

# Computational biology

## Sequence-structure-function paradigm

Clovis Galiez



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

November 12, 2019

# Goal

- Get an overview of computational biology topics
  - Topics (genomics, metagenomics, proteomics, etc.)
  - Know some important databases
  - Know standard tools (Blast, PyMol) 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
  - Case study of protein structure prediction

# Lecture organization

- Session I: some background, starting your project
- Session II hands-on: database mining, sequence searching
- Session III hands-on: protein structure prediction

Evaluation: project-based + bonus for participation

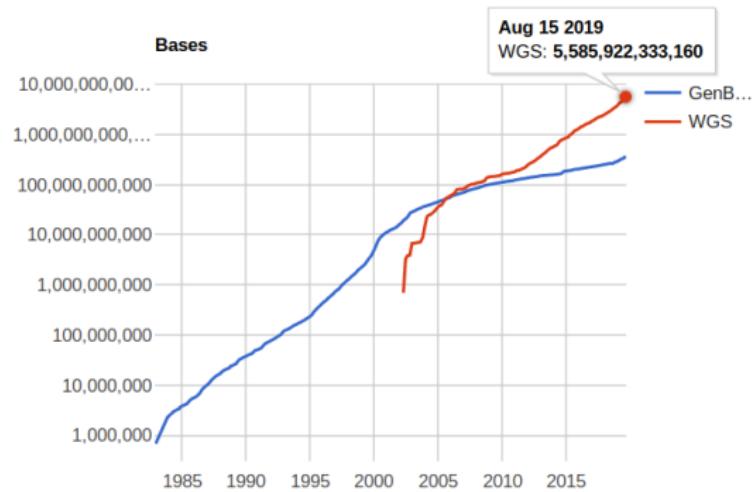
# Today's outline

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

# Why computational biology?

# Genomics: the biggest breakthrough

1977: first DNA sequencer.



# A larger scope: biology from the data perspective

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)

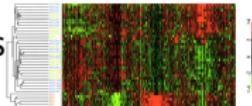
# A larger scope: biology from the data perspective

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

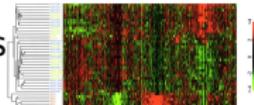
# A larger scope: biology from the data perspective

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

- Sequences **ATTCAGTACAT**

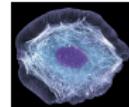
- Genomic (DNA sequence of one organism)
- Metagenomic (DNA sequence of a biological sample - many organisms)
- Proteomic (amino-acid sequences)

- Abundances



- Marker gene/species abundance
- Expression level of genes

- Images



- Neuroimaging
- Cell imaging

- Mass spectrometry

- ...

# Scope of applications for DNA sequence data

## Computational biology

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

Helps getting insights in many fields:

# Scope of applications for DNA sequence data

## Computational biology

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

Helps getting insights in many fields:

-  clinics

# Scope of applications for DNA sequence data

## Computational biology

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

Helps getting insights in many fields:

-  clinics

- ecology



# Scope of applications for DNA sequence data

## Computational biology

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

Helps getting insights in many fields:

-  clinics
-  ecology
-  biochemistry

# Scope of applications for DNA sequence data

## Computational biology

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

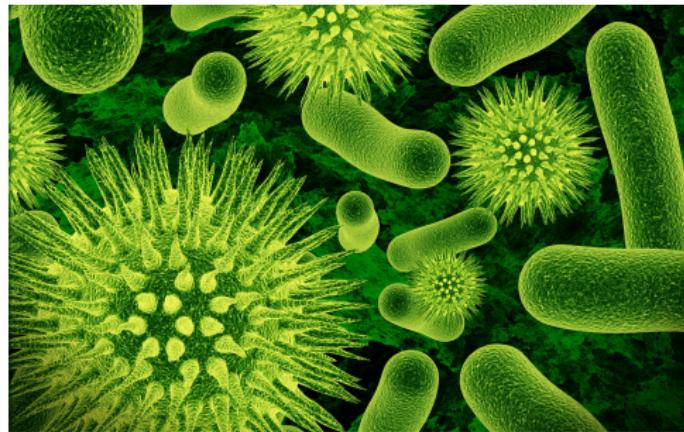
Helps getting insights in many fields:

