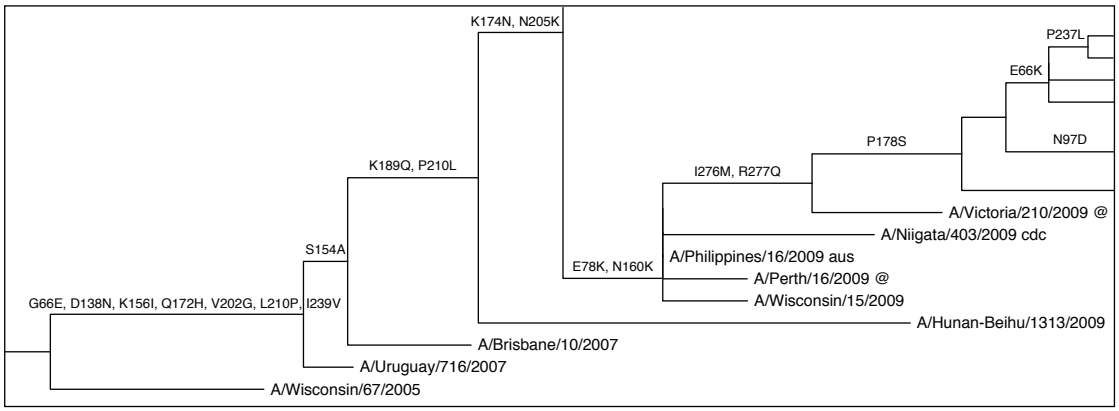
Annotating ancestral substitutions on a tree

Asif Tamuri (atamuri@nimr.mrc.ac.uk)

November 30, 2011



1 Introduction

This collection of software tools and utilities allows a user to take a codon alignment in FASTA format and produce an annotated phylogenetic tree showing substitutions along the tree. We use the following tools and libraries:

RAxML is a program for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees [1]. We use it to estimate the tree topology.

PAML is a suite of software tools for phylogenetic analysis [2]. We use it to estimate the branch lengths and infer the ancestral sequences.

PAL is a software library for molecular evolution and phylogenetics [3]. We use it to read, manipulate and write phylogenetic trees.

BioJava is a software framework for processing biological data [4]. We use it to read and parse FASTA format sequence files.

Figtree is a graphical viewer of phylogenetic trees. The annotated tree can be opened in Figtree and the tree can be prepared (as close as is possible) as you would like for printing or further editing in a program like Adobe Illustrator.

2 Installation

Currently, this application only runs on Mac OS X 10.6 and 10.7.

2.1 RAxML

RAxML can be downloaded from http://sco.h-its.org/exelixis/software.html. You need to compile from source. Please make sure that you use GNU GCC either from Xcode 4.1 (4.2 uses the Clang compiler, which doesn't work) or download from something like MacPorts.

2.2 PAML

PAML can be downloaded from http://abacus.gene.ucl.ac.uk/software/paml.html. You need to compile from source and, again, make sure you use the GNU GCC.

2.3 Java (JRE)

You need to be running JRE 1.6, which should already installed on Mac OS X 10.6 and 10.7.

2.4 Utilities

We recommend using Figtree to view the annotated trees. You can download it from http://tree.bio.ed.ac.uk/software/figtree/. The remaining libraries and small utilities are installed by hand by unzipping the 'annotator.zip' file.

3 Guide

The process has been designed to be as straightfoward as possible. However, we describe all the steps here to clarify the workflow. In theory, you only need to provide a sequence alignment (step 1 and 2) and hit 'RUN'!

3.1 Prepare the sequences and alignment

Your alignment must be a coding sequence alignment without any STOP codons. It is recommended that you remove any signal or trailing peptides, otherwise these will throw off the numbering of sites. The alignment must be in FASTA format. The first sequence in the alignment (i.e. the one at the top of the FASTA file) is used as your outgroup sequence. There is no requirement for any specific labeling. However, it is recommended that you avoid commas or parenthesis in your FASTA labels as these can cause issues when reading and writing tree files.

3.2 Create a new directory for this run

As we run several different programs that produce many different files, some with the same names, it is recommended that you create a new (i.e. empty) directory for each run. Put your FASTA alignment into this new directory.

3.3 Building the tree and ancestral reconstruction (RAxML and PAML)

You don't necessarily need to know all these details; they are provided here for your information and to assist with any troubleshooting. One you have saved your alignment in a new directory, start the Annotator program by double-clicking on annotator.jar. The first time you run the program, you will need to supply the path to the RAxML and PAML programs. Use the 'Browse...' buttons to select the raxmlHPC and baseml executables. These are saved and you don't need to enter them for subsequent runs. Finally, select your FASTA alignment using the 'Browse...' button and click 'Run'.

1. You can name your FASTA file as you like. However, before we start any processing, we rename the FASTA alignment to a file named 'alignment'. We do this because all the subsequent programs will expect a file with this name.

