1. Résolution structure 3D

1.1. Cristallographie à rayons x

L'observation d'un objet requière une source de lumière compatible avec la résolution qui est demandée.

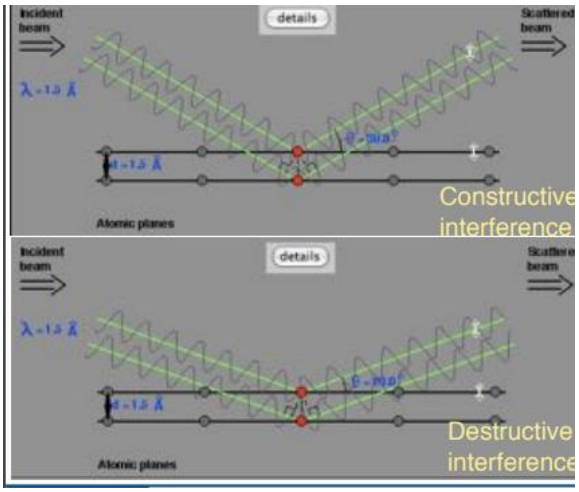
Principe : on projette un faisceau lumineux sur une molécule dans un cristal. Le résultat obtenu est un pattern de diffraction (spots lumineux ou non). L'intensité relative de ces spots donne l'information qui sera utilisée pour identifier l'arrangement moléculaire du cristal.

La source lumineuse utilisée doit avoir une longueur d'onde comparable à la distance entre les atomes (1,5 Angströms). Pour amplifier le signal, il faut utiliser un grand nombre de cette même molécule, sinon cela entrainera une diffusion trop petite.

Formation cristalline : les molécules sont ordonnées, avec un arrangement périodique. Pour que l'expérience se passe au mieux possible, il faut que le cristal soit pur et régulier. L'obtention de celui-ci reste néanmoins très complexe, notamment pour les protéines membranaires. Dans un cristal, seule une partie des molécules est en contact avec celui-ci, et interagissent indirectement avec le solvant.

L'étude du pattern de diffraction obtenue de ces rayons x est une résultante de l'interférence des ondes diffusées par chaque atome :

- Interférence constructive : les rayons ont un chemin de même longueur ou une longueur multiple de la longueur d'onde. Il en découle un spot lumineux qui donnera la position des électrons, et donc des atomes.
- Interférence destructive : le maximum de la longueur d'onde = le minimum de l'autre longueur d'onde. Il n'y a donc pas de spot lumineux.



Le résultat est donc une alternance d'interférences constructives et destructives. Seule une longueur d'onde doit être définie et doit irradier le cristal en tournant autour de lui.

LOI DE BRAGG : permet d'obtenir les conditions d'une interférence constructive :

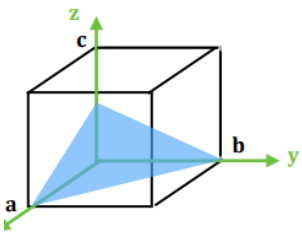
$$n\lambda = 2d\sin\Theta$$
$$\Leftrightarrow d = \frac{n\lambda}{2\sin\Theta}$$

where λ is the wavelength, Θ is the incident angle, n is any integer and d is the spacing between the diffracting plans.

Il y'a donc une proportionnalité inverse entre l'espace entre les plans de diffraction et l'angle d'incidence. A partir de ce pattern, la densité électronique peut être reconstruite.

Le cristal est obtenu par la répétition de cellules unitaires, qui est équivalent aux plus petits groupes de molécules qui caractérisent un lattice cristal. Le cristal peut être obtenu à partir de la traduction de la cellule unitaire. Les axes sont définis dans les 3 directions. La densité électronique peut être obtenue à partir de l'inverse de la transformée de fourrier, qui implique l'indice de Miller (mesure la direction du rayon du cristal), à partir des spots lumineux. L'amplitude de la longueur d'onde résultant de la lumière diffractée sera proportionnelle à hkl.

e Miller indices (hkl) indexes the lattice (diffraction) planes.



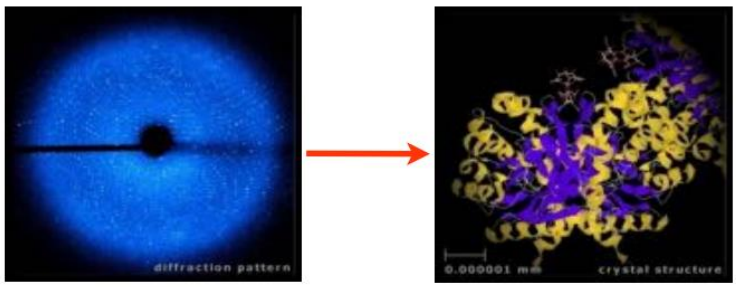
a,**b**,**c** are the length of the edges
Let's consider a plane within the unit cell (in blue)

Intercept with x axis: a
Intercept with y axis: b
Intercept with z axis: c/2

Reciprocal of these intercepts=Miller indices:
(1,1,2)

If they contain a fraction: multiply to obtain integer

The diffraction pattern corresponds to the Fourier transform of the electronic density. The goal is to reconstruct the electronic density from the diffraction pattern and to finally obtain a model of the structure:



The intensity of the spots, I(hkl), can be measured on the diffraction figure. The amplitude of the structure factor is proportional to I(khkl).

Au final, on obtient une carte de la densité électronique en 3D qui devra être en accord avec la modélisation de la structure. S'en suit ensuite un raffinement du modèle (Biophysics 2) pour correspondre au mieux à la carte.

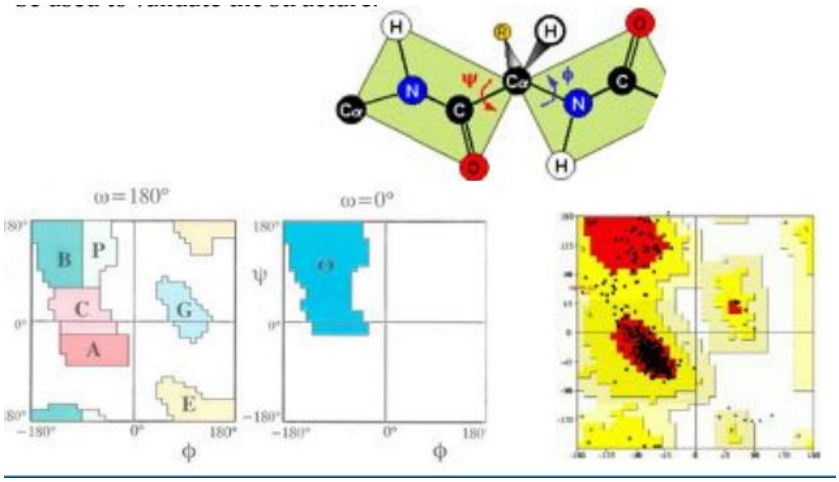
Les facteurs qualitatifs :

- R facteur : mesure si le modèle est en accord avec la densité électronique. 0 = parfait, la moyenne étant d'environ 0,20.

$$R = \frac{\sum_{h,k,l} |F_{obs} - F_{calc}|}{\sum_{h,k,l} |F_{obs}|}$$

- Résolution : donnée par la LOI DE BRAGG qui impose une condition pour avoir une interférence constructive.
  Si oméga approche 90°, on a accès à des détails plus précis avec comme limitation la moitié de la longueur d'onde. Ceci est seulement valide si les molécules sont parfaitement ordonnées et que la conformation est pareil que dans le cristal.
  En réalité, les motifs de diffraction disparaissent lorsque oméga augmente. La structure est donc limitée à un nombre N de détails, et certains atomes ne seront pas résolus.
  Au plus cette résolution est petite, au mieux c'est car moins de possibilités de placer des atomes à la mauvaise position. Elle doit être généralement en dessous de 2.5 Angströms pour avoir de bons résultats.

Pour valider la structure, nous avons la carte de Ramachandran qui va valider les angles dihédraux. Autour du carbone alpha, on peut définir des angles de rotation des chaines latérales. En les comparant à d'autres angles provenant d'autres structures de résolution comparable, on peut voir si la qualité stéréochimique de la structure est bonne ou pas.



1.2. Résonance magnétique nucléaire (NMR)

Etude de la biomolécule en solution, il n'y donc pas besoin de forme cristalline. Elle est utilisée pour les molécules de petite taille (max 400 aa) et pour les molécules qui sont partiellement repliées. Elle mesure la magnétisation globale ainsi que sa variation dans l'échantillon. Il faut donc une entité avec un moment magnétique et un moment angulaire.

Spectroscopie optique : émission ou absorption d'une onde électromagnétique par l'échantillon, qui correspond à la transition entre les différents niveaux d'énergie.

L'interaction matière-radiation se fait à travers le composant magnétique de l'onde électromagnétique. Cela implique des énergies plus basses comparé aux interactions à travers le composant magnétique.

La magnétisation est étudiée à la place de la transition entre les niveaux d'énergie.

Nuclear magnetic resonance requires that the components present a magnetic moment ($\boldsymbol{\mu}$) and an angular moment ($\mathbf{J}$). All nuclei do not present these characteristics.

$\boldsymbol{\mu} = \gamma \mathbf{J}$ where γ is the gyromagnetic ratio. It depends on the nucleus and is related to its sensitivity.

$\mathbf{J} = \hbar \mathbf{I}$ where $\hbar$ is the Planck constant divided by 2π and $\mathbf{I}$ is the spin of the nucleus.

Nuclei such as ^{12}C, ^{16}O, ^{2}H present a spin equal to 0, whereas ^{1}H, ^{13}C, ^{15}N, ^{31}P present an half-integer spin ($\mathbf{I}=1/2$) and can be used in NMR.

When a nucleus is placed in a magnetic field, $\mathbf{B_0}$, the energy is equal to:
$E = - \boldsymbol{\mu} \bullet \mathbf{B_0}$

If the magnetic field is oriented along the z-axis:
$E = - \mu_z \bullet B_0 = - \gamma \hbar I_z B_0$

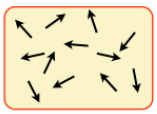
La plupart des atomes ont un nombre pair d'électrons : spin = 0. Il faut donc utiliser C13, H1,… Ici, on utilise souvent H1.

In quantum mechanics, the allowed values for I_z, m_s, are quantified: $m_s = \mathbf{I}, \mathbf{I}\text{-}1, \mathbf{I}\text{-}2, \ldots. \text{-}\mathbf{I}$.

In this case, $\mathbf{I}=1/2$, which means that there are two energetic levels:
$E_1 = + 1/2 \gamma \hbar B_0$ et $E_2 = - 1/2 \gamma \hbar B_0$

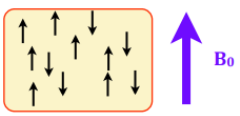
Without magnetic field, the nuclei are unordered.

There is no energy difference between the states.



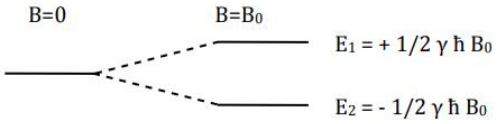
With a magnetic field, the nuclei with a magnetic moment are oriented.

There is an energy difference between the states.



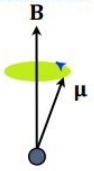
There is a small excess of nuclei oriented along the field.

There is an energy difference between the levels:



$\Delta E = \gamma \hbar B_0 = h \omega_0$ where ω_0 is the Larmor frequency. It corresponds to the frequency of precession around the z-axis, or to the resonance frequency. In a magnetic field of 18,7 Tesla, the resonance frequency of a proton is equal to 800 MHz.

Precession



Note that the energy difference between two levels depends on the intensity of the magnetic field. The frequency of Larmor depends on B_0 and γ.

Le ratio entre la population de deux niveaux est donné par la loi de boltzmann :

$$\frac{N1}{N2} = \exp(-delta\ E/kT)$$

Pour H1, gamma, est connu, et k est une constante de boltzmann. En augmentant B0, le ratio augmente (certes de peu mais impact négatif). Le moment magnétique de chaque noyau s'additionne, menant à une magnétisation macroscopique. Cette magnétisation est orientée le long du champ magnétique (axe z) et x et y sont nuls. Ensuite, un champs magnétique perpendiculaire B1 est appliqué pour perturber le système. Si la fréquence est proche de la fréquence de Larmor, le moment magnétique va switcher dans le plan XY. Après, on stoppe le champs magnétique B1, le moment angulaire va donc revenir vers l'axe Z : la NMR va suivre cette évolution de magnétisation (relaxation).

Il y aura deux constantes de temps :

- T1 = relaxation de la matrice de spin, mesure le retour vers Mz
- T2 = relaxation spin-spin, mesure l'évolution vers 0 de Mx et My.

Shift chimique : la fréquence de résonnance des différents h1 va dépendre de leurs environnements. Les autres noyaux, aussi chargés, vont influencer le champ magnétique et donc les noyaux environnants -> Une correction peut être obtenue via M0 (1-sigma) où sigma = 10^{-6}, avec importance d'avoir un champ magnétique très homogène, sinon sigma passe inaperçu. En pratique, toutes les fréquences ne sont pas testées pour B1. En travaillant avec des temps courts de pulse pour scanner un large spectre de fréquences, on excite tous les noyaux.

Les différents spectres obtenus :

- 1D NMR : 2 étapes (préparation avec impulsion de 90° par ex, et détection). Le spectre obtenu est assez difficile à interpréter.
- 2D NMR : caractérisé par une séquence d'impulsion différente. On obtient un spectre sur deux axes de fréquence. Il est plus facile de détecter les H1 qui sont en contact dans la protéine avec ce spectre. Les contraintes de distance sont dérivées à partir du spectre et celles-ci vont être utilisées pour construire un modèle 3D. L'information donnée par la NMR sera donc les contacts entre les atomes spécifiques H1, qui donnera une idée de la distance entre les aa. En bout de chaine, les contraintes sont moins grandes, on peut donc aboutir à différents modèles de structure.

2. Informations sur la structure des biomolécules

Méthodes complémentaires à NMR, XRay, …
Le principe de spectroscopie d'émission et d'absorption est de détecter les longueurs d'onde auxquelles une molécule émet ou absorbe une radiation électromagnétique. Cette absorption est quantifiée, elle correspond à la transition entre les différents niveaux d'énergie.

- La transition entre les niveaux électroniques moléculaires ont lieu dans le spectre UV-visible. Au plus la longueur d'onde est petite, au plus l'énergie est grande.
- La transition entre les niveaux vibrationnels moléculaires a lieu dans l'infrarouge. Petite énergie mais grande longueur d'onde.

Chaque niveau électronique peut être excité pour passer à un niveau d'énergie supérieur.

2.1. Dichroïsme circulaire (utilise les UV)

Spectroscopie d'absorption qui utilise le niveau d'énergie électronique. Elle est rapide (30 min) et nécessite peu de matériel (100 microgrammes).

But :

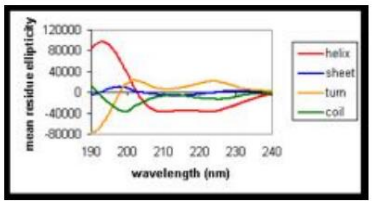
- Evalue le % de structure secondaire d'une protéine
- Suit le changement de conformation
- Etudie le repliement et la dénaturation des protéines
- Etudie la dynamique des molécules

Le CD mesure l'absorption différentielle de la lumière circulaire polarisée droite et gauche. Certaines structures présentes dans les biomolécules ont cette absorption chirale (présentes dans structures secondaires, ADN, …) et donc un spectre CD. La différence d'absorbance entre la lumière polarisée gauche et droite est donnée par la loi de Beer-Lambert : Delta A = (Kgauche – Kdroite) x C x L

Où K et le coefficient d'extinction molaire, C la concentration et L le chemin optique.

In CD, the molar ellipticity (θ) is measured; it depends on the molar extinction coefficients ε_G and ε_D ($\theta \div \Delta\varepsilon$).

The ellipticity depends on the type of secondary structure in a protein:



The spectrum is in the wavelength range 190 - 300 nm (UV). This range corresponds to electronic transitions in main chains and side chains of peptides, and in purines and pyrimidines in DNA.

Le spectre obtenu pour une protéine est la somme de la contribution des différentes structures secondaires présentes dans la protéine.

2.2. Spectroscopie à infrarouge

Elle permet d'étudier l'absorption ou l'émission des groupements NH et CO des chaines principales. Les radiations infrarouges correspondent à la transition entre les niveaux vibrationnels. La longueur d'onde à laquelle le groupe chimique Amide absorbe dépend de la structure secondaire dans laquelle il se trouve. A noter aussi que l'eau absorbe dans les UV, il faut donc travailler avec une combinaison H2O-D2O (eau lourde)

3. Stabilité thermodynamique des molécules

Rappels de thermodynamique :

- Energie interne (ΔU) = q (chaleur) + W (travail)
- W = -P x ΔV (P = pression, V = volume)
- La différence d'enthalpie ΔH = ΔU + Δ(PV)
- Le changement d'énergie libre ΔG = ΔH – T x ΔS (entropie)

> At constant volume: $q_V=\Delta U$
> At constant pressure: $q_p=\Delta H$
>
> The heat is also defined by:
> $q_V = n\ C_V\ \Delta T$ et $q_p = n\ C_p\ \Delta T$
> where C_p and C_V are the constant-pressure and the constant-volume heat capacity, respectively.

