

# Computational biology

## Sequence-structure-function paradigm

Clovis Galiez



Grenoble  
Statistiques pour les sciences du Vivant et de l'Homme

October 5, 2021

# Goal

- Get an overview of computational biology topics
  - Topics (genomics, metagenomics, proteomics, etc.)
  - Know basic elements in biology (gene to function)
  - Know some important databases
  - Know standard tools (Blast) and libraries (BioPython)
- Have a basic culture of order of magnitude in computational biology
  - Quantity of data
  - Size of genomes
  - Size of organisms
- Toward autonomy for design and implementation of methods
  - Case study of SNP detection

# Lecture organization

- Session I: some background, starting your project
- Session II hands-on: development, simulation
- Session III hands-on: application: database mining, sequence searching
- Bonus: application to SARS-CoV-2

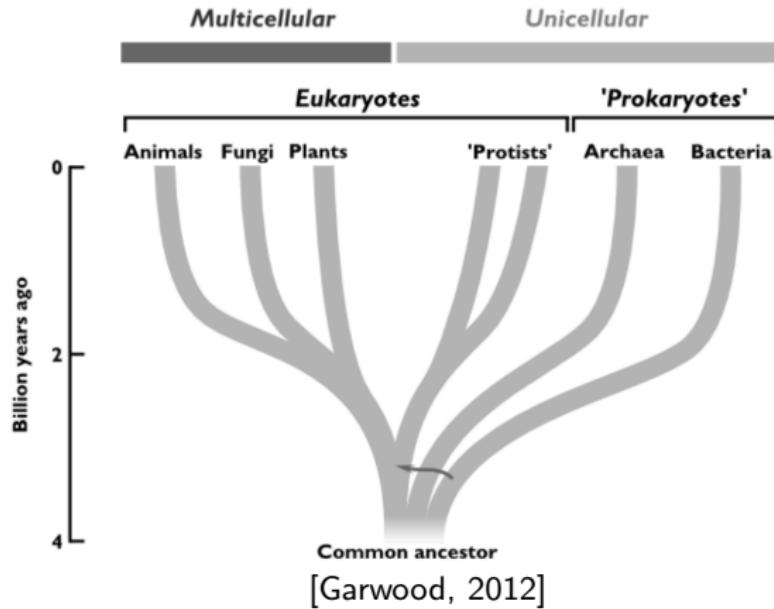
Evaluation: project-based + bonus for participation

# Today's outline

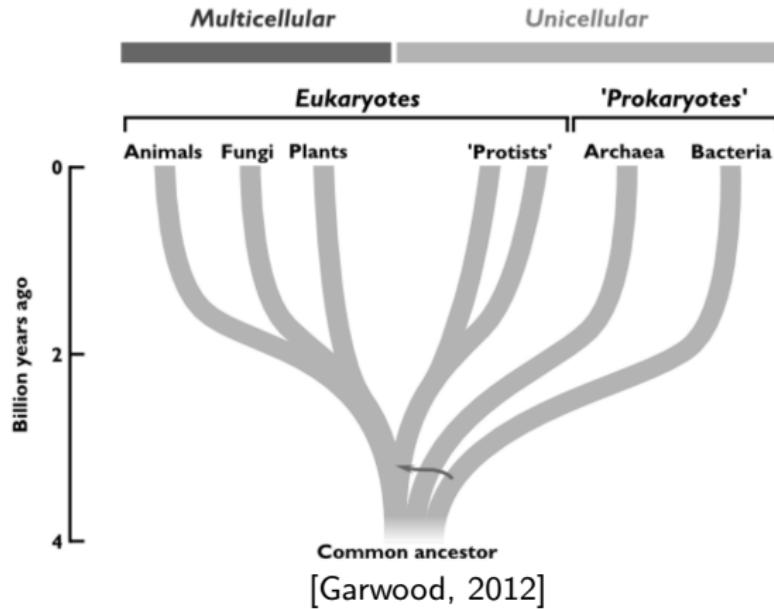
- Elements of biology
  - DNA and organisms
- Why computational biology?
  - Context: a world of data
  - Applications
- From genomics to biology
  - Sequence-structure-function paradigm

