Computational biology

Co-evolution to predict protein structures

Clovis Galiez

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

January 11, 2022

Topic: protein structure prediction

Topic of this series of lecture

- Introductory lectures about protein structure prediction
- Project (hands-on) of *de novo* protein structure prediction

Evaluation:

- Project-based (3-5 pages report)
- Code and tests
- Clarity, trustworthiness of the tests and method will be the most important criteria for the evaluation.

Today's outline: from gene sequence to protein structure

- Reminder about the central dogma
  - Genomes, genes, proteins
- Protein structure prediction methods
- Focus on *de novo* from covariation
  - Sequence evolution and selective pressure
  - Multiple sequence alignment
  - Residue co-variation

Context

Global pandemic of Salmonella.

- A team of biologists managed to identify two strains: one highly resistant to tetracyclin, one not.
- A team of computational biologists managed to identify the mutations between the two strains: it affects the tetR gene.

Context

Global pandemic of Salmonella.

- A team of biologists managed to identify two strains: one highly resistant to tetracyclin, one not.
- A team of computational biologists managed to identify the mutations between the two strains: it affects the tetR gene.

We want here to model the 3D structure of the protein associated to this gene **without** relying on X-Ray cristallography (too much time-consuming).

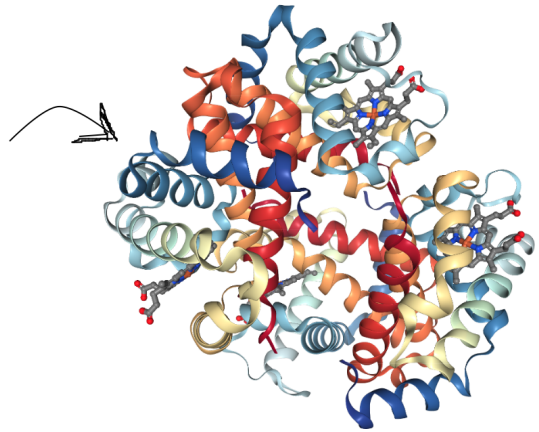
Context

Global pandemic of Salmonella.

- A team of biologists managed to identify two strains: one highly resistant to tetracyclin, one not.
- A team of computational biologists managed to identify the mutations between the two strains: it affects the tetR gene.

We want here to model the 3D structure of the protein associated to this gene **without** relying on X-Ray cristallography (too much time-consuming).

```
>1A3N:A|PDBID|CHAIN|SEQUENCE
VLSPADKTNVKAAWGKVGAHAGEYGAEALER
MFLSFPTTKTYFPHFDLSHGSAQVKGHGKKV
ADALTNAVAHVDDMPNALSALSDLHAHKLRV
DPVNFKLLSHCLLVTLAAHLPAEFTPAVHAS
LDKFLASVSTVLTSKYR
```



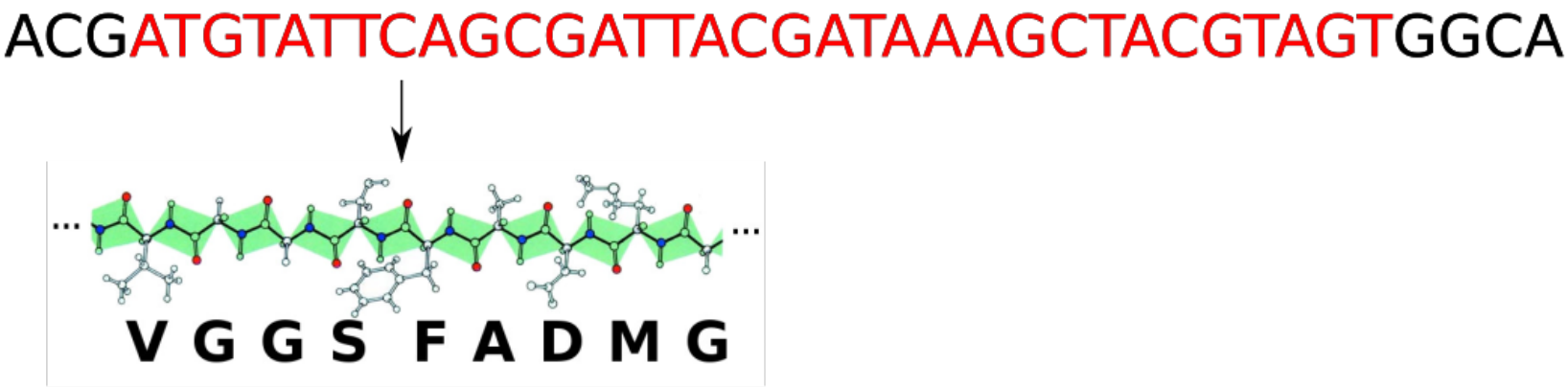
Some background

From genome to function, the very big picture

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

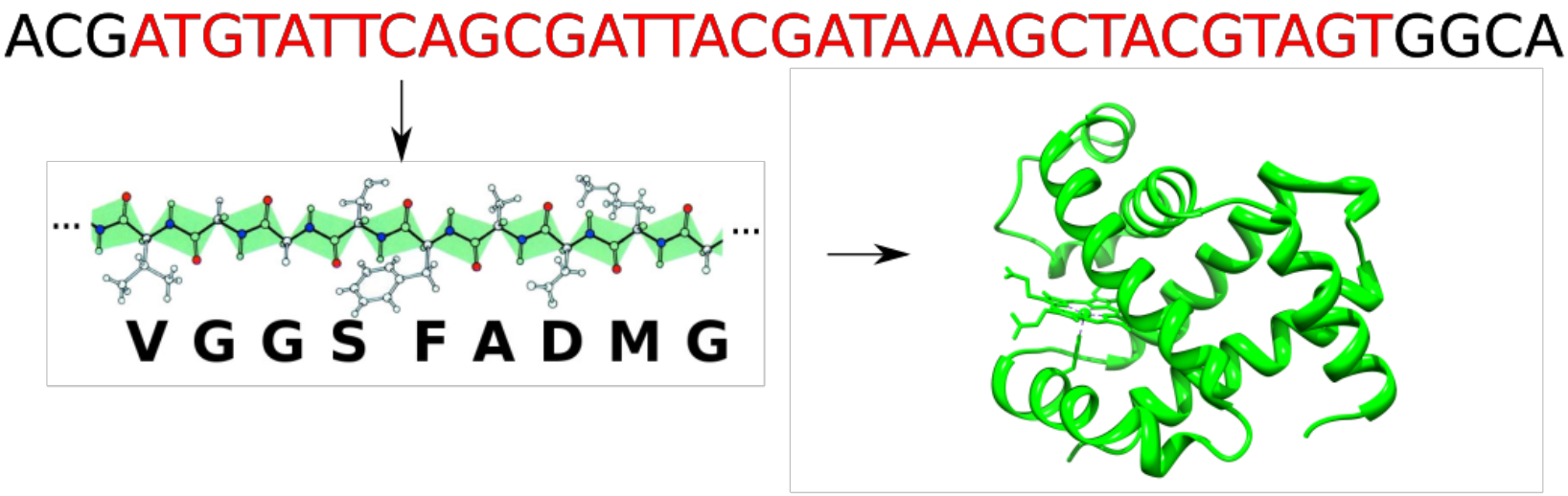
On a genome (∼5Mbp), specific motifs define begining and end of a gene

