TUTORIAL

- Learning basic operations in collecting data from sequence databases.
- Aligning the data so that the result is a resonable set forming material for other statistical analyses, such as clustering , phylogenetic trees
- Basic UPGMA-clustering with MEGA5-software.

- Time schedule:
  Proceed so that you have data collection done when you come to the next computer class session Thursday 12. September when we align your data and learn the clustering step.
- Recommendation is that you don´t work alone, **form groups of 2-3 students.**
- Assignment 1, which you are supposed to submit to pass the course, will be given Thursday 12. September. It will be basicly similar than this tutorial-example, but using different material (the p53 gene).
- Assignment 2, which you are also supposed to pass the course will be given Tuesday 17. September.
- Assignment submission deadline: Monday 23. September.

TUTORIAL - INSTRUCTIONS

● The initial dataset 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.

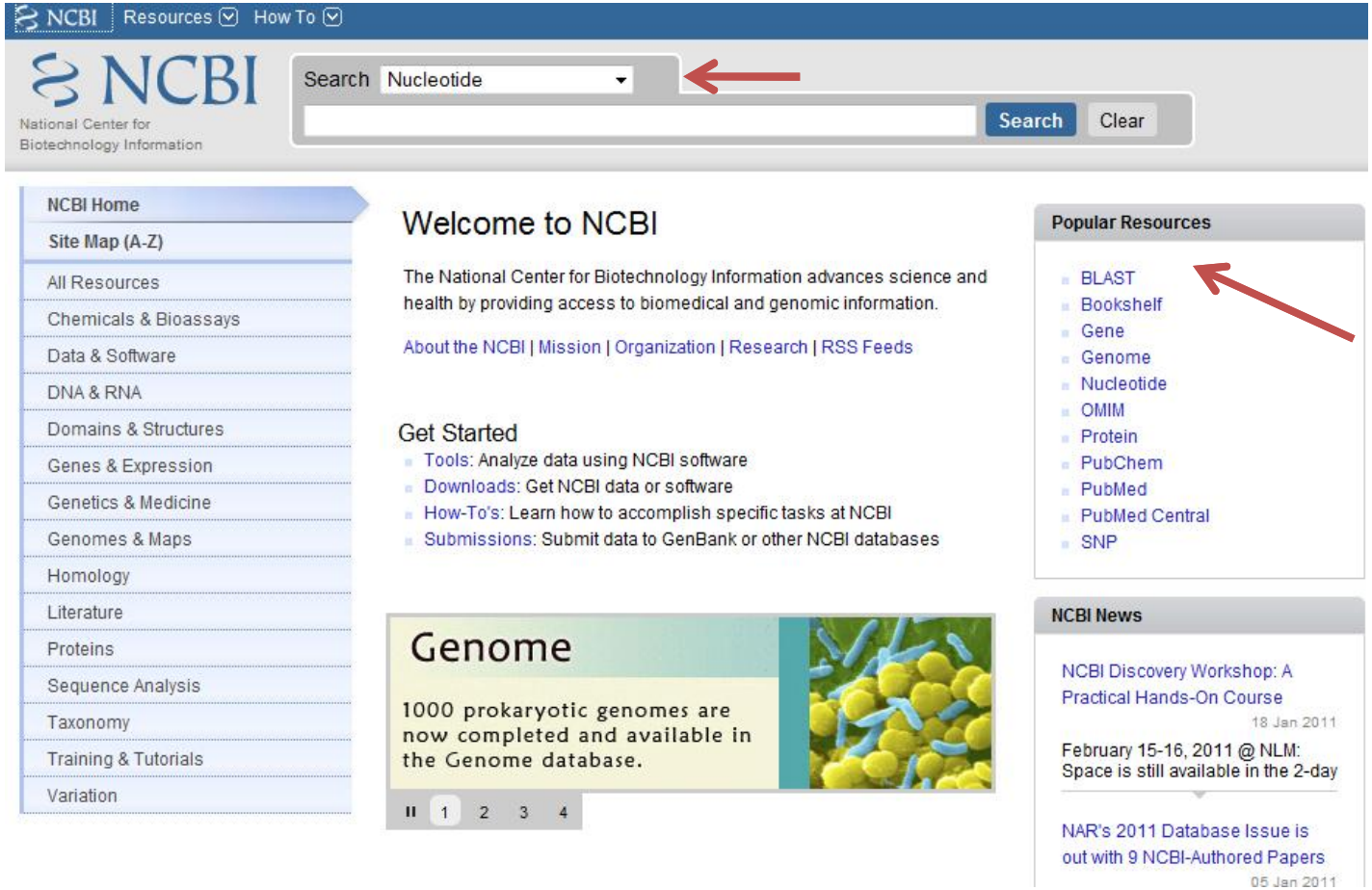
● **Expand the dataset by collecting at least 15 additional animals.**

