Computational biology

Co-evolution to predict protein structures

Clovis Galiez

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

October 19, 2021

Today's outline: from gene sequence to protein structure

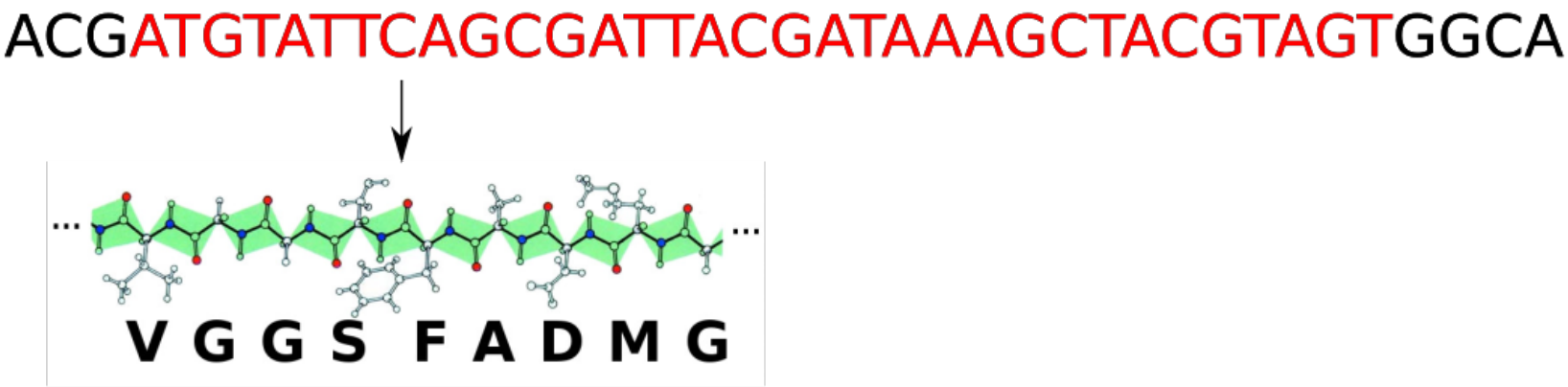
- Sequence-structure-function paradigm
  - Genomes, genes, proteins
  - Databases
- Evolution
  - Selective pressure
  - Multiple sequence alignment
  - Co-evolution

From genome to function, the very big picture

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

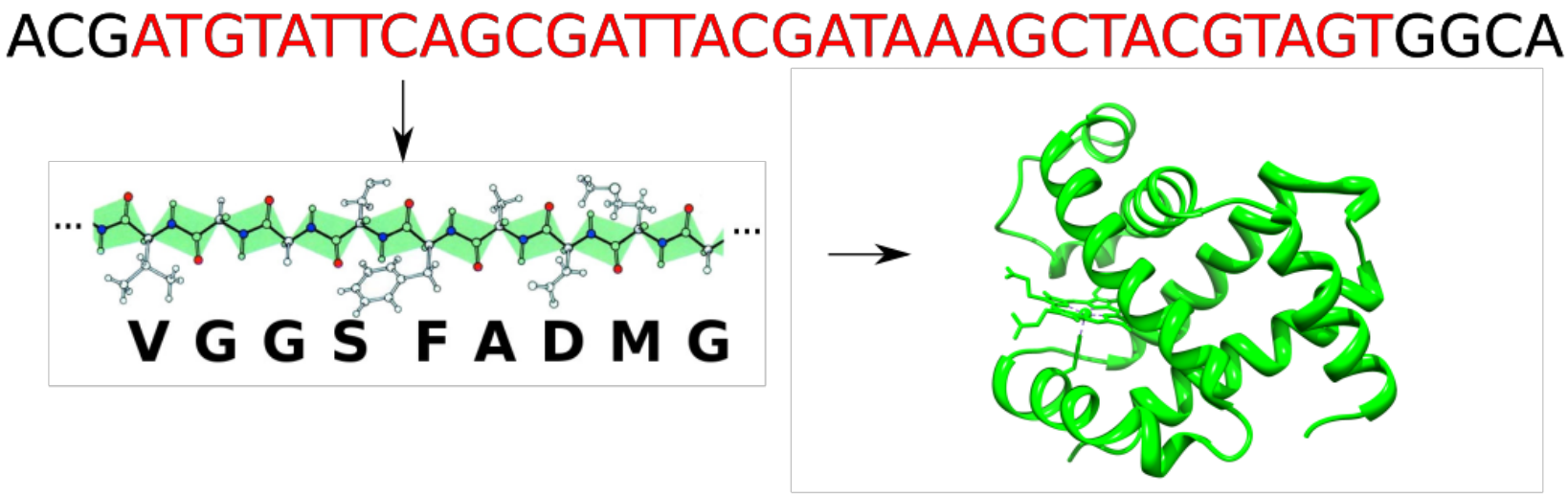
On a genome (∼5Mbp), specific motifs define begining and end of a gene

