

La vie résulte d'un extraordinaire bricolage moléculaire que l'évolution a rendu étonnement varié et fantastiquement complexe, tout particulièrement au niveau des protéines.

**Du gène à la protéine fonctionnelle:
les mécanismes moléculaires d'un double
décodage.**

Quelques rappels élémentaires de biologie

Notre corps est fait d'organes solides (cœur, poumons, muscles, ...) ou liquides (sang, lymphe, ...)

Nos organes sont faits de cellules, différentes pour chaque organe.

De quoi sont faites nos cellules ?

La cellule eucaryote

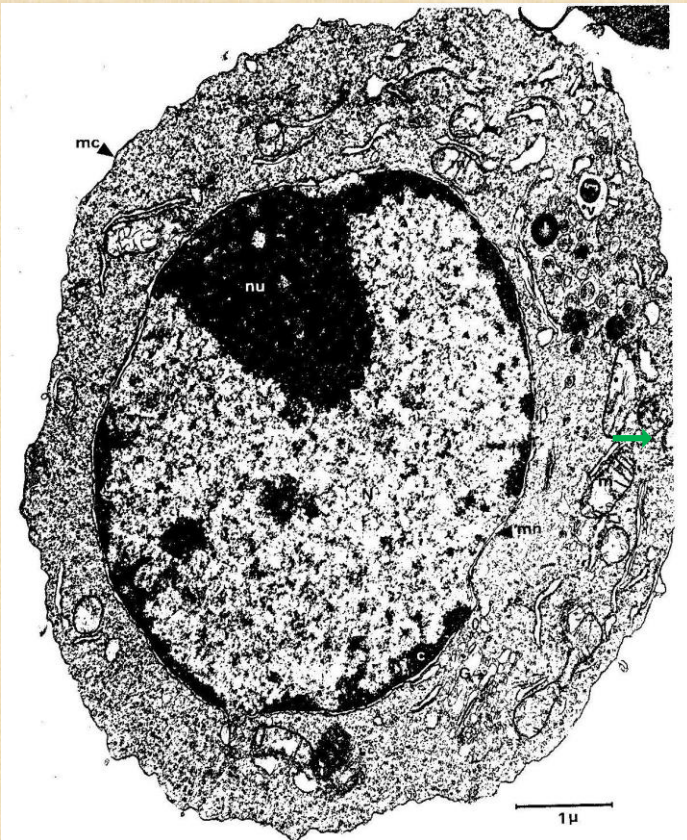
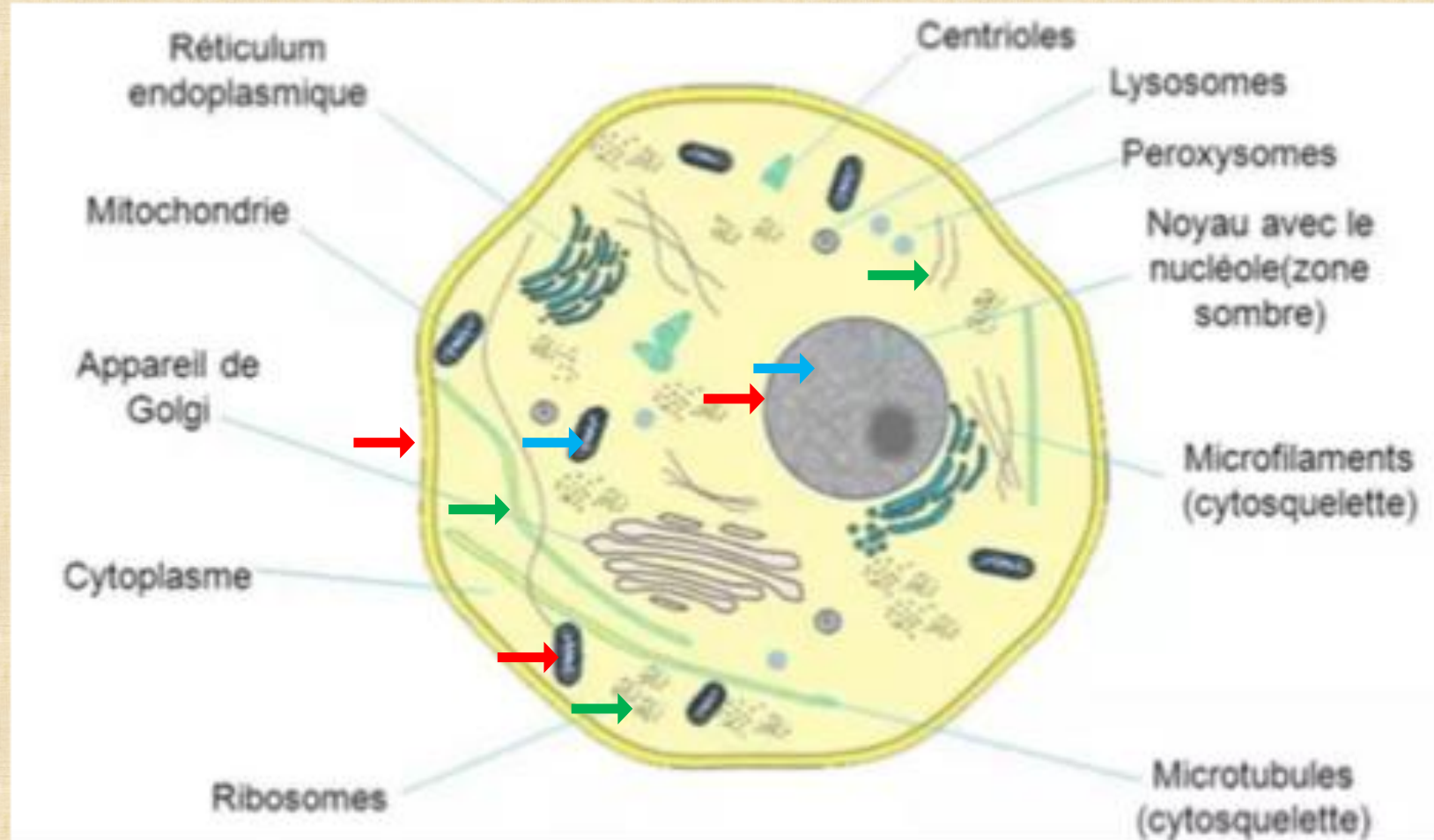


Fig. - Coupe d'une cellule eucaryote. $\times 25\,000$ (Institut Pasteur de Lille, Laboratoire de microscopie électronique, Pr VIVIER).
mc : membrane cytoplasmique ; m : mitochondries ; v : vacuole ; l : lysosome ; mn : membrane nucléaire ; nu : nucléole ; r : ribosomes ; N : noyau ; m : mitochondrie ; c : chromatine ; re : réticulum endoplasmique ; G : appareil de Golgi.

Principaux éléments



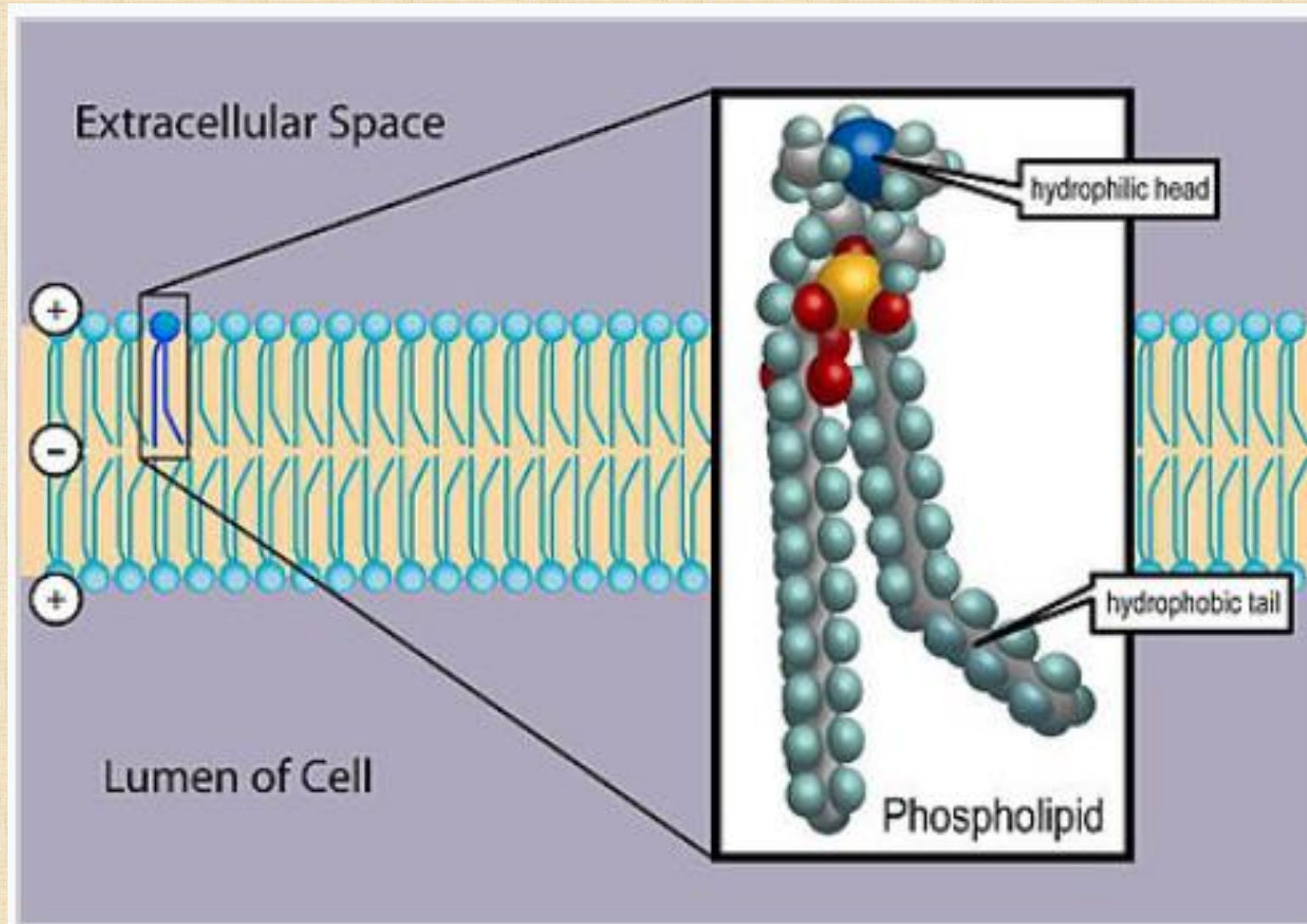
Membranes (*cytoplasmique, nucléaire, mitochondriale, ...*)

Matériel génétique (*chromosomes, mitochondries*)

Protéines (*plusieurs milliers de protéines différentes*)

De quoi sont faits ces trois éléments ?

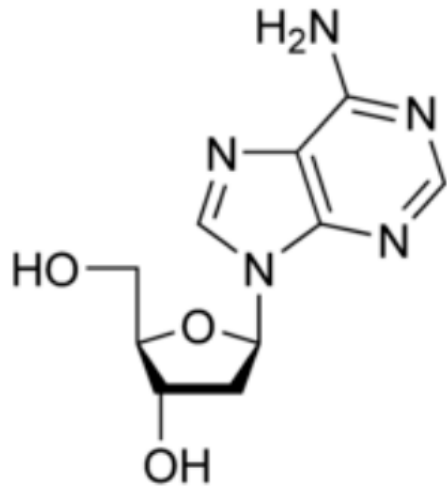
Les membranes sont faites de phospholipides: bicouche lipidique



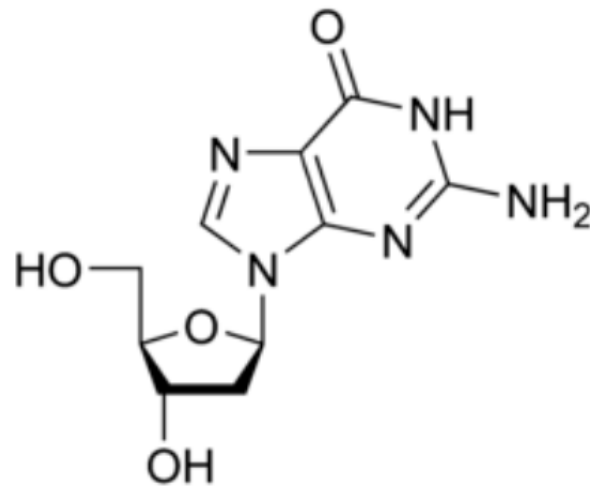
Le matériel génétique est constitué d'ADN

L'ADN est un enchaînement de millions de quatre désoxyribonucléosides

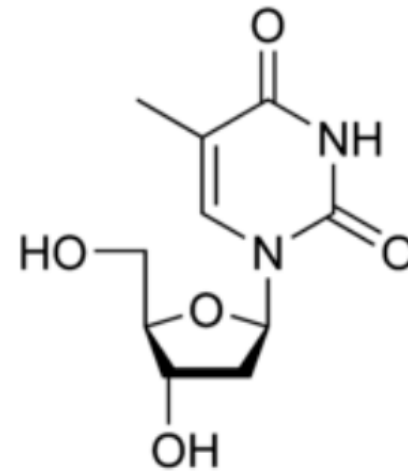
Désoxyribonucléoside



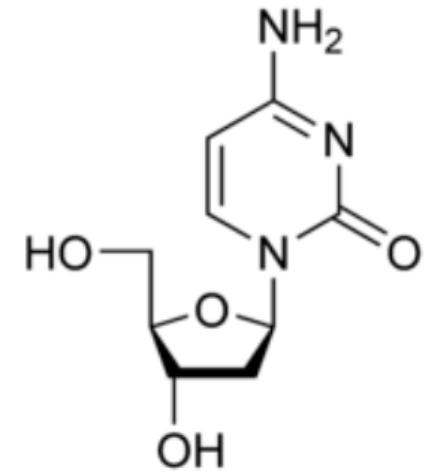
Désoxyadénosine



Désoxyguanosine



Thymidine



Désoxycytidine

A

G

T

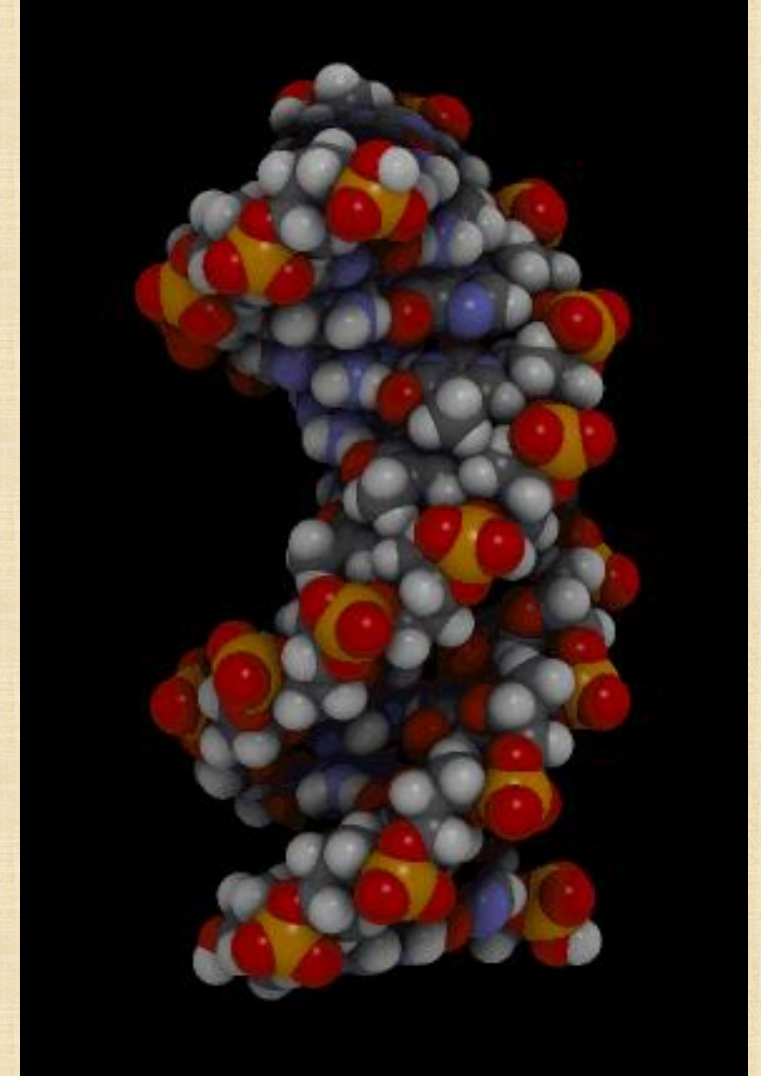
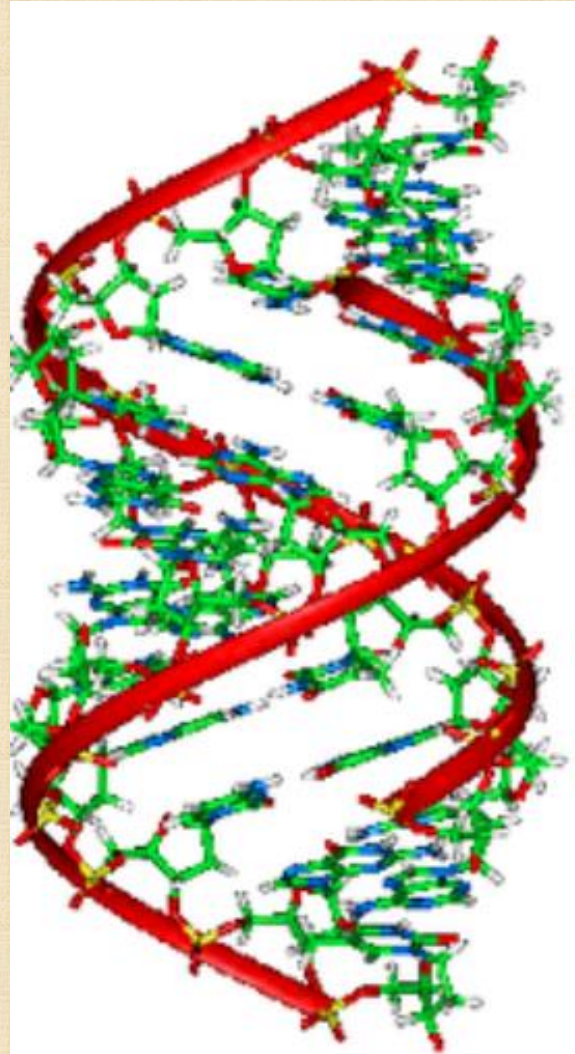
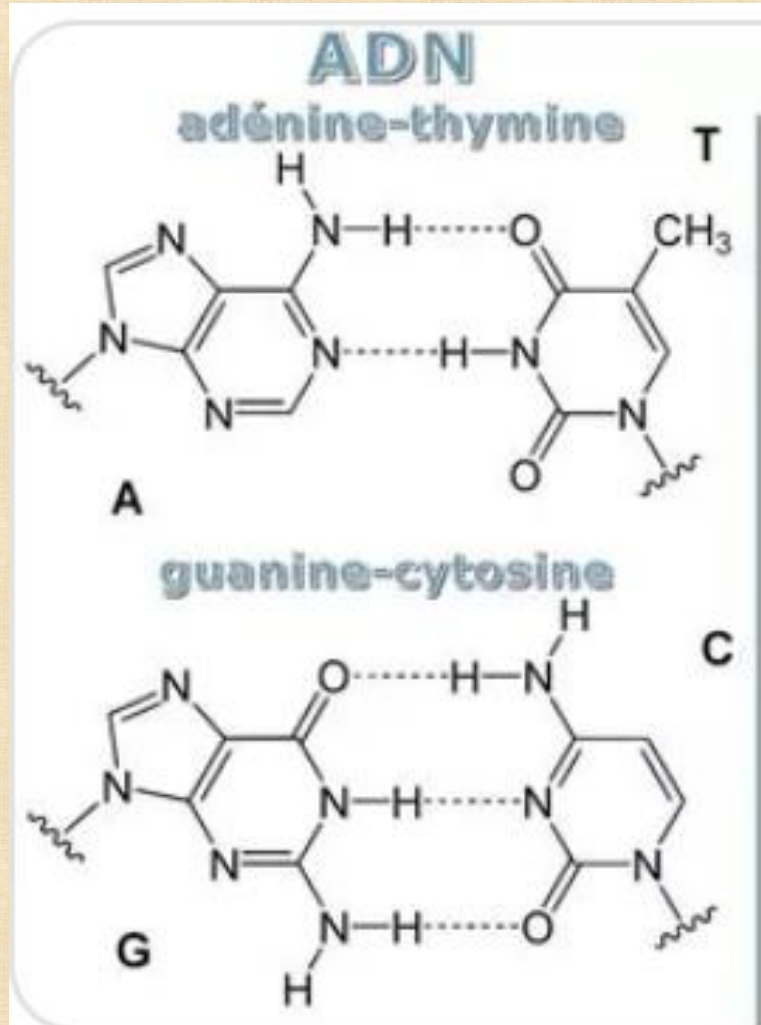
C

Les désoxyribonucléosides s'enchaînent dans un ordre précis pour faire un « **brin** » d'acide nucléique:

.....CTGGTCACTGAAtAGGCTATAACTGGGTCTGA

L'ADN des chromosomes est fait de deux brins **complémentaires**

Structure de l'ADN: double hélice faite de brins complémentaires



ET QU'EN EST-IL DES PROTÉINES ?

