

Introduction à la bioinformatique (SSV3U15)

Chapitre 2. Séquence - structure - fonction

Jacques van Helden (Aix-Marseille Université)

ORCID [0000-0002-8799-8584](https://orcid.org/0000-0002-8799-8584)

Caractérisation de la structure tridimensionnelle des protéines

Table des matières du chapitre “Séquence - structure - fonction”

1. Les premières structures de protéines (Kendrew 1957, Perutz 1959)
2. Méthodes de caractérisation des structures protéiques
3. Eléments de structure
4. Ressources bioinformatiques pour l'analyse des séquences, structures et fonctions des protéines
5. Visualisation des structures protéiques
6. Relations séquence - structure - fonction : quelques exemples
7. Prédiction de la structure des protéines à partir de la séquence
8. Evaluation des outils de prédiction: CASP
9. Utilisation de l'intelligence artificielle pour prédire les structures protéiques

Les premières structures de protéines

- 1957: John Cowdery Kendrew résout la structure tridimensionnelle de la myoglobine de cachalot
- 1959: Max Ferdinand Perutz résout la structure 3D de l'hémoglobine

The Nobel Prize in Chemistry 1962

Summary

Laureates

Max F. Perutz
John C. Kendrew

Speed read

Perspectives

Award ceremony video

Presentation Speech

Share this

Photo from the Nobel Foundation archive.
Max Ferdinand Perutz
Prize share: 1/2

Photo from the Nobel Foundation archive.
John Cowdery Kendrew
Prize share: 1/2

The Nobel Prize in Chemistry 1962 was awarded jointly to Max Ferdinand Perutz and John Cowdery Kendrew "for their studies of the structures of globular proteins"

To cite this section
MLA style: The Nobel Prize in Chemistry 1962. NobelPrize.org. Nobel Prize Outreach AB 2024. Sun, 28 Jul 2024.
<<https://www.nobelprize.org/prizes/chemistry/1962/summary/>>

- 1.Kendrew, J. C. et al. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* 181, 662–666 (1958).
- 2.Perutz, M. F. et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. *Nature* 185, 416–422 (1960).
- 3.<https://www.nobelprize.org/prizes/chemistry/1962/summary/>

Structure de l'hémoglobine de cheval (Perutz, 1958)

Voici le premier modèle publié de la structure tridimensionnelle de l'hémoglobine.

Haut : complexe protéique composé de 2 sous-unités alpha et 2 sous-unités beta (empilements). Les disques gris représentent le groupe hème.

Bas: schéma représentant la configuration des deux sous-unités de face. Les cylindres représentent les groupes hème.

Cet article couronne un travail de longue haleine

- 1936 : Perutz démarre une thèse de doctorat visant à caractériser la structure de l'hémoglobine du cheval
- 1940 : il n'a pas encore atteint cet objectif, mais ses premiers résultats lui permettent de soutenir une thèse
- Pendant les années 1950, il continue à progresser
- 1959 : il caractérise la structure (publication 1960)
- 1962: prix Nobel de chimie

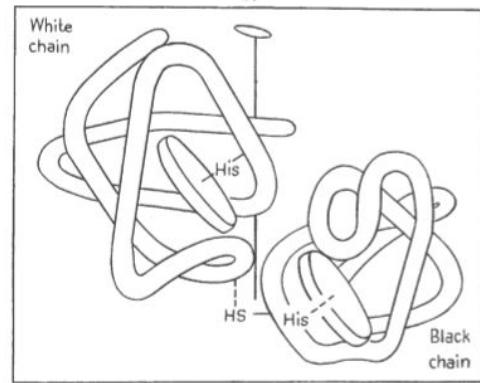
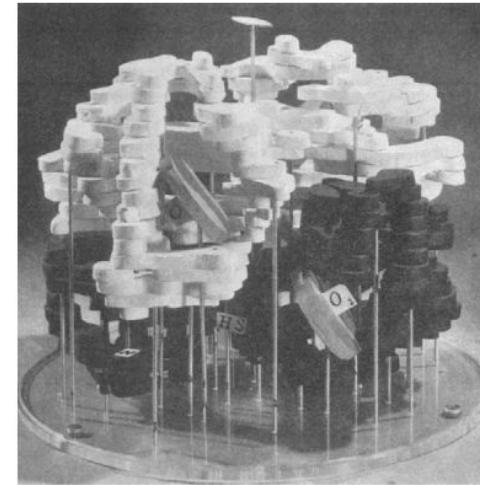


Fig. 8. (a) Hemoglobin model viewed normal to α . The heme groups are indicated by grey disks. (b) Chain configuration in the two sub-units facing the observer. The other two chains are produced by the operation of the dyad axis

Structure de la myoglobine de cachalot (Kendrew, 1958)

Voici le premier modèle publié d'une structure complète de protéine (Kendrew, 1958) : la myoglobine du cachalot.

La figure montre 4 photographies (sous des angles différents) d'un modèle de la molécule.

Kendrew a entrepris son projet en 1947.

Le choix de la protéine était judicieux :

- La myoglobine ne comporte qu'une seule chaîne polypeptidique, alors que l'hémoglobine est un tétramère (polymère de 4 polypeptides).
- La myoglobine, qui permet la rétention de l'oxygène dans les tissus musculaires, est plus abondante et chez les organismes marins.
- Par chance, Kendrew a eu l'occasion de disposer d'un gros échantillon de muscle de de cachalot en provenance du Pérou, et il a saisi cette opportunité.

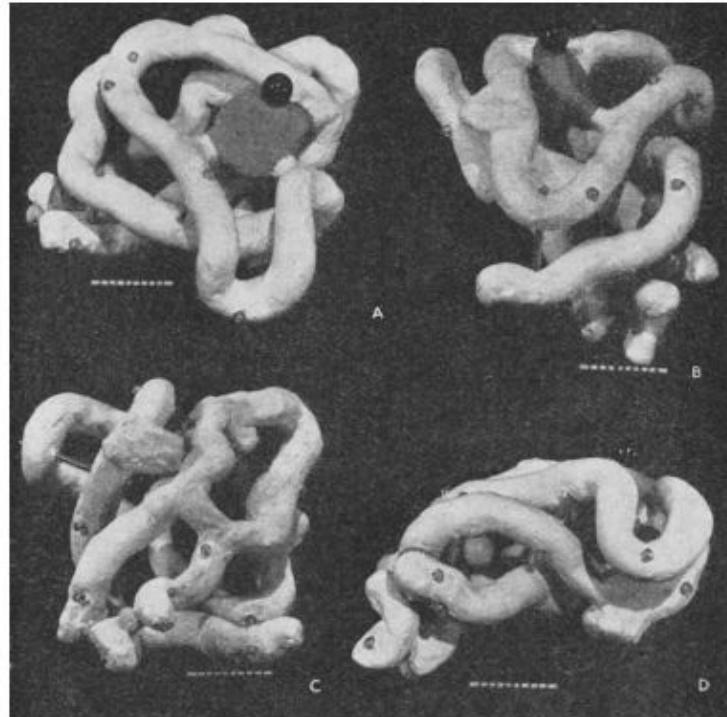


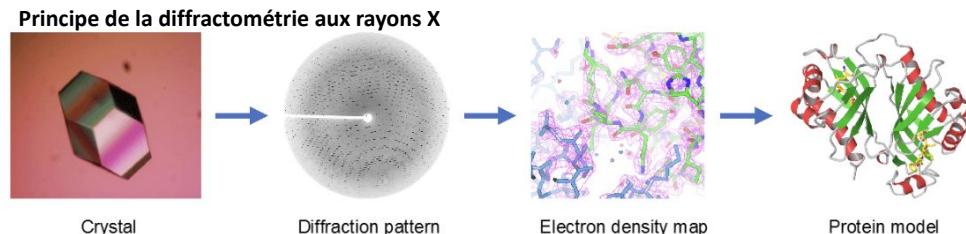
Fig. 2. Photographs of a model of the myoglobin molecule. Polypeptide chains are white; the grey disk is the heme group. The three spheres show positions at which heavy atoms were attached to the molecule (black: Hg of *p*-chloro-mercuri-benzene-sulphonate; dark grey: Hg of mercury diammine; light grey: Au of auri-chloride). The marks on the scale are 1 Å. apart

Kendrew et al. (1958). <https://doi.org/10.1038/181662a0>

Méthodes de caractérisation des structures de protéines

Diffractométrie aux rayons X (cristallographie)

- Méthode “historique” (Sanger, Kendrew, Perutz, ...)
- Bonne résolution (Angstrom)
- Fonctionne pour des protéines de grande taille ou pour des complexes (protéine, protéine-ADN)
- Requiert un travail de biochimie conséquent pour obtenir un cristal de haute qualité, ce qui peut représenter des années d'efforts, sans succès garanti



https://www.creative-biostructure.com/x-ray-crystallography-platform_60.htm

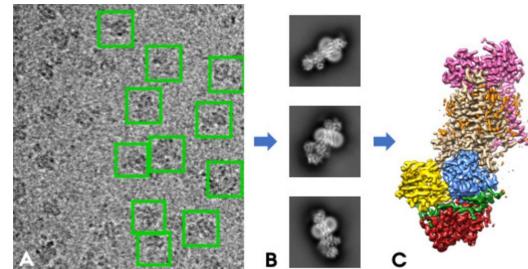
Spectroscopie RMN

- Analyse de protéines en solution
- Fournit des informations dynamiques (flexibilité, conformations alternatives de la protéine)
- Limitation de la taille des protéines étudiées

Cryo-microscopie électronique

- Pas besoin de cristal
- Nécessite beaucoup moins de quantité de protéine
- Initialement, résolution inférieure à la cristallographie mais progrès récents → égale voire dépasse les rayons X
- Ne fonctionne qu'avec protéines de grande taille (pour pouvoir les détecter)

Principe de la cryo-microscopie électronique. Reconstruction d'une structure 3D à partir d'un très grand nombre de photos 2D de piétre résolution révélant la protéine sous différents angles.



