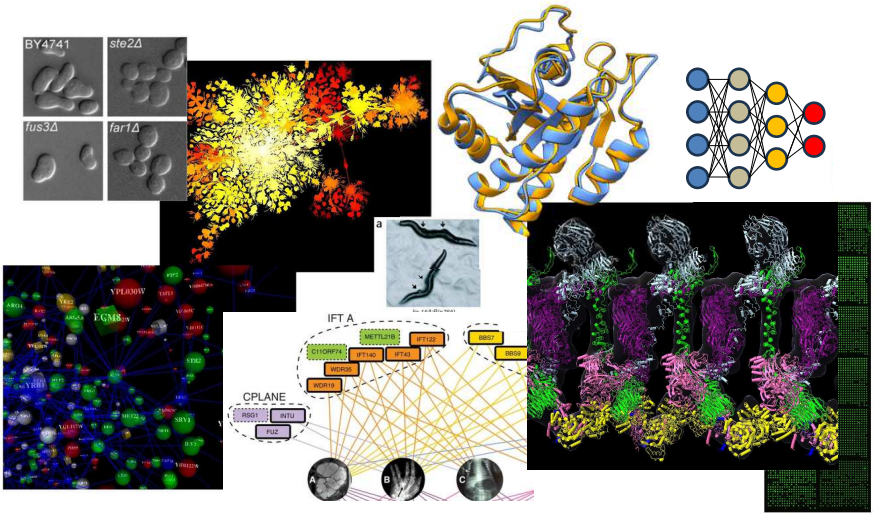
BCH394P/BCH364C Systems Biology & Bioinformatics
(course # 57450 / 57345)
Spring 2026 Tue/Thu 9:30 – 11:00 AM WEL 2.246



1

Instructor: Prof. Edward Marcotte marcotte@utexas.edu
Zoom office hours: Mon 4 – 5
The zoom channel will be posted on Canvas.

TA: Zoya Ansari zansari@utexas.edu
Coding/problem set help hours:
Tues 11:30-12:30 in MBB 3.204
Wed 1-2 in MBB 3.304
or by appointment on zoom

After hours Q/A, discussion: Canvas

2

Probably the most important slide today!

Course web page:

http://www.marcottelab.org/index.php/BCH394P_BCH364C_2026

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, 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.

4

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

5

Books

The lectures will be from research articles and slides.
For basic 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!

6

Grading

No exams. Grades will be based on:

- **Class attendance** (randomly assessed throughout the semester and counting 12% of the final grade)
- **3 online programming homework assignments** (6 points each and counting 18% of the final grade)
- **3 problem sets** (15 points each and counting 45% of the final grade)
- **A course research project** that you will develop over the semester & present in the last 3 days of class (25% of final grade)

The course research project will be 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 15, 2026.
The last 3 classes will be spent presenting your projects.

7

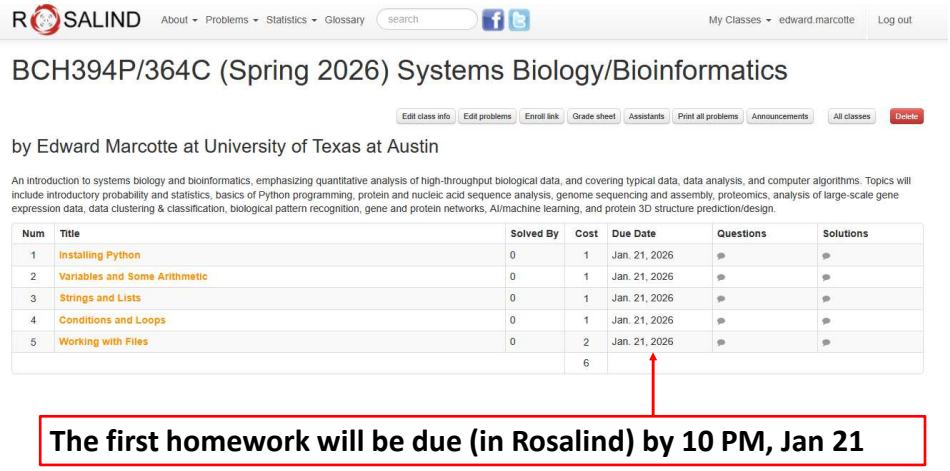
Late policy

- **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 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.

