专题: 软物质研究进展

蛋白质结构预测*

邓海游[1)] 贾亚[2)] 张阳[3)†]

1) (华中农业大学理学院, 武汉 430070)

2) (华中师范大学物理科学与技术学院, 武汉 430079)

3) (Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 45108, USA)

(2016 年 6 月 22 日收到; 2016 年 7 月 21 日收到修改稿)

从氨基酸序列出发预测蛋白质的三维结构是目前计算生物学和生物物理学领域最具挑战性和影响力的研究方向之一. 本文从结构预测的研究背景出发, 简要介绍了它的理论意义、应用需求及基本现状; 并根据结构预测的一般步骤, 依次介绍了构象初始化、构象搜索、结构筛选、全原子结构重建、结构优化等基本预测过程; 随后分基于模板和无模板两类, 各列举了几种具有代表性的结构预测方法; 最后对该领域的盛事——国际蛋白质结构预测技术评估大赛 (CASP) 做了简单介绍.

关键词: 蛋白质结构预测, 同源建模, 从头预测, 结构优化

PACS: 87.14.E–, 82.30.–b **DOI:** 10.7498/aps.65.178701

1 引 言

二十世纪末以来, 生命科学领域飞速发展, 越来越多不同学科背景的研究者参与到生命科学相关的研究中. 而作为生物体内分布最广、功能最复杂的一类大分子, 蛋白质受到了尤为广泛的关注和研究. 氨基酸是构成蛋白质的基本单元, 20 种不同氨基酸以 mRNA 为模板经脱水缩合形成首尾相连的氨基酸序列, 是为蛋白质的一级结构 (蛋白质序列). 不同的蛋白质拥有不同的氨基酸序列, 所有蛋白质都会在其一维序列的基础上折叠形成特定的三维结构, 了解蛋白质的三维结构是研究其生物功能及活性机理的基础. 然而至今为止, 人们仍不清楚蛋白质究竟如何由其一维序列折叠形成具有特定生物功能的三维结构. 相对于翻译过程中的三联体遗传密码, 蛋白质序列与其空间结构的对应关系常被称之为第二遗传密码 (或折叠密码) [1].

实验上测定蛋白质序列和结构的方法已经历了数十年的发展, 相关的人力物力投入也在逐年增加, 尤其在过去二十年, 蛋白质序列数据库 UniProt [2] 以及结构数据库 PDB [3] 中的数据条目几乎都在以指数形式增长. 尽管如此, 获得蛋白质序列数据要比获得结构数据简单得多, DNA 测序技术的突飞猛进更使得可直接通过翻译、推导得到大量的蛋白质序列. 而目前蛋白质结构数据库 PDB 中所存储的蛋白质三维结构主要通过 X 射线晶体衍射和核磁共振成像技术得到, 两种实验方法均成本不菲, 且有各自的应用局限. 截止 2016 年 5 月, PDB 数据库中存储了 11 万余条蛋白质结构数据, 而这只占 UniProt 中所有蛋白质序列数据的 1/600, 也就是说只有不到 0.2% 的蛋白质序列拥有实验测定的三维结构.

在理论探索和应用需求的双重推动下, 以序列为起点、三维结构为目标的蛋白质结构预测自 20 世纪末以来蓬勃发展. 美国科学家安芬森 1973 年在《科学》杂志发文指出 [4], 给定环境下蛋白质折叠所需的全部信息都包含在其氨基酸序列中 (称为

* 国家自然科学基金 (批准号: 11547255, 11474117)、中央高校基本科研业务费专项资金 (批准号: 2662015BQ045) 和美国国立卫生研究院 (批准号: GM083107, GM116960) 资助的课题.

† 通信作者. E-mail: zhng@umich.edu

Anfinsen's dogma). 从物理学角度来看, 氨基酸序列确定了蛋白质的基本分子组成, 它最终的折叠结构将对应于给定环境下该分子热力学最稳定、自由能最低的构象态. 尽管蛋白质的折叠本质上是受各种物理规律的支配, 但对于如此复杂的生物大分子(包括它与周围溶剂分子的相互作用), 目前人们不仅无法给出精确的物理描述, 而且也不具备足够强大的计算条件去完成相关运算. 这促使研究者们积极运用各种生物信息学方法、粗粒化物理模型以及优化的构象搜索算法, 发展出了一系列卓有成效的结构预测方法.