<https://mbg.au.dk/en/news-and-events/news-item/artikel/forskere-bestemmer-foerste-struktur-af-protein-som-opretholder-cellemembranen/>

Eléments de structure

Structure primaire

- Séquences des acides aminés

Structure secondaire

- Reploiement local de la chaîne des acides aminés
- Structures secondaires fréquentes (illustrées plus loin)
 - Hélice alpha
 - Feuillet bêta (anti-parallèle ou parallèle)

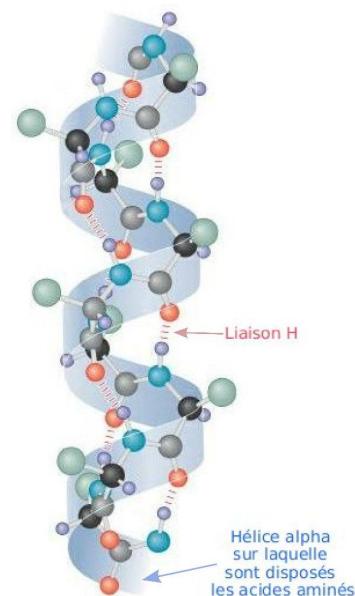
Structure tertiaire

- Structure de la protéine dans l'espace (=structure tridimensionnelle)
- Résulte des interactions entre les atomes de la protéine, et d'interactions avec son environnement (cytoplasme, membrane)

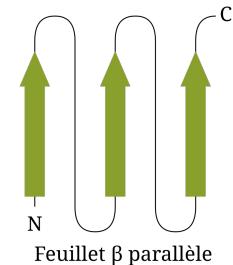
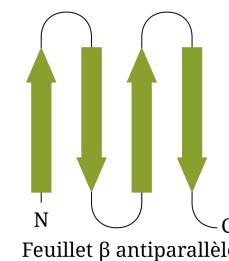
Structure quaternaire

- Structure tridimensionnelle résultant de l'agrégation de chaînes polypeptidiques en complexes protéiques

Hélice alpha



Feuilles bêta



Ressources bioinformatiques pour l'analyse des protéines



De Swiss-prot à Uniprot-KB

- En 1984, Amos Bairoch entreprend de créer une base de données de séquences protéiques qui puisse être utilisée sur un ordinateur personnel (à l'époque les données n'étaient accessibles que sur des serveurs).
- Il ajoute aux données de séquence des “annotations” : une fiche qui synthétise, de façon structurée, les informations biologiques de chaque protéine (fonction, domaines structurels, mutations, ...).
- 1986: ouverture de la base de données **Swiss-Prot**, avec ~ 3900 protéines
- 1988 : collaboration entre Swiss-Prot et le European Molecular Biology Laboratory (EMBL).
- Fin des années 1990: l'équipe Swiss-Prot emploie ~100 personnes, réparties entre la Suisse et l'European Bioinformatics Institute (situé en Angleterre)
- 2004: Swiss-Prot est renommée Uniprot

Uniprot aujourd’hui

- Uniprot (www.uniprot.org) continent aujourd’hui aujourd’hui (4 septembre 2024) la séquence de 245 millions de protéines.
- Lénorme majorité de ces protéines n'a jamais fait, et ne fera jamais, l'objet d'études expérimentales pour caractériser leur fonction.
- Parmi ces protéines, seule une “petite” fraction de 571.864 ont été annotées par des êtres humains dans la base de connaissance Swiss-Prot.
- Les autres sont annotées automatiquement (TrEMBL), par des méthodes bioinformatiques qui reposent sur leur similarité de séquence avec des protéines connues.



Status

Reviewed (Swiss-Prot)
(571,864)
Unreviewed (TrEMBL)
(245,324,902)

Uniprot - Annotation de la chaîne alpha de l'hémoglobine de cheval

- <https://www.uniprot.org/uniprotkb/P01958/>
- Information détaillée, structurée en sections
 - Nom, taxonomie
 - Localisation cellulaire
 - Phénotypes et variants (génétiques)
 - ...

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search Help

P01958 · HBA_HORSE

Function	P01958 · HBA_HORSE	
Names & Taxonomy	Protein	Hemoglobin subunit alpha
Subcellular Location	Gene	HBA
Phenotypes & Variants	Status	UniProtKB reviewed (Swiss-Prot)
PTM/Processing	Organism	Equus caballus (Horse)
Expression	Amino acids	142 (go to sequence)
Interaction	Protein existence ⁱ	Evidence at protein level
Structure	Annotation score ⁱ	(56)
Family & Domains	Entry Variant viewer Feature viewer Genomic coordinates Publications External links Hi	
Sequence	Tools Download Add Add a publication Entry feedback	
Similar Proteins	Feedback Help	

Functionⁱ
Involved in oxygen transport from the lung to the various peripheral tissues.

Hemopressin
Hemopressin acts as an antagonist peptide of the cannabinoid receptor CNR1. Hemopressin-binding efficiently blocks cannabinoid receptor CNR1 and subsequent signaling. By Similarity

Features
Showing features for binding siteⁱ.

1 10 20 30 40 50 60 70 80 90 100 110 120 130 140 142

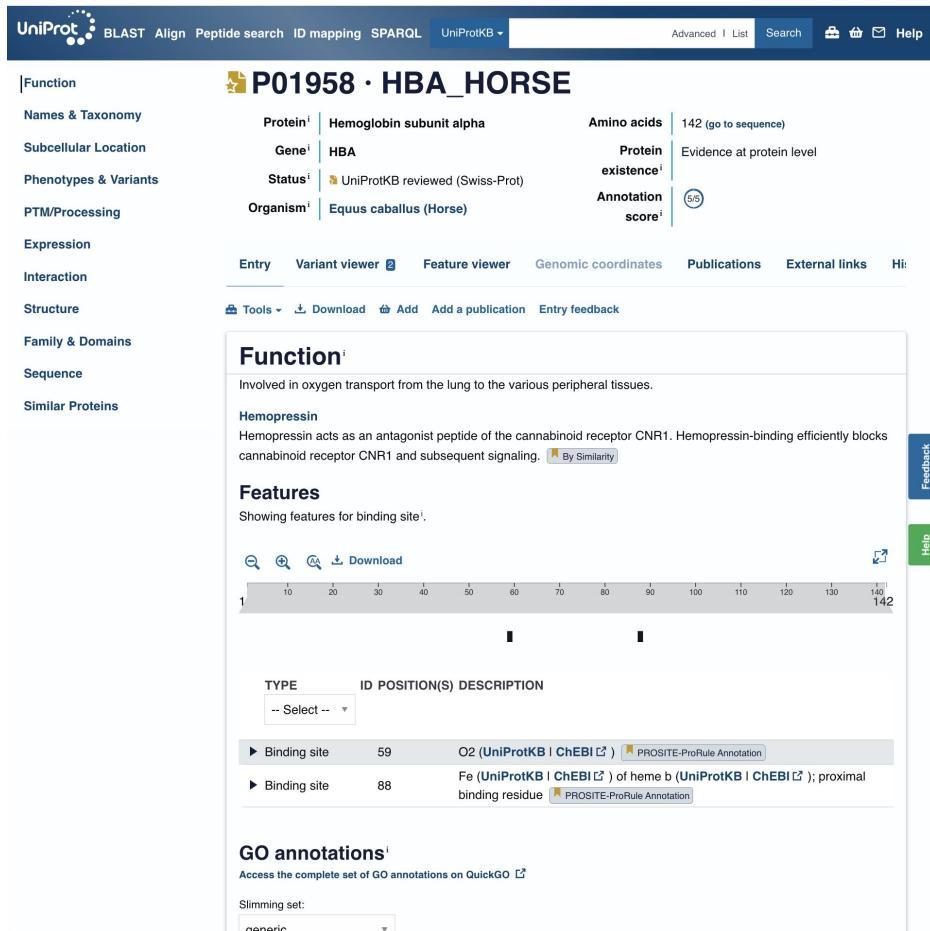
TYPE ID POSITION(S) DESCRIPTION
-- Select -- ▾

▶ Binding site 59 O2 (UniProtKB | ChEBI) PROSITE-ProRule Annotation

▶ Binding site 88 Fe (UniProtKB | ChEBI) of heme b (UniProtKB | ChEBI); proximal binding residue PROSITE-ProRule Annotation

GO annotationsⁱ
Access the complete set of GO annotations on QuickGO ▾

Slimming set:
generic ▾



<https://www.uniprot.org/uniprotkb/P01958/>

Uniprot - Annotation de la chaîne alpha de l'hémoglobine de cheval

- <https://www.uniprot.org/uniprotkb/P01958/>
- Information détaillée, structurée en sections
 - Nom, taxonomie
 - Localisation cellulaire
 - Phénotypes et variants (génétiques)
 - ...
- Liens vers les structures de PDB
- Annotation des domaines
- Séquence de la protéine
- Un tas d'autres informations pertinentes

Au TP, vous apprendrez à utiliser la base de données Swiss-Prot / Uniprot-KB.