Rappels sur la thermodynamique. Formules très importantes :

- Si on est à volume constant (ΔV = 0), ΔU = q (q = n.C.ΔT)
- Si on est à pression constante, ΔH = q (q = n.C. ΔT)

Microcalorimétrie : elle sert à évaluer la stabilité thermique d'une molécule, ou mesurer les interactions entre molécules.

2 types d'expérience : ITC (isothermal titration calorimetry) pour l'interaction entre 2 molécules, un COMPLEXE (2 protéines, protéine-ligand/ADN) et DSC (differential scanning calorimetry) pour la stabilité d'une macromolécule.

- ITC (isothermal titration calorimetry)

  2 cellules :

  - une de référence qui contient des protéines en solution.
  - une de test, qui contient les mêmes protéines en solution (même concentration).

  On injecte ensuite le ligand dans la cellule test. Après cela, on mesure la chaleur nécessaire pour garder les 2 cellules à même température en sachant qu'en injectant le ligand, on aura une interaction entre celui-ci et la protéine. La température va donc changer, et ce changement est expliqué directement par cette interaction.

  Le système possède un équipement pour soit changer ou soit réduire cette température de la cellule test

  Le ligand est injecté par pulses, ce qui permet d'obtenir un graphe qui nous donne des informations sur la force d'interactions, la stoechiométrie de l'interaction (une protéine se lie sur une, deux molécules de ligand? Vice-Versa).

- DSC (differential scanning calorimetry)

L'idée est un peu différente. On l'utilisera pour mesurer stabilité d'une macromolécule, ou du nombre de domaines d'une protéine. Ici, on n'injecte pas de ligand. On a encore 2 cellules dont une de référence et une de test.

Par exemple, on veut évaluer la stabilité thermique d'une protéine. Pour la cellule de test, des protéines sont en solution. Pour la cellule de référence, il n'y a que la solution. Une différence de stabilité (et donc de chaleur à fournir) est donc présente entre ces deux cellules. Il faut mesurer la chaleur à fournir pour égaliser la température entre celles-ci. Lors de l'analyse du graphe (slide 74), il faut réaliser une intégration de delta H/Delta T.

La stabilité de la molécule est dépendante des conditions du milieux : pH, salinité,... Les graphes vont dès lors changer. Plus le pic est grand, plus la molécule est stable car besoin d'une plus grande température pour changer cette thermostabilité.

4. Les différents types de liaison

- Liaison covalente : lien chimique où 2 électrons sont partagés entre deux atomes d'une molécule : eau, NH4. Pas cassé par la température. Il est plus fort que les que les autres types de liaisons et sa force baisse en présence de solvant. En présence d'eau, l'interaction ionique est plus faible que la covalente.
- Non covalente : Lien chimique où 2 atomes ne sont pas "connectés" entre eux : forces de Van der Waals, liens H, liaison ionique.

La force de liaison est différente : la covalente est la plus forte (90/100 kcal/mol), alors que les forces de Van der Waals sont les plus basses (1 kcal/mol). Quant aux liaisons ioniques, elles se situent entre les deux.

Certains acides aminés présentent des charges positives (Lysine, arginine) ou négatives (Aspartate, glutamate): des interactions électrostatiques sont donc présentes. La force entre ces interactions est exprimée par cette loi:

$$E = \frac{q1\ x\ q2}{\varepsilon 1\ x\ \varepsilon 0\ x\ r}$$ où:

- q1 q2 = les charges
- E1/0 sont des constantes diélectriques
- r est la distance

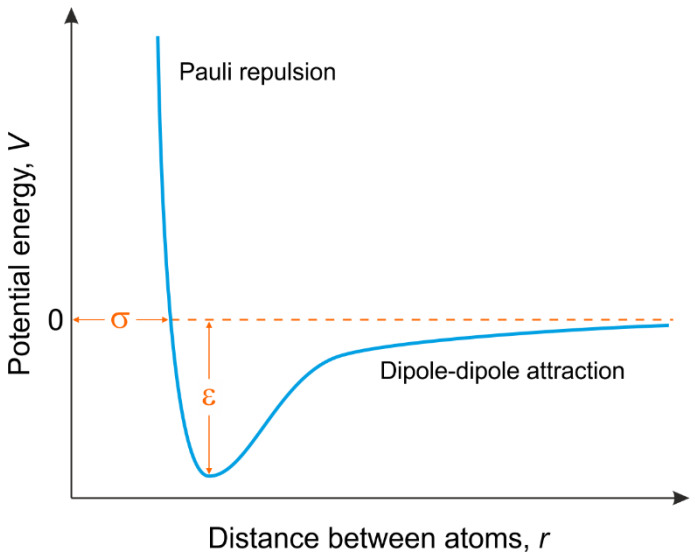
Ces constantes dépendent du milieux (solvant, salinité), qui aura un impact sur cette énergie d'interaction.

- Interactions de Van der Waals

  Se produit entre tous les atomes (voisins).

  -Part attractive : effet de polarisation de l'atome. Le nucléus est composé de protons/neutrons, avec des électrons en mouvement autour de ce nucléus. Cette densité d'électrons n'est pas tout le temps homogène. De temps en temps, une grande partie de ceux-ci peuvent se retrouver plus d'un côté que de l'autre, ce qui crée une polarisation. Si, dans un autre atome, une hétérogénéité se forme aussi, une liaison de faible énergie se formera entre ceux deux atomes. Elle est égale à 1/d^6

  -Part répulsive : à très courte distance (en dessous de 4 angströms), on remarque une répulsion entre les deux nucléus, par excès de charges positives (voir courbe du potentiel de Lennard-Jones).



- Lien hydrogène

Exemple avec la molécule d'eau. L'oxygène étant plus électronégatif que l'hydrogène, il attire à environ ⅔ de la distance l'électron des 2 hydrogènes (liaison covalente avec partage d'électrons, mais plus d'un côté que de l'autre). L'oxygène possède donc deux "moments" négatifs, alors que chaque hydrogène possède un moment positif. Cette légère polarisation induit la formation de 2 liens hydrogène entre l'oxygène d'une molécule (2 liaisons car deux paires électroniques libres) et 2 hydrogène d'une autre. Mais encore, les 2 hydrogènes de la première molécule vont eux aussi effectuer des liens H avec l'oxygène d'une autre molécule Ainsi, une molécule d'eau peut faire 4 liens H.

- Pi system

Dans le benzène (C6H6), on retrouve un mouvement continu d'électrons dans ce cycle aromatique, ce qui induit un système dynamique de doubles liaisons en position 1/3/5, puis en position 2/4/6, jusqu'à l'infini. Plus précisément, toutes les orbitales vont être partagées entre les 6 atomes de carbone. Les électrons sont délocalisés, ils sont donc nommés pi électrons (revoir orbitale atomique, moléculaire, orbitale pi, orbitale sigma, ...) Une interaction pi cation est donc la liaison temporaire entre une charge positive et un électron délocalisé. Cela est présent chez certains acides aminés comme le tryptophane (W), Phénilalanine (F/Phe).

- pi-pi stacking

Présent dans un pi system. Par exemple, on prend un system pi (molécule aromatique plane) avec un autre (au-dessus du premier par exemple). On aura donc une interaction entre ces deux systèmes, au niveau de leurs orbitales, par une sorte d'effet quantique. Il faut donc absolument avoir des plans parallèles avec des électrons délocalisés.

Cette interaction se retrouve beaucoup dans l'ADN via les cercles azotés (bases nucléiques) étant organisés de manière parallèle dans la double hélice.

Dans les protéines, certains acides aminés présentent un pi system (F/Phe, W,...) et peuvent donc réaliser ce type d'interaction, si les conditions sont réunies (//, …)

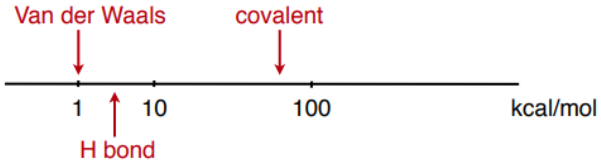
- Effet hydrophobe

Très important pour les protéines. Va induire un repliement des structures hydrophobes en milieu hydrophile (Hélice alpha).

Dans l'eau, on retrouve énormément d'interactions de Van der Waals, liens H, … Quand on place une molécule hydrophobe dans cette eau (huile par exemple), du fait que ces types de liaisons ne s'opèrent pas entre molécules hydrophobes, on aura une séparation (une sorte de repliement entre ces deux systèmes) pour diminuer la surface d'interaction.

Par exemple, le méthane (CH4) n'a pas assez de différence d'électronégativité entre ses 2 atomes pour créer la polarisation présente dans l'eau. Cette molécule ne sait donc pas faire de liens hydrogène ou liens électrostatiques, c'est pourquoi elle est qualifiée d'hydrophobe. Les chaines carbonées sont en général hydrophobes.

Dans une chaîne protéique non repliée, certains acides aminés sont hydrophobes, et vont essayer de réduire au maximum leur surface de contact avec les acides aminés hydrophiles, ce qui va provoquer une formation d'hélice alpha. Assez fréquent dans le cœur d'une protéine, ce qui laisse la surface pour des systèmes hydrophiles.



5. Repliement des protéines

Deux types de stade :

- Stade dénaturé (D) : stade désordonné, énergie libre élevée, set de conformations d'énergie similaire qui change tout le temps au cours de temps. Entropie et enthalpie élevée. Peut contenir des structures résiduelles secondaires ou groupements hydrophobiques et ponts disulfures qui réduisent la flexibilité de la chaine. Il existe des protéines qui ne se replient jamais
- Stade replié (N) : l'énergie enthalpique est basse avec une faible entropie puisque beaucoup d'interactions vont stabiliser la protéine, ce qui empêche une conformation différente. Dans cette conformation, la protéine joue un rôle biologique

5.1. Thermodynamique du repliement

Relation between ΔG and ΔG^0 :

$\Delta G = \Delta G^0$ + a term dependent on the concentration of reactants and products (here P const)

$= \Delta G^0 + RT \ln Q$

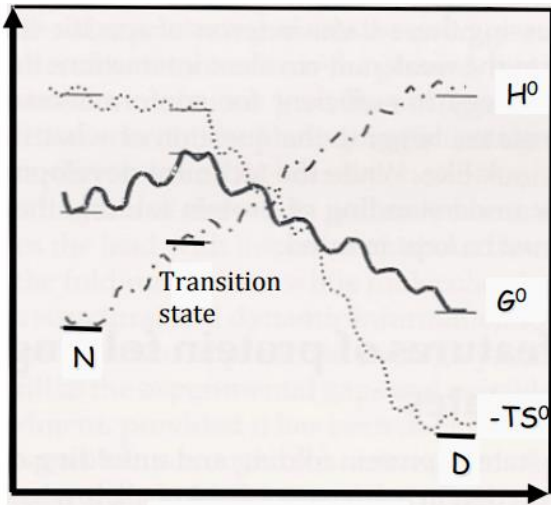
At equilibrium, $\Delta G = 0$ and $\Delta G^0 = - RT \ln K$
with K = equilibrium constant = [N] / [D]

- $\Delta G^0 = 0 \Rightarrow [N] = [D]$
- $\Delta G^0 < 0 \Rightarrow [N] > [D]$
- $\Delta G^0 > 0 \Rightarrow [N] < [D]$

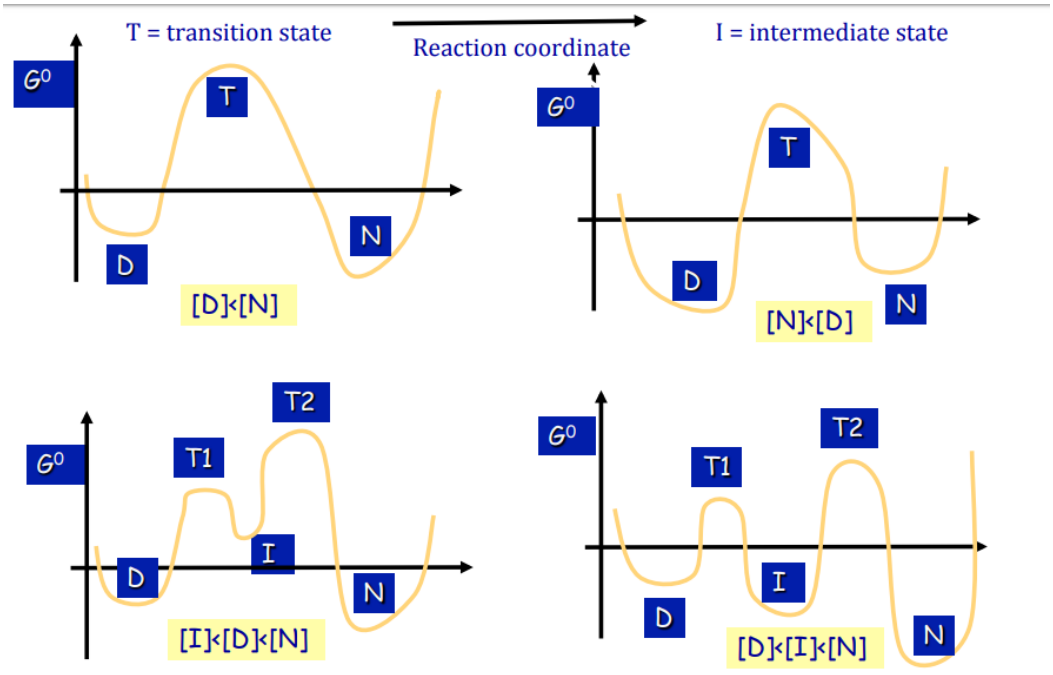
ΔG^0 depends on various parameters, T, pH, etc..

In a range of values of these parameters : $\Delta G^0 < 0$, N is more stable than D.
In another interval, $\Delta G^0 < 0$, and D is more stable than N.

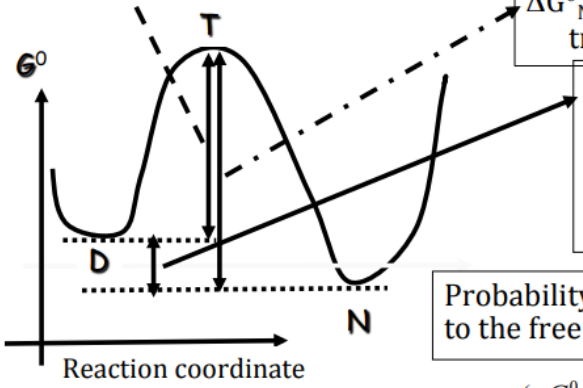
Dans l'état natif, l'enthalpie est petite (favorable) car stabilisation des interactions entre résidus et l'entropie est petite (défavorable) car peu de désordre
Dans l'état dénaturé, l'enthalpie est grande (défavorable) car peu ou pas d'interactions entre les résidus, et l'entropie est grande (favorable) car beaucoup de désordre.
L'état de transition entre D et N est localisé au point où la baisse d'entropie n'est pas compensée par la baisse en énergie.



Il existe un état intermédiaire entre D et N. Une barrière d'activation doit être franchie. Au plus le minimum d'énergie est profond, au plus la protéine va s'accumuler dans des états intermédiaires.
Au plus la différence d'énergie libre est grande, au plus la différence entre les stades est grande également, au plus l'enthalpie est positive, au plus la fraction des protéines au stades dénaturé est grande, et au moins la forme sera stable (et inversément). Au plus la barrière d'énergie vers T est grande, au plus il faudra attendre longtemps.