● 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….

TUTORIAL - INSTRUCTIONS

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

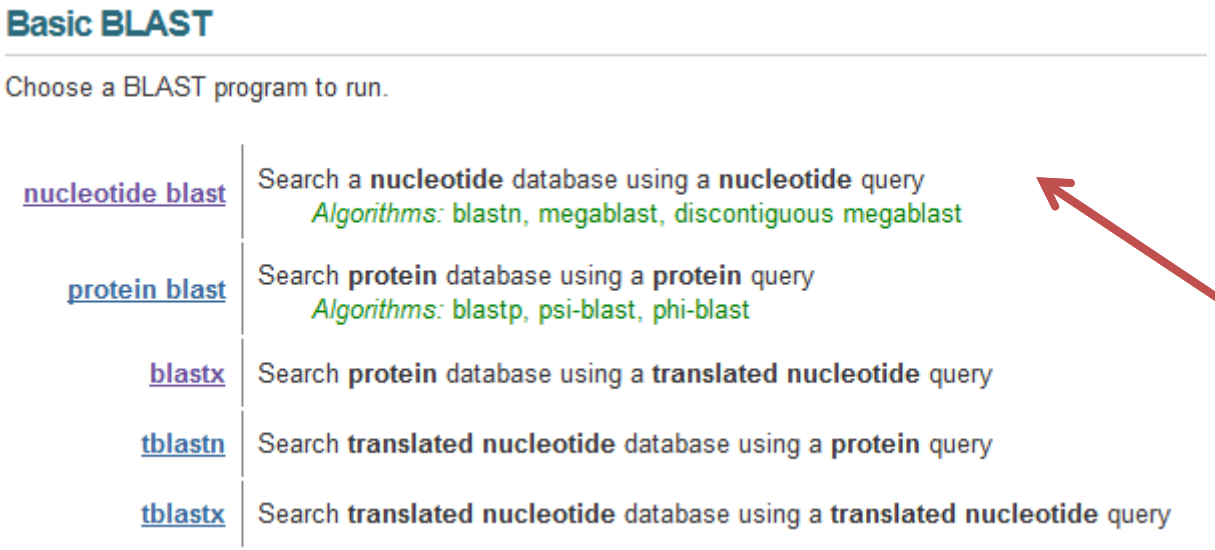


- **Search "nucleotide"** database because you are working with DNA-sequences (more of the like you already have...)

