This is an updated version and includes notes about 26.9 session when we started.

Assignment set 1 / Biometry and bioinformatics II / 2013

- Reports to be submitted to course Moodle Thursday 10.10.
- After that you will get personal comments from teacher about issues that might be "wrong" in your answers.
- Submissions after 10.10 are allowed, but very probably cannot lead to personal responses before the exam (18.10) which includes course assignment contents.
- In exam you have all course material available, including scientific papers which you should read, and which contain information for assignments. The spirit of this course is learning by doing - assignments, exercises, and self-learning by reading scientific papers.
- During the last course session, Wednesday 16.10, we discuss about all course assignments so that one student´s learning about assignments is not restricted only to his/her own solutions.
- Recommendation is that you do not work alone, work in groups. However, each student should submit *an own report which cannot be a copy of an other student´s report.*

The goal of assignment set 1 is to familiarize with different phylogeny methods by using different kind of data and research problems.

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

What to do first: Read the two general review-type scientific publications from course webpage: *Short tutorial article* and *Phylogenetics – principles and practice*

- Lectures on phylogenetic methods: during the lecture 26.9 we started, and we continue 1.10 when we also start with the topic *modeling nucleotide and amino acid substitutions.* One part of this topic, *model choice for phylogeny inference process,* is planned to be covered during 1.10 lecture.

- During the lecture 26.9 we had tutorial session with data_4.txt (see below. (There is also data_1.txt in course webpage for self-learning. This is the same initial data we have used in the course BB_I)). We constructed maximum parsimony and neighbor-joining phylogenies with and without bootstrapping.

- During 1.10 computer session: MrBayes *Tutorial, A simple analysis* (see page 8 in the manual). Last page here (page 7) includes some practical advise for starting with MrBayes.

Assignment set 1.1

1. 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.* 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. Preferably not as separate documents. Copy-paste them as pictures in your text.

Assignment set 1.2

2. Datafile data_3.txt is a set of complete mitochondrial sequences from a set of mammals (this is the same file as in course BB_I extra assignment 4). Datafile data_4.txt is one gene, cytochrome B, cut from this data.

   - Write an interpretation about phylogenetic relationships of the species
     a) on the basis of the whole mt-genome data
     b) on the basis of cytochrome B

     For a) do only neighbor-joining phylogeny; the data is so big that other analyses will be really slow.

     For b) do the same analyses as you did for question 1 (the bacteria data). In addition to writing an interpretation of the results, write about comparison between one-gene-data and whole-mt-genome-data (on the basis of neighbor-joining).

   - Inspect visually the whole mt-genome alignment and write about differences you notice between protein coding genes and D-loop: how useful are these two types of sequences in providing information for species comparisons by using phylogenetics (for which sequence alignment, which you are now asked to inspect, is the first step). Next page shows the coordinates for different genes and other pieces in the mt-genome. A verbal story is here enough.

   - NOTE: In MEGA, with cytB, you have to define the code: it is NOT "standard", but "vertebrate mitochondrial". When you work with the whole mt-genome data your answer to question "protein coding data" is NO because the data (although it contains protein coding genes, see the table next page) is not a clear 123 123 123…. (=the codons for amino acids). For example the first gene in the data is a transfer-RNA etc.

