

UNE MÉTHODE ADAPTATIVE POUR LE CALCUL D'ÉNERGIE LIBRE

Kimiya Minoukadeh

Résumé. — Nous nous intéressons au calcul par simulation moléculaire de certaines quantités macroscopiques liées à un système d'atomes. L'état du système est décrit par la position des atomes qui interagissent au travers d'une fonction d'énergie potentielle. L'objectif est de calculer des moyennes par rapport à la mesure de Boltzmann-Gibbs liée à ce potentiel, ce qui revient en pratique à échantillonner la mesure en utilisant une dynamique appropriée, par exemple la dynamique de Langevin. Cette approche naïve mène souvent à une convergence lente, due à des métastabilités liées au potentiel, où des zones de forte probabilité sont séparées par des zones de faible probabilité. Étant donné une coordonnée de réaction (une variable lente) décrivant les métastabilités du système, il existe une méthode, celle de la force biaisante adaptative [1, 2], pour remédier à ce problème. Cette méthode fournit un moyen pour échantillonner des mesures de probabilité multimodales, en biaisant la dynamique habituelle par une certaine force 'moyenne' dans la direction de la métastabilité. Il a été proposé récemment [4, 6] d'utiliser un système de répliques en interaction pour accélérer l'algorithme et surmonter les problèmes liés à d'autres métastabilités dans la direction orthogonale à la coordonnée de réaction. Cette approche, facilement parallélisable, permet de mettre en œuvre un processus de sélection pour accélérer davantage l'échantillonnage. Nous présentons une comparaison numérique de ces méthodes sur un système biologique prototypique.

1. Introduction

La simulation moléculaire est un outil indispensable pour comprendre le comportement des systèmes physiques (par exemple les matériaux) ou biologiques (par exemple les protéines) lorsque les expérimentations sont difficiles ou coûteuses. La simulation moléculaire donne un moyen pour calculer les propriétés moyennes des matériaux par les lois de la physique statistique.

On considère un système de N atomes, décrit par les configurations $q = (q_1, \dots, q_N)$, où $q_i \in \mathbb{R}^3$ désigne la configuration du i -ème atome. Les particules interagissent selon la fonction d'énergie potentielle $V : \mathbb{R}^{3N} \rightarrow \mathbb{R}$. La mesure canonique, qui donne l'information microscopique relative au système, est définie par

$$(1) \quad \mu(dq) = \phi(q) dq = Z^{-1} \exp(-\beta V(q)) dq,$$

où $\phi(q)$ est la densité de la mesure μ au point $q \in \mathbb{R}^{3N}$ et $Z = \int \exp(-\beta V(q)) dq$ est un facteur de normalisation, également appelé la fonction de partition. On voudrait calculer des moyennes par rapport à cette mesure, mais en grande dimension (typiquement $N \sim 10^5$), cela devient infaisable. On est donc davantage intéressé par le comportement macroscopique du système auquel est associée une variable lente ξ , qui représente souvent un angle structural ou bien la distance bout-à-bout d'une protéine. Dans la suite, cette quantité sera appelée *coordonnée de réaction* (RC), définie par une fonction dérivable $\xi : \mathbb{R}^{3N} \rightarrow \mathbb{R}$. L'image de la mesure canonique par ξ est donnée par

$$(2) \quad \mu^\xi(dz) = \phi^\xi(z) dz = \tilde{Z}^{-1} \exp(-\beta A(z)) dz,$$

où \tilde{Z} est la constante de normalisation, $\phi^\xi(z)$ est la densité de la mesure μ^ξ au point z et A est appelée l'énergie libre et est définie par

$$A = -\beta^{-1} \ln \left(\tilde{Z} \int_{\Sigma_z} \phi |\nabla \xi|^{-1} d\sigma_z \right)$$

où $d\sigma_{\Sigma_z}$ est la mesure de Lebesgue sur $\Sigma_z = \{x, \xi(x) = z\}$. L'objectif de ce travail est de calculer efficacement l'énergie libre A .

