Introduction to Bioinformatics: From the Bench to NGS Data Analysis

Tsai-Ming Lu

Assistant Research Scientist

Institute of Cellular and Organismic Biology

LSL Bioinformatics series

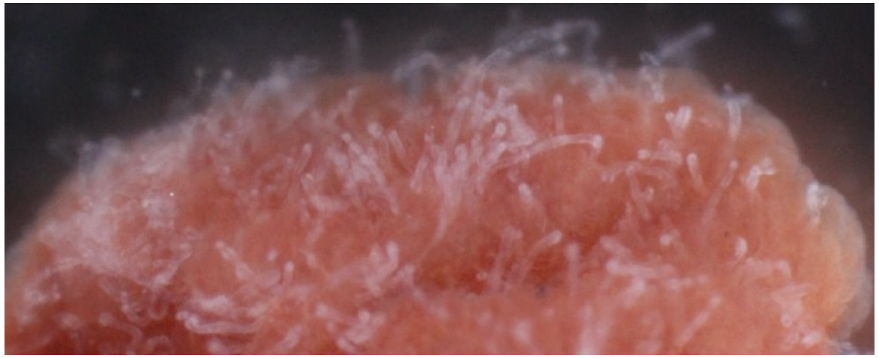
2024/05/31

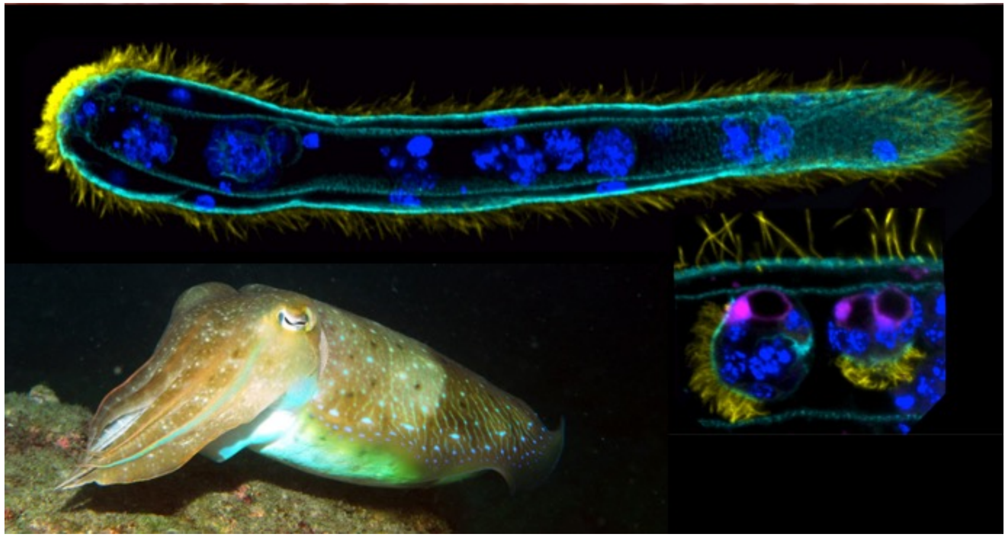
My background and research topic

ICOB Bioinformatics Core

I am currently employing bioinformatic approaches, combined with my background in the field of EvoDevo, to study the symbiotic relationships in animals.

