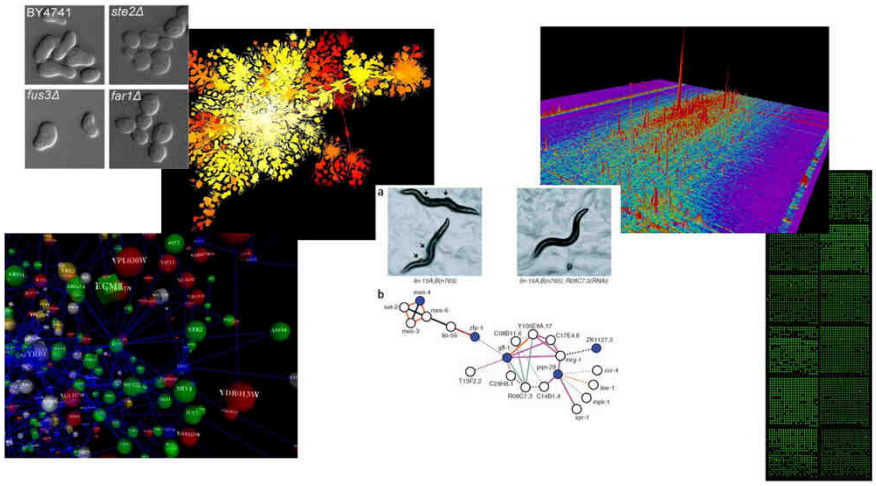
BCH339N Systems Biology/Bioinformatics
(course # 54040)
Spring 2016 Tues/Thurs 11 – 12:30 PM BUR 212



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

TA: Claire McWhite **claire.mcwhite@utexas.edu**
Office hours: Wed/Thurs 4 – 5 PM **MBB 3.128A**
Phone: 512-232-3919

Probably the most important slide today!

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.

Note that this is an UNDERGRADUATE class. There is a different version intended for graduate students in alternate years (CH391L).

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 3 days of class (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 (4% each) and presented in class (5%).

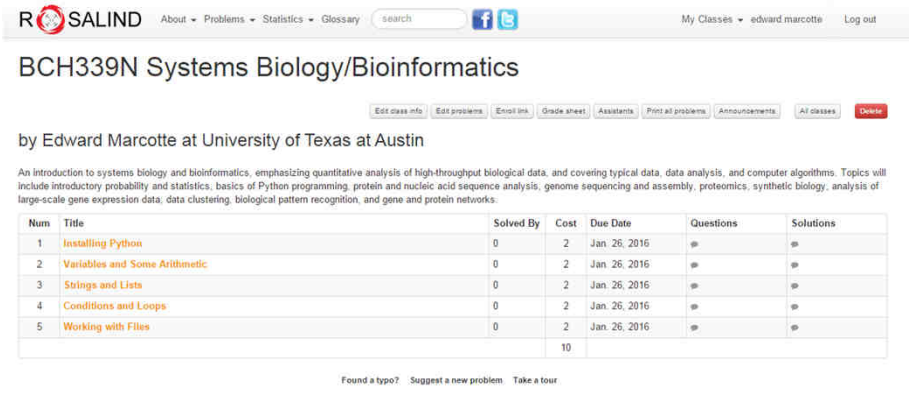
The project will be emailed as a web URL to the TA & I, developed through the semester and finished by midnight, April 27, 2016. The last three classes will be spent presenting your projects to each other.

