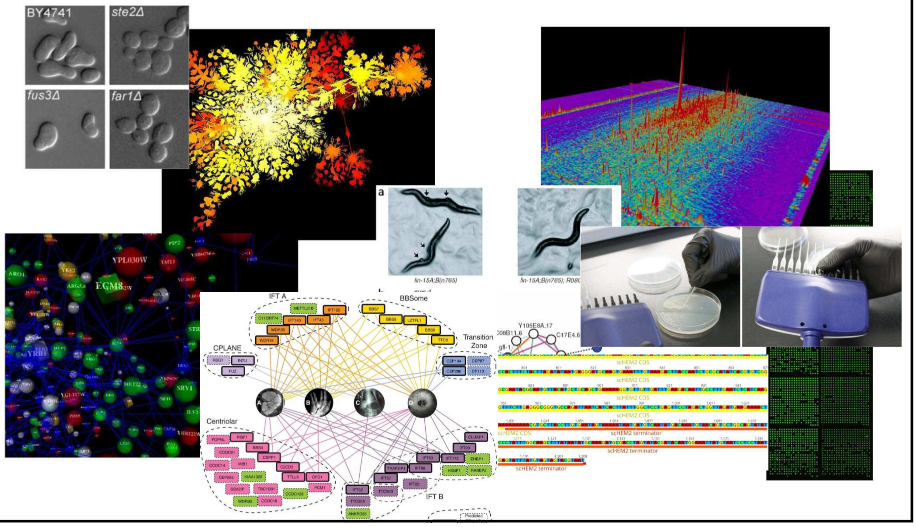
BCH394P/BCH364C Systems Biology & Bioinformatics

(course # 54540 / 54450)

Spring 2022 Tues/Thurs 11 – 12:30 PM

1st 2 weeks virtual, in person after in WEL 2.110



1

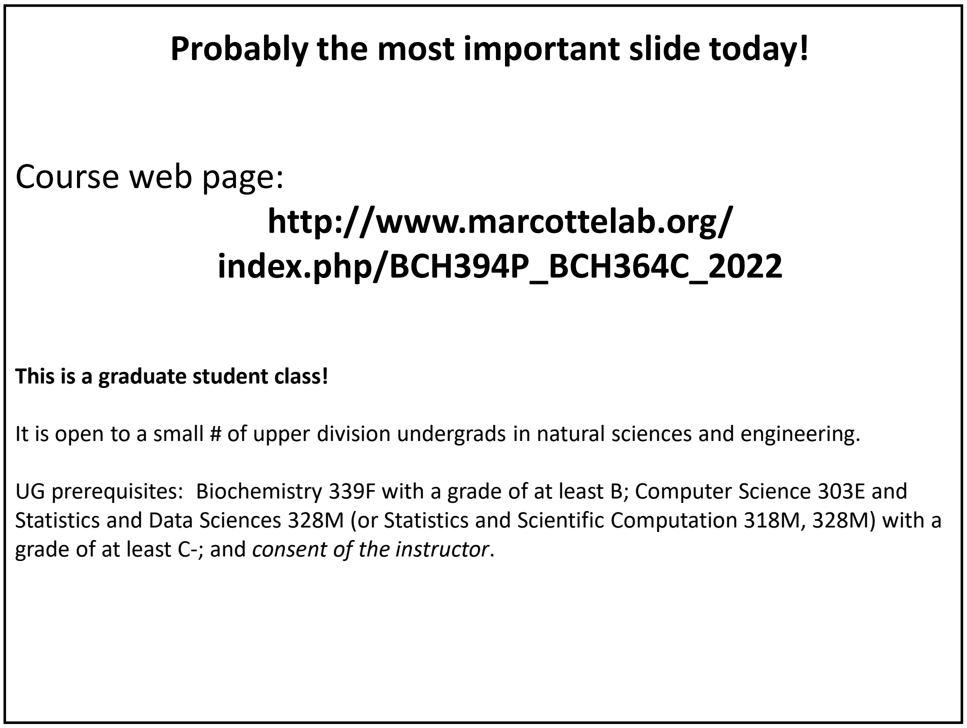
Instructor: Prof. Edward Marcotte **marcotte@utexas.edu**
Zoom office hours: Wed 11 – 12

