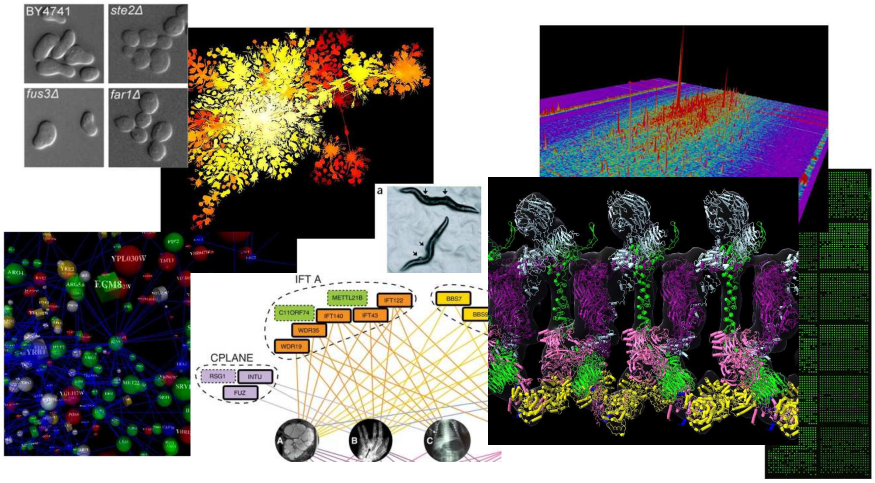
BCH394P/BCH364C Systems Biology & Bioinformatics

(course # 54960 / 54860)

Spring 2025 Tue/Thu 9:30 – 11:00 AM WEL 2.246

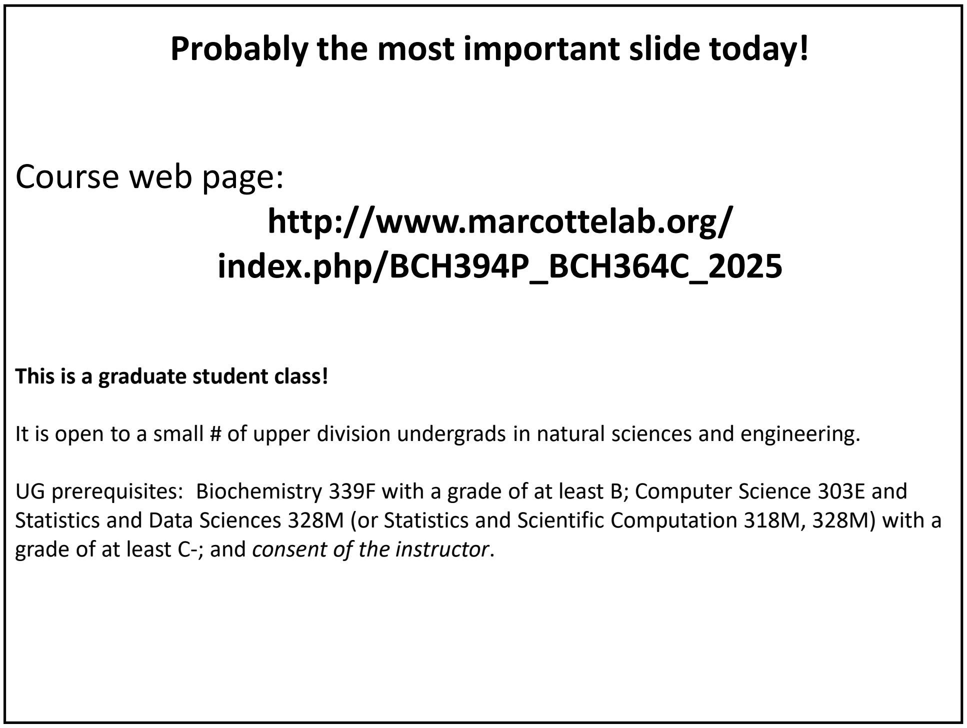


Instructor: Prof. Edward Marcotte **marcotte@utexas.edu**
Zoom office hours: Mon 4 – 5

TA: Zoya Ansari **zansari@utexas.edu**
Coding/problem set help hours:
Tues 1 – 2/Fri 1 – 2 in MBB 3.304
or by appointment on zoom

After hours Q/A, discussion: Canvas

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



Probably the most important slide today!

