物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

生物大分子多尺度理论和计算方法

李文飞　张建　王骏　王炜

Multiscale theory and computational method for biomolecule simulations

Li Wen-Fei　Zhang Jian　Wang Jun　Wang Wei

引用信息 Citation: Acta Physica Sinica, 64, 098701 (2015)　DOI: 10.7498/aps.64.098701
在线阅读 View online: http://dx.doi.org/10.7498/aps.64.098701
当期内容 View table of contents: http://wulixb.iphy.ac.cn/CN/Y2015/V64/I9

专题: 庆祝南京大学物理学科成立100周年

生物大分子多尺度理论和计算方法*

李文飞† 张建 王骏 王炜‡

(南京大学物理学院, 固体微结构国家实验室, 南京 210093)

(人工微结构科学与技术协同创新中心, 南京 210093)

(2015 年 1 月 19 日收到; 2015 年 3 月 5 日收到修改稿)

分子模拟是研究生物大分子的重要手段. 过去二十年来, 人们将分子模拟与实验研究相结合, 揭示出生物大分子结构和动力学方面的诸多重要性质. 传统分子模拟主要采用全原子分子模型或各种粗粒化的分子模型. 在实际应用中, 传统分子模拟方法通常存在精度或效率瓶颈, 一定程度上限制了其应用范围. 近年来, 多尺度分子模型越来越受到人们的关注. 多尺度分子模型基于统计力学原理, 将全原子模型和粗粒化模型相耦合, 有望克服传统分子模拟方法中的精度/效率瓶颈, 进而拓展分子模拟在生物大分子研究中的应用范围. 根据模型之间的耦合方式, 近年来发展起来的多尺度分子模拟方法可归纳为如下四种类型: 混合分辨多尺度模型、并行耦合多尺度模型、单向耦合多尺度模型、以及自学习多尺度模型. 本文将对上述四类多尺度模型做简要介绍, 并讨论其主要优缺点、应用范围以及进一步发展方向.

关键词: 生物大分子, 多尺度模型, 分子模拟, 粗粒化

PACS: 87.15.ap, 87.15.Cc, 87.18.–h, 87.16.A– **DOI:** 10.7498/aps.64.098701

1 引 言

蛋白质、核酸等生物大分子主要通过多个尺度上的相互作用和构象涨落运动行使其生物学功能[1]. 例如, 各种与 ATP 水解相关的蛋白质分子马达 (如 F_1-ATPase[2], DNA 解旋酶[3], 蛋白酶体中的转运马达等[4]) 的功能过程通常是通过蛋白质分子的全局大尺度构象运动和 ATP/ADP 与蛋白质分子在原子层次的局域相互作用紧密耦合、协同作用来完成的. 显然, 要完全理解蛋白质等生物大分子体系行使功能的物理机理, 需要从多个尺度上同时刻画其结构、相互作用与构象运动动力学. 然而, 要精确刻画这种多尺度的相互作用与功能运动是非常困难的. 实验上, 尽管 X 射线衍射、核磁共振等高分辨结构生物学方法能够给出蛋白质/核酸天然态结构的高分辨原子位置信息, 但通常不能直接提供分子功能运动的动态信息. 相反地, 以光谱学技术为代表的各种生物物理方法以及单分子实验技术能够给出分子功能运动的动态信息[5,6], 却很难同时提供功能运动过程中高分辨的原子位置等结构信息. 这些实验上的困难和局限性很大程度上限制了人们对生命过程分子机理的深入理解. 而以分子模拟为主的理论方法由于其能够同时提供结构和动力学等信息[7−19], 被认为是联系结构生物学方法与生物物理/单分子技术的重要桥梁, 在人们认识生物大分子功能运动机理中发挥着越来越重要的作用. 特别是近年来, 由于计算机技术的高速发展, 以分子模拟为主的理论方法已成为研究蛋白质等生物大分子功能运动的主要手段之一.

从物理的角度看, 生物大分子体系在多个尺度上的功能运动动力学完全由原子层次的微观相互

* 国家自然科学基金 (批准号: 11174134, 11334004, 11274157, 11174133) 和江苏省自然科学基金 (批准号: BK2011546) 资助的课题.

† 通信作者. E-mail: wfli@nju.edu.cn

‡ 通信作者. E-mail: wangwei@nju.edu.cn

作用(包括生物大分子内部原子之间以及生物大分子原子与溶剂等环境分子之间的相互作用)所决定. 因此, 原则上基于全原子水平的微观分子动力学(甚至量子力学方法), 人们可以模拟生物大分子的完整功能运动过程[7,11]. 然而, 由于生物大分子的多体复杂性以及构象运动的多尺度特征, 完全基于原子分辨的微观动力学模型所能描述的生物大分子体系功能过程十分有限. 特别是, 由于蛋白质/核酸等的全原子模型的庞大自由度数以及原子间复杂相互作用导致的粗糙能量面特征, 在合理的计算时间内仅能够模拟亚微秒尺度的蛋白质分子构象运动. 而典型的蛋白质/核酸功能运动时间涉及皮秒、纳秒、微妙、毫秒、甚至秒以上等多个时间尺度(见图1), 因此常规的全原子水平的微观分子动力学方法通常适合于描述局域的小尺度构象涨落运动或极端条件下的快动力学过程[20,21]. 在模拟更长时间尺度的动力学过程时, 人们需要引入粗粒化近似, 建立粗粒化的生物大分子理论模型[8,20,22].