Symbiotic adaptations between dicyemids and cephalopods



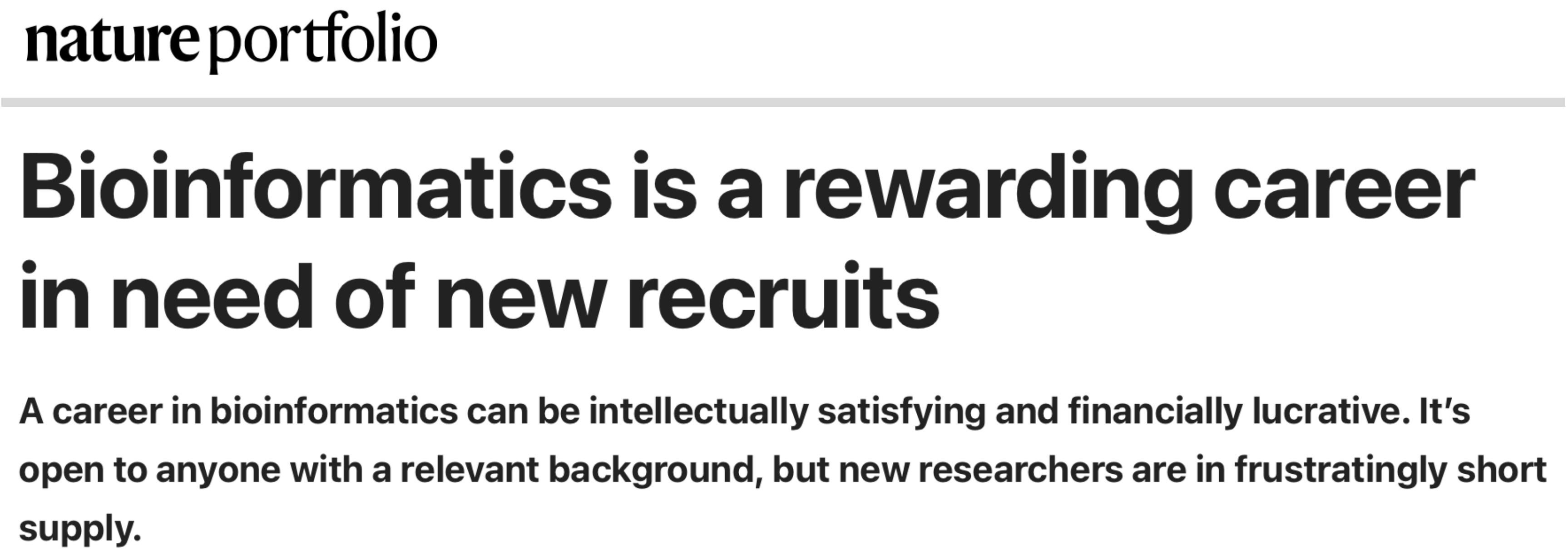


Bioinformatics is an interdisciplinary field

Bioinformatics is an interdisciplinary field that combines biology, computer science, statistics, and … to help answer questions in biological sciences.

Bioinformatics uses computers to store, process, analyze, manage, and retrieve large amounts of biologic data.

More people with bioinformatics skills are needed



nature portfolio

Bioinformatics is a rewarding career in need of new recruits

A career in bioinformatics can be intellectually satisfying and financially lucrative. It's open to anyone with a relevant background, but new researchers are in frustratingly short supply.

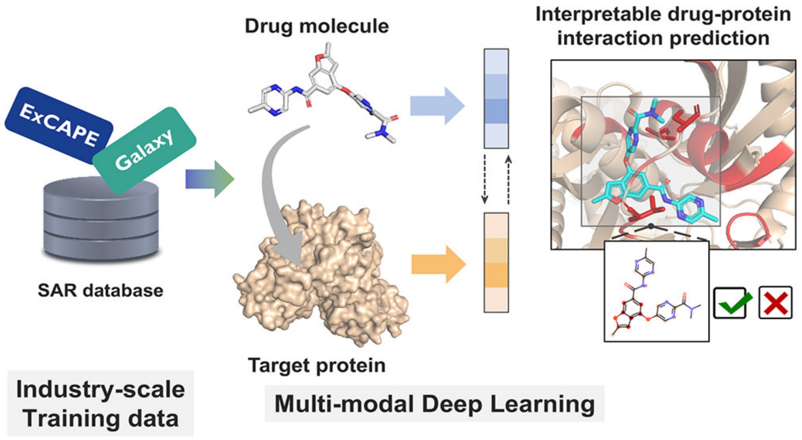
Bioinformatics is vital for advancements in other fields such as health and medicine, agriculture, conservation biology, and ...