# Elements of biology

# Tree of life

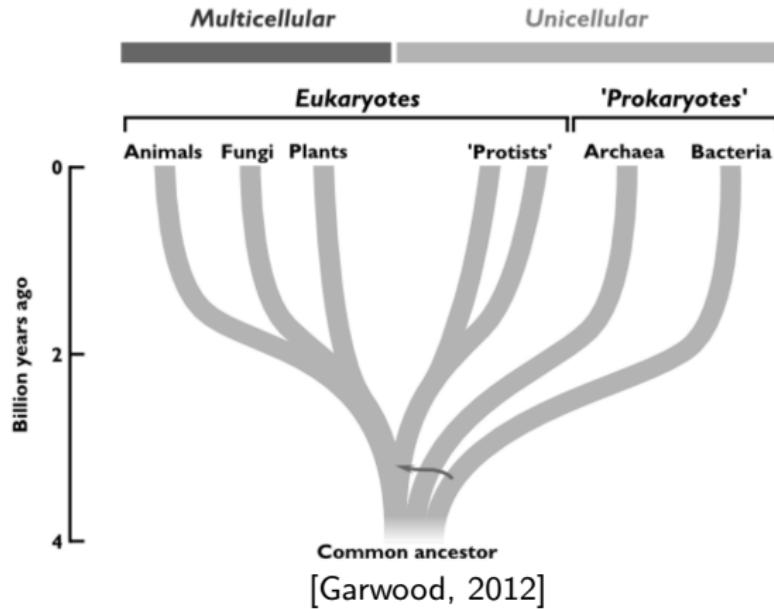


# Tree of life



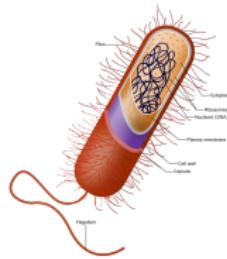
When was the split between *Homo* and apes?

# Tree of life

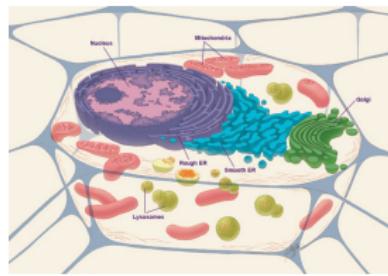


When was the split between *Homo* and apes?  $\approx 3$ M y. ago.

# Prokaryotes and eukaryotes

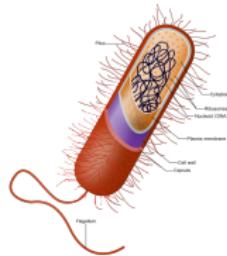


Prokaryotes  
"Simple", no nucleus

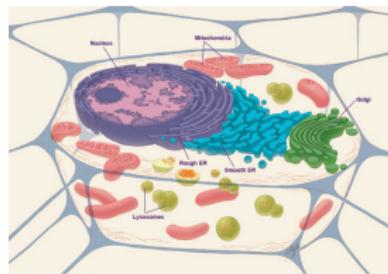


Eukaryotes  
**Advanced**, nucleus

# Prokaryotes and eukaryotes



Prokaryotes  
"Simple", no nucleus



Eukaryotes  
**Advanced**, nucleus

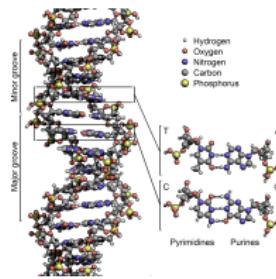
What are the main differences? What unifies those organisms?

# DNA: a universal way of coding (rather recent knowledge!)

## Universal code

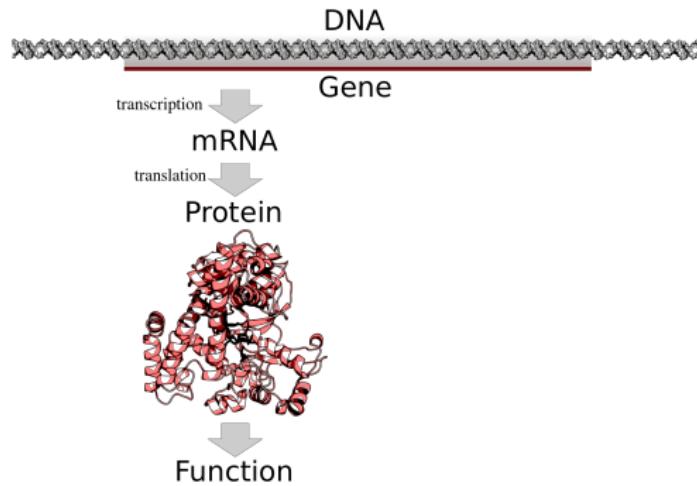
All known living organisms are coded through their DNA information. This determines to a large extent their morphologies and functions.

- 1952 Hershey and Chase: DNA is known to encode genetic information
- 1953 Physical structure (double-helix) of DNA is solved using X-Ray diffraction by Franklin (but that's Watson & Crick who got the awards)



# How DNA determines an organism?

The big picture (for computer scientists): see video.



Proteins are responsible for most of the biological functions in organisms (biochemical reactions (enzymes), nutrient transportation, structural proteins, etc.)

# Why computational biology?

# Two reasons triggering computational biology

## Computational biology

Data coupled to statistical models, machine learning, data visualization.

# Two reasons triggering computational biology

## Computational biology

Data coupled to statistical models, machine learning, data visualization.

- Availability of data
- Computing capacities

# Biology from the data perspective: not only DNA!

Biology brings various types of data, to get insights on various questions:

- Sequences **ATTCAGTACAT**

- Genomic (DNA sequence of one organism)
- Metagenomic (DNA sequence of a biological sample - many organisms)
- Proteomic (amino-acid sequences)

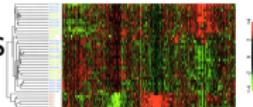
# Biology from the data perspective: not only DNA!

