Biomedical Data Science: Mining and Modeling

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 - ie 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.

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). This module also includes discussion of databases and knowledge representation issues.
- Module on Mining (Alignment & variant calling necessary for personal genomics; Basic multi-omics calculations; Supervised & unsupervised mining approaches towards multi-omic data; Networks)
- Module on Cell Modeling
- Module on Molecular Modeling

Lectures

- MW 1:00 - 2:15 PM, virtual using Zoom

Discussion Section

- F 9:30-10:30 AM or F 1:00-2:00 PM, virtual using Zoom

Different headings for this class (4 variants)

- CBB 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 CB752 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

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: Quiz

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

For references, please refer the previous Quiz Archive

Programming Assignments (Req'd 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 homework 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.

Pages from previous years

- 2021 Spring is the 24th time Bioinformatics has been taught at Yale. Pages for the 23 previous iterations of the class are available. Look at how things evolve!
- 2020 Spring
- 2019 Spring
- 2018 Spring
- 2017 Spring
- 2016 Spring
- 2015 Spring
- 2014 Spring
- 2012 Fall
- 2012 Spring
- 2011 Spring
- 2010 Spring
- 2009 and earlier (12 years of classes, staring in '98) (Note the pre-2010 course was Genomics & Bioinformatics; after 2010, the course contains all of the "Bioinformatics" of previous years and then more (!) with less "Genomics".)

Class data dump

- Syllabus and class info dump in single PDF file: PDF v1 or PDF v2
- Class poster: pdf

Syllabus

CBB752-spg-21-schedule-cbb752b21 : 2021 Public

#	Day	Date		Topic	Public Comment	URL	URL 2
	M	2/1	--	*YALE* Spring term classes begin, 8.20 a.m.			
				1st Half			
1	M	2/1	MG	Introduction			
2	W	2/3	MDS	DATA 1 - Genomics			
3	M	2/8	MDS	DATA 2 - Genomics			
4	W	2/10	JR	DATA 3 - Proteomics I			
5	M	2/15	JR	DATA 4 - Proteomics II			
6	W	2/17	KC	DATA 5 - Knowledge Representation & Databases			
--	M	2/22	--	*YALE* First break day			
7	W	2/24	MG	MINING 1 - Personal Genomes Intro. (with an individual's perspective)			
8	M	3/1	MG	MINING 2 - Seq. Comparison + Multi-seq Alignment + Fast Alignment			
9	W	3/3	MG	MINING 3 - Variant Calling (incl. a focused section on SVs) + Basic Multi-omics			
10	M	3/8	MG	MINING 4 - Basic Multi-Omics + Supervised Mining			
11	W	3/10	MG+TF	Quiz on 1st Half			
12	M	3/15	MG	MINING 5 - Unsupervised Mining #1 (SVD & Spectral Methods)			
13	W	3/17	MG	MINING 6 - Unsupervised Mining #2 (Extensions & Applications) & Network Analysis			
14	M	3/22	DG	Privacy			
--	W	3/24	--	*YALE* Third break day			
15	M	3/29	MG	TF/snow day			
				2nd Half			
16	W	3/31	RM	Deep Learning I			
17	M	4/5	RM	Deep Learning II			
18	W	4/7	RM	Deep Learning III			
19	M	4/12	MG+TF	TF			
20	W	4/14	CO	Protein Simulation I			
21	M	4/19	CO	Protein Simulation II			
22	W	4/21	CO	Protein Simulation III			
23	M	4/26	CO	Markov Models I			
24	W	4/28	CO	Markov Models II			
25	M	5/3	MG+TF	Quiz on 2nd Half			
26	W	5/5	MG	Final Presentations			
	F	5/7	--	*YALE* Classes end; Reading period begins			
	Th	5/13	--	*YALE* Final examinations begin			
	W	5/19	--	*YALE* Final examinations end			
				MG Lecture Slide Pack			
			#	**Topic**	**PDF**	**PPT**	
			1	Introduction to Biomedical Data Science			
			2	Introduction to Personal Genomes			
			3	Sequence Comparison			
			4	Multiple Sequence Comparison			
			5	Fast Alignment			
			6	Variant Identification, Focusing on SVs			
			7	Basic Multi-omics (pipeline processing)			
			8	Supervised Datamining			
			9	Unsupervised Datamining			
			10	Network Topology			
			11	Network Prediction			
			12	Hi-C Analysis			
			13	Privacy vs Sharing			
			14	Transition from Mining to Modeling			

