第38卷第8期 中国科学技术大学学报 Vol. 38, No. 8
2008年8月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Aug. 2008

文章编号:0253-2778(2008)08-0961-06

蛋白质结构与动力学的计算机模拟:从方法到应用

刘海燕

(合肥微尺度物质科学国家实验室,中国科学技术大学生命科学学院计算生物学实验室,安徽合肥 230027)

摘要:计算机模拟已发展为根据生物分子结构和动力学阐释生物功能的重要工具.同现有实验比较,计算机模拟不仅能提供结构的时空平均,而且可获得任意微观量的时空分布和演化轨迹.除结构之外,生物功能往往依赖于对动力学的控制.计算机模拟可用于重构构象跃迁路径,发现中间体和过渡态等.本文总结了作者实验室在这一领域的近期工作,特别是关于能量模型、酶催化模拟、构象空间采样等.

关键词:计算机模拟;分子动力学;蛋白质;力场;反应路径

中图分类号:Q71 **文献标识码**:A

Computer simulation of protein structure and dynamics: from method development to applications

LIU Hai-yan

(*School of Life Sciences, and Hefei National Laboratory of Physical Sciences at Microscale, University of Science and Technology of China, Hefei* 230027, *China*)

Abstract: Computer simulation has been developed into an important tool for the elucidation of biological functions from the atomic level structures and dynamics of biomolecules. Compared with current experimental techniques probing atomic level structures, simulations provide not only averages, but also distributions. Besides structures, biological functions often rely on sophisticatedly controlled dynamics of biomolecules, such as the allosteric effects in enzyme catalysis or the effects of ligand-receptor binding in signal transduction. To help understand and eventually control such processes, simulations can be used to reconstruct conformational pathways, identifying intermediates and transition states. This report highlights our recent work in this field. One focus of our research is on developing, testing and refining energy functions for protein simulations, including pure molecular mechanical models for modeling conformational dynamics and hybrid quantum mechanical/molecular mechanical models for modeling enzyme catalysis. Another focus is on developing methods for efficient sampling in the conformational space and for mapping conformational pathways.

Key words: computer simulation; molecular dynamics; protein; force field; reaction pathway

收稿日期:2008-06-28;**修回日期**:2008-07-05

基金项目:国家杰出青年科学基金(30025013),国家自然科学基金(90403120,36070485),中国高技术研究发展(863)计划(2004AA235110,2006AA02Z303)和中国科学院"百人计划"资助.

作者简介:刘海燕,博士/教授.国家杰出青年科学基金获得者,中国科学院"百人计划"入选者.1990年本科毕业于中国科学技术大学生物系,1996年博士毕业于中国科学技术大学生物系.1993~1995年瑞士苏黎世高工联合培养博士,1998~2000年美国杜克大学和北卡罗来那教堂山分校博士后.2001年获中国科学院"百人计划"和国家杰出青年科学基金.主要从事蛋白质结构、动力学的计算机模拟和分子设计的理论方法及实验应用的研究.E-mail: hyliu@ustc.edu.cn

0 引言

蛋白质三维结构与动力学对生物功能至关重要.对绝大多数蛋白质而言,决定其三维结构与动力学特性的全部信息完整地存在于蛋白质氨基酸序列之中.结构生物学的目标,是要解读这些信息,并建立结构、动力学与生物学功能之间的联系.然而,目前适用于大分子结构研究的实验手段受到种种限制.即使在可以通过实验建立原子分辨率三维结构模型时,现有实验手段提供的也仅是时间和空间上的平均结果,难以反映对功能至关重要的动力学过程及机制.

很久以来人们就认识到,同样的物理与化学原理适用于自然界中的生命分子与非生命分子.如同在物理和化学科学中那样,用理论方法分析大分子结构动力学,可以为生物学实验提供重要补充.然而,当我们试图用物理化学原理来解答蛋白质如何折叠、酶分子如何发挥催化功能、生物分子间如何通过特异性相互作用互相识别等问题,试图对序列、结构或功能进行定量预测或理性设计时,却发现我们面临一个巨大的方法或技术鸿沟.生物分子体系由成千上万的原子组成;其结构和动力学通过大量的范德华、疏水、氢键、盐键等物理化学意义上的弱相互作用来维系或支配;在微观上缺乏对称性;在介观或宏观上缺乏均一性.这使得很多经典的理论分析工具在生物大分子研究中失去了用武之地.

