

Protéines, hétéropolymères et systèmes désordonnés

Thomas Guérin

Master M2 SDSMC Systèmes Dynamiques et statistiques de la matière complexe
Parcours PTSC Physique théorique des systèmes complexes
Cours : Physique des systèmes désordonnés (L. Cugliandolo)

Résumé

Les protéines sont des chaînes de quelques centaines d'acides aminés. Dans les conditions physiologiques, elles parviennent presque toujours à se replier jusqu'à leur 'état natif' où elles sont fonctionnelles, et ce malgré le nombre astronomique de configurations accessibles. Dans ce document, nous présentons des approches physiques simples du repliement, en utilisant les concepts de la physique des systèmes désordonnés. Nous présentons essentiellement la théorie des hétéropolymères, et en particulier un modèle du type 'Random Energy Model'. Ce modèle permet de comprendre pourquoi les hétéropolymères aléatoires ne peuvent pas se replier. Ce résultat implique que les séquences d'acides aminés des protéines sont issues d'une évolution génétique. Enfin, nous présentons quelques approches plus récentes du repliement.

1. Introduction

A. Les protéines

Les protéines sont des chaînes d'environ une centaine d'acides aminés. Elles sont produites dans les ribosomes à partir du code génétique : 3 bases d'ARN messager (un codon) codent pour un acide aminé. Il existe 20 types d'acides aminés différents. Chaque protéine possède un état thermodynamiquement stable, appelé « état replié » ou « état natif », où elle est fonctionnelle. Une protéine qui n'est pas dans son état replié est dite dénaturée, et ne peut plus exercer sa fonction. Les fonctions des protéines sont très variées : certaines assurent le transport (moteurs moléculaires), la catalyse (de nombreuses enzymes sont des protéines), la communication, la reconnaissance (le système immunitaire possède des protéines spéciales - les immunoglobulines - qui permettent la reconnaissance moléculaire de formes « étrangères »), ou la structure (ainsi, les capsides virales sont formées de protéines).

Les protéines repliées sont compactes, d'un diamètre d'environ 4 nm. La configuration repliée est souvent unique, avec des positions d'atomes très bien définies. Ceci permet l'observation par diffraction aux rayons X de protéines cristallisées. L'enchaînement des acides aminés est codé par le génome, et forme la structure secondaire. La structure à plus grande échelle est appelée structure tertiaire. Parmi les formes secondaires les plus courantes, on trouve les hélices alpha ou les feuillets beta. La structure spatiale 3D à plus grande échelle est appelée structure tertiaire. Le repliement est la dernière étape de la formation d'une protéine.

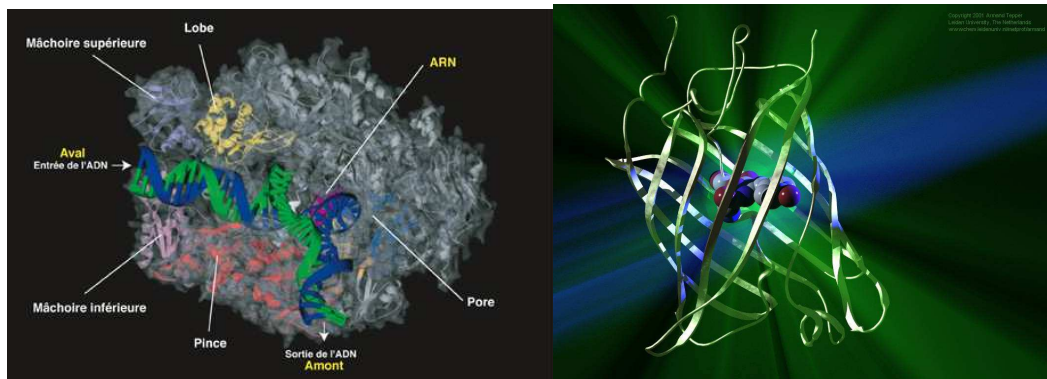


Fig. 1. Gauche : Exemple de protéine : l'ARN polymérase permet de fabriquer l'ARN à partir de l'ADN. Noter que sa structure 3D est intimement liée à sa fonction, puisqu'elle lui permet de s'enrouler autour de l'ADN, de l'ouvrir pour insérer de nouveaux nucléotides dans la chaîne d'ARN messager. Fig.1. Droite : un autre exemple de protéine célèbre : la GFP (Green Fluorescent protein), utilisé dans de nombreuses techniques de microscopie. Le chromophore est la partie de la protéine qui possède des propriétés optiques intéressantes (émission dans le vert, absorption du bleu). Noter l'importance de la structure : le chromophore est protégé par un 'tonneau' de feuillets beta.

