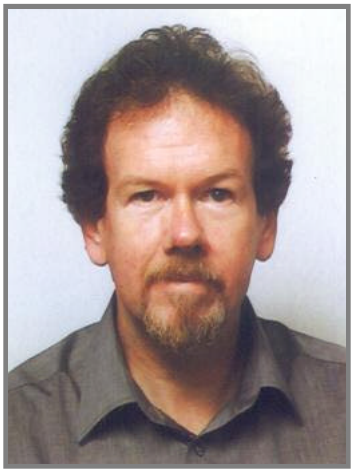
LA SIMULATION DES SYSTÈMES BIOLOGIQUES À L'AUBE DE L'AN 2000

SCIENCE

Richard LAVERY
Laboratoire de Biochimie
Théorique, CNRS UPR 9080
Institut de Biologie
Physico-Chimique
13 rue Pierre et Marie Curie
Paris 75005



Nous nous trouvons au cœur d'une explosion de nos connaissances biologiques, qui depuis son initiation il y a environ cinquante ans, est encore loin de montrer des signes de ralentissement.

Cette révolution, comparable à l'essor de la physique au début du siècle, nous a révélé le code génétique, nous approvisionne en structures macromoléculaires d'une étonnante complexité et nous laisse entrevoir la façon dont ces édifices interagissent pour donner lieu aux multiples processus du vivant.

Quel rôle joue la théorie dans cette révolution ?

Contrairement aux systèmes inorganiques, à quelques notables exceptions près, les systèmes vivants sont peu accessibles aux traitements théoriques simples. Leur hétérogénéité, aussi bien au niveau des biopolymères individuels qu'au niveau de leurs agencements et leurs interactions avec l'environnement, nous oblige généralement à construire des modèles détaillés, souvent au niveau atomique, pour en comprendre le fonctionnement.

Cette voie n'est pas sans difficultés. Les systèmes moléculaires biologiques sont caractérisés par trois facteurs majeurs :

(a) ils comportent beaucoup d'atomes,

(b) ils sont structurés par des interactions faibles,

(c) ils sont animés par des mouvements à plusieurs échelles de temps.

Pour donner quelques exemples de leur taille, nous pouvons commencer avec les protéines qui sont composées par des enchaînements d'environ 200 à 500 acides aminés et contiennent ainsi environ 3000 à 8000 atomes. Dans le domaine des acides nucléiques, un simple ARN de transfert, formé d'environ 80 nucléotides, contient déjà 3000 atomes (voir aussi, fig. 1).

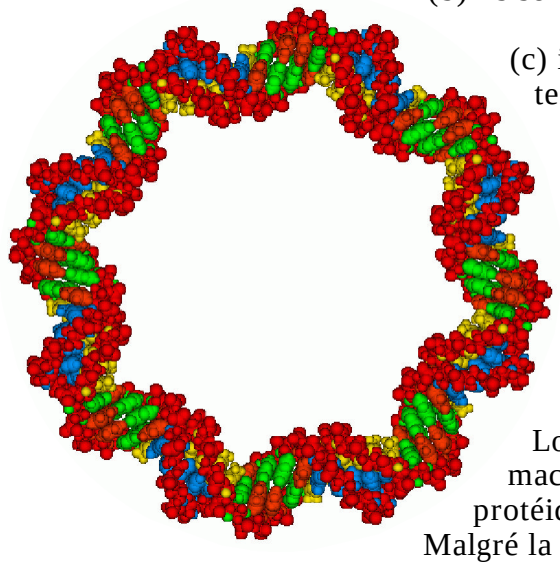


Fig 1

Modèle d'un minicercle de l'ADN contenant 80 paires de bases et plus de 5000 atomes.

Lorsqu'on assemble de telles molécules pour créer l'une des machines du vivant comme le ribosome, siège de la synthèse protéique, on atteint rapidement le quart de million d'atomes. Malgré la taille et la complexité de ces édifices, ils ne sont que marginalement stables. Le repliement des macromolécules et leur assemblage au sein de structures plus complexes mettent en jeu des interactions physiques plutôt que chimiques (liaisons hydrogènes, ponts salins, interactions van der Waals, ...). De surcroît, même si la somme de ces interactions représente des énergies importantes, des compensations lors de la structuration des édifices moléculaires (désolvatation, perte d'entropie, ...) font que les états finaux ne sont stabilisés que par quelques kcal/mol.

