfile:myOutput graphFile1.dot graphFile2.tre

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):

euncon +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

Adams, E. N. 1972. Consensus techniques and the comparison of taxonomic trees. *Syst. Zool.* 21:390–397.

Baroni, M., Semple, C., and Steel, M. 2005. A framework for representing reticulate evolution. *Annals of Combinatorics* 8:391–408.

Bremer, K. 1990. Combinable component consensus. *Cladistics* 6:369–372.

Cardona, G., Russelló, F., and Valiente, G. 2008. Extended newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics* 9. doi: 10.1186/1471-2105-9-532.

Margush, T. and McMorris, F. R. 1981. Consensus n-trees. *Bull. Math. Biol.* 43:239–244.

Miyagi, M. and Wheeler, W. C. 2019. Reconciling trees using phylogenetic edge union networks. 35:688–694.

Nelson, G. 1979. Cladistic analysis and synthesis: Principles and definitions, with a historical note on Adanson's Familles des plantes (1763-1764). *Syst. Zool.* 28:1–21.

Schuh, R. T. and Polhemus, J. T. 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera). *Syst. Zool.* 29:1–26.

Wheeler, W. C. 2020. Phylogenetic Supergraphs. in prep.