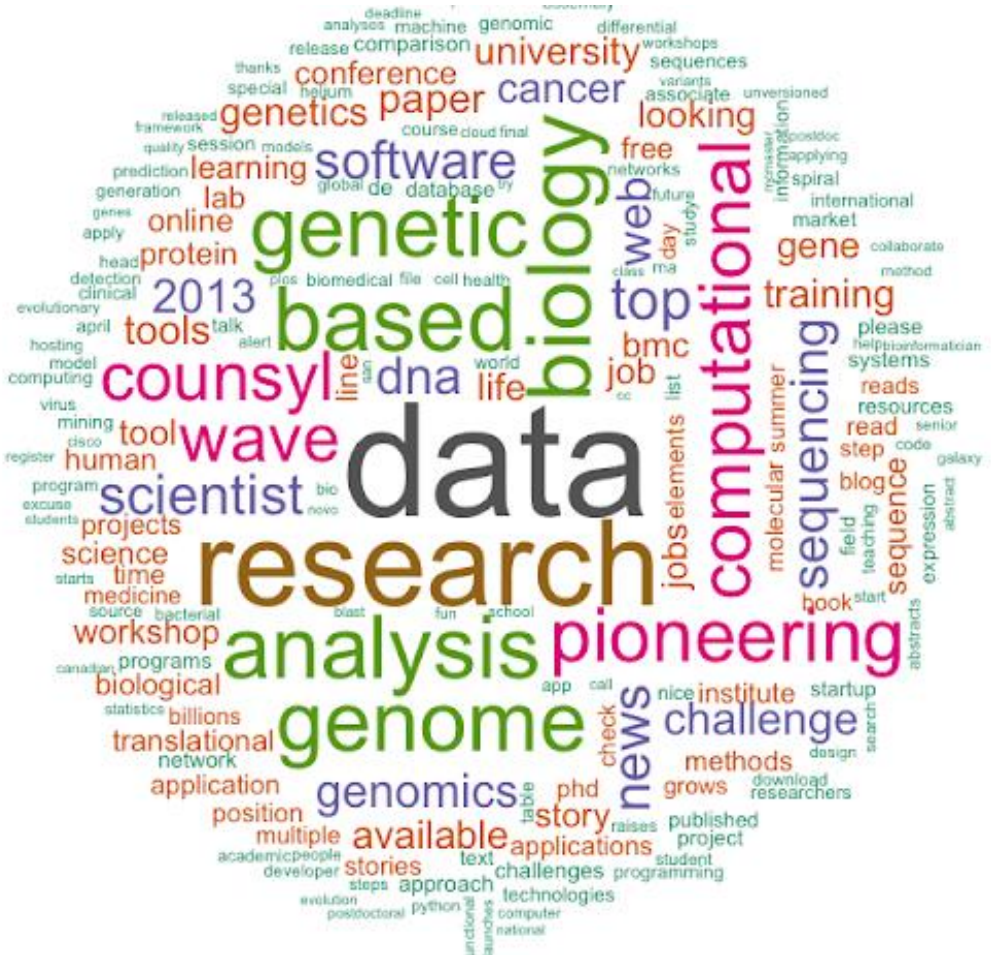
Introduction to Bioinformatics

Dr. Stanislav Kolenčík
stanislav.kolencik@famnit.upr.si



Main literature:

Pevsner, J. 2015. *Bioinformatics and Functional Genomics.* 3rd Edition. John Wiley & Sons , Ltd. 1124 str. . (www.bioinfbook.org)

Additional recommended literature:
https://www.biostarhandbook.com/

Last updated on **October 3, 2023**

Assessment:
Homeworks 30 %
Exam 70 %

The possibility to pass the exam with two midterms.

Responsibilities

- Attendance at the exercises is mandatory.
- Fulfilling of course activities.
- Passing the exam.

Syllabus

- History, Biological Databases and Access to Sequence Data
- Global/Local Alignment of Sequences
- Scoring Matrices for Alignment
- Multiple Sequence Alignment
- Tools for Searching Databases
- Base quality parameters
- Files with Biological Information & Tools
- Bioinformatics tools for operating with different file formats
- Sequencing Technologies

What will you learn?

- How the theoretical knowledge acquired during the lectures can be used?
- To use database and perform bioinformatic analyses.
- To interpret the results from those analyses.
- The importance of molecular tools for multiple research questions.

- Take Notes
- Ask Questions if you don't understand

The goal as a learning community is to create a safe environment that fosters open and honest dialogue.

Icebreakers

- Introduce yourself in 6 words
- Name tags
- Try to sit always next to someone else
- 3-mins talks at end of each class with person next to you –> what did you learn?

"In the past, having access to information meant power. Nowadays, you have to know which to filter"

-Alvin Toffler

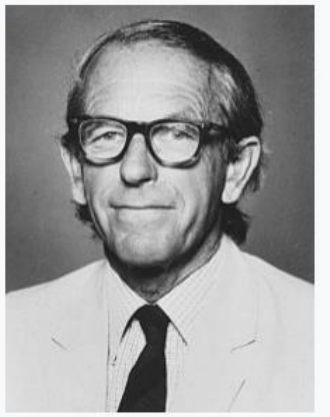
Chapter I: Introduction to Bioinformatics and Functional Genomics

Presentations use info from:

Jonathan Pevsner, Ph.D.
http://bioinfbook.org
pevsner@kennedykrieger.org
Bioinformatics and Functional Genomics
(3rd edition,

