De la séquence à la structure, (à la fonction)

Outils et méthodes de prédiction : application aux méthodes de prédiction

Stéphane Téletchéa

UFIP, Université de Nantes, CNRS, UMR 6286

Propriétés des AA

Pα			Pβ		
Glu	1.53	Hα Strong helix former	Met	1.67	Hβ Strong sheet former
Ala	1.45		Val	1.65	
Leu	1.34		Ile	1.60	
His	1.24	hα Helix former	Cys	1.30	hβ Sheet former
Met	1.20		Tyr	1.29	
Gln	1.17		Phe	1.28	
Trp	1.14		Gln	1.23	
Val	1.14		Leu	1.22	
Phe	1.12		Thr	1.20	
Lys	1.07	Iα Weak helix former	Trp	1.19	
Ile	1.00		Ala	0.97	Iβ Weak sheet former
Asp	0.98	iα Helix indifferent	Arg	0.90	iα Sheet indifferent
Thr	0.82		Gly	0.81	
Ser	0.79		Asp	0.80	
Arg	0.79		Lys	0.74	bβ Sheet breaker
Cys	0.77		Ser	0.73	
Asn	0.73	bα Helix breaker	His	0.71	
Tyr	0.61		Asn	0.65	
Pro	0.59	Bα Strong helix breaker	Pro	0.62	
Gly	0.53		Glu	0.26	Bβ Strong sheet breaker

Chou and Fasman, Biochemistry, 1974 ...

Grands principes du repliement des protéines

Fig. 2. Schematic representation of the reductive denaturation, in 8*M* urea solution containing 2-mercaptoethanol, of a disulfide–cross-linked protein. The conversion of the extended, denatured form to a randomly cross-linked, "scrambled" set of isomers is depicted at the lower right.

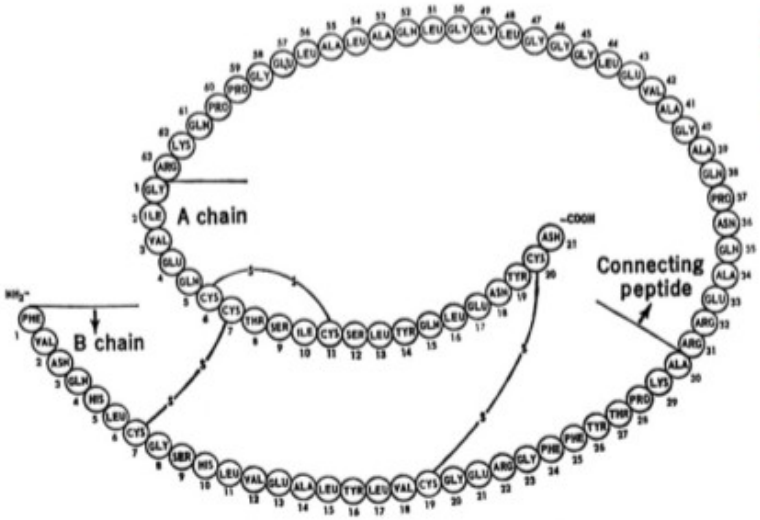


Fig. 3. The structure of porcine proinsulin (*51*).

Le repliement d'une protéine est sous contrôle thermodynamique

Toute l'information nécessaire au repliement est contenue dans la séquence

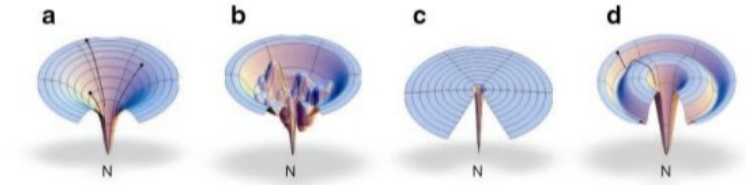
Anfinsen, 1973

Approche de « force brute »

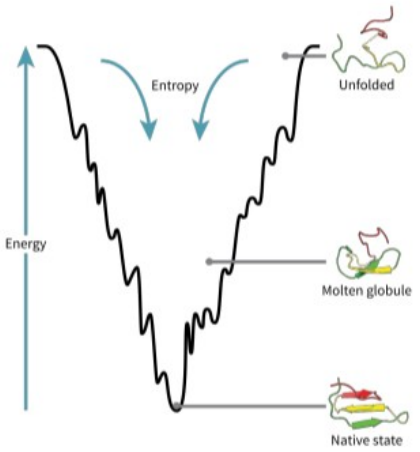
- **« All of the information necessary for folding the peptide chain into its "native" structure is contained in the amino acid sequence of the peptide. » (Anfinsen, 1960s,** https://www.nobelprize.org/prizes/chemistry/1972/press-release/**)**
- **Principe :** énumérer toutes les possibilités de structure
- **Impossible :** Paradoxe de **Levinthal**, protéine de 100 AA
  - **20^{100}** séquences possibles
  - 3^{100} conformations à explorer (hélice, feuillet, coil) : **5.10^{47}**
  - Si 1ns / conformation : 10^{38} s soit $1{,}6.10^{31}$ années …
- Il faut donc faire autrement...

Folding funnel

Grands principes du repliement (2)



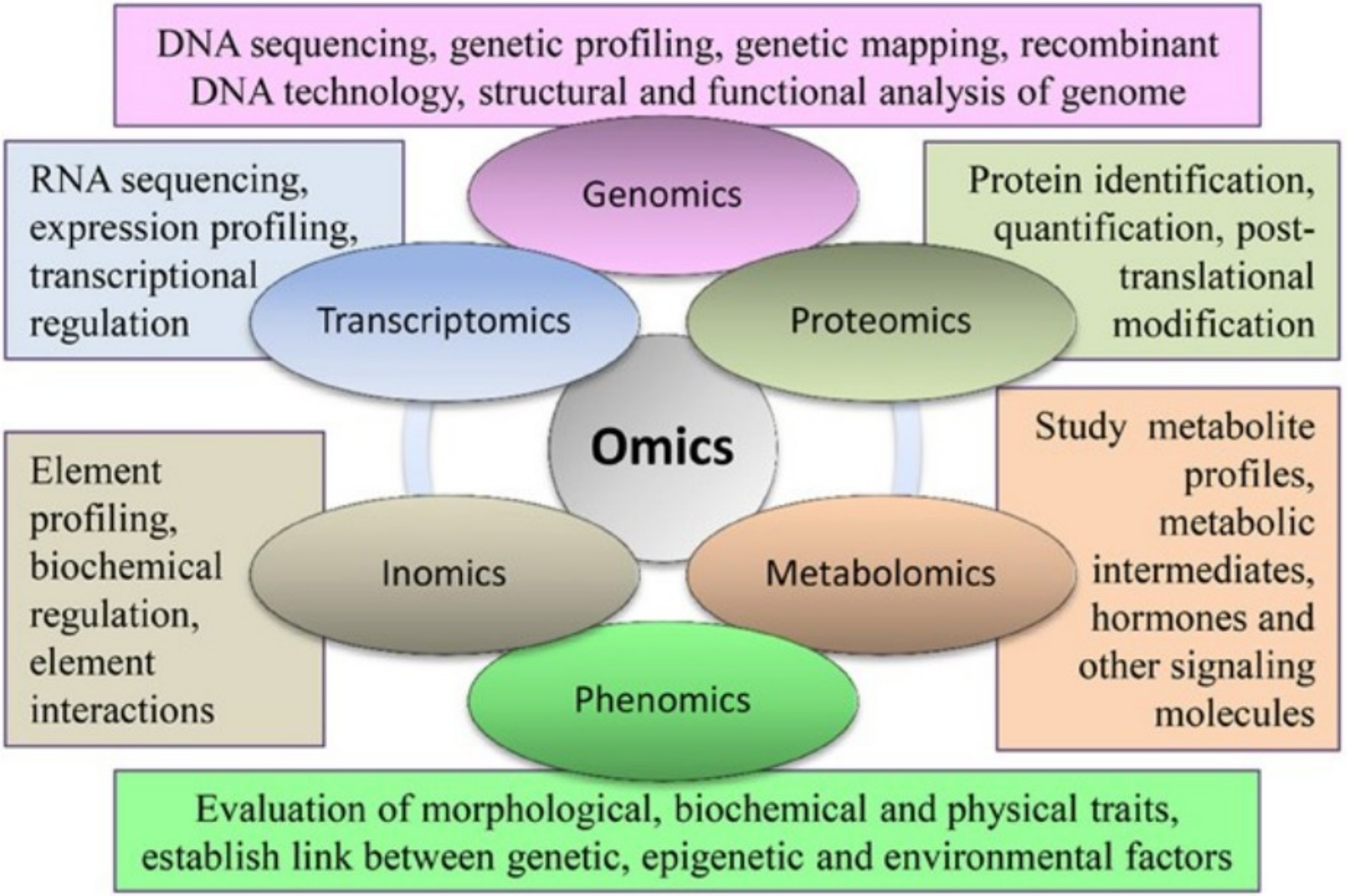
Dill KA, et al. 2008.
Annu. Rev. Biophys. 37:289–316.



