

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

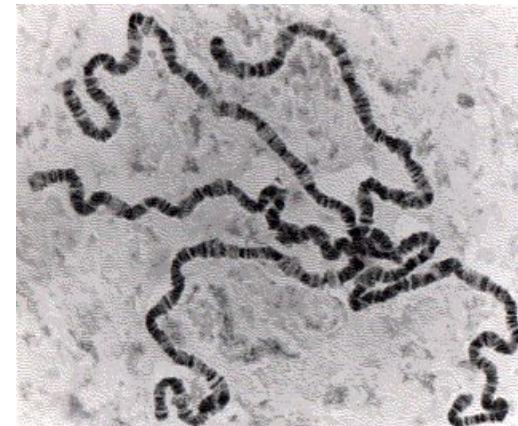
Les voies de l'information génétique : séquence, structure et fonction de l'ADN

Les chromosomes constituent le support physique de l'hérédité

- Les lois de Mendel (1865), redécouvertes en 1901, modélisent la progression des fréquences alléliques au fil des générations, mais ne fournissent aucune information quant aux mécanismes sous-jacents.
- Les travaux de Thomas Hunt Morgan ont permis d'établir que les chromosomes sont le support physique de l'hérédité.
- Dans un livre de 1915, il formule de façon générale sa **théorie chromosomique de l'hérédité**.
- Ses observations
 - Les 4 groupes de liaison génétiques de la drosophile correspondent aux 4 chromosomes.
 - Les chromosomes sont porteurs des caractères transmis de façon héréditaire.
 - Sur chaque chromosome, les gènes sont ordonnés de façon linéaire.
- Il en déduit que les chromosomes sont le support physiques des caractères héréditaires.



http://news.bbc.co.uk/olmedia/440000/images/_443673_drosgene.jpg



<http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Polytene.jpg>

Caryotype humain

Chez l'humain, les noyaux des cellules somatiques comportent 23 paires de chromosomes.

Les cellules somatiques sont diploïdes: chaque cellule comporte 2 copies de chaque chromosome (1 maternelle et 1 paternelle).

Photo de chromosomes étalés



Ces mêmes chromosomes regroupés pour mettre en évidence les paires homologues



Les chromosomes sont essentiellement composés d'ADN

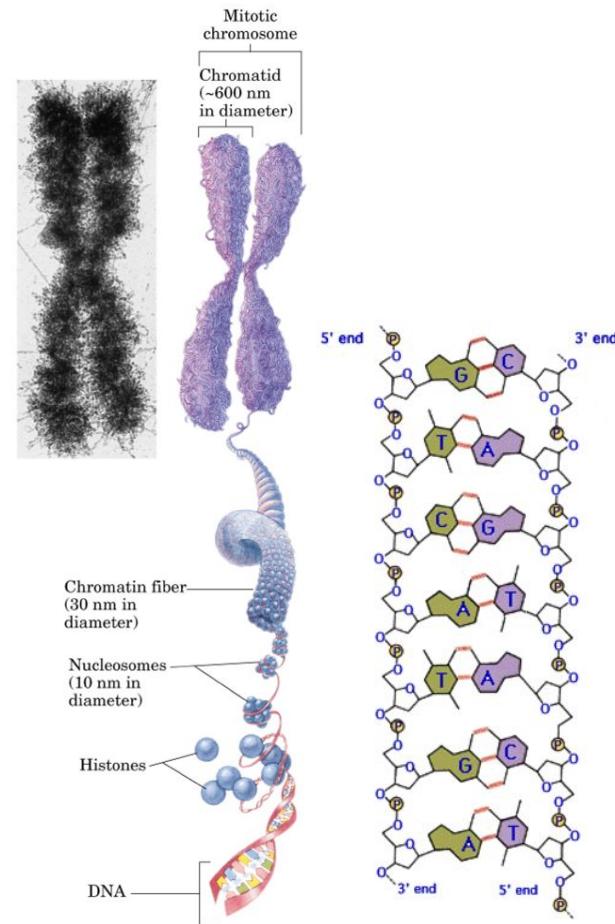
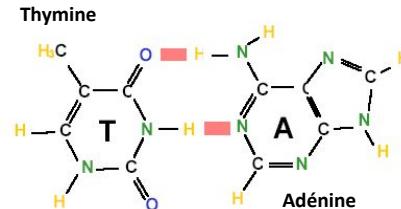
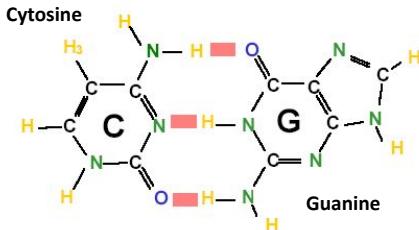
Chaque chromosome contient une chaîne extrêmement longue d'acide désoxyribonucléique (ADN).

L'ADN est composé d'une double hélice, qui porte 4 types de bases azotées.

- A Adénine
- C Cytosine
- G Guanine
- T Thymine

L'information génétique réside dans la succession de ces bases azotées.

Ces bases azotées sont appariées de façon spécifique dans la structure en double hélice.



Structure de l'ADN - la double hélice

- En 1953, en se basant sur une **image de diffraction des rayons X de fibres d'ADN** obtenue en 1952 par Rosalind Franklin et son étudiant Raymond Gosling, Watson et Crick proposent un **modèle en double hélice** pour la structure B de l'ADN.
- « Montants » : squelette sucre-phosphate ; successions de désoxyribooses liés de façon covalente par des groupes phosphate.
- “Barreaux” : paires de bases complémentaires, stabilisées par des liaisons hydrogène.
 - Guanine ↔ Cytosine (3 liaisons hydrogène)
 - Adénine ↔ Thymine (2 liaisons hydrogène)

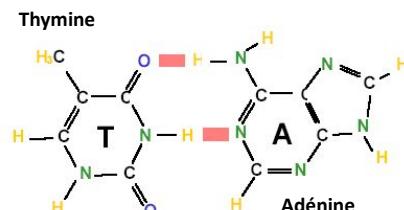
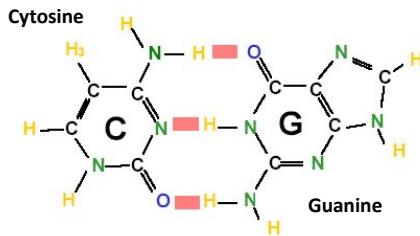
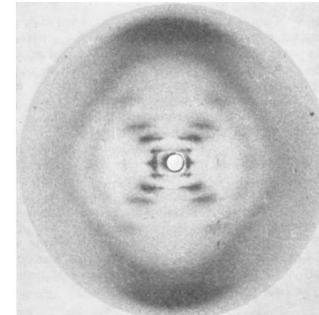


Image cristallographique l'ADN, par diffraction de rayons X
(R.E. Franklin and R. Gosling, 1953)



Sodium deoxyribose nucleate from calf thymus. Structure B

Modèle de la structure de l'ADN (Watson and Crick, 1953b)



- Franklin,R.E. and Gosling,R.G. (1953) Molecular configuration in sodium thymonucleate. doi.org/10.1038/171740a0
- WATSON,J.D. and CRICK,F.H. (1953a) The structure of DNA. Cold Spring Harb Symp Quant Biol, 18, 123–131. doi.org/10.1101/sqb.1953.018.01.020
- Watson,J. and Crick,F. (1953b) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171, 737–738. doi.org/10.1038/171737a0
- WATSON,J.D. and CRICK,F.H. (1953c) Genetical implications of the structure of deoxyribonucleic acid. Nature, 171, 964–967. doi.org/10.1038/171964b0

Implications de la structure de l'ADN

Dès 1953, Watson et Crick discutent de l'impact de leur modèle pour comprendre les mécanismes de réPLICATION de l'information génétique.

- *Il n'a pas échappé à notre attention que l'appariement spécifique que nous avons postulé suggère immédiatement un mécanisme possible de copie pour le matériel génétique. (Watson & Crick, 1953b)*
- *Notre modèle d'acide désoxyribonucléique constitue en fait une paire de modèles, chacun étant complémentaire de l'autre. Nous imaginons qu'avant la duplication, les liaisons hydrogène sont rompues et que les deux chaînes se déroulent et se séparent. Chaque chaîne sert alors de modèle pour la formation, sur elle-même, d'une nouvelle chaîne complémentaire, de sorte qu'on obtient finalement deux paires de chaînes, alors que nous n'en avions qu'une auparavant. De plus, la séquence des paires de bases aura été dupliquée exactement. (Watson & Crick, 1953c)*

Watson,J. and Crick,F. (1953b) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171, 737–738. doi.org/10.1038/171737a0

Watson,J.D. and Crick,F.H. (1953c) Genetical implications of the structure of deoxyribonucleic acid. Nature, 171, 964–967. doi.org/10.1038/171964b0

GENETICAL IMPLICATIONS OF THE STRUCTURE OF DEOXYRIBONUCLEIC ACID

By J. D. WATSON and F. H. C. CRICK

Medical Research Council Unit for the Study of the Molecular Structure of Biological Systems, Cavendish Laboratory, Cambridge

THE importance of deoxyribonucleic acid (DNA) within living cells is undisputed. It is found in all dividing cells, largely if not entirely in the nucleus, where it is an essential constituent of the chromosomes. Many lines of evidence indicate that it is the carrier of a part of (if not all) the genetic specificity of the chromosomes and thus of the gene itself.

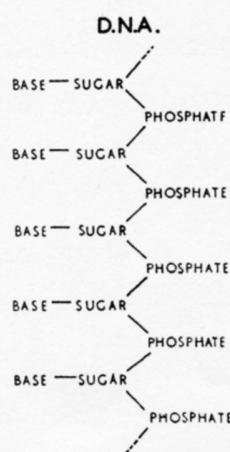


Fig. 1. Chemical formula of a single chain of deoxyribonucleic acid

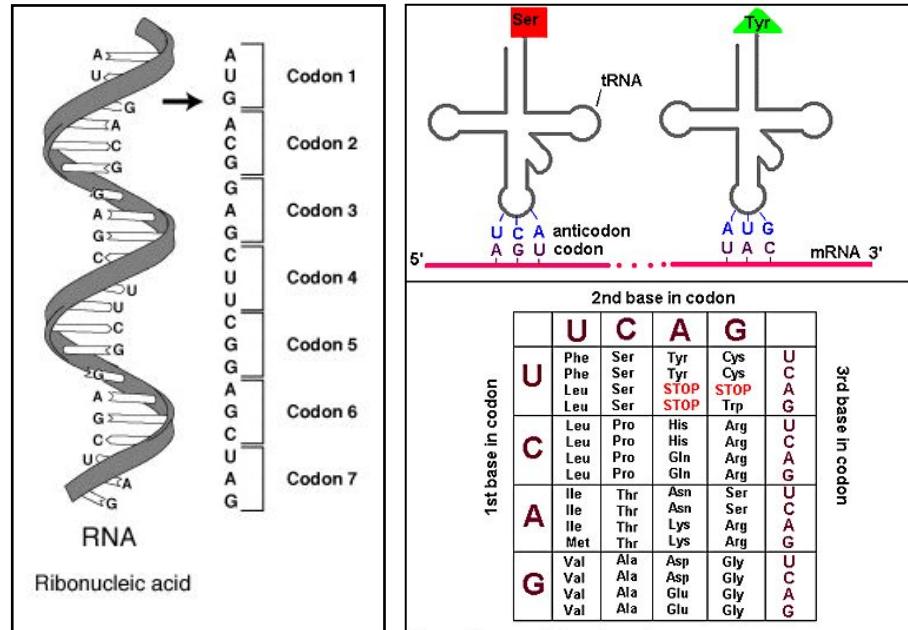
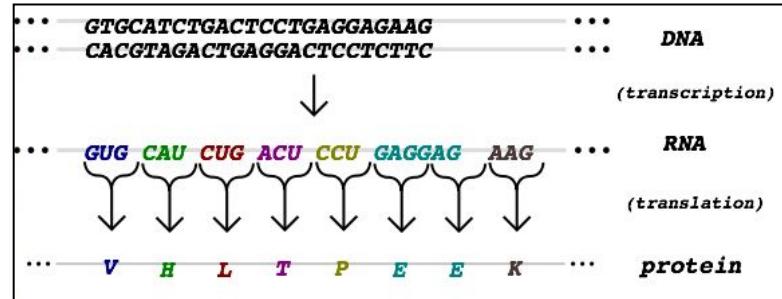


Fig. 2. This figure is purely diagrammatic. The two ribbons symbolize the two phosphate-sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

Le code génétique - Concepts de base

Bref rappel de concepts-clés vus lors des cours de biologie moléculaire

- Le **code génétique** a été élucidé au début des années 1960, par Marshall Nirenberg et son équipe.
- **Traduction:** les protéines sont synthétisées sur modèle de l'ARN.
- Contrairement à la transcription, il n'y a **pas de correspondance de un à un** entre les résidus de l'ADN (nucléotides) et de la protéine (acides aminés). En effet, l'ARN ne comporte que 4 nucléotides distincts (adénine, uracile, guanine et cytosine), tandis que les protéines sont formées de 20 acides aminés distincts.
- **Codons:** chaque acide aminé est spécifié par une succession de 3 nucléotides
- **Dégénérescence (redondance) du code:** Il y a 64 triplets de nucléotides possibles mais 20 acides aminés. **Plusieurs codons spécifient le même acide aminé.**



- Nirenberg, M. W. & Matthaei, J. H. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc. Natl. Acad. Sci. U.S.A. 47, 1588–1602 (1961). doi.org/10.1073/pnas.47.10.1588
- Nirenberg, M. et al. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. PNAS U.S.A. 53, 1161–1168 (1965). doi.org/10.1073/pnas.53.5.1161

Dégénérescence du code

AGA		UUA		AGC		GUA		UAA
AGG		UUG		AGU		GUC		UAG
GCA	CGA	CUA		ACA		GUG		UAA
GCC	CGC	CUC		ACC		GUU		UAG
GCG	CGG	AUA		ACG		GUU		UGA
GCU	CGU	GGG	CAC	UUC	CCU	UCA	UAC	UUA
		GGU	CAU	UUU	UCU	UCC	UCU	UUA
		GGG	AUC	AAA	UCG	UCC	ACU	UAG
		GGU	CUU	AAG	ACG	UCU	UGG	UGA
		GGG	CUG	AUG	ACG	UCU	UAC	UUA
Ala	Arg	Asp	Asn	Cys	Glu	Gln	Gly	His
A	R	D	N	C	E	Q	G	H
								I
								L
								K
								M
								F
								P
								S
								T
								W
								Y
								V
								stop

Figure 7-24 The nucleotide sequence of an mRNA is translated into the amino acid sequence of a protein via the genetic code. All the three-nucleotide codons that specify a given amino acid are listed above that amino acid,

Le code est dit dégénéré (redondant) : plusieurs codons correspondent à un même acide aminé

Synonymie non-aléatoire: la synonymie n'est pas distribuée au hasard entre codons. La dégénérescence se manifeste essentiellement au niveau du troisième nucléotide de chaque codon.

