ASSIGNMENT 1 / Biometry and bioinformatics I / 2014

- 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 and other statistical analyses.
- Elementary data clustering methods: UPGMA and neighbor joining

- Time schedule:

  Proceed so that you have some data collection done during the first week. When you come to the next session, 16.9, you should have at least something done.

- Recommendation is that you don´t work alone, instead form groups of 2-3 students for the data collection steps.

- Submit your aligned datafile to course Moodle. Your data will be checked and commented. Each member of a student group should do the submission so that all can read the comments. Please, include the information about group memebers!

ASSIGNMENT 1 - INSTRUCTIONS - DATASET 1

■ 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)

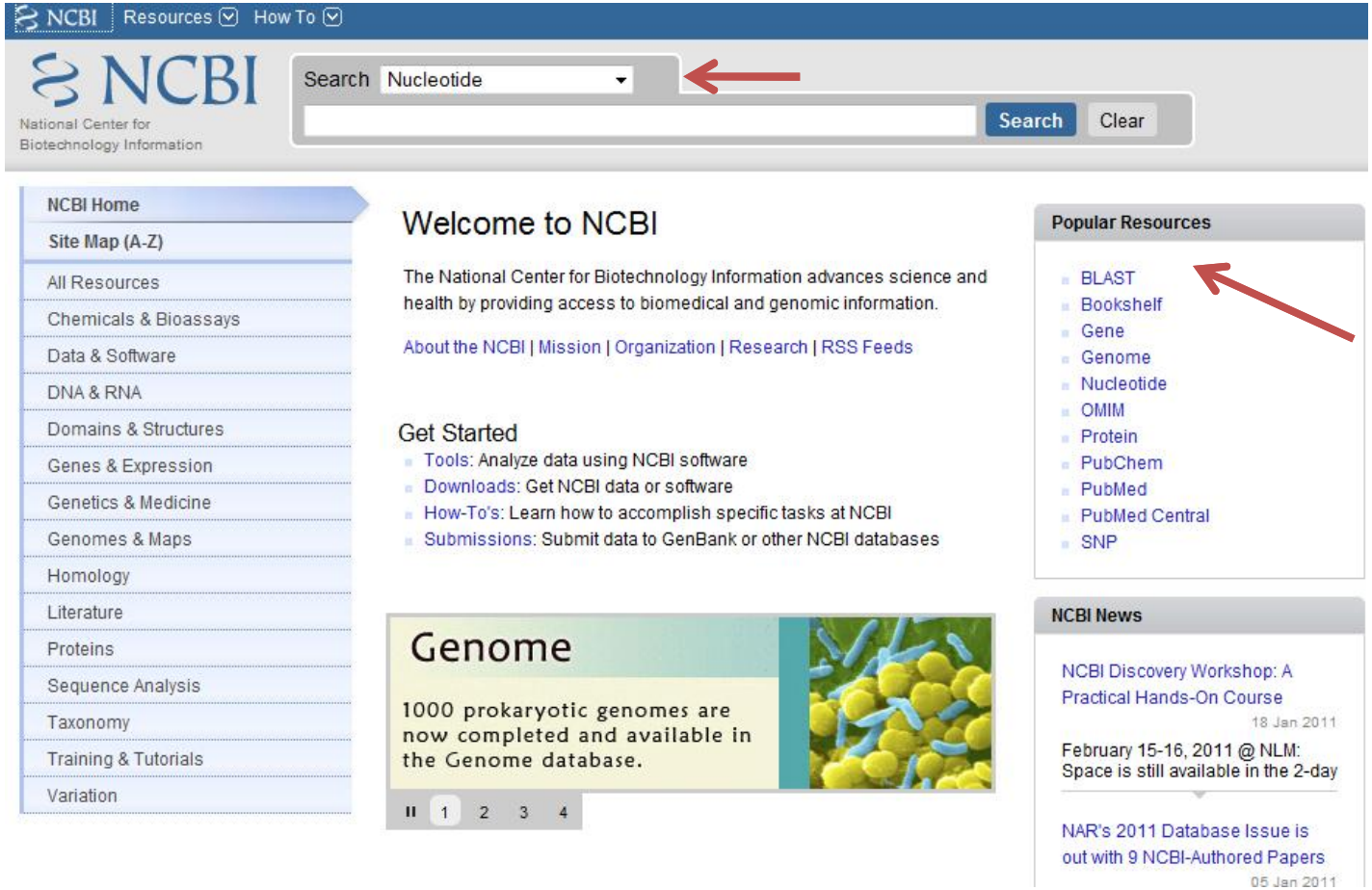
■ There is one bird (Gallus, chicken) and 11 mammals (two primates: human and chimpanzee, three Artodactyla: pig, cow, horse, two rodents: mouse and rat, the rest being Carnivora). Birds (Aves) and mammals are two “sister-groups” in animal kingdom.

