

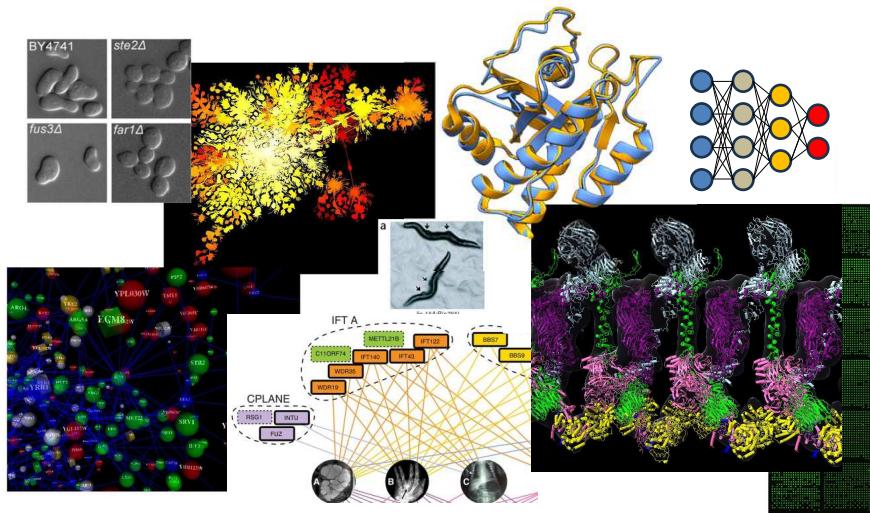
BCH394P/BCH364C Systems Biology & Bioinformatics

(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](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/>

ROSALIND About Problems Statistics Glossary search My Classes edward.marcotte Log out

BCH394P/364C (Spring 2026) Systems Biology/Bioinformatics

Edit class info Edit problems Enroll link Grade sheet Assistants Print all problems Announcements All classes Delete

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, AI/machine learning, and protein 3D structure prediction/design.

Num	Title	Solved By	Cost	Due Date	Questions	Solutions
1	Installing Python	0	1	Jan. 21, 2026	0	0
2	Variables and Some Arithmetic	0	1	Jan. 21, 2026	0	0
3	Strings and Lists	0	1	Jan. 21, 2026	0	0
4	Conditions and Loops	0	1	Jan. 21, 2026	0	0
5	Working with Files	0	2	Jan. 21, 2026	0	0
			6			

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

9

ROSALIND About Problems Statistics Glossary search My Classes edward.marcotte Log out

Installing Python

Problem 1 @ BCH394P/364C (Spring 2026) 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.

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

Questions

Found a typo? Take a tour

10

ROSALIND

About Problems Statistics Glossary search My Classes edward.marcoette Log out

Topics: Introductory Exercises, Programming

Installing Python

Problem 1 @ BCH394P/364C (Spring 2026) Systems Biology/Bioinformatics ↗

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

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 Python version 2, but we'll use version 3

Rosalind uses the “vanilla” installation of Python.

You're welcome to do it this way, but I recommend Anaconda/Jupyter as a nicer option

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.

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

Questions

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)

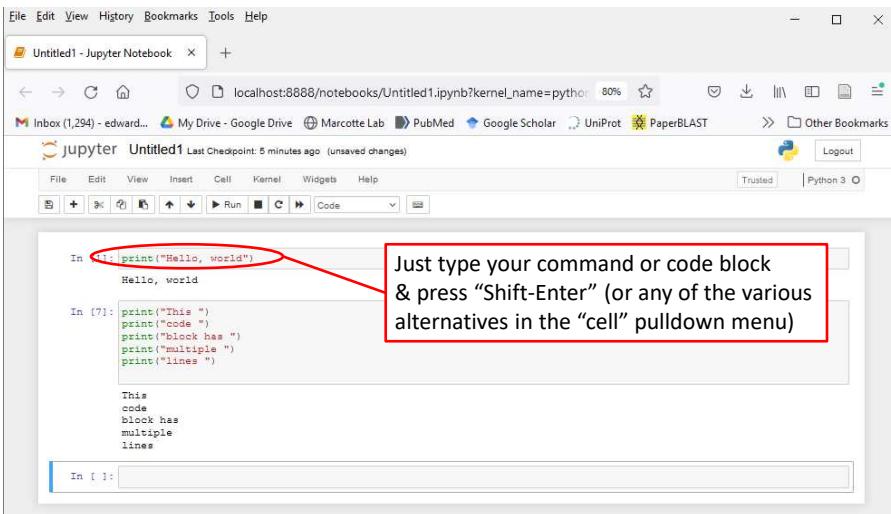
*****Get the latest Python 3 version*****
(but any version > 3.0 is probably fine)

