CS612 – Algorithms in Bioinformatics

Homework Assignment 1
Due Thu. Oct. 1 before midnight, on Gradescope

Objectives

- Learn how to perform sequence alignment and analyze the results
- Understand and apply the theory of sequence alignment

The FASTA format

FASTA (pronounced "fast-ay") is a text-based format for representing either nucleotide sequences or peptide sequences, in a single letter code. It is very simple. The top line starts with a $>$ followed by free text. The rest are the nucleotides or amino acid letter codes.

Part 1 – Practice

1. **Sequence alignment hands-on exercise.** You are given the following protein sequence:
   >protein
   `TCPFADPAALYSRQDTTSGQSPLAAYEVDDSTGYLTSDVGGPIQDQTSLKAGIRGPTLLEDFMFRQKIQHFDHERVPERAV`

   The sequence is also available at the course webpage: http://www.cs.umb.edu/ nurith/cs612/protein.fasta .

   (a) Go to the Blast website at http://blast.ncbi.nlm.nih.gov. Select "Protein Blast" to get to the BlastP website. Paste the above sequence to the top window. Use the default parameters – nr database (non-redundant protein sequence database). Hit "BLAST" – the search may take a few seconds. Save the search results using the "download" button and the "txt" option above. Submit a printout of the first 2 pages of the saved search as part of your homework. **DO NOT attach the entire search result – it's very large.**

   Notice: You can print screen and embed it as a picture into a word doc, later converting to PDF. You can also print the screen to a file (in pdf format) and merge it into your main document. Many online apps merge PDF documents.

   (b) Repeat the search above with the UniprotKB/SwissProt database and attach it first 2 pages as part of your homework (remember all the HW should be uploaded as a single PDF file).

   (c) Repeat search a. above with PAM30 as a substitution matrix. This can be done in the blastP homepage by opening "algorithm parameters" at the bottom of the page. Observe the changes between the results of a and c due to the change in the substitution matrix: Look at the first entry that differs between a and c. What is its rank in a and c? What is the name of the protein sequence in this entry?

2. **DNA sequence alignment:** The following sequence was constructed by NCBI scientist Mark Boguski for Michael Chrichton's "The Lost World" of the Jurassic Park series:
>DinoDNA from THE LOST WORLD p. 135
GAATTCCGGAAGCGAGCAAGAGATAAGTCCTGGCATCAGATACAGTTGGAGATAAGGACG
GACGTGTGGCAGCTCCCGCAGAGGATTCACTGGAAGTGCATTACCTATCCCATGGGAGCC
ATGGAGTTCGTGGCGCTGGGGGGGCCGGATGCGGGCTCCCCCACTCCGTTCCCTGATGAA
GCCGGAGCCTTCCTGGGGCTGGGGGGGGGCGAGAGGACGGAGGCGGGGGGGCTGCTGGCC
TCCTACCCCCCCTCAGGCCGCGTGTCCCTGGTGCCGTGGGCAGACACGGGTACTTTGGGG
ACCCCCCAGTGGGTGCCGCCCGCCACCCAAATGGAGCCCCCCCCACTACCTGGAGCTGCTG
CAACCCCCCCGGGGCAGCCCCCCCCATCCCTCCTCCGGGCCCCTACTGCCACTCAGCAGC
GGGCCCCCACCCTGCGAGGCCCGTGAGTGCGTCATGGCCAGGAAGAACTGCGGAGCGACG
GCAACGCCGCTGTGGCGCCGGGACGGCACCGGGCATTACCTGTGCAACTGGGCCTCAGCC
TGCGGGCTCTACCACCGCCTCAACGGCCAGAACCGCCCGCTCATCCGCCCCAAAAAGCGC
CTGCTGGTGAGTAAGCGCGCAGGCACAGTGTGCAGCCACGAGCGTGAAAACTGCCAGACA
TCCACCACCACTCTGTGGCGTCGCAGCCCCATGGGGGACCCCGTCTGCAACAACATTCAC
GCCTGCGGCCTCTACTACAAACTGCACCAAGTGAACCGCCCCCTCACGATGCGCAAAGAC
GGAATCCAAACCCGAAACCGCAAAGTTTCCTCCAAGGGTAAAAAGCGGCGCCCCCCGGGG
GGGGGAAACCCCTCCGCCACCGCGGGAGGGGGCGCTCCTATGGGGGGAGGGGGGGACCCC
TCTATGCCCCCCCCGCCGCCCCCCCCGGCCGCCGCCCCCCCTCAAAGCGACGCTCTGTAC
GCTCTCGGCCCCGTGGTCCTTTCGGGCCATTTTCTGCCCTTTGGAAACTCCGGAGGGTTT
TTTGGGGGGGGGGCGGGGGGTTACACGGCCCCCCCGGGGCTGAGCCCGCAGATTTAAATA
ATAACTCTGACGTGGGCAAGTGGGCCTTGCTGAGAAGACAGTGTAACATAATAATTTGCA
CCTCGGCAATTGCAGAGGGTCGATCTCCACTTTGGACACAACAGGGCTACTCGGTAGGAC
CAGATAAGCACTTTGCTCCCTGGACTGAAAAAGAAAGGATTTATCTGTTTGCTTCTTGCT
GACAAATCCCTGTGAAAGGTAAAAGTCGGACACAGCAATCGATTATTTCTCGCCTGTGTG
AAATTACTGTGAATATTGTAAATATATATATATATATATATATATCTGTATAGAACAGCC
TCGGAGGCGGCATGGACCCAGCGTAGATCATGCTGGATTTGTACTGCCGGAATTC

