Computational biology

Co-evolution to predict protein structures

Clovis Galiez

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

October 19, 2020

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, 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

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

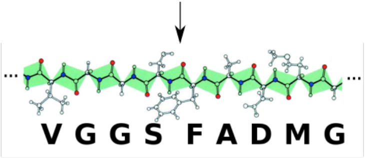
From genome to function, the very big picture

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

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

From genome to function, the very big picture

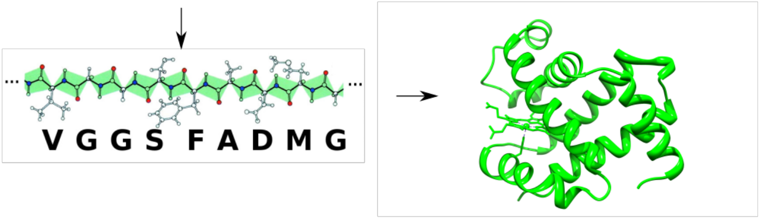
ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA



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

From genome to function, the very big picture

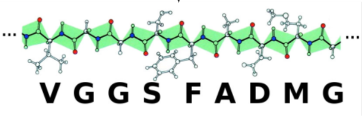
ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

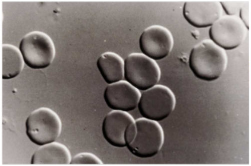


Protein folding under pysico-chemical interactions, diameter ∼ few nanometers

From genome to function, the very big picture

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA





O_2 transport

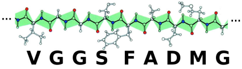
Protein endowed with a function (biochemical reactions, transport, etc.)

Data at every steps

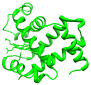
Nucleic seq.

..ATTGTCGATGAC..

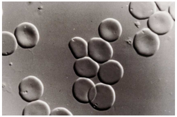
Amino acid seq.



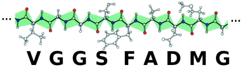
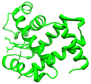
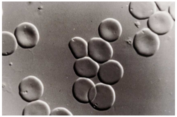
Protein



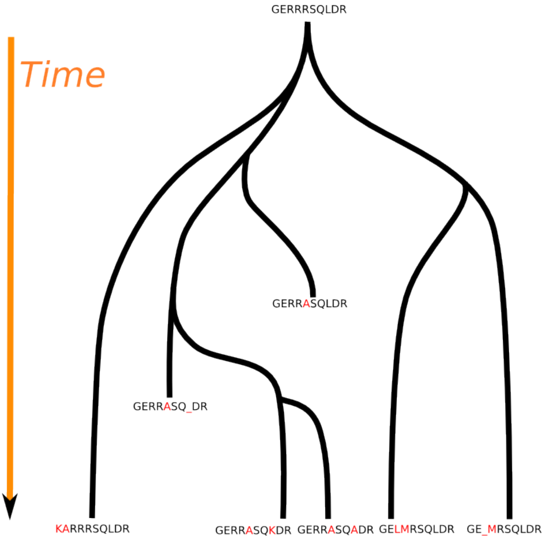
Function



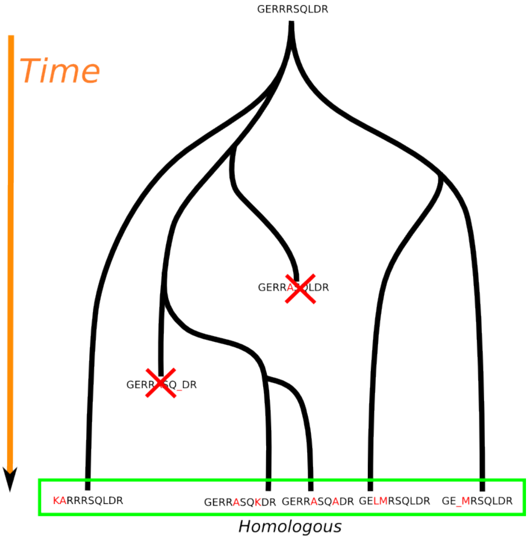
Data at every steps

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



Protein evolution through mutations



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.

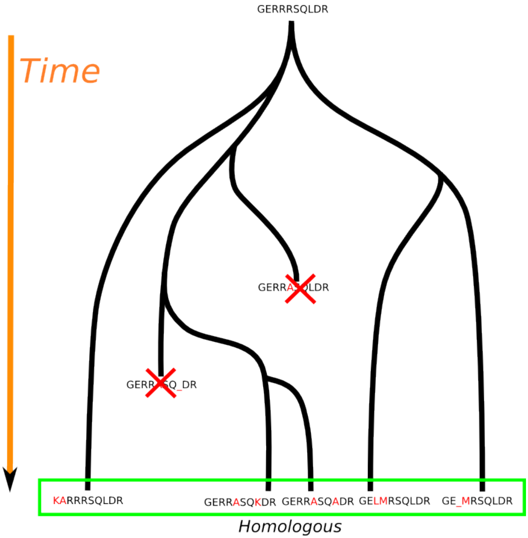
Big data: need for heuristic

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
- ...

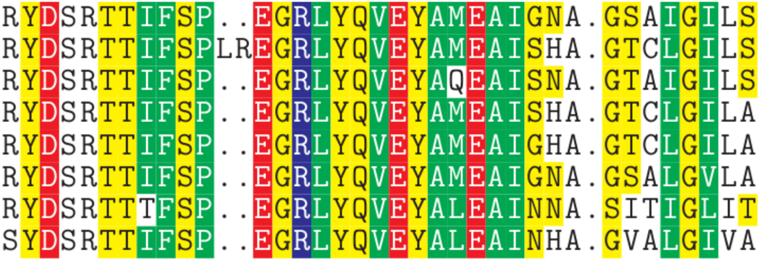
Heuristics are mostly based on similar k-mers, and efficiently filtering through hash tables.

