Motivation and background

This course will focus on the computational analysis of biological systems beyond genome sequence, and cover topics in functional genomics, population genetics, epigenetics and proteomics. Computational methods studied for this type of analysis include: network and graph algorithms, combinatorial algorithms, supervised and unsupervised learning, large data/network visualization, statistical modeling and inference, probabilistic graphical models, sparse methods in data analysis, numerical optimization. These methods are complementary to those used for analyzing biological sequences (on which CMSC 423 has historically focused).

Course Description

An introduction to the fundamental concepts in the computational analysis of biological systems with applications to: functional genomics, population genetics, proteomics and epigenetics. Computational concepts covered: network and graph algorithms, combinatorial algorithms, machine learning, large data/network visualization, statistical modeling and inference, probabilistic graphical models, sparse methods in data analysis, numerical optimization. No prior knowledge of biology required.

Prerequisites
CMSC 330 and 351, or permission of department.

Textbooks:

There is currently no single textbook that covers all the topics presented in the course. However, we will mainly use the following as a core textbook:


This textbook will be supplemented by relevant material from other books and made available to the students as handouts.

Tentative Syllabus

- Introduction (1 week)
  - A molecular biology for computer scientists refresher.

- Functional Genomics (3 weeks)
  - Differential analysis of gene expression by probabilistic modeling and statistical inference
  - Supervised and unsupervised learning models for expression data classification and clustering
  - Deriving medical diagnosis and prognosis models from expression data using sparse methods in machine learning
  - Probabilistic graphical models of gene co-regulation

- Proteomics (3 weeks)
  - Machine learning based algorithms for identification of proteins in biological samples
  - Spatial data structures and range queries for quantitative proteomics data
  - Combinatorial and graph-based algorithms for identify post-translational modifications of proteins
  - Computational methods for assessing statistical significance in proteomics data

- Networks (3 weeks)
  - Experimental and computational methods for detecting molecular interactions within the cell
  - Gene function prediction via network clustering, clique-finding, information flow heuristics
  - Identification of functional interaction patterns using network alignment and subgraph isomorphism

- Genetics and epigenetics (3 weeks)
• Genotype-phenotype association analysis by probabilistic modeling and statistical inference
• Genome-Environment-Clinical data integration models
• Emerging topics (1 week, time permitting), e.g. sequencing technologies, nanotechnology.