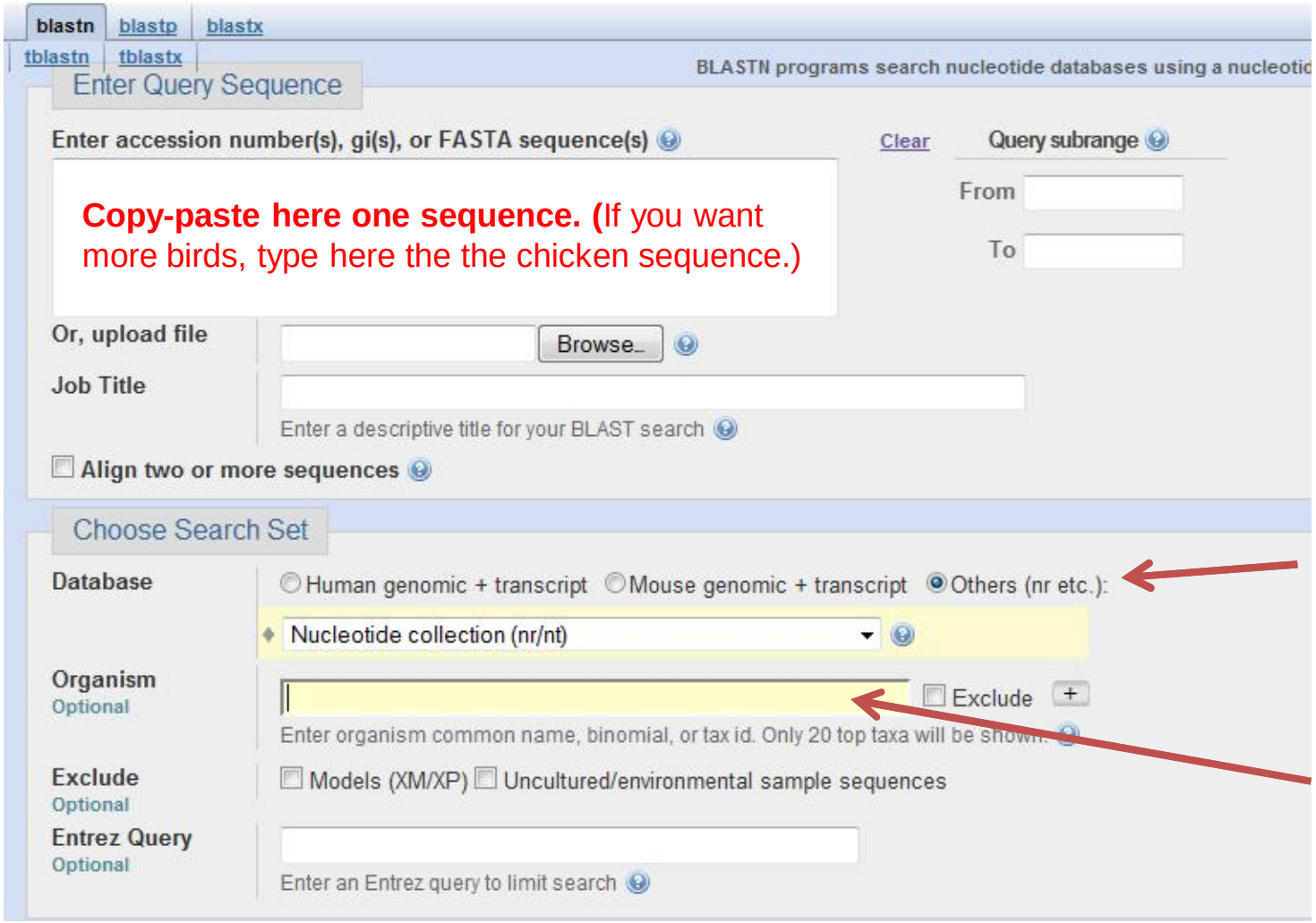
- **You do "BLASTing".** If you want to learn more about these algorithms (topics in other MBI-courses, not in this course), read here, everything is explained, and look at the paper in course webpage.

- 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.