■ **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 (see, however, page 11)…. (this is not a general rule or instruction, this is based on experience….)

ASSIGNMENT 1 - INSTRUCTIONS - WHAT YOU NEED FOR COLLECTING DATA

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



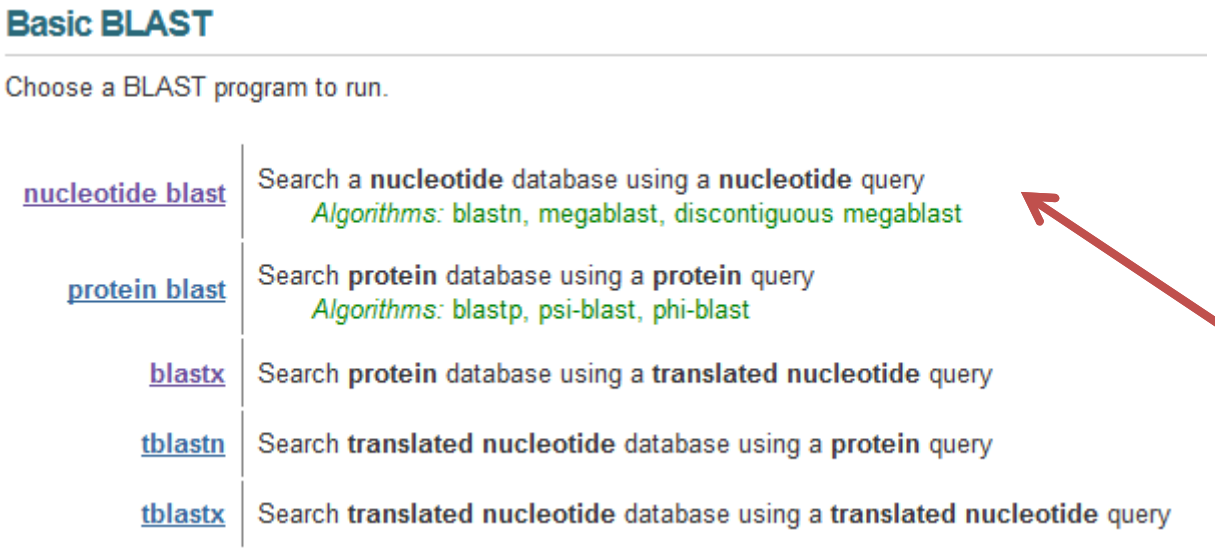
- **Search "nucleotide"** database because you are working with DNA-sequences (more of the like you already have...)

- **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.

