Computational biology

Homology and sequence alignment

Clovis Galiez

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

September 30, 2025

Today's outline: from gene sequence to protein structure

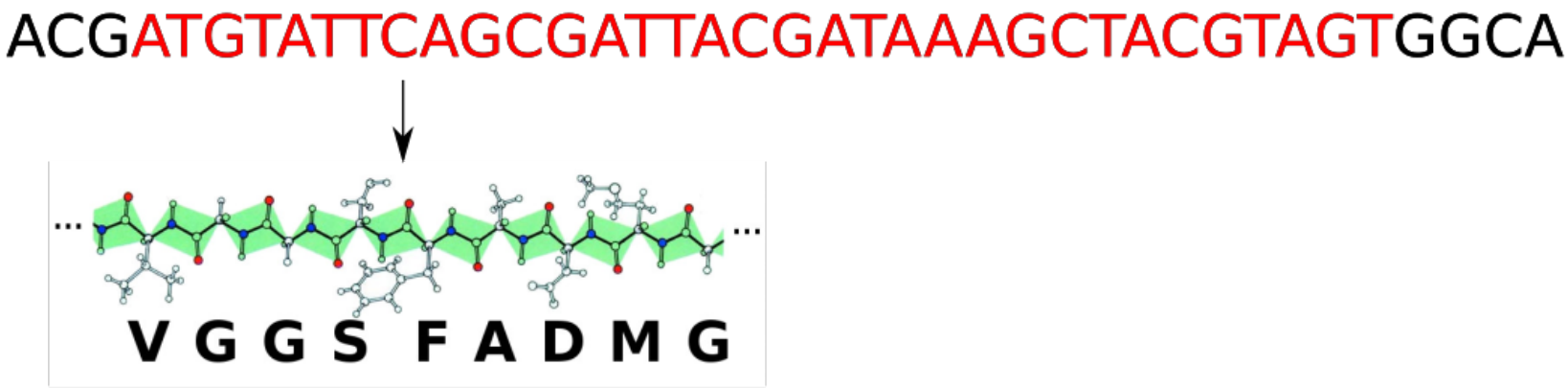
- Sequence-structure-function paradigm
  - Genomes, genes, proteins
  - Databases
- Evolution
  - Selection
  - Sequence homology
  - Multiple sequence alignment

From genome to function, the very big picture

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

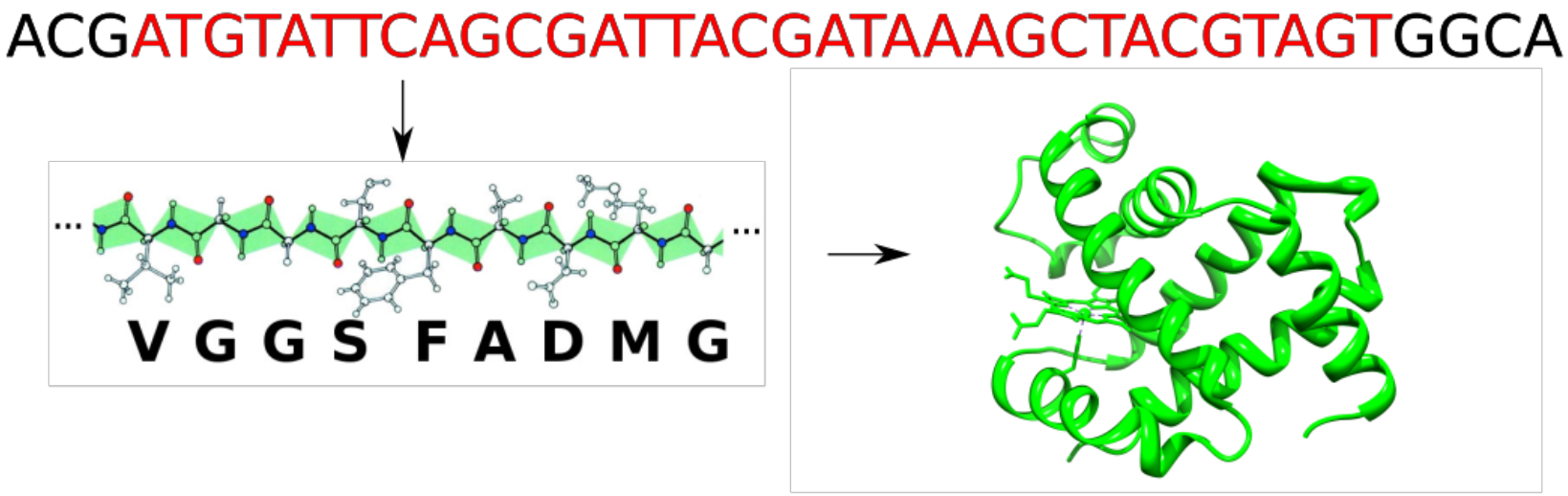
On a genome, specific motifs define begining and end of a gene