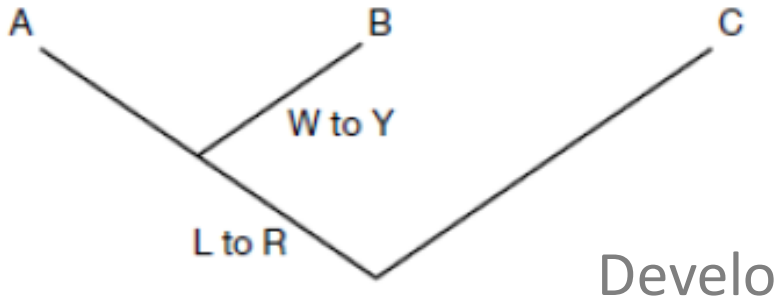
History and Development of Bioinformatics

- **1951:** F. Sanger and H. Tuppy -> First amino acid sequences (bovine insulin)
- **1964 - 1974:** Margaret Dayhoff et al. -> assembled databases of sequences into a protein sequence atlas -> the Protein Information Resource (PIR) (1984)
- **1978:** M. Dayhoff; point accepted mutation (PAM) matrix



Proteins were organized into families & superfamilies based on the degree of sequence similarity.

ORGANISM A	A	W	T	V	A	S	A	V	R	T	S	I
ORGANISM B	A	Y	T	V	A	A	A	V	R	T	S	I
ORGANISM C	A	W	T	V	A	A	A	V	L	T	S	I



Development of PAM matrix

History and Development of Bioinformatics

- **1979:** Walter Goad -> **assembled DNA sequence database**;

Los Almos National Laboratory (LANL), New Mexico.

  - Goad conceived **GenBank** prototype.
  - 1982-1992: LANL collected GenBank data;
  - GenBank now under the **NCBI.** http://www.ncbi.nlm.nih.gov
- **1980:** The **EMBL** Data Library was founded. http://www.ebi.ac.uk
- **1984:** DNA Data Bank of Japan – **DDBJ.** http://www.ddbj.nig.ac.jp

GenBank, EMBL, and DDBJ are part of the International Nucleotide Sequence Database Collaboration (**INSDC**) (http://www.ncbi.nlm.nih.gov/collab), which acts to facilitate exchange of data on a daily basis.

History and Development of Bioinformatics

Development of the first sequencing methods

- **1965:** First sequence of alanine tRNA
- **1970:** Ray Wu -> Primer extension approach, Ray Wu
- **1977:** Allan Maxam and Walter Gilbert -> DNA sequencing methods (chemical).
- **1977:** Frederic Sanger, chain termination method

Development of sequence search/comparison tools

- D. Benson, D. Lipman et al. -> GENINFO, NCBI
- ENTREZ -> web-based interface, NCBI

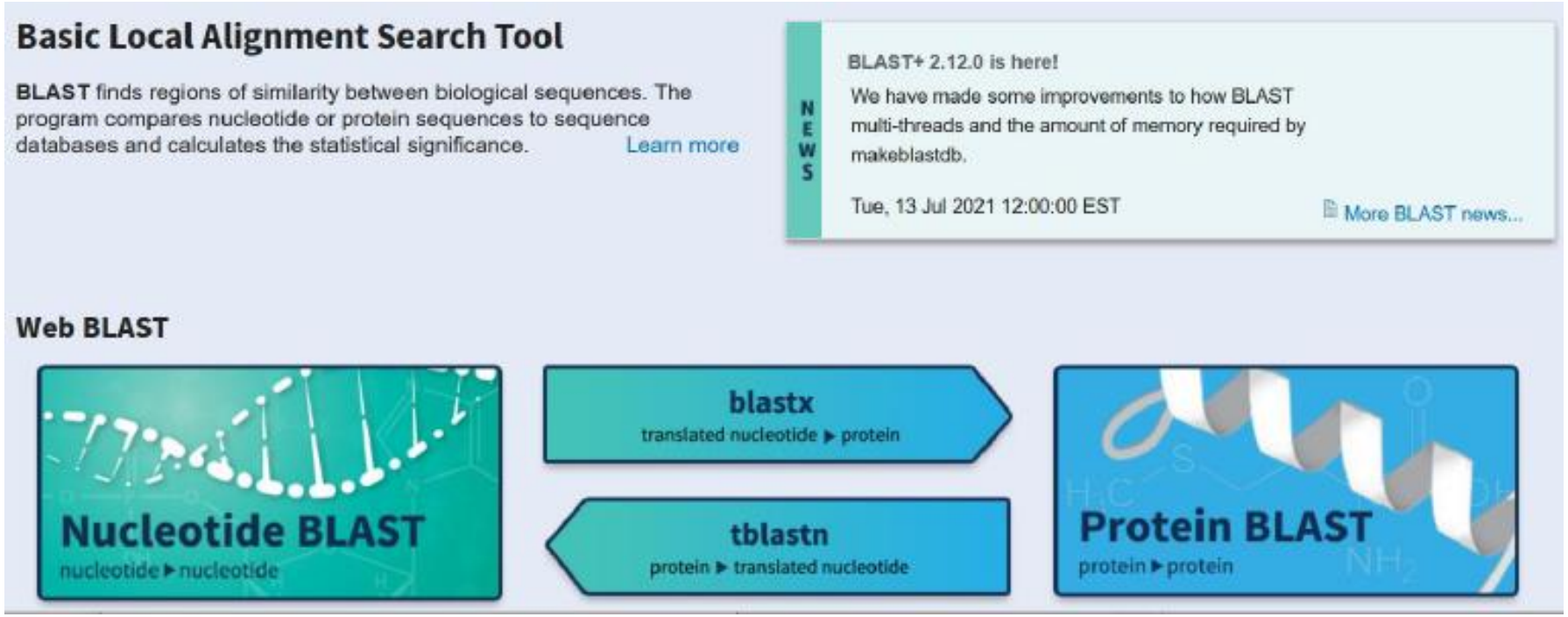
Sequence alignment tools

- Dot matrix or plot for sequence comparison, A. J. Gibbs and G. A. McIntyre (**1970**)