TA: Muyoung Lee **ml49649@utexas.edu**
Zoom office hours: Mon 1 – 2/Fri 11 – 12

Class Slack channel: ut-sp22-bioinfo.slack.com

The class zoom channel will be posted on Canvas.
It will be the same zoom for class and office hours.

2



Probably the most important slide today!

Course web page:

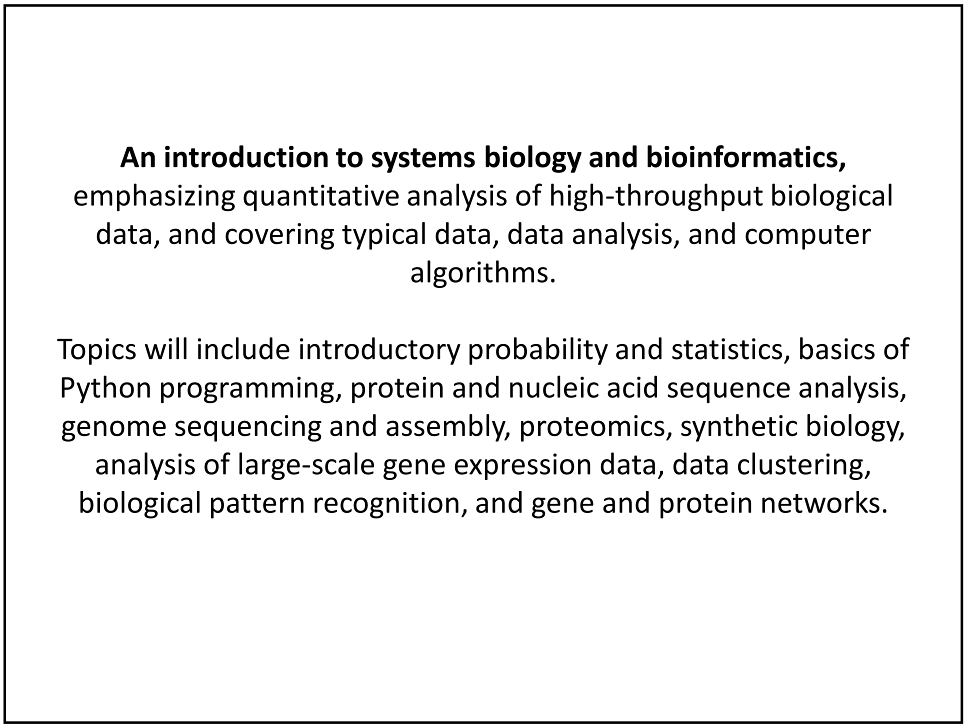
**http://www.marcottelab.org/
index.php/BCH394P_BCH364C_2022**

This is a graduate student class!

It is open to a small # 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.

4

Note that this is NOT a course on practical sequence analysis or using web-based tools. We'll use a number of these to help illustrate points, but the focus of the course will be on learning the underlying algorithms and exploratory data analyses and their applications, esp. in high-throughput biology.

By the end of the course, you will know the fundamentals of important algorithms in bioinformatics and systems biology, be able to design and implement computational studies in biology, and have performed an element of original computational biology research

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.
The course web site lists some recommendations to help you out, such as the free web course **Practical Python Programming**
https://dabeaz-course.github.io/practical-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 25, 2022. The last few classes will be spent presenting your projects.

