**CS 690U: Computational Biology and BioInformatics**

**Meeting Days:** T/Th
**Times:** 11:30 – 12:45
**Room: CS 140**
**Credits:** 3

**Instructor:**
Anna Green
annagreen@umass.edu
Office: CS 348
Pronouns: she/her

**TA:**
Juhyeon Lee
juhyeonlee@cs.umass.edu
Pronouns: she/her

**Office Hours:**
Anna Green: Tuesdays 10:30-11:20AM, CS 348
Juhyeon Lee: Mondays 3:00 – 4:00PM, LGRT 220

**Course Description:**
This course is designed to provide computer scientists with a comprehensive introduction to the field of computational biology. The course will cover the application of computational techniques to modern research challenges in biology, discussing both foundational algorithms and newly introduced methods. The necessary background on biology will be provided in order to understand the methods. The primary focus will be analysis of genomic data, including genome assembly, genome annotation, sequence alignment, phylogeny construction, mutation effect prediction, population genetics, and genotype-phenotype association studies. We will also cover gene expression analysis (RNA-seq and single-cell RNA-seq) and protein structure analysis and prediction. Throughout the course, we will emphasize the unique challenges to working with biological data. Through lectures and hands-on programming problem sets, students will develop the necessary skills to tackle computational challenges in the field of biology.

**Learning Objectives:**

- Gain an understanding of fundamental biological concepts and their relevance to computational biology.
- Develop proficiency in applying computational techniques and algorithms to analyze genomic data.
- Acquire practical skills in programming and data analysis relevant to computational biology.
- Explore and critically evaluate current research in the field of computational biology.
- Foster the ability to design and implement computational approaches for solving research problems in biology.

**Prerequisites:** None (this is a graduate-level CS course and thus proficiency in Python programming as well as undergraduate-level understanding of statistics, machine learning, and linear algebra is expected)

**Textbook:** None (Required readings will be free electronic materials provided by the instructor, including pdfs of journal and conference articles and pdfs of free textbooks)

**Eligibility:** This course is aimed at PhD students with a computer science background. The course assumes no prior knowledge of biology. The course may also be suitable for PhD students in biology-related disciplines who have strong computational skills.

**Grading Criteria:**
95% problem sets. The problem set with the lowest grade will not be counted.
5% participation – either via attendance, Canvas, or office hours

**Late work policy:** Problem sets will be due at 11:59pm on the day specified. If, due to extenuating circumstances, you are not able to meet the deadline, you may email me and I will adjudicate on a case-by-case basis. The lowest problem set grade will be dropped (ie, not counted towards your final average), even if that lowest grade is a 0. I encourage you to use this "drop" on a week that you will otherwise struggle to meet the deadline.

**Grading Scale:** Grading is on a letter scale, listed below. Grades will be rounded to the nearest integer.

Graduate students:

- A: 93-100%
- A-: 90-92%
- B+: 87-89%
- B: 84-86%
- B-: 80-83%
- C+: 77-79%
- C: 74-76%
- F: below 73.5

**Attendance policy:** Regular attendance and participation is critical to developing an understanding of the material and achieving success on the final project. There is no formal attendance policy.

**Collaboration policy:** In corporate and academic settings, it is encouraged that you collaborate with your colleagues and use available resources to complete work. In this spirit, I encourage you to discuss course material with your classmates and use online resources to extend your understanding. However, in order to fairly evaluate your understanding in a classroom setting, **I expect all versions of all assignments to be produced independently by you, in your own words (or code), and to reflect your own understanding of the problem.** Copying any component of an assignment from your fellow students or any other resource (including chatGPT and similar technology) is not permitted. Each assignment will include an option to describe any resources you used or fellow students you discussed with answers with, please answer these honestly.

**Accommodation Statement:** The University of Massachusetts Amherst is committed to providing an equal educational opportunity for all students. If you have a documented physical, psychological, or learning disability on file with Disability Services (DS), you may be eligible for

reasonable academic accommodations to help you succeed in this course. If you have a documented disability that requires an accommodation, please notify me within the first two weeks of the semester so that we may make appropriate arrangements. For further information, please visit Disability Services (https://www.umass.edu/disability/)

**Academic Honesty Statement:** Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst. Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent (http://www.umass.edu/dean_students/codeofconduct/acadhonesty/).

**Title IX Statement:** In accordance with Title IX of the Education Amendments of 1972 that prohibits gender-based discrimination in educational settings that receive federal funds, the University of Massachusetts Amherst is committed to providing a safe learning environment for all students, free from all forms of discrimination, including sexual assault, sexual harassment, domestic violence, dating violence, stalking, and retaliation. This includes interactions in person or online through digital platforms and social media. Title IX also protects against discrimination on the basis of pregnancy, childbirth, false pregnancy, miscarriage, abortion, or related conditions, including recovery. There are resources here on campus to support you. A summary of the available Title IX resources (confidential and non-confidential) can be found at the following link: https://www.umass.edu/titleix/resources. You do not need to make a formal report to access them. If you need immediate support, you are not alone. Free and confidential support is available 24 hours a day / 7 days a week / 365 days a year at the SASA Hotline 413-545-0800.

