

# TUTORIAL

- Learning basic operations in collecting data from sequence databases.
- Aligning the data so that the result is a reasonable 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/>

NCBI Resources How To

Search Nucleotide Search Clear

NCBI Home  
Site Map (A-Z)  
All Resources  
Chemicals & Bioassays  
Data & Software  
DNA & RNA  
Domains & Structures  
Genes & Expression  
Genetics & Medicine  
Genomes & Maps  
Homology  
Literature  
Proteins  
Sequence Analysis  
Taxonomy  
Training & Tutorials  
Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS Feeds](#)

### Get Started

- [Tools](#): Analyze data using NCBI software
- [Downloads](#): Get NCBI data or software
- [How-To's](#): Learn how to accomplish specific tasks at NCBI
- [Submissions](#): Submit data to GenBank or other NCBI databases

## Genome

1000 prokaryotic genomes are now completed and available in the Genome database.

### Popular Resources

- [BLAST](#)
- [Bookshelf](#)
- [Gene](#)
- [Genome](#)
- [Nucleotide](#)
- [OMIM](#)
- [Protein](#)
- [PubChem](#)
- [PubMed](#)
- [PubMed Central](#)
- [SNP](#)

### NCBI News

[NCBI Discovery Workshop: A Practical Hands-On Course](#)  
18 Jan 2011  
February 15-16, 2011 @ NLM: Space is still available in the 2-day

[NAR's 2011 Database Issue is out with 9 NCBI-Authored Papers](#)  
05 Jan 2011

- 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.

## TUTORIAL - STARTING BLAST

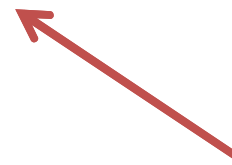
- Make sure that you know what is an accession number and fasta-format of a sequence.
- You have initial knowledge about the BDNF-sequences.
  - You can proceed by copy-pasting one sequence into BLAST-window (see next page), **or**
  - you can write to “search”-window (previous page) BDNF, you´ll get a long list of results, try by restricting the search BDNF primates, or BDNF aves etc.

### Basic BLAST

Choose a BLAST program to run.

<a href="#">nucleotide blast</a>	Search a <b>nucleotide</b> database using a <b>nucleotide</b> query <i>Algorithms: blastn, megablast, discontinuous megablast</i>
<a href="#">protein blast</a>	Search <b>protein</b> database using a <b>protein</b> query <i>Algorithms: blastp, psi-blast, phi-blast</i>
<a href="#">blastx</a>	Search <b>protein</b> database using a <b>translated nucleotide</b> query
<a href="#">tblastn</a>	Search <b>translated nucleotide</b> database using a <b>protein</b> query
<a href="#">tblastx</a>	Search <b>translated nucleotide</b> database using a <b>translated nucleotide</b> 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.....



# TUTORIAL - STARTING BLAST

blastn blastp blastx  
tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query sequence

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)  [Clear](#) Query subrange

From

To

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

Or, upload file

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.):

Organism   Exclude

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

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

Entrez Query

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.

# TUTORIAL - DATASET 1

**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!  
(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.



## TUTORIAL - some remarks on data collection

- Collect the sequences 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 ATGACCATCCTTTTCCTTACTATGGTTATTTCACTTTGGTTGCATGAAGGCTGCCCC 60
      |
Sbjct 247 ATGACCATCCTTTTCCTTACTATGGTTATTTCACTTTGGTTGCATGAAGGCTGCCCC 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.**

# TUTORIAL - some remarks on data collection

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 [v]  
Customize view [v]  
Analyze this sequence [v]

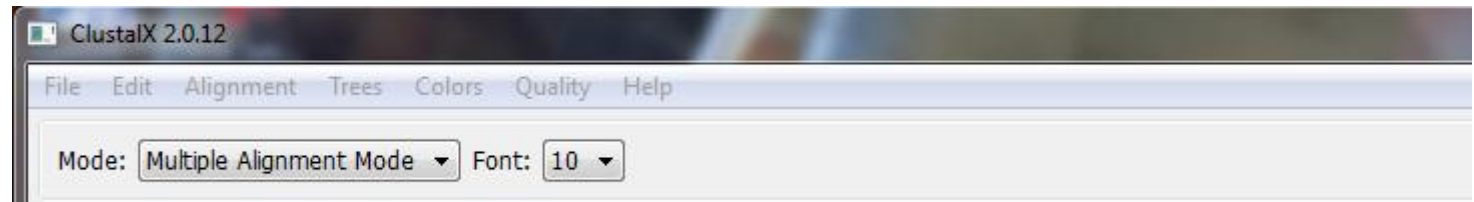
- You have now clicked from one result (from its accession number) and have this page including one sequence for your data collection. You need it in FASTA- format and get that from

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).



