BCH394P/BCH364C Systems Biology & Bioinformatics (course # 54540 / 54450)

Spring 2022 Tues/Thurs 11 – 12:30 PM 1st 2 wks virtual, then in WEL 2.110

Instructor: Prof. Edward Marcotte marcotte@utexas.edu
Office hours: Wed 11 AM – 12 noon On the class Zoom channel

TA: Muyoung Lee ml49649@utexas.edu
Office hours: Mon 1-2/Fri 11-12 On the class Zoom channel Slack: ut-sp22-bioinfo.slack.com
Course web page: **http://marcottelab.org/index.php/BCH394P_BCH364C_2022**

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, as 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, synthetic biology, analysis of large-scale gene expression data, data clustering, biological pattern recognition, and gene and protein networks.

** 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, be able to design and implement computational studies in biology, and 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 rusty on their stats, *The Cartoon Guide to Statistics* (Gonick/Smith) is very good (really!).

We will also be learning some Python programming. The class web site has a list of recommendations for books and resources to help you better learn Python.

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/3862a679ae/**

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 midnight, April 25, 2022. 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% ≤ A, 90% ≤ A- < 92%, 88% ≤ B+ < 90%, 82% ≤ B < 88%, 80% ≤ B- < 82%, 78% ≤ C+ < 80%, 72% ≤ C < 78%, 70% ≤ C- < 72%, 68% ≤ D+ < 70%, 62% ≤ D < 68%, 60% ≤ D- < 62%, F < 60%.

Since we will be in coronavirus lockdown at the start of this semester, this 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 timezones 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 by a student 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.

We'll cover the following topics, approximately in this order:

BASICS OF PROGRAMMING

Introduction to Rosalind
A Python programming primer for non-programmers
Rosalind help & programming Q/A

BIOLOGICAL SEQUENCE ANALYSIS

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

GENOMES, PROTEOMES, & "BIG BIOLOGY"

Gene finding algorithms
Genome assembly & how the human genome was sequenced
We'll (probably) attempt a live, in-class (or on zoom if need be) demo of nanopore DNA sequencing!
An introduction to large gene expression data sets
Promoter and motif finding, Gibbs sampling
Clustering algorithms, hierarchical, k-means, self-organizing maps, force-directed maps
Classifiers, k-nearest neighbors, Mahalonobis distance
Principal component analysis and data transformations

NETWORK & SYNTHETIC BIOLOGY

Biological networks: metabolic, signaling, graphs, regulatory
Deep homology and the evolution of traits
Designing, simulating, and building gene circuits
Genome design and synthesis

Also, we'll have several guest lectures sprinkled throughout the semester on:
NGS best practices; Mass spectrometry shotgun proteomics; Protein 3D modeling; Deep learning

***** THE FINAL COURSE PROJECT IS DUE by midnight, April 25, 2022 *****

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 15 & March 17).

We're also reserving the last class day, May 5, as an emergency flex day. The current plan is for classes to end on May 3 and for there to be NO CLASS on May 5, but if weather, the pandemic, etc, leads to loss of lecture days, we'll vote as a class to extend class to May 5.