2004 Short Course in Fungal Molecular Systematics
Hibbett lab, Biology Department, Clark University

Analysis of DNA Sequences for Molecular Ecology and Systematics

This handout describes several bioinformatics applications used in fungal molecular ecology and systematics, including BLAST searches, DNA sequence alignment, and phylogenetic analysis.

You are being provided with two different sets of sequences of nuclear ribosomal RNA genes:

1. Individual sequences of the internal transcribed spacers (ITS) from the mushrooms ("ITS mycorrhizae sequences.doc"), monotrope roots, and other mycorrhizae that you collected. These will be used in BLAST searches to identify these materials in the Friday morning session (see **BLAST search**, below).

2. Sequences of the nuclear large subunit rRNA gene sequences, including (1) a reference set of identified nuclear large subunit rRNA gene sequences ("nuc-lsu rRNA ref data.doc"); and (2) individual sequences from the mushrooms that you collected ("nuc-lsu rRNA mushroom data.doc"). You will combine these data and perform a phylogenetic analysis in the Friday afternoon session (see **CLUSTAL alignment**, and **Phylogenetic analysis**, below).

BLAST search:

BLAST is a family of algorithms used to search for similar sequences among sets of DNA or protein sequences. We will use BLAST to search the GenBank database for sequences that are similar to those obtained from the mushrooms, monotrope roots, and other mycorrhizal roots that you collected.

1. Go to the NCBI homepage (http://www.ncbi.nlm.nih.gov/) and click BLAST on the header.
2. Select nucleotide-nucleotide BLAST
3. Cut and paste the sequence of your collection into the search window and click BLAST!.
4. When the search is complete, you will be presented with a Results window showing a map of the sequences that have been retrieved, aligned to your query. The aligned sequences are color coded according to their similarity to the query. As you scroll down to the more distantly related hits, note that the 5.8S rDNA region, which lies between ITS1 and ITS2, is more conserved than the spacers (why?).
5. Click on "taxonomy report" to see a taxonomic list of the sequences that have been retrieved. The species are listed in order of similarity to your query. One of the species toward the bottom may be a good outgroup candidate—consult with David or Manfred.

CLUSTAL alignment:

Phylogenetic analysis of molecular data begins with the alignment nucleic acid or protein sequences. Within the alignment, each pair of nucleotides or amino acids represents the inferred history at that site in the sequence. There are three possible outcomes for each pair in an alignment: match, mismatch or gaps. Matches represent a pair that is assumed to be unchanged since the sequences diverged, mismatches indicate a substitution has occurred in at least one of the sequences, and a gap (-) indicates that an insertion or deletion occurred in one of the

sequences. The optimal alignment between two (or more sequences) occurs when the numbers of mismatches and gaps are minimized. We will perform alignments using CLUSTALX, which is a multiple alignment program that runs on both PC and Mac platforms and is distributed freely.

1. Add your individual sequences to those in the reference sequences in FASTA format, with extraneous text removed. This can be done in Word (save file in text-only format) or the PAUP* data editor. Note that there is a ten-character limit on sequence names. Example of file containing sequences in FASTA format:

```
>Species1
ACGTGATGCTGACGTAGCTGC
>Species2
ACGTGATGCTGACGTAGCTGC
>Species3
ACGTGATGCTGACGTAGCTGC
```

2. Launch CLUSTALX.
3. Open your file with the sequences in FASTA format (File→Load Sequences). The sequences should appear in the alignment window.
4. Set the output format to NBRF/PIR (Alignment→Output Format→select NBRF/PIR format; deselect CLUSTAL format).
5. Perform alignment (Alignment→Do Complete Alignment). This may take some time if you have a lot of sequences in your file. The output file will be placed in the same folder as your file of sequences in FASTA format, with the same name as the file of sequences, with a ".aln" suffix.

Conversion to nexus format (and optional manual adjustment) in MacClade

Your CLUSTAL output must be converted into the "nexus" file format before it can be analyzed using PAUP*. For this purpose, we will use a data-editing tool called MacClade. MacClade also provides an opportunity to adjust the alignment by hand, which you may want to consider if CLUSTAL has made mistakes (what does that mean?). In these cases, you may be able to improve the alignment manually.

1. Launch MacClade.
2. Find and open your datafile: FILE→OPEN FILE.
3. MacClade has two main windows: the data editor and the tree window; go to the data editor: WINDOWS→DATA EDITOR
4. MacClade should recognize your NBRF-formatted alignment file from CLUSTAL and allow you to import it.
5. Once MacClade opens your file, it will place you into a data editor, which presents your alignment. You may now rename the sequences by adding species names to the GenBank accessions (the 10-character limit no longer applies). An example of a good, informative name for a terminal could be: "Russula_bicolor_DH101".
6. Select molecular format to show nucleotides in color coded form: DISPLAY→DATA MATRIX STYLES→PLAIN MOLECULAR (the Bird's Eye View option may provide you with a helpful perspective on your data as well)
7. Examine your alignment by scrolling in the data matrix.
8. Do you see areas where your alignment appears to be incorrect? If so, you may modify the alignment using the tools in the tool pallette (see the PDF MacClade manual, pp. 244-251).

9. Save your manually adjusted dataset with a new name FILE→SAVE FILE AS. This will be saved in Nexus format, which is used by a phylogenetic analysis using PAUP*.

Phylogenetic analysis.