2021 Public

Biomedical Data Science: Mining and Modeling

Contact Info

For general correspondence and questions, please contact us at:

cbb752 (at) gersteinlab.org

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 Laboratory	corey.ohern (at) yale.edu
Jesse Rinehart	JR	West Campus	jesse.rinehart (at) yale.edu
Matthew Simon	MDS	West Campus	matthew.simon (at) yale.edu
Kei Cheung	KC	300 George St	kei.cheung (at) yale.edu
Martin Renqiang Min	RM		renqiang (at) gmail.com
Dov Greenbaum	DG		dov.greenbaum (at) gmail.com
Carl Zimmer	CZ	266 Whitney Ave	carl.zimmer (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 (TF)

Name	Abbr	Office	Email
Jiahao Gao	JG	Bass 437	jiahao.gao (at) yale.edu
Tianxiao Li	TXL	Bass 437	tianxiao.li (at) yale.edu

General Course Policy

First Meeting

The first lecture will be held on Mon. Feb 1st, 2021.

Grading Policy

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

Category	% of Total Grade
Quiz 1	15%
Quiz 2	15%
Discussion Section	20%
Homeworks	20%
Final Project	30%

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.

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). "

Also, it might be of interest to people, to look at this recent article regarding academic dishonesty.

Weekly Discussion Sections & Readings

Time and Location

Session	Time	Location	Note
Section 1	9:30-10:30 AM	Zoom	
Section 2	1:00-2:00 PM	Zoom	

Format

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**.

For write-ups and presentation, think about the following:

- What was missing in the field? (introduction/background)
- What were the questions the paper aim to address? (hypothesis)

- What they did and what was the breakthrough? (method/results)
- Conclusion and future direction (discussion/conclusion)

Section Readings

Reading assignments for discussion sessions are listed below:

(Optional) Suggested Reading for Week 1

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

Session 1, 2/5, Zoom (First Discussion Sessions)

Topic

- Next-Gen Sequencing and database

Reading Assignment

- 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, 2/12, Zoom

Topic

- Proteomics

Reading Assignment

- 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, 2/19, Zoom

Topic

- Debate I

Reading Assignment

- 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, 2/26, Zoom

Topic

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

Reading Assignment

- 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, 3/5, Zoom

Topic

- Debate II - Phylogenetics

Reading Assignment

- 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, 3/12, Zoom

Topic

Debate III - Cancer incidence

Reading Assignment

- 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 7, 3/19, Zoom

Topic

- Deep learning for genomics

Reading Assignment

- A primer on deep learning in genomics [PDF]
- Deep learning for biology [PDF]

Session 8, 3/26, Zoom

Topic

- Immune system modelling and dynamics

Reading Assignment

- 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 9, 4/2, Zoom

Topic

- Protein structure and biophysics

Reading Assignment

- 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, 4/9, Zoom

Topic

- Help session on HW 2 / final project

Reading Assignment

- (no reading assignment this week)

Quiz Archive

- Spring 2020: Main Page, Quiz 1 key, HW1, HW2
- Spring 2019: Main Page, Quiz 1 key, Quiz 2, HW1, HW2
- Spring 2018: Main Page, Quiz 1, Quiz 1 Key, Quiz 2, Quiz 2 Key
- Spring 2017: Main Page, HW 1, HW 2, Midterm, Final
- Spring 2016: Main Page, Midterm, Midterm Key, Final Quiz, Final Quiz Key
- Spring 2015: Main Page
- Spring 2014: Main Page
- Fall 2012: Main Page, Quiz 1-4
- Spring 2012: Main Page, Quiz 1-4
- Spring 2011: Main Page, Quiz 1
- Spring 2010: Main Page, Quiz 1, Quiz 2, Quiz 3+4
- Spring 2009: Main Page, Quiz 1, Quiz 2
- Spring 2008: Main Page
- Fall 2006: Main Page, Quiz 1, Quiz 2
- Fall 2005: Main Page, Quiz 1, Quiz 2
- Spring 2005: Main Page, Quiz 1, Quiz 2
- Fall 2003: Main Page, Quiz 1, Quiz 1 Key, Quiz 2, Quiz 2 Chart, Quiz 2 Key
- Fall 2002: Main Page, Quiz 1, Quiz 1 Key, Quiz 2, Quiz 2 Key
- Fall 2001: Main Page, Quiz 1, Quiz 1 Key, Quiz 2
- Fall 2000: Main Page
- Fall 1999: Main Page, Quiz 2 with Key
- Spring 1999: Main Page
- Spring 1998: Main Page