2. Calcul d'énergie libre

Il est important de noter qu'en pratique on souhaite calculer la différence d'énergie libre et non l'énergie libre absolue. La différence d'énergie libre entre deux points z_1 et z_0 dans la RC est définie par $\Delta A = A(z_1) - A(z_0) = \int_{z_1}^{z_0} A'(z) dz$. On sait que l'énergie libre est définie, à une constante additive près, par $A = -\beta^{-1} \ln(\phi^\xi)$. Après dérivation de cette quantité, il s'avère que A' s'exprime sous la forme d'une espérance conditionnelle

$$(3) \quad A'(z) = \mathbb{E}_\mu[F^V(q) \mid \xi(q) = z] = \frac{\int_{\Sigma_z} F^V \exp(-\beta V) |\nabla \xi|^{-1} d\sigma_{\Sigma_z}}{\int_{\Sigma_z} \exp(-\beta V) |\nabla \xi|^{-1} d\sigma_{\Sigma_z}},$$

où $F^V = \frac{\nabla V \cdot \nabla \xi}{|\nabla \xi|^2} - \beta^{-1} \operatorname{div} \left(\frac{\nabla \xi}{|\nabla \xi|^2} \right)$. Pour déduire cette dernière, voir les références [7, 8]. Dans la suite A' est appelée *force moyenne*. Une manière de calculer A' est d'utiliser une dynamique qui échantillonne efficacement la mesure μ pour toute valeur de la RC.

Une manière naïve de procéder est d'utiliser une dynamique de Langevin sur-amortie, pour laquelle on sait qu'en temps long les positions de la trajectoire sont distribuées selon la mesure canonique. Cette dynamique s'écrit sous la forme d'une équation différentielle stochastique :

$$(4) \quad dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}}dW_t,$$

où X_t est un vecteur de longueur $3N$, représentant la trajectoire du système et W_t est un mouvement brownien standard de dimension $3N$. Cette dynamique, cependant, se heurte au phénomène de métastabilité : des zones de \mathbb{R}^{3N} de forte probabilité sont séparés par des zones de faible probabilité. Puisque ξ décrit les degrés de liberté lents du système, ces barrières à franchir sont liées à celles de l'énergie libre A . La méthode de la force biaisante adaptative, détaillée dans la section suivante, propose une façon de remédier à ce problème.

3. Force biaisante adaptative

L'idée de la méthode force biaisante adaptative (ABF) [1, 2] est de biaiser la dynamique habituelle (4) par adjonction d'une force A'_t variant avec le temps. La dynamique biaisée est ainsi

$$(5) \quad \begin{cases} dX_t = -\nabla(V - A_t \circ \xi)(X_t)dt + \sqrt{2\beta^{-1}}dW_t, \\ A'_t(z) = \mathbb{E}[F^V(X_t) | \xi(X_t) = z] \end{cases}$$

où $'$ signifie une dérivée par rapport à z . La force biaisante A'_t est ainsi une approximation de la force moyenne A' , calculée en utilisant l'information obtenue par la trajectoire du processus X_t . Dans la suite ψ représente la densité de la loi du processus X_t et ψ^ξ celle de la marginale en ξ . Il a été montré que $A'_t \rightarrow A'$ exponentiellement vite, au sens $L^2(\psi^\xi dx)$ [5].

A l'origine de la méthode ABF [1], une unique réalisation de la dynamique (5) était considérée. La force moyenne était donc calculée par une moyenne trajectoirelle. Une nouvelle formulation d'ABF [4] propose d'utiliser plusieurs répliques indépendantes qui suivent toutes la dynamique (5). Le processus X_t dans (5) est ainsi remplacé par X_t^i et W_t par W_t^i , où l'indice $i \in \{0, \dots, R-1\}$ indexe la i -ème réplique. L'espérance conditionnelle est alors calculée comme une moyenne sur les trajectoires et sur toutes les répliques

$$A'_t(z) \approx \frac{\sum_{i=0}^{R-1} \int_0^t F^V(X_s^i) \delta_{\xi(X_s^i)}(dz) ds}{\sum_{i=0}^{R-1} \int_0^t \delta_{\xi(X_s^i)}(dz) ds}.$$