Late policy

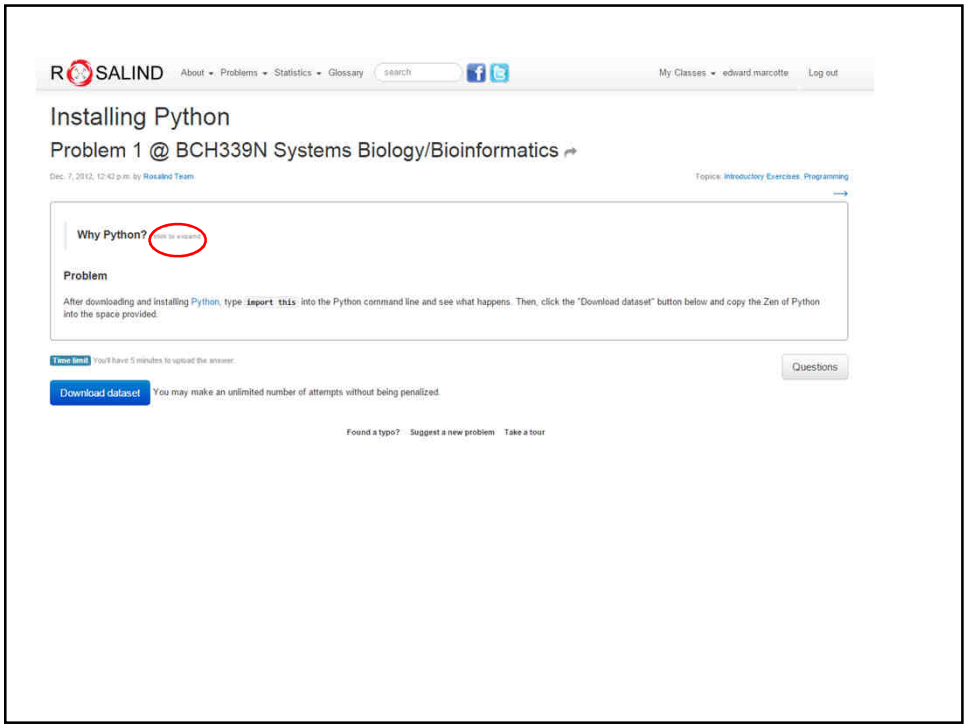
- **All projects and homework will be turned in electronically and time-stamped.**
- **No makeup work will be given.**
- **Instead, all students have 5 days of free "late time".**
  This is for the entire semester, NOT per project, and counting weekends/holidays just like any other day.
  - For projects 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.

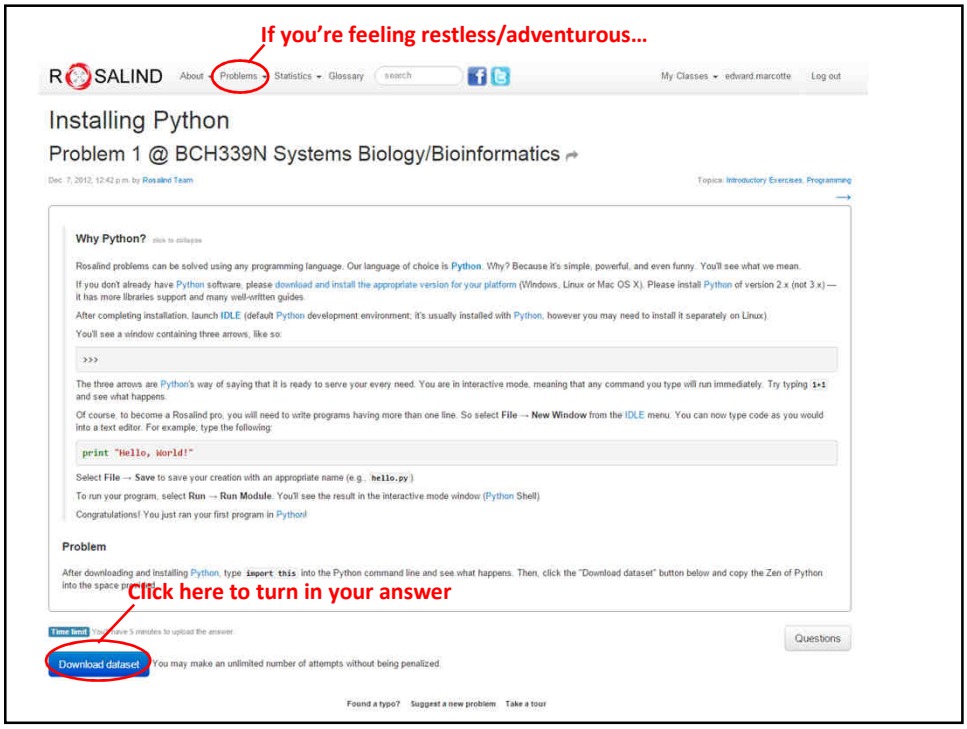
Online homework will be via *Rosalind*: http://rosalind.info/faq/

Enroll specifically for BCH339N at:
http://rosalind.info/classes/enroll/c5be9c4629