ASSIGNMENT 1 - INSTRUCTIONS - STARTING BLAST

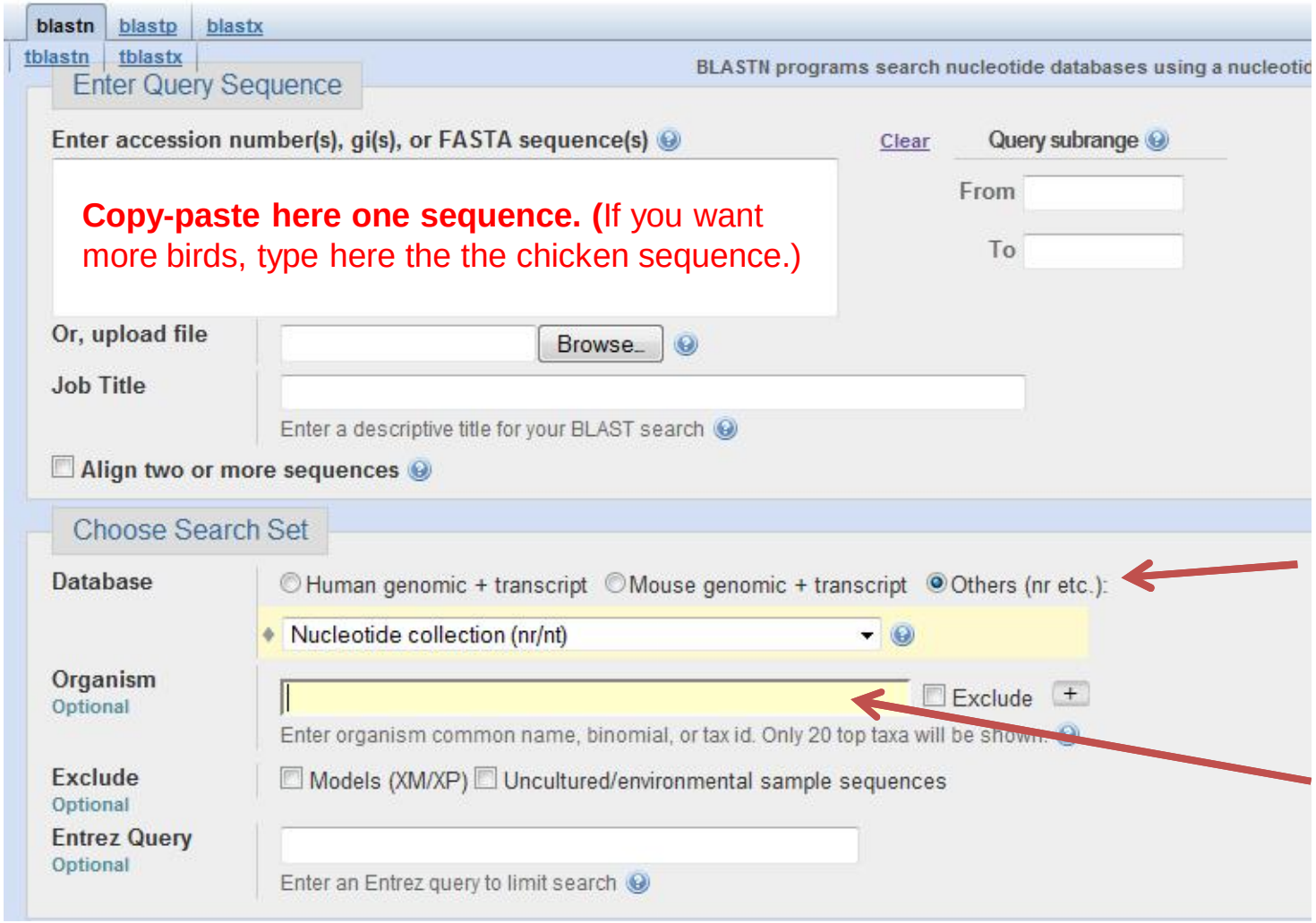
- Make sure that you know what is an accession number and fasta-format of a sequence.

- 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.