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.



The screenshot shows a Jupyter Notebook interface. At the top, there's a menu bar with File, Edit, View, History, Bookmarks, Tools, and Help. Below the menu is a toolbar with various icons. The main area is a notebook with two cells. The first cell, labeled 'In [1]:', contains the Python code `print("Hello, world")`. The output 'Hello, world' is shown in the cell below. The second cell, labeled 'In [2]:', contains the code `print("This ")
print("code ")
print("block has ")
print("multiple ")
print("lines")`. The output 'This code block has multiple lines' is shown in the cell below. A red box highlights the first cell, and a red arrow points from the text 'Just type your command or code block & press "Shift-Enter" (or any of the various alternatives in the "cell" pulldown menu)' to the cell content.

It's widely used in bioinformatics and data visualization.

13

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



If you're feeling restless/adventurous...

Installing Python

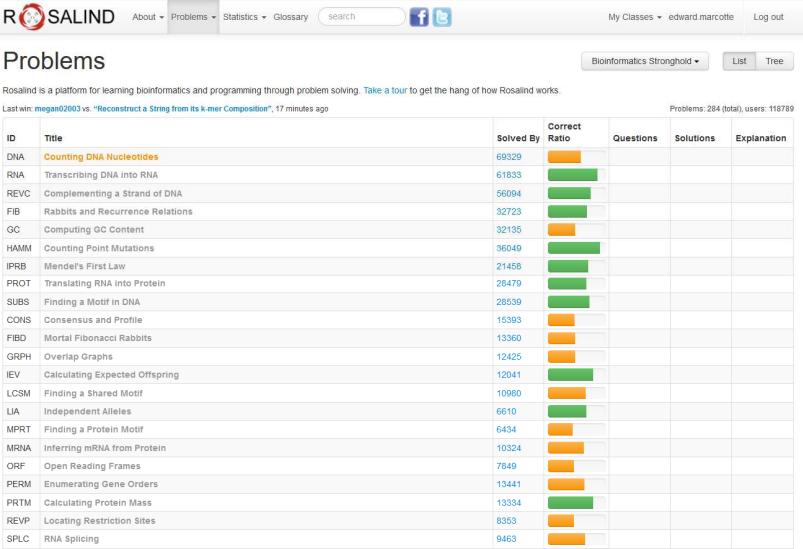
Problem 1 @ BCH394P/364C (Spring 2026) Systems Biology/Bioinformatics ↗

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

Topics: Introductory Exercises, Programming

14

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



The screenshot shows the Rosalind platform interface. At the top, there is a navigation bar with links for 'About', 'Problems', 'Statistics', 'Glossary', 'search', 'My Classes', 'edward.marcotte', and 'Log out'. Below the navigation bar, a red banner displays the text '...there are quite a few good bioinformatics problems in the archives.' The main content area is titled 'Problems' and shows a table of 40 bioinformatics problems. Each row in the table includes the problem ID, title, solved count, correct ratio (represented by a bar chart), and links for 'Questions', 'Solutions', and 'Explanation'. The problems listed include: Counting DNA Nucleotides, Transcribing DNA into RNA, Complementing a Strand of DNA, Rabbits and Recurrence Relations, Computing GC Content, Counting Point Mutations, Mendel's First Law, Translating RNA into Protein, Finding a Motif in DNA, Consensus and Profile, Mortal Fibonacci Rabbits, Overlap Graphs, Calculating Expected Offspring, Finding a Shared Motif, Independent Alleles, Finding a Protein Motif, Inferring mRNA from Protein, Open Reading Frames, Enumerating Gene Orders, Calculating Protein Mass, Locating Restriction Sites, and RNA Splicing. The table is paginated with 'Problems: 284 (total), users: 118789' and 'List' and 'Tree' buttons at the top right.

