CSE 8803/ CX4803 Machine Learning in Computational Biology

Lecture 1: introduction

Xiuwei Zhang

School of Computational Science and Engineering

Outline

Course logistics

Biological background & computational topics

History of computational biology

Team

Instructors:

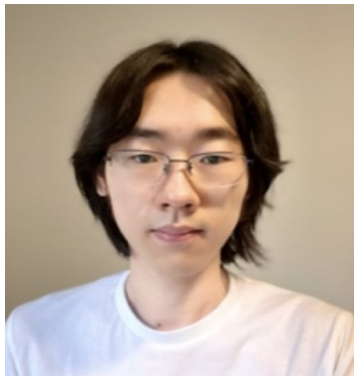
Xiuwei Zhang xiuwei.zhang@gatech.edu
Yunan Luo yunan@gatech.edu
Office hour: Monday 2:10pm-3:10pm (BlueJeans link on Canvas -> Calendar)

Teaching Assistant:

Office hours:

Wed 3-4 (Hechen)
Thu 3-4 (Hira)
Fri 12-1 (Hechen)
Fri 3-4 (Hira)
(no OH on holidays and spring break)

Hira Anis hanis3@gatech.edu



Hechen Li hli691@gatech.edu

Requirements and evaluation

Homeworks (50%)

5 homeworks in total

Due Friday night, two days of grace period

Paper presentations (20%)

One papers each team (two students form a team); may change if number of students changes drastically

Class participation (5%)

Answering questions on Piazza, discussion during paper presentations

Final exam (25%)

Grade curving

Paper presentations

Form a team of 2 students.

Each team presents one paper with around 20~25min

- Instructions in upcoming lectures

Deadline to form a team and select slots: Feb. 2nd.

- You can use Piazza to find teammates; Time slot selection will be announced later.
- Requirements on paper presentation will be introduced during lectures on 1/31 or 2/2

Pandemic related

Please wear masks!

We will try to make recordings (slides and voice)

Background/Prerequisites

- Forgot your high-school biology? Shouldn’t be an issue.
- Probability and statistics, algorithms, and linear algebra
- Prior courses on machine learning or data analytics will be a plus
- Programming language: Python, PyTorch

Textbooks, reading material

Sequence alignment, HMM application in DNA sequences:

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. (Cambridge University Press, 1998).

Machine learning methods:

Murphy, K. P. *Machine Learning: A Probabilistic Perspective*. (MIT Press, 2012).

Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. (Springer Science & Business Media, 2009).

Goodfellow, I., Bengio, Y., Courville, A. & Bengio, Y. *Deep learning*. vol. 1 (MIT press Cambridge, 2016).

Papers

Course management and communications

Canvas

https://gatech.instructure.com/courses/247960
Homework, slides, lecture recordings, exam, grades

Piazza

piazza.com/gatech/spring2022/cse8803cx4803mlb

Course website

https://cse8803mlb.github.io/spring2022/

Bioinformatics vs computational biology

BIOINFORMATICS & COMPUTATIONAL BIOLOGY = SAME? NO.

I spent the first 15 years of my professional life unwilling to recognize a difference between bioinformatics and computational biology. It was not because I didn't think that there was or could be a difference, but because I thought the difference was not significant. I have changed my position on this. I now believe that they are quite different and worth distinguishing. For me,

- the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.
- the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about discovery.

Bioinformatics vs computational biology

BIOINFORMATICS & COMPUTATIONAL BIOLOGY = SAME? NO.

I spent the first 15 years of my professional life unwilling to recognize a difference between bioinformatics and computational biology. It was not because I didn't think that there was or could be a difference, but because I thought the difference was not significant. I have changed my position on this. I now believe that they are quite different and worth distinguishing. For me,

- *Bioinformatics* = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.
- *Computational biology* = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about discovery.
- Or the opposite?... Used interchangeably in this course

PLOS | BIOLOGY

RESEARCH MATTERS

All biology is computational biology

Florian Markowetz*

University of Cambridge, Cancer Research UK Cambridge Institute, Cambridge, United Kingdom

