BCH364C/BCH394P Systems Biology/Bioinformatics
(course # 55120/55210)
Spring 2017 Tues/Thurs 11 – 12:30 PM GDC 4.302

Instructor: Prof. Edward Marcotte marcotte@icmb.utexas.edu
Office hours: Wed 11 AM – 12 noon MBB 3. 148BA

TA: Azat Akhmetov azat@utexas.edu
Office hours: Mon/Wed 3 – 4 PM MBB 3.128A or adjacent conference room
Phone: 512-232-3919
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.

Note that this is a GRADUATE class, with a few spots open to advanced undergraduates. There is a different version (CH339N) for undergraduate students in alternate years.

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.

** 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. **
Books

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 $26.85)

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. I highly recommend...

**Python programming for beginners:**
http://www.codecademy.com/tracks/python

Grading

No exams. Instead, grades will be based on:

- **Online programming homework**
  (10 points each and counting 30% of the final grade)
- **3 problem sets**
  (15 points each and counting 45% of the final grade)
- **A course project** that you will develop over the semester & present in the last 2 days of class (25% of final grade)

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.) turned in as a web URL (20%) and presented in class (5%).

The project will be emailed as a web URL to the TA & I, developed through the semester and finished by midnight, April 27, 2017. The last 2 classes will be spent presenting your projects.
Late policy

• 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”. This is for the entire semester, NOT per project, and counting weekends/holidays just like any other day.
  • For projects turned in late, days will be deducted from the 5 day total (or what remains of it) by the # of days late.
  • Deductions are in 1 day increments, rounding up e.g. 10 minutes late = 1 day deducted.
  • Once the 5 days are used up, assignments will be penalized 10% / day late (rounding up), e.g., a 50 point assignment turned in 1 ½ days late would be penalized 20%, or 10 points.

Online homework will be via Rosalind:  http://rosalind.info/faq/

Enroll specifically for BCH364C/BCH394P at:
http://rosalind.info/classes/enroll/b22533ccd7/

The first homework will be due (in Rosalind) by midnight, Jan 26.
If you’re feeling restless/adventurous...

Click here to turn in your answer
...there are quite a few good bioinformatics problems in the archives.

Expectations on working together

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 presentation.

tl;dr: study/discuss together
do your own programming/writing/project

 collaborate on the final presentation
• “By submitting *as your own work* any unattributed material that you obtained from other sources, you have committed plagiarism.”
• Copying homework solutions from other students or internet sources is cheating, collusion, and/or plagiarism.
• Software and computer code are legally considered in the same framework as other written works. Copying code directly without attribution is plagiarism.

http://deanofstudents.utexas.edu/sjs/acadint_plag_collab.php
http://deanofstudents.utexas.edu/sjs/acadint_whatis.php
• You can use the internet to get ideas, programming suggestions and syntax, but downloading completed answers to assigned questions and submitting these as your own work is cheating/plagiarism.

• Copying entire programs verbatim from marked repositories offering Rosalind homework solutions is cheating and plagiarism.

Similarly, downloading or otherwise obtaining solutions to homework problems from previous students (or Coursehero/similar sites) and turning these in as your own work is cheating, collusion, and/or plagiarism.
http://deanofstudents.utexas.edu/sjs/acadint_conseq.php

Why are we here? (practically, not existentially)
Our current knowledge of human metabolism...

- Total number of reactions: 7,440
- Total number of metabolites: 5,063
- Number of unique metabolites: 2,626
- Number of metabolites in extracellular space: 642
- Number of metabolites in cytoplasm: 1,878
- Number of metabolites in mitochondrion: 754
- Number of metabolites in nucleus: 165
- Number of metabolites in endoplasmic reticulum: 570
- Number of metabolites in peroxisome: 435
- Number of metabolites in lysosome: 302
- Number of metabolites in Golgi apparatus: 317
- Number of transcripts: 2,194
- Number of unique genes: 1,789

Pales beside the phenomenal drop in DNA sequencing costs...

Cost per Raw Megabase of DNA Sequence

& the corresponding explosion of DNA sequencing data...

Growth of GenBank (1982 - 2008)

& the corresponding explosion of DNA sequencing data...

Here are the latest statistics...

December 2016:
224 billion bp
 +
1.8 trillion bp DNA whole genome shotgun sequencing

Which basically means GenBank is falling behind more every year!


We have no choice!

Biologists are now faced with a staggering deluge of data, growing at exponential rates.

Bioinformatics offers tools and approaches to understand these data and work productively, and to build algorithmic models that help us better understand biological systems.

We’ll learn some of the important basic concepts in this field, along with getting exposed to key technologies driving the field forward.
Specifically...

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
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
Network alignment and comparisons, network organization
Deep homology and the evolution of traits
Designing, simulating, and building gene circuits
Genome design and synthesis
& 5 expert guest lectures on:

Homologs, orthologs, and paralogs
Next- (& next-next-) generation DNA and RNA sequencing
Overview of mass spectrometry shotgun proteomics
Protein 3D structural modeling
Genome engineering

THE FINAL COURSE PROJECT IS DUE by midnight, April 27, 2017

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