2. We run a small utility program that reads the FASTA alignment and produces (a) a file called 'alignment.names', which is a list of all the labels for your sequences and (b) a PHYLIP formatted alignment called 'alignment.raxml.phylip'. PHYLIP formatted alignments are required by RAxML and PAML. Additionally, the sequences are renamed to 'seq_1', 'seq_2', 'seq_3' and so on. This is to ensure that there are no name conflicts or errors in accepted name format in the subsequent programs.

3. RAxML is run with 'alignment.raxml.phylip' to estimate the tree topology by maximum likelihood. We use the GTRGAMMA model.

4. A small utility then roots the resulting RAxML tree by your outgroup sequence, which should have been the first sequence in your original FASTA alignment file (hence, we know this as 'seq_1'). This is required to make sure that the ancestral reconstruction goes in the right direction.

5. We run the PAML program 'baseml' to estimate the branch lengths and perform the ancestral reconstruction.

3.4 Producing the annotated tree

Once we have run RAxML and PAML, the program is ready to produce the annotated tree file. It will read the optimised trees and ancestral substitutions from the PAML output and the sequence names from the 'alignment.names' file and write a new tree called 'substitutions.tree'. It will also create a table of

all synonymous and non-synonymous substitutions in a tab-separated values files called 'substitutions.tsv'. It specifies the node/branch number, codon change and amino acid change for the substitutions. This can be opened in a text editor or a program like Excel.

You may find it useful to take a look at the list of your sequences in 'alignment.names'. Each line has the name of one sequence. You can edit the names of your sequences in this file. The only hard rules are to keep the same order of sequences and to avoid using commas and parentheses. For example, you could rename "A-URUGUAY-716-2007" to "A/Uruguay/716/2007 cdc @". You may find it easier to do this kind of renaming in this file rather than in a tree viewing program or graphics editing program. If you make changes to the alignment.names and save the file (it must keep the same name), you can generate a new tree file by clicking the 'Re-annotate' button. Note that you do not have to do this re-annotating at the time of your initial analysis. Simply open the Annotator program (by double-clicking on annotator.jar), select the same alignment file (in the same directory, with all the miscellaneous results files) and click 'Re-annotate'.

3.5 Viewing the annotated tree

Start the Figtree viewing program and open the 'substitutions.tree' file. There are many options in Figtree to manipulate the rendering of the tree. The following are an example of the types of things you can do using the panels in the sidebar:

1. In the 'Layout' panel, move the 'Expansion' slider to spread out the taxa so the labels do not overlap.

2. In the 'Trees' panel, tick the 'Order nodes' checkbox and set ordering to 'decreasing'.

3. The 'Tip Labels' panel allows you to choose what to display as the taxa name: the names, all substitutions, non-synonymous substitutions, node number, or 'FULL', which displays the node number, taxon name and non-synonymous substitutions together.

4. Tick the 'Node Labels' panel and in the panel select 'NUMBER' for display. This will show the internal node numbers which are needed to decipher the list of substitutions in the 'substitutions.tsv' file.

5. Tick the 'Branch Labels' panel and in the panel selection 'NONSYNSUBS' for display. This shows any non-synonymous substitutions that occurred along the branch.

6. Finally, you can use the 'Layout' panel again to spread out the tree to avoid overlapping labels by using the 'Zoom' and 'Expansion' sliders.

Each time you have the tree as you like, you can export to PDF or a graphics image. If you export to PDF (or EPS) you can edit the result in Adobe Illustrator (or Inkscape).