L'intérêt de cette formulation est d'une part, d'accélérer la convergence de la méthode, surtout dans le cas où la coordonnée de réaction ne décrit pas toutes les métastabilités du système. D'autre part, elle permet d'effectuer une sélection des répliques selon une fonction d'utilité $S(t, z) = c \frac{\partial_{zz} \psi^\xi}{\psi^\xi}$, de sorte que $S(t, z) < 0$ (respectivement $S(t, z) > 0$) si l'état z dans la RC est sous-échantillonné (respectivement sur-échantillonné) par rapport à sa moyenne locale. La constante $c > 0$ représente ainsi l'intensité de cette fonction d'utilité. À chaque réplique est assigné un poids w_t^i , défini par

$$w_t^i = W_t^{-1} \exp \left(\int_{t_0}^t S(s, \xi(X_s^i)) ds \right),$$

où est le facteur de normalisation et $t_0 < t$ désigne le dernier temps du rééchantillonnage. Au moment du rééchantillonnage, la i -ème réplique est dupliquée en moyenne $Rw_{t^-}^i$ fois, où l'indice t^- signifie le temps juste avant le rééchantillonnage. Les poids sont ensuite distribués uniformément, c'est à dire $w_t^i = 1/R, \forall i \in \{0, \dots, R-1\}$, et évoluent jusqu'au prochain rééchantillonnage, selon $\dot{w}_t^i = S(t, \xi(X_t^i))w_t^i$. Pour plus de détails sur la mise en œuvre, voir la référence [6].

4. Résultats numériques

Cette section présente des tests numériques effectués sur un système prototypique, la déca-alanine, de $N = 104$ atomes. La RC est choisie comme la distance bout-à-bout de la protéine.

4.1. Les structures dépliées. — On étudie d'abord les états dépliés de la protéine, pour lesquels la RC choisie décrit bien les phases de transition du répliage. En imposant des conditions de réflexion aux bords, la distance bout-à-bout du système est restreinte entre 12 et 32 Å. La figure 1(a) montre les estimations de A' , obtenues par des simulations d'une réplique et de 16 répliques, après 0,25 ns. Les courbes représentent la moyenne de 20 simulations et des intervalles de confiance à 95%. Les simulations avec une réplique n'ont pas réussi à échantillonner au-delà d'une élongation de 22 Å, alors que les simulations avec 16 répliques ont pu obtenir une meilleure estimation de A' , nous permettant de mieux biaiser la dynamique et atteindre les états dépliés. La figure 1(b) illustre l'influence du mécanisme de sélection. On obtient un meilleur échantillonnage des états dépliés pour les simulations avec sélection, ce qui se reflète dans le profil d'énergie libre dans l'encart de la figure 1(b).

4.2. Les structures repliées. — Dans les états repliés, la distance bout-à-bout de la protéine ne décrit plus toutes les métastabilités du système. Dans de tels cas, l'usage de plusieurs répliques s'avère indispensable pour avoir une convergence à une échelle de temps raisonnable. La RC est restreinte entre 4 et 16 Å.

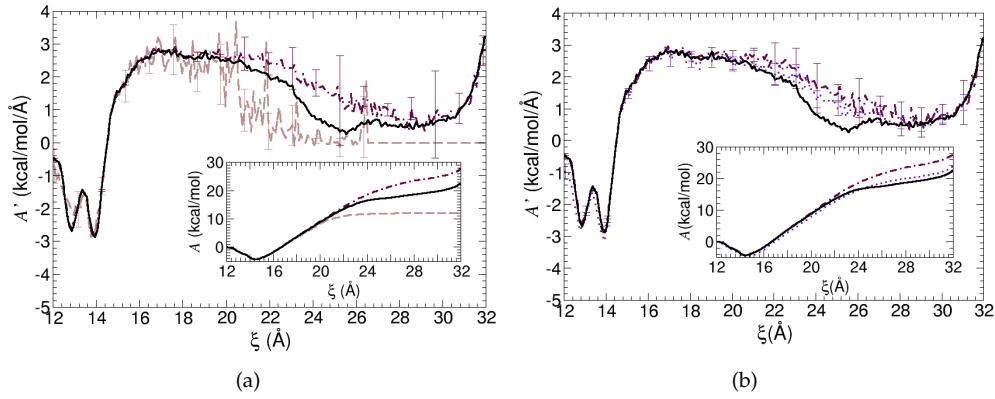


FIGURE 1. **Estimations de la force moyenne après 0,25 ns, pour les états dépliés** : Les lignes solides représentent les courbes de référence. (a) Simulations avec 1 réplique (lignes de tirets) et 16 répliques sans sélection (lignes de tirets/pointillés). (b) Simulations avec 16 répliques sans sélection (lignes de tirets/pointillés) et avec sélection (lignes de pointillés).