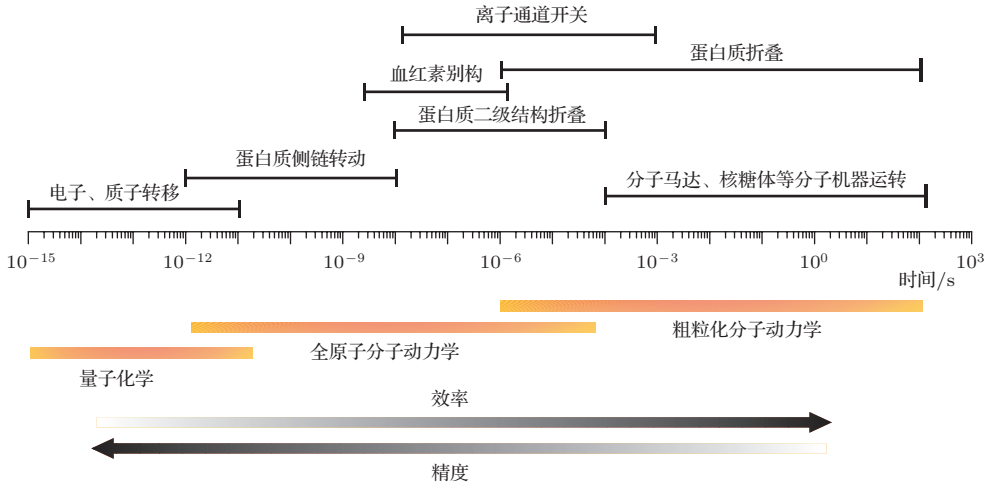

图1 典型生物大分子动力学的时间尺度以及常用模拟计算方法

Fig. 1. Time scale of typical motions of biomolecules and the related simulation methods.

在粗粒化模型中, 通常每个残基由一个(或几个)相互作用粒子代替, 从而很大程度上降低了相关的自由度数[20,23]. 同时, 在粗粒化模型下, 蛋白质功能运动能量面更为光滑. 这些特征使得粗粒化模型具有较高的构象空间采样效率, 因而适用于描述蛋白质分子体系的大尺度、长时间构象运动(图1). 近年来, 在高精度结构数据以及单分子实验观测的基础上, 人们利用粗粒化层次的分子模型揭示了一系列蛋白质/核酸分子体系功能过程的物理机理[9,23–32]. 然而, 建立能够合理反应真实相互作用特征的粗粒化力场是十分困难的. 目前, 人们尚缺乏有效而通用的方法来确定粗粒化模型的力场参数. 这将不可避免地导致对一些关键的特异性相互作用的描述缺乏精确性. 另一方面, 生物大分子体系的功能过程涉及多个层次的相互作用与构象运动, 粗粒化模型由于丢失了原子层次的相互作用和动力学等信息, 因此通常不能够全面刻画蛋白质等生物大分子体系的多尺度耦合动力学特征.

总的来说, 常规的全原子模拟方法具有力场精确度高、采样效率低的特征. 相反地, 粗粒化模型具有采样效率高、力场精度低的特征(图1). 显然, 要实现高精度、高效率的分子模拟, 需要分别克服全原子模型的效率瓶颈和粗粒化模型的精度瓶颈. 理想的做法是, 将全原子模型与粗粒化模型相耦合, 结合全原子力场的高精度特性和粗粒化模型的高效率特性, 实现高效率、高精度分子模拟, 此即为多尺度模型的核心思想[33–47] (见图2), 近年来受到了高分子物理以及生物物理领域科学家的重视, 被认为是最有望突破生物大分子理论模拟精度与效

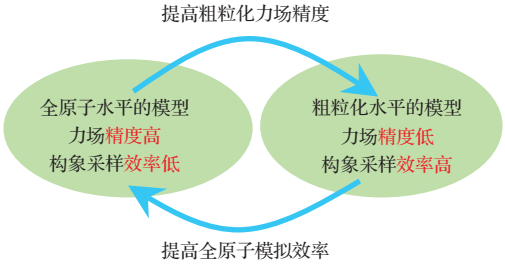

图2 生物大分子多尺度模型示意图

Fig. 2. Schematic of multiscale models for biomolecule simulations.

率瓶颈的计算策略. 通过将全原子模型与粗粒化模型相耦合, 一方面可以提高分子模拟的精度和效率; 另一方面, 可以更真实地刻画原子层次的局域相互作用与大尺度构象运动的耦合, 这对理解一些生物大分子体系的功能过程至关重要.

2 生物大分子多尺度模拟方法简介

结合全原子模型和粗粒化模型的多尺度方法在高分子物理研究中已经是比较成熟的理论方法了. 但是, 类似的多尺度方法在蛋白质、核酸等生物大分子模拟中的应用还处于理论与计算方法的探索阶段. 根据全原子模型和粗粒化模型之间的耦合方式, 我们可以将目前文献中已报道的典型多尺度模型归纳为如下四类: 1) 混合分辨多尺度模型[44,48−50]; 2) 并行耦合多尺度模型[39,41−43,45,51,52]; 3) 单向耦合多尺度模型[34,40,46,47,53−55]; 4) 自学习多尺度模型[35,36]. 以下将对上述四类多尺度模型做简要介绍, 重点强调这些多尺度模型的基本物理思想、模型的优缺点和应用范围, 期望能提供一个关于多尺度模型在生物大分子研究领域应用的概貌, 更详尽的介绍请读者参考相关文献.