Folding funnel, hydrophobic collapse

Folding funnels, Leopold, PNAS 1992
Dill, Ann. Rev. Biophys. 2008

Genomes, Proteomes, OMICS...



http://journal.frontiersin.org/article/10.3389/fpls.2014.00244/full (Soja)

De plus en plus de génomes ...

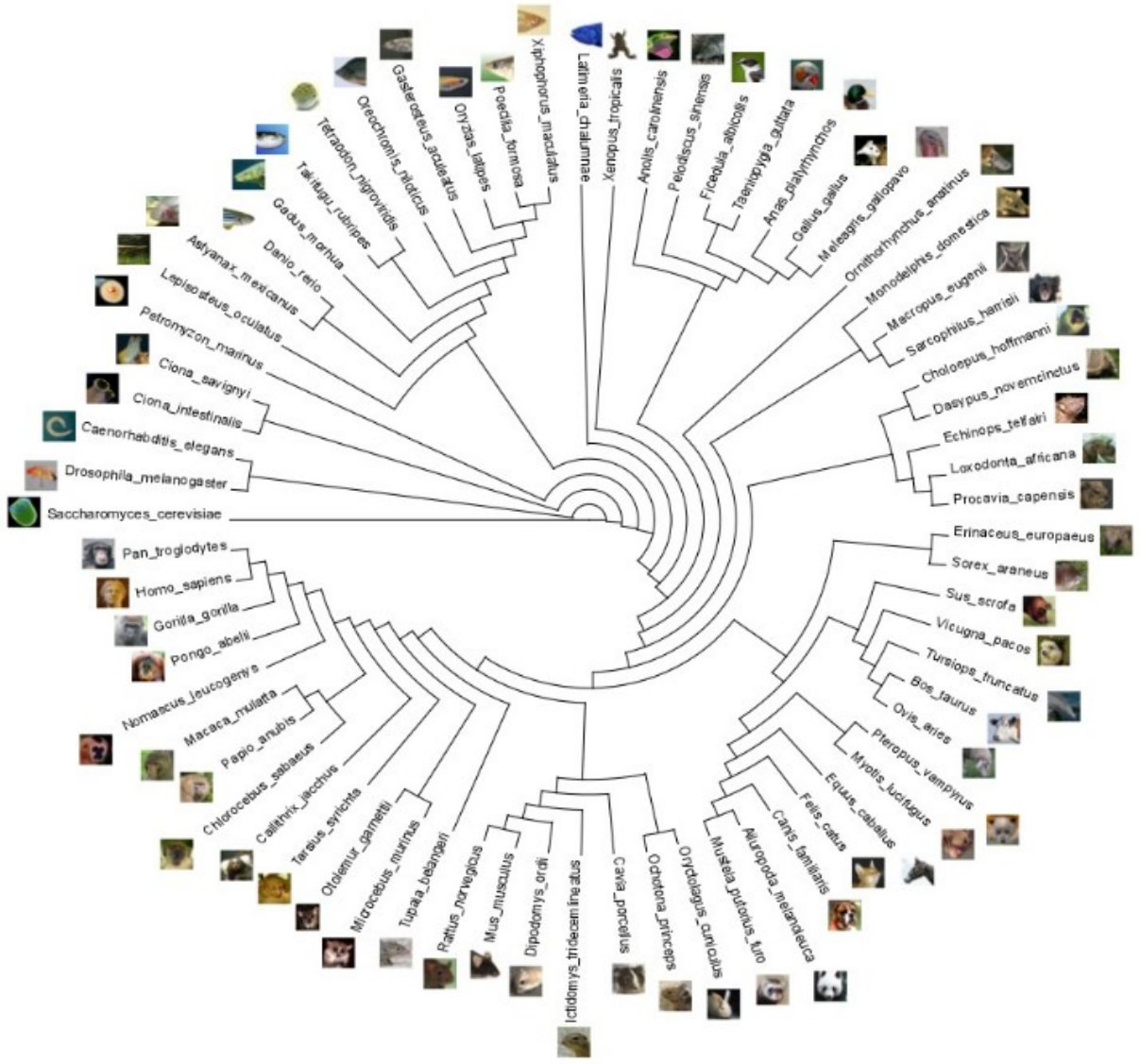


Image obtained using Dendroscope (D.H. Huson and C Scornavacca, Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks, Systematic Biology, 2012)

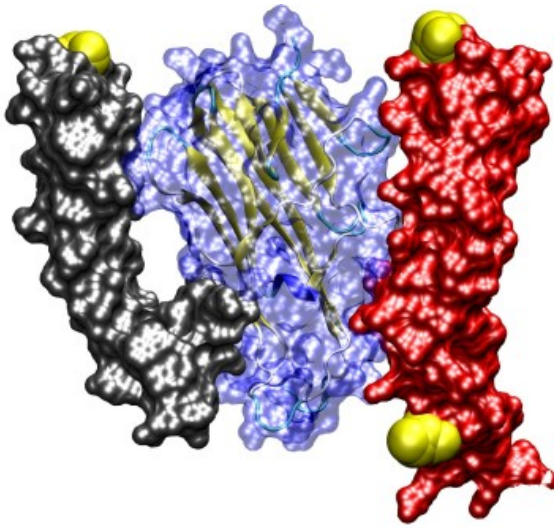
75 genomes complets

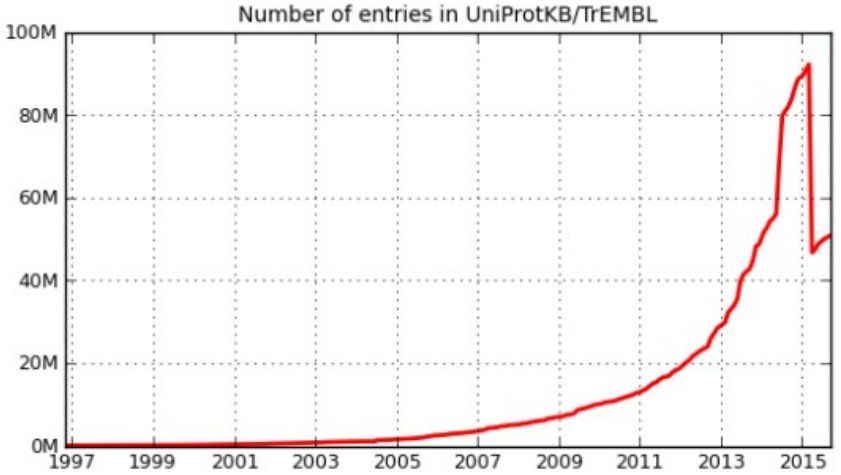
1000 genomes humains
Issus de 14 populations

