Biol 381 - Bioinformatics Lab
Lab 9: Phylogenetic Analysis

Over the past few months, you've used several tools for sequence analysis - BLAST for finding sequences that are homologous to a query sequence, and multiple sequence alignment tools for examining similarity levels among several sequences. These tools have many important uses, but one thing they don't do well is tell you about the evolutionary relationships between sequences. For example, look at the MSA below. You can see where conserved domains might be, and that is quite interesting, but what if you wanted to know about how the sequence changed over time, or about how great the evolutionary distance is between sequences? It would be difficult-to-impossible to infer this information just by looking at groups of letters.

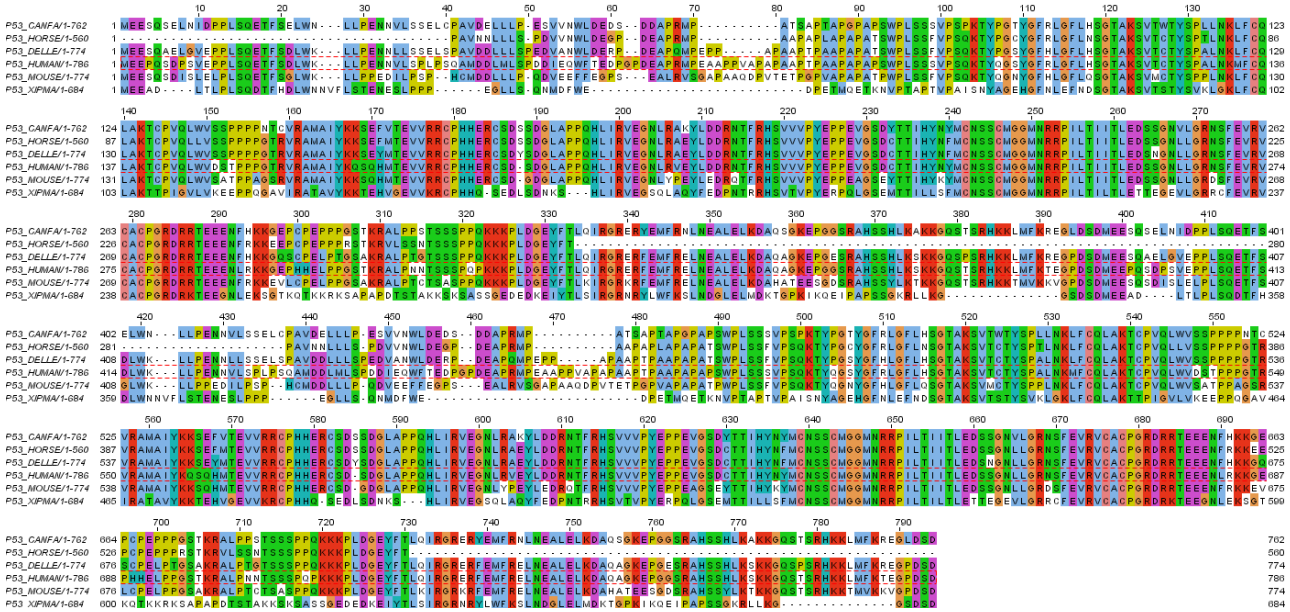

Figure 1: MSA Example. From http://ntoc.wordpress.com/2010/03/18/hello-world/.

With the goal of studying evolutionary relationships in mind, today we'll be working with another type of analysis that is much better suited to the task, phylogenetic analysis. Phylogenetic trees can be used to get a sense of how a sequence may have changed over time, and provide insight into the factors that may have affected its evolutionary path. You have already had some practice constructing phylogenetic trees using the phylogeny.fr tool for your homework in this course. In today's lab, we'll be using another phylogeny tool, MEGA, to construct phylogenetic trees. MEGA is often used in real research, so it is a very useful tool to have experience with!

Your write-up for this week will consist of answers to the questions in the handout and copies of all the files you generate during the lab.

Part 1: MEGA Basics

In this section, you'll work through several MEGA tutorials to help you get used to the program.

Step 1:
Read over the Hall MEGA paper and answer the following questions.

1. When building a **nucleotide** multiple sequence alignment in MEGA, what option should you always choose?

2. Which phylogenetic methods can only estimate unrooted trees?

3. When examining bootstrapped confidence levels, what is the typical percent confidence cutoff for useful information?

Step 2:
First, construct a multiple sequence alignment. Complete Examples 2.1, 2.2, and 2.3 in the Walk Through MEGA tutorials. (you can also access these example exercises by opening MEGA and clicking the "Tutorials" button on the bottom of the screen.)

Step 3:
Next you'll use MEGA's Sequence Data Explorer to visualize variable sites.
1. What are parsim-informative sites?
2. Explain what 0-, 2- and 4- fold degenerate sites are.

Complete MEGA tutorial Example 9.2 to visualize the variable sites in the sequences.