Biology brings various types of data, to get insights on various questions:

- Sequences **ATTCAGTACAT**

- Genomic (DNA sequence of one organism)
- Metagenomic (DNA sequence of a biological sample - many organisms)
- Proteomic (amino-acid sequences)

- Abundances



- Marker gene/species abundance
- Expression level of genes

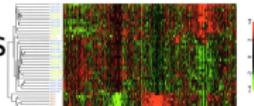
# Biology from the data perspective: not only DNA!

Biology brings various types of data, to get insights on various questions:

- Sequences **ATTCAGTACAT**

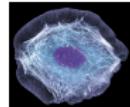
- Genomic (DNA sequence of one organism)
- Metagenomic (DNA sequence of a biological sample - many organisms)
- Proteomic (amino-acid sequences)

- Abundances



- Marker gene/species abundance
- Expression level of genes

- Images



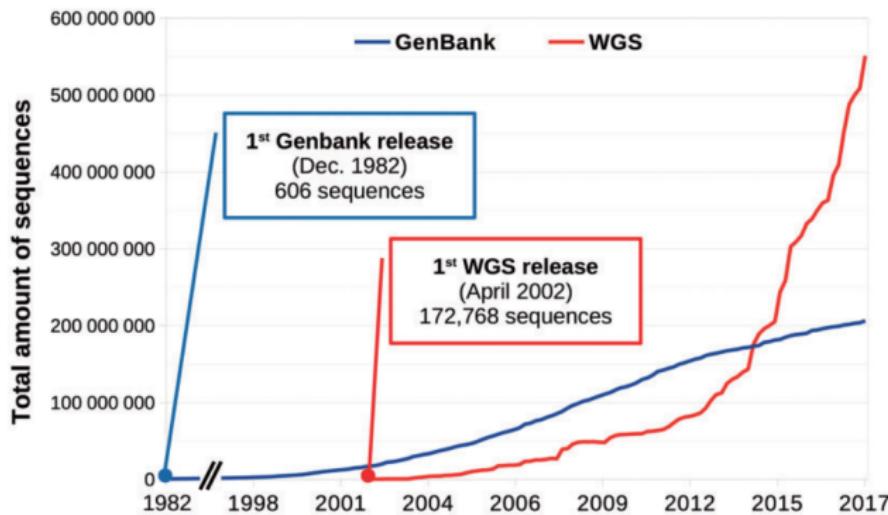
- Neuroimaging
- Cell imaging

- Mass spectrometry

- ...

## Example: genomics, the biggest breakthrough

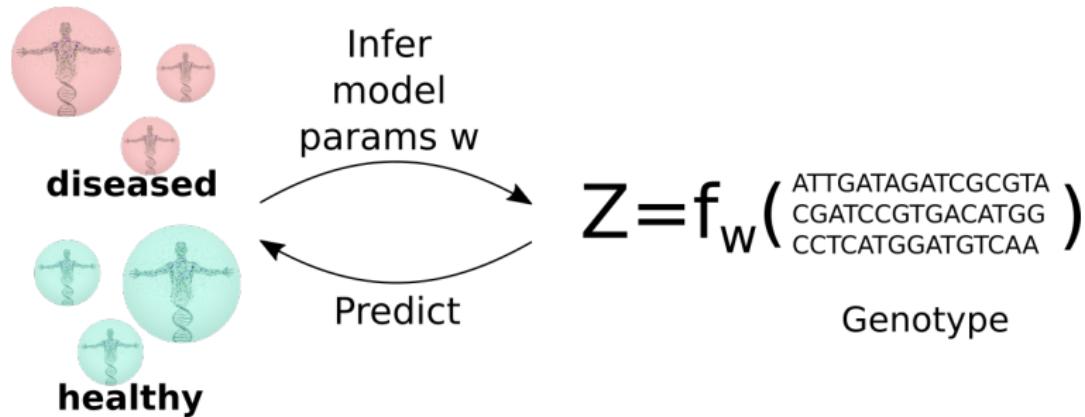
1977: first DNA sequencer.



You can now sequence a human cell for less than a thousand euros.

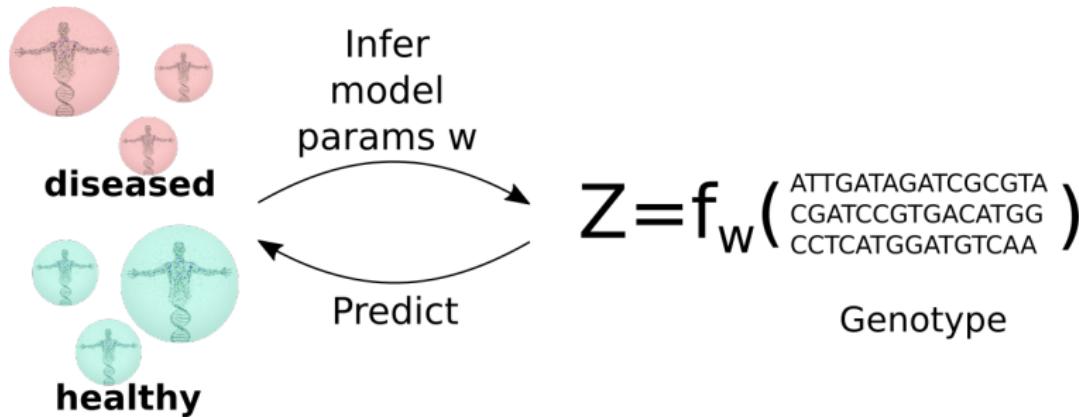
# Statistical association studies

Relates the variation of the genome to the phenotype.