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 **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.** Asking AI chatbots to answer your homework for you is too.

19



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.

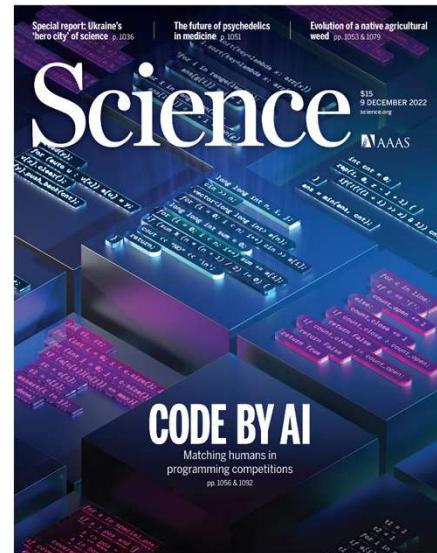
Grade-related penalties are routinely assessed ("F" in the course is not uncommon), but students can also be suspended or even permanently expelled from the University for scholastic dishonesty.

<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.

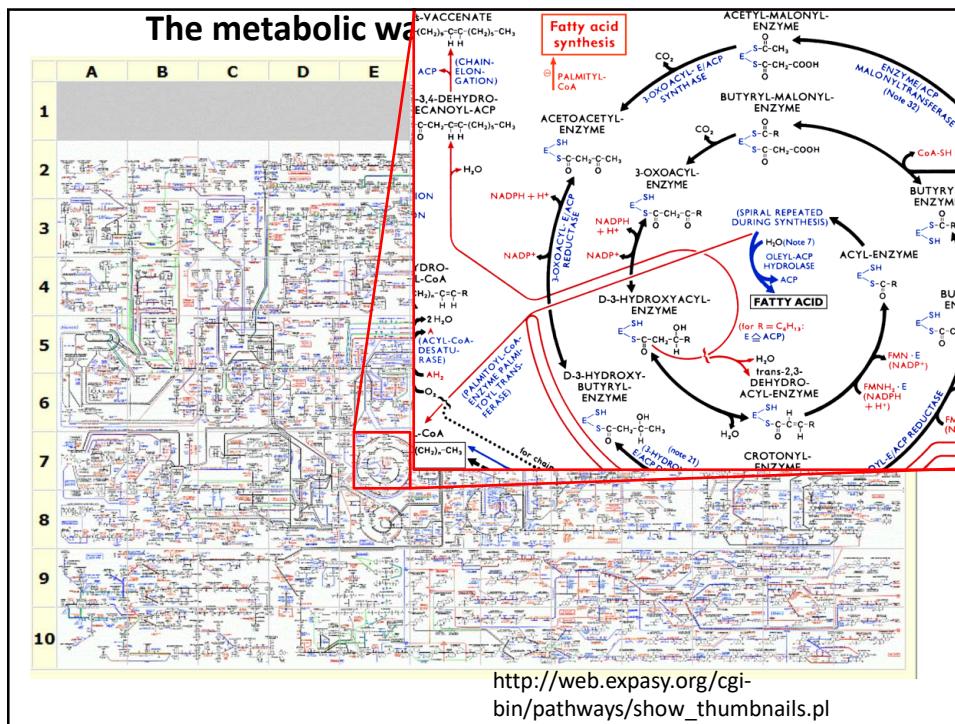


21

Why are we here?

