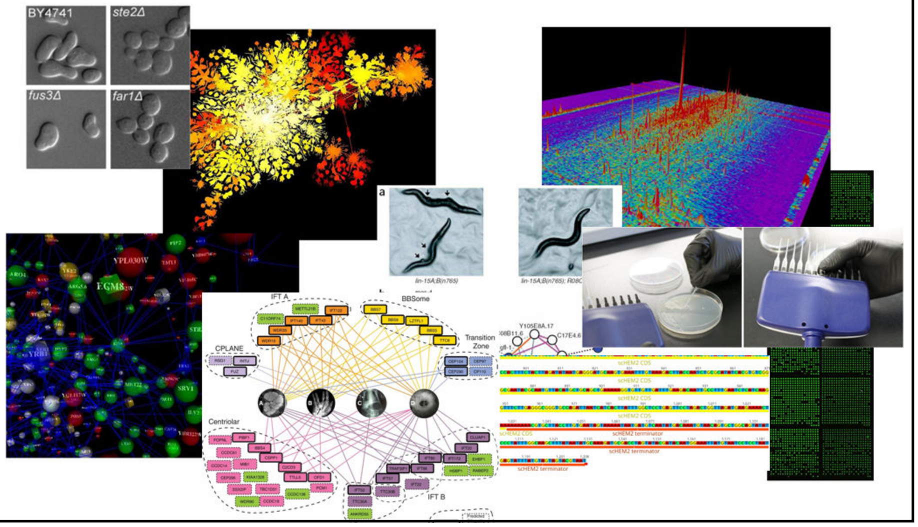
BCH394P/BCH364C Systems Biology & Bioinformatics

(course # 54044/53945)

Spring 2019 Tues/Thurs 11 – 12:30 PM JGB 2.202



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

TA: Caitie McCafferty clmccafferty@utexas.edu
Office hours: Mon 11 – 12/Fri 2 – 3 NHB 3.400B atrium
(or MBB 3.128B)

Phone: 512-232-3919

Probably the most important slide today!

Course web page:

http://www.marcottelab.org/index.php/BCH394P_BCH364C_2019

This is a graduate student class!

It is open to a small # (<10) of upper division undergrads in natural sciences and engineering.

UG prerequisites: Biochemistry 339F with a grade of at least B; Statistics and Data Sciences 328M (or Statistics and Scientific Computation 318M, 328M) with a grade of at least C-; and *consent of the instructor*. ← MOST IMPORTANT (no one currently has consent!)

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:
https://www.codecademy.com/learn/learn-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 2.5 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%).

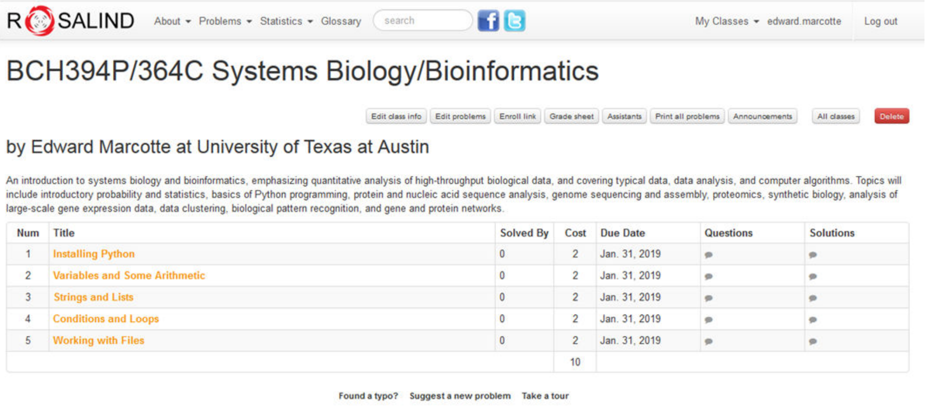
The project will be emailed as a web URL to the TA & I, developed through the semester and finished by midnight, April 29, 2019. The last few classes will be spent presenting your projects.