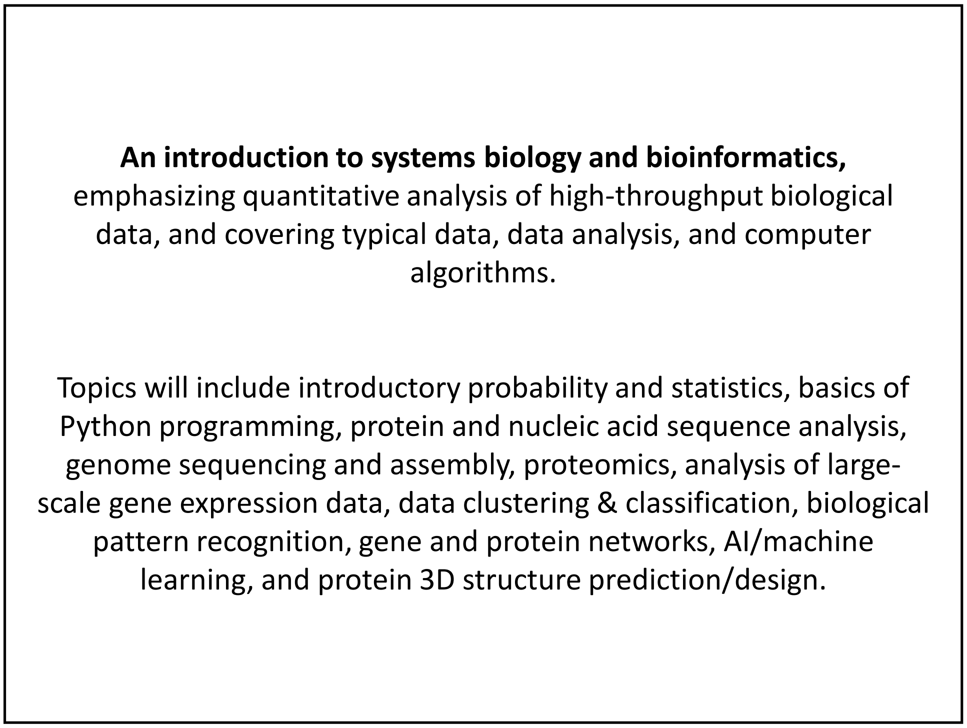
Course web page:

http://www.marcottelab.org/ index.php/BCH394P_BCH364C_2025

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*.



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, analysis of large-scale gene expression data, data clustering & classification, biological pattern recognition, gene and protein networks, AI/machine learning, and protein 3D structure prediction/design.

Note: it's NOT really a course on practical sequence analysis or using web-based tools. We'll use these, but the focus will be on learning the underlying algorithms, exploratory data analyses, and their applications, esp. in high-throughput biology.

By the end of the course, you'll know the fundamentals of important algorithms in bioinformatics and systems biology, be able to design and run 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 intro 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/

Important: There are bi-weekly coding/problem set help sessions.
Plan to attend at least one per week!

Grading

No exams. 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 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%).

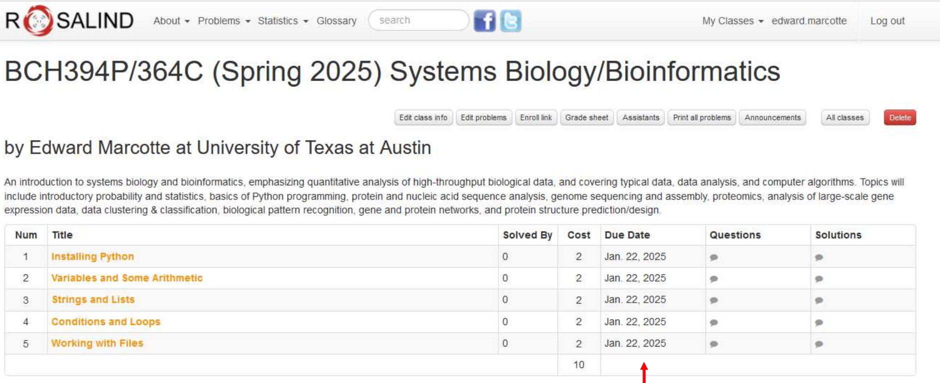
The project will be emailed as a web URL to the TA & I, developed through the semester and finished by 10 PM, April 16, 2025.
The last 3 classes will be spent presenting your projects.

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 <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.

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

Enroll specifically for BCH394P/364C at:
https://rosalind.info/classes/enroll/8cf0c8d95f/



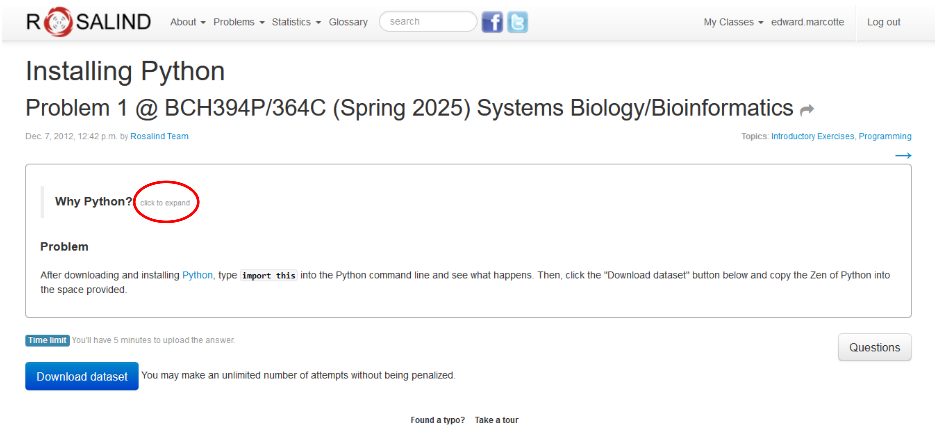
BCH394P/364C (Spring 2025) Systems Biology/Bioinformatics

