CS 4775 / BIOCB 4840 / 6840
Computational Genetics and Genomics

Syllabus: Fall 2023

General Course Information

Lectures:	Tue/Thu 10:10am–11:25am, Warren Hall 175
Discussion:	Fri 12:20pm–1:10pm, Warren Hall 175
Credit hours:	4 (S/U or letter)
Course web:	Canvas canvas.cornell.edu/courses/57066
Course forum:	Ed Discussion (linked from Canvas)
Assignments:	Gradescope (linked from Canvas)
Instructor:	Jaehee Kim (jaehee.kim@cornell.edu)
	Office hours: Wed 10–11am, Weill Hall 102C
	(Note: Jaehee's office hours are for course project related topics)
	jaeheekimlab.github.io
TAs:	Jeremy Jung (jj523@cornell.edu)
	Office hours: Mon 1–2pm, Warren Hall 113
	Zhaozhi Li (zl643@cornell.edu)
	Office hours: Tue 5–6pm, Warren Hall 113
	Wai Tung 'Jack' Lo (wl428@cornell.edu)
	Office hours: Wed 5–6pm, Warren Hall 113

Course Description

This course presents some of the most foundational computational methods for analyzing genetic and genomic data. Topics include sequence alignment, hidden Markov models for discovering sequence features, motif finding using Gibbs sampling, phylogenetic tree reconstruction, inferring haplotypes, genetic epidemiology, and DNA mixture analysis. Prior knowledge of biology is not necessary to complete this course. By the end of the course, students will 1) understand computational algorithms used for the analysis of genetic and genomic data; 2) formulate computational approaches for solving problems in computational genomics; and 3) understand challenges and limitations in inference methods used in computational genetics and genomics.

Prerequisites

The prerequisites for this course are listed as "BTRY 3010 and CS 2110 or equivalents." In practice, those who are willing to work at understanding algorithms or statistics can do well. However, students with strong backgrounds in these disciplines will find the course easier than those without such backgrounds. The following are general guidelines for a strong background in the three major topic areas in this course:

- Statistics. A general statistical methods course (e.g., BTRY 3010 or 6010) would provide helpful background. Those who have taken only a probability course will need to spend additional time understanding the statistical inference approaches covered in this course. However, such students can succeed with sufficient effort, including consulting outside textbooks where needed.
- Computer science. The ability to program is essential to completing this course, and problem set solutions written in Python will be accepted. Additionally, because we will discuss algorithms and

algorithm complexity in the class, a programming/algorithms course, such as CS 2110 or CS 3110, would be ideal. Students who have not taken these courses can succeed by taking the time to learn algorithm complexity and design.

- Molecular biology and genetics. All biological and genetics concepts that students need will be covered in this course within the lectures and sections. However, the course does not delve deeper into these topics. Occasional outside reading from available online resources will likely be beneficial to gain greater clarity on these concepts, although detailed understanding is not needed.

Textbooks

The primary textbook for this course is:

- Durbin R, Eddy SR, Krogh A, and Mitchison G, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, 1998.
- E-book available via Cornell library newcatalog.library.cornell.edu/catalog/15318860.
- Book errata: http://eddylab.org/cupbook_errata.html.

Other books on bioinformatics that may be worth referencing are:

- Jones NC, Pevzner PA, *An Introduction to Bioinformatics Algorithms*, MIT Press, 2004.
- Felsenstein J, *Inferring Phylogenies*, Sinauer, 2004.

The Felsenstein book contains very detailed information about methods for phylogenetic inference. It also contains introductions to continuous-time Markov models of DNA substitution, maximum likelihood phylogeny reconstruction, and other topics we will cover. The Jones and Pevzner book focuses on algorithms, with a clear introduction to dynamic programming.

Students who wish to study more on probability and statistics, algorithms, or molecular biology can explore books on these topics. Some relevant books include:

- Wasserman, Larry *All of statistics: a concise course in statistical inference*. Springer, 2004. E-book available via Cornell library newcatalog.library.cornell.edu/catalog/12462121.
- Hogg RV, Craig AT. *Introduction to Mathematical Statistics*. Prentice Hall, 1995.
- Casella G, Berger RL. *Statistical Inference*. Duxbury Press, 2001.
- Cormen TH, Leiserson CE, Rivest RL, Stein C. *Introduction to Algorithms*. MIT Press, 2001.

Grading

Grades will be based on a combination of five problem sets, a class project, and discussion section participation; each of these elements will contribute to final grades as follows:

- Assignments (45% total): Assignment 1–3 (12% each), Assignment 4 (9%)
- Discussion section attendance: 5%
- Class project: 50%
- (optional) Extra credit, class project difficulty: 3% max
- (optional) Extra credit, participation (answering Ed questions throughout the semester): 3% max

Problem Sets

Problem sets will be somewhat challenging and take some time to complete. Start early, and make use of discussion sections and office hours. The problem sets will be assigned roughly every other

week, and you will have two weeks to complete each set. All assignments must be submitted electronically to Gradescope. Programming assignments must include readable code with reproducible results. Assignments are due at 11:59 pm (Eastern Time) on the assigned due date.

Late policy: Late problem sets will have their score reduced by 10% per day late. Three total free late days for the entire semester will be granted without any reduction in score. The final project report must be in on time.

Class Project
The project is an opportunity to apply the concepts and skills you have learned in this course to a real research question. The project should be substantial. Start thinking about the project early even if you do not decide what you will do for the project until later in the semester. You will be asked to submit a project proposal in late-September and a progress report in late-October. The project must be done in a group of four students for undergraduates (CS 4775/BIOCB 4840). For graduate students (BIOCB 6840), the project can be done individually as well.

