评述

蛋白质空间结构的实验技术和理论方法*

江 凡†

（中国科学院物理研究所 北京凝聚态物理国家实验室 北京 100080）

摘 要 文章主要介绍几种蛋白质空间结构的实验测定方法. 在现代生物学研究中，最常用的方法包括 X 射线晶体学、二维核磁共振(2D-NMR)和低温冷冻电镜. 近几年发展起来的单分子技术在生物大分子动态结构的研究中应用越来越多. 这些方法都有它们特定的时间和空间分辨率，所测定的结构及其动力学受环境热运动涨落的影响也非常不同，文章对这些问题作了较详细的分析. 在蛋白质结构的理论方法方面，介绍了一个新的折叠理论及其与现有折叠模型的关系. 讨论了模拟计算在研究蛋白质构象变化和动力学方面的应用，同时强调了分子动力学和蒙特卡罗方法. 指出粗粒化模型是研究的热点之一，对生物学中经常遇到的多长度多时间尺度问题提供了一个可行的解决方案.

关键词 蛋白质，X 射线晶体学，构象模拟，折叠路径，中间态，瞬态

Experimental and theoretical methods for the analysis of the spatial structure of protein

JIANG Fan†

(*Beijing National Laboratory for Condensed Matter Physics, Institute of Physics, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract This review describes several experimental methods of determining the three-dimensional structures of proteins. In modern biological structure research, the most widely used methods are X-ray crystallography, 2D-nuclear magnetic resonance, and cryo-electron microscopy. The single – molecule technique developed in recent years is finding more and more applications in the studies of dynamics of the biological macromolecules. These methods all have their specific resolutions in temporal and spatial dimensions. The observed structures and dynamics are affected to different degrees and in different ways by the surrounding thermal motions and fluctuations. Therefore, a more detailed analysis of these different effects is presented here. With regard to theoretical methods for protein structure analysis, a new folding theory and its relationship to other folding theories is described. In the discussion of the application of simulation methods in studying protein conformational change and dynamics, both molecular dynamics and Monte Carlo methods are emphasized. It is pointed out that biological problems often involve the treatment of multitime and multilength scale phenomena. Thus, the coarse – grain approach is a feasible solution and a focus of recent development. The author hopes that this review will inspire more collaboration between theoreticians and experimentalists so that we may obtain a better understanding of how protein works.

Keywords protein, X-ray crystallography, conformation simulation, folding pathway, intermediate state, transient state

1 蛋白质空间结构基础知识

1.1 什么是蛋白质

蛋白质是具有特定氨基酸序列的生物大分子，是细胞基因的表达产物，是细胞功能的重要执行者. 21 世纪初，各种重要物种的基因组(全 DNA 序列，包括有功能的基因编码序列和功能未知的“垃圾”序列)的测定工作或已基本完成，或正在开展. 系统

* 国家自然科学基金(批准号：10674172)资助项目
2007 – 01 – 09 收到初稿，2007 – 01 – 20 收到修改稿

† Email：fjiang@aphy.iphy.ac.cn

地解读 DNA 编码,由此理解细胞功能的分子机制和调控机制,不仅有了基本的条件,而且正在成为越来越重要的工作. 由于细胞的大部分功能是由蛋白质来行使的,而蛋白质的氨基酸序列是由基因编码的,因此,解读 DNA 编码的任务之一是研究蛋白质序列、结构和功能三者之间的关系. 例如,为什么某一个蛋白质的序列突变会导致某种癌症?为什么某些病毒能够轻易地侵入人体并进行繁殖?为什么人体免疫系统会失去功效?我们如何更有效地制造疫苗来产生人工免疫反应?这些问题的解决都与蛋白质的研究有关. 所以,蛋白质科学的研究有着重要的科学意义,是后基因组时代完成从基因到破解生命机制的宏伟跨越的关键环节,我国已经把蛋白质科学研究列为“十一五”国家重大研究计划之一.

1.2 为什么需要测定蛋白质分子的(原子)结构和建立蛋白质分子结构的理论模型

蛋白质的化学结构是由 20 种氨基酸线性连接形成的“高”分子链. 链的长度一般在几十到几百残基不等. 链上的相邻氨基酸的氨基和羧基通过脱水反应而形成肽键链,肽链上氨基酸的剩余部分称为残基. 蛋白质链是柔性的,可以采取很多构象. 因此,一个展开的长链可以在特定条件下卷曲成一个直径很小的球形链. 一般我们称展开的状态为“去折叠态”,而称卷曲的状态为“折叠态”. 折叠态也称为“天然态”,因为大部分蛋白质(至少我们现在了解最多的蛋白质)大多只在折叠态才具有正常的生物功能. 在这里,我们说实验测定蛋白质分子的结构,特别是它的原子结构,指的是测定天然态的结构. 蛋白质分子的结构可以帮助我们解释和理解蛋白质是如何行使它的生物功能的. 例如,如何参与催化反应(催化酶),如何把化学能转换成机械能(分子马达)等等.

