CBB752a12 Homework Assignment 1

(DUE DATE: 3 March 2014 11.59pm)

Choose to do either MCDB&MBB or CBB&CS homework, depending on your academic affiliation. **Please zip up all the files to be submitted, with filename according to the format: netID_firstNameLastName_cbb752a12_assignment1.zip. No late submissions will be accepted.**

This zipped file should be emailed to **cbb752@gersteinlab.org**.

MCDB & MBB (choose 3 of 4)

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.

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

3. Genome-wide association studies (GWAS) assume a notion of "common disease common variants". But those studies have reached a bottleneck and many studies have realized that rare variants might be more important. However, rare variants are hard to pick out.

What are the challenges facing rare variants identification? One way to tackle the problem is using exome sequencing. Describe exome sequencing and discuss how it could help.

4. 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?

CBB & CSPC

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 pre-existing 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 **cbb752a12_assign1.zip**, which can be downloaded from the class wiki.

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.