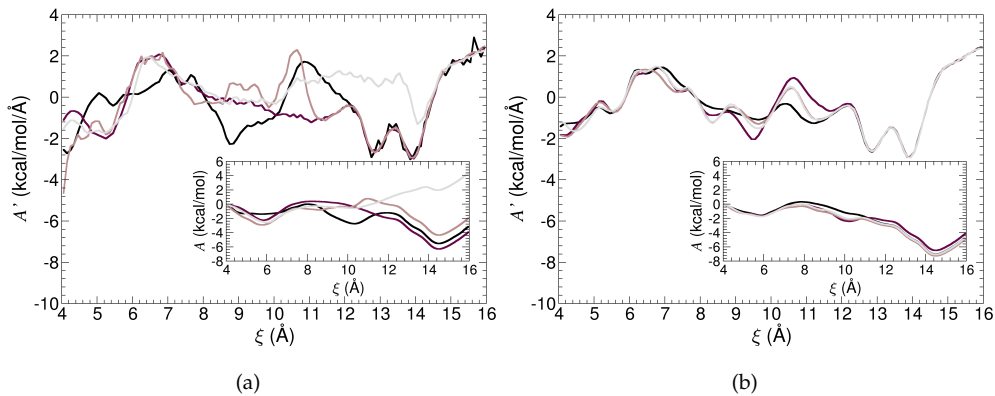


FIGURE 2. **Estimations de la force moyenne après 100 ns, pour les états repliés** : Les courbes représentent les résultats de quatre réalisations indépendantes. (a) Simulations avec une réplique. (b) Simulations avec 32 répliques : les estimations sont plus précises et les courbes sont qualitativement proches l'une de l'autre.

Les figures 2(a) et 2(b) montrent les estimations de A' , obtenues après 100 ns, par quatre réalisations indépendantes d'ABF, en utilisant 1 réplique et 32 répliques respectivement. On observe sur la figure 2(a), des incohérences entre les estimations obtenues par chaque simulation. Une explication possible est la suivante : dans les états compacts il existe, pour une valeur fixe de ξ , une multitude de conformations

de basse énergie séparées par des barrières à haute énergie. La méthode ABF habituelle n'est plus capable de calculer la force moyenne d'une manière efficace. Ce problème, cependant, peut être surmonté lorsqu'on utilise plusieurs (et éventuellement un grand nombre de) répliques. On observe sur la figure 2(b) que les estimations obtenues par les simulations de 32 répliques sont cohérentes l'une avec l'autre. Ceci est en accord avec la spéculation que la surface de l'énergie libre contient de nombreuses vallées de basse énergie dans la direction orthogonale à l'isosurface $\xi(q) = z$. En utilisant plusieurs répliques, elles ont une chance de parcourir chacune une vallée différente. Il est, par conséquent, souhaitable d'utiliser un nombre important de répliques lorsque la RC est sous optimale.

Références

- [1] E. Darve and A. Pohorille *Calculating free energies using average forces*, J. Chem. Phys. 2001.
- [2] J. Héning and C. Chipot, *Overcoming free energy barriers using unconstrained molecular dynamics simulations*, J. Chem. Phys., 2004.
- [3] C. Chipot and J. Héning, *Exploring the free-energy landscape of a short peptide using an average force*, J. Chem. Phys., 2005.
- [4] T. Lelièvre, M. Rousset and G. Stoltz, *Computation of free energy profiles with parallel adaptive dynamics*, J. Chem. Phys., 2007.
- [5] T. Lelièvre, M. Rousset and G. Stoltz, *Long-time convergence of an Adaptive Biasing Force method*, Nonlinearity., 2008.
- [6] K. Minoukadeh, C. Chipot and T. Lelièvre, *Potential of mean force calculations : a multiple-walker adaptive biasing force approach*, Journal of Chemical Theory and Computation, 6(4), 1008-1017, (2010).
- [7] W. K. den Otter and W. J. Briels, *The calculation of free energy differences by constrained molecular dynamics simulations*, J. Chem. Phys., 1998.
- [8] M. Sprik and G. Cicotti, *Free energy from constrained molecular dynamics*, J. Chem. Phys., 1998.

Kimiya Minoukadeh

CERMICS, École des Ponts ParisTech, 6-8 avenue Blaise-Pascal, 77455 Champs-sur-Marne, Marne-la-Vallée cedex 2, France.

E-mail : kimiya.minoukadeh@cermics.enpc.fr