From genome to function, the very big picture



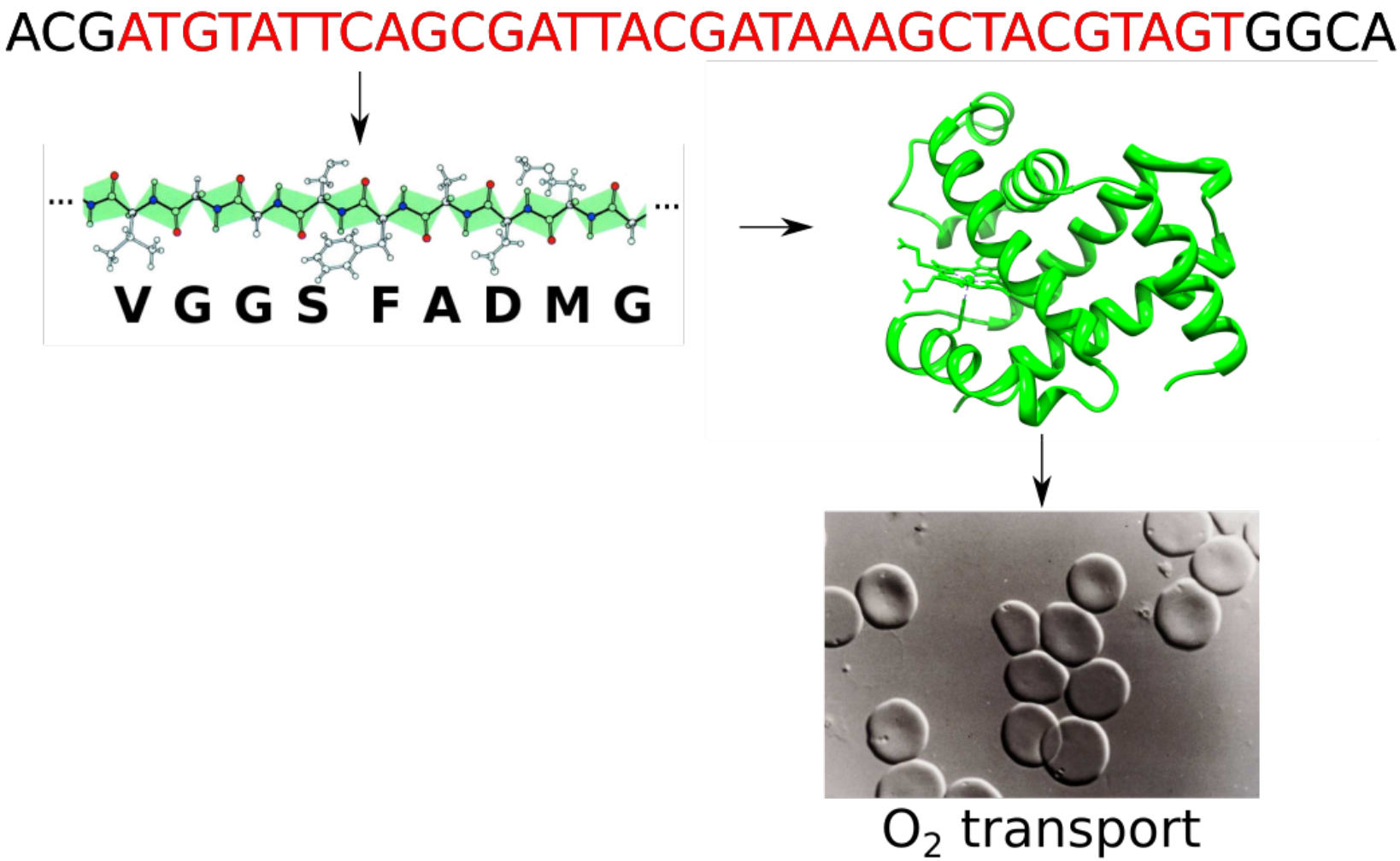
Transcription + *translation*, to form a chain of amino acids (∼300-3000AA)

From genome to function, the very big picture



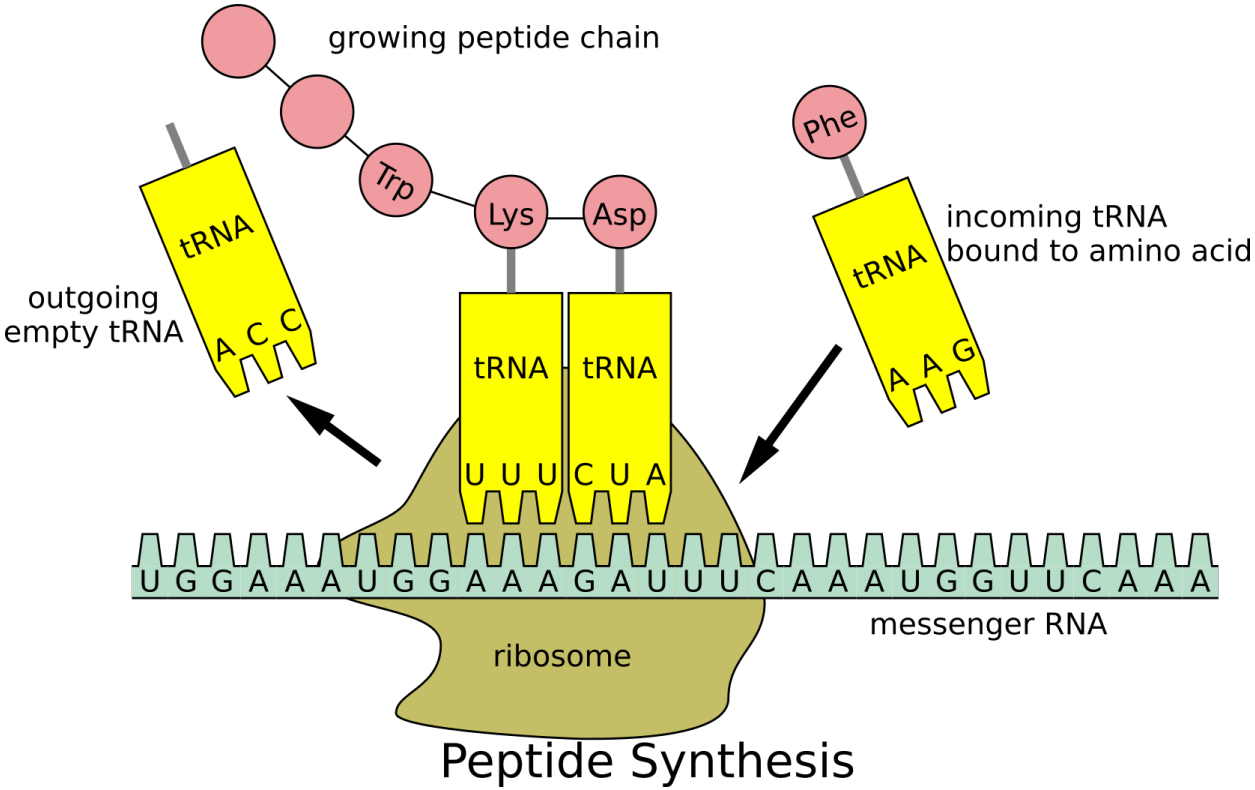
Protein folding under pysico-chemical interactions, diameter $\sim$ few nanometers