2.1 混合分辨多尺度模型

在很多情况下, 生物大分子只有特定区域的原子细节对其生物功能起关键作用, 而其他区域只起到结构支撑作用. 例如, 在酶识别底物时, 其活性位点原子层次的结构和动力学特性相比其他位点更为重要. 在蛋白质分子与靶蛋白形成复合物时, 其界面区域的物理化学性质最为关键. 因此, 在对此类过程做分子模拟时, 可以只对关键区域使用原子细节的模型, 而对其他区域使用粗粒化模型 (见图 3). 这样能够在确保一定精确度的前提下提高计算效率. 基于这一想法, 人们建立了一系列混合分辨多尺度模型. 代表性的工作包括 Neri 等建立的混合 MM/CG (hybrid MM/CG) 模型[44] 以及 Kremer 等建立的自适应分辨 (AdResS) 模型[37,48−50]. 混合 MM/CG 模型主要用于蛋白质活性位点构象涨落动力学的模拟[44]. 其中, 活性位点使用原子细节的分子力场模型, 其他区域使用残基水平的粗粒化模型. 对应的能量函数由下式给出[44]:

$$V = V_{\mathrm{MM}} + V_{\mathrm{CG}} + V_{\mathrm{I}} + V_{\mathrm{MM/I}} + V_{\mathrm{CG/I}} + V_{\mathrm{SD}}, \quad (1)$$

其中, V_{MM}, V_{CG}, V_{I} 分别表示活性位点区域, 粗粒化区域以及二者界面区域的相互作用势. 活性位点区域和界面区域包含所有原子细节, 并使用全原子分子力场. 在粗粒化区域, 每个残基只用 C_α 和 C_β 原子表示, 其粗粒化原子之间的相互作用使用天然态结构为最稳定态的 Morse 势, 以使整个分子体系能够维持在天然态活性结构附近涨落. $V_{\mathrm{MM/I}}$ 和 $V_{\mathrm{CG/I}}$ 项是为实现两种不同模型的自洽连接所引入的交叉相互作用. 最后一项描述随机和摩擦效应. 由于粗粒化模型的计算效率相比于全原子分子力场可以忽略, 因此最终的计算效率取决于只占整个蛋白质分子小部分的活性位点的全原子分子模拟, 从而能够提高计算效率, 扩展分子模拟能够实现的时间和空间尺度范围. 对两个重要蛋白质 (HIV-1 蛋白酶和人类 β 分泌酶) 活性位点结构和涨落动力学的测试模拟结果表明, 该混合 MM/CG 模型能够合理描述蛋白质分子的关键功能运动[44].

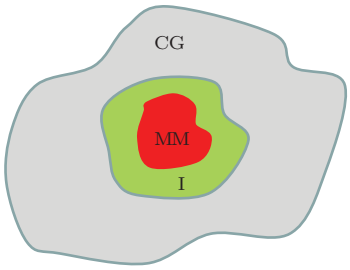

图 3 混合分辨多尺度模型示意图. 其中 CG, MM 和 I 分别代表粗粒化区域, 全原子分子力学区域以及二者界面区域

Fig. 3. Schematic of multiscale models with hybrid resolutions. The CG, MM, and I represent the coarse grained region, all-atom molecular mechanics region, and their interface, respectively.

在混合 MM/CG 方法中, 活性位点区域和粗粒化区域通常是固定的, 不允许两者之间有粒子交换. 然而, 有时我们需要在模拟系统中考虑多个底物或溶剂分子. 由于这些底物或溶剂分子随机地进入蛋白质的活性区域并参与功能过程, 因而需要全原子模型部分和粗粒化模型部分能够实现实时动态交换. Kremer 和合作者提出的自适应分辨模型 (AdResS) 能够合理描述不同分辨模型区域的粒子交换, 并能够根据问题需要在模拟中实时调整全原子模型区域的范围[37,49,50].

事实上, 上述混合分辨多尺度模型非常类似于

量子化学领域人们在描述涉及电子自由度的催化反应中所广泛使用的混合 QM/MM 方法 [56].

2.2 并行耦合多尺度模型

混合分辨多尺度模型要求模拟体系可划分为活性区域和非活性区域. 然而, 很多情况下, 模拟体系并非总能进行如此划分. 这时, 为保证足够精确度, 需要对整个模拟体系考虑全原子细节. 对此类情形, 上述混合分辨多尺度模型不再适用, 而各种并行耦合多尺度模型成为合适的选择 [39,41−43,45,52].

在并行耦合多尺度模型中, 对整个分子体系同时使用高分辨全原子模型和低分辨粗粒化模型. 不同分辨的模型之间通过统计力学原理引入适当耦合, 实现粗粒化模型对全原子模型构象搜索过程的引导. 由于粗粒化模型具有比全原子模型更高的采样效率, 两种模型的耦合能够对全原子模拟的构象搜索过程加速, 提高全原子模拟效率. 代表性的并行耦合多尺度模型包括分辨交换 (resolution exchange) 分子动力学 [41,42,45]、多粒化 (multigraining) 分子动力学 [43]、以及多尺度重要性采样法 (multiscale essential sampling) [51,52].

