Biomedical Data Science: Mining and Modeling

CB&B 752, CPSC 752, MB&B 752, MB&B 753, MB&B 754, MCDB 752, MB&B 452, MCDB 452, S&DS 352
Spring 2022

Instructor-in-Charge

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

Guest Instructors

Name	Abbr	Office	Email
Corey O'Hern	CO	Mason 203	corey.ohern@yale.edu
Jesse Rinehart	JR	West Campus, 3rd Flr. Ste.	jesse.rinehart@yale.edu
Matthew Simon	MDS	West Campus, Ste MIC312	matthew.simon@yale.edu
Kei-Hoi Cheung	KC	300 George St.	kei.cheung@yale.edu
Martin Renqiang Min	RM		renqiang@gmail.com
Carl Zimmer	CZ	Bass 3rd Flr.	carl.zimmer@yale.edu

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

Teaching Fellows (TFs)

Name	Abbr	Office	Email
Eric Ni	EN	Bass 437	eric.ni@yale.edu
Tianxiao Li	TXL	Bass 437	tianxiao.li@yale.edu

Lectures

MW 1:00 - 2:15 PM, BASS305

Discussion Section

Section 1	Fri 10:00-11:00 AM	BASS405
Section 2	Fri 1:00-2:00 PM	BASS405

Course Description

Rapid developments in bio- and information- technology and are changing the way that biomedical scientists interact with data. Traditionally, data were the end result of laborious

experimentation, and their interpretation mostly involved careful thought and background knowledge. Today, data are increasingly generated much earlier in the scientific workflow and are much larger in scale. Also, before the data can be interpreted, extensive computational processing is often necessary. Thus, the data deluge in biomedicine now requires mining and modeling on a large scale – i.e. biomedical data science.

This course aims to equip students with some of the concepts and skills relevant to biomedical data science, with an emphasis on bioinformatics, a sub-discipline of this broader field, through examples of mining and modeling of genomic and proteomic data. More specifically, 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, mining of functional genomics data sets, and machine learning approaches for data integration.

Course Format

Every week there will be two lectures and one discussion section (except for holidays and Yale break days, see Syllabus for details).

Enrollment Cap, Selection Process, and Notification

There is no enrollment cap for the course. There will be 4 different variants for this class, see details below:

CB&B 752 / CPSC 752 - Grad. with programming
This graduate-level version of the course consists of lectures, in-class tests, discussion section, programming assignments, and a final programming project.

MB&B 752 / MCDB 752 - Grad. without programming
This graduate-level version of the course consists of lectures, in-class tests, discussion section, written problem sets, and a final (semi-computational section and a literature survey) project. Unlike CBB752, there is no programming required.

MB&B 753b3 / MB&B 754b4 - 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.

MB&B 452 / MCDB 452 / S&DS 352 - Undergrad.
This undergraduate version of the course consists of lectures, in-class tests, discussion section, written problem sets, and a final (semi-computational section and a literature survey) project.

The programming assignments from CBB752 can be substituted for the written work by permission of instructor.

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 undergraduates and graduate students wishing to learn about types of large-scale quantitative analysis that whole-genome sequencing and forms of large-scale biological data will make possible. It would also be suitable for students from other fields such as computer science, statistics or physics wanting to learn about an important new biological application for computation.

Students should have:

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

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

Class Requirements and Assignments

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 Quizzes

- There will be a quiz covering the 1st half of the course.
- There will be a quiz covering the 2nd half of the course.

Quizzes will comprise simple questions that you should be able to answer from the lectures plus the main readings.

Programming Assignments (Required for CBB and CS grad. students)
There will be two 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.

Non-programming Assignments
There will be equivalent two assignments, particularly for MB&B and MCDB students without a programming background. The programming part will be replaced with assignments involving the use of web-based tools or essay questions.

General Course Policy

First Meeting

The first lecture will be held on Wed. Jan 19, 2022.

