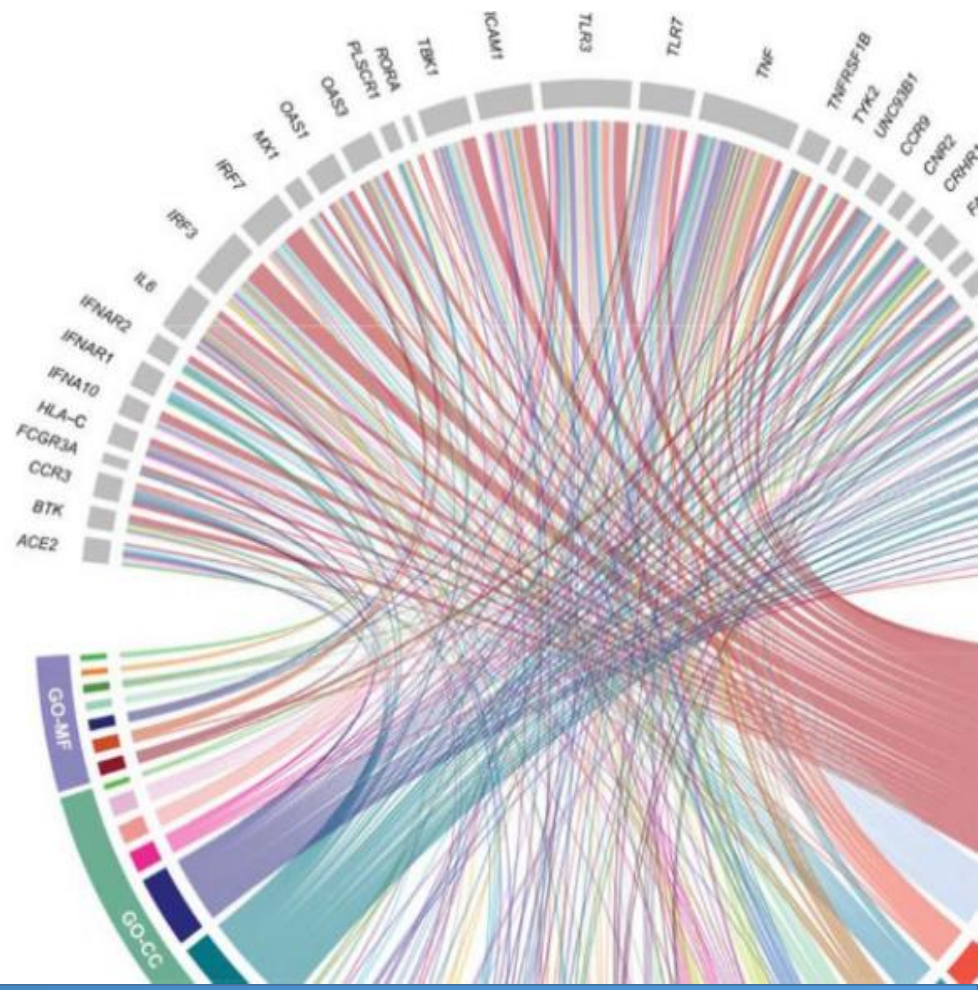
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Tècniques i Eines Bioinformàtiques

Introduction



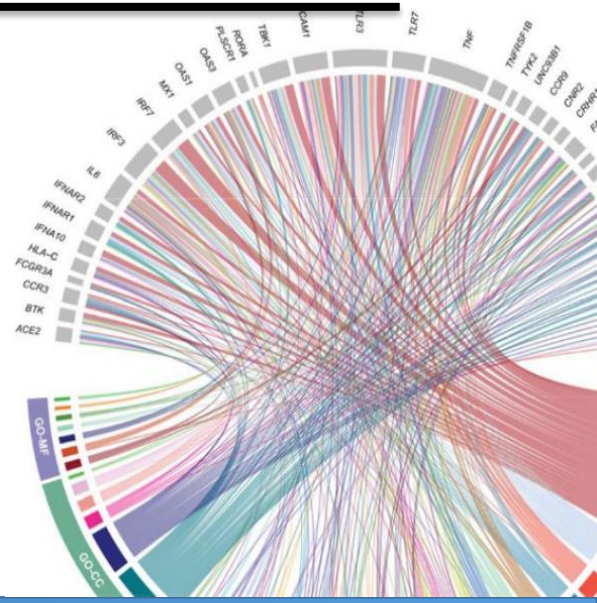
Santiago Marco-Sola (santiago.marco@upc.edu)