在 Zuckerman 等所建立的分辨交换分子动力学方法中, 模拟体系同时包含全原子副本和粗粒化副本, 其动力学演化分别由全原子分子力场和给定的粗粒化力场独立支配 [41,42]. 每隔一定时间间隔, 不同分辨的副本之间尝试进行构象交换. 接受交换与否由 Monte-Carlo 方法按如下 Metropolis 判据确定 [41,42]:

$$P(X_i \to X_j) = \min(1, \exp(-\Delta_{\rm AC})), \qquad (2)$$

其中

$$\Delta_{\rm AC} = \beta(E_{\rm CG}(X_j) - E_{\rm CG}(X_i)) + \beta(E_{\rm AA}(X_i) - E_{\rm AA}(X_j));$$

$\beta = 1/k_{\rm B}T$; $E_{\rm CG}(X_j)$ 和 $E_{\rm CG}(X_i)$ 分别表示结构为 X_j 和 X_i 的两个副本在粗粒化能量函数下对应的能量; $E_{\rm AA}(X_j)$ 和 $E_{AA}(X_i)$ 表示结构为 X_j 和 X_i 的两个副本在全原子能量函数下对应的能量. 在计算粗粒化副本的全原子能量时, 需要重建其全原子结构细节. 类似地, 在计算全原子副本的粗粒化能量时, 需要将其约化为粗粒化结构 (图 4). 以上 Metropolis 交换判据能够保证模拟体系在不同副本之间的跃迁满足细致平衡原理, 从而使各对应分辨水平下采样得到的构象满足正则分布. 由于粗粒化模拟能够遍历较大的构象空间范围, 通过构象交换, 实现对全原子副本构象搜索的引导, 从而提高了全原子模拟的采样效率. 在实际应用中, 还可以引入具有不同粗粒化程度的多个中间副本, 通过减小相邻副本的能量差异来提高交换成功率.

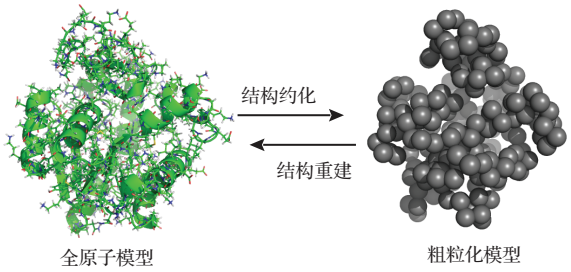

图 4 结构重建和结构约化示意图

Fig. 4. Schematic of the structure reduction from atomistic model to coarse grained model and structure reconstruction from coarse grained model to atomistic model.

在分辨交换分子动力学中, 每次尝试构象交换之前, 需要将粗粒化结构重建出对应的全原子细节 (图 4). 频繁地由粗粒化结构重建其全原子结构将大大降低模拟效率, 是分辨交换分子动力学方法的应用瓶颈之一. van Gunsteren 与其合作者提出的多粒化分子动力学可以避免频繁地进行结构重建 [43]. 在多粒化分子动力学中, 同一模拟体系以一定的权重同时在全原子分子力场和粗粒化力场支配下演化, 其能量函数由下式给出 [43]:

$$\begin{aligned} &V(r^{\rm FG}, \lambda) \\ &= V^b_{\rm FG}(r^{\rm FG}) + (1-\lambda)V^{nb}_{\rm FG}(r^{\rm FG}) \\ &\quad + \lambda V^b_{\rm CG}(M(r^{\rm FG})) + \lambda V^{nb}_{\rm CG}(M(r^{\rm FG})), \end{aligned} \qquad (3)$$

式中, λ 为 (0, 1) 之间的耦合参数, 决定了全原子力场和粗粒化力场的相对权重. $M(r^{\rm FG})$ 为结构约化算符, 将全原子分子结构约化为粗粒化结构. 在模拟中, 可以同时进行具有不同 λ 值的多个副本的分子动力学模拟, 相邻副本之间根据类似 (2) 式给出的 Metropolis 判据尝试构象交换. 其中, $\lambda = 0$ 的系综给出了不依赖于粗粒化力场的构象分布和统计性质.

Moritsugu 等发展的多尺度重要性采样法使用了另一种方法来实现全原子模型和粗粒化模型之间的耦合 [51,52]:

$$V = V_{\rm FG}(r^{\rm FG}) + k_{\rm CG}V_{\rm CG}(M(r^{\rm FG}))$$

$$+k_{\mathrm{FG/CG}}\sum_{i=1}^{L}\left[g_i(M(r^{\mathrm{FG}}))-g_i((r^{\mathrm{CG}}))\right]^2/2, \tag{4}$$

式中, 最后一项是全原子模型和粗粒化模型之间的耦合项, k_{CG} 和 $k_{\mathrm{FG/CG}}$ 分别是粗粒化模型以及耦合项的标度参数. g_i 是由分子结构计算出的几何量, 如粗粒化粒子对距离、二面角等. 粗粒化模型及其力场的选取要求尽量用较少的自由度能够描述重要的构象运动. 通过耦合项, 全原子结构的运动被高效率的粗粒化结构运动所加速, 从而能够在重要的构象空间充分采样. 类似地, 可以通过基于Metropolis的副本交换, 同时模拟多个具有不同标度参数的副本, 来消除由于粗粒化模型的耦合所导致的偏差. 值得一提的是, 在副本交换模拟中, 为保证足够的交换效率, 所需的副本数目随体系的自由度数呈指数增加, 这在很大程度上限制了副本交换方法在大体系中的应用. 在多尺度重要性采样法中, 不同副本的能量差只取决于包含较小自由度数目的粗粒化结构的差异, 因此由能量差所决定的交换效率随系统的尺寸变化较慢, 具有良好的系统标度行为, 因此可应用于较大的分子体系. 该模型能够成功描述小蛋白的折叠以及大蛋白的无序区域由无规结构状态向折叠状态的转变过程[52].