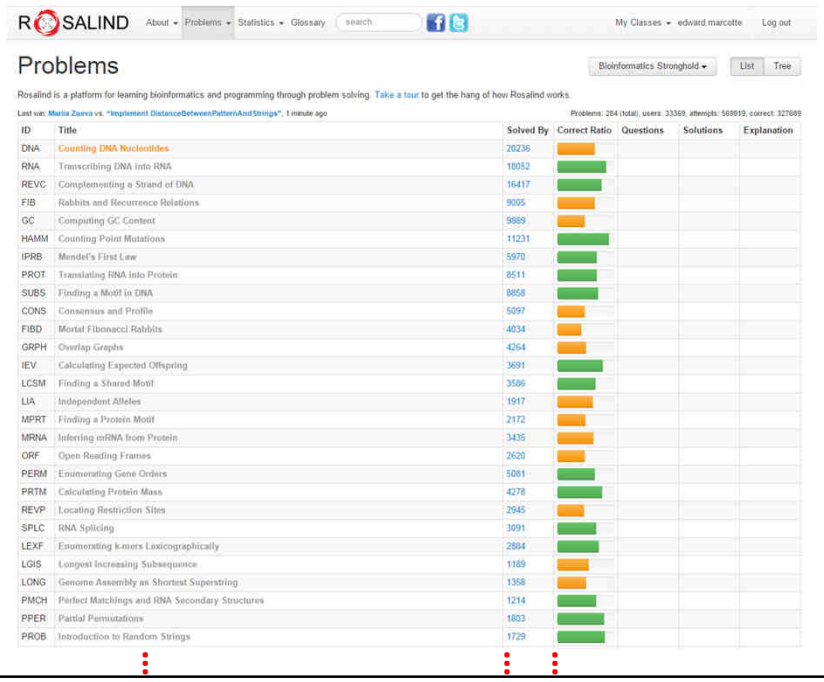


The first homework will be due (in Rosalind) by midnight, Jan 26.


ROSALIND
About
Problems
Statistics
Glossary
search
My Classes
edward.marcotte
Log out
Installing Python
Problem 1 @ BCH339N Systems Biology/Bioinformatics
Dec. 7, 2012, 12:42 p.m. by Rosalind Team
Topics: Introductory Exercises, Programming
Why Python?
Problem
After downloading and installing Python, type import this into the Python command line and see what happens. Then, click the "Download dataset" button below and copy the Zen of Python into the space provided.
Time limit: You'll have 5 minutes to upload the answer.
Questions
Download dataset
You may make an unlimited number of attempts without being penalized.
Found a typo?
Suggest a new problem
Take a tour


If you're feeling restless/adventurous…
ROSALIND
About
Problems
Statistics
Glossary
search
My Classes
edward.marcotte
Log out
Installing Python
Problem 1 @ BCH339N Systems Biology/Bioinformatics
Dec. 7, 2012, 12:42 p.m. by Rosalind Team
Topics: Introductory Exercises, Programming
Why Python?
Rosalind problems can be solved using any programming language. Our language of choice is Python. Why? Because it's simple, powerful, and even funny. You'll see what we mean.
If you don't already have Python software, please download and install the appropriate version for your platform (Windows, Linux or Mac OS X). Please install Python of version 2.x (not 3.x) — it has more libraries support and many well-written guides.
After completing installation, launch IDLE (default Python development environment; it's usually installed with Python, however you may need to install it separately on Linux).
You'll see a window containing three arrows, like so:
>>>
The three arrows are Python's way of saying that it is ready to serve your every need. You are in interactive mode, meaning that any command you type will run immediately. Try typing 1+1 and see what happens.
Of course, to become a Rosalind pro, you will need to write programs having more than one line. So select File → New Window from the IDLE menu. You can now type code as you would into a text editor. For example, type the following:
print "Hello, World!"
Select File → Save to save your creation with an appropriate name (e.g., hello.py).
To run your program, select Run → Run Module. You'll see the result in the interactive mode window (Python Shell).
Congratulations! You just ran your first program in Python!
Problem
After downloading and installing Python, type import this into the Python command line and see what happens. Then, click the "Download dataset" button below and copy the Zen of Python into the space provided.
Click here to turn in your answer
Time limit: You'll have 5 minutes to upload the answer.
Questions
Download dataset
You may make an unlimited number of attempts without being penalized.
Found a typo?
Suggest a new problem
Take a tour

…there are quite a few good bioinformatics problems in the archives.