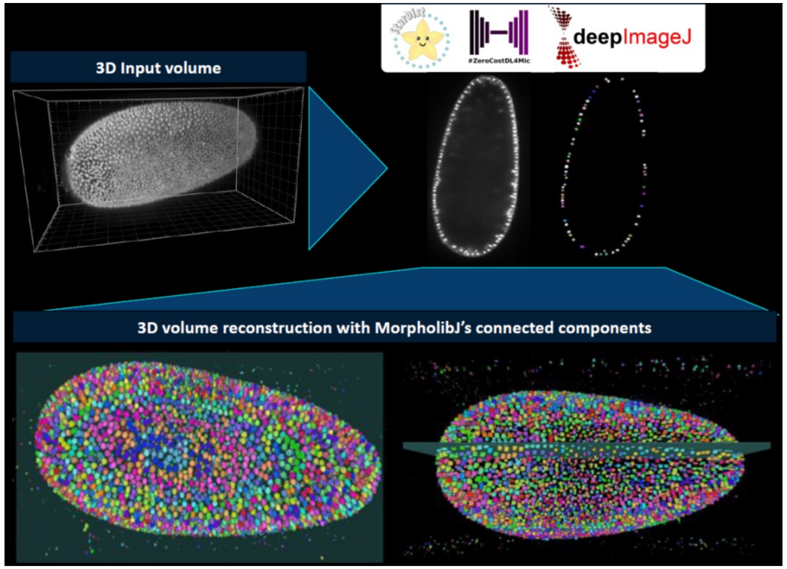
https://www.nature.com/articles/d42473-023-00243-0

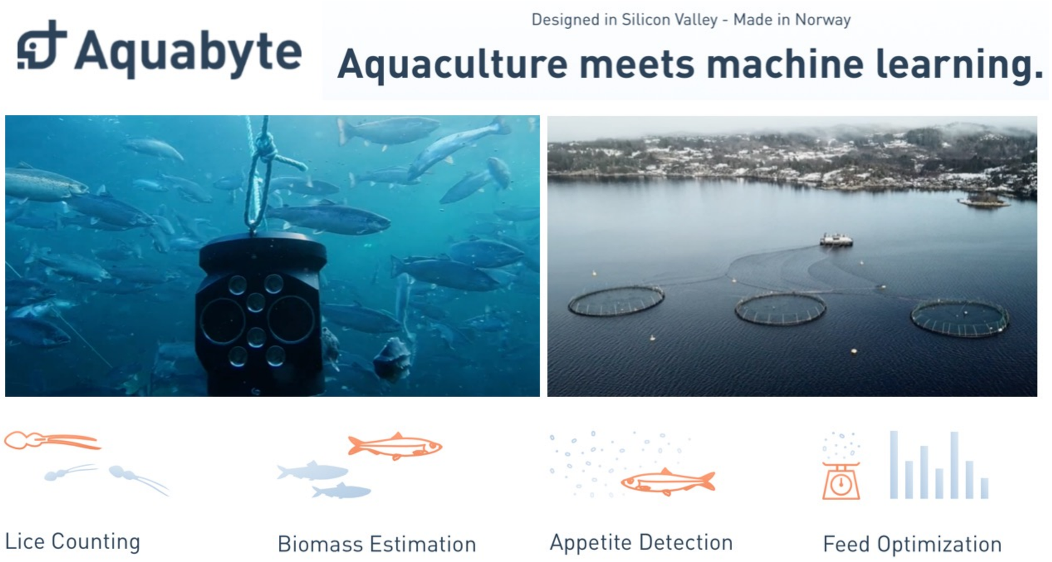
Types of bioinformatics data

Bioinformatics is a very large field.