由于相似的蛋白质序列往往拥有相似的三维结构, 这就有了以PDB数据库中的已知结构为模板的同源建模(homology modeling)方法, 它是迄今为止精度最高的一类结构预测方法. 随着PDB数据库的快速壮大, 会有越来越多的蛋白质可通过同源建模获得精确的预测结构. 而当PDB数据库中找不到与待预测蛋白质序列(下文中称为"目标蛋白"或"目标序列")具有显著序列相似性的蛋白质结构时, 此时通过穿线方法(threading)[5]仍有可能找出与目标蛋白具有结构相似性的已知结构. 穿线方法实际上是通过某种策略将序列与结构进行比对, 评估将序列以各种匹配方式"安放"到各个三维结构上的"舒适"程度, 因此也被称为折叠辨识(fold recognition)[6−8]. 该类方法之所以有效, 是因为蛋白质三维结构具有很强的保守性, 蛋白质折叠子(folds)的总数目是有限的[9,10]. 同源建模和穿线方法可以统称为基于模板的结构预测方法[11]. 不同于同源建模以及穿线方法, 从头预测方法(*ab initio* modeling)不依赖于任何已知结构, 而是以第一性原理构建蛋白质折叠力场(force field), 再通过相应的构象搜索(conformational search)方法搜寻目标蛋白的天然结构(native structure). 显然, 从头预测方法的发展也是对"第二遗传密码"的探索过程, 它具有非凡的理论意义. 尽管如此, 该类方法目前还面临着诸多困难和挑战, 纯粹的从头预测方法非常罕见, 几乎所有预测方法都会在不同程度上使用已知的蛋白质结构信息, 因此, 在对结构预测方法的分类介绍中, 通常把除同源建模和穿线方法之外的方法统称为无模板(template-free)的结构预测方法[12].

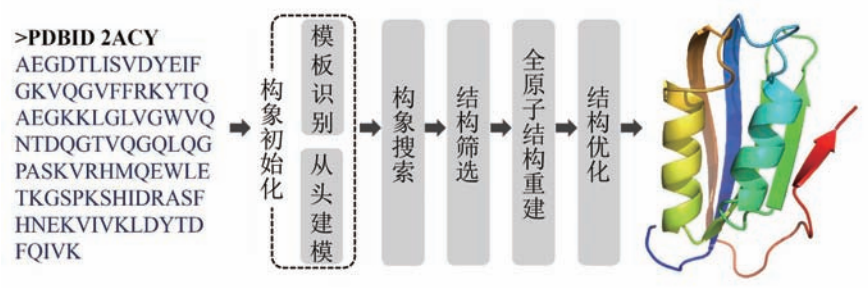

图1 蛋白质结构预测的一般步骤

Fig. 1. The general flowchart of protein structure prediction.

尽管不同的结构预测方法所涉及的具体预测过程千差万别, 但总的来说基本的步骤是一致的, 往往包括构象初始化(含模板识别和从头建模)、构象搜索、结构筛选、全原子结构重建、结构优化等(见图1). 本文将依次介绍这些步骤以及各步骤常用的方法, 而后对一些综合的结构预测方法进行简要介绍; 另外, 我们还将特别介绍过去二十余年里见证和促进了蛋白质结构预测领域发展的国际蛋白质结构预测技术评估大赛(CASP)[13,14]; 最后是对目前结构预测所面临的一些问题以及未来发展方向的讨论.

2 结构预测的一般步骤

2.1 构象初始化

如前面所说, 蛋白质结构预测的起点(输入)是目标蛋白一维的氨基酸序列, 终点(输出)是其折叠形成的三维结构. 然而对于任意一条目标序列, 其理论上可能的空间构象都有无穷多个, 但对于大多数蛋白质来讲, 真正的天然结构却只有一个. 纯粹从未折叠的一维序列开始构象搜索, 所面临的困难是巨大的, 一方面, 我们还远远无法构建起能引导目标序列朝正确方向折叠的足够精确的力场, 另一

方面, 完成如此浩瀚的构象搜索过程所涉及的计算量也是目前我们的计算机难以承受的. 事实上, "基于模板"和"无模板"两类结构预测方法的关键区别就在于所采取的构象初始化方法不同. 基于模板的预测方法通过搜索识别与目标蛋白具有同源性或结构相似性的已知结构作为模板而获得初始化构象; 无模板的预测方法则通常以小的结构片段为起点从头构建初始构象. 目前来看, 无论是哪种方法, 都会在不同程度上利用PDB数据库中已知的结构信息, 从而构建起一个远优于随机构象的初始构象, 由此显著降低构象搜索的压力, 包括减少计算量、减轻对高精度力场的依赖.