Màster en Enginyeria Informàtica, UPC
Departament of Computer Science
Facultat d'Informàtica de Barcelona (FIB), UPC

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Course Introduction



Instructor

- **Santiago Marco-Sola**
  - Professor at the Computer Science Department, UPC
  - Senior Researcher at Barcelona Supercomputing Center, BSC

- **Research Interests**
  - Algorithms in Bioinformatics
  - Bioinformatics Tool Development
  - Genome Sequencing Data Analysis
  - Software and Hardware Accelerators for Bioinformatics

- **Previuos Academic Teaching**
  - Algorithms, Programming, AI
  - Master in Bioinformatics

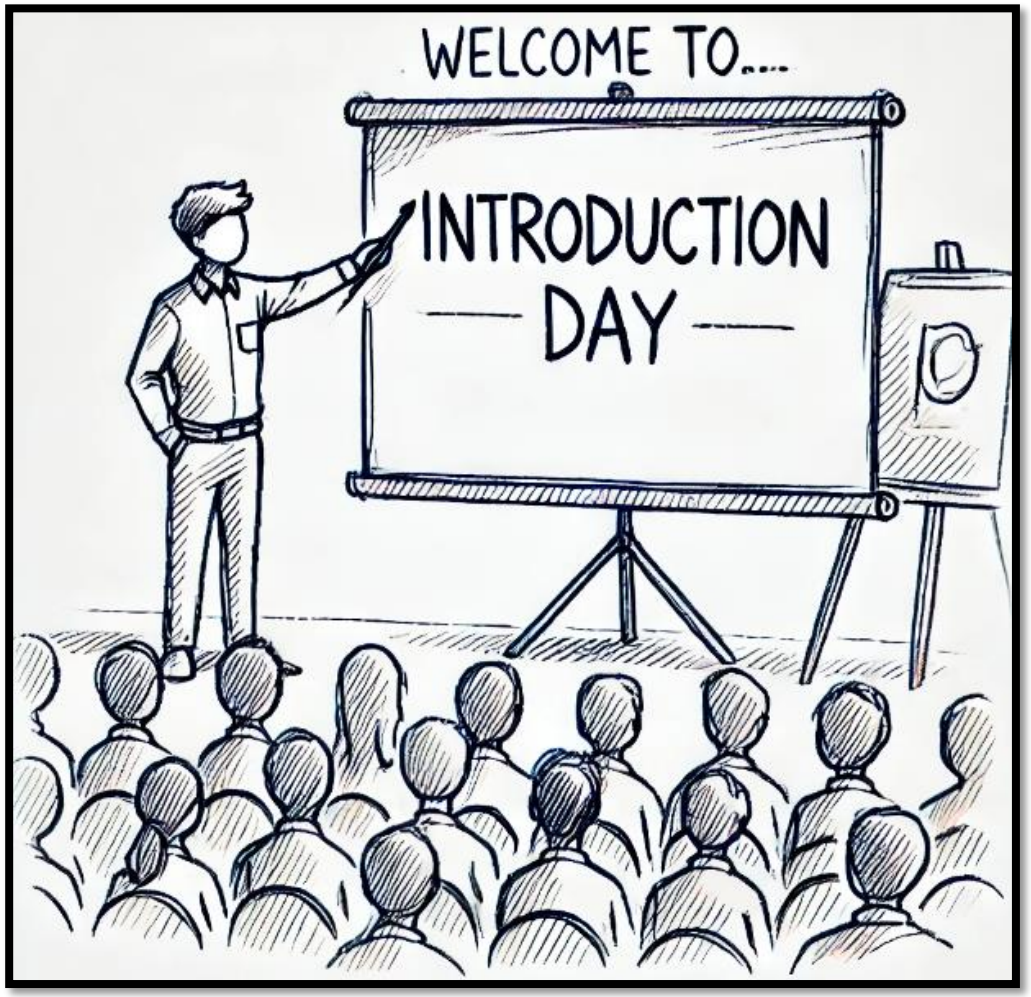
- Contact:
  - ***santiago.marco@upc.edu***

Today Schedule

- Course
  - The goal of this course
  - Content of the weekly lectures
  - Calendar
  - Grading

- What we expect from students
  - Course logistics
  - Communication with mentors