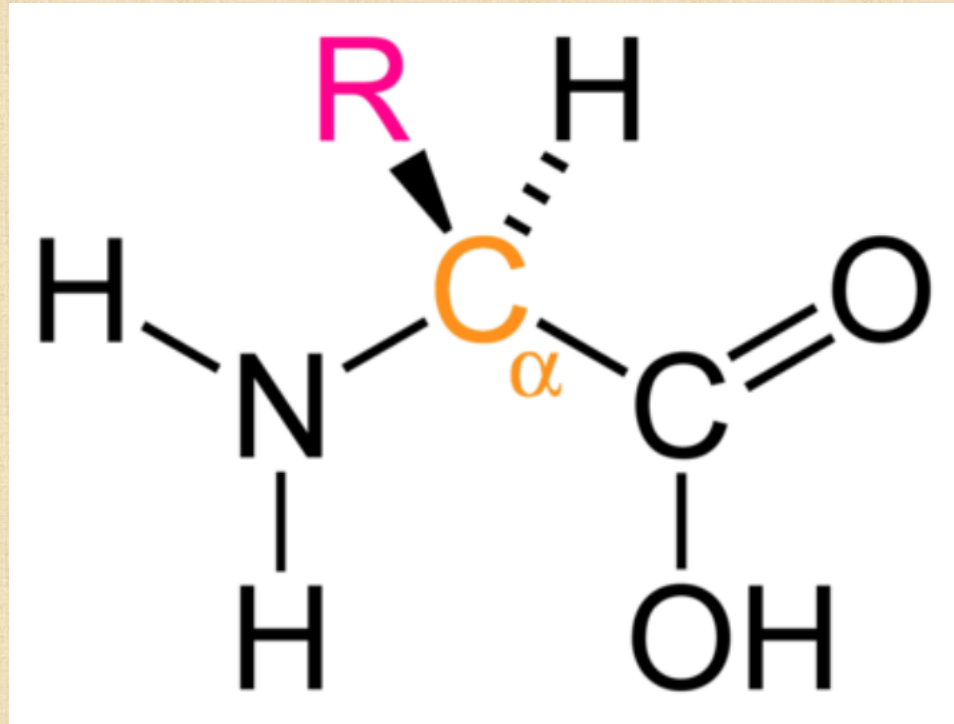
Il existe plusieurs milliers de protéines différentes dans un organisme vivant

Protéine	Fonction
Actine, ...	Squelette des cellules
Myosine, actomyosine, ...	Mouvements des cellules
Enzymes	Catalyseurs des réactions biochimiques <i>(métabolisme, catabolisme, énergie, synthèse des macromolécules)</i>
Hémoglobine, transferrine	Transporteurs de molécules <i>(oxygène, fer, ...)</i>
Anticorps	Défenses immunitaires
Hormones, cytokines	Vecteurs d'informations pour communications intercellulaires
Récepteurs	Capteurs d'informations
Adhésines	Fixation des cellules les unes aux autres ou sur matrice extracellulaire

DE QUOI SONT FAITES LES PROTÉINES ?

d'acides aminés

ils sont les constituants élémentaires des protéines



Les 21 acides aminés naturels

Polaires chargés

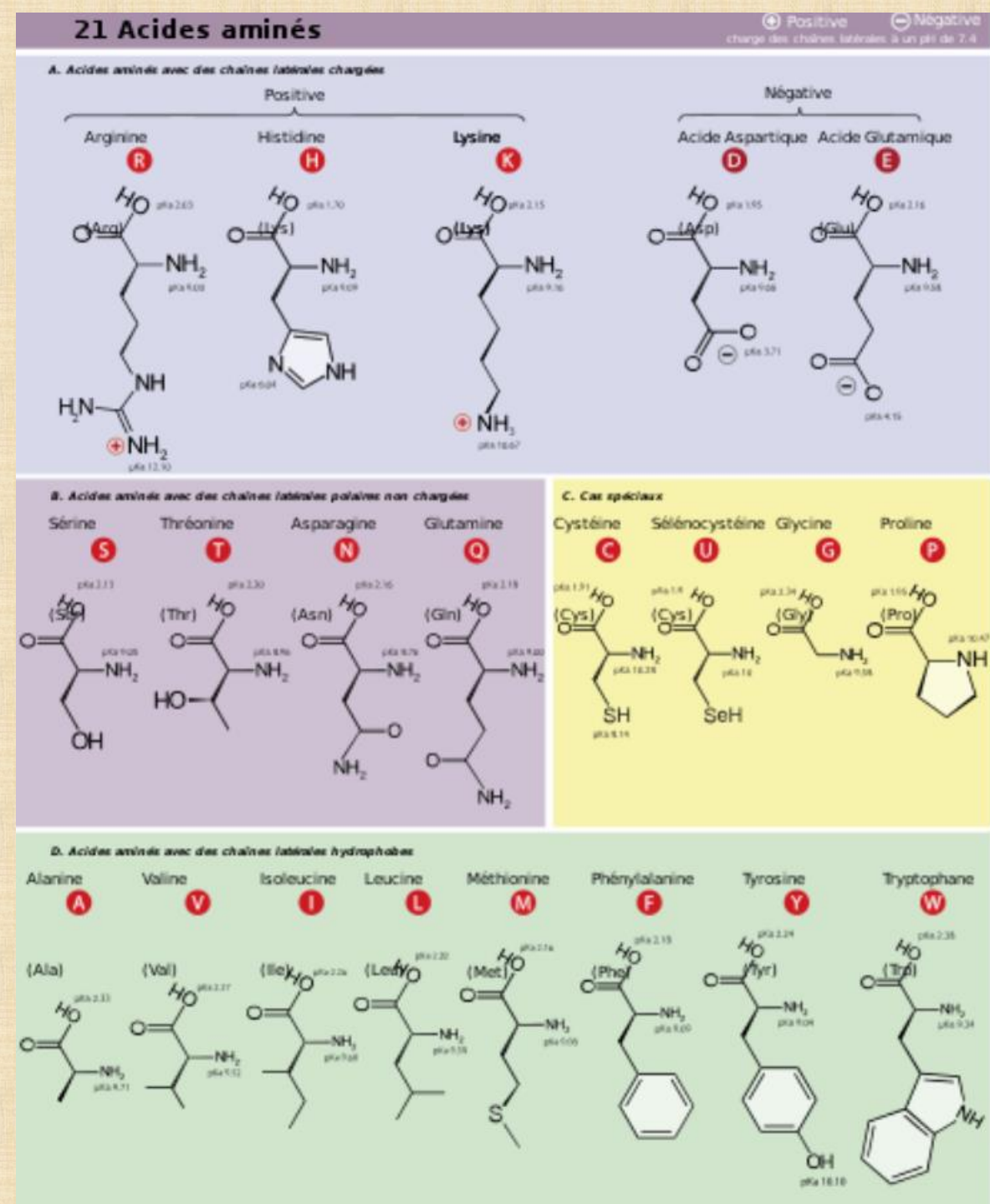
+ ou -

Polaires neutres

cas particuliers

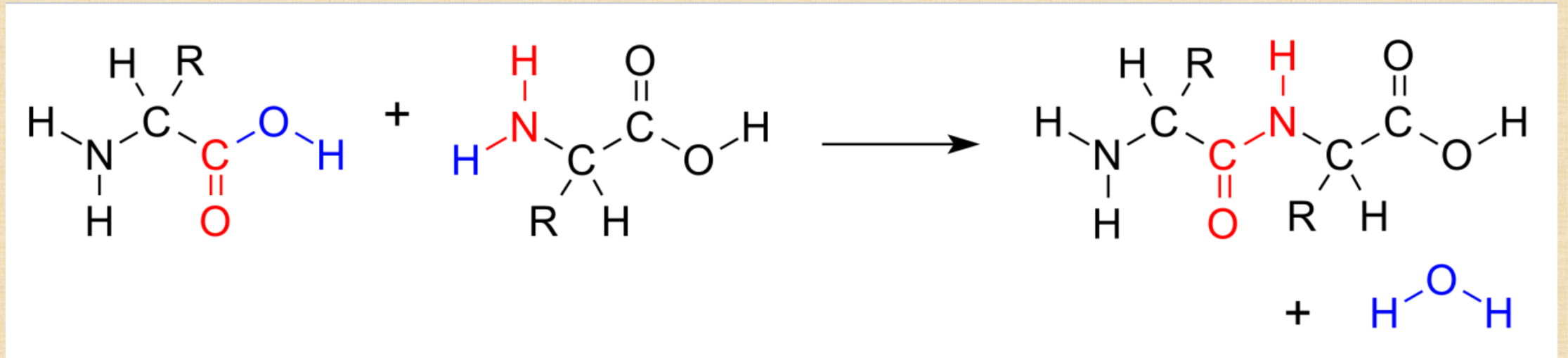
Hydrophobes

linéaires - aromatiques



Liaison peptidique

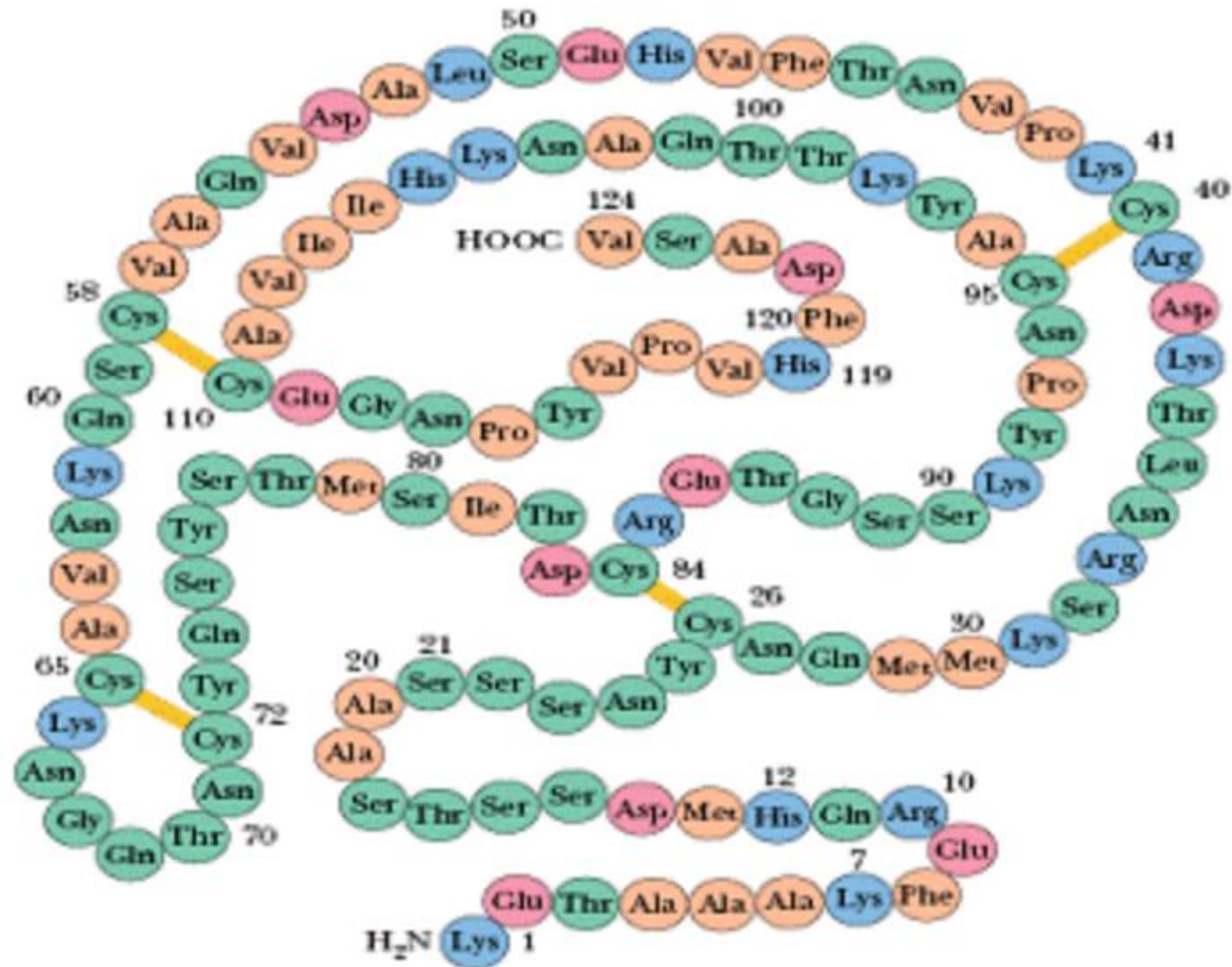
Élimination d'une molécule d'eau entre deux acides aminés



Une protéine est un enchaînement d'acides aminés liés par des liaisons peptidiques

*ici, la ribonucléase du pancréas de bœuf
124 acides aminés – 4 ponts disulfures*

124 acides aminés – 4 ponts disulfures

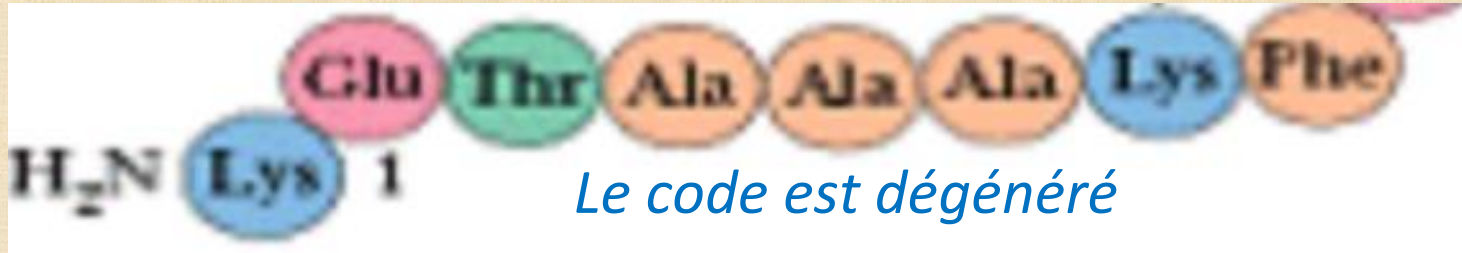


Chaque aminoacide est codé dans l'ADN par une séquence de trois lettres

Début de la séquence de la ribonucléase



ATG.AAA.GAG.ACC.GCG.GCC.GCG.AAA.TTC



Methionine

ATG code pour le début de la lecture; l'acide aminé du début est souvent coupé après synthèse

Le code génétique

LE CODE GENETIQUE							
		ARN messenger Codon : deuxième base azotée					
		U	C	A	G		
ARN messenger Codon : première base azotée	U	Phe	Ser	Tyr	Cys	U	ARN messenger Codon : troisième base azotée
		Phe	Ser	Tyr	Cys	C	
		Leu	Ser	STOP	STOP	A	
		Leu	Ser	STOP	Trp	G	
	C	Leu	Pro	His	Arg	U	
		Leu	Pro	His	Arg	C	
		Leu	Pro	Gln	Arg	A	
		Leu	Pro	Gln	Arg	G	
	A	Ile	Thr	Asn	Ser	U	
		Ile	Thr	Asn	Ser	C	
		Ile	Thr	Lys	Arg	A	
		START	Thr	Lys	Arg	G	
	G	Val	Ala	Asp	Gly	U	
		Val	Ala	Asp	Gly	C	
		Val	Ala	Glu	Gly	A	
		Val	Ala	Glu	Gly	G	