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	1NS9	X-ray	2.00 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	1Y8H	X-ray	3.10 Å	A/C	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	1Y8I	X-ray	2.60 Å	A/C	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	1Y8K	X-ray	2.30 Å	A/C	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	2D5X	X-ray	1.45 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	2DHB	X-ray	2.80 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	2MHB	X-ray	2.00 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	2ZLT	X-ray	1.90 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	2ZLU	X-ray	2.00 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	2ZLV	X-ray	2.00 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek

Protein Data Bank (PDB)

- La structure tridimensionnelle des protéines régit leur fonction : leur forme détermine la façon dont les acides aminés pourront interagir avec les autres molécules et composantes de la cellule.
 - Transporteurs: insertion dans la membrane et transport de petites molécules
 - Enzymes: interactions ave un groupe de molécules (substrats) et catalyse d'une réaction qui produira d'autres molécules
 - Polymérase de l'ADN : interaction avec l'ADN, "lecture" de sa séquence et catalyse de la biosynthèse de l'ARN.
 - Facteurs transcriptionnels: interaction avec l'ADN, et avec la polymérase de l'ARN
 - ...
- Protein Data Bank (www.rcsb.org)** contient à ce jour (4 septembre 2024)
 - 224,572 structures caractérisées expérimentalement
 - 1.068.577 modèles prédictifs



RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19 MyPDB Contact us

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

PDB-101 PDB EMDataResource NAKB wwwPDB Foundation PDB-Dev

Access Computed Structure Models (CSMs) of available model organisms Learn more

Welcome Deposit Search Visualize Analyze Download Learn

RCSB Protein Data Bank (RCSB PDB) enables breakthroughs in science and education by providing access and tools for exploration, visualization, and analysis of:

- Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive
- Computed Structure Models (CSM) from AlphaFold DB and ModelArchive

These data can be explored in context of external annotations providing a structural view of biology.

Explore NEW Features PDB-101 Training Resources

September Molecule of the Month

Carbon Capture Mechanisms

Latest Entries As of Tue Sep 03 2024

8WLT Cryo-EM structure of the membrane-anchored part of the flagellar motor-hook complex in the CCW state

Features & Highlights

Register for the Sept 18 Virtual Office Hour on Molecular Animations
Join us to learn more about RCSB PDB's molecular animations

Register for the Sept 12 Virtual Office Hour: Streamlining PDB Deposition
Join our Biocurators to learn about preparing files and supplementary information, starting a deposition session, navigating OneDep, and more.

News Publications

Poster Prize Awarded at ISMB 2024
Congratulations to Yehlin Cho for Enhancing Protein Design Robustness through Noise-Informed Sequence Design
x 09/03/2024

Biocurator Milestone: >10,000 Depositions Processed
Congratulations to wwPDB's Irina Persikova on processing more than 10,000 PDB structures
x 08/28/2024

Paper Published: Exploring protein 3D similarities via comprehensive structural alignments

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), Entrez Include CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-Resoures NAKB wwwPDB Foundation PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Phyceret catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.C.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

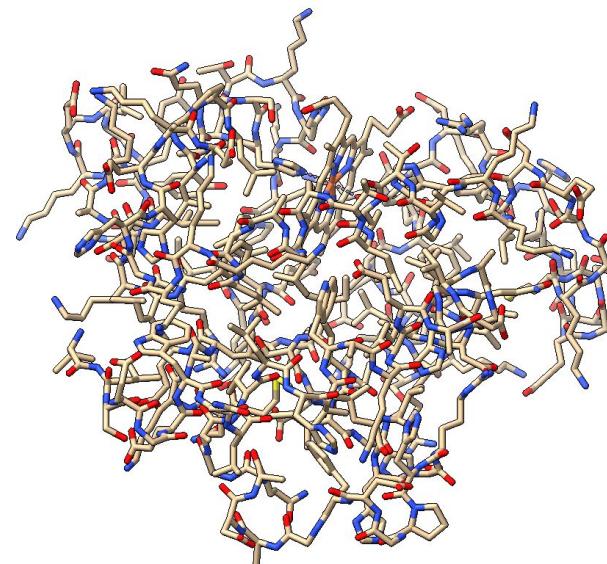
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN) Vue des liaisons atomiques en bâtonnets ("sticks")



Beige: carbone

Bleu: azote

Rouge: oxygène

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-Resou NAKB wwPDB Foundation PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Phycerter catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.C.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

Find Similar Assemblies

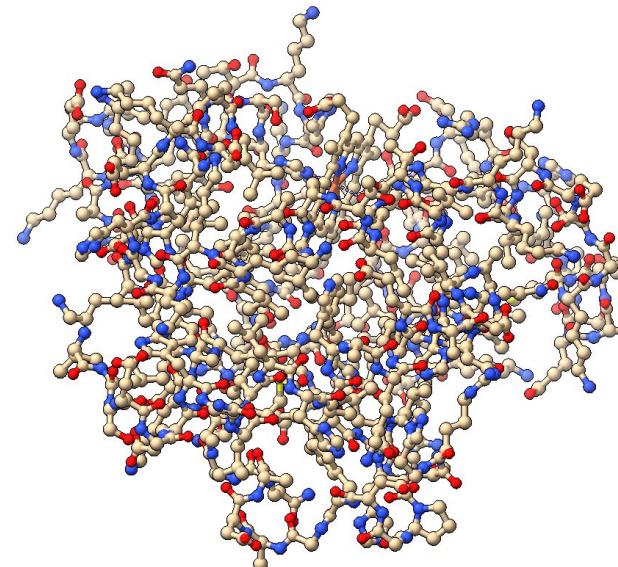
Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN)

Vue des atomes + liaisons atomiques
("balls and sticks")



Beige: carbone

Bleu: azote

Rouge: oxygène

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit ▾ Search ▾ Visualize ▾ Analyze ▾ Download ▾ Learn ▾ About ▾ Documentation ▾ Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez Include CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-Resourse NAKB wwwPDB Foundation PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Phyceret catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.C.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

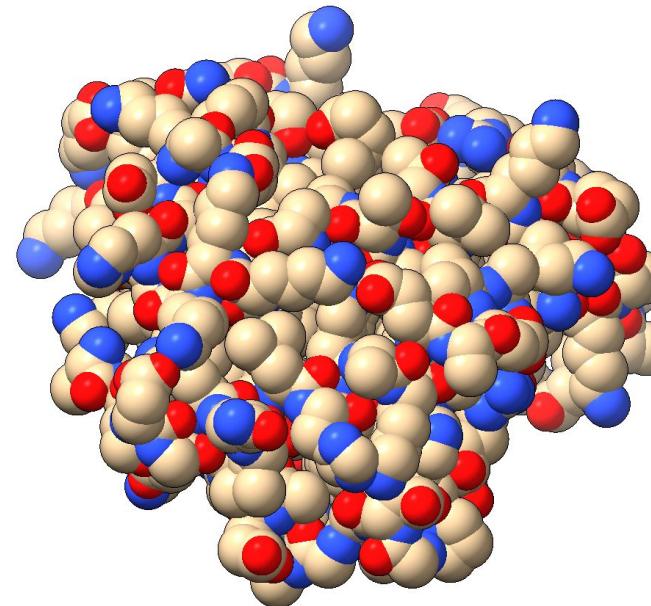
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN) Espace occupé par chaque atome ("spheres")



Beige: carbone

Bleu: azote

Rouge: oxygène

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), CSM Advanced Search | Browse Annotations Help

PDB-101 PDB EMDataResource NAKB wwwPDB Foundation PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Phyceret catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.C.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

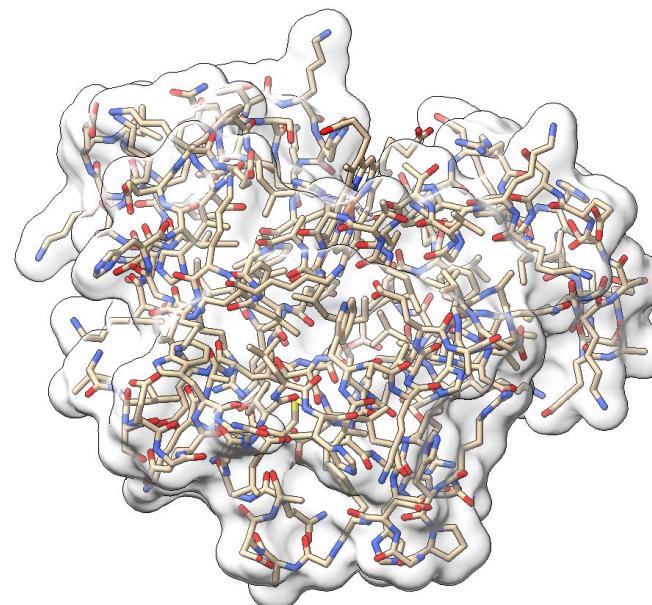
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN) ("Ghostly white")



Beige: carbone
Bleu: azote
Rouge: oxygène

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), Entrez Include CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-Reserve NAKB wwwPDB Foundation PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Physeter catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.G.

Experimental Data Snapshot

wwPDB Validation

Method: X-RAY DIFFRACTION Resolution: 2.00 Å

Metric	Percentile Ranks	Value
Claesson	3.34	3.34
Ramachandran outliers	3.3%	3.3%
Sidechain outliers	15.2%	15.2%

This is version 1.4 of the entry. See complete history.