# Statistical association studies

Relates the variation of the genome to the phenotype.



Define a predictor  $f : \mathbb{R}_+^M \rightarrow [0, 1]$  such that it minimizes a *loss* on a training set  $(\vec{x}_1, z_1), \dots, (\vec{x}_N, z_N)$ :

$$\min_f - \sum_{i=1}^N z_i \cdot \log f(\vec{x}_i) + (1 - z_i) \cdot \log(1 - f(\vec{x}_i))$$

# Lots of data is not just for fun!

Human genome: what size? 1 Kbp? 10 Mbp?

# Lots of data is not just for fun!

Human genome: 3Gbp.

# Lots of data is not just for fun!

Human genome: 3Gbp.

With  $f : \mathbb{R}_+^M \rightarrow [0, 1]$ , the theoretical number of input possibilities is:

# Lots of data is not just for fun!

Human genome: 3Gbp.

With  $f : \mathbb{R}_+^M \rightarrow [0, 1]$ , the theoretical number of input possibilities is:

$$4^{3 \cdot 10^9} =$$

# Lots of data is not just for fun!

Human genome: 3Gbp.

With  $f : \mathbb{R}_+^M \rightarrow [0, 1]$ , the theoretical number of input possibilities is:

$$4^{3 \cdot 10^9} = 10^{1806179974} \text{ possibilities :-/}$$

## Lots of data is not just for fun!

Human genome: 3Gbp.

With  $f : \mathbb{R}_+^M \rightarrow [0, 1]$ , the theoretical number of input possibilities is:

$$4^{3 \cdot 10^9} = 10^{1806179974} \text{ possibilities :-/}$$

In practice, one “reference genome” and “only”  $\approx 88 \cdot 10^6$  possible mutation places [The 1000 Genomes Project Consortium, 2015] .

Which mutation is responsible for a specific disease?

# Lots of data is not just for fun!

Human genome: 3Gbp.

With  $f : \mathbb{R}_+^M \rightarrow [0, 1]$ , the theoretical number of input possibilities is:

$$4^{3 \cdot 10^9} = 10^{1806179974} \text{ possibilities :-/}$$

In practice, one “reference genome” and “only”  $\approx 88 \cdot 10^6$  possible mutation places [The 1000 Genomes Project Consortium, 2015] .

Which mutation is responsible for a specific disease?

Needs a **lot of** data... and fine statistics.

/

## Scope of applications for DNA sequence data

Computationally processing DNA sequences has a huge number of applications, in particular in the fields of:

# Scope of applications for DNA sequence data

Computationally processing DNA sequences has a huge number of applications, in particular in the fields of:

-  clinics

# Scope of applications for DNA sequence data

Computationally processing DNA sequences has a huge number of applications, in particular in the fields of:

-  clinics
-  ecology

# Scope of applications for DNA sequence data

Computationally processing DNA sequences has a huge number of applications, in particular in the fields of:

-  clinics
-  ecology
-  biochemistry

# Scope of applications for DNA sequence data

Computationally processing DNA sequences has a huge number of applications, in particular in the fields of:

-  clinics
-  ecology
-  biochemistry

...and sometimes involving the three at the same time!

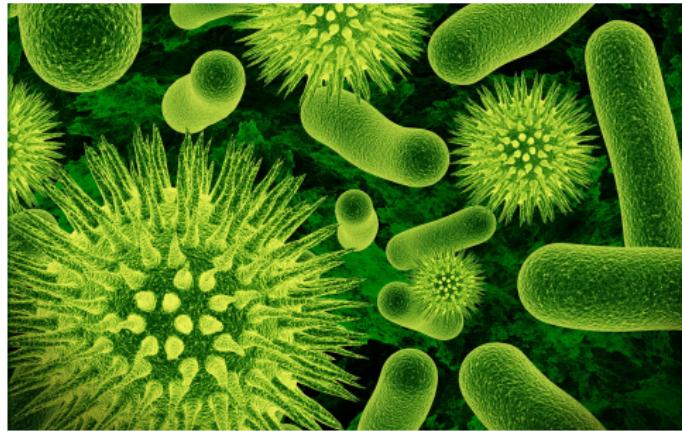
# Focus on the microbial world

# The microbial world

They are everywhere... they work hard 24h a day... they fight against each other... and they collaborate.