2556 genomes en tout …
(bactéries, OTU ...)

Progression exponentielle

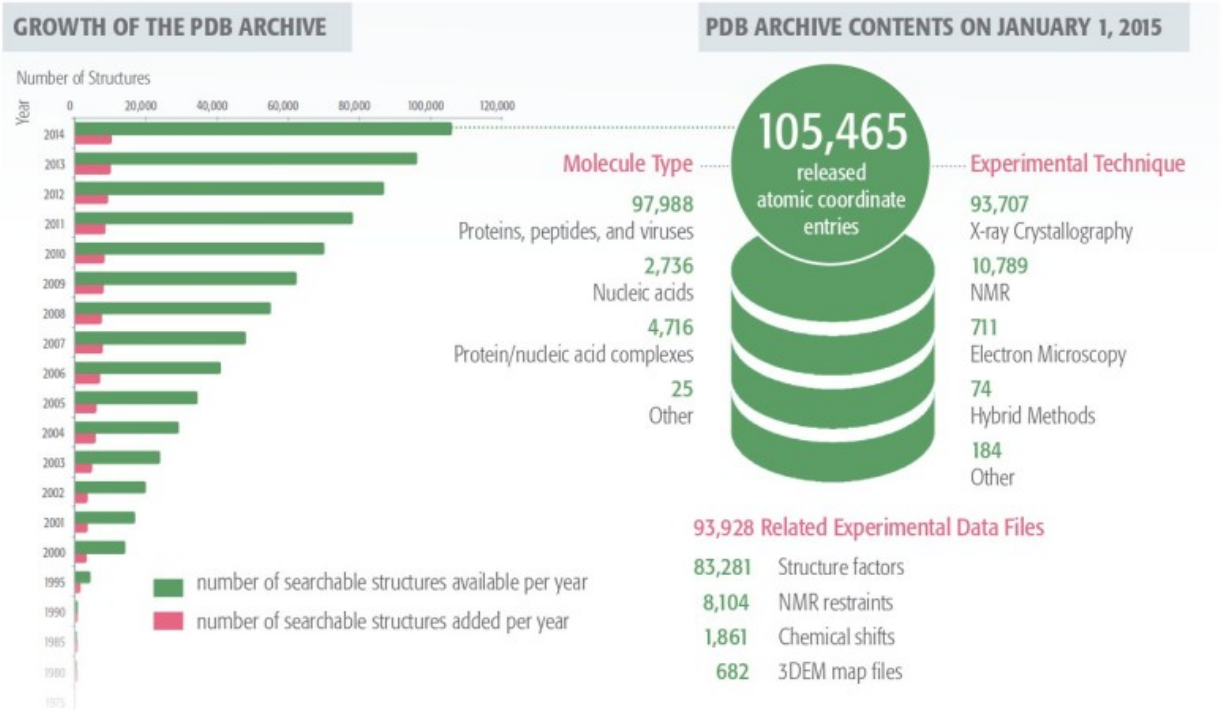
De l'ADN à la protéine …





http://www.uniprot.org

50.10^6 protein sequences



http://www.rcsb.org/pdb/general_information/news_publications/annual_reports/annual_report_year_2014.pdf

And the Dark Proteome ...

50 % des protéines détectées ont une fonction inconnue ...

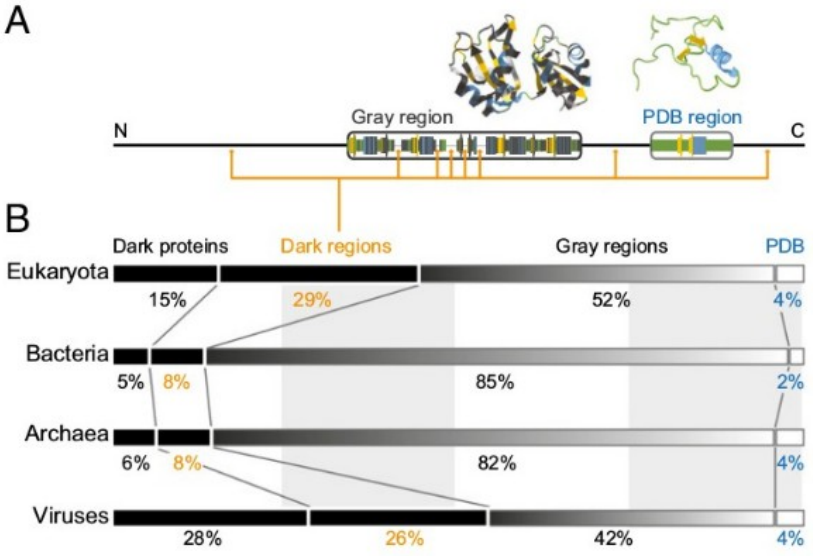


Fig. 1. Mapping the dark proteome. (*A*) For all proteins in Swiss-Prot, each residue was classified into one of four categories: (*i*) PDB regions—residues exactly matched to a PDB entry in Aquaria; (*ii*) gray regions—residues aligned to at least one PDB entry in Aquaria but always with amino acid substitutions (dark gray); (*iii*) dark regions—residues with no matching PDB entry in Aquaria; and (*iv*) dark proteins, where a single dark region spans the entire sequence. (*B*) We then calculated the total fraction of residues in each of the above four categories for all proteins in eukaryotes, bacteria, archaea, and viruses. The dark proteome (i.e., the fraction of residues in dark proteins or dark regions) varies from 13% (bacteria) to 54% (viruses).