2.3 单向耦合多尺度模型

如本文引言部分所述, 粗粒化模型具有较高的构象采样效率, 但通常缺少有效的手段提取粗粒化力场. 另外, 上述分辨交换多尺度方法对粗粒化力场的精度有较高要求. 根据Metropolis判据, 只有当粗粒化力场与全原子力场的平均行为比较接近时, 才能够保证较高的副本交换效率. 但由于通常的粗粒化力场精确度很低, 与全原子力场的平均行为偏差较大, 使得对于真实系统, 实现高效率分辨交换十分困难, 从而限制了其应用范围. 因此提取较高精度的粗粒化力场非常重要.

以往, 人们在发展粗粒化力场时, 主要采取由实验数据以及高精度理论模型给出的宏观性质来优化粗粒化模型. 近年来, 人们开始使用另一条途径来提取粗粒化力场, 即由全原子分子力场自下而上提取粗粒化力场的单向耦合多尺度模型. 代表性的方法包括Bolzmann inversion[47], 力匹配(force match)[46,57], 涨落匹配(fluctuation matching)[26,40,58,59]以及能量分解(energy decomposition)等[60]. 在Bolzmann inversion中, 对给定的生物大分子做一定时间尺度的全原子分子模拟, 得到对应粗粒化自由度q的统计分布函数$P(q)$, 如粗粒化粒子对的距离分布函数, 粗粒化键角、二面角分布函数等. 通过选取合适的参考态分布函数$P_R(q)$, 由下式给出对应粗粒化自由度的能量函数[47]:

$$V_{\mathrm{CG}}(q)=-k_{\mathrm{B}}T\ln(P(q)/P_{\mathrm{R}}(q)). \tag{5}$$

由此得到的粗粒化力场可用于更大体系或更长时间尺度的分子模拟.

不同于Boltzmann inversion, 在力匹配和能量分解中, 粗粒化力场中的相关参数通过匹配全原子力场给出的粗粒化粒子的受力或粗粒化粒子对之间的相互作用能来确定. 而涨落匹配是通过匹配对应自由度在粗粒化模型中的涨落幅度和全原子模型中的涨落幅度来确定. 在具体应用中, 单向耦合多尺度模型中通常需要同时用到多个方法来由全原子模拟提取粗粒化力场.

最近, 本文作者与合作者利用单向耦合多尺度策略建立了一套"基于原子相互作用的粗粒化(AICG)"模型[34,53−55,61], 并成功应用于具有复杂拓扑结构的蛋白质折叠以及蛋白质大尺度功能运动研究中. 在AICG模型中, 残基用C_α原子表示(也可用$C_\alpha+C_\beta$原子表示), 并使用了如下基于结构的能量函数[53]:

$$\begin{aligned}V=&\sum_I k_b(r^I-r_0^I)^2+\sum_I V_a^I(\theta^I)+\sum_I V_{\mathrm{dih}}^I(\phi^I)\\&+\sum_{J=I+2}\varepsilon_{1,3}^{IJ}\exp\left(-\frac{(r^{IJ}-r_0^{IJ})^2}{2w^2}\right)\\&+\sum_{J=I+3}\varepsilon_{1,4}^{IJ}\exp\left(-\frac{(\phi^I-\phi_0^I)^2}{2w_\phi^2}\right)\\&+\sum_{I>J+3}^{\mathrm{native}}\varepsilon_{\mathrm{nloc}}^{IJ}[5(r_0^{IJ}/r^{IJ})^{12}-6(r_0^{IJ}/r^{IJ})^{10}]\\&+\sum_{I>J+3}^{\mathrm{non-native}}\varepsilon(C/r^{IJ})^{12},\end{aligned} \tag{6}$$

其中r^I, θ^I以及ϕ^I是粗粒化键长、键角以及二面角. r^{IJ}是粗粒化粒子之间的距离. r_0^I, ϕ_0^I和r_0^{IJ}是对应变量在天然态结构中的取值. (6)式中的$V_a^I(\theta^I)$和$V_{\mathrm{dih}}^I(\phi^I)$由蛋白质结构数据库中的无规结构片段通过Boltzmann inversion给出, 能合理描

