Computational Biology Methods (COMP-462; 3 credits)
Computational Biology Methods and Research (COMP-561; 4 credits)

Fall 2023

Instructors:
David Becerra, David.becerra@mcgill.ca
Mathieu Blanchette, Mathieu.blanchette@mcgill.ca

Teaching assistants: See MyCourses

Course Abstract:
Computational biology is the sub-discipline of Bioinformatics that is closest in spirit to pure computer science. The main efforts in this field are two-fold. Firstly, we are concerned with creating models for problems from the biosciences (biology, biochemistry, medicine) that are both biologically and mathematically sound. Secondly, we are interested in the design and analysis of efficient, and accurate algorithms that solve these problems in practice and strategies for validation of results.

This course is designed to introduce upper-year undergraduate students and graduate students to this area by examining several classic problems from the field. The intention of the course is to act as a gateway whereby, upon completion of the course, students will have the necessary biology, mathematics and computer science background to attend graduate level courses in bioinformatics geared towards specific topics (phylogenetics, genomic evolution, functional genomics, proteomics). The course is designed in such a manner that no previous formal training in biology is required of the students.

The necessary mathematical background consists of the lower-level discrete structures and probability courses, since topics such as maximum likelihood estimation, hidden Markov models, and dynamic programming will be used repeatedly throughout the material. (Both maximum likelihood and hidden Markov models will be introduced at a basic level however.) Students will be required to have already taken the lower-level algorithms/data-structures, numerical computing and theoretical computer science courses.

Computer Science/mathematics topics	**Biology and biotechnologies topics**
Basic probability and statistics (ubiquitous)	Genomics
Dynamic programming (sequence alignment)	Models of evolution and phylogenetics
Approximation algorithms (string alignment)	Sequence comparison
Advanced data structures (suffix trees)	Gene expression and regulation
Numerical techniques (least squares fits)	DNA sequencing
Basic machine learning	Peptide identification
Experimental design	RNA and protein structure
Programming	Population genetics

Prerequisites:

COMP 251 Data Structures and Algorithms
MATH 323 Probability Theory (or equivalent)

Office hours: See MyCourses

Books: No book is required, but the following books contain material covered in this course.

- Bernhard Haubold and Thomas Wiehe. Introduction to Computational Biology: An Evolutionary Approach , Burkhauser Basel, 2006
  Probably the best introductory book out there. Its level is ideal for the course, but it does not go much beyond this.

- Durbin, Eddy, Krogh, Mitchison, Biological Sequence Analysis, Cambridge, 1998.
  This book is particularly good for learning some of the basics of statistical inference.

- Jones, N.C. and Pevzner, P. An Introduction to Bioinformatics Algorithms, MIT press, 2004
  Good book for understanding some of the classic problems in computational biology and the algorithms used to solve these problems.

- Alberts, Heald, Johnson, Morgan, Raff, Roberts, Walter, Wilson Molecular Biology of the Cell (7th Edition), Norton.
  This is a widely used and comprehensive book covering the biology of the cell. It is a good place to start when you want to explore a new topic.

- Jan Gorodkin and Walter Russo, *RNA Sequence, Structure, and Function: Computational and Bioinformatics Methods*, Humana Press, 2014.
  A book which covers the Introduction to RNA structure prediction.

- *Peter Clote and Rolf Backofen,* Computational Molecular Biology: An Introduction*,* Wiley, 2000.
  A good book to review the Protein structure prediction topic.

Evaluation for COMP 462:

4 assignments: Best 3 out of 4, weighed at 13.33% each	40%
In-class Midterm exam, November 1	20%*
Final Exam, during exam period, date TBD	40%

(*) If the final exam grade is better than the midterm grade,
the final exam grade replaces the midterm grade

Evaluation for COMP 561:

3 assignments (the first three): Best 2 out of 3, weighted at 15% each	30%
Final project (teams of at most 3)	20%

In-class Midterm exam, November 1	20%*
Final Exam, during exam period, date TBD	30%

(*) If the final exam grade is better than the midterm grade,
the final exam grade replaces the midterm grade

Missed assignments or midterm exams due to COVID: The evaluation scheme has built-in flexibility to accommodate students who may get sick with COVID. *If you have COVID or suspect you may have it, stay home and take care of yourself.* If you miss an assignment or the midterm exam, no need for a doctor's note; the flexible grading scheme will automatically adjust. If you miss two or more assignments due to illness, or the final project, please alert me immediately. If you miss the final exam due to illness, you can take the deferred exam.