History and Development of Bioinformatics

Tools for searching similar sequences in databases:

- FASTA (Pearson and Lipman, **1988**)
- BLAST (Altschul et al., **1990**)



History and Development of Bioinformatics

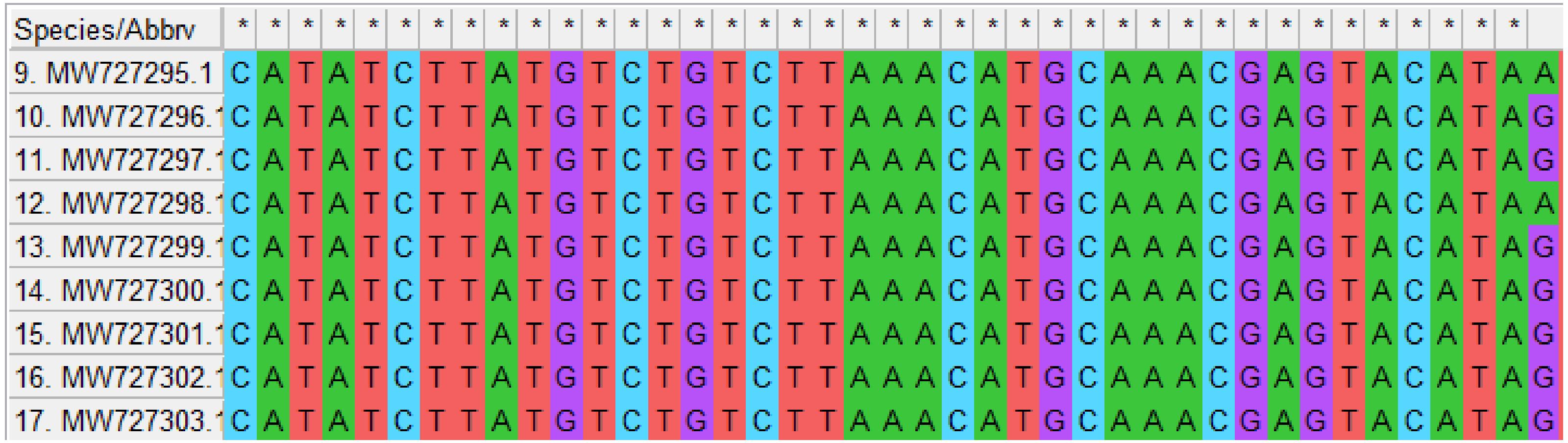
Sequence alignment using dynamic programming

- **1970:** A.J. Gibbs and G.A. McIntyre -> the dot matrix or diagram method for comparing sequences.
- **1970:** Needleman and Wunsch -> comparison of two sequences - determination of the most optimal alignment considering the entire sequence - global alignment
- **1981:** Smith and Waterman -> local alignment - determination of biologically relevant "subregions" of DNA or proteins

History and Development of Bioinformatics

Alignment of multiple sequences

- **1986**: Johnson and Doolittle

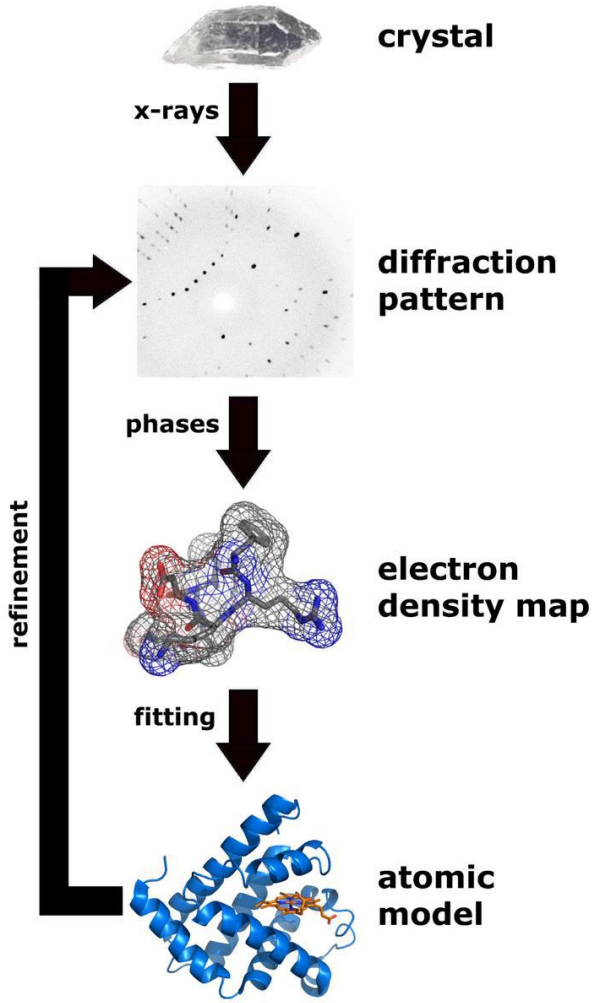


Species/Abbrv	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	
9. MW727295.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	A
10. MW727296.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G
11. MW727297.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G
12. MW727298.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	A
13. MW727299.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G
14. MW727300.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G
15. MW727301.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G
16. MW727302.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G
17. MW727303.1	C	A	T	A	T	C	T	T	A	T	G	T	C	T	G	T	C	T	T	A	A	A	C	A	T	G	C	A	A	A	C	G	A	G	T	A	C	A	T	A	G