## TUTORIAL - ALIGNMENT

- 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



# TUTORIAL - ALIGNMENT

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

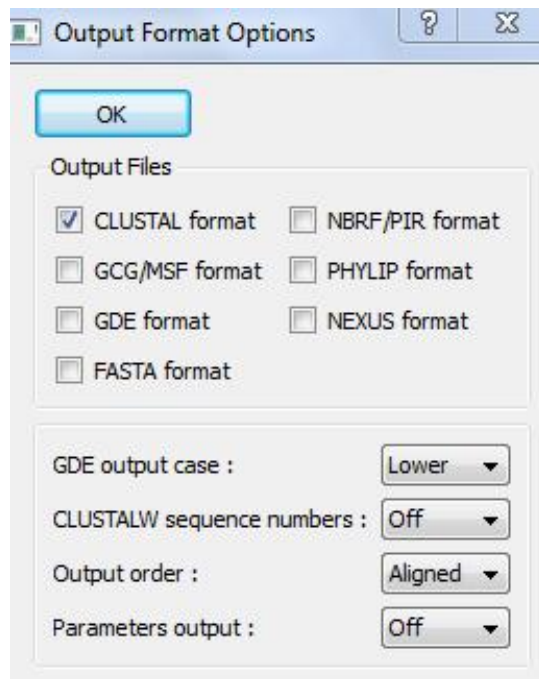
	*** ** * * ***** * ** ** ** * **** *
Gallus_chi	GAAAGCCTAACTGGGCCCCAATGCTGGTTCAAGAGGACTGACATCACTGGCGGACACTTTTGAACACGTGATAGAGGAGCTTCTAGATGAAGATCAGC
Bos_taurus_ca	GAGAGCATGAATGGGCCCCAAGGTGGGTTCAAGAGGCCTGACGTCCTCGTCGTCGTTGGCTGACACTTTTGAACACGTGATCGAAGAGCTGTTGGAC
Sus_scrofa	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCAAGAGGCCTGACATCGTCGTCATCGTCGTTGGCGGACACTTTTGAACACGTGATCGAGGAGCTGT
Ursus_arctos	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCCGAGAGGCCTGACTTCCTTGGCTGACACTTTTGAACACGTGATAGAAGAGCTGCTGGACGAGGACCAG
elanoleuca_giant_p	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCCGAGAGGCCTGACTTCCTTGGCTGACACTTTTGAACACGTGATAGAAGAGCTGTTGGACGAGGACCAG
Felis_catus	GAGAGCGTGAACGGGCCCCAAGGCAGGTTCCGAGAGGCCTGACATCCCTTGGCTGACACTTTTGAACACGTGATAGAAGAGCTGTTGGACGAGGACCAG
Equus_caballus_h	GAGAGCGTGAACGGGCCCCAAGGCAGGTTCCGAGAGGCCTGACCTCGTTGGCTGACACTTTTGAACACGTGATAGAAGACCTGTTGGATGAGGGCCAG
Canis_lupus	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCCGAGAGGCCTGACGTCGTTGGCCGACACTTTTGAACACGTGATAGAAGAGCTGTTGGACGAGGACCAG
Homo_sapiens_h	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCAAGAGGCCTTACATCATTGGCTGACACTTTTGAACACGTGATAGAAGAGCTGTTGGATGAGGACCAG
troglodytes_chimpa	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCAAGAGGCCTTACATCGTTGGCTGACACTTTTGAACACGTGATAGAAGAGCTGTTGGATGAGGACCAG
Rattus_norvegicus	GAGAGCGTGAATGGGCCCCAGGGCAGGTTCCGAGAGGTCTGACGACGACGTCCTGGCTGACACTTTTGAACACGTGATCGAAGAGCTGCTGGATGAGC
Mus_musculus_m	GAGAGCGTGAATGGGCCCCAGGGCAGGTTCCGAGAGGTCTGACGACGACATCACTGGCTGACACTTTTGAACACGTGATCGAAGAGCTGCTGGATGAGC

- ... and after alignment

	*** ** * * ***** * ** ** ** * **** * ** **** ***** ** ***** ** ** ** **
Gallus_chicken	GAAAGCCTAACTGGGCCCCAATGCTGGTTCAAGAGGACTGAC-----ATCACTGGCGGACACTTTTGAACACGTGATAGAGGAGCT
Bos_taurus_cattle	GAGAGCATGAATGGGCCCCAAGGTGGGTTCAAGAGGCCTGAC-----GTCCTCGTCGTCGTTGGCTGACACTTTTGAACACGTGATCGAAGAGCT
Sus_scrofa_pig	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCAAGAGGCCTGACATCGTCGTCATCGTCGTCGTTGGCGGACACTTTTGAACACGTGATCGAGGAGCT
Ursus_arctos_bear	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCCGAGAGGCCTGAC-----TTCCTTGGCTGACACTTTTGAACACGTGATAGAAGAGCT
leuca_giant_panda	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCCGAGAGGCCTGAC-----TTCCTTGGCTGACACTTTTGAACACGTGATAGAAGAGCT
Felis_catus_cat	GAGAGCGTGAACGGGCCCCAAGGCAGGTTCCGAGAGGCCTGAC-----ATCCTTGGCTGACACTTTTGAACACGTGATAGAAGAGCT
us_caballus_horse	GAGAGCGTGAACGGGCCCCAAGGCAGGTTCCGAGAGGCCTGAC-----CTCGTTGGCTGACACTTTTGAACACGTGATAGAAGACCT
Canis_lupus_wolf	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCCGAGAGGCCTGAC-----GTCGTTGGCCGACACTTTTGAACACGTGATAGAAGAGCT
omo_sapiens_human	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCAAGAGGCCTTACATCATTGGCTGACACTTTTGAACACGTGATAGAAGAGCT
odytes_chimpanzee	GAGAGCGTGAATGGGCCCCAAGGCAGGTTCAAGAGGCCTTACATCATTGGCTGACACTTTTGAACACGTGATAGAAGAGCT
us_norvegicus_rat	GAGAGCGTGAATGGGCCCCAGGGCAGGTTCCGAGAGGTCTGACG-----ACGACGTCCTGGCTGACACTTTTGAACACGTGATCGAAGAGCT
us_musculus_mouse	GAGAGCGTGAATGGGCCCCAGGGCAGGTTCCGAGAGGTCTGACG-----ACGACATCACTGGCTGACACTTTTGAACACGTGATCGAAGAGCT

## TUTORIAL - 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 multiples of 3) gaps (deletions/insterions) are reasonable.