ECE398BD: Making Sense of Big Data

Lab 11

Handed Out: April 4, 2017 *Due: 2pm, April 13, 2017*

In this lab, you will familiarize yourself with various applications of dynamic programming in bioinformatics and the BLAST tool. Grading is based on the lab report which is due at the beginning of the Lecture on Thursday, April 13th. We will deduct 20 points from your score for each day after the due date (late penalty). You may submit a printed report in person or an electronic report via email to the TAs, Vida and Hussein. Be sure to include all of your codes along with the report.

1. **Sequence Alignment – (60 points)**

In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than (" > ") symbol in the first column. The word following the " > " symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). There should be no space between the " > " and the first letter of the identifier. The sequence ends if another line starting with a" > "appears; this indicates the start of another sequence. A simple example of one sequence in FASTA format:

```
>AAB Human gene example
GGCAGATTCCCCCTAGACCCGCCCGCACCATGGTCAGGCATGCC
CGCTGGGCACAGCCCAGAGGGTATAAACAGTGCTGGAGGCTGGC
CAGGCCAGCTGAGTCCTGAGCAGCAGCCCAGCGCAGCCACCGA
>AAC Mouse gene example
AAAAACCCCCTGATTTTAGTCCCCCCCGATCTTAGTCCCGTAGCG
AAAAAATCTTTCTATTCTTATTCTTAGTCCCGTAGCGCTTTTAGTCG
```

A fasta file can have any of the following extensions: .fasta, .fas, .fa, .seq, .fsa, .fna, .ffn, .faa, .frn.

a. Calculate, by hand, the dynamic programming matrix and an optimal alignment for the DNA sequences CTTAAG and CTAATG, scoring +2 for a match, -1 for a mismatch, and with a gap penalty of 2 (i.e., each gap column contributes -2).

b. Calculate, by hand, the dynamic programming matrix and an optimal alignment for the DNA sequences CTTAAG and CTAATG, scoring +1 for a match, -1 for a mismatch, and with a gap penalty of 3 (i.e., each gap column contributes -3). Did you get the same result as in a)?

c. Using your favorite programming language, implement a dynamic programming algorithm (Needleman Wunsch) that finds the optimal global alignment of two input DNA sequences.

   The program should take as input: 1) fastafile1.fna containing sequence 1, 2) fastafile2.fna containing sequence 2, 3) subs.txt containing the 4 by 4 substitution score matrix, 4) a negative integer representing the "gap penalty".

   It should then output an optimal global alignment (if there is more than one optimal alignment, the program should pick any one of them).

   For example, the program will be run from the Linux command-line as:

   <programfilename><fastafilename1><fastafilename2><substitutionmatrixfile>
   <gappenalty>

   The output should be a textual display of the optimal global alignment as follows:

   "The optimal alignment between given sequences has score X."

   A C C C – – A C ... A T

   A C – C G G A C .. A T

   Since the global alignment algorithm is a dynamic programming algorithm, your program should align two sequences of length n in $O(n^2)$ time.

   Run your program on seq1.fna, seq2.fna, sub.txt, gap penalty -8 and include the output in your report.

d. Run your program on BRCA1_part_mutated.fna, BRCA1_part.fna, sub.txt, and gap penalty -8. Include the output in your report.

e. Run your program on BRCA1_part_mutated.fna, BRCA1_part.fna, sub2.txt, and gap penalty -1. Include the output in your report.

2. **BLAST (Basic Local Alignment Search Tool) – (20 points)**

   BRCA1 (breast cancer 1, early onset) is a human gene normally expressed in the cells of breast and other tissue, where it helps repair damaged DNA. If *BRCA1* is mutated, then DNA damage may not be repaired properly, and this increases the risk for breast cancer. The human *BRCA1* gene is located on the long (q) arm of chromosome 17 at region 2 band 1, from base pair 41,196,312 to base pair 41,277,500. The sequence can be found in BRCA1.fasta file, or here: `http://www.ncbi.nlm.nih.gov/nuccore/NC_000017.11?report=fasta&from=41277500&to=41196312&strand=true`.

   BLAST is a software tool used to find regions of "local similarity" between sequences. Given that dynamic programming is computationally expensive, it uses a suboptimal procedure for finding multiple alignments and for evaluating their p-values. Link: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=OGP__9606__9558`

   a. Using BLAST, align "BRCA1_part.fna" against the human genome (GRCh38 reference assembly top-level, Annotation Release 106) using the default setting. Include the "Sequences producing significant alignments" table provided by BLAST.

   b. Repeat part a) using "BRCA1_part_mutated.fna".

3. **RNA fold prediction – (20 points)**
   Vienna RNAfold (`http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi`) is a software that predicts minimum energy secondary structures of RNA molecules, as well as pairing probabilities. Again, the method used is dynamic programming.

   a. Predict the secondary structure of "BRCA1_part.fna" on Vienna. Include the result (graphics) in your report.

   b. Predict the secondary structure of "BRCA1_part_mutated.fna" on Vienna. Include the result (graphics) in your report.

   c. Are the results from part a and b the same? Note that changes in shapes of macromolecules may cause severe disruptions in cellular function.

   d. Repeat steps a)-c) with a change in the temperature of the folding environment from 37C to 100C? What do you think is preventing the strings to fold?