Exercise 5: Prepare data sets and partitions in a nexus file

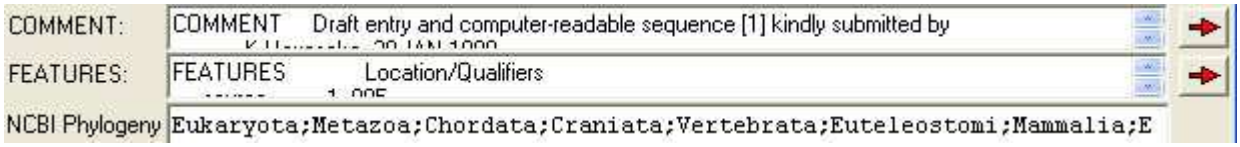
(BIO332 Phylogeny 2007. E.Willassen) PDF version

In the Introducing PAUP* and in Exercise 1 we saw that nexus files may contain very different kinds of information. The basic element is certainly the matrix, and a matrix is simply what it takes do a phylogeny reconstruction. However, the matrix may contain data from different genes, some regions that are noisy and hard to align, sites with diverging evolutionary rates etc. That is why we may want to separate the data in sets and partitions.

A plain nexus data file called Exercise5.nex is intended as the starting point for this exercise. By studying genbank information for Lemur_catta in the file Exercise5.gb you should be able to map out sets and partitions for the data. Consult the chapter Introducing PAUP* or the PAUP* manual to see the PAUP* notation for character sets and character partitions.

Consult Genbank information

- Start BIOEDIT
- Select File / Open Exercise5.gb
- use left mouse button to double clickLemur_catta. The pop-up window displays genbank information.
- Mouse-click the red arrow to the right of the FEATURES field.



The Genbank information about the positions of genes (and other features) is now displayed in detail. You may now use this information to partition your data set. DO NOT FORGET that the nucleotide positions in the reference sequence (Lemur_catta) and the site positions in the alignment may not correspond exactly if the alignment has gaps. Fortunately, the mouse pointer in BIOEDIT gives you the position of a sequence residue, so you should be able to work out the spans of genes in the alignment.

Writing character sets in the nexus file

- Leave the file Exercise.gb open in BIOEDIT
- Use Windows Note Pad to open the file Exercise5.nex

Notice the format of this simple non-interleaved nexus data file. Comments may be written in [brackets]. Brackets may also be used to make PAUP* commands (and other information) passive during the execution of the file.

The Dimensions line gives the number of taxa and the numbers of character sites. The Format line defines the range of character state symbols in the matrix. You may experience sensitivity of some applications to the order of information in this line, so just make sure that the data type info comes before gap symbols etc. on this line. Notice the semicolon after the last sequence and the end; after the matrix block.

You will now write a SETS block with Character sets and Character partitions under the matrix block.

Testing the SETS block

- Start PAUP*
- Select File / Open and find the folder with your modified Exercise5.nex
- Notice the choice between two open modes
- Select Execute
- Select File name (by mouse-click) and
- Hit the Open/Execute button

Messages in the run buffer window will now tell you if your nexus file loads with no problems. If not, you may correct errors by reopening the file in Edit mode. When finished:

- select File / Save and
- select File / Execute

A solution

Here is an example of how your sets block could look like when you have finished specifying the contents of the columns in the data matrix:

```
begin sets;
    charset div = 1 458;
    charset ND4 = 2-457; [! NB 458 + extra AA make stop codon]
    charset tRNA1 = 459-528; [tRNA_His]
    charset tRNA2 = 529-588; [tRNA_Ser]
    charset tRNA3 = 589-659; [tRNA_Leu]
    charset ND5 = 660-898;

    charset 1st = 2-457\3 660-898\3;
    charset 2nd = 3-457\3 661-898\3;
    charset 3rd = 4-457\3 662-898\3;

    charset coding = 2-457 660-898;
    charset trna = 459-659;

end;

begin paup;
    exclude div;
    charpartition genegroup = 1:coding, 2:trna;
end;
```

The last part of this screen print indicates how we can use character set names to include or exclude parts of

the data in the analysis and also to define character partitions to PAUP*. You will later see that the syntax of these types of commands is very similar in MrBayes language, but that these programs differ slightly in the way character partitions are defined.

Having completed this exercise you have made a considerable step towards understanding nexus files. You will soon be ready to use PAUP*, MrBayes, Mequite, and several other program packages for phylogentics that uses nexus files.