CBB752 - 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.

Timing & Location

Class: Meeting from 1:00-2:15 pm on Monday and Wednesday, in Bass 305. (First meeting will be on Mon. 12 Jan 2015, immediately followed by an office hour for Prof Gerstein, for those with individual questions about the class)

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
This 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.

We realize that students with diverse backgrounds will be taking the course and will be willing to recommend supplementary reading and/or MOOCs to help with the background to specific topics.

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.

Bioinformatics quizzes

There will be four short quizzes (25 minutes) in class comprising SIMPLE questions that you should be able to answer from the lectures plus the main readings.

Answer keys to Quizzes 1-4 in the fall of 2012 can be found here

Programming Assignments (CBB and CS) and Programming issues

There will be several short programming assignments required for CBB and CS students taking this course. Acceptable languages and submission requirements will be discussed prior to the first assignment. These assignments are NOT required for students not taking the CBB or CS sections of the course.

These are the programming languages that we permit in the programming assignments and final project: Perl, Python, C, C++, MATLAB and R. If you really feel more comfortable with other languages, please email the TFs to discuss. Also, packages such as BioPerl and BioPython are not allowed in the assignments and final project. If in doubt, please consult the TFs.

We recommend the use of PERL for most of the programming. A useful resource is the following book: Programming Perl, 3rd Edition in the O' Reilly series, by Larry Wall, Tom Christiansen, Jon Orwan . The Yale Library has also older editions, which would work too. We would also recommend the following online resources: http://www.perlmonks.org and http://stackoverflow.com/ . Otherwise, Google is your best friend.

Pages from previous years

2015 is the 18th time Bioinformatics has been taught at Yale. Pages for the 16 previous iterations of the class are available. Look at how things evolve! 2014 spring, 2013 fall, 2012 spring, 2011 spring, 2010, 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".)

Assignments

Assignments #1

Posted Oct 14, 2014, 5:47 AM by Yao Fu

Showing posts **1 - 1** of **1**. View more »

Homework

Name	Due Date	Description

Showing **0** items from page Final Project sorted by Due Date, edit time. View more »

Materials

Showing **0** files from page Section Readings.

CBB752 - Bioinformatics: Practical Application of Data Mining & Simulation

Syllabus

cbb752b15-schedule

#	Day	Date		Topic	
				Data Mining (1st Half)	
1	M	1/12/2015	MG	[INTRODUCTION]	
2	W	1/14/2015	MDS	Genomics I	
3	F	1/16/2015	MDS	Genomics II	
	M	1/19/2015	--	(no class, MLK)	
4	W	1/21/2015	JR	Proteomics I	
5	M	1/26/2015	JR	Proteomics II	
6	W	1/28/2015	MG	[ALIGNMENT] (seq. comparison & multiple-seq. alignment)	
7	M	2/2/2015	MG	[UNSUPERVISED MINING] (focusing on spectral methods such as SVD)	
8	W	2/4/2015	TAs	Quiz only - no lecture	QUIZ 1
9	M	2/9/2015	KC	Application of Semantic Web to Biology	
10	W	2/11/2015	MG	[SUPERVISED MINING] (focusing on Trees & SVMs)	
12	M	2/16/2015	MG	Mining #3 - Processing Next-Gen sequencing data (w/ MRS)	
13	W	2/18/2015	MG	Mining #4 - Practical Tips in Building Models ("MOOC class")	
14	M	2/23/2015	MG	Analysis of [NETWORK TOPOLOGY]	
15	W	2/25/2015	MG	Practical Network & NGS Analysis (Cancer Genomics Application w/ YF)	
16	M	3/2/2015	MG	[NETWORK PREDICTION]	Homework 1 Due
17	W	3/4/2015	MG	Flexible Class	QUIZ 2
	M	3/9/2015	--	(no class)	
	W	3/13/2015	--	(no class)	
	M	3/16/2015	--	(no class)	
	W	3/18/2015	--	(no class)	
				Simulation (2nd Half)	
18	M	3/23/2014	SK	Cell/Immune Simulation I	
19	W	3/25/2014	SK	Cell/Immune Simulation II	
20	M	3/30/2014	SK	Cell/Immune Simulation III	
21	W	4/1/2014	CO	Protein Simulation I	QUIZ 3
22	M	4/6/2014	CO	Protein Simulation II	Homework 2 Due
23	W	4/8/2014	CO	Protein Simulation III	
24	M	4/13/2014	CO	Markov Models I	
25	W	4/15/2014	CO	Markov Models II	
26	M	4/20/2014	CO	Markov Models III	Homework 3 Due
27	W	4/22/2014	CO	Protein Aggregation	QUIZ 4
	F	4/26/2014			FINAL PROJECT DUE
		5/8/2014			NO FINAL EXAM

Private | Public

Comments

You do not have permission to add comments.

CBB752 - Bioinformatics: Practical Application of Data Mining & Simulation

Search this site

Home | Syllabus | Announcements | Instructors | Section Readings | Assignments | Final Project | Grading

Announcements

First Meetings

First meeting will be on Mon. 12 Jan 2015, immediately followed by an office hour for Prof Gerstein, for those with individual questions about the class. The class will meet again Wed. and Fri. of that week.