Predicting Protein Secondary Structure

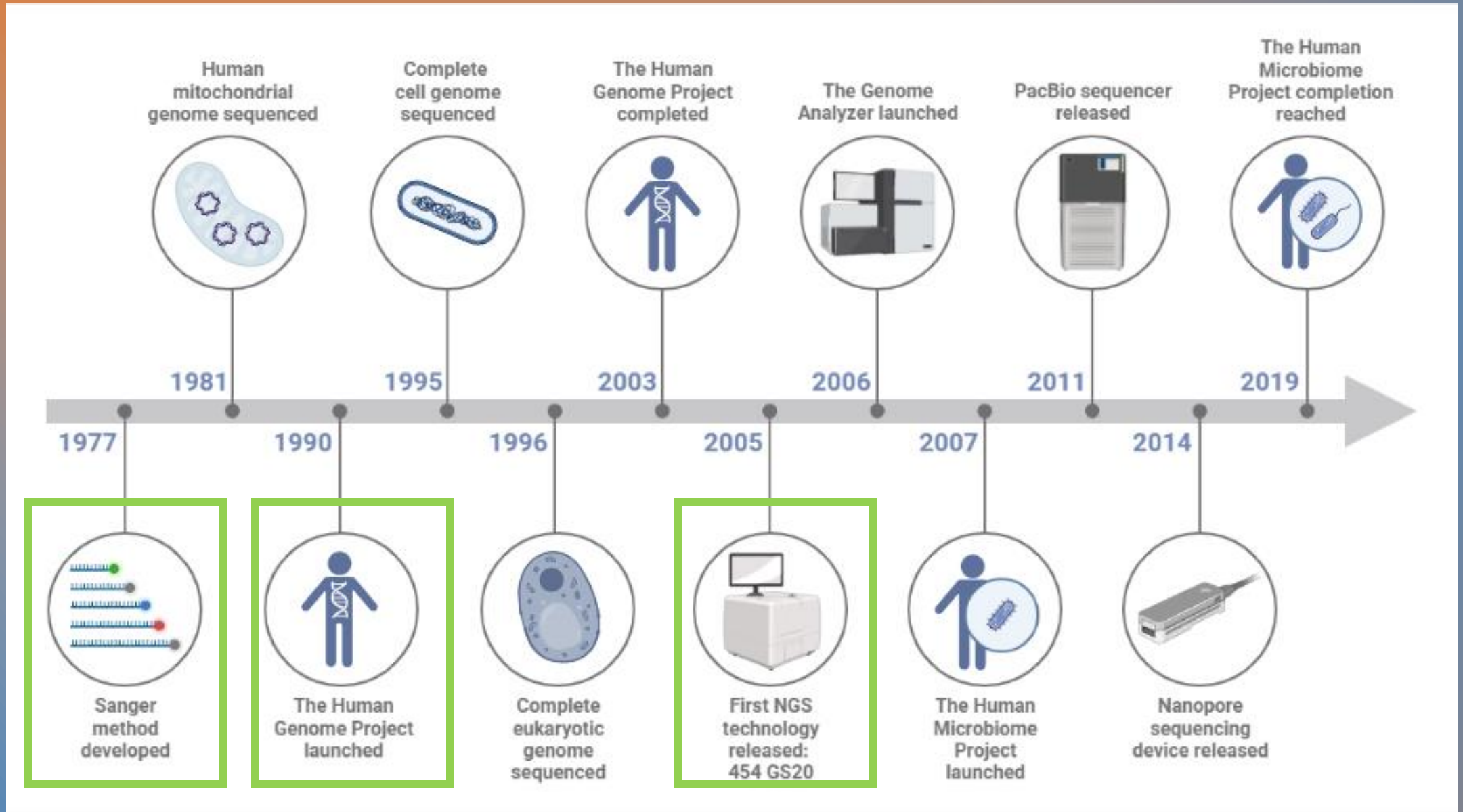
1976: Walter Fiers et al. -> the first published genome sequence of single cell organism - bacteriophage MS2 (shorter length ~3.6k bp).

1995: The **first COMPLETE GENOME SEQUENCE** of multicellular organism - circular chromosome of bacterium, *Hemophilus influenzae* (Fleischmann et al. 1995) - led by researcher Craig Venter (very long length – 1.8 million).