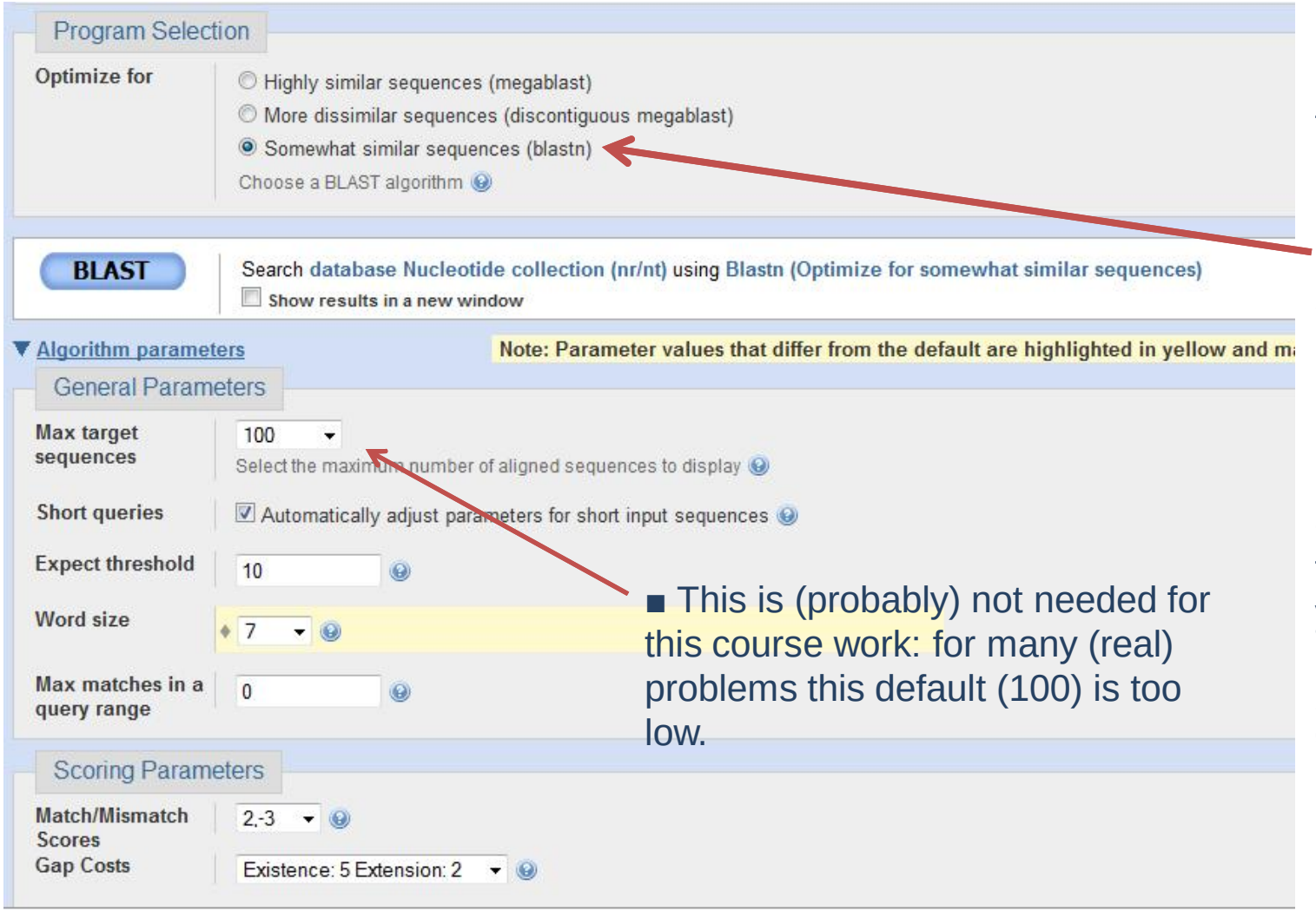
- When you proceed by using a sequence that you already have in the initial file, and you have clicked “BLAST” from the previous page, you are now here and you continue by “nucleotide blast” to the next page……

ASSIGNMENT 1 - INSTRUCTIONS – STARTING BLAST



- ….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.

ASSIGNMENT 1 - INSTRUCTIONS - DATASET 1



■ 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.

ASSIGNMENT 1 - INSTRUCTIONS - some remarks on data collection

■ 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).

…….

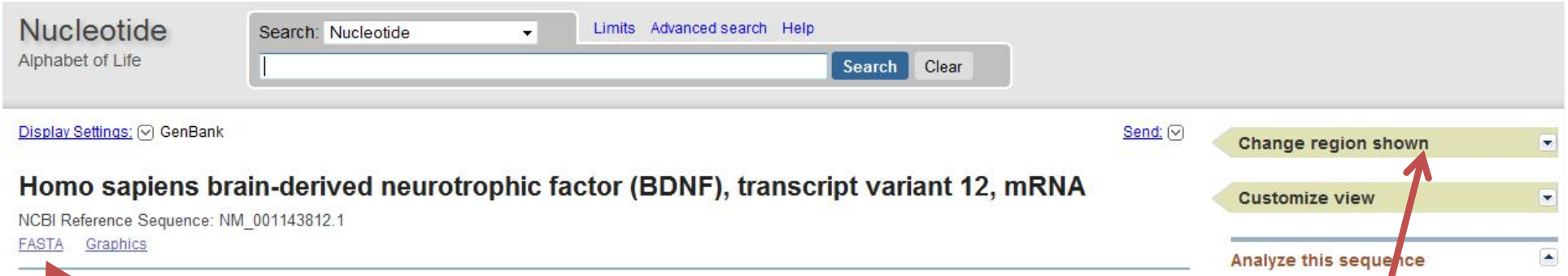
```
Query  721  TTGACCATTAAAAGGGGAAGATAG  744
            |||||||||||||||||||||||
Sbjct  967  TTGACCATTAAAAGGGGAAGATAG  990
```