-  clinics
-  ecology
-  biochemistry

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

# The microbial world

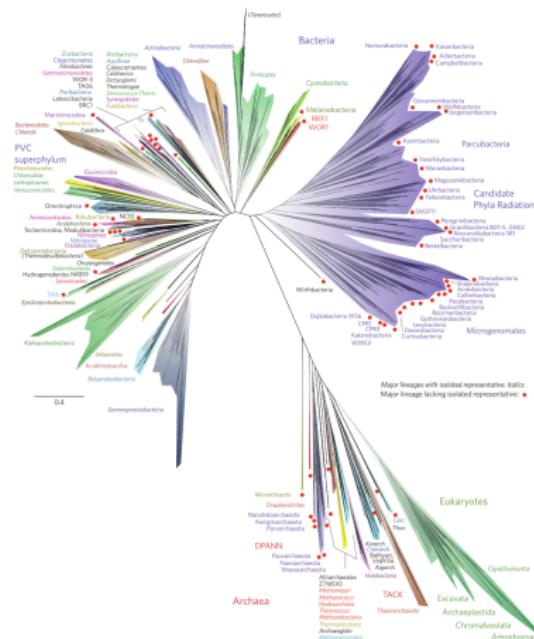
They are everywhere... they work hard 24h a day... they fight against each other... and they collaborate.



Bacteria, phages, picoeukaryotes, viruses, etc.

# Origins and evolution of micro-organisms

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



[Hug et al. 2016]

Some facts are known, but the deep origins are still debated.

# 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]



# Microbiome importance in human health

The bright side:



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

## Germany: Ten die from E.coli-infected cucumbers

⌚ 28 May 2011

f t m ↗ Sh

The dark side:

The death toll in Germany from an outbreak of E.coli caused by infected cucumbers has risen to at least 10.

The cucumbers, believed to have been imported from Spain, were contaminated with E.coli which left people ill with hemolytic-uremic syndrome (HUS).

Hundreds of people are said to have fallen sick.



It is unclear whether the cucumbers were infected at source or in transit

[Karch et al. EMBO Mol. Med. 2012]

# The human gut microbiome

2000's

Human genome



2010's

Gut metagenomes



≈ 20k protein-coding genes

# 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!

# 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

$\downarrow^1$

---

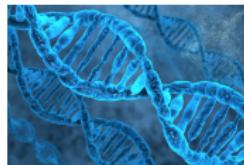
<sup>1</sup>Need some computational biology magics

# Studying biological function through DNA information

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



Organism



DNA



↓<sup>1</sup>

---

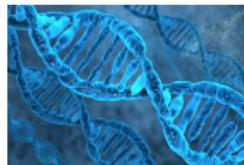
<sup>1</sup>Need some computational biology magics

# Studying biological function through DNA information

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



Organism



DNA



Illumina/Nanopore

↓<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



ATTCGA  
GACTACGGCGA  
GATCTGTGCAGCTGA  
CTAGACTACGACGGAT  
ACTACGGCGATCTACG  
ATCTGTGCAGCTGAT  
CTACGACGTTCA  
ATCGT

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

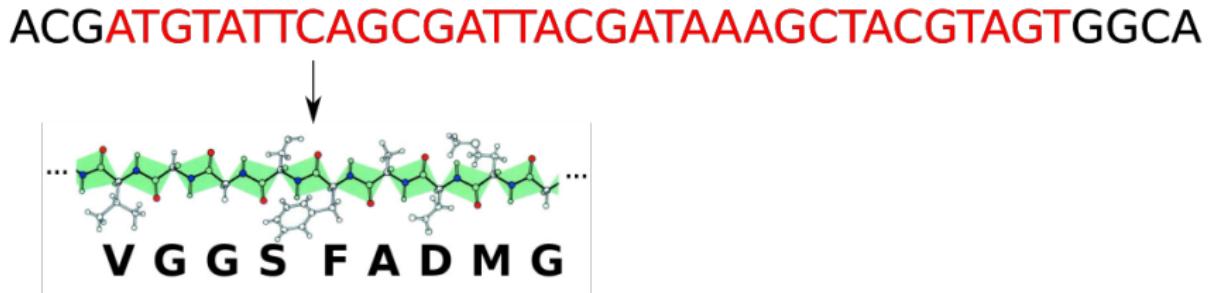
How does it help?