From genome to function, the very big picture



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

Zoom: genes to proteins



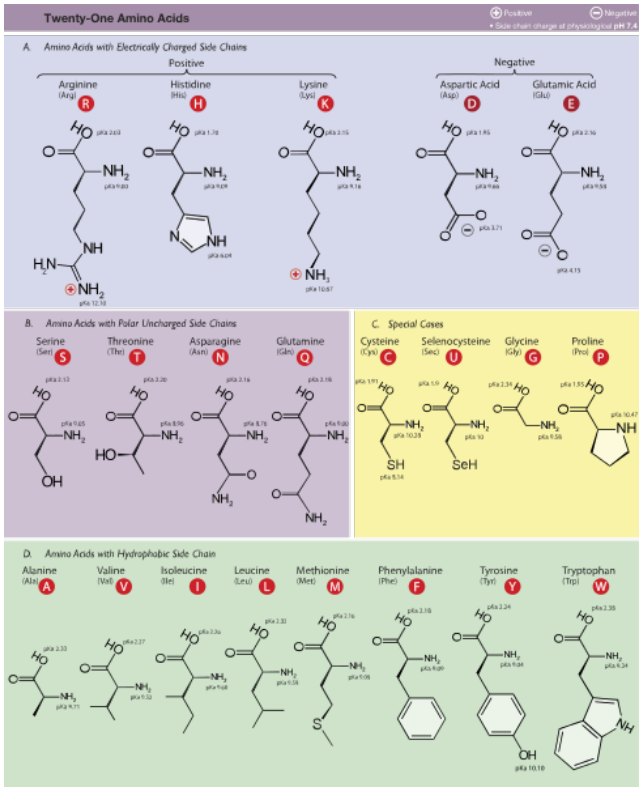
RNA codon table

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

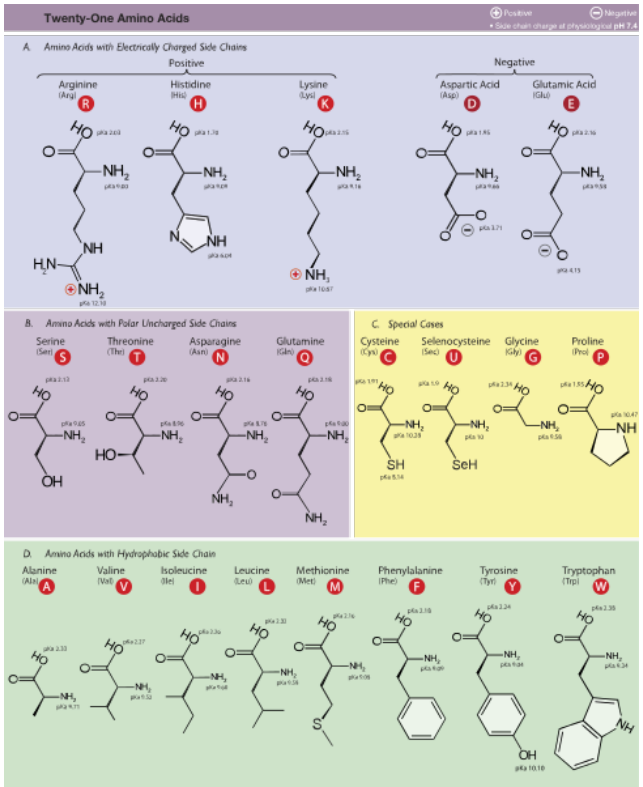
Amino Acids

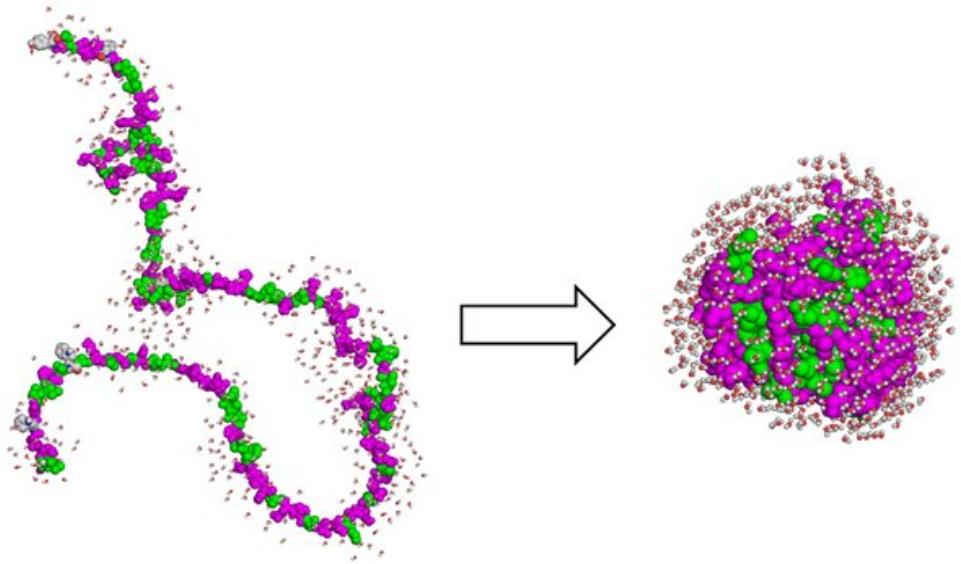
Ala: Alanine
Arg: Arginine
Asn: Asparagine
Asp:Aspartic acid
Cys:Cysteine
Gln: Glutamine
Glu: Glutamic acid
Gly: Glycine
His: Histidine
Ile: Isoleucine
Leu: Leucine
Lys: Lysine
Met: Methionine
Phe: Phenylalanine
Pro: Proline
Ser: Serine
Thr: Threonine
Trp: Tryptophane
Tyr: Tyrosisne
Val: Valine

Primary to tertiary protein structure

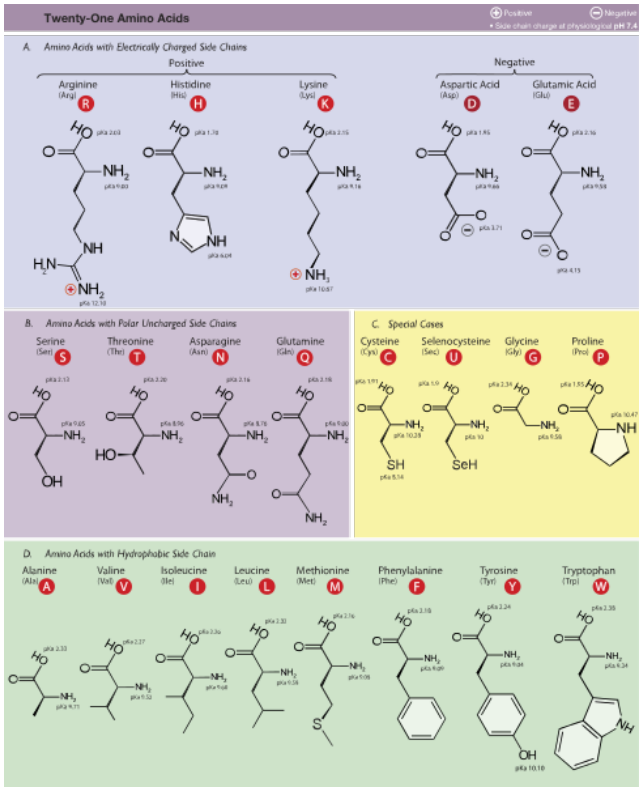


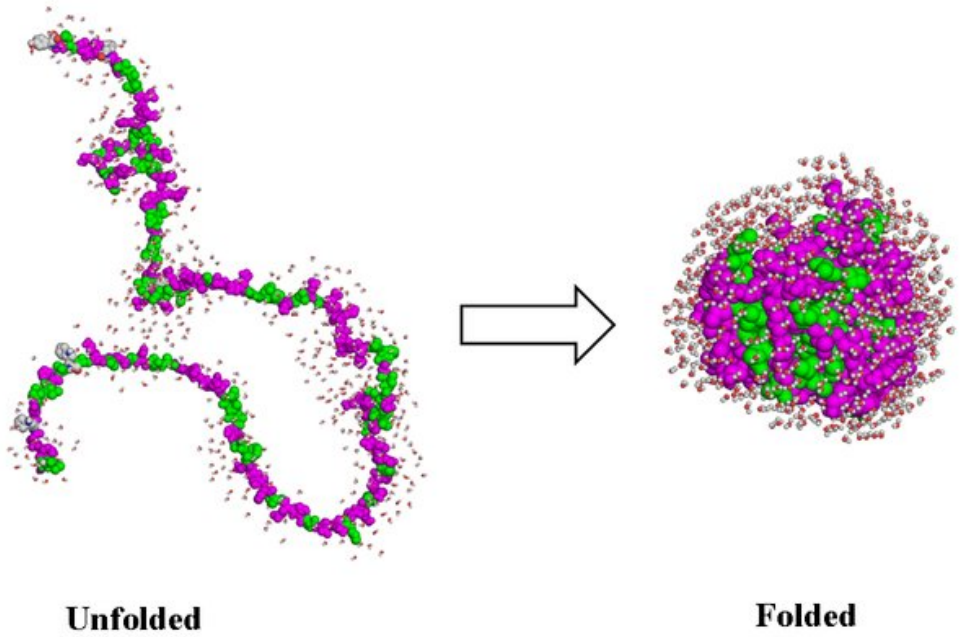
Primary to tertiary protein structure





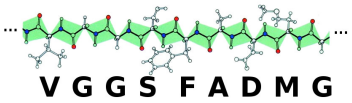
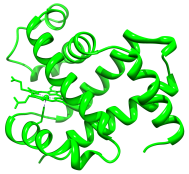
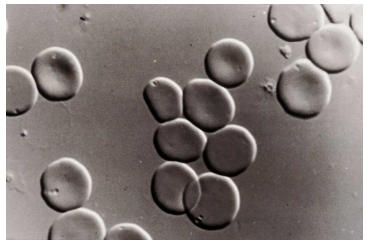
Primary to tertiary protein structure



