

Computational biology

Co-evolution to predict protein structures

Clovis Galiez



Grenoble

Statistiques pour les sciences du Vivant et de l'Homme

October 21, 2025

Topic: protein structure prediction

Topic of this series of lecture

- Introductory lectures about protein structure prediction
- Project: I picked up one protein I found in marine viruses with unidentified function, will you determine what function this protein is likely to carry out?

Evaluation:

- Project-based (3-5 pages report)
- Explanations of the methods used, with documented references 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

Some background

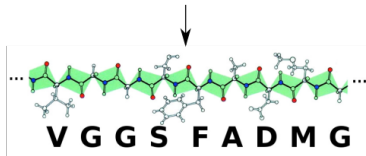
From genome to function, the very big picture

ACGATGTATTTCAGCGATTACGATAAAGCTACGTAGTGGCA

On a genome ($\sim 5\text{Mbp}$), specific motifs define beginning and end of a gene

From genome to function, the very big picture

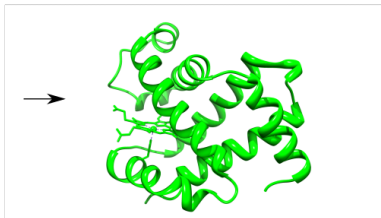
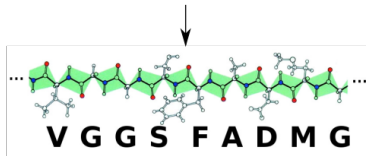
ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA



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

From genome to function, the very big picture

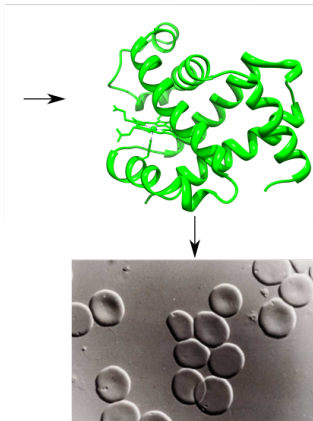
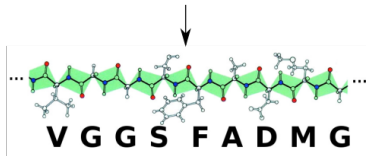
ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA



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

From genome to function, the very big picture

ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA



O₂ transport

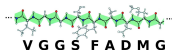
Protein endowed with a function (biochemical reactions, transport, etc.)

Data at every steps

Nucleic seq.

..ATTGTCGAAC..

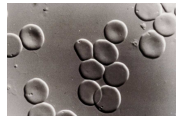
Amino acid seq.



Protein



Function



Data at every steps

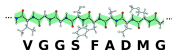
Nucleic seq.

..ATTGTCGAAC..



ncbi.nlm.nih.gov

Amino acid seq.



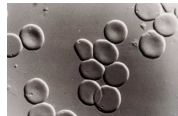
uniprot.org

Protein



rcsb.org

Function



ebi.ac.uk/interpro

How to predict gene function?

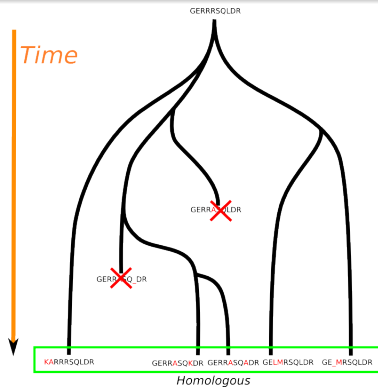
Some gene functions have been previously identified by biologists.

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.

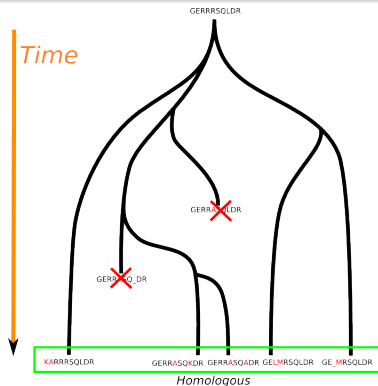
When having an unknown sequence, how can you guess its function?



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 **homologous** genes are already known.

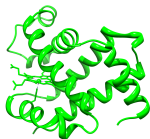
Get insights from the structure

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?

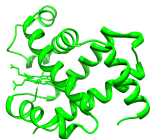
Look at the structure!



Get insights from the structure

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

Look at the structure!

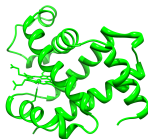


The bad news is...

Get insights from the structure

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

Look at the structure!



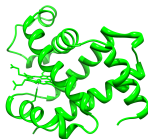
The bad news is...

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!