Les deux étapes de la synthèse des protéines

ADN double brin

transcription



ARN messenger

traduction



Protéine

TAC. TTT . CTC. TGG.CGC.CGG.CGC. TTT. AAG
ATG. AAA.GAG.ACC.GCG.GCC.GCG.AAA.TTC

AUG. AAA.GAG.ACC.GCG.GCC.GCG.AAA.UUC

Met. Lys. Glu. Thr. Ala. Ala. Ala. Lys. Phe

La 1^{ère} traduction du message génétique

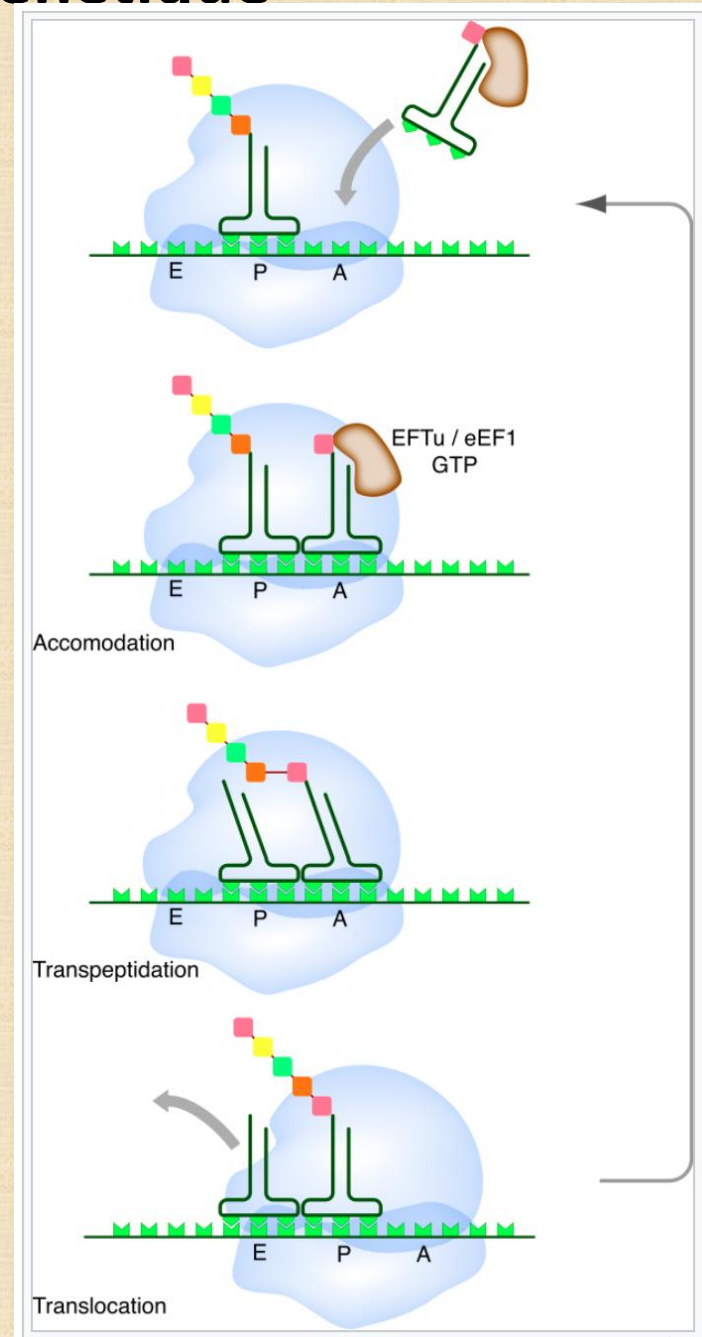
Les partenaires:

TRANSCRIPTION

- l'ADN *(le message archivé)*
- l'ARN polymérase *(copie le message quand il le faut)*

TRADUCTION

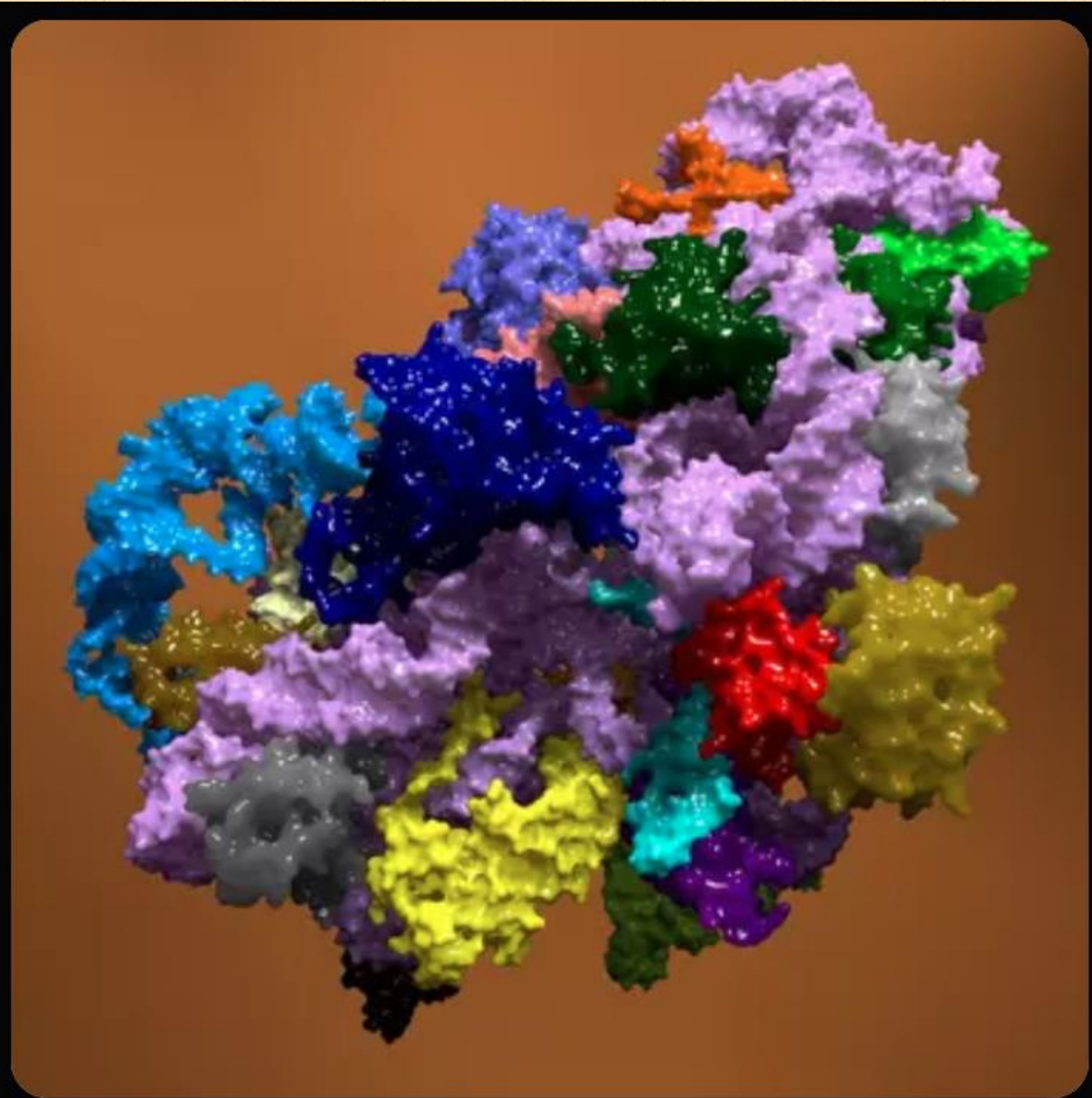
- l'ARN messenger *(fac-similé du message)*
- le ribosome *(la machine)*
- les ARN de transfert *(les détenteurs du code)*



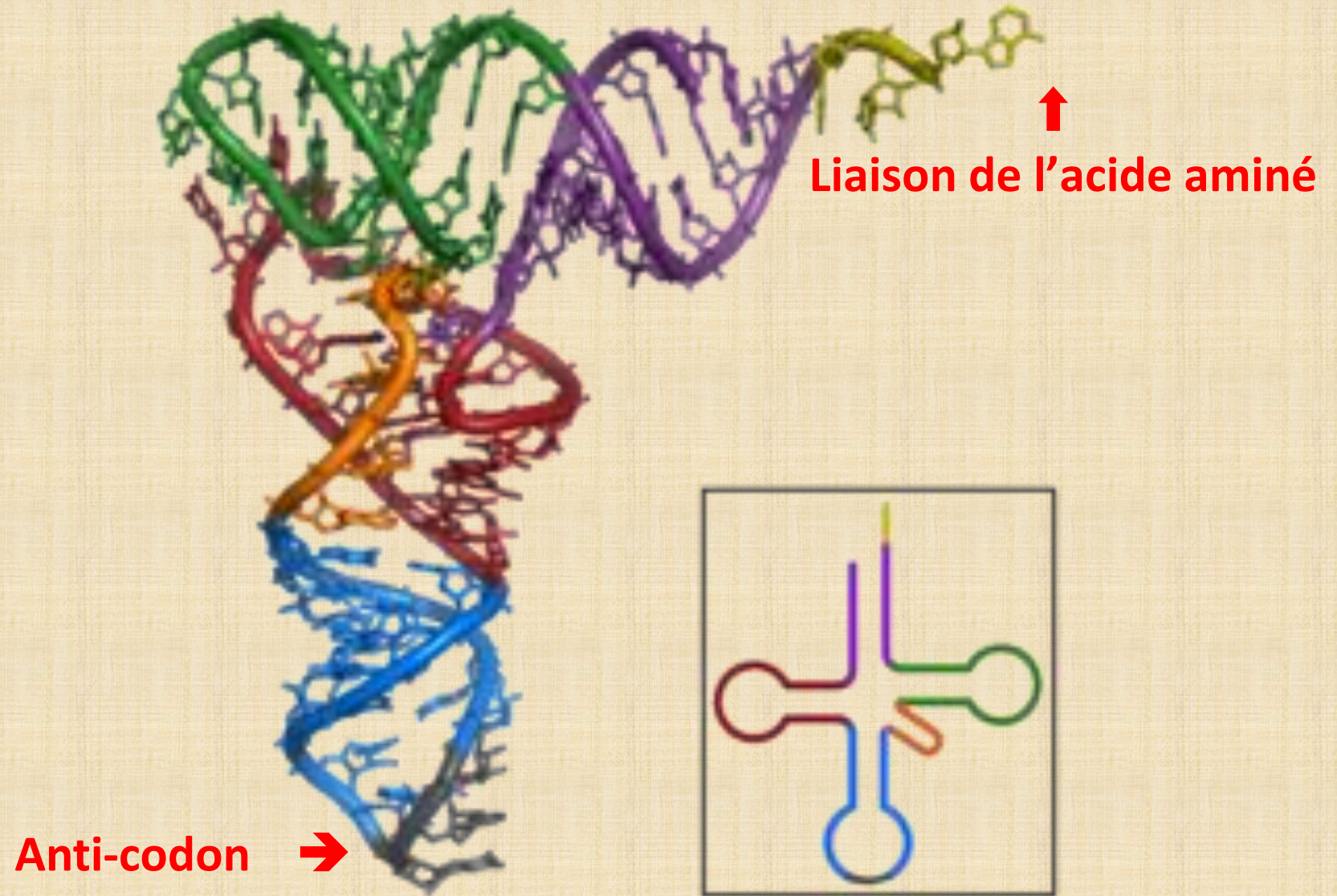
Structure 3D du ribosome

4 ARN ribosomiques
(squelette)

82 protéines



ARN de transfert, ou tARN: Les vrais traducteurs du code génétique



Le ribosome en action

On connaît maintenant de manière approfondie le rôle de chaque composant du ribosome, et le fonctionnement précis de cette machinerie de biosynthèse des protéines.

On a même une image dynamique détaillée du ribosome en train de fabriquer une protéine:

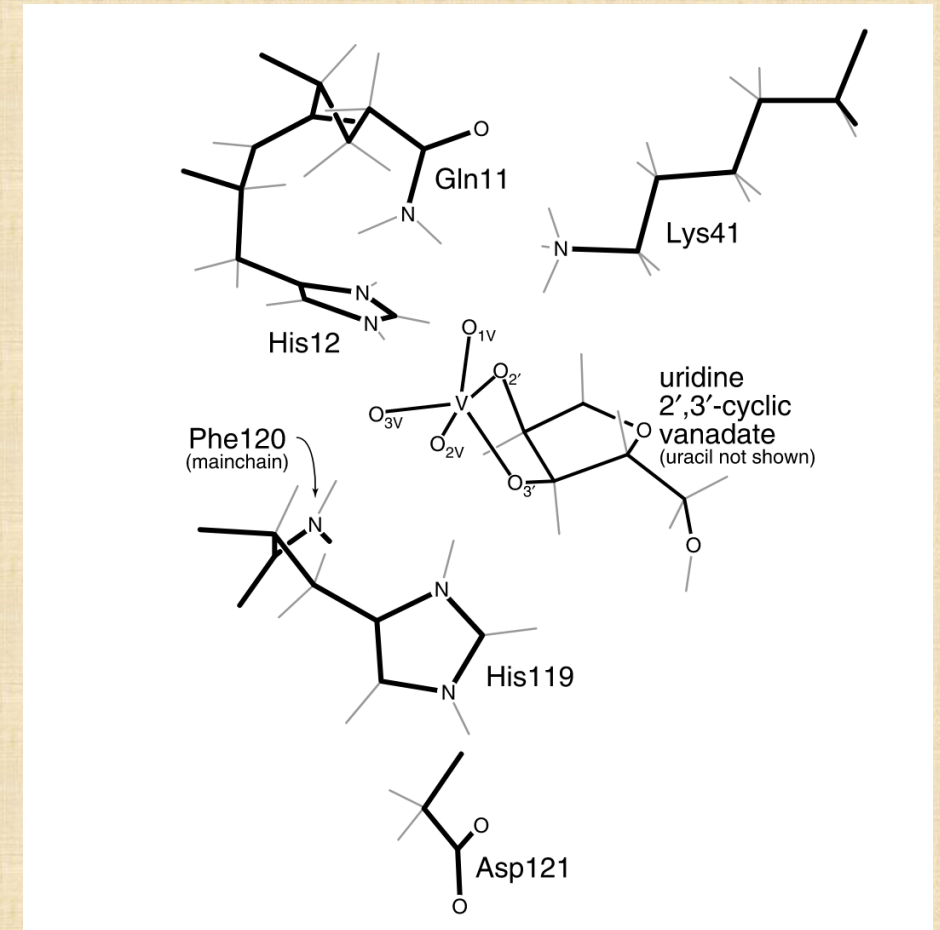
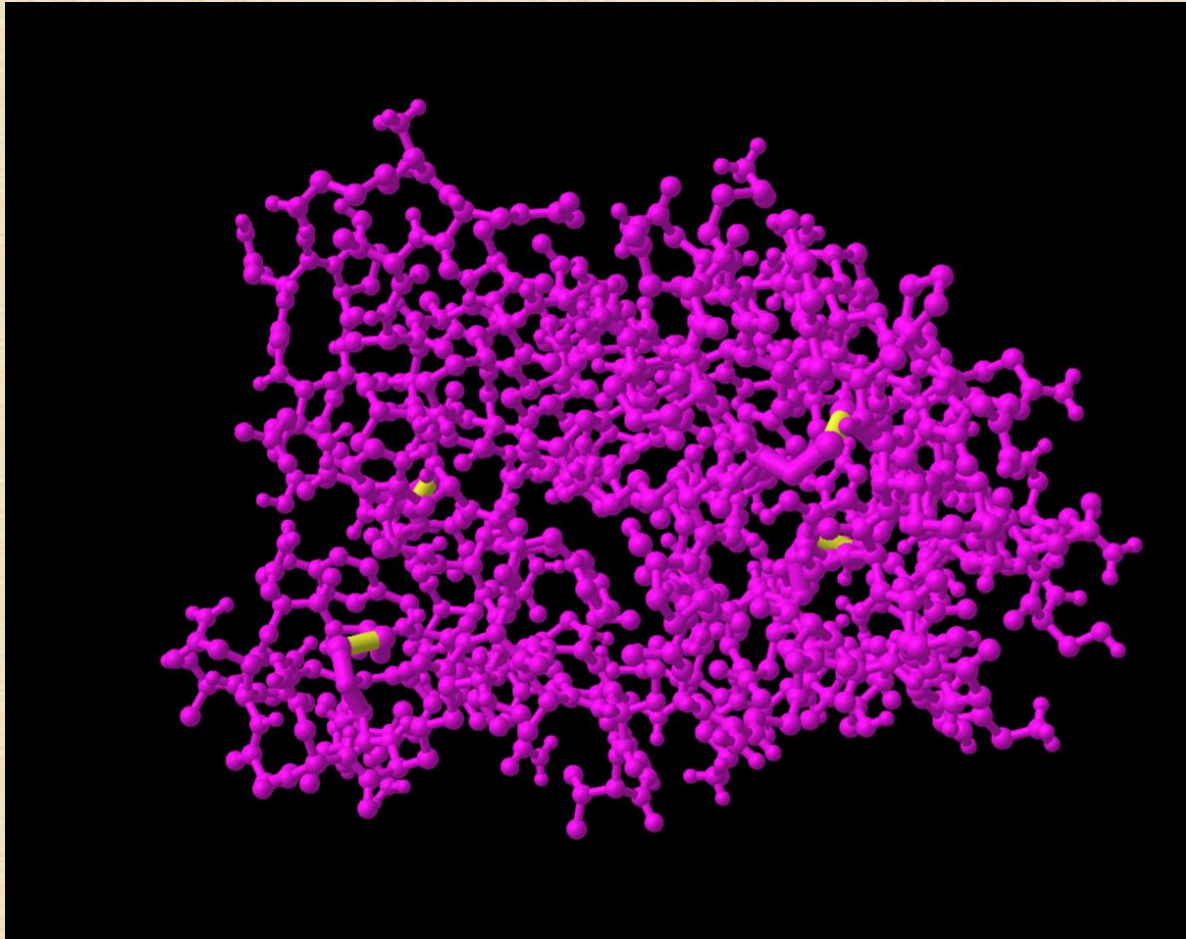
<https://youtu.be/Jml8CFBWcDs>

Mais on n'a résolu qu'une partie du problème:

Comment cette chaîne s'enroule-t-elle sur elle même pour donner à la protéine sa structure tridimensionnelle fonctionnelle ?

*Structure dite « **native** », précise à la fraction d'Angström près, requise pour l'activité biologique*

Le repliement de la Ribonucléase A



Le repliement est indispensable pour rapprocher et positionner dans l'espace les chaînes latérales du site actif:

Comment décrire une structure si complexe ?