对于大部分目标蛋白来说, 都可以通过序列比对 (sequence alignment) [15−17] 从已知结构数据库中识别出与之具有同源性的蛋白质, 并进一步获得目标序列与模板序列精确的残基匹配信息. 序列的相似性往往意味着结构的相似性, 依据残基匹配信息从模板拷贝而来的空间结构 (不一定完整) 有时已非常接近于目标蛋白的天然结构, 甚至于使后续的构象搜索过程变得可有可无, 这是同源建模方法特有的现象. 因此, 提高序列比对方法的效率、精度是促进同源建模方法发展的关键. 由于蛋白质结构远比序列具有更高的保守性, 毫无序列相似性的两个蛋白质也可能拥有相似的结构, 这是穿线法 [5] 发挥作用的领域. 过去二十年里, 穿线法被广泛研究和应用, 由此也大大提高了结构预测方法对PDB数据库中已知结构的利用率. 总的来说, 这类由模板拷贝而来的初始化构象可靠性颇高、信息量极大, 可急剧缩短后续构象搜索的进程.

然而, 并不是所有的目标蛋白都能通过序列比对或穿线法找到满意的结构模板, 此时就需要用到无模板的结构预测方法. 最简单的办法是随机生成目标蛋白的初始构象, 但正如前面提到的, 这样一来构象搜索的任务会异常艰巨, 加上目前蛋白质力场的精度远远不够, 想要完成由随机构象到天然构象如此大跨度的模拟过程极为困难. 事实上蛋白质复杂的、多层次的结构特征给我们提供了更多选择. 我们可以单独对蛋白质的二级结构、主链二面角、溶剂可及性 (solvent accessibility)、接触图 (contact map) 等各种结构特征进行预测, 这些预测并不依赖与目标蛋白同源或相似的结构模板, 而且能获得较高的预测精度. 许多无模板的结构预测方法 [18,19] 所采用的结构片段装配 (fragment assembly) 就是以预测得到的二级结构、主链二面角以及溶剂可及性为依据, 从PDB数据库中截取一系列小的结构片段, 通过片段装配得到目标蛋白完整的初始构象. 相对于随机构象, 结构片段装配方法显然大大缩小了构象搜索的空间, 通常也更有利于提高蛋白质局部结构 (local structure) 的质量.

2.2 构象搜索

获得初始化构象之后, 通常还需要进行构象搜索, 即在一定力场的指导下, 采用某种搜索策略不断改变分子构象, 以搜寻更接近天然的构象. 作为典型的生物大分子, 一个蛋白质分子包含了千百个原子, 构象自由度极高, 往往需要做某种近似处理才能进行有效的构象搜索, 这就有了蛋白质表达 (protein representation) 的问题. 其实在前面的构象初始化中已经涉及蛋白质表达问题, 譬如蛋白质序列比对都是残基水平上的比对, 比对过程中允许两个不同的残基进行匹配, 这样一来, 模板残基的侧链构象就不能为目标蛋白所用, 能够拷贝的只有匹配区域的主链构象. 可见, 由模板得到的初始化构象已经是一种简化的构象. 目前常见的结构预测方法在构象搜索过程中也都使用了某种简化的模型来表示蛋白质, 例如以Cα-原子和虚拟的侧链中心来表示每一个残基, 或只考虑残基的主链原子, 由此构象的自由度得以显著降低.

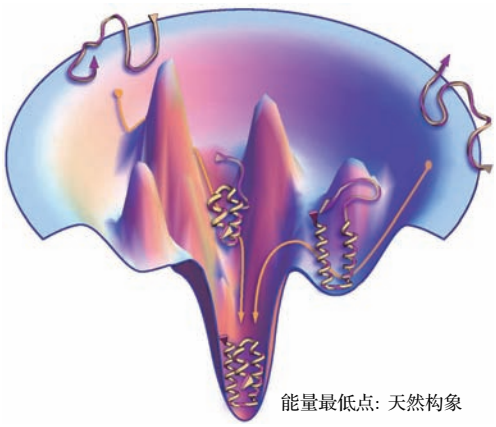

图 2 能量景貌引导的蛋白质折叠示意图 [20]

Fig. 2. Protein folding guided by funnel-shaped energy landscape.

进行构象搜索必须要有一个能描述蛋白质构象能量景貌 (energy landscape) 的力场, 目标蛋白的任意一个构象都对应能量景貌上的一个点, 对于设计精良的力场来说, 目标蛋白的天然构象应当位

