From Sequence to Structure, From Structure to Function

Predicting macromolecular structures: Methods Strenghts and Weaknesses

Stéphane Téletchéa

US2B, Nantes University, CNRS, UMR 6286, France

What is structural bioinformatics?

- **Structural biology:** study of biological macromolecues using advanced **experimental methods**, X-Ray cristallography, NMR, EM

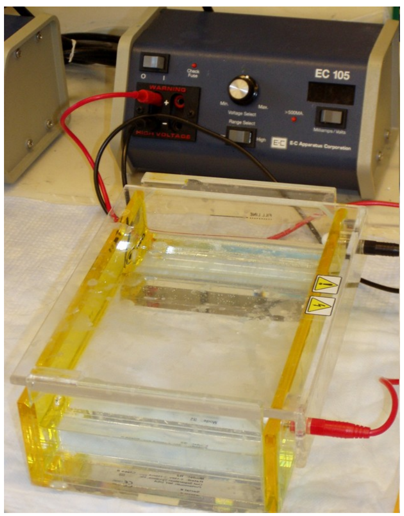
- **Structural bioinformatics:** study of biological macromolecules using **advanced prediction methods**, Molecular Modelling, Protein-Protein Interactions, Energy Estimation, Molecular Motions

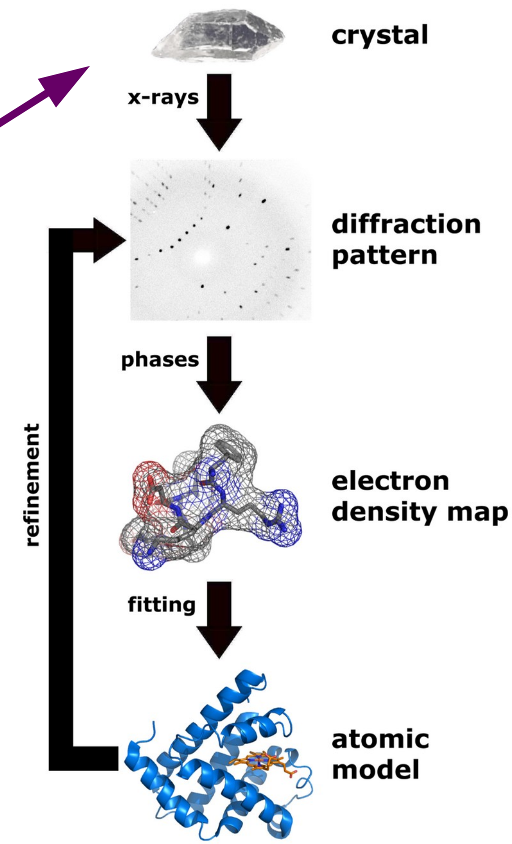
Some milestones in Structural Biology

- 1938 : Chymotrypsine et hémoglobine (M. Perutz)
- 1951 : structure de l'ADN (Watson & Crick)
- 1981 : haemagluttinin Influenzae
- 1984 : nucléosome
- 1997 : ATP synthase (à droite)
  https://pdb101.rcsb.org/motm/72
- 2000 : Sous-unité 30S
- 2007 : Structure Beta-2 adrénergique (B. Kobilka)
  https://med.stanford.edu/kobilkalab/research.html
- 2010 : Complexe respiratoire 1
- 2011 : Canal sodique
- 2020 : « Méthode d'édition du génome », E. Charpentier, J.Doudna

https://www.nature.com/milestones/milecrystal/library/structural-biology/index.html
https://pdb101.rcsb.org/learn/flyers-posters-and-other-resources/other-resource/structural-biology-and-nobel-prizes

X-Ray cristallography





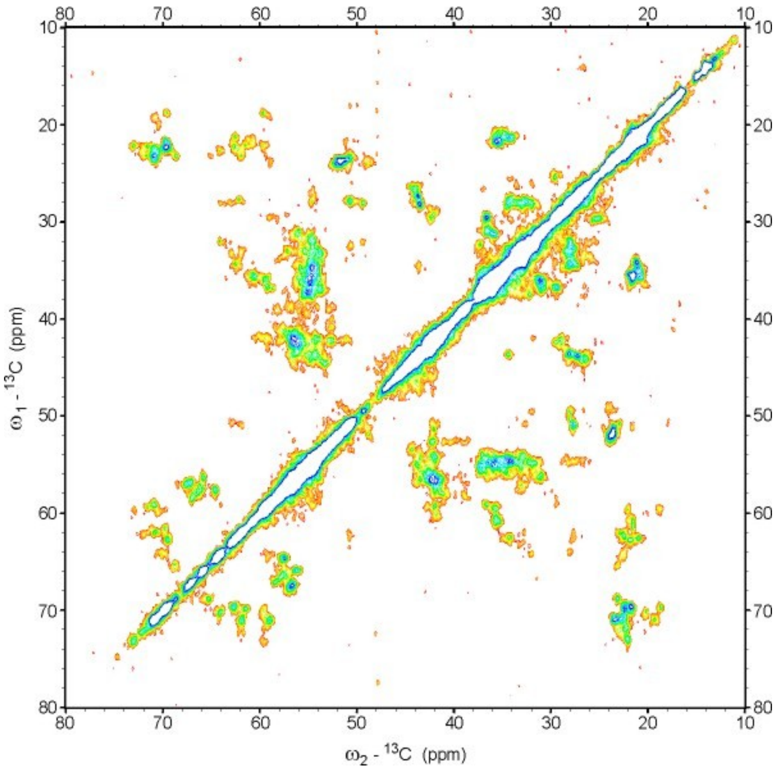
https://www.jove.com/science-education/10216/growing-crystals-for-x-ray-diffraction-analysis

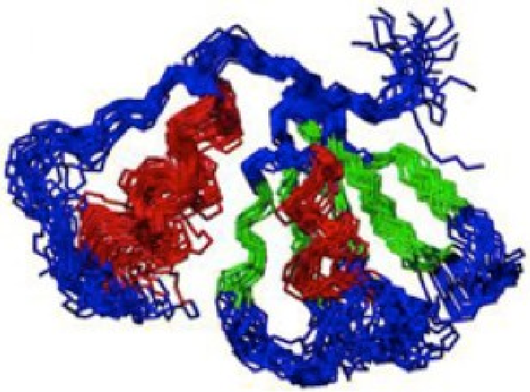
http://www.nature.com/news/crystallography-atomic-secrets-1.14608

Notions importantes :

- **résolution (Angstroems), le plus bas possible**
- **R-free (adéquation au maillage), <20 %**

Nuclear Magnetic Resonance (NMR)



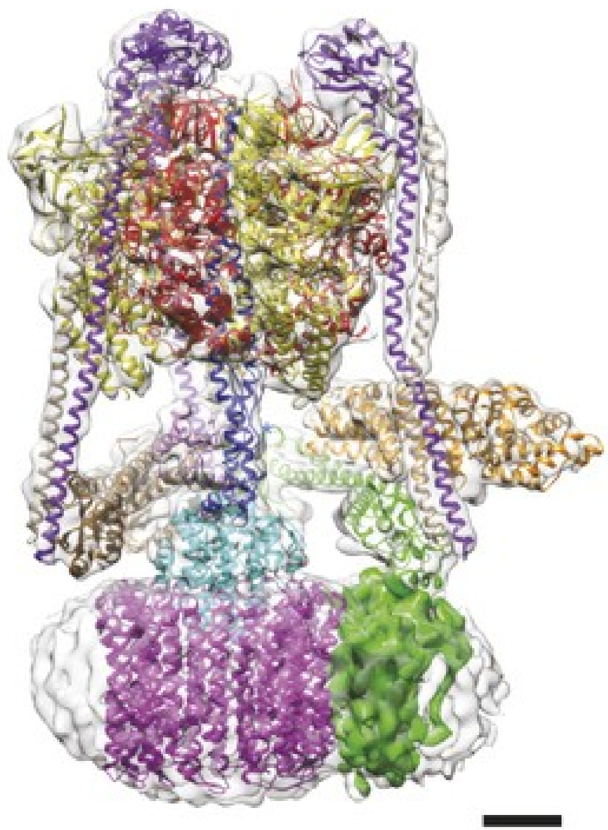


Notions importantes :

flexibilité des éléments de structure

précision moindre que la cristallographie

Electron Miscroscopy (EM)



Volume global

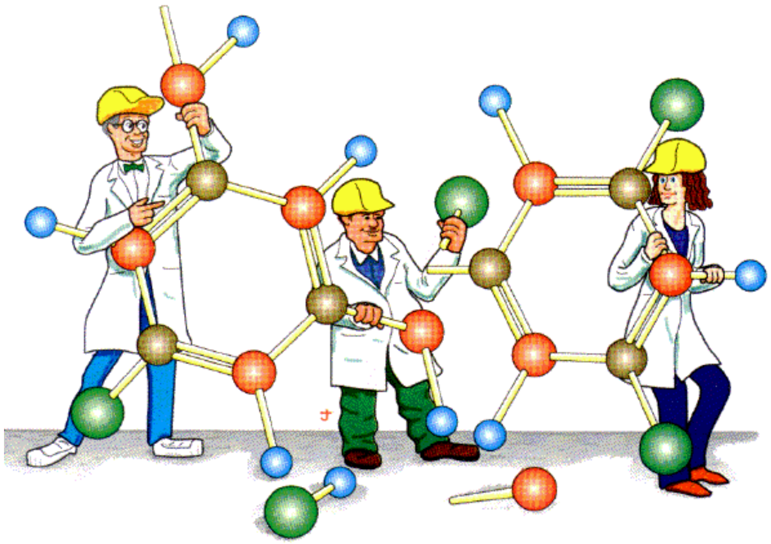
Pas de détail atomique

Architecture macromoléculaire

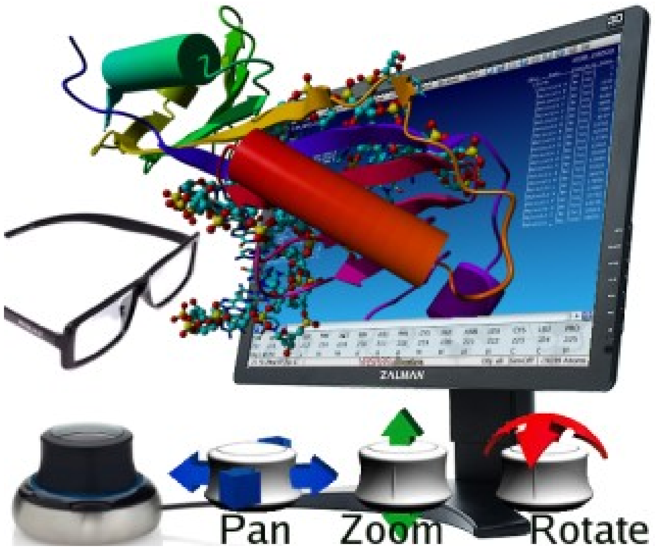
En développement ++

http://www.nature.com/news/the-revolution-will-not-be-crystallized-a-new-method-sweeps-through-structural-biology-1.18335