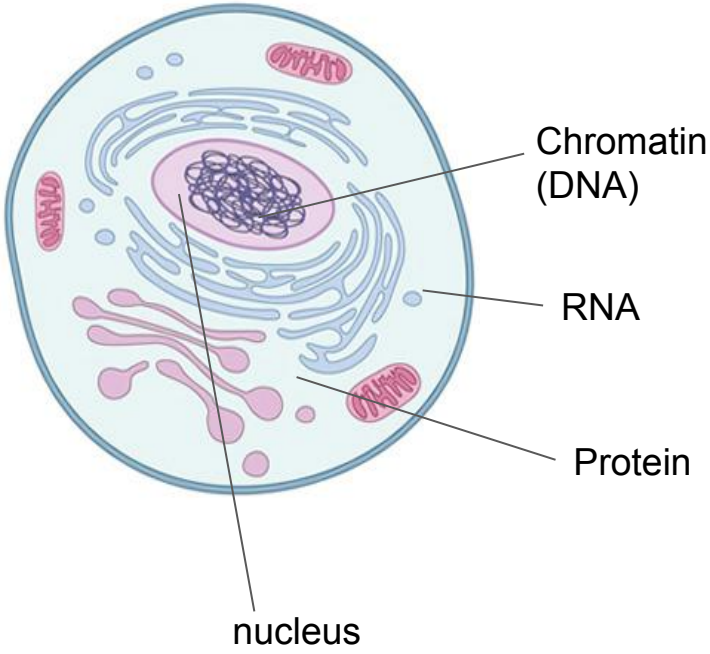
* florian.markowetz@cruk.cam.ac.uk

Abstract

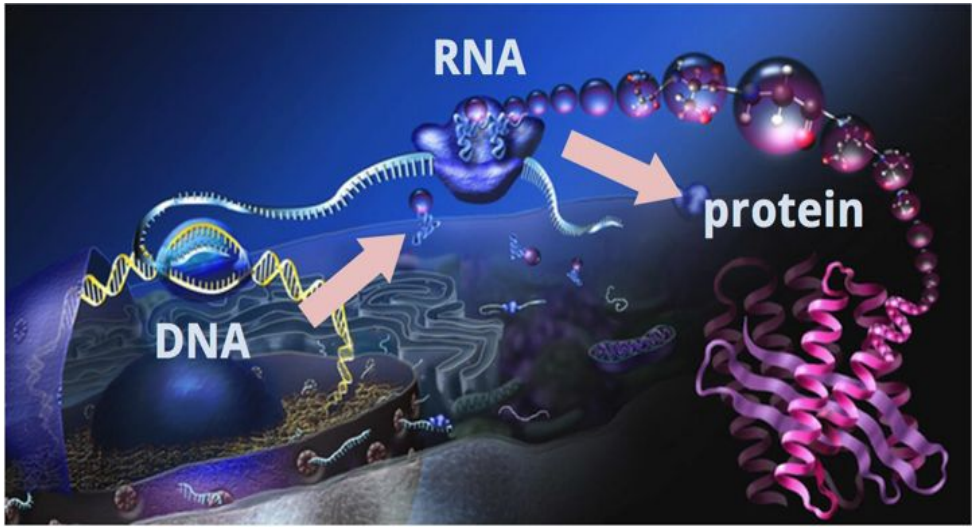
Here, I argue that computational thinking and techniques are so central to the quest of understanding life that today all biology is computational biology. Computational biology brings order into our understanding of life, it makes biological concepts rigorous and testable, and it provides a reference map that holds together individual insights. The next modern synthesis in biology will be driven by mathematical, statistical, and computational methods being absorbed into mainstream biological training, turning biology into a quantitative science.

