ASSIGNMENT 1

Dataset 1
■ Learning basic operations in collecting data from sequence databases, aligning the data so that the result is a resonable set forming the material for phylogeny analyses.

■ Phylogeny analyses by MEGA-software (neighbor-joining and UPGMA trees and parsimony cladogram), including bootstrapping as the method for confidence and bayesian phylogeny inference by mrBayes-software including MCMCMC.

■ Time schedule:

■ Wed 2. Feb from 12 to 20 there will be extra help for you in C128. This concerns data collection and alignments for dataset 1 and 2. The week after (maybe in Wed 9. Feb, need to be agreed) you can again come to meet extra help and check your alignments.
■ Alignments and MEGA-analyses should be ready during weeks 3 and 4, so that you can proceed to bayesian analyses with dataset 1. Planned mrBayes-demo will be during the lecture Tue 8. Feb. **Work should be ready in 22. Feb**.

■ Recommendation is that you don´t work alone, instead form groups.

■ The initial dataset 1 in course webpage is a textfile in fasta-format from the gene brain-derived neurotrophic factor (BDNF) from 12 vertebrate animals (Vertebrates = the animal group which has bones, invertebrates are animals without skeleton, i.e. insects and crustaceans)

■ There is one bird (Gallus, chicken) and 11 mammals (two primates: human and chimpanzee, three Artodactyla: pig, cow, horse, two rodents: mouse and rat, the rest being Carnivora). Birds (Aves) and mammals are two "sister-groups" in animal kingdom.

■ **Expand the dataset by collecting at least 15 additional animals.**
■ Some suggestions which contribute for making the data a bit more presentable throughout vertebrates and also highlight differences between animal "groups".
  - ■ Take more birds.
  - ■ Take also frogs (Amphibia)
  - ■ Take more primates (i.e. relatives of human and chimp)
  - ■ Take also the "almost-mammal-animals" = those that do not carry their baby inside, but outside their body (like kangaroo), i.e. Marsupiala.
  - ■ If you want to make a challenging alignment work, take fishes….. but then you need to do lots of alignment editing…. (this is not a general rule or instruction, this is based on experience….)

■ Go to NCBI, http://www.ncbi.nlm.nih.gov/

■ **Search "nucleotide"** database because you are working with DNA-sequences (more of the like you already have...)

■ **You do "BLASTing".** If you want to learn more about these algorithms (topics in other MBI-courses, not in this course), read here, everything is explained, and look at the papers in course webpage.

■ Make sure that you know what is an accession number and fasta-format of a sequence.

■ You have initial knowledge about the BDNF-sequences.
- ■ You can proceed by copy-pasting one sequence into BLAST-window (see next page), **or**
- ■ you can write to "search"-window (previous page) BDNF, you´ll get a long list of results, try by restricting the search BDNF primates, or BDNF aves etc.

Basic BLAST

Choose a BLAST program to run.

nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

■ When you proceed by using a sequence that you already have in the initial file, and you have clicked "BLAST" from the previous page, you are now here and you continue by "nucleotide blast" to the next page……

blastn | blastp | blastx
tblastn | tblastx

BLASTN programs search nucleotide databases using a nucleotid

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) ⓠ   Clear   Query subrange ⓠ

From [   ]

Copy-paste here one sequence. (If you want more birds, type here the the chicken sequence.)

To [   ]

Or, upload file   [   ] Browse_ ⓠ

Job Title   [   ]

Enter a descriptive title for your BLAST search ⓠ

☐ Align two or more sequences ⓠ

Choose Search Set

Database   ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.): ←

◆ Nucleotide collection (nr/nt)   ▾ ⓠ

Organism
Optional   [   ] ← ☐ Exclude [+]

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. ⓠ

Exclude
Optional   ☐ Models (XM/XP) ☐ Uncultured/environmental sample sequences

Entrez Query
Optional   [   ]

Enter an Entrez query to limit search ⓠ

■ ….now you are here (many kind of options….)

■ When you enter this page, the default is that you are interested in "Human genomic + transcript" but that is not true: remember to click "others"

■ When you want to get results from a restricted source, you type here for example primates or aves or amphibia or marsupiala, etc.