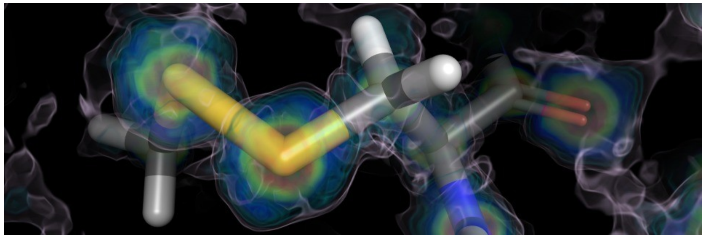
Molecular Modelling



https://dasher.wustl.edu/tinker/



http:/www.yasara.org/



http://pymol.org/

Expertise +++

Précision inconnue

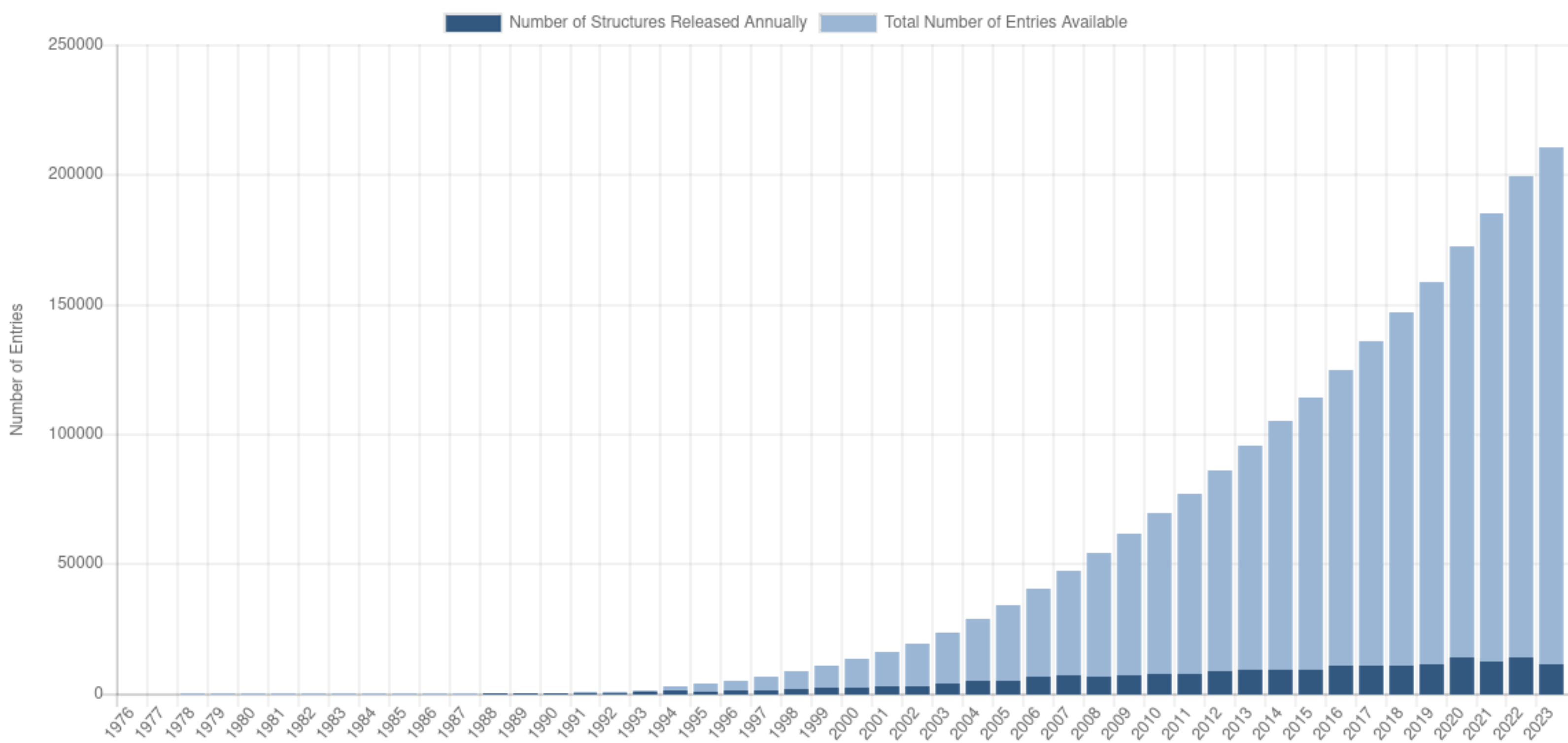
3D Protein Availability

PDB : https://www.rcsb.org (Oct 20th, 2023)

210,836 Structures from the PDB

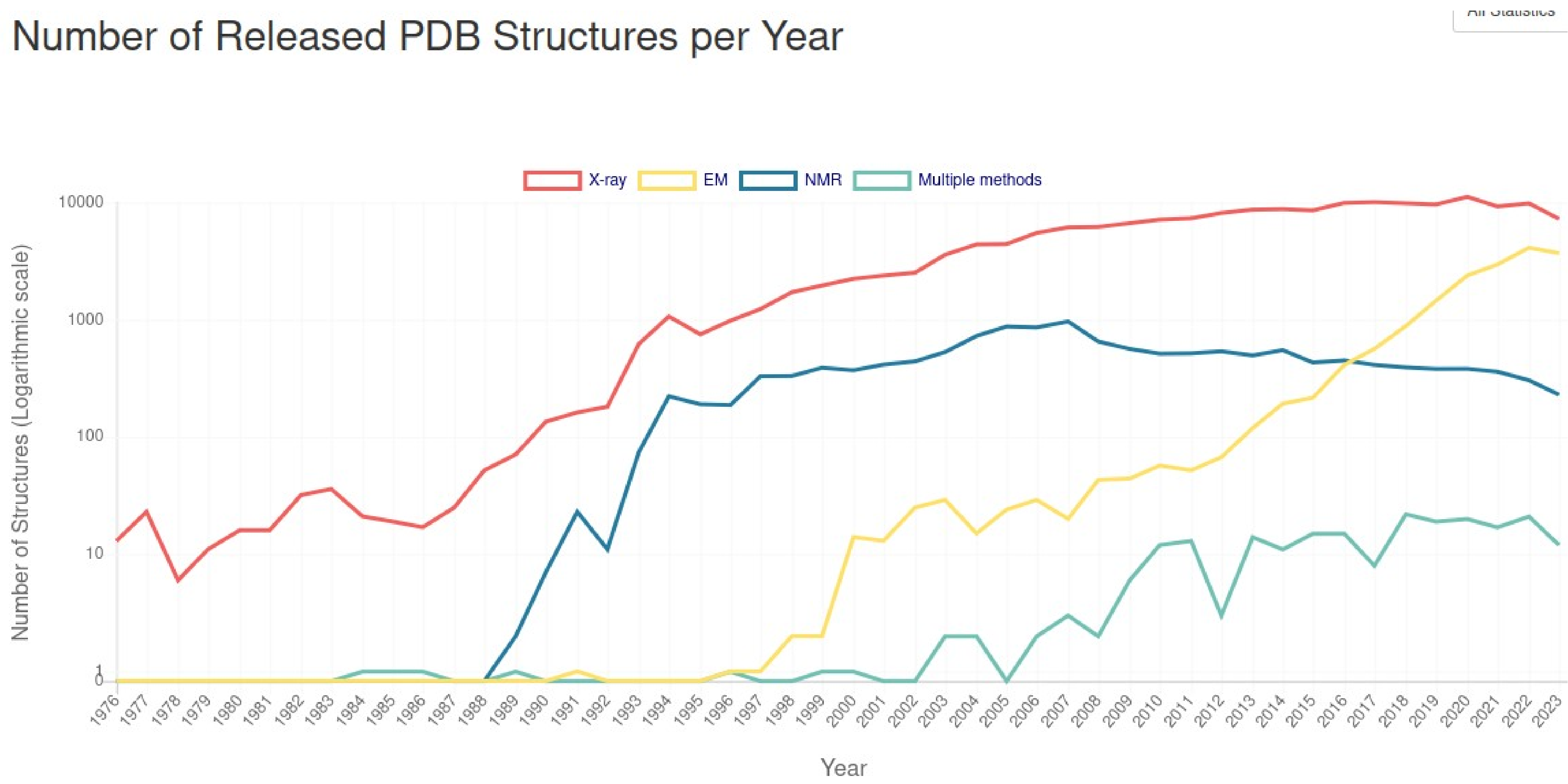
1,068,577 Computed Structure Models (CSM)

PDB Statistics: Overall Growth of Released Structures Per Year



https://www.rcsb.org/stats/growth/growth-released-structures

Experimental methods



- X-ray: X-RAY DIFFRACTION, FIBER DIFFRACTION, or POWDER DIFFRACTION
- NMR: SOLUTION NMR or SOLID-STATE NMR
- EM: ELECTRON MICROSCOPY or ELECTRON CRYSTALLOGRAPHY or ELECTRON TOMOGRAPH
- MULTIPLE METHODS: Multiple experimental methods. For example, if a structure is solved by X-RAY DIFFRACTION AND NEUTRON DIFFRACTION, it will be counted only in this category.

