Exercise 0

Fasta format

Hands up if you don't know what FASTA format is!

Exercise 1.

BLAST

Find out which protein this sequence belongs to:

```
>protein
MFHPGMTSQPSTSNQMYYDPLYGAEQIVQCNPMDYHQANILCGMQYFNNSHNRYPLLPQMPPQFTNDHPY
FPNVPTISTLDEASSFNGFLIPSQPSSYNNNNISCVFTPTPCTSSQASSQPPPTPTVNPTPIPPNAGAV
LTTAMDSCQQISHVLQCYQQGGEDSDFVRKAIESLVKKLKDKRIELDALITAVTSNGKQPTGCVTIQRSL
DGRLQVAGRKGVPHVVYARIWRWPKVSKNELVKLVQCQTSSDHPDNICINPYHYERVVSNRITSADQSLH
VENSPMKSEYLGDAGVIDSCSDWPNTPPDNNFNGGFAPDQPQLVTPIISDIPIDLNQIYVPTPPQLLDNW
CSIIYYELDTPIGETFKVSARDHGKVIVDGGMDPHGENEGRLCLGALSNVHRTEASEKARIHIGRGVELT
AHADGNISITSNCKIFVRSGYLDYTHGSEYSSKAHRFTPNESSFTVFDIRWAYMQMLRRSRSSNEAVRAQ
AAAVAGYAPMSVMPAIMPDSGVDRMRRDFCTIAISFVKAWGDVYQRKTIKETPCWIEVTLHRPLQILDQL
LKNSSQFGSS
```

Idea: Carry out a Blast search and choose the best match

On the page http://blast.ncbi.nlm.nih.gov/Blast.cgi choose the protein blast option.

Which database to use?

Normally use would use the *nr database* which is the biggest. However, the large the dataset, the longer your search will take.

Can you identify your protein?

Can you find related proteins in chicken? Why do you think that is?
Check the taxonomy distribution.

Go back to the search page and increase the number of alignments to 1000.
Check again the taxonomy distribution.
Save the RID link!

Click on the name of the best hit. On this page search for link “Blink” on the right side.

This page contains previous run Blast search results.

What is the taxonomic distribution of this protein?

Exercise 2.

Repeat the search but now search in chicken only

```
Type in the box Entrex Qyery :
Gallus gallus[organism]
```

What is the best hit, what is its score and E-value? How does it relate to the previous search (you can go back to you previous result using the RID link).

Exercise 3.

PSI-BLAST

Search again with your sequence, but use the Swissprot database.
Run 2 rounds of PSI-Blast. (search for Run PSI-Blast iteration round …)
Can you find new homologues? What is their e-value?

Save the PSSM!

(Download,
PSSMwithparameters)

You can look at your PSSM using this server:
http://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi

Upload the downloaded PSSM here.

Select the Matrix view

Compare the scores for positions 19 and 292.
What is the score of D-D and D-A in these two cases?
What is the reason for this difference?

You can also save the found sequences aligned:

Click on Multiple_alignment
When it is ready click on the Download option and save in Fasta Plus format.

Exercise 4.

You can also carry out Blast searches and alignments directly through Uniprot.

Search for creatine kinase in human. There are different types. Select and M-type and S-type.
Place them in the “Basket” and align them.

Which regions differ in the alignment? What could be the reason for that?
Hint: Use the Highlight function!

Exercise 5.

ALIGNMENT editing

Start the Jalview program!