Polls

Go to http://goo.gl/zkhxBJ for students' sign up and good times for the weekly discussion section

Changes relative to Last Year

18th year of the course. You are encourage to last year's site as a guide. In particular, many of the PDFs & PPTs for lectures will be similar to those last year. New developments for this year include :

(1) inclusion of one inside-out-class using a MOOC on machine learning
(2) Different lectures on "Genomics Data," now done by M Simon, rather than J Noonan
(3) New website look
(4) Additional emphasis on cancer genomics

Snow Days (general policy)

We have built into the class schedule the potential for snow days. To avoid last minute uncertainty and confusion, we will not wait until Yale officially closes the university for snow (which only happens in the most extreme of blizzards). If the weather looks particularly problematic a few days before (eg on Sat. for a Mon. class), we will preemptively cancel via the class email list, which means it important for all to be on this list.

Comments

You do not have permission to add comments.

Sign in | Recent Site Activity | Report Abuse | Print Page | Powered By Google Sites

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	300 George St	jesse.rinehart(at)yale.edu
Matthew Simon	West Compus	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

Name	Office	Email
Michael Rutenberg Schoenberg	Bass 437	michael.rutenbergschoenberg(at)yale.edu
Yao Fu	Bass 437	yao.fu(at)yale.edu

Comments

You do not have permission to add comments.

Section Readings

Each section will include discussion of papers assigned (below). Students are expected to submit 1-2 paragraph summaries of each paper before the section. The written assignment will be the same, and students will be graded on a combination of the written assignments and your participation in discussions.

Session 1: Next Gen Sequencing
Metzker ML. "Sequencing technologies - the next generation" Nature Reviews Genetics. 11 (2010) PDF
Wheeler DA et al. "The complete genome of an individual by massively parallel DNA sequencing," Nature. 452:872-876 (2008) PDF

Session 2: Proteomics/Sequence Alignment
T.F. Smith and M.S. Waterman. (1981) Identification of common molecular subsequences. Journal of Molecular Biology,147(1): 195-7. PMID: 7265238. PDF Nevan J. Krogan et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae Nature 440, 637-643 (30 March 2006) PDF
Additional readings suggested by Professor Rinehart

Session 3: Sequence Alignment/Machine learning
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 Yip, KY, Cheng, C, Gerstein, M (2013). Machine learning and genome annotation: a match meant to be?. Genome Biol., 14, 5:205. PDF

Session 4: Bioinformatics for Next-Gen Sequencing
Rozowsky, J, Euskirchen, G, Auerbach, RK, Zhang, ZD, Gibson, T, Bjornson, R, Carriero, N, Snyder, M, Gerstein, MB (2009). PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. Nat. Biotechnol., 27, 1:66-75 PDF
Cooper, GM, Shendure, J (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet., 12, 9:628-40 PDF

Session 5: Bioinformatics for Next-Gen Sequencing 2
Lior Pachter. Models for Transcript Quantifications from RNA-Seq (2011) ArXiV PDF

Session 6: Networks
Ekman D, Light S, Björklund AK, Elofsson A. (2006) What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? Genome Biol. 2006;7(6):R45. PDF
Barabási, AL, Oltvai, ZN (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet., 5, 2:101-13. PDF

Session 7: Immunological Modeling/Semantic Web
Perelson AS. Modelling viral and immune system dynamics. Nat Rev Immunol. 2002 Jan;2(1):28-36. PDF
Antezana E, Egaña M, Blondé W, Illarramendi A, Bilbao I, De Baets B, Stevens R, Mironov V, Kuiper M. (2009) The Cell Cycle Ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome Biol. 2009;10(5):R58. Epub 2009 May 29. PDF

Session 8: Protein Simulation 1
Martin Karplus and J. Andrew McCammon. (2002) Molecular dynamics simulations of biomolecules. Nature Structural Biology,9, 646-52. PMID: 12198485.PDF
Zhou, AQ, O'Hern, CS, Regan, L (2011). Revisiting the Ramachandran plot from a new angle. Protein Sci., 20, 7:1166-71 PDF

Session 9: Protein Simulation 2
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

Assignments

Assignments #1

posted Mar 19, 2012, 9:34 AM by Yao Fu **[updated Oct 14, 2014, 5:47 AM]**

1-1 of 1 ‹ ›

Home | Syllabus | Announcements | Instructors | Section Readings | Assignments | **Final Project** | Grading

Final Project

Showing **0** items

Name	Due Date	Description
Sort	Sort	Sort

Showing **0** items

Comments

You do not have permission to add comments.

CBB752 - Bioinformatics: Practical Application of Data Mining & Simulation

Grading

Elements Contributing to Grade

The following is the approximate way that the grade will be determined:

CBB and CPSC Sections:

Category	% of Total Grade
Quizzes	33%
Final Project	33%
Discussion Section	9%
Programming Assignments	25%

MBB and MCDB Sections:

Category	% of Total Grade
Quizzes	33%
Final Project	33%
Discussion Section	17%
Problem Sets	17%

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://yalecollege.yale.edu/content/reading-period-and-final-examination-period
http://yalecollege.yale.edu/content/completion-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 Dean Mary Miller 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.