7

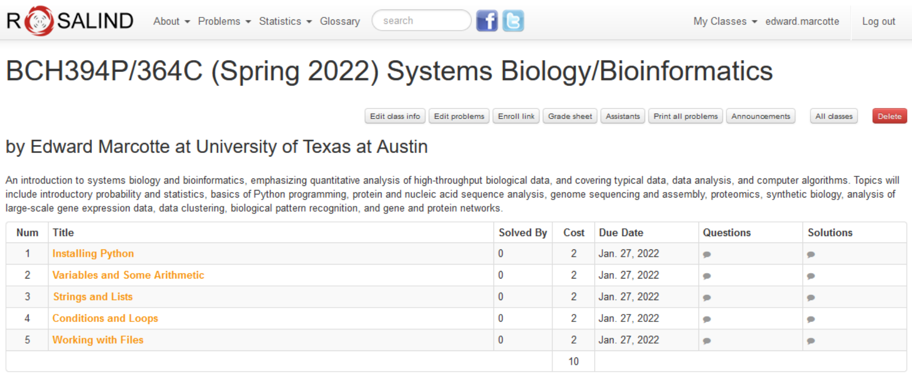
Late policy

- **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 <u>entire semester</u>, 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, <u>rounding up</u>
    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.

8

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

Enroll specifically for BCH394P/364C at:
https://rosalind.info/classes/enroll/3862a679ae/



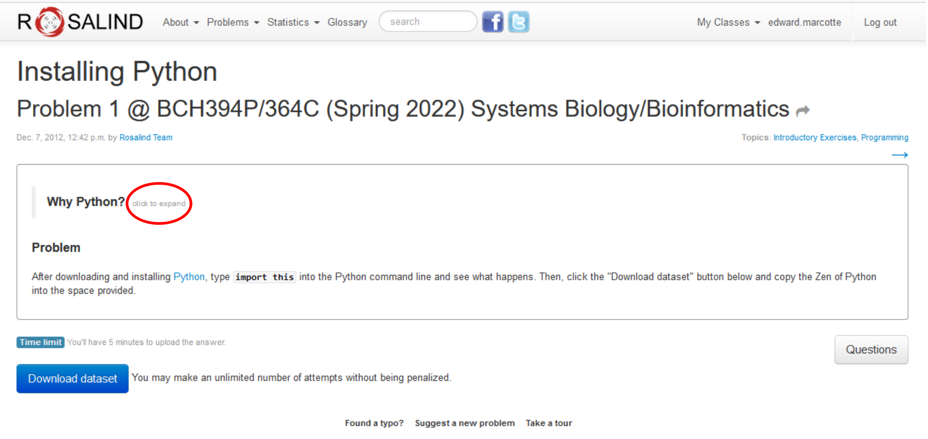
BCH394P/364C (Spring 2022) Systems Biology/Bioinformatics

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. 27, 2022	💬	💬
2	Variables and Some Arithmetic	0	2	Jan. 27, 2022	💬	💬
3	Strings and Lists	0	2	Jan. 27, 2022	💬	💬
4	Conditions and Loops	0	2	Jan. 27, 2022	💬	💬
5	Working with Files	0	2	Jan. 27, 2022	💬	💬
			10			

The first homework will be due (in Rosalind) by midnight, Jan 27.