From genome to function, the very big picture



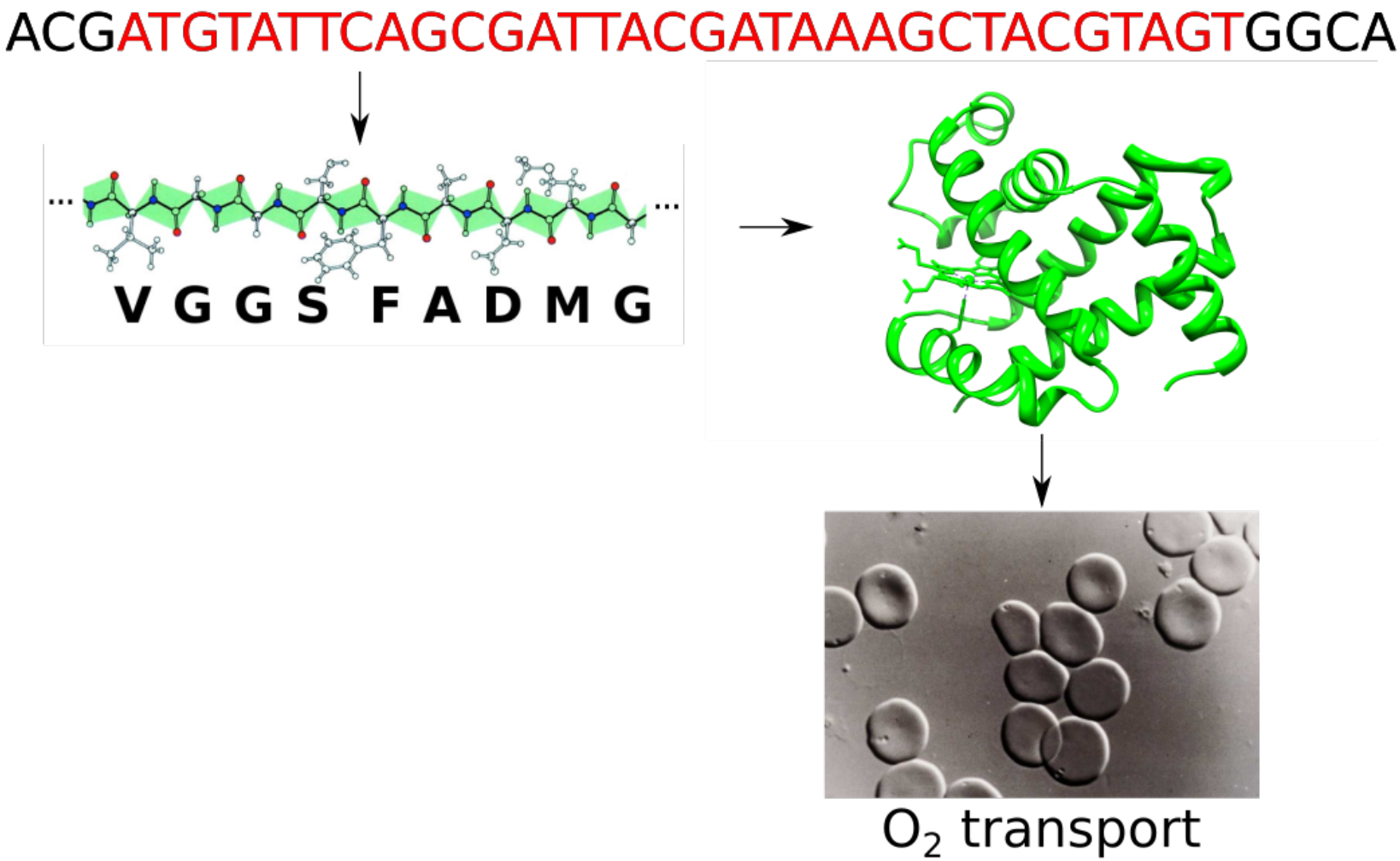
Transcription + *translation*, to form a chain of amino acids (∼300-3000AA)

