EC Innovative Training Network MATURE-NK

Bioinformatics Course

Exercises Day 1:

BLAST:

1) Go to the BLAST page at NCBI: https://blast.ncbi.nlm.nih.gov/Blast.cgi
2) Enter the protein sequence you saved before as query. You want to search against the “non-redundant protein dadabase (nr), what BLAST type do you have to use?
3) How many hits do you get reported? What range are their scores, what their query coverage and what range the E-value? How many of the hits would you regard as being significant?
4) Mark and save a number of sequences as FASTA. Try to get a good distribution of species/genera (not all the same).
5) Restrict the search to Mus musculus. How does the results change?
6) Restrict the number of hits in range to 10. How does the result change?
7) What happens if you reduce the “Max matches in a query range” to 10? Do you still get all the best scores?

HMMs, SMART, PFAM:

8) Go the the SMART webpage at http://smart.embl-heidelberg.de/
9) Enter your protein sequence into “Sequence analysis” and include the PFAM domains and signal peptides.
10) In how many domains do you get?
11) Are there domains not not shown? If so, determine why.

12) Go the the PFAM webpage at http://pfam.org/ and do a similar analysis.
13) What domains do you find here?
14) Go to the PFAM entry and have a look at the HMM logo. What do you see. (We will take a closer look later.)
15) Check the architecture – how many sequences show the same domain organization?

Multiple sequence alignment and sequence logos:

16) Go the the MAFFT web interface at https://mafft.cbrc.jp/alignment/server/index-rawreads.html
17) Enter you sequences as FASTA file and align them with the automatic option.
18) Take a look at the alignment does it look reasonable?

19) Go to the sequence logos page at https://weblogo.berkeley.edu/logo.cgi

20) Create a sequence logo. Which positions are most informative?