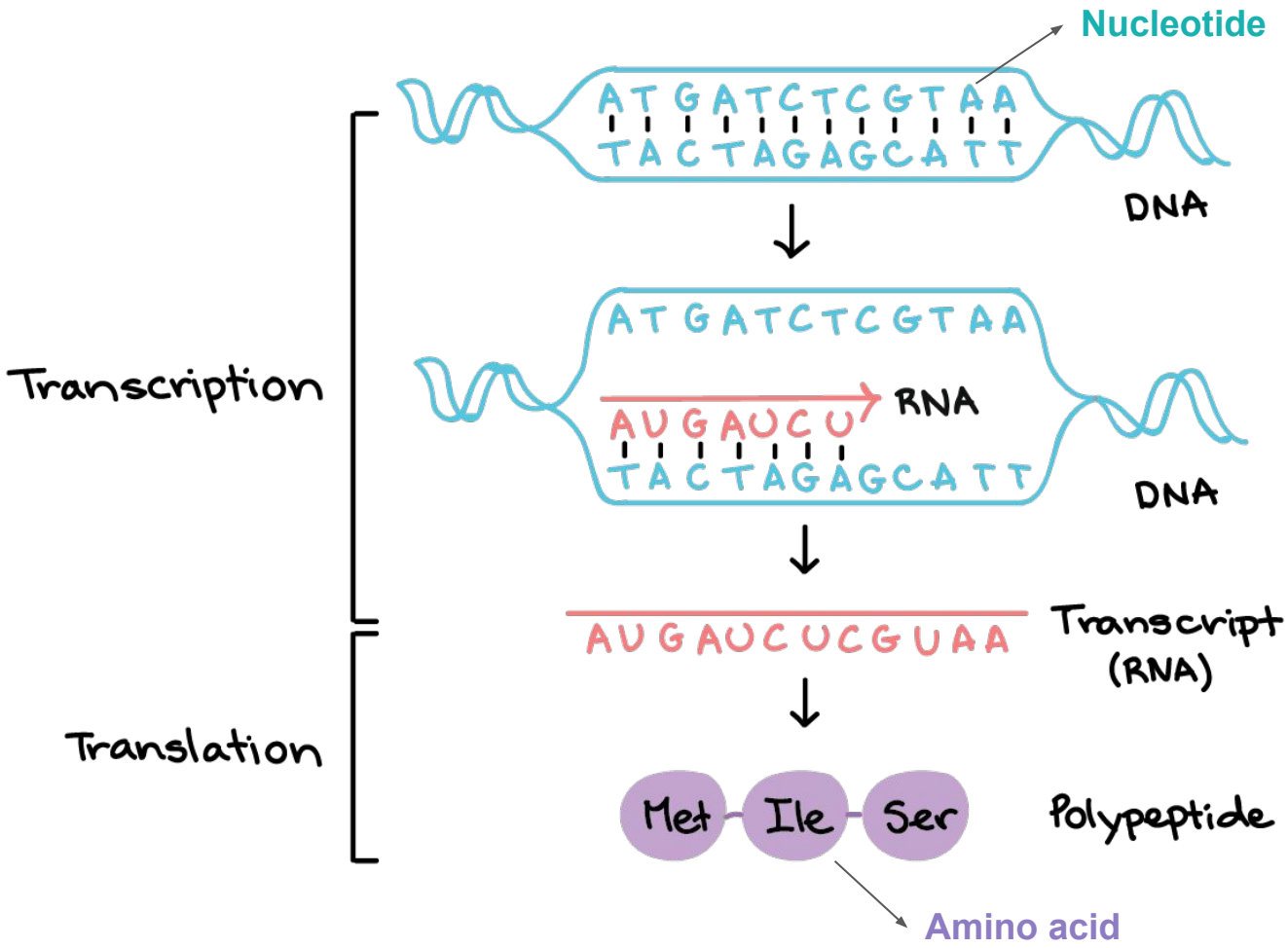
Cell

Image from https://www.nature.com/scitable/topicpage/what-is-a-cell-14023083/

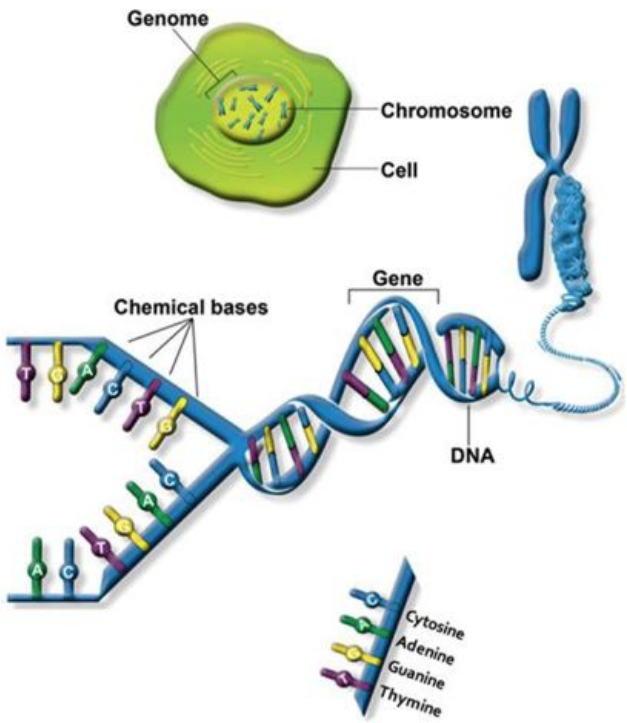


Central Dogma of molecular biology




Nucleotide
ATGATCTCGTAA
TACTAGAGCATT
DNA
Transcription
ATGATCTCGTAA
AUGAUCU
RNA
TACTAGAGCATT
DNA
AUGAUCUCGUAA
Transcript (RNA)
Translation
Met
Ile
Ser
Polypeptide
Amino acid