From genome to function, the very big picture



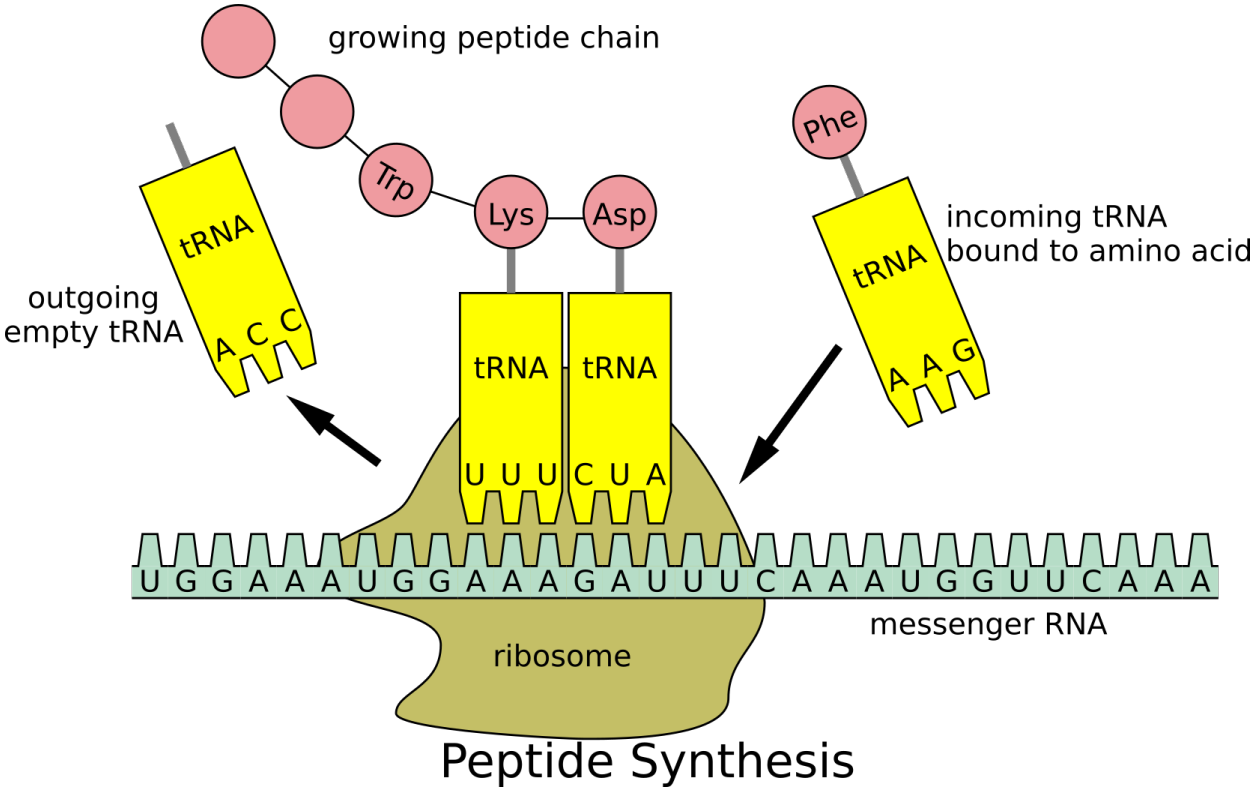
Protein folding under pysico-chemical interactions, diameter $\sim$ few nanometers

From genome to function, the very big picture



Protein endowed with a function (biochemical reactions, transport, etc.)