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 BCH394P/364C at:
http://rosalind.info/classes/enroll/127bf319a3/



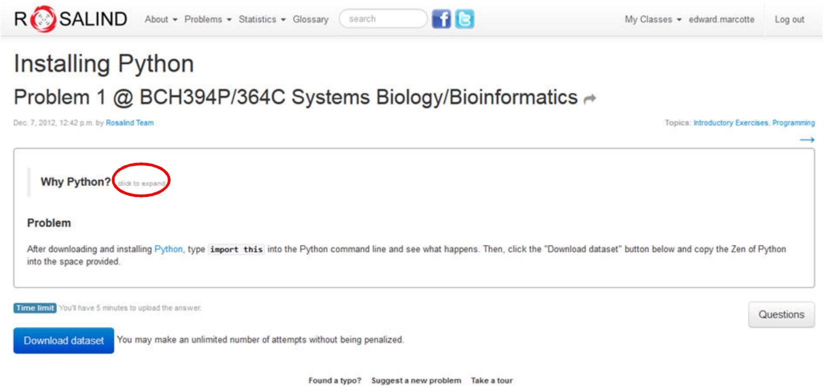
BCH394P/364C Systems Biology/Bioinformatics

by Edward Marcotte at University of Texas at Austin

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.

Num	Title	Solved By	Cost	Due Date	Questions	Solutions
1	Installing Python	0	2	Jan. 31, 2019	💬	💬
2	Variables and Some Arithmetic	0	2	Jan. 31, 2019	💬	💬
3	Strings and Lists	0	2	Jan. 31, 2019	💬	💬
4	Conditions and Loops	0	2	Jan. 31, 2019	💬	💬
5	Working with Files	0	2	Jan. 31, 2019	💬	💬
			10			

The first homework will be due (in Rosalind) by midnight, Jan 31.



Installing Python

Problem 1 @ BCH394P/364C Systems Biology/Bioinformatics

Dec. 7, 2012, 12:42 p.m. by Rosalind Team

Topics: Introductory Exercises, Programming

Why Python? click to expand

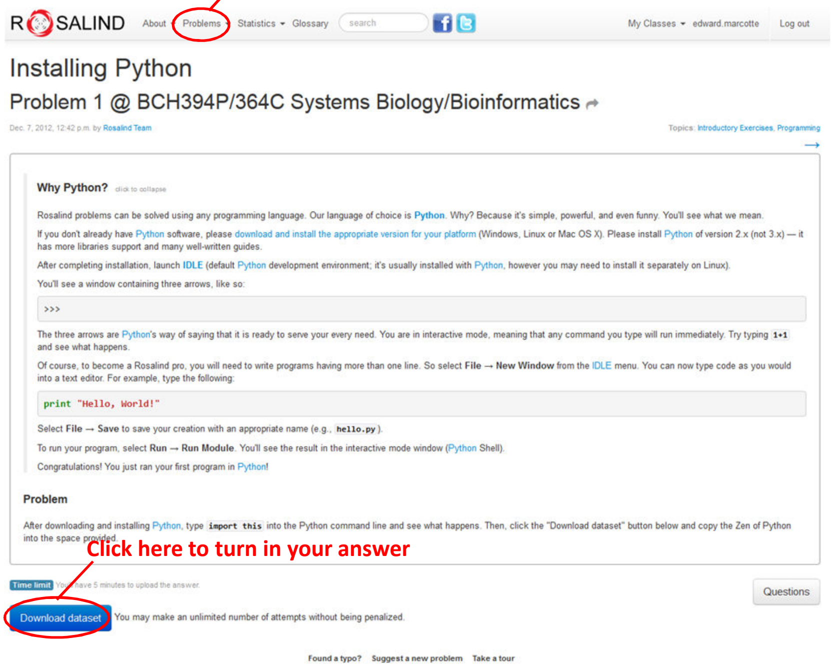
Problem

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.

You may make an unlimited number of attempts without being penalized.