ROSALIND About Problems Statistics Glossary search My Classes edward.marcotte Log out

Problems

Bioinformatics Stronghold List Tree

Rosalind is a platform for learning bioinformatics and programming through problem solving. Take a tour to get the hang of how Rosalind works.

Last win: Marija Zueva vs. "Implement DistanceBetweenPatternAndStrings", 1 minute ago

Problems: 284 (total), users: 33369, attempts: 569919, correct: 327669

ID	Title	Solved By	Correct Ratio	Questions	Solutions	Explanation
DNA	Counting DNA Nucleotides	20236				
RNA	Transcribing DNA into RNA	18052				
REVC	Complementing a Strand of DNA	16417				
FIB	Rabbits and Recurrence Relations	9005				
GC	Computing GC Content	9889				
HAMM	Counting Point Mutations	11231				
IPRB	Mendel's First Law	5978				
PROT	Translating RNA into Protein	8511				
SUBS	Finding a Motif in DNA	8858				
CONS	Consensus and Profile	5097				
FIBD	Mortal Fibonacci Rabbits	4034				
GRPH	Overlap Graphs	4264				
IEV	Calculating Expected Offspring	3691				
LCSM	Finding a Shared Motif	3586				
LIA	Independent Alleles	1917				
MPRT	Finding a Protein Motif	2172				
MRNA	Inferring mRNA from Protein	3435				
ORF	Open Reading Frames	2620				
PERM	Enumerating Gene Orders	5081				
PRTM	Calculating Protein Mass	4278				
REVP	Locating Restriction Sites	2945				
SPLC	RNA Splicing	3091				
LEXF	Enumerating k-mers Lexicographically	2884				
LGIS	Longest Increasing Subsequence	1189				
LONG	Genome Assembly as Shortest Superstring	1358				
PMCH	Perfect Matchings and RNA Secondary Structures	1214				
PPER	Partial Permutations	1803				
PROB	Introduction to Random Strings	1729				

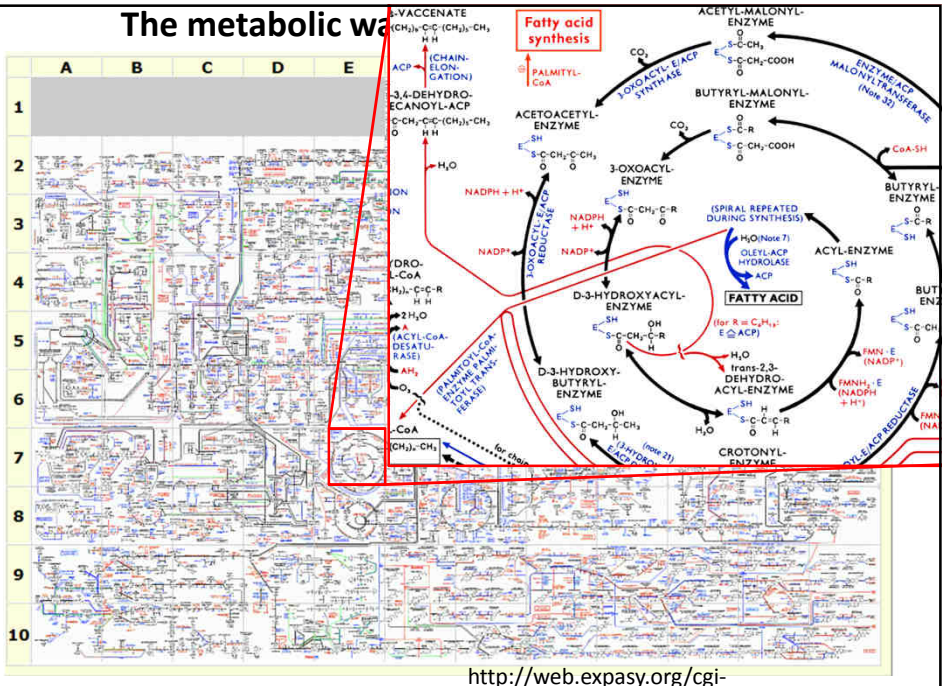
Expectations on working together

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

Why are we here? (practically, not existentially)