From genome to function, the very big picture



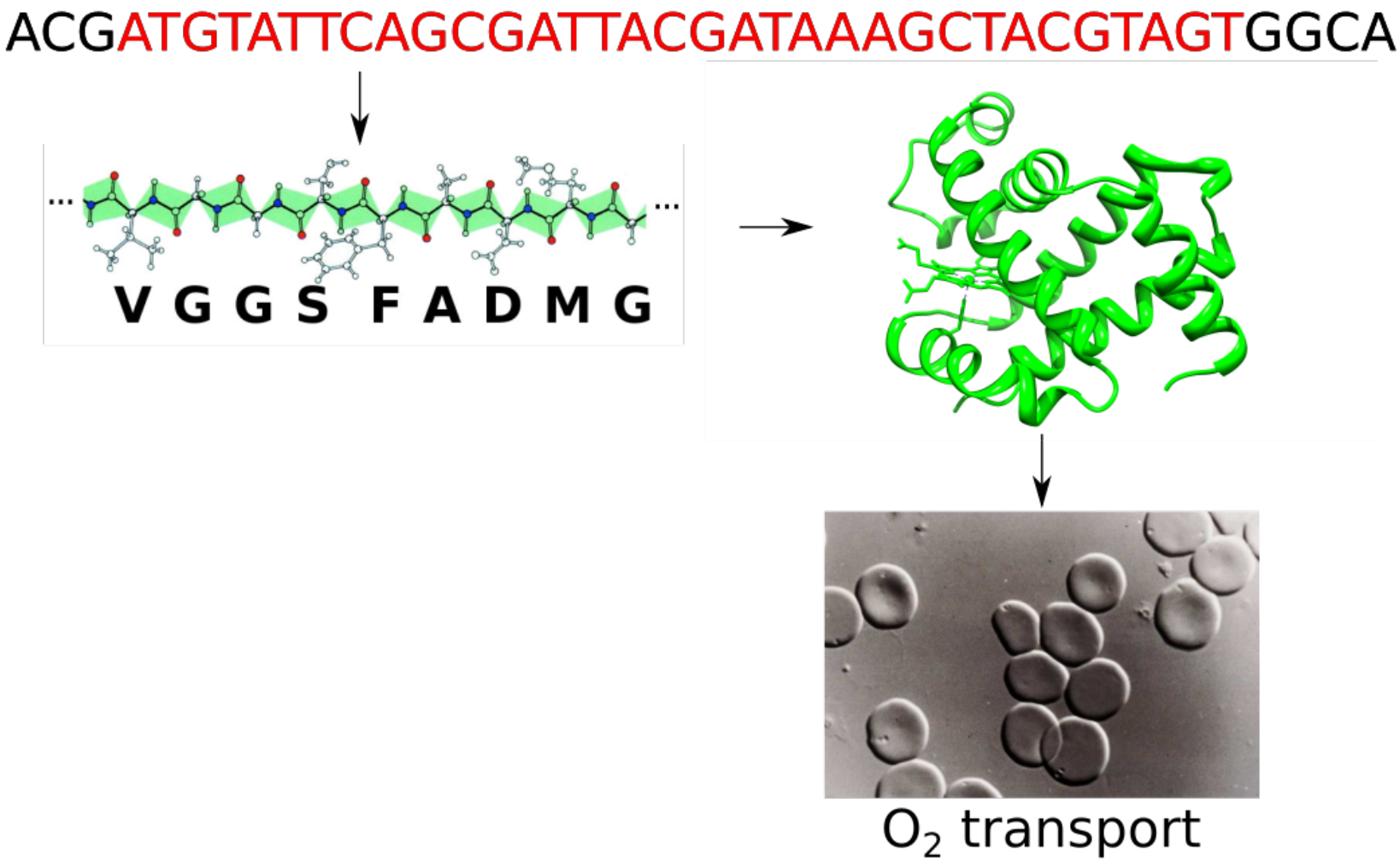
Transcription + *translation*, to form a chain of amino acids (∼300-3000AA)