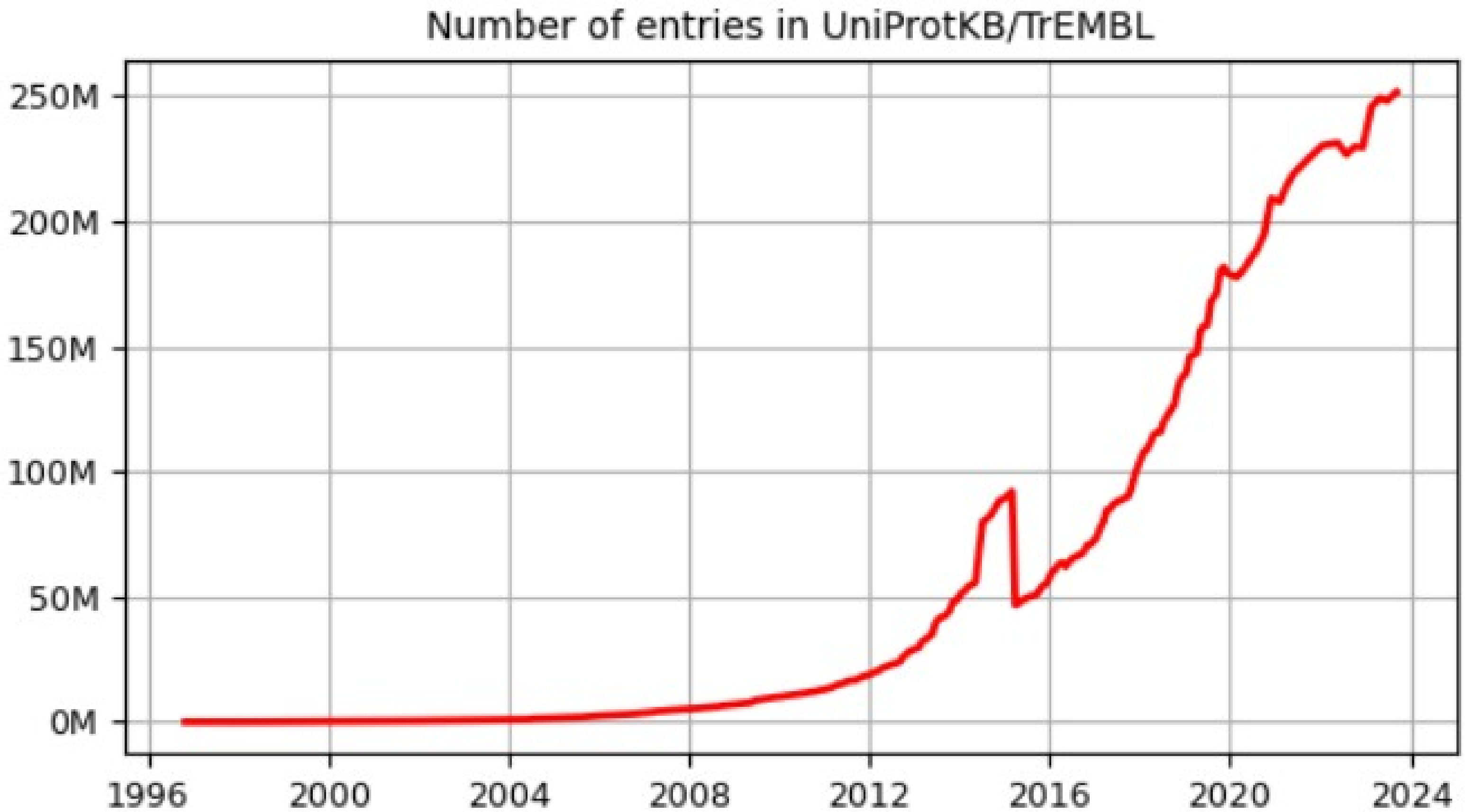
Data availability

- PDB : RX, RMN, Small-Angle X-Ray Scattering (SAXS), modèles http://www.wwpdb.org/
- EMDB https://www.ebi.ac.uk/pdbe/emdb/
- Serveurs de prédiction
  - Proteinmodelportal http://www.proteinmodelportal.org/
  - Swiss-Model https://swissmodel.expasy.org/)
- Protéines membranaires https://blanco.biomol.uci.edu/mpstruc/

Littérature

PubMed : https://pubmed.ncbi.nlm.nih.gov/
Google Scholar : https://scholar.google.fr/
Semantics Scholar : https://www.semanticscholar.org
CNRS : https://www.cnrs.fr/
INSERM : https://www.inserm.fr/
Maladies rares et médicaments orphelins : https://www.orpha.net

210,000 protein structures…



260,000,000 protein sequences …

Number of fragments: 26324883

Protein existence (PE):	entries	%
1: Evidence at protein level	216396	0.09%
2: Evidence at transcript level	1374760	0.55%
3: Inferred from homology	81622144	32.44%
4: Predicted	168387468	66.93%
5: Uncertain	0	0.00%

Predicting a protein structure

Propriétés des AA

	$P\alpha$		
Glu	1.53	$H\alpha$	Strong helix former
Ala	1.45		
Leu	1.34		
His	1.24	$h\alpha$	Helix former
Met	1.20		
Gln	1.17		
Trp	1.14		
Val	1.14		
Phe	1.12		
Lys	1.07	$I\alpha$	Weak helix former
Ile	1.00		
Asp	0.98	$i\alpha$	Helix indifferent
Thr	0.82		
Ser	0.79		
Arg	0.79		
Cys	0.77		
Asn	0.73	$b\alpha$	Helix breaker
Tyr	0.61		
Pro	0.59	$B\alpha$	Strong helix breaker
Gly	0.53		

	$P\beta$		
Met	1.67	$H\beta$	Strong sheet former
Val	1.65		
Ile	1.60		
Cys	1.30	$h\beta$	Sheet former
Tyr	1.29		
Phe	1.28		
Gln	1.23		
Leu	1.22		
Thr	1.20		
Trp	1.19		
Ala	0.97	$I\beta$	Weak sheet former
Arg	0.90	$i\alpha$	Sheet indifferent
Gly	0.81		
Asp	0.80		
Lys	0.74	$b\beta$	Sheet breaker
Ser	0.73		
His	0.71		
Asn	0.65		
Pro	0.62		
Glu	0.26	$B\beta$	Strong sheet breaker

Chou and Fasman, Biochemistry, 1974 ...

Grands principes du repliement des protéines

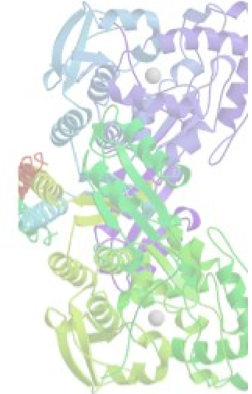


Fig. 2. Schematic representation of the reductive denaturation, in 8*M* urea solution containing 2-mercaptoethanol, of a disulfide–cross-linked protein. The conversion of the extended, denatured form to a randomly cross-linked, "scrambled" set of isomers is depicted at the lower right.

Le repliement d'une protéine est sous contrôle thermodynamique

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

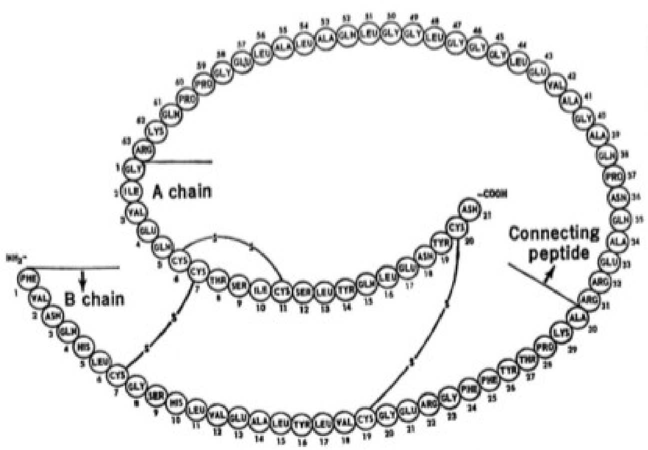


Fig. 3. The structure of porcine proinsulin (*51*).

Anfinsen, 1973

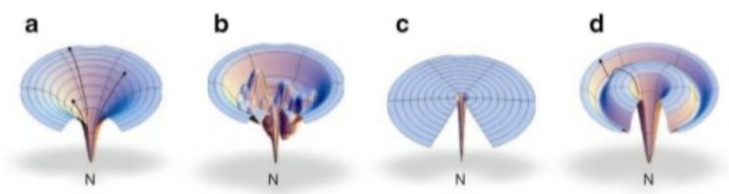
« All of the information necessary for folding the peptide chain into its "native" structure is contained in the amino acid sequence of the peptide. » (Anfinsen, 1960s, https://www.nobelprize.org/prizes/chemistry/1972/press-release/)

Approche de « force brute »

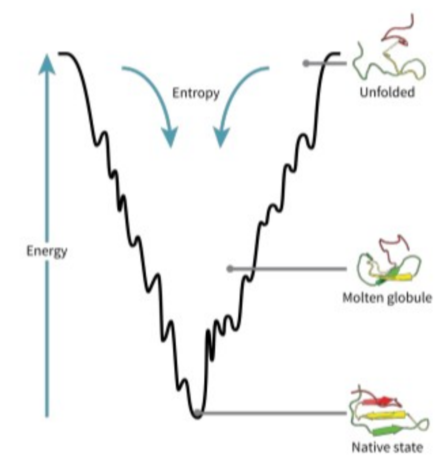
- **Principe :** énumérer toutes les possibilités de structure
- **Impossible :** Paradoxe de **Levinthal**, protéine de 100 AA
  - **20^{100}** séquences possibles
  - 3^{100} conformations à explorer (hélice, feuillet, coil) : **5.10^{47}**
  - Si 1ns / conformation : 10^{38} s soit $1,6.10^{31}$ années …
- 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

But the Dark Proteome ...