Workflow by Thomas Splettstoesser

History and Development of Bioinformatics

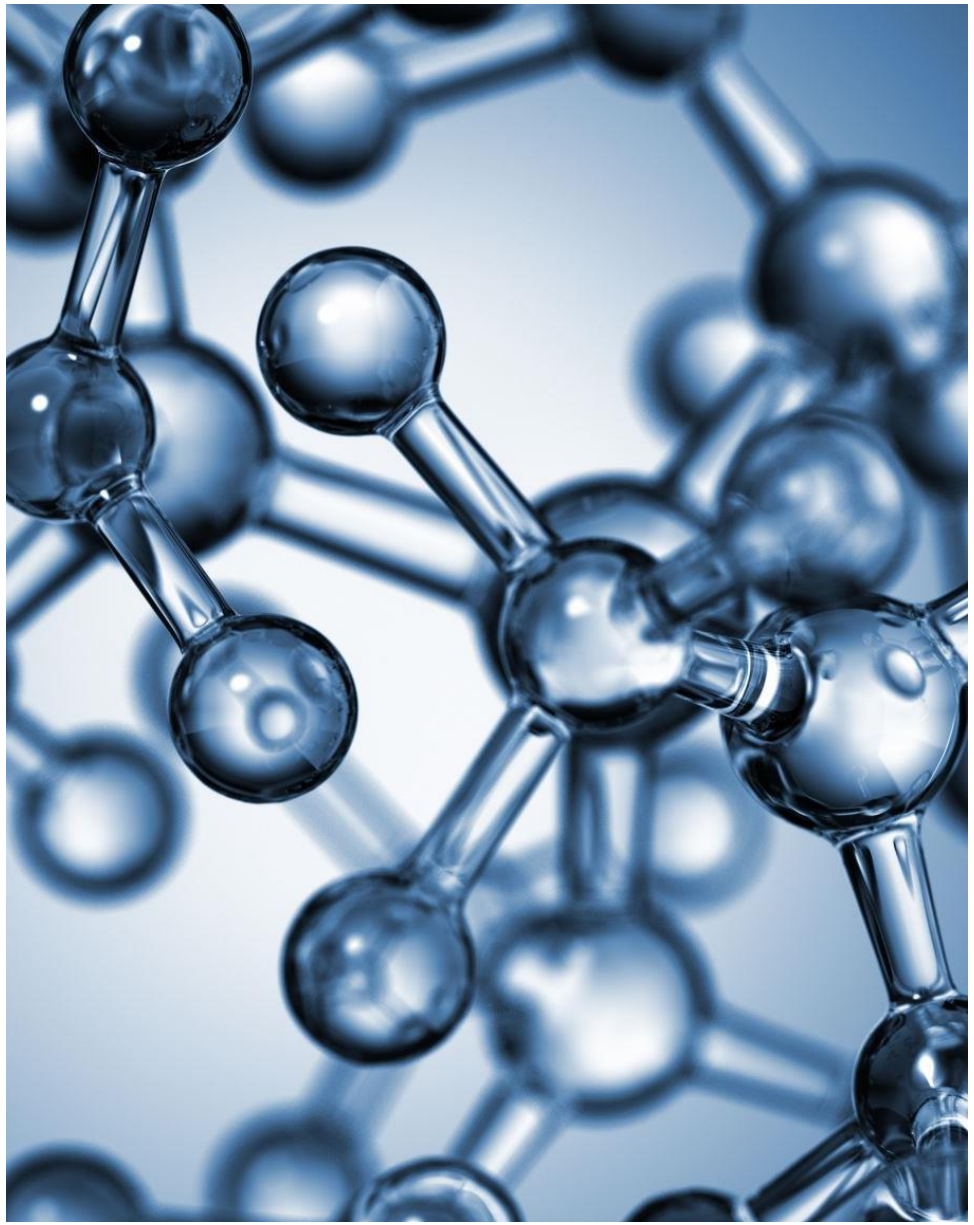


History and Development of Bioinformatics

- 2003: the human genome

Development of NGS technology

- 2000: Massively Parallel Signature Sequencing (MPSS) Lynx Therapeutics (USA) (later acquired by Illumina)
- 2004: Roche 454, the first NGS technology on the market
- 2005: Genome Analyzer, Solexa (2007 acquisition of Illumina)
- 2007: SOLiD technology, Applied Biosystems
- 2009: Helicos single molecule sequencing
- 2011: Ion Torrent, Life Technologies
- 2011: Pacific Biosciences SMRT technology
- 2012-2013: Oxford Nanopore Technologies ->  (portability) .....
- 2021: Element Biosciences .....->..... "delivering industry-leading accuracy at low run costs"



Definitions of bioinformatics and genomics

Genome = operating manual (all your DNA)

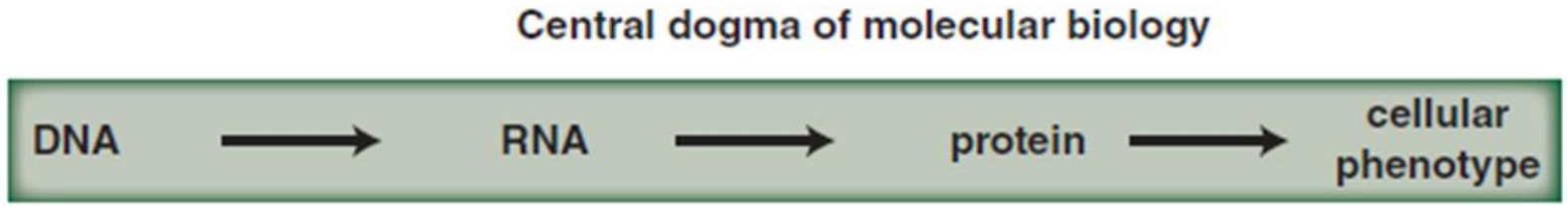
Genomics *-> study of genome*

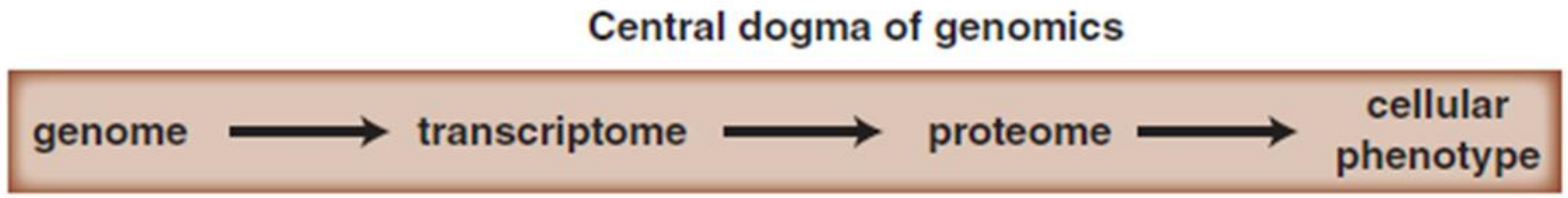
Bioinformatics, *as related to genetics and genomics, is a scientific subdiscipline that involves using computer technology to collect, store, analyze and disseminate biological data and information, such as DNA and amino acid sequences or annotations about those sequences.* [NIH NHGRI]