**Course Inclusiveness Statement:** It is important to me that this course be a welcoming environment to people of all backgrounds. My goal as an instructor is to help you learn the subject material in a way that is useful and empowering, and I believe that the best learning happens on a foundation of mutual respect.

This course will discuss subject matter related to human genetics and evolution. While in an ideal world science would be objective, the reality is that false beliefs about genetic differences between humans have been used to justify racism and oppression, and continue to fuel hateful ideologies today. I will strive to teach accurate information about the data and techniques used in human genomics while also acknowledging their potential for problematic misinterpretations.

I anticipate that students in this interdisciplinary course may come from different intellectual backgrounds, and thus there may be substantial differences in terms of familiarity with the concepts. I expect you to be patient with your fellow students, and hope that you will help one another in learning the material. I also hope that you will ask questions when something in the course is confusing or unfamiliar! There are no stupid questions.

Please know that my door is open to you if you wish to bring any issues to my attention. I especially encourage you to come to me if something that I say or do as an instructor makes you feel unwelcome in the course.

**Pronouns Policy Statement:** Everyone has the right to be addressed by the name and pronouns that they use for themselves. You can indicate your preferred/chosen first name and pronouns on SPIRE, or indicate them to me directly. I will do my best to ensure that I address you with your chosen name and pronouns. Please remember: A student's chosen name and pronouns are to be respected at all times in the classroom.

**Syllabus:**

**Module 0: Introduction to computational biology and relevant concepts in biology**
Overview of computational biology and its interdisciplinary nature
Introduction to basic biological concepts, including genes, genomes, and genetic variation

**Module 1: Sequence search and alignment**
Concepts underlying sequence alignment
Probabilistic foundations of alignment
Sequence scoring matrices
Needleman-Wunsch algorithm and extensions (eg Smith-Waterman algorithm)
Fast sequence database search using BLAST
*Models and Algorithms*: Needleman-Wunsch algorithm

**Module 2: From sequencing technology to genome sequences**
Introduction to DNA sequencing technologies and read formats
*De novo* DNA sequence assembly algorithms (for short-read sequencing)
Analysis of long-read sequencing data
Reference-based assembly algorithms and tools
*Models and Algorithms*: Burroughs-Wheeler transforms, De Bruijn Graphs

**Module 3: Annotating genome sequences with functional information**
Introduction to genome annotation and its importance
Gene prediction algorithms: *ab initio* and comparative methods
PFAM domains
*Models and Algorithms*: Hidden Markov models

**Module 4: Phylogeny Construction**
Introduction to phylogenetics and evolutionary relationships
Phylogenetic reconstruction algorithms
Phylogenetic tree visualization and interpretation
Gene trees vs. species trees, horizontal gene transfer, and the pangenome
*Models and Algorithms:* Binary trees, continuous-time Markov chains, Jukes-Cantor model

**Module 5: Population Genetics and Tests for Selection**
Basic concepts in population genetics: allele frequencies, genetic drift, natural selection
Detecting signatures of selection (eg, dN/dS, linkage disequilibrium)
Analysis of genomic variation in populations.
*Models and algorithms:* random walks, Wright-Fisher Model

**Module 6: Genome-Wide Association Studies (GWAS)**

Introduction to GWAS and its role in identifying genetic factors associated with traits and diseases
Statistical methods for identifying significant genetic variants
Heritability and genetic risk scores
*Models and algorithms:* linear and logistic regression, significance testing, variance

**Module 7: Mutation Effect Prediction in proteins**
Introduction to genetic mutations and their impact on proteins and non-coding regions
Functional consequences of genetic variants: missense, nonsense, frameshift, etc.
Prediction of protein structure and function changes
Recent deep learning methods used to predict mutation effects
*Models and algorithms:* auto-encoders, transformer-based models

**Module 8: Mutation effect prediction in non-coding regions**
Introduction to non-coding region functions, gene regulation, epigenetics, and chromatin
Laboratory methods for measuring function of non-coding regions
The ENCODE project to annotate non-coding DNA
Machine learning methods to predict function in non-coding regions
*Models and algorithms:* convolutional neural network, transformers

**Module 9: Gene Expression Analysis: RNA-Seq and Single-Cell RNA-Seq**
Introduction to gene expression analysis using RNA sequencing (RNA-Seq) data.
Preprocessing and quality control of RNA-Seq data.
Differential gene expression analysis and functional enrichment analysis.
Introduction to single-cell RNA-Seq analysis and clustering techniques.
*Models and algorithms:* PCA, t-SNE

**Module 10: Protein Structure Prediction**
Introduction to protein structure and its importance
Protein structure prediction methods: homology modeling, ab initio methods
Recent advances: AlphaFold
Evaluation of predicted protein structures
Disordered proteins and the challenges still to come
*Models and algorithms:* AlphaFold

Class Schedule:

**Feb 1**