Homework 1: Foundational Concepts and Personal Interest

Zackery Field, Section: vargas 002

January 27, 2014

1. Data Science

Data in and of itself does not provide any insight or knowledge. Instead, what is often sought is the information within a dataset. The process of extracting this information is not always straightforward. If we have a huge mass of text (the data) and we seek to find all of the phone numbers within that text (the information) it is fairly easy to use a tool like grep to extract that information. What if instead the dataset is a genome, and the information that we seek to extract is the location of the protein coding sequences in that genome? Data scienctists seek to extract information from data; always trying to make their process in keeping with the scientific method.

The general public often confuses data science with big data. The top few hits of a google search of "data science npr" are actually links to articles on big data, not data science. In business, data science often refers to a method of extracting information that is useful, but often not scientific. A more apt term for the 'data science' that businesses utilize is analytics.

2. Computational biology and bioinformatics

My interpreation of the bioinformatics wiki page is that bioinformatics is a subfield of data science. In particular, bioinformatics seeks to identify and classify the information contained within data collected from biological systems. Additionally, it is a subfield of database science as bioinformaticists are also concerned with the storage, retrieval, and organization of biological.

On the other hand, computational biology is concerned with the computational modeling of biological systems in order to develop some new insight into how these systems function. This blog post does a good job of differentating between these two fields, both involving computation and biology. A biologist would participate in the field of computational biology, using tools provided to them by engineers to uncover new knowledge. An engineer would create the tools that the biologist utilizes, this is the field of bioinformatics.

3. Ideal computational biology graduate program

Given the above definition of computational biology as being a scientific field that utilizes tools that come out of bioinformatics, the ideal program would not be solely computational. The program would require the use of some computational tools however. These tools would be necessary to extract some novel information about the biological system in question. Many courses in biology would be necessary in order to understand the complexities of the system being studied. In order to tweak the tools to meet the needs of a particular project, some programming experience would be necessary. This programming experience could come from a few CS courses, if the student has not already gained this experience elsewhere. Since computational biology is a scientific field, there is no need for the student's project to be practical. Any thesis topic that led to some new insight into a biological system, and used computational tools to develop that insight would be acceptable.

4. Lecture notes and commentary

Computational

1. K-means and hierarchical clustering
2. Supervised learning methods
3. Graph theory and network theory

Bioinformatics

1. Genome annotation pipelines
2. Pathway databases and prediction
3. Protein functional site prediction