蛋白质结构是三维的空间结构. 在去折叠态时,蛋白质链上相距较远的氨基酸残基之间的物理相互作用较少. 在折叠态时,则存在很多这样的长程特异相互作用,它们实际上定义了蛋白质的三级结构. 所谓蛋白质结构主要是针对这种长程的特异相互作用而言,它们本质上是三维空间中原子和原子基团的相对位置和取向,测定蛋白质的空间结构就是要通过物理(化学)手段确定这些相对位置和取向. 为什么空间结构重要?因为实验发现蛋白质所参与的化学反应经常取决于反应中心的立体化学特性. 由于蛋白质空间结构的存在和作用,化学反应的机制和路径是被严格地控制的. 蛋白质的作用不仅是提供了一个特异调节的立体微环境,而且直接参与与反应物(底物)的物理化学作用. 因此,有关蛋白质的空间结构信息是理解反应机理的必不可少的重要基础.

实验测定的空间结构不足以描述蛋白质结构的复杂特性. 这是因为实验测定的结构经常是一个静止的结构,这个结构一般对应于一个热力学自由能极小点. 在有些实验条件下,这个静止结构可能是一个在时间或空间上的系综平均. 有些实验手段甚至可以给出系综或系统的分布函数. 但是,我们知道蛋白质结构不是静止的. 蛋白质在溶液中受溶液热运动的影响总是在不停地运动的. 蛋白质的化学结构是共价结构,可以用弹簧模型构建. 蛋白质的非共价结构比较复杂,也突出反映了蛋白质的特性,也是蛋白质空间结构的主要关注点. 非共价结构主要是由非共价相互作用形成的. 这些相互作用包括离子键(即盐桥),电荷相互作用(静电和极化,包括盐桥),氢键,范德瓦尔斯相互作用(体积排斥和形状),疏水相互作用(水分子引起的熵变化),和溶剂效应(溶质引起的水结构变化和溶质与溶剂参与的相互作用). 构建非共价三级相互作用的模型比较复杂,最简单的模型采用离散的二进制,将相互作用分为吸引和排斥(非吸引)两种情况. 相应的非共价三级结构的构建也比较难,最简单的模型仍然采用二进制,将原子或基团的相对位置分为有连接和没有连接(或接触)两种情况. 由此得到的三级结构是一个三维相互作用网路,这也刚好反映了非共价结构的复杂性和流动性. 所以,仅用一个静止结构描述蛋白质的分子结构显然是不够的,我们必须在实验结构的基础上构建可以描述蛋白质复杂特性的理论结构模型.

1.3 蛋白质结构原理及其动力学

蛋白质结构的基本原理是序列决定结构. 这个原理是分子生物学中心法则一个重要组成部分. 中心法则的内容是 DNA 序列被首先转录成 mRNA 序列,mRNA 序列经过加工再被翻译成氨基酸序列,最后,氨基酸序列可以自动地折叠成功能蛋白质. 20 世纪 50 年代,Christian Anfinsen 首先完成了一个经典实验,最先证明一条蛋白质肽链的氨基酸序列包含了决定其天然结构的所有信息[1]. 实验显示,通过改变试管里溶液的条件,蛋白质可以从折叠态过渡到去折叠态,再回到折叠态,以至于完全恢复生物活性. Anfinsen 原理被称为蛋白质结构的热力学原理. 它说明蛋白质结构在给定溶液环境条件下是处于自由能最低点或者至少是一个相对稳定的极小点.

蛋白质热力学原理的发现立刻引出一个蛋白质折叠的动力学悖论. 它首先由 Cyrus Levinthal 于 1968 年提出[2]. 它的基本思路是假定蛋白质的自发折叠是随机进行的,那么蛋白质肽链将尝试分子链中每个单键的所有构象,直到找到接近自由能最小点的构象. 由于需要尝试的构象数目呈指数增长,因此完成折叠的时间会很长,估算结果是一个天文数字. 因此 Levinthal 推断蛋白质折叠不可能是一个完全随机的和反复尝试的过程.

如果蛋白质的自发的折叠过程是按步进行、按级进行的,那么,蛋白质的具体折叠路径是如何由它的氨基酸残基序列决定的呢?对这个问题的回答将有助于理解序列是如何决定结构的,有助于预测蛋白质的结构和功能,有助于发现折叠的中间态和亚稳态,有助于描述结构的动力学等等. 在进一步分析序列和结构的关系之前,让我们来仔细考察一下决定蛋白质结构及其稳定性的几个原子水平的微观相互作用. 它们主要是非共价相互作用. 比较常温下的热力学能量单位(kT 或摩尔单位 RT)与每种相互作用的平均强度会很有意义. $T = 298\mathrm{K}$ 时,$RT = 2.4789$ kJ/mol. 单个共价单键的能量 = 200—450 kJ/mol,水分子的氢键能量 = 20 kJ/mol,蛋白质氢键的能量 = 8—20 kJ/mol,离子键的能量 = 20 kJ/mol,范氏作用的能量 = 4 kJ/mol,疏水作用的能量 = 0—40 kJ/mol. 所以,由这些非共价相互作用形成的结构或构象一般来说是比较稳定的,不会因为热运动的涨落而失去确定的结构,从有序的状态转变为完全无序的状态. 特别是由氢键和离子键形成的局部结构,具有比较确定的几何构型,因此,它们一般比较固定,由热运动引起的转换或交换比较少. 它们有支撑蛋白质结构骨架的作用,使蛋白质具有较固定的和一定刚性的构象. 相对而言,范氏作用和疏水作用对几何构型的要求不是很严格,不同构型的能量可以很接近. 相应的能量差值应该在热运动涨落的量级,使得构象具有一定柔性或流动性,表现出随机涨落运动. 因此,这些相互作用对蛋白质构象的限制相对较小,使能量相近的构象之间的转换或交换变得很容易. 由此可以推论,氢键和离子键可能有固定结构的作用,而范氏作用和疏水作用有稳定结构和降低自由能的作用. 这样一个微观相互作用的变化可能导致整个蛋白质(介观)状态的转变和整体动力学特性的变化. 因此,原子水平的结构在大多数蛋白质研究中,尤其是在与生物功能相关的研究中,应该是相关的、重要的. 如果我们关心的是蛋白质的某个热力学量,那么它对原子水平的结构细节的依赖性要小得多.