Installing Python

Problem 1 @ BCH394P/364C (Spring 2022) Systems Biology/Bioinformatics

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

Topics: Introductory Exercises, Programming

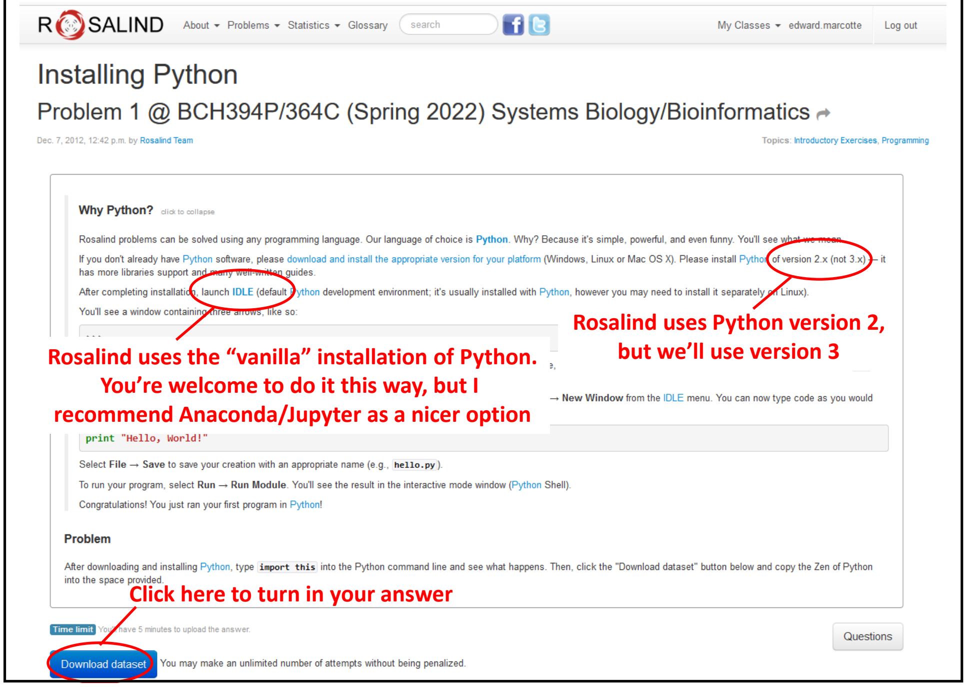
Why Python?

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.

You'll have 5 minutes to upload the answer.

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



Installing Python

Problem 1 @ BCH394P/364C (Spring 2022) 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:

Rosalind uses the "vanilla" installation of Python. You're welcome to do it this way, but I recommend Anaconda/Jupyter as a nicer option

Rosalind uses Python version 2, but we'll use version 3

→ **New Window** from the IDLE menu. You can now type code as you would

```
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

You have 5 minutes to upload the answer.

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

Installing Anaconda/Jupyter

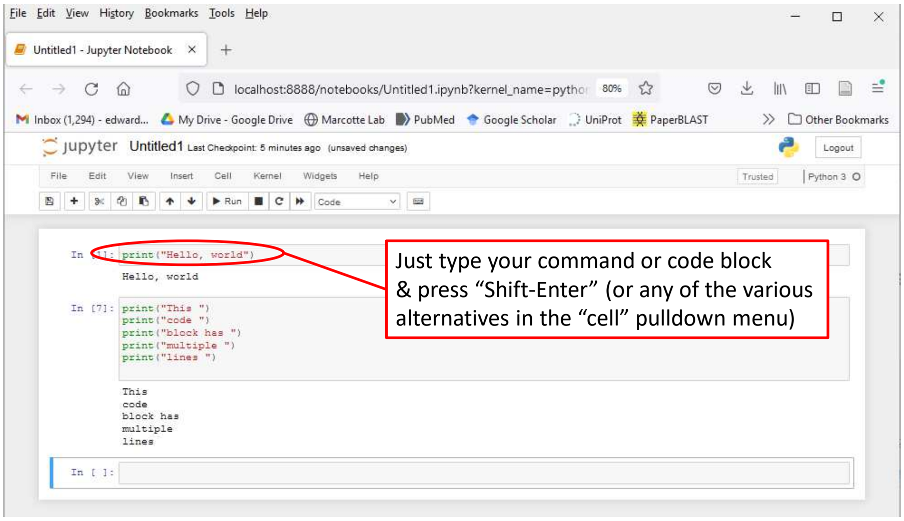
My recommendation for a good, all-round Python installation is ***Anaconda***, available free to individuals here: https://www.anaconda.com/products/individual

*****Get the latest Python 3 version** (currently 3.9)*****

Anaconda is a general management system for the various Python libraries and packages you might need, with >7,500 data science, visualization, and machine learning packages

Anaconda also provides multiple Python interfaces. For this course, I recommend using ***Jupyter Notebook***, which can be launched directly from the main Anaconda navigation window.

Jupyter is an interactive Python interface that shows your code & its output in successive entries in a shareable, archivable notebook viewable in any web browser, e.g.


File Edit View History Bookmarks Tools Help
Untitled1 - Jupyter Notebook
localhost:8888/notebooks/Untitled1.ipynb?kernel_name=python
80%
Inbox (1,294) - edward... My Drive - Google Drive Marcotte Lab PubMed Google Scholar UniProt PaperBLAST Other Bookmarks
jupyter Untitled1 Last Checkpoint: 5 minutes ago (unsaved changes)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code

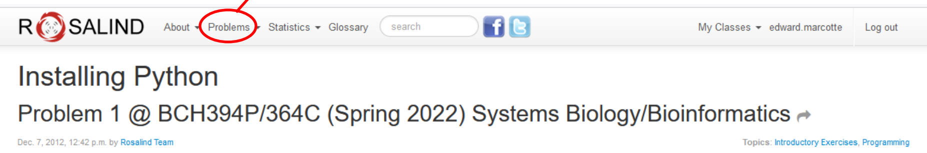
```
In [1]: print("Hello, world")
        Hello, world