(practically, not existentially)

22



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

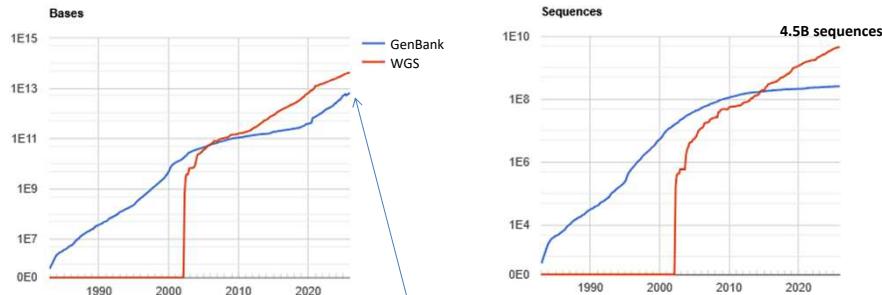
Published online 16 November 2022
https://doi.org/10.1002/anie.202208008

HMDB 5.0: the Human Metabolome Database for 2022

David S. Wishart^{1,2,3,4,*}, AnChi Guo¹, Eponine Ofer¹, Fei Wang¹, Afia Anjum¹, Harrington Peters¹, Raymund Duman¹, Zhenyu Li¹, Sheng Tang¹, Bridget L. Lee¹, Mark A. Lachance¹, Michael J. Lachance¹, Michael J. Lachance¹, Olga Tsvetkova¹, Mikel Hisham-Giesbrecht¹, Vicki W. Lu¹, Donna Varszegi¹, Dorota Varszegi¹, Dana Allen¹, David Amdt¹, Nitin Khetarpal¹, Aadhyava Sivakumaran¹, Karvena Harford¹, Selena Santoro¹, Kristen Yee¹, Xuan Cao¹, Zeljko Buncic¹, Jianus Ligand¹, Lin Peng¹, Jianlin Zheng¹, Rajneeti Mandel¹, Neema Karu¹, Majda Demirov¹, Heigl B. Schiess¹, Russell Greiner¹ and Vasak Gaudan¹

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!

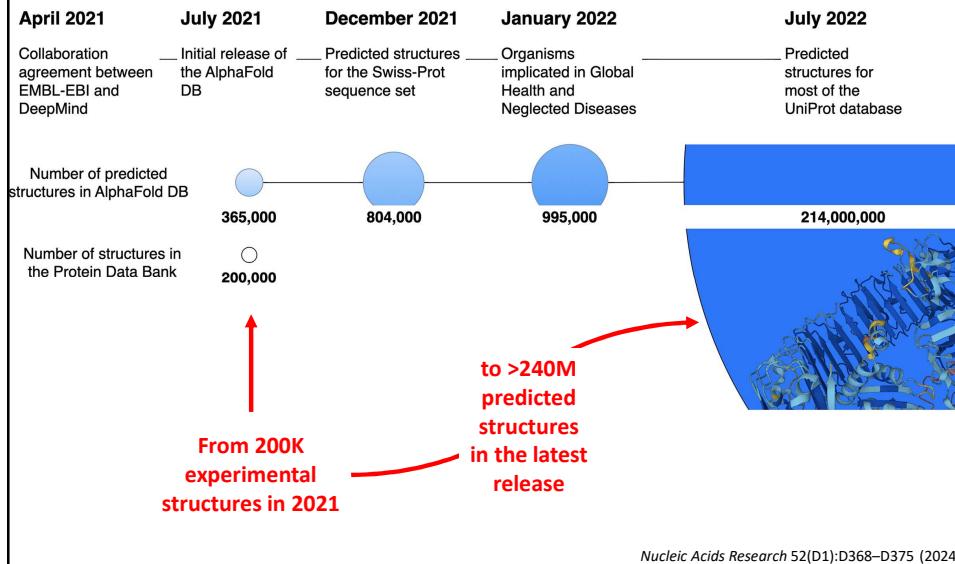
<http://www.ncbi.nlm.nih.gov/genbank/statistics>