述序列相关的链柔性[62]. 第4, 5, 6项代表天然态接触相互作用, 即只有在天然态结构中距离比较靠近的残基对之间才有相互吸引的相互作用. 最后一项表示体排斥效应. 以上粗粒化能量函数能够确保天然态结构具有最低能量, 通常称为Gō 模型[29,30,32], 是近年来最为成功和广泛使用的蛋白质折叠模型, 其理论基础是蛋白质折叠的最小阻挫原则[31,63]. 在AICG能量函数中, 蛋白质的氨基酸序列信息主要体现在第4, 5, 6项中依赖于残基对的相互作用系数$\varepsilon_{1,3}^{IJ}$ (1–3局域项), $\varepsilon_{1,4}^{IJ}$ (1–4局域项), 和$\varepsilon_{\mathrm{nloc}}^{IJ}$(非局域项), 并由全原子分子力场给出. 具体地, 对给定的蛋白质分子, 基于全原子分子力场在常温下做一定时间尺度的分子模拟, 并计算粗粒化自由度的涨落幅度. 另外, 基于全原子分子力场, 由能量分解法计算天然态结构中残基对之间的全原子相互作用强度, 称为接触能. (6)式中第4, 5, 6项相互作用系数的相对权重由归一化的接触能给出. 相互作用系数的平均强度根据涨落匹配法由全原子模拟得到的涨落幅度确定. 由此得到的粗粒化模型参数能够合理描述氨基酸的序列信息以及氨基酸链柔性分布. 而这两点特性对正确描述蛋白质折叠和功能运动问题至关重要. 例如, 传统的未显式包含氨基酸链序列信息以及氨基酸链柔性分布的蛋白质折叠模型在描述多域蛋白和打结蛋白折叠时遇到困难. 对具有多结构域的腺甘酸激酶的折叠模拟研究表明, 基于单向耦合多尺度策略所建立的AICG模型能够正确重现单分子实验观察到的多折叠路径行为[54,64]. 另外, 对具有三叶草结构的打结蛋白2 ouf-knot的折叠研究也取得了成功, 能够在很大程度上提高打结蛋白折叠的成功率[54,65]. 随着人们对更为庞大的生物大分子机器关注程度的提高, 此类具有一定精确度和较高效率优势的粗粒化模型将会得到越来越广泛的应用. 目前, AICG模型已成为通用粗粒化生物分子模拟软件CafeMol的重要部分[66], 可以方便地下载使用.

2.4 自学习多尺度模型

上述单向耦合多尺度方法中, 粗粒化力场的精度完全由全原子模拟所确定. 但是, 由于全原子模拟的低效率特性, 其构象采样只能局限在有限的构象空间. 由此提取的粗粒化力场原则上只在全原子模拟所覆盖的构象范围内才是精确的, 这在很大程度上限制了所提取粗粒化力场的应用范围. 由引言中所讨论的粗粒化模型和全原子模型的优缺点, 一个理想的做法是: 所有的构象采样都由粗粒化模型完成, 而所有的能量估算都由全原子模型给出. 基于这一原则构建的算法才能最大程度地体现多尺度模型的优势. 基于该原则, 我们最近发展了一套自学习多尺度算法[34–36], 其主要实现过程由图5给出. 首先, 基于任意的初始粗粒化力场(其精度可能很低)进行粗粒化分子模拟. 对采样得到的粗粒化构象系综重建其全原子结构细节, 并进行短时间尺度的全原子平衡模拟. 对任一粗粒化结构, 可获得一系列相对应的全原子结构和能量. 通过整合所得粗粒化结构、能量以及全原子结构、能量信息, 并利用统计力学原理(reweighting)可将粗粒化构象系综转化为全原子构象系综[67]. 在此基础上, 利用2.3小节所讨论的单向耦合多尺度模型, 可由全原子构象系综提取新的粗粒化力场(自学习过程). 由于此过程有全原子能量信息的输入, 因此所得到的粗粒化力场将比初始的粗粒化力场有更高的精度. 重复以上过程, 直到所得粗粒化力场不再有显著改变, 这时所得粗粒化力场即为最终的优化力场. 基于短肽和小蛋白的测试计算表明, 以上自学习多尺度方法得到的粗粒化力场能够达到和长时间全原子分子模拟直接提取的平均力势相比拟的精确度, 同时能够极大提高计算效率. 由粗粒化分子模拟计算所给出的统计分布和最稳定结构性质与全原子分子模拟结果接近[35,36].

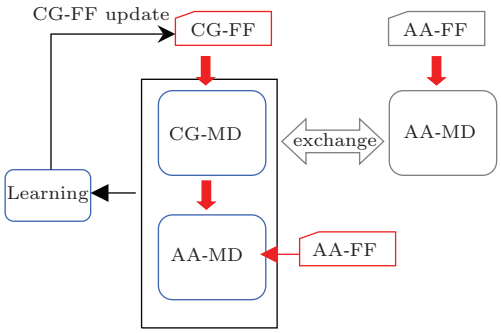

图5 自学习多尺度分子模拟流程图(其中CG, AA, FF和MD分别代表粗粒化, 全原子, 力场, 以及分子动力学)
Fig. 5. Flow chart of the self-learning multiscale simulations. The CG, AA, FF, MD represent coarse grained, all-atom, force field, and molecular dynamics, respectively.

以上提取的较高精度粗粒化力场, 除直接用于粗粒化分子模拟外, 也可用于分辨交换分子动力学模拟中. 由自学习多尺度算法优化的粗粒化力场接近全原子力场的平均行为, 因此在分辨交换分子模

拟中能够得到较高的交换成功率, 从而克服以往分辨交换分子模拟中低交换率困难.

3 结 论

得益于理论方法和计算机硬件的发展, 近年来分子模拟已经成为除实验方法外的另一种研究生物大分子的重要手段, 并已经取得了很大成功. 常规的全原子模型和粗粒化模型分别存在效率和精度瓶颈. 基于多尺度的策略, 人们能够克服常规分子模拟的主要瓶颈, 从而同时实现高精度高效率分子模拟. 目前, 耦合全原子模型和粗粒化模型的多尺度方法在生物大分子模拟中的应用刚刚起步, 但已经展现出其独特的优势和应用前景. 未来一段时间, 人们需要集中解决多尺度模型中尚存在的主要技术困难. 例如, 大多数多尺度方法依赖粗粒化模型的全原子细节重建过程. 目前虽然已有一些重建算法可供使用 [59,68−70], 但结构重建效率仍然是多尺度模型应用的主要障碍. 因此, 人们需要发展更高效率的全原子结构重建算法. 另外, 如何更有效地实现模型之间的耦合仍是需要进一步探索的问题. 随着这些关键技术的改进, 未来多尺度模拟将会在生物大分子结构和动力学研究中得到越来越广泛的应用.