Program Selection

Optimize for
- ○ Highly similar sequences (megablast)
- ○ More dissimilar sequences (discontiguous megablast)
- ◉ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

BLAST Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
☐ Show results in a new window

▼ Algorithm parameters   Note: Parameter values that differ from the default are highlighted in yellow and m

General Parameters

Max target sequences   100
Select the maximum number of aligned sequences to display

Short queries   ☑ Automatically adjust parameters for short input sequences

Expect threshold   10

Word size   ♦ 7

Max matches in a query range   0

Scoring Parameters

Match/Mismatch Scores   2,-3

Gap Costs   Existence: 5 Extension: 2

■ This is the bottom half of the page (see the previous page here)

■ Choose this algorithm! Difficult to explain, but compare the results from a given BLASTing experiment by the three algorithms, you´ll get some practical experience and understanding "by doing".

■ This is (probably) not needed for this course work: for many (real) problems this default (100) is too low.

■ Collect the sequencies so that they are of comparable lengths already before alignments (which is then fine-tuning of gaps).

■ A result might be like this:

```
Query   1    ATGACCATCCTTTTCCTTACTATGGTTATTTCATACTTTGGTTGCATGAAGGCTGCCCCC   60
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct   247  ATGACCATCCTTTTCCTTACTATGGTTATTTCATACTTTGGTTGCATGAAGGCTGCCCCC   306
```

(only the first and last row of a result query are shown).

…….

```
Query   721  TTGACCATTAAAAGGGGAAGATAG   744
             ||||||||||||||||||||||||
Sbjct   967  TTGACCATTAAAAGGGGAAGATAG   990
```

■ "Query" is your sequence and you are interested only on this part.

■ "Sbjct", a given sequence item (with a given accession number, its identifier from which you get it), has the relevant part beginning from **its nucleotide 247 and spanning to its 990. Take only this part** (see next page).

■ You can delete the extra parts (here the 246 first nucleotides, and something after 990) after aligning you whole set. **HOWEVER, it is advisable to do this kind operations before alignments => less "thinking" for the alignment program.**

Nucleotide
Alphabet of Life

Search: Nucleotide ▼   Limits Advanced search Help

[   ] Search Clear

Display Settings: ☑ GenBank   Send: ☑

Homo sapiens brain-derived neurotrophic factor (BDNF), transcript variant 12, mRNA

NCBI Reference Sequence: NM_001143812.1

FASTA   Graphics

Change region shown ▼

Customize view ▼

Analyze this sequence ▲

■ You have now clicked from one result (from its accession number) and have this page including one sequence for your data collection. You need it in FASTA- format and get that from

here, but you don´t want to take the the whole sequence behind this accession number and thus you use this and type the region you want (for example 247-990).
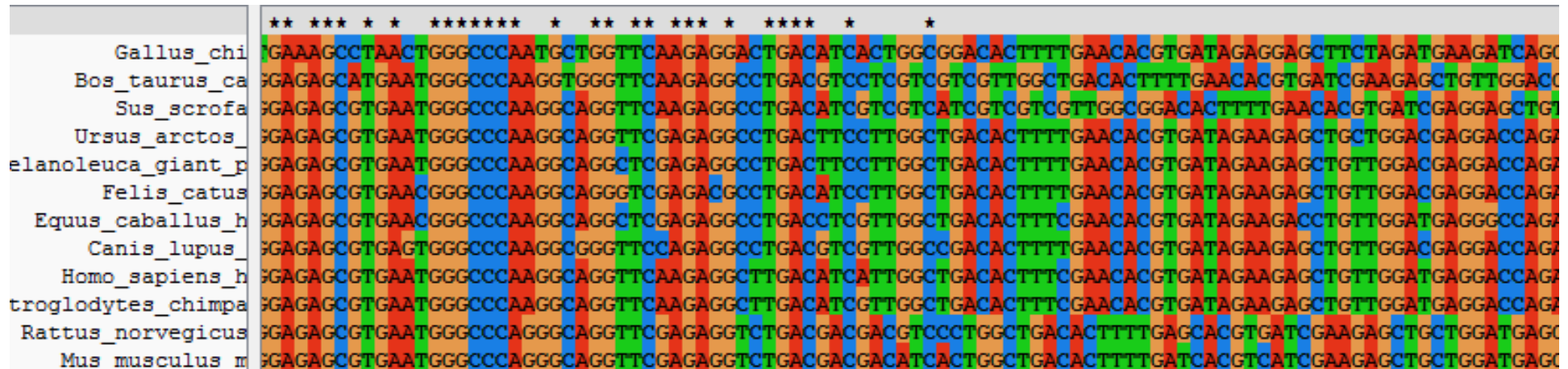
■ The default in computer class C128 is that you use the installed programs ClustalX for alignments and Genedoc for editing the alignments

■ Course webpage has an example of an aligned FASTA-file (you must do that for the expanded dataset) and a MEGA-file (= aligned FASTA with some changes).