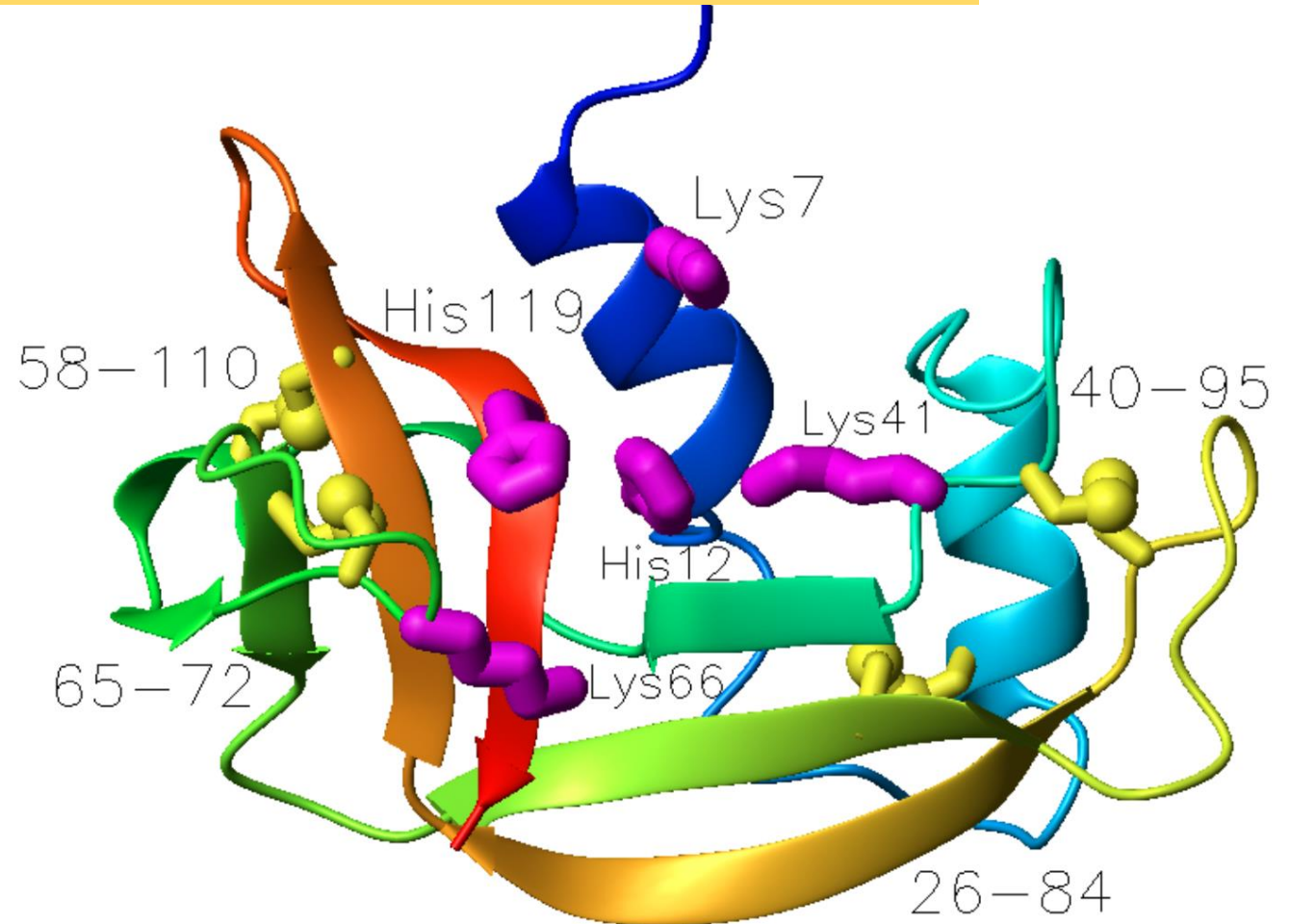
Structures régulières

α -hélices; brins et coudes β

Structure secondaire

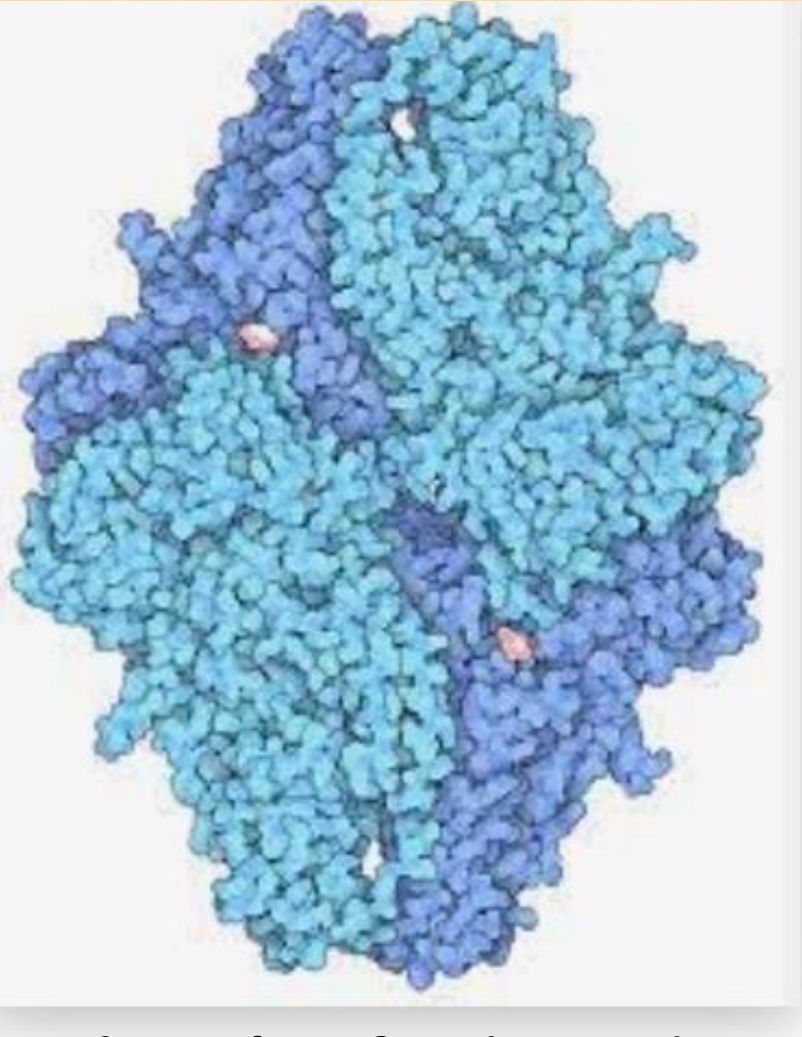
Agencement

Structure tertiaire

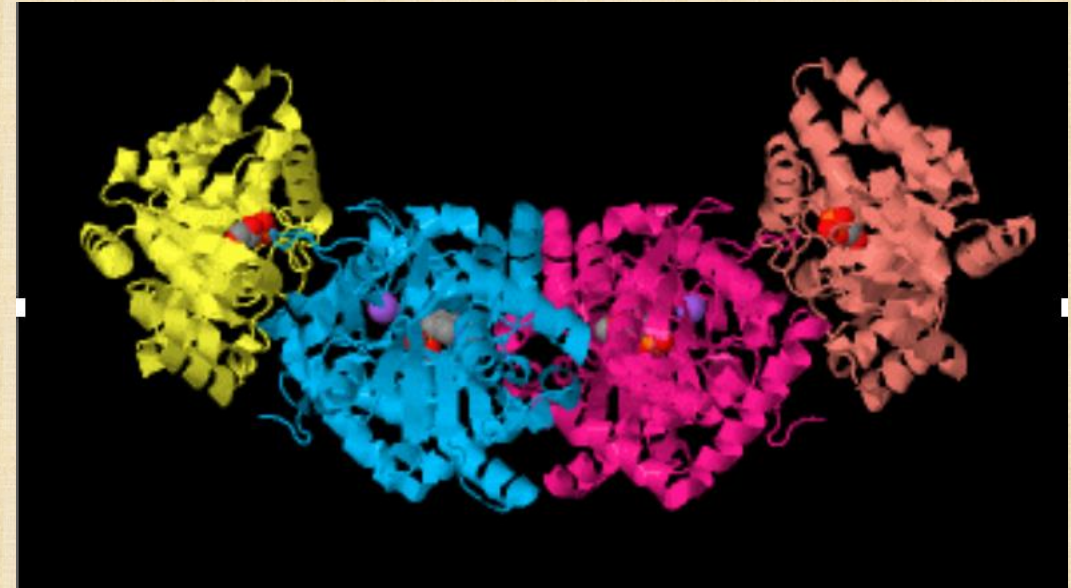


En mauve, les acides aminés du site actif

Assemblage en oligomères = **structure quaternaire**



Homo-oligomère: β -galactosidase
4 chaînes identiques



Hétéro-oligomère: tryptophane synthase
2 chaînes α + 2 chaînes β

Décembre 1966

Question de M. Monod pour ma « petite thèse »:

« Peut-on, à partir de la séquence d'acides aminés d'une protéine, prédire sa structure tridimensionnelle ? »

Et d'abord, la séquence d'acides aminés contient-elle toute l'information requise pour le repliement ?

Expériences de Christian Anfinsen sur la « **renaturation** » de la Ribonucléase (1964) - Prix Nobel en 1972

RNAase pure (*aucune autre information que la protéine elle-même*)

Incubée en Urée 8M : perte totale de structures secondaire et tertiaire: (*aucune autre information que la séquence des aminoacides*)

Diluée fortement (urée ε M): **Toute l'activité enzymatique initiale, donc la structure native, est retrouvée.**

LA SÉQUENCE CONTIENT BIEN TOUTE L'INFORMATION

Comment la **séquence** des acides aminés, une information unidimensionnelle, est-elle traduite en **structure native**, une information tridimensionnelle ?

C'est le problème de la seconde traduction du message génétique

CARACTÉRISTIQUES GÉNÉRALES OBSERVÉES EN 1965-1980

Thermodynamique

Études de l'équilibre

Dénaturation $\longleftrightarrow$ Renaturation

Les courbes de changement de:

- Fluorescence (*exposition des chaînes latérales aromatiques enfouies*)
- Dichroïsme circulaire (*disparition des structures secondaires*)
- Fixation d'un ligand (*perte de conformation native*)
- Diffusion de la lumière (*perte de compacité*)
-

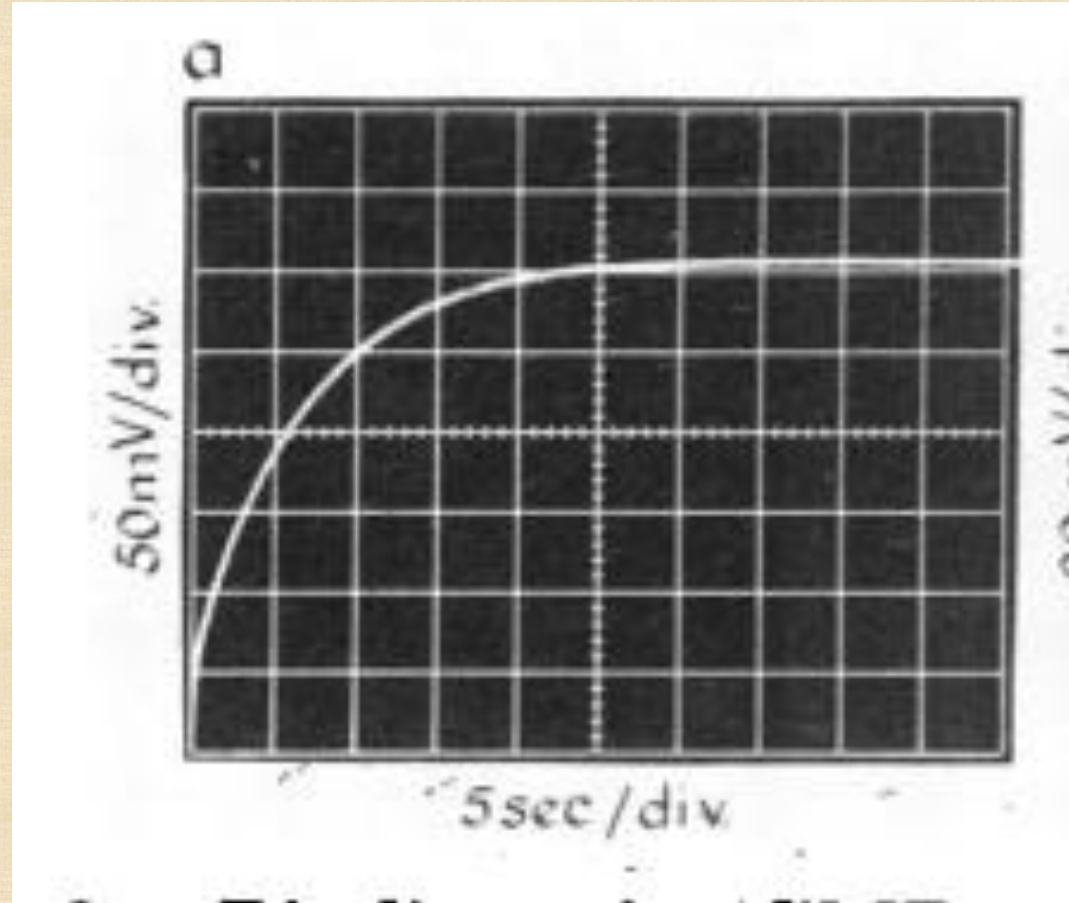
sont superposables.

Conclusion: *Équilibre entre deux états, sans intermédiaires*

→ Postulat: *État Natif = État fondamental*

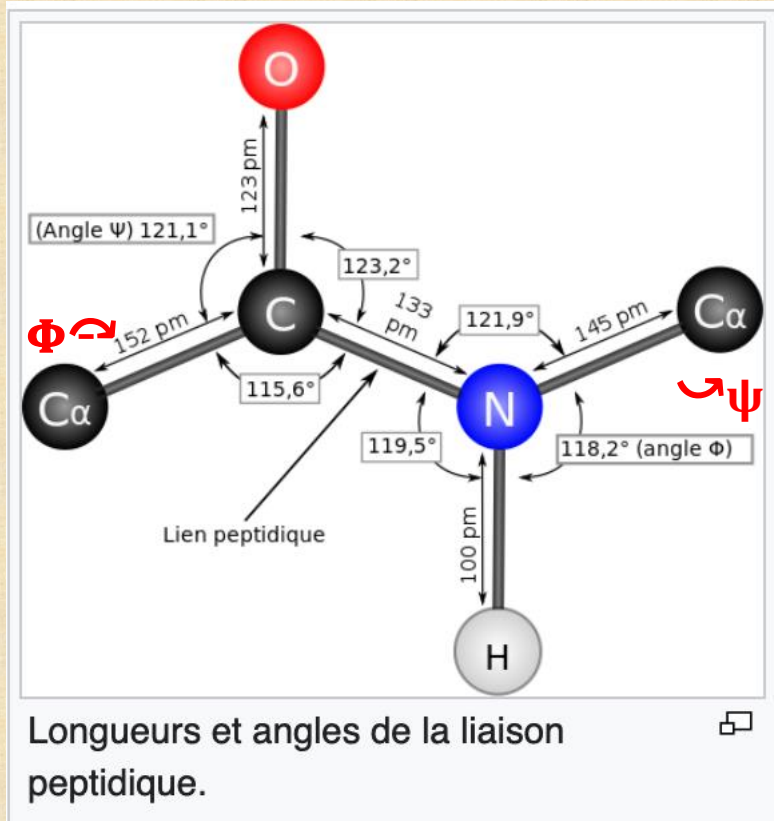
Cinétique:

Exemple: Renaturation de la RNAase
(avec ponts SS intacts – ralentie à 4°C)



Repliement très rapide: de quelques msecondes à quelques dizaines de minutes

Le paradoxe de Levinthal (1967)



Les 6 atomes du motif peptidique sont dans un même plan.

Φ_{i+1} et Ψ_i positionnent le résidu $i+1$ par rapport au résidu i

Pour chaque liaison, 3 Φ et 3 Ψ stables

Pour une chaîne de 300 aminoacides,
 $3^{2 \times 299}$ conformations théoriques

En estimant à quelques picosecondes le temps de test d'une conformation, la recherche aléatoire de l'état natif prendrait plus de temps que l'âge de l'univers.

Comment aborder le problème du repliement ?

Approche expérimentale consistant à observer le repliement pour le comprendre

Approche énergétique

Déterminer les champs des forces qui stabilisent la structure et rechercher le minimum énergétique (état fondamental).

Exploitation des bases de données structurales (1D et 3D)

Trouver des relations entre séquences locales et structures secondaires puis tertiaires

L'approche expérimentale, du plus gros au plus petit ...

Mes premiers pas au laboratoire (1962-1963) avec la β -galactosidase microbienne, *l'enzyme du Prix Nobel de Monod-Jacob-Lwoff (1965)*

Enzyme qui hydrolyse le lactose en galactose + glucose

4 chaînes de **1.021** acides aminés.

Masse moléculaire: **464.000** daltons pour le tétramère.

Repliement en une étape invraisemblable

Hypothèse: Les monomères se replient avant leur association

- **Dénaturation en urée 8M** → *dissociation en monomères et déstructuration complète des monomères*
- **Renaturation soit en une étape (8M → 0M urée) soit en deux étapes (8M → 2M → 0M)**
- **La comparaison des cinétiques de repliement montre que**
 - le repliement du monomère est cinétiquement limitant
 - l'association suit le repliement
 - le monomère replié est *inactif*.

Pourquoi le monomère replié est-il inactif ? A-t-il ou pas sa conformation native ?

Réponse apportée par études sur la **Tryptophanase**, une enzyme bactérienne qui dégrade le tryptophane lorsqu'il est en excès

- La tryptophanase a besoin d'un co-facteur, le Pyridoxal-phosphate (vitamine B6) pour être actif .
- L'apo-enzyme (sans B6) est en équilibre dimère-tétramère alors que l'holo-enzyme (B6) est un tétramère.
- Dès qu'on ajoute B6, association en tétramère très rapide.

La mise au point d'un test hypersensible d'activité nous a permis d'étudier la cinétique de tétramérisation à très faible concentration de protéine et de montrer que:

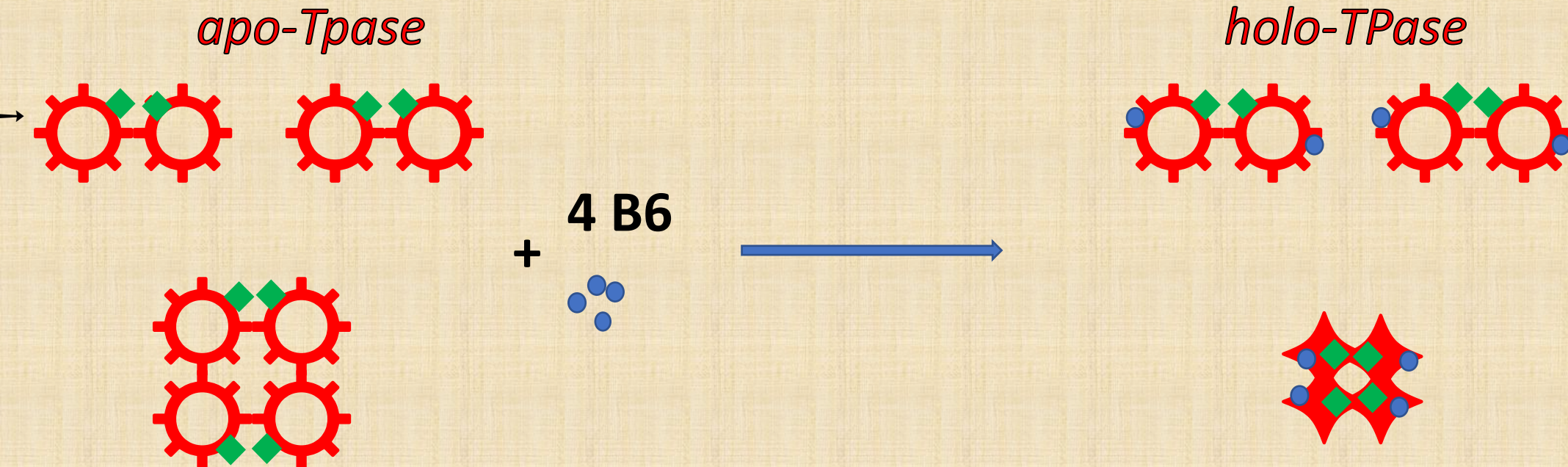
comme pour la β -galactosidase, la forme dissociée (dimère + B6) est inactive.

