CSE 182 Syllabus

1. The title " Biological Databases" is a misnomer. In fact, most Biological databases are simply structured as text files, but maintain a lot of complexity. This course is about understanding sources for Biological Data, and the tools used to query them. We therefore call it "Biological Data Analysis".

2. The course synthesizes material from other classes. In Math 186, you learned the basics of statistical analysis; in CSE 181, you learned the algorithmic principles used to build Bioinformatics tools; in BENG 183 explains how modern instruments generate biological data. In this class, you will learn how tools are constructed from these principles, and how they can be used to make biological discoveries.

3. The course is conceptually divided into two parts: Static Data Analysis and Dynamic Data Analysis.

   Static data: We use this term to refer to data that is determined once for an organism. For example, it could refer to a reference genomic sequence (minus variation), a list of genes, a list of proteins and so on. We learn to query large databases of these objects. Topics:

   (a) Pairwise sequence alignment (global and local alignment algorithms)
   (b) An overview of BLAST with alignments (duplicated from CSE 181).
   (c) Scoring matrices for DNA and Proteins (PAM, Blosum)
   (d) Measuring the 'significance of database hits' via E-values and P-values (Loose similarity to topics in Math 186)
   (e) Database filtering as a principle; Trade-offs in filtering; Fast keyword matching (Aho-Corasick algorithm) as a filtering strategy (Minor overlap with CSE 181).
   (f) Protein sequence/structure analysis using regular expressions.
   (g) Protein sequence/structure analysis using profiles.
   (h) Hidden Markov Models: Viterbi algorithm and parameter estimation.
   (i) Genomes and the encoding of proteins in genes. Introduction to the central dogma.
   (j) Basics of Eukaryotic gene structure
   (k) Discriminating Exons from Introns using HMMs
   (l) Gene finding using HMMs and Dynamic programming

   Dynamic data: This term refers to data from an organism that changes from experiment to experiment. This includes lists of active RNA (transcripts), lists of active proteins, variation in population samples, etc.

   (a) Overview of RNA-seq (duplicated in CSE 181)
   (b) Creating an abstract expression matrix from transcripts
   (c) Basics of Linear algebra
   (d) Supervised Classification (Clustering was covered in 181 and will not be covered again)
   (e) Perceptron algorithm for identifying a separating hyperplane
   (f) Fisher's Linear Discriminat Analysis for separating classes
   (g) Protein expression using Mass Spectrometry
   (h) Basics of Mass Spectrometry, Bottom up mass spectrometry, tandem mass spectrometry
   (i) Isotope analysis and charge determination
   (j) *de novo* peptide identification