We will use PAUP* to estimate the evolutionaryu relationships among the sequences in your CLUSTAL aligned dataset. Afterwards, we will compare the trees obtained with molecular data in PAUP* with those that you generated "by eye" based on morphology. PAUP* is a complex program that can run many different kinds of phylogenetic analyses. We will perform parsimony analyses, which seek to find the tree (or trees) that requires (invokes) the fewest possible mutations. In other words, parsimony attempts to find the "simplest" explanation of the dataset. The macintosh version of PAUP* has a menu-driven interface that is easy to use, but the options can be a bit overwhelming. The instructions in this handout lead you through the process of running parsimony analyses, saving results, and printing trees. Please read this entire handout carefully and follow the instructions exactly.

I. Performing a parsimony analysis

1. Launch PAUP*. Find and execute your datafile: FILE→OPEN→EXECUTE. PAUP* should process your datafile, report the number of taxa (sequences), datatype (DNA or protein), and characters, and indicate that it is ready for your next command.
3. Determine how your tree will be rooted. For simplicity, I suggest that you root your trees by mid-point rooting: OPTIONS→ROOTING→MIDPOINT ROOTING.
4. Run your analysis. If you have more than 20 taxa, you should run a heuristic search, which is not guaranteed to find the most parsimonious tree(s), but if you have 20 or fewer taxa, you may use a branch-and-bound search, which is guaranteed to find the most parsimonious tree(s). Before starting make sure that the search criterion is set to parsimony: ANALYSIS→PARSIMONY.

5a. Heuristic search: ANALYSIS→HEURISTIC SEARCH puts you into a series of dialog windows for heuristic search settings:

- GENERAL SEARCH OPTIONS select Keep Best Trees Only (i.e., suboptimal trees will not be retained). SET MAXTREES→select Automatically Increase (all the most parsimonious trees will be retained, as memory permits; with large or "messy" datasets there may be a huge number of equally parsimonious trees, and MAXTREES will have to be limited).
- STARTING TREE OPTIONS→select Get By Stepwise Addition (it is also possible to preselect a starting tree).
- STEPWISE ADDITION OPTIONS→select Random, with 10 replicates (i.e., you will perform 10 heuristic searches, each starting with a tree generated by stepwise taxon addition, with a random taxon addition sequence).
- BRANCH SWAPPING OPTIONS→select TBR, which is the most exhaustive swapping routine (with large datasets it may be necessary to use a less thorough routine to speed the analysis). Leave other settings alone.
- MISCELLANEOUS OPTIONS→no need to change these→SEARCH.
- Follow the progress of your search in the Heuristic Search Status box. When search is complete→CLOSE. Note results in display buffer.
- Go to 6.

5b. Branch-and-bound search: ANALYSIS→BRANCH AND BOUND SEARCH→SEARCH (default settings are OK).

- Follow the progress of your search in the Branch-and-Bound Search status box.
- When search is complete→CLOSE. Note results in display buffer (compare Time Used to heuristic search)
- Go to 6.

6. Check the tree scores: TREES→TREE SCORES→PARSIMONY→select Treelength. This is the number of inferred "steps" (mutations). Remember, shorter trees are preferred under the parsimony cirterion.
7. Save your trees: TREES→SAVE TREES TO FILE→save as Nexus file.
8. Examine your trees: TREES→SHOW TREES. Observe trees on screen.
9. Print your trees (if you have many trees, print only a few): TREES→PRINT TREES. Here you have many options for how your trees will be presented. You can also add a title, which can include analysis details or some other label. In the PREVIEW window you can see what your printed tree will look like, and you will also have the option of saving your tree as a PICT file, which can be manipulated in many different graphics packages. Print (and view) at least one of your trees as a phylogram with branch lengths indicated. Play with the options here.

II. Calculating consensus trees

10. If you obtained multiple equally parsimonious trees, you should construct a consensus tree, which summarizes the information common to all the trees. You should calculate a strict consensus tree, which shows only the branches common to ALL the equally parsimonious trees, and a majority-rule consensus tree, which shows the branches found in the MAJORITY of the equally parsimonious trees, with an indication of their frequency: TREES→COMPUTE CONSENSUS→select strict and majority rule/50%, select "include other compatible groupings".
11. View the consensus trees on the screen, and print them in the same manner as the individual most parsimonious trees: TREES→PRINT CONSENSUS TREES. You can also save the consensus tree graphic, as above.

III. Bootstrap analysis.

12. You will use a bootstrap analysis to assess confidence in your results on a node-by-node basis. You will run 100 bootstrap replicates, i.e., 100 searches. So, you will need to use analytical settings that allow the program to run relatively quickly. If your branch-and-bound search took five minutes, then 100 bootstrap replciates with branch-and-bound will take about eight hours! You might need to run a heuristic search with a small number of replicates in each bootstrap replicate. Discuss analytical settings with David or Wang.
13. To start the bootstrap: ANALYSIS→BOOTSTRAP/JACKNIFE, select "bootstrap" and either "branch and bound" or "full heuristic", then click "continue".
14. You will now be placed into the dialog boxes for either a branch-and-bound search or a heuristic search, as before. If you are running a heuristic search, select one random taxon addition sequences per replicate. Other settings should be as above. When you are done with the settings, click "search".
15. View the bootstrap majority-rule consensus trees on the screen, print it, and save the tree graphic as above: TREES→PRINT BOOTSTRAP CONSENSUS TREES.
16. Save your display buffer: EDIT→EDIT DSIPLAY BUFFER→FILE→SAVE. The display buffer retains a record of all your analysis settings. It can beuseful when you are writing up your analyses. It is a text file that can be opened in a word processor.