Est-elle « native » ?

Le groupement –SH d’une cystéine de chaque monomère réagit très vite avec N-éthyl-maléimide (NEM), donc est **exposé à la surface** de la protéine dans l’apo-dimère (sans B6), l’holo-dimère (avec B6) et l’apo-tétramère (sans B6).

Donc **ni au site de liaison** de B6 **ni à l’interface** entre dimères au sein du tétramère

Il réagit très lentement dans l’holo-tétramère → la tétramérisation provoque un **changement de structure qui masque le groupement –SH.**



Conclusions sur la formation des structures quaternaires

Les monomères **se replient avant de s'associer**.

Ce repliement génère une structure **proche de l'état natif**, possédant déjà les **aires spécifiques d'association** entre sous-unités: les monomères se reconnaissent entre eux même au sein d'un mélange de protéines

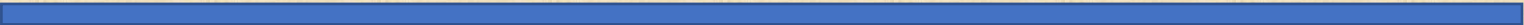
Les sous unités isolées ne sont en général **pas fonctionnelles**.

Leur structure finale fonctionnelle n'est acquise qu'après un **changement conformationnel imposé par les interactions entre sous-unités**.

Comment le monomère fait-il pour se replier ?

Les grandes chaînes peptidiques (>100-150 acides aminés) se replient-elles **d'un coup** ? Ou bien autour de **« centres de nucléation » indépendants**

Modèle né des études génétiques de François Jacob sur la β -galactosidase: la complémentation « intracistronique »

Mutant		
« sauvage »		Actif
accepteur		Inactif
donneur		Inactif
«complémenté»		Actif

Etudes (de 1968) sur la β -galactosidase complémentée

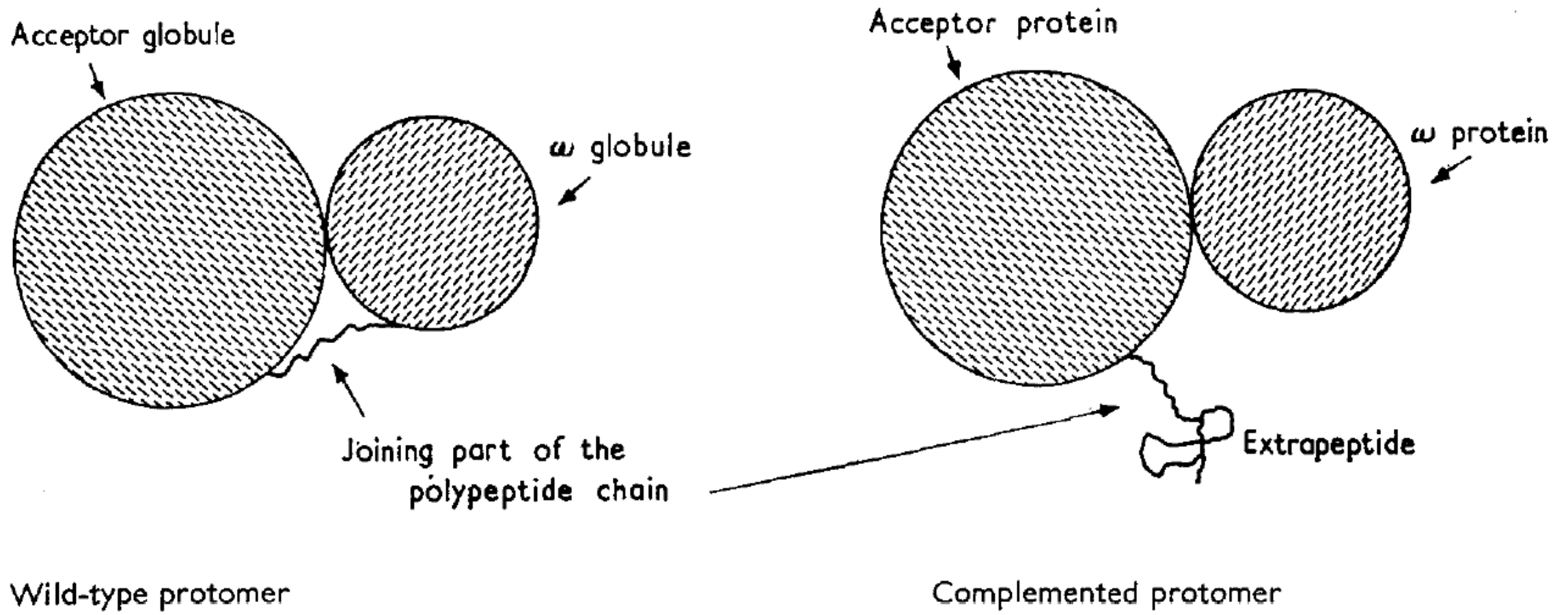
L'activité enzymatique de l'enzyme complémentée est identique à celle de l'enzyme « sauvage »: **elles ont des structures similaires**

L'enzyme complémentée est aussi un **tétramère**, mais son monomère est fait des deux fragments. (*ultracentrifugation analutique*)

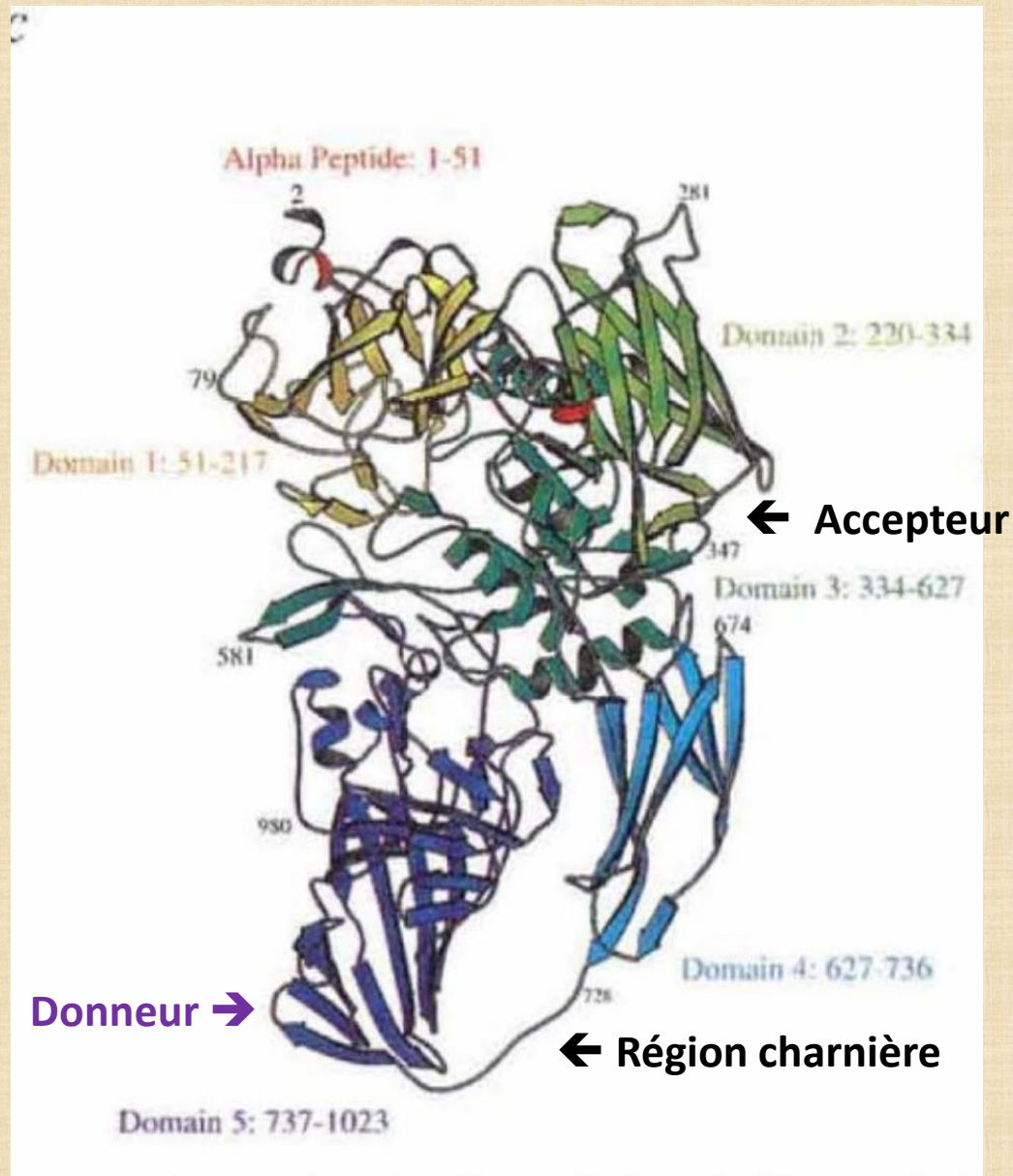
Les fragments isolés **sont globulaires et compacts**. (*propriétés hydrodynamiques et optiques*)

Les fragments isolés se replient spontanément en une structure proche de leur structure native. (*immunoréactivité*)

le modèle des globules



Confirmation 25 ans plus tard par diffraction des rayons X



- Le globule « donneur » apparaît clairement.
- L'accepteur est lui-même fait de plusieurs globules.
- Le terme « **domaine** » a remplacé celui de « globule »

Le rôle de la région charnière est de faciliter l'association entre les domaines.

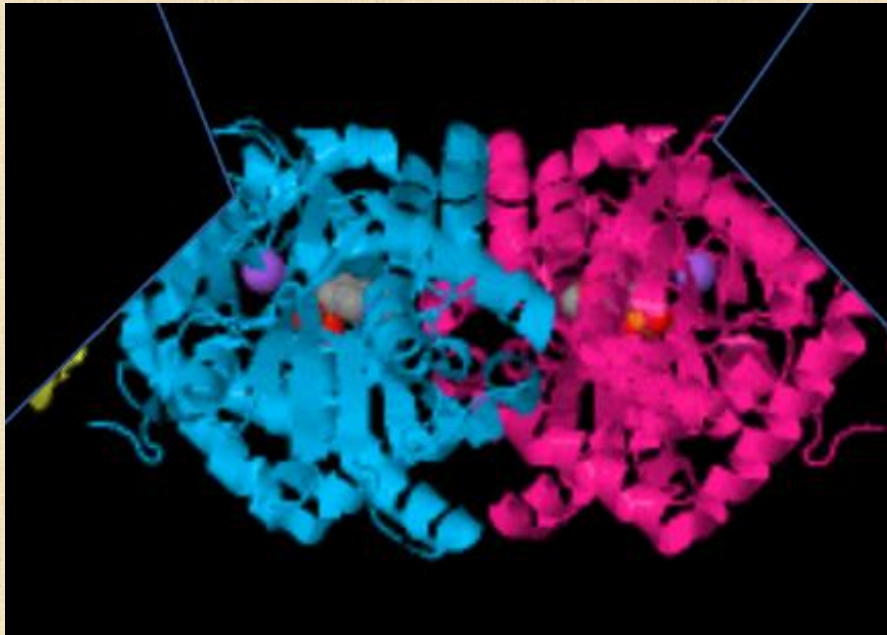
La protéolyse ménagée pour disséquer une protéine en « globules »

Coupure de la charnière, puis dénaturation, séparation des fragments, renaturation et caractérisation des fragments isolés pour:

- β -galactosidase

Mêmes résultats qu'avec l'enzyme complétementée

- sous-unité $\beta 2$ de la tryptophane synthase:



Protéolyse de $\beta 2 \rightarrow \beta 2$ -coupée **native, quasi-active**

Dénaturation libère F1 (29 kD) et F2 (12 kD)

Renaturation séparée $\rightarrow$ F1 et F2 globulaires condensés

Mélange F1 + F2 $\rightarrow$ $\beta 2$ -coupée **reconstituée est identique** à $\beta 2$ -coupée **initiale** (*structure et fonction*)

Les fragments se replient séparément

Caractérisation des fragments repliés isolés

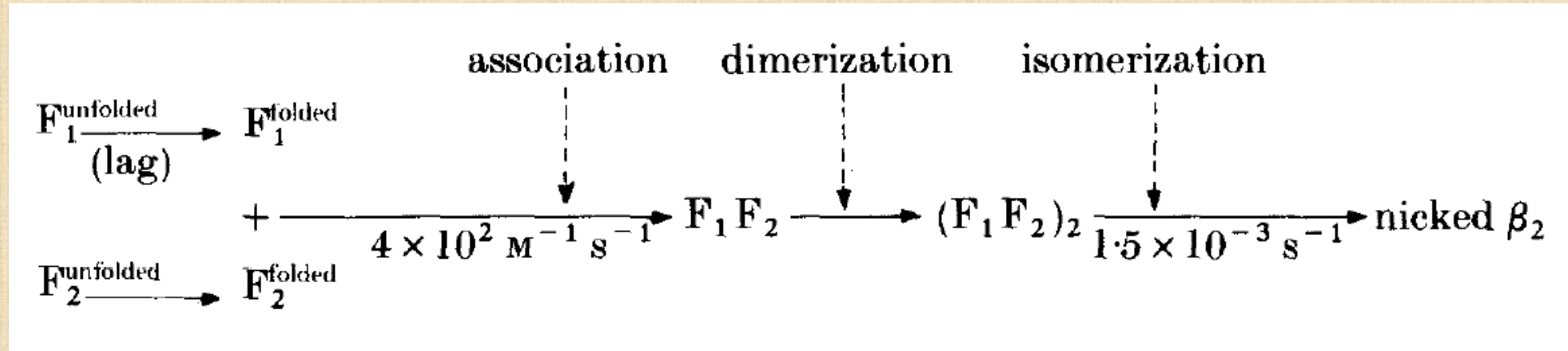
F1 est proche de l'état natif

F2 présente des similitudes avec l'état natif, mais possède plus de structure secondaire qu'à l'état natif

Conclusion: la chaîne β de la tryptophane synthase, comme la β -galactosidase, se replie autour de centres de nucléation indépendants.

Cinétique de renaturation de β 2-coupée

Aboutit au modèle suivant:



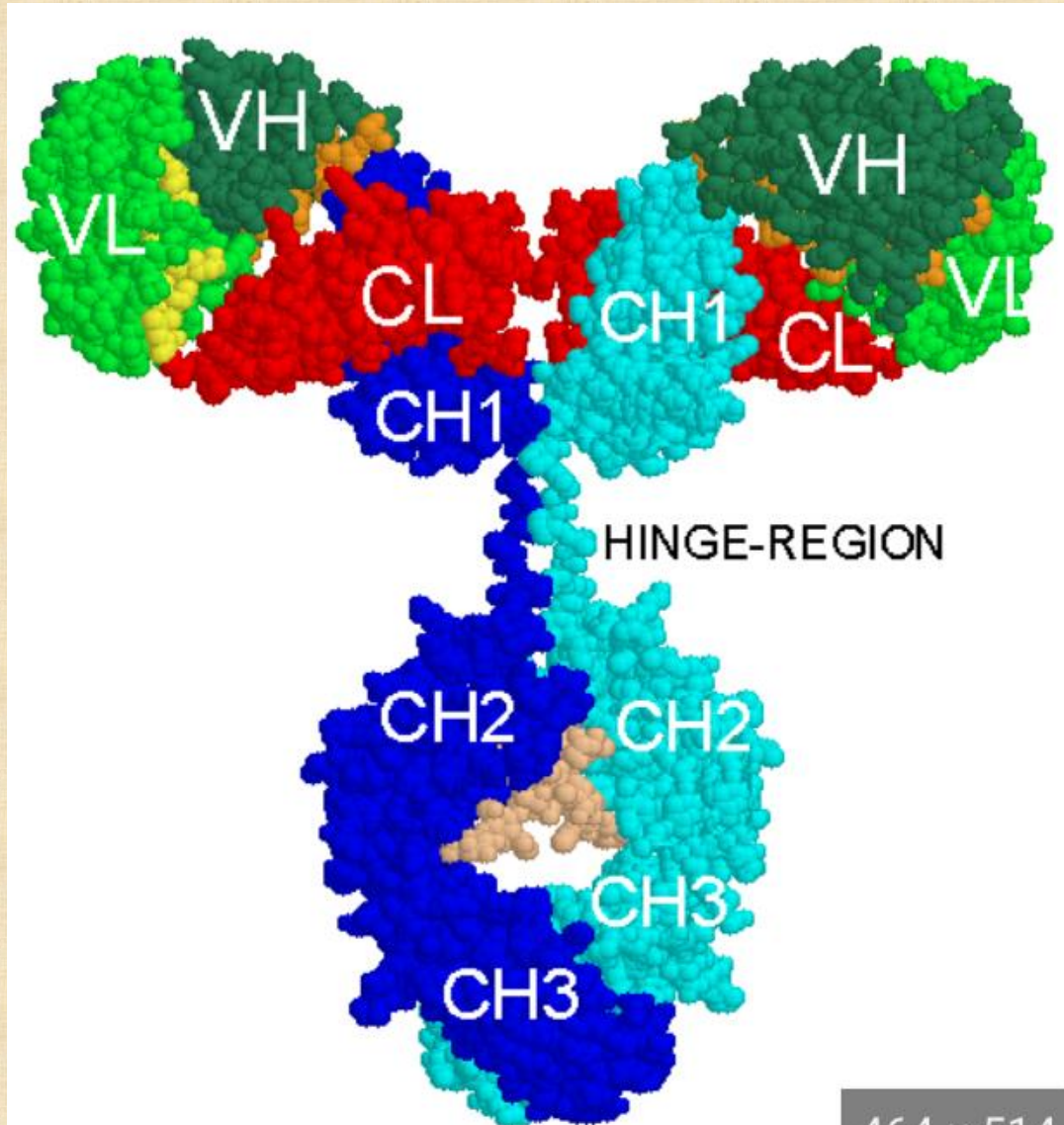
→ F1 replié est un intermédiaire dans le repliement de β

→ le repliement de la chaîne β est modulaire

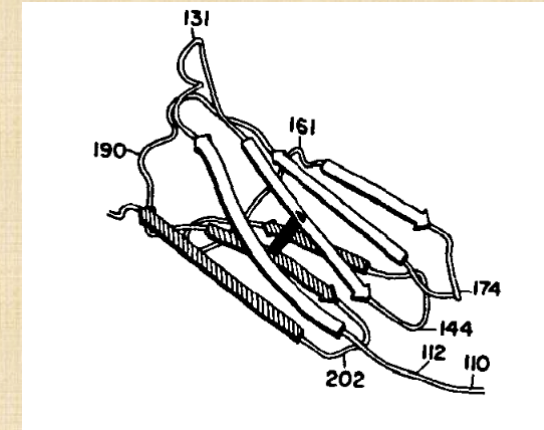
→ L'association module la structure de F2 replié

le modèle des globules est-il généralisable ?