B. Le paradoxe de Levinthal

En 1969, lors d'une conférence, Levinthal donne une estimation du nombre configurations accessibles pour une protéine d'environ 150 acides aminés, en tenant compte de contraintes angulaires. Il trouve un résultat astronomique de 10^{300} configurations*. Il faut en déduire que la recherche de l'état natif n'est pas une recherche au hasard : énumérer toutes les configurations en espérant trouver la configuration optimale prendrait un temps absolument astronomique ! Le paradoxe de Levinthal est le suivant : si toutes les configurations sont énumérées, les protéines ne peuvent pas se replier.

En fait, dans son calcul, C. Levinthal ne tient pas compte du poids de Boltzmann. Or, plusieurs arguments semblent indiquer que l'état replié est un minimum d'énergie libre. Le premier argument est que les protéines atteignent systématiquement l'état natif, où les atomes sont placés avec une grande précision. De plus, si l'on ajoute un dénaturant, la protéine se déplie, mais dès que l'on enlève le dénaturant, elle reprend sa forme native. Ceci semble donc indiquer que la forme native est un minimum d'énergie libre.

On peut faire un parallèle avec la méthode Monte Carlo. Cette méthode sert à évaluer des intégrales de fonctions (typiquement des fonctions de partition ou des moyennes) définies sur des espaces de grande dimension, donc avec un grand nombre de configurations accessibles. Effectuer un échantillonnage direct de l'espace des configurations serait catastrophique. En revanche, l'algorithme de Metropolis permet un échantillonnage efficace de l'espace des configurations en tenant compte du poids de Boltzmann. Le fait qu'il y ait un grand nombre d'états accessibles ne signifie pas qu'ils sont équiprobables, et donc ne signifie pas qu'il faut les énumérer un par un.

Cependant la dynamique de repliement des protéines est complexe. Des expériences montrent que les lois sont de type multiexponentielles, avec des constantes de temps qui varient de 1 ms à 100 s. C'est trop court pour être une énumération de toutes les configurations, et trop long pour être une simple descente dans un puits d'énergie libre. En fait, la diversité des interactions en jeu dans les protéines implique un paysage d'énergie libre complexe, avec de nombreux minima, comme nous allons le constater en étudiant la théorie des hétéropolymères.

* Depuis, le texte de la conférence est cité dans un nombre lui aussi astronomique d'articles. De façon surprenante, les auteurs reportent des estimations variées de nombres de configurations : 10^{30} , 2^{300} , etc. Heureusement, ce nombre reste astronomique, donc ne change pas les conclusions.

2. Théorie des hétéropolymères

A. Introduction

Une première approche pour l'étude théorique des protéines est de voir celles-ci comme des hétéropolymères, c'est-à-dire des polymères, dont les monomères sont puisés dans un ensemble de plusieurs types de monomères appelés « lettres ». La vision des protéines en tant qu'hétéropolymères fait suite à l'approche développée dans les années 1960, qui était essentiellement biochimique. Les scientifiques voulaient connaître la structure détaillée de chaque protéine.

Quelle est la différence fondamentale entre un homopolymère et un hétéropolymère ? L'énergie d'une configuration compacte d'un homopolymère dépend essentiellement du nombre de contacts de la chaîne avec elle-même, et est donc identique pour des chaînes avec même nombre de contacts. Les minima d'énergie libre ont donc globalement la même valeur. Pour les hétéropolymères, la situation est différente, conséquence de la diversité des interactions entre monomères. En effet, un petit décalage de la moitié des monomères, par exemple, va induire une redéfinition des interactions et mener à une énergie différente. Les minima d'énergie libre auront donc des valeurs très diverses. Une approche théorique intéressante est d'exploiter cette diversité des interactions pour relier le repliement des protéines à la physique des systèmes désordonnés. Dans la section suivante, nous nous intéressons à deux exemples de modèles décrivant les hétéropolymères comme des systèmes désordonnés.