1.4 蛋白质空间结构国内外研究动态

在国际上,美国首先提出大规模测定蛋白质结构的计划,现在已经进入第二期的产出阶段. 其他发达国家(欧盟和日本)也相继启动自己的结构基因组计划. 我国也不甘落后. 根据美国第一期的试验计划,发现 X 射线晶体学仍然是测定结构的主要手段,这与预期的结果相符. 过去和现在情况都是这样,蛋白质结构数据库中的 80% 的结构来自 X 射线衍射. 其他有重要贡献的手段有核磁共振和低温冷冻电镜(cryo-EM). 由于这三种方法的重要性,最近几年,它们都有很大的改进. 本文将主要介绍 X 射线晶体学方面的进展.

理论方法的进展也非常快,与实验的结合也越来越紧密. 尤其是蛋白质折叠的机制成为研究的热点和焦点. 有关文献数量巨大,各种理论众说纷纭. 本文将根据作者自己这两三年的工作,提出新的折叠理论. 指出这一理论与现有理论模型的关系,明确阐述折叠中成核过程与疏水塌陷(凝聚或蜷曲)过程之间的关系. 我们可以进一步发现,在蛋白质动力学中,经常涉及到多长度多时间尺度问题. 目前处理这类问题的方法之一是采用粗粒化的蒙特卡罗模拟. 本文将建议进一步发展这类模拟方法,因为它们在相关的物理问题和生物问题研究中都会有重要的应用.

2 测定蛋白质空间结构的实验技术

2.1 蛋白质晶体学和实验步骤

蛋白质晶体学是 X 射线晶体学的延伸. 利用 X 射线衍射测定一个晶体的原子结构有非常长的历史. 它逐步地应用于金属学、固体晶体、化学有机分子晶体和生物大分子. 20 世纪 60 年代左右,重要生物大分子的结构,如 DNA 双螺旋和血红蛋白,对分子生物学和生物化学的发展起到了巨大的推动作用.

X 射线衍射方法的特点是可以确定原子精度的结构. 对于有机分子和蛋白质,可以给出几百到上万个原子的相对坐标. 这些结果与其他方法得到的结果一致,如谱学方法和量子力学计算. 衍射方法的空间分辨率是由 X 射线源的波长决定的. 测得的原子位置的精度还受到晶体衍射能力的限制.

蛋白质晶体的特点是晶胞中含有很多水分子,它们是无序的,与水溶液的液体状态类似. 它们经常

占到晶胞的 50% 到 70%. 这保证了蛋白质的结构与溶液中的结构非常相似. 的确,有实验表明,蛋白质在晶体中仍然可以进行正常的催化反应. 而且蛋白质晶体非常脆弱,很小的温度升高(几度)就可以使晶体融化. 这说明晶格堆积的能量很小,这样小的能量一般不会对蛋白质的结构产生本质的重大改变. 多年的实验结果表明,晶体结构均有很强的生物相关性,非常可靠. 到目前为止,尚未发现任何重要的反例. 二维核磁共振技术发明以来,同样证明溶液结构与晶体结构有很强的一致性. 因此,测定晶体结构是蛋白质结构测定的最重要的手段之一.

通过严格控制溶液条件,蛋白质的中间态和亚稳态结构是可以被结晶的. 蛋白质行使生物功能时一般经历一系列离散的中间态结构. 这些中间态结构一般具有一定的稳定性. 只要稳定自由能远大于 kT(3—5 倍),就可以通过找到合适的结晶条件获得晶体结构. 这是符合热力学原理的. 实验也证明如此,只是结晶条件比较难找. 实际上,大量的结构生物学研究是通过测定蛋白质的一系列中间态结构,以构建蛋白质执行生物功能的动态过程,犹如用结构动画或结构电影演示蛋白质的工作过程和机制. 这种构建结构电影的方法部分弥补了晶体结构是时间平均结构的内在缺陷.