25

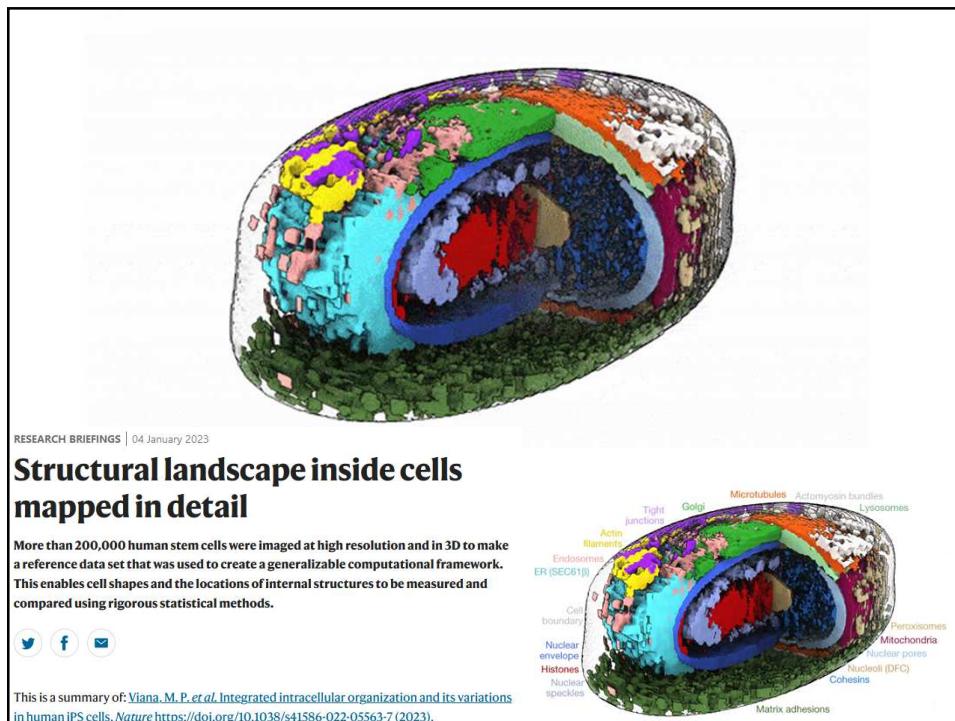


26

Resulting in huge growth in 3D structural data:

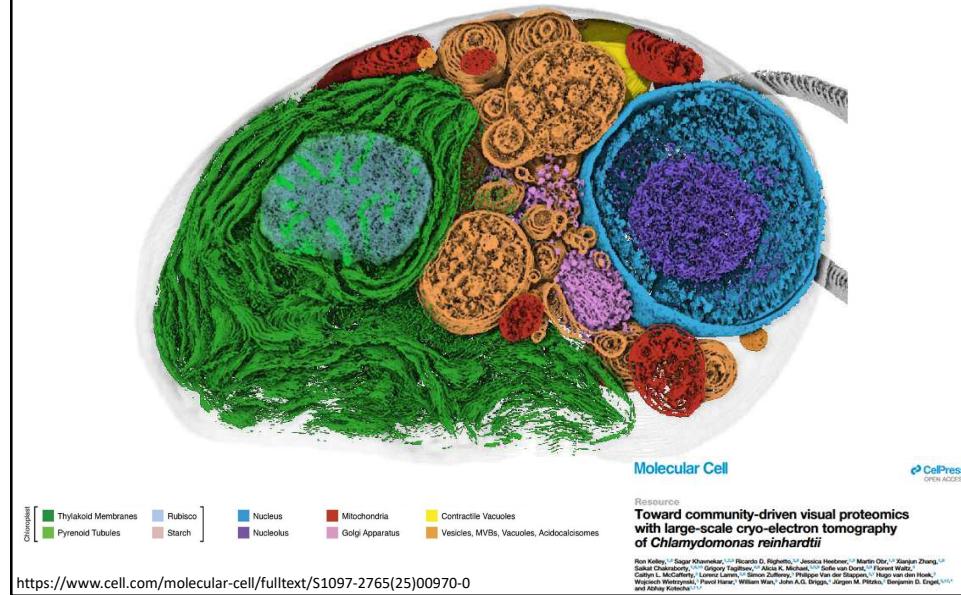


27



28

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



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 (BLOSUM, 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