

Phylogenetic Tutorial

Survey Course in Bio Crystallography and Bioinformatics
Lima, Peru, March 17-29, 2009

Tutorial on Phylogenetic analysis

Outline of document:

1. Questions
2. Schematic of generating a phylogenetic tree
3. Ensure Alignment is correct
4. Generate evolution Matrix
5. Draw a phylogenetic Tree
6. Generate support

This tutorial uses the modules in phylip version 3.85 as well as clustalx release 2.0.9 (www.clustal.org). The files needed are 1) fasta file containing all the sequence from the *M tuberculosis* FABG BLAST search and 2) the clustalw alignment file generated from past tutorial.

1. Questions:

- a. Test if the added information in a maximum likelihood tree is any different then a Neighbor joining tree.

2. Schematic of generating a phylogenetic tree from protein sequences

There are many ways to make a phylogenetic tree. We are going to explore ways of doing this. What you need to start with is a sequence alignment. Then you have to know how all the sequences compare with each other. Then you need to plot them on a “tree” that shows the comparison.

a. FASTA -> Clustalx -> Ctree

The first way of doing this is using 2 programs. Clustalx will do most of the work in aligning the sequences and comparing them for us. Ctree just shows us how the sequences compare.

b. Alignment File -> ProML -> seqboot -> consense -> Ctree

This schematic represents the steps needed to generate a phylogenic tree with the modules in Phylip. It all starts with a good alignment file of your sequences. If the alignment is bad the end tree will be bad. In the following example we are going to use the alignment file generated from clustalx. However you could use any alignment file you have on hand. The alignment file is the only piece of the schematic requiring user input. From there Proml generates a matrix of evolution from your alignment. This will output a tree file in newick format. To add support to say your tree is correct we will run Seqboot. To display the tree in a GUI, Ctree will be used.

3. Ensure Alignment is Correct

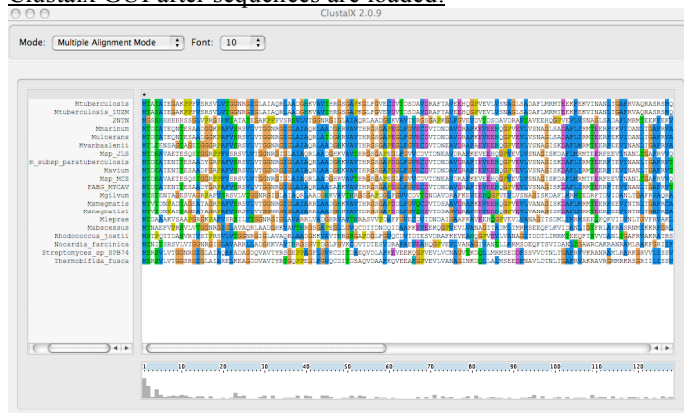
1. We already have an alignment file but this time we are going to regenerate it with a stand-alone program, not a web form. As such we are going to start with the fasta file again. However this time we need to set up our fasta file with simpler identification headings
2. Copy the 20 sequences from FABG BLAST search into the 5_phylo directory.
3. Make new copy of 20sequence.fasta file as 20sequence.fasta
4. Edit 20sequence.fasta so species is right after the > symbol, ensure all the names are unique. If any of the sequence you downloaded are from the same species give them a 1 or 2 *before* the name (important as phylip format will truncate the length of the name at 10 characters). As well make sure there are no spaces in the name.

```
>gi|15608621|ref|NP_215999.1| 3-oxoacyl-[acyl-carrier protein] reductase FabG1  
[Mycobacterium tuberculosis H37Rv]
```