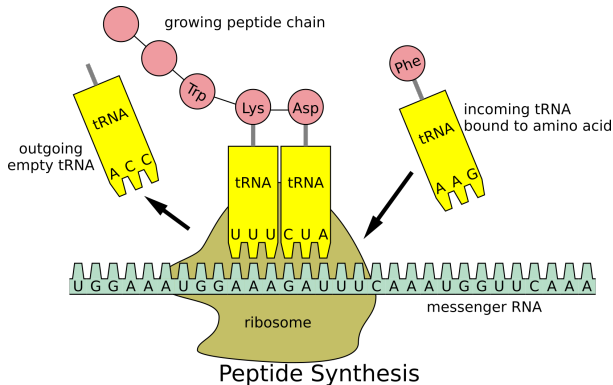


The bad news is...

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

... so have to predict the structure

Zoom: genes to proteins



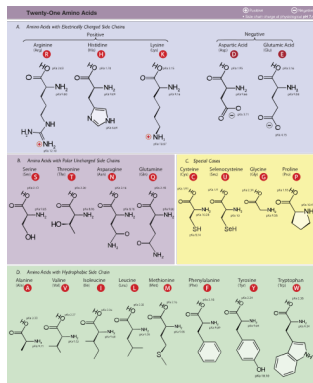
RNA codon table

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

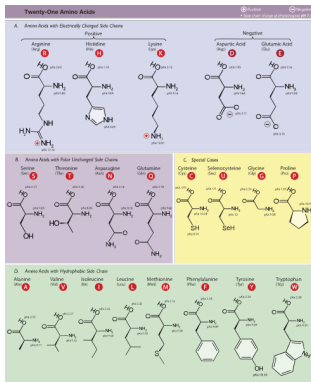
Amino Acids

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: Tryptophan
 Tyr: Tyrosine
 Val: Valine

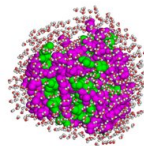
Primary to tertiary protein structure



Primary to tertiary protein structure

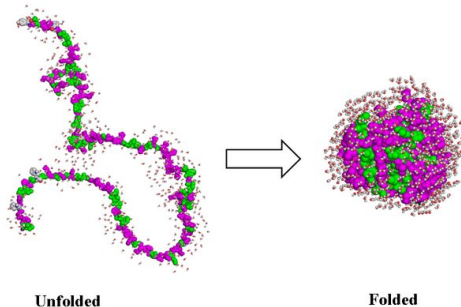
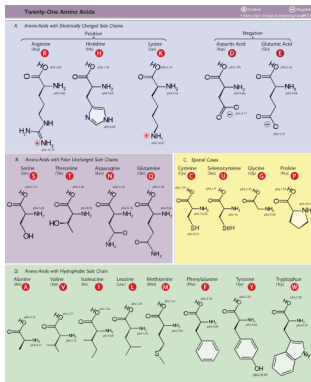


Unfolded



Folded

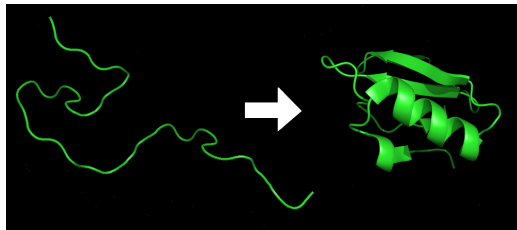
Primary to tertiary protein structure



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

Protein folding

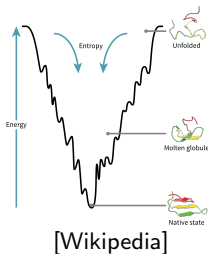
Protein folds to their stable structure in milliseconds [Karplus, 97] under interactions between their amino acids, as well with the environment (mostly water).



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:



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

Predicting the structure

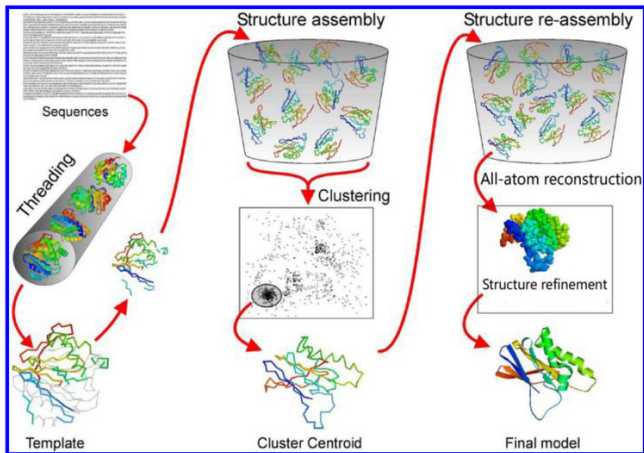
Several types of approaches:

- 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.