作者感谢京都大学 Shoji Takada 教授在发展生物大分子多尺度理论方面的合作和讨论.

参考文献

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P 2007 *Molecular Biology of the Cell* (1st Ed.) (New York: Garland Science, Taylor & Francis Group)

[2] Abrahams J P, Leslie A G W, Lutter R, Walker J E 1994 *Nature* **370** 621

[3] Sun B, Wei K J, Zhang B, Zhang X H, Dou S X, Li M, Xi X G 2008 *Embo. J.* **27** 3279

[4] Glynn SE, Martin A, Nager AR, Baker TA, Sauer RT 2009 *Cell* **139** 744

[5] Stigler J, Ziegler F, Gieseke A, Gebhardt J C, Rief M 2011 *Science* **334** 512

[6] Lv C, Gao X, Li W, Xue B, Qin M, Burtnick L D, Zhou H, Cao Y, Robinson R C, Wang W 2014 *Nat. Commun.* **5** 4623

[7] Lindorff-Larsen K, Piana S, Dror RO, Shaw D E 2011 *Science* **334** 517

[8] Zhang J, Li W F, Wang J, Qin M, Wu L, Yan Z Q, Xu W X, Zuo G H, Wang W 2009 *Iubmb Life* **61** 627

[9] Levitt M, Warshel A 1975 *Nature* **253** 694

[10] Li W F, Zhang J, Wang J, Wang W 2008 *J. Am. Chem. Soc.* **130** 892

[11] Duan Y, Kollman P A 1998 *Science* **282** 740

[12] Zhao G P, Perilla J R, Yufenyuy E L, Meng X, Chen B, Ning J Y, Ahn J, Gronenborn A M, Schulten K, Aiken C 2013 *Nature* **497** 643

[13] Guo C, Luo Y, Zhou R H, Wei G H 2012 *ACS Nano* **6** 3907

[14] Xie L G, Luo Y, Lin D D, Xi W H, Yang X J, Wei G H 2014 *Nanoscale* **6** 9752

[15] He J B, Zhang Z Y, Shi Y Y, Liu H Y 2013 *J. Chem. Phys.* **119** 4005

[16] Li W F, Zhang J, Su Y, Wang J, Qin M, Wang W 2007 *J. Phys. Chem. B* **111** 13814

[17] Bian Y, Tan C, Wang J, Sheng Y, Zhang J, Wang W 2014 *PLoS Comput. Biol.* **10** e1003562

[18] Inanami T, Terada T P, Sasai M 2014 *Proc. Natl. Acad. Sci. USA.* **111** 15969

[19] Huang Y D, Shuai J W 2013 *J. Phys. Chem. B* **7** 11

[20] Takada S 2012 *Curr. Opin. Struct. Biol.* **22** 130

[21] Vendruscolo M, Dobson CM 2011 *Current Biology* **21** R68

[22] Tozzini V 2010 *Q. Rev. Biophys.* **43** 333

[23] Tozzini V 2005 *Curr. Opin. Struc. Biol.* **15** 144

[24] Xu W X, Lai Z Z, Oliveira R J, Leite V B P, Wang J 2012 *J. Phys. Chem. B* **116** 5152

[25] Yao X Q, Kenzaki H, Murakami S, Takada S 2010 *Nature Commun.* **1** 1116

[26] Moritsugu K, Smith J C 2007 *Biophys. J.* **93** 3460

[27] Marrink S J, Risselada H J, Yefimov S, Tieleman D P, de Vries A H 2007 *J. Phys. Chem. B* **111** 7812

[28] Zuo G H, Wang J, Wang W 2006 *Proteins* **63** 165

[29] Koga N, Takada S 2001 *J. Mol. Biol.* **313** 171

[30] Clementi C, Nymeyer H, Onuchic J N 2000 *J. Mol. Biol.* **298** 937

[31] Onuchic J N, Luthey-Schulten Z, Wolynes P G 1997 *Annu. Rev. Phys. Chem.* **48** 545

[32] Go N 1983 *Annu. Rev. Biophys. Bioeng.* **12** 183

[33] Zhou H X 2014 *Curr. Opin. Struct. Biol.* **25** 67

[34] Li W F, Yoshii H, Hori N, Kameda T, Takada S 2010 *Methods* **52** 106

[35] Li W F, Takada S 2010 *Biophys. J.* **99** 3029

[36] Li WF, Takada S 2009 *J. Chem. Phys.* **130** 214108

[37] Praprotnik M, Delle Site L, Kremer K 2008 *Annu. Rev Phys. Chem.* **59** 545