8

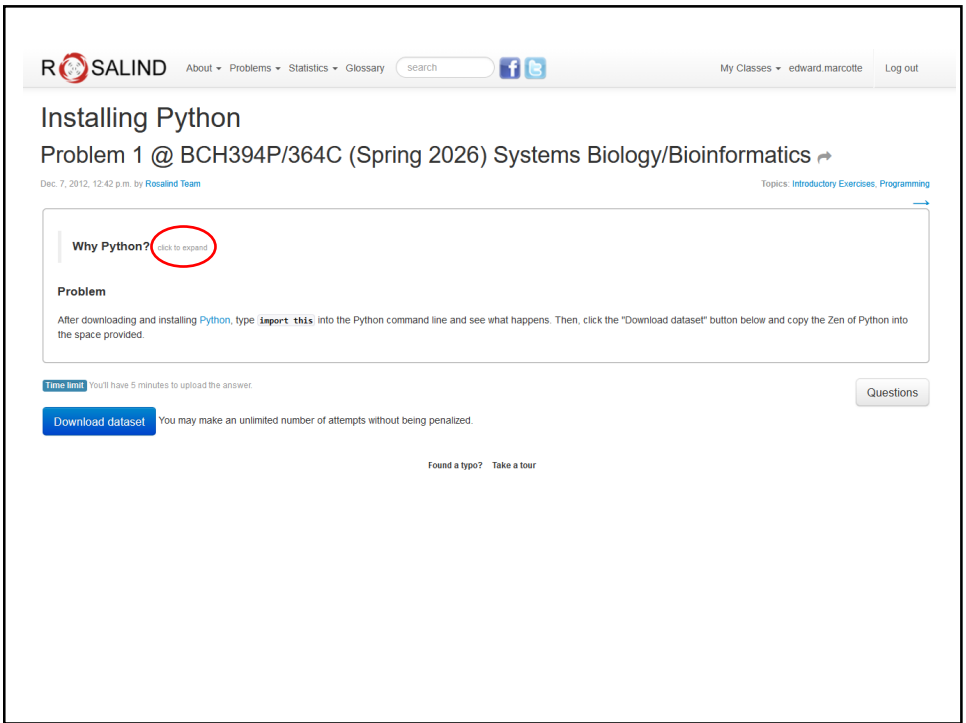
Online homework will be via *Rosalind*: http://rosalind.info/faq/

Enroll specifically for BCH394P/364C at:
https://rosalind.info/classes/enroll/2fa64f76b9/

