**Alignment and Maximum Likelihood Phylogenetic Inference**

Data from:

Lutzoni FM. 1997. Phylogeny of Lichen- and Non-Lichen-Forming Omphalinoid Mushrooms and the Utility of Testing for Combinability among Multiple Data Sets. Syst Biol, 46:373-406.

Before you begin, you may want to download ClustalX:

<http://www.clustal.org/clustal2/>

You can use Clustal omega or Clustal W webservers if you prefer.

You will also need to download FigTree:

<http://tree.bio.ed.ac.uk/software/figtree/>

Your goal is to generate and use multiple sequence alignments to infer phylogenies of omphalinoid mushrooms. Start by opening the ClustalX application. This is a friendly user interface for the wildly popular clustal algorithm for multiple sequence alignment, which has a command line interface.

1. Load the sequences in the lutzoni1997.fasta file. These are sequences of the internal transcribed spacer 2 (ITS2) of the ribosomal RNA of mushrooms. It is located between the 5.8S and the 28S ribosomal RNAs. The spacers are spliced before the ribosomal RNAs fold to make up the ribosome in eukaryotic cells. What do the lengths of the different ITS2 sequences in this data set indicate regarding the homology of different nucleotides?

2. Go to the Alignment header, and click on “Set all parameters to default”. Under the same header click on alignment parameters, and click on “Reset all gaps before alignments”. Now, open “Multiple alignment parameters” in this last header. Focus on the gap opening and gap extension penalties. These are expressed as costs from 0-100. Using the default parameters for these penalties, click on Alignment, Do complete alignment. What is the length of the alignment? What do the asterisks indicate? How are those sites relevant to the alignment? Use the File header to save your alignment in fasta format.

3. Sensitivity analysis of gap penalties. Change the global gap opening and extension penalties first to 5 and 1, and then to 50 and 15. Save the resulting alignments in fasta format. What is the relationship between alignment length and gap opening penalty? Which one of the 3 alignments is “right”? Explain.

4. Inferring ML phylogenies. Use this web page: <http://www.hiv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html> to convert your fasta alignments to relaxed phylip format. Use a text editor to copy-paste your fasta alignments onto the input box. You may only convert 1 alignment at a time. If it gives you an error message, try reloading the page before submitting your fasta data again. Open the phylip file in a text editor before proceeding. The first line should list the number of taxa and the number of characters. If it has any random text in it, delete that text.

When you have your relaxed phylip format data, go to this page: <http://www.atgc-montpellier.fr/phyml/> to infer your ML phylogenies.

Before you run the analysis, look at the different parameters that you can set. If you are unsure of the meaning of one, try to determine what it means and what the choice are before moving on. You can go ahead and use the default settings for now

Results are sent via email. If you choose your Emory email address, you may need to log into the spam filtering system to get the results.

Each phyml run produces 3 output files. Use the stats file and previous observations to complete the following table. The ML tree length is given under “Tree size”.

|  |  |  |  |
| --- | --- | --- | --- |
| Gap:extension penalty | Length of alignment | Tree length (parsimony) | Tree length (ML) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Use the Figtree application to open the tree files and visualize each of the resulting phylogenies. You may have to change the file extension from .txt to .tree to do this. To facilitate visual comparison among trees, root the trees at the branch that unites the two *Multiclavula* outgroups

5. Why do phylogenies resulting from alignments using different gap opening and extension penalties differ? How does inferred change in the phylogeny relate to the penalties? Explain. What would you recommend if a colleague asked you what to do given the different results from different parameters? Assume collecting data from another gene is not an option.