	Nucleotides in AB499817, the first sequence in datafile	Nucleotides taking into account gaps in aligned file
tRNA-Phe	1-69	1-80
12S ribosomal RNA	70-1023	81-1090
tRNA-Val	1024-1090	1091-1161
16S ribosomal RNA	1091-2670	1162-2840
tRNA-Leu	2671-2745	2841-2917
gene ND1	2748-3704	2919-3882
tRNA-Ile	3704-3722	3882-3901
tRNA-Gln	3769-3843	3948-4025
tRNA-Met	3845-3914	4028-4097
gene ND2	3915-4958	4098-5143
tRNA-Trp	4957-5024	5142-5215
tRNA-Ala	5038-5106	5232-5301
tRNA-Asn	5108-5179	5310-5386
tRNA-Cys	5213-5280	5419-5495
tRNA-Tyr	5281-5348	5496-5572
gene COI	5350-6894	5574-7140
tRNA-Ser	6892-6962	7132-7216
tRNA-Asp	6967-7034	7222-7292
gene COII	7035-7718	7293-7977
tRNA-Lys	7736-7802	7995-8066
gene ATPase subunit 8	7804-8007	8068-8276
gene ATPase subunit 6	7965-8645	8234-8914
gene COIII	8645-9428	8914-9697
tRNA-Gly	9429-9496	9698-9770
gene ND3	9497-9843	9771-10117
tRNA-Arg	9843-9911	10117-10187
gene ND4L	9914-10210	10191-10487
gene ND4	10204-11581	10481-11858
tRNA-His	11580-11650	11857-11930
tRNA-Ser	11651-11710	11931-11995
tRNA-Leu	11711-11780	11996-12067
gene ND5	11781-13601	12068-13895
gene ND6	13585-14112	13879-14406
tRNA-Gln	14111-14181	14405-14476
gene cytB	14186-15325	14482-15625
tRNA-Thr	15326-15395	15626-15703
tRNA-Pro	15395-15460	15703-15772
D-loop	15461-16741	15773-18424

Assignment set 1.3

This set of questions is based on data_4.txt

Take a subsample of four animals from the data and by using this data, answer the following questions.

- A) How many conserved sites, how many variable sites, how many parsimony informative sites?

- B) What is relationship between nucleotide differences (among the four animals you are now comparing) at sites which do alter amino acid and those that do not alter amino acid (i.e. is there a difference between observed mutations leading to amino acid differences and those mutations that do not lead to amino acid difference).

- C) Pick up one variable site which is not parsimony informative and explain why it is not parsimony informative.

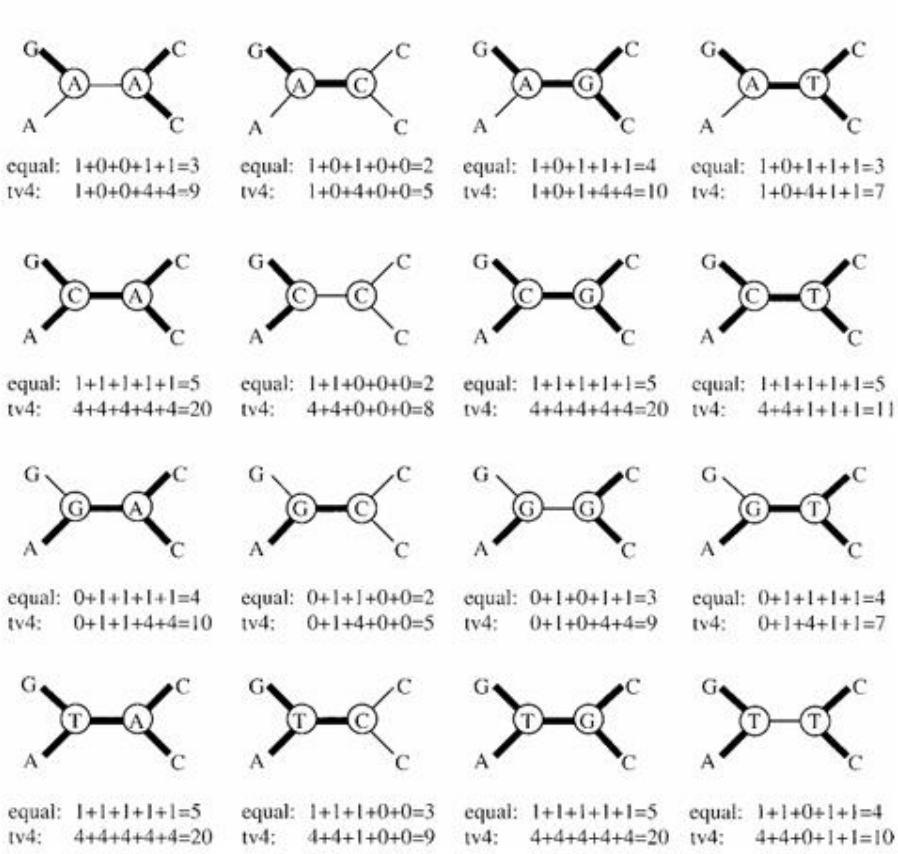
- D) Pick up three parsimony informative sites and construct the unrooted maximum parsimony trees on the basis of these sites. Then combine the information from these separate sites: what is the maximum parsimony tree?