ΔG^0_{DT} = activation free energy = G^0 (T)-G^0(D) of the transition D → N
= - RT ln $k_{D\to N}/k_0$= RT ln $t_{D\to N}/t_0$
$k_{D\to N}$ = rate constant of the reaction D→N ; k_0 = k of « elementary » step
$t_{D\to N}$ =duration of the transition D→N ; t_0 = t of « elementary » step
The reaction is faster if the free energy barrier is small / i.e. if ΔG^0_{DT} is small
(→ **kinetics of the process**)



ΔG^0_{NT}= activation free energy of the transition N → D

ΔG^0=G^0 (N)-G^0(D)=-RT ln([N]/[D])
N is more stable thermodynamically if ΔG^0 is more negative
(→ **thermodynamics of the process**)

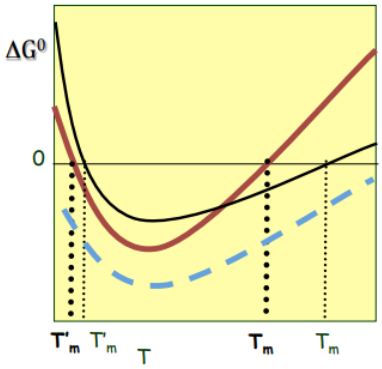
Probability of occurrence of a state (P) is related to the free energy by the Boltzmann law :

$$P(N) = \frac{\exp(-G^0(N)/RT)}{Z} \quad \text{avec} \quad Z = \sum_A \exp(-G^0(A)/RT)$$

Z : partition function

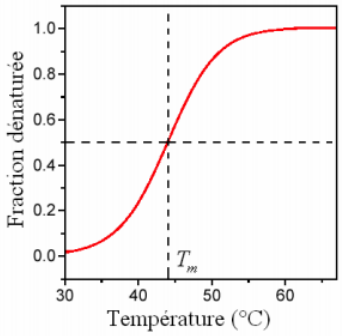
En bas de l'image, la probabilité de Boltzmann de trouver une protéine dans son état natif. Z = sommes des facteurs de boltzmann sur toutes les conformations.

Dependence of the free energy on the temperature T



Reversed bell-like shape
T such that $\Delta G^0(T)=0$: Melting/denaturation temperatures T_m.
In general, a protein has 2 T_m, one at high and one at low temperature.

$$\Delta H^0(T_m) = T_m \Delta S^0(T_m)$$



at T = T_m, there are as many proteins in the states N and D :

$$\Delta G^0(T_m) = -RT_m \ln[N]/[D] = 0$$

- A gauche : T° de fusion due au froid, même énergie libre pour N et D (0)
- Au milieu en bas : T° idéale, Delta G° < 0 : N >>> D. Plus c'est bas, plus la stabilité thermodynamique est grande
- A droite : T° de fusion : même énergie libre pour N et D (0). Au plus c'est loin sur l'axe x, signifie plus stable thermiquement.

Heat capacity : constante qui dépend de notre échantillon. C'est la quantité de chaleur nécessaire pour augmenter la température d'1° d'une mole de notre échantillon.

A P° constante :

$$C_P^0 = \left.\frac{dH^0}{dT}\right|_{Pcte} = T\left.\frac{dS^0}{dT}\right|_{Pcte}$$

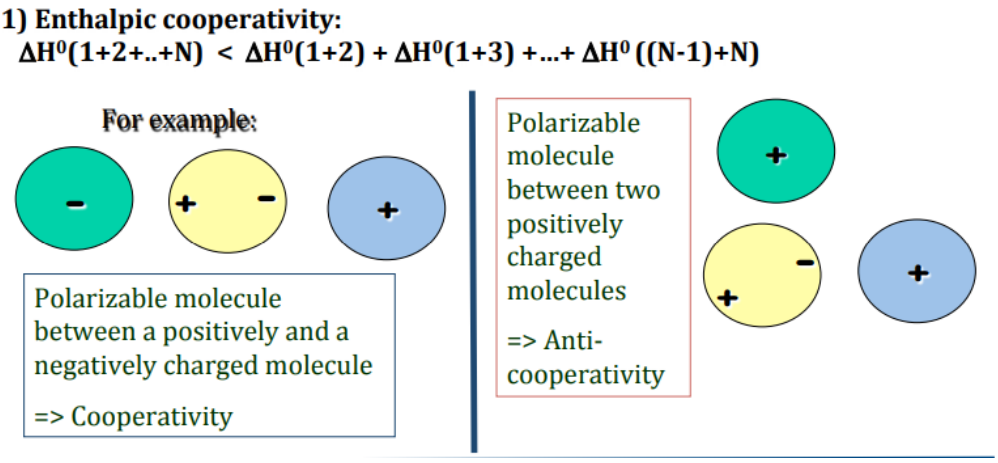
A pression constante, Cp est la dérivée de l'enthalpie en fonction de la température.
L'énergie libre G°(T) est de H-TS
Delta C°p varie très peu : même constante peut être sortie de l'intégrale (voir slide 2 :14 si t'es chaud)

Expérience d'Afinsen : protéine repliée en solution. On ajoute un agent réducteur (mercaptans) qui rompt les ponts disulfures, puis on ajoute un agent dénaturant (chaleur ou urée).

➔ La protéine est dépliée, et on voit qu'elle va se replier toute seule (si elle n'est pas trop grande). En conclusion, les protéines contiennent toute l'information nécessaire dans leur séquence primaire pour leur repliement.

La dénaturation des petites protéines est souvent une transition coopérative :

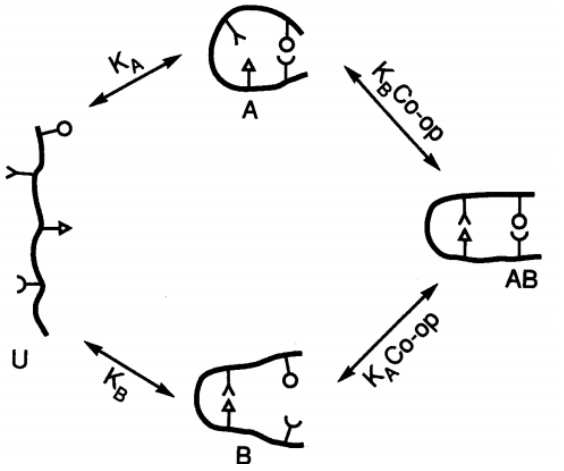
- Coopérativité : Molécule polarisée entre deux molécules chargées de manière opposées
- Anti-coopérativité : Molécule polarisée entre 2 molécules chargées pareilles : les interactions se gênent l'une l'autre
- Coopérativité enthalpique : dans le cas d'une protéine neutre mais polarisable de par son entourage en mettant 2 molécules chargées différemment. Cela favorise le dipôle.



- Coopérativité entropique :
  1) Entropie conformationnelle comme la chaine latérale, rotation, translation
  2) Solvatation : organisation moléculaire autour de la protéine. La perte d'entropie pour une des interactions implique automatiquement une perte d'entropie pour la seconde interaction

2) Entropic cooperativity: $\Delta S(A+B) < \Delta S(A) + \Delta S(B)$

2 types of entropy: conformational (rotation/ translation) and solvation



Loss of entropy for one of the two interactions (rotation, translation, and solvation for hydrophobic residues), automatically implies entropy loss for the second interaction

Little additional entropy loss upon formation of the second interaction.

Transition tout ou rien : Pour les petites protéines, seuls les états initiaux (N et D) sont détectables.

En fait, il existe un état intermédiaire (surtout dans les grosses protéines) car les protéines ne peuvent pas passer comme ça sans ces états. Ces états sont transitoires et ne sont jamais ou très rarement accumulés. La calorimétrie étudie la dénaturation par la chaleur à différents pH.

Critère de van't Hoff :

On the other hand: measure the heat absorbed in the calorimeter:

$$\Delta q_{TOT} / N$$

where Δq_{TOT} is the heat absorbed by the set of N molecules

So:
- when $\Delta H^0 = \Delta q_{TOT}/ N$: denaturation of the whole protein is all-or-none
- when $\Delta H^0 < \Delta q_{TOT}/ N$: denaturation unit is smaller than the whole protein
- when $\Delta H^0 > \Delta q_{TOT}/ N$: denaturation unit is larger than the whole protein – i.e. it is an aggregate of protein molecules rather than a single protein

This is van 't Hoff criterion

This is the hot denaturation (melting). The existence of cold denaturation has been shown (for some proteins - for all? ; sometimes water freezes first, so difficult to show).
- Cold denaturation is also a all-or-none transition.
- Due to the hydrophobic interactions that become less/not favorable at low T

!!! $\Delta G(T) = \Delta H(T) - T\,\Delta S(T)$!!!!!

How does protein folding proceed ?

Levinthal paradox (1967)

Estimation of N, the number of conformations accessible to a protein

Say : L = 150 residues, and counting n ≈ 3 conformations per residue

⇒ N is of the order of: $N = n^L \approx 3^{150} \approx 10^{72}$
(we can take n = 10, the conclusion remains unchanged)

Say: duration of elementary transitions is of the order of $\tau \approx 10^{-12}$ sec

⇒ the time required to sample all conformations is: $t = N\tau \approx 3^{150}\, 10^{-12}$ sec
$\approx 10^{48}$ years

But real folding takes : $t_{real} \approx 1$ msec to 1 sec!

⇒ Apparent paradox

Number of conformations sampled by a natural protein during refolding is of the order of: $N_{real} = t_{real} / \tau \approx 10^9\text{-}10^{12} << 10^{72}$!!

How does protein folding proceed ?

Levinthal calculation is an overestimate :

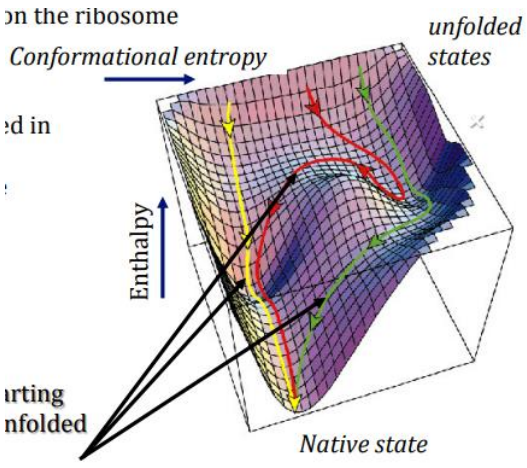
- it is unrealistic that a protein must adopt all possible conformations before finding its native state – more realistically, it can perform a random walk in the conformational space towards its final structure
- ➔ takes less time but it does not solve the paradox: it is too long compared to the time observed.

  • Resolution of the paradox (by Levinthal by itself, in the paper where he presented his paradox) : there are (one or several) folding pathway(s)

  If there are multiple pathways: possibility to have key points where all pathways cross
  For example, formation of certain secondary structures, or some contacts.

Processus de repliement des protéines

- Folding tunnel : l'espace conformationnel est représenté en forme de tunnel. Assure l'indépendance à partir des conditions initiales. Le nombre de conformations est une petite fraction des conformations possibles. On n'explique pas toutes les conformations possibles mais seulement celles qui baissent le plus en enthalpie et donc en énergie.



Etats intermédiaires de repliement : Etat instable qui est transitoire (Car E libre est moindre. Il existe différents chemins par lesquels on passe de différents états intermédiaires. Ils sont éparpillés et rapidement en équilibre avec d'autres structures, donc très difficiles à détecter. Mais, parfois, sous conditions particulières de dénaturation (pH bas, …), il est possible de les accumuler.

Exemple : Molten globule : plus ou moins la bonne structure mais à cause de la présence d'encore quelques molécules de solvant, impossible à voir très précisément et rarement actif.

Détection des états intermédiaires de repliement

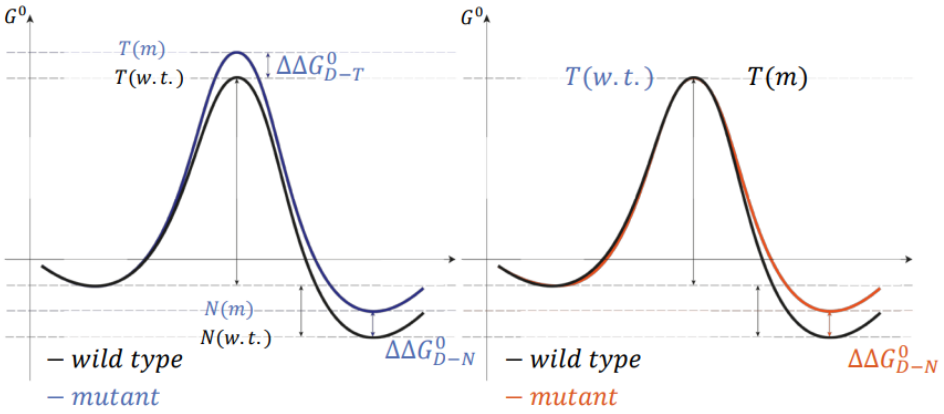
- **Utilisation des ponts disulfures comme sonde** : uniquement pour petites protéines avec des ponts disulfures. Le principe est basé sur l'interaction entre deux cystéines à travers la formation / la cassure des ponts disulfures par un processus redox. Pour cette réaction, il faut des électrons donneur (formation) et accepteur (cassure). En variant la quantité de ceux-ci, on peut évaluer le processus et piéger le plus d'états intermédiaires.
- **Fast mixing techniques** : la protéine est en condition N ou D. Petit à petit, on dilue la solution avec un agent dénaturant (ou l'inverse si en condition D), changeant graduellement le pH ou la T°. Cela piège les états intermédiaires. Caractérisation par NMR, fluorescence via cycles aromatiques, IR spectro, CD ou spectrométrie de masse. Avec les résultats, on peut dire qu'une partie des structures secondaires se forment avant les contacts tertiaires.
- **Fast mixing techniques couplés à échange de H^+** : méthode de marquage qui utilise l'échange de H^+ de la protéine comme une sonde de changement structural. Exploite le fait que la réaction d'échange des protons du groupe amide NH avec les protons du solvant, et le fait que la réaction est plus lente au stade replié qu'au stade dénaturé puisqu'il y'a implication des H (peut-être ponts H) au stade N. Lors du processus de repliement, les protons deviennent résistants à l'échange avec le solvant car participation à une structure secondaire, ou ancré dans le cœur. On utilise du deutérium pour mesurer l'échange avec le solvant après repliement partiel : ratio H/D. On commence donc avec un stade D avec solvant contenant du deutérium, échange H/D entre amide et solvant et dilution de l'agent dénaturant pour que le repliement se fasse, puis ratio H/D. Souvent réalisé après une NMR.
- **Analyse des fragments protéiques, modèles peptides :** on coupe la protéine en petits morceaux et on voit la préférence des petits morceaux à être stables. Si on voit que la plupart ont tendance à former des structures, on peut déduire qu'il va se replier à un stade précoce et former des intermédiaires de repliement très tôt.

Détection des états transitoires de repliement

Etat de transition correspondant à un maximum d'énergie : impossible à piéger, comparé aux états intermédiaires qui eux ont un niveau d'énergie plus bas, donc plus facilement détectables. Ils sont caractérisés à travers la mesure cinétique. Dans ce cas, on prend la protéine sans état intermédiaire.

- **Protein engineering techniques** : fournit l'information sur les interactions formées dans l'état de transition mais seulement pour une interaction à la fois.
  1) On spécifie une interaction pour laquelle on veut savoir si elle fait partie de l'état de transition
  2) On mute un des résidus de cette interaction pour la casser
  3) Dénaturation ou renaturation pour mesurer la stabilité et le taux/ratio de réaction des WT (sauvage) et mutants

  On va donc mesurer la différence d'énergie libre de repliement entre le WT à partir de la constante d'équilibre K



-The effect of the **mutation** on the energy of the **transition state** is defined as:
$\Delta\Delta G^0_{D-T} \coloneqq \Delta G^0_{D-T}$(mutant) - ΔG^0_{D-T}(wild type)

-The effect of the **mutation** on the energy of the **native state** is defined as:
$\Delta\Delta G^0_{D-N} \coloneqq \Delta G^0_{D-N}$(mutant) - ΔG^0_{D-N}(wild type)

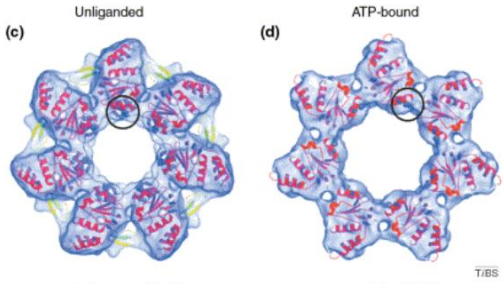
$$\Phi \coloneqq \frac{\Delta\Delta G^0_{D-T}}{\Delta\Delta G^0_{D-N}}$$

Si égal à 1 : la mutation affecte à la fois l'état N et T, l'interaction est donc déjà formée à l'état T.
Si égal à 0 : la mutation n'affecte pas l'état natif, et n'est pas formée à l'état de transition.
Il peut y avoir des valeurs comprises entre 0 et 1, où l'interaction est à moitié formée à l'état T ou interactions faibles.