If you’re interested in self-learning, it’s better to focus on one topic at a time.


Drug molecule
Interpretable drug-protein interaction prediction
ExCAPE
Galaxy
SAR database
Target protein
Industry-scale Training data
Multi-modal Deep Learning


deepImageJ
#ZeroCostDL4Mic
3D Input volume
3D volume reconstruction with MorpholibJ’s connected components


Aquabyte
Designed in Silicon Valley - Made in Norway
Aquaculture meets machine learning.
Lice Counting
Biomass Estimation
Appetite Detection
Feed Optimization

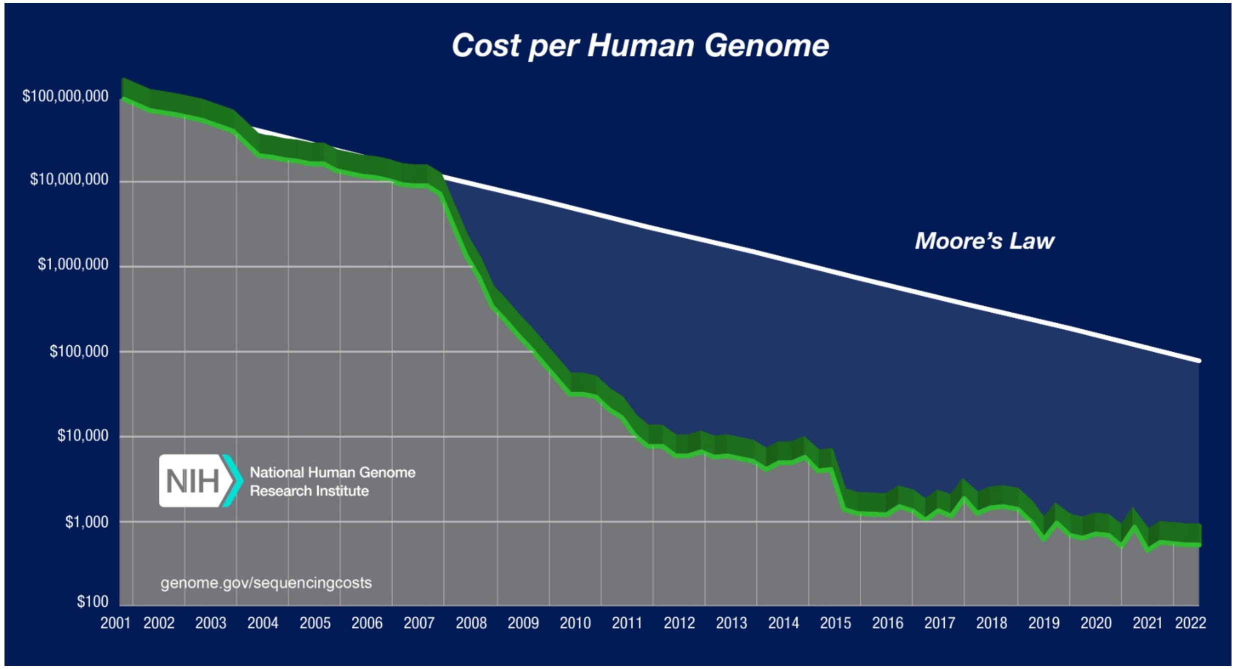
J. Chem. Inf. Model. 2022; 62(5): 1308–1317

Why Bioinformatics?