- Brief introduction:
  - Motivation
  - Genomics
  - DNA Sequencing
  - Exercises



Prerequisites of the Course

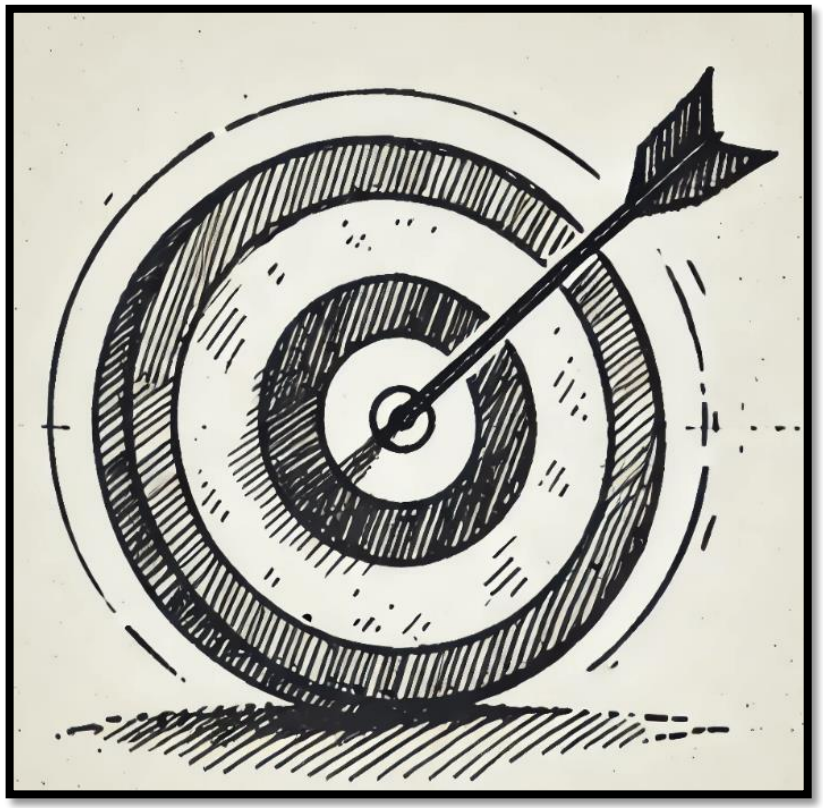
- No prior knowledge in bioinformatics or genome analysis is needed.
- Interest in algorithms and data-structures.
- Interest in optimizing efficiency and solving complex problems.
- Basic knowledge:
  - C or C++ programming language
  - Python

Objectives

- **Algorithms in Bioinformatics**
  - The course presents some of the main algorithms, techniques, and tools used in genomics and bioinformatics, such as exact and approximate string matching, pairwise and multiple sequence alignment, pattern matching, and degenerate string alignment.

- **Goal:** To be a gentle **introductory course in bioinformatics algorithm** for genome sequence analysis.
  - Understand **genome-scale algorithms and data structures** used for genome analysis.
  - **Analyze and optimize** computational performance.
  - Develop algorithmic **problem-solving skills**.

  - Motivate technological and biological context and the need for efficient algorithms and implementations.
  - Encourage research and independent learning skills.



Contents and Schedule

- This course is about **algorithms in bioinformatics** and **algorithm design techniques**.
  - What are the problems?
  - What algorithms are developed for what problem?
- This course is **not** about how to analyze biological data using tools, AI, or other methods.

	M	T	W	T	F	S
February	10	11	12	13	14	15
	17	18	19	20	21	22
	24	25	26	27	28	1
March	3	4	5	6	7	8
	10	11	12	13	14	15
	17	18	19	20	21	22
	24	25	26	27	28	29

