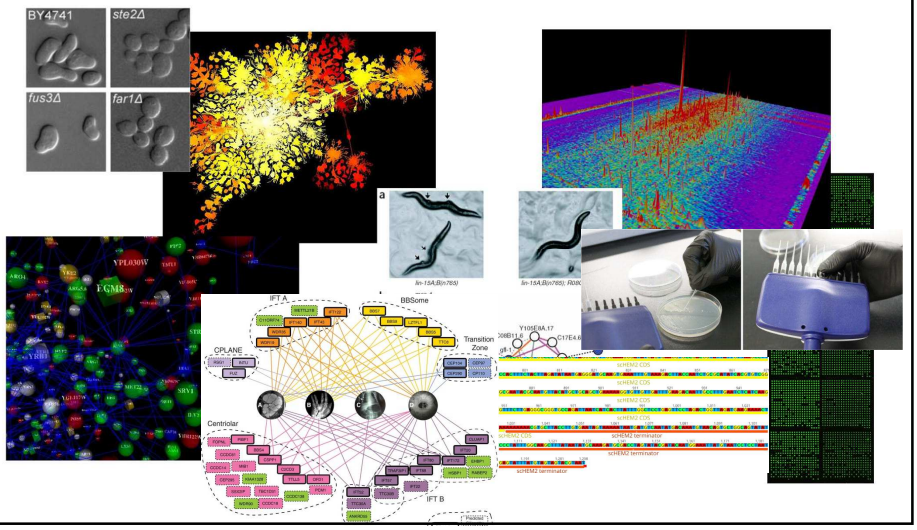
**BCH394P/BCH364C Systems Biology & Bioinformatics**
(course # 53545/53436)
**Spring 2020 Tues/Thurs 11 – 12:30 PM JGB 2.202**




1

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

**TA: Brendan Floyd** **bmfloyd@utexas.edu**
**Office hours: Mon 1 – 2/Fri 1:30 – 2:30** **NHB 3.400B atrium**
**(or MBB 3.128B)**

**Phone: 512-232-3919**

2

## Probably the most important slide today!

Course web page:

**http://www.marcottelab.org/index.php/BCH394P_BCH364C_2020**

**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; Computer Science 303E and Statistics and Data Sciences 328M (or Statistics and Scientific Computation 318M, 328M) with a grade of at least C-; and *consent of the instructor*.

3

**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 those too), but rather on algorithms, exploratory data analyses and their applications in high-throughput biology. **

4

## 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 & ebook)

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**

5

## 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 27, 2020. The last few classes will be spent presenting your projects.**

6

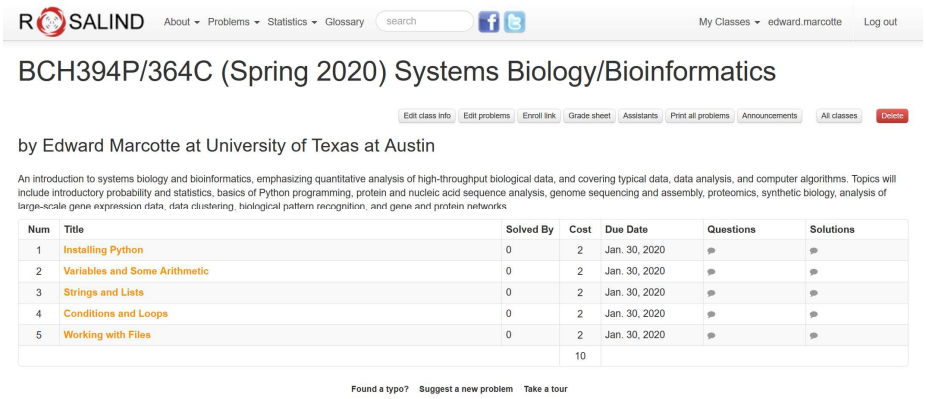
## Late policy

- **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.