最近几年,由于结构基因组学的大量投入,蛋白质晶体学的实验方法得到飞速的发展. 主要实验步骤如图 1 所示. 第一排 A 方框对应实验第一阶段,即目标基因的选择. 根据研究目的适当选择目标基因不仅可以省钱省时,还可以事半功倍,达到有效地利用有限资源的目的. 目前测定一个可溶蛋白的晶体结构平均仍要花费 10 万到 25 万美元左右. 其他实验方法更加昂贵. 测定膜蛋白的费用目前很难估算. 图 1 的第二排 B 方框表示实验第二阶段,即蛋白质的生产和结晶(图 2 显示几张晶体照片). 蛋白质晶体学的主要瓶颈都在这个阶段. 实验方法的改进和探索主要集中在这个阶段的 4 个 B 方框. 越是后期的步骤自动化程度越高. 比如很多实验室现在配备有结晶机器人. 因此,生产足够量的、可溶的、稳定的、有生物功能和活性的蛋白质是晶体学目前最大的难题. 图 1 中的第三阶段用 C 方框表示,主要包括衍射数据收集和结构解析(蛋白质晶体的衍射图案如图 3 所示). 虽然这些步骤自动化程度已经很高,但新的发明仍然继续出现,例如,从 X 射线光源的改进到数据收集自动化,到晶体相位的确定,到结构模型的构建和优化,等等. 因为,后期步骤的结果和信息可以直接反馈到任何一个前期步骤(图 1,某一方框 A_2, B_2, C_1),对相应步骤加以修饰. 用于这种信息反馈和信息管理的项目管理系统(主要使用 IT 技术)正在应用于晶体学. 总而言之,图 1 中的每一个方框都可以有新的方法和算法的发展,从而改进甚至根本改变晶体学的实验步骤和产出效率.

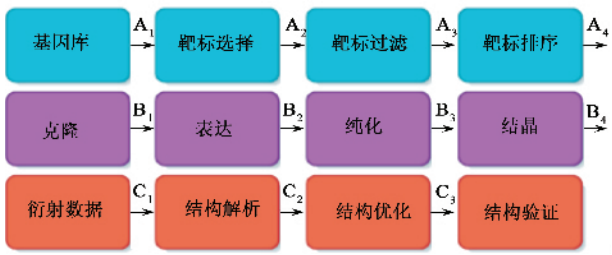

图 1　蛋白质晶体学流程图

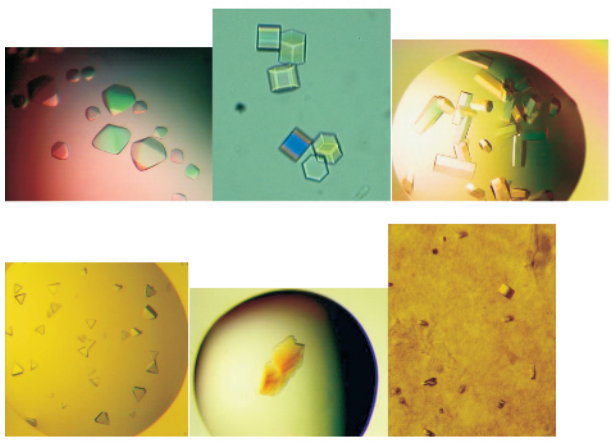

图 2　蛋白质晶体照片

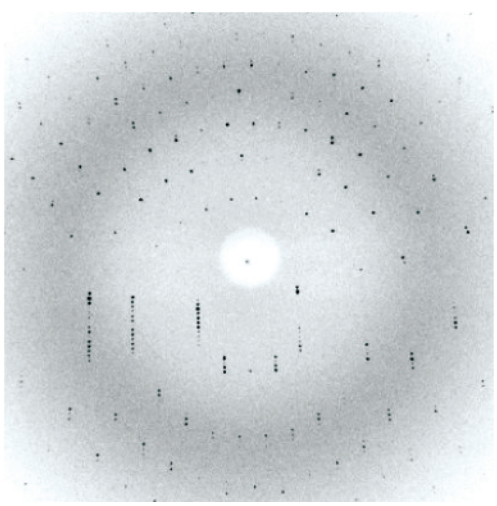

图 3　蛋白质晶体衍射图案

图 1 方框 C 主要涉及计算晶体学. 研究者需要掌握较多较深入的晶体学知识才能正确操作计算软件. 但是由于晶体学的基本原理比较清楚,有严格的数学基础,因此便于编写计算机程序代替人工操作.

随着计算能力的提高,复杂的计算算法得以实现,晶体学计算软件不断向智能化发展,人工干预的必要性在逐渐减小.通过计算可以快速评估实验结果,及时反馈给前期步骤,以提高实验效率.在第三阶段的另一个值得研究的课题是如何从一颗晶体获得它的结构.这是因为蛋白质可能由于偶然因素结晶,但重复性较差.如果是结构生物学研究,能够准确地控制结晶条件也许很重要,但如果是结构基因组学研究,而且目的只是想得到新结构,那么,只使用一颗晶体就能解析结构的能力将非常重要.要想实现这一目标,主要还在于新的计算方法的发展.例如中科院物理研究所范海福发展的蛋白质直接法[3,4]就是一个很有前途的方法.