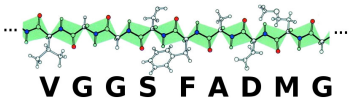
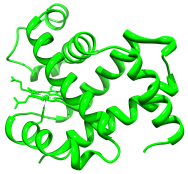
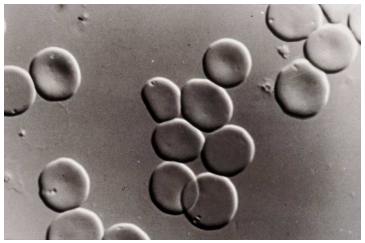


The amino acid sequence is called the **primary** structure of the protein and the final structure is called the **tertiary** protein structure.

Data at every steps

Nucleic seq.	Amino acid seq.	Protein	Function
..ATTGTCGAAC..			

Data at every steps

Nucleic seq.	Amino acid seq.	Protein	Function
..ATTGTCGAAC..			
NCBI	UniProt	RCSB PDB PROTEIN DATA BANK	
`ncbi.nlm.nih.gov`	`uniprot.org`	`rcsb.org`	`ebi.ac.uk/interpro`

How to predict gene function?

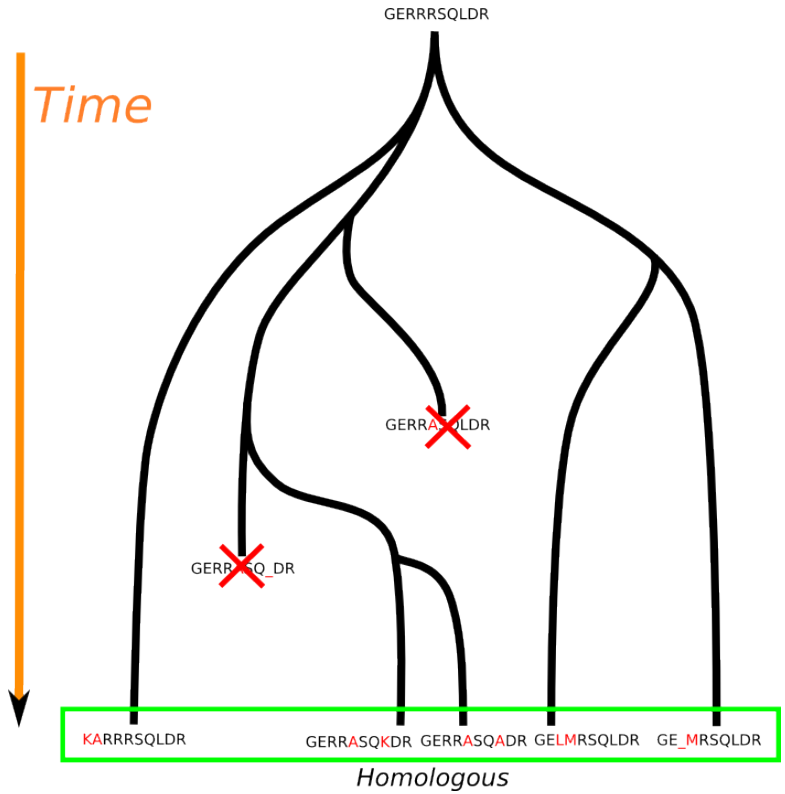
Some gene functions have been previously identified by biologists.

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.

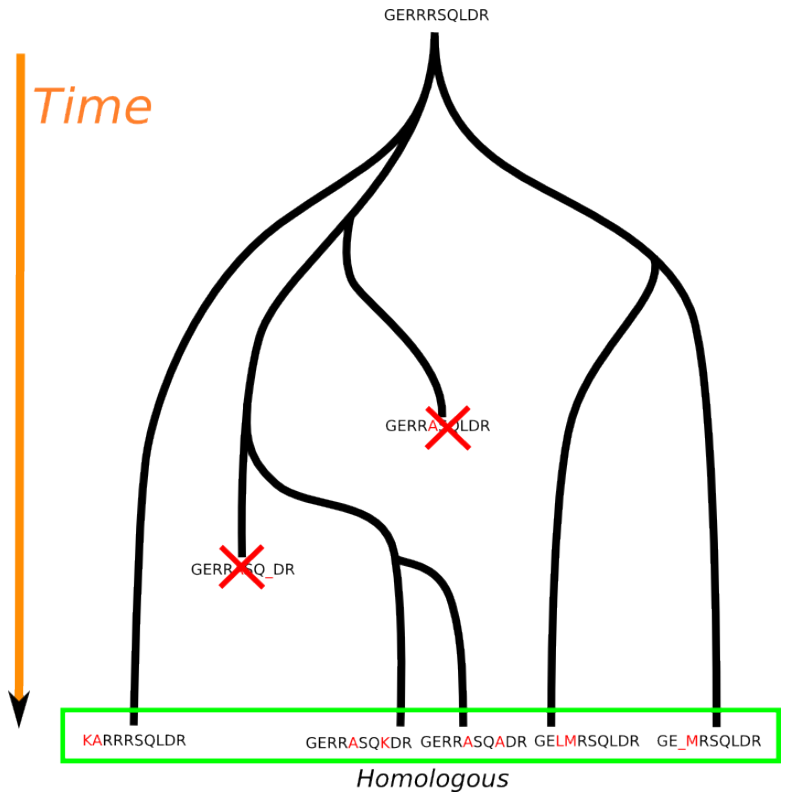
When having an unknown sequence, how can you guess its function?



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.

Get insights from the structure

What if no homologous sequences or if they have no functional annotation?

Get insights from the structure

What if no homologous sequences or if they have no functional annotation?

Look at the structure!

Get insights from the structure

What if no homologous sequences or if they have no functional annotation?

Look at the structure!

The bad news is...

Get insights from the structure

What if no homologous sequences or if they have no functional annotation?

Look at the structure!

The bad news is...

Ok, but most of the time, when we have the structure, we have the function :-/

Get insights from the structure

What if no homologous sequences or if they have no functional annotation?

Look at the structure!

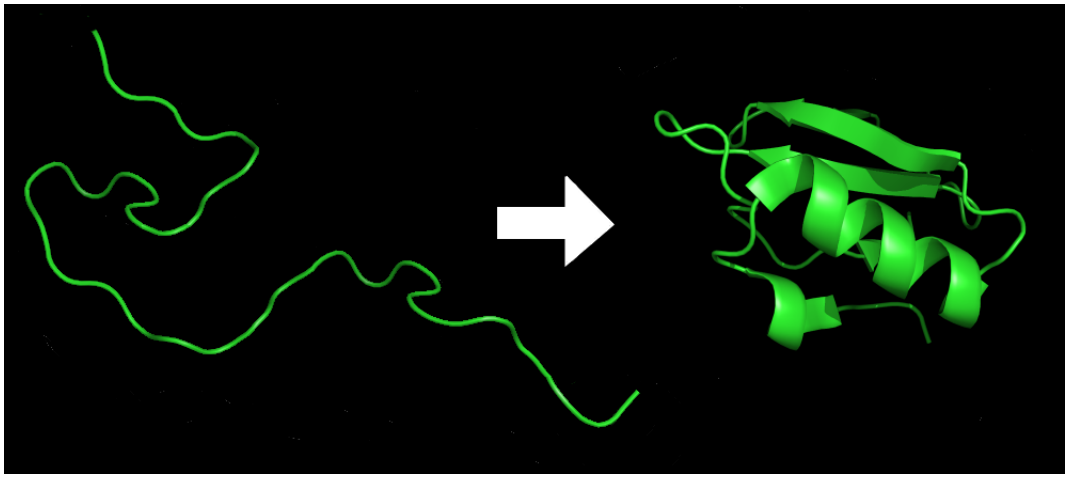
The bad news is...

