科学通报 2016 年 第 61 卷 第 24 期: 2670 ~ 2680
《中国科学》杂志社
SCIENCE CHINA PRESS
Science125个科学前沿问题系列解读(XII)
Can we predict how proteins will fold?

蛋白折叠预测

马彬广

华中农业大学信息学院, 武汉 430070
E-mail: mbg@mail.hzau.edu.cn

2016-06-01 收稿, 2016-06-29 修回, 2016-06-29 接受, 2016-08-02 网络版发表
国家自然科学基金(31570844, 31100602)资助

摘要 蛋白折叠是指蛋白质分子从线性高分子形成有生物学功能的三维结构的过程. 迄今为止, 蛋白折叠已经有50余年的研究历史, 成为一个十分宽广而活跃的研究领域. 本文简要回顾了蛋白折叠问题的提出背景和研究历程, 并从蛋白折叠过程预测、折叠过程相关参数的预测、蛋白折叠结果预测、折叠结果相关参数的预测等四个方面介绍了蛋白折叠问题的研究进展. 最后, 指出了蛋白折叠研究领域值得关注的一些未来的发展方向.

关键词 蛋白折叠, 蛋白折叠模拟, 蛋白结构预测, 深度学习, 系统生物学

自从最早的蛋白质结构被解析以来[1,2], 这种看上去与传统晶体迥然不同、近乎无规则的结构是怎样形成的, 一直是令生物物理学家着迷的问题. 从组成上来看, 蛋白质是由20余种氨基酸通过肽键连接而成的高分子, 那么, 第一个问题就是, 这种多肽链与其三维结构之间是否存在确定的对应关系? 20世纪60年代前后, Anfinsen等人[3]通过对核酶蛋白的复性研究, 提出了蛋白结构形成的热力学观点; 该观点认为, 蛋白质的天然结构具有热稳定性, 对应于自由能的一个全局最小点, 而蛋白质的结构信息可以由其氨基酸序列所决定(Anfinsen Principle)[4]. 后来, Levinthal[5]在分析了这种热力学观点之后注意到, 蛋白质高分子的构象数目非常之多, 即使对较小的蛋白(比如101个氨基酸), 若以各态遍历的方式找到能量的最小点, 所需时间也超过了宇宙的年龄[6], 明显与实际情况不符. 因此, 蛋白质结构的形成过程不会只是受到热力学规律的支配, 一定还存在动力学机制[7].

蛋白折叠问题, 就是指蛋白质分子如何从线性高分子形成有生物学功能的三维结构的问题. Dill等人[8,9]曾将蛋白折叠问题表述成3个相关问题: (ⅰ) 物理折叠码问题, 蛋白质的天然三维结构是如何由其一维氨基酸序列所编码的物理化学性质所决定的; (ⅱ) 折叠机制问题, 蛋白质是如何从一维的线状分子快速折叠成三维天然结构的; (ⅲ) 蛋白结构预测问题, 能否从蛋白质的氨基酸序列预测其三维结构. 这3个相关问题反映了认识蛋白折叠的不同角度. 对于*Science*在2005年发表的125个科学问题中的"蛋白折叠能否被预测的问题"[10], 答案是肯定的. 鉴于蛋白折叠是一个宽广而活跃的研究领域, 本文不可能也无意覆盖此领域的方方面面, 而是仅就作者的理解对上述问题的解答给出线索. 本文将从蛋白折叠过程预测、折叠过程相关参数的预测、蛋白折叠结果预测、折叠结果相关参数的预测等4个方面对*Science*所提出的问题进行解答, 并指出蛋白折叠研究领域未来可能的发展方向.

1 蛋白折叠预测的研究

1.1 折叠过程: 蛋白折叠模拟

蛋白折叠模拟是指在计算机中模拟蛋白质从无

引用格式: 马彬广. 蛋白折叠预测. 科学通报, 2016, 61: 2670–2680
Ma B G. Protein folding prediction (in Chinese). Chin Sci Bull, 2016, 61: 2670–2680, doi: 10.1360/N972016-00658

© 2016《中国科学》杂志社

结构的伸展状态到有结构的天然构象的折叠过程(而模拟相反的过程称为“去折叠”模拟). 在20世纪70年代前后, 蛋白折叠问题刚刚提出之时, 计算机辅助的蛋白质建模与折叠过程模拟就已经开始了[7,11]. 对蛋白折叠过程的模拟, 不仅可以检验蛋白质折叠理论[12], 也可以辅助蛋白质结构和功能设计[13~16], 是蛋白折叠研究领域不可或缺的手段.

当前的蛋白折叠模拟, 按蛋白模型的精细程度, 大体可以分为3类[17]: (ⅰ) 格子模型(lattice model), 其中, 每个氨基酸被表示成一个珠子, 其运动限定在空间的网格中, 且一般来说, 氨基酸的属性也被简化成亲水和疏水2种类型; 该模型虽然简单, 但对于认识蛋白质的可设计性起到了重要的作用[18]; (ⅱ) 去格子模型(off-lattice model), 这种模型也是把氨基酸表示成珠子, 但其运动不再受限于空间网格, 而是可以在空间中自由移动; 此类模型中, 通常会考虑氨基酸之间的键、角和范德华力等作用; (ⅲ) 全原子模型(all-atom model), 这类模型包含了每个氨基酸中的原子细节, 也包括离子和水分子; 通常, 所有的物理相互作用, 包括键长、键角、扭角、静电力和范德华力等都会包含在模型中, 有时甚至会加上极化效应、配位键、质子化过程和电荷转移过程. 一般来说, 模型越精细, 越能提供细节知识, 但模拟所需的计算资源也就越多.