Zoom: genes to proteins (translation)



RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

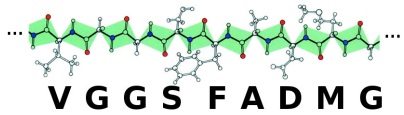
Amino Acids

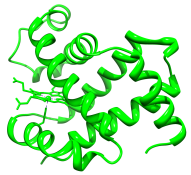
Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

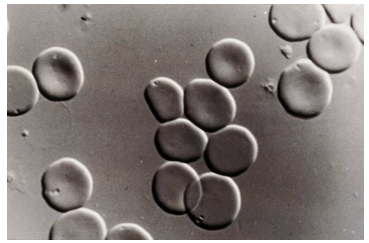
Data at every steps

Nucleic seq. | Amino acid seq. | Protein | Function

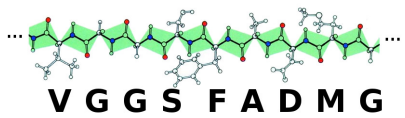
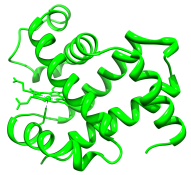
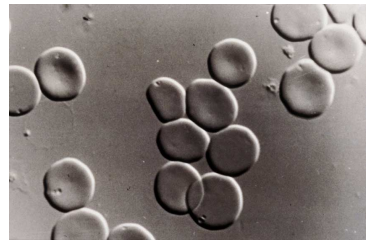
..ATTGTCGATGAC..



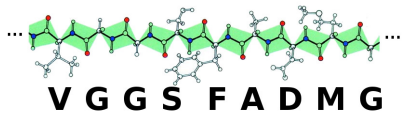
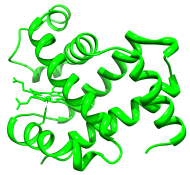
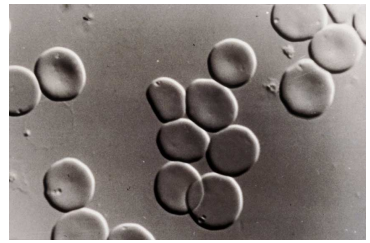




Data at every steps

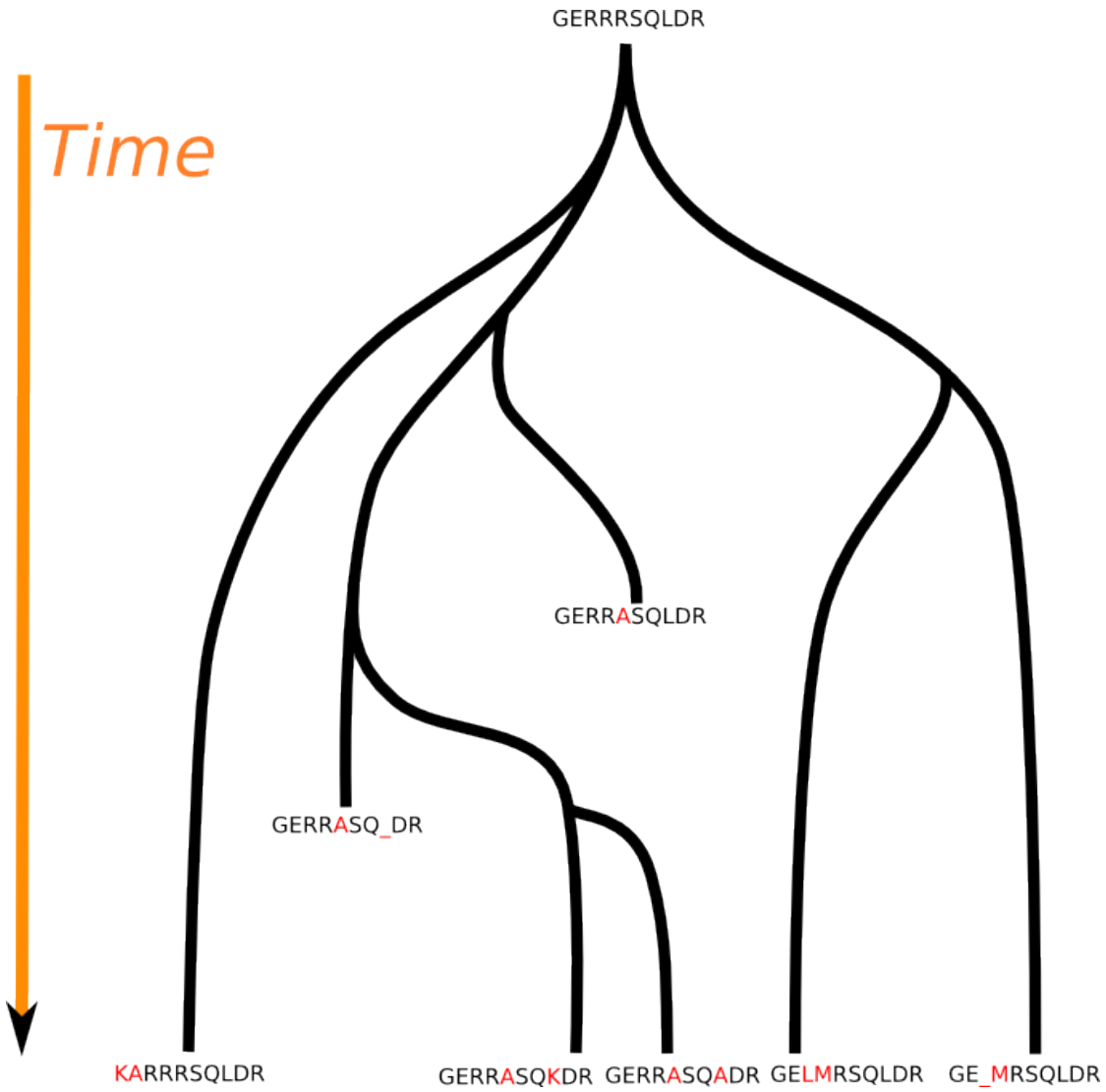
Nucleic seq.	Amino acid seq.	Protein	Function
..ATTGTCGATGAC..			
ncbi.nlm.nih.gov	uniprot.org	rcsb.org	geneontology.or