The genetic code is triplet							
		Second base					
		U	C				
		A	G				
U	UUU	Phe	UCU	UAU	Tyr	UGU	Cys
	UUC		UCC	UAC	Ser	UGC	
	UUA	Leu	UCA	UAA	STOP	UGA	STOP
	UUG		UCG	UAG		UGG	Trp
C	CUU		CCU	CAU	His	CGU	
	CUC	Leu	CCC	CAC	Pro	CGC	
	CUA		CCA	CAA	Gln	CGA	
	CUG		CCG	CAG		CGG	
A	AUU		ACU	AAU	Asn	AGU	
	AUC	Ile	ACC	AGC	Ser	AGC	
	AUA		ACA	AGA	Arg	AGA	
	AUG	Met	ACG	AAG	Lys	AGG	
G	GUU		GAU	GAU	Asp	GGU	
	GUC	Val	GCU	GAC		GGC	
	GUU		GCC	GAA		GGG	
	GUG		GCA	GAG	Glu	GGA	Gly

FIGURE 9.1 All the triplet codons have meaning: Sixty-one represent amino acids, and three cause termination (STOP). Source: Genes IX

Le “dogme central”

- Le « dogme central » a été formulé en 1958 par Francis Crick. Je recommande également de lire cette discussion ultérieure (Crick, 1970).
- On le résume souvent par la formule concise
“DNA makes RNA makes protein”
 (“L’ADN fait l’ARN faire la protéine”)
- Cette phrase, subtile syntaxiquement et sémantiquement, est souvent mal comprise. Le dogme ne restreint pas les voies de l’information à cette succession particulière, mais établit tous les **transferts d’information possibles** (schéma du haut) ou **impossibles** (schéma du bas) entre séquences nucléiques et peptidiques.
- Le “dogme” a été critiqué par des gens qui n’avaient pas lu sa formulation exacte, évoquant par exemple
 - La transcription réverse (“RNA makes DNA”)
 - Les modifications des prions (“protein changes protein”)
- La formulation de Crick est pourtant sans ambiguïté, et elle a conservé toute sa validité.
- Il ne s’agit pas d’un dogme mais d’une **théorie scientifique** rationnelle et logique. L’impossibilité de transfert de protéine à acide nucléique résulte directement de la dégénérescence du code.

Crick, F. H. (1958). [On protein synthesis](#). Symp Soc Exp Biol 12, 138-63.

Crick, F. (1970). Central dogma of molecular biology. Nature 227, 561-3.

<https://doi.org/10.1038/227561a0>

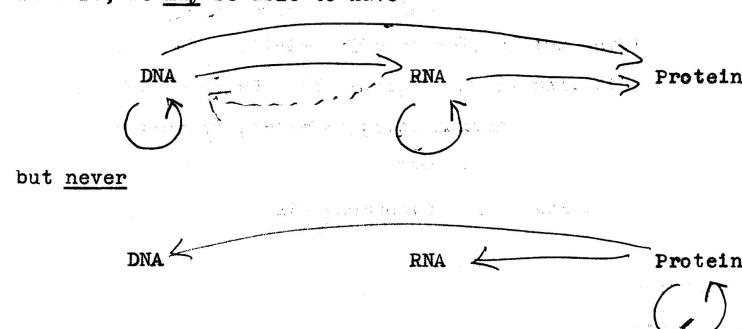
Le dogme central stipule que, une fois que l’« information » est passée dans la protéine elle ne peut pas en ressortir.
Plus précisément, le transfert d’information serait possible d’acide nucléique à acide nucléique, ou d’acide nucléique à protéine, mais le transfert de protéine à protéine, ou de protéine à acide nucléique est impossible. Information signifie ici la détermination précise de la séquence, soit des bases dans l’acide nucléique, soit des résidus aminoacides dans la protéine.

Crick, F. H. (1958). On protein synthesis. Symp Soc Exp Biol 12, 138-63.

The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have



where the arrows show the transfer of information.

Schéma de 1956, préfigurant la publication du dogme central.

<http://resource.nlm.nih.gov/101584582X65>

Le dogme central a-t-il été réfuté ?

On a à plusieurs reprise affirmé que le dogme central avait été réfuté :

- découverte de la transcription réverse.
- découverte du prion.

En 1970 Crick publie une clarification, pour rappeler ce que dit le dogme central, et explique pourquoi la réverse transcription ne le réfute pas.

Fig 1 : toutes les voies a priori envisageables pour les flux de l'information contenue dans les séquences de macromolécules

Fig 2 : état des connaissances au moment de la publication du dogme central (1958)

- flèches pleines: transferts “probables”, pour lesquels on disposait d’indications directes ou indirectes. Rappelons qu’à l’époque le code n’avait pas encore été décrypté.
- flèches pointillées : transferts “possibles” mais sans aucune indication d’existence à l’époque.
- flèches absentes: transferts considérés impossibles.

Fig 3 : mise à jour en 1970

- flèches pleines: transferts “généraux”, qui ont lieu dans toutes les cellules (Crick excepte des réticulocytes, qui n’ont pas de noyau)
- flèches pointillées : transferts “spéciaux”, spécifiques à certains virus (ARN → ARN, ARN → ADN) ou éventuellement dans des systèmes artificiels (ADN → protéine).
- flèches absentes : transferts non détectés, et dont la découverte remettrait complètement en cause les bases de la biologie moléculaire.

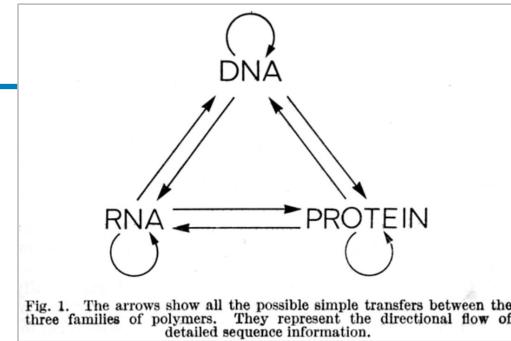


Fig. 1. The arrows show all the possible simple transfers between the three families of polymers. They represent the directional flow of detailed sequence information.

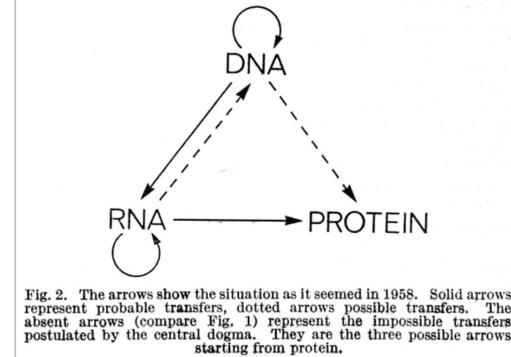


Fig. 2. The arrows show the situation as it seemed in 1958. Solid arrows represent probable transfers, dotted arrows possible transfers. The absent arrows (compare Fig. 1) represent the impossible transfers postulated by the central dogma. They are the three possible arrows starting from protein.

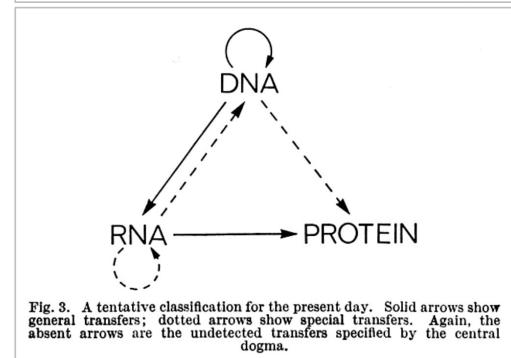


Fig. 3. A tentative classification for the present day. Solid arrows show general transfers; dotted arrows show special transfers. Again, the absent arrows are the undetected transfers specified by the central dogma.

Voies connues et possibles de l'information

Le dogme central ne se limite pas aux transferts connus à un moment ou à un autre, mais couvre tous les transferts d'information qui sont conceptuellement envisageables.

En 1958, on n'avait pas encore élucidé les mécanismes de la traduction (le code a été complètement décrypté en 1960), et l'on n'avait pas encore découvert la transcription réverse. Depuis lors, on a démontré l'existence des transferts d'information suivants.

- ADN → ADN: **réPLICATION DE L'ADN** commune à tous les organismes cellulaires, catalysée par la polymérase de l'ADN.
- ADN → ARN: **transcription**, via la polymérase de l'ARN dépendante de l'ADN, présente chez tous les organismes cellulaires.
- ARN → ADN: **rétrotranscription**, assurée par la transcriptase réverse qu'on trouve chez les rétrovirus, découverte en 1970 (doi.org/10.1038/2261209a0).
- ARN → ARN: **réPLICATION DIRECTE DE L'ARN**, via une polymérase de l'ARN dépendante de l'ARN, qu'on trouve chez des virus de bactérie ou d'eucaryotes (doi.org/10.1073/pnas.51.3.450).
- ARN → protéine: **traduction**, assurée par les ribosomes chez tous les organismes cellulaires.

On ne connaît à ce jour aucun mécanisme qui permet de synthétiser directement des protéines à partir d'ADN, mais en principe un tel transfert d'information serait compatible avec le dogme central.

Voies impossibles et voies existantes qui ne relèvent pas du dogme central

Dans la formulation du dogme central, Crick prend bien soin de définir ce qu'il entend par "information : il s'agit de la détermination précise de la séquence des résidus qui constituent une macromolécule. Donc, la question est de savoir s'il est possible de générer une macromolécule en spécifiant sa séquence par la séquence d'une autre macromolécule.

Voies inconcevables d'après le dogme central

Ce qui serait impossible serait la **synthèse d'un ADN ou d'un ARN dont la séquence serait spécifiée sur modèle d'une séquence**

protéique. La raison est que la séquence protéique ne contient que 20 "signes" (les 20 acides aminés) alors que les séquences nucléiques en contiennent 61 (les codons qui correspondent à un acide aminé). Donc, une polymérase de l'ARN ou de l'ADN qui se baserait sur les 20 acides aminés ne pourrait pas indiquer lequel des codons synonymes il faudrait utiliser. Par exemple, si on observe une leucine à une position donnée d'une protéine, il existerait une ambiguïté entre 6 trinucléotides (les 6 codons correspondant à la leucine).

Crick exclut également la possibilité d'un **transfert d'information (de séquence) de protéine à protéine.** Cette dernière option me semble moins évidente à exclure, même si on ne dispose pas d'exemple et si on n'a aucune raison de penser qu'un tel processus existe.

Voies existantes mais qui ne relèvent pas du dogme central

Les phénomènes suivants n'entrent donc pas dans le cadre du dogme central

- La modification post-traductionnelle des protéines, notamment le changement de conformation du prion, car il ne s'agit pas d'un transfert d'information concernant la séquence de la protéine.
- La maturation de l'ARN n'est pas l'objet du dogme central, car il ne s'agit pas d'un transfert d'information de séquence.

En tout état de cause, l'objet principal du dogme central était d'attirer l'attention sur la perte d'information quand on passe d'une séquence à 61 à 20 signes.

Résumé : les voies de l'information génétique

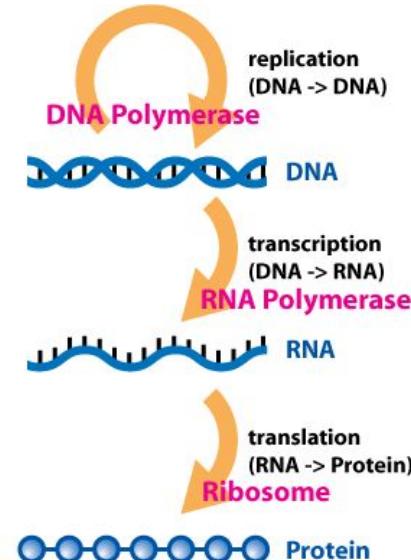
L'ADN est le support de l'information génétique, et ceci de deux façons

- **Hérédité**, via la réplication de l'ADN, qui précède la mitose ou la méiose
- **Information fonctionnelle**: les protéines et certains ARN sont les effecteurs moléculaires des fonctions biologiques

Voies de l'information (schéma simplifié) :

"DNA makes RNA makes protein"

- ADN –[transcription]→ ARN
- ARN –[traduction]→ protéine



Les premières structures de protéines

Les premières structures de protéines

- 1957: John Cowdery Kendrew résout la structure tridimensionnelle de la myoglobine de cachalot (publication 1958)
- 1959: Max Ferdinand Perutz résout la structure 3D de l'hémoglobine (publication 1960)

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 sous la direction de Lawrence Bragg (fondateur de la cristallographie à rayons X), visant à caractériser la structure de l'hémoglobine du cheval
- 1940 : il n'a pas encore résolu la structure, mais ses premiers résultats sur la cristallisation lui permettent de soutenir une thèse
- Pendant les années 1950, il continue à progresser, et monte une équipe de recherche
- 1959 : il caractérise la structure de l'hémoglobine (publication 1960)
- 1962: prix Nobel de chimie

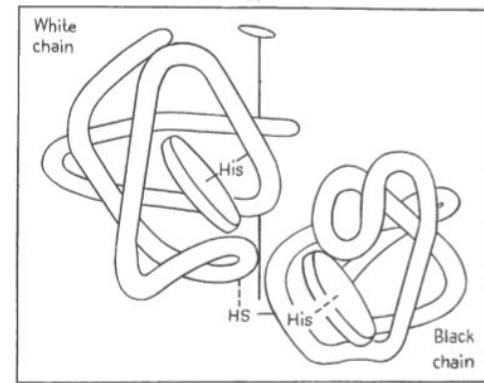
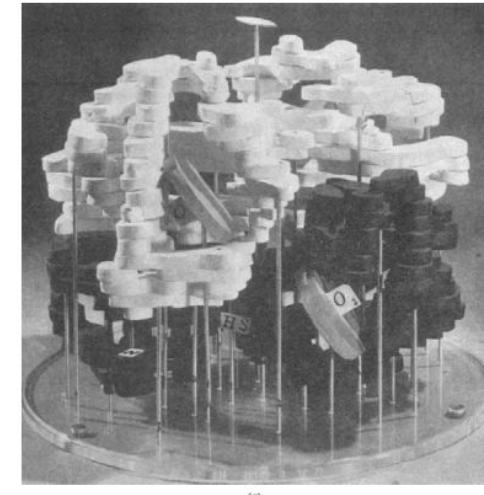


Fig. 8. (a) Hemoglobin model viewed normal to α . The heme groups are indicated by grey disks. (b) Chain configuration in the two sub-units facing the observer. The other two chains are produced by the operation of the dyad axis

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).

doi.org/10.1038/185416a0

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 d'un modèle de la molécule (sous des angles différents).