Literature

The Stereochemistry of the Protein Myoglobin
Watson, H.C.
(1969) Prog Stereochem 4: 299

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

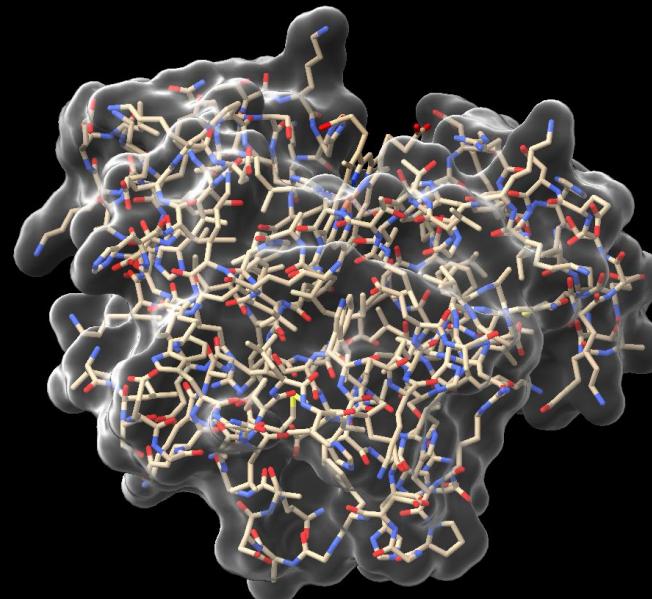
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN) ("Ghostly white")



Beige: carbone
Bleu: azote
Rouge: oxygène

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), Entrez Include CSM Advanced Search | Browse Annotations Help

PDB-101 PDB EMDataResource NAKB wwwPDB Foundation PDB-Dev

Facebook Twitter LinkedIn

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Phyceret catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.G.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

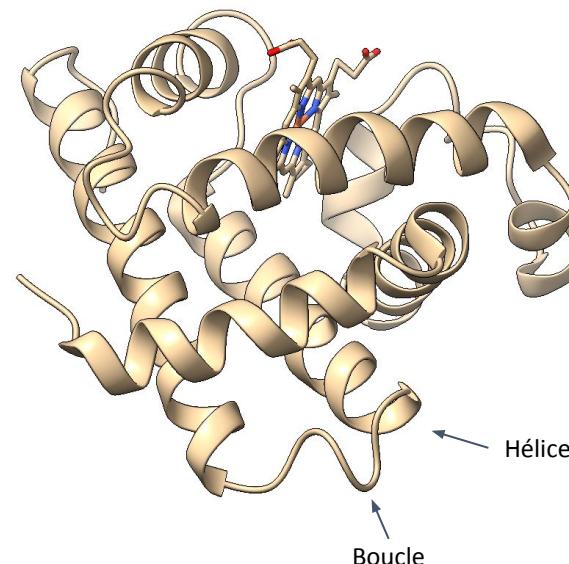
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN) Structures secondaires (vue "ribbons/slabs")



Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

PDB PROTEIN DATA BANK 224,572 Structures from the PDB 1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), Entrez Include CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-Resou NAKB wwwPDB PDB-Dev

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1 1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE Organism(s): Phyceret catodon Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19 Deposition Author(s): Watson, H.C., Kendrew, J.C.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1 Global Stoichiometry: Monomer - A1

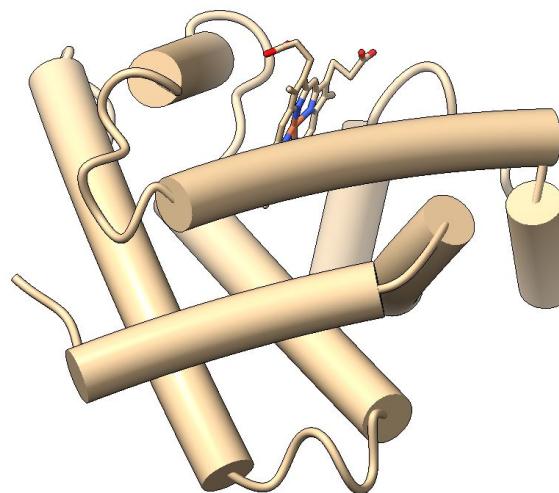
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

Structure de la myoglobine (1MBN) Hélices schématisées en cylindres (“cylinders/stubbs”)



Beige: carbone
Bleu: azote
Rouge: oxygène

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

Deposit Search Visualize Analyze Download Learn About Documentation Careers COVID-19

224,572 Structures from the PDB
1,068,577 Computed Structure Models (CSM)

3D Structures Enter search term(s), Entrez search term(s), Entrez Include CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-DataResource NAKB wwwPDB Foundation PDB-Dev

Facebook Twitter LinkedIn

Structure Summary Structure Annotations Experiment Sequence Genome Versions

Biological Assembly 1

1MBN

The stereochemistry of the protein myoglobin

PDB DOI: <https://doi.org/10.2210/pdb1MBN/pdb>

Classification: OXYGEN STORAGE
Organism(s): Phyceret catodon
Mutation(s): No

Deposited: 1973-04-05 Released: 1976-05-19
Deposition Author(s): Watson, H.C., Kendrew, J.G.

Explore in 3D: Structure | Sequence Annotations | Validation Report | Ligand Interaction (HEM)

Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

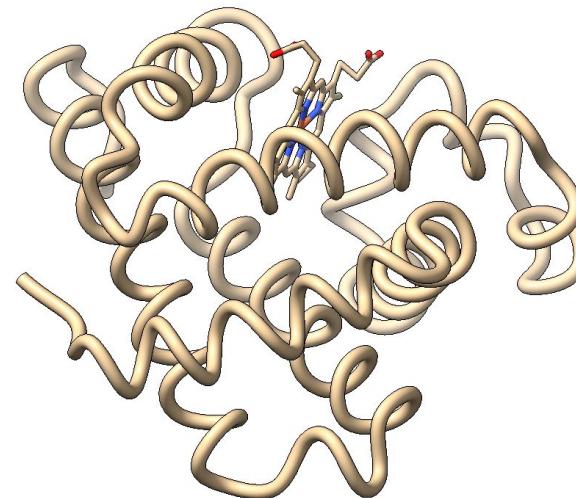
Find Similar Assemblies

Biological assembly 1 assigned by authors.

Macromolecule Content

- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

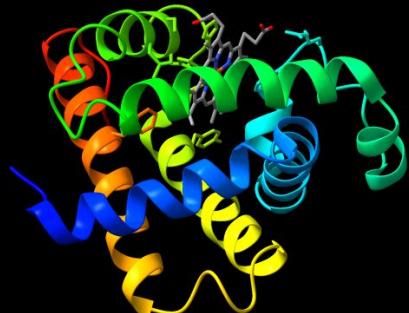
Structure de la myoglobine (1MBN) ("Licorice / ovals")



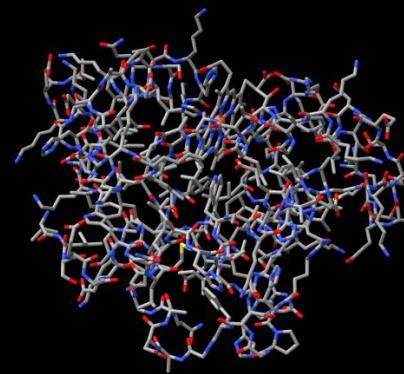
Beige: carbone
Bleu: azote
Rouge: oxygène

6 vues de la myoglobine

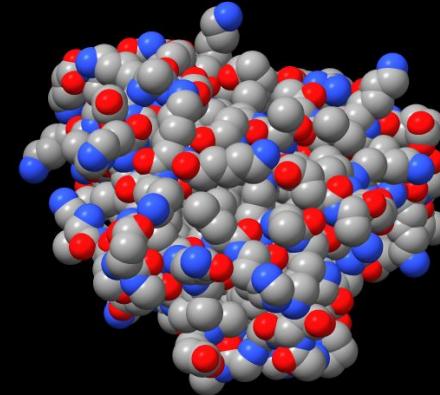
Rainbow



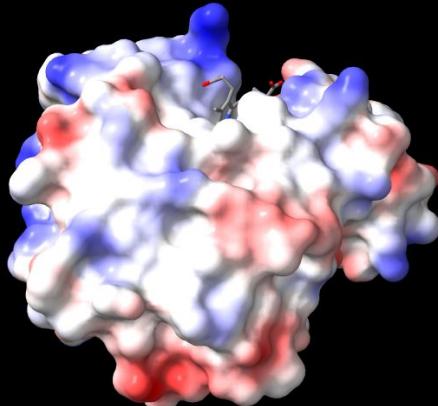
Atoms



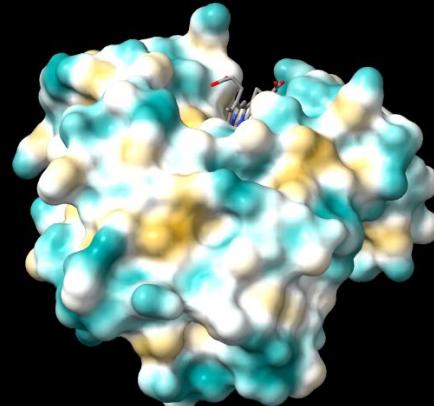
Spheres



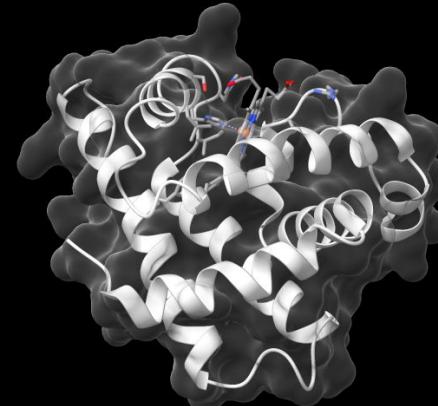
Electrostatic



Hydrophobicity



Surfaces



Structure tridimensionnelle de l'hémoglobine de cheval dans PDB

Structure Summary Structure Annotations Experiment Sequence Genome
Versions

Display Files Download Files Data API

4HHB

THE CRYSTAL STRUCTURE OF HUMAN DEOXYHAEMOGLOBIN AT 1.74 ANGSTROMS RESOLUTION

PDB DOI: <https://doi.org/10.2210/pdb4Hhb/pdb> Entry: 4Hhb supersedes: 1Hhb

Classification: OXYGEN TRANSPORT

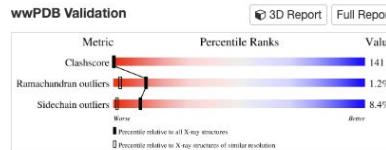
Organism(s): Homo sapiens

Mutation(s): No

Deposited: 1984-03-07 Released: 1984-07-17

Deposition Author(s): Fermi, G., Perutz, M.F.

Experimental Data Snapshot



This is version 4.2 of the entry. See complete history.