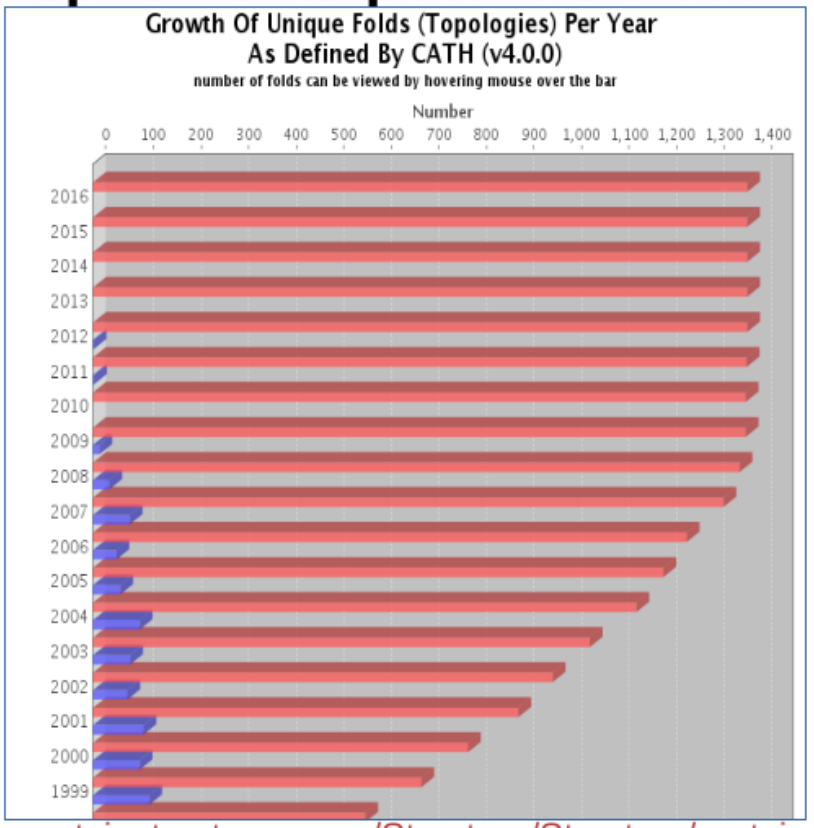
Perdigao, PNAS, 2015

Impossible?

I) « Structure is three to ten times more conserved than sequence »

http://onlinelibrary.wiley.com/doi/10.1002/prot.22458/full

II) Le nombre de repliements potentiels est fini (<1400)



http://www.proteinstructures.com/Structure/Structure/protein-fold.html

Les paradigmes de la bioinformatique structurale

I) « Structure is three to ten times more conserved than sequence »

http://onlinelibrary.wiley.com/doi/10.1002/prot.22458/full

II) Le nombre de repliements potentiels est fini (<1400)

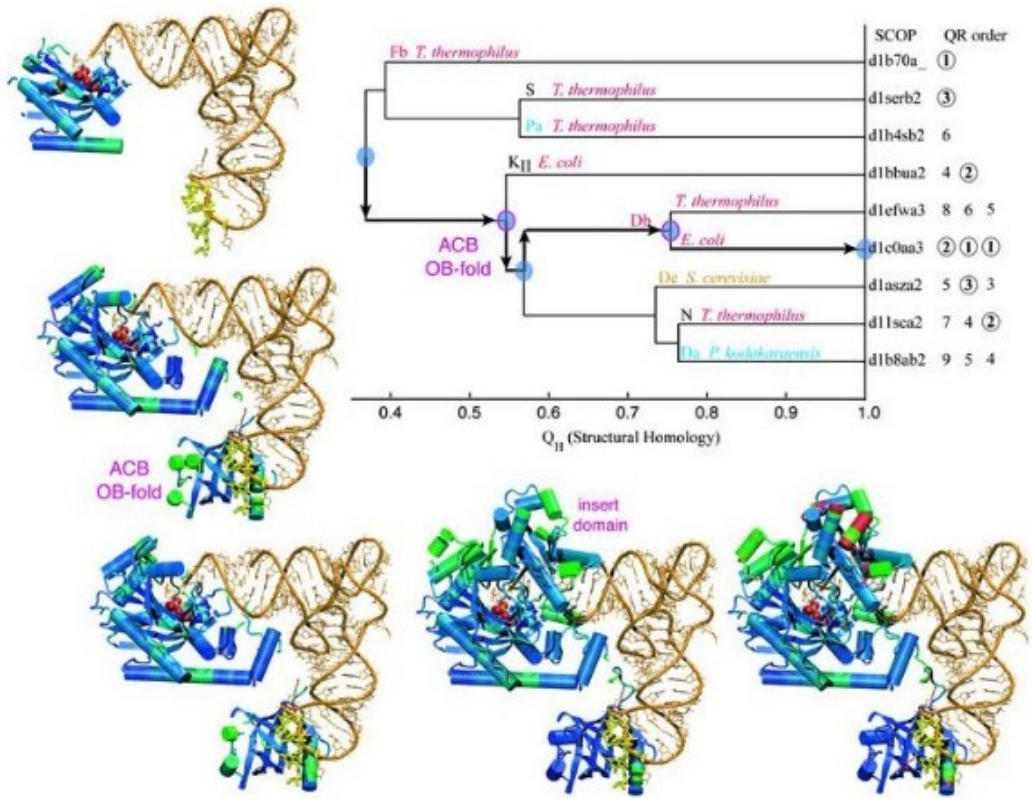
http://www.proteinstructures.com/Structure/Structure/protein-fold.html

III) Les propriétés des constituants individuels sont connues (acides aminés, acides nucléiques, sucres, lipides, …)

IV) Les macromolécules sont reliées entre elles,
ce qui permet de bénéficier de la transitivité (séquence, structure),
c'est à dire la possibilité de transférer certains paramètres aux molécules proches

Evolution of Protein Structure

Aspartyl-tRNA Synthetase




```
>Proteine1
MEGTIPANAK ...
>Proteine2
MDGSIGGRAK ...
```

% identité séquence :
30 %

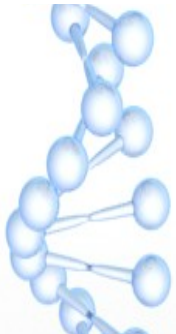
Proximité « structurale » :
90 %

http://www.ks.uiuc.edu/Training/TutorialsOverview/science/aars/aars_html/index.html

100 000 structures connues → il est possible d'en prédire beaucoup plus ...

To sum up

- Données de séquence disponibles ++++
- Données de structure disponibles
- Plusieurs méthodes à exploiter
- Besoin d'une vue statique
- Besoin d'une vue dynamique



Quelle méthode, pour quel besoin ?

En fonction des données expérimentales disponibles, trois grandes catégories :

- comparaison (*comparative*)
- enfilage (*threading*)
- *ab initio / de novo*

Pour chaque catégorie, des dizaines de méthodes disponibles : comment choisir la « meilleure » ?

Il faut une évaluation indépendante :

Critical Assessment of protein Structure Prediction (CASP)

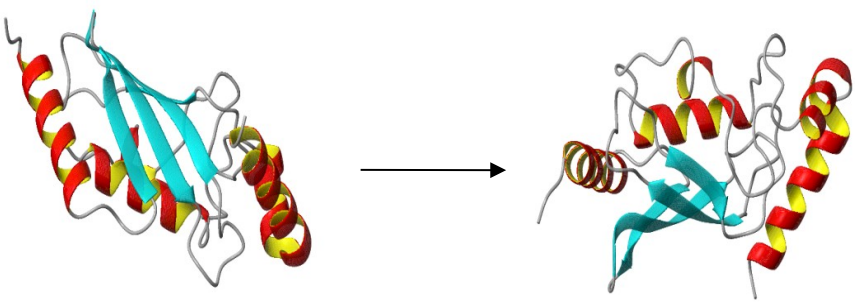
http://www.predictioncenter.org/

Comment ?