Kendrew a entrepris son projet en 1947.

Le choix de la protéine était judicieux :

- Par chance, Kendrew a l'occasion de disposer d'un gros échantillon de muscle de cachalot en provenance du Pérou. Le muscle du cachalot, adapté aux longues plongées, présente une forte concentration en myoglobine, ce qui facilite l'obtention de cristaux.
- La myoglobine ne comporte qu'une seule chaîne polypeptidique, alors que l'hémoglobine est un tétramère (polymère de 4 polypeptides). La structure de la myoglobine était donc plus simple à cristalliser et à résoudre.

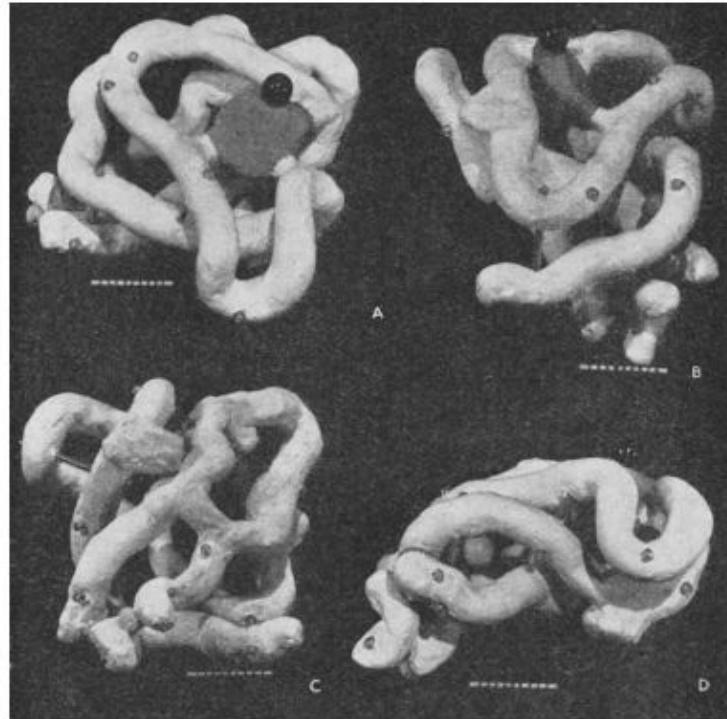


Fig. 2. Photographs of a model of the myoglobin molecule. Polypeptide chains are white; the grey disk is the heme group. The three spheres show positions at which heavy atoms were attached to the molecule (black: Hg of *p*-chloro-mercuri-benzene-sulphonate; dark grey: Hg of mercury diammine; light grey: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

Méthodes de caractérisation des structures de protéines

Diffractométrie aux rayons X (cristallographie)

- Méthode "historique" (Sanger, Kendrew, Perutz, ...)
- Bonne résolution (au niveau de l'ångström)
- Fonctionne non seulement pour des petites molécules, mais également 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 garantie de succès

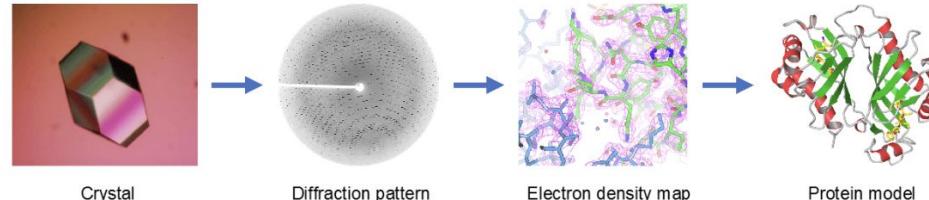
Spectroscopie à résonance magnétique nucléaire (RMN)

- Analyse de protéines en solution
- Fort champ magnétique + impulsions radio-fréquences → spectre (1D, 2D) → calcul des structures 3D.
- Fournit des informations dynamiques (flexibilité, conformations alternatives de la protéine)
- Limitation de la taille des protéines étudiées

Cryo-microscopie électronique (Cryo-EM)

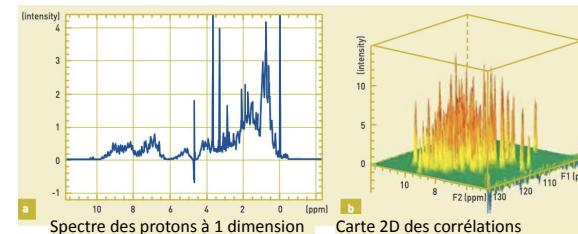
- Plus besoin de cristal
- Nécessite beaucoup moins de quantité de protéine
- Initialement, résolution inférieure à la cristallographie mais progrès récents → égale voire dépasse les rayons X
- Ne fonctionne qu'avec protéines de grande taille (pour pouvoir les détecter)

Principe de la diffractométrie aux rayons X

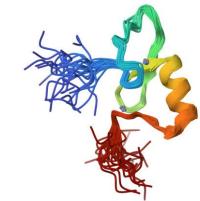


https://www.creative-biostructure.com/x-ray-crystallography-platform_60.htm

Principe de la résonance magnétique nucléaire



<https://www.rcsb.org/structure/2ECM>

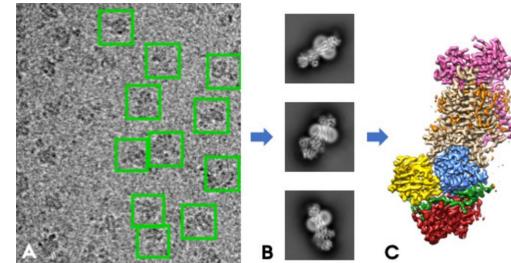


https://www.cea.fr/multimedia/Documents/publications/clefs-cea/archives/en/Clefs_56_p52_55_OchsenbeinGB.pdf

Principe de la Cryo-microscopie électronique

Reconstruction d'une structure 3D à partir d'un très grand nombre de photos 2D de piètre résolution révélant la protéine sous différents angles.

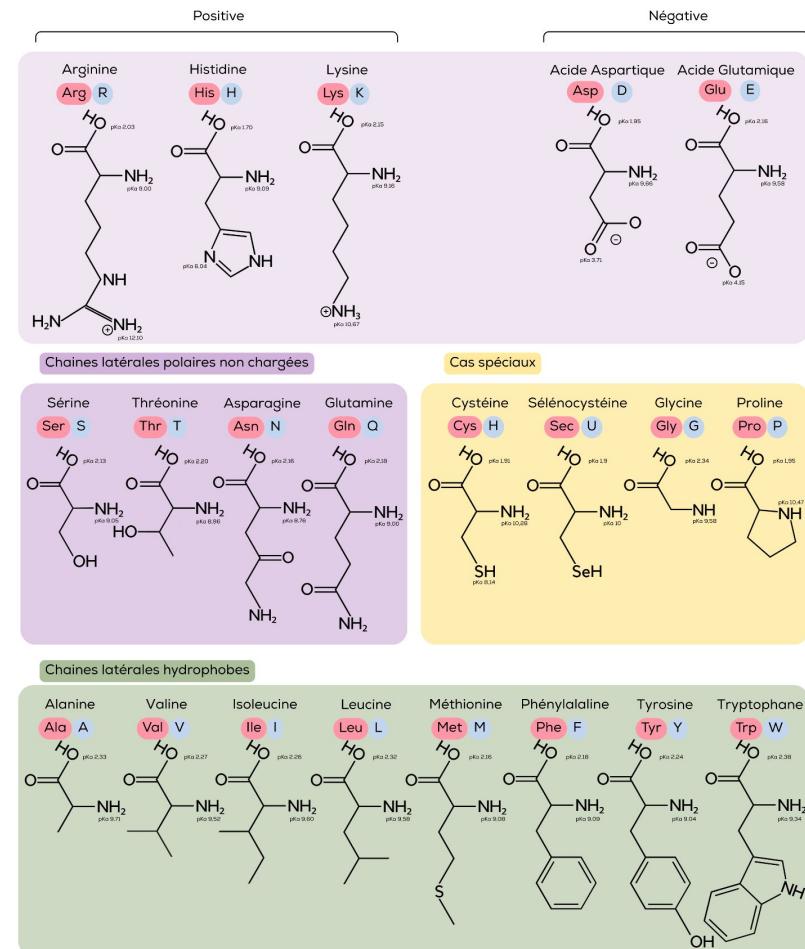
<https://mbg.au.dk/en/news-and-events/news-item/artikel/forskere-bestemmer-foerste-s-truktur-af-protein-som-opretholder-cellemembranen/>



Relation séquence → structure

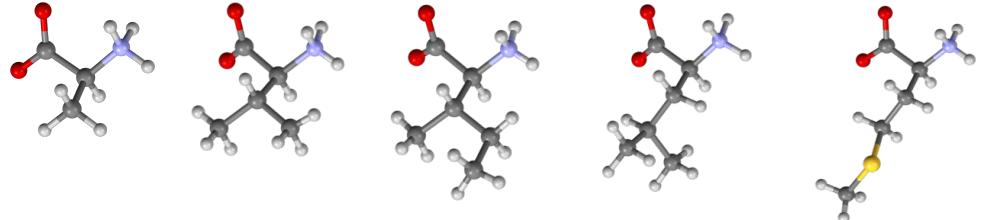
Rappel – Nomenclature et composition des acides aminés

Amino Acid	Abbrev	1-letter	Codon(s)
Alanine	Ala	A	GCA, GCC, GCG, GCT
Arginine	Arg	R	CGA, CGC, CGG, CGT, AGA, AGG
Aspartic acid	Asp	D	GAC, GAT
Asparagine	Asn	N	AAC, AAT
Cysteine	Cys	C	TGC, TGT
Glutamic acid	Glu	E	GAA, GAG
Glutamine	Gln	Q	CAA, CAG
Glycine	Gly	G	GGG, GGC, GGA, GGT
Histidine	His	H	CAC, CAT
Isoleucine	Ile	I	ATT, ATC, ATA
Leucine	Leu	L	CTA, CTC, CTG, CTT, TTA, TTG
Lysine	Lys	K	AAA, AAG
Methionine	Met	M	ATG
Phenylalanine	Phe	F	TTC, TTT
Proline	Pro	P	CCA, CCC, CCG, CCT
Serine	Ser	S	TCA, TCC, TCG, TCT, AGC, AGT
Threonine	Thr	T	ACT, ACC, ACG, AAT
Tryptophan	Trp	W	TGG
Tyrosine	Tyr	Y	TAC, TAT
Valine	Val	V	GTA, GTC, GTG, GTT
STOP	-	-	TAG, TAA, TGA



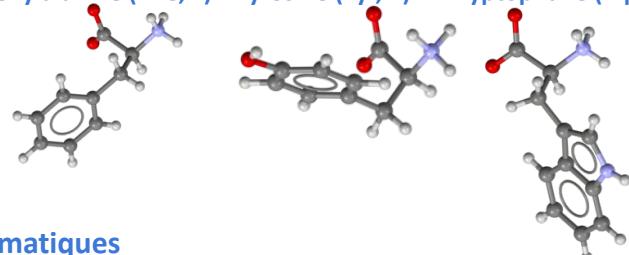
Propriétés biochimiques de quelques acides aminés entrant dans la composition des protéines

Alanine (Ala, A) Valine (Val, V) Isoleucine (Ile, I) Leucine (Leu, L) Méthionine (Met, M)



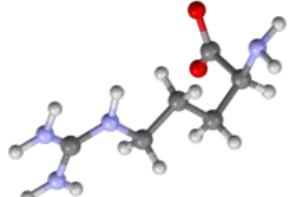
Hydrophobes

Phénylalanine (Phe, F) Tyrosine (Tyr, Y) Tryptophane (Trp, W)

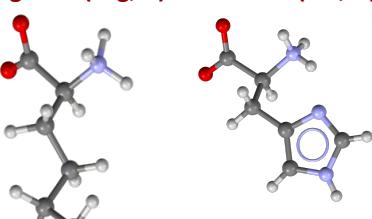


Aromatiques

Lysine (Lys, K)



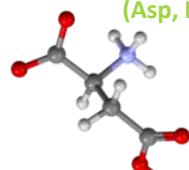
Arginine (Arg, R)



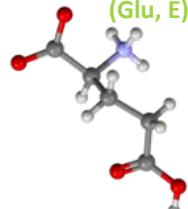
Histidine (His, H)