DNA

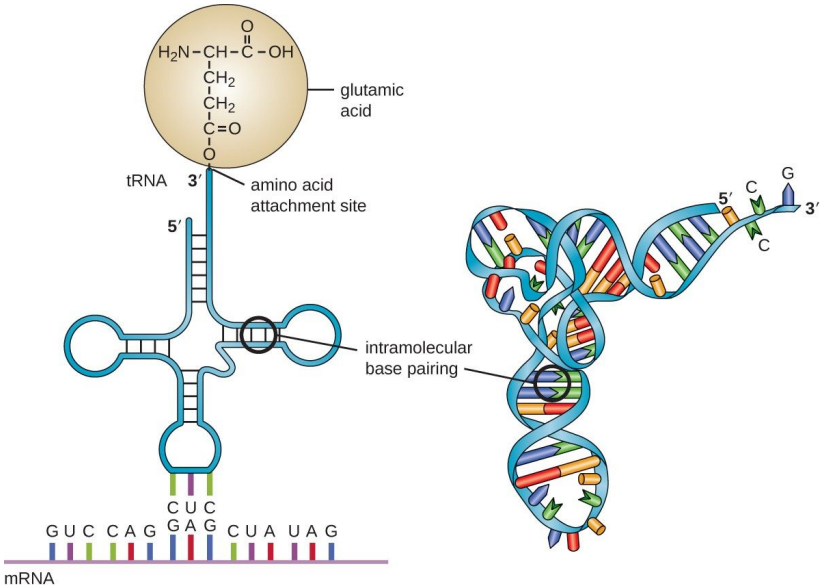



```
GCGTCTGACGGCGCACCGTTCGCGCTGCCGGCACCCCGGGCTCCATAATGAAAATCATGT
TCAGTAAGCTACACTCTGCATATCGGGCTACCAACGAAATGGAGTATCGGTCATGATCTT
GCCAGCCGTGCCTAAAAGCTTGGCCGCAGGGCCGAGTATAATTGGTCGCGGTCGCCTCGA
AGTTAGCTTATGCAATGCAGGAGGTGGGGCAAAGTTCAGGCGGATCGGCCGATGGCGGGC
GTAGGTGAAGGAGACAGCGGAGGCGTGGAGCGTGATGACATTGGCATGGTGGCCGCTTCC
CCCGTCGCGTCTCGGGTAAATGGCAAGGTAGACGCTGACGTCGTCGGTCGATTTGCCACC
TGCTGCCGTGCCCTGGGCATCGCGGTTTACCAGCGTAAACGTCCGCCGGACCTGGCTGCC
GCCCGGTCTGGTTTCGCCGCGCTGACCCGCGTCGCCCATGACCAGTGCGACGCCTGGACC
GGGCTGGCCGCTGCCGGCGACCAGTCCATCGGGGTGCTGGAAGCCGCCTCGCGCACGGCG
ACCACGGCTGGTGTGTTGCAGCGGCAGGTGGAACTGGCCGATAACGCCTTGGGCTTCCTG
TACGACACCGGGCTGTACCTGCGTTTTCGTGCCACCGGACCTGACGATTTCCACCTCGCG
...
```

Computational problems covered:

Motif/gene finding (HMM)
Sequence alignment (dynamic programming algorithms)

RNA



tRNA 2-dimensional structure

tRNA 3-dimensional structure

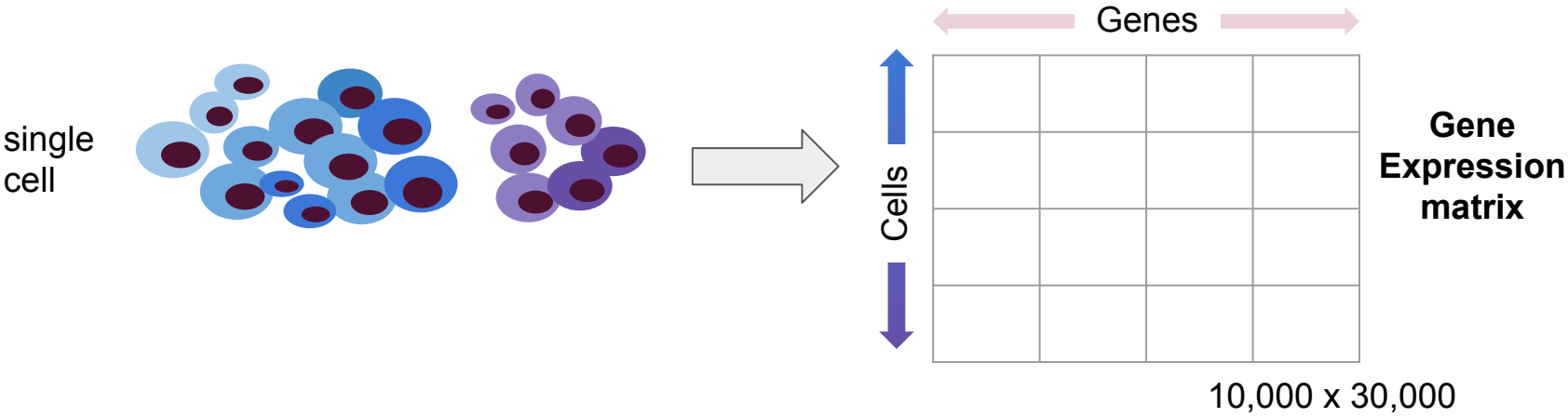
Computational problems covered:

- RNA structure prediction
  Dynamic programming algorithm
  Probabilistic graphical model (Stochastic Context-Free Grammar)
  Deep learning methods

- RNA-expression (gene-expression analysis)

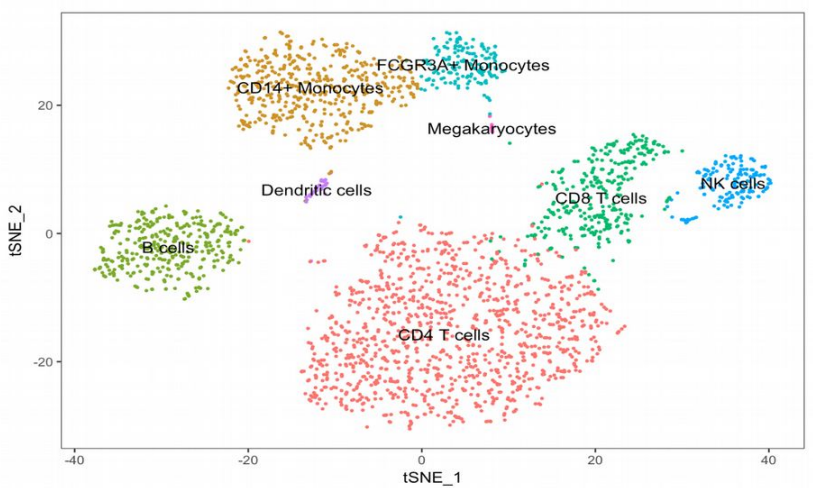
  DNA remains (almost) the same across cells, but RNA can be different across tissues/conditions/cells

Gene expression analysis

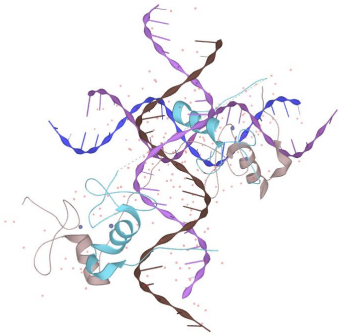


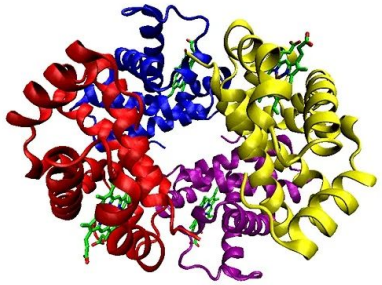
Gene-expression analysis

- Dimension reduction methods (PCA, MDS, auto-encoders, VAE, visualization in low dimensions using tSNE or UMAP, diffusion maps)
- Clustering cells to find new cell types (k nearest neighbor graphs, graph based clustering methods, matrix factorization)



Protein





DeepMind's program, called AlphaFold, outperformed around 100 other teams in a biennial protein-structure prediction challenge called CASP, short for Critical Assessment of Structure Prediction. The results were announced on 30 November, at the start of the conference – held virtually this year – that takes stock of the exercise.

Protein sequence:

Made of 20 amino acids

MEVTADQPRWVSHHHPAVLNGQHPDTHHPGLSHSYMDAAQYPLPEEVDV
LFNIDGQGNHVPPYYGNSVRATVQRYPPTHHGSQVCRPPLLHGSLPWLDG
GKALGSHHTASPWNLSPFSKTSIHHGSPGPLSVYPPASSSSLSGGHASPHL
FTFPPTPPKDVSPDPSLSTPGSAGSARQDEKECLKYQVPLPDSMKLESSHS
RGSMTALGGASSSTHHPITTYPPYVPEYSSGLFPPSSLLGGSPTGFGCKSR
PKARSSTGRECVNCGATSTPLWRRDGTGHYLCNACGLYHKMNGQNRPLIK
PKRRLSAARRAGTSCANCQTTTTTLWRRNANGDPVCNACGLYYKLHNINR
PLTMKKEGIQTRNRKMSSKSKKCKKVHDSLEDFPKNSSFNPAALSRHMSSL
SHISPFSHSSHMLTTPTPMHPPSSLSFGPHHPSSMVTAMG