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

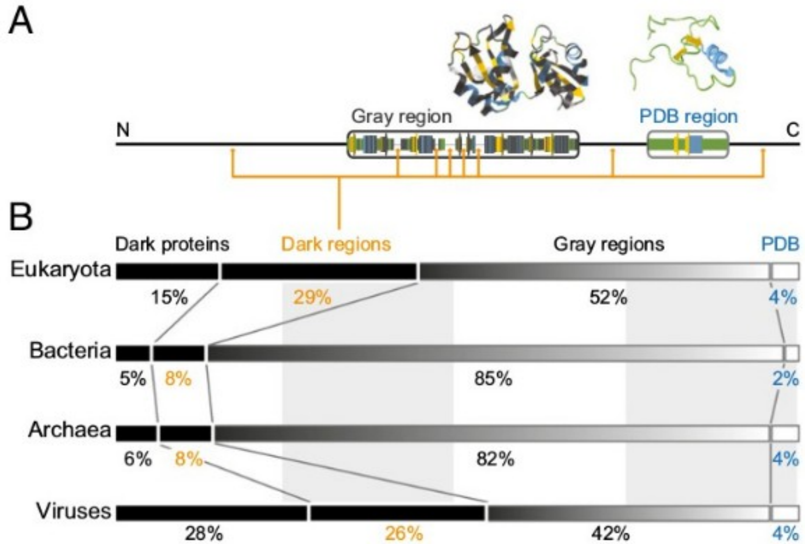


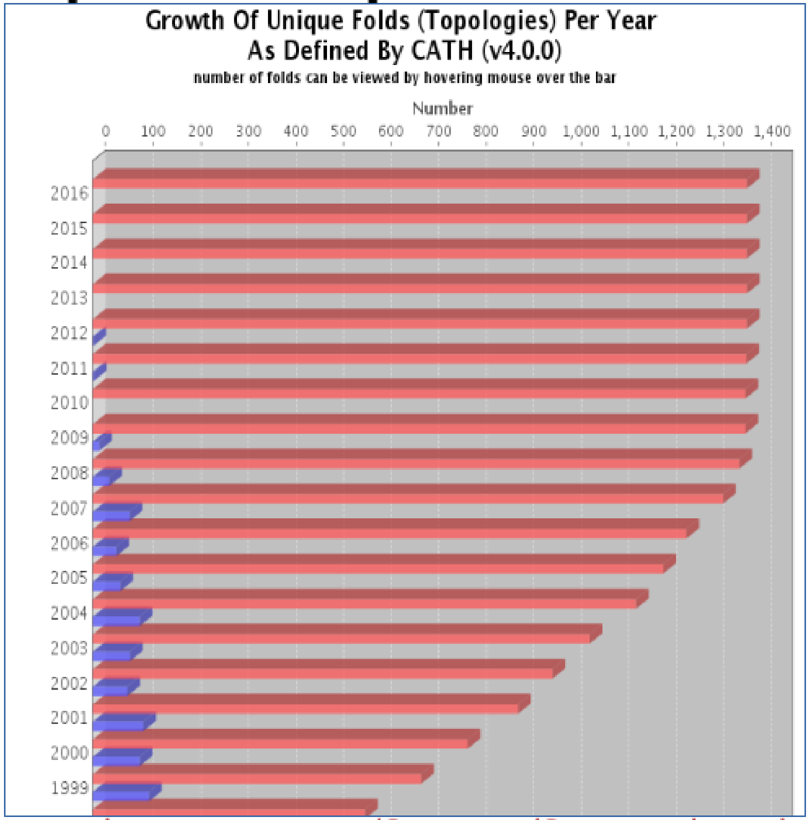
Fig. 1. Mapping the dark proteome. (*A*) For all proteins in Swiss-Prot, each residue was classified into one of four categories: (*i*) PDB regions—residues exactly matched to a PDB entry in Aquaria; (*ii*) gray regions—residues aligned to at least one PDB entry in Aquaria but always with amino acid 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, …)

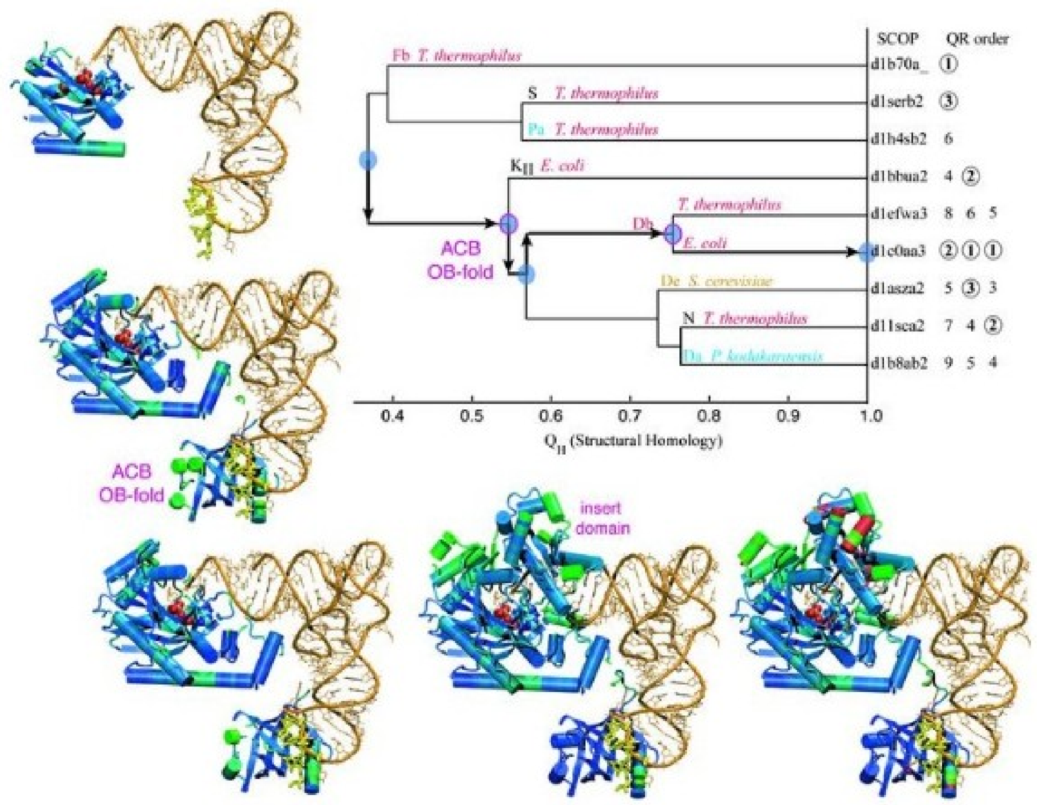
IV) Les macromolécules sont reliées entre elles,

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

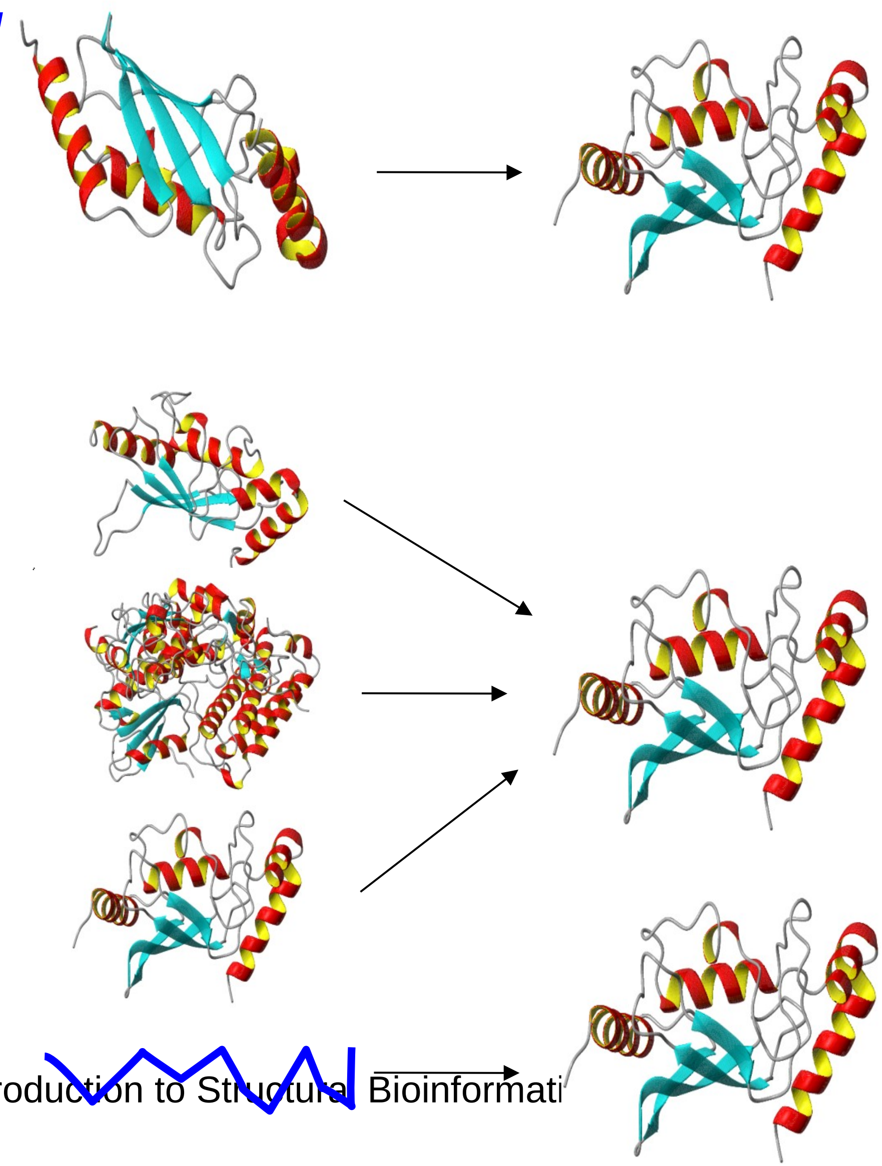
" Threading”