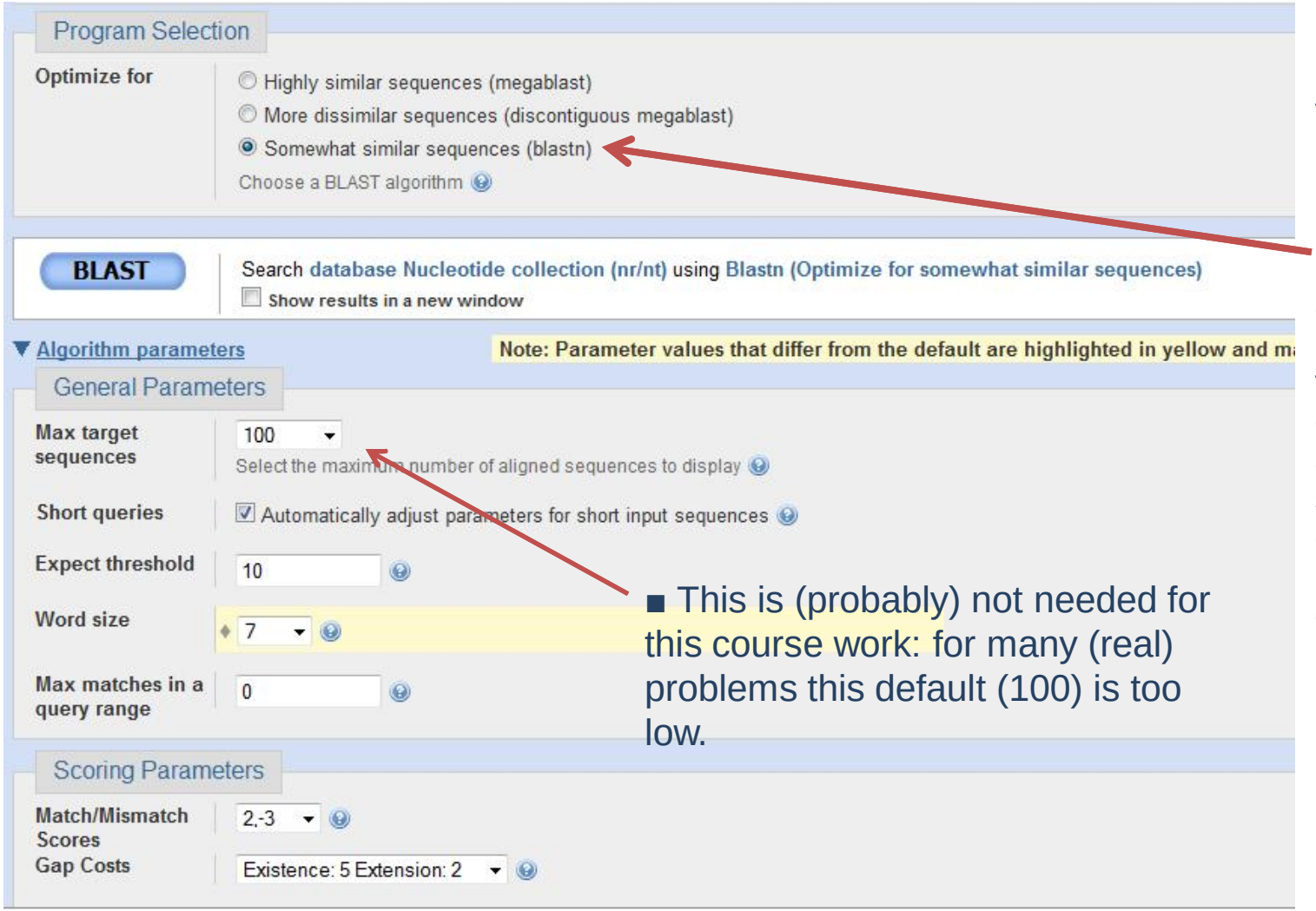
- When you proceed by using a sequence that you already have in the initial file, and you have clicked “BLAST” from the previous page, you are now here and you continue by “nucleotide blast” to the next page……



- ….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.



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

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

● Choose this algorithm!
(Compare the results from a given BLASTing experiment by the three algorithms, you´ll get some practical experience and understanding ”by doing”.)

TUTORIAL - 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).