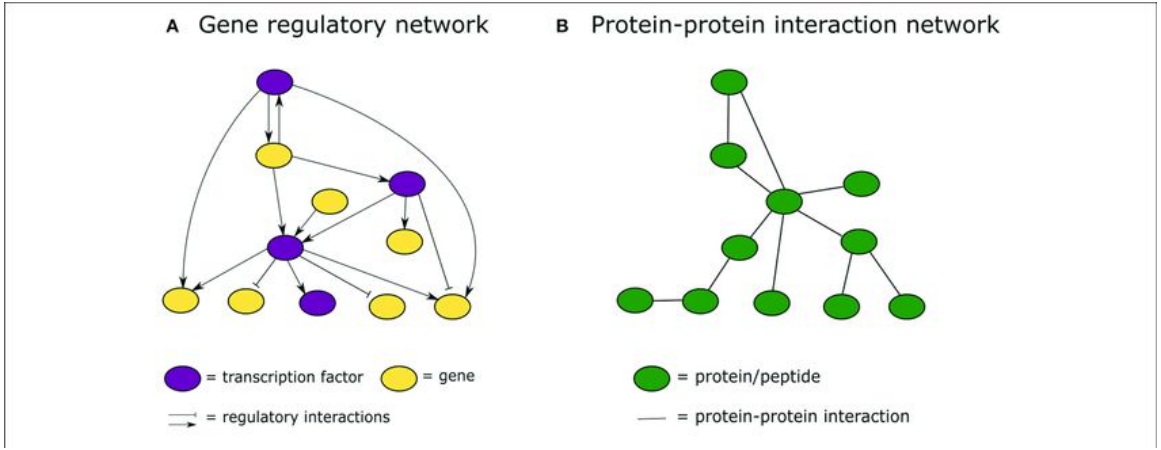
Computational problems covered:

Protein structure prediction (deep neural networks)

Traditional methods

Interactions between molecules (biological networks)

- Learning network structure and causality between molecules (Bayesian networks, decision trees, random forests)
- Comparison between multiple networks (probabilistic graphical models)
- Deep learning models for graphs



Topics

Week	Date	Topic	Contents	Instructor
1	1/10/2022	Introduction	Course intro & how to present papers	Zhang
	1/12/2022	Learning from sequence data	Dynamic programming & sequence alignment I	Zhang
2	1/17/2022		No class (MLK Day)	
	1/19/2022		Sequence alignment II	Zhang
3	1/24/2022		HMM & gene/motif finding	Zhang
	1/26/2022		HMM & Profile HMM	Zhang
4	1/31/2022		Deep learning for DNA/protein sequence	Luo
	2/2/2022	Learning from high-dim data	Learn from high-dim data: PCA, autoencoder & VAE	Luo
5	2/7/2022		Learn from high-dim data: MDS, tSNE, UMAP	Zhang
	2/9/2022		Clustering I	Zhang
6	2/14/2022		Clustering II	Zhang
	2/16/2022		Clustering III	Zhang
7	2/21/2022		Student presentation 1-3	
	2/23/2022		Student presentation 4-6	
8	2/28/2022	Learning from structure data	RNA structure prediction	Luo
	3/2/2022		Deep learning for structures (protein structure prediction)	Luo
9	3/7/2022		Student presentation 7-9	
	3/9/2022	Learning from network data	Network basics & traditinal ML for graphs	Luo
10	3/14/2022		Network embeddings	Luo
	3/16/2022		Student presentation 10-12	
11	3/21/2022		No class (Spring Break)	
	3/23/2022		No class (Spring Break)	
12	3/28/2022		Graphical Models	Luo
	3/30/2022		Deep learning for networks (graph neural networks)	Luo

Course plan

In terms of computational methods:

- Dynamic programming
- HMM (Hidden Markov Model)
- PCA (Principal Component Analysis)
- MDS (MultiDimensional Scaling)
- NMF (Non-negative matrix factorization)
- Autoencoders & VAE (variational autoencoders)
- tSNE / UMAP
- Clustering k-means
- Clustering graph based clustering Louvain/Leiden clustering
- Dynamic Programming for RNA structure prediction
- Deep learning method for RNA/protein structure prediction
- Inferring biological networks with linear regression and decision trees
- Bayesian networks and probabilistic graphical models
- Graph neural networks

History of computational biology

When could we do what?