蛋白折叠的模拟, 一方面可以促进对蛋白折叠过程的理解, 另一方面, 也是基于对蛋白折叠机理的认识实现的. 对蛋白折叠机理的认识, 经历了从热力学稳态到动力学途径再到折叠能量漏斗这样一个过程[19,20]. 现在普遍认可的是折叠漏斗模型, 因为它较好地综合了前人的观点. 折叠漏斗模型的要点如下: (ⅰ) 蛋白质的天然状态对应于一个能量极小点; (ⅱ) 蛋白质非折叠态的数目要远大于折叠态的数目; (ⅲ) 蛋白质折叠的能量曲面总体是粗糙的, 可能存在多条折叠途径; (ⅳ) 因折叠途径的不同, 可能采用单态折叠(downhill folding, 无过渡态)、两态折叠(有过渡态、无中间体)和多态折叠(有过渡态和中间体)等形式. 至于具体的折叠机制, 主要有3类模型: (ⅰ) 层级折叠模型, 即, 蛋白折叠过程中先形成局部的二级结构, 而后, 这些二级结构元件扩散和碰撞, 最终形成三级结构; (ⅱ) 疏水塌缩模型, 即, 蛋白折叠过程中, 在疏水力的作用下, 先形成熔球体的三级结构, 局部的二级结构是在三级结构的驱动下形成的; (ⅲ) 成核凝缩模型, 即, 前面2种极端情况的混合, 二级结构和三级结构同时形成[21]. 不同的蛋白质依据不同的条件, 可能采取不同的折叠机制. 因此, 目前的观点是, 蛋白的折叠机制不唯一, 在层级折叠和成核凝缩之间存在连续统[22,23].

随着计算技术的进步和计算机性能的提高, 蛋白折叠模拟的硬件资源也在发生着变化. 蛋白折叠模拟的硬件平台经历了从最初的单机模拟到并行化和分布式计算的发展过程, 且大部分已经完成的蛋白折叠模拟研究都是基于多核(CPU)并行的计算机集群实现的. 2000年前后, 斯坦福大学的Pande研究组推出了蛋白质折叠模拟的分布式计算程序Folding@home[24], 将大规模的蛋白折叠模拟问题分割成小块任务; 全世界感兴趣的用户都可以下载其计算机屏幕保护程序, 利用闲暇的计算资源运行分块的模拟任务, 并将计算结果自动上传到斯坦福大学的服务器进行汇总, 解决蛋白折叠模拟中的问题; 迄今, 该项目已发表了100多篇研究论文. 2010年前后, Shaw研究组[25]推出了专门针对蛋白模拟(分子动力学)问题而设计的计算机平台ANTON, 显著提高了蛋白折叠模拟的蛋白大小和时间长度[26]. 此外, 通用图形处理单元(general-purpose GPU)计算技术的发展使得基于GPU的“众核”模拟程序和CPU、GPU联合使用的异构运算平台的开发成为蛋白折叠模拟领域的新研究方向[27,28], 而此方向上着重要解决的是GPU内存小且与CPU之间通讯存在速度瓶颈所带来的算法设计问题.

同样, 蛋白折叠模拟的方法和软件也在不断推陈出新[29~32]. 当前, 最流行的模拟方法是基于全原子模型的分子动力学(MD)模拟方法, 它涉及3个主要问题: 准确的力场模型、充分有效的构象空间采样方法和稳健的模拟数据分析方法[33]. 分子动力学模拟中最为常用的力场(所谓“力场”, 可以理解为原子间相互作用的模型)是CHARMM和AMBER系列[34~36]. 这两种力场模型, 虽然实现原理上不尽相同, 但都能对蛋白的折叠态结构和折叠速率给出比较准确的预测, 但对“去折叠”过程中的一些动力学细节以及去折叠状态下结构性质的预测, 与实验相比还有一些差距, 比如, MD模拟所给出的去折叠状态, 与实验比较起来, 总是更为紧凑, 含有更多的螺旋构象等[37]. 当前, 如何在这两个力场中加入和完善原子的极化模型是一个重要的发展方向[35,38,39]. 除了蛮力

的动力学模拟之外, 充分有效的构象空间采样方法有助于加快模拟过程和提高处理问题的规模. 当前的分子动力学模拟中, 采样方法大体上可以归结为两类[40]: 连续方法和离散方法, 前者除了蛮力的分子动力学模拟之外, 还包括转换途径采样、转换界面采样、前向流采样和加权系综采样等方法, 而后者主要有马尔科夫状态模型(MSM)和里程碑方法(Milestoning)两种. 对于大分子体系来说, 分子动力学模拟所产生的数据也是生物大数据的一种. 如何从这些大数据中, 提取信息, 转化成人类的知识, 也是蛋白折叠模拟研究中的一个重要问题. 当前的模拟数据分析方法主要有两种: 反应坐标方法和马尔科夫状态模型方法[33], 前者是一种传统方法, 把整个蛋白折叠过程标注在一维的反应坐标轴上, 可以定义过渡态等概念, 而后者是近几年才出现的新方法, 可以较好地反映多条折叠途径. 这两种方法不仅反映了对蛋白折叠过程的不同理解, 而且其分析效果也各有优劣, 可以就所研究的问题类型, 酌情采用.

蛋白折叠模拟中的一个关键问题就是如何突破局部的能量障碍, 找到能量的全局最小点, 从而加快折叠模拟的速度. 针对这一问题, 近年来也出现了一些解决办法. 在2000年前后, 出现了多副本交换(replica exchange)方法[41,42]; 该方法在不同温度下对同一体系构建多个模拟副本, 利用低温副本获得过程细节, 而用高温副本突破能量障碍, 通过高温和低温条件下构象信息的交换来达到加速蛋白模拟的目标. 后来, 又出现了多尺度建模(multi-scale modeling)的模拟方法, 它的核心思想是把高分辨率尺度上的全原子模型和低分辨率尺度上的粗粒化模型相耦合, 从而得到高精度且高效的模拟; 此类方法包括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和自学习多尺度模型等几类[43]. 近年来, 把外部信息和经验规则引入蛋白模拟过程, 从而提升蛋白模拟的规模, 加快蛋白模拟的速度, 则代表了另一个重要的发展方向. 例如, 最近出现的MELD (modeling employing limited data)方法, 就是利用贝叶斯推断把一些外部结构信息和启发式的经验规则引入蛋白折叠模拟, 可以显著加快构象搜索的速度, 且依然可以给出准确的自由能变化[44,45]. 另外, 将人的智力引入蛋白折叠模拟和结构预测(如Foldit游戏[46])也是一个有趣的尝试.