# The microbial world

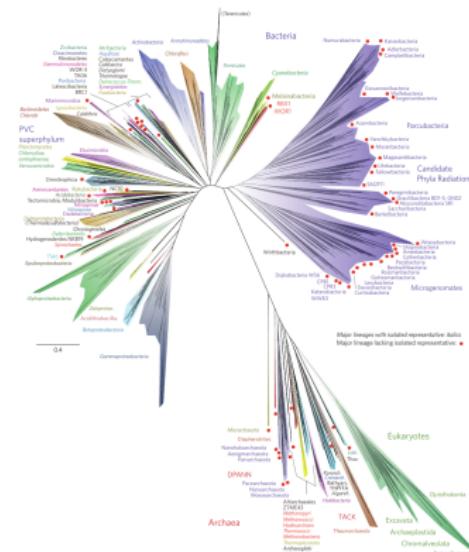
They are everywhere... they work hard 24h a day... they fight against each other... and they collaborate.



There are very diverse in terms of morphology, mechanisms, and genetics: bacteria, fungus, viruses, picoeukaryotes, etc.

# Origins and evolution of micro-organisms

Not a fixed knowledge: **we still continue to discover new branches of life:**



[Hug et al. 2016]

The Candidate Phyla Radiation (top right, in purple) has been discovered in 2016!

# Microbiome importance in biogeochemical cycles



Nitrogen cycle [Canfield et al., Science 2010]

CO<sub>2</sub> turnover: viruses kill 20% of the living biomass in the ocean every day! [Suttle, Nat. Microbiol. 2007]

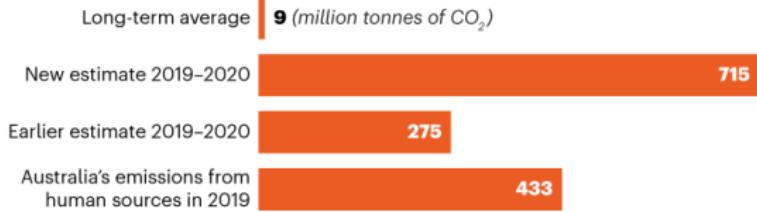


# A recent example: Wildfires in Australia



## RECORD EMISSIONS

Devastating fires in southeastern Australia in the summer of 2019–2020 released almost 80 times as much carbon dioxide into the atmosphere as a typical summer bush-fire season.



©nature

# A recent example: Wildfires in Australia

## RECORD EMISSIONS

Devastating fires in southeastern Australia in the summer of 2019–2020 released almost 80 times as much carbon dioxide into the atmosphere as a typical summer bush-fire season.



Long-term average | 9 (million tonnes of CO<sub>2</sub>)

New estimate 2019–2020 | 715

Earlier estimate 2019–2020 | 275

Australia's emissions from  
human sources in 2019 | 433

©nature

95% of emitted CO<sub>2</sub> has been pumped down by planktonic bloom.

[Nature 597, 459–460 (2021)]

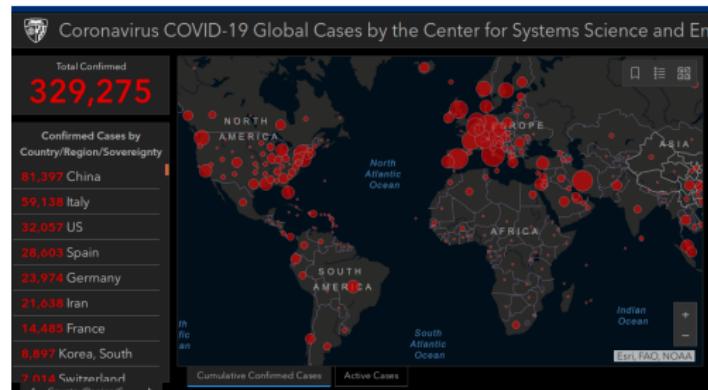
# Microbiome importance in human health

The bright side:



Health status highly correlated with the diversity of the gut microbiome [Valdes et al. 2018]

The dark side:

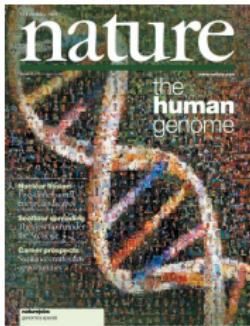


Covid-19

# The human gut microbiome

2000's

Human genome



≈ 20k protein-coding genes

2010's

Gut metagenomes



# The human gut microbiome

2000's

Human genome



≈ 20k protein-coding genes

$\times 100$

2010's

Gut metagenomes



≈ 2M protein-coding genes

Human gut microbiome is rich! What microbes do there is absolutely necessary to keep alive!