100%
id

Comparative / Homology Modelling

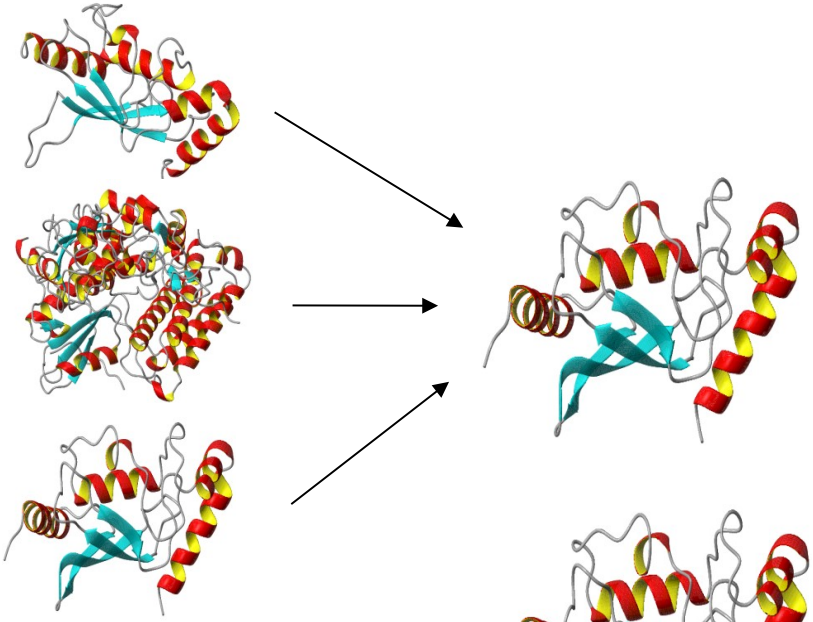
```
ATPLGLPTHVVVAGLNPHTRESD
ATPLGIPTHVPPAGLNPHTRESD
 !!!!! !!!!  !!!!!!!!!!!
```



40

" Threading"

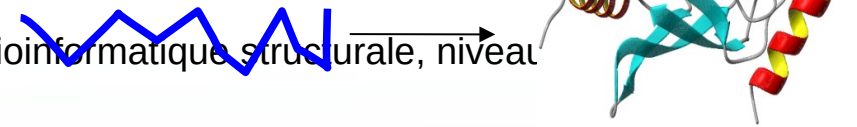
```
ETPLGLPTHVVVEGLNPHTRESD
IRVLGIPTHVPPIGLNPHTRIID
   !! !!!!   !!!!!!!  !
```



25

ab initio

```
ETPLGLPPHVVVEGLNPPPRESD
IRVLGIPVHVPPIGPNVVVRIID
   !! ! !!   ! !   !  !
```



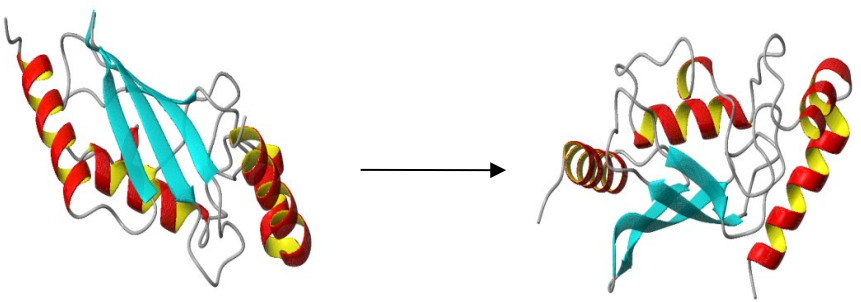
Comment ?

Fiabilité

100% id

Comparative / Homology Modelling

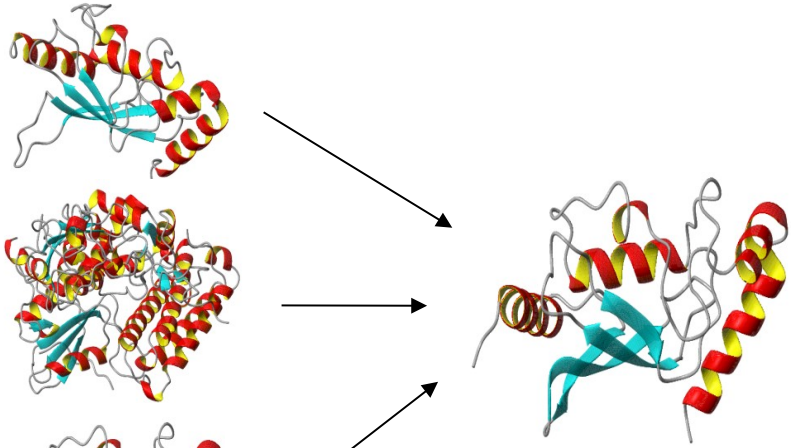
```
ATPLGLPTHVVVAGLNPHTRESD
ATPLGIPTHVPPAGLNPHTRESD
 !!!!! !!!!  !!!!!!!!!!!
```



40

" Threading"

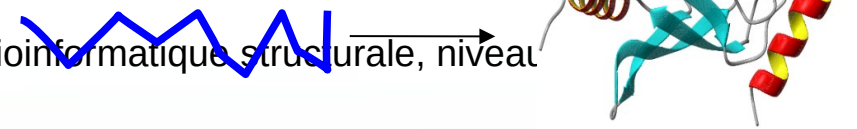
```
ETPLGLPTHVVVEGLNPHTRESD
IRVLGIPTHVPPIGLNPHTRIID
   !! !!!!   !!!!!!!  !
```



25

ab initio

```
ETPLGLPPHVVVEGLNPPPRESD
IRVLGIPVHVPPIGPNVVVRIID
   !! ! !!   ! !   !  !
```



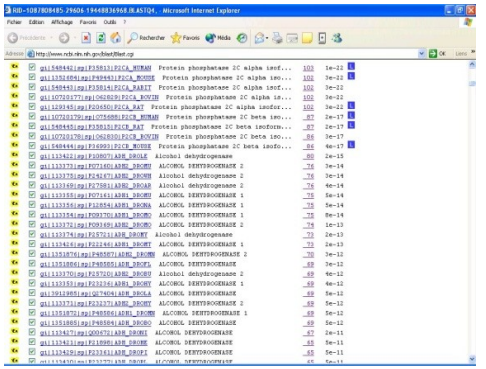
la bioinformatique structurale, niveau

Modélisation par comparaison : MODELLER (%id 40 -100)

Sequence

```
SFITPVPGGVGPMTVFLEMDLTN
KNVIFVADKRKGGPGGIIANICV
HTFNSWLDVEPRVAIEANKNGAI
WKLDLAIWKLDLGTLEAIEWWDS
HIGAFLDKPKMENAQGQGNGLRY
GLSSDAHTAVIGLPSGLESAVIG
LPSGLESWSFFFAVYDGHAGSQV
AKY...
```