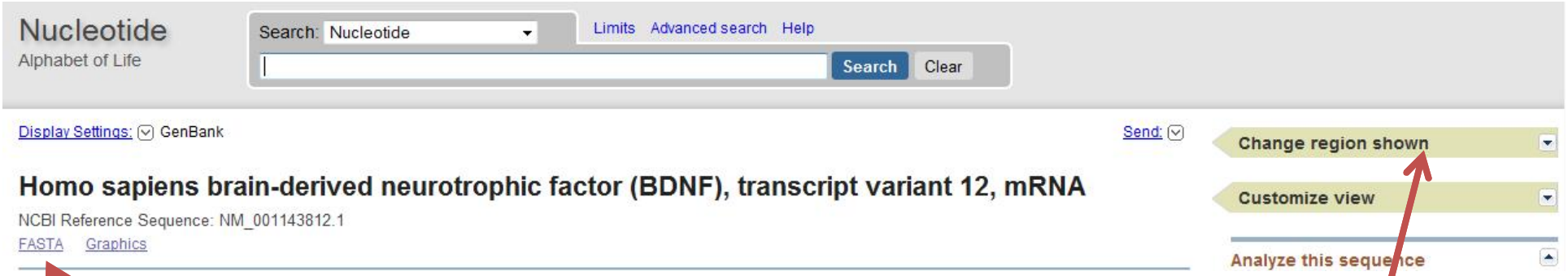
…….

```
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.**



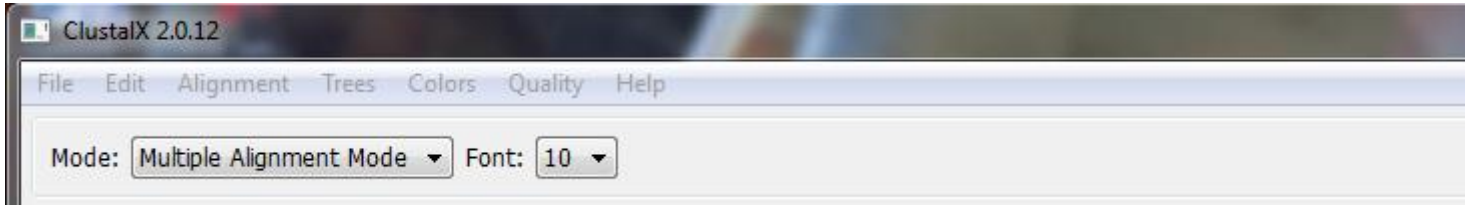
- You have now clicked from one result (from its accession number) and have this page including one sequence for your data collection. You need it in FASTA- format and get that from

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).

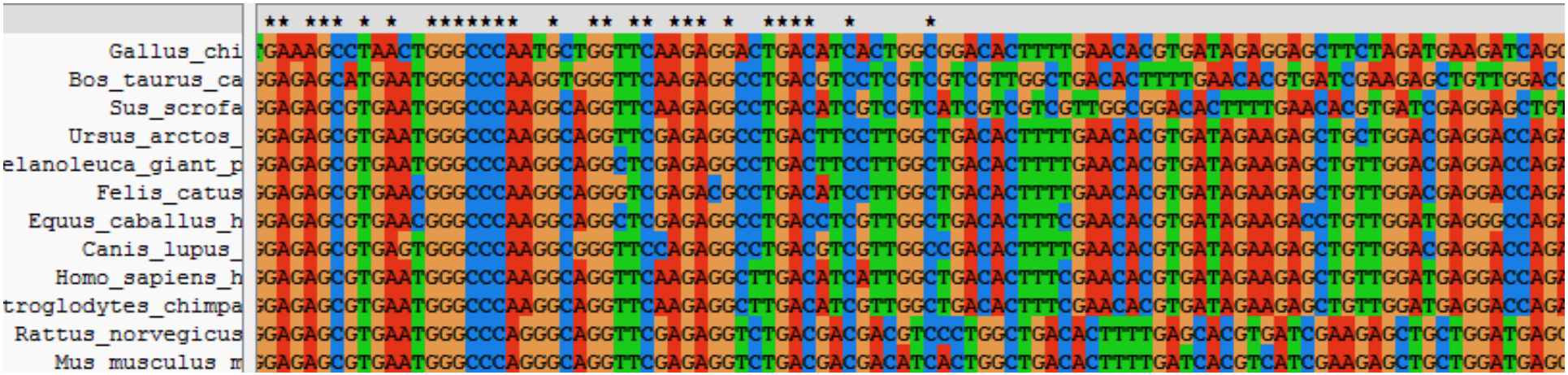
- The default in computer class C128 is that you use the installed programs ClustalX for alignments