Ainsi, ils sont fragiles et facilement influencés par leur environnement physico-chimique (température, pH, salinité, ...).

Cette fragilité, essentielle pour le contrôle biologique de leur activité, explique aussi l'importance de leur comportement dynamique. La dynamique est présente dans tous les processus du vivant, facilitant l'interaction des molécules, animant l'action catalytique des enzymes ou, à plus grande échelle, permettant la fabrication et le déplacement des acteurs moléculaires de la cellule. Elle couvre un énorme registre de temps, allant des femtosecondes pour les vibrations des liaisons chimiques, jusqu'aux millisecondes ou aux secondes pour les processus les plus complexes (repliement, synthèse de biopolymères, ...).

Ces multiples dynamiques ne font que traduire la complexité des surfaces énergétiques qui sous-tendent les interactions macromoléculaires et sont parsemées de barrières de hauteurs très variables.

Ces trois caractéristiques (taille, stabilité marginale et dynamique) sont également les trois facteurs qui compliquent la simulation des systèmes biologiques au niveau atomique (fig. 2).

En premier lieu, la nécessité de prendre en compte plusieurs milliers d'atomes nous contraint à adopter une représentation classique de leurs interactions, en employant des champs de force empiriques à la place de calculs quantiques plus exacts. En même temps, la faiblesse des interactions mises en jeu et la délicatesse des équilibres nous amènent à calculer les interactions aussi précisément que possible. Finalement, la complexité des hypersurfaces énergétiques et une dynamique multi-échelles nous obligent à trouver des moyens efficaces pour parcourir l'espace des conformations et celui des phases.

Aujourd'hui, nous pouvons placer la barre des simulations aux alentours de quelques dizaines de milliers d'atomes en terme de taille, d'une dizaine de kcal/mol en terme de précision et d'une dizaine de nanosecondes en terme de temps. Ces limites sont très directement couplées aux progrès de l'informatique. Quand des moyens exceptionnels sont disponibles, des performances exceptionnelles peuvent être atteintes, comme dans un cas récent où plusieurs mois de calcul sur un T3E à 256 processeurs ont permis de simuler le repliement d'une petite protéine pendant une microseconde (1, 2).

Malgré l'importance des moyens de calcul, nous devons aussi compter sur des progrès algorithmiques pour repousser les limites des simulations. Ainsi, pour les simulations par dynamique moléculaire (3), de nouvelles techniques permettent déjà d'éviter la troncature des interactions électrostatiques qui ont été, jusqu'à récemment, une source importante d'instabilité (4).

Dans ce même domaine électrostatique, nous pouvons aussi compter sur des représentations hiérarchiques à base de multipôles pour accélérer la simulation des très grands systèmes. D'autre gains viendront du remplacement des molécules de solvant explicites par des repré-

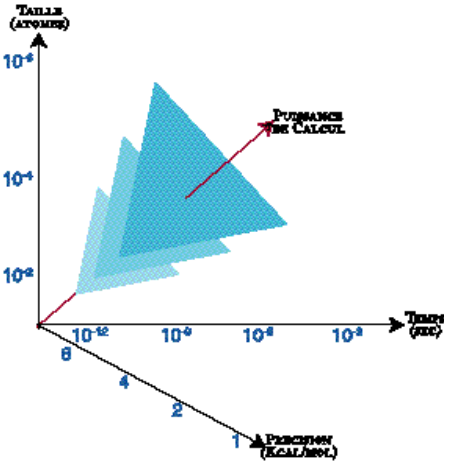


Fig 2

L'espace de simulation défini par les axes : taille, temps de simulation et précision.