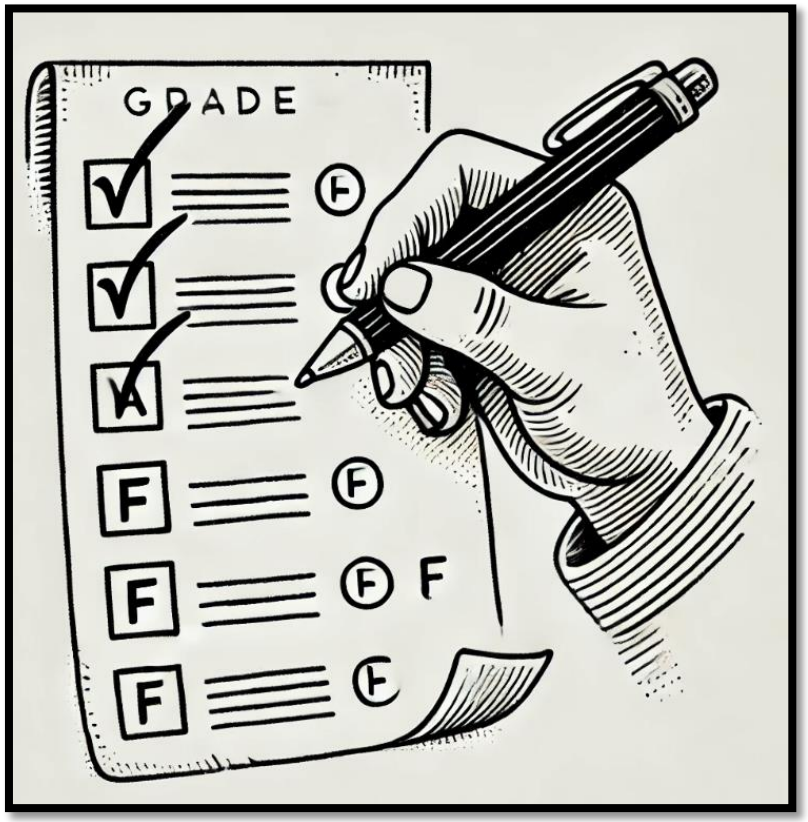
1. Genome Analysis and Sequencing
2. Exact String Matching
3. Indexing Data Structures
4. Approximate String Matching
5. Sequence Alignment

	DILLUNS	DIMARTS	DIMECRES	DIJOUS	DIVENDRES
18:00-19:00	TEB-MEI 10 T [A6206]				TEB-MEI 10 P [A6206]
19:00-20:00	TEB-MEI 10 T [A6206]				TEB-MEI 10 L [A6206]
20:00-21:00					

Grading and Research-Lectures Presentations

- **Option A (Officially):** There will be a (mid-term) final exam, in which the students will explain advanced algorithms and data structures from the research literature. The final grade will be just the final exam grade.

- **Option B:** Grading will be based on a research paper presentation. A selected list of papers will be published and each student will select a paper to present and discuss with the class for 30 mins.
  - **Optional (+2p).** A small programming project related to all the algorithms and techniques presented in class.
- Let me know your preference (*santiago.marco@upc.edu*)

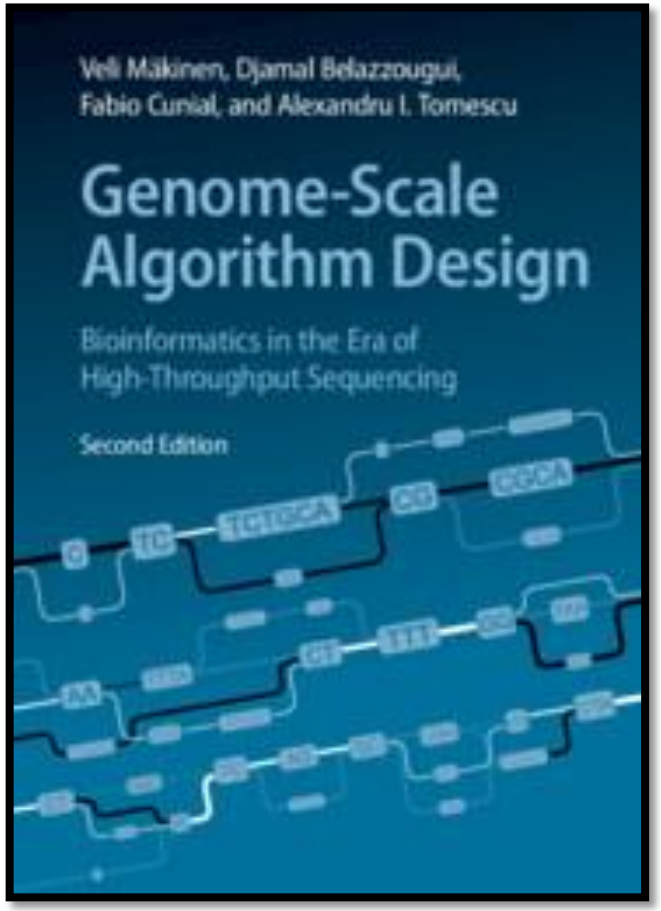


Resources

- **Slides and Course Materials:** UCP RACO
https://raco.fib.upc.edu/

- **Bibliography:** Genome-scale algorithm design : bioinformatics in the era of high-throughput sequencing - Mäkinen, Veli; Belazzougui, Djamal; Cunial, Fabio; Tomescu, Alexandru I, Cambridge University Press, 2023. ISBN: 9781009341233.
https://discovery.upc.edu/permalink/34CSUC_UPC/8e3cvp/alma991005219277706711