近20年来, 蛋白折叠模拟的能力在逐渐提高. 1998年, 首次实现了对35个氨基酸的绒毛蛋白(villin)的全原子显式溶剂模型的模拟[47], 尽管只产生了1 μs的折叠轨迹, 而未到达最终的折叠状态, 但却是全原子模拟技术发展过程中的一个重要里程碑. 2011年, Shaw研究组[26]基于模拟专用机器ANTON对12个常用模型蛋白的折叠过程的模拟则代表了另一个重要的里程碑. 据统计, 在过去几十年中, 蛋白折叠模拟能力的增长速度是大于摩尔定律的, 因此, 蛋白折叠模拟能力的提高, 不仅仅是由于硬件水平的提升而带来的, 还有模型和算法改进方面的贡献[29]. 现在, 对于几十个氨基酸组成的蛋白, 其折叠过程的模拟已经可以到毫秒的长度, 而有效模拟的最大蛋白已经达到了100个氨基酸左右的长度[33,37]. 当前的模拟能力已经可以模拟预测PDB数据库中大约10%的单链蛋白结构[37]. 据估计, 若仅依靠遵循摩尔定律的硬件计算能力的增长, 将需要约25年才能实现对长度为140个氨基酸的蛋白的折叠模拟, 但若将外部信息和经验规则引入模拟过程, 则有望将这一目标缩短为5年的时间[29].

除了蛋白模拟尺度的稳步增长之外, 蛋白折叠模拟的研究对象也越来越多. 例如, 除了模拟单链的模型蛋白之外, 小分子与蛋白乃至蛋白与蛋白之间的相互作用, 也是近几年来蛋白折叠模拟的研究对象[29,48]. 蛋白复合物形成过程的模拟, 涉及折叠过程(folding)与绑定过程(binding)的相互耦合, 不仅需要更多的计算资源, 也需要新的理论方法和建模手段. 除了体外折叠过程的模拟之外, 体内折叠过程的模拟, 比如, 与翻译过程偶联的蛋白折叠(co-translational folding)[49]和分子伴侣协助下的蛋白折叠[50], 也是近年来重要的研究领域[51,52]. 此外, 膜蛋白通道与转运蛋白的模拟研究[53]、配体诱导的蛋白折叠[54]以及金属离子参与的蛋白折叠[55]等, 则代表了另外一些重要的研究方向. 总之, 蛋白折叠模拟研究的问题多样性也在逐步提高.

1.2 折叠过程: 过程相关参数预测

蛋白折叠过程研究的最重要目标就是要理解蛋白为何折叠得这么快[8], 因此, 折叠速率决定因素的研究以及折叠速率的预测, 必定是引人注目的研究方向. 蛋白折叠速率可以直接用蛋白折叠模拟的方法进行预测, 但所需计算资源巨大, 因此, 近年来出现了很多统计和机器学习类的预测方法. 与蛋白折

叠速率关系最为密切的因素是蛋白的序列长度[56,57]. 除此之外, 还有哪些因素决定了蛋白的折叠速率? 解答这一问题的一个重要突破是1998年接触序参数(contact order)的发现[58]. 对于当时已知折叠速率的蛋白, 基于天然结构定义的一个简单的拓扑特征参数, 接触序参数, 就可以给出很好的折叠速率预测. 这一发现启发人们从蛋白天然结构的角度去理解和预测蛋白的折叠速率. 后来, 又出现了类似的参数, 如长程序参数(long-range order)[59]、总接触距离(total contact distance)[60]等.

因为大部分蛋白的三维结构是未知的, 基于蛋白三级结构的折叠速率预测受到很大的制约. 2004年, Ivankov等人[61]在*Proc Natl Acad Sci USA*上发表论文, 提出了"有效折叠长度"的概念, 而在其有效折叠长度的定义中, 最重要的因子是螺旋二级结构的含量, 因此, 可以从序列出发, 先预测蛋白的二级结构, 再进一步预测折叠速率. 于是, 各类二级结构对蛋白折叠速率的影响研究以及基于蛋白二级结构的折叠速率预测方法逐渐出现[62,63]; 例如, 研究表明, 两态折叠先形成α螺旋和β折叠片, 然后再缓慢形成loop结构, 而多态折叠则相反, 先形成稳定的loop结构, 再缓慢形成α螺旋和β折叠片[62]; 随后, 通过简化蛋白二级结构的类型集合, Huang等人[63]发现, 只需3种二级结构类型的组合, 就可与蛋白的折叠速率高度相关, 它们是: 延伸的β串(extended β strand, E)、α螺旋(α-helix, H)和弯折(bend, S).

从蛋白二级结构预测折叠速率也不是最直接的方法. 最直接的蛋白折叠速率预测方法, 应该是从一维氨基酸序列出发的, 即所谓的"蛋白折叠速率的从头预测"(*ab initio* folding rate prediction). 笔者在2006年发表文章, 提出了蛋白折叠速率的从头预测问题, 通过考察蛋白质的氨基酸序列组成与其折叠速率之间的关系, 提出了预测蛋白折叠速率的组成指标(composition index)[64]. 基于组成指标的预测方法, 虽然未必是准确率最高的方法(因为忽略了氨基酸序列的排列信息), 但迄今为止依然是数据集依赖性最小、泛化能力最强、最为稳健的预测方法[65]. 后来, 又出现了诸多从蛋白质序列出发预测蛋白折叠速率的算法和软件(或网络服务), 详见文献[65]. 最近, 又出现了用精简的氨基酸和二级结构元件集来预测蛋白折叠速率的方法, 结果表明, 折叠速率只与20种氨基酸之中的8种、7种二级结构类型中的2种密切相关[66].

蛋白折叠过程中还有一个重要的问题就是蛋白折叠过程的类型(folding type), 即单态折叠(down hill folding)、两态折叠(存在过渡态)、还是三态折叠(存在亚稳定的中间体), 因此, "折叠过程类型的预测"是另一个重要的预测问题. 笔者在2007年系统地研究了蛋白折叠类型的决定因素, 并提出了基于蛋白序列和蛋白结构的折叠类型预测方法, 达到了80%以上的准确率[22]. 随后, 又出现了多种蛋白质折叠类型的预测算法和网络服务; 例如, 2008年Huang等人[67]就采用了logistic回归和支持向量机方法基于序列来判别蛋白的两态和多态折叠类型; 更多内容参见文献[65]. 有人对这些预测方法进行比较之后, 发现有些预测方法存在夸大准确率的情况[68].

