CMSC 498Y
Statistical Inference and Machine Learning for Genomics Data

Syllabus for Spring 2022
Last updated: 2022-01-24 22:54:01 -05:00

1 Course Information

- **Number:** CMSC 498Y  
- **Section:** 0101  
- **Title:** Statistical Inference and Machine Learning for Genomics Data  
- **Credits:** 3  
- **Instructor:** Erin Molloy (she/her) ([https://ekmolloy.github.io](https://ekmolloy.github.io))  
- **Teaching Assistant:** Tobias Rubel (they/them)  
- **Term:** Spring 2022 (January 24th to May 18th)  
- **Times:** Tuesdays and Thursdays 9:30–10:45 AM  
- **Location:** IRB-2207 unless otherwise indicated on the schedule  
- **Course Materials:** ELMS ([http://elms.umd.edu](http://elms.umd.edu))  
- **Communication:**  
  - CampusWire ([https://campuswire.com](https://campuswire.com)) – For questions/comments on course material, assignments, exams, etc.  
  - ELMS (send message to instructors) – For excused absences and other urgent / personal matters  
- **Office Hours:**  
  - Erin Molloy – Thursdays 3:30-4:30 PM in IRB-3256 / IRB-3236 (same hallway)  
  - Tobias Rubel – Mondays 4:00–5:00 PM in AVW-4166  
  - It may be possible to meet with us via Zoom during these times, please message us in advance to check.

2 Course Description

The objective of this course is to provide an understanding of statistical and machine learning algorithms that have been developed or applied to problems in **comparative genomics**, for example building multiple sequence alignments, estimating phylogenies (evolutionary trees), and inferring population structure from single nucleotide polymorphisms (SNPs). Example topics include Hidden Markov Models (HMMs), Expectation-Maximization (EM), Markov Clustering (MCL), Principal Component Analysis (PCA), Matrix Factorization, Maximum Likelihood (ML) and Bayesian methods (including composite and pseudo-likelihood
functions), and Markov Chain Monte Carlo (MCMC). This course (CMSC 498Y) is intended to be complementary to CMSC 423 Bioinformatic Algorithms, Databases, and Tools, which largely focuses on discrete algorithms (e.g., suffix trees, Burrows-Wheeler transform, FM-index, de Bruijn graphs, etc.); see the syllabus from Fall 2021 here: https://rob-p.github.io/CMSC423_F21/. Please keep in mind that CMSC 498Y is a new course being offered for the first time at UMD in Spring 2022. The result is that course materials (e.g., lecture slides and any accompanying notes) will be made available as the semester progresses instead of weeks in advance.

3 Course Prerequisites

The target audience is upper-level undergraduate students majoring in computer science and/or other quantitative disciplines as well as graduate students in bioinformatics-related disciplines. The prerequisites are familiarity with probability, statistics, and linear algebra as well as programming ability in either C/C++ or Python. No prior knowledge of biology is assumed. Please contact me via email (ekmolloy@umd.edu) if you are unsure as to whether you should enroll in the course or if you don’t exactly meet the prerequisites but are interested in enrolling.

4 Course Overview

- **Module #1: Background in statistics, biology, etc (~2 weeks)**
- **Module #2: Ortholog Identification and Multiple Sequence Alignment (~3 weeks)**
  - Models:
    * Sequence Profiles
    * Hidden Markov Models (HMMs)
  - Methods:
    * Viterbi Algorithm (dynamic programming)
    * Baum-Welch Algorithm (expectation-maximization)
    * Ensemble Methods
    * Markov Clustering
- **Module #3: Phylogeny Estimation (~2 weeks)**
  - Models:
    * Cavendar-Farris-Neyman (CFN) model
    * Generalized Time Reversible (GTR) model
  - Methods:
    * Felsenstein’s pruning algorithm (dynamic programming)
    * Phylogenetic placement
    * Maximum Likelihood (ML) search heuristics (randomized taxon addition & tree space moves)
    * Bootstrapping
- **Module #4: Species (or Population) Tree Estimation (~2.5 weeks)**
  - Models:
    * (Multi-Species) Coalescent model
    * Infinite Sites model
    * Wright-Fischer model
  - Methods:
    * ML and Bayesian methods, with pseudo-likelihood and composite likelihood functions
• Markov Chain Monte Carlo (MCMC)
• Statistical tests for non-tree-like evolution (e.g., admixture or hybridization)
• Jackknife resampling

• Module #5: Population Structure Inference (~2.5 weeks)
  – Models:
    * STRUCTURE/ADMIXTURE model
  – Methods:
    * Principal Component Analysis (PCA)
    * Supervised and unsupervised approaches based on matrix factorization
    * Mailman algorithm for matrix-vector multiplication

Note that this list of topics is not exhaustive. There may be time permitting for additional topics to be covered or guest lectures.