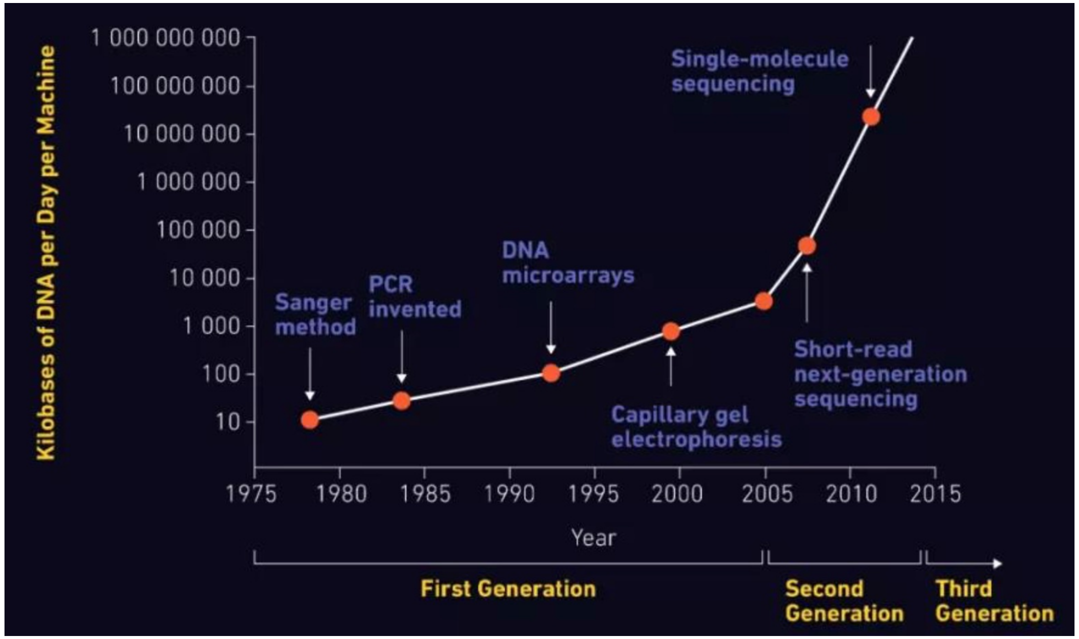
Why Bioinformatics? Biology’s Growing Data

NGS fundamentally changed the ways of many research fields.

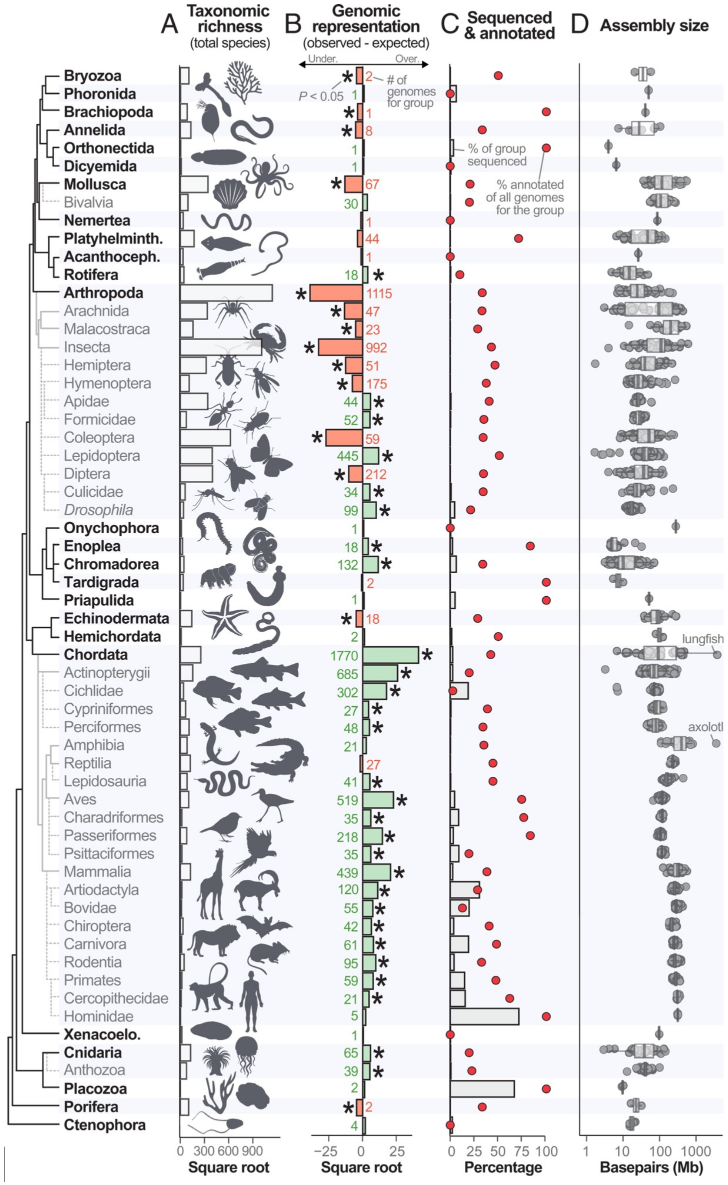
Sequencing costs dropped drastically, allowing researchers to utilize NGS data to help answer important biological questions.