In [7]: print("This ")
        print("code ")
        print("block has ")
        print("multiple ")
        print("lines ")

        This
        code
        block has
        multiple
        lines

In [ ]:
```

Just type your command or code block & press “Shift-Enter” (or any of the various alternatives in the “cell” pulldown menu)

It’s widely used in bioinformatics and data visualization.

Back to Rosalind, for those of you that are a bit more advanced:

If you’re feeling restless/adventurous…



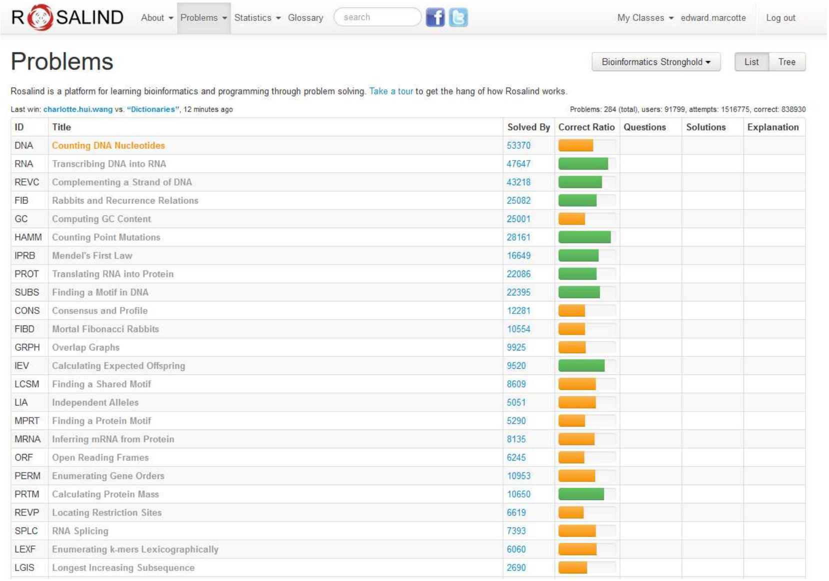
Installing Python

Problem 1 @ BCH394P/364C (Spring 2022) Systems Biology/Bioinformatics

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

Topics: Introductory Exercises, Programming

…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: charlotte.hui.wang vs. "Dictionaries", 12 minutes ago

Problems: 284 (total), users: 91799, attempts: 1516775, correct: 838930

ID	Title	Solved By	Correct Ratio	Questions	Solutions	Explanation
DNA	Counting DNA Nucleotides	53370	🟧			