Ok, but most of the time, when we have the structure, we have the function :-/

... so have to predict the structure

Protein folding

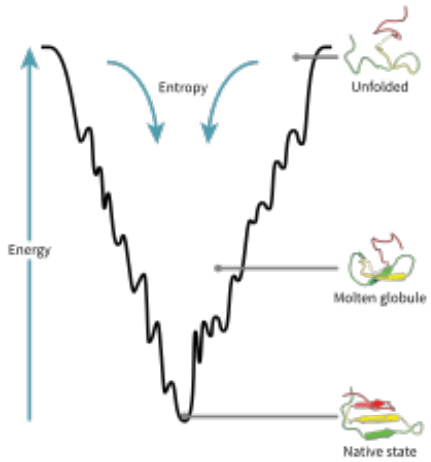
Protein folds to their stable structure in milliseconds [Karplus, 97] under interactions between their amino acids, as well with the environment (mostly water).



Given the large number of possible conformations, the energy landscape cannot be flat (*aka* Levinthal's paradox), and it hints to be a problem computationally tractable.

Not a single perfect model

Several models have been proposed for the folding mechanism, like the funnel energy landscape (source Wikipedia):



No model gives full satisfaction on all aspect of folding (folding times, physically realistic, computationally tractable, etc.)

Predicting the structure

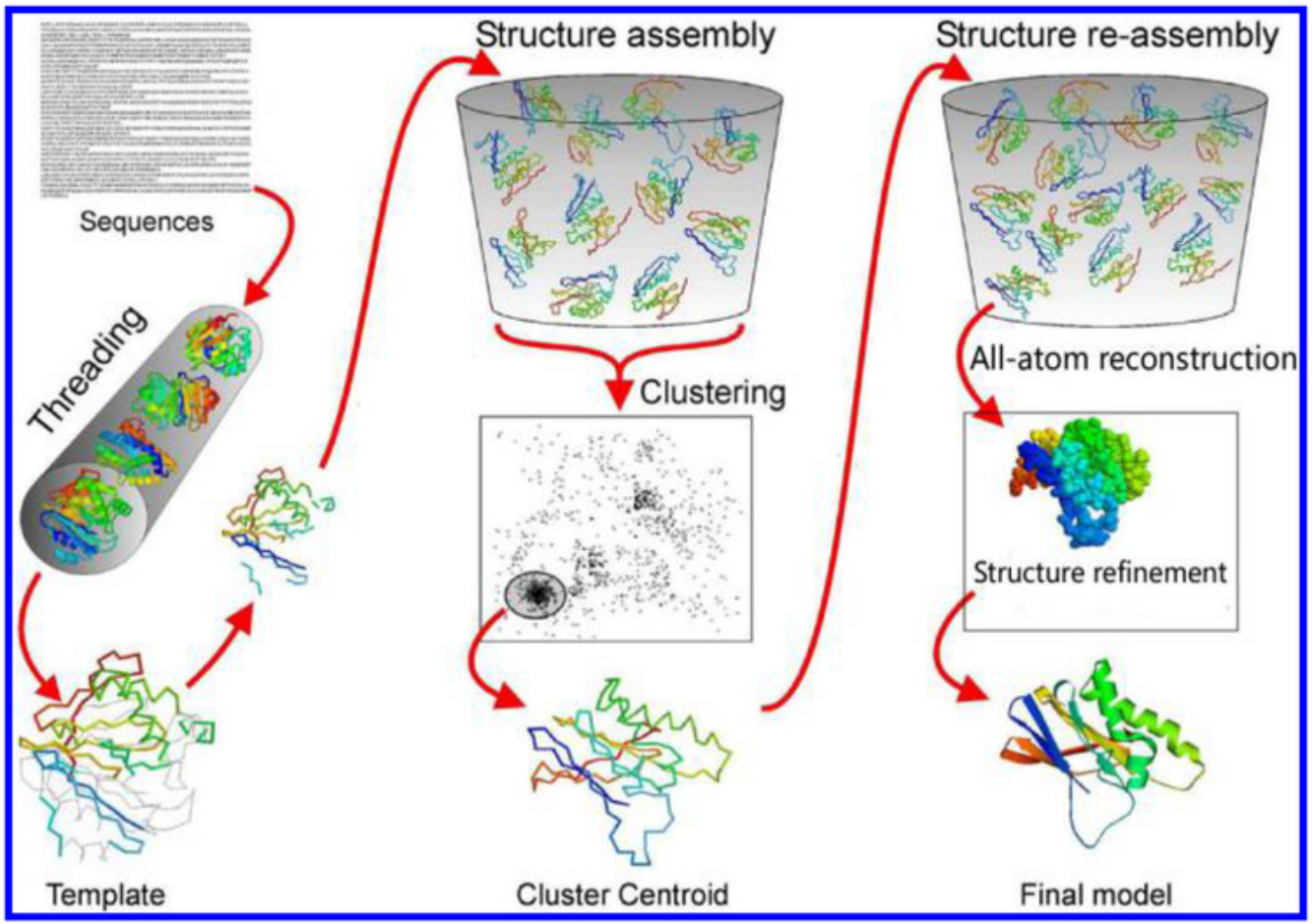
Several types of approaches:

- Molecular dynamics (much costly!): simulate force field between amino acids
- Fragment assembly: Rosetta [Baker Lab 2004]
- Template-based modelling: use known structures with similar sequences [Zhang Lab 2010]
- Coevolution based: [Weigt et al 09], [Jones et al. 2012]
- Hybrid+machine learning methods: AlphaFold2 [Jumper et al. 2021]

The last mentioned method has been a *revolution* for science in 2021. It combines **threading** and **co-evolution** based methods in a **modern AI framework** with attention mechanisms.