Oui, d'après de nombreuses études cristallographiques: exemple des IgG



464 x 514



Le « domaine immunoglobuline »

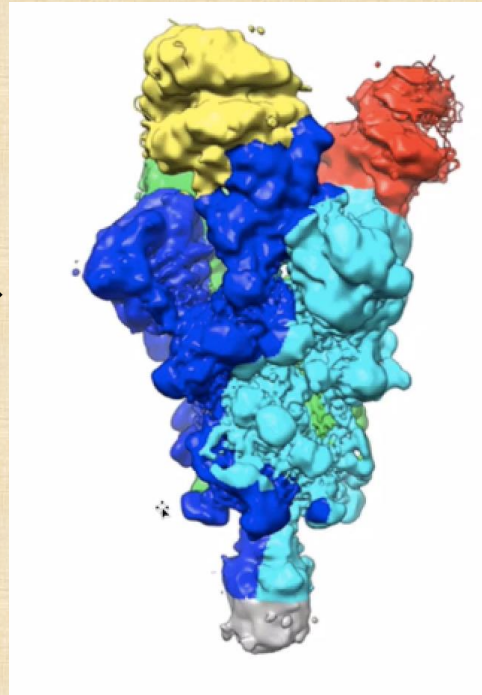
2 chaînes lourdes (H) – 4 domaines chacune
2 chaînes légères (L) - 2 domaines chacune

Soit 12 domaines presque identiques

Depuis l'avènement du génie génétique, de très nombreux domaines ont été produits *in vivo*, purifiés et caractérisés

Leur repliement autonome en structures **pseudo-natives** ainsi que leur **réarrangement structural** lors de leur association sont des caractéristiques très générales.

La protéine « Spike », un trimère →



← RBD = receptor binding domain →



Le vaccin anti-COVID est le domaine de « Spike » qui attache le virus aux cellules cibles.

COMMENT SE REPLIE UN GLOBULE ?

Le modèle du «molten globule»

Mis en évidence par **Oleg Ptitsyn**(*Pustchino – URSS*):

α -lactalbumine à l'équilibre en conditions de dénaturation douce:

Structure globulaire condensée (*diffusion de la lumière*)

Structure secondaire native (*dichroïsme circulaire «natif»*)

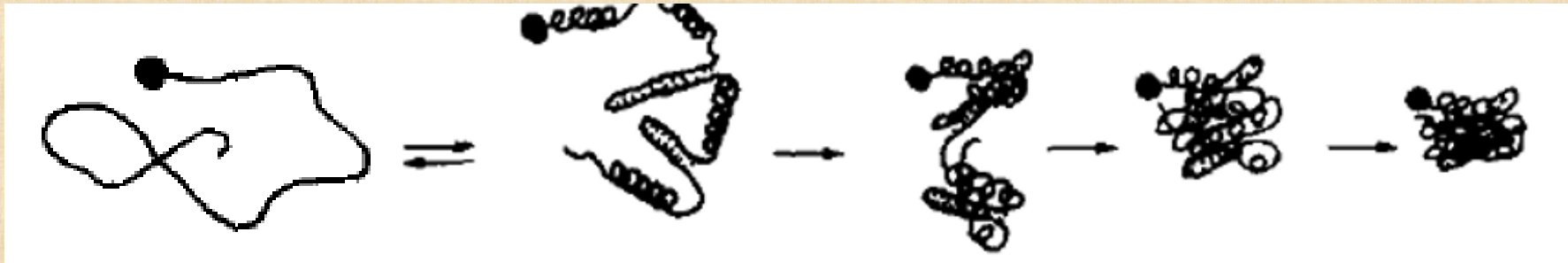
Cœur hydrophobe formé (*accessible aux colorants fluorescents; aromatiques enfouis*)

Structure non compacte (*rayon hydrodynamique un peu plus grand que forme native, chaînes latérales mobiles*)

Hypothèse de Ptitsyn:

Le molten globule est un intermédiaire cinétique pour le repliement d'un « domaine »

Le modèle « consensus » en 1980-1990: un chemin unique de repliement



Structure

primaire

secondaire

tertiaire partielle

tertiaire

tertiaire compacte

État

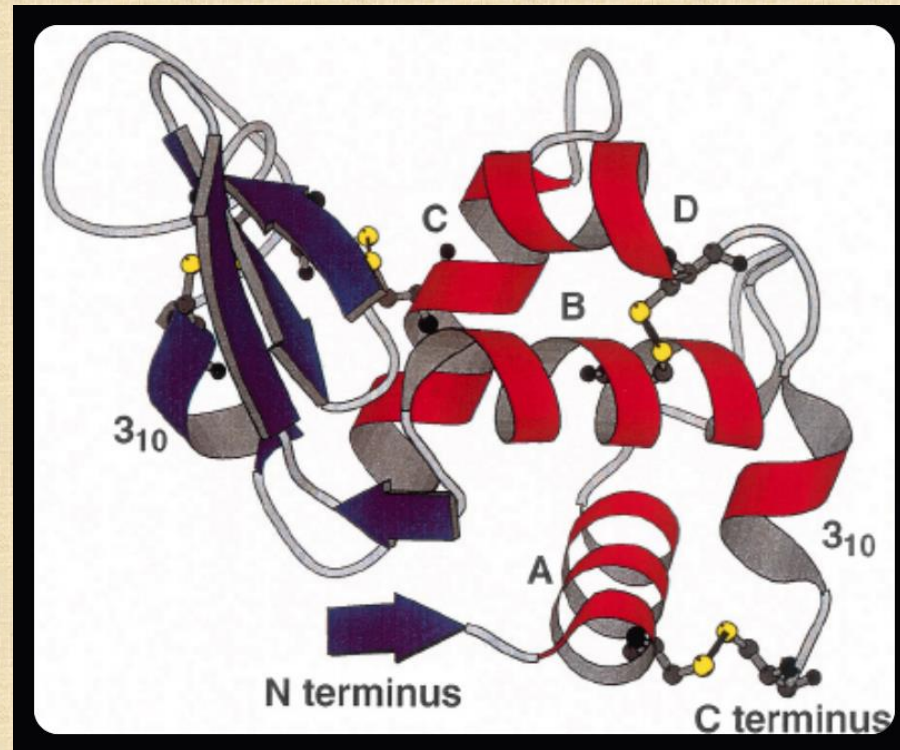
dénaturé

Molten Globule

Natif

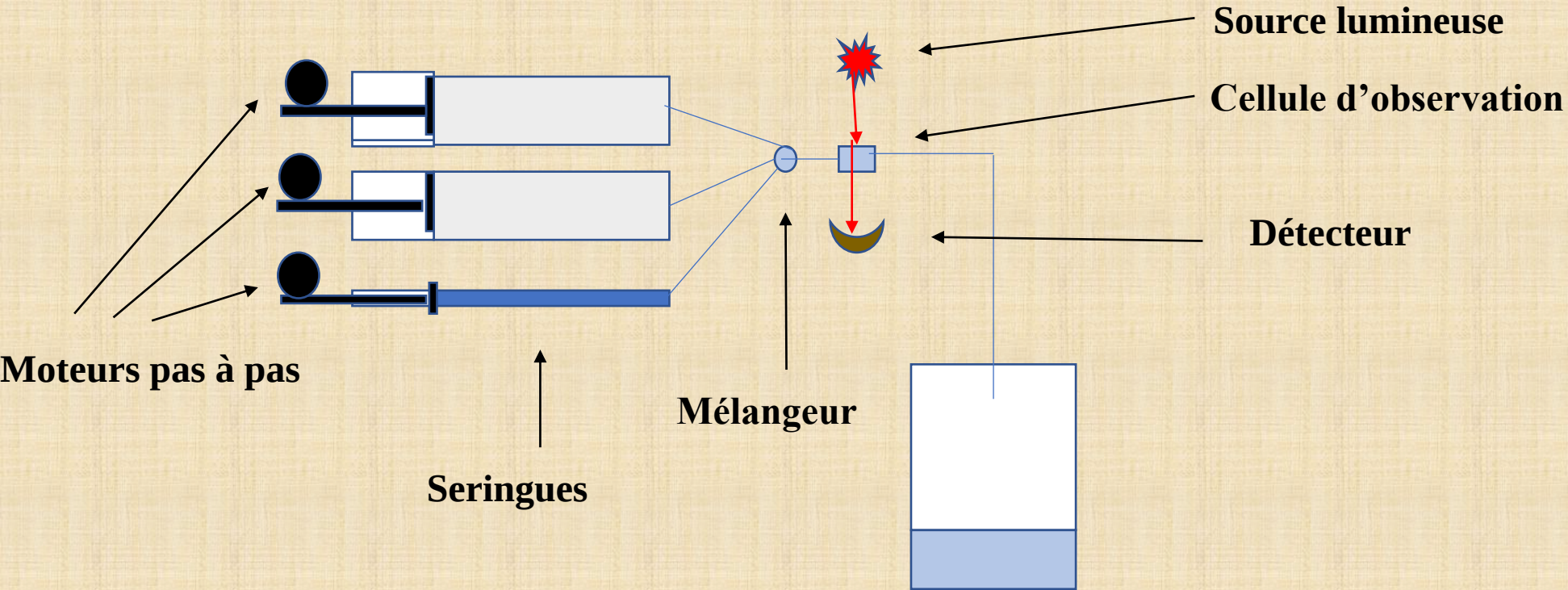
Modèle testé sur la renaturation du lysozyme de poule

HEWL: chaîne de 129 acides aminés – 4 ponts disulfures



HEWL dénaturé en Guanidine 6 M sans rompre les ponts S-S se renature à 100% très vite par dilution de la Guanidine

Le stopped-flow: outil d'étude de cinétiques rapides

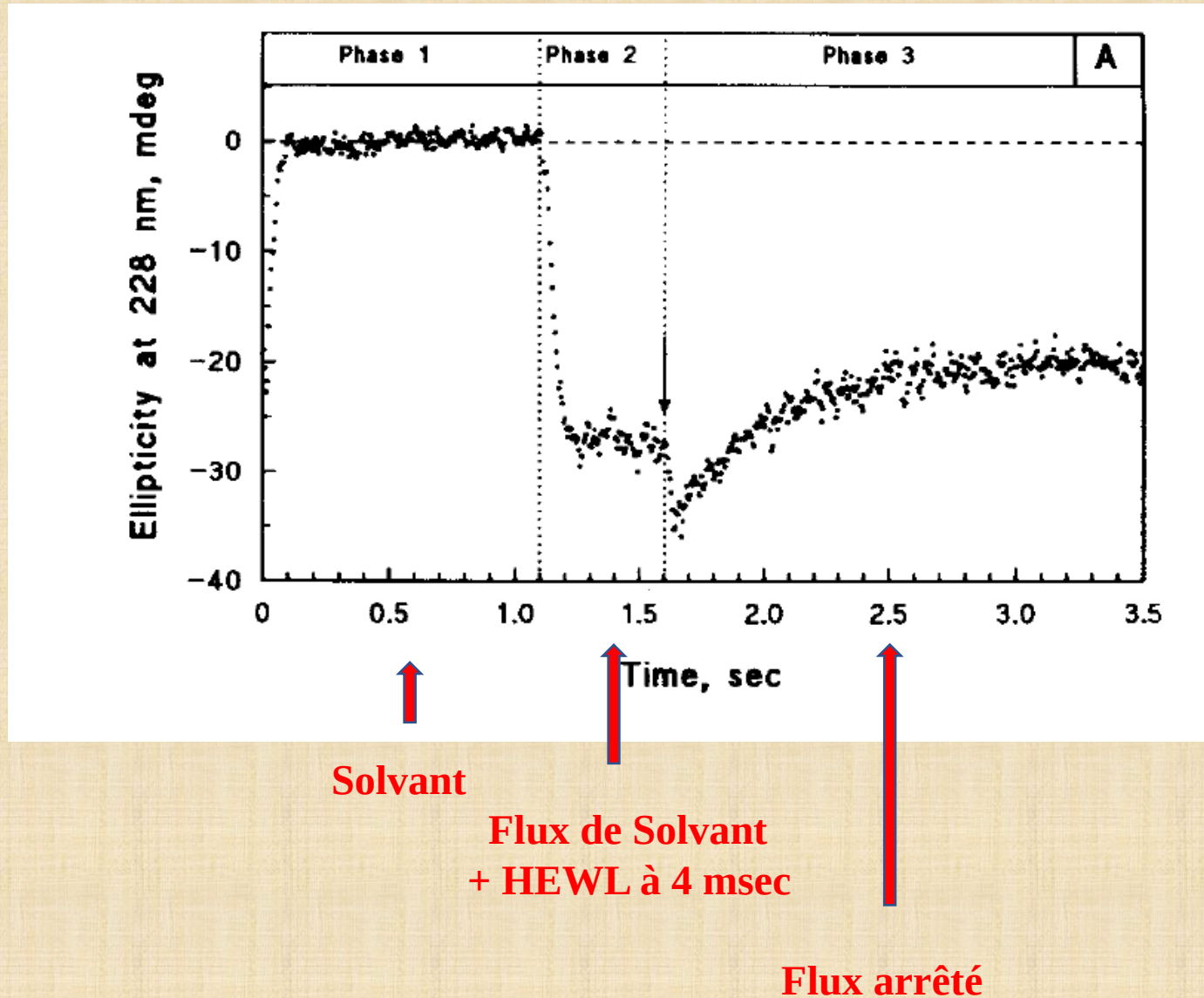


La solution diluée arrive dans la cellule 4 msec après le mélange

Signal observé: absorption, fluorescence, dichroïsme circulaire

Renaturation du lysozyme à 20°C

Cinétique de regain du dichroïsme circulaire à 228 nm



L'état atteint à 4 msec est un molten globule

- Le spectre de dichroïsme circulaire à 4 msec est celui calculé pour le contenu en structure secondaire de HEWL natif: **mêmes structures secondaires**
- Les changements ultérieurs de ce spectre sont dus à des contraintes sur les ponts disulfures
- La fluorescence des acides aminés aromatiques montre qu'ils sont enfouis dans un **environnement hydrophobe**
- Un fluorochrome hydrophobe (ANS) pénètre dans le cœur de l'intermédiaire 4 msec : **le cœur hydrophobe de l'intermédiaire 4 msec n'est pas compact**
- **L'ANS** est expulsé ultérieurement : **la molécule devient compacte lors de l'ajustement définitif de la structure tertiaire.**

Le molten globule est bien un intermédiaire du repliement

Résultats obtenus avec le lysozyme avec 4 ponts S-S

les ponts S-S sont un «souvenir» de l'état natif
Comment se replie le lysozyme réduit ?

Comme pour la ribonucléase, le repliement de HEWL réduit est lent

Ne se fait bien qu'à très faible concentration:

- *formation d'agrégats insolubles, inactifs, irréversibles*
- *plus la concentration est grande, plus la proportion d'agrégats est élevée*

À quel stade du repliement les ponts S-S jouent-ils un rôle ?

Étude de la renaturation oxydative par nombreuses sondes

*Dichroïsme circulaire – Fluorescence intrinsèque – Formation des S-S - Fixation de ligands –
Anticorps monoclonaux spécifiques du domaine α natif ou du domaine β natif*

Ni structure secondaire ni cœur hydrophobe ne sont détectés à 4 msec de repliement

Les ponts S-S accélèrent ou stabilisent le repliement dès le stade molten globule

La structure secondaire native n'est observée qu'à partir de 3 ponts S-S

Le repliement est séquentiel

$$U \rightarrow I_0 \rightarrow I_1 \rightarrow I_2 \rightarrow N.$$

- Le domaine α se replie en premier, lentement, avec formation de ses deux ponts S-S,
- Le domaine β se replie ensuite rapidement avec formation du 3^{ème} pont S-S, ➔ **structure native**
- Enfin le 4^{ème} pont S-S se forme et **stabilise** la structure native.

**Le modèle séquentiel à un seul chemin semble vérifié
mais ...**

***nous avons démontré sur la tryptophanase l'existence de
chemins alternatifs aboutissant à la forme native***

Selon les conditions de dénaturation (pH acide vs Guanidine) les cinétiques de renaturation dans exactement les mêmes conditions ne sont pas identiques.

Depuis, le dogme du chemin unique de repliement a été abandonné suite à de nombreuses observations sur d'autres protéines.

Depuis le début des années 2000, aucun progrès conceptuel significatif n'est venu de l'expérimentation.

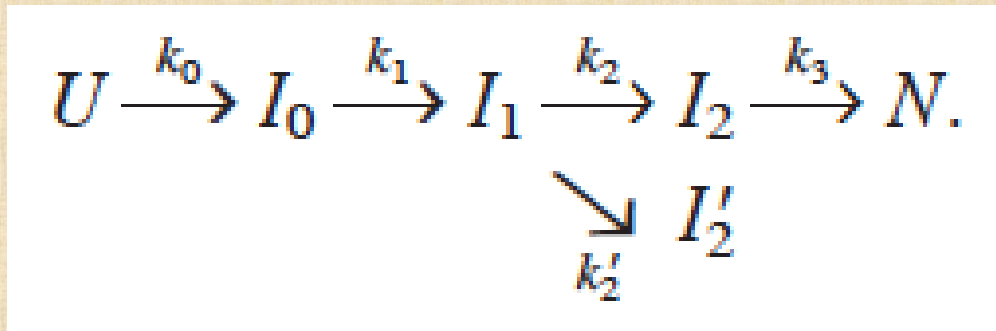
Qu'en est-il de l'étude du repliement par calcul de l'énergie libre du système protéine/solvant ?

L'approche par calcul de l'énergie libre

- Hypothèse de base: **État Natif = État Fondamental**
- Stratégie:
 - déterminer le champs des forces qui stabilisent la structure native
 - chercher l'état fondamental par calcul de **minimisation d'énergie** des conformations

Mais d'abord, l'état **Natif** est-il l'état **Fondamental** ?

