PhylogeneticGraph
Program Documentation
Version 0.1

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

August 19, 2021

Contents

1 Introduction

This is the initial version of documentation for the program PhylogeneticGraph (`PhyG`). This program is designed to produce phylogenetic graphs from input data and graphs via heuristic searching of general phylogenetic graph space.

`PhyG` is a sucessor program to `POY` (Gladstein and Wheeler, 1997; Wheeler et al., 2005; Varón et al., 2008, 2010; Wheeler et al., 2013, 2015) https://github.com/wardwheeler/POY5, containing much of its functionality but extended to broader classes of input data and phylogenetic graphs. The phylogenetic graph inputs and outputs of `PhyG` include not only trees, but other forms including forests and both soft and hard-wired networks.

2 Overview of program use

At present, `PhyG` is operated solely via command-line interface.

Commands are entered via a scriptfile that contains commands, which may themselves contain references to other script files. These commands specify input files, output files and formats, graph type and search parameters.

```
phyg script.pg
```

3 QuickStart

3.1 Requirements

3.2 Obtaining and Installing `PhyG`

All source code, precompiled binaries, test data, and documentation are available from https://githib.com/wardwheeler/PhyGraph.

4 Commands

4.1 Command Structure

4.2 Commands

4.2.1 read

4.2.2 rename

4.2.3 report

5 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.

5.1 fasta

Single character sequence input (Pearson and Lipman, 1988).

5.2 fastc

Multicharacter sequence input. (Wheeler and Washburn, 2019).

5.3 TNT

The TNT (Goloboff et al., 2008) format is accepted here for specification of qualitative, measurement, and prealigned molecular sequence data. PhyG does not parse all the diversity of options that can be specified in TNT input files. 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.'-')

6 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, 2021) 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.

7 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.

8 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).

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

9 Program Use

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

9.1 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.

10 Bibliography

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.

Gladstein, D. S. and Wheeler, W. C. 1997. POY version 2.0. program and documentation available at http://research.amnh.org/scicomp/projects/poy.php. American Museum of Natural History, New York.

Goloboff, P., Farris, J. S., and Nixon, K. 2008. Tnt, a free program for phylogenetic analysis. *Cladistics* 24:774–786.

Pearson, W. R. and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *PNAS* 85:2444–2448.

Varón, A., Vinh, L. S., Bomash, I., and Wheeler, W. C. 2008. Poy 4.0. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.

Varón, A., Vinh, L. S., and Wheeler, W. C. 2010. POY version 4: Phylogenetic analysis using dynamic homologies. *Cladistics* 26:72–85.

Wheeler, W. C. 2021. Phylogenetic Supergraphs. Cladistics, in press.

Wheeler, W. C., Gladstein, D. S., and De Laet, J. 1996-2005. POY version 3.0. program and documentation available at http://research.amnh.org/scicomp/projects/poy.php (current version 3.0.11). documentation by D. Janies and W. C. Wheeler. commandline documentation by J. De Laet and W. C. Wheeler. American Museum of Natural History, New York.

Wheeler, W. C., Lucaroni, N., Hong, L., Crowley, L. M., and Varón, A. 2013. POY version 5.0. American Museum of Natural History. http://research.amnh.org/scicomp/projects/poy.php.

Wheeler, W. C., Lucaroni, N., Hong, L., Crowley, L. M., and Varón, A. 2015. POY version 5: Phylogenetic analysis using dynamic homologies under multiple optimality criteria. *Cladistics* 31. 189-196.

Wheeler, W. C. and Washburn, A. J. 2019. Fastc: a file format for multi-character sequence data. *Cladistics* 35:573–575.