Rapid accumulation of NGS data



https://medicaltrend.org/2021/03/16/overview-of-next-generation-sequencing-technology/



The gap between the rapid accumulation of data and the ability to effectively utilize data continues to widen.

Briefings in Bioinformatics 2019; 20(2): 398–404. PNAS 2021; 118(52): e2109019118.

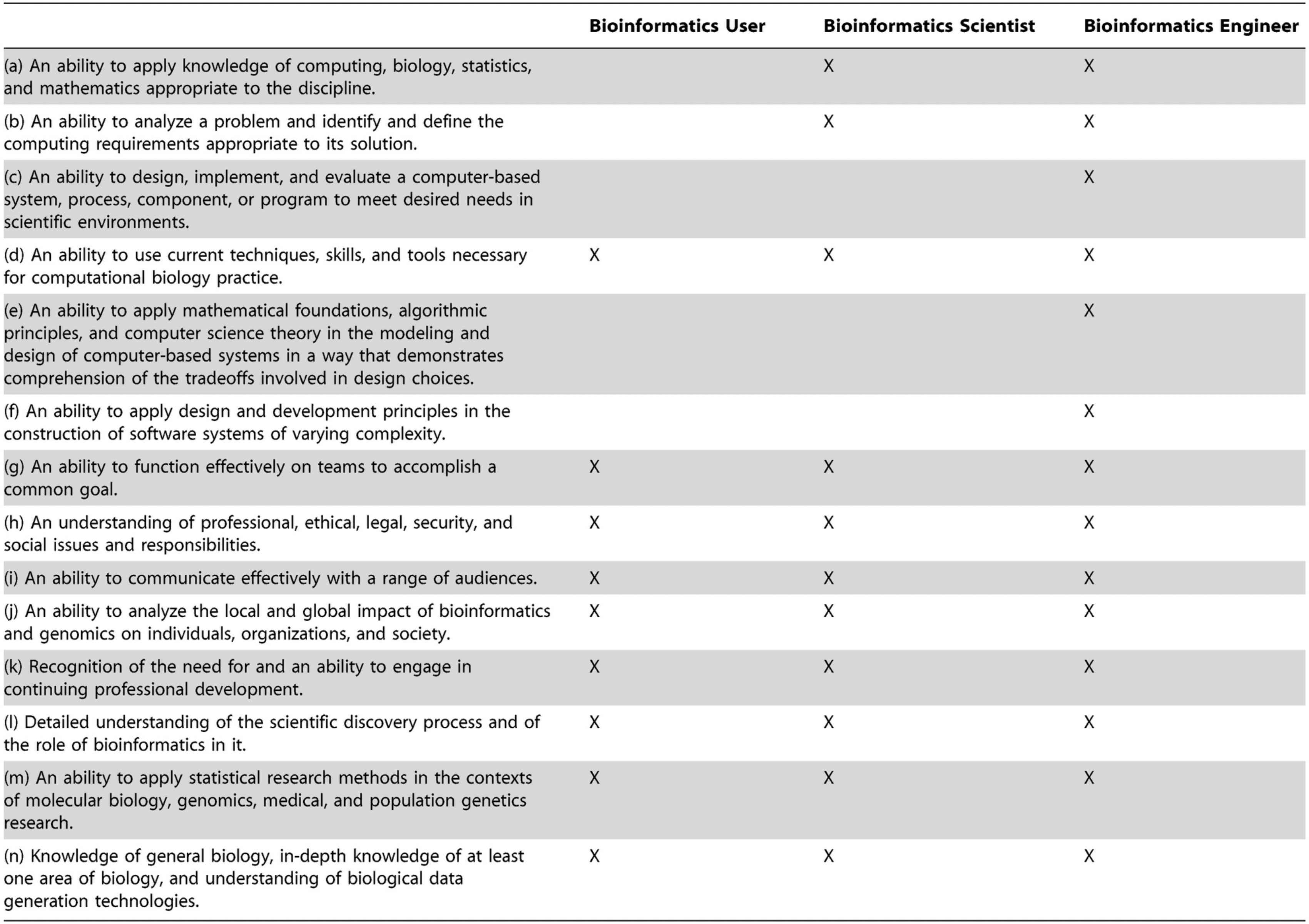
Learning core bioinformatics skills

With the nature of biological data changing so rapidly,
how are you supposed to learn bioinformatics?

A survey in 2013, ~200 responses, from students to senior researchers:
Most (76%) considered their bioinformatics skills to be self-taught.
Only 20% reported having acquired their skills during university courses.

Now, scientists not directly involved in a bioinformatics program need to be skilled at basic concepts of bioinformatics tools to avoid misuse and erroneous interpretations of the results.

Three bioinformatics user categories



	Bioinformatics User	Bioinformatics Scientist	Bioinformatics Engineer
(a) An ability to apply knowledge of computing, biology, statistics, and mathematics appropriate to the discipline.		X	X
(b) An ability to analyze a problem and identify and define the computing requirements appropriate to its solution.		X	X
(c) An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs in scientific environments.			X
(d) An ability to use current techniques, skills, and tools necessary for computational biology practice.	X	X	X
(e) An ability to apply mathematical foundations, algorithmic principles, and computer science theory in the modeling and design of computer-based systems in a way that demonstrates comprehension of the tradeoffs involved in design choices.			X
(f) An ability to apply design and development principles in the construction of software systems of varying complexity.			X