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

25

ab initio

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



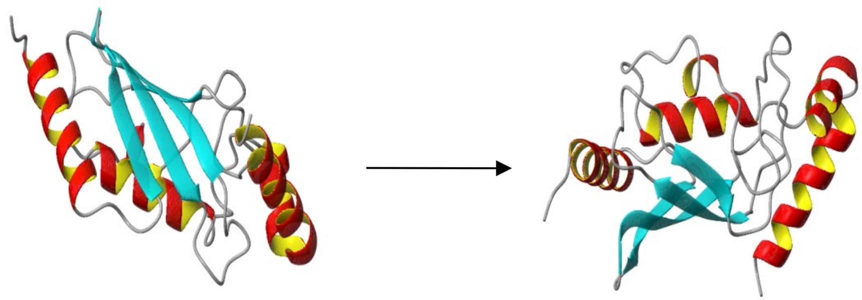
Comment ?

Fiabilité

100% id

Comparative / Homology Modelling

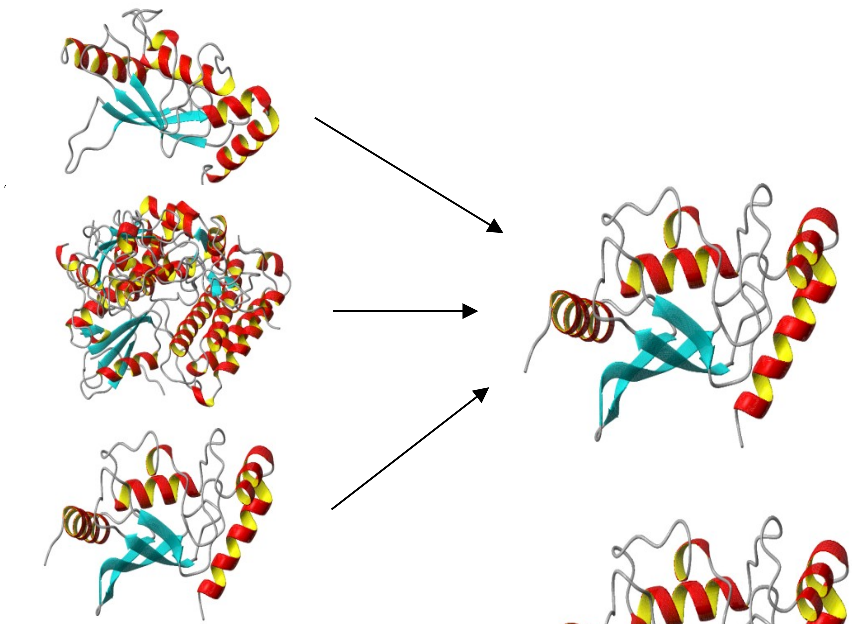
```
ATPLGLPTHVVVAGLNPHTRESD
ATPLGIPTHVPPAGLNPHTRESD
!!!!! !!!!   !!!!!!!!!!!
```



40

" Threading"

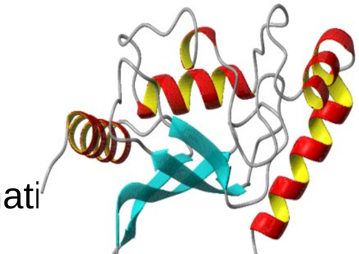
```
ETPLGLPTHVVVEGLNPHTRESD
IRVLGIPTHVPPIGLNPHTRIID
   !! !!!!   !!!!!!!   !
```



25

ab initio

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

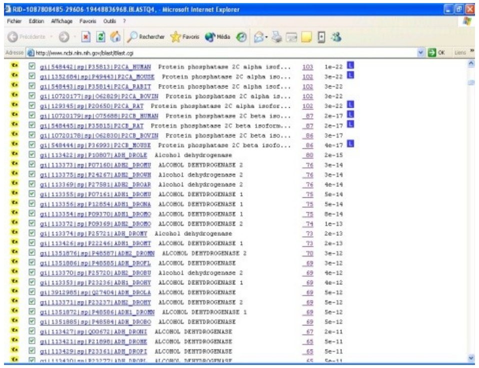


Modélisation par comparaison : MODELLER (%id 40 -100)

Sequence

SFITPVPGGVGPMTVFLEMDLTN
KNVIFVADKRKGGPGGIIANICV
HTFNSWLDVEPRVAIEANKNGAI
WKLDLAIWKLDLGTLEAIEWWDS
HIGAFLDKPKMENAQGQGNGLRY
GLSSDAHTAVIGLPSGLESAVIG
LPSGLESWSFFFAVYDGHAGSQV
AKY...