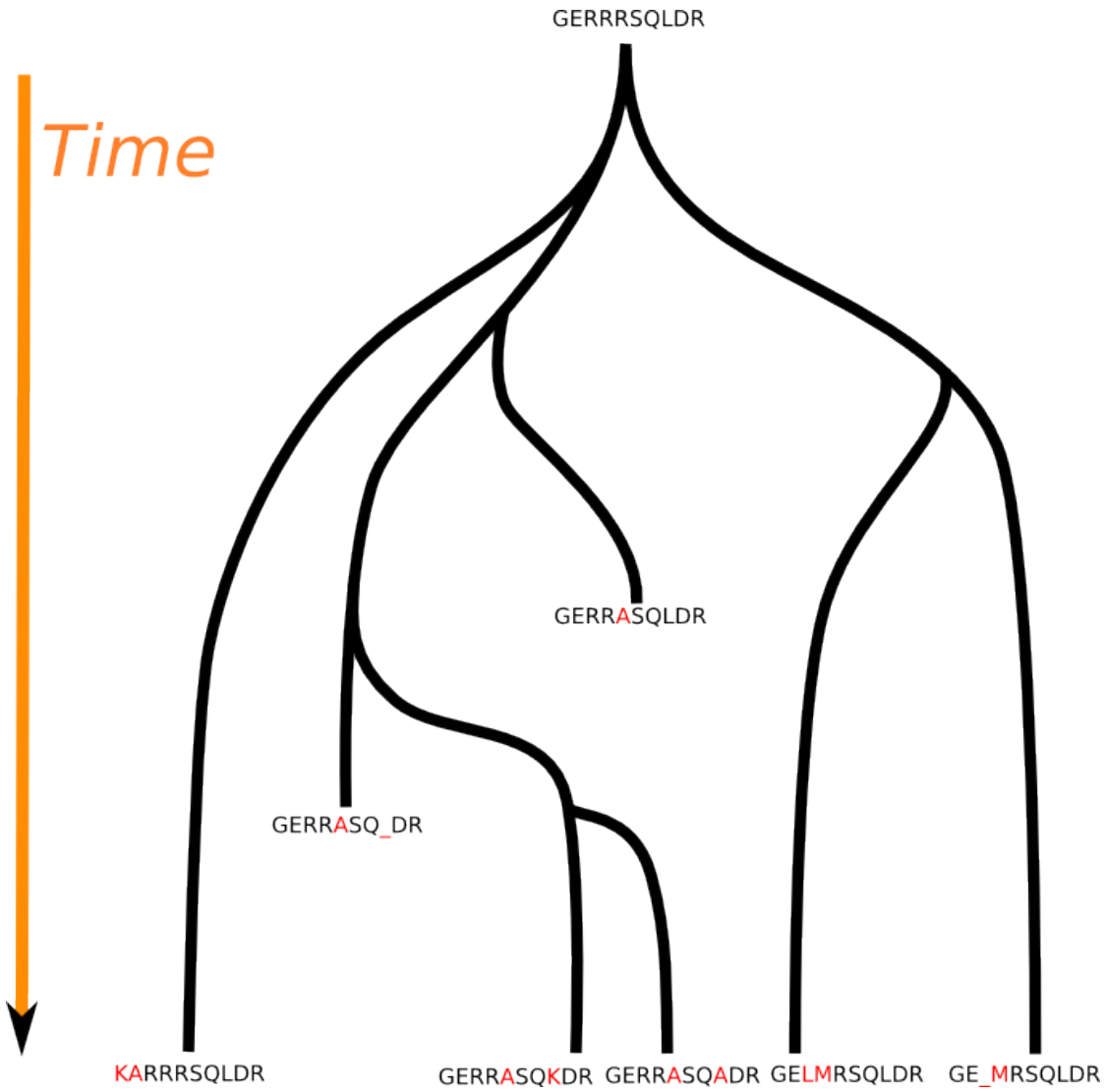
Threading

I-TASSER [Roy et al. 10]:

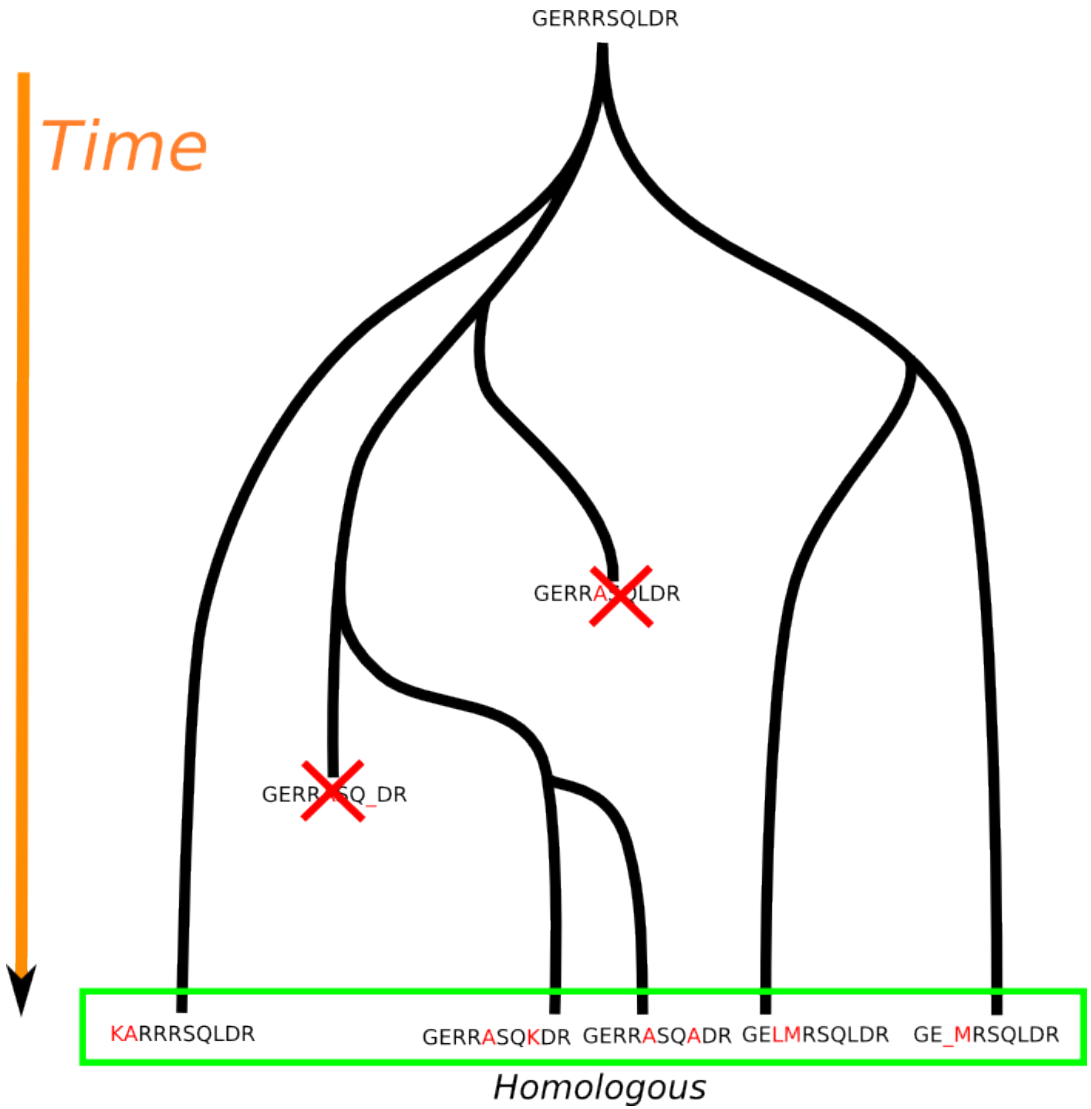


Covariation for structure prediction

Sequence evolution



Sequence evolution



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):

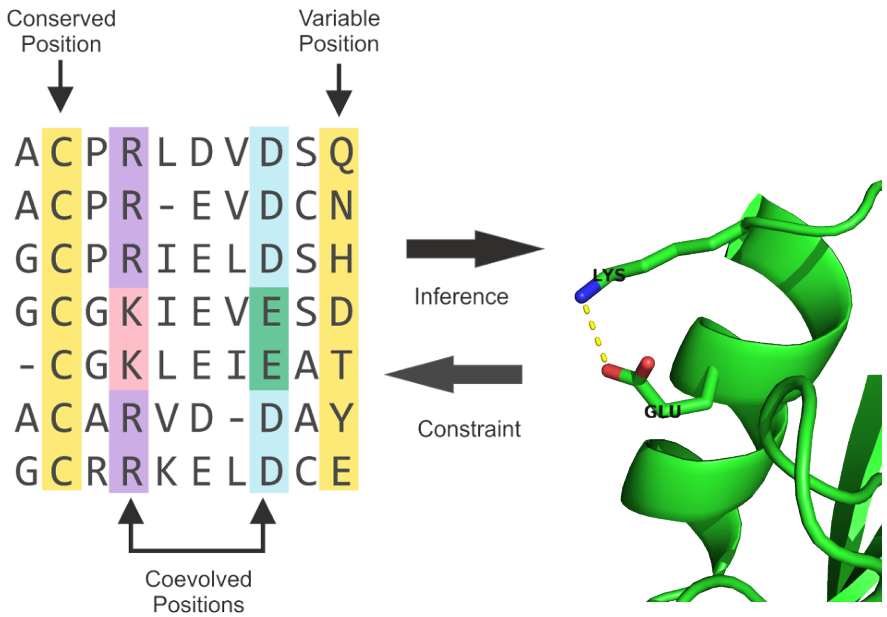
Tools	Database
ClustalW [Larkin et al. 07]	Pfam `pfam.xfam.org`