B. « Random Energy Model » et Protéines

1. Le Random Energy Model (REM)

Rappelons ici la définition du modèle de Random Energy Model (REM) introduit par Derrida en 1980. Le REM peut être vu comme la limite du modèle d'interactions à p spins pour p tendant vers l'infini. C'est un modèle très simple, qui a l'avantage de pouvoir être résolu analytiquement. Les hypothèses sont les suivantes : on considère un système de N spins à 2^N états, dont les énergies sont d'une part indépendantes les une des autres, et d'autre part tirées avec une loi gaussienne. Le nombre d'états d'énergie comprise entre E et $E+dE$ est en moyenne sur le désordre :

$$[n(E)] \propto \gamma^N \exp(-E^2 / (NJ^2)) = \exp(N(\ln \gamma - (E/N)^2 / J^2)) \quad (1)$$

Si on prend la limite thermodynamique N tend vers l'infini, on constate que le nombre d'états termes d'énergie supérieure en valeur absolue à une énergie critique $E_0 / N = J\sqrt{\ln 2}$ est nul. Comme les fluctuations autours de $n(E)$ sont d'ordre $\sqrt{n(E)}$, qui est grand pour les énergies plus petites que E_0 , on peut approximer l'entropie moyennée sur le désordre par $[S(E)] = k_B [\ln n(E)] \sim k_B \ln [n(E)]$. L'entropie a donc la forme donnée en figure 2. On utilise ensuite la relation $1/T = \partial S / \partial E$ pour trouver la température, qui s'interprète donc comme l'inverse de la pente de la courbe de la figure 2. A température infinie, le système a une énergie nulle, l'entropie est maximale. Lorsque T décroît, l'énergie décroît. Lorsque T atteint la valeur T_0 correspondant à l'énergie $-E_0$, l'entropie devient nulle, le système choisit un état de basse énergie et y reste bloqué aussi pour les températures inférieures à T_0 .

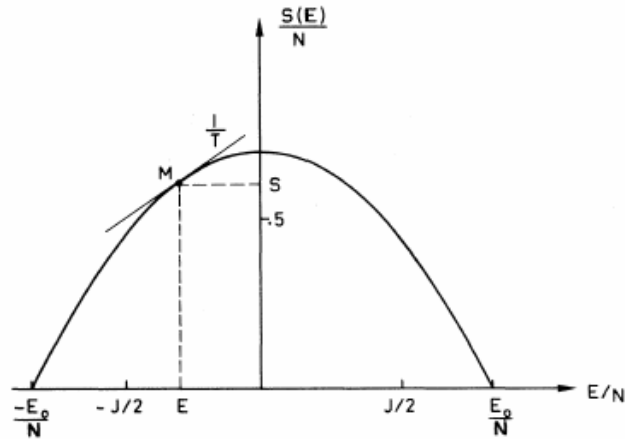


Fig.2. Entropie pour le Random Energy Model.

2. Application naive au cas des hétéropolymères aléatoires

Pour adapter ce modèle aux hétéropolymères, nous prenons à la place des N spins N acides aminés qui peuvent avoir γ configurations. Chaque acide aminé a z voisins. On suppose que les énergies d'interaction ont une moyenne $\langle U \rangle$ et un écart type σ_U . En faisant ces hypothèses, on oublie la connectivité de la chaîne (ceci ne donne pas forcément des mauvais résultats, ainsi les exposants de Flory sont une excellente estimation des exposants des polymères). On a de plus supposé qu'il y a suffisamment de « lettres » (types d'acides aminés) pour que différents états soient d'énergie indépendantes les uns des autres.

En faisant le même raisonnement que précédemment, on trouve la présence d'une énergie critique $E_0 = -z \langle U \rangle + \sigma_U \sqrt{2z \ln \gamma}$, et donc une température critique $T_0 = \sigma_U \sqrt{z / (2 \ln \gamma)}$. En dessous de cette température, le polymère est dans une configuration d'entropie nulle, et choisit un des états de plus basse énergie (pas forcément l'état natif). La protéine est donc gelée et mal repliée pour $T < T_0$.

Supposons à présent qu'il existe un état natif en plus des états du spectre continu. Cet état devrait avoir une énergie basse $-E_N$, et une entropie nulle. Alors il devient thermodynamiquement favorable lorsque l'énergie libre correspondant à la phase désordonnée devient plus haute que $-E_N$, ce qui se produit à une température T_r que l'on peut lire graphiquement sur la figure 3. (r signifie repliement dans T_r)

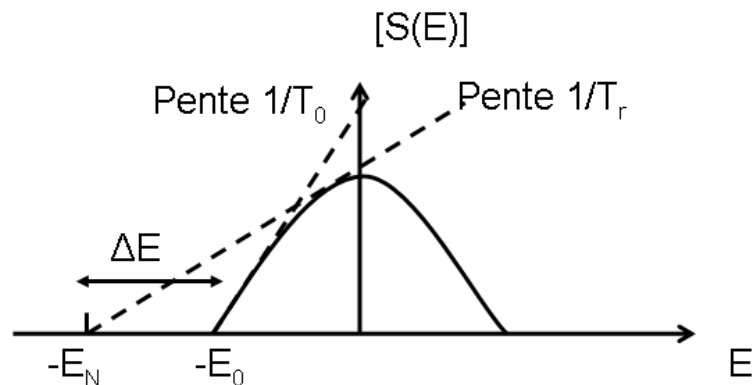


Fig. 3. Détermination graphique de la température de repliement.