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

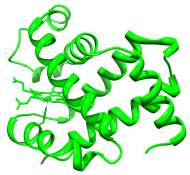


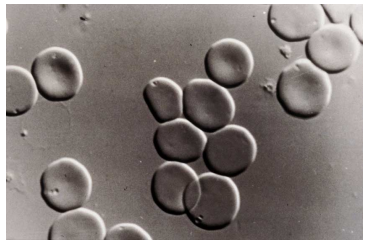
Protein endowed with a function (biochemical reactions, transport, etc.)

Data at every steps

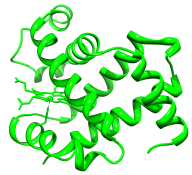
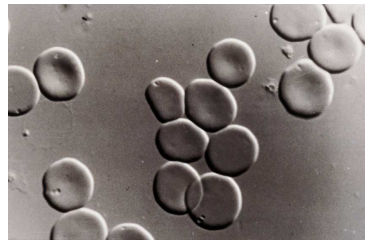
Nucleic seq. | Amino acid seq. | Protein | Function

..ATTGTCGATGAC..



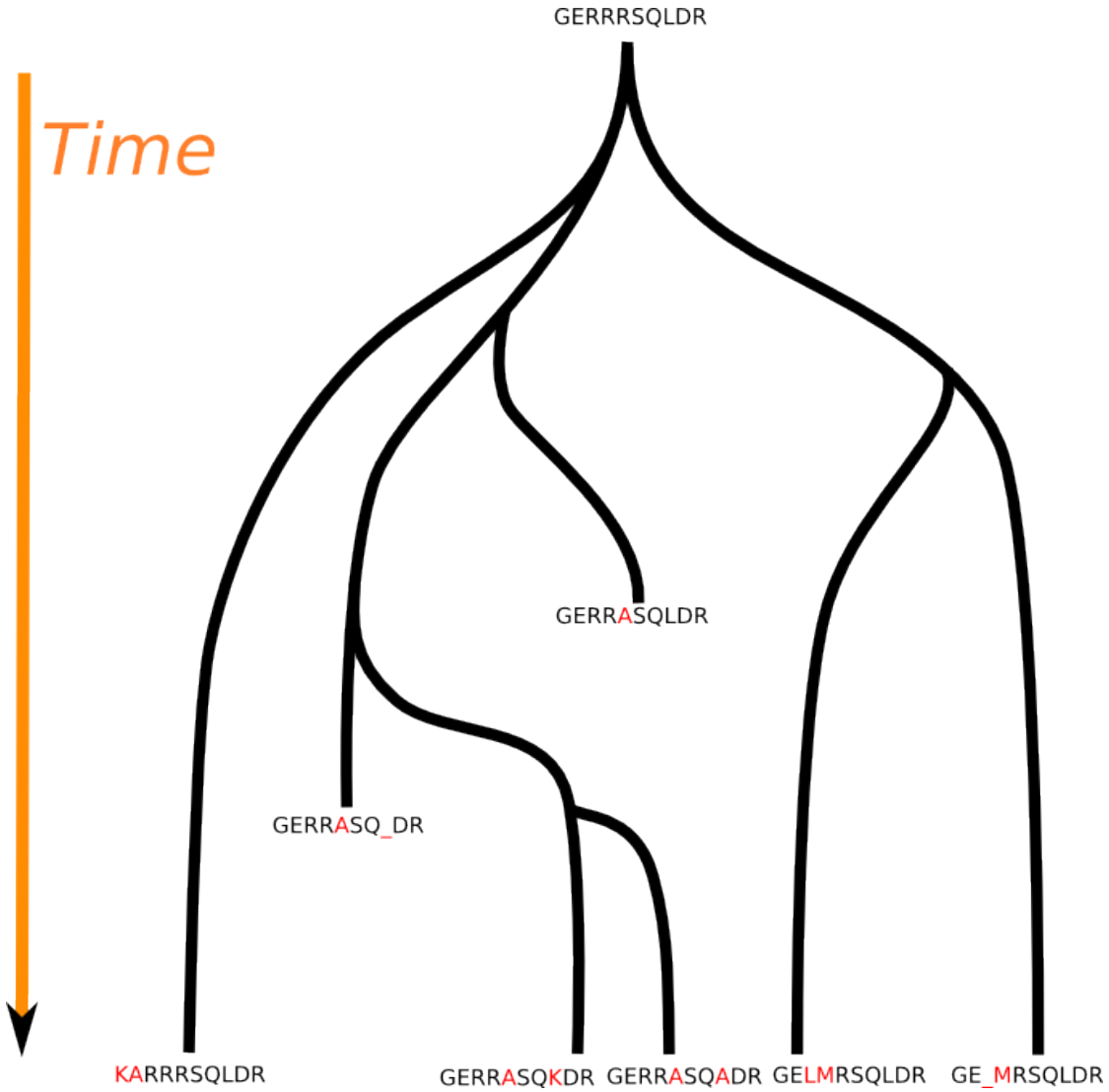


Data at every steps

Nucleic seq.	Amino acid seq.	Protein	Function
..ATTGTCGATGAC..			
NCBI	UniProt	RCSB PDB PROTEIN DATA BANK	-
`ncbi.nlm.nih.gov`	`uniprot.org`	`rcsb.org`	-