5 Course Grading

40% – Exams
  • 10% – Midterm Exam #1 on Tuesday, March 8th 9:30–10:45 AM
  • 10% – Midterm Exam #2 on Tuesday, April 26th 9:30–10:45 AM
  • 20% – Final Exam on Friday, May 13th 8:00–10:00 AM

60% – Assignments
  • Five mini-projects, one for each module

Mini-projects. For each of the four main modules, we will release a mini-project that includes some mix of solving math problems, programming, and analyzing data. We will use publicly available genomic data sets in addition to simulations. The programming assignments can be completed in either Python or C++ using libraries/modules that are specified in the assignment. Some data analyses will be performed with existing bioinformatics tools that can be run through the command line. There will be some in-class demos and further questions can be asked on CampusWire or during office hours. You will have between 2 weeks (between 1.5 and 3 weeks) to complete each mini-project.

6 Textbook

There is no required textbook for the class. We will post PDFs or links to recommended readings to accompany lectures on ELMS. These will typically be from

• Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids (Durbin, Eddy, Krogh, Mitchinson, 1998) [Available online]

or open-source textbooks, for example:

• Phylogenetics in the Genomic Era [Download here: https://hal.inria.fr/PGE/]
• Population Genetics Notes (Coop, 2020) [Download here: https://github.com/cooplаб/popgen-notes]

Please reach out to the course staff (e.g., via CampusWire) if you are having trouble understanding a topic after listening to the lecture and reviewing the slides and other course materials (e.g., the recommended readings).
7 Course Policies

7.1 Excused Absences

Please review the university’s policy on excused absences for what constitutes a reason for an excused absence and the expectations for requesting accommodations. All requests accommodations for excused absences must be made in writing via ELMS. If, for a health-related or medical reason, you will miss two or more consecutive classes, or will miss class on a recurring basis, or were unable to meet a particular academic obligation of this course, I will require a written note from the Student Health Service or a healthcare provider documenting the range of dates for which you were unable to meet your academic obligations. This note need not contain any diagnostic information. If you are exhibiting any symptoms indicative of COVID-19 or believe that you have been exposed, please do not attend class in person. Lastly, report absences on https://grades.cs.umd.edu/classWeb/viewGrades.cgi and through ELMS to the instructors.

7.2 Religious Holidays

Excused absences for religious holidays will be made on an individual basis. You must request accommodations for religious reasons within the first two weeks of the semester.

7.3 Academic Accommodations

Any student eligible for and requesting academic accommodations due to a disability must provide a letter of accommodation from Accessibility & Disability Service (ADS) within the first two weeks of the semester (or within a reasonable timeframe of the disability becoming known). Letters must be sent via email (ekmolloy@umd.edu).

7.4 Mask Wearing

President Pines provided clear expectations to the University about the wearing of masks for students, faculty, and staff. Face coverings over the nose and mouth are required while you are indoors at all times. KN95 masks are required in all classroom settings and recommended everywhere. There are no exceptions when it comes to classrooms, laboratories, and campus offices. Students not wearing a mask will be given a warning and asked to wear one, or will be asked to leave the room immediately. Students who have additional issues with the mask expectation after a first warning will be referred to the Office of Student Conduct for failure to comply with a directive of University officials.

7.5 Other COVID-19 policies and protocols

The University has put forth policies related to COVID-19, many of which can be found on the Provost’s website. For information on COVID-19 protocols, including requirements for coming to campus, visit this website. It is our shared responsibility to stay up-to-date on the University’s COVID-19 protocols and policies throughout the semester.

7.6 Academic Integrity

Short form: Don’t cheat. If you are wondering whether something constitutes cheating, ask the instructor or TA. Long form: The university’s Code of Academic Integrity is a nationally recognized honor code, administered by a Student Honor Council. Any of the following acts, when committed by a student, shall constitute academic dishonesty:

- Cheating: Fraud, deceit, or dishonesty in any academic course or exercise in an attempt to gain an unfair advantage and/or intentionally using or attempting to use unauthorized materials, information, or study aids in any academic course or exercise.
Fabrication: Intentional and unauthorized falsification or invention of any information or citation in any academic course or exercise.