Dans ce modèle, la protéine se replie donc pour des températures T vérifiant $T_0 < T < T_r$. La gamme de température pour laquelle le repliement est favorable est donc $[T_0, T_r]$. Cette gamme est d'autant plus grande que le gap d'énergie ΔE est grand, ce qui se voit assez intuitivement sur la figure 3 (ΔE est la différence entre le plus bas niveau du spectre continu et l'énergie de l'état

natif, cf figure 3). Avec ce modèle simple, l'importance de la taille du gap apparaît. Notons que le gap est ici introduit artificiellement. Dans le cas d'un hétéropolymère aléatoire, E_N est de l'ordre de E_0 , donc T_f est proche de T_0 et le polymère ne trouve son état fondamental que pour une petite gamme de température : *un hétéropolymère aléatoire ne se replie pas*. C'est un résultat que l'on pouvait attendre : en effet, on peut supposer que l'évolution sélectionne des séquences qui se replient bien, c'est-à-dire qui ont un état fondamental d'énergie bien plus basse que la plus basse énergie du spectre continu.

Le fait d'inclure artificiellement un état natif de basse énergie est fait un peu plus proprement dans l'analyse de Bryngelson et Wolynes (1994), où ils considèrent des acides aminés possédant un état natif de basse énergie et m autres états non natifs. Leur Hamiltonien comprend les énergies des structures primaires, secondaires (interactions de proches voisins sur la chaîne) et tertiaire (interactions à longue portée). Ils invoquent un *principe de frustration minimale*, selon lequel l'énergie de toutes les structures (primaires, secondaires, tertiaire) est minimisée pour l'état natif. Le nombre d'acides aminés dans l'état natif est noté N_0 . L'hypothèse simplificatrice est encore du type REM : à N_0 fixé, les énergies des états sont indépendantes. Un point important est que l'énergie de ces états natifs est pris en compte pour définir une statistique effective (gaussienne) pour $E(N_0)$. L'énergie libre est alors calculable, et l'on distingue les phases bien pliées ($N_0=N$) des phases non pliées ; ainsi que les phases vitreuses d'entropie nulle. Dans celles-ci, la protéine est gelée dans les états de basse énergie, mais non nécessairement dans l'état natif.

C. Le paramètre d'ordre pertinent

Considérons le modèle suivant, qui compte un coût ou gain d'énergie pour chaque contact:

$$H = \sum_{i,j} B(\sigma_i, \sigma_j) \delta(\vec{r}_i - \vec{r}_j) \quad (2)$$

On suppose que les $B_{ij} = B(\sigma_i, \sigma_j)$ sont tirés suivant une loi gaussienne, et on veut calculer l'énergie libre moyennée sur le désordre (car elle est « self-averaging », à la différence de Z). L'utilisation de la méthode des répliques fait apparaître des puissances de la fonction de partition Z . L'intérêt de la méthode des répliques ici est de faire apparaître un paramètre d'ordre pertinent. On écrit :

$$\left[Z^n \right] = \int \prod_{i=1}^N \prod_{\alpha=1}^n d\vec{r}_i^\alpha \int \prod_{i,j} dB_{ij} e^{-\left(\sum_{i,j} (B_{ij} - B_0)^2 / 2B^2 + \sum_{i,j,\alpha} B_{ij} \delta(\vec{r}_i^\alpha - \vec{r}_j^\alpha) \right) / kT} \quad (3)$$

Et on obtient après avoir effectué des intégrales gaussiennes :

$$\left[Z^n \right] = \int \prod_{i=1}^N \prod_{\alpha=1}^n d\vec{r}_i^\alpha \prod_{i,j} e^{-\left(\frac{B_0^2}{2B^2} - \frac{1}{2B^2} \left(B_0 - 2B^2 \sum_{\alpha} \delta(\vec{r}_i^\alpha - \vec{r}_j^\alpha) \right)^2 \right) / kT} = \int \prod_{i,\alpha} d\vec{r}_i^\alpha e^{-H_{\text{eff}} / kT} \quad (4)$$

Ceci fait apparaître un Hamiltonien effectif pour les répliques, qui contient le paramètre d'ordre :

$$q_{\alpha\beta} = \sum_i \delta(\vec{r}_i^\alpha - \vec{r}_i^\beta) \delta(\vec{r}_i^\beta - \vec{r}_j^\beta) \quad (5)$$

Ce paramètre d'ordre compte le nombre de contacts communs aux chaînes alpha et beta. S'il est nul, alors c'est que la phase est désordonnée. Nous ne nous étendrons pas plus sur le traitement de ce problème compliqué.