Literature

Download Primary Citation ▾

The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution

Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R.
(1984) J Mol Biol 175: 159-174

PubMed: 6726807 Search on PubMed

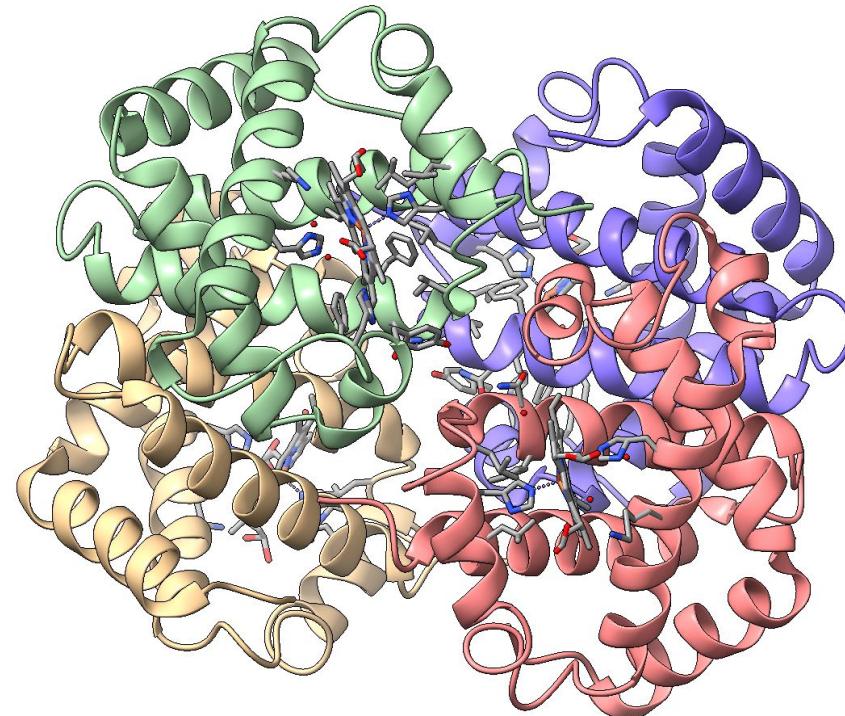
DOI: [https://doi.org/10.1016/0022-2836\(84\)90472-8](https://doi.org/10.1016/0022-2836(84)90472-8)

Primary Citation of Related Structures:

4Hhb 4Hhb 4Hhb

Structure de l'hémoglobine de cheval (4Hhb)

Coloration par chaîne



[Animation: rotation de la structure de l'hémoglobine](#)

Relations séquence - structure - fonction

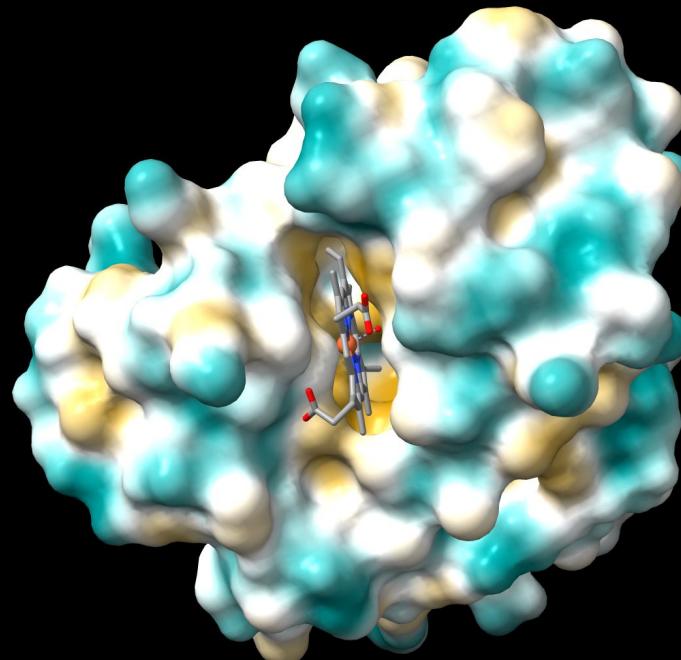
Quelques exemples illustratifs

Structure tridimensionnelle de la myoglobine dans PDB

- La séquence de la myoglobine détermine la structure
 - les structures secondaires (hélices)
 - Les structures tertiaires
(agencement des hélices → protéine globulaire)
 - Le profil électrostatique (propriétés des résidus)
 - Le profil d'hydrophobicité (idem)
- La structure détermine la fonction
 - Poche où s'insère l'hème
 - Échanges d'oxygène

Profil d'hydrophobicité et site de liaison avec l'hème

(animation créée avec ChimeraX)



Animation:

<https://drive.google.com/file/d/131LT0IPD0bWxUheNbFJVUnL9EzmscW8N>

Facteur transcriptionnel PAX6 : Interaction protéine/ADN

- La protéine humaine PAX6 (en bleu sur la figure) est un **facteur transcriptionnel** responsable de la formation des yeux lors du développement embryonnaire.
- Sa séquence détermine la formation de 4 hélices alpha, et un petit feuillet beta antiparallèle
- Deux des hélices ont la capacité de **reconnaître des séquences spécifiques d'ADN** (brins réverse complémentaires marqués en rose et vert sur la figure)
- Les autres hélices de la protéine interagissent avec la polymérase de l'ARN, et **régulent la transcription des gènes voisins** des sites de liaison de PAX6.
- Les **mutations de PAX6** provoquent des malformations de l'oeil et la cécité (maladie aniridia: absence d'iris)
- Nous reviendrons sur le gène PAX6 lors de prochaines séances consacrées à la structuration et la régulation des génomes, et à l'évolution biologique



Animation:

https://drive.google.com/file/d/172qRrH0OqMQCHEjpuFheZY7Jo4F_G7pa

Porine

Source des données: [PDB 1A0S](#)

Cartoon (haut)

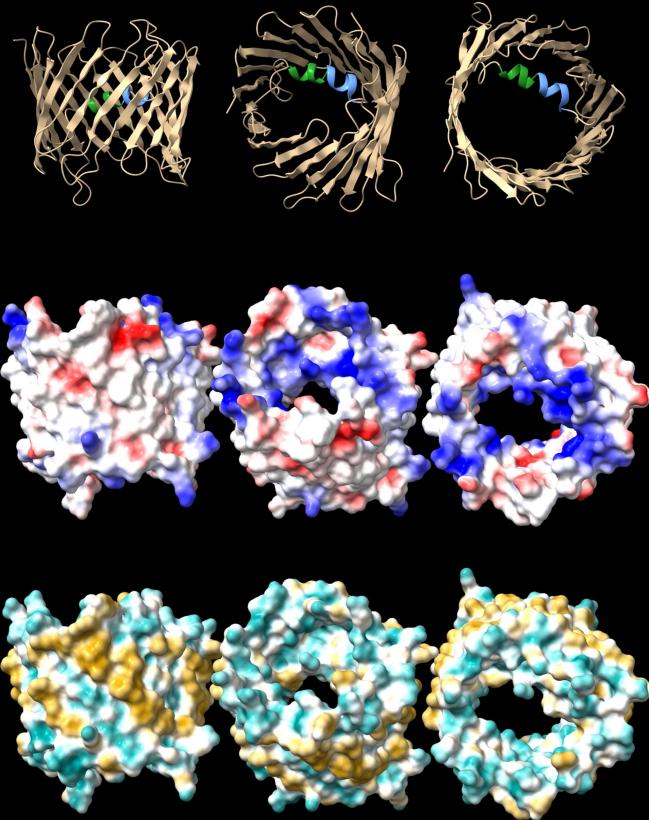
- La **porine du sucre** est une protéine formée en majorité par des feuillets bêta antiparallèles, dont l'agencement forme un cylindre. Cette topologie est dénommée **tonneau bêta** (*beta barrel* en anglais)
- L'intérieur du cylindre est partiellement occupé par deux petites hélices alpha

Profil hydrostatique (haut à)

- L'extérieur du tonneau bêta est essentiellement non-polarisé (blanc)
- L'intérieur est chargé positivement (bleu)

Profil d'hydrophobicité (bas)

- Extérieur : surfaces hydrophobes, qui stabilisent la protéine dans la membrane
- Intérieur: ouverture traversant la membrane, surfaces intérieures hydrophiles, qui permet au sucre de passer



Animation:

https://drive.google.com/file/d/1SIC_YTOVOIKBA_KHxF0CvYmUeITbvmi7

Prédiction de la structure tridimensionnelle des protéines

Méthodes pour la prédition de structures tridimensionnelle

La prédition de structures à partir de séquence a fait l'objet intense de recherche depuis les années 1990.

On distingue deux types de **situations**

1. Il existe une séquence similaire dont la structure a été caractérisée expérimentalement → on recourt à la **Modélisation par homologie**. On aligne la séquence de la protéine sur la structure connue, et on adapte les positions des résidus pour tenir compte des acides aminés qui diffèrent entre les deux séquences
2. **Modélisation *ab initio*** : pour certaines protéines on ne dispose d'aucun homologue de structure connue.

Approches algorithmiques pour la prédition de structures

Depuis les années 1990, de nombreuses équipes de recherche ont développé des logiciels pour prédire la structure d'une protéine à partir de sa séquence.

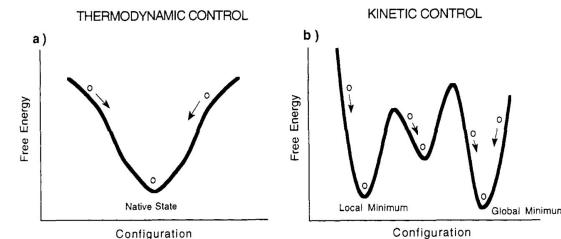
Voici les principales approches

- Minimisation d'énergie
- Dynamique moléculaire
- Recuit simulé

La description de ces approches sort largement du cadre d'un cours d'introduction à la bioinformatique, elles seront présentées dans des cours de biochimie/bioinformatique structurale. Les figures sont fournies uniquement à titre d'illustration.