Data at every steps

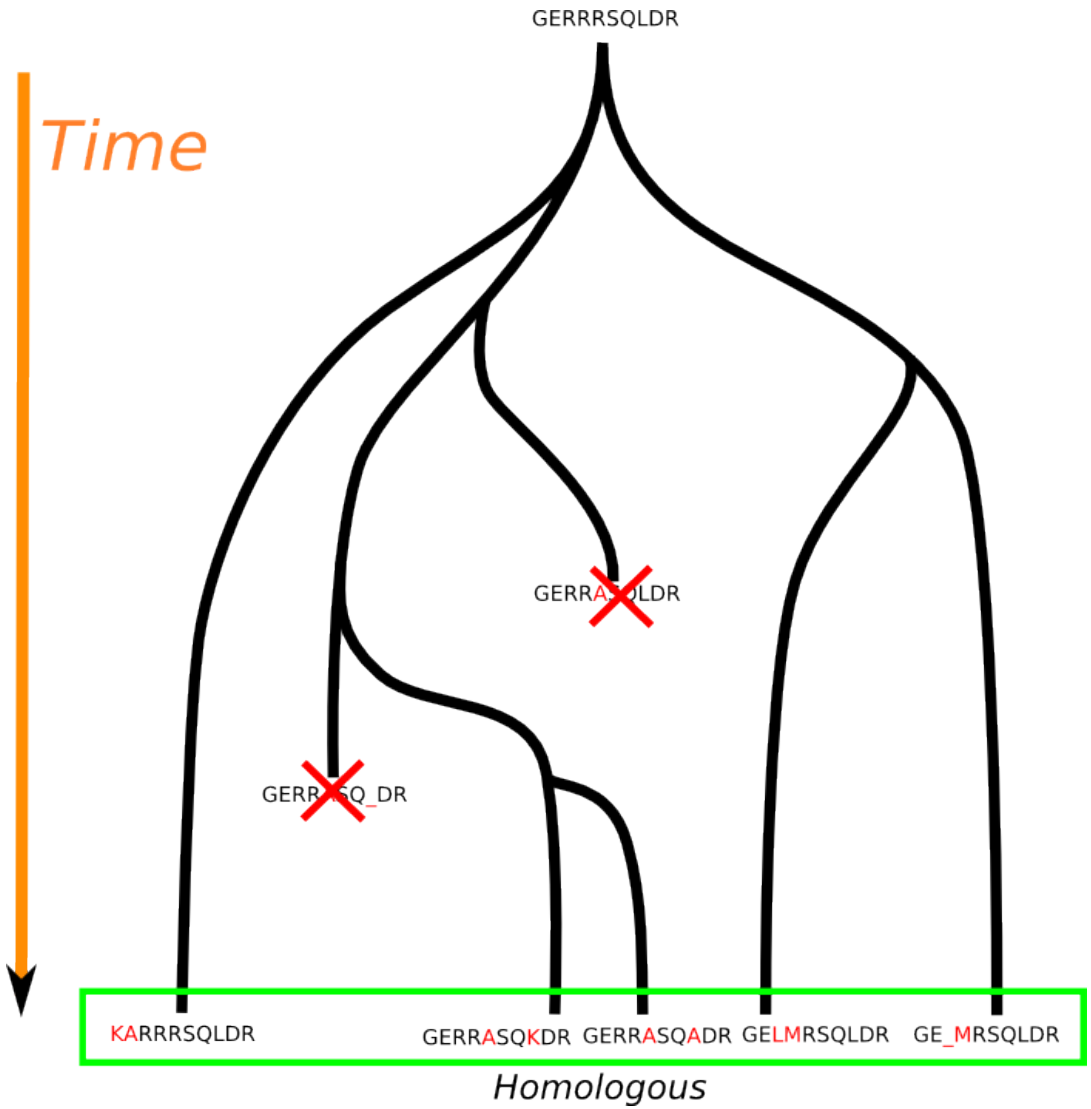
Nucleic seq.	Amino acid seq.	Protein	Function
..ATTGTCGATGAC..			
ncbi.nlm.nih.gov	uniprot.org	rcsb.org	geneontology.or