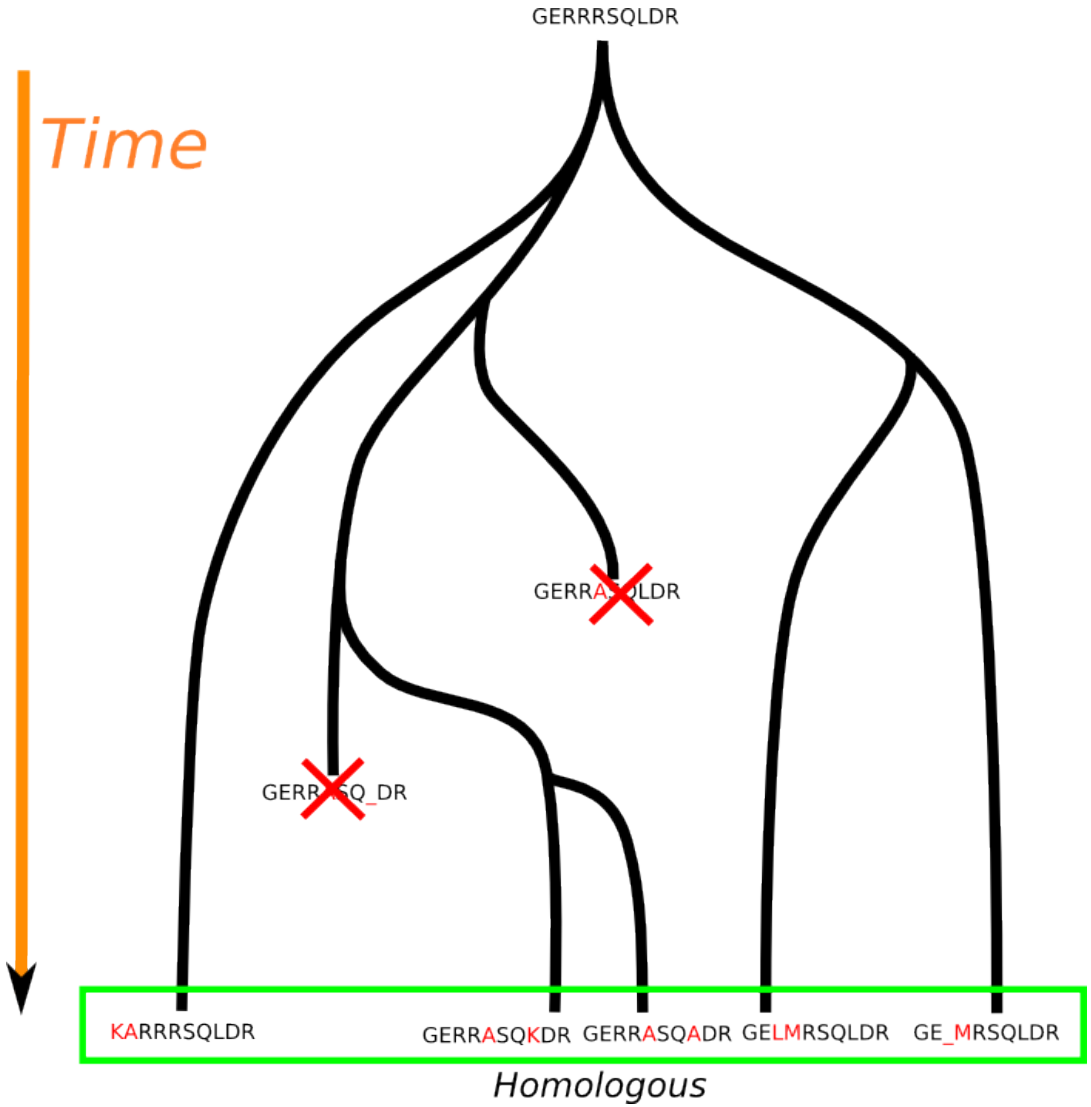
Protein evolution through mutations

We arrange sequences in a phylogenetic tree:



Protein evolution through mutations

We arrange sequences in a phylogenetic tree:



How to predict gene function?

Some gene functions have been previously identified by biologists.

When having an unknown sequence, how can you guess its function?

How to predict gene function?

Some gene functions have been previously identified by biologists.

When having an unknown sequence, how can you guess its function?

By comparing to millions of existing sequences and hope that **homologuous** genes are already known.

How to predict gene function?

Some gene functions have been previously identified by biologists.

When having an unknown sequence, how can you guess its function?

By comparing to millions of existing sequences and hope that **homologuous** genes are already known.

How to compare sequences?

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.