Chargés positivement

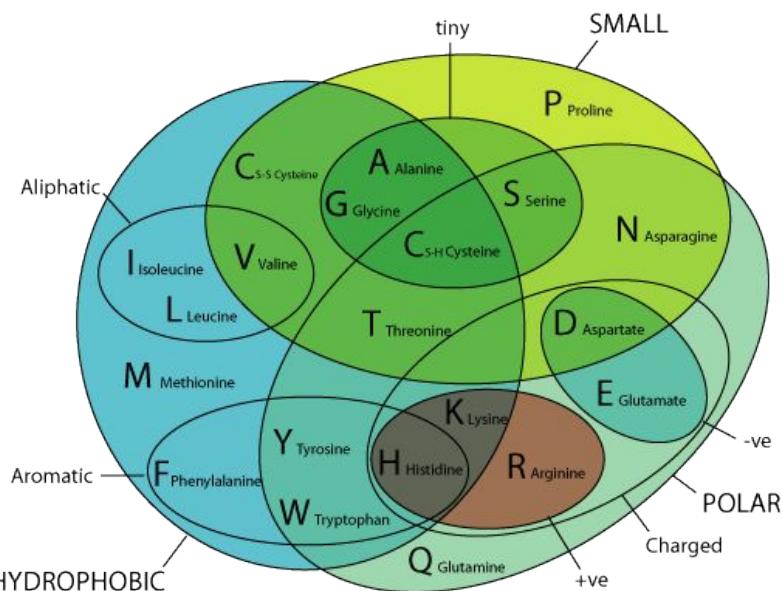
Acide aspartique (Asp, D)



Acide glutamique (Glu, E)



Chargés négativement



Structure primaire

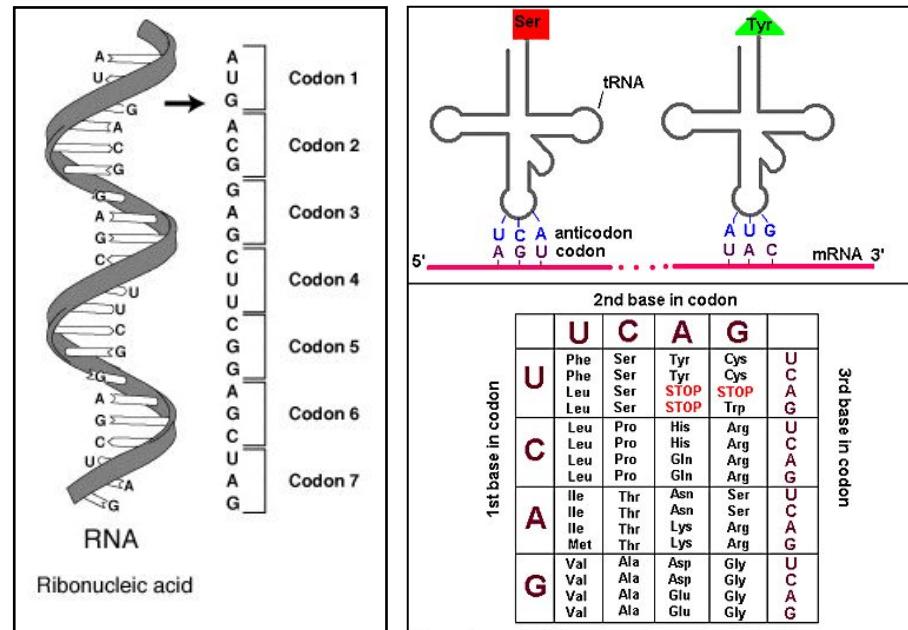
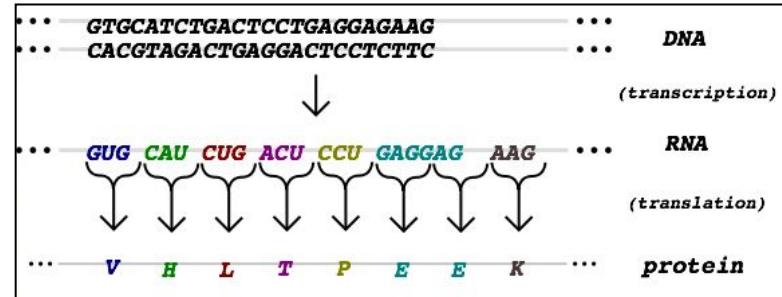
La structure primaire est la **succession linéaire des acides aminés** qui constituent un polypeptide.

Elle se forme durant la **traduction**, au niveau du ribosome, qui catalyse la formation de liens dipeptides suivant la succession des codons de l'ARN messager.

La succession des codons résulte de

- la séquence de l'ADN (transcription)
- la maturation de l'ARN primaire (épissage) en ARN messager
- la position des codons start et stop sur l'ARN messager

La transcription et la maturation de l'ARN font l'objet de différents types de régulation, enseignés lors des cours de biologie moléculaire.



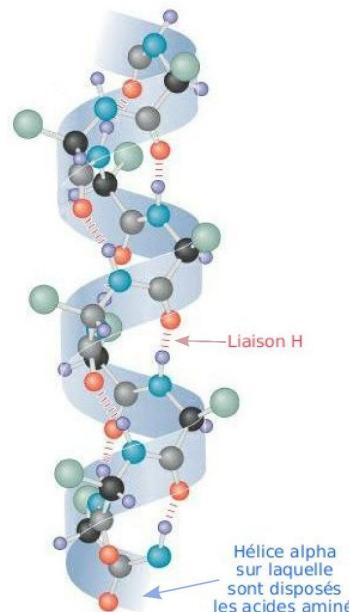
Structure secondaire

La structure primaire détermine à son tour la formation d'éléments de **structure secondaire**, par interactions entre résidus plus ou moins distants :

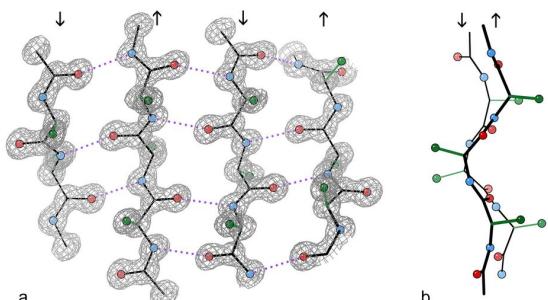
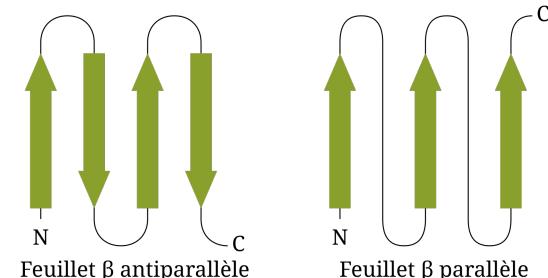
- Reploiement local de la chaîne des acides aminés
- Structures secondaires fréquentes (illustrées plus loin)
- hélices alpha : 3,6 acides aminés par tour → chaque acide aminé interagit, de façon non covalente, avec 2 acides aminés distants (à 3 et 4 pas de distance) dans chaque direction (en amont et en aval dans la chaîne)
 - feuillets beta : interactions entre acides aminés

La formation de ces éléments de structure se fait de façon **dynamique et progressive**, au fil de la concaténation des acides aminés par le ribosome.

Hélice alpha



Feuillets bêta

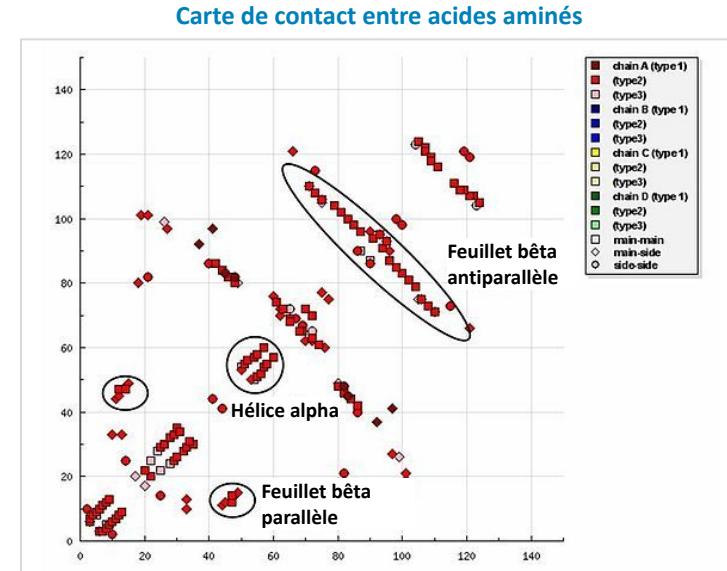
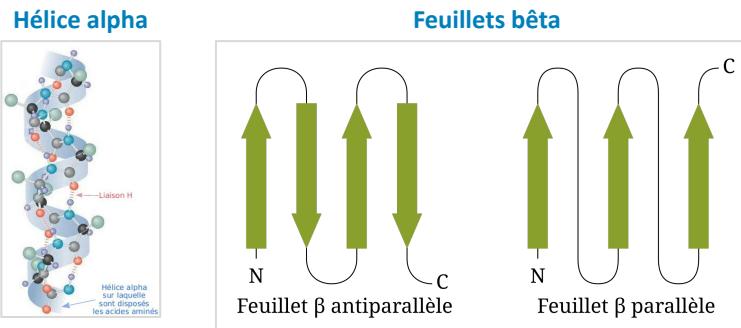


Carte des contacts entre résidus

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

* On considère que deux acides aminés sont en contact si leurs C_α sont séparés de moins de 6 Å (le seuil peut éventuellement varier).

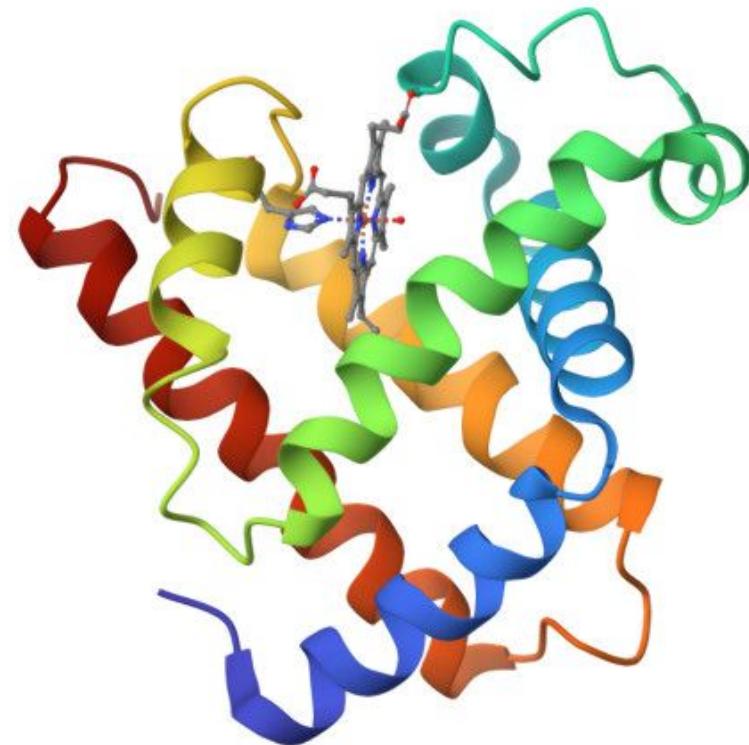


Structure tertiaire

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

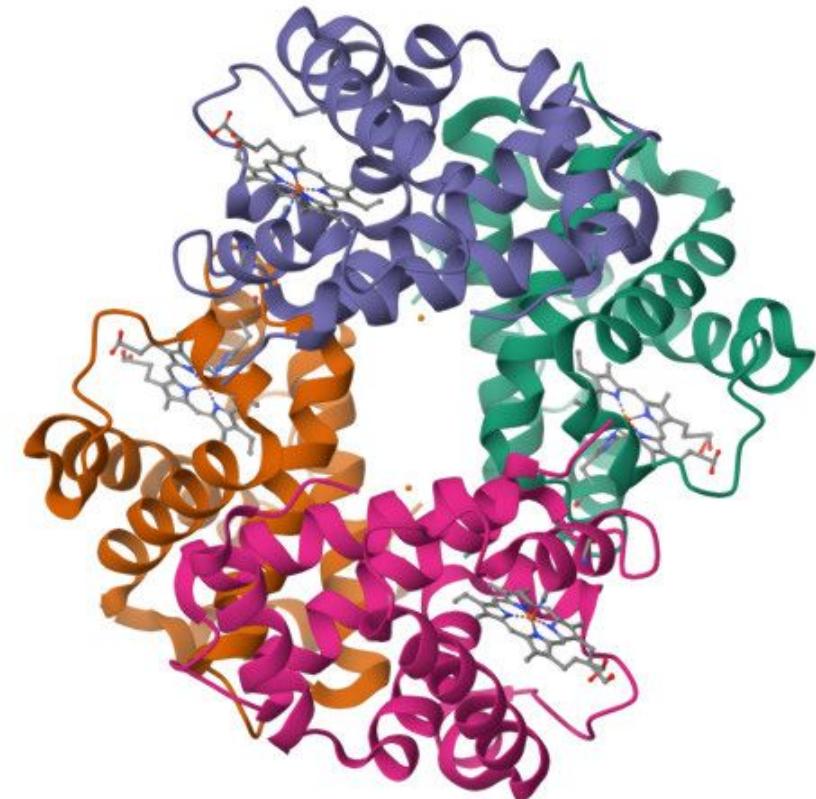
Figure: structure tertiaire de la myoglobine, montrant l'agencement des hélices α autour du groupe hème (structure PDB 1MBN).

- Succession d'**hélices alpha** (marquées par des couleurs différentes), reliées par des **boucles non structurées**.
- L'ordre des éléments est mis en évidence par un dégradé de couleurs.



Structure quaternaire

- La structure quaternaire résulte de l'agrégation
 - de polypeptides avec des groupements prosthétiques (molécules non-protéiques complexées avec un polypeptide)
 - de plusieurs chaînes polypeptidiques en complexes protéiques
- Exemple : Structure quaternaire de l'hémoglobine de cheval, avec coloration par chaîne (structure PDB 4HHB).
 - Noter la présence de **4 chaînes**, correspondant aux 2 sous-unités alpha et aux 2 sous-unités bêta de l'hémoglobine.
 - Chaque chaîne polypeptidique est liée à un groupe prosthétique **hème**, qui assure la liaison avec l'oxygène ou le CO₂.



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

- Au 12 août 2025, UniProt (www.uniprot.org) continent la séquence de 253,6 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.
- **Reviewed (Swiss-Prot)** : parmi les protéines d'UniProt, seule une "petite" fraction de 573 661 (0,22%) ont été annotées par des êtres humains. Elles se trouvent dans la **base de connaissance Swiss-Prot**.
- **Unreviewed (TrEMBL)** : protéines annotées automatiquement, par des méthodes bioinformatiques qui reposent sur leur similarité de séquence avec des protéines connues, accessibles dans la **base de données TREMBL**.



