CBB752b15 Homework Assignment 1

(DUE DATE: 6 March 2015 Friday 11.59pm)

Choose to do either MCDB&MBB or CBB&CS homework, depending on your academic affiliation. **No late submissions will be accepted.**

File should be emailed to cbb752@gersteinlab.org

MDCB 752|MBB 752/753

Complete 2 of the first 3 problems and problem 4. Submit your file according to the format: netID_firstNameLastName_cbb752b15_assignment1. Please scan answers to problem 4, or turn in a hard copy.

1. Multiple sequence alignments (MSA) cannot be efficiently handled using purely dynamic programming. Choose one existing MSA software and describe how it implements MSA. (for example Muscle, clustalW, Kalign, MView, T-coffee...)

2. ChIP-seq is a common method to determine protein-DNA interaction on a genome-wise scale. The exact sites of binding must be inferred from sequence reads of the DNA that is purified along with the protein of interest. Describe an algorithm for determining protein-DNA binding sites from ChIP-seq data. See the following citation for a list of example algorithms: Wilbanks, EG, Facciotti, MT (2010). Evaluation of algorithm performance in ChIP-seq peak detection. PLoS ONE, 5, 7:e11471.

You may also choose an algorithm targeted to finding enriched regions in a different type of functional genomics experiment involving high-throughput sequencing.

(Please do not use PeakSeq, which will be discussed in section on Feb 14)

3. Machine learning approaches are becoming extremely useful in the analysis of genome-scale data, as reviewed in the following paper (which we will discuss in section on Feb 14): Yip, KY, Cheng, C, Gerstein, M (2013). Machine learning and genome annotation: a match meant to be?. Genome Biol., 14, 5:205. Choose one article that describes the application of supervised machine learning to genomics and answer the following:

- What are the researchers trying to predict/infer?
- What information in being used for the prediction? What is the logic behind using these data?
- What preprocessing steps are used to prepare the data for machine learning?
- What is the model the researchers use, and why did they select their particular method?
- How do the researchers evaluate their predictions? Were they effective? What biological insight was gained?

4. Align the following two sequences using the Smith-Waterman algorithm (local alignment), with the following scores:

Match: 2
Mismatch: 0
Gap: -1

In addition to filling out the alignment matrix, indicate the traceback and write out the final alignment.

		A	A	A	A	C	G	C	T	T
	0	0	0	0	0	0	0	0	0	0
T	0									
T	0									
T	0									
A	0									
A	0									
T	0									
C	0									
G	0									
C	0									

CBB & CSPC

Please zip up all the files to be submitted, with filename according to the format: netID_firstNameLastName_cbb752b15_assignment1.zip.

Choose one of the following programming languages: Perl, Python, C, C++, MATLAB or R for this programming assignment. Scripting must be done from scratch, without the use of any preexisting packages. In your ZIPPED email submission, include input file(s), source code, output file(s) and a short README file on how to execute your program.

The first programming task is to implement the Smith-Waterman local alignment algorithm for protein sequences.

Gap penalties: opening gap -2, extension gap = -1

Requirements:
The program should automatically read in the similarity matrix file called "blosum62.txt" and input sequences in "input.txt", where each line is a sequence. These 2 files can be found in **cbb752b15_assign1.zip**, which can be downloaded from the class website.

The output should contain a human-readable alignment such as the following:

```
T  C  W  A
   |     |
S  C  -  A
```

where | represents amino acid identity and - represents a sequence gap.

For each sequence pair, the output must include the completed scoring matrix (including the sequences themselves) in tab-delimited format (akin to the hand-drawn DP scoring matrix), best-scoring local alignment(s) and the score. (Just to be precise, the completed scoring matrix contains the best score in the alignment up to this point.) These will constitute 90% of your grade, with the remaining 10% coming from your programming style (e.g. clear comments). Also, clearly document how your script works (README.txt) in order for us to successfully run your script.

Programs that do not compile will get an immediate 0. To receive partial credit, please make sure your program is well-commented.