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, analysis of large-scale gene expression data, data clustering & classification, biological pattern recognition, gene and protein networks, and protein structure prediction/design.

Num	Title	Solved By	Cost	Due Date	Questions	Solutions
1	Installing Python	0	2	Jan. 22, 2025	💬	💬
2	Variables and Some Arithmetic	0	2	Jan. 22, 2025	💬	💬
3	Strings and Lists	0	2	Jan. 22, 2025	💬	💬
4	Conditions and Loops	0	2	Jan. 22, 2025	💬	💬
5	Working with Files	0	2	Jan. 22, 2025	💬	💬
			10			

The first homework will be due (in Rosalind) by 10 PM, Jan 22



Installing Python

Problem 1 @ BCH394P/364C (Spring 2025) Systems Biology/Bioinformatics

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

Topics: Introductory Exercises, Programming

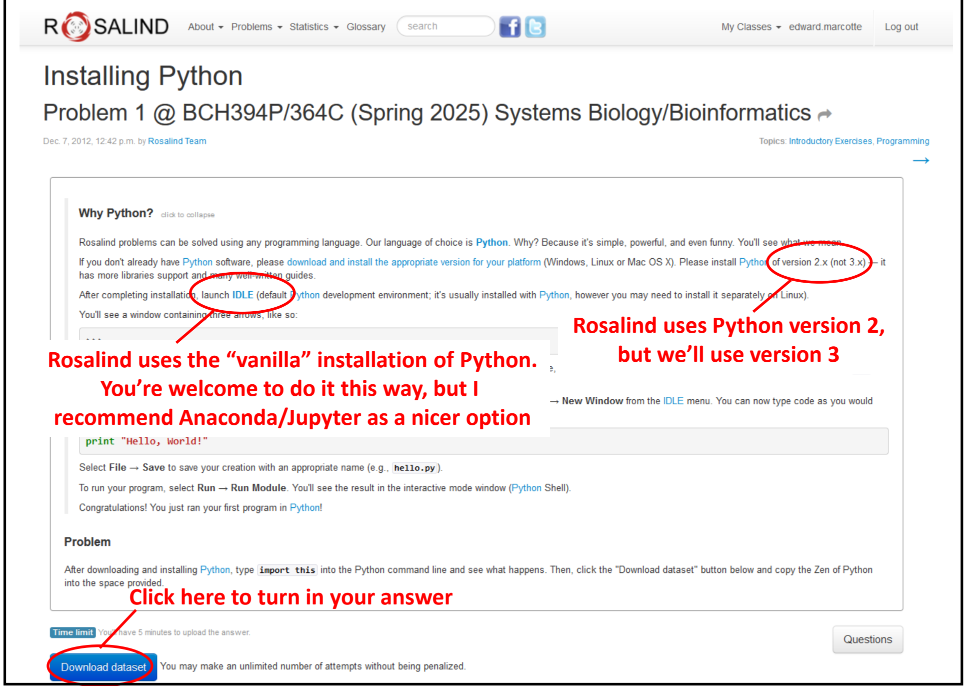
Why Python? click to expand

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.



Installing Python

Problem 1 @ BCH394P/364C (Spring 2025) 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:

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

Time limit You have 5 minutes to upload the answer.