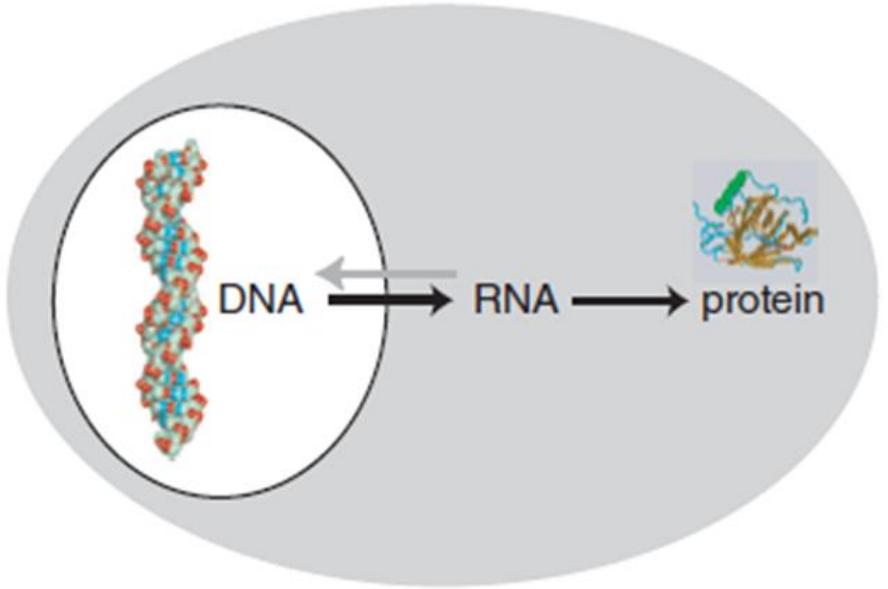
Task: Find online a definition of bioinformatics/proteomics/genomics and share it with the group.

Sequence Analysis <- **2000s** -> Next Gen. Sequencing

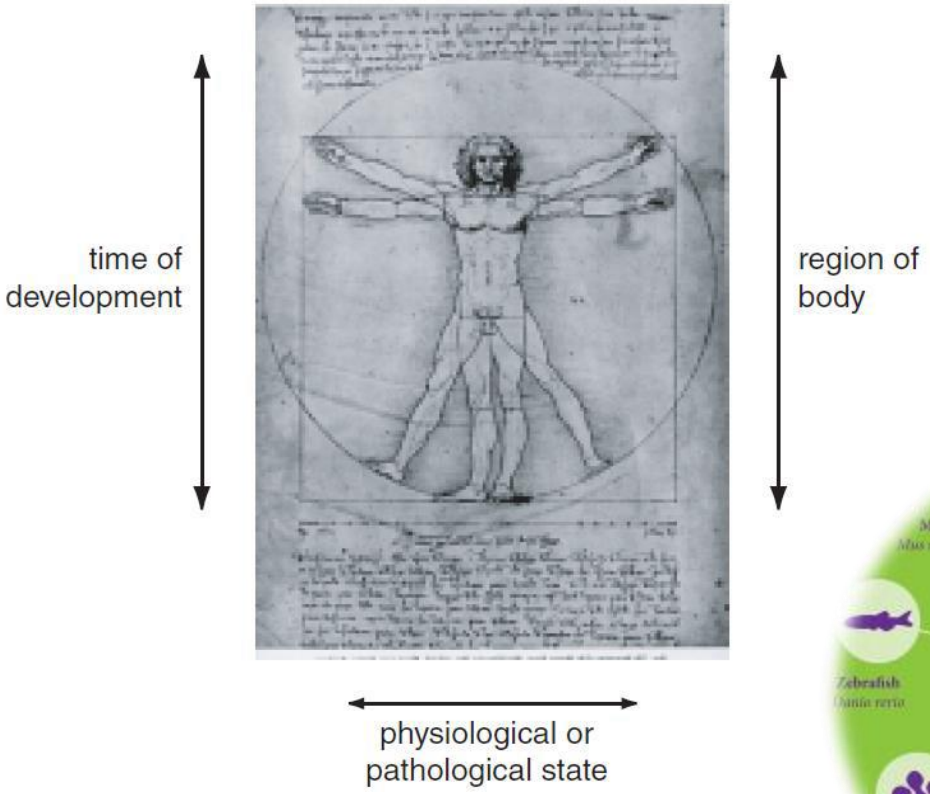
Central Dogma of Molecular Biology & Genomics