How do we predict the function from the sequence?

Protein evolution through mutations



Protein evolution through mutations



Sequence alignement: algorithm and p-value

Find the best alignment between your query sequence S_Q and a reference sequence S_R:

```
MEAIGNA.GSAI
QEAIGNAMGSNI
```

Sequence alignement: algorithm and p-value

Find the best alignment between your query sequence S_Q and a reference sequence S_R:

```
MEAIGNA.GSAI
QEAIGNAMGSNI
```

Algorithm (sketch):

- given a 20×20 matrix of scores between amino-acids, set gap penalties
- find the alignment maximizing the total score.

Can be solved by **dynamic programming** in $\mathcal{O}(L^2)$ (see *Smith-Waterman algorithm*).

Sequence alignement: algorithm and p-value

Find the best alignment between your query sequence S_Q and a reference sequence S_R:

```
MEAIGNA.GSAI
QEAIGNAMGSNI
```

Algorithm (sketch):

- given a 20×20 matrix of scores between amino-acids, set gap penalties
- find the alignment maximizing the total score.

Can be solved by **dynamic programming** in $\mathcal{O}(L^2)$ (see *Smith-Waterman algorithm*). An approximate **p-value** can be derived to assess the significance of the alignment.

Under a given p-value threshold we estimate the function to be similar.

Big data: need for heuristic

Uniprot (amino acid db) have more than 200M sequences.

Big data: need for heuristic

Uniprot (amino acid db) have more than 200M sequences. Even with optimized versions of Smith-Waterman, it is still too heavy to compare sequences to all known sequences.