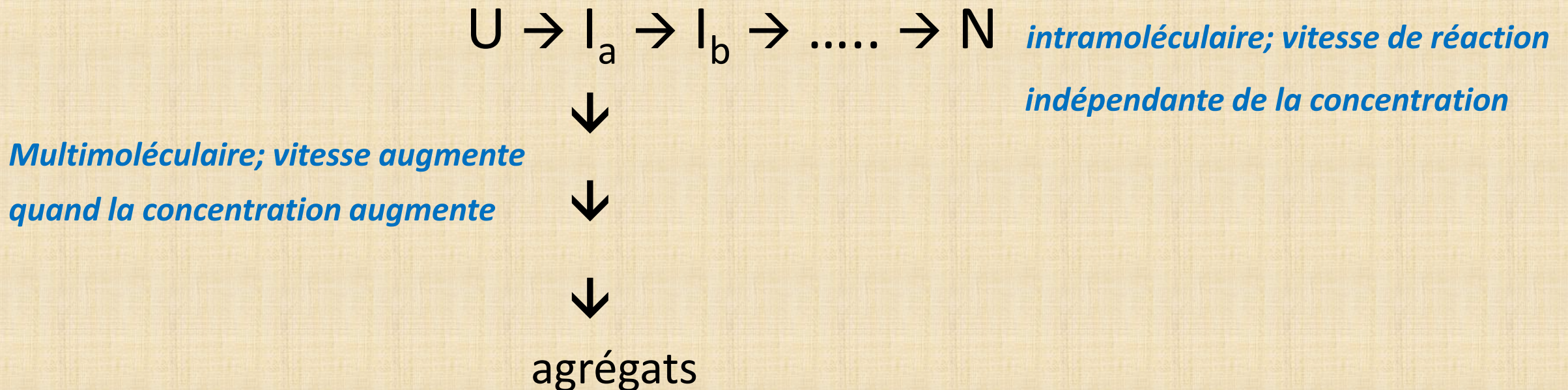
Lors de la renaturation oxydative du lysozyme réduit, il se forme aussi une espèce à 3 S-S, non native, inactive, stable



Qui de I'_2 ou de N est le plus stable ?

Pour le lysozyme réduit, et **d'une manière très générale**, il y a aussi **formation d'agrégats** stables.

Le résultat de la renaturation dépend d'une **compétition cinétique** entre renaturation et agrégation:



Qui est le plus stable, natif ou agrégats ?

La question n'est pas que théorique: les pathologies liées à l'agrégation

- **Maladie « de la vache folle » - Creutzfeldt-Jakob chez l'homme**
*PrPc se change en Prion par contact entre protéine « native »
et protéine pathologique*
- **Alzheimer**
formation de plaques amyloïdes
- **Parkinson**
accumulation d' α -synucléine (type prion)

Certaines protéines natives sont dans un état métastable

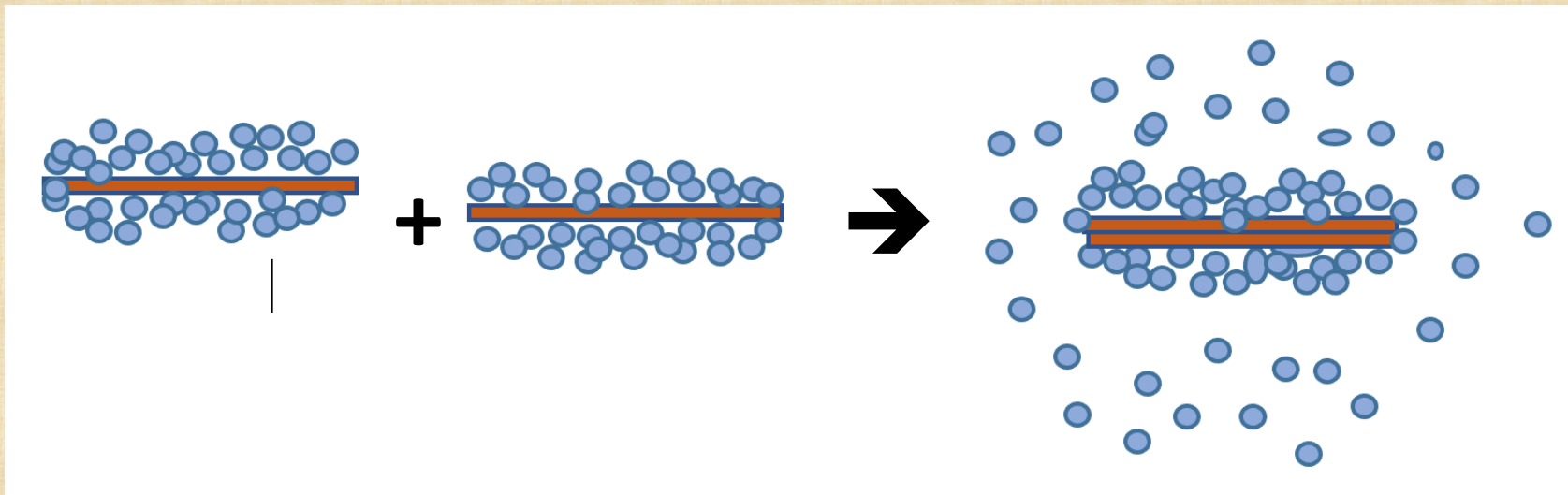
Calcul de l'énergie – le champ de forces

Quelles sont les interactions **non-covalentes** qui stabilisent les protéines:

- **Liaisons hydrogène** (e^- d'un atome d'hydrogène partagé entre deux atomes électronégatifs comme O ou N) : géométrie très précise
- **Interactions électrostatiques** entre aminoacides chargés : varie avec $1/r^2$
 - acides aspartique et glutamique: charge -*
 - arginine, histidine et lysine: charge +*
- **Forces de London - Van der Waals** : varie en $1/r^6$ – très courte distance
attraction puis répulsion électrostatique dues aux déformations des orbitales électroniques de deux molécules très proches



- Interactions hydrophobes: *libération d'H₂O glaciale:*



Forte augmentation du nombre de degrés de liberté = gain d'entropie

Courte distance

Interactions de faible énergie: $\Delta G = -1,4$ kCalorie/mole pour liaison hydrogène

Mais, le secret de la biologie moléculaire:

$$K_{\text{éq}} = e^{-\Delta G/RT}$$

R= constante des gaz parfaits

T= température absolue (310°K)

si 1 liaison « H », $K_{\text{éq}} = 10$

si 2 liaisons $K_{\text{éq}} = 10^2$

si 5 liaisons $K_{\text{éq}} = 10^5$

si 10 liaisons $K_{\text{éq}} = 10^{10}$

pratiquement irréversible à l'échelle de temps de la cellule

Le champ de forces de CHARMM

Algorithme développé à Harvard par **Martin Karplus***.

Fondé sur les travaux pionniers de:

Lifson, Warshel* (Institut Weizmann) : *champ de force CCF pour petites molécules*

Sheraga (Cornell): *rôle des interactions hydrophobes dans le repliement*

Levitt* (Cambridge/Institut Weizmann, puis Stanford/Weizmann): *modélisation des conformations*

Prix Nobel de chimie en 2013

CHARMM intègre l'ensemble des forces, en calcule l'énergie pour chaque conformation de la protéine, et explore l'espace conformationnel par « **dynamique conformationnelle** »: *petites déformations spontanées de chaque conformation successive lors d'une élévation de température (agitation thermique)*

CHARMM a été appliqué d'abord avec succès à de petites hormones peptidiques

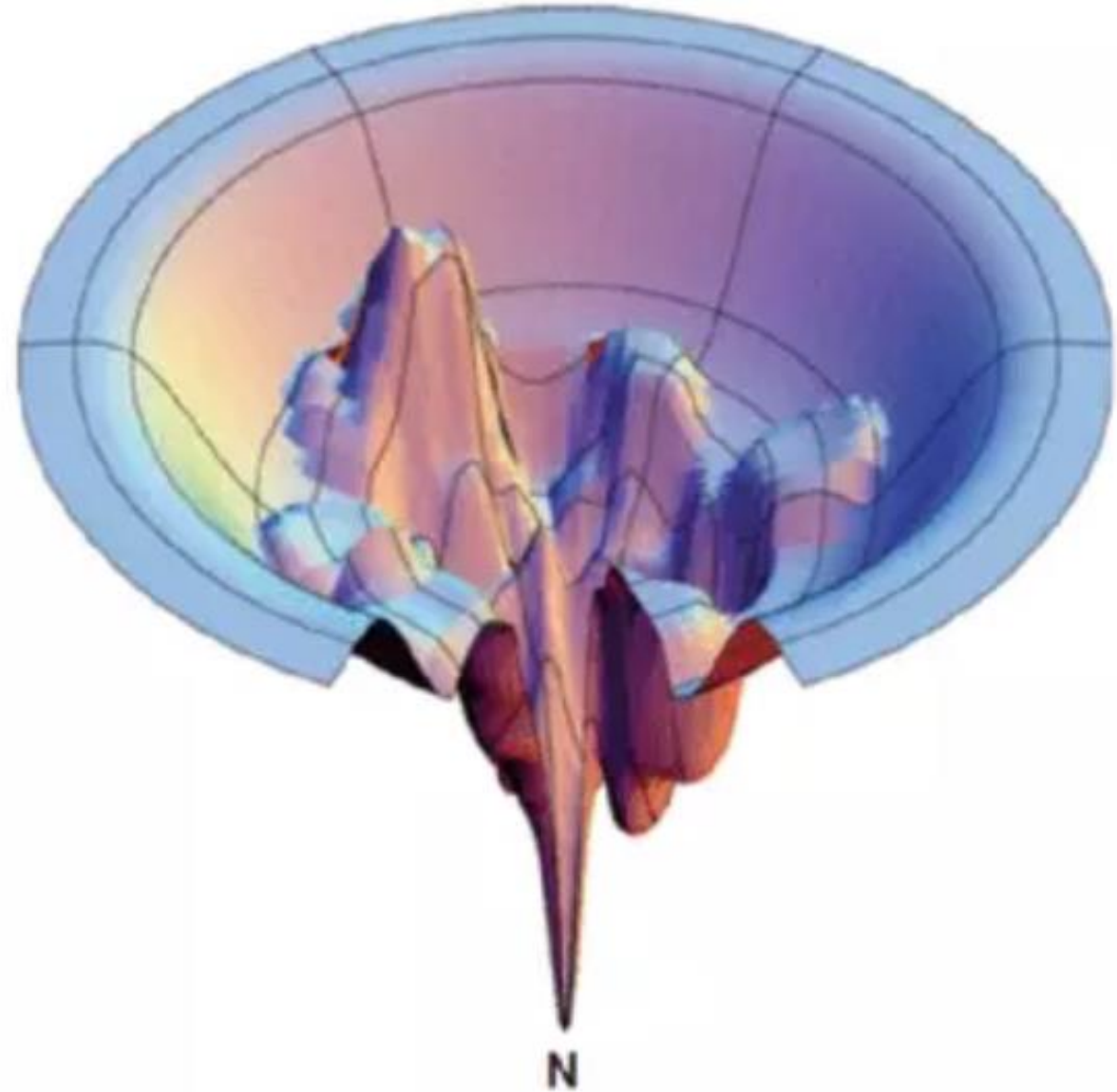
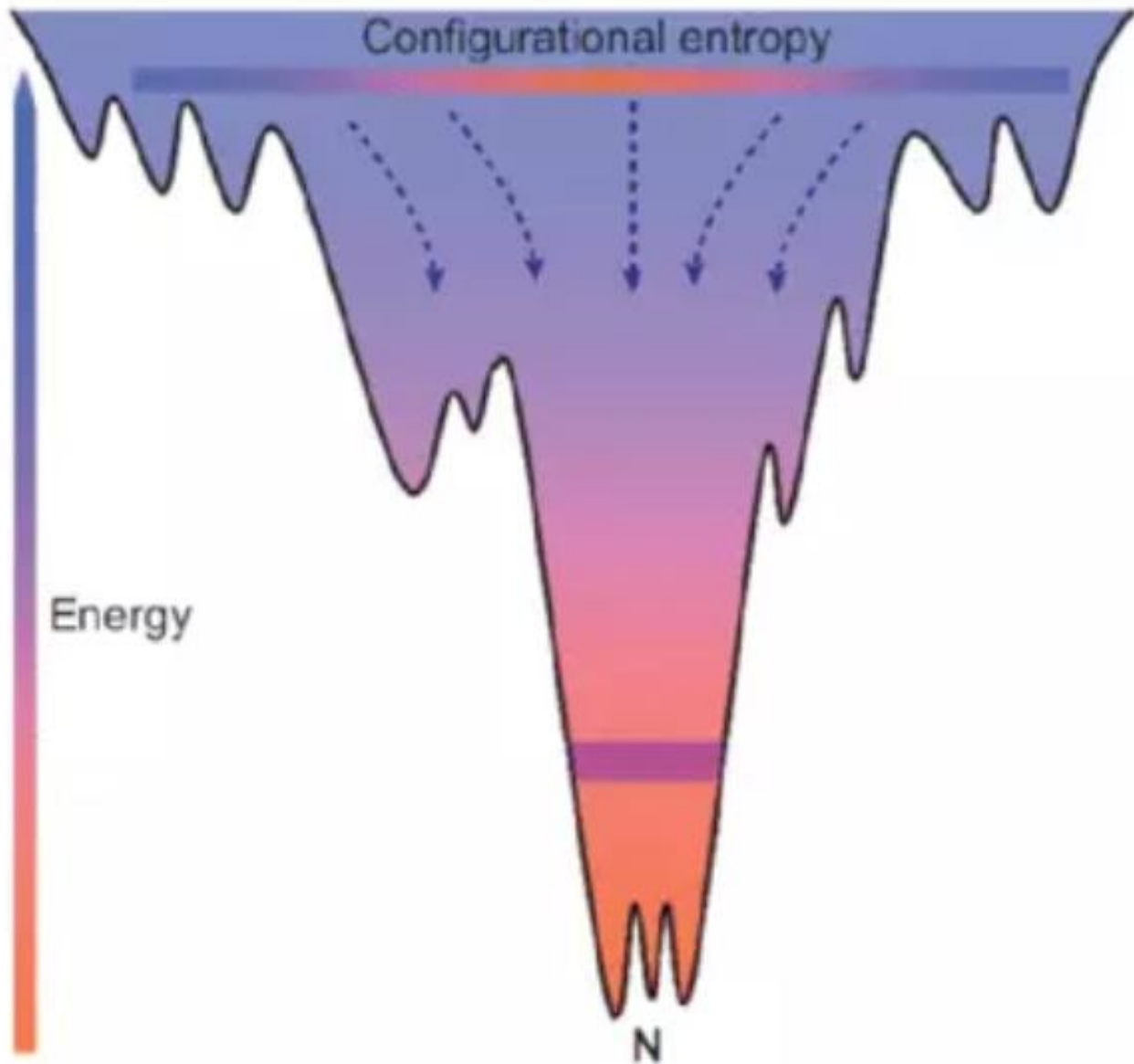
Puis à des protéines un peu plus grandes

Très peu de succès pour la prédiction des structures pendant des années

Mais contribue largement à briser le dogme du « chemin de repliement unique »

Aboutit au modèle de l' « **entonnoir énergétique** » pour le repliement

L'entonnoir énergétique



Implications de ce modèle

- Il existe des chemins multiples de repliement
- Il existe des « pièges énergétiques » (*états métastables*)
- Ne démontre pas que l'état Natif est l'état Fondamental: **Pour des raisons « cinétiques »** (*pour le calcul, durée et énergie d'activation autour d'un piège – pour le repliement, durée de l'observation*) **le calcul et la renaturation peuvent systématiquement mener au même état métastable.**

Pourquoi le calcul d'énergie ne réussit pas à modéliser le repliement

- Même pour 100 acides aminés, le calcul de l'énergie de chaque conformation devient long ...
- La puissance des ordinateurs limite à quelques nanosecondes la durée des « trajectoires » simulées: **trop court pour aller de l'état dénaturé à l'état natif**

s'améliorent petit à petit avec les progrès de l'informatique

- Par contre, CHARMM marche convenablement si le programme démarre d'une conformation déjà proche de la forme native

Comment choisir la conformation de départ ?

La modélisation à partir des structures 3D connues

Essais de prédiction des structures secondaires

Recherche de séquences spécifiques formant des hélices α , des brins β ou des coudes β

1- Synthèse de peptides ayant la séquence étudiée

Mesures par dichroïsme circulaire en milieux de polarité variable (intérieur ou surface)

Pas de corrélation évidente entre la structure du peptide isolé ou dans la protéine.

2- Comparaison de courtes séquences présentes dans plusieurs protéines de structures 3D connues

Une même séquence peut être soit α , soit β , soit non répétitive.

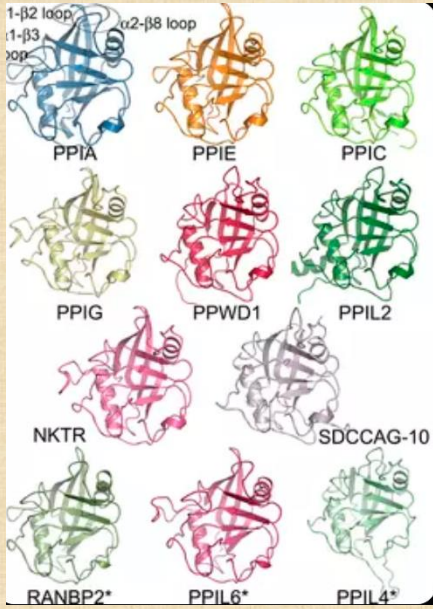
Pas de valeur prédictive de la séquence locale seule ...