Phylogenetic Tutorial

>1_Mtuberculosis

5. Open clustalx GUI
6. File -> load sequence -> 20sequences.fasta
Clustalx GUI after sequences are loaded.

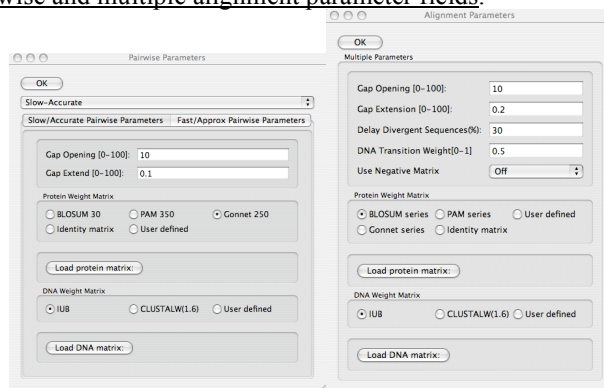


Sequences

Histogram of conservation

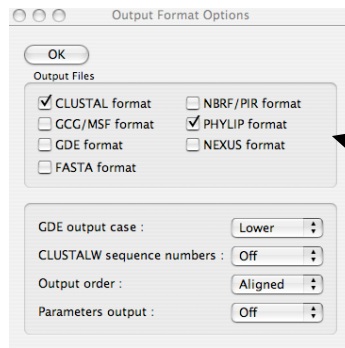
7. To set the alignment parameters.
 - a. Alignment -> alignment Parameters -> Pairwise alignment parameters: Multiple alignment parameters.

GUI representation of pairwise and multiple alignment parameter fields.



- b. Clustal does its alignments by comparing each protein with every other protein (pairwise alignment) then uses those to generate a larger multiple alignment. Both parameters are important.
8. Ensure to output the alignment file for PHYLIP
 - a. Alignment -> Output Format options

Phylogenetic Tutorial



9. Run alignment with default values to start.
 - a. Alignment -> Do complete alignment
 - b. Save output guide file and both alignment files (.aln, .phy) to the same directory.



GUI representation of alignment in clustalx

- c. The colors in this alignment represent columns of conservation in the FABG sequences. There is a gray histogram at the bottom of the display that also represents conservation.
 - d. What regions do you see are not well conserved in the FABG sequences?
 - e. What regions are highly conserved?
10. The sequence looks really good. There does not need to be any manual editing. We can be going to run the evolution matrix

4. Generate evolution Matrix

1. Clustalx can generate a tree based on the guide tree from aligning the sequences. This is a neighbor-joining tree. Which means it is going to take the 2 sequences with the highest identity and put them next to each other on a tree. Then do this recursively till all the sequences are placed. This is an easy way to create the tree so let us do this.
 - a. Click Trees=> DrawTree.
 - i. Save phymlip tree as default (.ph extension).
2. Next we will generate a maximum likelihood tree. This tree is generated not on sequence identity between 2 sequences but by an expected pattern of mutational changed from one amino acids to another then determines the most likely arrangement of branches on the tree.
 - a. Using the .phy file generated alignment copy the file and change the name of the file to infile
 - i. 20sequencesn.phy => infile
 - b. Copy the infile to the /phymlip3.65/exe directory
 - c. Open "proml" (double click the icon)

GUI of proml

Phylogenetic Tutorial

```

proml.out

Maximum Likelihood method, version 3.65

Settings for this run:
U Search for best tree? Yes
P JTT, PMB or PAM probability model? Jones-Taylor-Thornton
C One category of sites? Yes
R Rate variation among sites? constant rate of change
M Sites weighted? No
S Speedier but rougher analysis? Yes
G Global rearrangements? No
J Randomize input order of sequences? No, Use input order
O Outgroup root? No, use as outgroup species 1
I Analyze multiple data sets? No
I Input sequences interleaved? Yes
O Terminal type (IBI, PC, ANSI, none)? (none)
1 Print out the data at start of run No
2 Print indications of progress of run Yes
3 Print out tree Yes
4 Write out trees onto tree file? Yes
5 Reconstruct hypothetical sequences? No

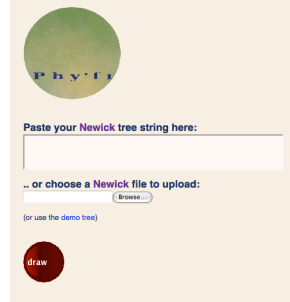
Y to accept these or type the letter for one to change

```

- d. Type “Y” into the consol and press enter.
- e. This will generate 2 files an “outfile” and a “outtree”
 - i. outtree is the .ph newick form tree
 - ii. outfile contains the log data from the maximum likelihood estimate.

