CBB/CPSC/MBB 752, MBB 452 Course Syllabus

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:

- TBD

Different headings for this class (4 variants)

- **CB&B752/CPSC752 - Grad. w/ programming**

  This graduate-level version of the course consists of lectures, in-class tests, programming assignments, and a final programming project.
- **MB&B452/MCDB452 - Undergrad.**

  This undergraduate version of the course consists of lectures, in-class tests, 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, in-class tests, 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 - Biomedical Data Science: Mining (1st half of term)

754 - Biomedical Data Science: Modeling (2nd half of term)
Each module consists of lectures, in-class tests, written problem sets, and a final, graduate level written project that is half the length of the full course's final project.

- **Auditing**

This is allowed. We would strongly 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 analysis 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 including assignment 0. 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 (including assignment 0) 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.

Class Schedule:

#	Day	Date		Topic
	M	01-16	--	(MLK)
	T	01-17	--	(Spring-term classes begin)
Data Mining (1st Half)				
1	W	01-18	MG	Introduction
2	F	01-20	MDS	The Data 1 - Genomics
3	M	01-23	MDS	The Data 2 - Genomics
4	W	01-25	JR	The Data 3 - Proteomics
5	M	01-30	JR	The Data 4 - Proteomics
6	W	02-01	KC	Knowledge Representation & Databases
7	M	02-06	MG+guest	Introduction to personal genomes
8	W	02-08	MG+guest	MINING 1 - Alignment (seq. comparison & multiple-seq. alignment)
9	M	02-13	MG	MINING 2 - Variant Calling (including a focused section on SVs)
10	W	02-15	MG	MINING 3 - Unsupervised Mining (focusing on spectral methods, eg SVD)
11	M	02-20	MG	MINING 4 - Supervised Mining (focusing on DTrees, SVMs, NNs)
12	W	02-22	MG	MINING 5 - Mining continued (deep learning)
13	M	02-27	MG	Mid-term Exam on 1st Half
14	W	03-01	MG	MINING 6 - Analysis of Network Topology
15	M	03-06	MG	MINING 7 - Network Prediction
	W	03-08		OPEN DAY FOR SNOW

			--	(Spring recess)
Simulation (2nd Half)				
16	M	03-27	SK	Cell/Immune Modeling I
17	W	03-29	SK	Cell/Immune Modeling II
18	M	04-03	SK	Cell/Immune Modeling III
19	W	04-05	CO	Protein Simulation I
20	M	04-10	CO	Protein Simulation II
21	W	04-12	CO	Protein Simulation III
22	M	04-17	CO	Markov Models I
23	W	04-19	CO	Markov Models II
24	M	04-24	CO	Markov Models III / Protein Aggregation
25	W	04-26	CO	Quiz

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

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 (TA)

Name	**Office**	**Email**
Mengting Gu	**Bass 437**	**mengting.gu (at) yale.edu**
Paul Muir	**Bass 437**	**paul.muir (at) yale.edu**

For general correspondence and questions, please contact us at: cbb752 (at) gersteinlab.org

Grading Policy

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

Category	% of Total Grade
Midterm	15%
Quiz	15%
Discussion Section	10%
Homeworks	20%
Final Project	40%

Relevant Yale College Regulations

Students may have questions concerning end-of-term matters. Links to further information about these regulations can be found below:

- http://catalog.yale.edu/ycps/academic-regulations/reading-period-final-examination-period/
- http://catalog.yale.edu/ycps/academic-regulations/completion-of-course-work/
- Brief presentation on how to cite correctly :
  http://archive.gersteinlab.org/mark/out/log/2012/06.12/cbb752b12/cbb752_cite.ppt

Plagiarism

Below is a message from the Dean of Yale College about citing your references and sources of information and plagiarism:

" *You need to cite all sources used for papers, including drafts of papers, and repeat the reference each time you use the source in your written work. You need to place quotation marks around any cited or cut-and-pasted materials, IN ADDITION TO footnoting or otherwise marking the source. If you do not quote directly – that is, if you paraphrase – you still need to mark your source each time you use borrowed material. Otherwise you have plagiarized. It is also advisable that you list all sources consulted for the draft or paper in the closing materials, such as a bibliography or roster of sources consulted.*
You may not submit the same paper, or substantially the same paper, in more than one course. If topics for two courses coincide, you need written permission from both instructors before either combining work on two papers or revising an earlier paper for submission to a new course.
It is the policy of Yale College that all cases of academic dishonesty be reported to the chair of the Executive Committee.... "

" *Academic integrity is a core institutional value at Yale. It means, among other things, truth in presentation, diligence and precision in citing*
works and ideas we have used, and acknowledging our collaborations with others. In view of our commitment to maintaining the highest standards of academic integrity, the Graduate School Code of Conduct specifically prohibits the following forms of behavior: cheating on examinations, problem sets and all other forms of assessment; falsification and/or fabrication of data; plagiarism, that is, the failure in a dissertation, essay or other written exercise to acknowledge ideas, research, or language taken from others; and multiple submission of the same work without obtaining explicit written permission from both instructors before the material is submitted. Students found guilty of violations of academic integrity are subject to one or more of the following penalties: written reprimand, probation, suspension (noted on a student's transcript) or dismissal (noted on a student's transcript). "