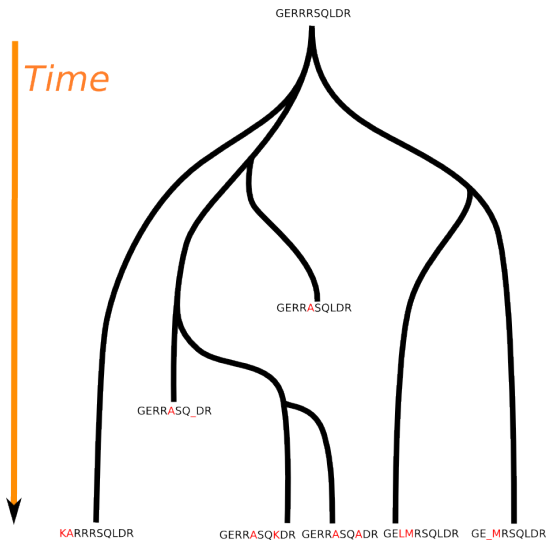
Threading

I-TASSER [Roy et al. 10]:

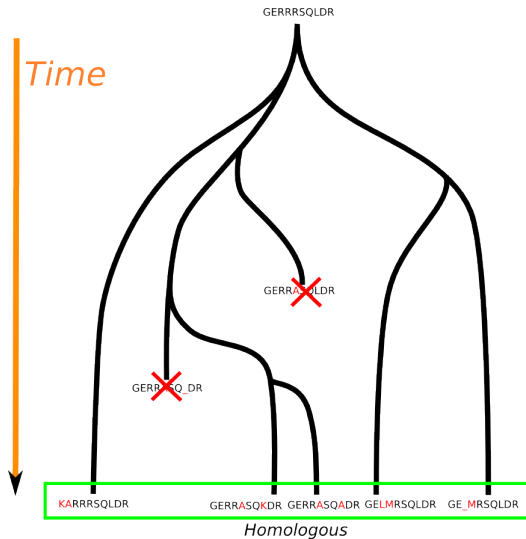


Covariation for structure prediction

Sequence evolution



Sequence evolution



Sequence conservation

Aligning the sequences (MSA, multiple sequence alignment):

```
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I G N A . G S A I G I L S
R Y D S R T T I F S P L R E G R L Y Q V E Y A M E A I S H A . G T C L G I L S
R Y D S R T T I F S P . . E G R L Y Q V E Y A Q E A I S N A . G T A I G I L S
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I S H A . G T C L G I L A
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I G H A . G T C L G I L A
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I G N A . G S A L G V L A
R Y D S R T T T F S P . . E G R L Y Q V E Y A L E A I N N A . S I T I G L I T
S Y D S R T T I F S P . . E G R L Y Q V E Y A L E A I N H A . G V A L G I V A
```

Tools	Database
ClustalW [Larkin et al. 07]	Pfam ebi.ac.uk/interpro/entry/pfam

Sequence conservation

Aligning the sequences (MSA, multiple sequence alignment):

```
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I G N A . G S A I G I L S
R Y D S R T T I F S P L R E G R L Y Q V E Y A M E A I S H A . G T C L G I L S
R Y D S R T T I F S P . . E G R L Y Q V E Y A Q E A I S N A . G T A I G I L S
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I S H A . G T C L G I L A
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I G H A . G T C L G I L A
R Y D S R T T I F S P . . E G R L Y Q V E Y A M E A I G N A . G S A L G V L A
R Y D S R T T T F S P . . E G R L Y Q V E Y A L E A I N N A . S I T I G L I T
S Y D S R T T I F S P . . E G R L Y Q V E Y A L E A I N H A . G V A L G I V A
```

Tools	Database
ClustalW [Larkin et al. 07]	Pfam ebi.ac.uk/interpro/entry/pfam

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

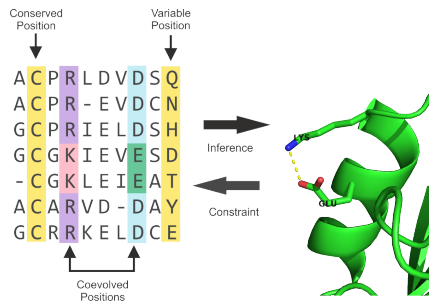
Structure is determined by amino acid interactions



Preserving the function: coevolution of residues

As protein function is vital, **evolution** selects mutations preserving **structures**.