7

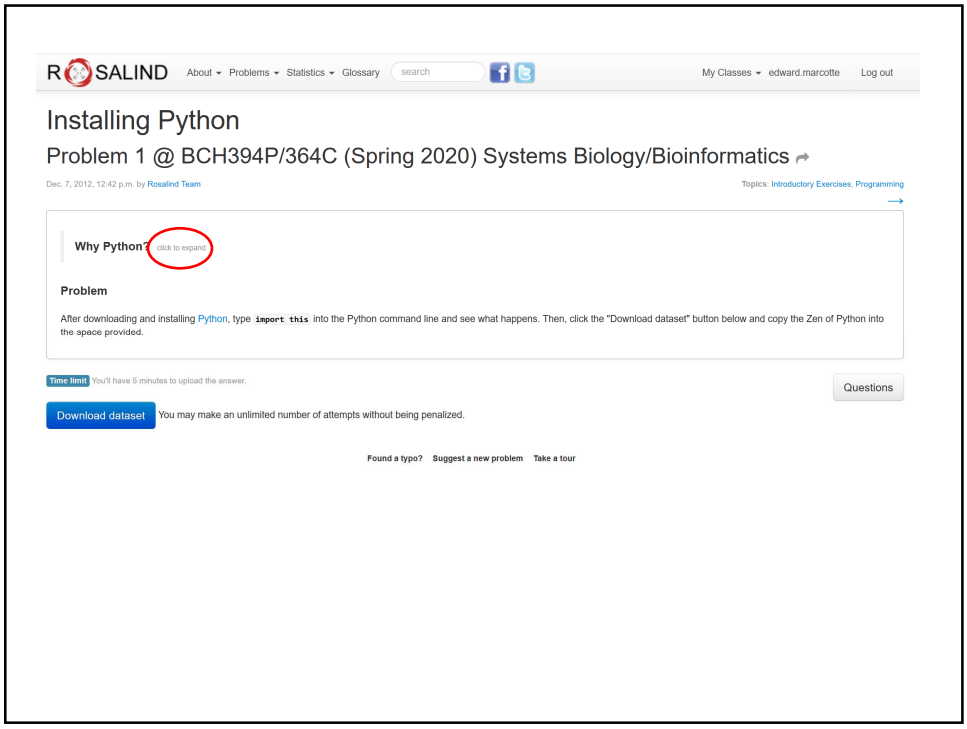
**Online homework will be via *Rosalind*: http://rosalind.info/faq/**