[38] Liu P, Shi Q, Lyman E, Voth G A 2008 *J. Chem. Phys.* **129** 114103

[39] Liu P, Voth G A 2007 *J. Chem. Phys.* **126** 045106

[40] Chu J W, Ayton G S, Izvekov S, Voth G 2007 *Mol. Phys.* **105** 167

[41] Lyman E, Zuckerman D M 2006 *J. Chem. Theory Comput.* **2** 656

[42] Lyman E, Ytreberg F M, Zuckerman D M 2006 *Phys. Rev. Lett.* **96** 028105

[43] Christen M, van Gunsteren W F 2006 *J. Chem. Phys.* **124** 154106
[44] Neri M, Anselmi C, Cascella M, Maritan A, Carloni P 2005 *Phys. Rev. Lett.* **95** 218102
[45] Lwin T Z, Luo R 2005 *J. Chem. Phys.* **123** 194904
[46] Izvekov S, Voth G A 2005 *J. Phys. Chem. B* **109** 2469
[47] Reith D, Putz M, Muller-Plathe F 2003 *J. Comput. Chem.* **24** 1624
[48] Peter C, Kremer K 2010 *Faraday Discuss* **144** 9
[49] Peter C, Kremer K 2009 *Soft Matter* **5** 4357
[50] Praprotnik M, Delle Site L, Kremer K *J. Chem. Phys.* **123** 224106
[51] Moritsugu K, Terada T, Kidera A 2010 *J. Chem. Phys.* **133** 224105
[52] Moritsugu K, Terada T, Kidera A 2012 *J. Am. Chem. Soc.* **134** 7094
[53] Li W F, Wang W, Takada S 2014 *Proc. Natl. Acad. Sci. USA* **111** 10550
[54] Li W F, Terakawa T, Wang W, Takada S 2012 *Proc. Natl. Acad. Sci. USA* **109** 17789
[55] Li W F, Wolynes P G, Takada S 2011 *Proc. Natl. Acad. Sci. USA* **108** 3504
[56] Warshel A, Levitt M 1976 *J. Mol. Biol.* **103** 23
[57] Thorpe I F, Zhou J, Voth G A 2008 *J. Phys. Chem. B* **112** 13079
[58] Trylska J, Tozzini V, McCammon J A 2005 *Biophys. J.* **89** 1455
[59] Hori N, Takada S 2012 *J. Chem. Theory Comput.* **8** 3384
[60] Gohlke H, Kiel C, Case D A 2003 *J. Mol. Biol.* **330** 891
[61] Li W F, Wang J, Zhang J, Wang W 2014 *Curr. Opin. Struct. Biol.* **30** 25
[62] Terakawa T, Takada S 2011 *Biophys. J.* **101** 1450
[63] Bryngelson J D, Onuchic J N, Socci N D, Wolynes P G 1995 *Proteins* **21** 167
[64] Pirchi M, Ziv G, Riven I, Cohen SS, Zohar N, Barak Y, Haran G 2011 *Nat. Commun.* **2** 493
[65] King N P, Jacobitz A W, Sawaya M R, Goldschmidt L, Yeates T O 2010 *Proc. Natl. Acad. Sci. USA* **107** 20732
[66] Kenzaki H, Koga N, Hori N, Kanada R, Li W, Okazaki K I, Yao X Q, Takada S 1992 *J. Chem. Theory Comput.* **7** 1979
[67] Kumar S, Bouzida D, Swendsen R H, Kollman P A, Rosenberg J M 2013 *J. Comput. Chem.* **13** 1011
[68] Heath A P, Kavraki L E, Clementi C 2007 *Proteins* **68** 646
[69] Gront D, Kmiecik S, Kolinski A 2007 *J. Comput. Chem.* **28** 1593
[70] Canutescu A A, Shelenkov A A, Dunbrack R L 2003 *Protein Sci.* **12** 2001

SPECIAL ISSUE — Celebrating 100 anniversary of physical science in Nanjing University

Multiscale theory and computational method for biomolecule simulations*

Li Wen-Fei† Zhang Jian Wang Jun Wang Wei‡

(*National Laboratory of Solid State Microstructure, Department of Physics, Nanjing University, Nanjing 210093, China*)

(*Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing 210093, China*)

(Received 19 January 2015; revised manuscript received 5 March 2015)

Abstract

Molecular simulation is one of the most important ways of studying biomolecules. In the last two decades, by combining the molecular simulations with experiments, a number of key features of structure and dynamics of biomolecules have been revealed. Traditional molecular simulations often use the all-atom model or some coarse grained models. In practical applications, however, these all-atom models and coarse grained models encounter the bottlenecks in accuracy and efficiency, respectively, which hinder their applications to some extent. In recent years, the multiscale models have attracted much attention in the field of biomolecule simulations. In the multiscale model, the atomistic models and coarse grained models are combined together based on the principle of statistical physics, and thus the bottlenecks encountered in the traditional models can be overcome. The currently available multiscale models can be classified into four categories according to the coupling ways between the all-atom model and coarse gained model. They are 1) hybrid resolution multiscale model, 2) parallel coupling multiscale model, 3) one-way coupling multiscale model, and 4) self-learning multiscale model. All these multiscale strategies have achieved great success in certain aspects in the field of biomolecule simulations, including protein folding, aggregation, and functional motions of many kinds of protein machineries. In this review, we briefly introduce the above-mentioned four multiscale strategies, and the examples of their applications. We also discuss the limitations and advantages, as well as the application scopes of these multiscale methods. The directions for future work on improving these multiscale models are also suggested. Finally, a summary and some prospects are presented.

Keywords: biomolecules, multiscale model, molecular simulations, coarse grained

PACS: 87.15.ap, 87.15.Cc, 87.18.–h, 87.16.A– **DOI:** 10.7498/aps.64.098701

* Project supported by the National Natural Science Foundation of China (Grant Nos. 11174134, 11334004, 11274157, 11174133), and the Natural Science Foundation of Jiangsu Province (Grant No. BK2011546).

† Corresponding author. E-mail: wfli@nju.edu.cn

‡ Corresponding author. E-mail: wangwei@nju.edu.cn