Search in Structure database

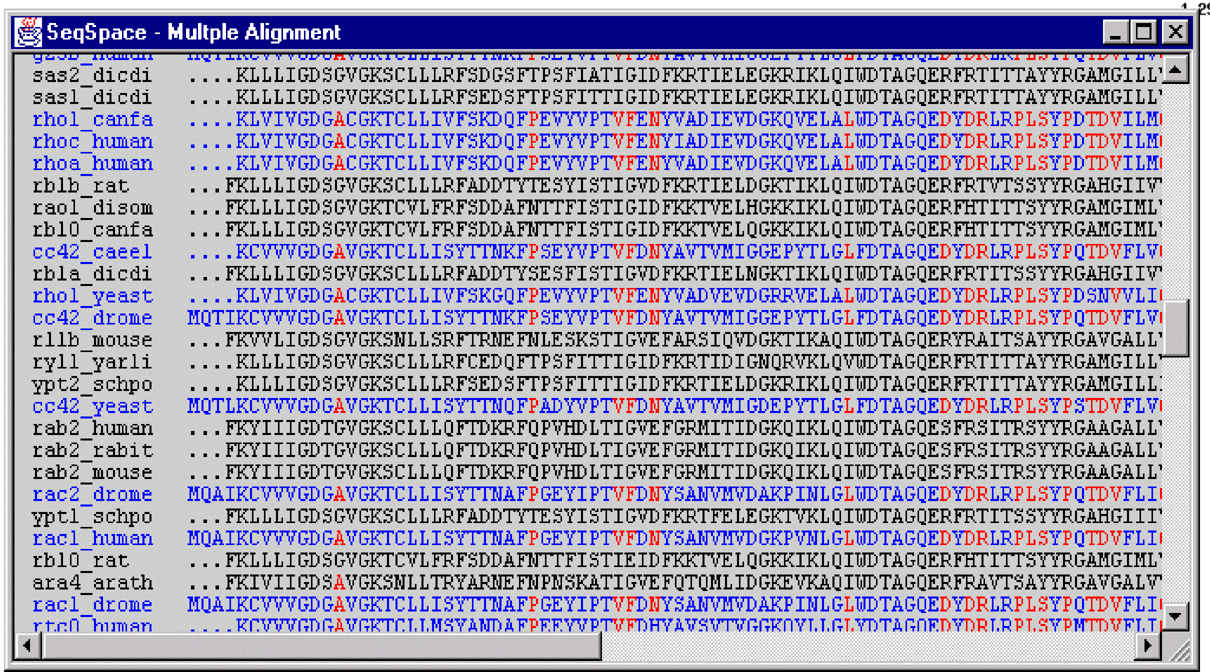


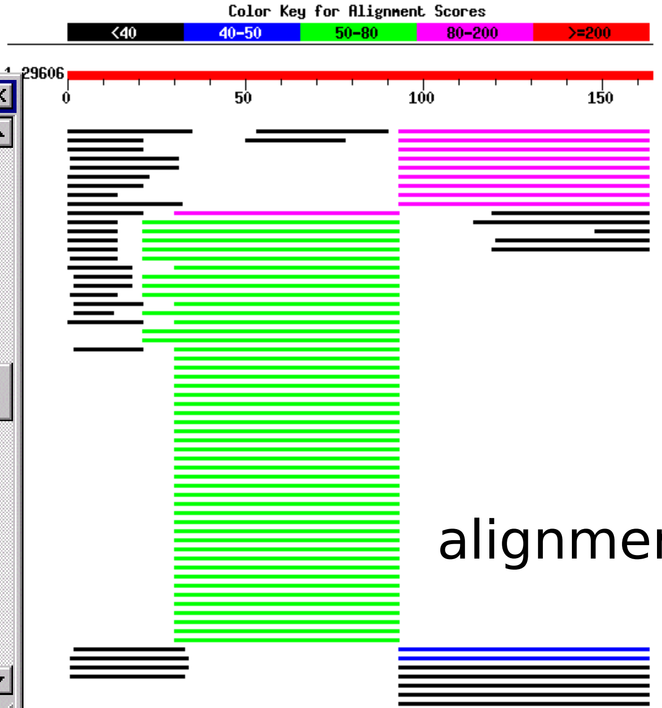
PDB

Multiple Sequences Alignment

N sequences

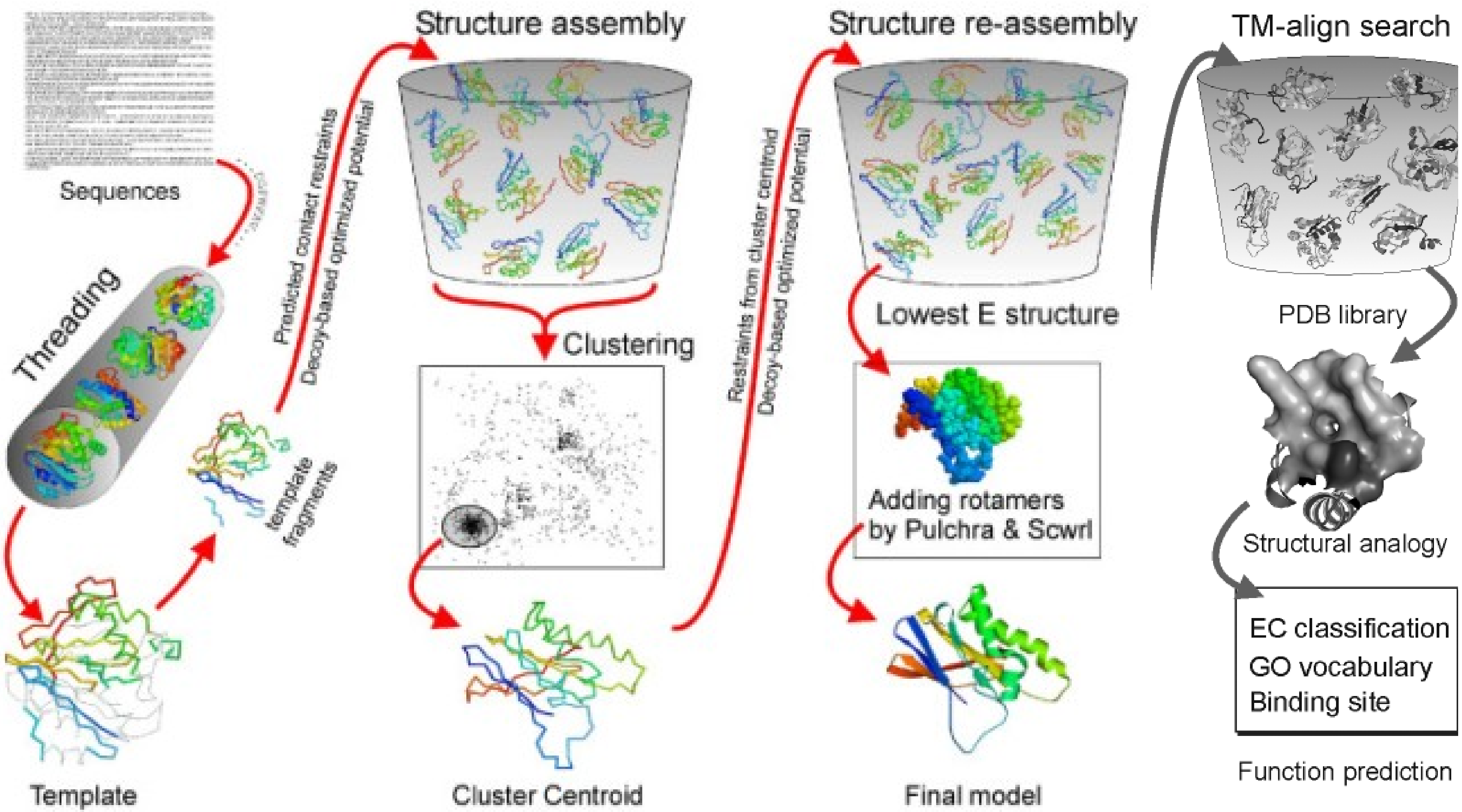
profile





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

Current Topic/Perspective

Biochemistry, Vol. 49, No. 14, 2010 2989

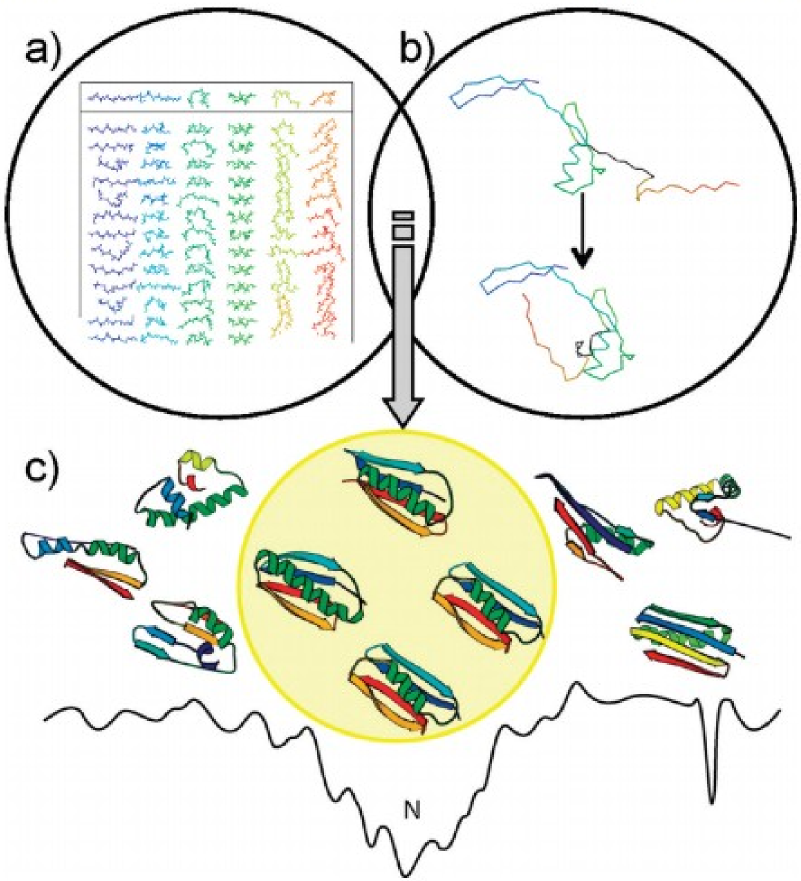
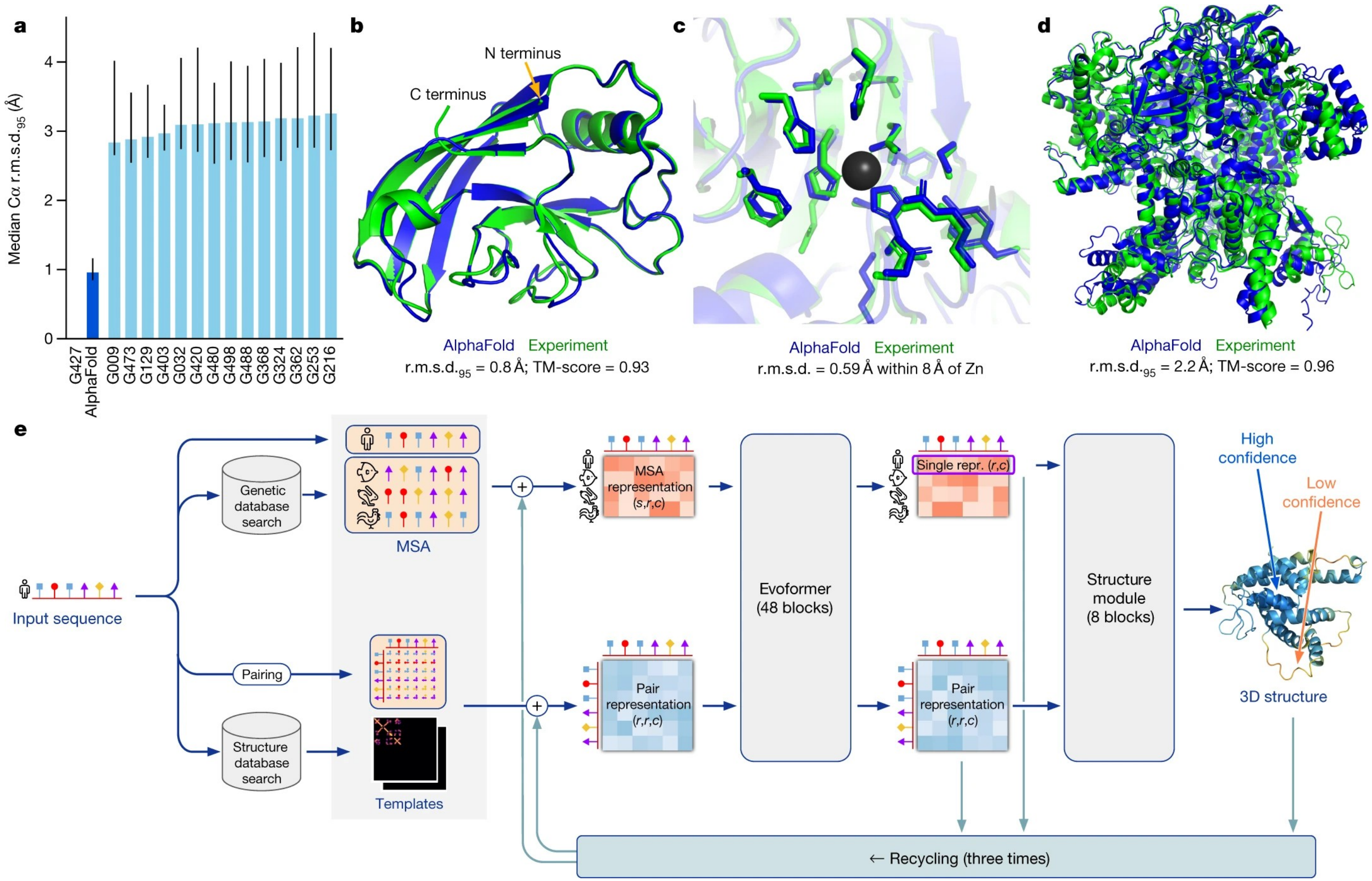