The sequence is also available at the course webpage: http://www.cs.umb.edu/ nurith/cs612/dino.fasta

Perform a Blast search using blastn (nucleotide search) and the default non-redundant (nt) nucleotide database.

(a) What are the two main species used to construct the dinosaur DNA sequence?

(b) Repeat the search with blastx (DNA vs. protein sequence) using the default non-redundant protein sequence database. Look at the top sequence alignment and retrieve the hidden message there (hint: look at the gaps...).

Part 2 – Theory

1. Lesk book question 5.3: The edit distance between the strings `agtcc` and `cgctca` is 3, consistent with the following alignment:

```
ag-tcc
cgctca
```

Find the sequence of three edit operations that convert `agtcc` to `cgctca`.

2. **Dynamic programming:**

(a) Use the Needleman Wunsch global alignment Dynamic programming formula in slide set no. 2 to find the sequence alignment score of the two DNA sequences `TACGGGTAT` and `GGACGTACG`.

Show the filled dynamic programming matrix using +1 for a match, -1 for a mismatch and -1 for a gap penalty in a way similar to the slide sets.

(b) Repeat with the Smith-Waterman local alignment algorithm and the same scoring scheme.

3. **Substitution matrices:**

   (a) Given the BLOSUM-62 matrix (see sequence class notes or `http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/BLOSUM62`), find the score of the following alignment (assume this is the optimal alignment):
   `THISSEQ`
   `THATSEQ`

   (b) Repeat with the PAM-250 matrix. It can be found in Lesk, page 257, or here: `http://www.ncbi.nlm.nih.gov/IEB/ToolBox/C_DOC/lxr/source/data/PAM250`

4. **Multiple Sequence Alignment:** Extend the dynamic programming formula to 3 dimensions. What is the run time in this case? How many cases do we have to compare this time?

   Hint: this time the matrix is cubic since instead of a 2-dimensional matrix we need to run on a cube of $m \times n \times k$ where m,n, and k are the lengths of the three sequences. Every path goes from one vertex of the cube and traveling inside the cube to another vertex. Try to count how many such paths there can be.

Part 3 – Programming (optional)

You do not have to submit this part (you could, it won't count towards your grade but you may get helpful feedback) For this part you need the biopython module installed (https://biopython.org/docs/1.75/api/index.html) and in particular - BioBlast. In particular – look at the NCBIWWW module. Here is an extensive tutorial: http://biopython.org/DIST/docs/tutorial/Tutorial.pdf

- Write a python program that implements questions 1 and 2 using BioBlast. Using NCBIWWW it should not be difficult. Use the web run and do not try to install and run standalone BLAST! It will take a lot of space to install the databases locally. Try to have your program obtain the results in a human-readable format. By default the results are in xml, but you can change the format or use the parse tools from that library. **Notice:** The BLAST run may take a while (up to 30–60 seconds).
- If you feel more comfortable with R you can use R, but I can't recommend a specific package that does that. You are on your own here...