

ATV: display and manipulation of annotated phylogenetic trees

Christian M. Zmasek and Sean R. Eddy

Howard Hughes Medical Institute

Department of Genetics

Washington University School of Medicine

St. Louis, MO 63110 USA

Email: {zmasek,eddy}@genetics.wustl.edu

Key words: tree display, tree viewer, phylogenetic tree, java, phylogenomics

Abstract

Summary: *ATV (A Tree Viewer) is a Java tool for the display and manipulation of annotated phylogenetic trees. It can be utilized both as a standalone application and as an applet in a web browser.*

Availability: *ATV is available via WWW at <http://www.genetics.wustl.edu/eddy/atv/> and via FTP at <ftp://ftp.genetics.wustl.edu/pub/eddy/software/forester.tar.Z>*

Contact: *eddy@genetics.wustl.edu*

Introduction

Many proteins belong to large families consisting of subfamilies with different biological functions. This complicates efforts to infer the function of new proteins by computational sequence analysis. Neither of the two main sequence analysis methods handle large protein families satisfactorily in high-throughput automated annotation. Pairwise sequence similarity searches, exemplified by BLAST (Altschul *et al.*, 1990), lead to overly specific annotations. A new sequence in a protein family is always “most similar” to something, so it is difficult to recognize when the new sequence is the pioneer member of a novel functional subfamily. Profile search methods, exemplified by

HMMER (Eddy, 2000) and the Pfam database (Bateman *et al.*, 2000), lead to overly general annotations. They recognize that a new sequence fits a general profile of a family, but do not attempt to subclassify the sequence at all.

Phylogenetic inference is a sensible approach to subclassifying sequences, by grouping them hierarchically into evolutionary clades. The use of phylogenetic inference to improve genome sequence annotation has been termed “phylogenomics” by Eisen (1998). A key idea of phylogenomics is to distinguish sequences that have diverged by speciation (orthologues) from sequences that have diverged by duplication (paralogues). Although orthology does not equate with functional conservation, as is sometimes assumed, orthologues often do conserve more aspects of a protein’s function than paralogues do. We are working on automating a phylogenomic approach to improve Pfam-based annotations.

During phylogenomic analysis, gene trees are annotated with various data. Nodes are annotated as either a gene duplication or a speciation, and subtrees are annotated according to sequence function (as description and/or EC number). In addition, information about species (as name and/or taxonomy ID) and sequence names, branch lengths, and bootstrap values are likely to be present. We needed a tool for visualizing heavily annotated phylogenetic trees. Although a variety of excellent tree browsers exist, including DRAWTREE from the PHYLIP package (Felsenstein, 1993), TREEVIEW (Page, 1996), NIFAS (<http://www.cgr.ki.se/Pfam/nifas.html>), NJPLOT (Perriere and Gouy, 1996), and Phylodendron (<http://iubio.bio.indiana.edu/soft/molbio/java/apps/trees/>) none of them exactly suited our annotation needs. Hence, we developed our own design. The purpose of this application note is to make our viewer, ATV (A Tree Viewer), available to the community.

Features

ATV is mouse and menu driven. The user can choose which data elements to display on the tree. All the data fields associated with nodes can be edited. The tree can be rerooted on any branch. ATV allows visualization of very large trees (>500 sequences): the user can display any subtree of the tree, zoom in or out, or collapse

any subtree into a single node. The applet hyperlinks to SWISS-PROT entries for sequences with a SWISS-PROT name. Branches can be colored according to likelihood values associated with them. The Swing version (see below) of the application allows printing trees in color. Depending on the user's environment, it also allows tree images to be exported as PostScript or PDF files (which in turn gives the user the opportunity to employ graphics software to manipulate tree images beyond the capabilities of ATV). An example of ATV displaying an annotated tree is shown in Figure 1.

Trees can be read and saved in the standard "New Hampshire" format (Felsenstein, 1993), but this format is not suitable for storing annotated trees. Currently we use a simple extension of the format that we call "New Hampshire eXtended" format (NHX). In NHX, additional tag/value pairs are used to associate annotation with nodes (e.g. ":E=" is a tag for a EC number, ":S=" is a tag for a species name). In the long term, we envision replacing NHX with a structured markup language, such as the XML document type definition for the description of taxonomic relationships described in Gilmour (2000).

Implementation

ATV is coded in Java, for portability reasons. ATV can be used either as an applet in a web browser or as a standalone application. ATV should run on any platform for which a Java 1.1.x runtime environment is available. It has been tested on Red Hat Linux 6.1, SGI IRIX 6.5, Sun Solaris 5.6, and Microsoft Windows 95B and Windows NT Workstation 4.0 using various Java runtime environments from Sun Microsystems and Silicon Graphics. The ATV distribution includes the "forester" class library which we use as a toolkit for phylogenomics. Two versions of ATV exist. One version uses Swing graphics classes, and is less portable but more aesthetically pleasing. The other version uses basic AWT (Advanced Windowing Toolkit) and is more portable. It is straightforward to incorporate ATV and forester into other Java applications.

ATV is freely available under a BSD open source license. The ATV distribution includes all source code files, as well as extensive documentation (including a definition of the NHX format).

Acknowledgements

We would like to thank Peter Ernst for useful additions. This work was supported primarily by a grant from Pharmacia Corporation (formerly Monsanto), and also by the Howard Hughes Medical Institute and the NIH National Human Genome Research Institute.