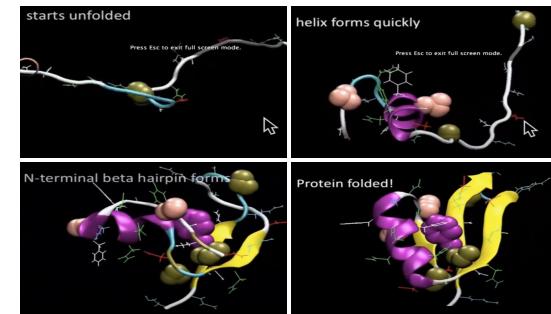
Ces méthodes ne font pas partie de la matière d'examen.

Minimisation d'énergie



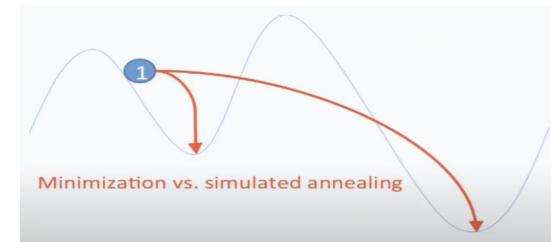
Baker, D. & Agard, D. A. [Biochemistry 33, 7505–7509](#) (1994).

Dynamique moléculaire



MIT OpenCourseWare "[Predicting Protein Structure](#)"

Recuit simulé



● Baker, D. & Agard, D. A. Kinetics versus thermodynamics in protein folding. Biochemistry 33, 7505–7509 (1994).

doi.org/10.1021/bi00190a002

● MIT OpenCourseWare "Predicting Protein Structure". <https://youtu.be/j1s9JfZKFqU?t=806>

Critical Assessment of protein Structure Prediction (CASP)

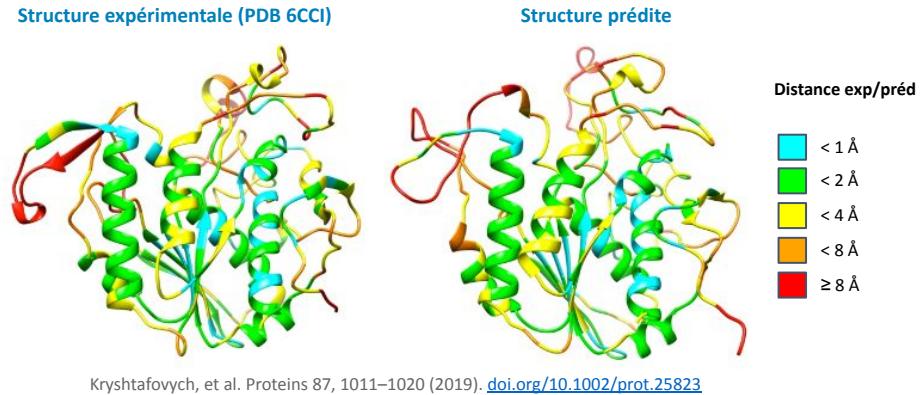
Critical Assessment of Structure Prediction (CASP)

Depuis 1996, la communauté de structuralistes organise une évaluation objective de la valeur prédictive des modèles de structures protéique, via un événement intitulé “Critical Assessment of Structure Prediction (CASP)” (évaluation critique de la prédiction de structure).

Principe

- Des chercheurs qui viennent de caractériser expérimentalement une structure l'envoie aux organisateurs avant de la publier dans un journal.
- A l'ouverture du challenge, les organisateurs mettent en public les séquences de ces protéines, mais (évidemment) pas les structures.
- Les bioinformaticiens structuralistes utilisent leurs différentes méthodes pour prédire la structure à partir de chaque séquence.
- Les organisateurs mesurent ensuite le degré de correspondance entre structures expérimentales et structures prédites

Ce challenge est organisé tous les deux ans depuis 1996.



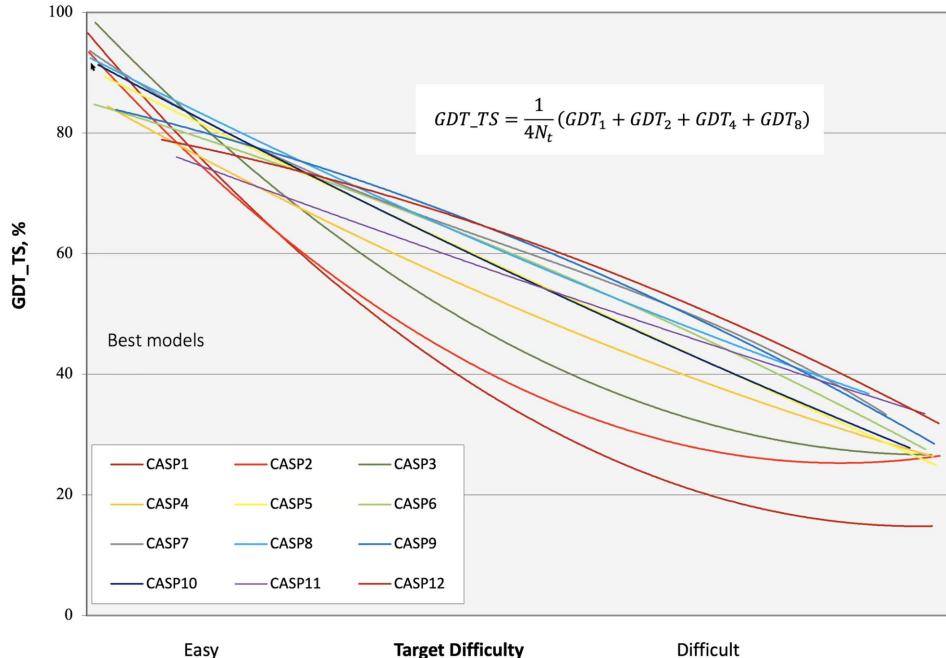
Qualité des prédictions lors de CASP

CASP permet de mesurer l'évolution des performances des outils de prédiction au fil des années.

- Abscisse: degré de difficulté (protéine expérimentale plus ou moins complexe)
- Ordonnée: précision des prédition
 - 100%: tous les atomes prédits sont exactement à la position des atomes de la structures expérimentale
 - 90% - 100% : prédiction du même niveau que la structure expérimentale (les structures expérimentales varient selon les conditions)
 - ~50% : prédiction qui identifie correctement la topologie générale de la protéine (hélices, feuillets beta) mais avec de grosses imprécisions sur les positions des résidus et atomes
 - 10% - 20% : précision attendue au hasard (aucune valeur prédictive)

On constate

- Amélioration progressive de CASP1 à CASP5
- Relative stagnation de CASP5 à CASP12

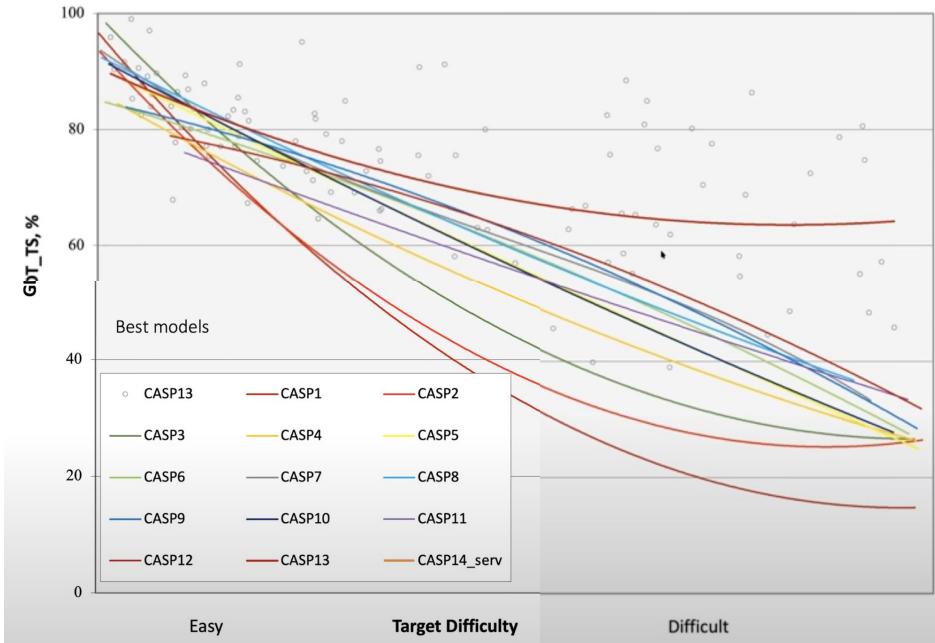


Kryshtafovych, et al. Proteins 87, 1011–1020 (2019). doi.org/10.1002/prot.25823

Quand l'IA est entré dans le jeu

CASP13 – Les réseaux neuronaux profonds convolutifs

- Amélioration progressive de CASP1 à CASP5
- Relative stagnation de CASP5 à CASP12
- CASP13 : **bond quantitatif**
 - Les prédictions dépassent 60% de précision pour tous les niveaux de difficulté → prédiction correcte de la topologie des protéines, mais incertitudes sur les positions précises des atomes
 - Nouvelle approche: application de **réseaux de neurones convolutifs** (NNC) pour prédire la structure sur base de cartes d'interactions entre acides aminés (voir diapos suivantes)

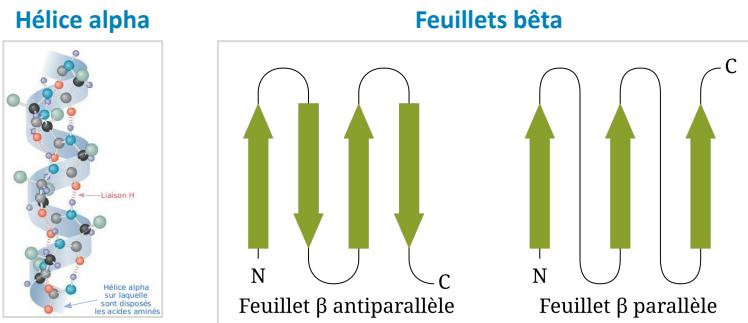


Carte des contacts entre résidus

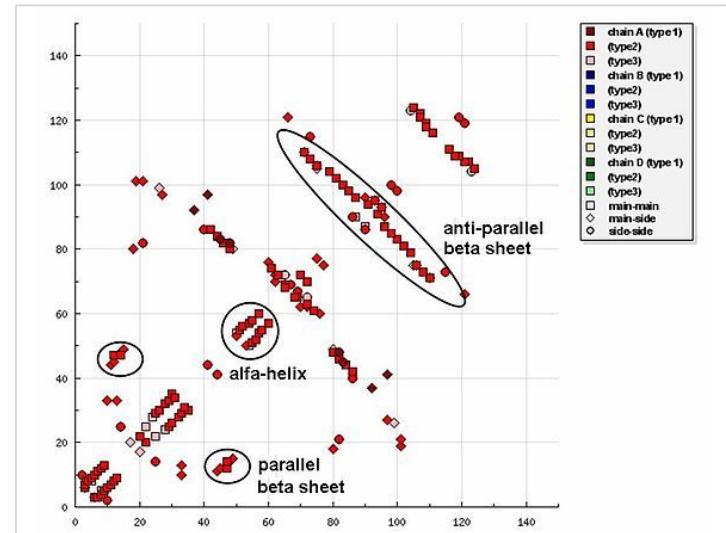
La carte ci-contre (contact plot) représente les contacts entre résidus.