现在第二阶段的实验步骤中有很多先进技术正飞速发展.方框 B_1 可以使用基因的全 DNA 序列化学合成,这在时间和费用上都是可行的,比 PCR 克隆还要灵活方便.方框 B_2 可以发展体外无细胞表达系统,这样的系统对表达真核物种的蛋白、糖蛋白和膜蛋白都会有很大帮助.方框 B_3 可以使用融合蛋白作亲和纯化和使用荧光标记蛋白检验表达蛋白的可溶性、稳定性和折叠状态.检验蛋白的物理化学和生物化学性质还有其他方法,如质谱、光谱等等.这些步骤的目的都是为了保证结晶条件的筛选.蛋白构象的纯一性是结晶的最好保证.

2.2 中子散射

用于晶体衍射的辐射光源不仅可以是 X 射线,中子也可以是很好的光源.中子的散射截面比较小,穿透能力强,一般适用于弹性或非弹性的散射实验.随着散裂中子源的发展,中子光源的强度得到几个数量级的提高,使得衍射实验变得可行.针对生物大分子(如蛋白质),中子散射的结构信息得益于氢原子与氘原子散射截面的巨大差别,而且符号相反.因此,晶体结构的 X 射线衍射和中子衍射结合起来可以分析蛋白质的氢键,蛋白质表面的溶剂结构,以及与溶剂的相互作用.中子散射与 X 射线散射类似,可以获得蛋白质等生物大分子在水溶液中的聚合状态.更多关于散射方法的内容就不在此详述.

2.3 低温冷冻电镜

低温冷冻电镜(cryo-EM)最近有很大发展.快速冷冻的样品制备方法与传统的干燥方法不同,主要优点是生物样品可以保持在含水的状态下,这更接近天然的生理状态.快速冷冻可以避免冰晶的形成,使生物大分子的结构不被破坏.冷冻速度在毫秒量级,可以用来观察瞬时过程.测定多亚基蛋白质聚合体的三维结构的电镜方法主要有二维晶体的三维重构法和单粒子法.结构分辨率有时可以达到0.4 nm.它们的计算方法一般比较复杂.电镜法可以与来自晶体或 NMR 的原子结构结合,构建庞大生物分子的高分辨率结构.这一方法近年来在结构生物学中的应用备受关注.

2.4 二维核磁共振

二维核磁共振的二维频域谱中的交叉峰反映某个核与相邻的另一核之间的相互作用(见图 4).使用二维核磁作结构测定时,主要使用两种谱 相关谱(correlated spectroscopy, COSY)和 NOESY(nuclear overhauser enhancement spectroscopy)谱.使用 COSY 测量的是通过键(through bond)的自旋 - 自旋耦合,而 NOESY 测量的是通过空间(through space, $< \sim 0.5$nm)的自旋 - 自旋耦合.COSY 可以确定氨基酸序列在二维谱中的峰位,NOESY 则可以确定序列的二级结构和三级结构.核磁共振的优点是在溶液中测定结构,无需生长晶体.但由于交叉峰之间的信号重叠,使得可测定的蛋白质分子量不能太大,一般在 30kDa(Da 的中译名为道尔顿,是原子质量单位,$1\text{Da} = 1.66 \times 10^{-27}$ kg),有的可以达到 40—50kDa.而晶体结构没有分子量的限制.由于同样需要大量的可溶蛋白样品,而且数据收集时间长,因为没有类似于同步辐射光源这样大型的实验装置,核磁方法的产出率是有限的.尽管如此,核磁共振在膜蛋白结构测定中将会有更多的应用.

2.5 单分子技术

最近单分子技术有飞速的发展,如荧光共振能量转移(FRET)、原子力显微镜(AFM)、单分子操作、荧光光谱、拉曼光谱和快速光谱等.单分子技术有一些共同特点,这里主要简单地介绍一下 FRET.FRET 的原理如图 5 所示,受体由于供体与受体之间的能量转移而发射荧光.能量转移对供体 - 受体的距离和取向敏感.发光效率与 Foster 距离有关,一般需 ≤ 0.5nm.被荧光基团标记的生物大分子因构象不同而改变距离,使发光效率不同,因此可以通过探测单光子荧光随时间的变化观测到单个生物大分子的构象随时间的变化.单分子探测的优点是避免了系综平均、时间或空间的平均,可以直接观测单分子的涨落现象.通过建模分析构象的涨落和概率分