(g) An ability to function effectively on teams to accomplish a common goal.	X	X	X
(h) An understanding of professional, ethical, legal, security, and social issues and responsibilities.	X	X	X
(i) An ability to communicate effectively with a range of audiences.	X	X	X
(j) An ability to analyze the local and global impact of bioinformatics and genomics on individuals, organizations, and society.	X	X	X
(k) Recognition of the need for and an ability to engage in continuing professional development.	X	X	X
(l) Detailed understanding of the scientific discovery process and of the role of bioinformatics in it.	X	X	X
(m) An ability to apply statistical research methods in the contexts of molecular biology, genomics, medical, and population genetics research.	X	X	X
(n) Knowledge of general biology, in-depth knowledge of at least one area of biology, and understanding of biological data generation technologies.	X	X	X

Briefings in Bioinformatics 2019; 20(6): 1981–1996. PLoS Comput Biol 2014; 10(3): e1003496.

Three bioinformatics user categories

Core competencies for all three user categories:

knowledge of general biology, in-depth knowledge of at least one area of biology

using current techniques, skills, and tools necessary for computational biology practice

applying statistical methods in the contexts of molecular biology, genomics, medical, and population genetics research

Briefings in Bioinformatics 2019; 20(6): 1981–1996. PLoS Comput Biol 2014; 10(3): e1003496.

How to self learn Bioinformatics

Start reading lots of academic papers

- Try to reproduce these results yourself
- Learning by doing

Learn programming languages,

- python and R
- Fill the gap between users and developers

Soon or later you are going to end up working in a Linux/Unix environment.