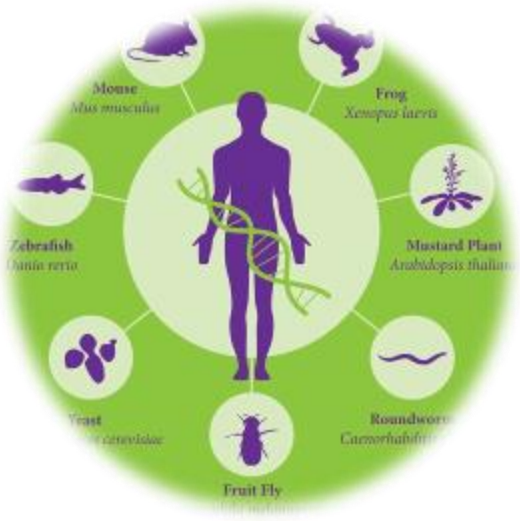


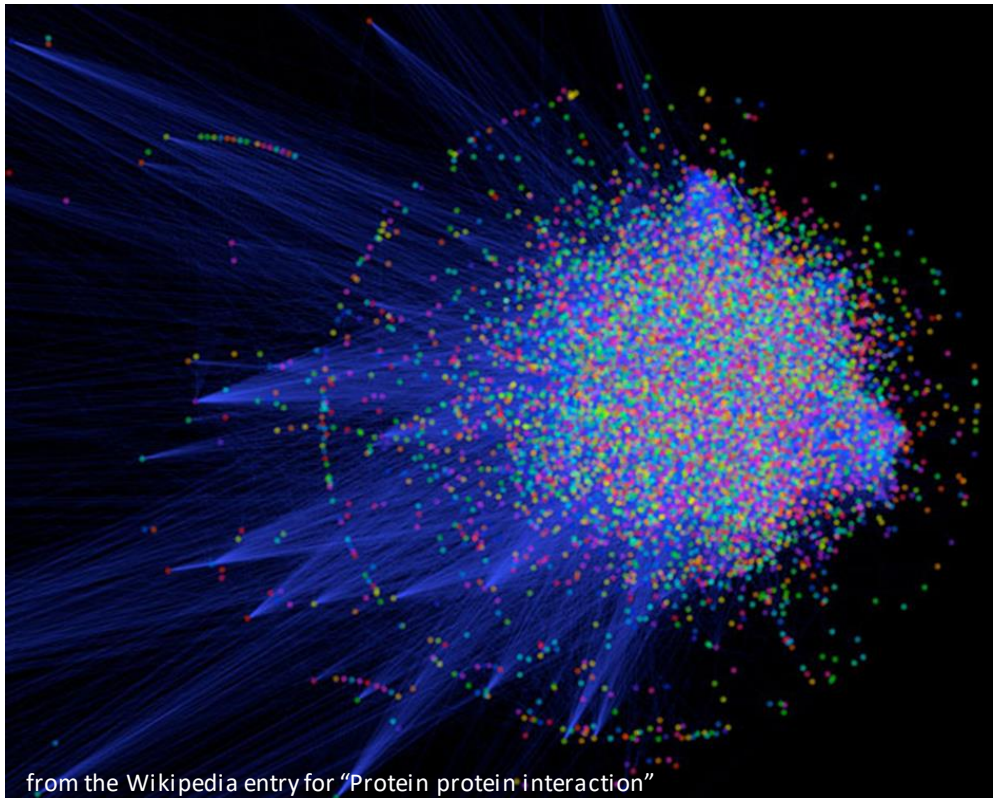




Analyses at the level of the whole genome, transcriptome, proteome



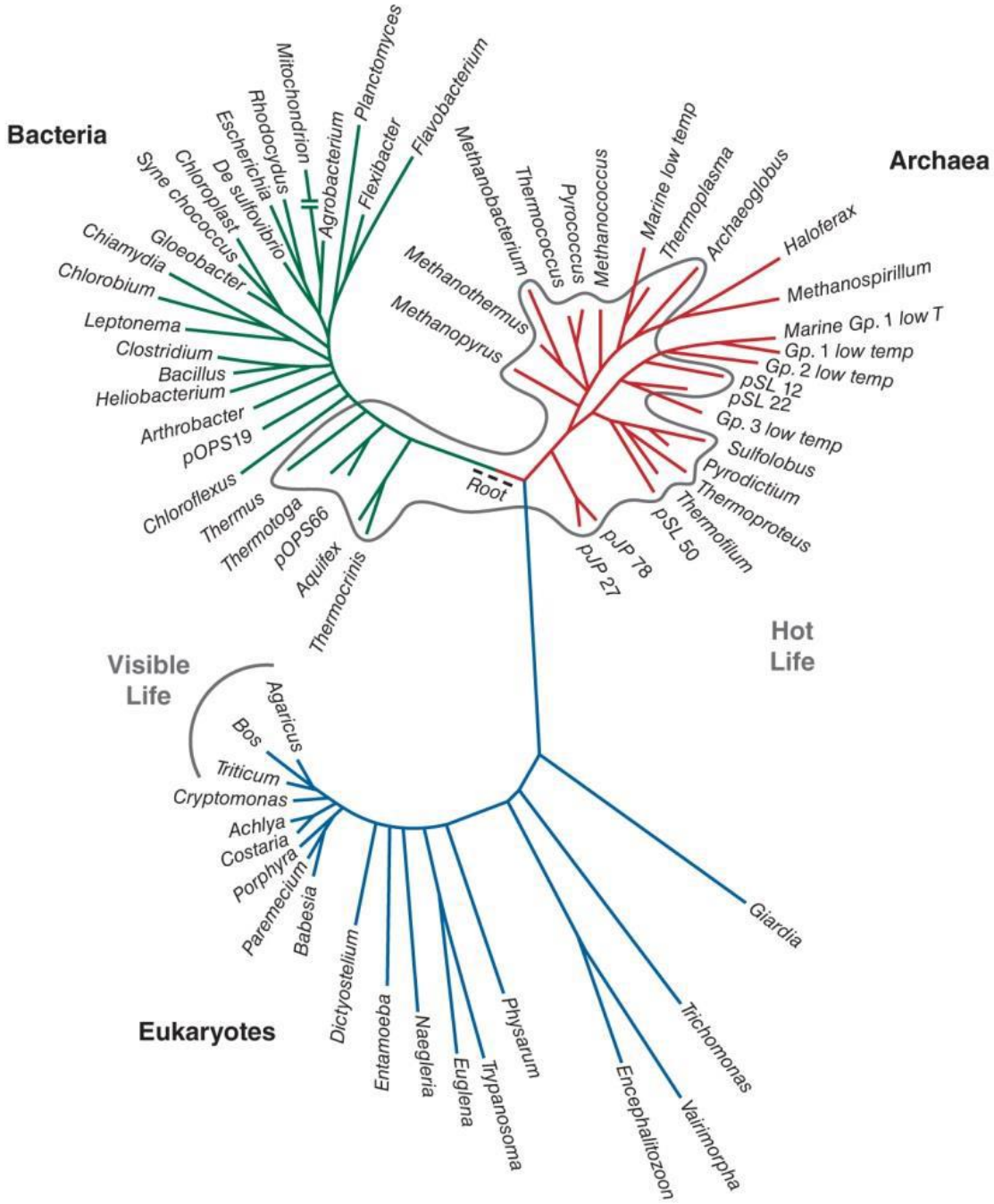




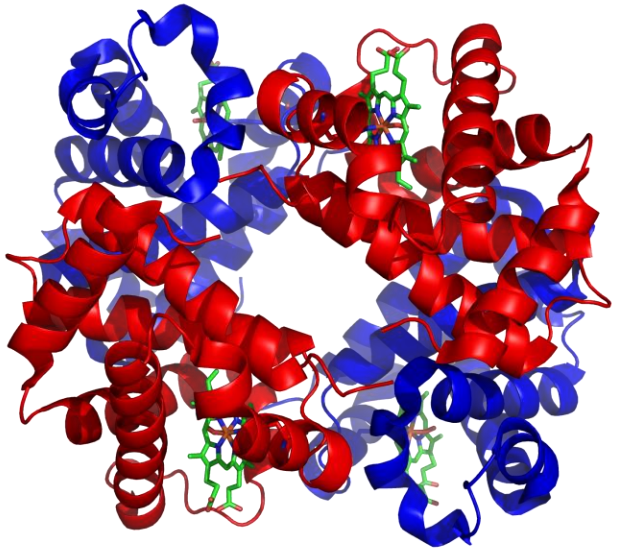
A map of human protein interaction

Three domains of life

2.13 million+++



A Consistent Example: Globins

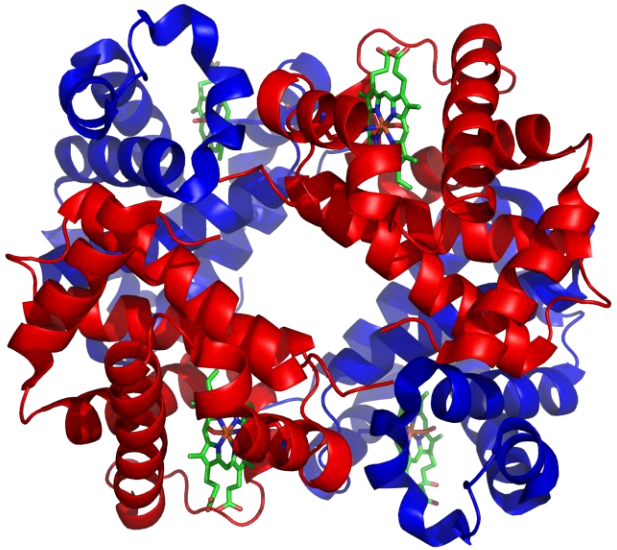


- **Hemoglobin** - one of the first proteins to be studied (1830s);
  - the main oxygen carrier in the blood of vertebrates;
  - structure was one of the earliest to be described.

- **Myoglobin** - a globin that binds oxygen in the muscle tissue;
  - the first protein to have its structure resolved by X-ray crystallography.

- **The comparison of myoglobin, alpha globin, and beta globin protein sequences --->>>** one of the earliest applications of multiple sequence alignment (development of amino acid substitution matrices used to score protein relatedness).

A Consistent Example: Globins



- The globin loci on human chromosomes 16 (for α globin) and 11 (for β globin) were among the first to be sequenced and analyzed.

- **The globin genes are exquisitely regulated across time** and with tissue-specific gene expression.

- **The family of homologous proteins extends to separate classes** of plant globins, invertebrate hemoglobins , bacterial homodimeric hemoglobins, and flavohemoglobins (bacteria, archaea, and fungi).

The globin family is therefore useful as we survey the tree of life.

Bioinformatics and genomics: two cultures

Web-based or