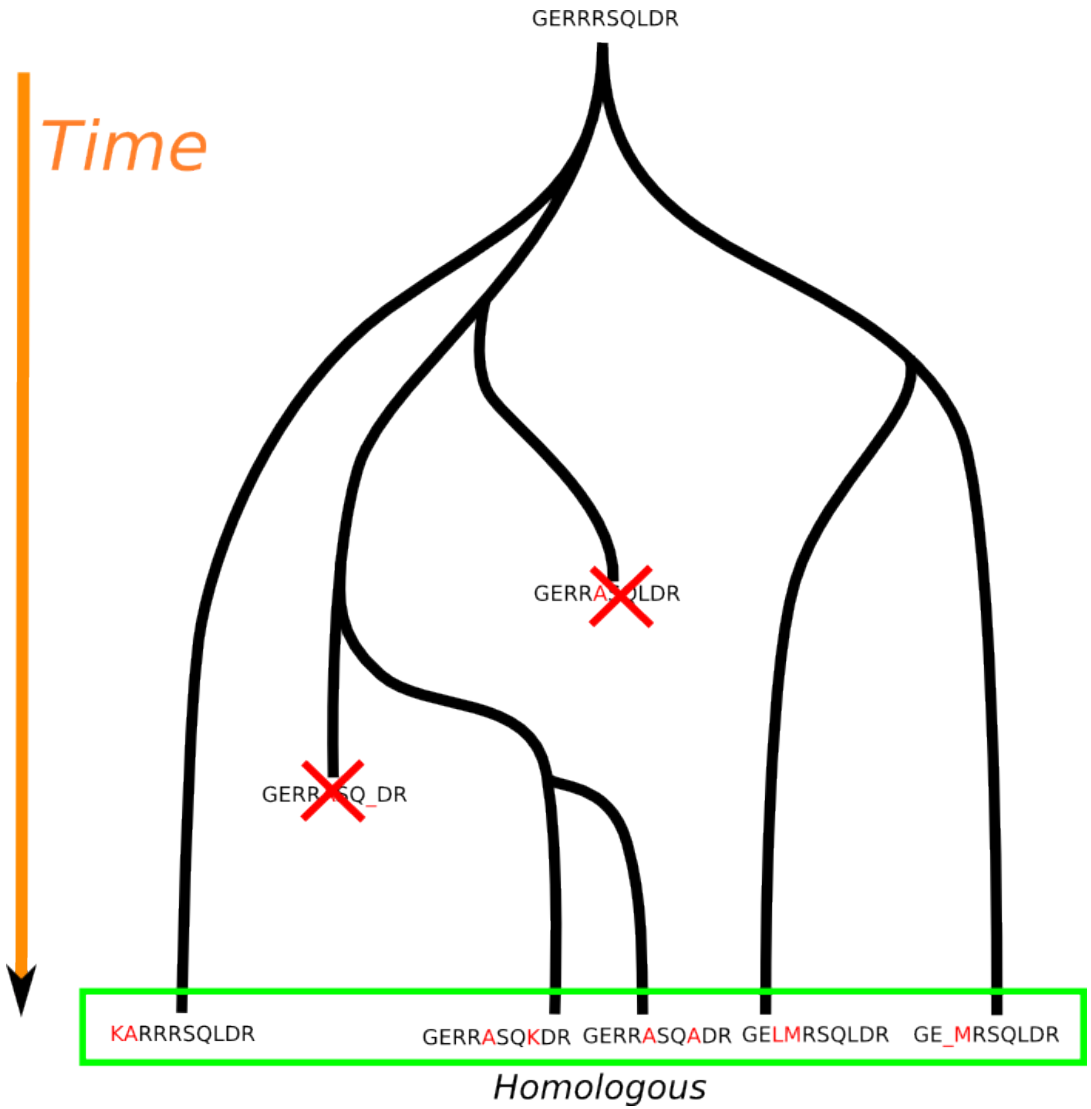
Big data: need for heuristic

Even with optimized versions of Smith-Waterman, it is still too heavy to compare sequences to all know sequences.
Tools have developed heuristics to filter down the possible target sequences:

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

Heuristics are mostly based on similar k-mers, and efficiently filtering through 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):