# Gut microbiota and higher order diseases

- **Autism**  
spectrum disorder (ASD), but the underlying mechanisms are unknown. Many studies have shown alterations in the composition of the fecal flora and metabolic products of the gut microbiome in patients with ASD. The gut microbiota influences brain development and behaviors through the neuroendocrine, neuroimmune and autonomic nervous systems. In addition, an abnormal gut microbiota is associated with several diseases, [Li et al. *Front. in Cell. Neur.* 2017]
- Type II diabetes (50 microbial genes → AUC ROC 0.81)  
[Qin et al. *Nature* 2012]
- Parkinson's differential abundance of gut microbial species  
[Heintz-Buschart et al. *Mov. Disord.* 2018]

# Gut microbiota and higher order diseases

- **Autism**

spectrum disorder (ASD), but the underlying mechanisms are unknown. Many studies have shown alterations in the composition of the fecal flora and metabolic products of the gut microbiome in patients with ASD. The gut microbiota influences brain development and behaviors through the neuroendocrine, neuroimmune and autonomic nervous systems. In addition, an abnormal gut microbiota is associated with several diseases, [Li et al. *Front. in Cell. Neur.* 2017]

- Type II diabetes (50 microbial genes → AUC ROC 0.81)

[Qin et al. *Nature* 2012]

- Parkinson's differential abundance of gut microbial species

[Heintz-Buschart et al. *Mov. Disord.* 2018]

A tiny fraction of microbes are cultivable in a lab...  
But how to deal with this with data science point of view?

# Sequence-structure-function paradigm

# Studying biological function through DNA information

From an organism to its **genome**...



Organism

$\downarrow^1$

---

<sup>1</sup>Need some computational biology magics

# Studying biological function through DNA information

From an organism to its **genome**...



Organism



DNA

$\downarrow^1$

---

<sup>1</sup>Need some computational biology magics

# Studying biological function through DNA information

From an organism to its **genome**...



Organism



DNA



Illumina/Nanopore

↓<sup>1</sup>

---

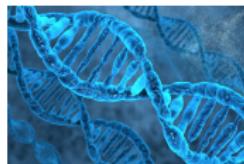
<sup>1</sup>Need some computational biology magics

# Studying biological function through DNA information

From an organism to its **genome**...



Organism



DNA



Illumina/Nanopore



A vertical sequence of DNA base pairs (A, T, C, G) in a light blue font, representing a portion of the genome.

5kbp - 5Mbp

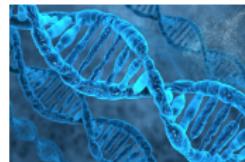
<sup>1</sup>Need some computational biology magics

# Studying biological function through DNA information

From an organism to its **genome**...



Organism



DNA



Illumina/Nanopore



5 kbp - 5 Mbp

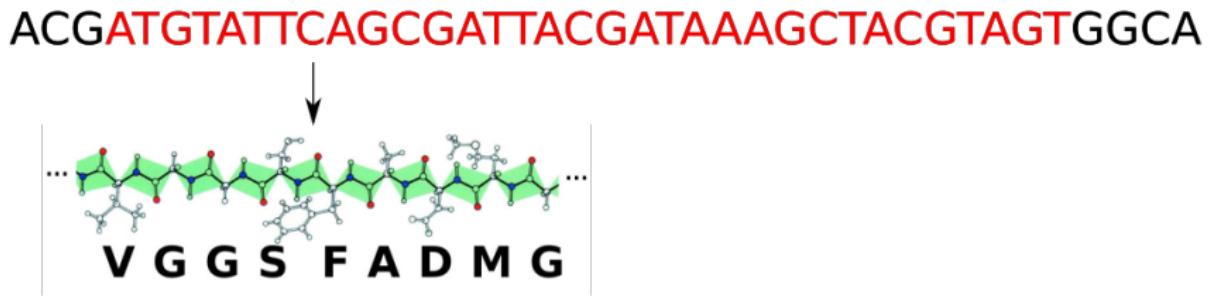
How does it help?

<sup>1</sup>Need some computational biology magics

# Bioinformatics: from genome to function

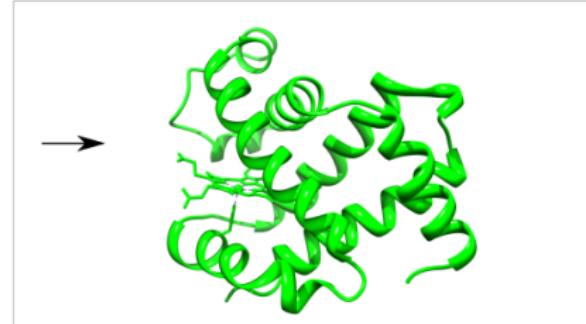
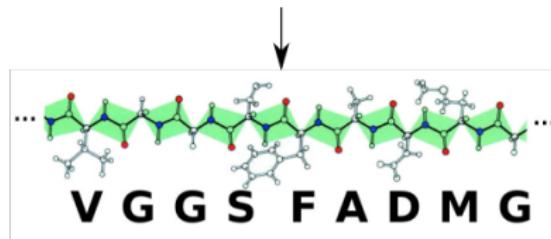
ACGATGTATTCA  
GCGATTACGATAAAGCTACGTAGTGGCA

