BIOINFORMATICS APPLICATIONS NOTE Vol. 14 no. 9 1998
Pages 819–820

# GeneTree: comparing gene and species phylogenies using reconciled trees

Roderic D. M. Page

*Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, UK*

Received on June 10, 1998; accepted on July 22, 1998

## Abstract

***Summary:*** *GeneTree is a program for comparing gene and species trees using reconciled trees. The program can compute the cost of embedding a gene tree within a species tree, visually display the location and number of gene duplications and losses, and search for optimal species trees.*

***Availability:*** *The program is free and is available at {{http://taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html}}.*

***Contact:*** *r.page@bio.gla.ac.uk*

The concept of reconciled trees dates from Goodman *et al.*'s (1979) attempts to reconcile disagreements between then accepted mammalian evolutionary relationships and those obtained from haemoglobin genes. Largely neglected until recently, reconciled trees are now receiving renewed attention from biologists and mathematicians (Guigó *et al*., 1996; Mirkin *et al*., 1995; Page, 1994; Page and Charleston, 1997a,b). However, the use of this technique in molecular biology and systematics has been hampered by a lack of software implementing the method. COMPONENT (Page, 1993) includes the ability to compute reconciled trees, but the algorithms that the program uses (Page, 1994) are slow and inefficient, and the program is only available on PCs. GeneTree was written to provide a fast, user-friendly implementation of reconciled trees on both PCs and Apple Macintosh computers. The program uses a modification of Eulenstein's (1997) linear time algorithm to perform the tree-mapping. This results in a considerable increase in computational speed compared to the implementation in COMPONENT.

A reconciled tree is constructed by embedding a gene tree within a tree for the species from which the gene sequences were obtained (Page and Charleston, 1997a). This embedding is unique, and corresponds to a mapping of each node in the gene tree onto a node in the species tree (hence the technique is also referred to as 'tree-mapping') (Page and Charleston, 1997b). The embedding can be represented by a reconciled tree which depicts the complete history of the gene within the species tree. Discordances between gene and species trees require a combination of gene duplications and losses to be postulated. The reconciled tree can be used to depict these in an intuitive manner compared to other equivalent procedures, such as tree 'annotating' (Eulenstein *et al*., 1997).

GeneTree is available for both Mac OS and Windows 95/NT 4.0 computers. The user interface includes a log window to record results of analyses, a text editor, and various tree and chart display windows. The program supports drag-and-drop opening of data files, printing, copy and pasting of graphics, text editing of data files, tree editing, and Balloon Help (Mac OS only). GeneTree uses essentially the same variant of the NEXUS format (Maddison *et al*., 1997) used by COMPONENT 2.0 (Page, 1993). Data files can be created within the program using its built in text editor. GeneTree can also save trees in NEXUS format. A manual is available in Adobe Acrobat PDF format from the GeneTree Web site.

GeneTree has two major uses. The first is to quantify and visualise the fit between gene and species trees. By displaying the reconciled tree, the program enables the user to locate and identify instances of gene duplication and loss in the evolution of a gene family. The program distinguishes between 'obvious' duplications where more than one sequence has been obtained from the same species, and duplications inferred from incongruence between gene and species trees. The later represent previously undetected instances of paralogy (Fitch, 1970). Due to uneven taxonomic sampling in the sequence databases, many gene losses may be more apparent than real. Hence, losses identified by the reconciled tree may represent genes that are actually present but as yet undetected. In these cases the reconciled tree can be viewed as a predictive tool for directing the search for these 'missing' genes.

The second use of GeneTree is to infer species phylogeny from trees for one or more genes. The optimal species tree is that in which the gene trees can be embedded with the least cost. Three optimality criteria are available for selecting species trees: numbers of duplications and losses, duplications only, and deep coalescences (Maddison, 1997). The program uses a heuristic search employing three tree perturbation strategies to attempt to find the optimal species tree