If you're feeling restless/adventurous…



Installing Python

Problem 1 @ BCH394P/364C Systems Biology/Bioinformatics

Dec. 7, 2012, 12:42 p.m. by Rosalind Team

Topics: Introductory Exercises, Programming

Why Python? click to collapse

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

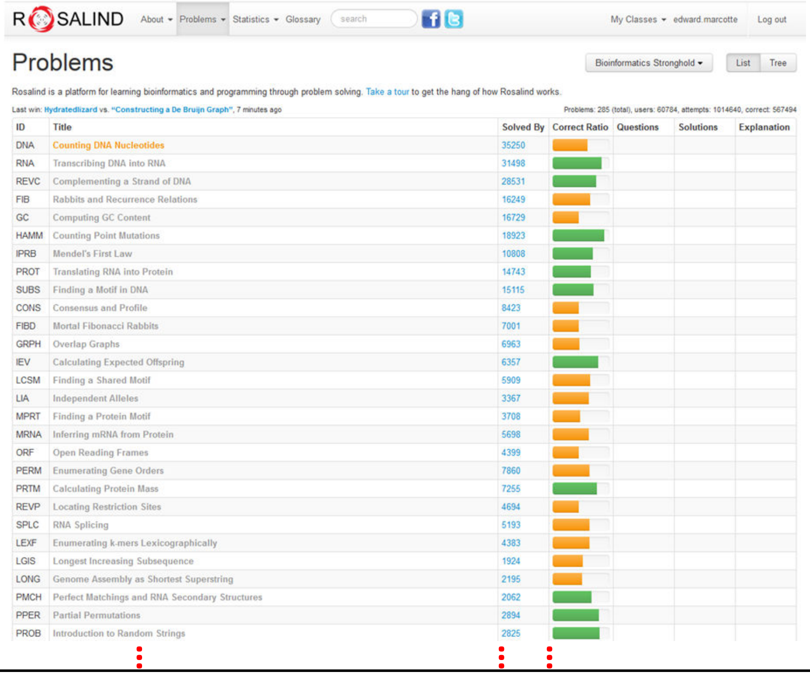
After downloading and installing Python, type `import this` into the Python command line and see what happens. Then, click the "Download dataset" button below and copy the Zen of Python into the space provided.

Click here to turn in your answer

Time limit You'll have 5 minutes to upload the answer.

You may make an unlimited number of attempts without being penalized.

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



Problems

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: Hydratedlizard vs. "Constructing a De Bruijn Graph", 7 minutes ago