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

Installing Anaconda/Jupyter

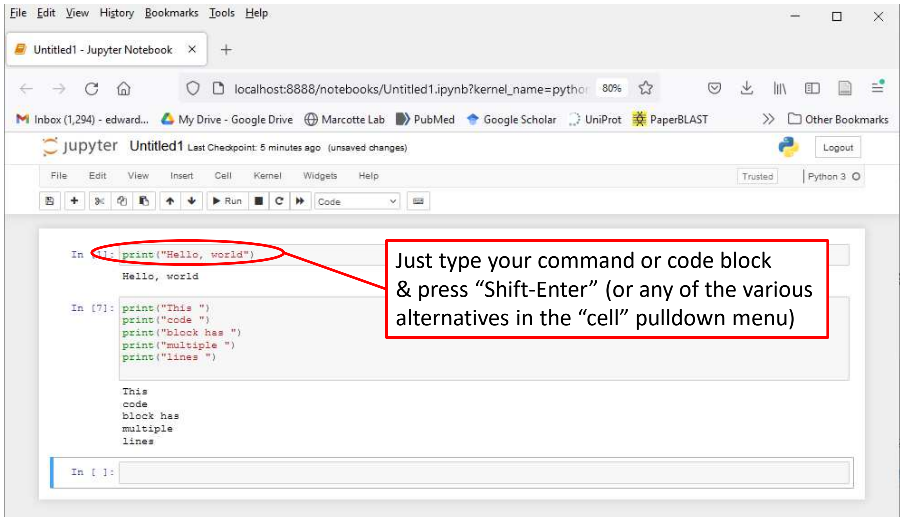
My recommendation for a good, all-round Python installation is ***Anaconda***, available free to individuals here:
https://www.anaconda.com/download
(note you can "skip registration" if you prefer that)

*****Get the latest Python 3 version*****
(but any version > 3.0 is probably fine)

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=pythor
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