References

- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.
- Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) The Pfam protein families database. *Nucleic Acids Res.* 28:263-266.
- Eddy,S.R. (2000) HMMER: Profile hidden Markov models for biological sequence analysis. Washington University School of Medicine, St. Louis, MO (<http://hmmer.wustl.edu/>).
- Eisen,J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8, 163-167.
- Felsenstein,J. (1993) PHYLIP: Phylogeny Inference Package, Version 3.5. University of Washington, Seattle, WA (<http://evolution.genetics.washington.edu/phylip.html>).
- Gilmour,R. (2000) Taxonomic markup language: applying XML to systematic data. *Bioinformatics* 16, 406-407.
- Page,R.D.M. (1996) TREEVIEW: An application to display phylogenetic trees on personal computers. *Comput. Applic. Biosci.* 12, 357-358.
- Perriere,G. and Gouy,M. (1996) WWW-query: an on-line retrieval system for biological sequence banks. *Biochimie* 78, 364-369.

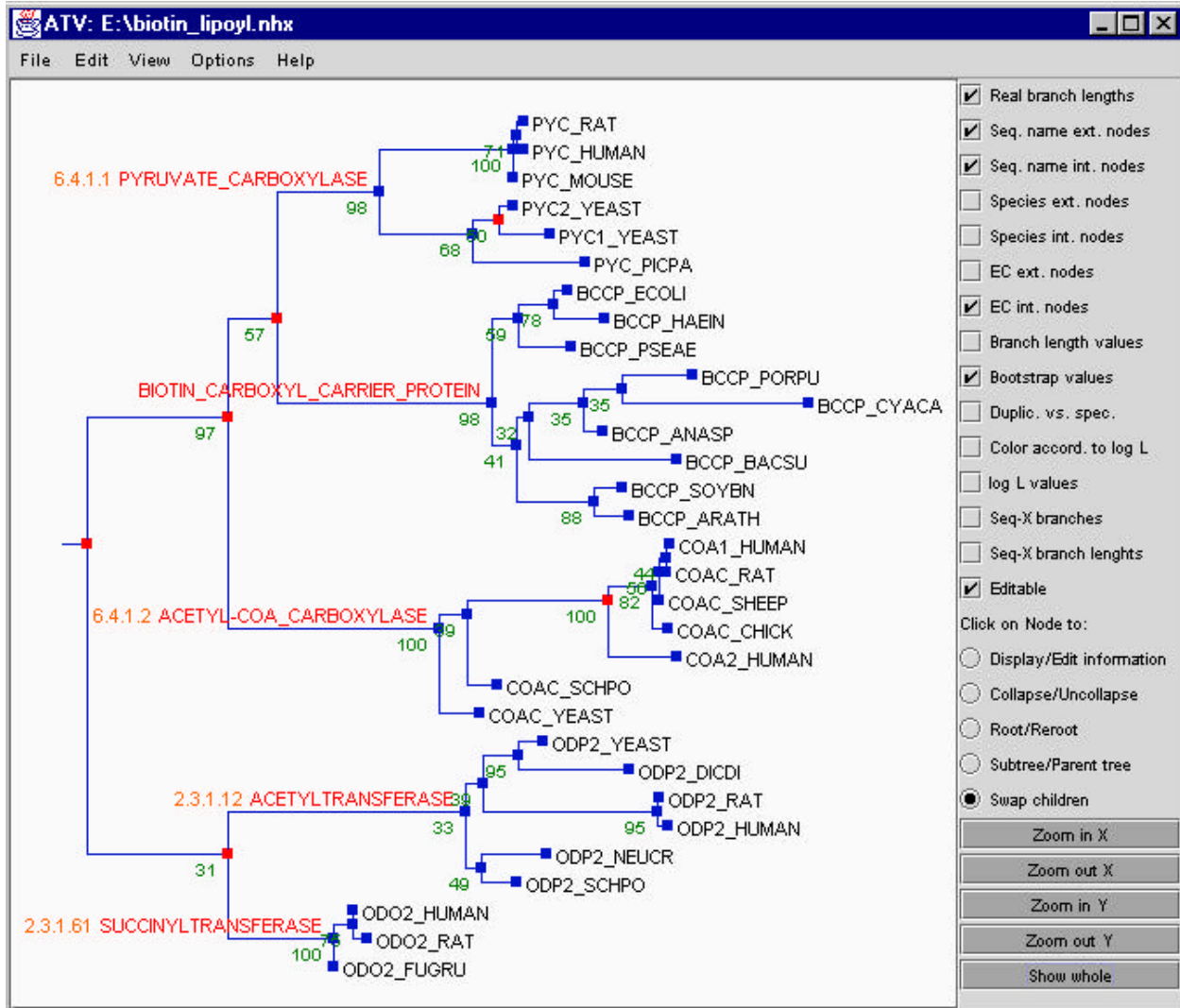


Figure 1

ATV displaying a phylogenetic tree of biotin-requiring enzymes. Red nodes indicate duplications, green numbers represent bootstrap values, orange numbers are EC numbers, and the functional description of subtrees is in red. The check boxes in the right side panel are used to choose which information is displayed, whereas the radio buttons are used to determine the behavior for node clicking.