计算机分子模拟技术的发展,为填平这道鸿沟提供了机会[1].在基于多维核磁共振或 X 射线晶体衍射实验数据的结构修正中,分子模拟已成为常规手段;基于同源结构的三维结构建模、蛋白质复合物结构预测(如文献[2])、小分子药物设计(如文献[3])等分子模拟技术能够为实验优化设计提供重要信息;近年来,人工蛋白的序列设计、具有催化功能的人工酶的设计等模拟方法与定向进化等实验技术的结合,取得了一系列在生物分子工程领域影响深远的成果,将会对合成生物学等新兴学科领域产生巨大推动作用.

分子模拟在大分子体系中的成功应用,至少依赖于两个方面:一是模拟中能量模型能够准确地刻画与大分子折叠、相互作用、动力学变化相关的分子间相互作用;二是分子动力学或蒙特卡洛等模拟技术能够对大分子构象空间相关区域进行充分采样.后者与目前计算机模拟的时间尺度密切相关:现有计算能力仅限于模拟大分子体系在几十到数百纳秒量级的时间演化,而相当多的生物学过程如蛋白质折叠、酶催化功能运动等在毫秒或更长的时间尺度上发生.为了研究这些过程,必须发展采样加速等技术使得数十到数百纳秒的模拟能够有效模拟采样相关的功能运动过程.从 2001 年以来,计算生物学实验室在蛋白质分子的能量模型、加速采样模拟方法、酶催化路径的量子力学/分子力学优化等方面开展了一系列方法学与应用研究.

1 分子力学势能函数

1.1 有效溶剂化自由能

溶剂效应是蛋白质折叠、相互作用等过程的主要驱动力之一,必须包含在能量模型中.计算机模拟中处理溶剂效应包括两种方案:显式溶剂化模型,即把每个溶剂分子的坐标都包括在能量函数中;隐式模型,即在能量函数中不包括溶剂分子自由度,但包括代表溶剂平均效应的有效能量项.

显式溶剂化模型的优点是对溶质和溶剂使用统一的模型,能够较好刻画分子间氢键等特异性溶剂效应.其缺点包括:计算量大,相当一部分计算时间用于刻画溶剂分子内部的相互作用;大量溶剂分子的涨落给能量、温度或其他热力学量的计算带来比较大的统计误差;对静电等长程相互作用只能近似处理并且需要特殊的模拟技术.隐式溶剂化模型能够克服显式模型的上述缺点,当然其代价是忽略了溶剂的涨落,忽略了特异性溶剂效应,并带有更多的经验性.目前应用于蛋白质体系最成功的隐式溶剂模型是基于数值求解 Poisson-Boltzmann 方程的方法(如基于有限差分的 FDPB 方法).该方法计算量大,不适用于在结构优化或分子动力学模拟中使用.广义波恩模型是对 PB 模型的解析近似,具有恰当的精度,能够解析求导.

我们完成了 GROMOS 分子力学力场下广义波恩模型(GB/SA)的参数化[4]:通过拟合氨基酸侧链类似物的实验溶剂化自由能获得初始参数集;通过考虑模型中参数的物理化学意义,特别是有效溶剂化项对分子内氢键的影响选择参数优化策略;通过对两种不同结构类型蛋白分子的长时间动力学模拟优化参数;通过对第三个蛋白分子的长时间动力学模拟对模型进行了检验.以多个蛋白质错误折叠构象集检验了 GB/SA 模型区分蛋白质正确/错误折

叠构象的能力,并分析发现在错误折叠构象中残基间接触出现频率与天然构象下残基间接触存在性存在一定关联[5]. 该模型在我们的后续工作中反复应用. 我们也完成了在半经验的自洽紧束缚密度泛函模型下广义波恩模型的参数化,可以用线性标度的半经验量子力学模型描述溶液中大分子的结合、构象变化、化学反应等过程[6].