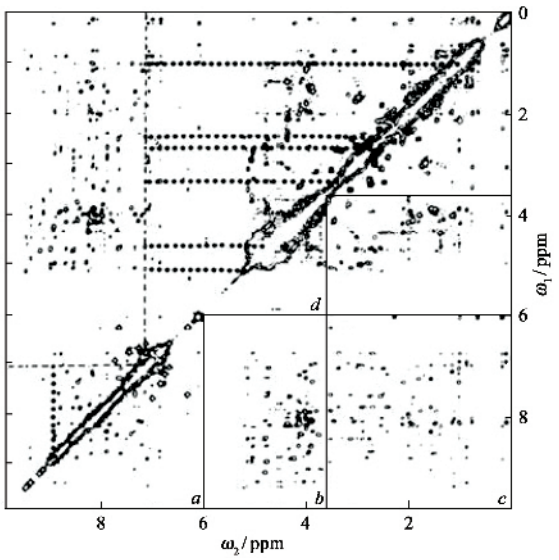

图4　蛋白质二维核磁共振 H-NOESY 谱

布，可以构建或勾画出系统的势景（energy landscape）函数. 系统的概率分布应有一定的特征，它们与系统的特定状态有关，由此可以计算某些事件的概率或速率，如反应速率. 由于构象变化可以实时观测，反应过程的多个途径的随机选取可以被记录下来. 这对理解反应的机理和动力学非常重要. 所以，FRET 方法刚好弥补了晶体学静态结构的缺陷. FRET 方法对数量级为 kT 的能量变化敏感，受环境热运动产生的涨落的影响较大. 在微观尺度，相互作用能量的波动应该在相近的数量级. 当然，还应该考虑非平衡态动力学的作用，尤其他所产生的可能的大的涨落现象还有待进一步研究.

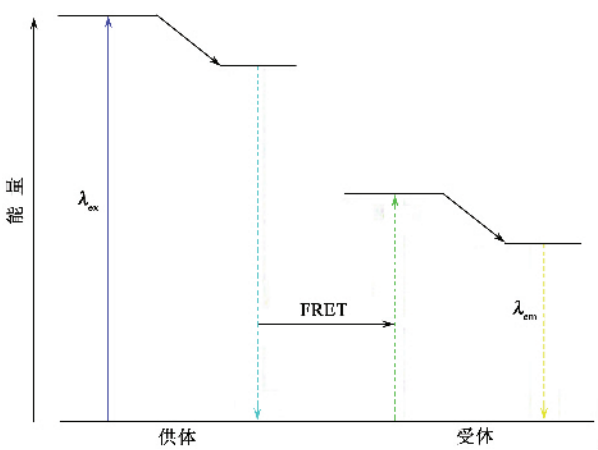

图5　蛋白质折叠理论和蛋白质动力学的模拟方法

3.1　蛋白质折叠的成核理论

最近作者提出一个新的蛋白质折叠的成核假设[5—7]. 这个假设的想法很简单，现表述如下：在蛋白质中有规则的二级结构和复杂的三级结构，前者主要由短程的局域相互作用决定，而后者由长程的全局相互作用决定. 折叠的成核假设主要强调蛋白质结构是逐级形成的，先是局域作用决定的某些二级结构，然后是全局作用决定的三级结构和其他二级结构. 因为局域作用决定的二级结构与蛋白质的总体氨基酸序列没有特别的相关性，这类二级结构中序列与结构关系具有普遍性，应该存在于所有蛋白质序列中. 由于这类二级结构的普适性，它们可以称为所谓的‘折叠密码”. 由此引出蛋白质折叠成核理论的基本假设，即在所有蛋白质序列中存在一些保守的序列段，它们对应的二级结构也具有很强的保守性.

根据这一假设，作者设计了一个新的近邻法来预测蛋白质二级结构. 结果发现，预测准确度与序列的保守程度直接相关[5]. 如图6所示，竖轴是二级结构预测准确度，横轴是预测可信度 Z 值，正比于序列保守度. 从图中可以推断，一个蛋白质序列中的不同部分保守程度不同，因而它们在折叠过程的角色也会不同. 我们发现一个序列最保守的 20% 残基有一个有趣的特性. 它们符合疏水氨基酸被埋藏、亲水氨基酸被暴露的基本规律. 这一规律与公认的 HP 模型是一致的，即蛋白质折叠主要是由疏水相互作用驱动的. 我们在三个独立的结构数据库中验证了这一规律.

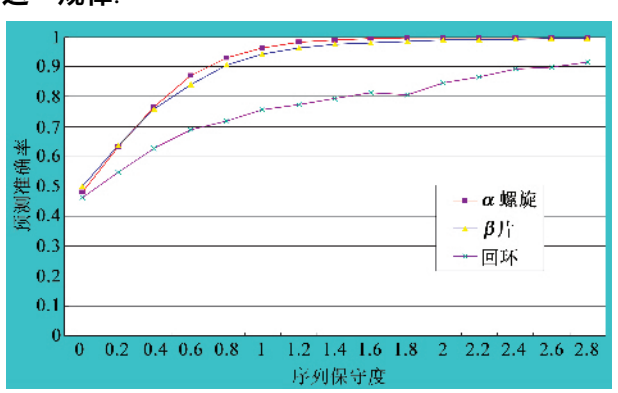

图6　蛋白质二级结构预测准确率与序列保守的关系

我们进一步在蛋白质中发现了两个幂次率，如图7所示[6]. 如果我们把蛋白质的自发折叠看作连续相变的过程，这两个幂次率就不难理解. 其中一个对应成核过程，另一个对应疏水塌陷（凝聚）过程. 计算出这两个过程的空间维数会对理解过程的机制有很重要的启发. 图8为蛋白质在有限尺寸下相变的空间维数. 我们利用高分子物理的标度理论[8]和

Flory 理论[9]，对连续相变的指标作了推导和计算. 结果表明，成核过程是一个三维过程，而塌陷过程是一个二维过程. 而且，前者对应的相关长度较短（~0.5nm），后者则较长（~蛋白分子的大小，>1.0 nm）. 所以这个理论结果很说明问题，符合一般的基本的物理思想，即塌陷过程中蛋白质肽链是凝聚到成核过程已经形成的晶核或模版骨架表面上的. 我们同时发现蛋白质肽链的运动与一个接近理想链的真实链相似，这将帮助我们构建蛋白质的动力学模型. 比如由于蛋白质的稳定度比较低（与理想链模型一致），所以有较大的柔性，比较容易发生构象变化，对微观相互作用的变化比较敏感（见引言部分）.