The metabolic wa



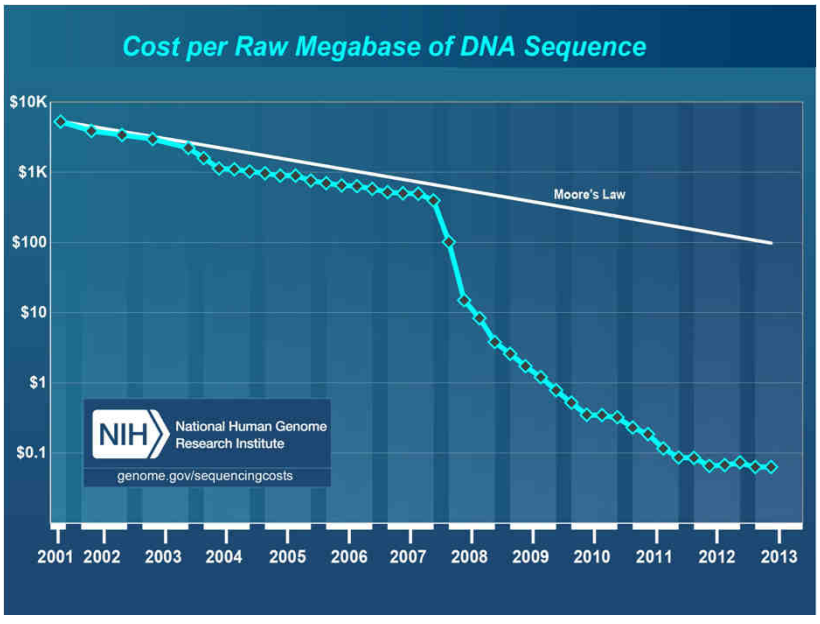
http://web.expasy.org/cgi-bin/pathways/show_thumbnails.pl

Our current knowledge of human metabolism…

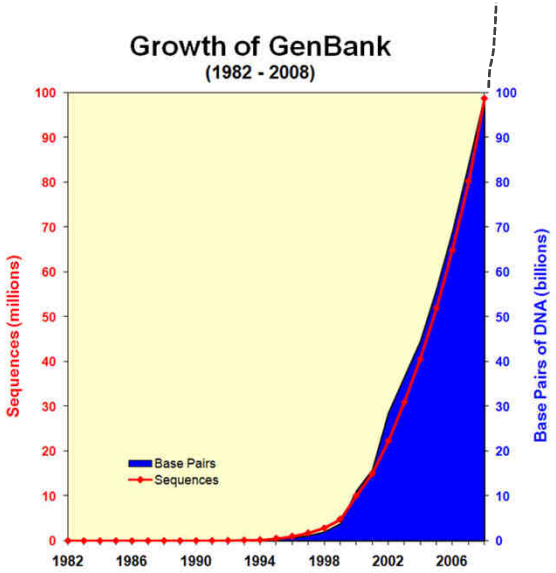
Total number of reactions	7,440
Total number of metabolites	5,063
Number of unique metabolites	2,626
Number of metabolites in extracellular space	642
Number of metabolites in cytoplasm	1,878
Number of metabolites in mitochondrion	754
Number of metabolites in nucleus	165
Number of metabolites in endoplasmic 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

Nat Biotechnol. 2013 May;31(5):419-25

Pales beside the phenomenal drop in DNA sequencing costs…

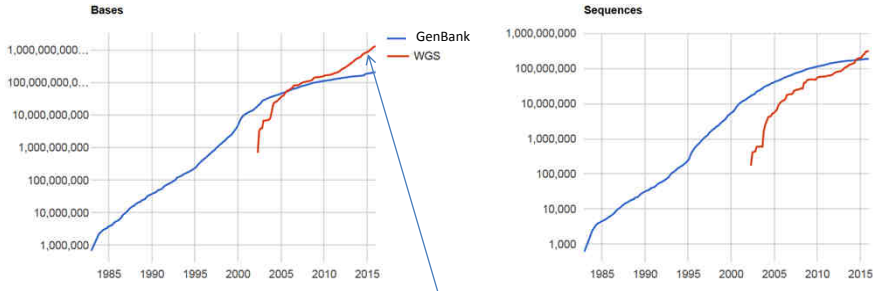


& the corresponding explosion of DNA sequencing data…



http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/
ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

& the corresponding explosion of DNA sequencing data…



Here are the latest statistics…

December 2015:
203 billion bp
+
1.3 trillion bp DNA
whole genome
shotgun sequencing

Which basically means GenBank is falling behind more every year!

http://www.ncbi.nlm.nih.gov/genbank/statistics

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

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

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.