3. Approches récentes pour étudier le repliement : les séquences sont issues de l'évolution

Les études sur les hétéropolymères aléatoires ont montré qu'ils ne se comportent pas comme des protéines : non seulement leur température de repliement est proche de la température de transition vitreuse, mais en plus, une seule mutation a pour conséquence une

modification totale de l'état natif, ce qui n'est pas favorable à une évolution efficace. Nous avons vu des modèles où un état natif de basse énergie était introduit artificiellement. Une question qui se pose est : peut on donner des règles de sélection de séquences pour évoluer vers des protéines qui peuvent se replier ? Une façon simple de traiter ce problème est de faire tourner un algorithme de Metropolis pour optimiser l'énergie de l'état natif dans l'espace des séquences en gardant la composition de la protéine constante (Shakhnovich, Gutin, 1993). Les protéines obtenues par cette méthode se replient bien. Cette méthode fonctionne bien car la partie discrète du spectre dépend de la composition, alors que la partie continue est self averaging et ne dépend donc que de la séquence. En travaillant dans l'espace des séquences, on peut donc maximiser le « gap », donc la facilité à se replier.

Une théorie récente est celle de l'entonnoir : l'énergie sélectionnée au cours de l'évolution a la forme d'un entonnoir, ce qui permet aux protéines de se replier efficacement. Une autre approche est celle du noyau de nucléation, qui serait l'étape cinétique limitante pour le repliement. Une fois passée cette étape, la protéine se replie rapidement, même si elle doit passer plusieurs petites barrières d'énergie libre. Ceci est à rapprocher de la création de gouttes dans une transition liquide-gaz, la nucléation devant faire face à une compétition entre énergie surfacique (défavorable) et volumique (favorable).

4. Conclusion

La théorie des hétéropolymères fut une des premières approches pour modéliser une protéine, et met en lumière le concept de gap entre l'état natif et l'ensemble continu des états mal pliés, ainsi que son importance pour le repliement. Il montre que les protéines ne sont pas des hétéropolymères aléatoires, ce qui indique que les séquences ont dû être sélectionnées par l'évolution afin de produire des protéines qui se replient bien. Ces mécanismes de sélection sont de plus en plus mis en lumière par des théories récentes.

Bibliographie

- Protein folding thermodynamics and dynamics : where physics, chemistry, and Biology meet*, E. Shakhnovich, Chem. Rev., 106, 1559-1588 (2006)
- Engineering of stable and fast folding sequences of model proteins*, E. Shakhnovich, A.M. Gutin, Proc. Natl. Acad. Sci. USA, 90, 7195-7199 (1993)
- Mössbauer Spectroscopy in Biological Systems Proceedings: "How to fold gracefully"*, C. Levinthal, 67, 41, 22-24 (1969)
- Theory of protein folding*, J.N. Onuchic, P. Wolynes, Current Opinion in Structural Biology 14, 70-75 (2004)
- Statistical mechanics of random heteropolymers*, C.D. Sfatos, E.I. Shakhnovich, Physics Reports 288, 77-108 (1997)
- Protein folding theory: from lattice to all-atom models*, Annu. Rev. Biophys. Biomol. Struct. 30,361-96 (2001)
- Specific Nucleus as the transition state for protein folding: evidence from the lattice model*. V.I. Abkevich, A.M. Gutin, E.I. Shakhnovich, Biochem. 33, 10026-10036 (1994)
- Spin glasses and the statistical mechanics of protein folding*, J.D. Bryngelson, P.G. Wolynes, Proc. Natl. Acad. Sci. USA, 84, 7524-7528 (1987)
- Requirements for protein folding potentials: Perspectives from lattice model simulations*, A Dissertation Presented to the Faculty of the Graduate School of Arts and Sciences, Antônio F. Pereira de Araújo, (Brandeis University , Graduate Program in Biophysics and Structural Biology)(1998)
- Random energy model: limit of a family of disordered models*, B. Derrida, Phys. Rev. Letters, 45,2,74-82,(1980)
- Random-energy model: an exactly solvable model of disordered systems*, B. Derrida, Phys. Rev. B. 24,5,2613-2626 (1981)