graphical user interface (GUI)

Central resources (NCBI, EBI,)

Genome browsers (UCSC, Ensembl)

GUI software (Partek, MEGA, RStudio, BioMart, IGV)

Galaxy (web access to NGS tools, browser data)

Command line (often Linux)

Biopython, Python, BioPerl, R: manipulate data files

Data analysis software: sequences, proteins, genomes

Next generation sequencing tools

Online Resources

Massive Open Online Courses (MOOCs)

https://software-carpentry.org/

https://www.datacamp.com/

http://rosalind.info/problems/locations/

https://www.udemy.com/

https://www.coursera.org/courses?query=computational%20biology

https://www.coursera.org/specializations/bioinformatics

https://www.edx.org/learn/bioinformatics

https://github.com/mdozmorov/Python_notes

https://r4ds.had.co.nz/

Biostars - an online forum in which you can post questions, get answers from the community, explore tutorials, and more...

Try joining Biostars or other bioinformatics forums to find others who have questions similar to yours.

Modern tools -> AI-powered language models

Common Rules

- The analysis process should be well documented (e.g., a text document with command line commands and additional comments).
- Good organization is important when storing data and other files related to the analysis.
- Provide public access to raw data (GEO, SRA, ENA,...) in addition to scientific publication.
- Data about data (**metadata**) - as important as data!
- Documented versions of used databases and tools -> **Methods**

Research should be conducted in a way that is reproducible.