Search in Structure database

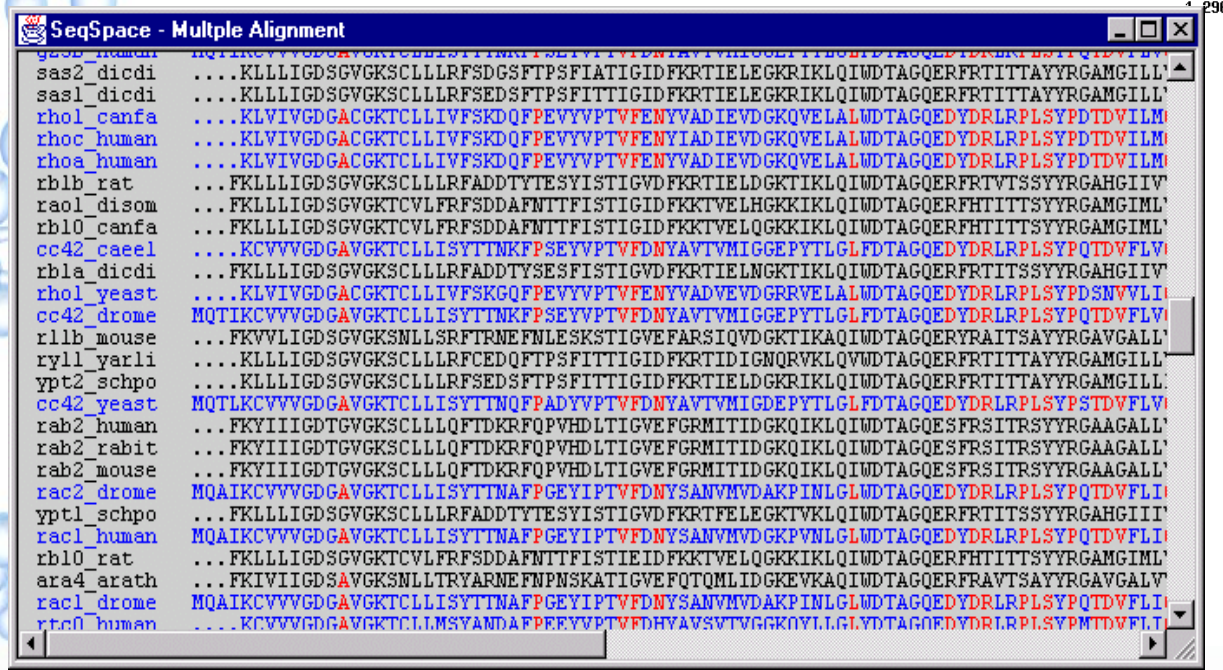


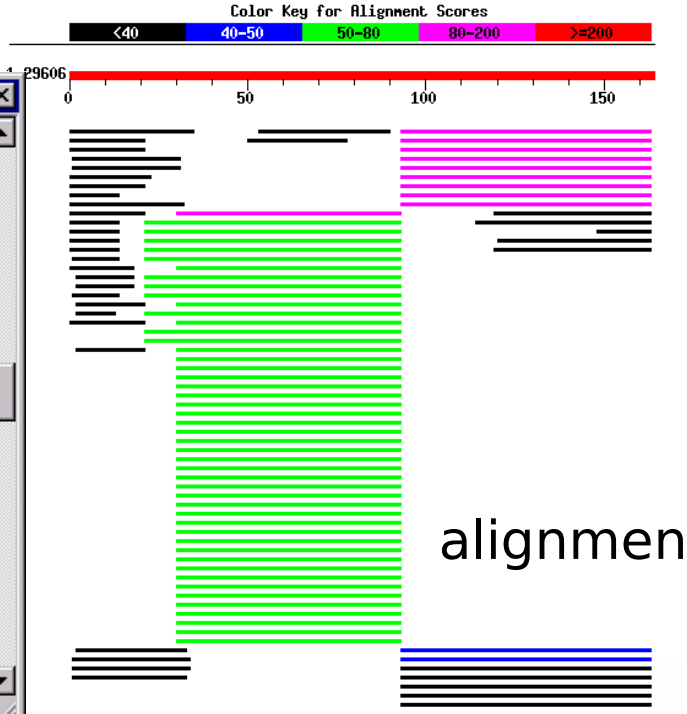
PDB

Multiple Sequences Alignment

N sequences

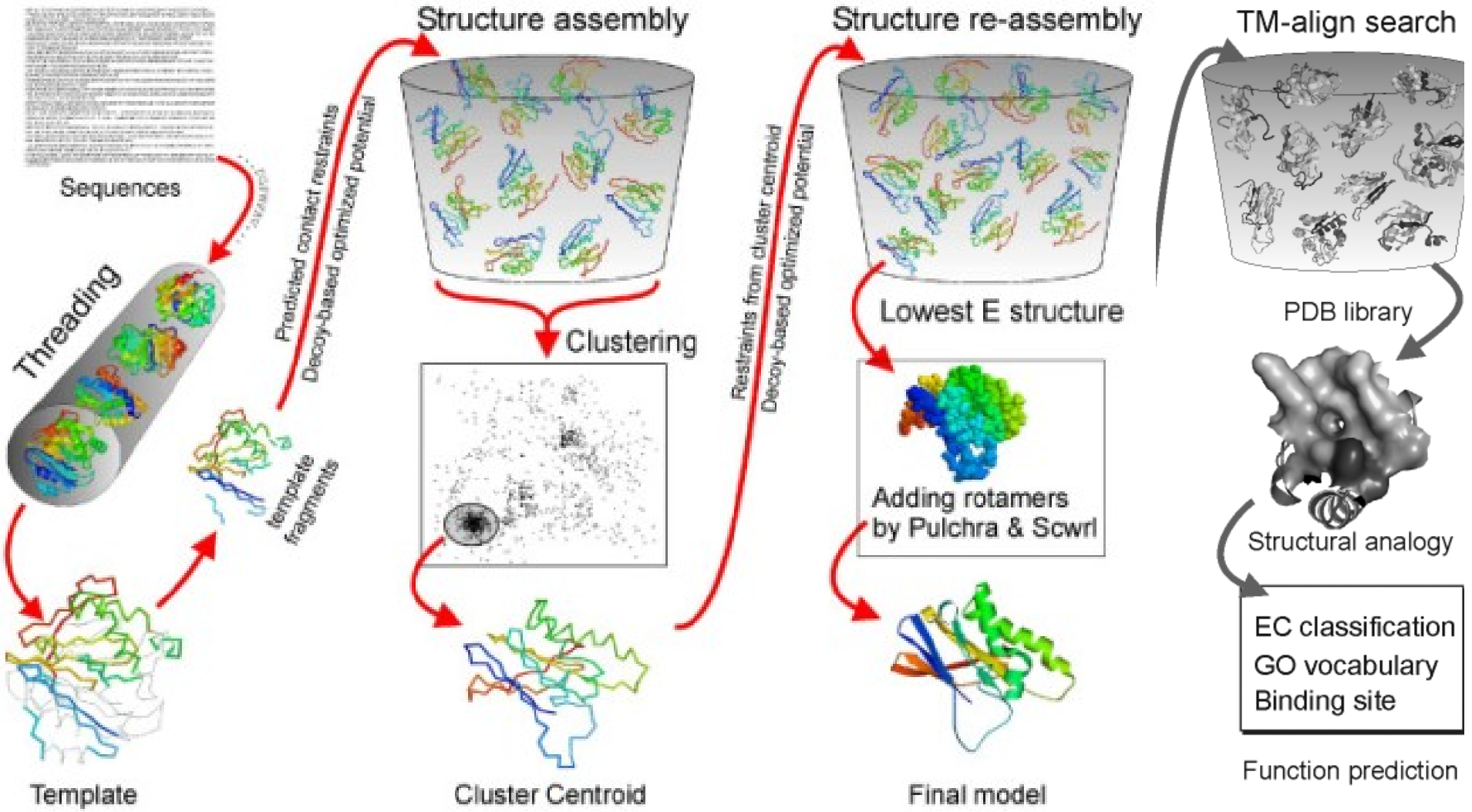
profile





alignment

THREADING: I-Tasser



de Novo / Ab initio (< 25%)

