COMP 364: Class work

EMBOSS (sequence alignment)

March 15, 2013

We will experiment the sequence alignment programs implemented in the EMBOSS software suite. This overview is not exhaustive and I encourage you to explore the suite by yourself as well. We will also interface this software with biopython.

1 Prepare the data

Download the sequences at : `http://www.cs.mcgill.ca/~jeromew/comp364/data/PF00870_seed.txt` Using Biopython, create a script that reads the file `PF00870_seed.txt` and store each sequences individually in a file named from the ID stored in the header of each sequence (E.g. `P53_DANRE/63-257.fasta`).

2 Running Needleman-Wunch Algorithm

In a terminal, run the program `neddle` and compare the sequence with the ID's `P53_DANRE/63-257` and `P53_ONCMY/83-278`. What is their sequence identity and similarity? What is their alignment score?

3 Automatic generation of commandline with Biopython

We will use the function `NeedleCommandline` of Biopython. First, you will need to import this function with : "`from Bio.Emboss.Applications import NeedleCommandline`". Then create a command line using : `needle-cline = NeedleCommandline()`. You will indicate the FASTA file to use using the `.asequence` and `.bsequence` fields. E.g.: `needle-cline.asequence=<filename1>`.

Similarly, you will set up the gap open penalty and gap extend penalty with `needle-cline.gapopen=10` and `needle-cline.gapextend=0.5`, and the output file with `needle-cline.outfile="needle.txt"`.

Create and print your command line.

4 Run EMBOSS with Biopython

Next we want to use Python to run this command for us. For full control, we recommend you use the built in Python `subprocess` module, but for simple usage the wrapper object usually suffices: `stdout, stderr = needle_cline()`

5 Retrieve EMBOSS output

Load the output file with `Bio.AlignIO`. import first the module with : `from Bio import AlignIO`. Next, call the function `read` and store your result in a variable `align = AlignIO.read("needle.txt", "emboss")`.

Print your output.

6 Iterating

Enumerate all pairs of sequences found in `PF00870_seed.txt` and compute a pairwise for each of them. Find the two most and least related sequences.