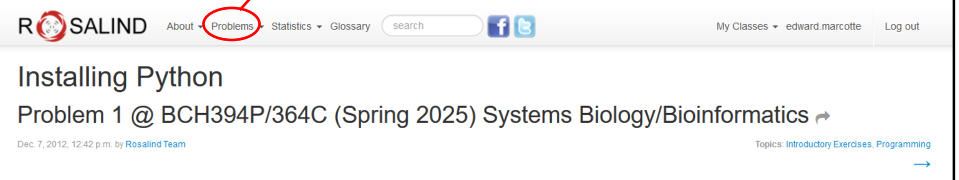
```
In [1]: print("Hello, world")
        Hello, world

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…



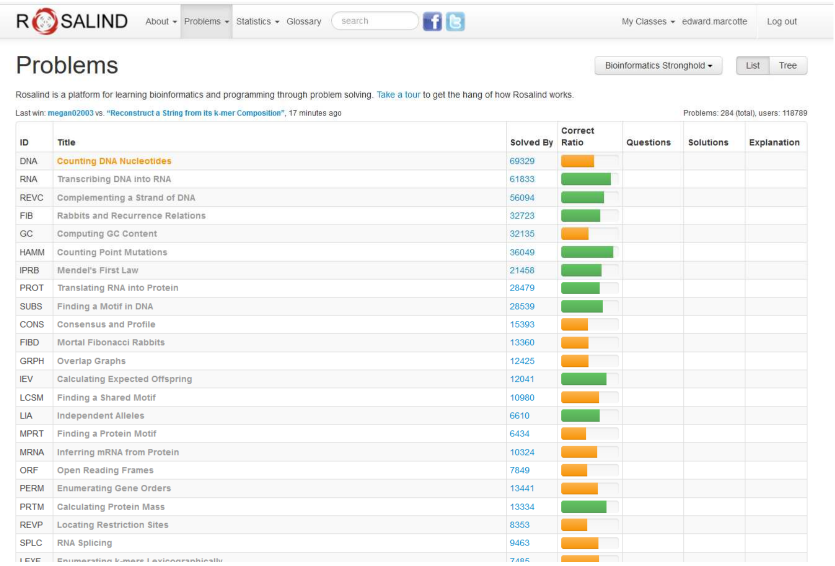
Installing Python

Problem 1 @ BCH394P/364C (Spring 2025) 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: megan02003 vs. "Reconstruct a String from its k-mer Composition", 17 minutes ago

Problems: 284 (total), users: 118789

ID	Title	Solved By	Correct Ratio	Questions	Solutions	Explanation
DNA	Counting DNA Nucleotides	69329	🟧			
RNA	Transcribing DNA into RNA	61833	🟩			
REVC	Complementing a Strand of DNA	56094	🟩			
FIB	Rabbits and Recurrence Relations	32723	🟩			
GC	Computing GC Content	32135	🟧			
HAMM	Counting Point Mutations	36049	🟩			
IPRB	Mendel's First Law	21458	🟩			
PROT	Translating RNA into Protein	28479	🟩			
SUBS	Finding a Motif in DNA	28539	🟩			
CONS	Consensus and Profile	15393	🟧			
FIBD	Mortal Fibonacci Rabbits	13360	🟧			
GRPH	Overlap Graphs	12425	🟧			
IEV	Calculating Expected Offspring	12041	🟩			
LCSM	Finding a Shared Motif	10980	🟧			
LIA	Independent Alleles	6610	🟩			
MPRT	Finding a Protein Motif	6434	🟧			
MRNA	Inferring mRNA from Protein	10324	🟧			
ORF	Open Reading Frames	7849	🟧			
PERM	Enumerating Gene Orders	13441	🟧			
PRTM	Calculating Protein Mass	13334	🟩			
REVP	Locating Restriction Sites	8353	🟧			
SPLC	RNA Splicing	9463	🟧			
LEXF	Enumerating k-mers Lexicographically	7485	🟧			

⋮ ⋮ ⋮

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

A reminder about academic integrity

- 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.

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

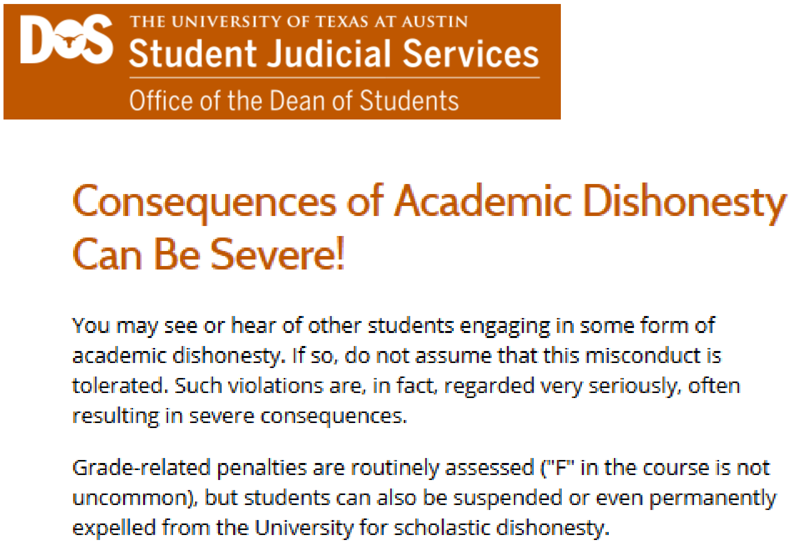
DOS THE UNIVERSITY OF TEXAS AT AUSTIN
Student Judicial Services
Office of the Dean of Students

- 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/

- 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**.



DOS THE UNIVERSITY OF TEXAS AT AUSTIN

Student Judicial Services

Office of the Dean of Students

Consequences of Academic Dishonesty Can Be Severe!

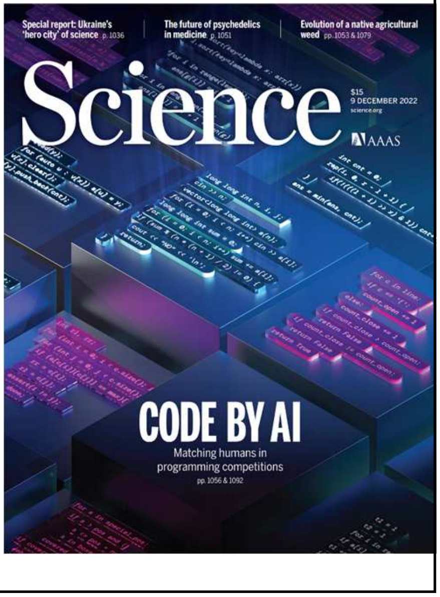
You may see or hear of other students engaging in some form of academic dishonesty. If so, do not assume that this misconduct is tolerated. Such violations are, in fact, regarded very seriously, often resulting in severe consequences.

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.

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

Yes, but …

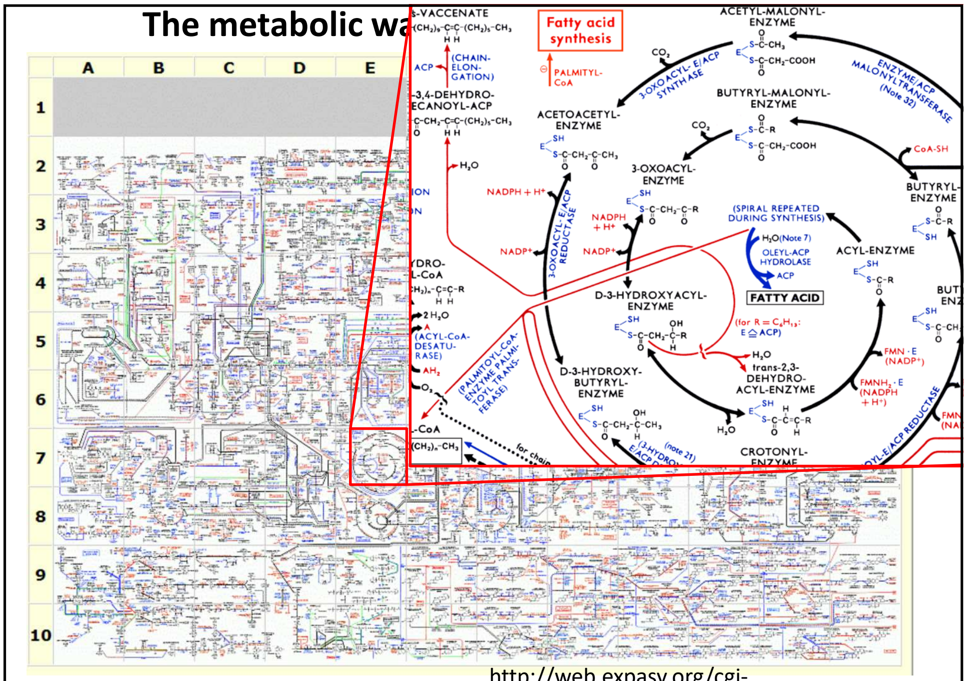