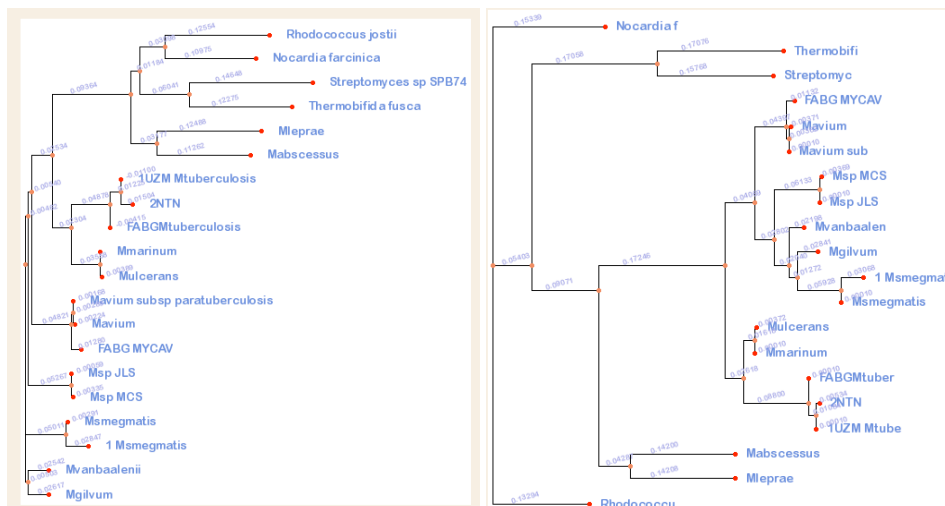
5. Draw the Tree

1. We are going to generate trees using 2 approaches. One will be through a website and the other will be with a downloaded java app.
2. Online generator of trees: PHY-FI <http://cgi-www.daimi.au.dk/cgi-chili/phyfi/go>
 - a. Go to the website above



PHY-FI homepage

- b. Open the .phy files generated from clustalx(.ph file) and proml (outtree) as text documents.
- c. Paste them one at a time into the generator and save the output
- d. Click “draw”



left is the clustalx tree data, right is the ML tree data

- e. How are these to trees the same and how are they different?

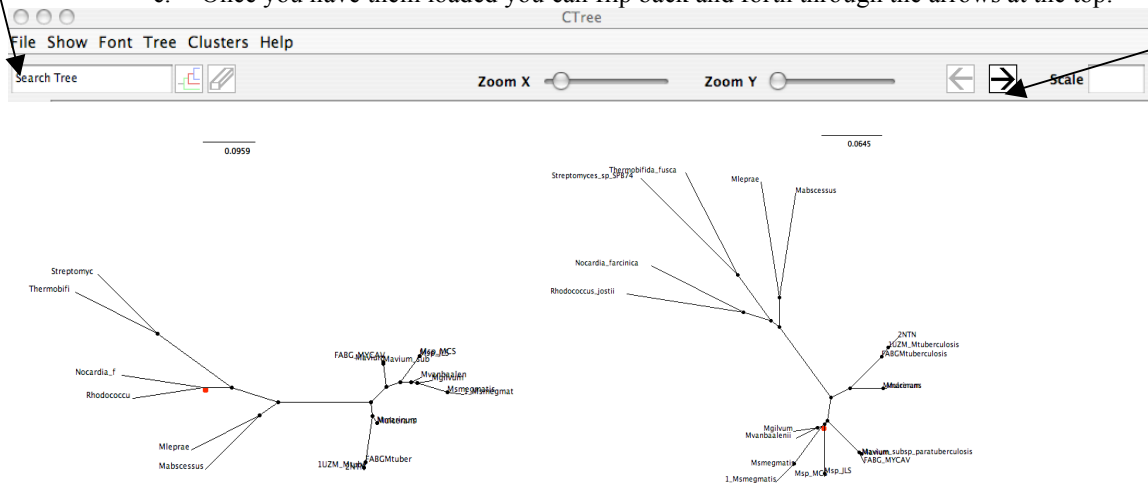
Phylogenetic Tutorial

f. What can we learn from them?

3. The other tree generator, Ctree <http://www.bioinf.manchester.ac.uk/ctree/>, should have already been downloaded and you just need to locate it.
 - a. Open the clustalx file (.ph) in ctree
 - i. File-> load Newick Tree(s)
 - b. Open the outtree file in ctree.
 - i. File-> load Newick Tree(s)
 - c. Once you have them loaded you can flip back and forth through the arrows at the top.

Load newick trees

Flip between trees



Left (ML tree), Right (guide tree from clustalx alignment)

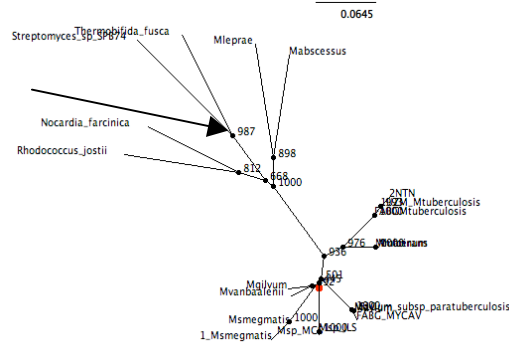
6. How are the trees the same?
7. How are they different?
8. Are these trees the same or different then the ones produced online?