RNA	Transcribing DNA into RNA	47647	🟩			
REVC	Complementing a Strand of DNA	43218	🟩			
FIB	Rabbits and Recurrence Relations	25082	🟩			
GC	Computing GC Content	25001	🟧			
HAMM	Counting Point Mutations	28161	🟩			
IPRB	Mendel's First Law	16649	🟩			
PROT	Translating RNA into Protein	22086	🟩			
SUBS	Finding a Motif in DNA	22395	🟩			
CONS	Consensus and Profile	12281	🟧			
FIBD	Mortal Fibonacci Rabbits	10554	🟧			
GRPH	Overlap Graphs	9925	🟧			
IEV	Calculating Expected Offspring	9520	🟩			
LCSM	Finding a Shared Motif	8609	🟧			
LIA	Independent Alleles	5051	🟧			
MPRT	Finding a Protein Motif	5290	🟧			
MRNA	Inferring mRNA from Protein	8135	🟧			
ORF	Open Reading Frames	6245	🟧			
PERM	Enumerating Gene Orders	10953	🟧			
PRTM	Calculating Protein Mass	10650	🟩			
REVP	Locating Restriction Sites	6619	🟧			
SPLC	RNA Splicing	7393	🟧			
LEXF	Enumerating k-mers Lexicographically	6060	🟧			
LGIS	Longest Increasing Subsequence	2690	🟧			

⋮ ⋮ ⋮

15

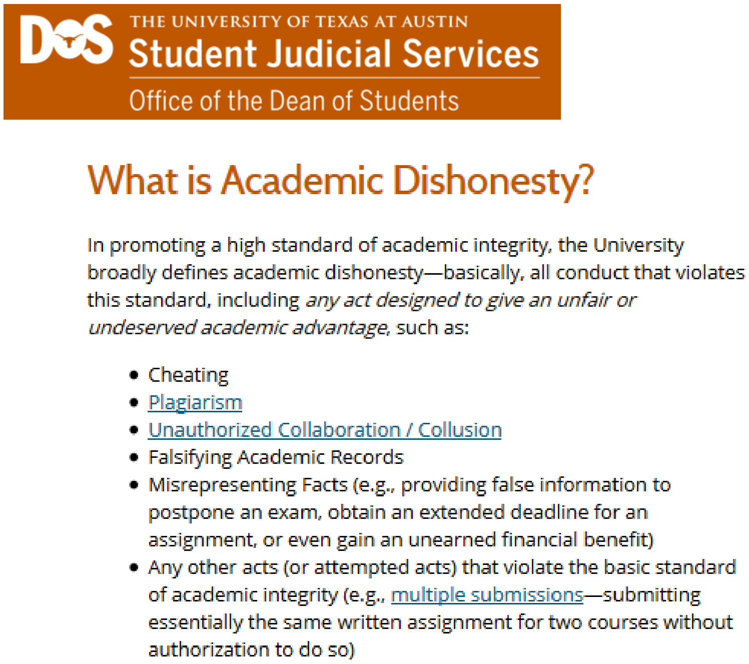
Expectations on working together

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