1.2　优化分子力学能量函数对多肽构象平衡的刻画精度

在蛋白质结构模拟中,度量能量函数精度的主要指标,是不同构象态间的自由能差. 这首先需要完善分子体系构象自由能的精确计算方法. 我们对用于计算自用能面的自适应伞形采样方法进行了改进,通过一系列模拟证明,对真空和水溶液中的丙氨酸十肽体系,以该十肽偏离理想二级结构的程度为反应坐标,采用改进后的自适应伞形采样方法能获得收敛很好的自由能面,据此可以计算给定力场下 α 螺旋和 β 折叠两种理想构象状态间的相对自由能[7]. 在此基础上我们以溶液中的丙氨酸二肽的量子化学计算结果为参照,改进了 GROMOS 分子力场中依赖主链二面角的能量项. 对多个蛋白质分子的长时间分子动力学模拟结果证明改进后的能量函数显著减小了用该分子力学力场模拟蛋白质天然结构时的系统误差[8]. 我们还发展了一种新的哈密顿量副本交换分子动力学模拟方法. 证明用该方法能够获得不同长度的多肽的高精度 Ramachandran 自由能面;我们将其应用于量化研究多肽中近邻残基侧链类型和构象对特定残基主链构象偏好性的影响,发现多肽中表现出来的多类近邻效应与对蛋白质结构数据库进行统计得到的效应一致,即局部氨基酸序列的构象偏好性在多肽和蛋白质中具有一定的普遍性[9].

分子力学能量函数的参数化主要是基于拟合小分子体系在不同状态(气相、晶体、液态、溶液等)下的结构、光谱与热力学特性. 由于能量模型具有高维参数空间,不同参数在决定分子性质时相互耦合. 不同参数的协同改变往往能够同样程度地拟合小分子体系性质. 这种变化对大分子模拟的影响往往无法预见. 因此,既有力场在描述大分子结构时的缺陷通常只是在长时间后续应用与实验结果的比较中被发现. 我们提出了将自由能微扰方法与对多肽体系的增强采样模拟技术结合,定量或半定量地预测力场参数在一定范围内变化对刻画多肽构象平衡的影响. 该方法在应用于隐式溶剂化能量模型时取得了合理的结果[10]. 正对该方法应用于显式溶剂化模型进行检验. 利用该方法将有可能高效利用大量的多肽构象平衡的实验数据,对能量模型的参数空间进行约束,将可能提高模型参数化的效率,提高模型刻画大分子体系的精度,而不是局限于用多肽构象平衡的实验数据对模型进行后续检验.

2　蛋白质大范围构象变化、酶催化等功能相关运动的模拟方法与应用

常规的动力学模拟的时间尺度不能涵盖蛋白质大的构象变化、酶催化等功能性运动. 用于克服这一困难的手段包括两类:一类是在不扭曲对逃逸机制的刻画的情况下,设法加速体系从构象空间中不同的能量或自由能极小状态中逃逸的速率;另一类手段适用于已知初态和终态构象,需要重构连接二者的最低能量路径或最低自由能路径的情况. 前者包括需要从单个构象亚态出发,发现体系的其他构象亚态(如可以以不同构象与不同分子结合的蛋白质). 后者包括酶催化过程或蛋白质折叠、蛋白质在不同已知构象间的转变机制等.

2.1　蛋白质大范围构象变化的模拟方法与粗粒化构象采样方法

蛋白质分子正确的低频集合运动自由度一般对应大分子大尺度构象转化(折叠态变化、功能运动等)过程. 我们认为,大分子体系的动力学行为本质上是非谐性的,准谐近似下获得的低频集合自由度只在构象空间局部的能量超曲面上有效. 其他大部分基于集合自由度的加速采样方法大都违背了这一原则. 我们采用粗粒化的弹性网络模型在局部能量超曲面获得低频集合运动自由度空间,在动力学模拟过程中选择性升高体系在低频集合运动自由度空间的模拟温度,从而缩短在原子水平的模拟中大分子体系逃逸局部极小的时间尺度. 其优点包括:采用粗粒化模型,低频集合运动自由度空间可以随动力学模拟过程不断更新,而无需不合理地假定该空间与构象无关;同时,由于模拟过程中高频运动维持常温,保证了采样过程集中在物理上可能的构象转化路径上. 我们分别以溶菌酶结构域间的大规模相对运动、小肽折叠过程为例检验了 ACM(amplified collective motions)方法的有效性[11]. 以 ACM 方法为基础,我们发展了在动力学模拟过程中自动对动

态的低频集合运动自由度空间选择性地重复模拟退火的 ACM-AME 方法,保证体系不会在模拟中长时间陷于局部极小,而是沿物理上可能的构象转化路径进行大范围采样[12].