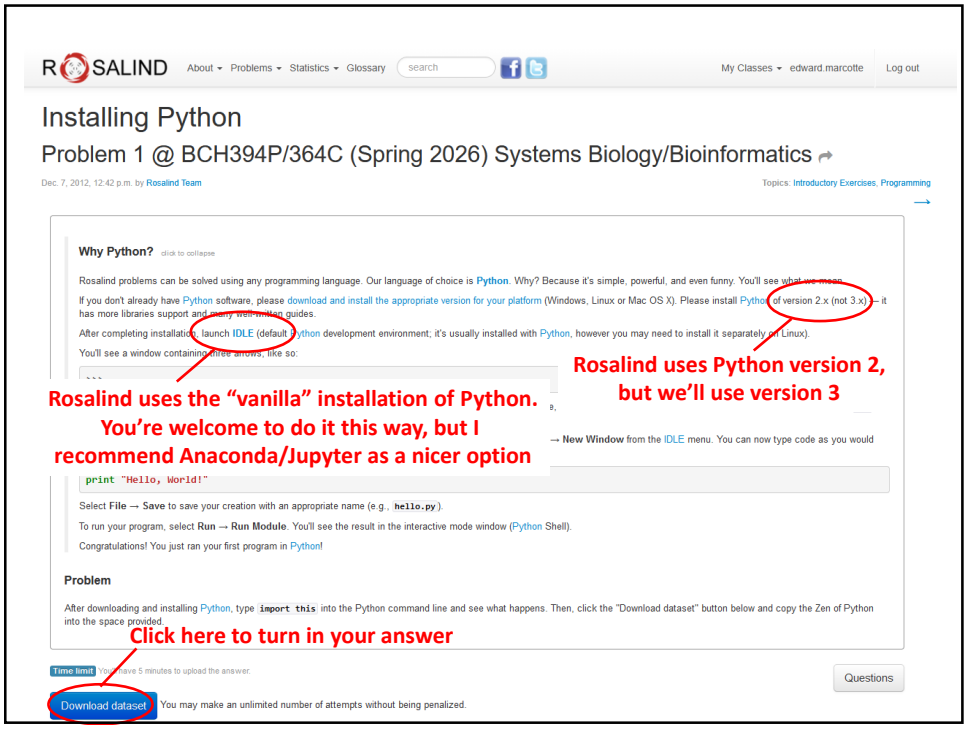


The first homework will be due (in Rosalind) by 10 PM, Jan 21

9



10



11

Installing Anaconda/Jupyter

My recommendation for a good, all-round Python installation is ***Anaconda***, available free to university students here:

https://www.anaconda.com/download

(note you can "skip registration" if you prefer that)

*****Get the latest Python 3 version*****

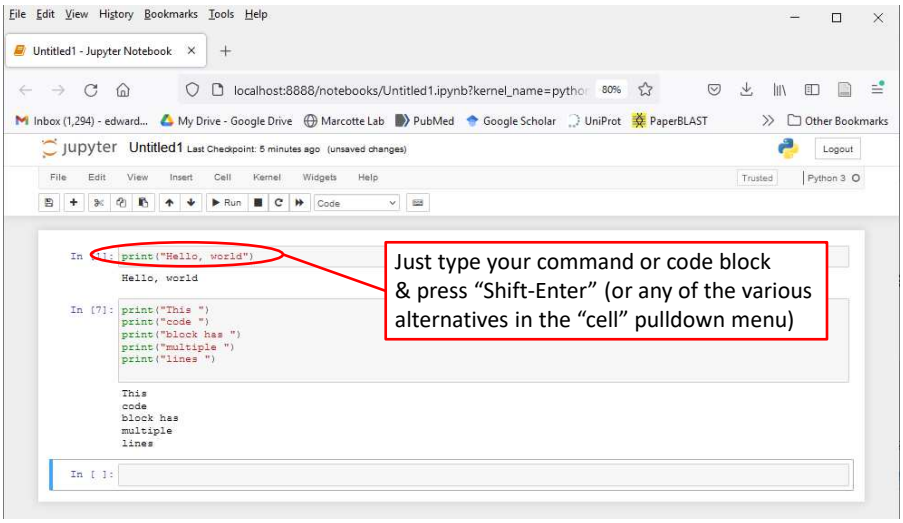
(but any version > 3.0 is probably fine)

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.

12

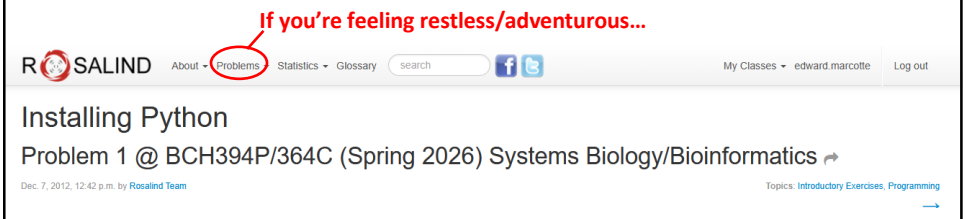
Jupyter is an interactive Python interface that shows your code & its output in successive entries in a shareable, archivable notebook viewable in any web browser, e.g.



It’s widely used in bioinformatics and data visualization.

