BCH339N Systems Biology & Bioinformatics (course # 54040)

Spring 2016 Tues/Thurs 11 – 12:30 PM BUR 212

Instructor: Prof. Edward Marcotte marcotte@icmb.utexas.edu
Office hours: Mon 4 – 5 PM MBB 3. 148BA

TA: Claire McWhite claire.mcwhite@utexas.edu
Office hours: Wed/Thurs 4–5 PM MBB 3.128A TA Phone: 512-232-3919
Course web page: **http://www.marcottelab.org/index.php/BCH339N_2016**

Open to biochemistry majors. Prerequisites: Biochemistry 339F or Chemistry 339K with a grade of at least C-. Requires basic familiarity with molecular biology & basic statistics, although varied backgrounds are expected.

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'll use a few), but rather on algorithms, exploratory data analyses and their applications in high-throughput biology. **

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
(available from Amazon, used from $15.00)

For students rusty on their stats, *The Cartoon Guide to Statistics* (Gonick/Smith) is very good (really!).

We will also be learning some Python programming. The following is highly recommended:
Python programming for beginners: http://www.codecademy.com/tracks/python
Online homework will be assigned and evaluated using the free bioinformatics web resource Rosalind (http://rosalind.info/faq/). **Enroll here: http://rosalind.info/classes/enroll/c5be9c4629/**

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 an independent course project** (25% of final grade). The course project will be focused on a specific gene & will involve bioinformatics research (e.g. calculation, programming, database analysis, etc.) developed over the semester in 5 mini-assignments (worth 4% each), which will be turned in as a link to a web page. **The project will be emailed as a web URL to the TA & Prof, developed through the semester and finished by midnight, April 27, 2016. The last three classes will be spent presenting your projects to each other. (The presentation will be worth 5%.)**

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%.

Students are welcome to discuss ideas and problems with each other, but **all programs, Rosalind homework, and written solutions should be performed independently** (except the final collaborative project).

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

BASICS OF PROGRAMMING

Introduction to Rosalind
A Python programming primer for non-programmers

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
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

Guest lectures will also be given by researchers from the UT Center for Systems and Synthetic Biology (CSSB) on topics at the cutting edge of systems biology and bioinformatics, including:

Homology, orthology, and evolutionary trees
3D modeling of protein structures
Next- (& next-next-) generation DNA and RNA sequencing
Mass spectrometry shotgun proteomics
Genome engineering

***** THE FINAL GENE PAGES PROJECT IS DUE by midnight, April 27, 2016 *****

The last three 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).