Facilitating academic dishonesty: Intentionally or knowingly helping or attempting to help another to violate any provision of the Code of Academic Integrity.

Plagiarism: Intentionally or knowingly representing the words or ideas of another as one’s own in any academic course or exercise.

Academic integrity is a very serious issue. Any assignment or exam that you complete in this course is expected to be your own work. If you are allowed to discuss the details of or work together on an assignment, this will be made explicit. Otherwise, you are expected to complete the work yourself. Plagiarism is not just the outright copying of content. If you paraphrase someone else’s thoughts, words, or ideas and you don’t cite your source, this constitutes plagiarism. It is always much better to turn in an incorrect or incomplete assignment representing your own efforts than to attempt to pass off the work of another as your own. Please ensure that you fully understand this code and its implications because all acts of academic dishonesty will be dealt with in accordance with the provisions of this Code.

7.7 Late Assignment Policy
Assignments that are turned in late will be docked 1% for each hour they are late up to the first 48 hours. After 48 hours, late assignments will not be accepted. If you have a reason for an excused absence (e.g., sickness, death in the family, etc.), you may choose to contact the course instructor via ELMS regarding an extension. Do not assume that an extension will be granted given that the majority of assignments (i.e., mini-projects) have several weeks to be completed.

7.8 Regrade Policy
All requests to re-grade, re-check, or re-mark an assignment or exam question must be sent through ELMS within two weeks of receiving the grade. When the assignment is re-graded, it will be re-checked in its entirety. This means that it is possible to lose points on other problems if they were graded incorrectly or too leniently the first time. Therefore, I urge you to thoroughly consider each regrade request you make.

7.9 Names/Pronouns and Self-Identifications
I invite you, if you wish, to tell us how you want to be referred to both in terms of your name and your pronouns (he/him, she/her, they/them, etc.). I will work to address and refer to all students accordingly, and I ask you to do the same for others. Visit this website to learn more.

7.10 Course Concerns and Incident Reporting
Please review this information about reporting course concerns as well as incidents of sexual harassment, student contact, and hate bias. Any behavior (including harassment, sexual harassment, and racially and/or culturally derogatory language) that threatens this atmosphere will not be tolerated. I invite you to come to me if you have concerns about such behavior during the semester.

7.11 Course Evaluation
Please submit a course evaluation through CourseEvalUM in order to help faculty and administrators improve teaching and learning at the University. All information submitted to CourseEvalUM is confidential. Campus will notify you when CourseEvalUM is open for you to complete your evaluations. Please go directly to the Course Eval UM website to complete your evaluations. By completing all of your evaluations each semester, you will have the privilege of accessing through Testudo, the evaluation reports for the thousands of courses for which 70% or more students submitted their evaluations.
7.12 Other Campus Policies

It is our shared responsibility to know and abide by the University of Maryland’s policies that relate to all courses, which include topics like:

- Academic integrity
- Student and instructor conduct
- Accessibility and accommodations
- Attendance and excused absences
- Grades and appeals
- Copyright and intellectual property

Please visit this website for undergraduate student policies and this website for graduate student policies.

8 Resources & Accommodations

8.1 Accessibility and Disability Services

The university is committed to creating and maintaining a welcoming and inclusive educational, working, and living environment for people of all abilities. The University of Maryland is also committed to the principle that no qualified individual with a disability shall, on the basis of disability, be excluded from participation in or be denied the benefits of the services, programs, or activities of the university, or be subjected to discrimination. The Accessibility & Disability Service (ADS) provides reasonable accommodations to qualified individuals to provide equal access to services, programs and activities. ADS cannot assist retroactively, so it is generally best to request accommodations several weeks before the semester begins or as soon as a disability becomes known. For assistance in obtaining an accommodation, contact ADS via email adsfrontdesk@umd.edu or phone 1-301-314-7682.