Later in the semester, we’ll try co-programming with AI using chatGPT, where the goal is to make the computer write the code for you


Special report: Ukraine's 'hero city' of science p. 1036
The future of psychedelics in medicine p. 1051
Evolution of a native agricultural weed pp. 1053 & 1079
Science
$15
9 DECEMBER 2022
science.org
AAAS
CODE BY AI
Matching humans in programming competitions
pp. 1056 & 1092

Why are we here?

(practically, not existentially)

The metabolic wa

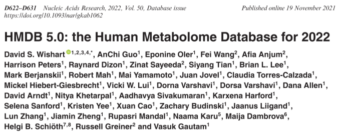


http://web.expasy.org/cgi-bin/pathways/show_thumbnails.pl

Our current-ish knowledge of human metabolism…

A few statistics from the Human Metabolome Database (https://hmdb.ca/):

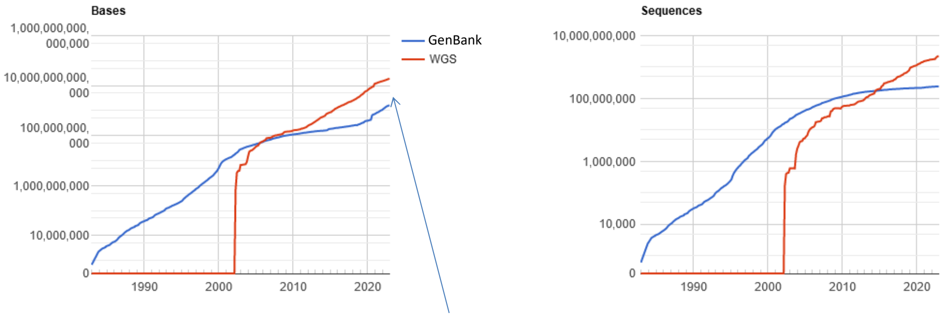
Total Number of Metabolites	253,245
Total Number of Expected Metabolites	98,257
Total Number of Predicted Metabolites	130,679
Total Number of Endogenous Metabolites	222,860
Total Number of Metabolites Having Associated Proteins (Enzymes and Transporters)	71,168
Total Number of Metabolites with Synthesis Records	1,608
Total Number of Compounds Detected and Quantified for Normal Individuals	3,292
Total Number of Compounds Detected and Quantified for Abnormal Conditions	1,791
Total Number of Different Diseases	657
Total Number of Metabolites Associated with Diseases	22,600
Total Number of Metabolite Concentrations for Diseases	32,087
Total Number of NMR or GC-MS or MS/MS Spectra	2,732,152
Total Number of Compounds with NMR or GC-MS or MS/MS Spectra	211,527
Total Number of NMR Spectra	242,268
Total Number of Compounds with NMR Spectra	12,345


D622–D631 Nucleic Acids Research, 2022, Vol. 50, Database issue *Published online 19 November 2021*
https://doi.org/10.1093/nar/gkab1062

HMDB 5.0: the Human Metabolome Database for 2022

David S. Wishart[1,2,3,4,*], AnChi Guo[1], Eponine Oler[1], Fei Wang[2], Afia Anjum[2], Harrison Peters[1], Raynard Dizon[1], Zinat Sayeeda[2], Siyang Tian[1], Brian L. Lee[1], Mark Berjanskii[1], Robert Mah[1], Mai Yamamoto[1], Juan Jovel[1], Claudia Torres-Calzada[1], Mickel Hiebert-Giesbrecht[1], Vicki W. Lui[1], Dorna Varshavi[1], Dorsa Varshavi[1], Dana Allen[1], David Arndt[1], Nitya Khetarpal[1], Aadhavya Sivakumaran[1], Karxena Harford[1], Selena Sanford[1], Kristen Yee[1], Xuan Cao[1], Zachary Budinski[1], Jaanus Liigand[1], Lun Zhang[1], Jiamin Zheng[1], Rupasri Mandal[1], Naama Karu[5], Maija Dambrova[6], Helgi B. Schiöth[7,8], Russell Greiner[2] and Vasuk Gautam[1]

Pales beside the phenomenal explosion of DNA sequencing:



Here are the latest statistics…

December 2024:
5 trillion bp Genbank
+
33 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

Science
BREAKTHROUGH OF THE YEAR
2021
nature
PROTEIN POWER
AI network predicts highly accurate 3D structures for the human proteome
nature methods
Method of the Year 2021: Protein structure prediction
Science
PREDICTING STRUCTURES
Deep learning accurately folds proteins

NOBEL PRIZE IN CHEMISTRY 2024

AWARDEES

David Baker
"For computational protein design."

Demis Hassabis and John M Jumper
"For protein structure prediction."

PRIZE AMOUNT
11 MILLION
SWEDISH KRONOR ($1 MILLION).

One half awarded to David Baker and the other half jointly to Demis Hassabis and John Jumper.

In 2003, David Baker succeeded with the almost impossible feat of **building entirely new kinds of proteins.** Since then his research group has produced one imaginative protein creation after another, including proteins that can be used as pharmaceuticals, vaccines, nanomaterials and tiny sensors.

In 2020, Demis Hassabis and John M Jumper **developed an AI model called AlphaFold2** to predict proteins' complex structures. Among a myriad of scientific applications, AlphaFold2 has helped researchers better understand antibiotic resistance and create images of enzymes that can decompose plastic.

Network18 creative

Resulting in huge growth in 3D structural data:

April 2021

Collaboration agreement between EMBL-EBI and DeepMind

July 2021

Initial release of the AlphaFold DB

December 2021

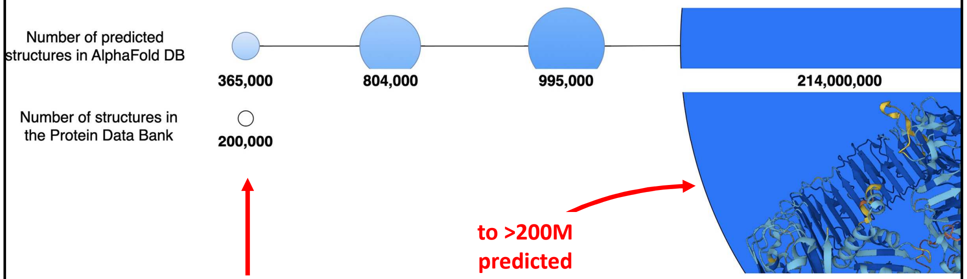
Predicted structures for the Swiss-Prot sequence set

January 2022

Organisms implicated in Global Health and Neglected Diseases

July 2022

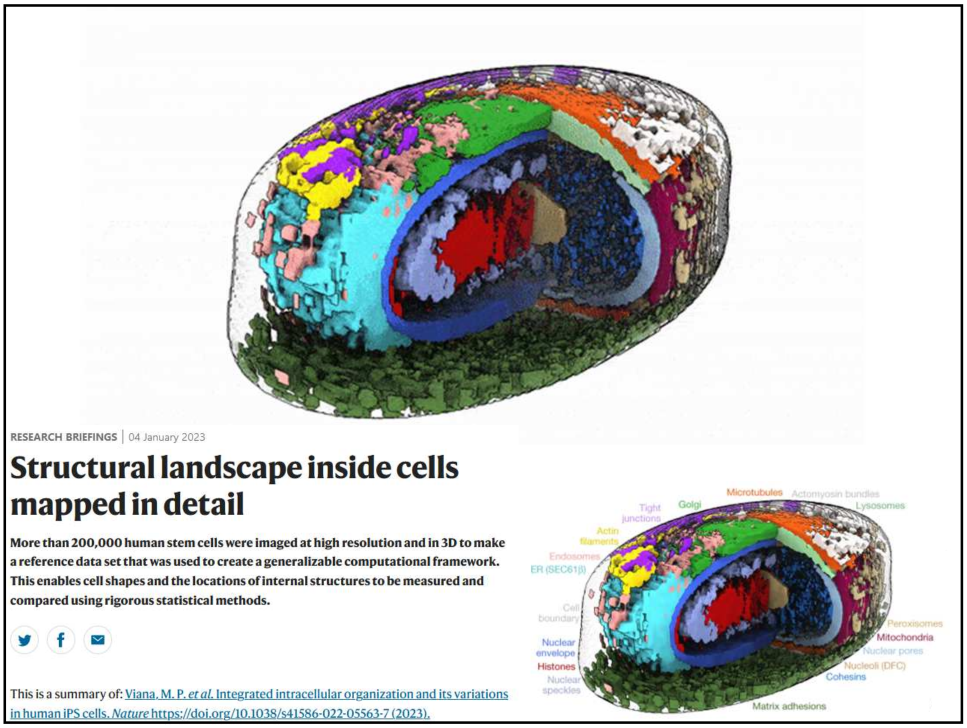
Predicted structures for most of the UniProt database



From 200K experimental structures in 2021

to >200M predicted structures in the latest release

Nucleic Acids Research 52(D1):D368–D375 (2024)



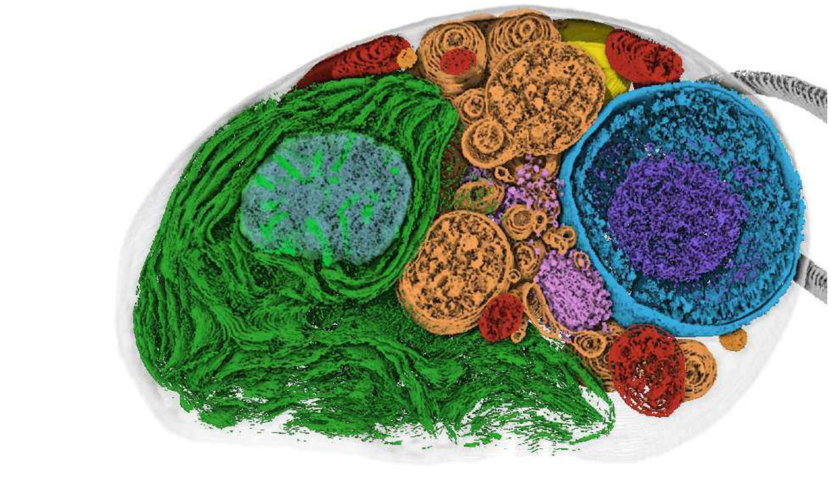
RESEARCH BRIEFINGS | 04 January 2023

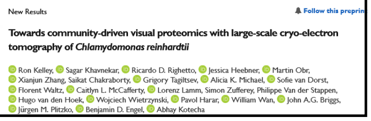
Structural landscape inside cells mapped in detail

More than 200,000 human stem cells were imaged at high resolution and in 3D to make a reference data set that was used to create a generalizable computational framework. This enables cell shapes and the locations of internal structures to be measured and compared using rigorous statistical methods.

This is a summary of: Viana, M. P. *et al.* Integrated intracellular organization and its variations in human iPS cells. *Nature* https://doi.org/10.1038/s41586-022-05563-7 (2023).

