CBB752 (Spr. '16) - Bioinformatics: Practical Application of Data Mining & Simulation

Course Description

Bioinformatics encompasses the analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. It represents a major practical application for modern techniques in data mining and simulation. Specific topics to be covered include sequence alignment, large-scale processing, next-generation sequencing data, comparative genomics, phylogenetics, biological database design, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, normalization of microarray data, mining of functional genomics data sets, and machine learning approaches for data integration.

Overall Flow of the Class:
(Module = Group of Lectures)

- Introduction
- Module on "the Data" (Genomic, Proteomic & Structural Data), introducing the main data sources (their properties, where you access, &c)
- Module on Databases & Data Science Issues (Knowledge Representation incl. Sem. Web & Privacy, Provenance & Standards)
- Module on Mining (Alignment & Variant Calling, Supervised & Unsupervised Approaches, Networks)
- Module on Cell Modeling
- Module on Molecular Modeling

Lectures:

- MW 1:00 - 2:15 PM, Bass 305 (plus some Fridays at same place and location)

Discussion Section:

- F 1:00 - 2:00 PM, Bass 405

Different headings for this class (4 variants)

- **CB&B752/CPSC752 - Grad. w/ programming**
  This graduate-level version of the course consists of lectures, programming assignments, and a final programming project.
- **MB&B452/MCDB452 - Undergrad.**
  This undergraduate version of the course consists of lectures, written problem sets, and a final (semi-computational section and a literature survey) project.
- **MB&B752/MCDB752 - Grad. w/o programming**

This graduate-level version of the course consists of lectures, written problem sets, and a final (semi-computational section and a literature survey) project. Unlike CBB752, there is no programming required.

- **MB&B 753a3/MB&B 754a4 - Modules**
  For graduate students the course can be broken up into two "modules" (each counting 0.5 credit towards MB&B course requirement):
  753 - Bioinformatics: Practical Application of Data Mining (1st half of term)
  754 - Bioinformatics: Practical Application of Simulation (2nd half of term)
  Each module consists of lectures, written problem sets, and a final, graduate level written project that is half the length of the full course's final project.
- **Auditing**
  It is allowed, but we'd prefer if you would register for the class.

Prerequisites

The course is keyed towards CBB graduate students as well as advanced MB&B undergraduates and graduate students wishing to learn about types of large-scale quantitative analyses that whole-genome sequencing will make possible. It would also be suitable for students from other fields such as computer science or physics wanting to learn about an important new biological application for computation.

Students should have:

1. A basic knowledge of biochemistry and molecular biology.
2. A knowledge of basic quantitative concepts, such as single variable calculus, basic probability and statistics, and basic programming skills.

These can be fulfilled by: MBB 200 and Mathematics 115 or permission of the instructor.

Class Requirements

Discussion Section / Readings

Papers will be assigned throughout the course. These papers will be presented and discussed in weekly 60-minute sections with the TFs. A brief summary (a half-page per article) should be submitted at the beginning of the discussion session.

In-class tests: Midterm & Quiz

- There will be a midterm covering the 1st half of the course.
- There will be a quiz covering 2nd half of the course comprising SIMPLE questions that you should be able to answer from the lectures plus the main readings.

For references, please refer the previous quizzes and answer keys from Fall 2012

Programming Assignments (Req'd for CBB and CS students)

- There will be **FOUR** homework assignments. We will try to promote the idea of reproducible research and using version control system, specifically GitHub, in facilitating the process of homework submission.
- For Homework 1, you will be given an opportunity to get familiar with GitHub and programming with version control. You can choose to either submit your homework through GitHub **OR** through email. However, for the later assignments, you will only be able to submit homework through GitHub.
- For the programming assignments, you can use either R or Python. However, if you would like to use other programming languages, please contact the TAs and request for a permission.
- For detailed instruction and information, please refer the Start up for Homework 1 & Homework Submission Instructions.

Non-programming Assignments (For MB&B and MCDB students)

- There will be equivalent FOUR homework assignments for MB&B and MCDB students without a programming background. Programming part will be replaced with assignments involving the use of web-based tools or essay questions.

Grading Policy

The following grade distribution will be used for all students (CBB/CPSC/MBB/MCDB):

Category	% of Total Grade
Midterm	20%
Quiz	10%
Discussion Section	10%
Homeworks	40%
Final Project	20%

Instructors

Use **cbb752(at)gersteinlab.org** for general correspondence and questions.

Instructor-in-Charge

Name	Office	Email
Mark Gerstein	Bass 432A	contact.gerstein.info

Guest Instructors

Name	Office	Email
Corey O'Hern	Mason Laboratory	corey.ohern (at) yale.edu
Jesse Rinehart	West Campus	jesse.rinehart (at) yale.edu
Matthew Simon	West Campus	matthew.simon (at) yale.edu
Kei Cheung	300 George St	kei.cheung (at) yale.edu
Steven Kleinstein	300 George St	steven.kleinstein (at) yale.edu
Dov Greenbaum	266 Whitney Ave	dov.greenbaum (at) yale.edu

Consultation is available UPON REQUEST or according to times stipulated by the individual instructors. Prof. Gerstein's office office hours will usually be right after some the classes.

Teaching Fellows

Name	Office	Email
Donghoon Lee	Bass 437	donghoon.lee (at) yale.edu
Xiu Huang	300 George Suite 503	xiu.huang (at) yale.edu

Schedules

Time	Teacher(s)	Topic
01/20	MG	Introduction
01/22	MDS	The Data 1 - Genomics
01/25	MDS	The Data 2 - Genomics
01/27	JR	The Data 3 - Proteomics

02/01	JR	The Data 4 - Proteomics
02/03	KC	Knowledge Representation & Databases
02/08	DG	Issues in Data Science (Privacy, Standard, Provenance)
02/10	MG	MINING 1 - Alignment (seq. comparison & multiple-seq. alignment)
02/15		President's day - No Class (Buffer)
02/17	MG/XH	MINING 2 - Variant Calling
02/22	MG	MINING 3 - Data Integration & Machine Learning
02/24	MG	MINING 3 - Unsupervised Mining (focusing on spectral methods i.e., SVD)
02/29	MG/DL	MINING 4 - Supervised Mining (focusing on Trees, SVMs, NNs)
03/02	MG	MINING 5 - Analysis of Network Topology
03/07	MG	MINING 6 - Network Prediction
03/09		(Buffer)
	--	(Spring recess)
03/28	MG	Mid-term Exam on 1st Half
03/30	SK	Cell/Immune Modeling I
04/04	SK	Cell/Immune Modeling II
04/06	SK	Cell/Immune Modeling III
04/11	CO	Protein Simulation I
04/13	CO	Protein Simulation II
04/18	CO	Protein Simulation III
04/20	CO	Markov Models I

04/25	CO	Markov Models II
04/27	CO	Markov Models III / Protein Aggregation
	--	(Classes end)