我们还尝试利用基于数据库的能量模型在更粗的粒度上对蛋白质主链可能构象进行采样.我们以主链二面角为自由度,采用进化算法对多肽链构象空间进行采样.发展了以下技术:①在进化算法中,在评价每一个体适应性之前对适应性函数进行随机优化,一定程度上克服了常规进化算法局部优化能力不足的弱点;②采用小生境技术在适应性函数中考虑同一代个体在构象空间中的分布密度,保持进化过程中个体构象的多样性,避免在构象空间局部采样.这一方法提供了高效率地产生蛋白主链构象的工具,所产生的大量构象符合多肽共价结构对主链局部构象的约束(包括 Ramachandran 二面角分布约束)、肽链在三维空间紧密堆积、满足空间位阻约束.这为我们进一步发展基于知识的能量函数和提供序列设计的候选目标等后续工作打下了基础[13].

2.2 基于已知初态和终态的路径优化方法

在 ACM 基础上,我们发展了用分布式计算对蛋白质结构域折叠或其他大尺度构象变化的连续反应路径进行采样的模拟方法,该方法以构象变化的终态为导向,在物理势能面上获得折叠过程的连续轨迹,避免了将多维空间投影到一个或数个反应坐标.我们以细菌表面免疫原性蛋白 A 的 B 结构域为导向获得了该蛋白从去折叠结构到折叠结构的连续路径.据此我们分析了该蛋白折叠过渡态系综,该模型与实验对过渡态的描述一致,同时揭示出序列上对折叠过渡态系综形成起关键作用的片段[14].

酶催化的高效性、专一性是由酶活性中心催化官能团特有的空间排布、大分子和溶液环境对过渡态的选择性稳定作用决定的.目前,量子力学/分子力学结合模型为阐明酶催化机理、推断过渡态结构提供了有力的理论工具.酶体系中存在大量柔性自由度,它们随催化反应进程的弛豫过程对催化机理有不可忽略的影响.另一方面,它们的存在对根据量子力学/分子力学结合的方法预测合理的反应路径和过渡态模型提出了挑战,近年来在计算化学其他领域中一些较为成功的路径优化方法应用到酶体系中出现了一系列问题.针对酶催化过程的特殊性,我们对目前最为成功的反应路径优化方法 NEB (nudged elastic band)进行了改进,以 β-内酰胺酶为例证明了改进后的 NEB 方法能够克服上述困难.同时,我们发展了一种适用于从头算量子力学/分子力学结合研究酶催化机理模拟的高效率迭代优化方法[15,16].

3 大分子模拟方法的应用

3.1 环境因素对大分子结构与动力学的影响

我们应用计算机模拟方法研究了特定蛋白质体系的结构、动力学及其与功能的关系.通过不同环境条件(不同 pH、不同温度)下的分子动力学模拟推测了人源朊病毒蛋白 PrP 受环境影响从正常构象态转变为致病构象态过程的初始变化事件,以及该过程中一些关键性残基间的相互作用变化[17].用计算机模拟方法比较了同源的非嗜热和嗜热核糖核酸酶 HI 在不同温度下的分子动力学特性.模拟结果表明,无论在低温或高温条件下,嗜热核糖核酸酶都表现出比非嗜热蛋白更大的结构柔性.非嗜热蛋白的柔性呈现出更强的温度依赖性.这一结果不支持嗜热蛋白的结构稳定性可能与其结构刚性相联系的观点.随着温度升高,嗜热和非嗜热蛋白天然结构中的静电相互作用更为稳定.这与升高温度能够减弱溶剂分子对静电相互作用的屏蔽效应的理论相符[18].