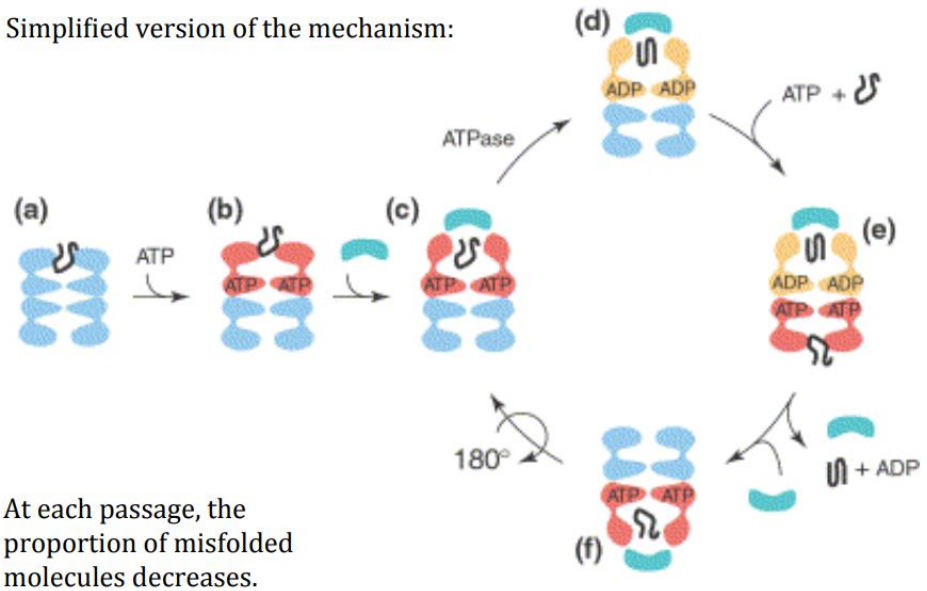
Repliement in vivo : protéines chaperonnes

In vivo, gros challenge pour les protéines de se replier. Elles sont synthétisées sur le ribosome et doivent se replier correctement. L'environnement cellulaire est très dense : on retrouve des interactions avec d'autres molécules, et pas de repliement. Elles peuvent aussi se déplier dû à la fluctuation de la température et la composition chimique du milieu. C'est pour tout cela qu'il y'a eu le développement de protéines chaperonnes.

- GroEL : spécifique aux protéines hydrophobes. Il contient une poche hydrophobe qui attire et accueille des chaines hydrophobes, ce qui déplie totalement la protéine. De l'ATP peut venir se lier à GroEL, initiant un changement conformationnel de la proche GroES qui vient alors se fixer à une extrémité de GroEL. On a alors une expansion de la poche qui devient également hydrophile. La protéine se replie dans cette poche avant d'être libérée.
  Il est possible qu'il y ait plusieurs passages de la protéine dans une chaperonne. A chaque passage, la protéine se replie un peu plus. C'est le cas pour les protéines à domaines multiples qui ont besoin de plus d'aides. En conclusion, bien que la structure finale soit encodée dans la séquence protéique, la protéine peut recourir à une certaine aide, de par ces chaperonnes.



without ATP with ATP



6. Les maladies conformationnelles

Maladies qui sont dues à une perte d'activité protéique, dues à la formation d'agrégats amyloïdes (Alzheimer) ou non amyloïdes.

6.1. Perte de l'activité protéique

Peut aussi impliquer l'accumulation d'agrégats.

- Anémie falciforme : les globules rouges sont anormaux. En effet, on observe une mutation de Glu6 en Valine dans l'hémoglobine. Après désoxygénation du globule rouge, on a une exposition d'une poche hydrophobe qui peut interagir avec la Valine (aa hydrophobe), ce qui conduit à une déformation du GR et son agrégation dans les vaisseaux sanguins.
- Certaines mutations bloquent la protéine p53 qui est un suppresseur de tumeurs. Sa conformation se voit donc inactivée.
- Désordre conformationnel : implique des inhibiteurs de sérines protéases (serpines). Cette famille d'inhibiteurs présente une boucle reconnue par les sérines protéases. Lorsque la protéase se lie à cette boucle et la clive, l'inhibiteur subit un changement de conformation, conduisant au piège de la protéase. Problème possible : agrégation due à l'insertion de la boucle dans une autre serpine.

6.2. Agrégations amyloïdes

Ces fibres sont des structures très organisées, avec un diamètre d'environ 100 A°. Celles-ci sont composées d'au moins 2 fibres de 25 à 35 A° de diamètre qui tournent pour former la fibre mature. Les amyloïdes correspondent à des feuillets Beta qui sont perpendiculaires à l'axe des fibres. Cela est possible pour un grand nombre de protéines. Néanmoins, la propensité à adopter une conformation en feuillet Beta semble être un paramètre important.

- Alzheimer : l'agrégat fibrillaire est composé d'un B-amyloïde peptide de 42 aa, alors que le clivage normal donne 40 aa. Il y'a donc un mauvais clivage.
- Maladie à Prion : agrégation de fibrilles amyloïdes dans le cerveau. Elles sont principalement composées de la protéine Prion qui présente une forme cellulaire (beaucoup d'hélice alpha) et d'une forme amyloïdogénique (beaucoup de feuillets Beta)

6.3. Agrégations non amyloïdes

- Cataracte : agrégations de protéines dans la lentille du cristallin. Différents types de protéines se retrouvent en grand nombre dans cette partie. Ces cristallines sont cassées, mal repliées, oxydées (dû à l'âge), ce qui forme des agrégats opaques.
- Parkinson : agrégats de protéines dans les neurones dopaminergiques

L'habilité à former des fibres amyloïdes dépend entre autres de la séquence, de la charge, de la propensité à adopter une structure secondaire étendue, de l'hydrophobicité. L'état d'une protéine dépend de sa stabilité thermodynamique et de la cinétique de transition entre ces stades. Certaines mutations augmentent l'état d'une population et de la propensité à s'agréger. Durant l'évolution, les systèmes biologiques sont devenus robustes contre l'agrégation.

7. Les différents types d'interactions

7.1. Les interactions protéine-ADN

Les fonctions principales sont la liaison/libération de molécules variées (ADN), ou la transformation chimique, conformationnelle, cinétique de la protéine ou du substrat dans l'espace.
Pour lier l'ADN, une partie de la surface protéique doit être complémentaire à la surface de la double hélice, en prenant compte de la flexibilité. Lorsqu'il y'a complémentarité, la chaine latérale de la protéine pénètre profondément dans l'ADN : reconnaissance spécifique. On obtient donc un complexe ADN-Protéine en équilibre.

$$\text{Protein} + \text{DNA} \leftrightarrow \text{Protein-DNA complex}$$
$$\text{Affinity :} \quad K_{binding} = [\text{protein-DNA}] / [\text{protein}]\,[\text{DNA}]$$
$$\text{Specificity} = K_{binding}\,(\text{target}) / K_{bindng}\,(\text{non-target})$$

Parfois, un cofacteur est nécessaire pour faire passer la protéine d'un état inactif à actif. Exemple avec le TRP-repressor

Contrôle transcriptionnel chez les eucaryotes
La boite TATA est la séquence promotrice générale chez les eucaryotes. Elle est reconnue par TBP, un composant central d'un complexe protéique qui lie différents facteurs pour former le PIC (complexe de pré-initiation).
En quelque sorte, TBP ne reconnait pas la séquence mais la structure caractéristique riche en AT.
Ce complexe de transcription contient :

- TBP
- Des activateurs liant des enhancers, qui sont parfois distants
- Des répresseurs qui se lient à certains sites de l'ADN
- Co-activateurs qui lient des facteurs basaux et activateurs

Les activateurs se lient à l'ADN à des séquences spécifiques. Sa spécificité dépend de la séquence d'ADN. Ils agissent coopérativement, et contiennent une partie de leurs surfaces qui se lient à des facteurs de transcription (augmentation de la spécificité).

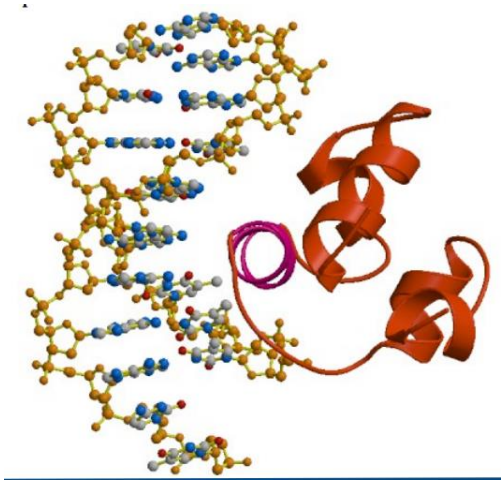
Les forces électrostatiques sont de longue distance ne sont pas très spécifiques. La surface de l'ADN est chargée négativement à cause des groupements phosphates. Cela attire des protéines chargées positivement. Ces forces permettent le rapprochement pour après rendre effectives les forces de plus courte distance :

- Ponts H entre aa et les bases nucléiques
- Cation-pi interaction entre les aa chargés positivement et les bases nucléiques
- Amino-pi entre les aa partiellement chargés et les bases nucléiques
- Pi-pi stacking entre les aa aromatiques et les bases nucléiques

L'ADN est accessible par interactions spécifiques : elles se font à l'intérieur, dans le bas du sillon. Dans le grand sillon, il y'a plus d'opportunités, ce qui offre une meilleure spécificité car plus de ponts H peuvent être réalisés. La spécificité de la liaison détermine quels gènes sont transcrits.

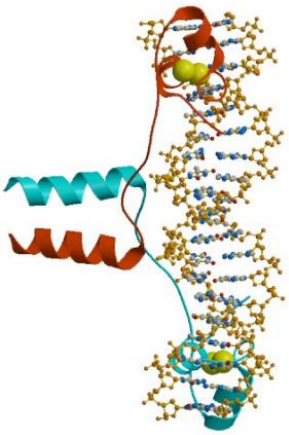
Classification des motifs protéines-ADN

- Domaines HTH : ils régulent la transcription des eucaryotes, impliqués dans la différenciation cellulaire durant le développement des eucaryotes, et jouent le rôle d'empactement de l'ADN avec les histones. Ils possèdent 2 formes caractéristiques : 2 hélices formant un angle de 120°, et souvent connectées par un tournant spécifique de 3 résidus avec une glycine en position 1, où l'angle dihédral est positif. La seconde hélice du motif sert de reconnaissance. Elle sort de la surface de la protéine et entre dans le grand sillon de l'ADN.

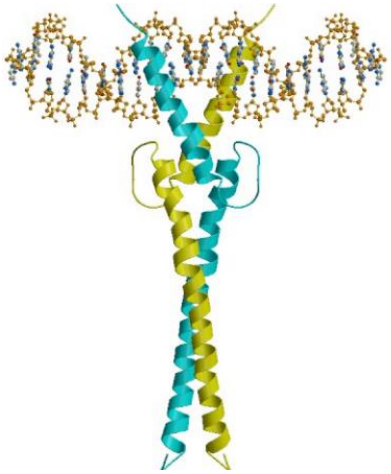


Comme exemple, on retrouve les homéodomaines, qui lient des domaines de facteurs de transcription, et les gènes homéotiques, qui régulent les protéines de manière majeure dans le développement des différentes parties du corps, notamment chez les insectes.

- Les doigts de zinc : dimère de protéines qui reconnait chacun l'ADN de son côté.



- Les leucines zippers : Obligatoirement en dimère. Partie N-terminale est basique et positivement chargée. Elle interagit avec l'ADN et est structurée sur contact avec l'ADN.
  Les hélices des deux monomères entrent dans la rainure principale de chaque côté de l'ADN : ils le tiennent comme une pince.
  Elles servent à reconnaître des séquences palindromiques de l'ADN dans deux rainures majeures successives.



- Les Beta-ribbons groups : Lient l'ADN via des feuillets Beta. Le bras de la protéine entre dans le sillon mineur et s'étend en s'intercalant dans les résidus hydrophobes. L'ADN enveloppe alors la protéine et le feuillet Beta enroule l'ADN. Rôle dans la reconnaissance de l'ADN (TATA box binding sequence,…)

Spécificité de l'interaction ADN-protéine

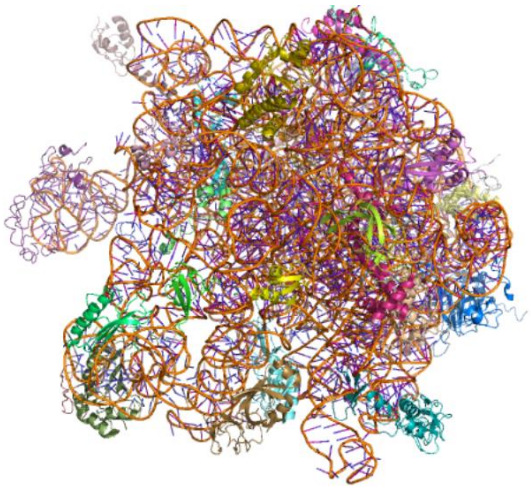
Certaines protéines sont capables de reconnaitre une séquence ADN parmi plus de 1000 autres (spécificité d'un FT est de 10^6). D'autres protéines sont spécifiques de la structure plutôt que de la séquence, comme les protéines de réparation, topoisomérases, histones, HGM1/2 (recombinaison entre deux brins)

Les mécanismes responsables de la spécificité :

1) Niveau 1 : prédisposition intrinsèque d'un motif à reconnaitre l'ADN. Parfois, les protéines ont des contacts adjacents dans le sillon, ce qui augmente l'affinité et la spécificité. Exemple avec les Homéodomaines.
2) Niveau 2 : augmentation de la spécificité en assemblant des motifs variés de liaison à l'ADN. Exemple avec l'association d'une tirette à leucine avec NFAT
3) Niveau 3 : spécificité acquise indirectement quand il y'a une liaison d'une protéine spécifique à une protéine non spécifique

7.2. Interactions protéines-RNA

Exemple avec le ribosome.



Ribosome: translation mRNA -> protein

Complex with thousands of nucleobases and dozens of proteins

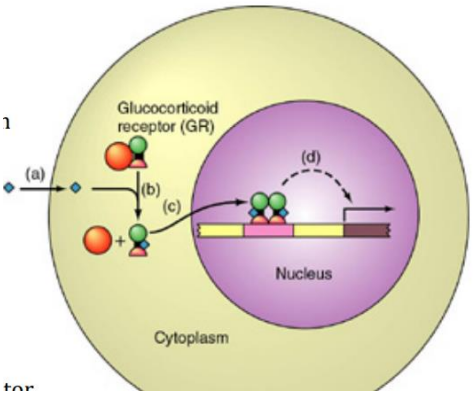
Basic principles identical for protein-DNA and protein-RNA.

But differ because of DNA-RNA differences:

Flexibility ...

7.3. Les interactions protéines-ligand

Un ligand est un peptide ou une petite molécule organique. Son rôle est d'activer/inhiber des protéines récepteurs (modulation de l'activité, drug design). Un exemple est le récepteur aux hormones stéroïdes : il est inactif dans le cytoplasme. Quand l'hormone s'y lie, le complexe va jusqu'au noyau. Là, le domaine en doigt de zinc se lie à des sites spécifiques de l'ADN, ce qui active l'expression du gène et les conséquences qui en découlent : régulation continue des tissus, reproduction etc. Durant l'évolution, des familles de récepteurs se sont liées. Cela suggère que les domaines peuvent agir comme des domaines indépendants et peuvent être interchangés dans la membrane entre membres de la même famille.



8. Structure des protéines : classification

L'interaction entre les différents domaines forme une structure quaternaire. Cela permet une fonction protéique et une interaction entre protéines. En général, une séquence = une structure 3D, qui représente l'énergie libre minimale sauf dans le cas où trop grande barrière d'énergie de transition qui est impossible à franchir.

8.1. Comparaison de structure protéique

Le meilleur outil pour la détection d'homologie est l'alignement de structure. La structure diverge moins que la séquence. L'alignement de structure résidus par résidus est une méthode puissante d'alignement de séquence. Au-delà de 40% de similarité séquentielle, les structures sont très similaires, et ça arrive aussi parfois avec 20%.

Cet alignement permet d'identifier des régions qui ont des structures identiques, les classifier, les grouper en familles. Après superposition des atomes, on mesure le RMSD (Root Mean Square Deviation). Pour cela, on considère certains atomes, on cherche les atomes correspondants dans les 2 protéines, puis on regarde la distance moyenne entre ceux-ci (entre atome 1 prot 1 et atome 1 prot 2). On prend en compte seulement la valeur minimale pour toutes les rotations et translations possibles. Puis, calcul de la superposition où le RMSD est minimisé. Voici la formule :

$$\mathrm{rms} = \underset{\substack{\text{all rotations}\\ \text{and translations}}}{Min} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(x_1^i - x_2^i\right)^2 + \left(y_1^i - y_2^i\right)^2 + \left(z_1^i - z_2^i\right)^2\right)}$$

N et le nombre total d'atomes considérés, le reste sont les coordonnées cartésiennes des différents atomes.