Status

■ Reviewed (Swiss-Prot)
(573,661)

■ Unreviewed (TrEMBL)
(253,061,696)

Exemple de page UniprotKB - 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)
 - ...

The screenshot displays the UniProtKB entry for P01958 · HBA_HORSE. At the top, the UniProt logo and navigation links (Tools, UniProtKB, Advanced, Search, Help) are visible. The main header is "P01958 · HBA_HORSE". Below the header, the protein is identified as "Hemoglobin subunit alpha" from "Equus caballus (Horse)". The "Function" section describes its role in oxygen transport. The "Features" section shows a sequence alignment with positions 54 and 83 highlighted, and two binding sites at positions 59 and 88. The "Gene Ontology" section lists annotations and models.

Function

Names & Taxonomy

Subcellular Location

Phenotypes & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

P01958 · HBA_HORSE

Proteinⁱ Hemoglobin subunit alpha

Geneⁱ HBA

Statusⁱ UniProtKB reviewed (Swiss-Prot)

Organismⁱ Equus caballus (Horse)

Amino acids 142 (go to sequence)

Protein existenceⁱ Evidence at protein level

Annotation scoreⁱ 55

Entry Variant viewer Feature viewer Genomic coordinates Publications External links

Tools Download Add Add a publication Entry feedback

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.

Download

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

54 83

A Q V K A H G K K V G D A L T L A V G H L D D L P G A L S N

Type	ID	Position(s)	Description
Binding site	59	O2 (UniProtKB ChEBI)	[PROSITE-ProRule Annotation]
Binding site	88	Fe (UniProtKB ChEBI)	: proximal binding residue [PROSITE-ProRule Annotation]

Gene Ontologyⁱ

GO annotations GO-CAM models New

Exemple de page UniprotKB - 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 / UniProtKB.

The screenshot shows the UniProtKB entry for P01958. The top navigation bar includes links for BLAST, Align, Peptide search, ID mapping, SPARQL, UniProtKB, Advanced List, Search, and Help. The left sidebar has categories like Function, Names & Taxonomy, Subcellular Location, Phenotypes & Variants, PTM/Processing, Expression, Interaction, Structure (which is selected), Family & Domains, Sequence, and Similar Proteins. The main content area is titled 'Structure' and displays a 3D ribbon model of the protein structure. Below the structure is a table with columns: SOURCE, IDENTIFIER, METHOD, RESOLUTION, CHAIN, POSITIONS, and LINKS. The table lists various PDB entries with their details and links to other databases like PDBe, RCSB-PDB, and PDBsum. At the bottom, there's a note about the 3D structure database.

SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
PDB	1NSQ	X-ray	2.00 Å	A	2-142	PDBe · RCSB-PDB · PDBj · PDBsum · Foldseek
PDB	1NS9	X-ray	1.60 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

3D structure database

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 (12 août 2025)
 - 240 177 structures caractérisées expérimentalement (protéines, ADN, ARN, ...)
 - 1.068.577 modèles prédictifs



240,177
Structures from the PDB archive

1,068,577
Computed Structure Models (CSM)

Polymer Entity Type
<input type="checkbox"/> Protein (235,819)
<input type="checkbox"/> DNA (12,263)
<input type="checkbox"/> RNA (8,954)
<input type="checkbox"/> NA-hybrid (286)
<input type="checkbox"/> Other (12)

The screenshot shows the RCSB PDB homepage with the following sections:

- Header:** Deposit, Search, Visualize, Analyze, Download, Learn, About, Careers, COVID-19, Help, Contact us, MyPDB.
- Statistics:** 240,685 Structures from the PDB archive, 1,068,577 Computed Structure Models (CSM).
- Search Bar:** Enter search term(s), Ligand ID or sequence, Include CSM, Advanced Search, Browse Annotations.
- Navigation:** Welcome, Deposit, Search, Visualize, Analyze, Download, Learn.
- Features & Highlights:** Redesigned PDB Statistics Support Enhanced Functionality, Experimentally-determined 3D structures from the Protein Data Bank (PDB) archive, Integrative 3D Structures from the PDB Archive, Computed Structure Models (CSM) from AlphaFold DB and ModelArchive, NEW Explore Integrative Structures, PDB-101 Training Resources.
- August Molecule of the Month:** Arc, a complex protein structure.
- Latest Entries:** As of Tue Aug 12 2025, 9D3N, 167-bp 5S rRNA nucleosome cross-linked with glutaraldehyde.
- Features & Highlights:** Register for the Sept 25 Virtual Office Hour on PDB Policies, Register for the Aug 11 Virtual Office Hour on Pairwise Alignment, Register Now for Webinar: Searching the PDB for high-quality ligand-bound structures.
- News:** Paper Published: PDB-IHM, Paper Published: rcsb-api.
- Publications:** Paper Published: PDB-IHM, Paper Published: rcsb-api.

Structure tridimensionnelle de la myoglobine dans PDB

RCSB PDB

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

PDB 101 · www.PDB.org · EMDataResource · NAKB · wwPDB Foundation · PDB-IHM

Help · Contact us · MyPDB ·

PDB PROTEIN DATA BANK

240,665 Structures from the PDB archive · 1,068,577 Computed Structure Models (CSM)

Enter search term(s), Ligand ID or sequence · Include CSM · Help

Advanced Search | Browse Annotations

Structure Summary · Structure · Annotations · Experiment · Sequence · Genome · Versions

Biological Assembly 1

1MBN | pdb_00001mbn ⓘ

The stereochemistry of the protein myoglobin

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

Classification: OXYGEN STORAGE
Organism(s): *Physeter macrocephalus*
Mutation(s): No

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

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.00 Å

wwPDB Validation

Metric	Percentile Ranks	Value
Chirality	54	9.3%
Ramachandran outliers	1.3%	15.2%
Sidechain outliers	80	0.0%

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

Literature · Download Primary Citation

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

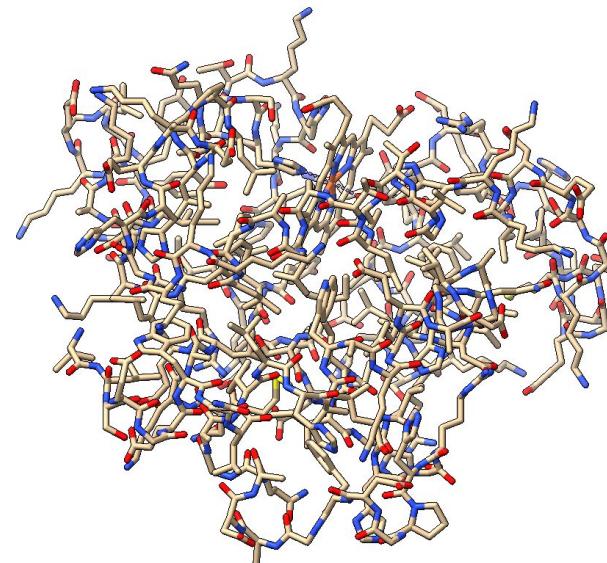
Macromolecule Content

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

Find Similar Assemblies

Biological assembly 1 assigned by authors.

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



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

Figure réalisée avec le logiciel ChimeraX
Source des données : <https://www.rcsb.org/structure/1MBN>

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 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): Phyceter catodon
Mutation(s): No

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

Experimental Data Snapshot

wwPDB Validation

Method: X-RAY DIFFRACTION Resolution: 2.00 Å

Metric Percentile Ranks Value

Claesson	34	3.3%
Ramachandran outliers	3.3%	15.2%
Sidechain outliers	0	0

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.

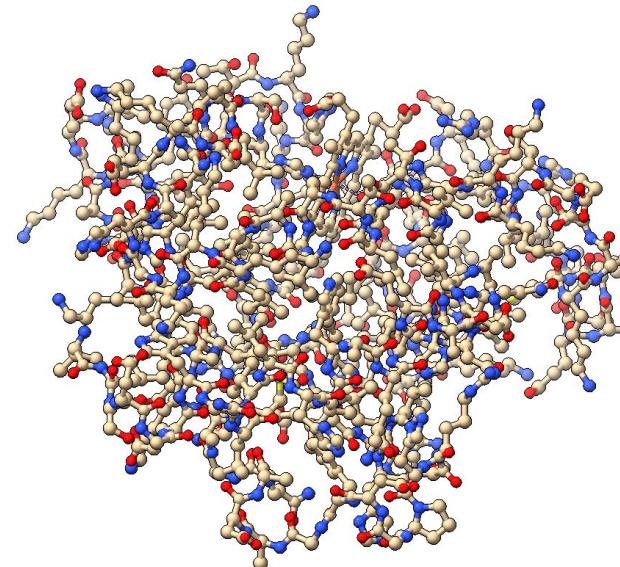
Macromolecule Content

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

<https://www.rcsb.org/structure/1MBN>

Structure de la myoglobine (1MBN)

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



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

Figure réalisée avec le logiciel ChimeraX

Source des données : <https://www.rcsb.org/structure/1MBN>

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-Resoures 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): Phyceret catodon
Mutation(s): No

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

Experimental Data Snapshot

wwPDB Validation

Method: X-RAY DIFFRACTION Resolution: 2.00 Å

Metric	Percentile Ranks	Value
Claesson	34	3.3%
Ramachandran outliers	3.3%	15.2%
Sidechain outliers	N/A	N/A

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

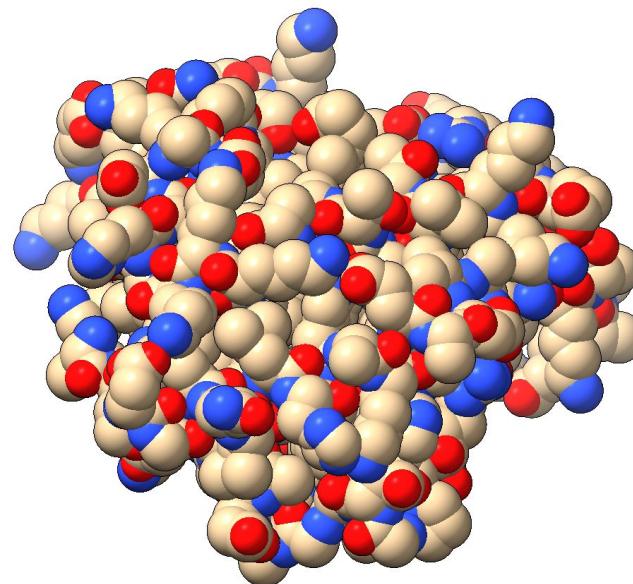
- Total Structure Weight: 17.87 kDa
- Atom Count: 1,260
- Modelled Residue Count: 153
- Deposited Residue Count: 153
- Unique protein chains: 1

<https://www.rcsb.org/structure/1MBN>

Structure de la myoglobine (1MBN)

Modèle compact ("Space-filling")

Espace occupé par chaque atome (sphères de van der Waals)



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

Figure réalisée avec le logiciel ChimeraX

Source des données : <https://www.rcsb.org/structure/1MBN>

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 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): Physeter catodon
Mutation(s): No

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

Experimental Data Snapshot

wwPDB Validation

Method: X-RAY DIFFRACTION Resolution: 2.00 Å

Metric Percentile Ranks Value

Claesson	34	3.3%
Ramachandran outliers	3	15.2%
Sidechain outliers	0	N/A

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

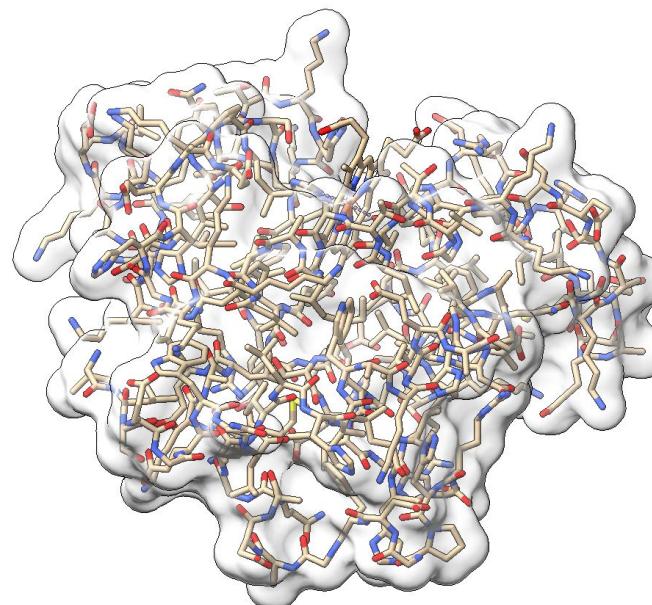
Biological assembly 1 assigned by authors.