Concept :

la protéine peut être décomposée en un ensemble de fragments

- ☞Il faut une bibliothèque de fragments
- ☞Il faut une méthode pour les combiner
- ☞Il faut un score (« fonction objective »)

ROSETTA

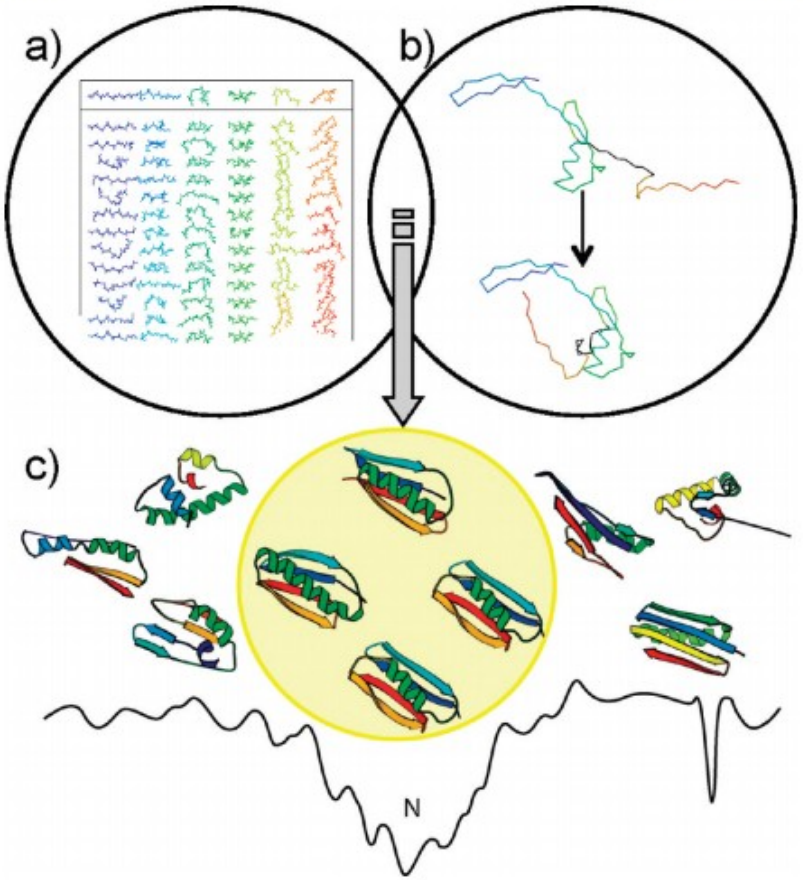
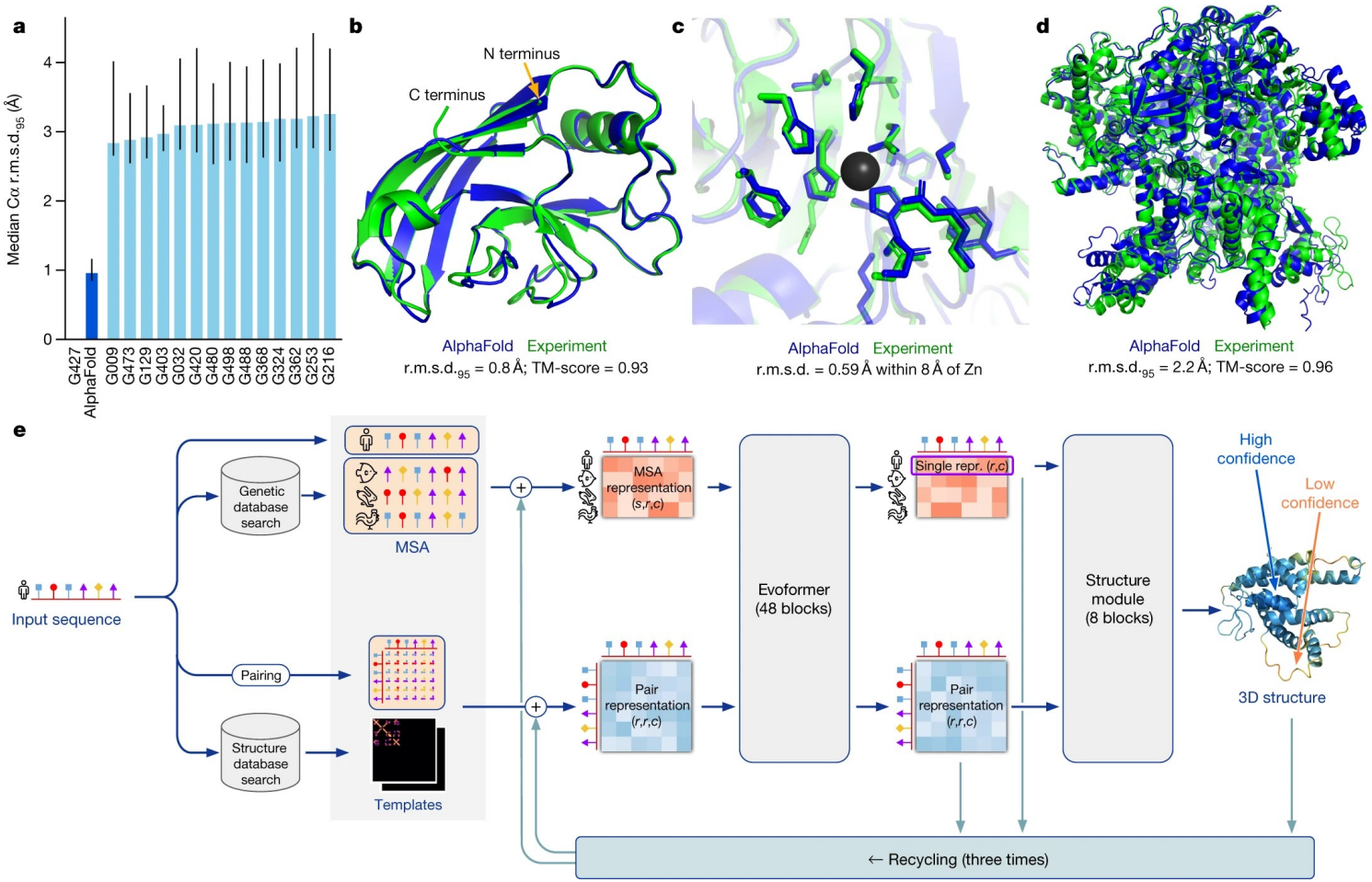


FIGURE 1: *De novo* folding algorithm. ROSETTA starts from (a) fragment libraries with sequence-dependent (φ and ψ) angles that capture the local conformational space accessible to a sequence. (b) Combining different fragments from the libraries folds the protein through optimization of non-local contacts. The low-resolution energy function depicted in panel c smoothes the rough energy surface, resulting in a deep, broad minimum for the native conformation. Metropolis Monte Carlo minimization drives the structure toward the global minimum.

And AlphaFold ?



And AlphaFold ?