Exemple d'algorithme pour aligner deux structures protéiques

1) On identifie les segments protéiques de conformation similaire. Pour cela on divise la protéine :
    - Soit en segments se chevauchant de N résidus
    - Soit en éléments de structures secondaires
    - Soit en portion d'élément de structures secondaires

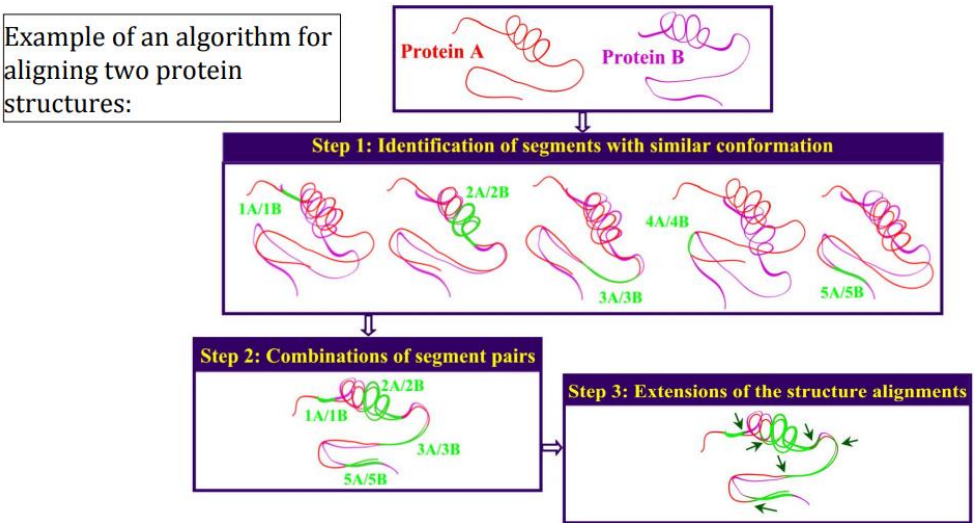
    La similarité des segments est estimée par :

    - Filtre sur la distance fin-fin
    - Filtre sur la distance à partir de la terminaison N-terminale
    - La valeur du seuil de RMSD

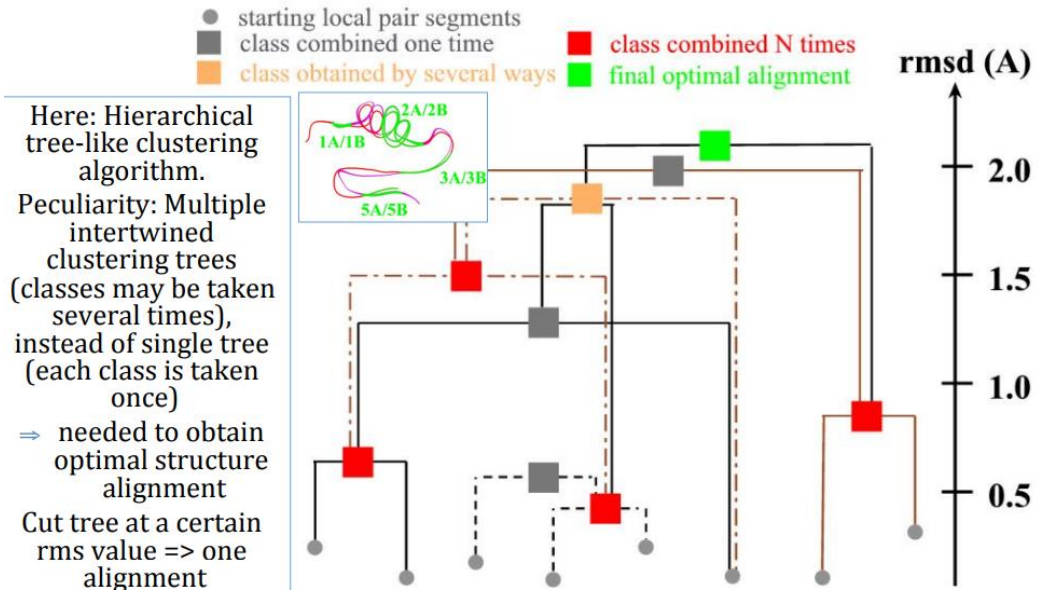
2) Combinaison des paires de segments, similarité de structure globale. On recherche le plus grand nombre de paires de séquences compatibles pour avoir l'alignement recouvrant le plus les protéines. Cela peut être réalisé par un

algorithme hiérarchique de regroupement par arbre. On construit un arbre en mettant des paires de séquences ensemble, on considère que ces paires ne sont plus séparables. On remonte petit à petit. Selon le type d'arbre, une classe/nœud peut être repris plusieurs fois. Selon le treshold défini, on coupe l'arbre à une certaine valeur de RMSD. Finalement, on obtient un alignement.

3) Extension des régions alignées pour améliorer l'alignement. On rajoute 1 résidu à la fois (en N ou C terminal), pour ensuite calculer le RMSD et la superposition avec la valeur la plus basse est gardée.



Step 2: Combination of segment pairs => Global structure similarity



Algorithme DALI

Procède par comparaison de matrice de distances. En effet, chaque conformation est caractérisée par une matrice de distance C alpha – C alpha d_{ij} (en A°). Cet algo est capable de comparer une structure avec toutes les structures du PDB. Après comparaison des résidus semblables dans les deux protéines, on attribue un faible poids aux résidus plus éloignés de la séquence.

En pratique, les matrices sont divisées en sous-matrices de taille fixée. Ensuite, on compare les sous-matrices 2 à 2 pour chaque protéine, pour assembler les paires de sous-matrices pour l'alignement global.

How to compare the different submatrices? Measure of structural similarity

$$Score = \sum_{i \in core} \sum_{j \in core} (\Theta - \Delta(d_{ij}^A, d_{ij}^B))\ \omega(d_{ij}^A, d_{ij}^B)$$

- 'core' corresponds to the set of equivalent residues in proteins A and B
- Δ corresponds to the deviation of the distances d_{ij}^A and d_{ij}^B with respect to their arithmetic mean: $\Delta = \left| d_{ij}^A - d_{ij}^B \right| / \bar{d}_{ij}$
- Θ is a similarity threshold, determined empirically : Θ = 0.2 (to ensure that known structural similarities are recovered
- $\omega = \exp(\bar{d}_{ij}^2 / r^2)$ with r=20Å => gives less weight to longer distances ; 20 Å is the typical size of a protein domain

How to assemble the different submatrices ?

-> Non trivial optimization problem

Requires specific and fast algorithms (cf previous alignment method)

Algorithms used: branch & bound and Monte Carlo (see later)

8.2. Classification des protéines

Les motifs structuraux et certaines organisations spatiales sont retrouvés dans des protéines de séquence différentes. La classification permet de comprendre la structure, la fonction et l'évolution des protéines. Il existe différents niveaux de similarités :

1) Via les structures secondaires : hélices a, 3-10, feuillets beta, …
2) Assignement de domaines/unité de repliement. Classification de domaine. Identification de groupes de résidus de manière à ce que le nombre de contact entre les groupes soit minimal
3) Assignement d'un domaine à une classe structurelle : Tout Alpha ou tout Beta, mix entre les deux, ponts disulfures, ou par metal binding sites
4) Affectation de repliement : défini par le nombre, le type, la connectivité et l'arrangement des structures secondaires. Pour cela, on peut utiliser un programme d'alignement de structure et peut être fait visuellement.
5) Superfamilles : groupes de protéines homologues. Evolution convergente (à bien différencier d'évolution divergente).

Deux protéines sont considérées comme homologues lorsque :

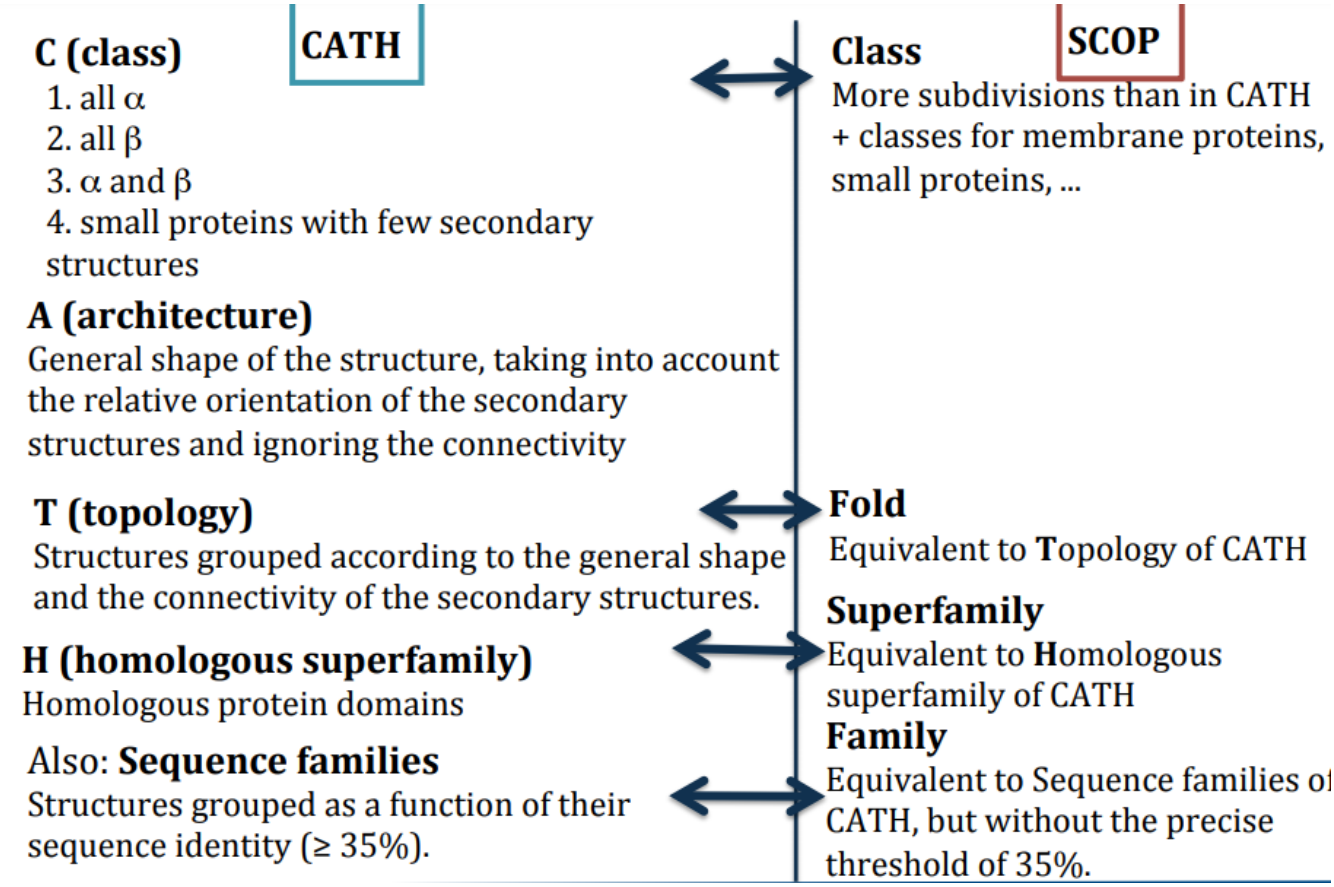
- Elles sont au-dessus d'un certain degré de similarité
- Avec des caractéristiques de structures inhabituelles conservées
- Avec une identité significative (même si elle est basse)
- Avec conservation des résidus clés dans les sites actifs
- Par transitivité : A et B sont homologues, B et C aussi, donc A et C le sont aussi

6) Super repliement et supersites : certains repliements sont peuplés par différentes superfamilles. Cela suggère que le repliement est apparu plusieurs fois par évolution convergente. Les protéines appartenant au même super repliement lient les ligands aux mêmes sites = supersites. Les super repliements semblent dicter la meilleure région de liaison (sans se soucier de l'origine évolutionnaire). Il est possible de prédire les sites de liaison même en absence d'informations sur l'ancêtre commun.

Prédiction de fonction à partir de la similarité structurelle

Pour 50% des nouvelles structures, le site de liaison peut être prédit à partir de comparaison de structures.

Programmes CATH et SCOP



9. Méthode prédiction des structures secondaires

Ces méthodes prédisent la localisation des hélice alpha et feuillets beta à partir de la séquence différente des aa qui ont différentes prédispositions pour certaines structures secondaires.

Elles sont utiles pour :

- Le design de nouvelles protéines : connaitre les règles qui gouvernent la stabilité des hélices et des brins aide à sélectionner des mutants
- Confirmer une relation structurelle et fonctionnelle entre deux protéines quand l'identité de séquence est faible.
- Aider à obtenir la structure 3D à partir des contraintes NMR
- Permettre le raffinement d'un alignement de séquence d'identité faible
- 1ère étape de la prédiction 3D

Il existe de nombreux algorithmes pour prédire la structure secondaire.

9.1. Méthode de prédiction statistique

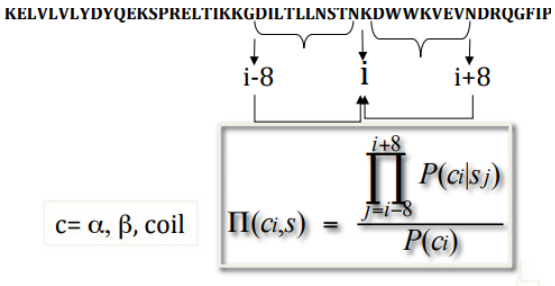
Elle se base sur l'étude d'un dataset de structures de protéines connues (primaires et secondaires) et la recherche de relations statistiques entre ces structures. Elle exploite les données pour calculer la prédisposition d'un acide aminé ou d'un motif à adopter une structure particulière.

L'avantage est que cela est explicite et consistant puisqu'elle exploite une grande database de structures protéiques

Le désavantage est qu'elle ignore les propriétés physiques et a un petit pouvoir explicatif

2 algorithmes :

- **Chou et Fasman** : Compute la prédisposition d'un acide aminé à adopter une conformation. Le score est limité car le résidu n'influence que lui-même et pas l'autre
- **GOR (meilleur)** :
  - Se base sur l'information d'un résidu i sur sa propre structure secondaire
  - Se base aussi sur l'information d'un résidu i sur la structure secondaire d'un résidu j, indépendamment de sa nature
  - Se base sur l'information d'un résidu sur la structure secondaire d'un autre résidu en tenant compte de la nature de l'autre résidu.
  - La probabilité qu'un acide aminé en position i le long de la séquence adopte une structure secondaire c est calculée comme le produit des probabilités conditionnelles d'avoir la structure secondaire c à i sachant que le résidu s occupe la position j (i ± 8)



Score, or folding free energy:

$$\Delta G = -RT \ln \Pi$$

9.2. Méthode de prédiction physico-chimique

Elle se base sur l'observation de protéines connues et sur la connaissance des caractéristiques physiques et chimiques des structures protéiques. Par exemple, les résidus hydrophobes sont enterrés dans le noyau.

1 algorithme :

- **HCA (Analyse de cluster d'hydrophobicité)** : méthode toujours utilisée mais qui ne marche pas forcément toujours. La séquence hydrophobique est représentée comme si c'était helical ou cylindrique. Ce cylindre est coupé parallèlement à son axe et déroulé en un diagramme 2D. Ce diagramme est dupliqué pour restaurer complètement l'environnement de chaque acide aminé. Ces aa hydrophobes forment des clusters, et c'est la forme de ces clusters qui va déterminer le type de structure secondaire (Cluster horizontal = hélice, vertical = feuillet).

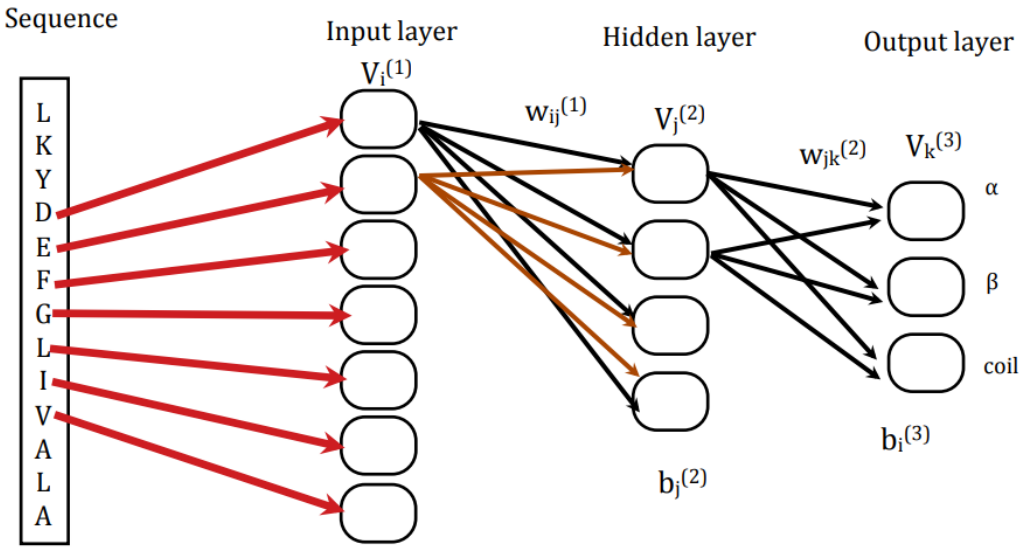
9.3. Méthode hybride et de consensus

Très bonne méthode. Elle utilise différentes méthodes de prédiction et fait un consensus général.