Problems: 285 (total), users: 60784, attempts: 1014640, correct: 567494

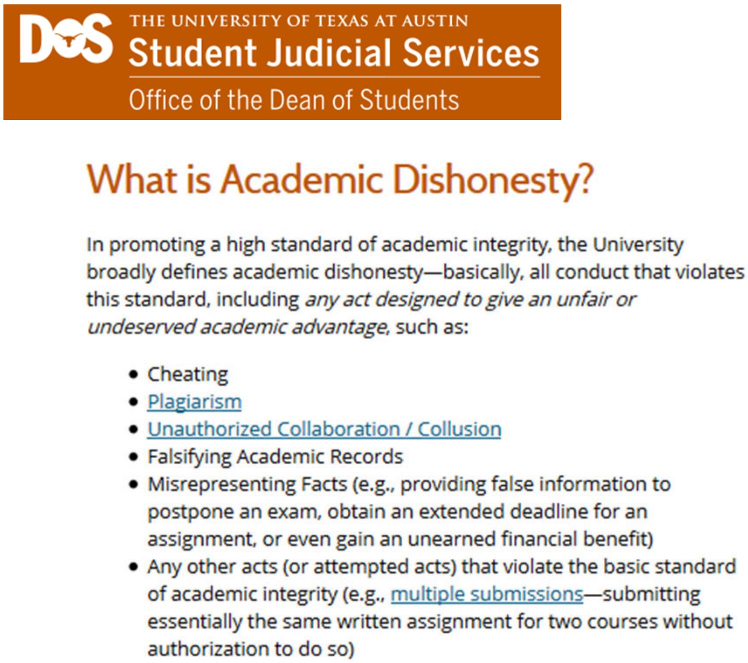
ID	Title	Solved By	Correct Ratio	Questions	Solutions	Explanation
DNA	Counting DNA Nucleotides	35250	🟧			
RNA	Transcribing DNA into RNA	31498	🟩			
REVC	Complementing a Strand of DNA	28531	🟩			
FIB	Rabbits and Recurrence Relations	16249	🟧			
GC	Computing GC Content	16729	🟧			
HAMM	Counting Point Mutations	18923	🟩			
IPRB	Mendel's First Law	10808	🟩			
PROT	Translating RNA into Protein	14743	🟩			
SUBS	Finding a Motif in DNA	15115	🟩			
CONS	Consensus and Profile	8423	🟧			
FIBD	Mortal Fibonacci Rabbits	7001	🟧			
GRPH	Overlap Graphs	6963	🟧			
IEV	Calculating Expected Offspring	6357	🟩			
LCSM	Finding a Shared Motif	5909	🟧			
LIA	Independent Alleles	3367	🟧			
MPRT	Finding a Protein Motif	3708	🟧			
MRNA	Inferring mRNA from Protein	5698	🟧			
ORF	Open Reading Frames	4399	🟧			
PERM	Enumerating Gene Orders	7860	🟧			
PRTM	Calculating Protein Mass	7255	🟩			
REVP	Locating Restriction Sites	4694	🟧			
SPLC	RNA Splicing	5193	🟧			
LEXF	Enumerating k-mers Lexicographically	4383	🟧			
LGIS	Longest Increasing Subsequence	1924	🟧			
LONG	Genome Assembly as Shortest Superstring	2195	🟧			
PMCH	Perfect Matchings and RNA Secondary Structures	2062	🟩			
PPER	Partial Permutations	2894	🟩			
PROB	Introduction to Random Strings	2825	🟩			
	⋮	⋮	⋮			

Expectations on working together

Students are welcome to discuss ideas and problems with each other, but **<u>all programs, Rosalind homework, problem sets, and written solutions should be performed independently</u>**,

→ except the final presentation.

tl;dr: study/discuss together
do your own programming/writing/project
collaborate on the final presentation



THE UNIVERSITY OF TEXAS AT AUSTIN
Student Judicial Services
Office of the Dean of Students

What is Academic Dishonesty?

In promoting a high standard of academic integrity, the University broadly defines academic dishonesty—basically, all conduct that violates this standard, including *any act designed to give an unfair or undeserved academic advantage*, such as:

- Cheating
- Plagiarism
- Unauthorized Collaboration / Collusion
- Falsifying Academic Records
- Misrepresenting Facts (e.g., providing false information to postpone an exam, obtain an extended deadline for an assignment, or even gain an unearned financial benefit)
- Any other acts (or attempted acts) that violate the basic standard of academic integrity (e.g., multiple submissions—submitting essentially the same written assignment for two courses without authorization to do so)

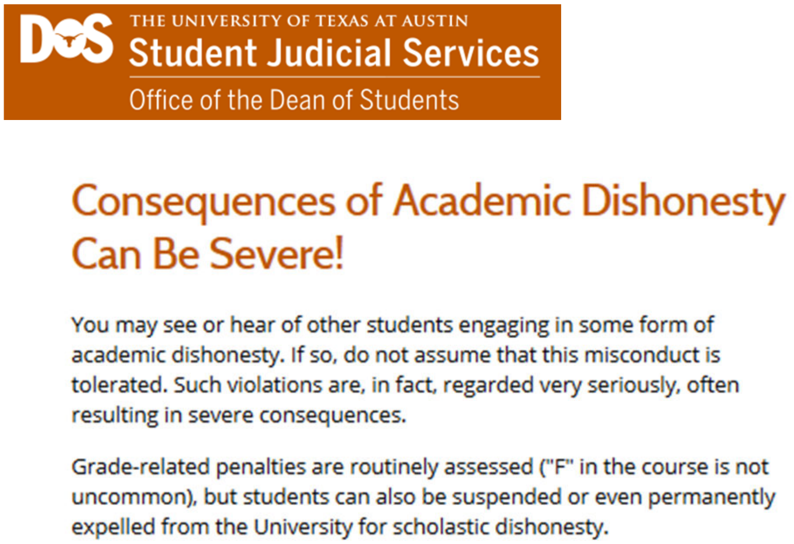
http://deanofstudents.utexas.edu/sjs/acadint_whatis.php

- “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.