3.2 量子力学/分子力学模拟阐明酶催化机制

我们将路径优化方法与从头算量子力学/分子力学结合应用于研究嗜热金属蛋白酶(TLN)大肠杆菌去甲酰化酶(ecPDF)具有不同金属离子偏好性的机制[19].TLN 和 ecPDF 都具有典型的锌离子结合序列模式,二者的活性中心,包括金属离子的配位环境、广义酸碱催化基团的位置等极为相似,可以使用同样的化学步骤催化肽键的水解.然而,TLN 用锌离子作为有活性的催化金属离子,ecPDF 使用亚铁离子作为活性催化离子.我们的计算表明,二者的确使用同样的化学催化步骤,且二者不同的金属离子偏好性,取决于肽键羰基接受被金属离子活化的氢氧根对的亲核进攻时不同的难易程度.后者可能取决于底物羰基与有不同特性的酶环境基团相互作用,而不取决于金属离子的配位环境或其他化学催化基团的位置.

在一项与国外实验室合作的工作中,我们通过了量子力学/分子力学结合模拟提出 4-

Oxalocrotonate Tautomerase 催化过程中不存在广义酸催化官能团[20].

4 展望

生物大分子模拟技术诞生 30 年来,在模拟精度、模拟体系规模、模拟时间尺度等方面都取得了长足发展,在结构生物学研究中得到了广泛应用. 量子力学/分子力学结合的模拟方法得到理论化学界的广泛认可. 随着人们对分子模拟技术的优点与局限性的认识更加深入,分子力学能量函数的进一步优化、采样时间尺度的进一步扩展、粗粒度模拟模型的发展等将拓展分子模拟的应用范围. 基于分子模拟的结构预测技术和 NMR、电镜技术等的结合,将可能为大的或溶液中的蛋白质复合物结构建模提供新的机遇.

分子模拟、分子设计与分子生物学、结构生物学乃至生物功能研究的结合将变得日趋紧密. 这主要包括两个方面,一是对结构、功能等的预测,二是设计具有新的结构或功能的分子. 特别是后者,将设计、合成、功能验证结合为一体来发现具有新结构或功能的蛋白质分子,进而通过实验室定向进化等实现结构和功能优化,正成为这一领域的发展趋势,成为"合成生物学"这一新兴学科领域的一个重要研究分支.

在大分子序列设计方面,从头设计能够折叠成特定结构的蛋白质序列,或者通过序列设计在既有蛋白质结构框架上引入全新的功能,不再是遥不可及的设想. 高通量、高并行 DNA 合成技术的应用,可以同时大量检验不同设计方案,完全控制实验室定向进化的突变方向等. 在理论设计方面,我们建立了能够在多肽片段层次上反应蛋白质序列-结构关系的新的统计能量函数[21]. 下一步将把它和描述长程相互作用的分子力学能量项结合,用于研究结构模块的序列并进行实验验证. 在一项尝试性的工作中,我们首先通过结构分析与定向进化,获得了具有不同操纵子序列识别特异性的乳糖阻遏蛋白(LacR)突变体. 利用两个不同 LacR 二聚体可以识别转录起始位置上恰当距离的操纵子序列,并通过蛋白-蛋白相互作用形成四聚的特性,我们用这些突变体构建并验证了能够按非与(NAND)、非或(NOR)等逻辑规则控制基因表达的逻辑元件. 在后续工作中,我们将结合我们在大分子结构建模方面的经验,和实验科学家合作,尝试建立设计与合成整合的蛋白质序列实验室进化方法,通过序列设计和实验室进化控制蛋白质参与的分子间相互作用,设计和合成能够用作生物网络元件的人工组件.

致谢　文中提及的研究工作是在朱江、张志勇、谢黎、程善美、王骏、杨跃东等前实验室成员和曹占霞、赵铮、李泉、詹剑、丁博、林志雄等共同努力下完成的.

参考文献(References)

[1] Van Gunsteren W F, Bakowies D, Baron R, et al. Biomolecular modelling: goals, problems, perspectives [J]. Angew Chem Int Ed Eng, 2006, 45: 4 064-4 092.