1.3 折叠结果: 蛋白结构预测

对蛋白质折叠过程的预测更多地是为了加深对蛋白折叠机理的认识, 而对蛋白折叠的结果, 即蛋白结构的预测, 则具有更强的实用性. 作为蛋白折叠过程的终点, 蛋白质天然结构的最直接预测方法就是蛋白折叠过程模拟. 然而, 受限于当前的计算能力, 可以有效模拟的蛋白大小一般不超过100个氨基酸, 且需要很多计算资源才能完成折叠过程的模拟. 因此, 发展基于经验规则的蛋白质结构预测方法也是重要的研究方向. 蛋白结构预测问题和蛋白折叠问题几乎是同时提出来的[5,7]. 根据蛋白质的结构层次, 蛋白结构预测包括蛋白的二级结构预测、三级结构预测和四级结构(蛋白复合物或相互作用)预测.

蛋白质二级结构是指序列上相邻的氨基酸所形成的局部结构, 其定义方式并不唯一, 如DSSP[69], DEFINE[70], STRIDE[71]等, 而目前广为接受的是DSSP的定义. DSSP定义了8种类型的二级结构, 分别为H (a-helix或4-helix), B (b-bridge), E (b-strand), G (3,10-helix), I (p-helix或5-helix), T (b-turn), S (bend), C或_(Coil), 其中, I极少出现. 在蛋白质二级结构预测中, 通常将上述8种类型归并为3种(E, H, C), 常用的归并方法为: E和B归为E, G和H归为H, 其余归为C; 预测准确性评价就是基于这3类(Q3)或8类(Q8)进行的, 而最常用的是Q3. 蛋白质二级结构预测的问题在20世纪60年代前后就已经开始出现了[72]. 迄今为止, 蛋白质二级结构预测大致经历了5代方法. 第一代方法, 以Chou-Fasman方法为代表[73], 它预测蛋

白二级结构主要依靠单个氨基酸在各种二级结构中出现的倾向性. 第二代方法, 考虑了局部相邻氨基酸的影响, 准确性在略高于60%的水平. 第三代方法, 通过多序列比对, 引入了进化信息(全局信息), 使得蛋白质二级结构预测的准确率提高到70%以上, 约为75%[74]. 第四代方法, 有时也称为系综方法(ensemble method)或综合方法(meta-method), 它的基本思想是综合多种方法的预测结果, 采用"专家委员会(jury-of-experts)"或叫"投票(vote)"算法, 取其中占多数的结果(consensus)作为预测结果, 可以将预测的准确率提高到80%左右[75]. 第五代算法, 就是深度学习理论的应用, 有望将二级结构的预测准确率提高到85%[76,77]. 考虑到蛋白结构的内在变动性和长程相互作用以及环境条件对蛋白质二级结构形成的影响, 据估计蛋白二级结构预测准确率的极限在88%左右[74]. 因此, 当前的二级结构预测算法还有一定的上升空间. 蛋白质二级结构预测的软件或网络服务主要有: JPred[78], PSIPRED[79], PROTEUS[75], RaptorX-SS8[80]和SPIDER2[77]等.

蛋白质三级结构是指蛋白质在三维空间中形成的立体结构. 二级结构是三级结构的重要组成部分, 因此, 二级结构可以和三级结构一起进行预测[81]. 当前的蛋白质三级结构预测方法主要分为两大类[82]: 基于模板的建模(template-based modeling)和自由建模(free modeling). 基于模板的预测方法是当前最为成功、应用也最为广泛的预测方法. 此类方法要求在PDB数据库中存在一个与待预测蛋白结构相近的模板; 它的预测步骤大体分为4步: (ⅰ) 寻找模板; (ⅱ) 把待预测序列与模板的结构对齐; (ⅲ) 将比对上的模板结构片段映射到待预测的蛋白序列上; (ⅳ) 对未比对上的空白区域进行建模, 添加侧链原子, 并对所得结构进行能量优化. 目前, 基于模板的预测方法也可分成两类: 同源建模和弱同源建模(fold recognition), 而后者是近20多年来研究者重点关注的领域; 该类方法的常用软件或网络服务主要有SwissModel[83], HHpred[84], RaptorX[85], MODELLER[86]等. 自由建模方法不要求PDB数据库中存在待预测蛋白的相应模板, 因此, 应用范围更广. 自由建模方法的基本思想是把待预测蛋白的序列分割成一定大小的多肽片段, 然后, 去PDB数据库中寻找这些肽段的结构模式, 将这些模式进行组合、拼接, 补上空白区和侧链原子, 然后基于能量函数和随机构象采样方法对所得结构进行优化, 得到待预测蛋白的最终结构. 目前, 采用自由建模的预测软件和网络服务主要有: QUARK[87], Rosetta[88], Robetta[89]等. 目前, 基于模板的预测方法的精度要高于自由建模的预测方法. 因此, 当有蛋白结构预测任务时, 首选是基于模板的预测方法. 若不成功, 再采用自由建模的预测方法. 然而, 据称当前PDB中的蛋白数量已经足够多, 可以覆盖几乎所有的蛋白结构模式[90], 因此, 基于模板的预测方法的应用将会越来越广. 此外, 将基于模板的建模方法和自由建模方法相结合, 如I-TASSER[91]等, 也是一个选择. 从1994年开始, 每两年一届的CASP (Critical Assessment of Techniques for Protein Structure Prediction)竞赛见证了蛋白质结构预测方法的进展[92,93], 迄今, 已举行了十一届, 2016年是第十二届.

蛋白质四级结构预测(蛋白相互作用预测), 按所预测结果的信息细节多少, 可以分成2个层次: 蛋白复合物结构预测和蛋白能否发生相互作用的预测. 蛋白复合物结构预测又分成2种情形: 第一, 已知2个蛋白的三级结构, 预测其形成复合物的构象; 此时, 可以采用分子对接的方法(刚性对接或柔性对接)[94,95]; 第二, 2个待预测蛋白的三级结构未知; 此时, 可以采用基于模板的建模方法, 它的基本思想和算法步骤跟基于模板的三级结构预测一样. 据分析, 目前基于模板的蛋白质四级结构预测方法大体可以分成3类[96]: 二聚体穿线法(dimeric threading)、单体穿线与多聚体映射法(monomer threading and oligomer mapping)、基于模板的分子对接方法(template-based docking); 关于这些方法的含义, 请参阅文献[96]. 当前, 预测蛋白质四级结构的软件主要有PrePPI[97], SPRING[98], Struct2Net[99], Interactome3D[100]等. 蛋白能否发生相互作用的预测(PPI预测), 是另一个重要的预测问题, 同样可以采用上述基于模板的方法, 也可以不依赖模板进行预测. PPI预测的结果不是一个具体的复合物结构, 而是蛋白之间能否发生相互作用的一个关系网络(简称为"蛋白互作网络"). 近年来, 蛋白互作网络的预测方法也越来越多[101], 大体可以分成以下几类: 基于序列相似性的同源映射方法(interolog)[102]、基于序列统计特征的机器学习方法[103]、基于基因共表达模式的预测方法[104]、基于进化信息的预测方法[105,106]、基于基因组定位的预测方法[107]、基于亚细胞定位的预测方法[108]、基于结构域组合信息的预测方法[109]、以及基

于蛋白质3D空间结构的映射方法[110,111]等. 在当前的实际应用中, 通常将上述多种方法进行整合来确定预测结果.