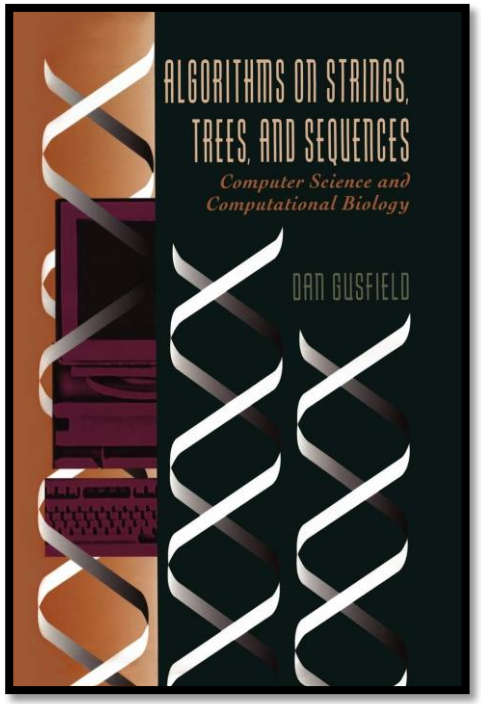
- **Complementary Online videos:**
https://www.youtube.com/playlist?list=PL2mpR0RYFQsBiCWVJSvVAO3OJ2t7DzoHA