DoS THE UNIVERSITY OF TEXAS AT AUSTIN
Student Judicial Services
Office of the Dean of Students

Consequences of Academic Dishonesty Can Be Severe!

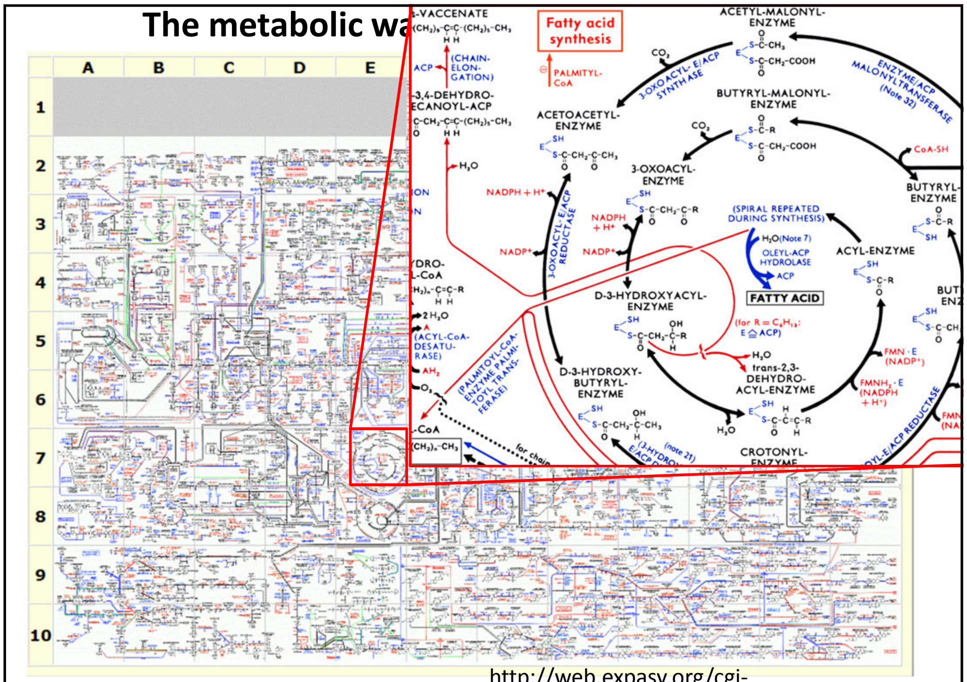
You may see or hear of other students engaging in some form of academic dishonesty. If so, do not assume that this misconduct is tolerated. Such violations are, in fact, regarded very seriously, often resulting in severe consequences.

Grade-related penalties are routinely assessed ("F" in the course is not uncommon), but students can also be suspended or even permanently expelled from the University for scholastic dishonesty.

http://deanofstudents.utexas.edu/sjs/acadint_conseq.php

Why are we here? (practically, not existentially)