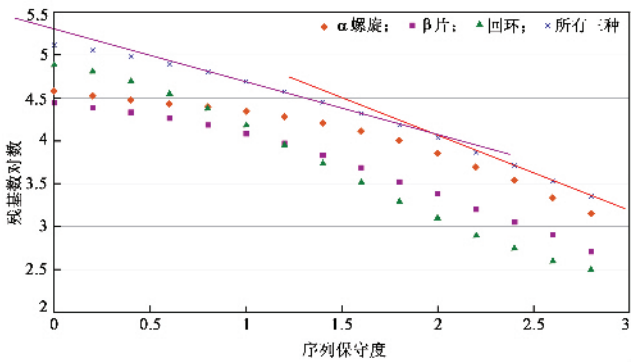

图7　蛋白质的成核与疏水塌陷模型中的幂次率（图中的两条直线分别是两个幂次率）

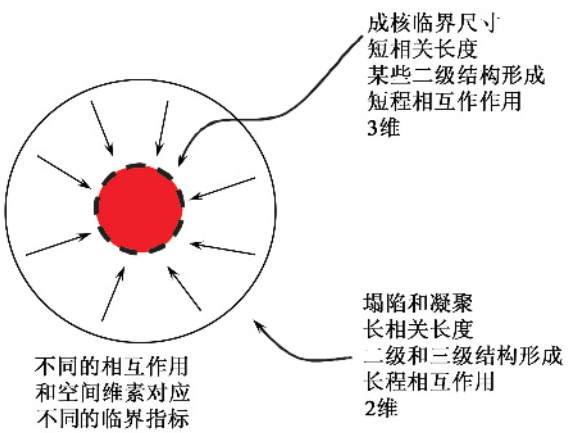

图8　蛋白质在有限尺寸下相变的空间维数

3.2　蛋白质折叠密码

通过以上的论述，我们知道成核假设与折叠密码的概念是一致的. 既然我们已经设计了衡量序列保守度的 Z 值，而且发现 20% 最保守序列在折叠中有特殊的成核作用，我们可以在这些保守序列中寻找折叠密码. 我们使用序列聚类法，结果发现，每种二级结构都有大约 100 个左右的聚类. 它们对应的结构即是结构密码[7]. 我们发现，在 α 螺旋和 β 片的结构密码中，序列与结构之间没有明显的相关性. 这是因为这两种二级结构有规则的氢键模式，不同序列的结构非常相似. 但是它们的结构密码还是有很强的序列模式. 如 α 螺旋有两面性，一个侧面疏水，另一侧面亲水. 相反，回环或转弯的结构密码中存在线性序列与结构的相关性，相关系数为 0.65（见图 8）. 序列越保守则结构越保守，一般具有固定规则的氢键，如 I 类 β 转弯. 因此推论结构保守性一般与固定的氢键模式相关. 这些结构密码一般可以独立存在，形成稳定（高概率）结构，在折叠中起到产生晶核的作用[5—7].

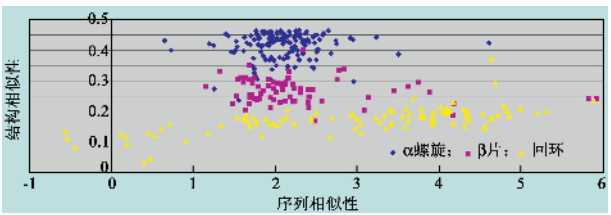

图9　蛋白质结构折叠密码的序列与结构相关性

用一个比喻可以清楚地说明折叠密码的意义和作用. 我们可以把蛋白质序列比作自然语言. 在自然语言中的一个词可以是语法词和主义词. 语法词决定一个句子的结构，而语义词决定这个句子的意思. 对应于蛋白质，折叠密码相当于语法词，它们决定蛋白质的全局结构或三级结构. 而序列的其他部分，非折叠密码部分（相对不太保守的序列部分），决定蛋白质的功能（蛋白质的‘语义’）. 因此，蛋白质折叠密码相当于文言文中的‘之乎者也’. 当然，知道‘之乎者也’还不能读懂文言文，但能够帮助我们开始理解意思. 理解蛋白质功能也要从理解折叠密码开始.

蛋白质折叠的成核理论的重要推论是：蛋白质三级结构是在折叠过程中分级形成的，折叠路径是有限发散的而不是完全随机的，因此我们可以通过计算机模拟和构建的方式预测三级结构.