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

Structure is determined by amino acid interactions

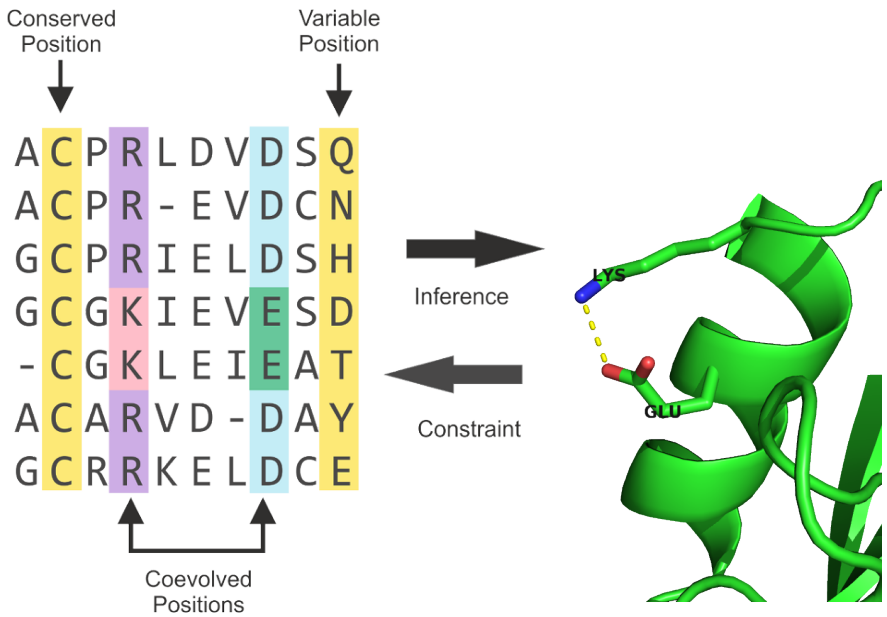
Preserving the function: coevolution of residues

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



Preserving the function: coevolution of residues

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



To predict a structure:

- Build or get multiple amino acid sequence alignments
- Infer what are the position in contact using machine learning

Conservation vs. co-evolution

How to measure co-variation?

Conservation vs. co-evolution

How to measure co-variation? A standard approach is to measure it through Mutual Information:

$$MI(i,j) = \sum_{a,b} p(x_i = a, x_j = b) \log \frac{p(x_i=a, x_j=b)}{p(x_i=a)\, p(x_j=b)}$$

Where

- x_i is the amino acid at position i
- $p(x_i = a)$ is estimated in the MSA by $\frac{\#\text{sequences having "a" at position } i}{N}$
- N the number of sequences in the MSA
- $p(x_i = a, x_j = b)$ is estimated in the MSA by $\frac{\#\text{sequence having "a" at } i \text{ and "b" at } j}{N}$

In practice you need $N > 1,000$ to have reasonable estimation of $p(x_i = a, x_j = b)$.

Issue: indirect dependencies

The later approaches suffer from indirect dependencies.
Proposed solutions:

Issue: indirect dependencies

The later approaches suffer from indirect dependencies.
Proposed solutions:

- Direct Coupling Analysis: infer J, h by maximizing the likelihood $P(x|J,h) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(x_i, x_j) + \sum_{i=1}^{N} h_i(x_i)\right)$
- Sparse Inverse Covariance matrix: the precision matrix ($\Lambda = \Sigma^{-1}$) represents the partial correlations ($\rho_{x_i x_j|\text{other positions}} = -\frac{\Lambda_{ij}}{\sqrt{\Lambda_{ii}\Lambda_{jj}}}$), infer it with a Lasso regularization.

Toward machine learning

That was the state-of-the-art until ≈ 2018.

Toward machine learning

That was the state-of-the-art until ≈ 2018.
Critics for the previous approaches:

- The link: covariation $\rightarrow$ contact in 3D may be suboptimal
- There are a lot of parameters to infer (at least 20x20 amino acids x length of the sequence^2) $\rightarrow$ need for a lot of sequences in the MSA

Machine learning models to the rescue to cope with these 2 issues.

CASP competition

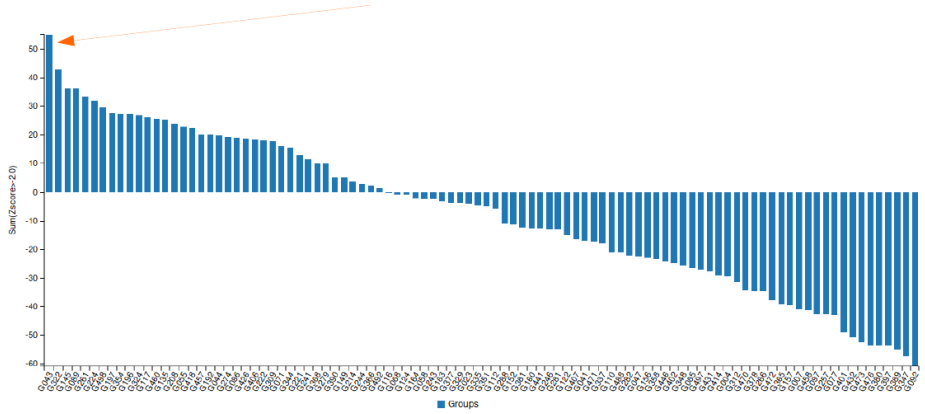
Blind competition. Simple principle:
- a sequence is given
- have to predict the structure.

Prior to 2018 it used to be (pseudo) physical models that where best performing.

CASP13 (2018)

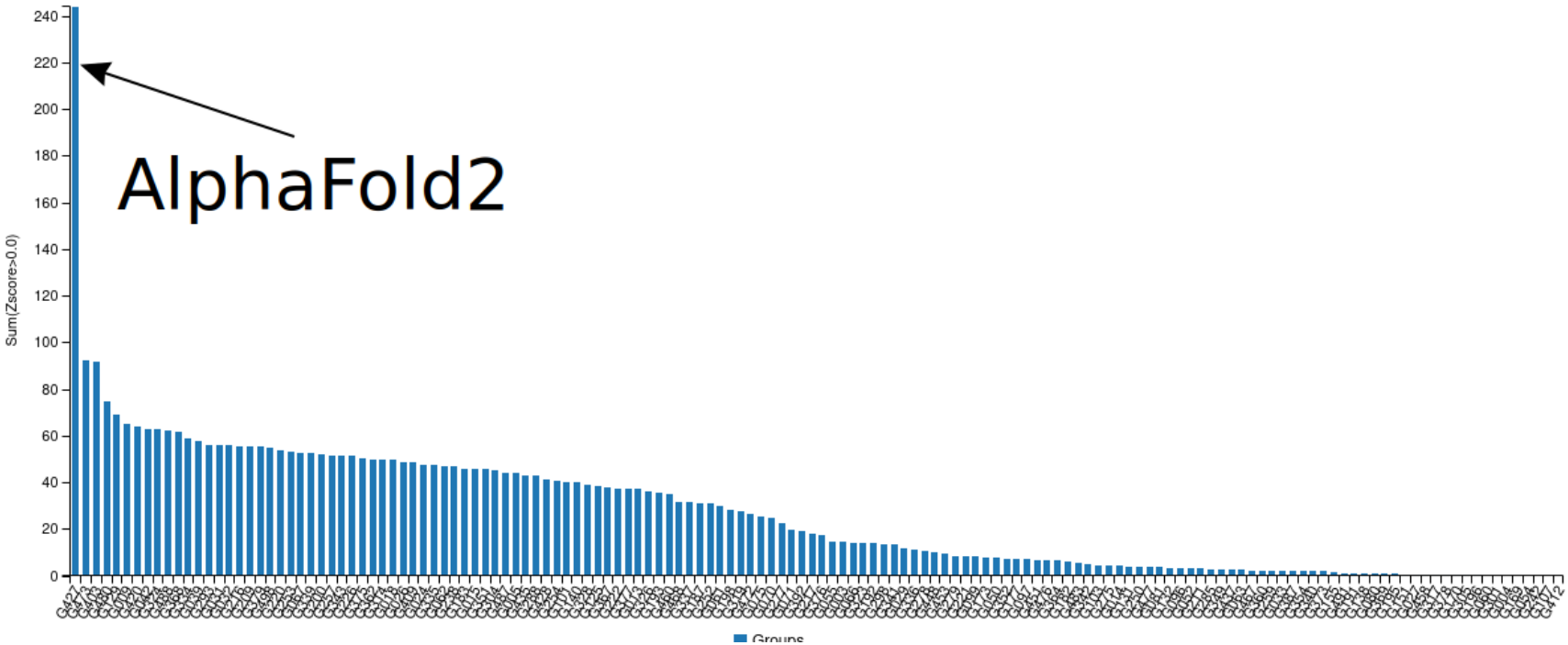
AI wins the challenge for the first time.