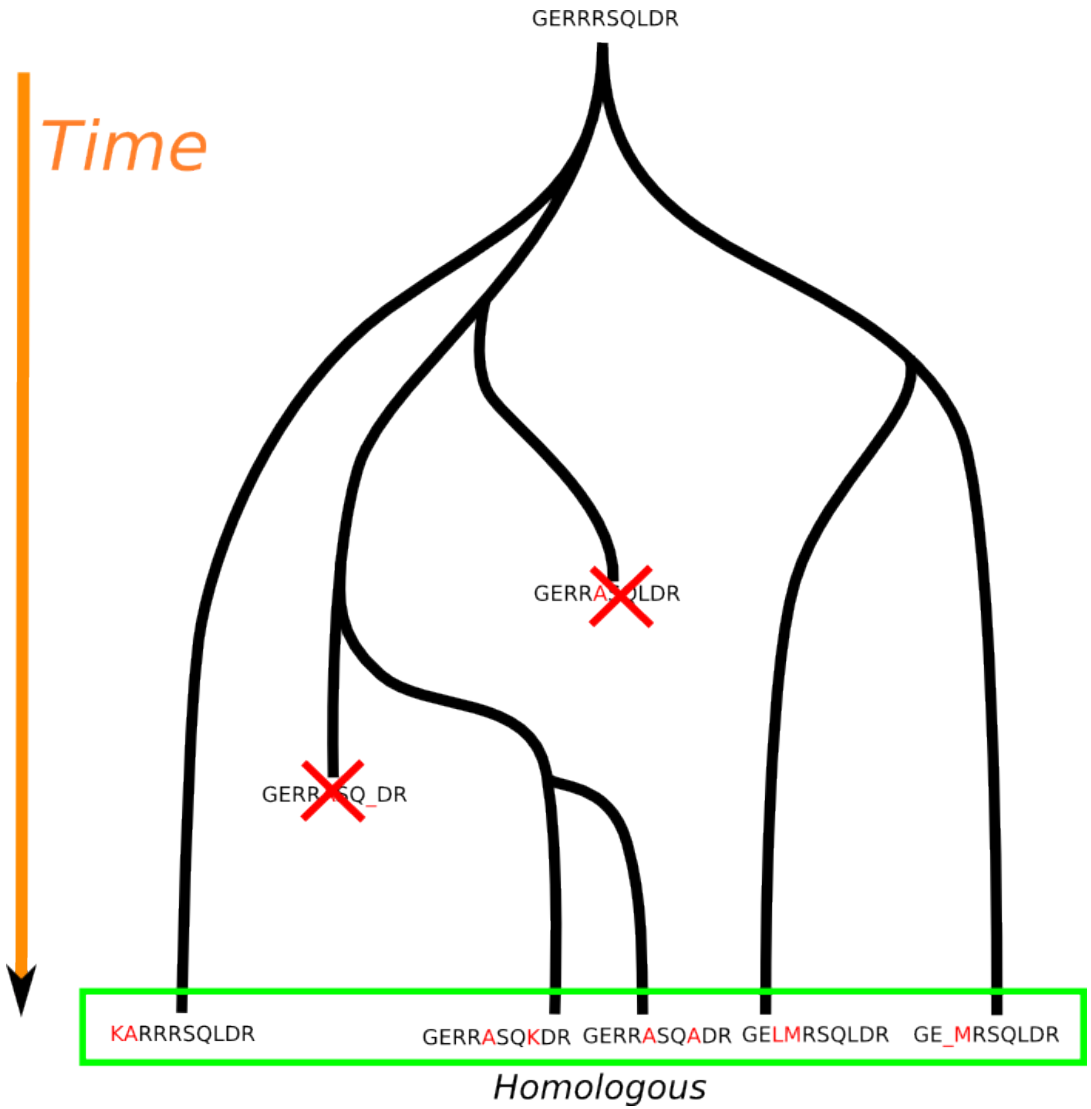
Big data: need for heuristic

Uniprot (amino acid db) have more than 200M sequences. Even with optimized versions of Smith-Waterman, it is still too heavy to compare sequences to all known sequences.
Tools have developed heuristics to filter down the possible target sequences:

- Blast (the historical tool)
- Diamond
- MMseqs2
- ...

Heuristics are mostly based on efficient pre-filtering (often using similar k-mers, with constant time looks up in hash tables).

Sequence conservation



Sequence conservation

Aligning the sequences (MSA, multiple sequence alignment):

Tools	Database
ClustalW [Larkin et al. 07]	Pfam `pfam.xfam.org`

Sequence conservation

Aligning the sequences (MSA, multiple sequence alignment):

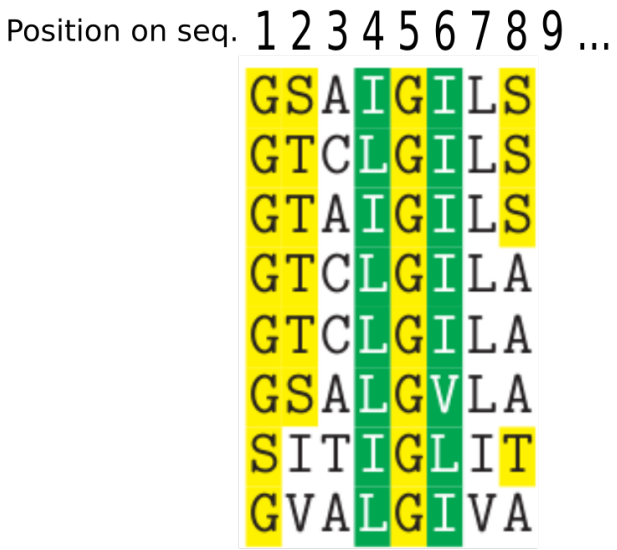
Tools	Database
ClustalW [Larkin et al. 07]	Pfam `pfam.xfam.org`

Why some positions are conserved, some other aren't?