1.4 折叠结果: 结构特征参数预测

鉴于蛋白结构预测难度较高, 若能预测与蛋白结构有关的一些特征参数, 对于理解蛋白结构的形成机制也有帮助, 因此, 便出现了蛋白质二级结构含量预测[112], 蛋白质结构类预测[113,114], 蛋白质天然结构中氨基酸接触模式的预测[115,116]、蛋白质无结构区(disordered region)的预测[117,118]、蛋白质相互作用界面上氨基酸的接触对预测[119]、溶剂接触表面预测[77,120]等多种与蛋白质各级结构特征相关参数的预测. 限于篇幅, 这里就不展开讨论了. 感兴趣的读者, 请直接参阅上述文献.

2 展望

迄今, 蛋白折叠问题的研究已有50多年的历史, 且已经取得了重要的进展, 正在逐步逼近问题的答案. 虽然如此, 人们对蛋白折叠问题的理解还远达到成熟的地步. 目前, 蛋白折叠仍是一个十分活跃的研究领域. 在接下来的数年里, 下述几个方向值得关注. 第一, 蛋白折叠(folding)与蛋白相互作用(binding)的偶联; 这个方向涉及到蛋白质相互作用、蛋白复合物的形成以及诱导折叠等问题, 目前所知, 相对较少. 第二, 蛋白折叠与系统生物学研究的融合; 在生物体内, 蛋白折叠过程不是孤立的, 它不仅与大分子的生成和正常发挥功能有关, 而且, 与生物体自稳态的维持密切相关[121]; 蛋白的错误折叠和聚集、沉淀, 会引起诸多疾病[122,123]; 再有, 如何将蛋白折叠的预测方法应用于重要物种的蛋白质组研究, 如Human Proteome Folding Project (HPF项目), Nutritious Rice for the World项目等, 也是与系统生物学融合的一个重要方向. 第三, 深度学习理论在蛋白折叠模拟和结构预测中的应用; 随着深度学习理论的提出与发展[124], 特别是谷歌的AlphaGo程序在围棋领域所取得的辉煌战绩[125], 使人们对于深度学习理论在同样有着巨大构象空间搜索问题的蛋白折叠模拟研究中的应用, 燃起了新的希望; 但不同于围棋有明确的胜负规则和众多高手的棋谱可供机器学习, 对蛋白折叠模拟来说, 尚需一些可靠的过程数据和评价方法. 当然, 上述几个研究方向并不能涵盖蛋白折叠研究的全部领域, 但在这些重要的研究方向上, 应该有不少新的挑战和机遇. 在这些方向上取得的突破, 无疑会进一步加深人们对蛋白折叠问题的认识和理解.

参考文献

1 Kendrew J C, Dickerson R E, Strandberg B E, et al. Structure of myoglobin: A three-dimensional Fourier synthesis at 2Å resolution. Nature, 1960, 185: 422–427
2 Perutz M F, Rossmann M G, Cullis A F, et al. Structure of haemoglobin: A three-dimensional Fourier synthesis at 5.5-Å resolution, obtained by X-ray analysis. Nature, 1960, 185: 416–422
3 Anfinsen C B, Haber E, Sela M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA, 1961, 47: 1309–1314
4 Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, 181: 223–230
5 Levinthal C. How to fold graciously. In: Debrunnder J T P, Munck E, eds. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois. Monticello, I L: University of Illinois Press, 1969. 22–24
6 Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci USA, 1992, 89: 20–22
7 Levinthal C. Are there pathways for protein folding? J Chim Phys, 1968, 65: 44–45
8 Dill K A, MacCallum J L. The protein-folding problem, 50 years on. Science, 2012, 338: 1042–1046
9 Dill K A, Ozkan S B, Shell M S, et al. The protein folding problem. Annu Rev Biophys, 2008, 37: 289–316
10 [No authors listed]. So much more to know. Science, 2005, 309: 78–102
11 Levitt M, Warshel A. Computer simulation of protein folding. Nature, 1975, 253: 694–698
12 Onuchic J N, Wolynes P G. Theory of protein folding. Curr Opin Struct Biol, 2004, 14: 70–75
13 Giri Rao V H, Gosavi S. Using the folding landscapes of proteins to understand protein function. Curr Opin Struct Biol, 2016, 36: 67–74
14 Salsbury F R Jr. Molecular dynamics simulations of protein dynamics and their relevance to drug discovery. Curr Opin Pharmacol, 2010, 10: 738–744

