Documentation for PhyGraph 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

April 20, 2021

Running Title: PhyGraph

0.1 Introduction

This is the first version of documentation for the program PhyGraph. This program is designed to produce a phylogenetic graphs from input data and graphs via heuristic searching of general phylogenetic graph space. All source code, precompiled binaries, test data, and documentation are available from `https://githib.com/wardwheeler/PhyGraph`.

This first version is brief.

0.2 Input Data Formats

Any character names in input files are (for now) ignored and internal names are created by appending the character number in its file to the filename as in "fileName:0". Qualitative data, and prealigned data include their index in their input files and unaligned data are treated as a single character.

fasta

fastc

TNT

Do not support everything, interleave yes, cc code and costs yes, but one set of commands per line. Costs A>B A/B syntax no spaces. Ambiguities not allowed. Must specify all transformations manually. ')' sets to non-additive, if wantadditive then need to reset to additive after
character state designations single characters
continuous and other multicharacter states–can be read, abinguitess or ranges (unlike tnt for continuous) are in squarebraskets with period as in [X.Y]
- ans ? are always missing/inapplicable. If DNA are not coded ACGT- =¿ 01234 but left as lettes, then to include gaps as a 5th state, just include another character, such the letter 'O' that is not an IUPAC ambiguity code for DNA (or amino acids for that matter) as an additional state. This can be used for matrix/Sankoff matrices as well. Amino acid sequences would be processed in the same way. Including gaps as information requires an extra state (such as letter 'O').
multi-character state designations (letters, numbers, etc) must be in their own "block" with spaces between them.
Continuous characters must be numbers only (float fine) and declares by cccode command as "additive", otherwise the number will be treated as non-additive character states.
nonAdd polymorphisms are [X.Y]–unless additive '-' for range
can't have '.' in multichar state (or single for that matter)
INherent ordering of DNA and AMino Acid codes is alphabetical (e.g. A,C,G,T.'-')

0.3 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 **?**, and Forest Enhanced Newick (defined by **?**) 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.

0.4 Output Graph 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

Mulitple "dot" graphs can be output in a single file. To create pdf and other formats the commandline would be (these files are named and numbered automatically):

dot -Tpdf -O myDotFile.dot

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.

0.5 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 **?**, CUN the Cluster Union Network (**?**), 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 (**?**). 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 (**?**). Adams denotes the Adams II consensus (**?**).

- 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 **?** [(A, (B,C)) consistent with (A,B,C)]. This option can be used to specify "semi-strict" consensus (**?**).

- 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:PhyGraph.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).

0.6 Program Use

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

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):

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

Acknowledgments

The author would like to thank DARPA SIMPLEX N66001-15-C-4039, the Robert J. Kleberg Jr. and Helen C. Kleberg foundation grant “Mechanistic Analyses of Pancreatic Cancer Evolution”, and the American Museum of Natural History for financial support.