Les axes X et Y représentent les coordonnées de chaque résidu. Les points indiquent que deux résidus sont en contact au sein de la protéine repliée.

- Autour de la diagonale: dans les hélices alpha, chaque acide aminé est en contact avec deux résidus:
 - Celui qui vient 3 ou 4 pas plus tôt dans la chaîne
 - Celui qui vient 3 ou 4 pas plus tard dans la chaîne
- Feuilles bêta parallèles
 - Succession de résidus qui interagissent chacun avec un résidu distant → ligne à +45° sur le graphique, éloignée de la diagonale
- Feuilles bêta antiparallèles
 - Contacts successifs entre deux chaînes d'acides aminés orientées en sens inverse → ligne perpendiculaire sur le graphique



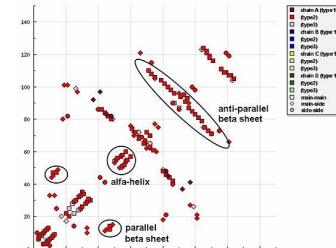
Carte de contact entre acides aminés



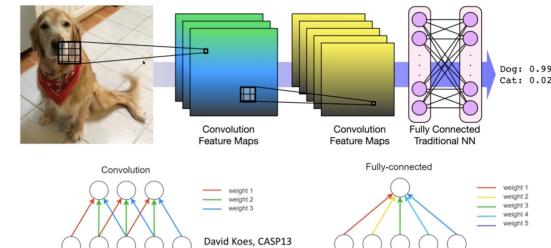
Prédiction de structure sur base des interactions entre résidus

Pour CASP 2013, plusieurs groupes de compétiteurs ont utilisé des cartes d'interactions entre acides aminés (haut) pour entraîner un **réseau neuronal convolutif** (milieu) à prédire les structures protéiques.

Carte de contacts entre acides aminés



Réseau neuronal de convolution



Prédiction de structure protéique

Un exemple de protéine présentée à CASP

LmrP est une protéine transmembranaire qui contribue à la résistance aux médicaments chez la bactérie *Lactococcus lactis* (pompe à efflux multi-drogues). L'analyse structurale a contribué à comprendre ses mécanismes d'action.

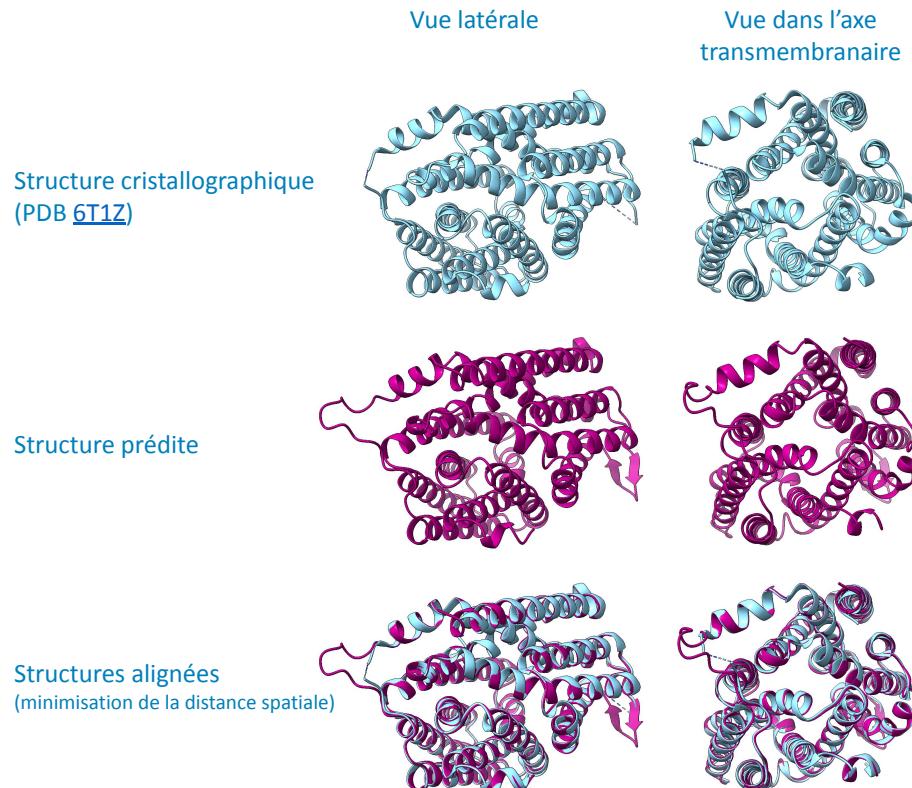
L'histoire d'une structure cristalline

- Fin 2006, Cédric Govaerts fait les premiers essais de cristallographie à San Francisco
- 2008: nouveaux essais de cristallo à l'Université Libre de Bruxelles (ULB)
- 2010 : premiers cristaux confirmés. Basse resolution (10A)
- 2012 : démarrage d'une thèse dédiée
- 2017 : structure caractérisée
- 2019 : thèse soutenue
- 2020 (avril) : article accepté pour publication (*Nat Struct Mol Biol*)
- 2020 (mai) : envoi de la structure à CASP 14
- 2020 (juin) : retour de Casp avec une modèle quasi identique ! Un niveau de similarité nettement supérieur à ce que permettaient les méthodes canoniques de prédiction de structure
- La structure quasiment parfaite avait été prédite par un logiciel reposant sur **AlphaFold2**, une intelligence artificielle (IA) spécialisée pour prédire les structures des protéines.
- Note : la **cristallographie** apporte cependant des informations additionnelles concernant les interactions entre la protéine, son ligand et des molécules lipidiques de la membrane.

Debruycker, V. et al. An embedded lipid in the multidrug transporter LmrP suggests a mechanism for polyspecificity. *Nat Struct Mol Biol* 27, 829–835 (2020).
doi.org/10.1038/s41594-020-0464-y

Merci à Cédric Govaerts pour la structure et pour la chronologie des événements

Protéine **LmrP** de la bactérie *Lactococcus lactis* (Uniprot [Q48658](#))



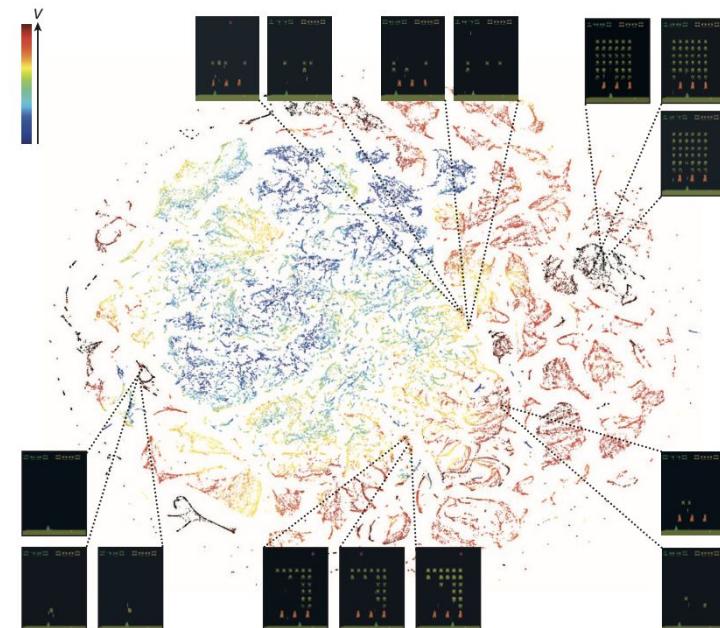
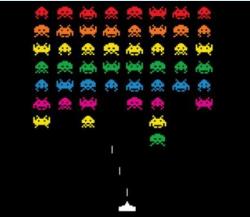
Et DeepMind créa AlphaFold

- 2010. Fondation de la compagnie DeepMind.
 - Objectif: développer des algorithmes d'apprentissage automatique basés sur des réseaux neuronaux profonds (une des approches d'intelligence artificielle)
- 2012-2015: apprentissage "renforcé" de **jeux vidéos des années 80** (space invaders, qBert, ...)
- 2015 - 2017: l'application **AlphaGo** bat les champions européens puis mondiaux de Go
- 2016: DeepMind est racheté par Google et devient Google DeepMind
- 2020: **AlphaFold** (version 1), logiciel de prédiction de structure des protéines
- 2021: **AlphaFold2**. Logiciel libre, 20.000 prédictions de structures de protéines
- 2023: **Gemini** (Generalized Multimodal Intelligence Network), grand modèle de langage capable de combiner une analyse de sons, vidéos, textes