Research

AlphaFold: a solution to a 50-year-old grand challenge in biology

November 30, 2020

A *SOLUTION* ...

https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

« Our AlphaFold AI system **solved** the 50-year-old challenge of protein structure prediction »

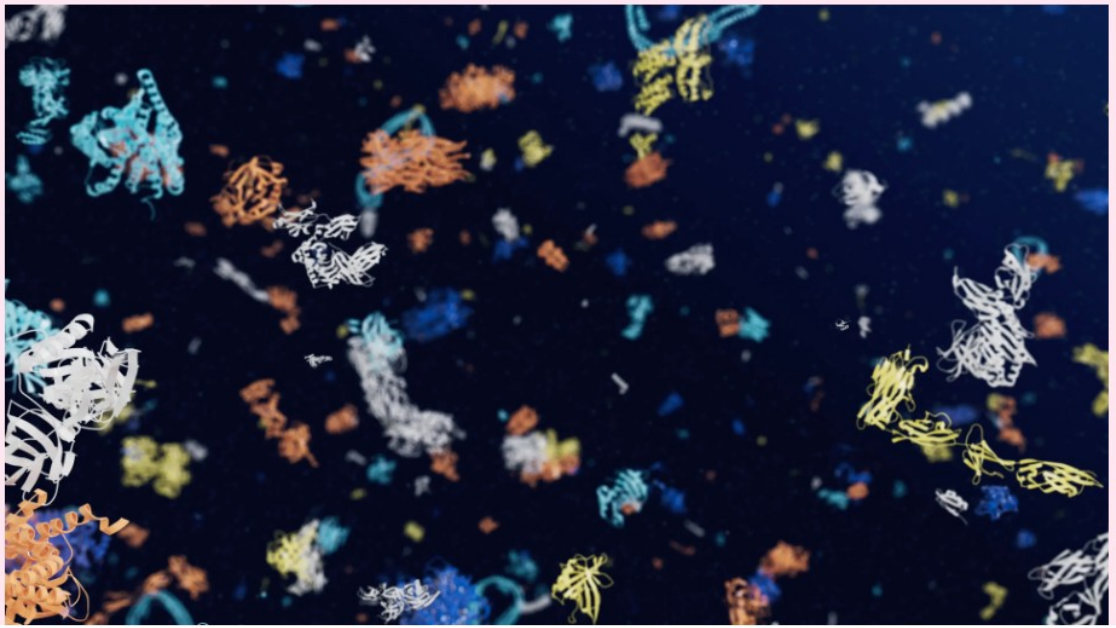
https://www.deepmind.com/blog/how-our-principles-helped-define-alphafolds-release **(September 14th, 2022)**

And for all the protein Universe…

Research

AlphaFold reveals the structure of the protein universe

July 28, 2022



https://www.deepmind.com/blog/alphafold-reveals-the-structure-of-the-protein-universe

AlphaFold heralds a data-driven revolution in biology and medicine

Janet M. Thornton ✉, Roman A. Laskowski & Neera Borkakoti

Nature Medicine **27**, 1666–1669 (2021) | Cite this article

Fig. 1: The good, the bad and the ugly.

From: AlphaFold heralds a data-driven revolution in biology and medicine

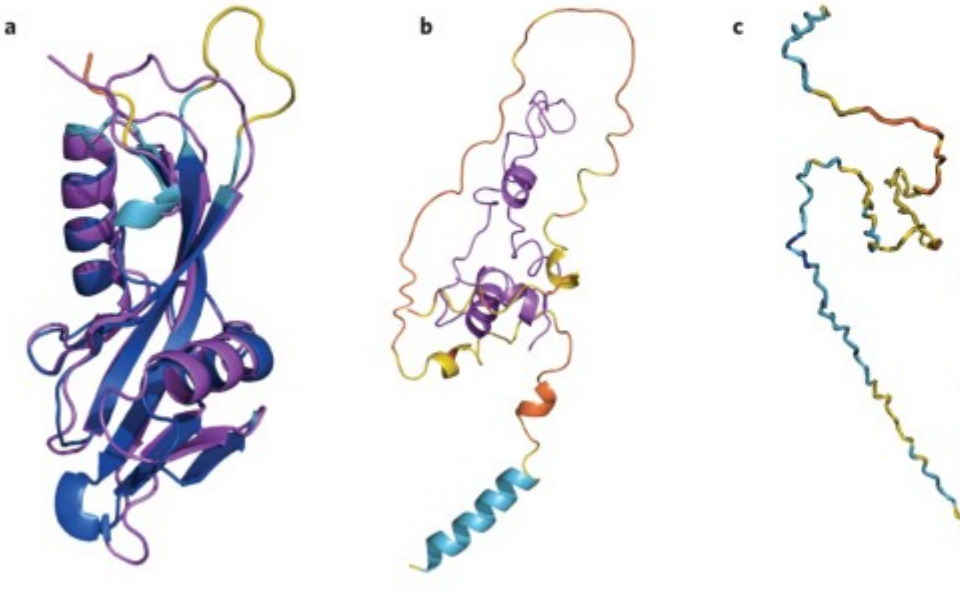
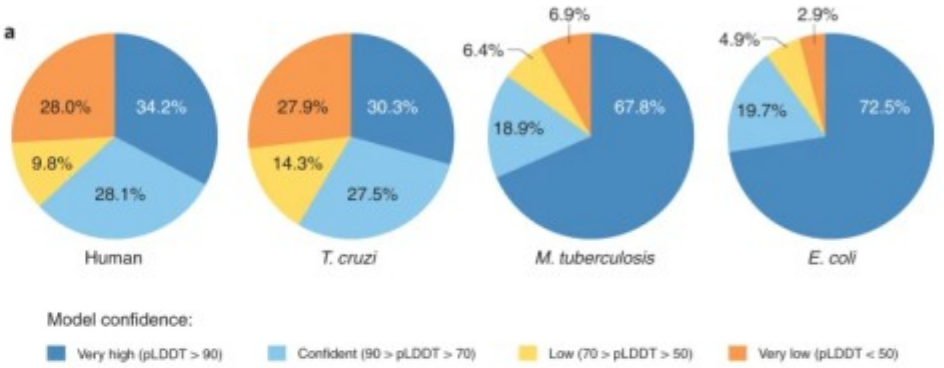


Fig. 2: Confidence scores for AlphaFold models.



Homo sapiens :

- 1/3 « grande confiance » (pLDDT > 90
- 2/3 confiance moyenne ou bonne (pLDDT >70)

Thornton, J.M., Laskowski, R.A. & Borkakoti, N. AlphaFold heralds a data-driven revolution in biology and medicine. Nat Med 27, 1666–1669 (2021). https://doi.org/10.1038/s41591-021-01533-0

Current limitations of the prediction method

Although the availability of predicted 3D models for the known “protein universe” is an exciting prospect with huge impact, there are nevertheless limitations to the AlphaFold method and resource, some of which may be addressed in the future:

- Many proteins **function as complexes** with other proteins
- Proteins are **dynamic systems** and adopt different structures depending on their environment or state
- For regions that are **intrinsically disordered or unstructured in isolation**, AlphaFold is expected to produce a low-confidence prediction
- AlphaFold **has not been trained or validated for predicting the effect of mutations**
- **Ligands are not included** in the structures (GTP, GDP, Mg^{2+}?)
- *PTM are not included in the predictions*
- **Caution must be taken about putative functions, they have to be tested by further experimentation**

https://www.ebi.ac.uk/about/news/perspectives/alphafold-potential-impacts/