Macromolecule Content

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

<https://www.rcsb.org/structure/1MBN>

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



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

Figure réalisée avec le logiciel ChimeraX
Source des données : <https://www.rcsb.org/structure/1MBN>

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 CSM Advanced Search | Browse Annotations Help

PDB-101 EMD-Reserve 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): 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	3.4	3.4
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

Find Similar Assemblies

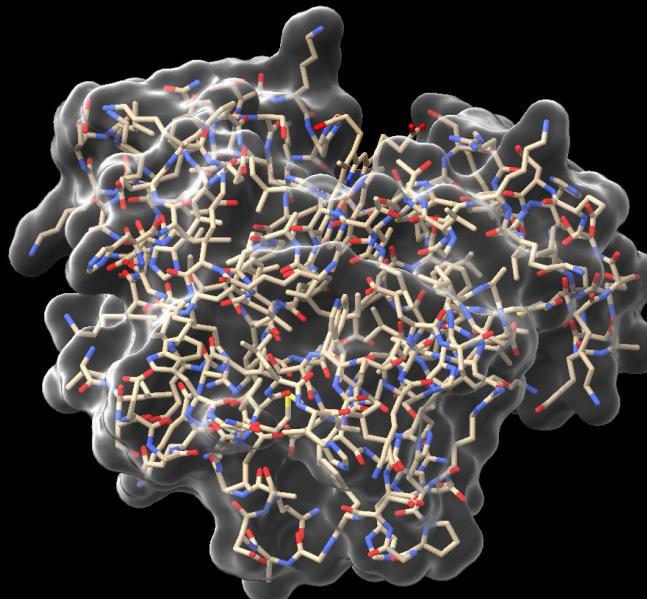
Global Symmetry: Asymmetric - C1
Global Stoichiometry: Monomer - A1

Macromolecule Content

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

<https://www.rcsb.org/structure/1MBN>

Structure de la myoglobine (1MBN) Surface accessible au solvant (affichage "Ghostly white")



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

Figure réalisée avec le logiciel ChimeraX

Source des données : <https://www.rcsb.org/structure/1MBN>

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 PDB EMDataResource 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): Physeter 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

wwPDB Validation

Method	Metric	Percentile Ranks	Value
X-RAY DIFFRACTION	Claesson	34	3.3%
	Ramachandran outliers	3.3%	15.2%
	Sidechain outliers	N/A	N/A

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

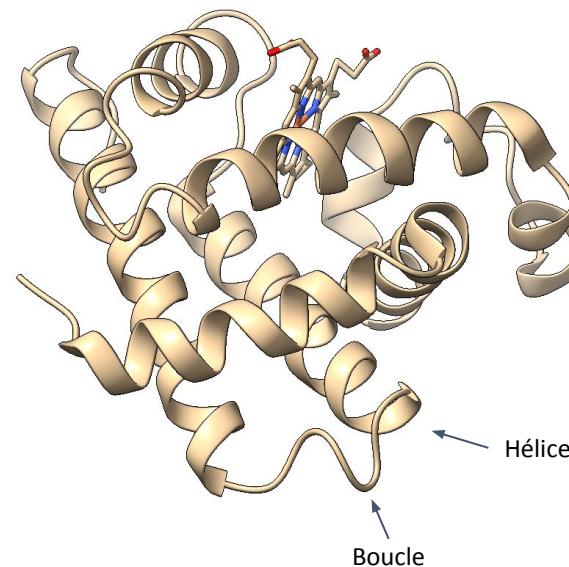
Literature

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

<https://www.rcsb.org/structure/1MBN>

Structure de la myoglobine (1MBN) Mode rubans (“ribbons”)

Mise en évidence des structures secondaires



Code couleur CPK

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

PDB PROTEIN DATA BANK 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 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): Phyceter 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

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 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): Phyceter 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

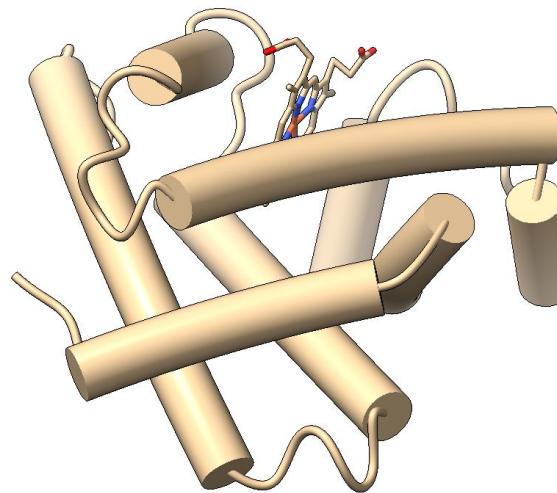
Find Similar Assemblies

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”)



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

<https://www.rcsb.org/structure/1MBN>

Figure réalisée avec le logiciel ChimeraX
Source des données : <https://www.rcsb.org/structure/1MBN>

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-Resource 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.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

Display Files ▾ Download Files ▾ Data API

wwPDB Validation

Metric	Percentile Ranks	Value
Claesson	34	3.3%
Ramachandran outliers	3.3%	15.2%
Sidechain outliers	0	N/A

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

Literature

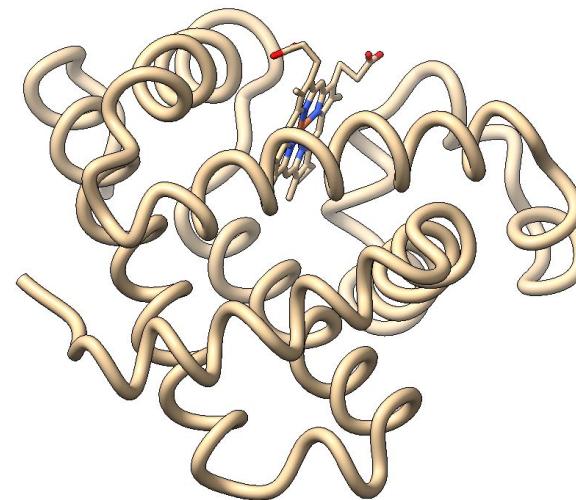
Download Primary Citation ▾

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

<https://www.rcsb.org/structure/1MBN>

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

* également orthographié liquorice (réglisse en français)



Code couleur CPK

Beige: carbone

Bleu: azote

Rouge: oxygène

Figure réalisée avec le logiciel ChimeraX
Source des données : <https://www.rcsb.org/structure/1MBN>

6 vues de la myoglobine

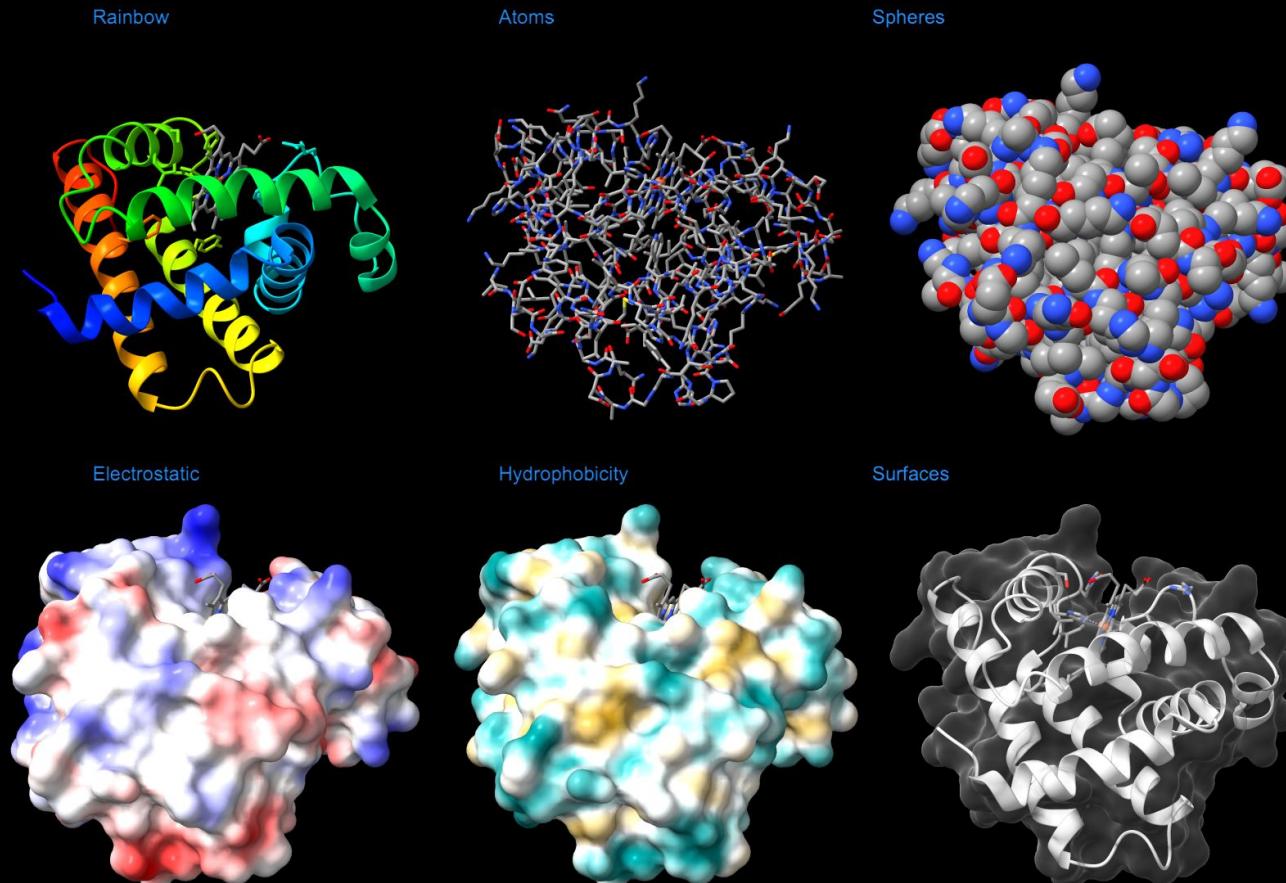


Figure réalisée avec le logiciel ChimeraX

Source des données : <https://www.rcsb.org/structure/1MBN>

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

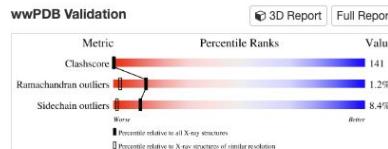
Organism(s): Homo sapiens

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

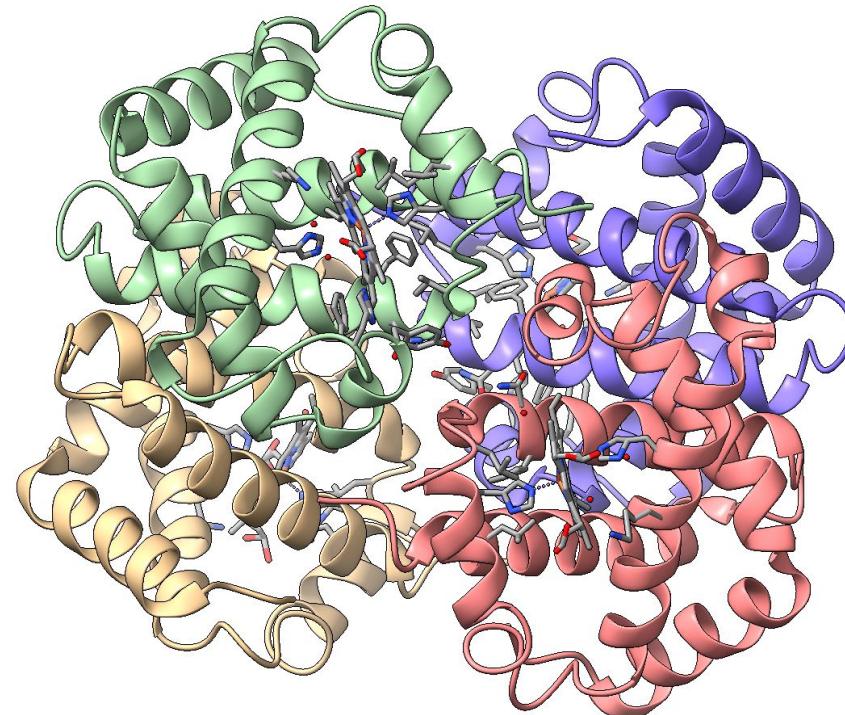
DOI: [https://doi.org/10.1016/0022-2836\(84\)90472-8](https://doi.org/10.1016/0022-2836(84)90472-8)

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](#)

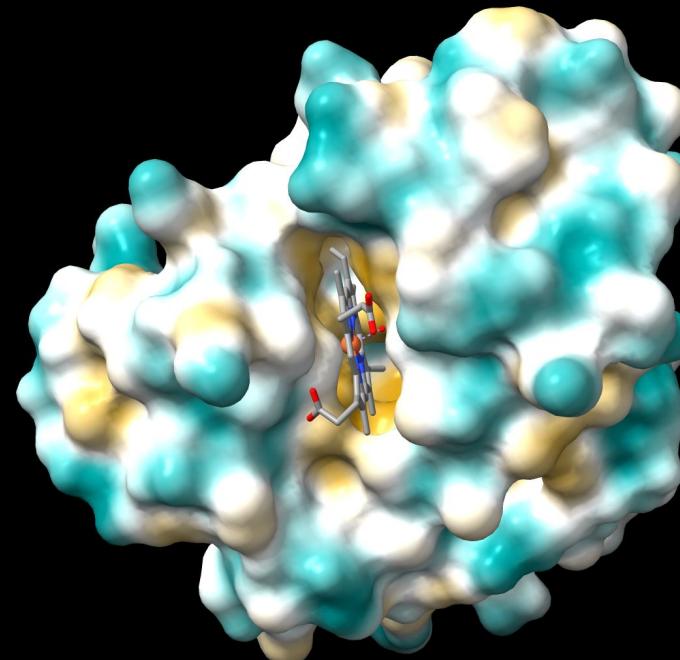
Relations structure - fonction : quelques exemples illustratifs

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)



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

Animation réalisée avec le logiciel ChimeraX

Source des données : <https://www.rcsb.org/structure/1MBN>

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

Cartoon (haut)

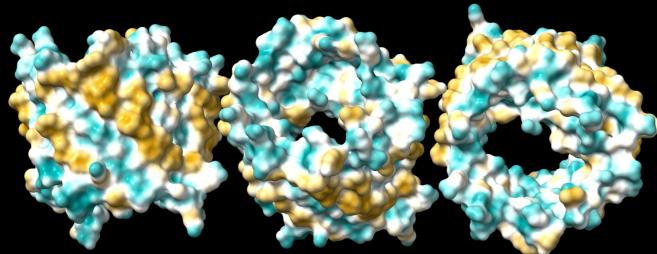
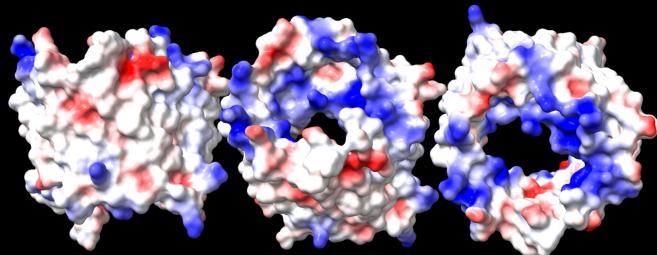
- La **porine du sucre** est une protéine formée en majorité par des feuillets bêta antiparallèles, dont l'agencement forme un cylindre. Cette topologie est dénommée **tonneau bêta** (*beta barrel* en anglais)
- L'intérieur du cylindre est partiellement occupé par deux petites hélices alpha

Profil hydrostatique (milieu)

- 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



Figures réalisées avec le logiciel ChimeraX

Source des données : structure [PDB 1A0S](#)

Animation:

https://drive.google.com/file/d/1SIC_YTOVOIKBA_KHxF0CvYmUelTbvmi7

Prédiction de la structure tridimensionnelle des protéines

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 (prise en compte de l'encombrement stérique du radical de chaque acide aminé, optimisation d'énergie).
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.