Illustration: Space Invaders

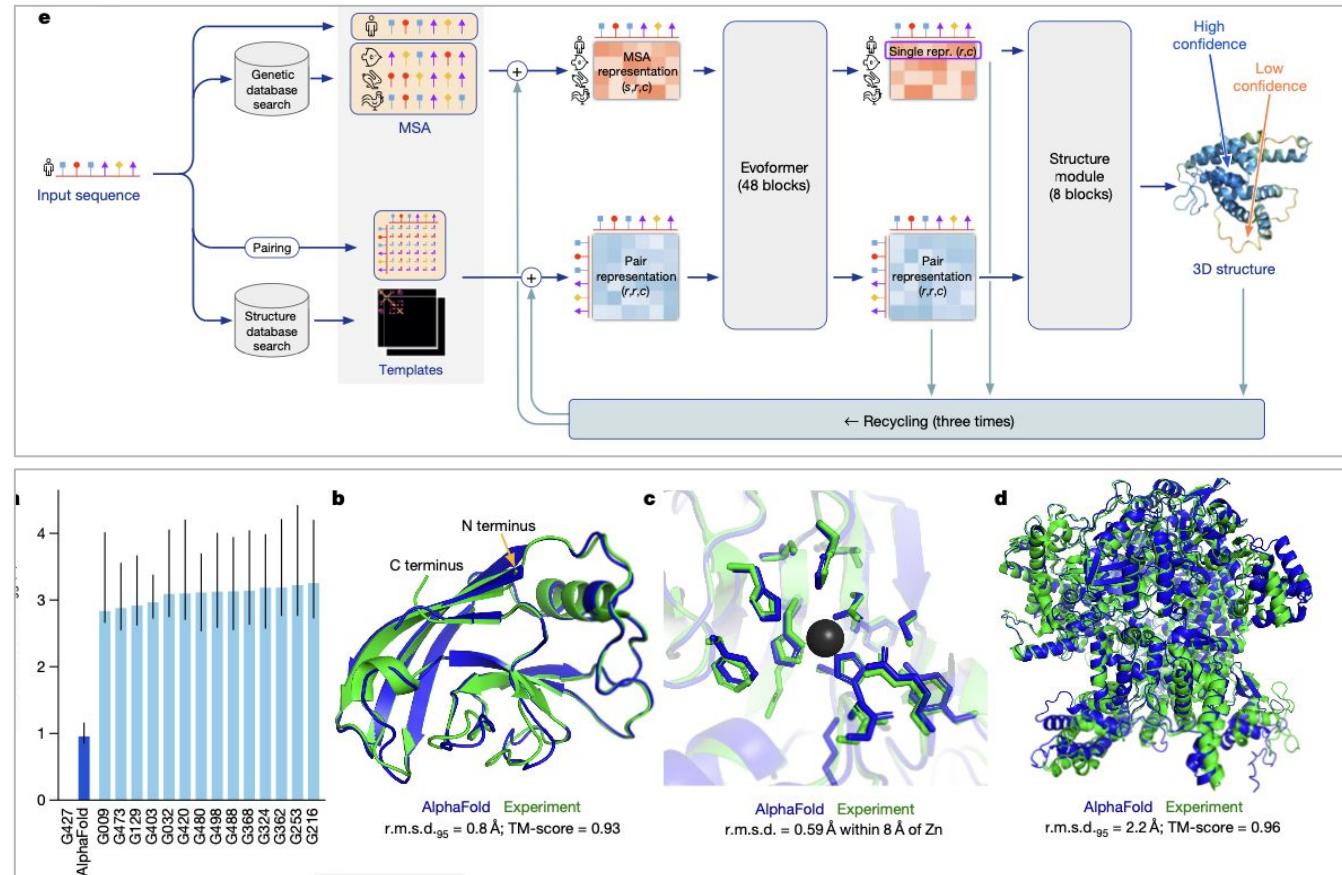
Haut: exemple d'écran du jeu Space Invaders

Bas: dernière "couche neuronale" (informatique) de DeepMind après quelques heures d'apprentissage de Space Invaders



AlphaFold2

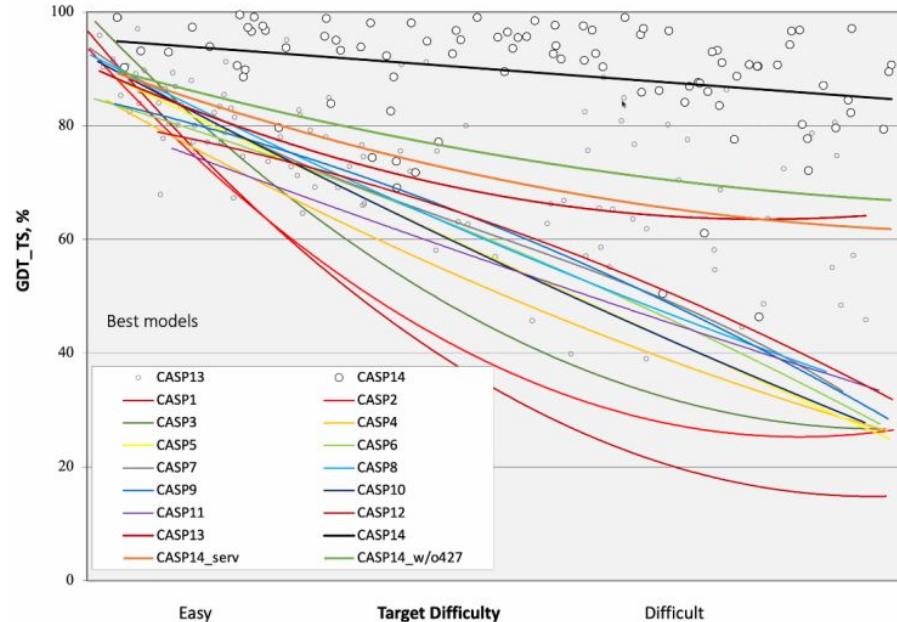
- **AlphaFold** est un logiciel d'intelligence artificielle (IA) spécialisée pour la prédiction de structures tridimensionnelles de protéines à partir de leurs séquences
- Haut: Schéma de la méthodologie (ne fait pas partie de la matière d'examen)
 - Modèle de type transformer, avec 2 modules de transformation
 - Apprentissage par réseau neuronal "profond" (48)
 - Données d'entraînement :
 - corpus complet des structures de PDB
 - Bases de données de séquences
- Performances (schéma du bas)
 - a. Distances entre les modèles et la structure expérimentale. AlphaFold est affiché en bleu foncé, les autres candidats en bleu pâle
 - b-d. Alignements entre la structure expérimentale (en vert) et la prédiction AlphaFold (en bleu)



Performances d'AlphaFold2 lors de CASP14

Pour la session **CASP14 (2020)**, la précision des prédictions dépasse de très loin celles de tous les algorithmes et experts humains, et ce pour tous les degrés de difficulté des protéines cibles.

Ces résultats sont ceux d'AlphaFold2, le logiciel d'IA spécialisé pour la prédiction de structures protéiques.



CASP14 Day 1 : Intro by John Moult : Alphafold2 results

(<https://youtu.be/EFwO1LX0eZY?t=1658>)

<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/>

Conclusion – l'arrivée de l'IA dans le domaine de la biologie structurale

Caractérisation expérimentale des structures

- La biologie structurale est un domaine de la biologie qui combine des méthodes de biochimie, biophysique, microscopie, informatique afin de caractériser ou de prédire les structures des molécules biologiques.
- Pendant 60 ans, les approches expérimentales (cristallographie, RMN, cryo-microscopie électronique) ont donné des résultats au prix d'efforts importants : pour schématiser, 1 thèse = 1 structure. De 1958 à ce 2024, ces efforts cumulés ont mené à caractériser **~225.000 structures protéiques**.

Prédiction de structure à partir des séquences

- La prédiction de structure est un problème notoirement difficile en bioinformatique structurale.
- L'initiative CASP a fourni une mesure objective de l'évolution des performances depuis 1996.
- La prédiction par homologie fonctionne relativement bien quand on dispose de modèles pour des protéines se séquences très similaires.
- Ceci a stimulé les nouveaux développement, et on observe
 - CASP1 (1994) à CASP5 (2002): augmentation progressive des performances
 - CASP6 (2004) à CASP12 (2016): Relative stagnation entre
 - CASP13 (2018): bond quantitatif avec lié aux réseaux neuronaux convolutifs (précision >70%)
 - CASP14 (2020): nouveau bond quantitatif avec AlphaFold2 (précision >90%)
- **2024: prédiction de >200.000.000 de structures protéiques** à partir de séquences (chaque séquence d'Uniprot), accessibles à partir d'Uniprot

AlphaFold a changé la donne

- AlphaFold2 (2020) puis AlphaFold3 (2024) ont changé la donne, en fournissant, pour la plupart des protéines, des prédictions de structures aussi bonnes que les méthodes expérimentales.
- Le code d'AlphaFold est public, et le logiciel est déployé sur des serveurs accessibles gratuitement (avec certaines limitations).
- Ceci a exercé une profonde transformation sur les pratiques les chercheurs, qui disposent désormais de prédictions relativement fiables pour toutes les protéines.
- 2024: AlphaFold3
 - Contrairement à AlphaFold2, pas (encore ?) dans le domaine public

Projet académique

- Projet [openfold.io](#) vise à développer un outil libre similaire à AlphaFold3

Limitations

- La précision des prédictions repose intrinsèquement sur la disponibilité d'un très grand nombre de structures connues, déterminées expérimentalement.
- On dispose de nombreuses structures expérimentales pour certains types de protéines, mais pour d'autres elles manquent encore → performances inégales selon le type de protéine.
- Nécessite de disposer d'un nombre suffisant de séquences homologues (nécessaires à la première phase de la prédition).

Défis (abordés par AlphaFold3)

- **Complexes** formés de plusieurs macromolécules
- Interactions **protéine - ligand** (petites molécules, notamment les médicaments)
- Interactions **protéine - ADN** (régulation transcriptionnelle)
- Prédiction de l'**effet des mutations** sur la structure