Phylogenetic Analysis of Molecular Data (Bot 563)

Computer Lab 1: Generating a data matrix

Additional documents: Common file format examples; MacClade tool helper.

Part 1. Downloading data from GenBank

1. Go to http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi
2. Search for: Adansonia [ORGANISM] AND internal transcribed spacer. Note the capitalized words! You should find 16 sequences (look for and go to the Nucleotide database).
   (For your own future searchers you might like to use the modifiers TAXONOMY or GENE to restrict searches.)
3. Download the last 15 sequences (we won't use the first one).

   Some tips:
   a. Check the box of each of the sequences 2-16
   b. Change display to: FASTA
   c. Change "send to" to "file".
   d. Downloading should start automatically.
   e. Rename the file as you wish.

   ALWAYS look at the file! For this you can use a text editor such as TextEdit. For your own work, you might want to trim some of the sequence names, so that they are more maneageable.
4. Search for the following sequence to use as an outgroup sequence: Bombax [ORGANISM] AND internal transcribed spacer. Save ONLY the third *Bombax* sequence to another file, with an appropriate name (you know what to do now!).

Part 2. Manipulating sequence data in MacClade

1. Open the *Adansonia* file in MacClade.
2. Add the outgroup sequence: Go to TAXA>Import Sequences> Fasta> Add as new sequences and select the *Bombax* file.
3. Try an automated alignment between *Bombax* and one of the *Adansonia* sequences using the alignment tool.
4. Practice using the various editing tools to modify the alignment (see MacCalde tool doc for help).
5. Save the file.
6. Make a second version of your file (File>Save as), naming it as: yourfilename.interleaved. Now, go to File>Options for saving>Nexus format. Check the interleave box.

Part 3. Opening the data file in PAUP*

1. Go to Applications>Classes. Launch PAUP and then open the file to edit.
2. Look at the file (if it is not visible look under the WINDOW menu). You should recognize this format by now. Note the number of taxa and characters.
3. Open the interleaved version of your file. What does the interleave option do for you?
4. EXCUTE either file (under the FILE menu).
5. To see how many informative characters there are go to DATA > INCLUDE-EXCLUDE CHARACTERS. In the pull-down menu select UNINF and then hit: EXCLUDE. The display will tell you how many characters (the uninformative) were excluded and how many remain (the parsimony informative characters). You can also go to DATA>SHOW CHARACTER STATUS.

Part 4. Specifying data partitions in PAUP*

1. Bring the data matrix to the front.
2. Insert a PAUP block at the end of the file and then type in character sets. For this data set, ITS1 corresponds to bases 1-350; 5.8S corresponds to bases 351-514, and ITS2 to 515-778. The PAUP block looks like this:

```
Begin PAUP;
Charset ITS1 = 1-350;
Charset 58s = 351-514;
Charset ITS2 = 515-778;
End;
```

3. Save the file and execute it. Now go to INCLUDE-EXCLUDE CHARATERS. The three partitions should be listed in the pull-down menu.
4. Other useful examples:

   a. Specifying introns:

```
Begin PAUP;
Charset introns = 1-100 200-310 700-825;
Charset exons = 101-199 311-699 826-1203;
End;
```

   b. Specifying codon positions:

```
Begin PAUP;
Charset firstpos = 1-5153\3;
Charset secondpos = 2-5154\3;
Charset thirdpos = 3-5155\3;
End;
```

Part 5. Getting data matrices from TreeBase

1. Open www.treebase.org
2. Go to SEARCH and select BROWSE WITH FRAMES. Enter Felis catus and then SUBMIT. Select SEARCH; the studies available for this taxon will show up.
3. You can look at the data for such studies using the "T" link. Select one and download the data using the "matrix" link.
4. Open the data in MacClade. Go to WINDOW>TREE and familiarize yourself with some tree editing tools (see MacCalde tool helper document).

Part 6. Some online multiple sequence alignment tools

Go to one or more of the following sites, and load the Adansonia-Bombax sequences. You can either upload the FASTA file or cut-and-paste it into a window (the first is recommended). Default settings are fine for now.

- ClustalW: http://Artedi.ebc.uu.se/programs/clustalw.html
- Muscle: http://www.ebi.ac.uk/Tools/muscle/index.html
- T-Coffee http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffee_cgi/index.cgi?stage1=1&daction=TCOFFEE::Regular

Part 7. Adding extra partitions to your dataset (Optional)

1. Do the searches for Adansonia and Bombax sequences as you did for ITS but now for rpl16.
2. Put all the 16 sequences (Adansonia and Bombax) in one file, and align this matrix.
3. Open both aligned data sets in Paup. Copy and paste the rpl16 data into your ITS matrix, just below the first sequences you already had. **Careful, taxa MUST be in the exact same order!**
4. Also, you need to make the following changes in the Paup "Begin data" block:
   a. change the `nchar=XX` to the correct number (ITS+rpl16).
   b. add `interleave=yes`; this tells the program that your data are split in different "blocks".