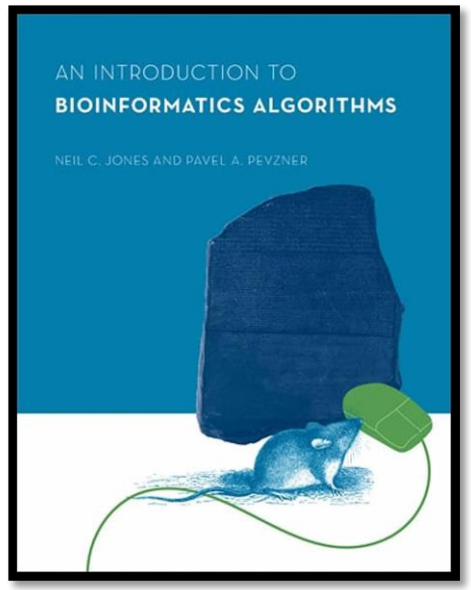


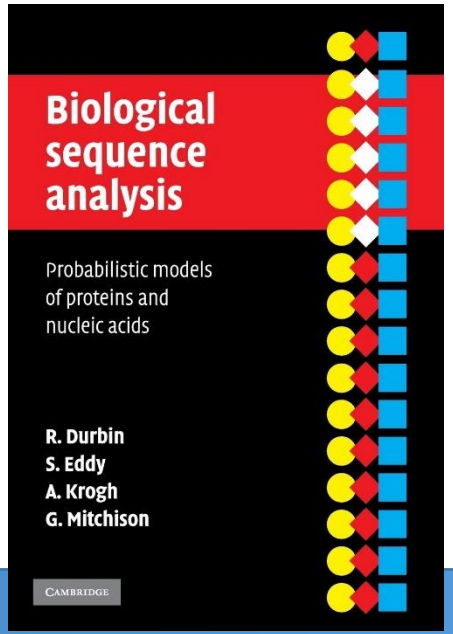
Complementary Resources

- **Problem sets on Rosalind:** http://rosalind.info/problems/locations/

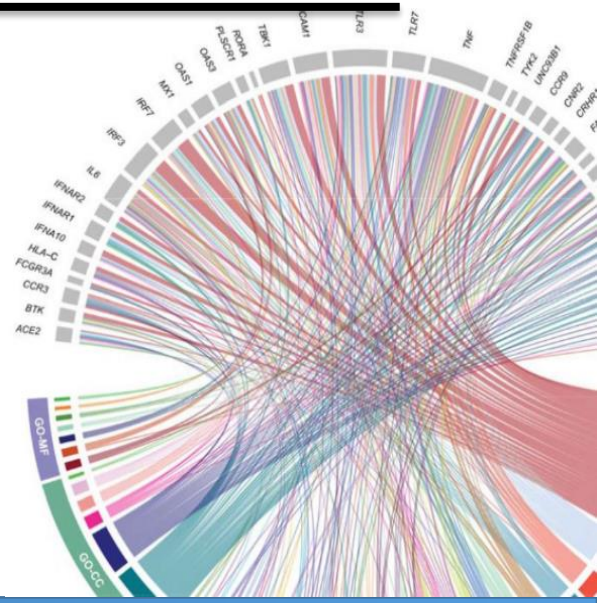
- **Algorithms on Strings, Trees, and Sequences**. Dan Gusfield. Cambridge University Press.
- **An Introduction to Bioinformatics Algorithms.** Neil C. Jones, Pavel A. Pevzner. MIT Press.
- **Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.** Richard M. Durbin, Sean Eddy. Cambridge University Press.







2 Motivation



Acknowledgements

Many pictures and materials are taken from **Ben Langmead's course**.

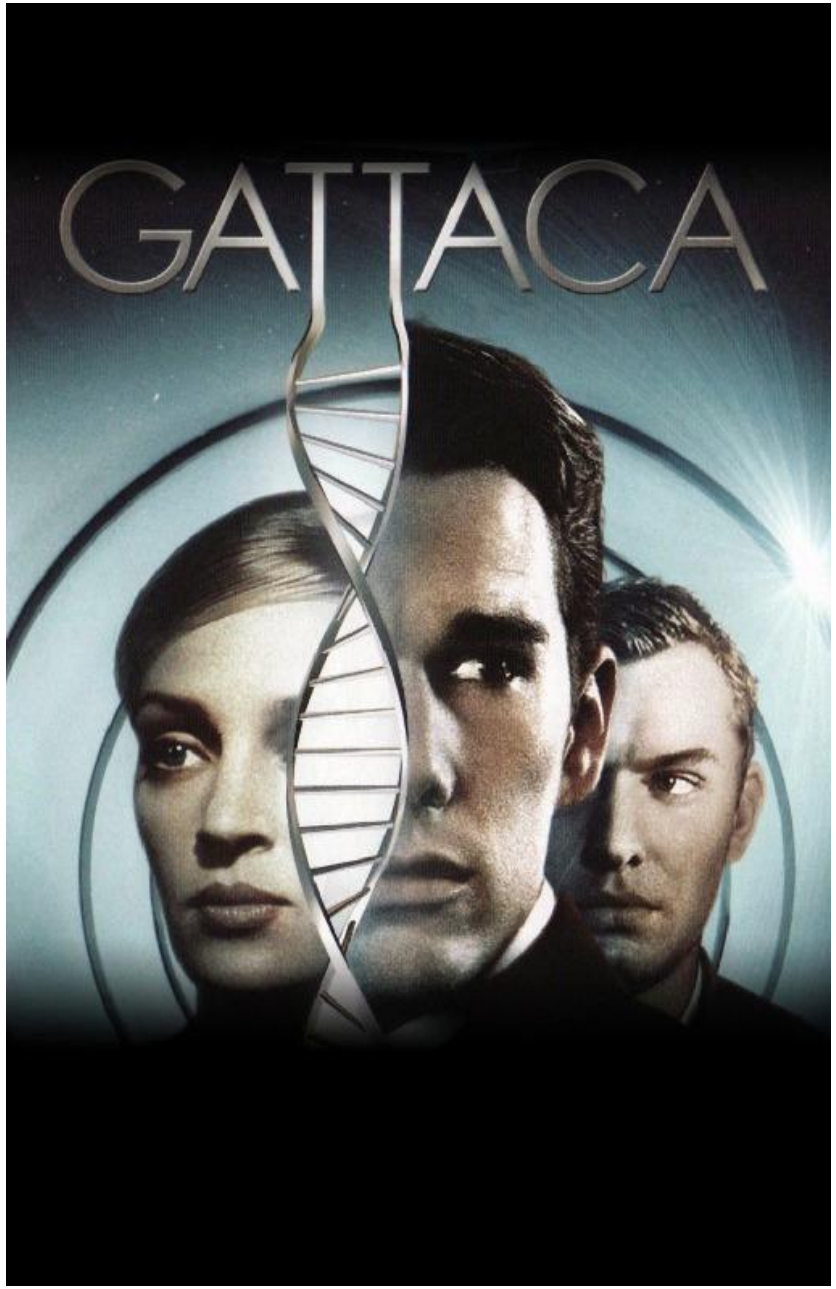
Course heavily inspired in:

- **Genome-Scale Algorithm Design**. Veli Mäkinen, Djamal Belazzougui, Fabio Cunial, Alexandru I. Tomescu. Cambridge University Press.
- **Algorithms on Strings, Trees, and Sequences**. Dan Gusfield. Cambridge University Press.
- **An Introduction to Bioinformatics Algorithms.** Neil C. Jones, Pavel A. Pevzner. MIT Press.

Genomics and Bioinformatics

- What do you know about genomes and genomics?
- Where did you hear about them?

1993



1997

READING THE BOOK OF LIFE: THE OVERVIEW

READING THE BOOK OF LIFE: THE OVERVIEW; Genetic Code of Human Life Is Cracked by Scientists

By NICHOLAS WADE
Published: June 27, 2000

Why Study DNA Sequencing and Computational Genomics?

- Intersection of **computer science and life sciences**.
- **DNA sequencing** has become cheap and efficient.
- Sequencing is widely used in life sciences and medicine.
- **Applications** of DNA Sequencing
  - Genetic diseases: Identifying rare conditions in children.
  - Ancient genomes: Understanding human origins and migration.
  - Cancer research: Analyzing tumor genomes for better treatments.
  - Microbiome studies: Exploring gut bacteria and their impact on health.
  - Basic genome research: Understanding genome function.

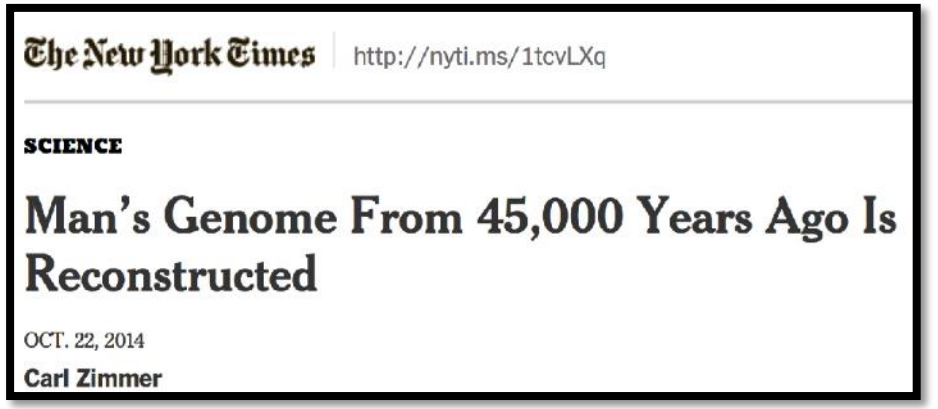

THE NEW YORKER

MEDICAL DISPATCH | JULY 21, 2014 ISSUE

ONE OF A KIND

What do you do if your child has a condition that is new to science?

BY SETH MNOOKIN


The New York Times http://nyti.ms/1tcvLXq

SCIENCE

Man's Genome From 45,000 Years Ago Is Reconstructed

OCT. 22, 2014

Carl Zimmer


The New York Times

June 27, 2013

Studying Tumors Differently, in Hopes of Outsmarting Them

By CARL ZIMMER

Sequencing Technologies (The Revolution)

GA II 1.6 billion nt/day (2008)

GA IIx 5 billion nt/day (2009)

HiSeq 2000 **75 billion** nt/day (2011)

HiSeq 2500 **120 billion** nt/day (2012)