The metabolic wa



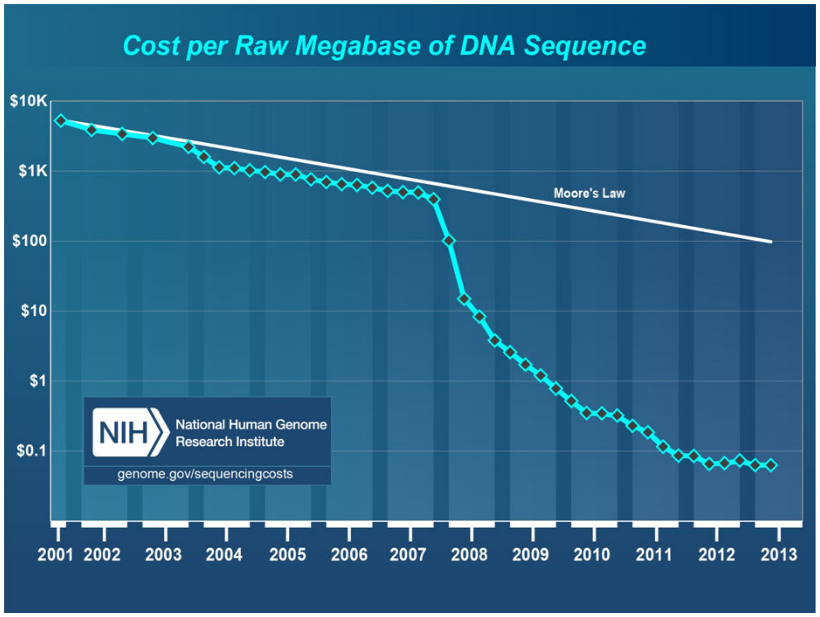
http://web.expasy.org/cgi-bin/pathways/show_thumbnails.pl

Our current knowledge of human metabolism…

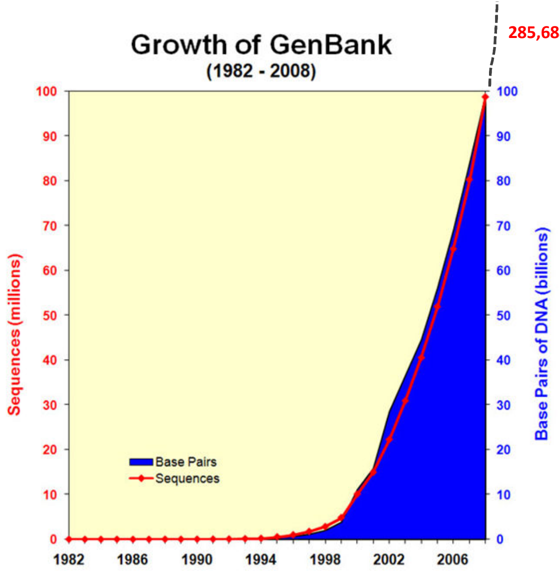
Total number of reactions	7,440
Total number of metabolites	5,063
Number of unique metabolites	2,626
Number of metabolites in extracellular space	642
Number of metabolites in cytoplasm	1,878
Number of metabolites in mitochondrion	754
Number of metabolites in nucleus	165
Number of metabolites in endoplasmic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


Cost per Raw Megabase of DNA Sequence

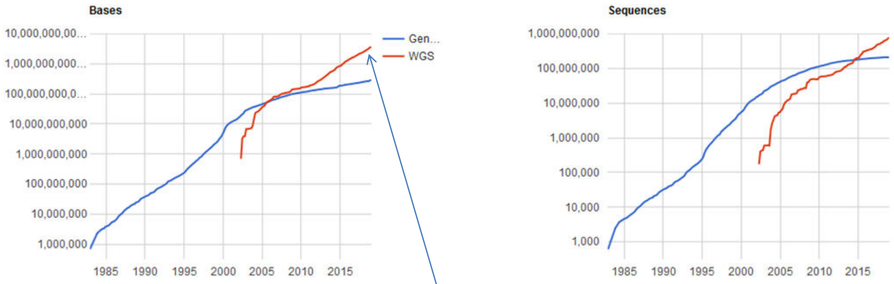
& the corresponding explosion of DNA sequencing data…


Growth of GenBank
(1982 - 2008)

As of Dec 2018:
285,688,542,186 base pairs from
211,281,415 sequences

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 2018:
286 billion bp
+
3.7 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
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
Classification algorithms
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

Plus, expert guest lectures on:

NGS best practices
Overview of mass spectrometry shotgun proteomics
Protein 3D structural modeling

Plus, plus:
we’ll attempt a live demo in-class of nanopore sequencing….

THE FINAL COURSE PROJECT IS DUE by midnight, April 29, 2019

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