Why Linux?

Many open-source bioinformatics tools are command-line and are only available in Linux.

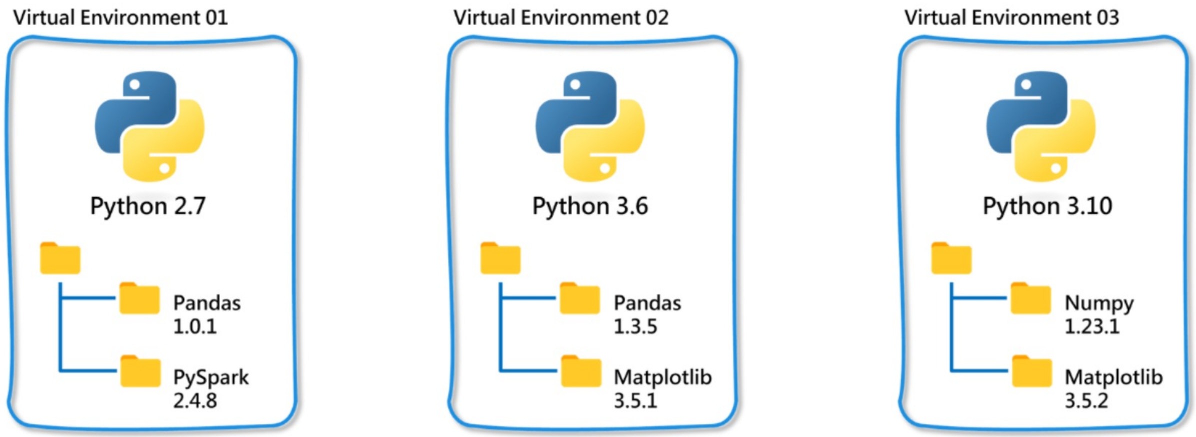
Free; easy to create analysis pipeline by integrating multiple tools

How to Practice Bioinformatics

If you get stuck, google!

Most common problems have solutions available online.

Conda is a powerful command line tool for package and environment management.



Prepare analysis pipeline

Identifying data requirements

Collecting data

Cleaning data (QC)

Data analysis

Interpreting data

Visualizing data

Pay attention to experimental design

No brilliant analysis can save an experiment with a bad design, especially necessary in high-throughput studies, as technical “batch effects” can significantly confound studies.

ACADEMIA SINICA DATA INFORMATION STATISTICAL COOPERATION CENTER
資料科學統計合作社 — 統計暨資料科學諮詢服務

To consult your statistician about any experimental design questions or concerns you may have in planning an experiment.

Choose appropriate tools

New tools are continually being created.

Organisms are all different, and their genomes are too.
Often, bioinformatics tools are not adequately benchmarked,
or if they are, they are only benchmarked in one organism.

Read the manual / usage documentation

Always be clear on the purpose of each step,
know what you are doing

Reproducible research; document everything

Documentation of methods, data versions, and code would have not only facilitated reproducibility, but it likely would have prevented serious errors.

Document each of your analysis steps, like a detailed lab notebook, a valuable record of your steps, where files are, where they came from, or what they contain.

Troubleshooting

Pay attention to error messages

Check spelling

Errors can be silent

Code and programs may produce output (rather than stop with an error), but the output may be incorrect.

Intermediate output that is too large and high dimensional to inspect or easily visualize.

Adopt a cautious attitude, and check everything between computational steps.

Golden Rule: Never ever trust your tools (or data).

Commenting code increases code readability

Readable code makes projects more reproducible, as others can more easily understand what scripts do.

It’s much easier to find and correct errors in readable, well-commented code than messy code.

Revisiting code in the future is always easier when the code is well commented and clearly written.

Thank you for your attention!

tmlu@gate.sinica.edu.tw