Google's DeepMind



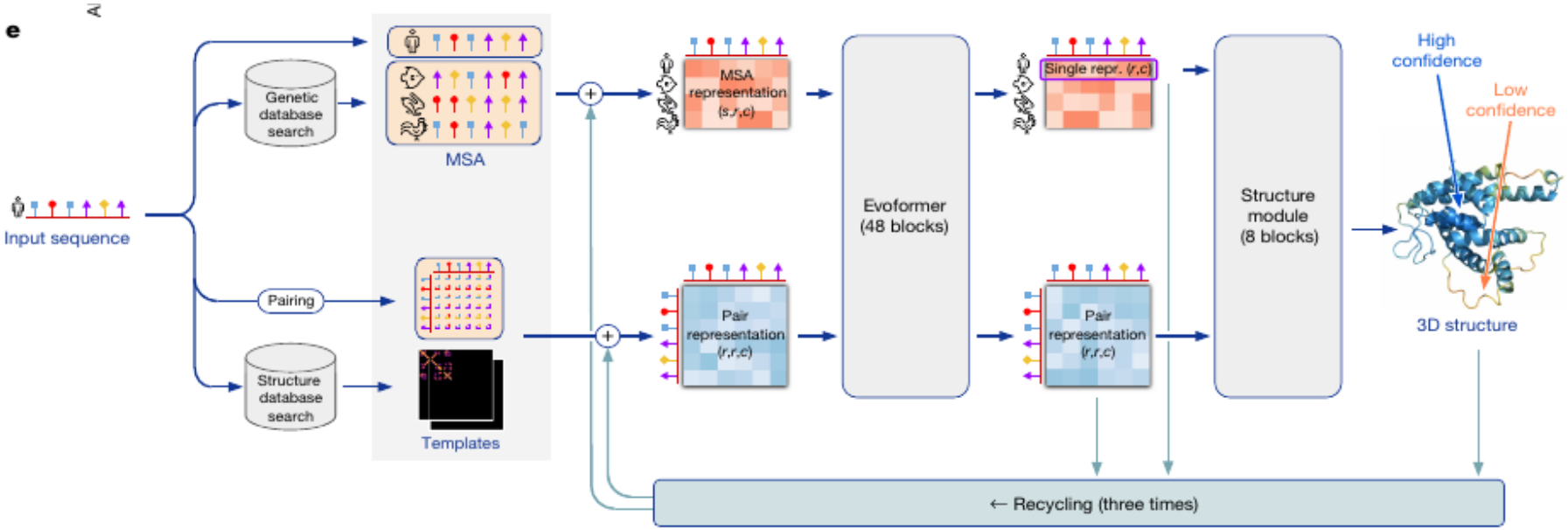
CASP14 (2020)

"The big leap forward"

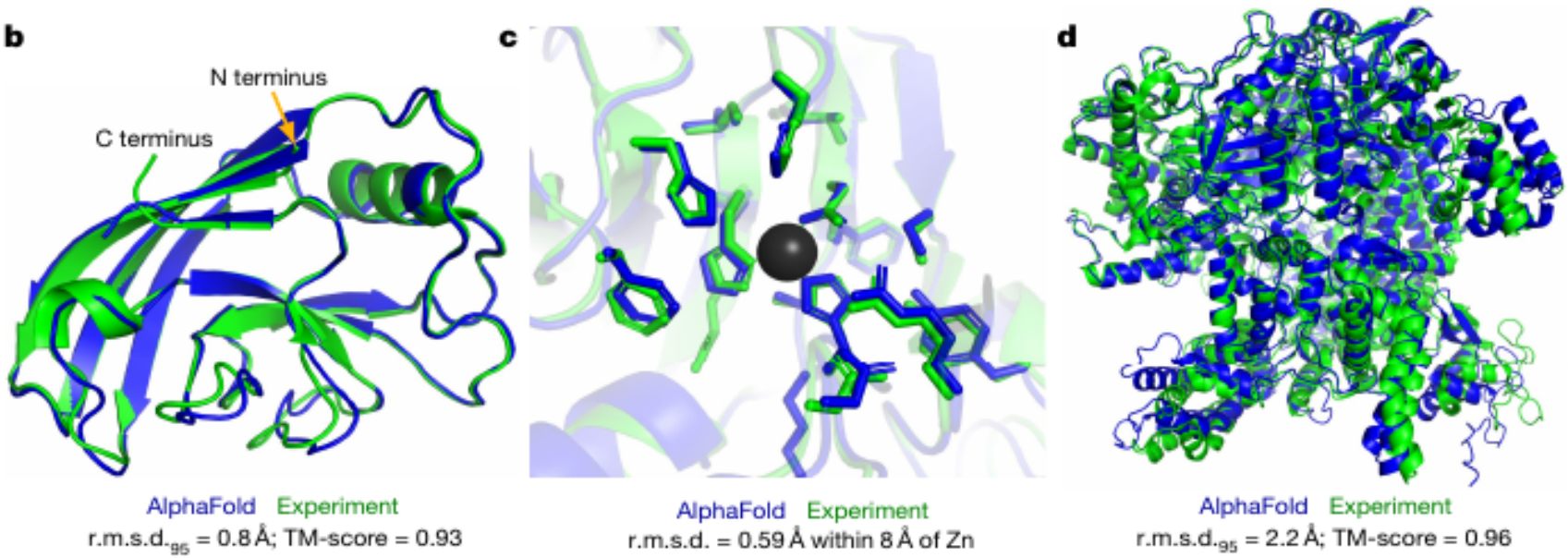


Alphafold2: attention-based learning on protein sequence alignments [Casp14.]
Nature's article.

AlphaFold2 architecture



AlphaFold2 results



Information for your project

Prediction by covariation

Proposed approach

We propose to implement a simple *de novo* protein structure prediction specific to the case of the tetR protein (small and many available sequences) using sequence covariations.

Tools and databases

Git of the project:
`https://gitlab.ensimag.fr/galiezc/prot-struct-pred`

- BioPython library
- Pfam alignments `http://pfam.xfam.org/`
- Protein structure PDB `https://www.rcsb.org/`
- Search protein sequence
  `https://www.ebi.ac.uk/Tools/sss/ncbiblast/`
- Visualize protein structure: PyMol, Chimera
- Contact Map visualizer:
  `https://pymolwiki.org/index.php/Contact_map_visualizer`
- From contact map to structure: FT-COMAR (see git)

Let's solve this structure!

Sequence comparison

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.

More on mutual information

Over-prediction at entropic position

When applying the rule

$$MI(i,j) > \tau \Rightarrow \text{contact between } i \text{ and } j$$

some positions predict too many contacts, often position with high entropy. Several corrections can be applied[1].

In your project

You can try using the simple correction:

$$MI'(i,j) = MI(i,j) - \frac{1}{N}\sum_k (MI(k,j) + MI(i,k))$$

and fix a τ to predict a contact as soon as:

$$MI'(i,j) > \tau$$

[1] See `https://doi.org/10.1093/bioinformatics/bti671`