COMP 364: Class work

EMBOSS (sequence alignment part 2)

March 18, 2013

1 Protein sequence databases

UniProt is one of the most popular protein sequence database: `http://www.uniprot.org/`. Search for all sequences in UniProtKB accepting the word "globin".

Refine your search and filter only sequences from "homo sapiens". Select the myoglobin from the manually curated database swissprot.

What is his function? In which part of the human body is it located? Download and store its sequence. Download and store its sequence.

Search for the human proteins with the same function located in the blood. What its common name? How many subunits did you find?

Download the sequences of the alpha and beta subunits.

2 Global alignment

We will align the subunits alpha and beta using the Neddleman-Wunch algorithm with the default parameters.

What is the score? The number of gaps? The sequence identity (i.e. number of similar positions)? An help of parameter setting can be found at: `http://www.ebi.ac.uk/Tools/psa/emboss_needle/help/` (look for the help on matrices). What are parameters are recommended for BLOSUM matrices? for PAM matrices. (NB: The web sever restricts the values you can use. Choose the closest one.)

Align the subunits alpha and beta using the new parameters (you will need to run `needle -h` to list the options and find how to format your command line). Compare results.

Perform the same operation (using defaults and new values) with the myoglogin and the subunit alpha. What do you observe?

Try to align the subunit alpha and the myoglobin using other BLOSUM matrices and parameters. Which ones will you choose? What do you observe?

3 Local alignment

You will now use an implementation of the Smith-Waterman algorithm for nucleotide sequences: `water`. Unlike the needleman-Wunch algorithm, this program intends to identify sub-regions with higher similarities. We are going to look at members of the rhodopsin family of G-protein coupled receptors. First, you collect data at: `http://www.ebi.ac.uk/embl/`. Extract the sequences with accession numbers L07770 and U23808. From which organism are they coming from? What type of molecules are they?

Align the two sequences using the Smith-Waterman algorithm with the default parameters. What do you notice? How many matching regions did you find? Can you guess what are these regions?