# Bioinformatics: from genome to function



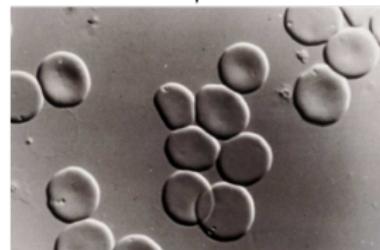
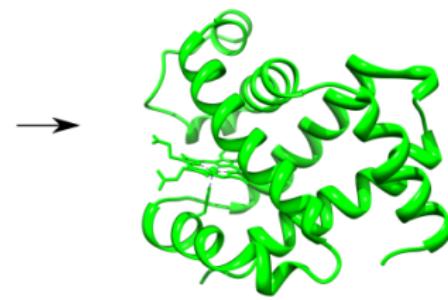
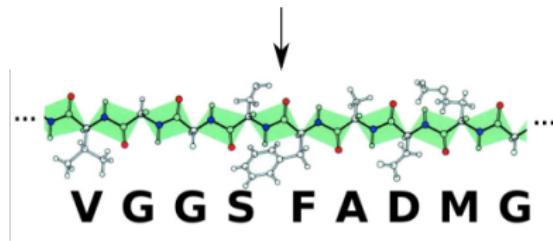
# Bioinformatics: from genome to function

ACG**ATGTATT**CAGCGATTACGATAAAGCTACGTAGT**GGCA**



# Bioinformatics: from genome to function

ACG**ATGTATTCA**GCAGATTACGATAAAGCTACGTAGT**GGCA**



O<sub>2</sub> transport

## Sequence-structure-function

This sequence-structure-function paradigm is the main motivation for studying biology from DNA information.

Why not directly carry studies at the structural or functional level?

# Sequence-structure-function

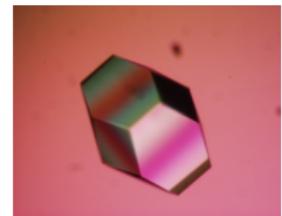
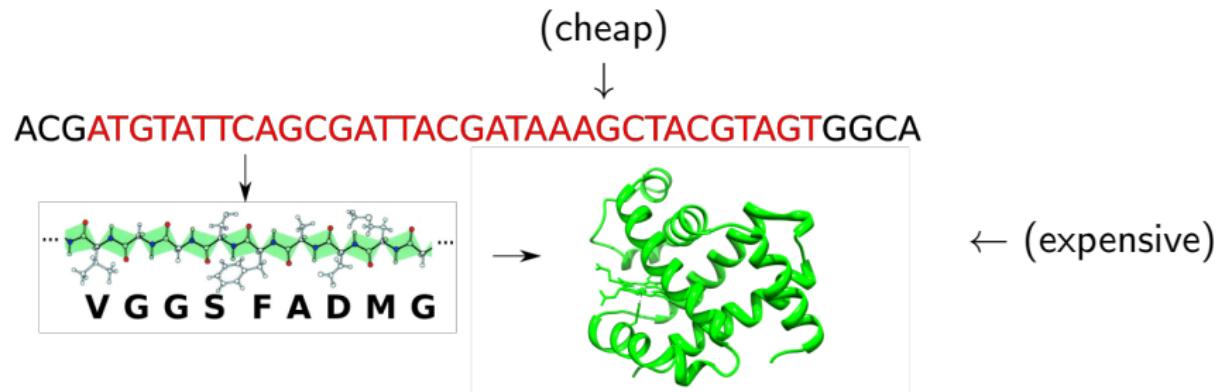
This sequence-structure-function paradigm is the main motivation for studying biology from DNA information.

Why not directly carry studies at the structural or functional level?



It's more work! It is actually done by biologist for specific case of interest.

# Computers and protein structure prediction



# Hands-on

# Hands-on

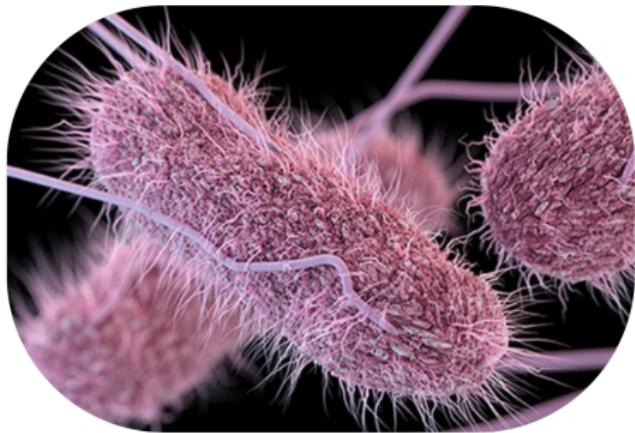
Disclaimer:

- **No fully guided syllabus**
- Act as a junior professional
  - Analyze provided information, think of a solution
  - Ask/discuss with your colleagues
  - Ask/discuss with your senior colleague (me)

Hands-on: realistic public health issue

**!ALERT!**

# Salmonella OUTBREAK



Breaking news

Bad infections kill many people. Antibiotics do nothing.

# Plan of this session

First, think and plan - 1h.

- Skim the context in the `hands-on/session1` on the git (15 min)
- Try to understand individually the work you will have to do and write down questions you have (10 min)
- Share your understanding with people in your group (10 min)
- We share together our understanding and elaborate a common strategy (20 min).