Conserved amino acids are essential for the structure/function

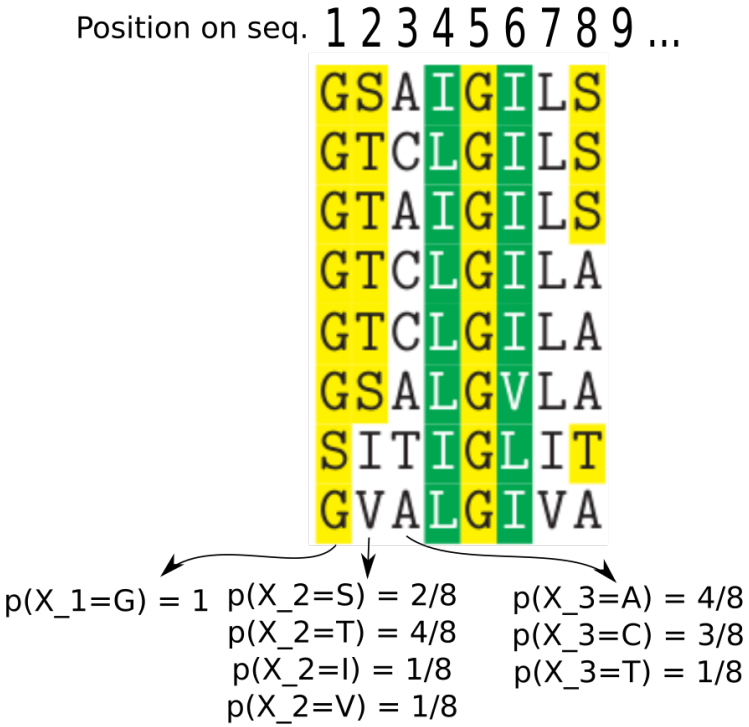
From sequence alignment to profile alignments

From an MSA, we can easily build a **probabilistic model** for modelling the sequence of the protein family:



From sequence alignment to profile alignments

From an MSA, we can easily build a **probabilistic model** for modelling the sequence of the protein family:



From sequence alignment to profile alignments

By assuming independence of positions, one find the best alignment σ that maximizes the likelihood of a given sequence s=GICLGILA:

$$\max_{\sigma} \prod P(X_i = s_{\sigma(i)})$$

From sequence alignment to profile alignments

By assuming independence of positions, one find the best alignment σ that maximizes the likelihood of a given sequence s=GICLGILA:

$$\max_{\sigma} \prod P(X_i = s_{\sigma(i)})$$

This can be solved again using Smith-Waterman algorithm again.
Matching important (=conserved) positions will play an important role in the likelihood $\rightarrow$ it finds homologs with matching conserved regions.

On-line tools and databases

- Blastn Nucl-Nucl comparison
  `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn`
- Blastx Nucl-Prot comparison
  `https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx`
- Pfam Prot-Prot comparison
  `http://pfam.xfam.org/search/sequence`
- Protein structure PDB `https://www.rcsb.org/`

Summary

Check what you've learn:

- What is a genome, a gene, a protein, its structure
- How real sequencing data look like
- What is a SNP, what can be the impact
- Main tools and databases in computational biology
- Potential application of computational biology for public health studies

The project involved basic skills from different area:

- biology
- statistics (Poisson distribution)
- algorithmics (linear time algorithms required)

Projects

Remember that your project should be like professional answers to the call:

- Clarity
- Fulfilment of the call
- Trustworthiness in the description of the approach

Projects

Remember that your project should be like professional answers to the call:

- Clarity
- Fulfilment of the call
- Trustworthiness in the description of the approach

You should send:

- a $\approx$ 5-page report, including:
  - description of the strategy
  - approximations and choices
  - application to the project data (what gene is impacted by the SNP)
- your code
- a step-by-step guide to reproduce the results of the report

The TATFAR
waits for
interesting answers to its call!