FIGURE 1: *De novo* folding algorithm. ROSETTA starts from (a) fragment libraries with sequence-dependent (φ and ψ) angles that capture the local conformational space accessible to a sequence. (b) Combining different fragments from the libraries folds the protein through optimization of non-local contacts. 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 »

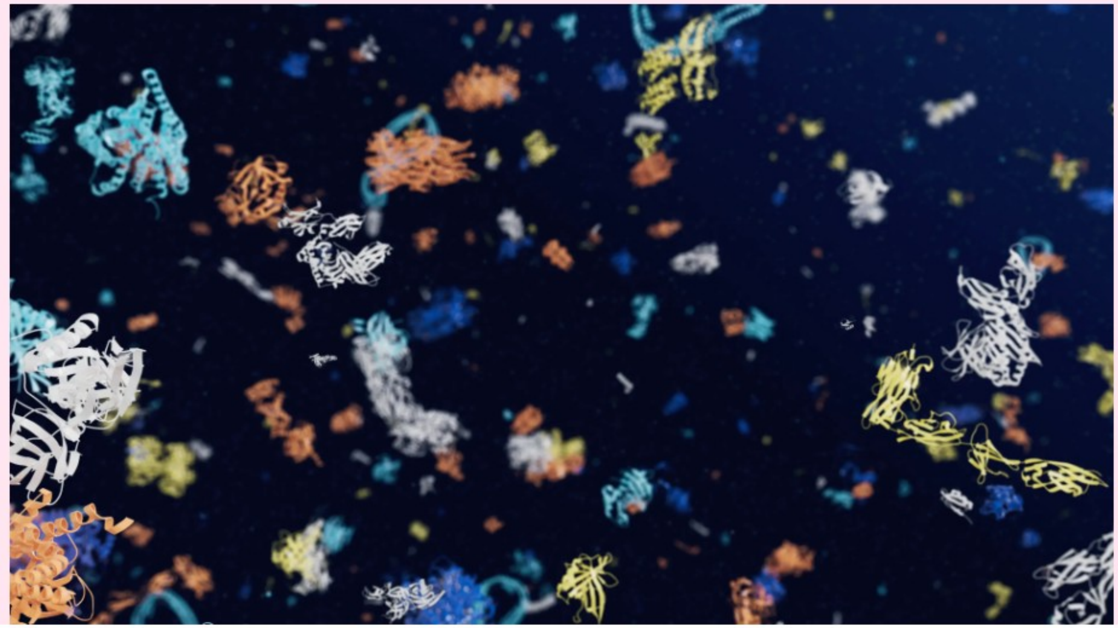
https://www.deepmind.com/blog/how-our-principles-helped-define-alphafolds-release **(September 14th, 2022)**

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

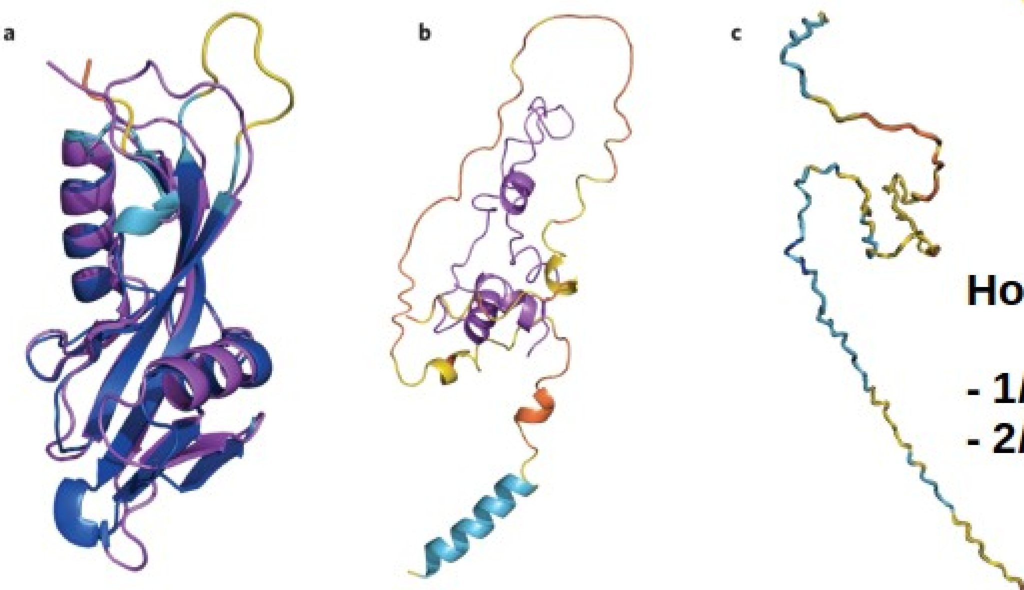
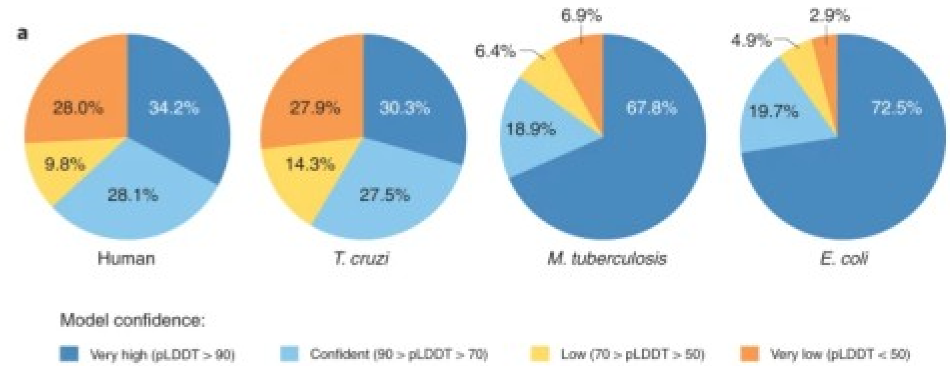


Fig. 2: Confidence scores for AlphaFold models.



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:

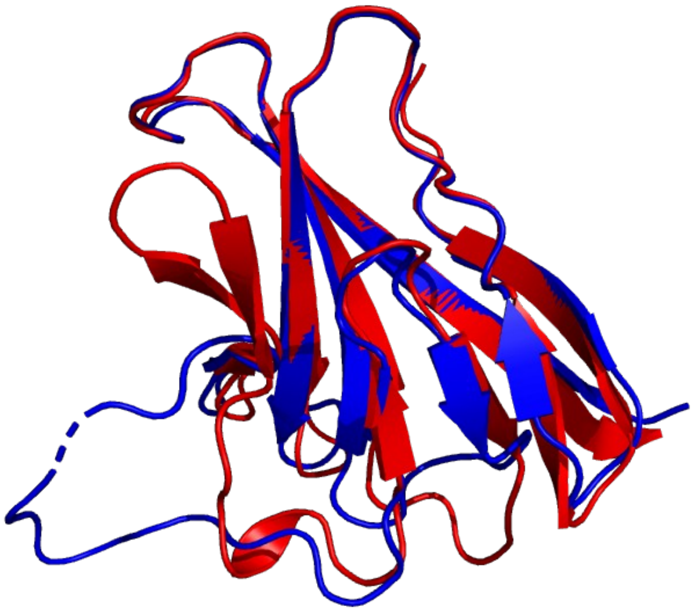
- Many proteins **function as complexes** with other proteins
- Proteins are **dynamic systems** and adopt different structures depending on their environment or state
- For regions that are **intrinsically disordered or unstructured in isolation**, AlphaFold is expected to produce a low-confidence prediction
- 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/

Once you have model…

Determine if a model is “good”

- Principle:
  from a known protein structure, compare the model to it (CASP)



Red: model predicted by AF (T1064)

Blue: experimental structure (7JTL)

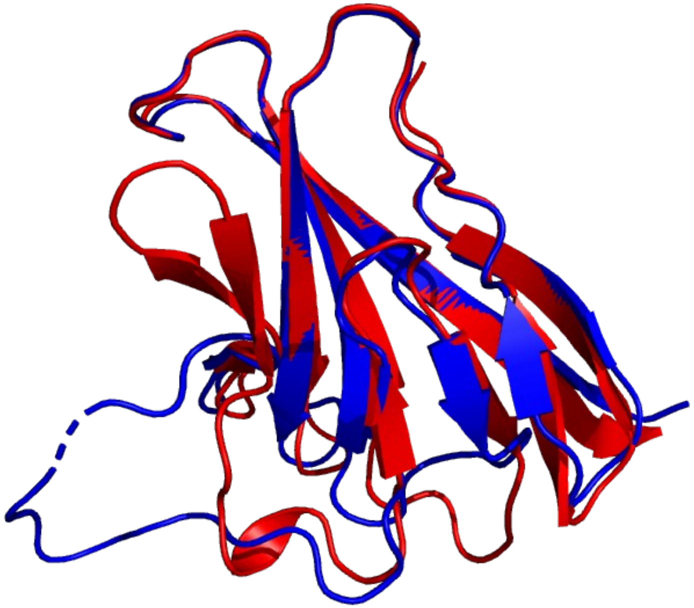
https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/

Metrics for comparison

- RMSD: Root Mean Square Deviation (Distance)
- GDT: Global Distance Test, RMSD per interval (<2, <4A, <6, …)
- Ramachandran Outliers
- Clashscores
- And more…

RMSD and GDT

- Compare the position of all atoms to see if the model and the reference structure deviate
- Advanteage: one measure (In Angstroems)
- Inconvenient: sensitive to small changes