**Enroll specifically for BCH394P/364C at:**
**http://rosalind.info/classes/enroll/deda74d0a0/**

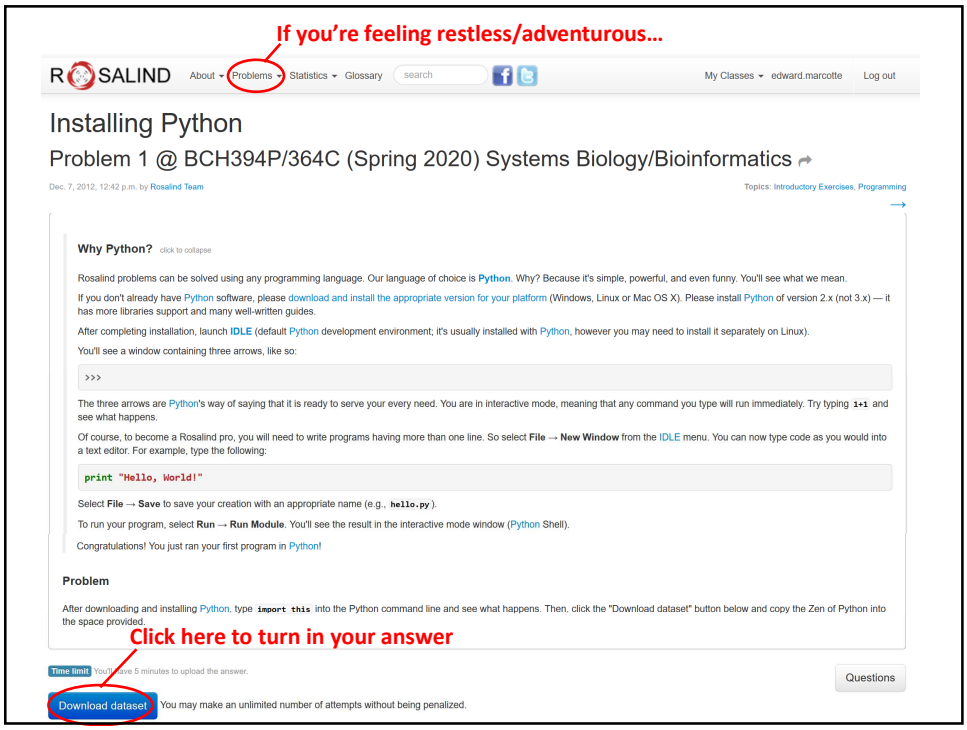



**The first homework will be due (in Rosalind) by midnight, Jan 30.**

8

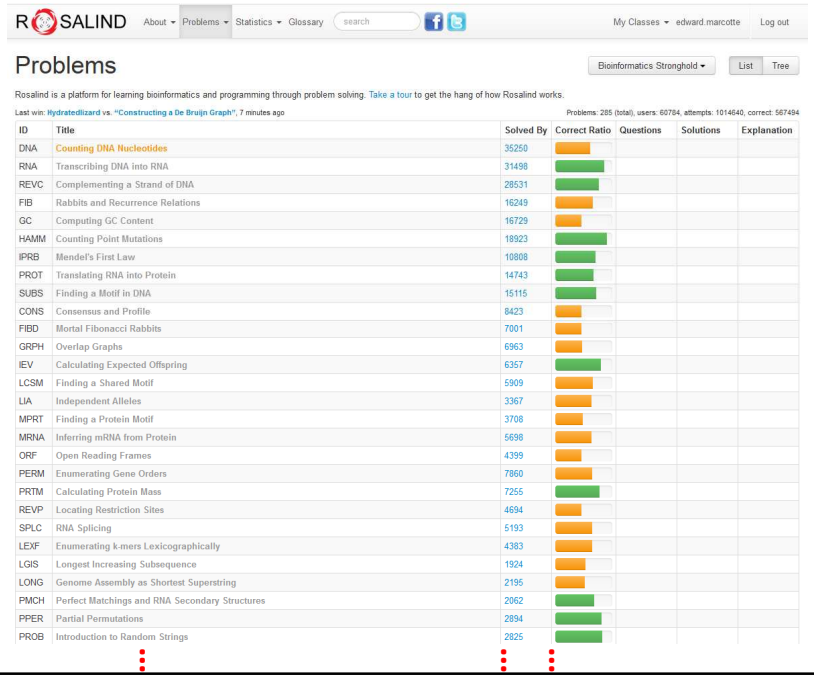

ROSALIND
About
Problems
Statistics
Glossary
search
My Classes
edward.marcotte
Log out
Installing Python
Problem 1 @ BCH394P/364C (Spring 2020) Systems Biology/Bioinformatics
Dec. 7, 2012, 12:42 p.m. by Rosalind Team
Topics: Introductory Exercises, Programming
Why Python?
click to expand
Problem
After 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.
Questions
Download dataset
You may make an unlimited number of attempts without being penalized.
Found a typo?
Suggest a new problem
Take a tour

9


If you're feeling restless/adventurous…
ROSALIND
About
Problems
Statistics
Glossary
search
My Classes
edward.marcotte
Log out
Installing Python
Problem 1 @ BCH394P/364C (Spring 2020) 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.
Questions
Download dataset
You may make an unlimited number of attempts without being penalized.