16



DOS
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)

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

17

- 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 (e.g. CourseHero) 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.

18

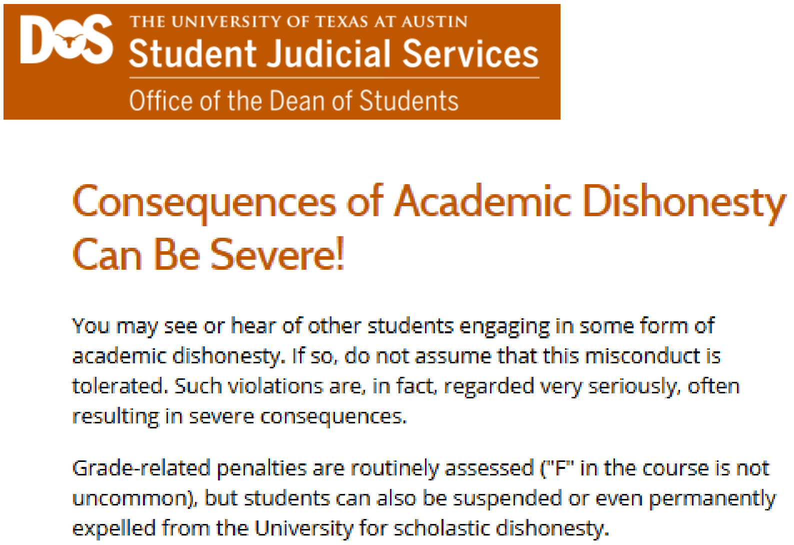
- Any materials found online (e.g. CourseHero) that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

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

19

- 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**.

20



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.

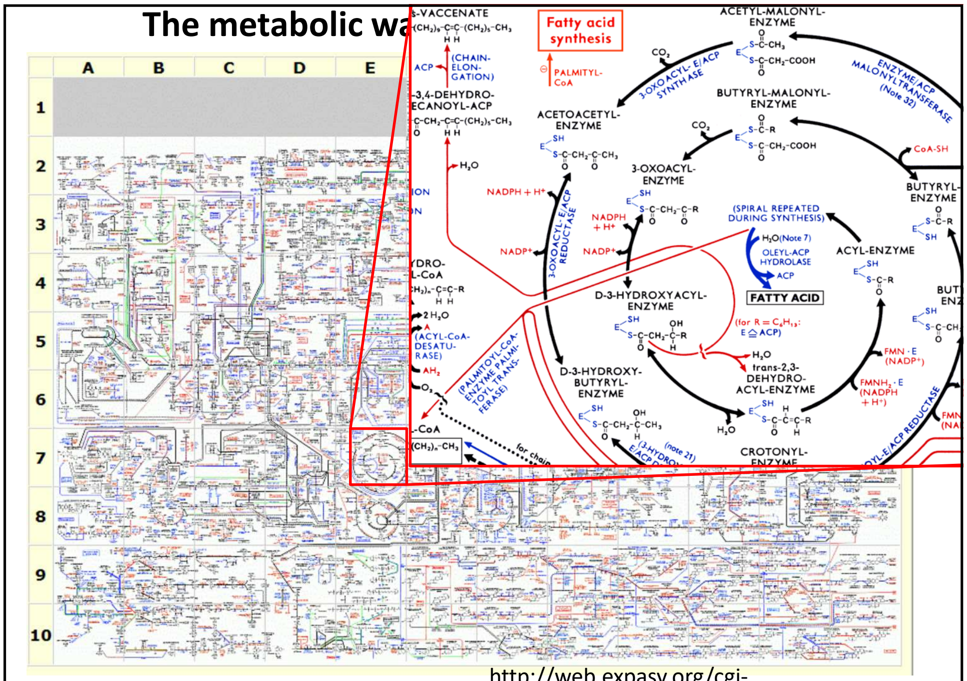
https://deanofstudents.utexas.edu/conduct/academicintegrity.php

21

Why are we here? (practically, not existentially)

22

The metabolic wa



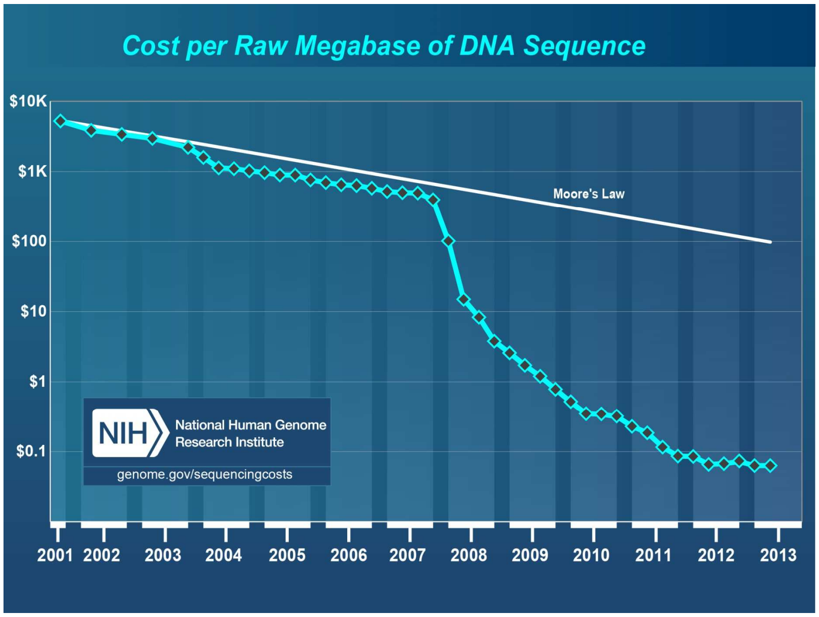
http://web.expasy.org/cgi-bin/pathways/show_thumbnails.pl

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

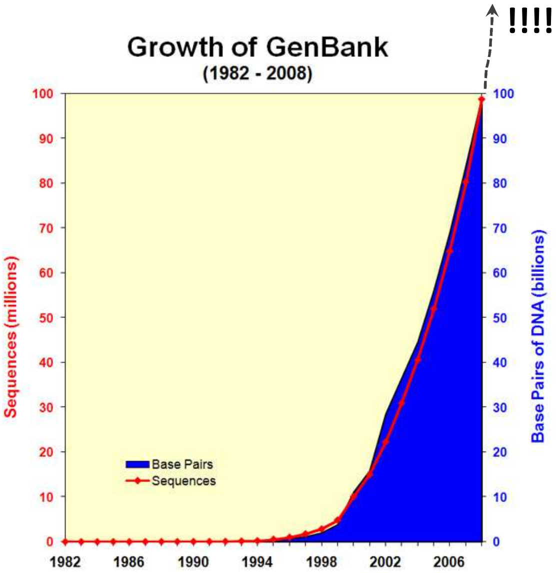
Nat Biotechnol. 2013 May;31(5):419-25