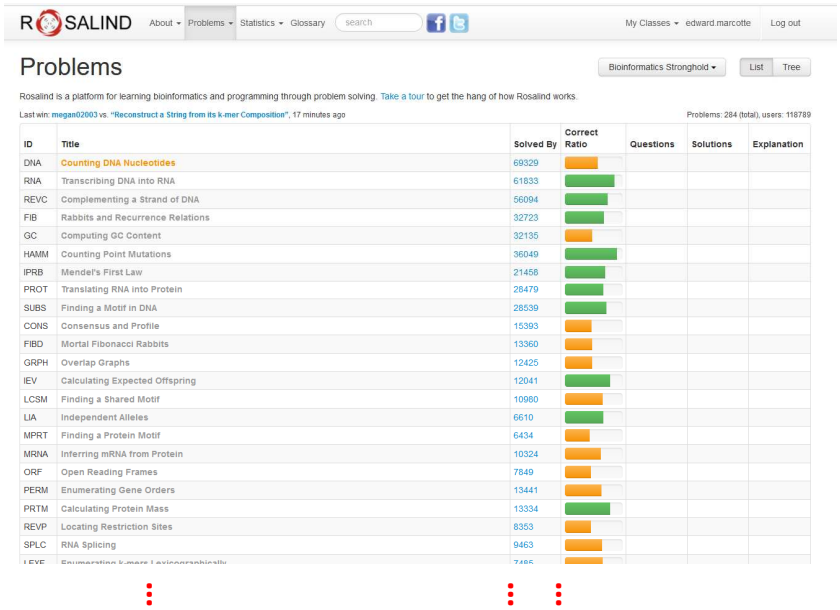
13

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



14

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



15

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

16

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/avoiding-academic-misconduct.php

17

- 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-conduct-and-academic-integrity/

18

- You can use the internet to get *ideas*, programming *suggestions* and *syntax*, but **<u>downloading completed answers to assigned questions and submitting these as your own work is cheating/plagiarism</u>**.

- **<u>Copying entire programs</u>** verbatim from marked repositories offering Rosalind homework solutions **<u>is cheating and plagiarism</u>**. Asking AI chatbots to answer your homework for you is too.

19

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.

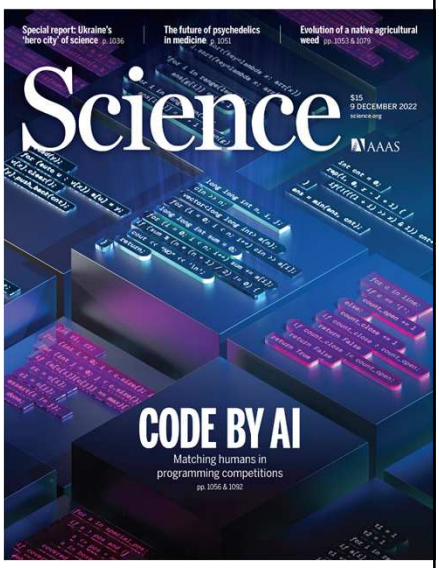
Grade-related penalties are routinely assessed ("F" in the course is not uncommon), but students can also be suspended or even permanently expelled from the University for scholastic dishonesty.

https://deanofstudents.utexas.edu/conduct/avoiding-academic-misconduct.php

20

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, but you need to build up a knowledge base to use these effectively.