[2] Ding H S, Yang Y D, Zhang J H, et al. Structural basis for SUMO-E2 interaction revealed by a complex model using docking approach in combination with NMR data[J]. Proteins, 2005, 61: 1 050-1 058.

[3] Liu H Y, Duan Z H, Luo Q M, et al. Structure-based ligand design by dynamically assembling molecular building blocks at binding site[J]. Proteins, 1999, 36: 462-470.

[4] Zhu J, Shi Y Y, Liu H Y. Parametrization of a generalized Born/solvent-accessible surface area model and applications to the simulation of protein dynamics [J]. J Phys Chem B, 2002, 106: 4 844-4 853.

[5] Zhu J, Zhu Q Q, Shi Y Y, et al. How well can we predict native contacts in proteins based on decoy structures and their energies[J]. Proteins, 2003, 52: 598-608.

[6] Xie L, Liu H Y. The treatment of solvation by a generalized Born model and a self-consistent charge-density functional theory-based tight-binding method [J]. J Comput Chem, 2002, 23: 1 404-1 415.

[7] Wang J, Gu Y, Liu H Y. Determination of conformational free energies of peptides by multidimensional adaptive umbrella sampling [J]. J Chem Phys, 2006, 125: 094907.

[8] Cao Z X, Lin Z X, Liu H Y. Refining the description of peptide backbone conformations improves protein simulations using the GROMOS 53A6 force field[J]. J Comput Chem, 2008(in press).

[9] Xu C, Wang J, Liu H Y. A Hamiltonian replica exchange approach and its application to the study of side chain type and neighbor effects on peptide backbone conformations [J]. J Chem Theo Comput, 2008(in press).

[10] Cao Z X, Liu H Y. Using free energy perturbation to predict effects of changing force field parameters on computed conformational equilibriums of peptides[J]. J Chem Phys, 2008(in press).

[11] Zhang Z Y, Shi Y Y, Liu H Y. Molecular dynamics simulations of peptides and proteins with amplified collective motions[J]. Biophys J, 2003, 84: 3 583-3 593.

[12] He J B, Zhang Z Y, Shi Y Y, et al. Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions[J]. J Chem Phys, 2003, 119: 4 005-4 017.

[13] Yang Y D, Liu H Y. Genetic algorithms for protein conformation sampling and optimization in a discrete backbone dihedral angle space[J]. J Comput Chem, 2006, 27: 1 593-1 602.

[14] Cheng S M, Yang Y D, Wang W R, et al. Transition state ensemble for the folding of B domain of protein A: A comparison of distributed molecular dynamics simulations with experiments[J]. J Phys Chem B, 2005, 109: 23 645-23 654.

[15] Xie L, Liu H Y, Yang W T. Adapting the nudged elastic band method for determining minimum-energy paths of chemical reactions in enzymes[J]. J Chem Phys, 2004, 120: 8 039-8 052.

[16] Liu H Y, Lu Z Y, Cisneros G A, et al. Parallel iterative reaction path optimization in ab initio quantum mechanical/molecular mechanical modeling of enzyme reactions[J]. J Chem Phys, 2004, 121: 697-706.

[17] Gu W, Wang T T, Zhu J, et al. Molecular dynamics simulation of the unfolding of the human prion protein domain under low pH and high temperature conditions [J]. Biophysical Chemistry, 2003, 104: 79-94.

[18] Tang L, Liu H Y. A comparative molecular dynamics study of thermophilic and mesophilic ribonuclease HI enzymes[J]. J Biomol Struct Dyn, 2007, 24: 379-392.

[19] Dong M H, Liu H Y. Origins of the different metal preferences of E. coli peptide deformylase and Bacillus Thermoproteolyticus thermolysin: a comparative QM/MM study[J]. J Phys Chem B, 2008(in press).

[20] Cisnero G A, Liu H Y, Zhang Y K, et al. Ab initio QM/MM study shows there is no general acid in the reaction catalyzed by 4-oxalocrotonate tautomerase[J]. J Am Chem Soc, 2003, 125: 10 384-10 393.

[21] Li Q, Liu H Y. Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities[J]. Proteins, 2008(in press).