BCH394P/BCH364C Systems Biology & Bioinformatics (course # 54960 / 54860)

Spring 2025 Tues/Thurs 9:30 – 11:00 AM WEL 2.246

Instructor: Prof. Edward Marcotte marcotte@utexas.edu
Office hours: Mon 4 PM – 5 PM On the class Zoom channel

TA: Zoya Ansari zansari@utexas.edu
Python/coding help hours: Tues 1-2/Fri 1-2 in MBB 3.304 or by appointment on Zoom
Discussion/Q&A on Canvas
Course web page: **http://marcottelab.org/index.php/BCH394P_BCH364C_2025**

Open to graduate students and upper division undergrads (with permission) in natural sciences and engineering.
Prerequisites: Basic familiarity with molecular biology, statistics & computing, but realistically, it is expected that students will have extremely varied backgrounds. UGs have additional prerequisites listed in the catalog.

An introduction to systems biology and bioinformatics, emphasizing quantitative analysis of high-throughput biological data, and covering typical data, data analysis, and computer algorithms. Topics will include introductory probability and statistics, basics of Python programming, protein and nucleic acid sequence analysis, genome sequencing and assembly, proteomics, analysis of large-scale gene expression data, data clustering & classification, biological pattern recognition, gene and protein networks, AI/machine learning, and protein 3D structure prediction/design.

** Note that this is not a course on practical sequence analysis or using web-based tools. Although we will use a number of these to help illustrate points, the focus of the course will be on learning the underlying algorithms and exploratory data analyses and their applications, esp. in high-throughput biology. By the end of the course, students will know the fundamentals of important algorithms in bioinformatics and systems biology, will be able to design and implement computational studies in biology, and will have performed an element of original computational biology research. **

Most of the lectures will be from research articles and slides. For sequence analysis, there will be an
Optional text: *Biological sequence analysis,* Durbin, Eddy, Krogh, Mitchison, Cambridge Univ. Press
(ebook available from Amazon, ~$13 to 32.00)

For biologists extremely rusty on their stats, *The Cartoon Guide to Statistics* (Gonick/Smith) is very good (really!), but a particularly good (and free online) intro stats text with Python examples is *An Introduction to Statistical Learning* (James/Witten/Hastie/Tibshirani/Taylor).

We will be learning some Python programming. The class web site has a list of recommendations for books and resources to help you better learn Python and we'll add more through the semester.

Online homework will be assigned and evaluated using the free bioinformatics web resource Rosalind (http://rosalind.info/faq/). **Enroll here: https://rosalind.info/classes/enroll/8cf0c8d95f/**

No exams will be given. Grades will be based on online homework (counting 30% of the grade), **3 problem sets** (given every 2-3 weeks and counting 15% each towards the final grade) **and a course project** (25% of final grade), which can be collaborative (1-3 students/project). The course project will consist of a research project on a bioinformatics topic chosen by the student (with approval by the instructor) containing an element of independent computational biology research (e.g., calculation, programming, database analysis, etc.). This will be

turned in as a link to a web page. **The final project is due by 10 PM, April 16, 2025. The last 3 classes will be spent presenting your projects to each other. (Presentations count for 5/25 points of the project grade.)**

All projects and homework will be turned in electronically and time-stamped. No makeup work will be given. Instead, all students have 5 days of free "late time" (for the entire semester, NOT per project, and counting weekends/holidays). For projects turned in late, days will be deducted from the 5-day total (or what remains of it) by the number of days late (in 1-day increments, rounding up, *e.g.* 10 minutes late = 1 day deducted). Once the full 5 days have been used up, assignments will be penalized 10 percent per day late (rounding up), *e.g.*, a 50-point assignment turned in 1.5 days late would be penalized 20%, or 10 points.