1. D.L. Wang, P.A. Kollman (1998) Proc. Natl. Acad. Sci. (USA) 95, 9897.

2. Y. Duan, P.A. Kollman (1998) Science 282, 740.

3. M. Karplus, G.A. Petsko (1990) Nature 347, 631.

4. P. Auffinger, E. Westhof (1998) Curr. Opinion Struct. Biol. 8, 227.

5. M.A. Cunningham, P.A. Bash Computer Simulation of Biomolecular Systems Vol.3 eds. W.F. van Gunsteren, P.K. Weiner, A.J. Wilkinson, Kluwer, Dordrecht pp177.

6. H.S. Chan, K.A. Dill (1998) Proteins 30, 2.

7. I. Lafontaine, R. Lavery (1999) Curr. Opinion Struct. Biol. 9, 170.

sentations continues, si des modèles suffisamment précis peuvent être développés. Finalement, pour dépasser les limites des champs de force classiques, et pour réintroduire de la chimie dans les simulations, il est désormais possible d'employer des méthodes hybrides qui permettent de traiter une partie du système au niveau quantique, tout en gardant des interactions avec la partie classique (5).

D'autres efforts aujourd'hui se concentrent sur le développement de modèles simplifiés et sur l'amélioration des algorithmes de recherche conformationnelle. Dans les deux cas, le but est de contourner les barrières qui rendent inefficace l'emploi de la dynamique moléculaire simple, qu'elles soient liées à la taille des systèmes ou à la lenteur des processus en question (repliement, assemblage, ...).
Les modèles simplifiés impliquent le remplacement de groupes d'atomes par des corps rigides ou élastiques. Ceci permet de réduire le nombre de particules dans le système, ainsi que le nombre de leurs degrés de liberté, jusqu'à restreindre dans certains cas les mouvements possibles aux déplacements sur un maillage de points prédéfinis.

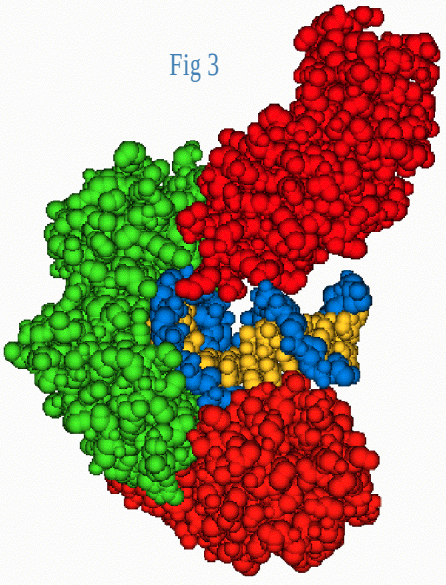
De tels modèles, qui ont déjà servi pour simuler le repliement des protéines (6), ou encore le comportement de fragments d'ADN composés de plusieurs milliers de nucléotides (7), nécessitent le développement de potentiels effectifs et soulèvent la question du passage ou de la cohabitation entre différents niveaux de représentation. Pour améliorer les recherches conformationnelles, plusieurs approches sont en développement, allant des extensions de la méthode Monte Carlo (recuit simulé, MC minimisation, « build-up » ...) aux algorithmes génétiques, ou encore aux techniques dites multi-copies ou au lissage artificiel des surfaces énergétiques.

Ces méthodes sont particulièrement importantes pour faciliter la recherche de complexes optimaux entre de petites molécules pharmacoactives et leurs cibles macromoléculaires ou pour décrire les modifications conformationnelles qui résultent de la mutation de quelques acides aminés lors de l'ingénierie des protéines. Beaucoup reste à faire dans ce domaine, particulièrement pour décrire la formation des complexes macromoléculaires, auquel cas il faut prendre en compte la plasticité des surfaces en contact et le rôle subtil de l'environnement (fig. 3). Pour ce faire, il ne faut pas ignorer l'apport du graphisme et surtout, depuis peu, la possibilité de coupler des calculs lourds aux représentations graphiques. Des interactions efficaces avec le système en temps réel sont ainsi tout à fait envisageables.

Quelles sont les applications de ces multiples techniques en biologie ?