Step 4:
Next you'll calculate the codon-usage and dinucleotide frequencies of the sequences by completing MEGA tutorial 9.3.

Step 5:
Construct a phylogenetic tree using the Maximum Likelihood option. Complete MEGA tutorial Examples 11.1 and 11.2.

Part 2: Multiple Sequence Alignment and Sequence Analysis

Step 1:
Choose 10 nucleotide sequences to construct a multiple sequence alignment. (You can use the same genes/proteins you used for past MSAs, but make sure you have their mRNA sequences for this exercise.) Make sure that you also include an outgroup.

1. What do you think might be different between a nucleotide and a protein phylogeny for the same gene?

2. Which do you think is likely to more accurately represent the evolutionary relationships among sequences, and why?

Step 2:
Open the Alignment Editor in MEGA and select the file containing your sequences as input. Align the sequences using one of the available MSA algorithms.

Step 3:
Look closely at the alignment to make sure it does not contain errors, and answer Question 1 below. If you do find a section that doesn't make sense, you can remove it or realign the sequences using different parameters. Once you are satisfied with the quality of your MSA, export it in MEGA format (.meg) so that it can be used to construct a phylogenetic tree.

1. Does the alignment seem reasonable?

    a. Are there any sequences that do not appear to be homologous?

    b. Do all the sequences contain roughly the same conserved regions?

    c. Are there any known homologs that contain large gaps in the alignment? If so, you may need to adjust the gap penalty.

Step 4:
Visualize the 0-, 2- and 4- fold degenerate sites in your sequences.

1. Do these sites occur with equal frequency along the length of your sequences? If not, note where the sites cluster.

Step 5:
Compute the codon-usage and di-nucleotide frequencies of your sequences. Be sure to save the results in a text file.

Part 3a: Choosing a Phylogenetic Method

In this section, you'll begin to construct a tree for your final project gene from the multiple sequence alignment you generated above. In MEGA, the first step in this process is to choose the method you'll use to build the tree.

Figure 1 shows the process of constructing a phylogenetic tree, as well as providing some insight into how you might select the most appropriate method for your sequence data. Other resources you can refer to for choosing a method include this page (under "Phylogenetic Analysis") and your textbook beginning on p. 255.

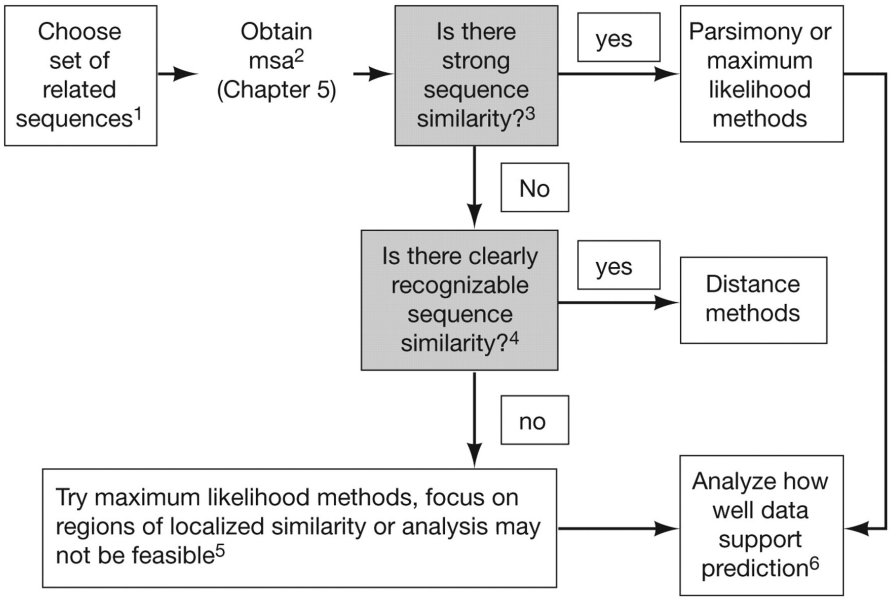

Figure 1: Flowchart showing process of phylogenetic tree construction. From http://cshprotocols.cshlp.org/content/2008/4/pdb.ip49.full

Step 1:
Start by selecting the method of construction from the Phylogeny menu.

1. Which method did you choose, and why?

2. Is this method a distance method or an inference method?

3. What are the potential downsides of this method?

Step 2:
Once you've selected a method, select the file containing your MSA from Part 2 as the input.

Part 3b: Choosing a Substitution Model and Constructing the Tree

Several different models for nucleotide substitution have been developed, and many of these are available in MEGA to be used in estimating trees. One main difference between the models lies in the rates of nucleotide change - given a mutation at particular position in a nucleotide sequence, are the probabilities of the three possible nucleotides appearing at that position equal or unequal? Another difference between the models is whether the four nucleotides are assumed to be present in the sequence in equal numbers.