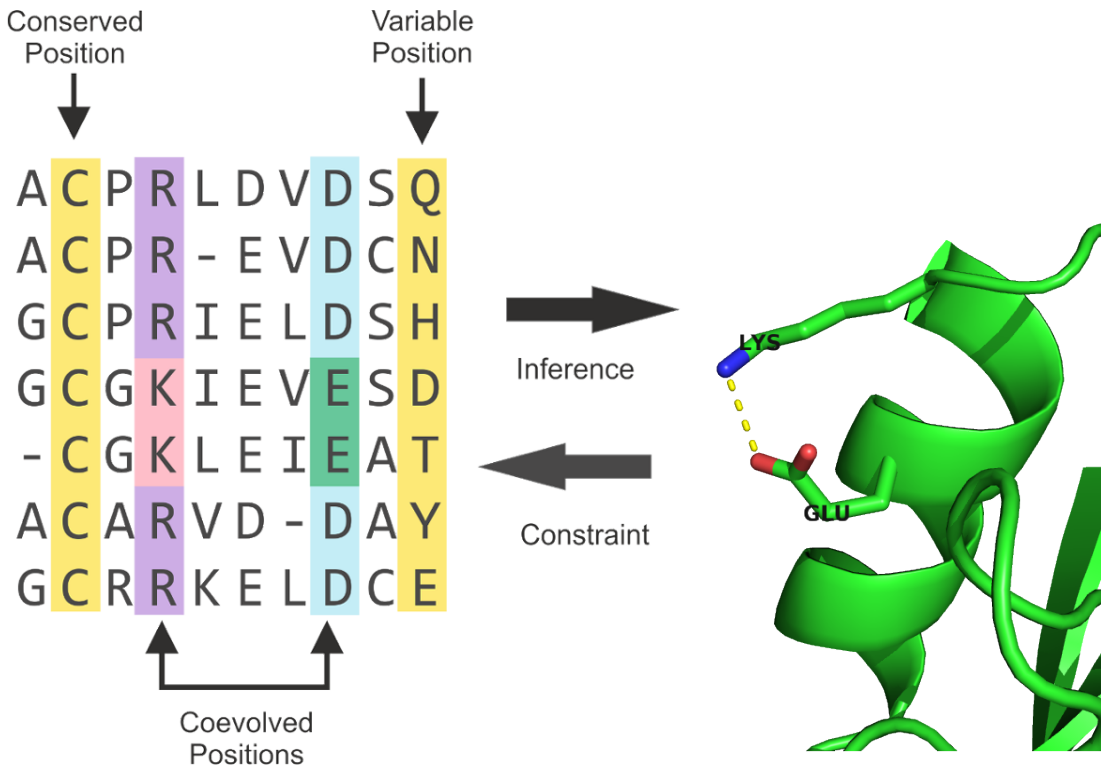
Tools	Database
ClustalW [Larkin et al. 07]	Pfam `pfam.xfam.org`

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

Structure is determined by amino acid interactions

Preserving the function: coevolution of residues

As protein function is vital, **evolution selects mutations preserving structures**.
Leading to **compensatory** mutations:

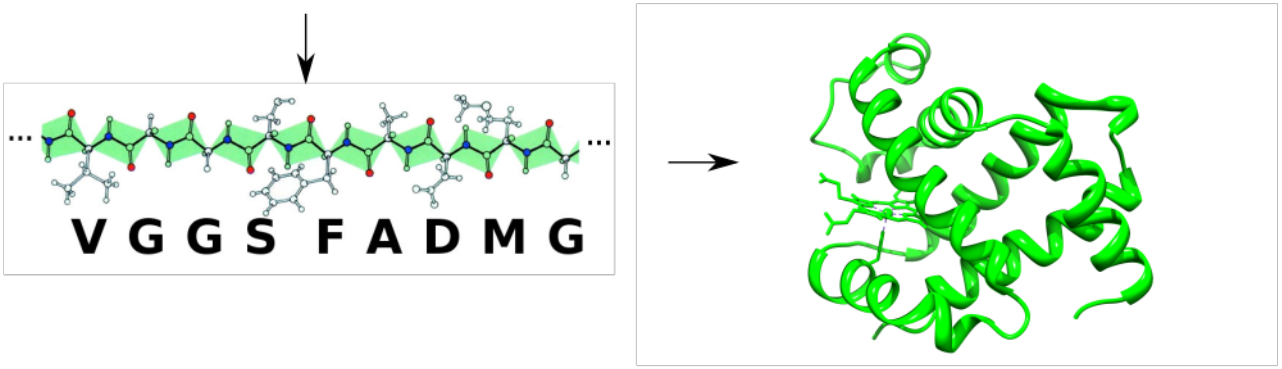


Computers and protein structure prediction

(cheap)

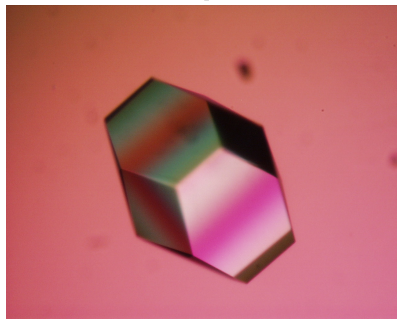
↓

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

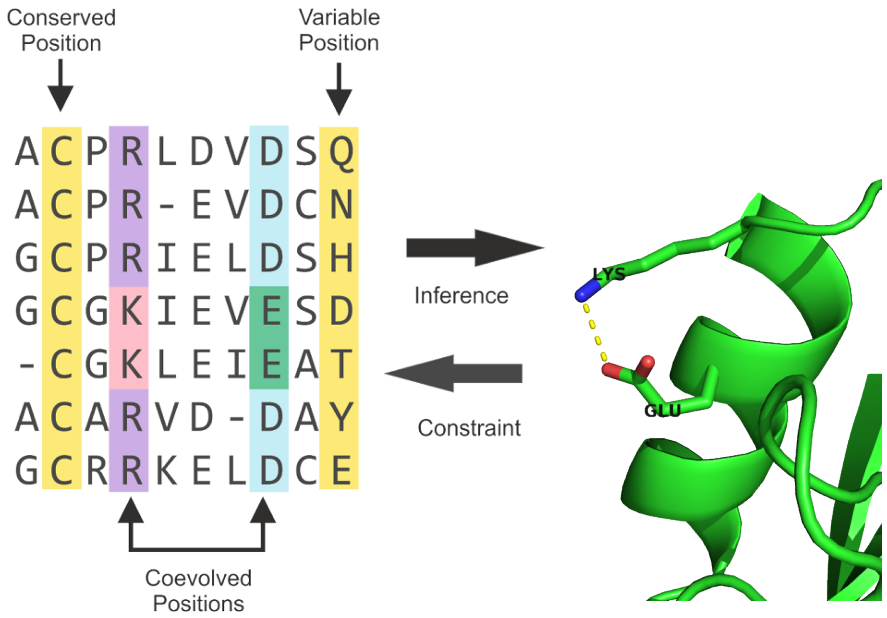


← (expensive)

Structure determined by X-Rays through a cristal of proteins



A simple approach for protein structure prediction



- Build or get multiple manio acid sequence alignments (**e.g.** in Pfam database)
- Infer what are the position in contact using machine learning

We'll check that in another module (January 2022)

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/`
- Covid-19 sequences `https://www.covid19dataportal.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
- Fulfillment of the call
- Trustworthiness in the description of the approach

Projects

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

- Clarity
- Fulfillment 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

Projects

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

- Clarity
- Fulfillment 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

Bonus: download closely related SARS-Cov2 sequencing data, and run your tool on it for recovering the SNPs (procedure to be explained in the report).

The TATFAR
waits for
interesting answers to its call!