- Course webpage has an example of an aligned FASTA-file and you should do that for the expanded dataset.

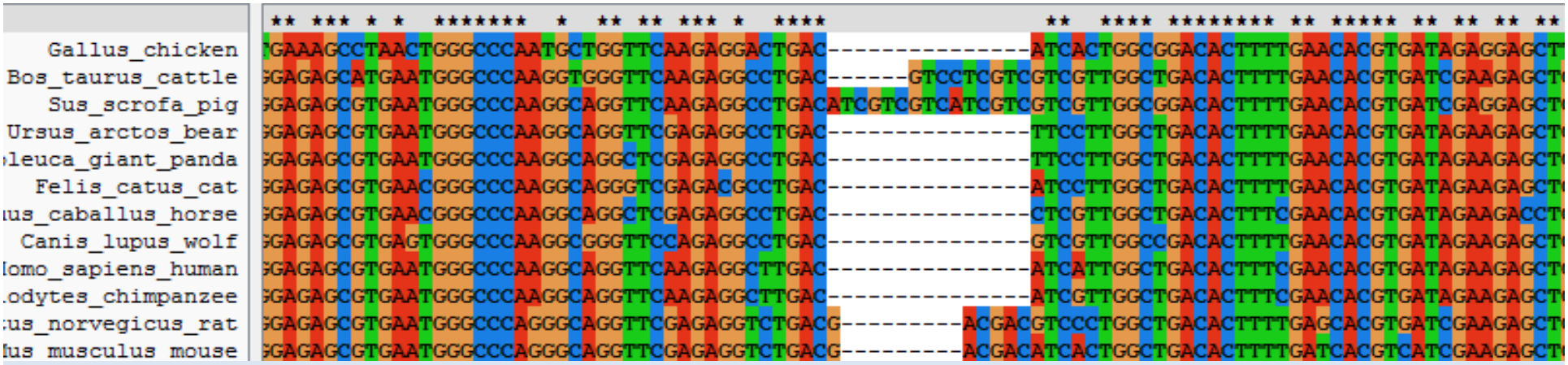
- Your FASTA-file here



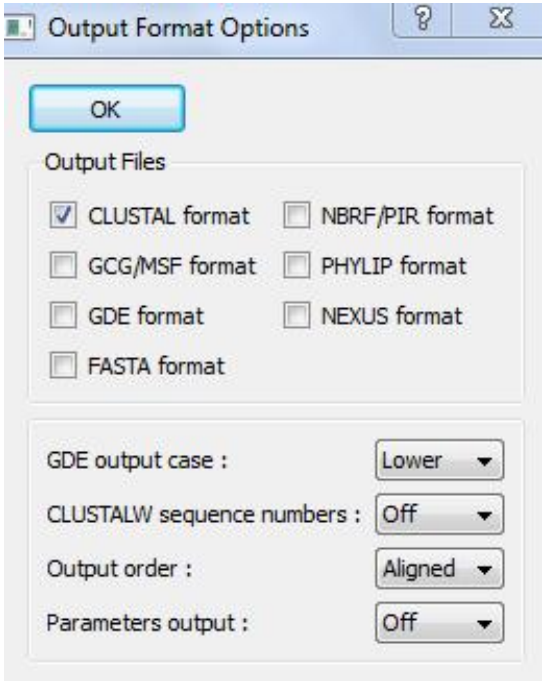
- Your data in Clustal, before alignment, looks like this…



- … 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 also a FASTA-format = aligned FASTA -> MEGA-format
- 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, you should 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.