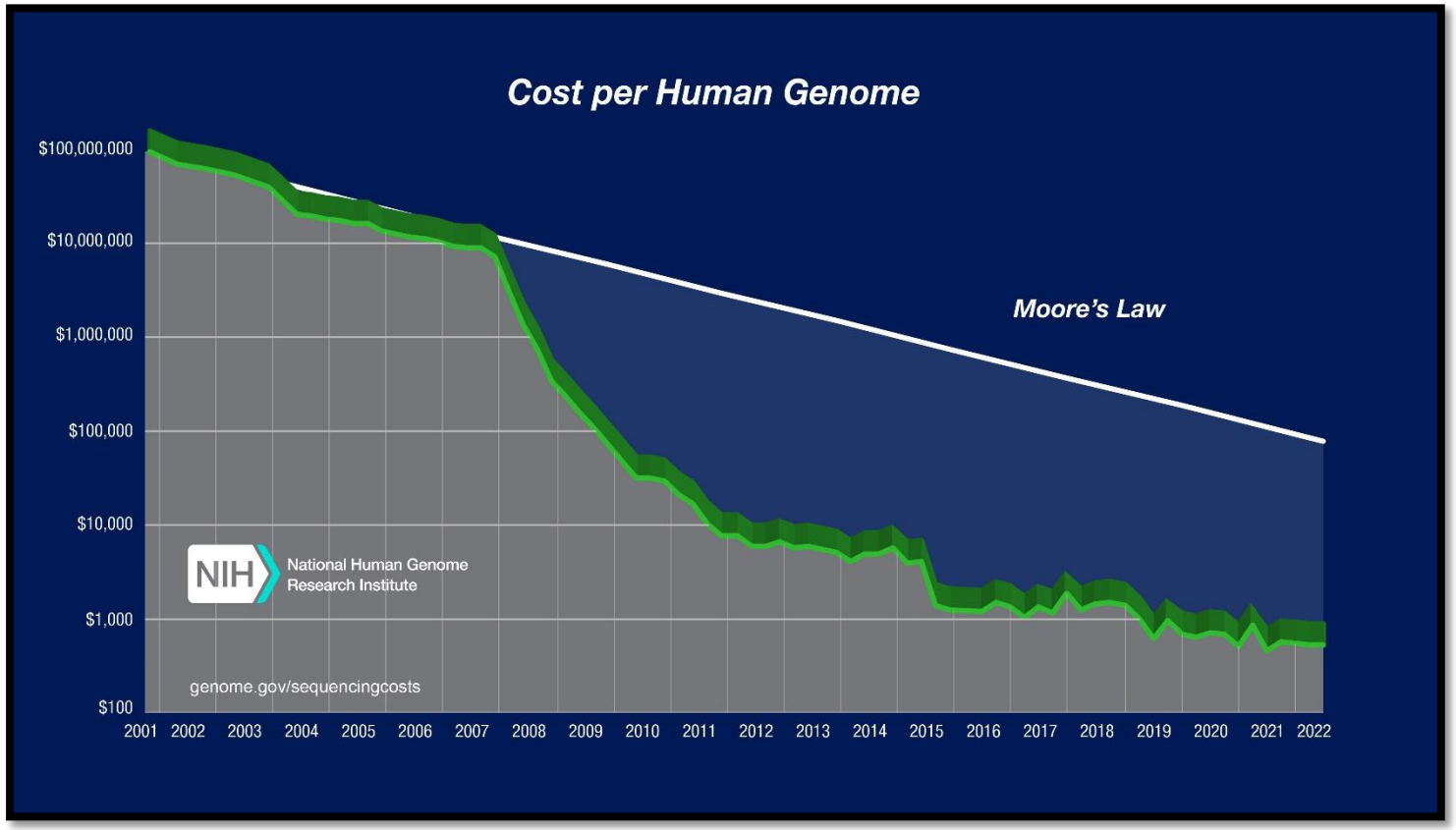
HiSeq 3000/4000 **200-400 billion** nt/day (2015)

NovaSeq 5000/6000 **1-3 trillion** nt/day (2017)

nt = nucleotide = **A**, **C**, **G**, or **T**

Sequencing Technologies (The Challenge)

- DNA Sequencing is Everywhere
  - Sequencing is as ubiquitous as computing.
  - Used in expected and unexpected ways.
  - Expanding beyond traditional biological research.



Computational Genomics and Bioinformatics Algorithms

- **Why Study Computational Genomics?**
  - Understanding algorithms helps determine their strengths and limitations.
  - Key example: De novo shotgun assembly problem in human genome sequencing.
  - Computational solutions enabled rapid genome assembly.
- **The Role of Algorithms in Genomics**
  - Algorithms define what's possible and what's practical.
  - Progress in computational genomics drives research and industry.
  - Genomics projects rely on computational experts and advanced algorithms.
- **A Growing Field with Opportunities**
  - Active research area in academia and industry.
  - Improving methods for analyzing large-scale sequencing data.
  - No large genomics project exists without computational tools.
  - Other Applications: Information retrieval, NLP, text analysis.

Questions