10

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




| 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 | | | | |

11

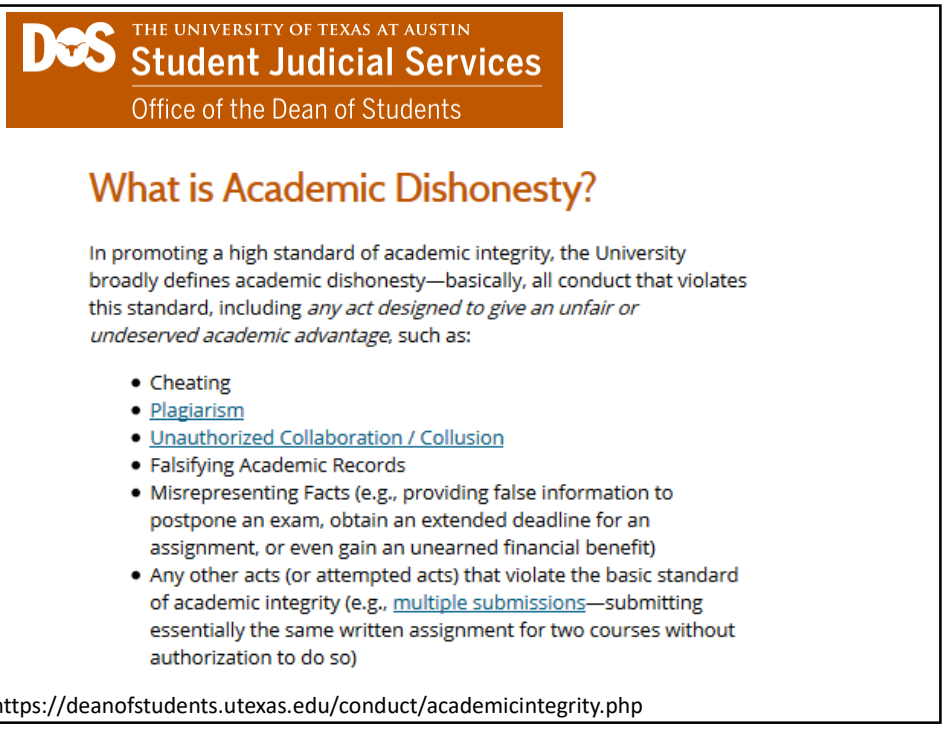
## Expectations on working together

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 presentation.

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

12




https://deanofstudents.utexas.edu/conduct/academicintegrity.php

13

- 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.

See the university's official policy on plagiarism here: https://catalog.utexas.edu/general-information/appendices/appendix-c/student-discipline-and-conduct/

14

- 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**.

15

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.

16

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.

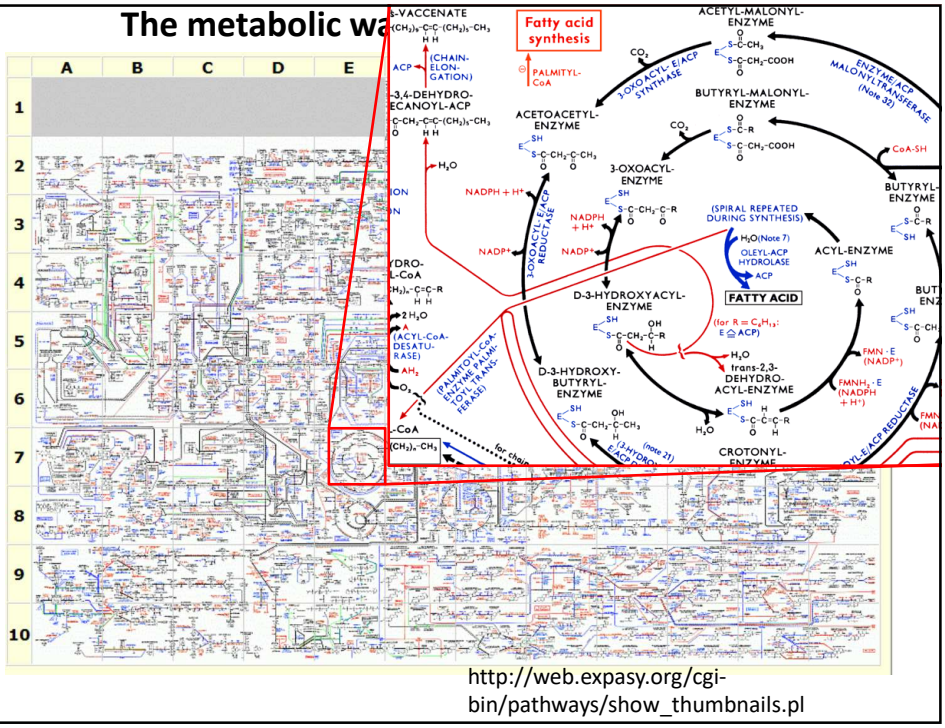
Grade-related penalties are routinely assessed ("F" in the course is not uncommon), but students can also be suspended or even permanently expelled from the University for scholastic dishonesty.