■ "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.**

ASSIGNMENT 1 - INSTRUCTIONS - some remarks on data collection



■ 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

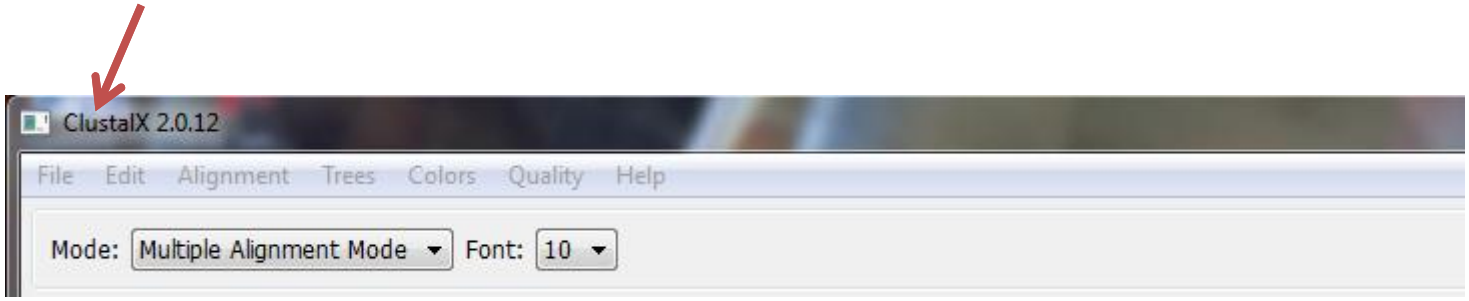
here, but you don´t want to take the the whole sequence behind this accession number and thus you use this and type the region you want (for example 247-990).

ASSIGNMENT 1 - INSTRUCTIONS - ALIGNMENT

■ The default in computer class C128 is that you use the installed programs ClustalX for alignments and Genedoc for editing the alignments

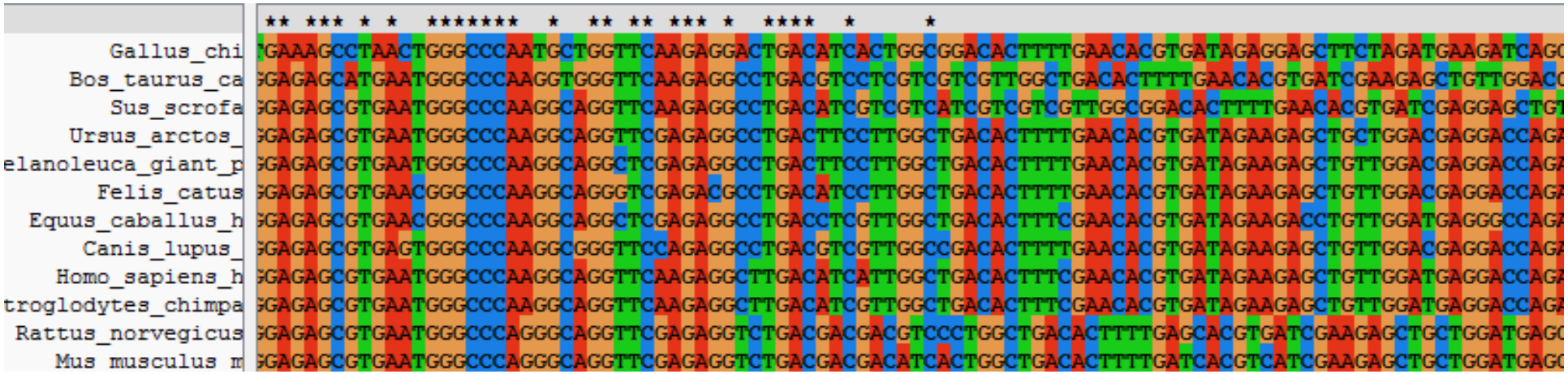
■ Course webpage has an example of an aligned FASTA-file (you must do that for the expanded dataset) and a MEGA-file (= aligned FASTA with some changes).