Updated in Metabolomics 2016 12:109

Pales beside the phenomenal drop in DNA sequencing costs…



25

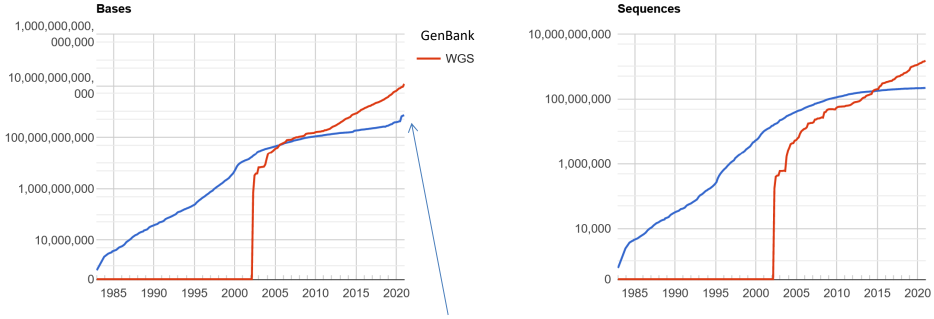
& the corresponding explosion of DNA sequencing data…



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

26

& the corresponding explosion of DNA sequencing data…



Here are the latest statistics…

December 2021:
1 trillion bp
+
14.9 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

29

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
Deep homology and the evolution of traits
Designing, simulating, and building gene circuits
Genome design and synthesis

30

Plus, expert guest lectures on:

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

Plus, plus:
we'll attempt a "live" (on zoom) demo in-class
of nanopore sequencing….

THE FINAL COURSE PROJECT IS DUE by midnight, April 25, 2022

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

& May 5 is reserved as a flex day. Current plan = NO CLASS
but we'll vote to revisit that if the pandemic/weather forces us to