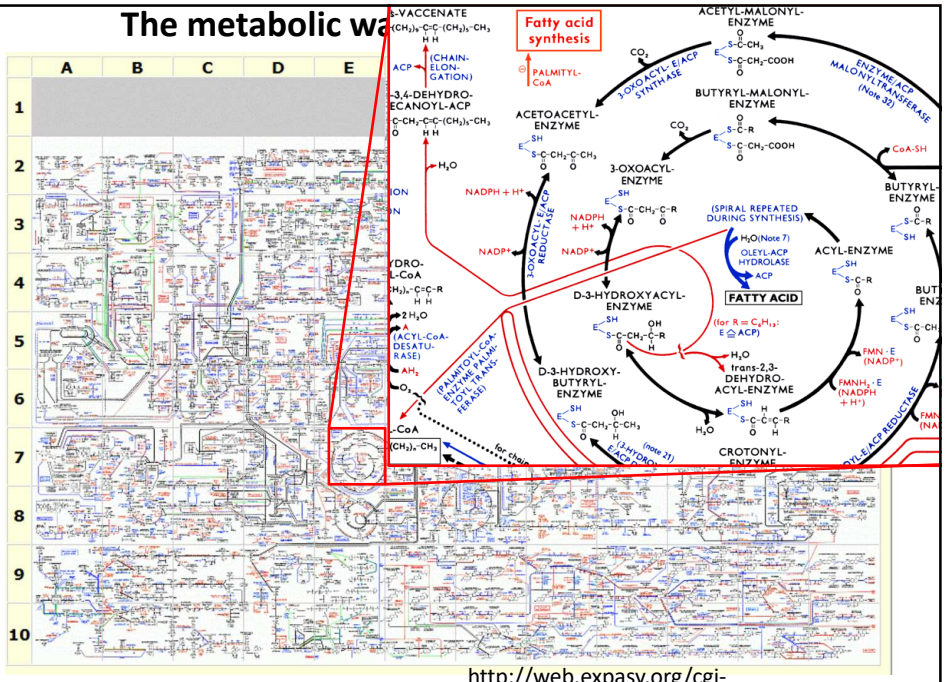
21

Why are we here?

(practically, not existentially)

22

The metabolic wa



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

23

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

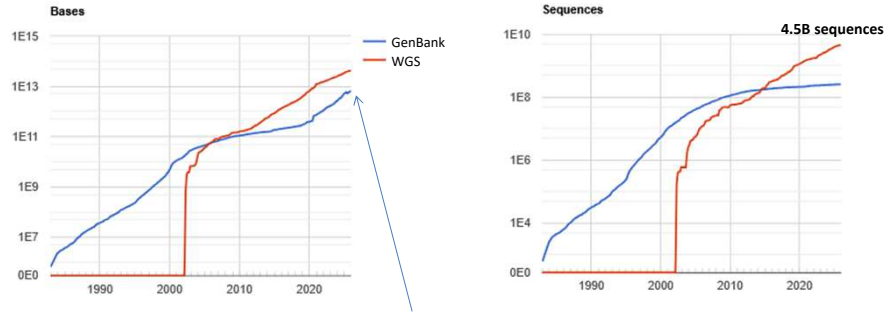
D622–D631 Nucleic Acids Research, 2022, Vol. 50, Database issue Published online 19 November 2021

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], Seiena Sanford[1], Kristen Yee[1], Xuan Cao[1], Zachary Budinski[1], Jaanus Liigand[1], Lucy Zhang[1], Jiamin Zhang[1], Rupasri Mandal[1], Naama Karu[2], Maija Dambrova[4], Helgi B. Schiöth[7,8], Russell Greiner[2] and Vasuk Gautam[1]

24

Pales beside the phenomenal explosion of DNA sequencing:



The latest statistics:

December 2025:
6 trillion bp Genbank
+
42 trillion bp DNA whole genome shotgun sequencing

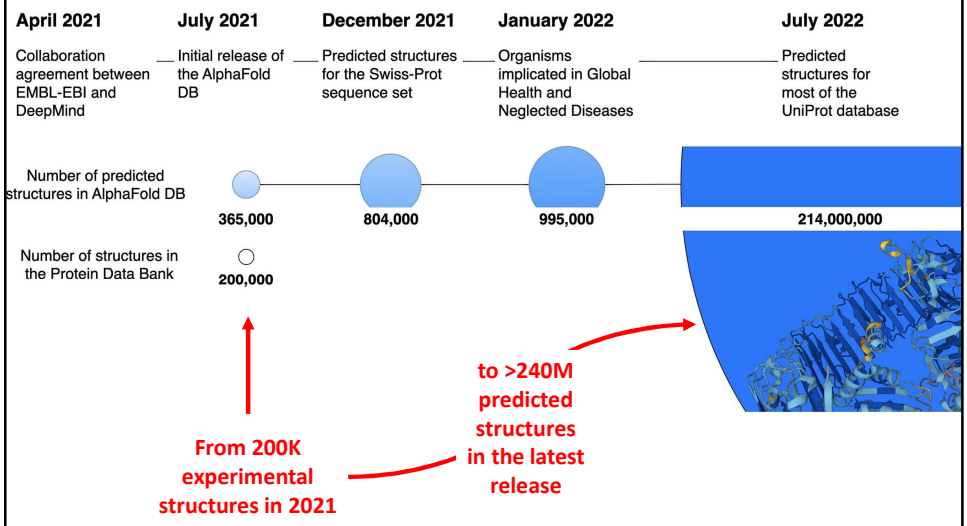
Which basically means GenBank is falling behind more every year!