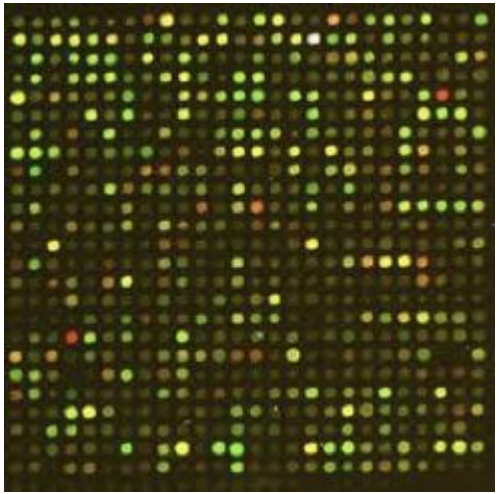
The protein wave

- 1955: Sanger sequenced bovine insulin (Nobel Prize)
- 1970: Needleman-Wunsch algorithm (dynamic programming)
- 1973: PDB (protein data bank)
- 1990: BLAST (Basic Local Alignment Search Tool)
- 1994-: CASP (Critical Assessment of protein Structure Prediction)
- 1997-: Proteomics
- 2017-: Proteomics through genomics

Slide credit: Shirley Xiaole Liu

The Microarray Wave

- Microarray contains hundreds to millions of tiny probes
- Simultaneously detect how much each gene is expressed



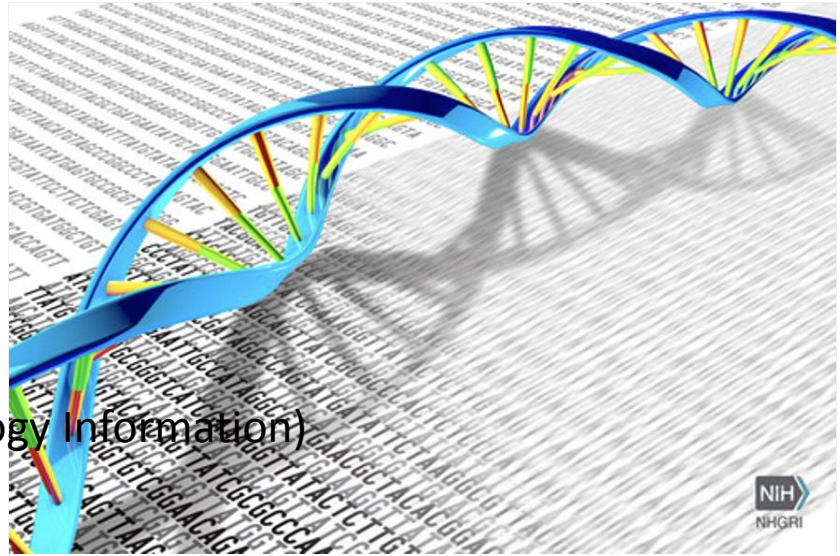
Slide credit: Shirley Xiaole Liu

The DNA Sequencing Wave

- 1953: DNA structure
- 1972: Recombinant DNA
- 1977: Sanger sequencing (Nobel Prize)
- 1985: PCR (Polymerase Chain Reaction)

  By GT alumni Kary Mullis (Nobel Prize)

- 1988: NCBI (National Center for Biotechnology Information)
- 1990: BLAST



Slide credit: Shirley Xiaole Liu

Sequencing in the 1970s

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 248, No. 11, Issue of June 10, pp. 3860-3875, 1973
Printed in U.S.A.

The Nucleotide Sequence of *Saccharomyces cerevisiae* 5.8 S Ribosomal Ribonucleic Acid

(Received for publication, November 20, 1972)

GERALD M. RUBIN*
From the Medical Research Council Laboratory of Molecular Biology, Cambridge, CB2 2QH, England