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

17

**Why are we here? (practically, not existentially)**

18




19

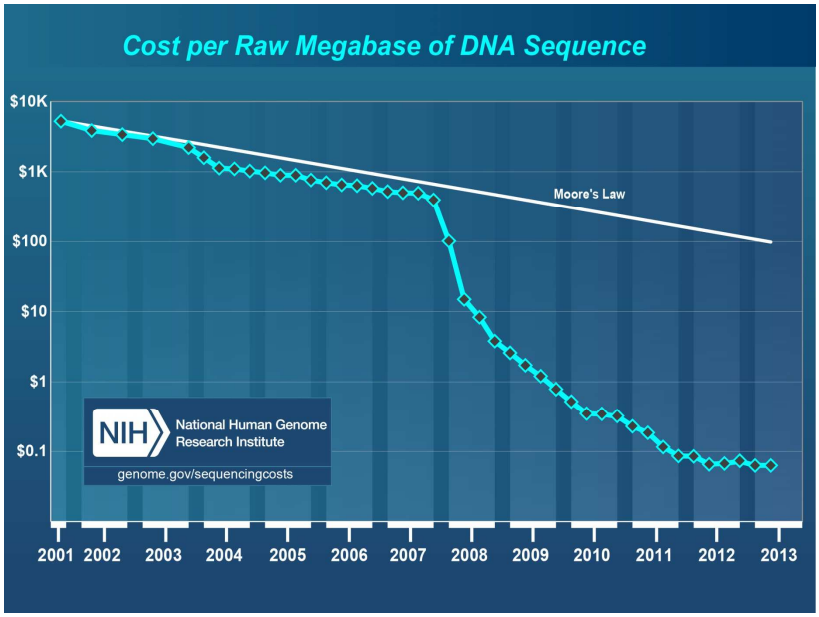
## Our current-ish knowledge of human metabolism…

| | |
|---|---|
| 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
Updated in Metabolomics 2016 12:109

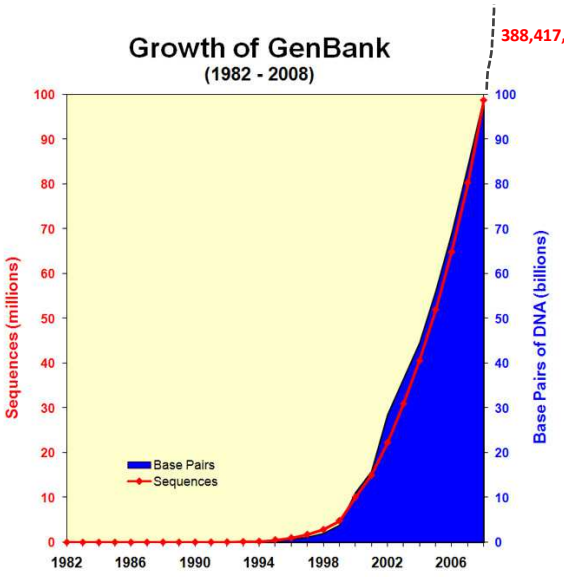
20

## Pales beside the phenomenal drop in DNA sequencing costs…




21

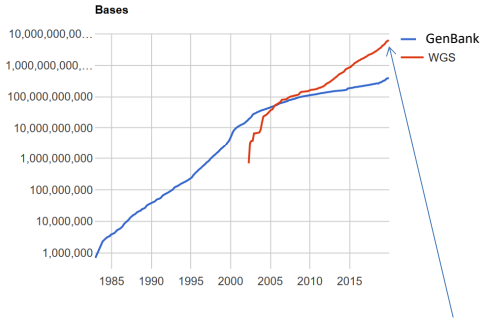
## & the corresponding explosion of DNA sequencing data…

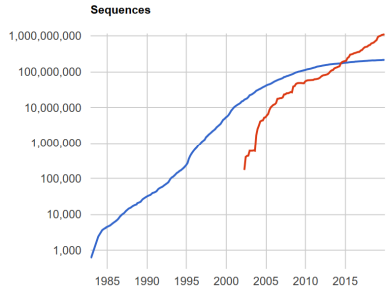



http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/
ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

22

## & the corresponding explosion of DNA sequencing data…




Here are the latest statistics…

**October 2019:**
386 billion bp
+
5.6 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

23

**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.**

24

## 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

25

**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

26

**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 27, 2020**

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

27