Homework, problem sets, and the project total to a possible 100 points. There will be no curving of grades, nor will grades be rounded up. We'll use the plus/minus grading system: A= 92 and above, A-=90 to 91.99, etc. Here are the grade cutoffs: $92\% \leq$ A, $90\% \leq$ A- $< 92\%$, $88\% \leq$ B+ $< 90\%$, $82\% \leq$ B $< 88\%$, $80\% \leq$ B- $< 82\%$, $78\% \leq$ C+ $< 80\%$, $72\% \leq$ C $< 78\%$, $70\% \leq$ C- $< 72\%$, $68\% \leq$ D+ $< 70\%$, $62\% \leq$ D $< 68\%$, $60\% \leq$ D- $< 62\%$, F $< 60\%$.

If, at some point, we have to go into coronavirus/flu/etc lockdown, that portion of the class will be web-based. We will hold lectures by Zoom during the normally scheduled class time. Log in to the UT Canvas class page for the link, or, if you are auditing, email the TA and we will send the link by return email. Slides will be posted before class so you can follow along with the material. We'll record the lectures & post the recordings afterward on Canvas so any of you who might be in other time zones or otherwise be unable to make class will have the opportunity to watch them. Note that the recordings will only be available on Canvas and are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of this restriction could lead to Student Misconduct proceedings.

Students are welcome to discuss ideas and problems with each other, but **all programs, Rosalind homework, problem sets, and written solutions should be performed independently** (except the final collaborative project). Students are expected to follow the UT honor code. **Cheating, plagiarism, copying, & reuse of prior homework, projects, or *programs* from CourseHero, Github, or any other sources are all *strictly forbidden* and constitute breaches of academic integrity and cause for dismissal with a failing grade, possibly expulsion (https://deanofstudents.utexas.edu/conduct/academicintegrity.php).** In particular, no materials used in this class, including, but not limited to, lecture hand-outs, videos, assessments (papers, projects, homework assignments), in-class materials, review sheets, and additional problem sets, may be shared online or with anyone outside of the class unless you have the instructor's explicit, written permission. Any materials found online (e.g., in CourseHero) that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

The use of artificial intelligence tools (such as ChatGPT or Github co-pilot) in this class shall be permitted on a limited basis for programming assignments. You are also welcome to seek my prior approval to use AI writing tools on any assignment. In either instance, AI writing tools should be used with caution and proper citation, and any use of AI should be properly attributed. Using AI writing tools without my permission or authorization, or failing to properly cite AI even where permitted, shall constitute a violation of UT Austin's Institutional Rules on academic integrity.

Students with disabilities may request appropriate academic accommodations from Disability and Access.

We will cover the following topics, approximately in this order:

BASICS OF PYTHON PROGRAMMING

Introduction to Rosalind
A Python programming primer for non-programmers
Rosalind help & programming Q/A, new aids for learning programming

BIOLOGICAL SEQUENCE ANALYSIS

Substitution matrices (BLOSSUM, PAM) & sequence alignment
Protein and nucleic acid sequence alignments, dynamic programming
Sequence profiles
BLAST! (the algorithm), MMSeqs2, & FoldSeek
Biological databases
Markov processes and Hidden Markov Models

GENOMES, PROTEOMES, & "BIG BIOLOGY"

Gene finding algorithms
Genome sequencing and assembly
An introduction to large gene expression data sets
Promoter and motif finding, Gibbs sampling
Guest lecture: Intro to NGS analysis and the CBRF core

MACHINE LEARNING/AI

Clustering algorithms, hierarchical, k-means, self-organizing maps, force-directed maps, UMAP/tSNE
Classifiers, k-nearest neighbors, precision/recall/ROC analysis
Principal component analysis and data transformations
Guest lecture: Protein 3D structure prediction, incl. AlphaFold
Guest lecture: AI/deep neural networks and large language models

SYNTHETIC BIOLOGY & PROTEIN DESIGN

Protein 3D design/engineering, RFDiffusion/ProteinMPNN, ColabFold
Synthetic biology & genome design

***** THE FINAL COURSE PROJECT IS DUE by 10 PM, April 16, 2025 *****

The last 3 class days will be devoted to presenting your projects to the rest of the class.

Note that there is NO CLASS over spring break (March 18 & March 20).