8.2 Student Resources and Services

I encourage you to visit the Student Academic Support Services website to learn more about the wide range of campus resources available to you. In particular, everyone can use some help sharpening their communication skills by visiting the Writing Center. There are a wide range of other resources available to support you. The Student Resources and Services website provides information on

- Resources for Students with Children
- Health Services
- Sexual Assault Resources
- Technology Resources
- University Libraries
- Financial / Resource Assistance
- COVID-19 Information

and more. Other important links include the LGBT Equity Center, International Student Scholars Services, and Veteran Student Life.
8.3 Mental Health Services

Diminished mental health, including significant stress, mood changes, excessive worry, substance/alcohol abuse, or problems with eating and/or sleeping can interfere with optimal academic performance, social development, and emotional wellbeing. If you or someone you know experiences any of the above mental health concerns, it is strongly encouraged to contact or visit any of the University’s resources, including UMD Mental Health Center or UMD Counseling Center. Getting help is a smart and courageous thing to do—for yourself and for those who care about you.

9 Copyright Notice

Students are permitted to use course materials for their own personal use only. Course materials may not be distributed publicly or provided to others, in any way or format.

10 Right to Change Information

This is the current version of the syllabus. Unforeseen circumstances arising during the semester could require the adjustment of any material given here. Therefore, I reserve the right to change any information on this syllabus (as well as the course schedule and the course materials) during the semester given due notice to students.
# 11 Tentative Course Schedule

Table 1: Tentative schedule for CMSC498Y. Topics covered will be added as we progress in the semester.

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Agenda</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Tu 01/25</td>
<td>Go through syllabus and give course overview</td>
</tr>
<tr>
<td></td>
<td>Th 01/27</td>
<td><strong>Lecture – Background Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>2</td>
<td>Tu 02/01</td>
<td><strong>Lecture – Background Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td></td>
<td>Th 02/03</td>
<td><strong>Lecture – Background Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>3</td>
<td>Tu 02/08</td>
<td><strong>Lecture – MSA + Ortholog Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td></td>
<td>Th 02/10</td>
<td><strong>Lecture – MSA + Ortholog Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>4</td>
<td>Tu 02/15</td>
<td><strong>Lecture – MSA + Ortholog Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td></td>
<td>Th 02/17</td>
<td><strong>Lecture – MSA + Ortholog Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>5</td>
<td>Tu 02/22</td>
<td><strong>Lecture – MSA + Ortholog Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td></td>
<td>Th 02/24</td>
<td><strong>Lecture – MSA + Ortholog Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>6</td>
<td>Tu 03/01</td>
<td><strong>Lecture – Phylogeny Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td></td>
<td>Th 03/03</td>
<td><strong>Review for Midterm Exam #1</strong></td>
</tr>
<tr>
<td>7</td>
<td>Tu 03/08</td>
<td><strong>Midterm Exam #1 (Background and MSA + Ortholog modules)</strong></td>
</tr>
<tr>
<td></td>
<td>Th 03/10</td>
<td><strong>Lecture – Phylogeny Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>8</td>
<td>Tu 03/15</td>
<td><strong>Lecture – Phylogeny Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td></td>
<td>Th 03/17</td>
<td><strong>Lecture – Phylogeny Module</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>Topics:</em></td>
</tr>
<tr>
<td>Week</td>
<td>Date</td>
<td>Agenda</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>------------------------------------------------------------------------</td>
</tr>
<tr>
<td>9</td>
<td>03/22</td>
<td>No class – spring break!</td>
</tr>
<tr>
<td></td>
<td>03/24</td>
<td>No class – spring break!</td>
</tr>
</tbody>
</table>
| 10   | Tu 03/29 | Lecture – Species Tree Module  
Topics: |
|      | Th 03/31 | Lecture – Species Tree Module  
Topics: |
| 11   | Tu 04/05 | Lecture – Species Tree Module  
Topics: |
|      | Th 04/07 | Lecture – Species Tree Module  
Topics: |
| 12   | Tu 04/12 | Lecture – Species Tree Module  
Topics: |
|      | Th 04/14 | Lecture – Population Structure Module  
Topics: |
| 13   | Tu 04/19 | Lecture – Population Structure Module  
Topics: |
|      | Th 04/21 | Review for Midterm Exam #2                                             |
| 14   | Tu 04/26 | Midterm Exam #2 (Phylogeny and Species tree modules)                   |
|      | Th 04/28 | Lecture – Population Structure Module  
Topics: |
| 15   | Tu 05/03 | Lecture – Population Structure Module  
Topics: |
|      | Th 05/05 | Lecture – Population Structure Module  
Topics: |
| 16   | Tu 05/10 | Review for Final Exam (cumulative but focus on population structure)   |
|      | Fr 05/13 | Final Exam 8:00-10:00AM (see here)                                    |