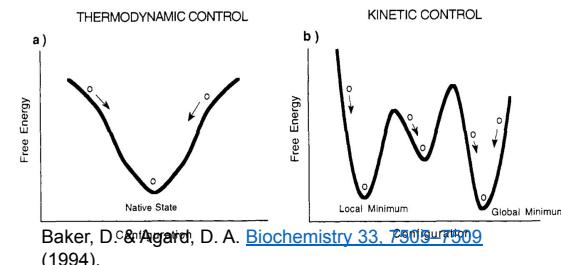
Voici les principales approches

- Minimisation d'énergie
- Dynamique moléculaire
- Recuit simulé

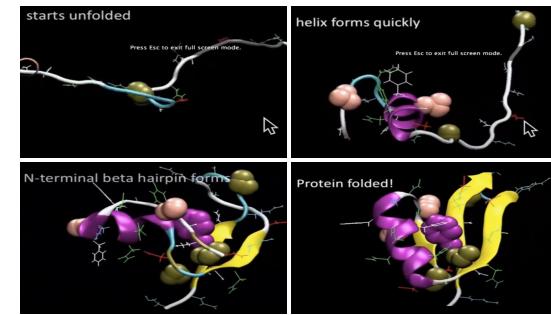
La description de ces approches sort largement du cadre d'un cours d'introduction à la bioinformatique, elles seront présentées dans des cours de biochimie/bioinformatique structurale. Les figures sont fournies uniquement à titre d'illustration.

Les méthodes présentées dans cette diapo ne font pas partie de la matière d'examen.

Minimisation d'énergie

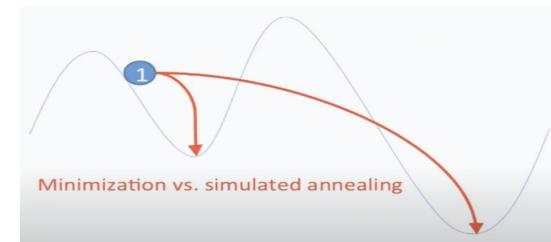


Dynamique moléculaire



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

Recuit simulé



Minimization vs. simulated annealing

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

doi.org/10.1021/bi00190a002

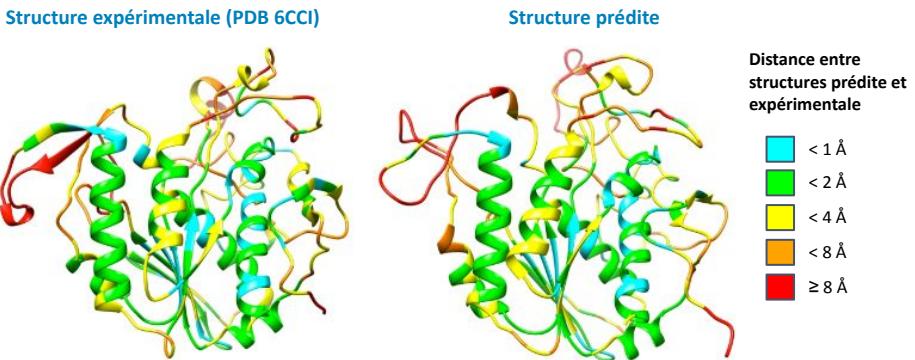
• MIT OpenCourseWare "Predicting Protein Structure". <https://youtu.be/j1s9JfZKFqU?t=806>

Critical Assessment of protein Structure Prediction (CASP)

Critical Assessment of Structure Prediction (CASP)

Depuis 1994, la communauté de structuralistes organise tous les deux ans 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).

- **Sélection des protéines cibles** : des chercheurs qui viennent de caractériser expérimentalement une structure s'engagent à ne pas la publier avant l'issue de cette session de CASP. Les organisateurs sélectionnent parmi ces propositions une cinquantaine de **protéines cibles**.
- **Principe du double aveugle** : les chercheurs qui ont caractérisé la structure communiquent les séquences aux organisateurs, qui les publient sur le site Web de CASP, mais ni les participants ni les organisateurs ne connaissent la structure réelle lors de la soumission des prédictions.
- **Prédictions** : les bioinformaticiens structuraux utilisent leurs différentes méthodes pour prédire la structure à partir de chaque séquence
- **Soumission des modèles** : les participants soumettent leurs prédictions via une plateforme dédiée dans le créneau de temps de 3 semaines.
- **Analyse et évaluation** : les structures expérimentales sont enfin révélées et comparées aux modèles soumis pour évaluation.



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

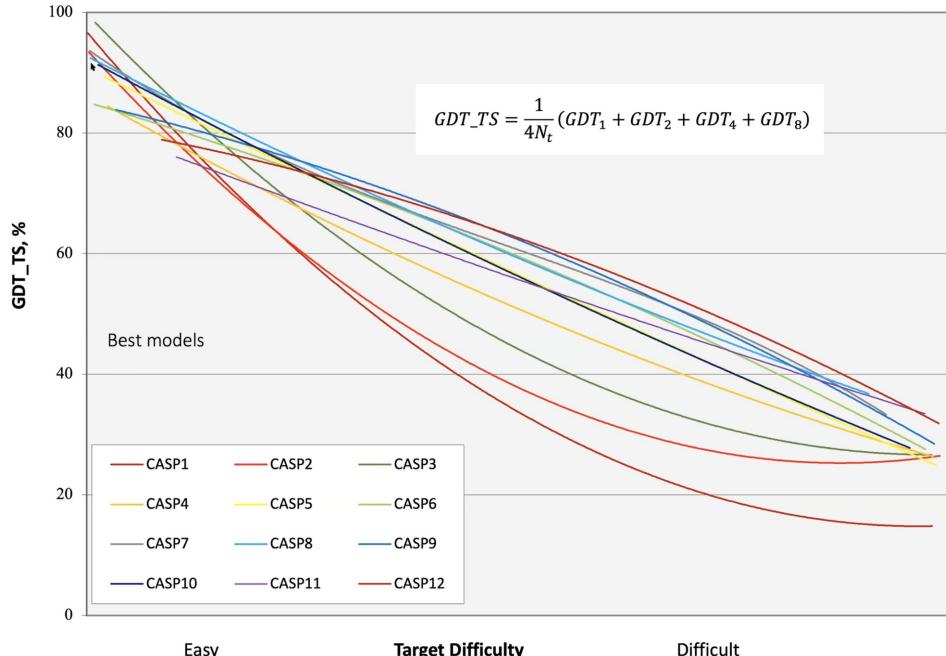
Qualité des prédictions de CASP

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

- Abscisse: degré de difficulté (protéine expérimentale plus ou moins complexe)
- Ordonnée: précision des prédition
 - 100: tous les atomes prédis sont exactement à la position des atomes de la structure expérimentale
 - 90 - 100 : prédition du même niveau que la structure expérimentale (les structures expérimentales varient selon les conditions)
 - ~50 : prédition 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 : niveau de précision attendu pour un modèle aléatoire (aucune valeur prédictive)

On constate

- Amélioration progressive de CASP1 (1994) à CASP5 (2002)
- Relative stagnation de CASP5 à CASP12 (2016)



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

Quand l'IA est entré dans le jeu

Définition

Ensemble des méthodes visant à faire exécuter par des ordinateurs des tâches qui relèvent de l'intelligence humaine.

Applications récentes populaires de l'IA

- Détection d'objets dans des images
- Reconnaissance faciale
- Transcription automatique de la parole
- Traduction instantanée de texte
- IA génératives capables d'entretenir un dialogue avec l'utilisateur

Toutes ces applications reposent sur un type particulier d'IA: les **réseaux neuronaux profonds**.

Le domaine de l'IA est cependant plus ancien, et inclut une variété de méthodes.

Approches en IA

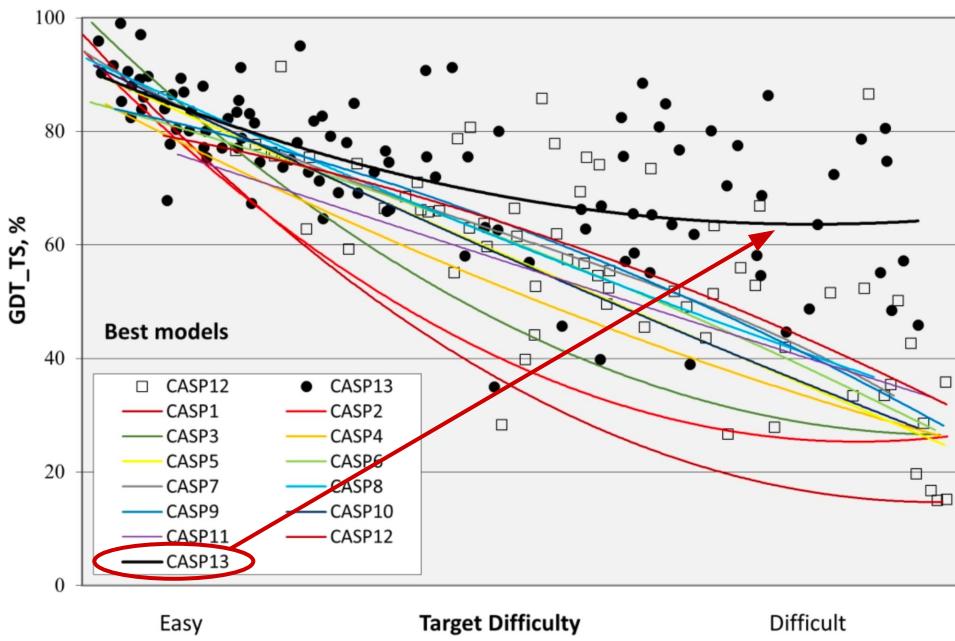
- **IA symbolique:** exploitation de connaissances explicites (bases de connaissances) au moyen de règles logiques
 - **Systèmes experts.**
 - Exemple: prescription d'antibiotique sur base de règles établies en collectant expertise de médecins
 - **Algorithmes de recherche et de planification**
 - Exemple: recherche d'itinéraire
- **IA statistique:** algorithmes d'**apprentissage automatique**
 - Méthodes "classiques": analyse discriminante, arbres de décision, support vector machines
 - Réseaux neuronaux, et en particulier **réseaux neuronaux profonds**.

Exemples d'apprentissage profond en biologie-santé

- Détection de tumeurs, anomalies, lésions dans des images (radiographie, histopathologie).
- Développement de nouvelles molécules à visées thérapeutiques (*drug design*).
- Prédiction de la structure 3D des protéines

CASP13 (2018) – Les réseaux neuronaux profonds convolutifs

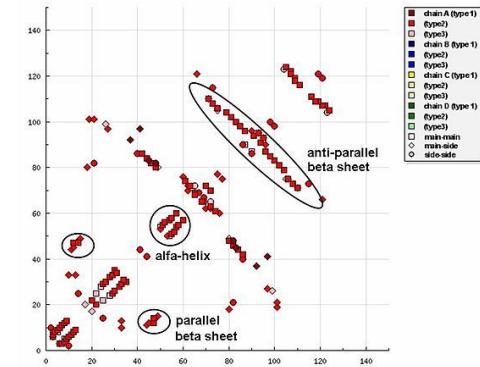
- Amélioration progressive de CASP1 à CASP5
- Relative stagnation de CASP5 à CASP12
- CASP13 (2018) : **bond quantitatif**
 - Les prédictions dépassent 60% de précision pour tous les niveaux de difficulté → prédiction correcte de la topologie des protéines, mais incertitudes sur les positions précises 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)



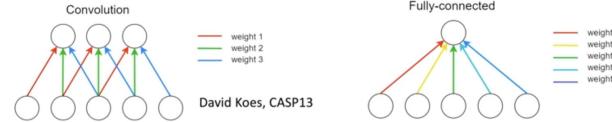
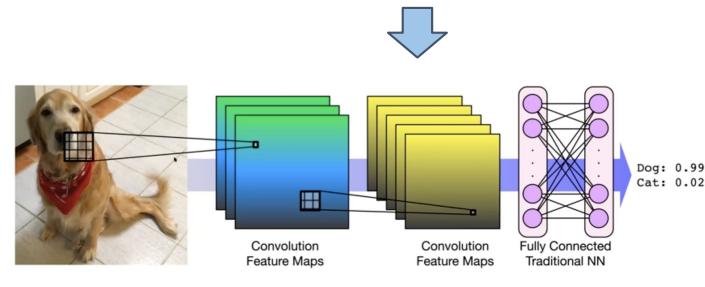
Prédiction de structure sur base des interactions entre résidus

Pour CASP13, 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 CASP14 (2020)

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.

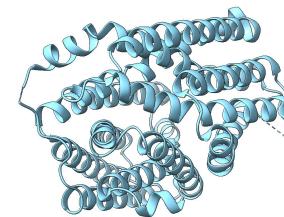
L'histoire d'une structure cristalline

- 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.