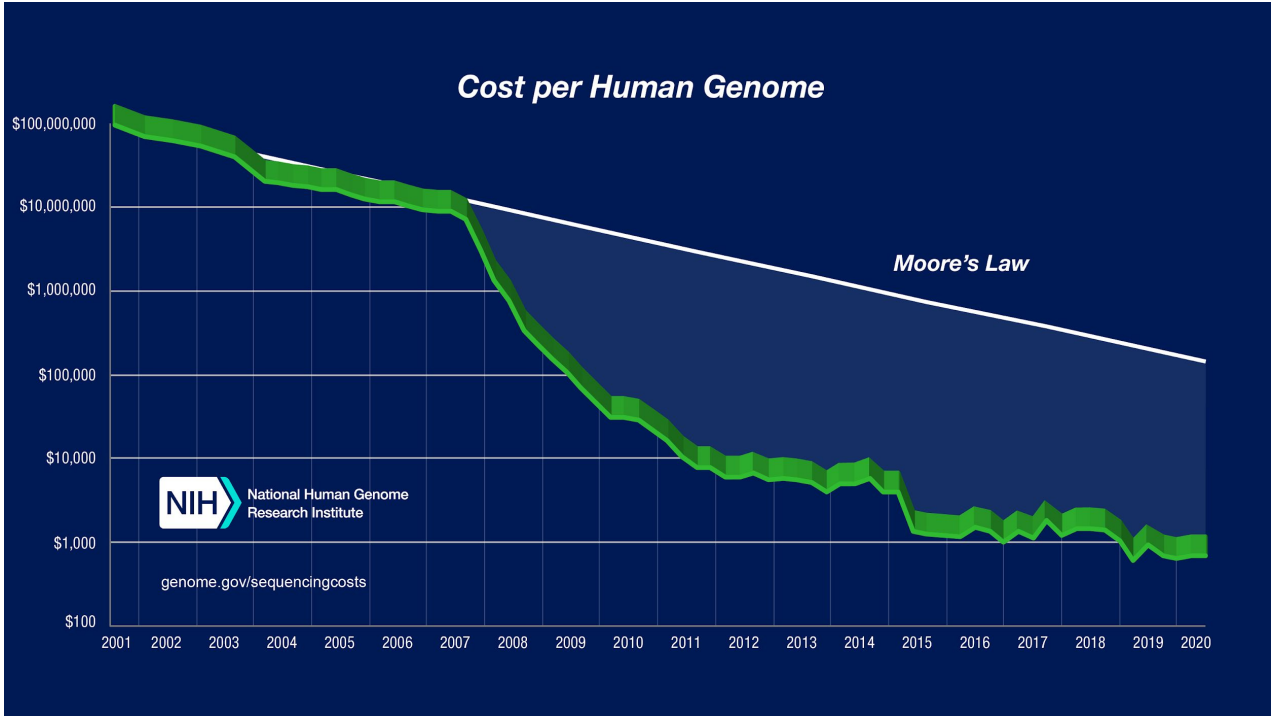
SUMMARY

The nucleotide sequence of *Saccharomyces cerevisiae* 5.8 S ribosomal RNA (also known as the 7 S or 1RNA species) has been determined to be pApApApCpUpUpUpCpApApCpA pApCpGpGpApUpCpUpCpUpUpGpGpUpUpCpUpCpGpC pApUpCpGpApUpGpApApGpApApCpGpCpApGpCpGpApA pApUpGpCpGpApUpApCpGpUpApApUpGpUpGpApAΨpUpG pCpApGpApApUpUpCpCpGpUpGpApApUpCpApUpCpGpA pApUpCpUpUpUpGpApApCpGpCpApCpApUpUpGpCpGpC pCpCpCpUpUpGpGpUpApUpUpCpCpApGpGpGpGpGpCpA pUpGpCpCpUpGpUpUpUpGpApGpCpGpUpCpApUpUpU.

Low Phosphate Medium—Inorganic phosphate was precipitated (as $MgNH_4PO_4$) from 10% Bacto-yeast extract and 20% Bacto-peptone by the addition of 10 ml of 1 M $MgSO_4$ and 10 ml of concentrated aqueous ammonia per liter. The phosphates were allowed to precipitate at room temperature for 30 min, and the precipitate was removed by filtration through Whatman No. 1 filter paper. The filtrate was adjusted to pH 5.8 with HCl and autoclaved. Sterile glucose was added to a final concentration of 2%.

Slide credit: Shirley Xiaole Liu

Cost of Sequencing Human Genomes



Moore's Law

is the observation that the number of transistors in a dense integrated circuit (IC) doubles about every two years. (wikipedia)

The single cell wave

Method of the Year 2013

Nature Methods **11**, 1(2014) | Cite this article

4967 Accesses | **24** Citations | **124** Altmetric | Metrics

Methods to sequence the DNA and RNA of single cells are poised to transform many areas of biology and medicine.

Method of the Year 2019: Single-cell multimodal omics

Nature Methods **17**, 1(2020) | Cite this article

28k Accesses | **9** Citations | **126** Altmetric | Metrics

Multimodal omics measurement offers opportunities for gaining holistic views of cells one by one.

Method of the Year 2020: spatially resolved transcriptomics

Nature Methods **18**, 1(2021) | Cite this article

7636 Accesses | **214** Altmetric | Metrics

Spatially resolved transcriptomics methods are changing the way we understand complex tissues.

Development cell by cell

AI and computational biology

Science Magazine:

2021 breakthrough of the year:

AI brings protein