http://www.jalview.org/
Launch Jalview Desktop

File
Input Aligment
From Texbox

Paste these two sequences and click on the “New Window” button.

```
>sp|Q71U36|TBA1A_HUMAN Tubulin alpha-1A chain OS=Homo sapiens
GN=TUBA1A PE=1 SV=1
MRECISIHVGQAGVQIGNACWELYCLEHGIQPDGQMPSDKTIGGGDDSFNTFFSETGAGK
HVPRAVFVDLEPTVIDEVRTGTYRQLFHPEQLITGKEDAANNYARGHYTIGKEIIDLVLD
RIRKLADQCTGLQGFLVFHSFGGGTGSGFTSLLMERLSVDYGKKSKLEFSIYPAPQVSTA
VVEPYNSILTTHTTLEHSDCAFMVDNEAIYDICRRNLDIERPTYTNLNRLIGQIVSSITA
SLRFDGALNVDLTEFQTNLVPYPRIHFPLATYAPVISAEKAYHEQLSVAEITNACFEPAN
QMVKCDPRHGKYMACCLLYRGDVVPKDVNAAIATIKTKRTIQFVDWCPTGFKVGINYQPP
TVVPGGDLAKVQRAVCMLSNTTAIAEAWARLDHKFDLMYAKRAFVHWYVGEGMEEGEFSE
AREDMAALEKDYEEVGVDSVEGEGEEEGEEY


>sp|Q8SRI6|TBA_ENCCU Tubulin alpha chain OS=Encephalitozoon cuniculi
GN=TUB1 PE=3 SV=1
MREIISLHIGQAGVQIGNACWELYCKEHGILPNGQLDQNKMDDESAESFFSPTSVGTYVP
RTLMVDLEPGVLDSIKTGKYRELYHPGQLISGKEDAANNYARGHYTVGKEIIEPAMEQIR
RMADSCDGLQGFLIYHSFGGGTGSGFASLMMDRLAAEFGKKSKLEFSVYPAPKIATAVVE
PYNSILTTHTTLDYSDCSFLVDNEAIYDMCRNLGIQRPYYTDINRVIAQVVSSITASLRF
```

PGSLNVDLTEFQTNLVPYPRIHFPLVAYSPMLSKEKAAHEKLSVQEITNACFEPQNQMVR
CDTRKGKYMACCLLFRGDVNPKEANNATANVKAKRTNQFVEWCPTGFKVGINSRKPTVLD
GEAMAEVSRAVCALSNTTAISEAWKRLNNKFDLMFSKRAFVHWYVGEGMEEGEFSEARED
LAMLEDDYERISSNAEPVDEY

Now these two sequences are just put underneath eash other. Are there any identities? How common are these (roughly)

Now align the sequences using the Clustalw method:

Web Service
Alignments
Clustal
With Defaults

It can take a while.,.

In the meantime, think about what are the characteristics of a good alignment based on properties such as:

- size of the gaps
- Number of gaps
- The properties of the amino acid in the same column (identical, similar? Different?)

Colour the alignment, this will make it much easier to see the similarities.
Choose the Clustlaw coloring.

It looks like that the C-terminal of the sequence is not conserved.

Remove this regions.
You can highlight this regions using the Left Mouse
Then right clikc:
Selection
Edit
Cut

You can also edit the sequence in a similar way.

Save the alignment as a picture.

First, wrap the lines:

Format

Wrap

Then:

File

Export Image

PNG

Exercise 6.

Now try one of the other nice examples of Aidan Budd multiple sequence alignment.
Copy the sequences from this link into the text window of Jalview:

http://www.embl.de/~seqanal/courses/commonCourseContent/sequences/src_human_ncbiBlastpDefaultsTop50Seqs.fasta

Try aligning the sequences using different program (Clustalw, Clustalw O, MAFTT, MUSCLE)

Do you get the same alignment?
Are there places where they disagree?

Remove empty columns.
Remove redundant sequences at 95%.

Exercise 7.

DNS polimerase I is an ancient protein, its catalytitc units can be found even in micro-organisms.

Load the sequences from 10 micro-organisms (Microorganisms_5-3_exonuc.fasta) into Jalview and align them. Use the MAFFT option.
It is known that for E.c coli the DNA binding regions is located between 184 – 189 with sequences "**GIGPKS**". Find the correspondong region in your alignment. How conserved are the residues in this regions. Which is the most conserved residue?

Use other alignment methods? How much do the alignments differ? Where can you see differences?

Compare the alignment to a manually curated reference alignment.
Microorganisms_5-3_exonuc-REFALIGNMENT.xml