9.4. Méthode de prédiction par apprentissage et intelligence artificielle

1) Apprentissage automatique des propriétés des acides aminés associés aux motifs de structure secondaires. Combinaison de règles, de généralisations, et de méta-règles.
2) **Réseau neuronal (important à comprendre !)**
   - Entrainement sur un set de protéines de structures secondaires connues

- L'algorithme apprend à reconnaitre des patterns complexes dans un set de données d'entrainement. Il apprend aussi à reconnaitre une association de structure secondaire avec la séquence à partir de la protéine dont la structure secondaire est connue. Il trouve le poids et les paramètres qui s'associent le mieux aux inputs et ouputs.
- Il y'a ensuite application pour tester le set : quelle structure secondaire va être prédite. Une fois les poids identifiés, le réseau neuronal peut être utilisé pour prédire la structure secondaire. On retrouve aussi une compilation de l'influence des résidus autour du résidu d'intérêt par ce système. Les valeurs de neurones d'une couche DEPENDENT de la couche précédente.



Generally, a window of 10-17 residues around a central residue, of which we consider the secondary structure, is considered.
-> influence of residues in an environment along the sequence on the structure of a residue
= learning motif ~ as many motifs as residues in the learning set

La valeur de l'input node est par exemple un 20-uples (1,0,0,0,…), (0,1,0,0,…). La node value de la couche suivante est obtenue une fonction des valeurs des nodes de la couche précédente, les poids w et les biais b (initialement random). Ils sont ajustés après que l'output ait été calculé. Quand la phase d'apprentissage est terminée, w et b sont gardés fixes. Ces poids sont utilisés pour calculer l'output du nouvel input.

$$V_j^{(a+1)} = \frac{1}{1+\exp\left(\sum_i w_{ji}^{(a)} V_i^{(a)} + b_j^{(a+1)}\right)}$$

Exemple avec PHF : plusieurs niveaux de programmation (seq-struct et struct-strcut) avec utilisation de plusieurs alignements de séquence comme input du premier réseau neuronal. Il est basé sur le fait que les séquences similaires (25% d'identité) adoptent des repliements et donc des structures secondaires similaires, ce qui apporte plus de précision dans la prédiction.

9.5. Evaluation de la performance de la prédiction

La prédiction de structure n'est pas 100% performante (max 75%) car on considère que la structure secondaire influence la structure 3D mais pas l'inverse. En réalité, c'est faux. La structure secondaire dépend de la structure tertiaire et pour certaines protéines, elles sont même la base de la structure secondaire et la détermine.

- **Résidu score Q :**
  Fraction des résidus correctement prédits dans chaque hélice, feuillet, coil.

$$Q_3 = \frac{q_\alpha + q_\beta + q_c}{N} \times 100$$

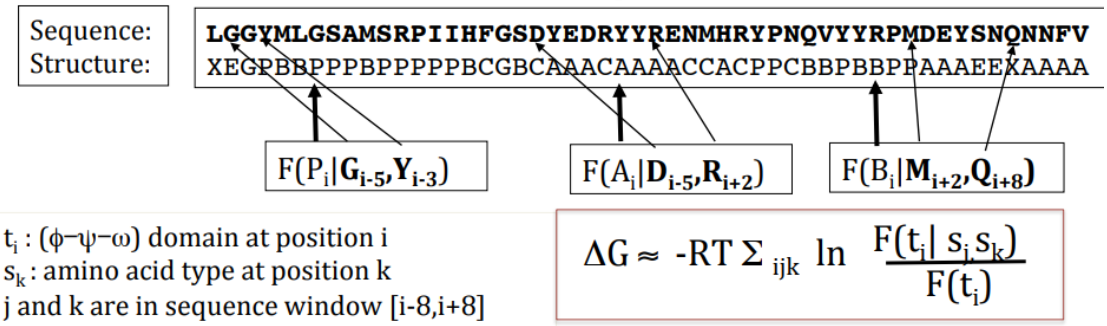
q = le nombre de résidus correctement prédits. Fréquence classique = 32%, 21% et 47%.

- **Score de segment :**
  Fraction d'élément de la structure secondaire prédite avec différents poids.
  Cross validation : on sépare les données en un set d'apprentissage pour identifier les paramètres et les optimiser, et d'un set de test pour évaluer les performances. Une protéine est retirée du set d'apprentissage et est utilisée comme une protéine test. On effectue une répétition pour toutes les protéines du set et on obtient un score moyen. Le meilleur est PHD avec 70%

9.6. Prédiction de structures secondaire locales

Basé sur les angles phi, psy et oméga. Basé sur la prédisposition des paires de résidus à une position J et K à être associés à un certain domaine d'angle phi, psy et oméga en positon i. Il y'a une prédisposition aux potentiels de force moyens. Globalement, cela donne une information un peu plus précise que les méthodes précédentes car donne aussi une information sur la structure tertiaire.
Il y'a 2 algorithmes :

- **Prélude** :
  Calcule les N conformations d'énergie libre la plus basse représentés par une succession de domaines d'angle phi, psy et oméga. La conformation à la différence d'énergie libre la plus basse équivaut à la conformation prédite. On l'utilise pour les petits segments, sinon on retrouve des erreurs de représentation. En général, ça prédit pas mal mais pas toujours la structure secondaire qui a la plupart des critères pour une conformation, mais quelques aa suffisant à en donner un autre.



- **Fugue** :
  Divise la séquence en segments se chevauchant de 5 à 15 résidus et de ces segments, on fait une prédiction Prélude. Il prédit seulement des portions de structures avec de bons scores, et prédit des régions pour lesquelles la structure est intrinsèquement préférée sur base des interactions locales le long de la séquence (intermédiaire de repliement ou fragments adoptant une conformation préférée, bien définie en solution).

10. Les fonctions d'énergie

Un score et une fonction d'énergie sont utilisés pour évaluer la compatibilité entre une séquence et sa structure.
On retrouve deux types de fonction d'énergie :

- Les potentiels semi-empiriques
- Les potentiels dérivés d'une base de données protéiques dont la séquence et la structure sont connues.

On retrouve aussi 2 types d'interactions :

- Les locales entre les résidus qui sont proches le long de la séquence
- Les non-locales entre les résidus distants le long de la séquence mais proches spatialement.

10.1. Les potentiels semi-empiriques

Correspond à une expression mathématique qui décrit les différentes interactions interatomiques. On prend en compte l'énergie d'acides aminés liés ou non de manière covalente.

Exemple d'expression : CHARMM (je ne l'écrirai pas ici!) qui contient différents termes, comme les changements d'angle, l'électrostatisme, ….

$$\boldsymbol{E} = \underbrace{\sum_{bonds} k_i^{bond}(r_i - r_0)^2}_{U_{bond}} + \underbrace{\sum_{angles} k_i^{angle}(\theta_i - \theta_0)^2}_{U_{angle}} +$$

$$\underbrace{\sum_{dihedrals} k_i^{dihe}[1 + \cos(n_i\phi_i + \delta_i)]}_{U_{dihedral}} +$$

$$\underbrace{\sum_i \sum_{j \neq i} 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}}_{U_{nonbond}}$$

Comment construire cette expression ?

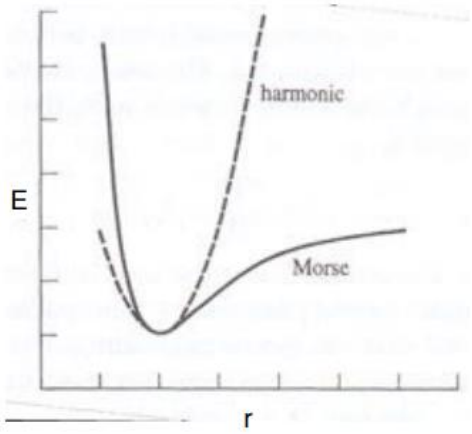
La distance entre deux atomes n'est pas toujours constante, il y'a des variations. On peut représenter cela par un graphique (slide 13) où un optimum de distance est présent (potentiel de Morse). Ce potentiel contient 3 paramètres pour chaque paire d'atomes :

- De
- a
- R0

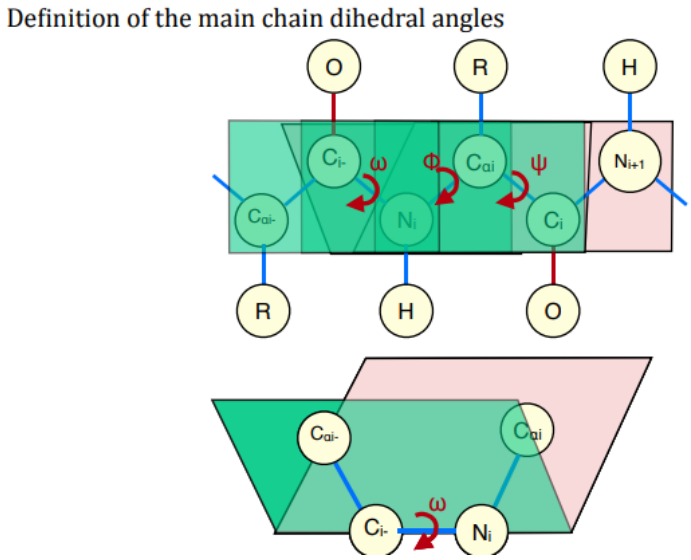
A savoir qu'il n'est pas nécessaire de modéliser à la perfection tous les atomes entre eux, mais juste se concentrer à ce point d'équilibre. Un modèle à large/proche distance est inutile car ça ne se rencontre pas. (En gros, on ne prend que la partie centrale de la courbe de Morse)

Un potentiel harmonique existe donc pour prédire cette courbe. Elle ne contient que 2 paramètres (k et r0): E = k * (r - rO)². CHARMM utilise ce potentiel harmonique. Certains potentiels sont plus adaptés à de l'ADN, ARN, protéines… Il faut donc parfois les optimiser selon le type de molécule.

$E = k\,(r-r_0)^2$



Potentiel des angles dihédraux



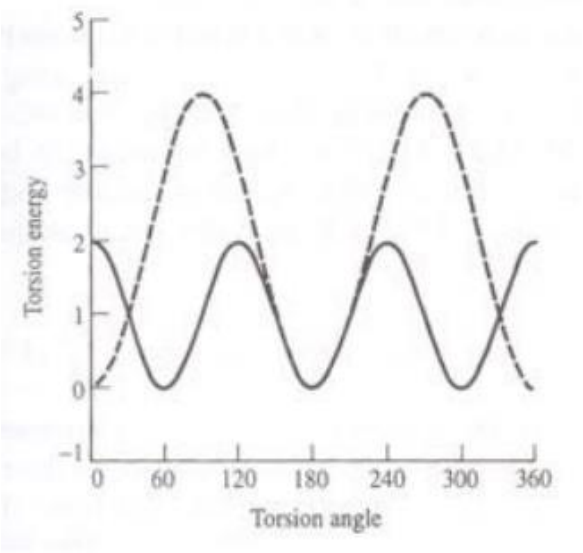
Potentiel de torsion

On a vu précédemment les angle phi, psy et oméga :

- phi: entre 2 plans
- psy: entre N et Carbone alpha
- oméga: entre lien peptidique

On a donc une variation périodique (et donc d'énergie) du fait de ces variations d'angles.

Par exemple, pour l'éthane (C2H6), la variation d'énergie est une fonction périodique. Suivant les rotations des carbones qui induiront des hydrogènes plus proches ou non entre eux (les hydrogènes du carbone 1 avec ceux du carbone 2), l'énergie va donc changer et sera donc périodique. Cela implique une fonction (slide 18) reprenant ce potentiel de fonction.

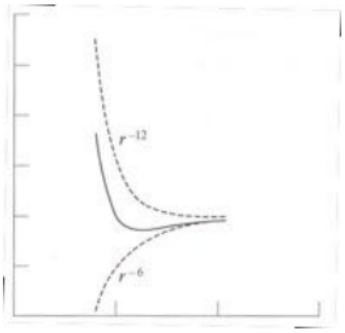


Interactions électrostatiques

Une loi vue précédemment : la loi de Coulomb, reprenant les charges, la constante diélectrique, la distance. Ce n'est pas vraiment un modèle mais une équation physique précise.

Interactions de Van der Waals

Pour ça, on a le potentiel de Lennard-Jones, qui représente la part répulsive (à gauche, positive, qui a été ajustée pour correspondre à la courbe) et attractive (à droite de l'équation, négative) (slide 19). Il y'a 2 paramètres, Epsilon et Sigma



The $1/r^6$ term is an attractive term: this analytical expression has a clear physical justification coming from the average of all the induced dipole-induced dipole geometries.

The $1/r^{12}$ term is the repulsive part of the interaction. This analytical expression has no physical justification. $1/r^{10}$ or $1/r^9$ terms are found in some potentials.

Effet hydrophobe

Dans une protéine, un effet important qui a été vu est l'effet hydrophobe. Le solvant est assez important pour le repliement des protéines. 2 différentes voies sont utilisées pour décrire cela :

- Le modèle explicite. Dans ce modèle, une protéine se retrouve dans une boite virtuelle avec présence d'eau pour atteindre une densité de 1. On analyse ensuite l'énergie du système avec toutes les interactions entre l'eau et les atomes protéiques. L'avantage de cela est qu'on a pas mal de détails au sujet de ces interactions, du système. Par contre, du fait qu'on ajoute beaucoup de molécules d'eau dans ce système, l'analyse est faite à partir d'un grand nombre de données, cela prend donc du temps.

- Le modèle implicite: plus rapide mais moins précis. Le solvant est considéré comme une perturbation. L'effet du solvant est décrit par une expression mathématique (slide 24).

Comment identifier la valeur (value) de ces paramètres ?

Encore deux approches : estimation empirique et approche quantique. Cette dernière est plus précise donc plus lente (On la fait seulement pour quelques parties de la protéine).

<u>Pourquoi semi empirique ?</u> Car on retrouve quelques paramètres empiriques et d'autres par approche quantique. Le transfert de ces paramètres n'est pas toujours possible et quelques potentiels sont optimisés pour certains types de molécules comme l'ADN, et d'autres pour les protéines. Tous ces potentiels (expressions mathématiques) existent déjà et sont à notre disposition

10.2. Les potentiels effectifs

Ils prennent en compte la présence des autres atomes à travers les paramètres. Ce genre de potentiel ne correspond pas à une interaction d'énergie réelle entre 2 atomes isolés mais est paramétrée pour prendre en compte les effets des autres atomes dans l'énergie pairée.
Il y'a computation entre les paires d'atomes des potentiels de VdW et électrostatiques. L'énergie est obtenue en sommant toutes les contributions des paires, mais les interactions des paires sont sous l'influence des autres atomes.

$$\sum_i \sum_{j \neq i} \frac{q_i q_j}{\epsilon r_{ij}}$$

$$\sum_i \sum_{j \neq i} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

10.3. Potentiels dérivés de databases

Ces potentiels utilisent des databases dont les structures et les séquences sont connues.
Il y'a deux approches :

- Formulation analytique qui décrit les interactions et optimise les paramètres de cette fonction en utilisant la Database de structures protéiques connues
- Dérivation de la fréquence d'association entre la séquence et l'élément de structure. Dans le cadre de la mécanique statistique, ces fréquences sont converties en énergies libres.

En mécanique statistique, la probabilité de distribution des conformations d'une molécule obéit à la loi de Boltzmann, qui est la probabilité d'observer une des conformations. Elle est égale à l'exponentielle de moins l'énergie de cette conformation divisée par K x T°, divisée encore par la somme des exponentielles de toutes les conformations d'énergie.