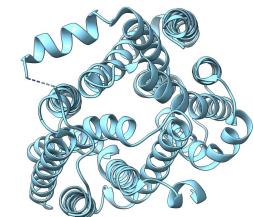
Debruycker, V. et al. An embedded lipid in the multidrug transporter LmrP suggests a mechanism for polyspecificity. *Nat Struct Mol Biol* 27, 829–835 (2020).
doi.org/10.1038/s41594-020-0464-y

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

Vue latérale

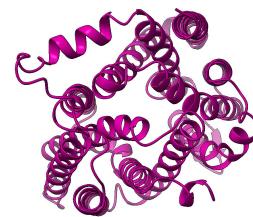
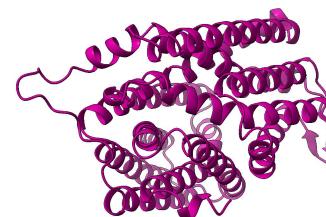


Vue dans l'axe transmembranaire

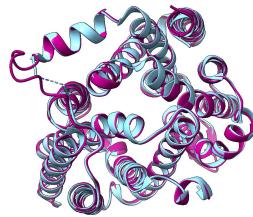
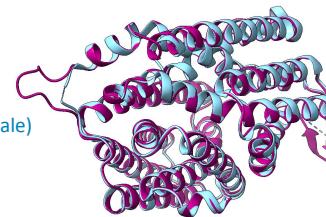


Structure cristallographique
(PDB [6T1Z](https://www.rcsb.org/structure/6T1Z))

Structure prédictive



Structures alignées
(minimisation de la distance spatiale)



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

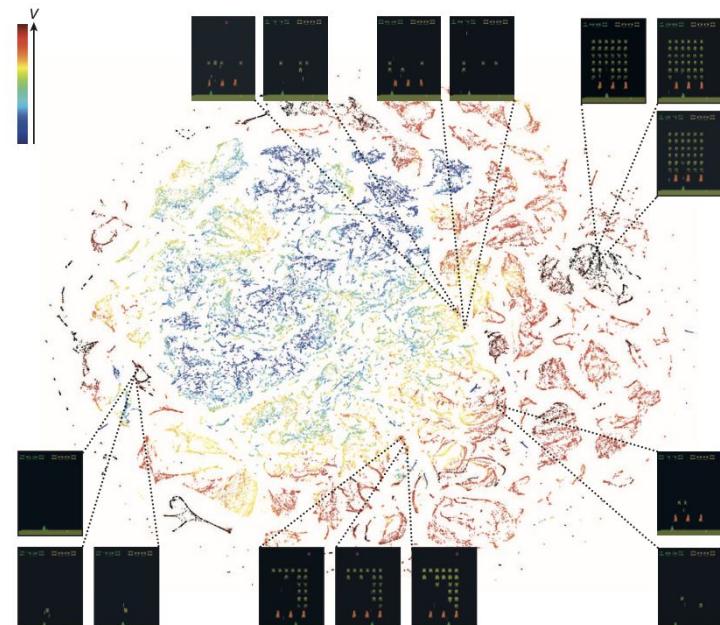
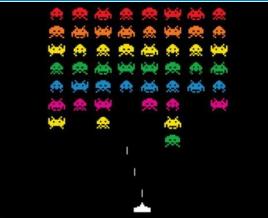
Et DeepMind créa AlphaFold

- 2010 : Fondation de la compagnie DeepMind à Londres (Hassabis, Legg, Suleyman).
 - Objectif : développer une IA générale basée sur l'apprentissage profond.
- 2012–2015 : **Apprentissage par renforcement** sur jeux Atari des années 80 (ex. Breakout, Space Invaders).
- 2014: DeepMind est racheté par Google
- 2015–2017 : **AlphaGo** bat successivement les champions mondiaux de Go.
- 2020 : **AlphaFold** (version 1), prédiction de structure des protéines sur base de leur séquence.
- 2021 : **AlphaFold2**, publié en open source ; plus de 200 millions de structures protéiques rendues disponibles.
- 2023: **Gemini** (Generalized Multimodal Intelligence Network), grand modèle de langage capable de combiner une analyse de sons, vidéos, textes.
- 2023 : Lancement de **Gemini** (modèle multimodal texte, image, audio, vidéo, code).

Illustration: Space Invaders

Haut: exemple d'écran du jeu Space Invaders

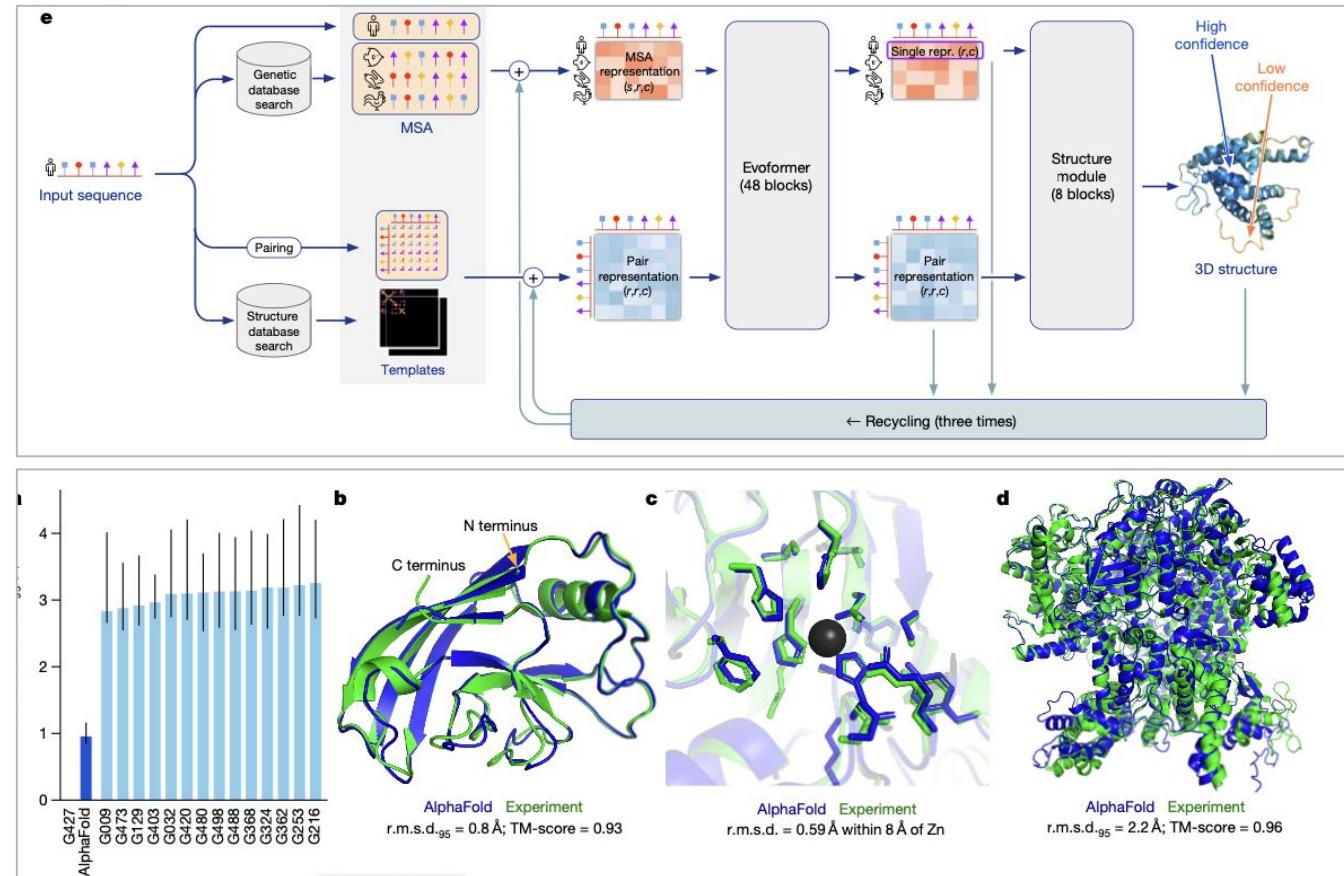
Bas: dernière “couche neuronale” (informatique) de DeepMind après quelques heures d'apprentissage de Space Invaders



Mnih, V. et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). doi.org/10.1038/nature14236

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 couches)
 - 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)

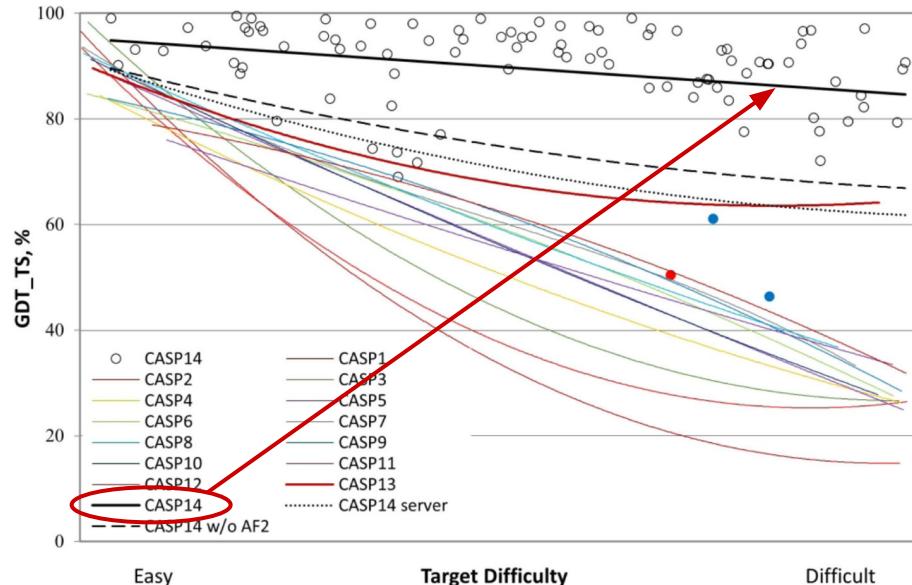


Performances d'AlphaFold2 lors de CASP14 (2020)

Pour la session **CASP14 (2020)**, la précision des prédictions dépasse de très loin celles de toutes les éditions précédentes de CASP.

- Les résultats sont excellents pour tous les degrés de difficulté des protéines cibles.
- Quelques cibles ont des résultats plus faible (marquées en couleur).

Cette amélioration spectaculaire des résultats provient d'**AlphaFold2**, IA spécialisée pour la prédition de structures protéiques.



Kryshtanovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins* 89, 1607–1617 (2021).
doi.org/10.1002/prot.26237

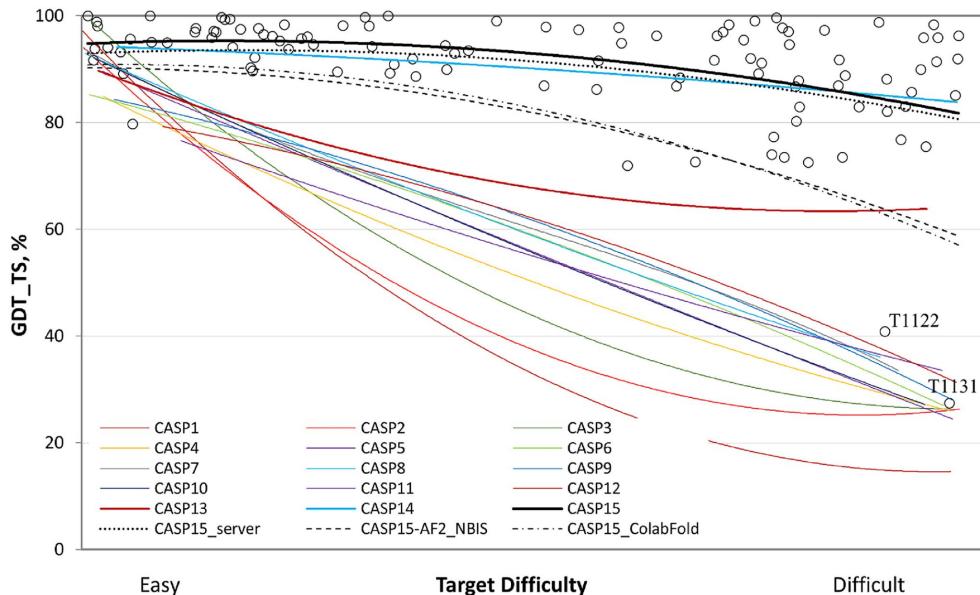
CASP14 Day 1 : Intro by John Moult : Alphafold2 results
(<https://youtu.be/EFwO1LX0eZY?t=1658>)

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

CASP15 (2022)

CAPS15 (2022) donne des résultats du même ordre que pour CASP14 (2020).

- Les prédictions atteignent généralement une excellente précision (>90%), sur toute l'échelle de difficulté.
- Seules 2 cibles donnent de très mauvais résultats.



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 – Limitations et défis

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.
- Ceci a exercé une profonde transformation sur les pratiques les chercheurs, qui disposent désormais de prédictions relativement fiables pour toutes les protéines.
- 2020: le code d'AlphaFold2 est immédiatement accessible publiquement, et le logiciel est déployé sur des serveurs accessibles gratuitement (avec certaines limitations).
- 2024: AlphaFold3, le code ne devient accessible quelques mois après la publication des résultats

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

Le prix Nobel de chimie 2024 décerné à l'utilisation d'IA pour la structure des protéines

Le prix Nobel de chimie 2024 est décerné à 3 personnes pour récompenser leurs travaux qui ont permis une percée sans précédent dans le domaine de la prédiction de structures tridimensionnelles des protéines en utilisant des méthodes d'intelligence artificielle (réseaux neuronaux profonds).

- **David Baker** : conception de nouvelles protéines (design)
- **Demis Hassabis & John Jumper** : concepteurs d'AlphaFold2, pour la prédiction de la structure des protéines à partir de leur séquence.

Nobel Prize in Chemistry 2024



© Nobel Prize Outreach. Photo:
Clément Morin

David Baker

Prize share: 1/2



© Nobel Prize Outreach. Photo:
Clément Morin

Demis Hassabis

Prize share: 1/4



© Nobel Prize Outreach. Photo:
Clément Morin

John Jumper

Prize share: 1/4

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John Jumper "for protein structure prediction"

Matériel supplémentaire (ne fait pas partie de la matière d'examen)

Une **base de données** est une ressource logicielle permettant de stocker et d'interroger des ensembles de données structurées, généralement homogènes par leur nature.

Elle peut être alimentée automatiquement, sans intervention humaine directe.

C'est le cas de TrEMBL, qui regroupe toutes les séquences protéiques obtenues par traduction de séquences nucléotidiques, avec des annotations générées automatiquement.

Une **base de connaissances** est une base de données qui intègre, en plus des données brutes, des informations interprétées, organisées et validées par des experts du domaine.

Ces annotations peuvent concerner la fonction biologique, la structure tridimensionnelle ou d'autres propriétés pertinentes.

C'est le cas de Swiss-Prot, dont chaque entrée est enrichie par un travail manuel de curateurs spécialisés.

Toute base de connaissances est donc une base de données, mais la réciproque n'est pas vraie.