MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 151–153 doi: 10.1111/j.1755-0998.2010.02890.x

COMPUTER PROGRAM NOTE

HapStar: automated haplotype network layout and visualization

A. G. F. TEACHER*‡ and D. J. GRIFFITHS†
*Department of Biological Sciences, Royal Holloway University of London, Egham Hill, Egham, Surrey TW20 0EX, UK,
†FoAM vzw, Koolmijnenkaai 30–34, 1080 Brussels, Belgium

Abstract

Haplotype networks are commonly used for representing associations between sequences, yet there is currently no straightforward way to create optimal layouts. Automated optimal layouts are particularly useful not only because of the time-saving element but also because they avoid both human error and human-induced biases in the presentation of figures. HapStar directly uses the network connection output data generated from Arlequin (or a simple user-generated input file) and uses a force-directed algorithm to automatically lay out the network for easy visualization. In addition, this program is able to use the alternative connections generated by Arlequin to create a minimum spanning tree. HapStar provides a straightforward user-friendly interface, and publication-ready figures can be exported simply. HapStar is freely available (under a GPLv3 licence) for download for MacOSX, UNIX and Windows, at http://fo.am/hapstar.

Keywords: drawing, haplotype, layout, minimum spanning, network, visualization

Received 9 March 2010; revision received 16 April 2010, 10 May 2010; accepted 13 May 2010

Introduction

Haplotype networks are commonly used as a method of visualizing differences between sequence data. Networks often more clearly represent these relationships than tree-formats, in part because they do not limit the connections to the linear, bifurcating mode of visualization that is used for trees (Posada & Crandall 2001). Networks also have a major advantage over trees in that they show the number of base-pair changes between sequences.

Arlequin is widely used freely available software for population genetics data analysis (Schneider *et al*. 2000), which can be used to identify haplotypes, and output connection distances between haplotypes. Alternative possible connections are also listed by Arlequin if the minimum spanning network option is chosen, although caution must be applied when using this option as connections are only generated among sampled haplotypes, without the possibility of inferring a missing haplotype node of degree three or above. At present, connection information from Arlequin can be imported into TreeView and visualized in a tree format (Page 1996), but to the best of our knowledge no software exists to view the data output as a network. Currently, the most widely used network software is TCS (Clement *et al*. 2000), which generates network connections from sequence data and shows the network in a viewing window. However, the networks generated are not always easy to decipher as there is no automatic-layout function and there is limited space in the viewing browser; haplotypes are frequently stacked on top of each other, and the manual states that the user will need to manually move them to achieve a desired layout. Sophisticated software called 'Network' (Fluxus-Engineering 1999-2009) is available for developing networks but again has no function for automatic layout and is very complicated to use, rendering it un-accessible to many users. Minspnet is software that can calculate a minimum spanning network from a distance matrix (Excoffier & Smouse 1994); however, the output generated is simply the connections between the haplotypes.

There is a clear need for a straightforward program to automatically draw and lay out standard haplotype networks. As no software is currently available to adequately draw such networks, many authors draw these out manually, potentially introducing human error. Furthermore, by manually drawing networks, an element of decision making can take place as to how best to lay out the network; this can lead to figures that artificially

Correspondence: A. Teacher, Fax: +358-9-19157694; E-mail: amber.teacher@helsinki.fi
‡ Current address: Ecological Genetics Research Unit, Department of Biosciences, P.O. Box 65, FI-00014 University of Helsinki, Finland.

emphasize aspects of the network (for example artificially centralizing a certain haplotype within the figure). The lack of automatic network lay-out is also becoming increasingly problematic as technology is advancing and allowing much larger data sets to be produced, and we are now at the stage where it is becoming unfeasible to manually lay-out such networks.

Program description

HapStar is written in Python and as such requires the prior installation of the latest version of Python on the computer on which it is being used (free to download; http://python.org/download). HapStar is free to download (under a GPLv3 licence) for MacOSX, UNIX and Windows, at http://fo.am/hapstar. The download includes a comprehensive user guide ('Readme' file) and several example input files. HapStar has been developed to read standard Arlequin output files in the form of connection data, which can be directly copied and pasted into a whitespace (space or tab), delimited text file. However, the input file is simple and easy to create (a list of haplotype pairs plus connection lengths) and so network connections generated by any other program can also be used. HapStar can also generate a minimum spanning tree (a network that connects all haplotypes in a single graph without cycles) if alternative connections are provided in the input file. A tick-box option is available to switch between a standard network and a minimum spanning tree. Minimum spanning trees are calculated using Prim's Algorithm (Prim 1957). If two connection routes are of equal length, this algorithm will choose arbitrarily but consistently which route to keep. Caution should be applied when using this option, as this will only produce one possible minimum spanning tree rather than all potential trees.