15 Klepeis J L, Lindorff-Larsen K, Dror R O, et al. Long-timescale molecular dynamics simulations of protein structure and function. Curr Opin Struct Biol, 2009, 19: 120–127
16 Head-Gordon T, Brown S. Minimalist models for protein folding and design. Curr Opin Struct Biol, 2003, 13: 160–167
17 Zhang J, Li W, Wang J, et al. Protein folding simulations: from coarse-grained model to all-atom model. IUBMB Life, 2009, 61: 627–643
18 Helling R, Li H, Melin R, et al. The designability of protein structures. J Mol Graph Model, 2001, 19: 157–167
19 Dill K A, Chan H S. From Levinthal to pathways to funnels. Nat Struct Biol, 1997, 4: 10–19
20 Bryngelson J D, Onuchic J N, Socci N D, et al. Funnels, pathways, and the energy landscape of protein folding: a synthesis. Proteins, 1995, 21: 167–195
21 Fersht A R, Daggett V. Protein folding and unfolding at atomic resolution. Cell, 2002, 108: 573–582
22 Ma B G, Chen L L, Zhang H Y. What determines protein folding type? An investigation of intrinsic structural properties and its implications for understanding folding mechanisms. J Mol Biol, 2007, 370: 439–448
23 White G W, Gianni S, Grossmann J G, et al. Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. J Mol Biol, 2005, 350: 757–775
24 Shirts M, Pande V S. COMPUTING: Screen Savers of the World Unite! Science, 2000, 290: 1903–1904
25 Shaw D E, Deneroff M M, Dror R O, et al. Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM, 2008, 51: 91–97
26 Lindorff-Larsen K, Piana S, Dror R O, et al. How fast-folding proteins fold. Science, 2011, 334: 517–520
27 Stone J E, Hardy D J, Ufimtsev I S, et al. GPU-accelerated molecular modeling coming of age. J Mol Graph Model, 2010, 29: 116–125
28 Stone J E, Phillips J C, Freddolino P L, et al. Accelerating molecular modeling applications with graphics processors. J Comput Chem, 2007, 28: 2618–2640
29 Perez A, Morrone J A, Simmerling C, et al. Advances in free-energy-based simulations of protein folding and ligand binding. Curr Opin Struct Biol, 2016, 36: 25–31
30 Best R B. Atomistic molecular simulations of protein folding. Curr Opin Struct Biol, 2012, 22: 52–61
31 Scheraga H A, Khalili M, Liwo A. Protein-folding dynamics: Overview of molecular simulation techniques. Annu Rev Phys Chem, 2007, 58: 57–83
32 Bowman G R, Voelz V A, Pande V S. Taming the complexity of protein folding. Curr Opin Struct Biol, 2011, 21: 4–11
33 Lane T J, Shukla D, Beauchamp K A, et al. To milliseconds and beyond: Challenges in the simulation of protein folding. Curr Opin Struct Biol, 2013, 23: 58–65
34 Lindorff-Larsen K, Maragakis P, Piana S, et al. Systematic validation of protein force fields against experimental data. PLoS One, 2012, 7: e32131
35 Freddolino P L, Harrison C B, Liu Y, et al. Challenges in protein folding simulations: Timescale, representation, and analysis. Nat Phys, 2010, 6: 751–758
36 Piana S, Lindorff-Larsen K, Shaw D E. How robust are protein folding simulations with respect to force field parameterization? Biophys J, 2011, 100: L47–L49
37 Piana S, Klepeis J L, Shaw D E. Assessing the accuracy of physical models used in protein-folding simulations: Quantitative evidence from long molecular dynamics simulations. Curr Opin Struct Biol, 2014, 24: 98–105
38 Shi Y, Xia Z, Zhang J, et al. The polarizable atomic multipole-based AMOEBA force field for proteins. J Chem Theory Comput, 2013, 9: 4046–4063
39 Lopes P E, Huang J, Shim J, et al. Force field for peptides and proteins based on the classical drude oscillator. J Chem Theory Comput, 2013, 9: 5430–5449
40 Zwier M C, Chong L T. Reaching biological timescales with all-atom molecular dynamics simulations. Curr Opin Pharmacol, 2010, 10: 745–752
41 Pitera J W, Swope W. Understanding folding and design: Replica-exchange simulations of “Trp-cage” miniproteins. Proc Natl Acad Sci USA, 2003, 100: 7587–7592
42 Mitsutake A, Sugita Y, Okamoto Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers, 2001, 60: 96–123
43 Li W F, Zhang J, Wang J, et al. Multiscale theory and computational method for biomolecule simulations. Acta Phys Sin, 2015, 64: 098701 [李文飞, 张建, 王骏, 等. 生物大分子多尺度理论和计算方法. 物理学报, 2015, 64: 098701]
44 Perez A, MacCallum J L, Dill K A. Accelerating molecular simulations of proteins using Bayesian inference on weak information. Proc Natl Acad Sci USA, 2015, 112: 11846–11851
45 MacCallum J L, Perez A, Dill K A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. Proc Natl Acad Sci USA, 2015, 112: 6985–6990

46 Cooper S, Khatib F, Treuille A, et al. Predicting protein structures with a multiplayer online game. Nature, 2010, 466: 756–760
47 Duan Y, Kollman P A. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science, 1998, 282: 740–744
48 Fitter J. The perspectives of studying multi-domain protein folding. Cell Mol Life Sci, 2009, 66: 1672–1681
49 Wang E, Wang J, Chen C, et al. Computational evidence that fast translation speed can increase the probability of cotranslational protein folding. Sci Rep, 2015, 5: 15316
50 Lin Z, Madan D, Rye H S. GroEL stimulates protein folding through forced unfolding. Nat Struct Mol Biol, 2008, 15: 303–311
51 Hingorani K S, Gierasch L M. Comparing protein folding *in vitro* and *in vivo*: Foldability meets the fitness challenge. Curr Opin Struct Biol, 2014, 24: 81–90
52 Kim Y E, Hipp M S, Bracher A, et al. Molecular chaperone functions in protein folding and proteostasis. Annu Rev Biochem, 2013, 82: 323–355
53 Khalili-Araghi F, Gumbart J, Wen P C, et al. Molecular dynamics simulations of membrane channels and transporters. Curr Opin Struct Biol, 2009, 19: 128–137
54 Rogers J M, Oleinikovas V, Shammas S L, et al. Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. Proc Natl Acad Sci USA, 2014, 111: 15420–15425
55 Li W, Wang J, Zhang J, et al. Molecular simulations of metal-coupled protein folding. Curr Opin Struct Biol, 2015, 30: 25–31
56 Galzitskaya O V, Garbuzynskiy S O, Ivankov D N, et al. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins, 2003, 51: 162–166
57 Ivankov D N, Garbuzynskiy S O, Alm E, et al. Contact order revisited: Influence of protein size on the folding rate. Protein Sci, 2003, 12: 2057–2062
58 Plaxco K W, Simons K T, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol, 1998, 277: 985–994
59 Gromiha M M, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol, 2001, 310: 27–32
60 Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J, 2002, 82: 458–463
61 Ivankov D N, Finkelstein A V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA, 2004, 101: 8942–8944
62 Huang J T, Cheng J P, Chen H. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins, 2007, 67: 12–17
63 Huang J T, Wang T, Huang S R, et al. Prediction of protein folding rates from simplified secondary structure alphabet. J Theor Biol, 2015, 383: 1–6
64 Ma B G, Guo J X, Zhang H Y. Direct correlation between proteins' folding rates and their amino acid compositions: an *ab initio* folding rate prediction. Proteins, 2006, 65: 362–372
65 Xu H R, Ma B G. Progress in the study of protein folding rate determinants and folding rate prediction. Acta Biophys Sin, 2013, 29: 192–202 [徐宏睿, 马彬广. 蛋白折叠速率决定因素与预测方法的研究进展. 生物物理学报, 2013, 29: 192–202]
66 Huang J T, Wang T, Huang S R, et al. Reduced alphabet for protein folding prediction. Proteins, 2015, 83: 631–639
67 Huang J T, Cheng J P. Differentiation between two-state and multi-state folding proteins based on sequence. Proteins, 2008, 72: 44–49
68 Corrales M, Cusco P, Usmanova D R, et al. Machine learning: How much does it tell about protein folding rates? PLoS One, 2015, 10: e0143166
69 Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 1983, 22: 2577–2637
70 Richards F M, Kundrot C E. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure. Proteins, 1988, 3: 71–84
71 Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins, 1995, 23: 566–579
72 Szent-Gyorgyi A G, Cohen C. Role of proline in polypeptide chain configuration of proteins. Science, 1957, 126: 697–698
73 Chou P Y, Fasman G D. Prediction of protein conformation. Biochemistry, 1974, 13: 222–245
74 Rost B. Review: Protein secondary structure prediction continues to rise. J Struct Biol, 2001, 134: 204–218
75 Montgomerie S, Sundararaj S, Gallin W J, et al. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics, 2006, 7: 301
76 Wang S, Peng J, Ma J, et al. Protein secondary structure prediction using deep convolutional neural fields. Sci Rep, 2016, 6: 18962
77 Heffernan R, Paliwal K, Lyons J, et al. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci Rep, 2015, 5: 11476