Different substitution models will present a different picture of sequence evolution, so it is important to consider the properties of the model you choose for building a phylogenetic tree.

Figure 3 below shows the characteristics of some of the more well-known nucleotide substitution models, so it may be helpful in making a decision. Additionally, the first several slides from this phylogenetic methods lecture provide useful information. You may find your textbook helpful as well - explanation of substitution models starts on p. 246.

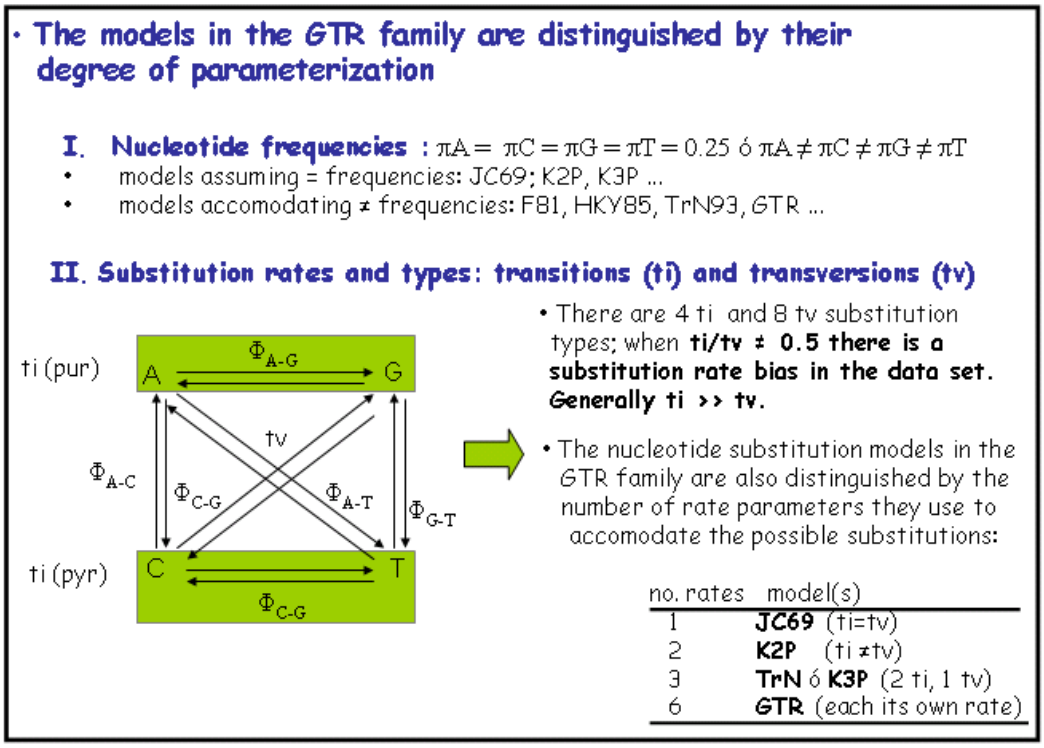

Figure 3: Common substitution models and their properties. From http://www.ccg.unam.mx/~vinuesa/Model_fitting_in_phylogenetics.html.

Step 3:
You should now see a box with options for the parameters of your tree. Make sure the Substitutions Type is set to Nucleotide, then select the options that you think will be best suited to your data for the model and Gap/Missing Data Treatment. You can change any other parameters you like as well, as long as you make a note of what you changed and why, and include it in your lab write-up.

(Note: Leave the Test of Phylogeny blank for now; we'll cover it in a later step.)

1. What model did you select for nucleotide substitution and why?

2. What option did you select for Gaps/Missing Data Treatment? Why?

3. List any other parameters you changed and explain your reasoning.

Step 4:
Compute the tree. Save the tree as a PDF by going to Image --> PDF File.

1. Does the tree seem reasonable when you examine it visually? Are there any branches or nodes that seem obviously out of place?

Part 4: Analyzing the Tree

In this last section, you'll analyze your tree. First, you'll get a sense of how likely the branches of your tree are to represent the true evolution of your sequence. Once you've done that, you'll examine the tree as a whole to see what it can tell you about the evolution of your gene.

Step 1:
Select the same method of tree construction you used in Part 3, and the same input MSA.

Step 2:
Make sure the parameters are the same as the ones you used for constructing your tree. Set the Test of Phylogeny to Bootstrap Method, and choose the number of replications.

Step 3:
Run the bootstrap analysis. Once your results have loaded, go to View --> Show/Hide and make sure Branch Lengths is checked - you'll include an analysis of branch lengths in your interpretation of the tree. Save the results as a PDF.

Step 4:
Examine the results and answer the questions below.

1. What are the confidence levels throughout the tree?

2. What can you say about the relative distances of the sequences by looking at the branch lengths

3. Are there any branches with low confidence levels? If so, what are they? What do you think the low confidence levels mean?

4. Overall, what can you conclude about the evolution of your gene?