Course outline

Lectures 1-2: Introduction to molecular biology and genomics.

Background Reading: Chapter 1 of Artificial Intelligence and Mol. Biology, by L. Hunter

Lectures 3-6: Sequence evolution and pairwise sequence alignment.

Topics: Introduction to sequence evolution. Global and local alignment; Gapping; BLAST algorithm; Multiple Alignments.
Applications: Sequence comparison, homology detection.
Background Reading: Chapter 6 of Jones, Pevzner; Chapter 6 of Durbin and Eddy.
Math/CS: Dynamic Programming, heuristics

Lecture 7-8: Genome sequencing&assembly, resequencing and read mapping

Topics: DNA sequencing; De novo genome assembly problem; Read mapping.
Background Reading: Chapter 8.10-8.15 of Jones, Pevzner.
Math/CS: Graph theory

Lecture 9-10: Genome sequencing&assembly, Introduction to population genetics

Topics: Polymorphisms, haplotypes. Genotyping by sequencing.
Background Reading: TBD
Math/CS: Probabilities

Lectures 11-13: Evolutionary models and phylogenetic inference.

Topics: Discrete and continuous nucleotide and amino acid substitution models; Distance-based methods; Parsimony; Maximum Likelihood; Tree-space search heuristics
Background Reading: Chap. 7-8 of Durbin et al. or Chap. 8 and 9.1-9.5 of Haubold&Wiehe
Math/CS: Discrete algorithm design; Maximum likelihood.

Lectures 14-16: Hidden Markov Models.

Topics: HHMs, Viterbi, Forward-Backward, Baum-Welch algorithms.
Applications: Gene-finding HMMs. Profile HMMs.
Background Reading: Chapters 3 and 5 of Durbin et al.

Math/CS: Markov processes; Dynamic programming; Parameter estimation.

Lecture 17: Midterm exam, Nov. 6, in class (open book).

Lecture 18-20: Gene Expression Analysis.

Topics: Class distinction; Class prediction; Class discovery (clustering)
Background Reading: Chapter 10 of Jones&Pevzner.
Math/CS: Differential expression; Intro. to supervised machine learning; Clustering;

Lecture 22-25: Computation Structural Biology

Topics: RNA secondary and tertiary structure; Protein secondary and tertiary structure;
De novo structure prediction; Homology-based modeling, Minimalist models.
Background Reading: Chap. 10.1 and 10.2 of Durbin et al..
Chap 2 and 4 of Gorodkin and Russo.
Chap 6 of Cote and Backofen.
Math/CS: Energy minimization, machine learning, Monte Carlo and Genetic Algorithms.

Lecture 26: Review session

Final exam: **During the exam session.**

+++++

Late submission policy

Late assignments will be deducted 10% each day or fraction thereof for which they are late, including weekend days and holidays. The instructors reserve the right to modify the lateness policy for a particular assignment; any such modifications will be clearly indicated at the beginning of the relevant assignment specifications.

Plagiarism Policy

McGill University values academic integrity. Therefore all students must understand the meaning and consequences of cheating, plagiarism, and other academic offenses under the Code of Student Conduct and Disciplinary Procedures (see www.mcgill.ca/integrity/ for more information).

Work submitted for this course must represent your own efforts. Assignments must be done individually; you must not work in groups. Do not rely on friends or tutors to do your work for you. You must not copy any other person's work in any manner (electronically or otherwise), even if this work is in the public domain or you have permission from its author to use it and/or modify it in your own work. Furthermore, you must not give a copy of your work to any other person, nor should you post your solutions on any publicly accessible repository.

The plagiarism policy is not meant to discourage interaction or discussion among students. You are encouraged to discuss assignment questions with instructors, TAs, and your fellow students. However, there is a difference between discussing ideas and working in groups or copying someone else's solution. A good rule of thumb is that when you discuss assignments with your fellow students, you should not leave the discussion with written notes. Also, when you write your solution to an assignment, you should do it on your own.

About posting solutions

The instructor will do their best to provide solutions to assignments and exams in a timely manner. These solutions are the property of McGill and must not be posted anywhere online. Posting questions and solutions online would be considered as facilitating plagiarism and would be treated as such.

Official language policy for graded work

In accordance with McGill University's Charter of Students' Rights, students in this course have the right to submit in English or in French any written work that is to be graded.

Mathieu Blanchette
David Becerra

2023-09-01