3.3　分子动力学

蛋白质动态构象模拟的最常用的计算方法是分子动力学. 分子动力学是在相空间（位置和动量空间）中计算系统随时间变化的轨迹. 对系统的系综平均就是对系统在时间轨迹上的平均. 由于分子动力学的广泛应用，相应的算法和软件在高速发展. 计算程序的并行化程度高，优化程度也高. 在蛋白质研究中，常用的软件有 CHARMM，AMBER，GRO-

MACS 和 NAMD[10] 等. 这几个软件包有较好的用户支持和用户群体. 软件说明和入门辅导也比较详细清楚. 这些软件对微机机群一般有很好的优化. 最近,发明了一种新的分布式计算模式,称为@ home 技术,例如 folding@ home[11], rosetta@ home,主要用于蛋白质折叠的长时间模拟计算和结构预测. 分子动力学的算法和势能函数仍在不断创新,例如复本交换方法[12]的应用也越来越多,能够利用小型微机集群有效地对构象空间和能量势景进行取样. 再例如,最近发明的 GB 模型(generalized Born model),可以在不使用水分子的情况下(即隐含溶剂模型)准确计算溶剂效应[13]. 这样可以大大减少被模拟系统的自由度,使运算速度加快. 总而言之,分子动力学已经成为蛋白质结构模拟的标准工具,而且仍然有待进一步改进.

3.4 蒙特卡罗方法

蒙特卡罗方法是另一种常用的在相空间中对系统取样的方法. 它的特点是只对系统中原子的位置取样,而不关心原子的速度. 通过计算蛋白质处于每一种构象的能量来计算系统物理量的系综平均. 通过蒙特卡罗方法可以获得各种系综平均的取样[14],一般常用的是正则系综(或等温系综). 蒙特卡罗方法在模拟某些系统时,有很强的灵活性,比较容易引进随机过程,特别是当使用粗粒化模型在构象空间中取样时.

3.5 构象空间的粗粒化

在生物大分子的模拟计算中,经常会遇到多长度、多时间尺度的生物现象和过程. 由于必须跨越几个数量级的时间尺度,才能观察到生物大分子的宏观效应(或介观效应),使得目前最快的计算机无法进行这样的计算. 所以,我们必须设计和发展新的算法,这是当前软物质物理和生物大分子模拟计算研究的一个热点[15]. 例如,我们可以对蛋白质的构象进行合理的离散化. 对晶体结构的分析表明,这是有根据的和可行的. 蛋白质的侧链分布有一定的聚集特征[16],因此可以用有限的离散态表示. 蛋白质的主链的二面角同样有非常聚集的分布[17],也可以用离散态表示. 对于不同时间尺度的模拟可以使用分级和分块模拟的方法,问题是如何整合所有模拟的结果,从而获得系统的完整物理图像. 在使用粗粒化模型时,不管是分子动力学模拟还是蒙特卡罗模拟,都需要设计与模型相匹配的势能函数. 势能函数的构建是生物大分子模拟计算中的另一个重要研究热点.

4 结束语

综合上述,我们可以看到,每一种实验方法都有它的优点与缺陷. 例如每个实验方法的时间和空间分辨率不同. 要想获得蛋白质结构的完整图像,必须集中几种实验方法的结果,使它们相互衔接. 它们之间的缺失部分只能通过计算模拟和理论模型来构建和预测. 因此发展有关蛋白质的一般性理论是非常重要的. 我们希望通过理论与实验的结合,不仅可以理解某个蛋白质的一般的物理化学性质,还可以理解它的特殊的生物功能. 所以计算模拟是不可或缺的预测工具之一.

参 考 文 献

[1] Anfinsen C B. Science, 1973, 181: 223
[2] Levinthal C. J. Chim. Phys., 1968 65: 44
[3] Fan HF, Gu YX. Acta Cryst., 1985, A41: 280
[4] Wang J W, Chen J R, Gu Y X *et al*. Acta Cryst., 2004, D60: 1244
[5] Jiang F. Protein Eng., 2003, 16: 651
[6] Jiang F. Sci. & Technol. Adv. Mater., 2005, 6: 860
[7] Jiang F, Li N. Chin. Phys., 2007, 16: 392
[8] de Gennes P-G. Scaling concepts in polymer physics. Ithaca and London: Cornell University Press, 1979
[9] Flory P J. Principles of Polymer Chemistry. Ithaca: Cornell Univ. Press, 1953
[10] http://www.charmm.org, http://amber.scripps.edu, http://www.gromacs.org, http://www.ks.uiuc.edu/Research/namd/
[11] http://folding.stanford.edu
[12] Mitsutake A, Sugita Y, Okamoto Y. Biopolymers, 2001, 60: 96
[13] Qiu D, Shenkin P S, Hollinger F P *et al*. J. Phys. Chem., 1997, 101: 3005
[14] Binder K, Heermann D W. Monte Carlo Simulation in Statistical Physics. New York: Springer-Verlag, 1992
[15] Karttunen M, Vattulainen I, Lukkarinen V. (Eds.) Novel methods in soft matter simulation. Berlin Heidelberg: Springer-Verlag, 2004
[16] Ponder J W, Richards F M. J Mol. Biol., 1987, 193: 775
[17] Nelson D L Cox M M. Lehninger Principles of Biochemistry. 3rd ed. New York: W. H. Freeman and Company, 2005