Assignment 1 / Biometry and bioinformatics II / 2014

The goal: To familiarize with different phylogeny methods

By using MEGA5-software (http://www.megasoftware.net/, installed in class C128)

- UPGMA
- Neighbor-joining
- Maximum parsimony
- Maximum likelihood

- Note that MEGA5 has a complete tutorial. Note also *A walk through MEGA,* Step-by-step instructions to learn how to use MEGA. Read *MEGA5 original paper* (in course webpape).

By using MrBayes 3.1.2 (installed in C128; for convenience, we do not use the most recent version) and it´s manual (in course webpage). Read *Original MrBayes paper* (in course webpage)

- Bayesian phylogeny inference

Datafile data_2.txt contains sequences from a gene related to so called *virulence* of the bacterium *Streptococcus pneumoniae,* which is a bad human pathogen, causing pneumonia etc., but lives also as a harmless "commensal" in our mouths and noses. The OTUs named by numbers or numbers+letters are all *Streptococcus pneumoniae* (different serotypes). Other *Streptococcus* species, *mitis, oralis, agalactiae, thermophilus, salivarius*, *suis, gordoniae, iniae* are included in the sequence set as a reference.

- Write a report about the evolutionary history/histories of *Streptococcus pneumoniae* serotypes by using phylogenetic inference. Use the methods included in MEGA5 software (UPGMA, neighbor-joining, parsimony, ML) and MrBayes.

- Tree confidence/credibility and nucleotide substitution model choice should be considered. At least some method should be performed by using simple models vs. a complex model. Choose neighbor-joining method for this experiment. Simple models = p-distance and Jukes-Cantor model. Complex model = the one you get from *model choice.* How does it matter (in this case), what is the model?

- Compare the results you get from five phylogeny inference methods. For example, do they result in different topologies?

- In addition to writing about clustering structure differences (topologies) you get (or: maybe get), include in your report explanations for these issues: Why is it that trees from different methods look different, for example all branches ending at the same vertical point / not ending like this. Why is it that some tree(s) are very "regular" so that the branch lengths appear very systematic. In addition, explain also a description about the ways (differences/similarities) how the methods use data, i.e. how do they "pick up" information from the data. What is the meaning of the horizontal axis (if there is an axis). What is its (= the axis) meaning in terms of (evolutionary) time?

  Include phylogenetic trees in your report: copy-paste them as pictures in your text.

Practical advise for working with MrBayes

The data must be converted to nexus-format and the the file (which you submit the program MrBayes) must be located in MrBayes home-folder!

- Course webpage has a file-converter link. You can also convert to nexus by Clustal. A file-converter does not (in all cases) produce an exactly "correct" form. When using the tool above, you get the nexus-file which you must edit a bit: you must add the text which is red here (this example has 988 seqs, 1737 nucleotides):

```
#NEXUS
begin data;
        dimensions ntax=988 nchar=1737;
        format datatype=dna interleave=no gap=-;
matrix
```

- We are using in class C128 an old version of MrBayes, 3.1.2. The reason is that the most recent version(s) might turn out to lead various practical problems.

- Use the FigTree –program (in computer class C128 machines) for visualization. The manual suggests TreeView.

- Use common sense in resolving, for example, this kind of questions: "should I continue running the program because…..???". You can very well report that "although better (?) results might have resulted from continuing, for convenience I stopped and collected the results at….." or something like this.

- It might be helpful/interesting to see what kind of problems MrBayes-users report: https://lists.sourceforge.net/lists/listinfo/mrbayes-users