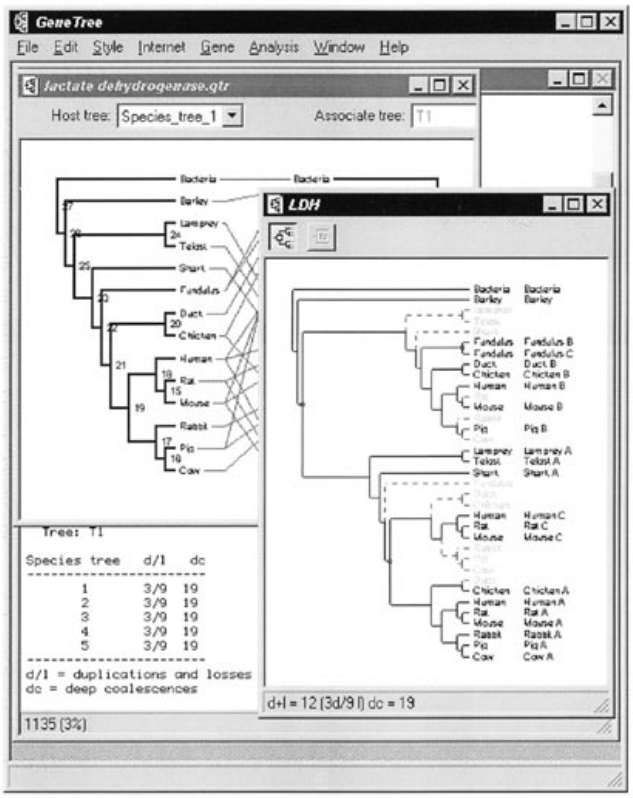


**Fig. 1.** A screen shot of GeneTree running under Windows 95 (the Mac OS version is essentially identical in appearance). Visible are the log window recording results of analyses, a window displaying the gene and species tree together and a window showing the reconciled tree.

(or trees). The relative performance of these strategies is discussed in Page and Charleston (1997a). The search can begin with either an existing tree, or one or more random trees. The user can also specify a constraint tree (Constantinescu and Sankoff, 1986) to restrict the search to a subset of the possible species trees (e.g. only those trees in which mammals are monophyletic).

Figure 1 shows a screen shot of a typical GeneTree session. The gene and species trees are displayed together in one window, and the corresponding reconciled tree is displayed in a separate window. For some applications of GeneTree see Page and Charleston (1997a,b) and Slowinski *et al.* (1997).

## Acknowledgements

I thank Joe Slowinski and Boris Mirkin for stimulating me to write the program, Mike Charleston for discussions on heuristic search strategies, and Oliver Eulenstein for sending me a manuscript describing his linear time algorithm for tree mapping. Part of this work was supported by NERC grant GR3/1A095.

## References

Constantinescu,M. and Sankoff,D. (1986) Tree enumeration modulo a consensus. *J. Classif.*, **3,** 349–356.

Eulenstein,O. (1997) A linear time algorithm for tree mapping. Arbeitspapiere der GMD No. 1046, St Augustine, Germany.

Eulenstein,O., Mirkin,B. and Vingron,M. (1997) Comparison of annotating duplications, tree mapping, and copying as methods to compare gene trees with species trees. In Mirkin,B., McMorris,F.R., Roberts,F.S. and Rzhetsky,A. (eds), *Mathematical Hierarchies in Biology*, American Mathematical Society, Providence, Rhode Island, pp. 71–93.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19,** 99–113.

Goodman,M., Czelusniak,J., Moore,G.W., Romero-Herrera,A.E. and Matsuda,G. (1979) Fitting the gene lineage into its species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, **28,** 132–168.

Guigó,R., Muchnik,I. and Smith,T.F. (1996) Reconstruction of ancient molecular phylogeny. *Mol. Phylogenet. Evol.*, **6,** 189–213.

Maddison,D.R., Swofford,D.L. and Maddision,W.P. (1997) NEXUS: An extensible file format for systematic information. *Syst. Biol.*, **46,** 590–621.

Maddison,W.P. (1997) Gene trees in species trees. *Syst. Biol.*, **46,** 523–536.

Mirkin,B., Muchnik,I. and Smith,T.F. (1995) A biologically consistent model for comparing molecular phylogenies. *J. Comput. Biol.*, **2,** 493–507.

Page,R.D.M. (1993) COMPONENT, Tree comparison software for Microsoft® Windows™, Version 2.0, The Natural History Museum, London.

Page,R.D.M. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, **43,** 58–77.

Page,R.D.M. and Charleston,M.A. (1997a) From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. *Mol. Phylogenet. Evol.*, **7,** 231–240.

Page,R.D.M. and Charleston,M.A. (1997b) Reconciled trees and incongruent gene and species trees. In Mirkin,B., McMorris,F.R., Roberts,F.S. and Rzhetsky,A. (eds), *Mathematical Hierarchies in Biology*, American Mathematical Society, Providence, Rhode Island, pp. 57–70.

Slowinski,J., Knight,A. and Rooney,A.P. (1997) Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. *Mol. Phylogenet. Evol.*, **8,** 349–362.