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

25

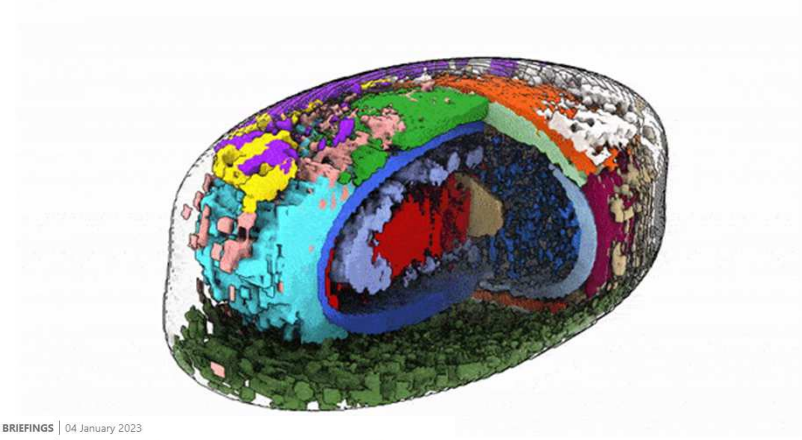
26

Resulting in huge growth in 3D structural data:



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

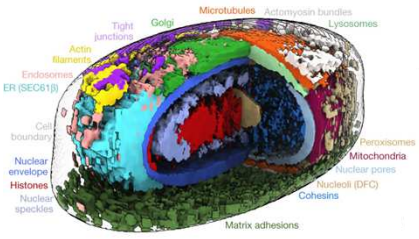
27



RESEARCH BRIEFINGS | 04 January 2023

Structural landscape inside cells mapped in detail

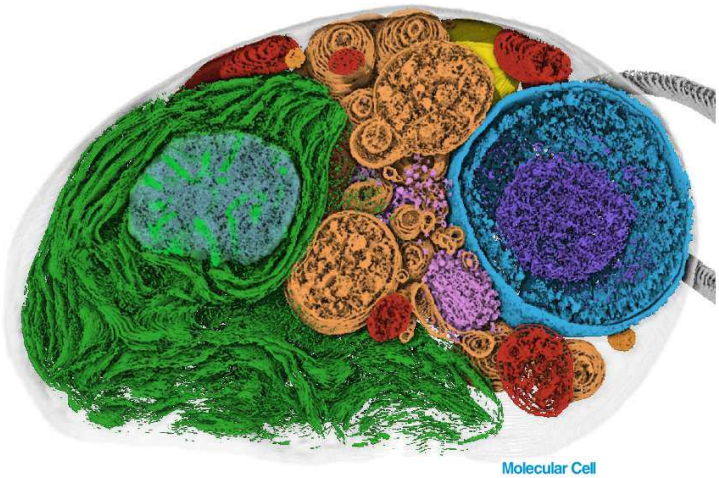
More than 200,000 human stem cells were imaged at high resolution and in 3D to make a reference data set that was used to create a generalizable computational framework. This enables cell shapes and the locations of internal structures to be measured and compared using rigorous statistical methods.



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).

28

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



https://www.cell.com/molecular-cell/fulltext/S1097-2765(25)00970-0

29

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**

30

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
BLAST! (the algorithm), MMSeqs2, & FoldSeek
Biological databases
Markov processes and Hidden Markov Models

31

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: Incorporating AI effectively into your coding habits
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, precision/recall/ROC analysis
Principal component analysis and data transformations
Guest lecture: Protein 3D structure prediction, incl. AlphaFold
Guest lecture: AI/deep neural networks and large language models

32

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

THE FINAL COURSE PROJECT IS DUE by 10 PM, April 15, 2026

The last 3 class days will be for presenting your projects

33