6. Generate Support

To generate support for the accuracy of our tree we are going to perform a bootstrap analysis. We are going to bootstrap both the clustalx tree and the phylip tree. By bootstrapping an alignment what you are doing randomly resembles the columns in the alignment and generating trees. Then you ask does this new arrangement of columns still generate the first tree. You get a number. What you are looking for is a high number between sequences (90%) indicating no matter how the data is arranged the sequences are related. The first tree accurately represents the data.

1. Bootstrap clustalx data.
2. With clustalx open,
 - a. Select Trees -> Bootstrap N-J tree
 - b. Random number seed can stay as default but number of trails should ≥ 1000 . The more the better the support for your tree.
 - c. Click ok
 - d. This will generate a file with the extension .phb (b for bootstrap)
 - e. This file can be opened in Ctree where you select
 - i. Show-> bootstrap values (numbers will appear at nodes)
3. Clustalx tree with bootstrap values

Phylogenetic Tutorial



4. To generate Bootstrap values with phylip on the maximum evolution tree
5. Run seqboot from Phylip
- 6.

Bootstrapping algorithm, version 3.65

```
Settings for this run:
D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
J      Bootstrap, Jackknife, Permute, Rewrite? Bootstrap
%      Regular or altered sampling fraction?  regular
B      Block size for block-bootstrapping?  1 (regular bootstrap)
R      How many replicates?  100
W      Read weights of characters?  No
C      Read categories of sites?  No
S      Write out data sets or just weights?  Data sets
I      Input sequences interleaved?  Yes
0      Terminal type (IBM PC, ANSI, none)?  (none)
1      Print out the data at start of run  No
2      Print indications of progress of run  Yes
```

Y to accept these or type the letter for one to change

```
Y to accept these or type the letter for one to change
r
Number of replicates?
1000
```

7. This generates a file “outfile” contains the 100 bootstrap runs.
8. Change outfile to infile
9. Run proml
10. change the option “number of dataset available” to yes

```
Y to accept these or type the letter for one to change
m
Multiple data sets or multiple weights? (type D or W)
d
How many data sets?
1000
Random number seed (must be odd)?
101
Number of times to jumble?
1
```

11. Then type “y” and enter
12. This run will take a bit, as each dataset undergoes a maximum likelihood tree generation.
13. Change the outtree to intree and rename the outfile to outfile_1
14. Open the phylip program “consense”
15. This will put all the trees together and give you a consensus tree (these are the bootstrap values)

Phylogenetic Tutorial

```

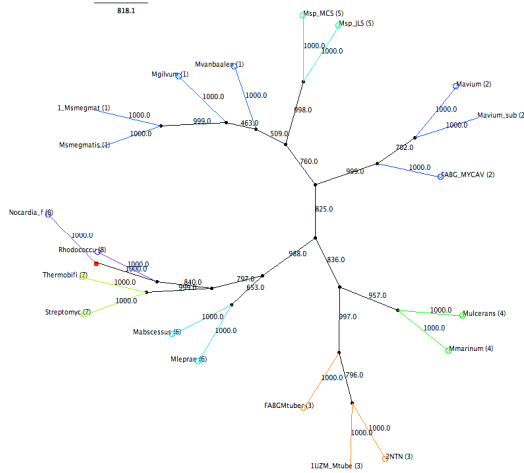
Consensus tree program, version 3.65

Settings for this run:
0      Consensus type (MR, strict, MR, ML): Majority rule (extended)
1      Outgroup root: No, use as outgroup species
2
3      Trees to be treated as Rooted: No
4      Terminal type (IBT PC, RNSI, none): (none)
5      Print out the sets of species: Yes
6      Print indications of progress of run: Yes
7      Print out tree: Yes
8      Write out trees onto tree file: Yes

Are these settings correct? (type Y or the letter for one to change)

```

16. Open the outtree in Ctree.



Based on the separation of FABG proteins what can we say about FABG from M tuberculosis?

References:

Phylogenetic trees made easy, second edition, Sinauer publishing, Barry Hall.

Phylip reference

Fredslund J (2006) "PHY·FI: fast and easy online creation and manipulation of phylogeny color figures" BMC Bioinformatics 2006, 7:315