78 Drozdetskiy A, Cole C, Procter J, et al. JPred4: A protein secondary structure prediction server. Nucleic Acids Res, 2015, 43: W389–W394
79 Buchan D W, Minneci F, Nugent T C, et al. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res, 2013, 41: W349–W357
80 Wang Z, Zhao F, Peng J, et al. Protein 8-class secondary structure prediction using conditional neural fields. Proteomics, 2011, 11: 3786–3792
81 Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. Proc Natl Acad Sci USA, 2003, 100: 12105–12110
82 Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struct Biol, 2008, 18: 342–348
83 Biasini M, Bienert S, Waterhouse A, et al. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res, 2014, 42: W252–W258
84 Soding J, Biegert A, Lupas A N. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res, 2005, 33: W244–W248
85 Peng J, Xu J. RaptorX: Exploiting structure information for protein alignment by statistical inference. Proteins, 2011, 79(Suppl 10): 161–171
86 Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Method. Enzymol, 2003, 374: 461–491
87 Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins, 2012, 80: 1715–1735
88 Simons K T, Bonneau R, Ruczinski I, et al. *Ab initio* protein structure prediction of CASP III targets using ROSETTA. Proteins, 1999, (Suppl 3): 171–176
89 Chivian D, Kim D E, Malmstrom L, et al. Automated prediction of CASP-5 structures using the Robetta server. Proteins, 2003, 53(Suppl 6): 524–533
90 Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci USA, 2005, 102: 1029–1034
91 Yang J, Yan R, Roy A, et al. The I-TASSER Suite: protein structure and function prediction. Nat Methods, 2015, 12: 7–8
92 Moult J, Pedersen J T, Judson R, et al. A large-scale experiment to assess protein structure prediction methods. Proteins, 1995, 23(3): ⅱ–ⅴ
93 Moult J, Fidelis K, Kryshtafovych A, et al. Critical assessment of methods of protein structure prediction (CASP) - progress and new directions in Round XI. Proteins, 2016, doi: 10.1002/prot.25064
94 Lensink M F, Wodak S J. Docking and scoring protein interactions: CAPRI 2009. Proteins, 2010, 78: 3073–3084
95 Vajda S, Camacho C J. Protein-protein docking: Is the glass half-full or half-empty? Trends Biotechnol, 2004, 22: 110–116
96 Szilagyi A, Zhang Y. Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol, 2014, 24: 10–23
97 Zhang Q C, Petrey D, Deng L, et al. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature, 2012, 490: 556–560
98 Guerler A, Govindarajoo B, Zhang Y. Mapping monomeric threading to protein-protein structure prediction. J Chem Inf Model, 2013, 53: 717–725
99 Singh R, Park D, Xu J, et al. Struct2Net: A web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Res, 2010, 38: W508–W515
100 Mosca R, Ceol A, Aloy P. Interactome3D: Adding structural details to protein networks. Nat Methods, 2013, 10: 47–53
101 Shoemaker B A, Panchenko A R. Deciphering protein-protein interactions. Part II. Computational methods to predict protein and domain interaction partners. PLoS Comput Biol, 2007, 3: e43
102 Walhout A J, Sordella R, Lu X, et al. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science, 2000, 287: 116–122
103 Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci USA, 2007, 104: 4337–4341
104 De Bodt S, Proost S, Vandepoele K, et al. Predicting protein-protein interactions in *Arabidopsis thaliana* through integration of orthology, gene ontology and co-expression. BMC Genomics, 2009, 10: 288
105 Valencia A, Pazos F. Prediction of protein-protein interactions from evolutionary information. Methods Biochem Anal, 2003, 44: 411–426
106 Hamp T, Rost B. Evolutionary profiles improve protein-protein interaction prediction from sequence. Bioinformatics, 2015, 31: 1945–1950
107 Dandekar T, Snel B, Huynen M, et al. Conservation of gene order: A fingerprint of proteins that physically interact. Trends Biochem Sci, 1998, 23: 324–328
108 Zahiri J, Mohammad-Noori M, Ebrahimpour R, et al. LocFuse: Human protein-protein interaction prediction via classifier fusion using protein localization information. Genomics, 2014, 104: 496–503