Then start developing - till the end :).

- Start developing T1, paying attention to pitfalls (noise in the data in particular)
- Build your own tests. You can make use of the data in the `hands-on/reference-data` directory

# Genomics

From DNA to **reads**...



DNA



Illumina



reads ( $\sim 250\text{bp}$ )

$$\eta_{err} \approx 1\%$$

# Genomics

From DNA to **reads**...



DNA



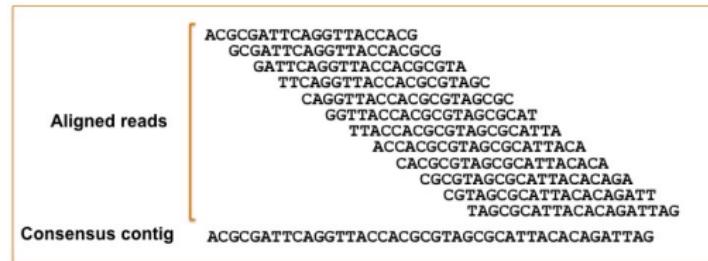
Illumina



reads ( $\sim 250\text{bp}$ )

$$\eta_{err} \approx 1\%$$

Assembly: from reads to **contigs**:



# Sequencing data

Two leading technologies:

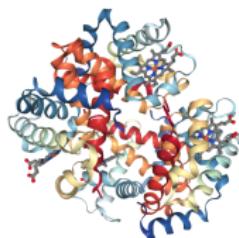
- Illumina: pieces of sequences (called **reads**, 150-250bp)
  - +: reliable, about 1% sequencing errors.
  - -: short reads, only have local view of the genome
  - Errors: rare (1 over 200 bases) almost uniformly distributed, almost all **mutations**.
- Nanopore: long reads, 10kb-100kb
  - +: long reads, easy to assemble, cheap and portable
  - -: high error rate
  - Errors: mostly insertion-deletion, mostly homopolymers (e.g. AAAAA)



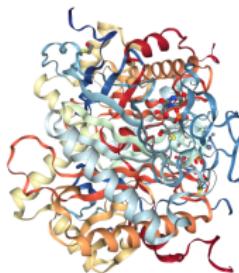


# Predict the structure from sequence: the data

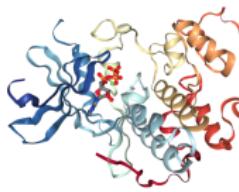
```
>1A3N:A|PDBID|CHAIN|SEQUENCE  
VLSPADTKNVKAANGKVGAAHGEYGAELER  
MPLSFPTTKTYFPHFQDLSHGSAQVKGKGKKV  
ADALTNAAVAVHDMPNALSALSDLHAHKLKV  
DPVNFKLLSHCLVTLA AHLPAEFTPAVHAS  
LDKFPLASVSTVLTSKYR
```



```
>1HXP:A|PDBID|CHAIN|SEQUENCE  
MTQFNPNVDPHPHRRYNPPLTCQWILVSPHRAKRWP  
EGAQETTPAKQVLPAHDPPDCFLCAGNVRTGDKN  
PDYTGTTFPTNDPAALMSDTPDAPESHDPLMRC  
QSAROTTSRVICSPDITKTLPELSVAALTEIVK  
TWQEQTEAELGKTYPFWVQVEENKAAMGCSNPHP  
HQGIWANSFLPKNEAEREEDRLQKEYFAEQKQSPML  
VDYVQRELADGSRTVVETEHNLAVVPVWAAMPPF  
ETLLLPPKAHVRLIRITDLTDAQRSDLALAKKLTS  
RYDNLPQCSPYNSMGWHGAPFNGEEENQHNQLHA  
HYFPPLRLSATVTRKFMVGYEMLAETQRDLTAEQ  
AAERLRAVSDIHPRESGV
```



```
>1HCK:A|PDBID|CHAIN|SEQUENCE  
MENFOKEVKIGEGTYGVYKARNKLTGEVVAL  
KKIRLDTEKTEGVPTSTAIREISLNLKELNHIPVAL  
KLLDVINTENKLYLWFEFLHQDJKKFMVASAL  
TGIPPLPLIKSYLPQLLQGLAFCHSHRVHLHRDL  
KPNQNLINTEGAIKLLADPGLARARAFGVPVRTYT  
HEVVTLYWRAPEIILLGCKYYSSTAVDIWSLGCY  
FAEMVTRRALPFGOSEIDQDLPFRIFRTLGT PDE  
VWWPGVTSMPDYKPSFPKWARQDFSKVVPPLD  
EDGRSLLSLSQMLHYDPNKRISAKAALAHPPFFQD  
VTKPVPHRLR
```



# CASP competition

Blind competition. Simple principle:

- a sequence is given
- have to predict the structure.

# CASP competition

Blind competition. Simple principle:

- a sequence is given
- have to predict the structure.

13th CASP...

... AI wins !

Google's DeepMind