& 3 weeks ago, >1,800 3D tomograms of green algae were released “as a community resource to … inspire biological discovery”





New Results

Follow this preprin

Towards community-driven visual proteomics with large-scale cryo-electron tomography of *Chlamydomonas reinhardtii*

Ron Kelley, Sagar Khavnekar, Ricardo D. Righetto, Jessica Heebner, Martin Obr, Xianjun Zhang, Saikat Chakraborty, Grigory Tagiltsev, Alicia K. Michael, Sofie van Dorst, Florent Waltz, Caitlyn L. McCafferty, Lorenz Lamm, Simon Zufferey, Philippe Van der Stappen, Hugo van den Hoek, Wojciech Wietrzynski, Pavol Harar, William Wan, John A.G. Briggs, Jürgen M. Plitzko, Benjamin D. Engel, Abhay Kotecha

https://www.biorxiv.org/content/10.1101/2024.12.28.630444v1.full

Why are we here? We have no choice!

- **Biologists are faced with a staggering deluge of data, growing exponentially**
- **Bioinformatics/comp bio tools and approaches help us understand these data and work productively, and to build increasingly powerful models of biological systems**
- **We’ll learn important basic concepts in this field and get exposed to key technologies driving the field**

Specifically…

We’ll cover the following topics, approximately in this order:

BASICS OF PYTHON PROGRAMMING
Introduction to Rosalind
A Python programming primer for non-programmers
Rosalind help & programming Q/A, new AI tools for learning programming

BIOLOGICAL SEQUENCE ANALYSIS
Substitution matrices (BLOSSUM, PAM) & sequence alignment
Protein and nucleic acid sequence alignments, dynamic programming
Sequence profiles
BLAST! (the algorithm), MMSeqs2, & FoldSeek
Biological databases
Markov processes and Hidden Markov Models

GENOMES, PROTEOMES, & "BIG BIOLOGY"
Gene finding algorithms
Genome sequencing & assembly
An introduction to large gene expression data sets
Promoter and motif finding, Gibbs sampling
Guest lecture: Intro to NGS analysis and the CBRF core

MACHINE LEARNING/AI
Clustering algorithms, hierarchical, k-means, self-organizing maps, force-directed maps, UMAP/tSNE
Classification algorithms
Principal component analysis and data transformations
Guest lecture: Protein 3D structure prediction, incl. AlphaFold
Guest lecture: AI/deep neural networks and large language models

SYNTHETIC BIOLOGY & PROTEIN DESIGN
Protein 3D design/engineering, RFDiffusion/ProteinMPNN, ColabFold
Synthetic biology & genome design

THE FINAL COURSE PROJECT IS DUE by 10 PM, April 16, 2024

The last 3 class days will be for presenting your projects