CS 4630/5630 Kresman Homework 5

Bioinformatics: When your boss asked about the internship experience, you said, ‘Frankly, I was not ready to take on the Kinematics project. However, as I look back, the opportunity to work with a couple of computational sciences’ teams in a span of few weeks has really perked up my curiosity and boosted my confidence!” Sure enough, boss was ecstatic at your response as another team, ‘bioinformatics team,’ that was firing on all cylinders due to the coronavirus outbreak was looking for quick learners like you to step in and assist in its investigations.

**Problem 1:** UniProt provides many services to researchers, among them: info on published literature for protein sequences; experimental data sets that are manually reviewed by scientists; entire set of proteins in a genome (for example, the list of all proteins in a virus). Write code to get information on a UniProt entry in XML format and **pretty print** all its references (just like the reference list in academic papers). Make sure the UniProt entry you pick has at least three references. [NCBI, like many other servers, provides both browser access and programmatic access its databases. For example, P59594 is a UniProt entry and one can use a browser to check the entry and its references (i.e. publications that refer to this) @ UniProt website.

**Problem 2:** Mass spectrometer separate the components of a sample by their mass and electrical charge and produces a mass spectrum that plots the mass-to-charge (m/z) ratio of compounds in a mixture. The diverse range of mass spectrometry (MS) instrumentation and its corresponding proprietary/nonproprietary data formats has generated a call by the proteomics community for a standardized format to facilitate management, processing, storing, visualization, and exchange of both wet lab data and processed data. One such format is mzXML, an XML based common file format for proteomics mass spectrometric data.

* Write code to count the number of spectra in the file "Homework 5 Input 2.mzXML" using ElementTree’s methods
* Print the following

How many MS/MS (attribute "msLevel" is 2) spectra (tag "scan") are there?

How many MS/MS spectra have precursor m/z value between 600 and 920 Da?

**Problem 3:** The bioinformatics team just obtained a genomic query sequence (‘Homework 5 Input 3.fasta’) from a species and wondered if this query sequence has regions with sequence similarities to known genes. BLAST to the rescue…

BLAST is a local search alignment tool (two major categories: local and global, and BLAST is local. In the next couple of weeks, we may discuss the technical details of these two categories, but for now we will just ‘use’ a local tool)

* Write code to run BLAST on the query sequence and save the BLAST output in XML format to ‘Homework 5 Output 3.XML.’
* The BLAST E-value is the number of expected hits of similar quality (score) that could be found just by chance. In other words, the smaller the E-value, the better is the match. Your bioinformatics team is very selective and wants to work with an E-value threshold of 1e-50.
* Parse the XML output file and print info of up to 100 hits with ‘expect’ value (aka E value) less than the threshold. P**rint** the alignment hits in a nice format (as a table) with at least these fields: title (10 characters), length, score, and expect. If you like add additional fields, but each hit printout is no more than 80 characters wide
* The NCBI BLAST server is a shared resource for scientists so do not query it more than 3 or 4 times please! BLAST takes a while to run and the output may be 20+Mb! So, please consider BLASTing just ONCE; since program development takes many runs, 1st run does the BLAST (and creates the output XML file), while subsequent runs just reuse the same (output) file of 1st run.

**Resource/notes**:

* CS 4630 Do Problem 1 OR 2, and Problem 3. CS 5630 Do all three problems. Console or GUI-based, whatever floats your boat
* Chapter 7 of Biopython tutorial is an amazing resource
* One program file with the problems in separate cells
* Consider commenting on the utility value of these problems for bioinformaticians/biologists - we are no domain experts, but googling can help, thanks.

**Canvas** submissions (do not zip): lastnameHw5.ipynb; Homework 5 Output 3.XML; a word doc of run snapshot