For graduate students (BIOCB 6840), the project should involve original research and, except in rare cases (e.g., a challenging theoretical project), should involve some programming and some analysis of real biological data. Undergraduates (CS 4775/BIOCB 4840) will not be expected to conduct original research, but should attempt a substantial programming project, a comprehensive literature review on a topic of interest, or something of similar scope. In all cases, presentation to the class will be due and take place during the last week of class, and a written project report will be due on Dec 16th. Some class projects may lead to thesis projects and/or conference or journal publications. For each submission, detailed instructions and LaTeX templates will be provided.

Grading breakdown for the class project (50% of total class grade):

- Initial project proposal: 5%
- Mid-term progress report: 10%
- Project presentation: 10% (7% on presentation, 3% attendance & participation)
- Final project report: 25%
- (optional) Extra credit, class project difficulty: 3% max

Participation
Note the discussion section participation component. The sections will cover examples and hands-on exercises of the materials covered in the lecture. You are encouraged to come to the section prepared, do your best to follow the lectures, keep up with the readings, ask good questions, and so on—in short, you are expected to be an active, inquisitive learner. Doing so will make the course more rewarding for you and all involved. Overall, you must attend 9 (out of 13) sections to receive the full credit for section participation. The section attendance will be tracked with Poll Everywhere during each section.

Lecture attendance will not be tracked and will not be factored into the final grade. However, class attendance is required on all final project presentation days (Nov 28, Nov 30, and Dec 01).

You can earn extra participation credit by actively (e.g., consistently throughout the semester) posting and responding to questions on Ed Discussion.

Collaboration and Academic Integrity

Collaboration is encouraged, but you must turn in your own work and must acknowledge all sources and collaborators. Be sure you understand the work you turn in.

You are expected to abide by the Cornell University Code of Academic Integrity and CS Department Academic Integrity Policy in this class. Any work you submit in this course for academic credit must be your own work. However, collaboration is allowed on the problem sets so long as you turn in your own work and list the individuals with whom you collaborated. If plagiarism is detected, you will receive a score of zero on the assignment in question.

Please note that some assignments are submitted to Turnitin. You agree that, by taking this course, all class project reports may be subject to submission for textual similarity review to Turnitin.com for the detection of plagiarism. All submissions will be included as source documents in the Turnitin reference database solely for the purpose of detecting the plagiarism of such papers. Use of the Turnitin service is subject to the Usage Policy posted on the Turnitin site.

Students with Disabilities

Your access in this course is important to me. Please request your accommodation letter early in the semester, or as soon as you become registered with Student Disability Services (SDS), so that adequate time exists to arrange your approved academic accommodations. If you have or think you may have a disability, please contact Student Disability Services for a confidential discussion: sds_cu@cornell.edu or visit sds.cornell.edu to learn more.

Mental Health and Well-being

There are services and resources at Cornell designed specifically to bolster undergraduate, graduate, and professional student mental health and well-being. Remember, your mental health and emotional well-being are just as important as your physical health. If you or a friend are struggling emotionally or feeling stressed, fatigued, or burned out, there is a continuum of campus resources available to you: mentalhealth.cornell.edu/get-support/support-students. Help is also available any time day or night through Cornell's 24/7 phone consultation (607-255-5155).

Tentative schedule

Date	Topic	Note
Aug 22	Lecture 1: Introduction	
Aug 24	Lecture 2. Probability and statistics background	
Aug 29	Lecture 3. Sequence alignment and dynamic programming	PS 1 assigned
Aug 31	Lecture 4. Global and local sequence alignment	
Sep 05	Lecture 5. Affine gaps, Markov chain	
Sep 07	Lecture 6. Hidden Markov model (HMM)	
Sep 12	Lecture 7. HMM: Viterbi, forward, and backward algorithms	PS 1 due; PS 2 assigned
Sep 14	Lecture 8. HMM: Simulating and supervised learning	
Sep 19	Lecture 9. HMM applications	
Sep 21	Lecture 10. Phylogenetic reconstruction: Parsimony, UPGMA, NJ	
Sep 26	Lecture 11. Phylogenetic reconstruction: Felsenstein's algorithm, Jukes-Cantor	PS 2 due; PS 3 assigned
Sep 28	Lecture 12. Phylogenetic reconstruction: Maximum likelihood	
Oct 03	Lecture 13. Phylogenetic reconstruction: Bayesian inference	
Oct 05	Lecture 14. MCMC I	**Initial project proposal due**
Oct 10	**No class: Fall break**	
Oct 12	Lecture 15. MCMC II	
Oct 17	Lecture 16. MCMC applications	PS 3 due; PS 4 assigned
Oct 19	Lecture 17. Expectation maximization I	
Oct 24	Lecture 18. Expectation maximization II	
Oct 26	Lecture 19. Multiple sequence alignment	PS 4 due
Oct 31	Lecture 20. Exact string matching, Heuristic alignment	
Nov 02	Lecture 21. Genome assembly	
Nov 07	Lecture 22. Dimensionality reduction	
Nov 09	Lecture 23. Machine learning	**Project progress report due**
Nov 14	Lecture 24. Special topic: Genetic epidemiology	
Nov 16	Lecture 25. Special topic: Forensic genetics	
Nov 21	Lecture 26. Special topic: RNA-seq data analysis	
Nov 23	**No class: Thanksgiving day**	
Nov 28	**Class project presentations**	
Nov 30	**Class project presentations**	
Dec 01	**Class project presentations**	
Dec 16	**Class project final report due**	