    - N.B. You should use a set of animals which are not *too similar* in order to get a reasonable set for answering to these questions. How to do that? Inspect your clustering! If your first trial set does not include any parsimony informative sites, then your set is too simple and you should take another sample!

- E) Below (next page) is an example using a cost scheme: transversions are weighted 4x. Re-consider your D) (= what you did above). Would you get the same result or a different result by using this kind of a cost scheme?

- F) Calculate the distance matrix by using p-distance and by Jukes-Cantor. You can do this by using MEGA5-facilities. However, in addition to this distance matrix, show at least one calculation by hand (= show, for one species pair, how is their p-distance and Jukes-Cantor –distance calculated).

- G) By using the distance matrix you made in F) (either p-distance or JC, does not matter), construct UPGMA by hand, NOT by MEGA5.



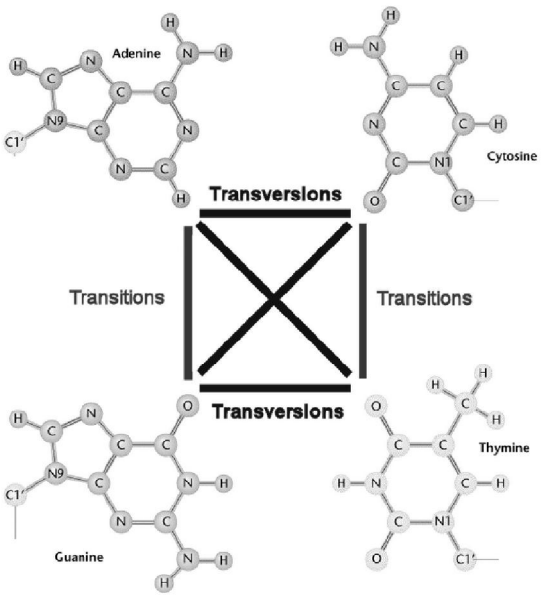
This picture is from *Lemey et al., The phylogenetic handbook, 2009,* www.cambridge.org/9780521877107

Two cost schemes:

- Equal vs. transversions 4x weighted. See next page for clarification on transitions and transversions.

- With equal costs, the minimum length in two steps and this length is achievable in three different ways: internal nodes assignment A-C, C-C and G-C. If a similar analysis for the other two possible trees, ((W,X),(Y,Z)) and ((W,Z),(Y,X)) is conducted, they are also found to have lengths of two steps. *Thus this character (state) does not discriminate among three tree topologies and is parsimony-uninformative under this cost scheme.*

- With 4:1 transversion:transition weighting the minimum length is five steps, achieved by two reconstructions: internal node assignments A-C and G-C. Similar evaluation of the other two trees finds a minimum of eight steps on both trees. This means that two transversions are required rather than one transition plus one transversion. *The character thus becomes informative as some trees have lower lengths than others.*



DNA substitution mutations are of two types:

Transitions are interchanges of two-ring purines (A ↔G) or of one-ring pyrimidines (C ↔T): they therefore involve bases of similar shape.

Transversions are interchanges of purine for pyrimidine bases, which therefore involve exchange of one-ring and two-ring structures.

It is well known that transitions are considerably more common than transversions.

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