ECS 124A Theory and Practice of Bioinformatics

Lab Assignment 1

Instructor: Ilias Tagkopoulos (iliast@ucdavis.edu)
TA: Linh Huynh (huynh@ucdavis.edu)

Key dates:

Assigned: 10/02/2014
Due by: 10/14/2014, 1:30pm

Scope:
Familiarize with (a) the PERL scripting language, (b) NCBI and its tools, (c) simple alignment models and their parameters.

Deliverables:

Answers for exercises 1.1, 1.2, 1.3, 1.5, 1.6, 1.7, 1.8, 1.9 from the “Introduction to Perl 1” notes (Smartsite or link below). Perl code and answers to the exercise in Part D. Also write 1-5 sentences for each of the questions below:

1. What is NCBI, Entrez, Gene Expression Omnibus (GEO)
2. What is Blast and what FASTA ? What are the differences of these two.
3. Write down the names of at least their sequence alignment algorithms and their main differences.
4. What is PubMed? What other ways exist to do the same thing?

Lab Instructions:

PART A: Introduction to Perl

You should be familiar with the Unix environment, if not, check out Prof. Matloff’s tutorial here:

http://heather.cs.ucdavis.edu/~matloff/unix.html

We will be using a document “Intro to Perl 1” that can be found under resources in SmartSite or here:

http://cs124.cs.ucdavis.edu/Workshop1/pb1.pdf

Follow the instructions and examples and complete the exercises as posted. The course TA will be there to help you.

PART B: Introduction to NCBI and PubMed

What is NCBI? Go to

http://www.ncbi.nlm.nih.gov/

to find out. Read its mission here:

http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html

What are the tools that it supports? Read more here:

http://www.ncbi.nlm.nih.gov/guide/dna-rna/

Click on relevant links and find out about Blast, FASTA, Entrez and Genebank. Once you are done, go and look online what PubMed and MEDLINE is and where they are used. Try to search for a scientific paper that includes the words “sequence alignment”.

PART C: Protein Sequence Alignment

Elvis Lives Example

It is reported that both the words "elvis" and "lives" each appear as part of protein sequences held in several protein databases (UniProtKB/SwissProt for example). Both of them appear multiple times, but they never appear together. We want to know if this is true.

To answer this question you would want to scan the sequence content of a protein database. Unfortunately Entrez does not allow you to answer this question (later we will use BLAST which can be used for related kinds of searches but doesn't work for this one – or does it?). However there is a web/database tool that will work for this and will be useful for several other tasks in the class.

http://myhits.isb-sib.ch/cgi-bin/index

Exercise: Find how many times ELVIS appears in the protein database SWISSPROT, how many times LIVES appears, and how many time they appear consecutively. Do ELVIS and DEAD appear together? How many times does PERL appear? Does PERLISGREAT appear? What would it mean if any of these longer statements did appear in the the protein files?

You can use the myhits tool to get these answers. To use my Myhits for this exercise, choose Pattern Search. Enter E-L-V-I-S in the Pattern Input window (unfortunately you do need to put the dashes between successive characters). Choose the database Swiss-Prot and hit search. The search can take a minute or so, but should bring up a Results page listing the number of matches and the details of each match. You might also want to determine if your name in the database.

PART D: Develop a FASTQ to FASTA format converter

Exercise: Write a Perl script, which converts a file from FASTQ format to FASTA format. For your tests use sample.fasq file (uploaded on Smartsite), which contains 10,000 reads from the Illumina sequencer. (2) Write a reverse converter script (FASTA→FASTQ). (3) If you open sample.fasq file in the editor you will see a lot of ‘B’ symbols in the “quality value’ lines. What does B mean? (*Hint: use Wikipedia; our fastq file is the output from the Illumina 1.5+ sequencer*). What is the probability that a nucleobase read at a position marked with B is incorrect? (Give a number or a range.)

END OF LAB #1