On compte deux étapes :

- Première étape :
    - Faible similarité de séquence < 25% pour éviter les biais
    - Structure bien résolue et raffinée (< 2A°)
    - Grande database pour avoir de bonnes stats

- Deuxième étape :
    - Diviser les séquences et les structures en séquence et l'élément de structure comme par exemple en éléments de séquence (résidus isolés ou par paire). Un élément de structure peut être vu comme des paramètres décrivant la structure (Distance entre les aa, les angles de torsion de la chaine principale, …)

10.3.1. Potentiels de distance (Dérivés de DB)

Le potentiel de l'état de référence représente un stade globulaire où les aa sont identiques, ce qui peut correspondre à un stade dénaturé. W est une énergie libre qui contient les contributions entropiques dues aux moyennes statistiques, à la présence implicite de molécules d'eau et à la discrétisation de l'espace conformationnel. L'équation découle de Boltzmann. ΔW entre l'état qui nous intéresse et l'état de référence. Les probabilités sont comparées à partir de fréquences d'observations (F), de paires d'aa séparées par une distance spatiale comprise entre r_{12} et $r_{12} + \Delta r_{12}$ dans une Database de structures connues.

$$\Delta W^{(2)}(\mathbf{r}_1,\mathbf{r}_2;\, s_1,s_2) = W^{(2)}(\mathbf{r}_1,\mathbf{r}_2;\, s_1,s_2) - w^{(2)}(\mathbf{r}_1,\mathbf{r}_2)$$

$F(r_{12})$ = fréq d'observat° d'1 dist de séparation 1-2
↳ à quelle fréq on observe S_1, S_2 à cette distance r_{12}

Avec $F(r_{12} | S_1, S_2) = \frac{F(r_{12}, S_1, S_2)}{F(S_1, S_2)}$

↳ fréq d'associat° de la paire des 2 aa qqsoit la dl qui les sépare

On se base donc sur des observations faites dans notre Database. De ces fréquences, on obtient un paramètre énergétique.

En pratique :

- Les distances entre C^a , C^b ou chaines latérales sont computées. Il est aussi possible de les computer entre atomes
- Les distances sont divisées en domaines (par exemple 0.2A° de large) ou peuvent correspondre à Contact/Pas contact
- Il est possible d'adoucir les potentiels en combinant les fréquences computées sur les domaines voisins.
- Il est possible de computer directement les fréquences pour les aa séparés par 2 ou 8 aa le long de la séquence et pour les aa séparés de plus de 8 aa. (Potentiels locaux et non locaux).

Ces potentiels vont donc dépendre de la taille moyenne des protéines dans la Database.

10.3.2. Potentiels de torsion (Dérivés de DB)

Les éléments de structure sont les domaines d'angle de torsion des chaines principales.

- Type 1 : influence d'une paire de résidus sur un domaine de torsion.
- Type 2 : influence d'un résidu sur une paire de domaines de torsion.

Certains acides aminés préfèrent se retrouver sous un angle particulier.

$F(t_i|s_j)$, $F(t_i,t_j|s_k)$, $F(t_i|s_j,s_k)$ is computed (t: torsion domain, s: amino acid type)

Type 1: influence of a residue pair on a torsion domain
$i-8 \le j \le k \le i+8$:

$$\Delta W^{(2)} \cong \sum_{i,j=1}^{N} \Delta W^{(2)}(t_i; s_j,s_k) \cong -kT \sum_{i,j=1}^{N} \frac{1}{\zeta_i} \ln \frac{F(t_i|s_j,s_k)}{F(t_i)}$$

ζ is a normalization factor, to count each contribution one time.

Type 2: influence of a residue on a pair of torsion domains

$$\Delta W^{(2)} \cong \sum_{i,j=1}^{N} \Delta W^{(2)}(t_i,t_j;s_k) \cong -kT \sum_{i,j=1}^{N} \frac{1}{\zeta_i} \ln \frac{F(t_i,t_j|s_k)}{F(t_i,t_j)}$$

10.3.3. Potentiels d'hydrophobicité (Dérivé de DB)

Dans ce potentiel, l'accessibilité du solvant de chaque résidu est calculée et la fréquence d'association entre les résidus et les domaines d'accessibilité au solvant est computée.

In this potential, the solvent accessibility of each residue is calculated and the frequency of association between residues and solvent accessibility domains is computed.

$$\Delta W^{(2)} \cong \sum_{i=1}^{N} \Delta W^{(2)}(h_i;s_i) \cong -kT \sum_{i=1}^{N} \ln \frac{F(h_i|s_i)}{F(h_i)}$$

10.4. Comparaison entre potentiels semi-empiriques et potentiels dérivés de Databases

Semi-empiriques	**Dérivés de Databases**
Avantage	Avantages
Interactions bien définies avec un background physique clair	Protéines simplifiées, représentation possible
	Prend en compte l'entropie et le solvant (de manière implicite)
Désavantage	Désavantages
Effet du solvant et de l'entropie	La contribution des différentes interactions est moins évidente
	Dépend de certaines caractéristiques de la database

10.5. Evaluation des performances de la fonction d'énergie pour la prédiction de structures

Il faut que la fonction d'énergie discrimine la structure native de toutes les structures non natives. La structure native doit correspondre au minimum d'énergie.
Les performances sont évaluées à partir des sets de decoy.
Un **set de decoy** est comme un set de protéines dont la structure est native ou non. Elles sont créées in silico :

- Via simulation : conformation obtenue durant la trajectoire de repliement
- Via modélisation comparative : decoy = mauvaise prédiction
- Via échange de séquences : séquence montée sur la structure d'une autre protéine.

Un bon Decoy set :

- Doit contenir des conformations proches de la structure native
- Doit contenir la structure de différentes protéines appartenant à différentes classes
- Doit contenir un grand nombre de structures
- Doit contenir des structures représentatives des différentes régions de l'espace conformationnel.

Une bonne fonction d'énergie :

- La structure native est computée avec l'énergie la plus basse
- Doit être capable de discriminer la structure native des decoys
- L'énergie des structures non natives augmente quand la similarité structurelle avec la structure native diminue

11. Prédiction de la structure 3D des protéines

On utilise la fonction d'énergie ou de score pour discriminer les différentes conformations possibles. Il existe un grand nombre de conformations possibles mais en général, les protéines vont adopter une conformation qui correspond à l'énergie minimale.

Espace conformationnel

Hypersurface multidimensionnelles obtenue en plottant l'énergie libre vs les coordonnées décrivant la conformation du système. L'espace conformationnel des protéines contient un grand nombre de minima.

11.1. Modélisation comparative (HHPred)

Elle se base sur la modélisation d'une structure protéique à partir d'une protéine résolue expérimentalement.

Principe :

L'espace des conformations possibles est plus petit que l'espace des séquences possibles. En effet, des séquences similaires adoptent des structures similaires. Premièrement, on recherche dans une base de données (BLAST) avec une protéine dont la structure a été résolue pour une protéine partageant 30 à 40% d'identité de leur séquence avec une protéine cible. Ces structures vont être les templates, et le choix va dépendre de l'identité, de la qualité du template, des conditions expérimentales utilisées pour obtenir la structure du template. L'alignement est crucial.

1) **Modélisation de la chaine principale** :

   Il y'a deux approches.

   - <u>Assemblage de rigid bodies</u>

     Correspond aux régions de séquences les plus conservées entre le template et la target. Si plusieurs templates sont sélectionnés et superposés, les régions similaires selon la structure vont correspondre aux rigid bodies. L'alignement de séquence avec la target va permettre d'assigner des rigid bodies à la cible.

   - <u>Modélisation par satisfaction de contrainte</u>

     Ces contraintes proviennent des structures des templates sélectionnées (longueur des liens, valeurs des angles, distance interatomique, …). L'alignement Template-Target est ensuite utilisé pour transférer ces contraintes du Template à la Target. On peut donc optimiser ces contraintes de sources différentes.

2) **Modélisation des boucles** :

   Etape difficile car tout n'est pas toujours obtenu expérimentalement (Xray, …). Il faut aussi tenir compte de la flexibilité, etc. De plus, il est difficile de modéliser une boucle des plus de 8 acides aminés

   Technique :

   - Chercher dans une Database de boucles connues pour des boucles qui ont une longueur et une distance entre les extrémités comparables avec notre modèle
   - Ab initio : utilisation de la mécanique moléculaire, dynamique moléculaire, méthode de Monte Carlo

3) **Modélisation des chaines latérales**
   Nécessite une fonction d'énergie pour sélectionner la meilleure solution.
   Les chaines latérales des différents résidus dans le template comparé à la cible sont modélisées.
   Les chaines latérales des résidus identiques dans le template et dans la cible sont modélisés à nouveau.
   Pour ce faire, on utilise une bibliothèque de rotamère : utilisation de la chaine latérale dont la conformation provient de protéines connues. Mais la chaine principale limite en général les degrés de liberté de la chaine latérale.

4) **Erreurs dans la modélisation comparative**
   Les erreurs vont dépendre du pourcentage d'identité de séquence. Si celle-ci est supérieure à 90%, il est possible que le modèle soit aussi bon qu'en XRay ! Si celle-ci se situe entre 50 et 90%, le RMSD global entre la structure prédite et la structure réelle peut être aux environ de 1.5A° (Parfois plus localement). En dessous de 25%, le plus compliqué est l'alignement de séquence. Et des erreurs là-dedans peuvent avoir de grosses répercussions sur la prédiction. Notons aussi des erreurs dans l'alignement, dans l'empaquetage des chaines latérales (embêtant si dans des régions fonctionnelles de la protéines), lorsqu'il n'y a pas de templates, ou un mauvais template (>25%), etc.

11.2. Fold recognition (Sparkt)

Dans le cas où il n'y a pas un partage d'une bonne identité de séquence avec notre target. On va considérer que la structure de notre cible peut être obtenue à partir d'un repliement connu.
On attribue notre séquence cible à la structure provenant d'une bibliothèque de repliement connus.
On compute l'énergie de toutes les associations séquence-structure, et on garde l'association pour laquelle l'énergie est la plus basse.

11.3. Validation

1) **Qualité d'un modèle : évaluation**
   - Sur base de la stéréochimie : longueur des liens, valeurs des angles, planéité des cycles aromatiques, chiralité, angles de torsion de la chaine principale. Tout ceci est le rôle du programme Procheck
   - Méthodes de profil 3D ou potentiels statistiques : l'environnement d'un résidu est comparé à celui trouvé dans une structure résolue expérimentalement (Xray)

2) **Utilisation :**
   - Dans le design d'un ligand qui lie le site actif. Doit être défini avec une erreur inférieure à 1A°.
   - Rationalisation de l'effet des mutations (encore définie avec <1A°)
   - Caractérisation des surfaces de liaisons

- Analyse des propriétés de surface

3) Performance d'une méthode de prédiction

- CASP, qui évalue aveuglément les performances des méthodes de prédiction
- Modélisation basée sur un template : les cibles ont des similarités de structure évolutionnaire avec des structures obtenues expérimentalement
- Modélisation libre : ab initio

11.4. Prédiction ab initio

C'est une prédiction à partir d'une séquence seule. Cette méthode est plus difficile est prend plus de temps. On recherche la structure native parmi toutes les structures possibles ou le long d'une voie de repliement

Les étapes :

- Choix d'une représentation structurale simplifiée. (1)
- Choix d'un algorithme pour la recherche dans l'espace conformationnel. (2)
- Choix de la fonction d'énergie pour évaluer la comptabilité séquence-structure, qui doit être adaptée au niveau de simplification. (3)

Concept de simplification

On considère que malgré la simplification de la structure, il est possible de trouver une structure proche de la structure native. La solution n'est pas forcément celle la plus basse, mais cette énergie reste toutefois minimale. Pour raffiner le modèle, on peut partir d'une structure à la première étape et simuler une représentation plus détaillée.

1) Choix d'une représentation structurale simplifiée

3 modèles :

- Modèle détaillé avec tous les atomes. Très précis mais prend beaucoup de temps
- Modèle très simplifié avec comme principe : un résidu : 1 point. Rapide mais peu précis
- Modèle intermédiaire : résidus représentés par la chaine principale et par un atome ou pseudo-atome représentant la chaine latérale. Généralement, on néglige les degrés de liberté des chaines latérales.

Représentation des conformations :

- Coordonnées cartésiennes de tous les atomes considérés, ou des points représentant les acides aminés. Le désavantage est qu'on ignore la chaine polypeptidique.
- Distance entre les atomes/résidus. La difficulté est que si la distance entre N atomes ou résidus est spécifiée aléatoirement, il n'est pas possible de faire une représentation en 3D.
- Les valeurs des angles de torsion de la chaine principale. La difficulté est que le changement d'angle d'un résidu entraine un mouvement de l'entièreté de la région de la chaine qui le suit. (Clash stérique)
- Matrice régulière : cubique, tétraèdre. La difficulté est qu'il est impossible de représenter une structure réelle sur une matrice régulière sauf si l'espace de matrice est réduit mais dans ce cas, on perd l'intérêt de la matrice

2) Choix d'un algorithme pour la recherche dans l'espace conformationnel

a) Dynamique moléculaire

Elle est basée sur l'équation de Newton F=m.a. C'est une résolution numérique qui prend en compte les conditions initiales + celles aux instants suivants : la position, trajectoire, etc.

Elle représente tous les atomes et les molécules de solvant. Le mouvement est très petit à chaque étape pour l'exploration de l'espace. En gros, il simule le repliement de la chaine polypeptidique à partir d'un coil random, mais n'est pas optimal.

b) Recherche systématique

Elle explore par changement régulier et prédictible de la conformation. Regarde toutes les conformations puis choisit l'énergie la plus basse. On a une explosion de combinaisons possibles même par petit nombre de résidus.

c) Amélioration de la recherche systématique

Il y'a une élimination de l'étape de minimisation de l'énergie pour les structures qui ont une énergie très haute. Cela permet à l'énergie semi-empirique de devenir très favorable.

d) Approche de construction de modèles

On divise la protéine en fragments et on prédit la structure ou les structures pour ces fragments qui ont déjà une structure 3D ou secondaire. On assemble ces fragments pour obtenir une structure 3D globale.

- La conformation des fragments est indépendante des autres. Néanmoins, il faut faire attention car l'interaction tertiaire affecte la structure secondaire.
- La solution est de garder plusieurs conformations pour chaque fragment et de construire différentes structures globales. On déduit que chaque groupe de conformation pour chaque fragment contient au moins la conformation native.

e) <u>Exploration aléatoire</u>

C'est la méthode la plus utilisée. Elle est opposée à la recherche systématique qui explore la surface d'énergie de manière prévisible. Ici, il n'y a pas d'ordre dans la recherche aléatoire. En une étape, on peut passer d'une région dans l'espace conformationnel à une autre région complètement différente. Différents mouvements sont appliqués durant la recherche :

- Dans les coordonnées cartésiennes : sélection d'un ou plusieurs résidus et ajoute une quantité random à ses coordonnées x,y,z

- Dans l'espace des angles de torsion de la chaine principale : on choisit de manière random un ou plusieurs angles de torsion et on ajoute une quantité random à l'angle, ou on transfère un angle de torsion à un autre.

- Processus similaire dans l'espace de distance excepté que la structure doit être possible.

Au niveau des étapes :

1) Conformation initiale
2) Conformation initiale d'itération
3) Génération de la nouvelle conformation par modification aléatoire
4) Estimation de l'énergie
5) Fin ? Impossible d'être certain d'avoir atteint l'énergie minimale car il existe des minima locaux
6) Si non, détermination de la <u>structure initiale de la prochaine génération</u>. On refait plusieurs fois pour voir si on obtient le même résultat. On compare tout ça et on garde les structures de plus basse énergie.

Il existe de nombreuses manières de <u>choisir la conformation pour la génération suivante</u>. On peut soit choisir une structure générée à l'étape précédente, soit on sélectionne aléatoirement une structure précédente en biaisant le choix vers celles qui ont été sélectionnées le moins souvent (protocole d'usage uniforme). On peut aussi choisir la structure d'énergie libre

la plus basse générée jusqu'ici, ou les biais vers structures énergétiques basses (critère de Metropolis). Toutes ces procédures se valent.

Algorithme de Monte Carlo (exploration aléatoire)

A chaque itération, une nouvelle conformation est générée en faisant une modification aléatoire, par exemple dans le cas des coordonnées cartésiennes où on modifie un résidu en utilisant un générateur de nombre aléatoire. Ensuite, on calcule la nouvelle énergie libre.

Critère de Metropolis : la nouvelle conformation est gardée comme un point initial pour l'itération si l'énergie libre est plus basse. Si pas, il y'aura quand même une probabilité de la garder, tout dépend de la valeur de cette énergie. Ceci permet de passer une barrière d'énergie (dans le cas d'un minimum local par exemple) pour atteindre un autre minimum. Cette probabilité se fait en fonction de Boltzmann :

B = exp (-ΔΔW/RT)

Ça dépend donc de la température. Si elle est faible, on aura peu d'explorations. Si elle est élevée, on aura beaucoup d'explorations. Cette température peut varier au fur et à mesure. Par exemple, elle peut être élevée au début, ce qui favorise l'exploration et le passage de barrière d'énergie. Ensuite, la température diminue pour explorer en détails.

Cet algorithme génère des états en chaines de Markov :

- Le résultat de chaque itération dépend de l'étape précédente
- Seulement un nombre fini de résultats possible pour chaque essai

Algorithmes évolutionnaires (exploration aléatoire)

Groupe de méthodes d'exploration dans l'espace conformationnel pour trouver la solution optimale. Ceci est basé sur l'idée de l'évolution biologique.