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

# Bioinformatics: from genome to function

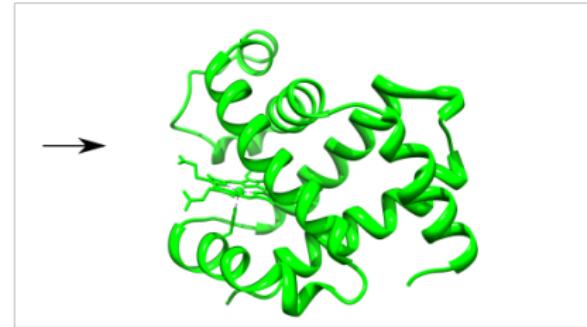
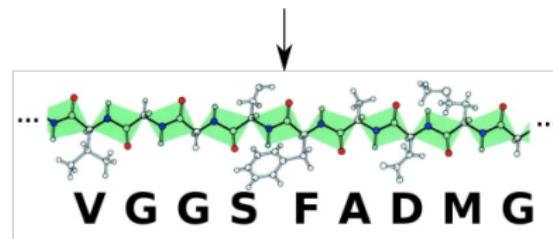
ACGATGTATTCAAGCGATTACGATAAAGCTACGTAGTGGCA

# Bioinformatics: from genome to function



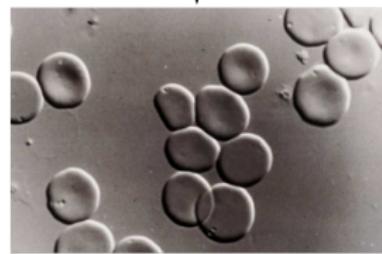
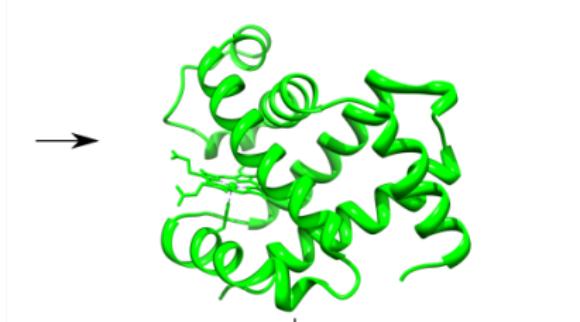
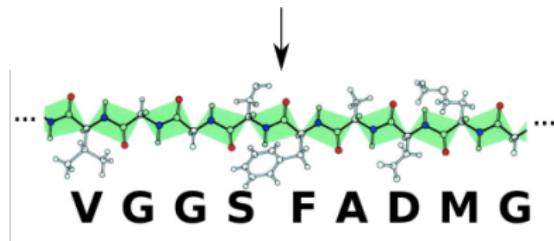
# Bioinformatics: from genome to function

ACG**ATGTATTCA**GCGATTACGATAAAGCTACGTAGT**GGCA**



# Bioinformatics: from genome to function

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA



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

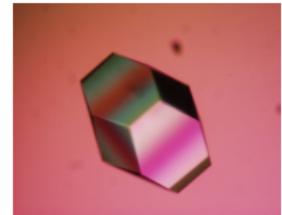
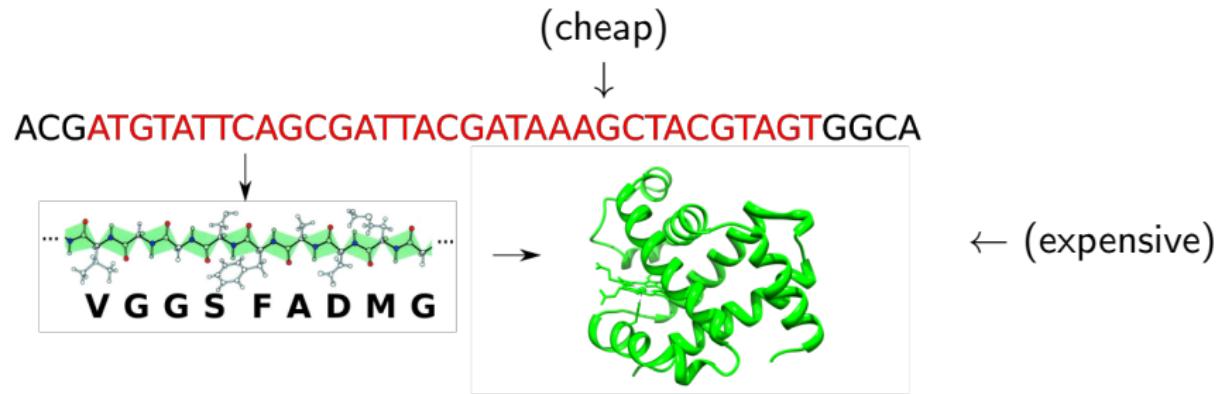
This sequence-structure-function paradigm is the main motivation for studying biology from DNA information.