Grading Policy
We expect that this year the weighting scheme will be to a first approximation:

Category	**% of Total Grade**
Midterm Quiz	15%
Final Quiz	15%
Discussion Section	20%
Homeworks	20%
Final Project	30%

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

- Reading Period and Final Examination Period
http://catalog.yale.edu/ycps/academic-regulations/reading-period-final-examination-period/
- Completion of Course Work
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

Useful Background Books & Websites
If you would like to get more background for the course, here are some resources. All of the following textbooks you can access online through Yale's e-library. Many of them cover the same material, but these should cover the core background that you need:
Biology related

- Essentials of Molecular Biology by Malathi V
- important: chapters 1-5, 7, 11
- recommended: everything else
- Biochemistry: Essential Concepts by Hardin, Charles
- chapters 1-6

Programming related
For the class you can choose to do assignments in either R or python. Here are some resources for both:

- The Python Tutorial - Official guide and documentation for Python3.
- Learn Python with Google Colab - An interactive guide to python. Useful if you are familiar with basic programming, but not python language.
- Learning Python: Powerful Object-Oriented Programming (5th edition) by Mark Lutz
- The R book (2nd edition) by Michael J. Crawley
- R for Data Science by Hadley Wickam

Accessibility Statement
We are committed to creating a course that is inclusive in its design. If you encounter barriers, please let us know immediately so that we can determine if there is a design adjustment that can

be made or if an accommodation might be needed to overcome the limitations of the design. We are always happy to consider creative solutions as long as they do not compromise the intent of the assessment or learning activity. You are also welcome to contact Student Accessibility Services to begin this conversation or to establish accommodations for this or other courses. We welcome feedback that will assist us in improving the usability and experience for all students.

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)."

Required Course Materials
There is no required course material.

Recording Policy
We will follow the default FAS policy on recording where the instructor's lectures will be recorded, and student contributions in seminars and sections will not be recorded.

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, etc)

- 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

Class Schedule

#	Day	Date		Topic
	T	1/18	--	*YALE* Spring term classes begin, 8.20 a.m.
1st Half				
1	W	1/19	MG	Introduction
2	M	1/24	MDS	DATA 1 - Genomics I
3	W	1/26	MDS	DATA 2 - Genomics II
4	M	1/31	JR	DATA 3 - Proteomics I
5	W	2/2	JR	DATA 4 - Proteomics II
6	M	2/7	KC	DATA 5 - Knowledge Representation & Databases
7	W	2/9	MG	MINING 1 - Personal Genomes Intro. (with an individual's perspective)
8	M	2/14	MG	MINING 2 - Seq. Comparison + Multi-seq Alignment
9	W	2/16	MG	MINING 3 - Fast Alignment + Variant Calling (incl. a focused section on SVs)
10	M	2/21	MG+TF	Quiz on 1st Half
11	W	2/23	MG	MINING 4 - Basic Multi-Omics + Supervised Mining #1
12	M	2/28	MG	MINING 5 - Supervised Mining #1 + Unsupervised Mining #1
13	W	3/2	MG	MINING 6 - Unsupervised Mining #2 + Network Analysis
14	M	3/7	MG+TF	TF short lecture + MG network
15	W	3/9	MG+TF	TF short lecture
--	F	3/11	--	Spring break begins
2nd Half				
16	M	3/28	RM	Deep Learning I
17	W	3/30	RM	Deep Learning II
18	M	4/4	RM	Deep Learning III
19	W	4/6	CO	Protein Simulation I
20	M	4/11	CO	Protein Simulation II
21	W	4/13	CO	Protein Simulation III
22	M	4/18	CO	Markov Models I
23	W	4/20	CO	Markov Models II
24	M	4/25	MG+TF	Quiz on 2nd Half
25	W	4/27	MG	Final Presentations
	F	4/29	--	*YALE* Classes end; Reading period begins
	F	5/6	--	*YALE* Final examinations begin
	W	5/11	--	*YALE* Final examinations end

Discussion Sections

The standard discussion section involves student presentations on 1 or 2 papers. Some discussion sections will involve hands-on skill-building demos taught by the teaching fellows, such as the use of R, High Performance Computing, and GitHub. The exact format will be determined based on the size of the class. However, we generally require the following:

- Each week, students should read the assigned papers and write at a minimum of 200 words (half a page, single-spaced, per paper) summaries of each paper (two articles = approx. 1 page). We would like to encourage electronic submission, via Canvas. For those who have trouble accessing canvas, we will also accept submission over email to cbb752 (at) gersteinlab.org BEFORE the start of each section.
- Each student will give one presentation about a selected paper (approx. 20 min) in one of the sessions.
- Students will be graded on a combination of the written summary, presentation, and participation in discussions.
- If you are presenting, you are exempt from writing a summary.
- Please notify TFs in advance if you cannot come to the discussion session. Student can miss up to one discussion section without a penalty.

Section Readings

Session 0

- How to (seriously) read a scientific paper, on your own. [Link]

Session 1, 1/28, BASS405 for both sessions

Next Generation Sequencing and database

- Goodwin S. et al. "Coming of age: ten years of next-generation sequencing technologies" Nature Reviews Genetics. 17 (2016) [PDF]
- Wheeler DA et al. "The complete genome of an individual by massively parallel DNA sequencing," Nature. 452:872-876 (2008) [PDF]

Session 2, 2/4, BASS405 for both sessions

Proteomics

- A draft map of the human proteome. Nature 509,575–581 (29 May 2014) [PDF]
- Mass-spectrometry-based draft of the human proteome. Nature 509, 582–587 (29 May 2014) [PDF]

Session 3, 2/11, BASS405 for both sessions

Debate I - Gencode vs Salzberg et al. debate

- (Main paper) Salzberg et al. CHESS paper using GTEx [PDF]
- (Main paper) GENCODE's rebuttal [PDF]
- (Optional) New human gene tally reignites debate [News Article]
- (Optional) Why most published research finding are false [PDF]

Session 4, 2/18, BASS405 for both sessions

Help session on Quiz 1 - TFs prepare materials on SW alignments and Q&A session

Sequence and Alignments

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990) Basic local alignment search tool. Journal of Molecular Biology, 215(3):403-10. PMID: 2231712. [PDF]
- T.F. Smith and M.S. Waterman. (1981) Identification of common molecular subsequences. Journal of Molecular Biology,147(1): 195-7. PMID: 7265238. [PDF]

Session 5, 2/25, BASS405 for both sessions
Debate II - Phylogenetics
- Jarvis ED et al. (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science, 346(6215), 1320-1331. [PDF]
- Mitchel KJ, Cooper A, Philips MJ (2015) Comment on "Whole-genome analyses resolve early branches in the tree of life of modern birds." Science, 349(6255) 1460 [PDF]

Session 6, 3/4, BASS405 for both sessions
Immune system modelling and dynamics
- Perelson AS. Modelling viral and immune system dynamics. Nat Rev Immunol. 2002 Jan;2(1):28-36. [PDF]
- Modeling the Spread of Ebola [PDF]

Session 7, 4/1, BASS405 for both sessions
Deep learning for genomics
- A primer on deep learning in genomics [PDF]
- Deep learning for biology [PDF]

Session 8, 4/8, BASS405 for both sessions
Debate III - Cancer incidence
- Debate reignites over the contributions of 'bad luck' mutations to cancer [Link]
- The simple math that explains why you may (or may not) get cancer [Link]

Session 9, 4/15, BASS405 for both sessions
Protein structure and biophysics
- Zhou, AQ, O'Hern, CS, Regan, L (2011). Revisiting the Ramachandran plot from a new angle. Protein Sci., 20, 7:1166-71 [PDF]
- Dill KA, Ozkan SB, Shell MS, Weikl TR. (2008) The Protein Folding Problem. Annu Rev Biophys,9, 37:289-316. PMID: 2443096. [PDF]
- Bowman GR, Beauchamp KA, Boxer G, Pande VS. "Progress and challenges in the automated construction of Markov state models for full protein systems," J. Chem. Phys. 131 (2009) 124101 [PDF]

Session 10, 4/22, BASS405 for both sessions
Help session on quiz 2 / final project