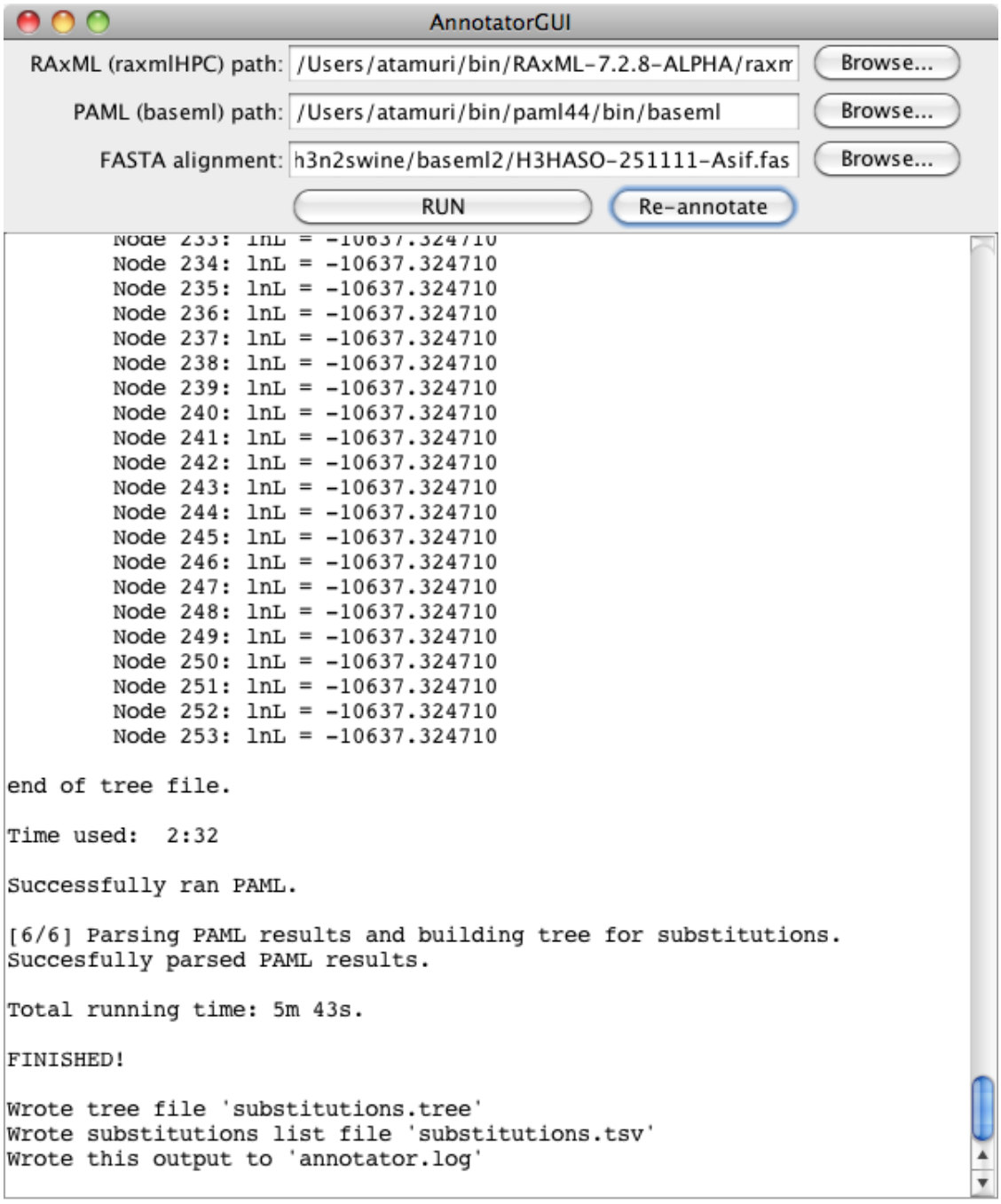
4 References

1. Stamatakis, A. 2006. RAxML-VI-HPC: Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models. Bioinformatics 22(21):2688–2690.

2. Yang, Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution 24: 1586-1591.

3. Drummond, A., and K. Strimmer. 2001. PAL: An object-oriented programming library for molecular evolution and phylogenetics. Bioinformatics 17: 662-663.

4. R.C.G. Holland; T. Down; M. Pocock; A. Prlić; D. Huen; K. James; S. Foisy; A. Dräger; A. Yates; M. Heuer; M.J. Schreiber. 2008. BioJava: an Open-Source Framework for Bioinformatics. Bioinformatics 24 (18): 2096-2097.


AnnotatorGUI
RAxML (raxmlHPC) path:
/Users/atamuri/bin/RAxML-7.2.8-ALPHA/raxm
Browse...
PAML (baseml) path:
/Users/atamuri/bin/paml44/bin/baseml
Browse...
FASTA alignment:
h3n2swine/baseml2/H3HASO-251111-Asif.fas
Browse...
RUN
Re-annotate
Node 233: lnL = -10637.324710
Node 234: lnL = -10637.324710
Node 235: lnL = -10637.324710
Node 236: lnL = -10637.324710
Node 237: lnL = -10637.324710
Node 238: lnL = -10637.324710
Node 239: lnL = -10637.324710
Node 240: lnL = -10637.324710
Node 241: lnL = -10637.324710
Node 242: lnL = -10637.324710
Node 243: lnL = -10637.324710
Node 244: lnL = -10637.324710
Node 245: lnL = -10637.324710
Node 246: lnL = -10637.324710
Node 247: lnL = -10637.324710
Node 248: lnL = -10637.324710
Node 249: lnL = -10637.324710
Node 250: lnL = -10637.324710
Node 251: lnL = -10637.324710
Node 252: lnL = -10637.324710
Node 253: lnL = -10637.324710
end of tree file.
Time used: 2:32
Successfully ran PAML.
[6/6] Parsing PAML results and building tree for substitutions.
Succesfully parsed PAML results.
Total running time: 5m 43s.
FINISHED!
Wrote tree file 'substitutions.tree'
Wrote substitutions list file 'substitutions.tsv'
Wrote this output to 'annotator.log'