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



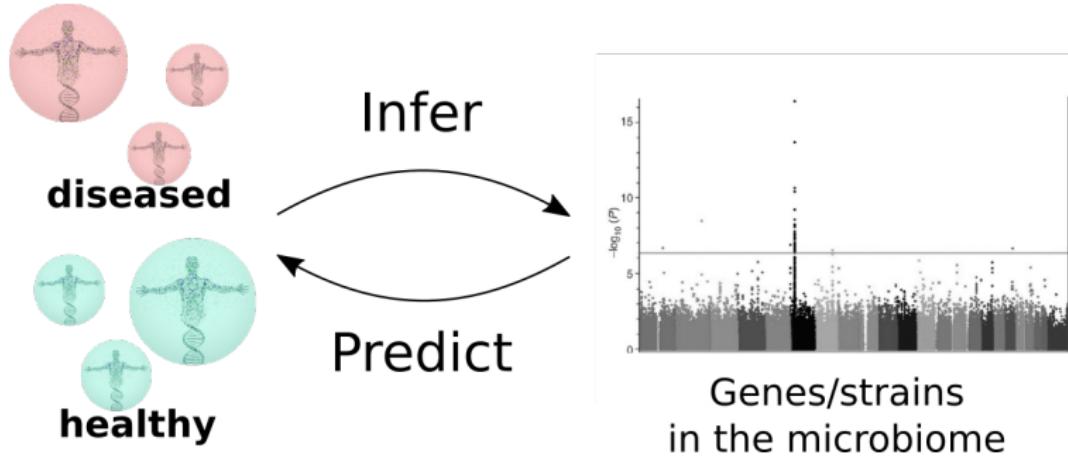
It's more work! It is actually done by biologist for specific case of interest.

# Computers and protein structure prediction



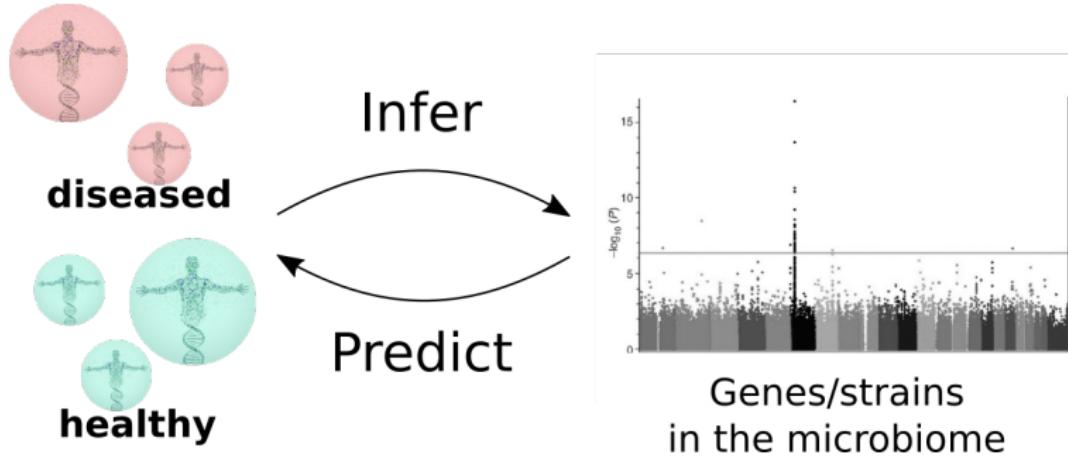
# Statistical association studies

Relates the variation of the microbiome to the phenotype.