■ Your FASTA-file here
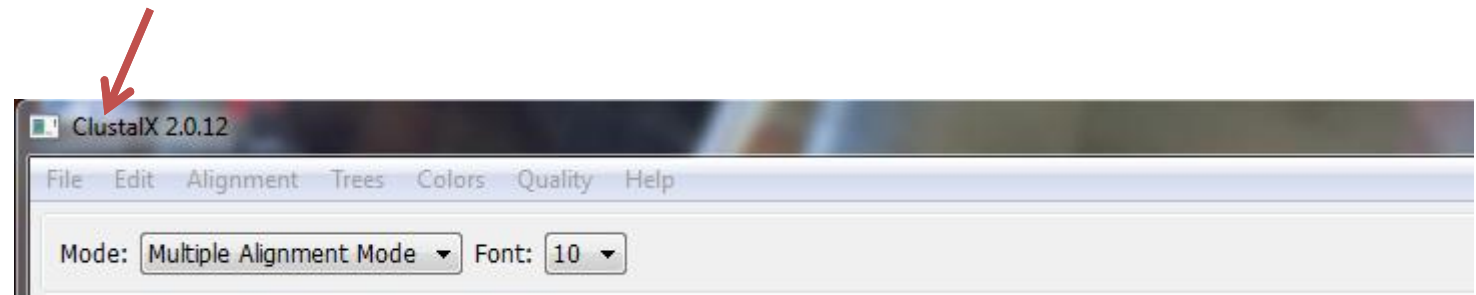
■ Your data in Clustal, before alignment…

■ … and after alignment

■ Before clicking "do complete alignment" (from Alignment), do the following:

■ Alignment -> Alignment parameters (depends on the case, set gaps..)

■ Alignment -> Output format options:
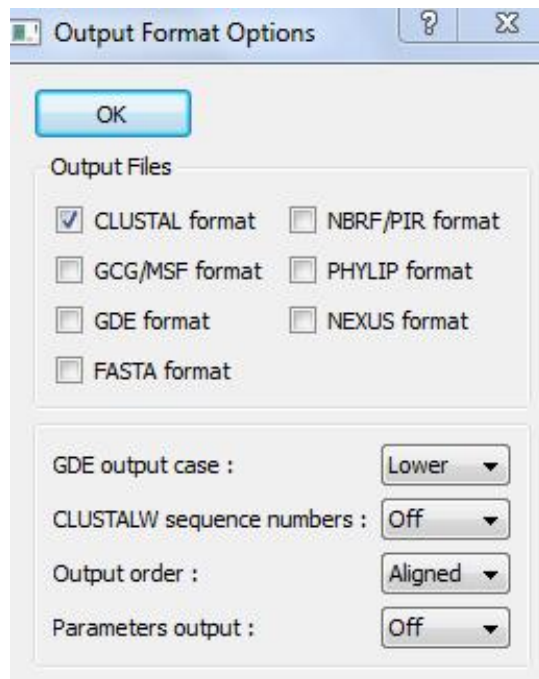
■ you need MSF-format for Genedoc (the program by which you edit = correct the alignment made by Clustal

■ you need FASTA-format = aligned FASTA -> MEGA-format

■ later during the course you also need NEXUS-format for mrBayes.

■ An alignment given by a program is always just a suggestion and must be inspected manually = by own eyes and brains. Depending on the case, corrections are needed / not needed.

■ When you get the alignment, open it by Genedoc (the MSF-file), and start thinking whether everything is okay, taking into account that the sequences should be from a protein coding gene (=> for example, only 3 nucleotide (or multiplies of 3) gaps (deletions/insterions) are reasonable.