CBB752 Quiz 1 Prep.

- *Databases*
  - Table joining in relational databases
  - Definitions of a key, a primary key, and a foreign key
  - The tradeoff between normalization and speed/efficiency

- *Genomics*
  - How are genomic sequencing data read out to make relevant biological outcomes?
  - List 3 metrics to measure the quality of sequencing technologies
  - Definitions of read coverage and deep sequencing
  - List 3 or more types of omics data used in functional genomics analyses
  - What is the main advantage of Third Generation Sequencing technologies over NGS?

- *Proteomics*
    - Definition of the field of Proteomics
    - Basic understanding of Mass Spectrometry, basic idea behind mass spectrum
    - How can you use Immunoprecipitation to detect multiple proteins using a spectrometer that can identify one peptide?
    - Limitations of MS and alternate approaches to quantify proteins
    - Listing 3 types of protein-protein interactions
    - Listing 3 methods of identifying protein structures

- *Alignment / Dynamic Programming*
    - The concept of optimal substructure in Dynamic Programming
    - Smith-Waterman and Needleman-Wunsch
        - How to apply algorithms on sequences: matrix calculation and alignment traceback
        - How similar are the algorithms? What is(are) the main difference(s) between them?

- *Multiple Sequence Alignment*
  - What is a multiple alignment?
  - How to convert multiple alignment on inspection into a simple profile?
  - How to convert it into a motif?
  - Sorting the following algorithms in increasing order of execution time (speed): BWA, Blast, FASTA, Smith-Waterman, PSI-Blast, HMMs
  - Similarity matrices and their relationship to profiles

- *Fast Alignment*
  - Hashing, hashtable, and how do they speed up alignments?
  - Time Complexity of alignment algorithms we discussed in class
  - Why are FastA and BLAST preferred to dynamic programming approaches to searching sequence databases?

- *SV/SNVs*
    - Approx. number of SNPs, indels, and SVs in a typical individual in 1000 Genomes
    - Ratio of rare variants in a typical human genome
    - Calling of SNVs from a read stack
    - A sense of how the read mapping changes for a split-read or paired-end calculation
    - Genome remodeling: duplication and retrotransposition

- *HMMs*
    - The goal and output of Viterbi algorithm
    - Difference between transition and emission probabilities in a HMM

- *Chip-Seq and RNA-Seq*
  - Definition of Chip-Seq
  - How does one do an aggregation plot for a ChIPseq factor around the TSS?
  - Describe roughly how peak calling is accomplished.
  - Describe allelic expression or eQTL, how does that work, and what are the differences that one is looking for?
  - When doing a simple gene expression clustering, how does one do a simple gene expression clustering, and interpret the resulting clusters in terms of modules?
  - Describe in simple terms how to convert a set of reads to gene expression measurements
- *Unsupervised Mining*
  - What is the difference between supervised, unsupervised, and semi-supervised learning?
  - What is the fundamental difference between PCA or SVD?
  - In particular, if one has a matrix of gene expression or a matrix of ChIPseq signal profiles over the genome, describe the results of doing SVD on this matrix in terms of the various icon vectors.