Sequence conservation



Sequence conservation

Aligning the sequences (MSA, multiple sequence alignment):

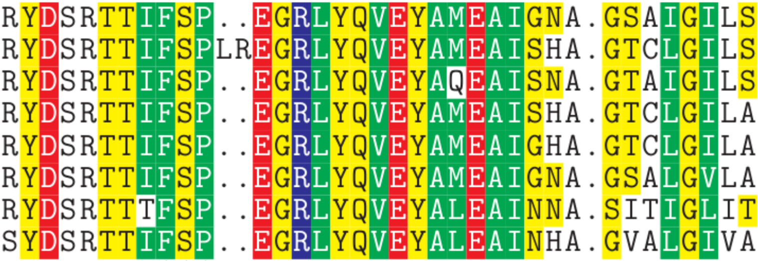


```
RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSAIGILS
RYDSRTTIFSPLREGRLYQVEYAMEAISHA.GTCLGILS
RYDSRTTIFSP..EGRLYQVEYAQEAISNA.GTAIGILS
RYDSRTTIFSP..EGRLYQVEYAMEAISHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEAIGHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSALGVLA
RYDSRTTTFSP..EGRLYQVEYALEAINNA.SITIGLIT
SYDSRTTIFSP..EGRLYQVEYALEAINHA.GVALGIVA
```

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

Sequence conservation

Aligning the sequences (MSA, multiple sequence alignment):



```
RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSAIGILS
RYDSRTTIFSPLREGRLYQVEYAMEAISHA.GTCLGILS
RYDSRTTIFSP..EGRLYQVEYAQEAISNA.GTAIGILS
RYDSRTTIFSP..EGRLYQVEYAMEAISHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEAIGHA.GTCLGILA
RYDSRTTIFSP..EGRLYQVEYAMEAIGNA.GSALGVLA
RYDSRTTTFSP..EGRLYQVEYALEAINNA.SITIGLIT
SYDSRTTIFSP..EGRLYQVEYALEAINHA.GVALGIVA
```

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

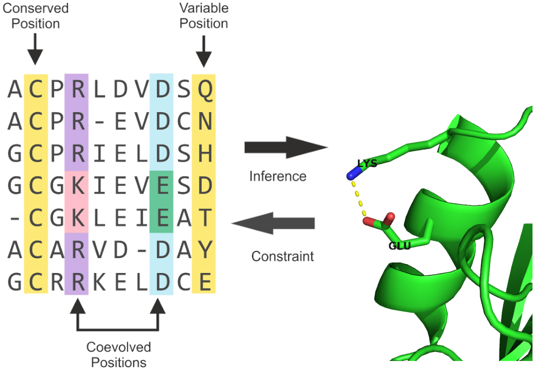
Why some positions are conserved, some other aren't?

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:



```
A C P R L D V D S Q
A C P R - E V D C N
G C P R I E L D S H
G C G K I E V E S D
- C G K L E I E A T
A C A R V D - D A Y
G C R R K E L D C E
```

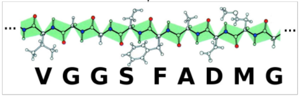
Computers and protein structure prediction

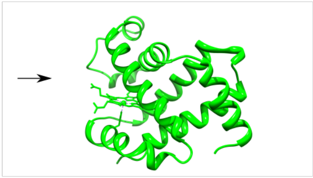
(cheap)

↓

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

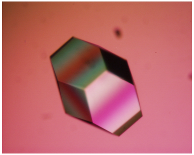
↓



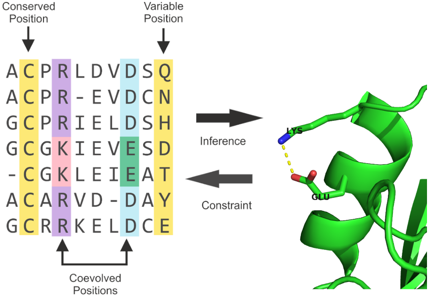


← (expensive)

Structure determined by X-Rays
through a cristal of proteins



A simple approach for protein structure prediction



```
A C P R L D V D S Q
A C P R - E V D C N
G C P R I E L D S H
G C G K I E V E S D
- C G K L E I E A T
A C A R V D - D A Y
G C R R K E L D C E
```

- Build or get multiple manio acid sequence alignments (**e.g.** in Pfam database)
- Quantify coevolution between positions in the sequence
- Infer what are the position in contact

What measure for co-evolution? Correlation would work?

Conservation vs. co-evolution

Conserved position carries no information in terms of co-evolution (entropy is zero).

Conservation vs. co-evolution

Conserved position carries no information in terms of co-evolution (entropy is zero). 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 paractice you need $N > 1,000$ to have reasonable estimation of $p(x_i = a, x_j = b)$.

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:

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

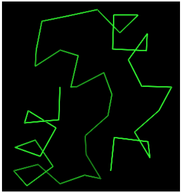
and fix a τ to predict a contact as soon as:

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

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

Hints concerning your Salmonella project

- The mutated gene you identified in the resistant bacteria have a large multiple sequence alignment in Pfam, search for its name in the Pfam browser.
- WP1.T2 consists in creating a tool taking as input an MSA in Fasta format and outputing a **contact matrix**[2].
- You can model the structure using FT-comar[3] software, and compare to the native structure [4] using RasMol or PyMol software:



[2] see the *cheatsheet* for details about the file format

[3] `clovisg.github.io/teaching/protein-structure-prediction/ft-comar.tgz`

[4] `clovisg.github.io/teaching/protein-structure-prediction/target.pdb`

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

You should send:

- a 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!