Elles sont très nombreuses. Certaines, comme la dynamique moléculaire sous contraintes, constituent désormais une des étapes incontournables dans l'obtention des structures biomacromoléculaires. D'autres, comme la conception assistée de médicaments, ou encore l'ingénierie des protéines, profitent déjà de l'apport de la modélisation, même si les techniques en question sont loin d'avoir atteint une maturité méthodologique. Il faut néanmoins signaler la contribution crois-

Fig 3



Un fragment du complexe protéine-ADN responsable de la transcription des gènes. Ce fragment a été construit par l'assemblage de deux structures cristallographiques (12,13) contenant la TBP (TATA-box binding protein, en vert) et les facteurs de transcription TF11A et TF11B (en rouge). La TBP est responsable de la reconnaissance de la bonne séquence de l'ADN (bases : jaune, brins : bleu). Malgré sa taille et la déformation importante de l'ADN qu'elle provoque (14), seulement 33 acides aminés sont directement en contact avec la double hélice.

sante des méthodes hybrides (5) ainsi que des calculs d'énergie libre (8) à la compréhension des mécanismes catalytiques des enzymes.

Certains domaines sont encore peu exploités, comme l'étude des membranes, des protéines qui les peuplent (9) et des processus de transport qu'elles gèrent (10), ou encore l'interaction des biomolécules avec des surfaces, vitale pour le développement des biomatériaux, des sondes et, depuis peu, des puces à ADN. Dans ces cas, la complexité d'un milieu hétérogène et la lenteur relative des processus mis en jeu sont des barrières importantes à l'utilisation des simulations à l'échelle atomique.

Dans la direction opposée, des développements expérimentaux continuent d'alimenter les progrès en modélisation. Pour ne citer qu'un cas, il est impressionnant de voir comment la manipulation de molécules individuelles a offert aux théoriciens une chance inespérée de sonder directement la mécanique d'une déformation ou d'une interaction moléculaire (11). Il ne fait aucun doute que la modélisation et la simulation sont désormais des outils puissants pour aider le biologiste dans la compréhension des systèmes vivants, permettant d'interpréter et de compléter des informations expérimentales.

Mais, dans de nombreux domaines nous sommes loin de pouvoir résoudre les problèmes les plus pressants.

Aujourd'hui, nous sommes confrontés à une nouvelle révolution biologique alimentée par les projets de séquençage. Ces projets ont déjà fourni les séquences complètes des génomes de plusieurs organismes simples et, dans quelques années, notre espèce sera décortiquée de la même manière. Posséder de telles informations nous permettra de comprendre non seulement qui sont les acteurs individuels des systèmes biologiques, mais aussi comment ils agissent et interagissent pour donner lieu à des organismes vivants.

Ce changement d'échelle dans la quantité de données disponibles appelle à des efforts renouvelés de la part des chimistes théoriciens pour aider à combler le gouffre entre séquence et structure, pour passer de l'étude des interactions binômes aux interactions multiples, ou encore pour mettre l'ingénierie des systèmes moléculaires complexes au service de l'homme et de ses industries.

Pour répondre à cet enjeu nous aurons besoin de tous les talents des nouvelles générations de chercheurs et de toute la puissance de calcul que l'informatique peut fournir.

8. T. Simonson dans Computational Biochemistry and Biophysics eds. O. Becker, A.D. Mackerell, B. Roux, M. Watanabe, Dekker, New York, sous presse.

9. C. Etchebest, J.L. Popot (1997) Membrane Protein Assembly ed. G. von Heijne, R.G. Landes Co. New York pp221.

10. G. Wipff (1998) Sciences Chimiques, Lettre du Département CNRS 66, 22.

11. D. Bensimon (1996) Structure 4, 885.

12. S.Tan, Y. Hunziker, D.F. Sargent, T.J. Richmond (1996) Nature 381, 127.

13. D.B. Nikolov, H. Chen, E.D. Halay, A.A. Usheva, K. Hisatake, D.K. Lee, R.G. Roeder, S.K. Burley (1995) Nature 377, 119.

14. A. Lebrun, Z. Shakked, R. Lavery (1997) Proc. Natl. Acad. Sci. USA 94, 757.