HapStar has a single viewing window in which the input file can be loaded through a 'Load' button and standard file viewing system. Several example input files are provided with the download. Once loaded, the starting network appears in the viewer. Clicking on ''Run'' starts the process of optimizing the layout, and the process can be stopped and started by toggling the 'Run' button. The program uses a force directed algorithm, also known as a spring model, which essentially repels the disconnected nodes of the branches while attracting the connected ones until they reach an optimal format (Di Battista *et al.* 1998). The links between nodes are represented as springs which constrain the node positions; the model is run for successive iterations, and as it does so the positions gradually relax into an optimal state given the lengths of the connections between them. This results in a better graph for viewing and understanding the network or tree.

Branch lengths are fixed (within slight flexibility to allow the force-directed algorithm to function) so that they are representative of the relative genetic distances. Black dots along the branches represent a 'missing haplotype' that was not included in the sample. Occasionally long branches can overlap, but if the optimization is temporarily stopped, the user is able to adjust such a branch by dragging it within the viewer – the optimization can then be re-started from this new position. The radius of the node circles and the text size can be individually controlled for the whole network to make it easier to view. Some larger networks can suffer from oscillation of the nodes when running, but this can be stopped using the 'Speed' slider to reduce the amount of force the algorithm uses. The 'Speed' slider is the only user-defined parameter that affects the layout optimization directly – lower speeds are better for large and complex networks with many nodes, whilst higher speeds achieve a faster optimization for small or straightforward networks. The speed can be adjusted while the optimization is running, and the effect is clearly visible to the user. The longer the algorithm is left to run, the better the result will be; the majority of networks will be at a near-optimal layout within two minutes.

The image can be zoomed in or out with the + and – buttons, which means that the whole network can be viewed in the window regardless of how big the network is, the window itself can also be re-scaled. The radius of the nodes and the text size within the nodes can be altered directly. The network image can be dragged around the viewing screen with the left mouse button if the user wants to focus on a particular area. Once the user is happy with the layout, the final image can be exported as an SVG file (see Fig. 1 for an example network output). If further editing of the image is required,

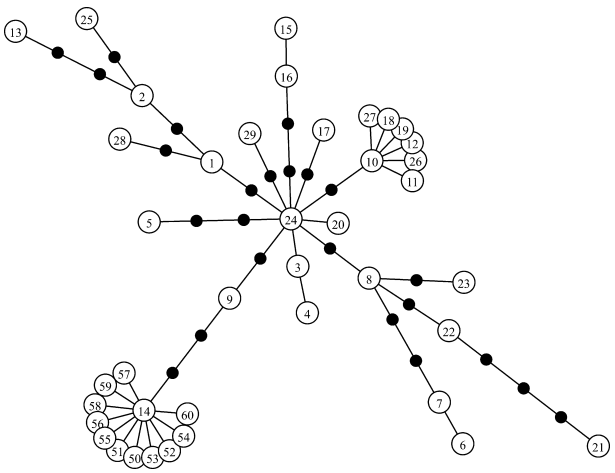

Fig. 1 A network exported directly from HapStar, with no additional editing. This can be recreated using the 'example.txt' test input file provided with the download.

for example the addition of colour or text, then SVG files are the most easily manipulated and this can be performed using software such as InkScape (free open-source SVG graphics editor; Bah 2007). HapStar also allows export of the networks in a form (DOT) that can be read by GraphViz (Gansner & North 2000), an open-source graph visualization toolkit.

Acknowledgements

Thanks to Ian Barnes, Meirav Meiri and Selina Brace for helpful suggestions during the development stages of HapStar. This work was funded by the European Union FP7 ERA-NET program, BiodivERsA.

References

Bah T (2007) *Inkscape: Guide to a Vector Drawing Program*. Prentice Hall Press, Upper Saddle River, NJ, USA.

Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9**, 1657–1660.

Di Battista G, Eades P, Tamassia R, Tollis IG (1998) *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Excoffier L, Smouse PE (1994) Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics*, **136**, 343–359.

Fluxus-Engineering (1999-2009) http://www.fluxus-engineering.com/sharenet.htm accessed February 26, 2010.

Gansner ER, North SC (2000) An open graph visualization system and its applications to software engineering. *Software: Practice and Experience*, **30**, 1203–1233.

Page RDM (1996) TreeView: an application to display phylogenetic trees on personal computers. *Bioinformatics*, **12**, 357–358.

Posada D, Crandall KA (2001) Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology and Evolution*, **16**, 37–45.

Prim RC (1957) Shortest connection networks and some generalizations. *Bell System Technical Journal*, **36**, 1389–1401.

Schneider S, Roessli D, Excoffier L (2000) *Arlequin: A Software for Population Genetics Data Analysis*. Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva, Geneva, Switzerland.