109 Han D S, Kim H S, Jang W H, et al. PreSPI: A domain combination based prediction system for protein-protein interaction. Nucleic Acids Res, 2004, 32: 6312–6320
110 Hue M, Riffle M, Vert J P, et al. Large-scale prediction of protein-protein interactions from structures. BMC Bioinformatics, 2010, 11: 144
111 Planas-Iglesias J, Marin-Lopez M A, Bonet J, et al. iLoops: A protein-protein interaction prediction server based on structural features. Bioinformatics, 2013, 29: 2360–2362
112 Liu W, Chou K C. Prediction of protein secondary structure content. Protein Eng, 1999, 12: 1041–1050
113 Chou K C, Zhang C T. Prediction of protein structural classes. Crit Rev Biochem Mol Biol, 1995, 30: 275–349
114 Ding S, Yan S, Qi S, et al. A protein structural classes prediction method based on PSI-BLAST profile. J Theor Biol, 2014, 353: 19–23
115 Jones D T, Singh T, Kosciolek T, et al. MetaPSICOV: Combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics, 2015, 31: 999–1006
116 Skwark M J, Raimondi D, Michel M, et al. Improved contact predictions using the recognition of protein like contact patterns. PLoS Comput Biol, 2014, 10: e1003889
117 Wang S, Weng S, Ma J, et al. DeepCNF-D: Predicting protein order/disorder regions by weighted deep convolutional neural fields. Int J Mol Sci, 2015, 16: 17315–17330
118 Eickholt J, Cheng J. DNdisorder: Predicting protein disorder using boosting and deep networks. BMC Bioinformatics, 2013, 14: 88
119 Ovchinnikov S, Kamisetty H, Baker D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife, 2014, 3: e02030
120 Ma J, Wang S. AcconPred: Predicting solvent accessibility and contact number simultaneously by a multitask learning framework under the conditional neural fields model. BioMed Res Int, 2015, 2015: 678764
121 Powers E T, Morimoto R I, Dillin A, et al. Biological and chemical approaches to diseases of proteostasis deficiency. Annu Rev Biochem, 2009, 78: 959–991
122 Herczenik E, Gebbink M F. Molecular and cellular aspects of protein misfolding and disease. FASEB J, 2008, 22: 2115–2133
123 Valastyan J S, Lindquist S. Mechanisms of protein-folding diseases at a glance. Dis Mod Mech, 2014, 7: 9–14
124 Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313: 504–507
125 Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search. Nature, 2016, 529: 484–489

马彬广

1979 年 2 月生人. 2007 年苏州大学毕业, 获得理学博士学位. 随后在挪威卑尔根大学计算科学中心从事博士后工作两年. 2009 年底回国, 到华中农业大学任教. 主要研究领域包括生物物理学和计算生物学, 重点关注系统与合成生物学方向. 长期从事蛋白折叠与相互作用的研究, 发表了一系列有一定国际影响的学术论文. 现担任多家国际刊物的编委或审稿人.

Protein folding prediction

MA BinGuang

College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Protein folding is the process that a protein molecule transforms from the linear polymer of peptides to a three-dimensional native structure with specific biological function. By now, the protein folding problem has been studied for more than 50 years and already became a broad and active research field. To answer the 58th question raised by *Science* in 2005, in this article we briefly reviewed the background and research history of the protein folding problem, and introduced the progresses of protein folding prediction research from four aspects: the protein folding process prediction (protein folding simulation), the folding process related parameter prediction, the protein folding result prediction (protein structure prediction), and the folding result related parameter prediction. The studies on the protein folding problem began in the 60s of 20th century, with the efforts to seek a solution to the paradox that a protein can actually form a native 3D structure in only several seconds but the time scale estimated by a thermodynamic ergodic hypothesis would be longer than the age of universe. Computer simulation is an important approach for protein folding study. The protein models can be classified into 3 categories: lattice model, off-lattice model and all-atom model. The current knowledge about protein folding mechanism is based on the concept of folding funnel on a free-energy landscape, and the current opinion is that the protein folding mechanism is not unique for the whole protein universe and that there may exist a continuum between the two extreme ends of hierarchical folding and nucleation folding scenarios. The hardware for protein folding simulation was becoming more powerful; distributed systems (e.g, Folding@home), special-purpose machines (e.g, ANTON), and GPU-based platforms have been developed for protein folding simulation. Meanwhile, the folding simulation software was continuously enhanced. An important issue in protein folding simulation is to overcome the local energy barrier to find the global energy minimum; several approaches such as replica-exchange, multi-scale modeling and Modeling Employing Limited Data (MELD) were developed to tackle this issue; human intelligence involvement (e.g, "Foldit" Game) is another interesting effort. During the past two decades, the ability of protein folding simulation was continuously rising. For now, the folding simulation for the proteins with dozens of amino acids can reach a time scale of millisecond, while the protein size able to do effective folding simulation is around 100 amino acids. The targets of protein folding simulation have been largely expanded and now include both the *in vitro* and the *in vivo* folding such as co-translational folding, chaperone-assistant folding, small-molecule-induced folding and metal-coupled folding. Folding rate and folding type are two important parameters related with the protein folding process and now they can be predicted by statistical and machine-learning approaches based on different levels of structural features such as the topological properties of tertiary structure, the contents of secondary structure and the amino acid frequencies of primary structure. The result of a protein folding process is the formation of a protein structure. According to the hierarchy of structural organization, the protein structure prediction problem includes secondary structure prediction, tertiary structure prediction and quaternary structure prediction. By now, the secondary structure prediction algorithm has experienced five generations and the current accuracy is about 80% for 3-classes prediction. The tertiary structure prediction approaches mainly include two categories: template-based modeling and free modeling, with the former having higher accuracy and the latter having larger application scope. The quaternary structure prediction includes the prediction of complex structure and the prediction of the possibility of protein-protein interaction, and these predictions can be performed based on protein 3D structure or merely amino acid sequence. Structure related parameter prediction also attracted research interests, including the predictions of protein structural classes, secondary structure contents, disordered regions, solvent accessible surface region and the amino acid contacting pairs in the interface of protein-protein interaction. In the end, some possible development directions worth noticing in the future of protein folding research were suggested and they are: the coupling between protein folding and binding, the fusion of protein folding research with systems biology and the application of deep-learning techniques in the field of protein folding prediction.

protein folding, protein folding simulation, protein structure prediction, deep learning, systems biology

doi: 10.1360/N972016-00658