# Statistical association studies

Relates the variation of the microbiome to the phenotype.



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

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

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

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

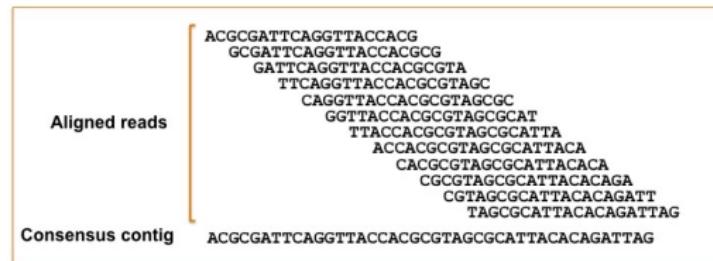
# Genomics

## From DNA to **reads**...



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

## Assembly: from reads to **contigs**:



# Sequencing data

Two leading technologies:

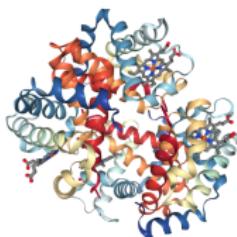
- Illumina: pieces of sequences (called **reads**, 150-250bp)
  - +: reliable, about 1% sequencing errors.
  - -: short reads, only have local view of the genome
  - Errors: rare (1 over 200 bases) almost uniformly distributed, almost all **mutations**.
- Nanopore: long reads, 10kb-100kb
  - +: long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



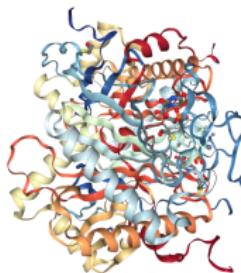


# Predict the structure from sequence: the data

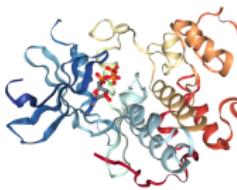
```
>1A3N:A|PDBID|CHAIN|SEQUENCE  
VLSPADTKNVKAANGKVGAAHGEYGAELER  
MPLSFPTTKTYFPHFDSLHSQAVKGHGKKV  
ADALTNAAVAHVDDMPNALSLSLHAKLRLV  
DPVNFKLLSHCLVTLA AHLPAEFTPAVHAS  
LDKFPLASVSTVLTSKYR
```



```
>1HXP:A|PDBID|CHAIN|SEQUENCE  
MTQFNPNVDHPHRRYNNPLTCQWILVSPHRAKRWP  
EGAQETTPAKQVLPAHDPPDCFLCAGNVRTGDKN  
PDYTGTTFPTNDPAALMSDTPDAPESHDPLMR  
QSAROTTSRVICSPDITKTLPELSVAALTEIVK  
TWQEQTEAELGKTYPFWVQVEENKAAMGCSNPHP  
HQGIWANSFLPKNEAEEREDRLQKEYFAEQKSPML  
VDYVQRELADGSRTVVETEHNLAVVPVWAAMPF  
ETLLLPPKAHVRLIRITDLTDAQRSDLALAKKLTS  
RYDNLPQCSPYNSMGWHGAPFNGEEENQHNQLHA  
HYFPPLRLSATVTRKFMVGYEMLAETQRDLTAEQ  
AAERLRAVSDIHPRESGV
```



```
>1HCK:A|PDBID|CHAIN|SEQUENCE  
MENFOKEVKIGEGTYGVYKARNKLTGEVVAL  
KKIRLDTEKTEGVPTSTAIREISLKLKELNHIPNIV  
KLLDVINTENKLYLWFELNQDJKKFMDSAL  
TGIPPLPLIKSYLPQLLQGLAFCHSHRVHLHRDL  
KPNQNLINTEGAIKLLADPGLARARAFGVPVRTYT  
HEVTVLWYRAPEIILLGCKYYSYSTAVDIWSLGC  
FAEMVTRRALPFGOSEIDQDLPFRIFRTLGT PDE  
VWWPGVTSMPDYKPSFPKWARQDFSKVVPPLD  
EDGRSLLSQMLHYDPNKRISAKAALAHPPFFQ  
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