3- Analyse des propriétés physico-chimiques des peptides (*distribution et alternance polaire/apolaire*)

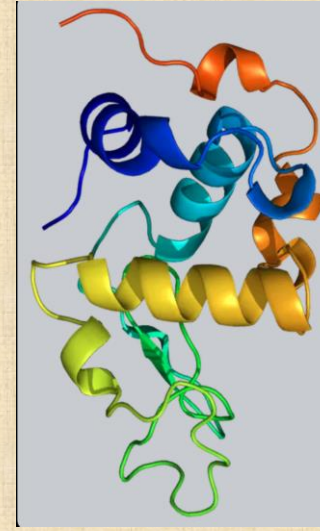
Permet au mieux d'estimer une propension à former une structure secondaire

Mais pas de prédiction sûre

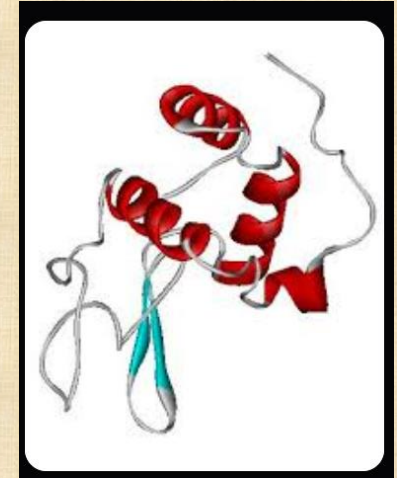
Recherche d'une protéine avec 3D connue et séquence homologue



Famille des cyclophilines



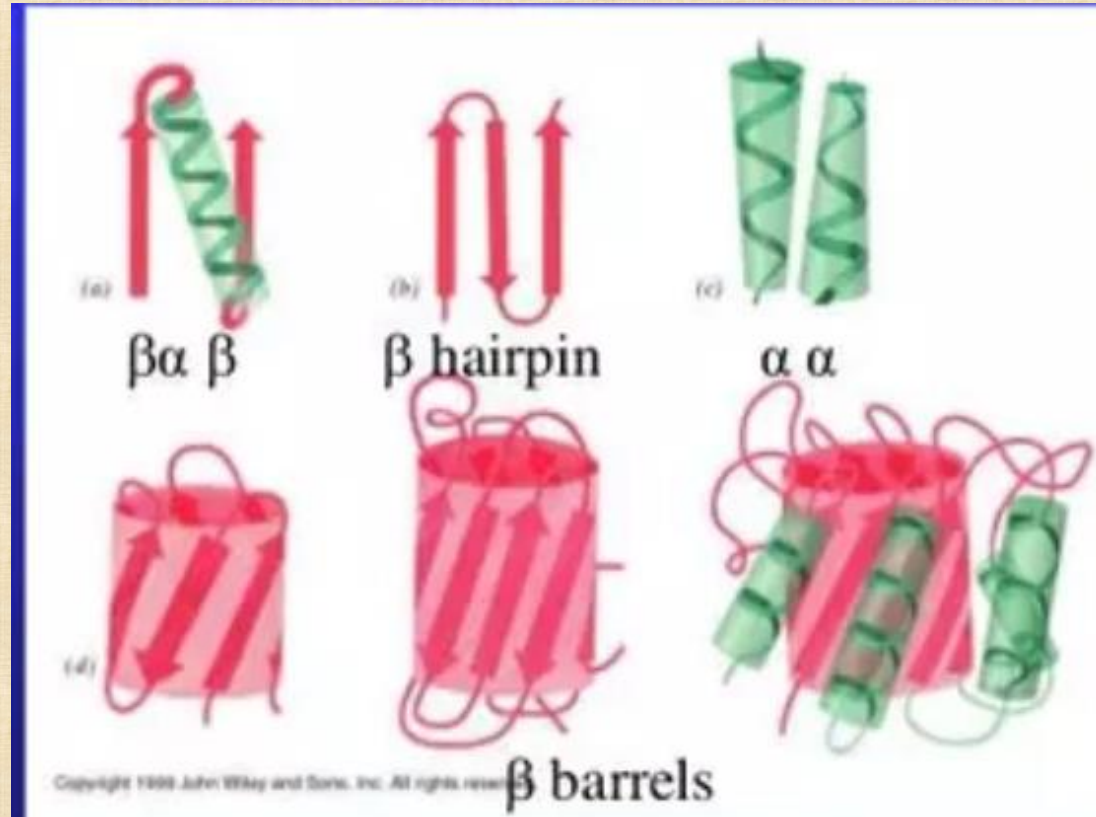
**α -lactalbumine
humaine**



**Lysozyme
de poule**

Ajustement de la chaîne peptidique à modéliser sur celle de la protéine connue, puis affinement de la conformation avec CHARMM

Recherche dans la protéine à modéliser de motifs structuraux connus



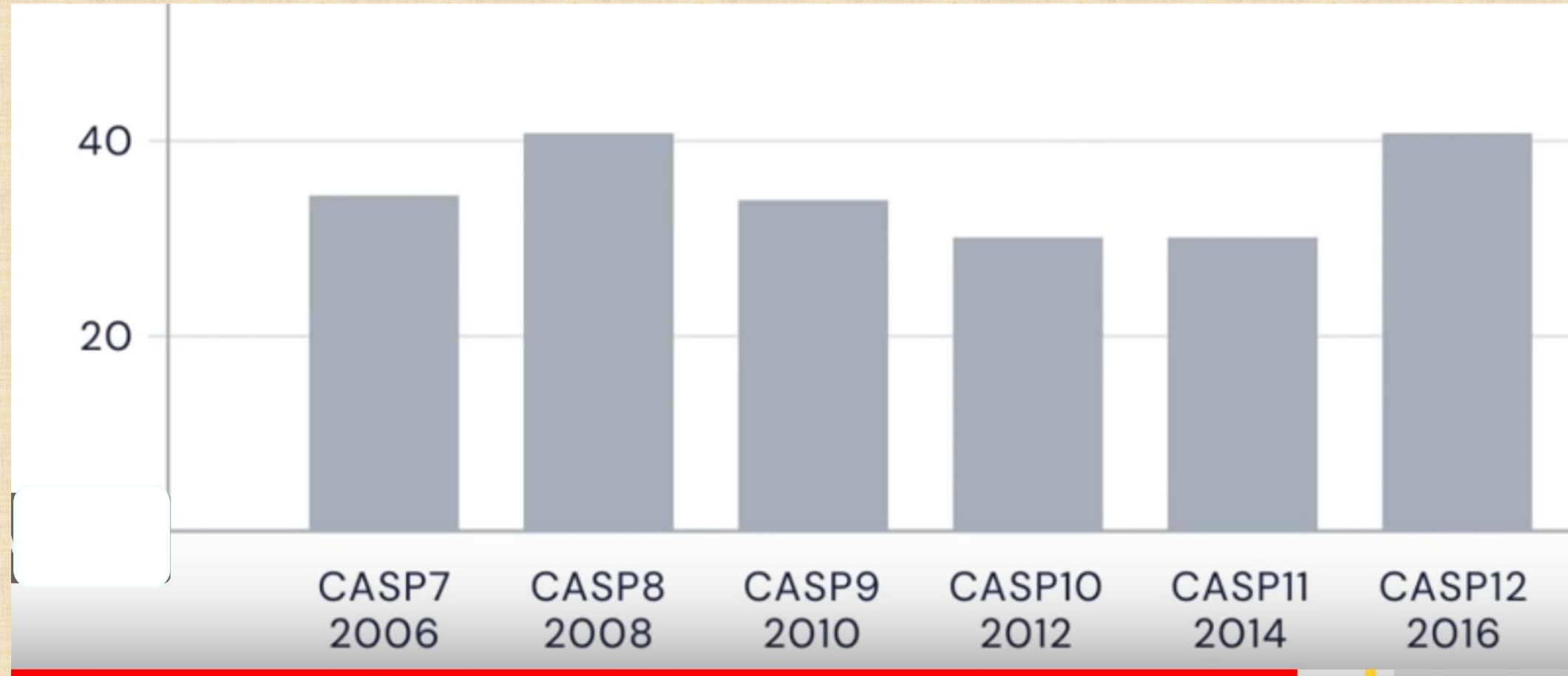
- et bien d'autres trouvés dans la séquence par divers algorithmes
- puis positionnement relatif des domaines par « docking »
 - puis affinement de la structure par CHARMM

Test des méthodes de modélisation: le concours CASP

Tous les deux ans depuis 1994

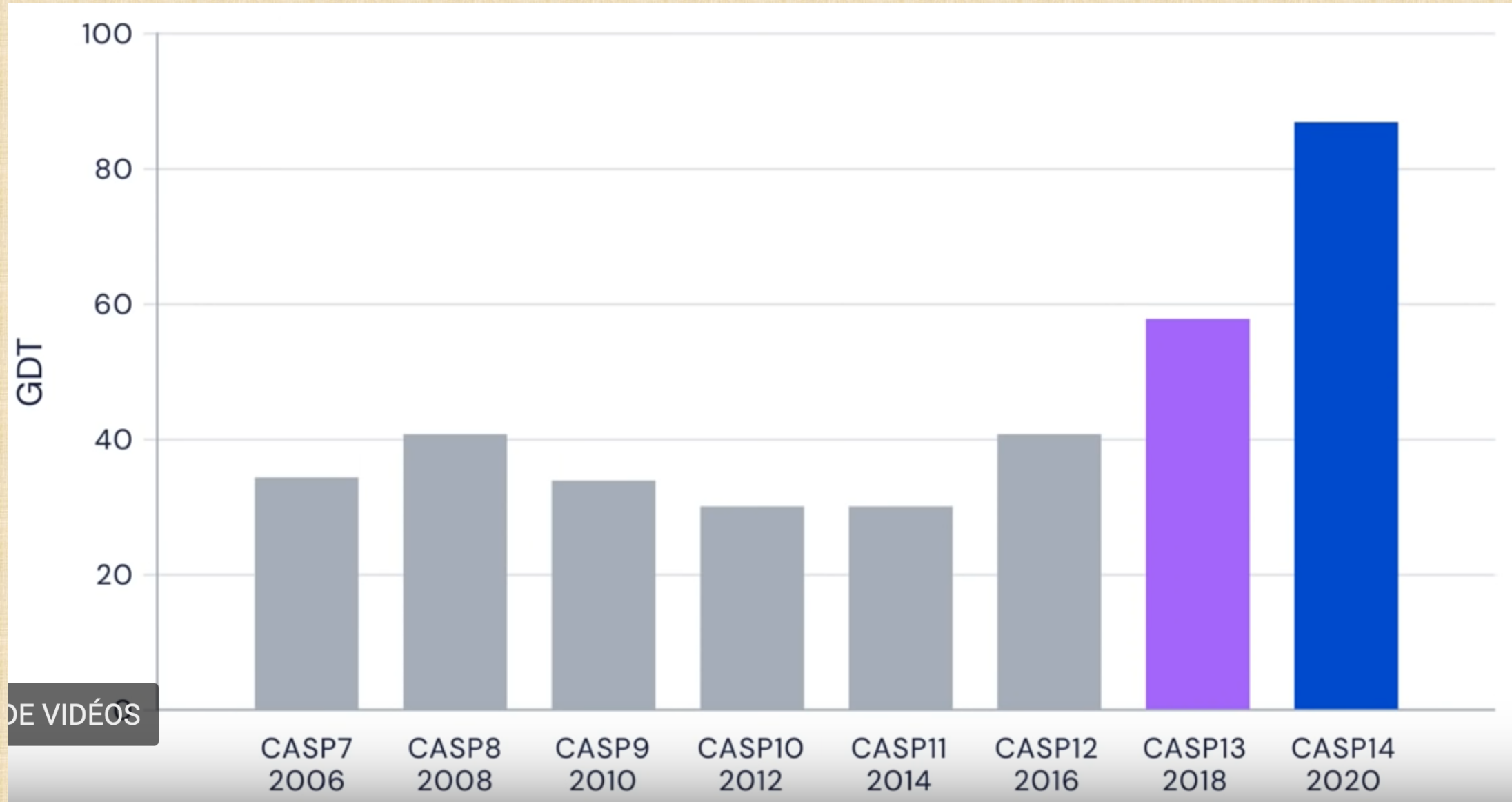
- choix secret de 84 (*en 2020*) protéines de structure 3D déterminée mais pas publiée
- annonce du choix à la centaine d'équipes qui participent au concours
- à la date limite, collecte des prédictions
- détermination du « score » de chaque modélisation (% d'atomes *du squelette peptidique* dont la distance entre position «native» et position «prédite» est $\leq$ à 1 Å)
- pour chaque méthode, moyenne des scores sur l'ensemble des protéines

Évolution de la moyenne du score final de la meilleure équipe (pour les protéines sans homologie)



Succès limité, et surtout, pas de progrès significatifs en 10 ans !!!

Mais en 2018 ... DeepMind est arrivé



2018: 58%

2020: 84%

2022: non publié mais a atteint 95 %

Comment fonctionne AlphaFold de DeepMind ?

Utilise des **réseaux neuronaux** avancés pour prédire:

- les distances entre les **C α** de chaque paires d'acides aminés
- les angles de torsion **Φ** et **Ψ** de chaque liaison peptidique

Les données expérimentales utilisées pour entraîner et tester les réseaux neuronaux sont extraites de la PDB (*Protein Data Bank*). Il existe actuellement près de 200.000 structures 3D et plus de 30.000.000 de séquences connues.

Des **domaines non-redondants** en sont extraits par un programme de recherche de **groupes de séquences** présentant 35% ou plus d'homologie:

➔ 31.247 domaines (29.427 domaines d'entraînement, et 1820 protéines pour tester).

Avec AlphaFold2 en 2022,

Les atomes du squelette peptidique de quasiment toutes les protéines soumises au concours sont positionnés à la précision de l'expérience ($\leq 1 \text{ \AA}$)

Les quelques unes « mal » prédites le sont, mais à $\leq 1,5 \text{ \AA}$ « seulement »

Les atomes des chaînes latérales peuvent être positionnés avec une bonne approximation.

Où en est-on aujourd'hui ?

- Un autre algorithme, **RoseTTaFold**, mis au point par une équipe de l'Université de Washington à Seattle, donne des résultats presque aussi bons que AlphaFold 2
- Les deux méthodes sont disponibles pour la communauté scientifique (*sous licence pour des fins commerciales*)
- On peut désormais considérer que la structure 3D de n'importe quelle protéine peut être déterminée à partir de sa séquence d'acides aminés.

Énorme pas en avant pour recherche, mais aussi pour médecine, industrie, environnement: *nouveaux médicaments, nouvelles fonctions enzymatiques, adaptation de plantes ...*

Oui mais ...

On prédit sans comprendre ...

sans même savoir si l'état natif est l'état fondamental

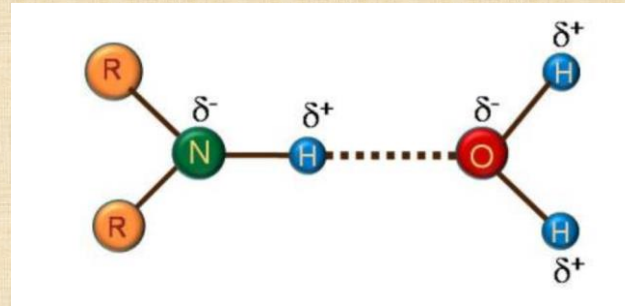
L'intelligence artificielle est-elle si intelligente ?

On le saura sans doute dans quelques années ...

Calcul de l'énergie – le champ de forces

Les interactions **non-covalentes** qui stabilisent les protéines:

- Liaisons hydrogène (*e^- d'un atome d'hydrogène partagé entre deux atomes électronégatifs comme O ou N*) – 1,4 kCal/mole



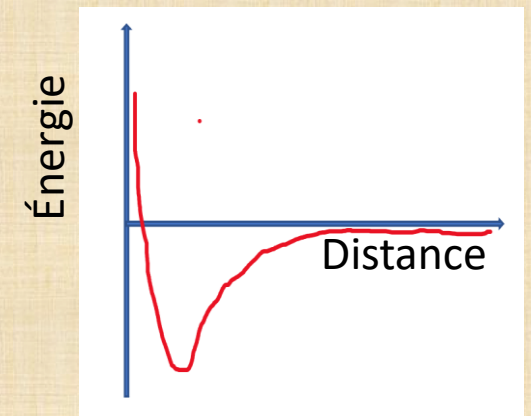
- Forces de van der Waals (*attraction puis répulsion électrostatique dues aux déformations des orbitales électroniques de deux molécules très proches*)

$$E_{\text{van der Waals}} = -\frac{1}{r^6} \left[\underbrace{\frac{2 \mu_1^2 \mu_2^2}{3 (4\pi\epsilon_0\epsilon_r)^2 k_B T}}_{E_{\text{Keesom}}} + \underbrace{\frac{\alpha_1 \mu_2^2 + \alpha_2 \mu_1^2}{(4\pi\epsilon_0\epsilon_r)^2}}_{E_{\text{Debye}}} + \underbrace{\frac{3h}{2} \frac{\alpha_1 \alpha_2}{(4\pi\epsilon_0\epsilon_r)^2} \frac{\nu_1 \nu_2}{\nu_1 + \nu_2}}_{E_{\text{London}}} \right]$$

charge/charge

charge/dipôle induit

dipôle induit/dipôle induit



AlphaFold (CASP13 – 2018) utilise des réseaux neuronaux convolutifs (CNN), la méthode « de pointe » qui a permis les percées spectaculaires de l'IA depuis 2005.

AlphaFold 2 (CASP 14 – 2020 et CASP15 – 2022) remplace les CNN par des réseaux de graphes combinés à l'**attention**.

L'attention est un processus qui donne **plus de poids** aux données qui semblent les **plus importantes**.

Ce **poids se définit et évolue** au cours de l'entraînement.

En août 2021, l’algorithme ALphaFold ainsi que les bases de données utilisées ont été publiés dans *Nature*.

Article

Highly accurate protein structure prediction with AlphaFold

<https://doi.org/10.1038/s41586-021-03819-2>

Received: 11 May 2021

Accepted: 12 July 2021

Published online: 15 July 2021

Open access

 Check for updates

John Jumper^{1,4}, Richard Evans^{1,4}, Alexander Pritzel^{1,4}, Tim Green^{1,4}, Michael Figurnov^{1,4}, Olaf Ronneberger^{1,4}, Kathryn Tunyasuvunakool^{1,4}, Russ Bates^{1,4}, Augustin Židek^{1,4}, Anna Potapenko^{1,4}, Alex Bridgland^{1,4}, Clemens Meyer^{1,4}, Simon A. A. Kohl^{1,4}, Andrew J. Ballard^{1,4}, Andrew Cowie^{1,4}, Bernardino Romera-Paredes^{1,4}, Stanislav Nikolov^{1,4}, Rishub Jain^{1,4}, Jonas Adler¹, Trevor Back¹, Stig Petersen¹, David Reiman¹, Ellen Clancy¹, Michal Zielinski¹, Martin Steinegger^{2,3}, Michalina Pacholska¹, Tamas Berghammer¹, Sebastian Bodenstein¹, David Silver¹, Oriol Vinyals¹, Andrew W. Senior¹, Koray Kavukcuoglu¹, Pushmeet Kohli¹ & Demis Hassabis^{1,4}

Proteins are essential to life, and understanding their structure can facilitate a mechanistic understanding of their function. Through an enormous experimental effort^{1–4}, the structures of around 100,000 unique proteins have been determined⁵, but this represents a small fraction of the billions of known protein sequences^{6,7}. Structural coverage is bottlenecked by the months to years of painstaking effort required to determine a single protein structure. Accurate computational approaches are needed to address this gap and to enable large-scale structural bioinformatics. Predicting the three-dimensional structure that a protein will adopt based solely on its amino acid sequence—the structure prediction component of the ‘protein folding problem’⁸—has been an important open research problem for more than 50 years⁹. Despite recent progress^{10–14}, existing methods fall far short of atomic accuracy, especially when no homologous structure is available. Here we provide the first computational method that can regularly predict protein structures with atomic accuracy even in cases in which no similar structure is known. We validated an entirely redesigned version of our neural network-based model, AlphaFold, in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)¹⁵, demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods. Underpinning the latest version of AlphaFold is a novel machine learning approach that incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm.