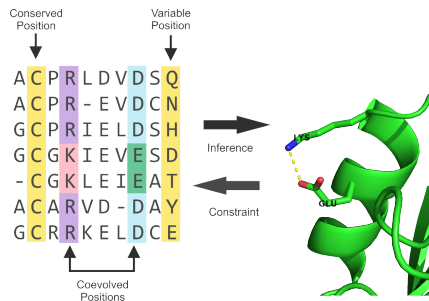
Leading to **compensatory** mutations:



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 \approx 2018.

Toward machine learning

That was the state-of-the-art until \approx 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 20×20 amino acids \times length of the sequence²) \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:

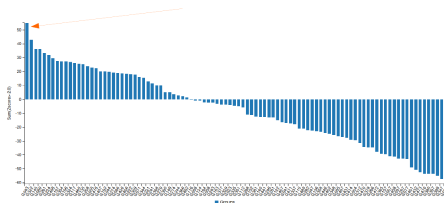
- a sequence is given
- have to predict the structure.

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

CASP13 (2018)

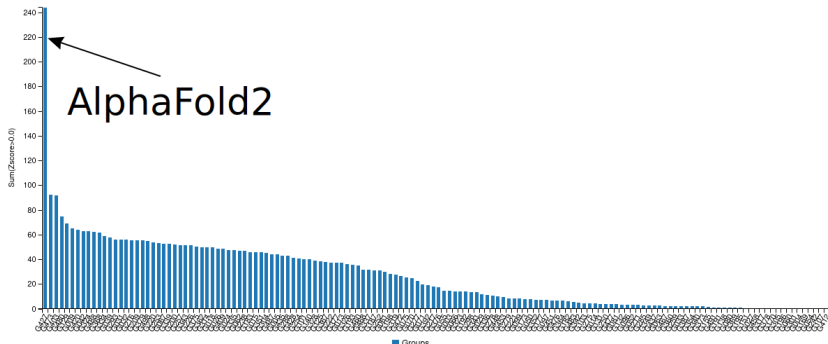
AI wins the challenge for the first time.

Google's DeepMind



CASP14 (2020)

“The big leap forward”

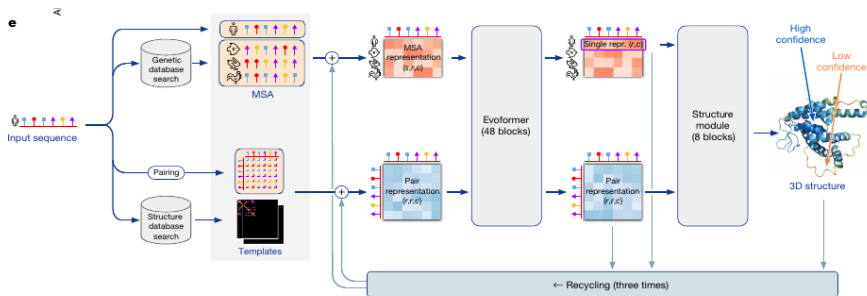


Alphafold2: attention-based learning on protein sequence alignments

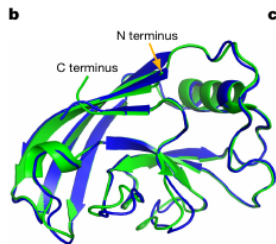
[Casp14.]

[Article in Nature]

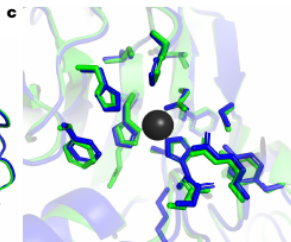
AlphaFold2 architecture



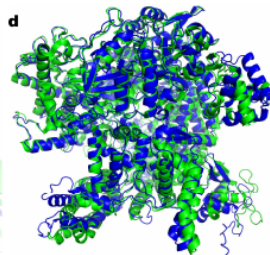
AlphaFold2 results



AlphaFold Experiment
r.m.s.d.₉₅ = 0.8 Å; TM-score = 0.93



AlphaFold Experiment
r.m.s.d. = 0.59 Å within 8 Å of Zn



AlphaFold Experiment
r.m.s.d.₉₅ = 2.2 Å; TM-score = 0.96

Information for your project

Find out the function of a gene extracted from a viral metagenome

See the git of the project here: <https://gitlab.ensimag.fr/galiez/predict-protein-x-function>.
Short syllabus and protein sequence is provided.

Let's find the function of protein
X!

Sequence comparison

Sequence alignment: algorithm and p-value

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

MEAIGNA.GSAI

QEAIGNAMGSNI

Sequence alignment: 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 alignment: 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 known 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 applied¹.

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:

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

¹See <https://doi.org/10.1093/bioinformatics/bti671>