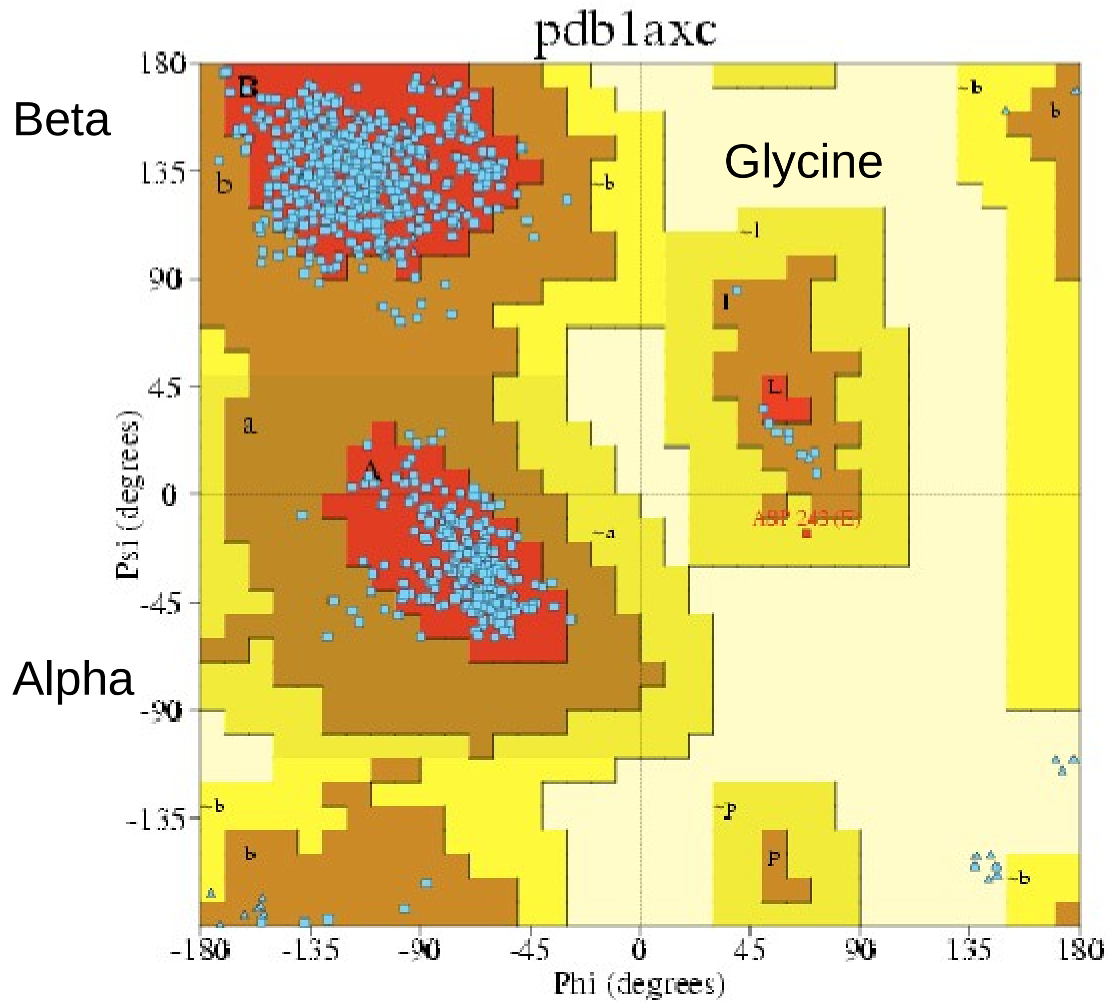
$$\sqrt{\frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{n}}$$

https://bioinfo-fr.net/comparaison-de-structures-le-rmsd

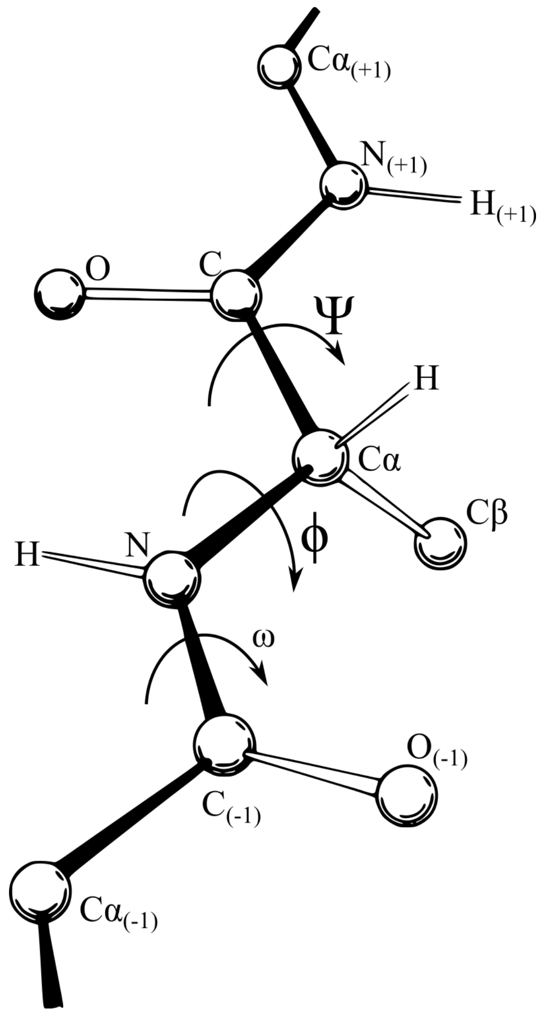
https://foldit.fandom.com/wiki/RMSD

https://foldit.fandom.com/wiki/GDT

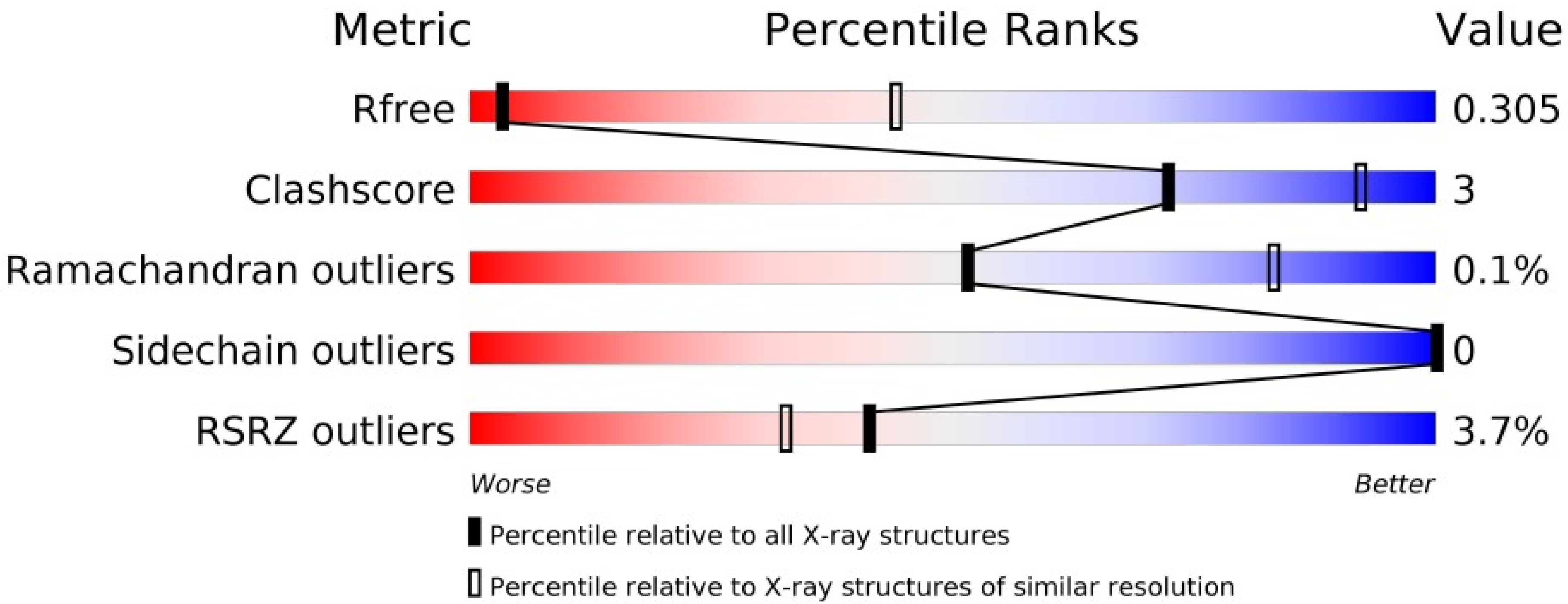
Ramachandran outliers (proteins)



https://en.wikipedia.org/wiki/Ramachandran_plot



Clashscores, and more…



Example for structure 5NWL, human RAD51-ATP filament

https://www.rcsb.org/structure/5NWL

How to assess them?

- MolProbity, ProCheck, Prosa-web, Verify3D, PDBsum, WHAT_IF, WHAT_CHECK
- And more…

Structure Validation and Quality

Access additional resources of interest, submitted by community members.

CheckMyMetal	A service checks metal binding site and validation
MolProbity	Structure validation on client-uploaded or PDB ID-specified files, using all-atom contact analysis tools and updated geometrical criteria for phi/psi, sidechain rotamer, and Cbeta deviation
NQ-Flipper	NQ-Flipper recognizes unfavorable rotamers of Asn and Gln residues in protein structures obtained from X-ray crystallography, NMR or modelling studies.
Procheck	A program that checks the stereochemical quality of a protein structure
Prosa-web	Quality scores of a protein are displayed in the context of all known protein structures and problematic parts of a structure are shown and highlighted in a 3D molecule viewer.
Verify3D Structure Evaluation Server	Determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D) by assigning a structural class based on its location and environment (alpha, beta, loop, polar, nonpolar etc) and comparing the results to good structures.
WHAT IF	A protein structure analysis program that may be used for mutant prediction, structure verification and molecular graphics
WHAT_CHECK	A system for protein structure validation derived from the WHAT IF program

https://www.rcsb.org/docs/additional-resources/structure-validation-and-quality

Why AlphaFold is good?

Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.

pLDDT: **predicted** local distance difference test