■ Your FASTA-file here

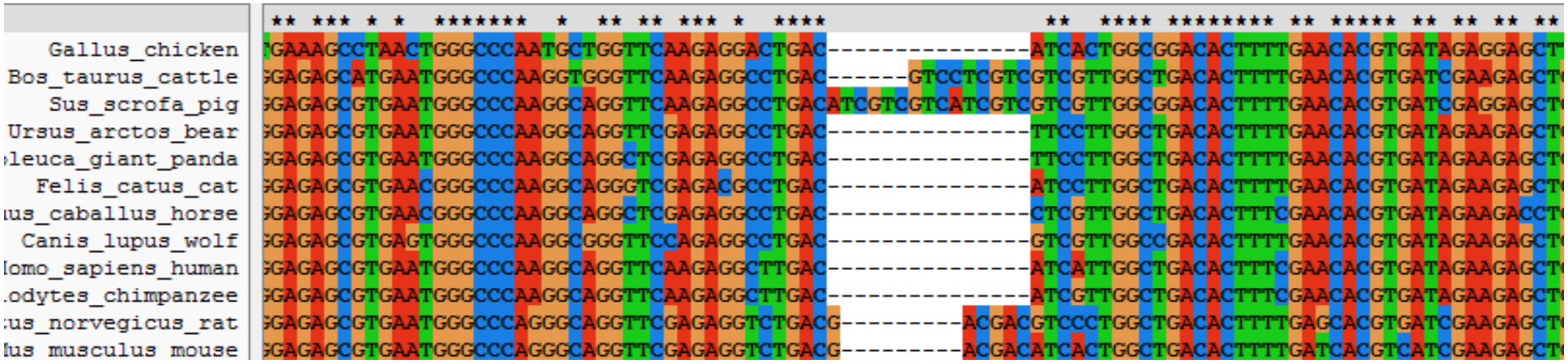


ASSIGNMENT 1 - INSTRUCTIONS - ALIGNMENT

- Your data in Clustal, before alignment, looks like this…



- … and after alignment

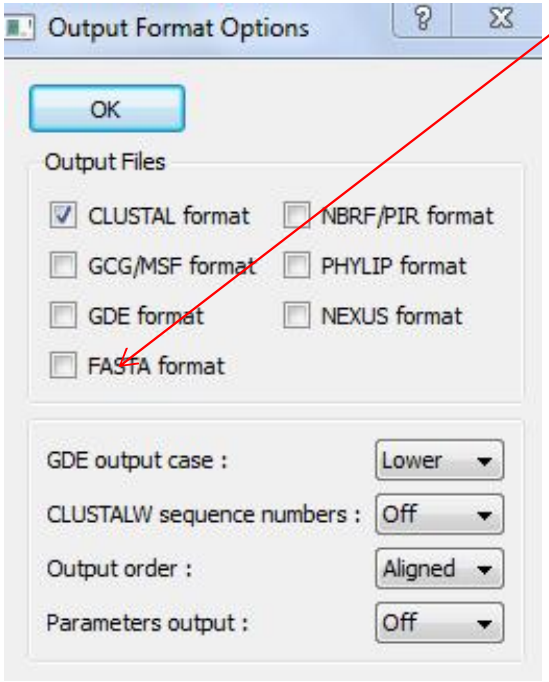


ASSIGNMENT 1 - INSTRUCTIONS - ALIGNMENT

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 FASTA-format = aligned FASTA -> MEGA-format

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

When you get the alignment, you should consider, whether everything is okay, taking into account that sequences should form a protein coding gene => only 3 nucleotide (or multiplies of 3) gaps (deletions / insertions) are reasonable. Why?

You don´t have to do editing because it might be too laborious! Include in your report, what kind of mistakes you have noticed! And proceed to following steps of the assignment (by using a wrong alignment). And: keep in mind that if you were doing real science, you should not do like this.

Advise for clustering by MEGA5 will be added here