COMP 364: Homework 3

EMBOSS (25 points)

Assigned: March 20, 2013
Due on April 1, 2013 by email before 11:59am

Submit all solutions in the form of python script files that the TA can run. Zip your solutions up into a compressed file called `<your name>_hw3.zip` *and email it to* `jerome.waldispuhl@mcgill.ca` *(indicate COMP364 in the subject).*

1 Protein sequence databases (4 pts)

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 UniProtKB/Swissprot database (with a yellow star).

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 which are located in the blood. To do so, come back to search interface and add research criteria. The function is reported with the GO term (Gene Ontology). You can also add a criterion for selected only reviewed entries. review each remaining protein and look at the "tissue specificity". What are their common name? How many subunits did you find?

Download the sequences of the alpha and beta subunits.

2 Global alignment (7 pts)

We will align the subunits alpha and beta using the Neddleman-Wunch algorithm with the default parameters. First, write the full command line for running the `needle` program with all arguments (i.e. when you will run it, the program will be completely executed and will not request more instruction from you). Run `needle -h` to list the options and find how to format your command line. What is the score of this alignment? The number of gaps? The sequence identity and similarity?

Read the instructions on matrices at `https://www.ebi.ac.uk/2can/tutorials/matrices.html`. By default, `needle` use the BLOSUM62 matrix. Here we consider that below 50% of sequence identity the sequence are divergent. What matrices should you use in that case? Align the subunits alpha and beta using the suggested matrices (run `needle -h` to list the options and find how to format your command line and). Report the sequence identity and similarity and compare results with the previous experiment. Try different matrices to find the substitution matrix that maximizes the sequence similarity.

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 (7 pts)

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/ena/`. 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?

4 Multiple Sequence Alignments (7 pts)

Visit the Protein Family database at `http://pfam.sanger.ac.uk/`. It gathers sequences with similar functions. Retrieve the "Kinase-like" family and download the seed alignment (8 sequences) and the same sequence *unaligned*. Align the unaligned sequences with the program MUSCLE `http://www.ebi.ac.uk/Tools/msa/muscle/`. Print the results using the `prettyplot` program from EMBOSS.

We will now compare these two alignments (Pfam and MUSCLE) with biopython. Use the `AlignIO` module, to read the alignments. Then, compute and compare the consensus sequences and the information contents (You will find instructions on the `AlignIO` module in section 6 & 18 at `http://biopython.org/DIST/docs/tutorial/Tutorial.html`).