3 classes :

- Algo génétique
- Programmation évolutionnaire
- Stratégie d'évolution

L'idée est de créer une population de solutions possibles. Les membres de la population sont évalués en utilisant une fonction d'énergie/de score/de coût mesurant leur qualité. La population change au cours du temps pour évoluer vers de meilleures solutions.

Algo génétique

1) On crée une population parent pour N conformations possibles. Ensuite, on calcule le score de chaque membre.
2) On crée après cela une nouvelle population où par exemple 50% des parents sont sélectionnés aléatoirement avec un bais/préférence pour les parents avec l'énergie la plus haute. Cette nouvelle population est sujette à de nombreux opérateurs génétiques : mutations, recombinaison homologue. Elle devient alors la nouvelle population parent.

Programmation évolutionnaire

Similaire à l'algo génétique mais on n'effectue pas de recombinaison homologue

Stratégie d'évolution

Ici, les recombinaisons sont permises. On sélectionne les meilleurs individus via un classement selon leur score d'énergie. C'est pour de l'optimisation globale mais ils contiennent un élément significatif. Souvent, la solution est proche du minimum global avec un temps de calcul pas trop long.

Utiliser des idées de la biologie : colonie de fourmis

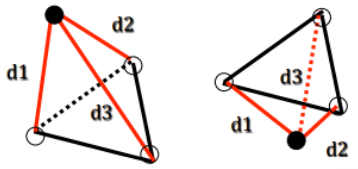
Imitation des fourmis. On explore l'espace et on recherche le chemin le plus court. Celui-ci aura le plus de phéromones et par conséquent, il sera le plus utilisé par les fourmis suivantes qui le renforceront.

- On établit une colonie artificielle. Chaque fourmi construit une partie de la conformation de la protéine.
- Dépôt de phéromones sur la meilleure solution (évaluation via fonction d'énergie)
- Création d'une nouvelle génération de fourmis artificielles qui construisent une nouvelle structure en prenant compte le dépôt de phéromones.

Géométrie de distance

***Ab initio* structure prediction - Distance geometry**

- Another way to describe the conformations of a molecule: in terms of distances between all pairs of atoms/residues.
- For N atoms/residues: N (N-1) / 2 distances => symmetric NxN matrix with 0 on the diagonal.
- Distance geometry explores the conformational space distances by generating a large number of distance matrices, which are then converted into conformations represented by Cartesian coordinates.
- But: inter-residue/atomic distances are correlated - many combination of distances are geometrically impossible.
  For example: for a triangle ABC, | AB | + | BC | ≥ | AC |



The position of each atom is determined by its distance to three other (non collinear) atoms – up to a reflection indeterminacy

On explore l'espace conformationnel en générant des matrices de distances qui sont converties en conformations représentées par coordonnées cartésiennes. Mais les distances entre les résidus sont corrélées et beaucoup de combinaison de distance son géométriquement impossibles.

1) 2 matrices (calculs à partir de principes chimiques simples) :
   - Calcul des limites basses de distance entre atomes
   - Calcul des limites hautes de distance entre atomes
2) Assignation aléatoire à chaque distance entre atomes entre limites hautes et basses
3) La matrice de distance est convertie en test set de coordonnées cartésiennes pour N atomes (processus appelé embedding = incorporation). Ensuite, recherche de structures 3D les plus compatibles avec les distances, et on incorpore les structures de N-1 dimensions à 3 dimensions
4) Raffinement des coordonnées

Cette technique est utilisée en NMR, qui fournit des distances de contraintes entre atomes, avec une certaine marge expérimentale : on recherche des structures 3D qui sont compatibles avec les contraintes de distances expérimentales, en prenant en compte les erreurs de marge.

Embedding

From the NxN distance matrix, composed of all distances between residues/atoms i and j (d_{ij}), compute the NxN real symmetric metric matrix :

$$G_{ij} = \tfrac{1}{2} (d_{io}^2 + d_{jo}^2 - d_{ij}^2)$$

The origin o is generally taken as one of the residues/atoms near the center of the molecule.

Diagonalize **G** => $\mathbf{G = V \Lambda V^{-1}}$; **V** et $\mathbf{\Lambda}$ are NxN matrices; $\mathbf{\Lambda}$ is diagonal.

G is square symmetric => $\mathbf{G = G^T}$ => $\mathbf{V^{-1} = V^T}$ => $\mathbf{G = V \Lambda V^T}$ (All real symmetric matrices are diagonalizable)

Express $\mathbf{\Lambda = L^2}$ => $\mathbf{G = (V L)(V L)^T = X X^T}$ where $\mathbf{X = V L}$ is an NxN matrix

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \lambda_N \end{pmatrix}$$

⇒ **X** "contains" the N atomic coordinates.

What does that mean?

In general, the conformation cannot be exactly embedded in 3D. The « best » 3D structure compatible with the distances d_{ij} is obtained from the eigenvectors that correspond to the three largest eigenvalues Λ, denoted $\lambda_1, \lambda_2, \lambda_3$

Ab initio structure prediction - Distance geometry

Embedding

$$X = VL = \begin{pmatrix} V_{11}\lambda_1^{1/2} & V_{12}\lambda_2^{1/2} & V_{13}\lambda_3^{1/2} & V_{14}\lambda_4^{1/2} & \dots & V_{1N}\lambda_N^{1/2} \\ V_{21}\lambda_1^{1/2} & V_{22}\lambda_2^{1/2} & V_{23}\lambda_3^{1/2} & V_{23}\lambda_3^{1/2} & \dots & V_{2N}\lambda_N^{1/2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ V_{N1}\lambda_1^{1/2} & V_{N2}\lambda_2^{1/2} & V_{N3}\lambda_3^{1/2} & V_{N3}\lambda_3^{1/2} & \dots & V_{NN}\lambda_N^{1/2} \end{pmatrix}$$

⇒ Coordinates of each residue/atom i are: $(\lambda_1^{1/2} V_{i1}, \lambda_2^{1/2} V_{i2}, \lambda_3^{1/2} V_{i3})$

⇒ Best way of embedding an object living in a space in N-1 dimensions to a space in 3 dimensions.

cf Principal component analysis –

Ab initio structure prediction - Distance geometry

Embedding

Check

If the object/conformation is already compatible with 3 dimensions:

$G_{ij} = \tfrac{1}{2} (d_{io}^2 + d_{jo}^2 - d_{ij}^2) = (x_i - x_0)(x_j - x_0) + (y_i - y_0)(y_j - y_0) + (z_i - z_0)(z_j - z_0) = \mathbf{i.j}$

where **i** and **j** are vectors linking the origin o to the residues/atoms i et j:

$\mathbf{i} = (x_i - x_0, y_i - y_0, z_i - z_0)$, $\mathbf{j} = (x_j - x_0, y_j - y_0, z_j - z_0)$

Diagonalize: $\mathbf{G = V \Lambda V^{-1} = V \Lambda V^T}$

Here $\mathbf{\Lambda}$ is the identity matrix and $\mathbf{X = V}$

$$\mathbf{X} = \begin{pmatrix} x_1 - x_0 & y_1 - y_0 & z_1 - z_0 & 0 & \dots & 0 & 0 \\ x_2 - x_0 & y_2 - y_0 & z_2 - z_0 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & 0 & 0 \\ x_N - x_0 & y_N - y_0 & z_N - z_0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

Evaluation des méthodes de prédiction

- En prédisant des structures déjà connues avec ces méthodes
- En faisant de la validation croisée mais embêtant car chaque prédiction prend beaucoup de temps

12. Les protéines membranaires

Elles sont associées à un grand nombre de fonction importantes dans la cellule. 30% du génome donne ces protéines. Malheureusement, peu de ces protéines ont été résolues. Ces protéines peuvent être hélicoïdale (hélices alpha) ou brins Beta.
Il existe :

1) Des protéines membranaires intégrales : elles sont insérées totalement et définitivement d'un côté de la membrane
2) Des protéines membranaires périphériques : elles sont attachées partiellement à la membrane ou aux protéines membranaires intégrales
3) Toxines polypeptidiques : elles sont solubles mais peuvent s'associer à la membrane et former des canaux.

12.1. Protéines membranaires hélicoïdales (helix bundles)

Elles sont plus faciles à prédire car les segments sont plus petits et parfois moins hydrophobes. Après sa synthèse, la protéine membranaire est transférée dans le bicouche lipidique via le translocon. Une fois insérée, l'hélice forme une structure compacte composée principalement de longs segments apolaires. Plusieurs segments sont incorporés dans la membrane. En général, quand une hélice entre d'un côté, elle ressort de l'autre. La plupart des segments sont perpendiculaires à la membrane et font 15 à 30 résidus. La plupart des outils de prédiction reposent sur des règles dérivées de structures expérimentales. Mais le nombre de structures 3D de protéines membranaires est bas.

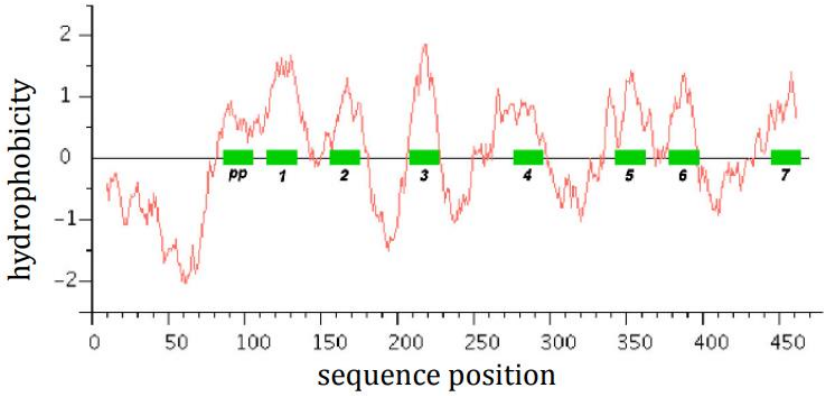
12.1.1 Paramètres utilisés par les outils de prédiction

1) Règles empiriques

- Les acides aminés positifs sont généralement intracellulaires. Ils sont donc à limite des régions apolaires et peuvent être utilisés pour identifier la topologie de l'insertion dans la membrane. Cependant, il y'a parfois exception à la règle
- Les hélices transmembranaires sont principalement apolaires et composées de 15 à 30 acides aminés.

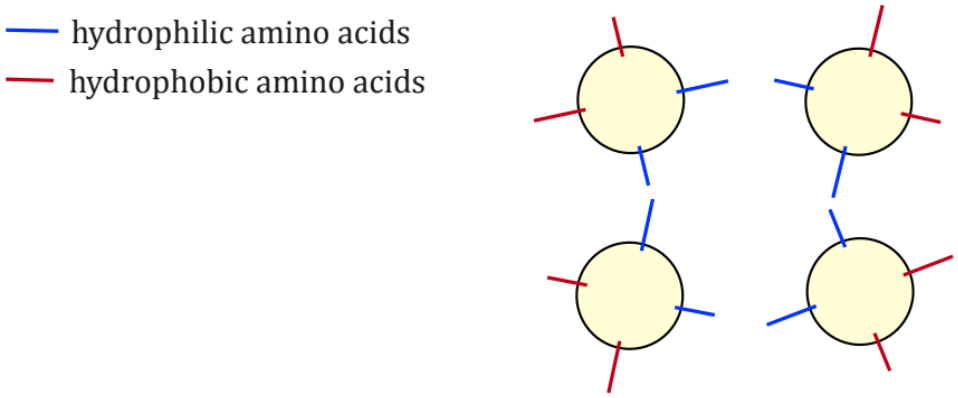
2) Plot d'hydrophobicité

Utilisation d'une échelle d'hydrophobicité et identification des hélices transmembranaires via l'hydrophobicité des acides aminés le long de la séquence.



3) Analyse du moment d'hydrophobicité

Calcul des hydrophobicités de faces. Les segments hélicoïdaux sont amphiphiles. On mesure la distribution des chaines latérales hydrophobes. Les hélices tendent à se regrouper de manière à exposer les résidus hydrophiles l'un vers l'autre (vers l'intérieur) et les résidus hydrophobes vers l'extérieur de la membrane



12.1.2 Programmes de prédiction (Protéines membranaires hélicoïdales)

1) TopPred

- Il compute le plot d'hydrophobiscité
- Il définit un seuil C1 au-delà duquel on est certains d'un segment transmembranaire
- Il construit toutes les topologies possibles incluant les fragments >C1 et incluant ou excluant les fragments < C1 mais > à C2
- Il évalue la différence entre le nombre d'ARG et LYS extracellulaire pour chaque topologie, et ne prend pas en compte les longues boucles
- Il choisit la topologie avec la différence la plus large

- Si grand nombre de longues boucles, leurs acides aminés peuvent indiquer leurs localisations intra ou extracellulaires

Evaluation de l'hydrophobicité :

- Soit par la mesure d'un coefficient de partage entre les phases polaires et apolaires. On mesure la prédisposition de chaque acide aminé à être concentré dans une phase que dans l'autre -> Echelle d'hydrophobicité pour chaque acide aminé.
- Soit à partir de protéines transmembranaires connues. On regarde les acides aminés qui ont tendance à se retrouver dans le cœur ou en surface du solvant.

2) TMHMM

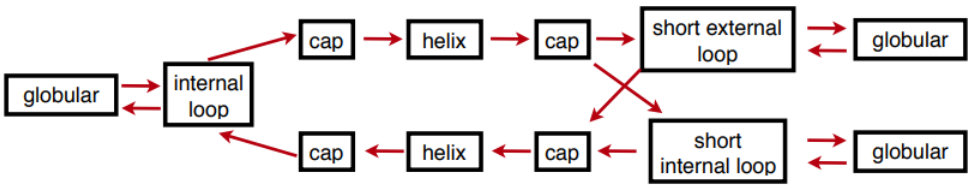
Repose sur le modèle de markov caché (HMM). Le programme Phobius prédit les peptides signaux et les segments transmembranaires. HMM inclut des informations évolutionnaires et des résultats expérimentaux. Il peut aussi prendre en compte des contraintes.

- Définition des différents états, chaque résidu peut être dans un de ces états
- La probabilité de distribution dans chaque état est évaluée pour les 20 acides aminés
- Une connectivité entre les états est définie en tenant compte de la biologie du système

- La probabilité de transition entre les stades est évaluée. Ces probabilités de transition entre les états pour les 20 acides aminés pour chaque état est évaluée pour un set de 160 protéines.

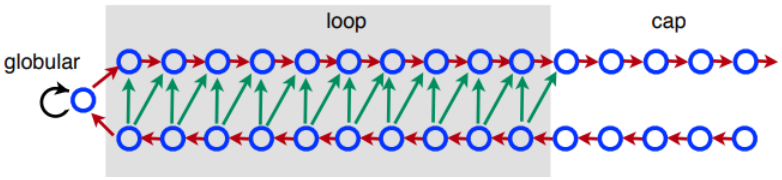
C'est une méthode performante

Example: considered states in *TMHMM*



Each box corresponds to one or several states.

Loop: maximum 20 residues; if it is larger, it is considered as a globular state:



12.2. Les protéines membranaires Beta Barrels

Plusieurs structures ont été résolues. La plupart ont un nombre égal de feuillet. En général, l'angle entre la membrane et le Beta strand est de 45°. Une face est composée d'acides aminés hydrophobes et fait face aux lipides. Un acide aminé sur deux est hydrophobe, l'autre hydrophile. Les acides aminés aromatiques sont souvent trouvés au début et à la fin.
Toute d'abord, évaluation d'un profil de la face hydrophobe. En effet, tous les 2 acides aminés, on change de face. Cela entraine une face plutôt hydrophobe au contact des lipides, et une face plutôt hydrophile à l'intérieur de la protéine.
Il y'a aussi des HMM développés pour cette prédiction

12.3. Performance des méthodes de prédiction

Les méthodes les plus performantes sont basées sur les modèles cachés de markov. Elles prédisent la topologie correcte de 70% des protéines membranaires, le nombre d'hélices et l'orientation dans la membrane. En ajoutant des contraintes, on a une amélioration expérimentale pour les deux types.

12.4. Approches pour la prédiction des boucles réentrantes

Moins fréquentes dans les récepteurs, elles apparaissent + fréquemment dans les canaux ioniques et les aquaporines.
En moyennes, elles contiennent des résidus plus petits, et présentent une composition d'acides aminés particuliers comparé aux autres régions transmembranaires. Il y'a aussi une séquence de motifs particuliers.

1) **Ab initio et méthode de fold recognition**
   Le plus difficile est de prendre en compte le milieu hydrophobe

2) **Modélisation comparative**
   Le problème est que peu de structures sont connues. Parfois, identité faible (<30%) entrainant un mauvais alignement. Il peut être amélioré en identifiant des positions fonctionnelles de séquence et en alignant ces positions sans prendre en compte le type d'acides aminés. Un autre problème est que pour la prédiction des boucles, si elles contiennent plus de 12 acides aminés, la prédiction est moins fiable.