于能量景貌的最低点(见图2). 蛋白质折叠是原子、分子水平上相互作用的结果, 要对这种相互作用进行精确描述自然要诉诸量子化学和量子物理理论. 然而目前只有很小的分子体系才可能进行基于量子理论的精确能量计算, 对于蛋白质这种环境因素复杂的大分子体系则几乎无能为力. 现有的很多力场都是基于经典的牛顿物理理论, 考虑键伸展能(bond-stretching energy)、角弯曲能(angle bending energy)、范德瓦尔斯相互作用、静电以及氢键相互作用等, 这类力场常被称为基于物理的(physics-based)力场, 分子动力学模拟中常用的AMBER力场[21], CHARMM力场等[22]都属于此类. 与之对应的, 另一类力场则绕过了第一性的物理原理, 转而通过挖掘和利用PDB数据库中丰富的已知结构信息来构建有效的评估函数, 因此也被称为基于知识(knowledge-based)的力场[23,24]. 对于蛋白质结构预测来说, PDB数据库可称得上是一座取之不竭的资源宝库, 前面讲到的模板结构搜索只是一个方面, 作为一个庞大的天然结构(实验结构)信息池, 从中可以挖掘出大量的天然结构特征信息用于构建基于知识的力场. 1990年, Sippl[25]通过对98个PDB结构中氨基酸残基对(residue pair)距离的统计, 构建起一种平均力势(potential of mean force)并利用它对低肽链的构象进行了详细的评估和分析. 基于相似方法, 后来陆续涌现出了一系列的基于原子对(atom pair)距离分布的统计势[26−29], 它们之间的关键区别在于对参考态的处理上[30]. 不止于此, 天然结构在主链二面角、溶剂可及性、侧链方位等各方面的特征均可用于构建起基于知识的力场, 构建过程还可用到如隐马尔可夫模型、人工神经网络、支持向量机等各类机器学习方法. 目前来看, 两类力场各有优劣. 基于物理的力场从第一性原理出发, 物理意义明确, 但精度不够, 实际应用效果往往并不理想. 基于知识的力场从已知结构归纳提取特征, 易于构建且性能表现不俗, 但却规避了对力场物理本质的探索. 很多结构预测方法采用了既有基于物理成分又包含基于知识成分的力场.

力场确定了蛋白质各个构象的能量值, 接着需要一定的方法来搜索低能量的构象. 分子动力学模拟[31,32]是常用的一种构象搜索方法, 它通过解牛顿运动方程来搜索构象, 该过程还能在一定程度上展现蛋白质折叠的动力学信息. 然而, 对于蛋白质这样的生物大分子, 采用分子动力学模拟的计算量非常巨大. 一般而言分子动力学模拟的时间步长在飞秒(10^{-15} s)量级, 而蛋白质折叠时间通常是在毫秒(10^{-3} s)量级, 二者之间存在十余个数量级的差距, 这在目前的计算发展水平下显然是一个极大的挑战. 相对而言, 基于蒙特卡罗模拟[33,34]的构象搜索过程就要快得多, 因为构象是在事先设计好的各种随机变动(movement)中不断变化折叠. 这些变动可以包括大到残基水平的结构片段变动, 小到单个原子的坐标变动, 可以包括连续变动, 以及在固定边长的网格上的变动(显著降低构象数量), 这些变动的尺度、幅度以及各种变动的出现频率都可以根据需要进行设置和调整. 此外, 为了跨越能量壁垒, 防止陷入局部能量最小态, 很多结构预测方法会在构象搜索过程中采用模拟退火[35]、副本交换等[36,37]策略, 同时还配合使用多个初始化构象, 从能量景貌的不同位置开始构象搜索.

2.3 结构筛选

构象搜索结束后, 通常会得到目标蛋白的大量构象各异的结构(这里的“构象”与“结构”都指的是目标蛋白的空间排布, 只不过前者凸显动态变化, 后者意含静态确定性). 就拿蒙特卡罗模拟来说, 构象的变动过程具有随机性, 尽管在力场的指引下总体会向低能态演变, 但难免会陷入局部能量最小态, 且由于力场自身的缺陷, 能量最低态也并不一定就最接近天然结构. 因此, 模拟过程中通常会不断输出一些能量较低的中间结构以供后续筛选. 结构筛选其实与结构评估对应着一个相同的问题, 即如何正确区分不同质量(与天然结构具有不同差异程度)的预测结构. CASP竞赛中把那些专用于结构筛选及评估的方法称为模型质量评估方法(MQAP)[38], 并提供了评估这些方法的平台. 需要说明的是, 构象搜索后得到的一般只是蛋白质的简化结构, 所能用到的通常也是针对此类简化结构的筛选和评估方法. 发展优秀的结构筛选及评估方法是蛋白质结构预测中极为重要的研究方向, 历届CASP竞赛保存的大量预测结构也常被作为假结构(structural decoy)而广泛用于训练或测试新的结构评估方法, 此外还有很多途径和方法[39−41]均可获得这类假结构.

其实, 力场本身就能筛选结构, 构象搜索过程输出的结构可以说是相应力场筛选的结果. 然而

力场的设计需要考虑构象搜索的特定需求, 如能量景貌的全局性、能量壁垒的可逾越性以及能量计算的复杂度等, 尤其为了保证构象搜索的速度, 力场的精细度往往受到限制. 相对于力场所需处理的海量构象, 构象搜索所输出的结构数目显得微不足道, 因此便可利用更为复杂、精细的打分函数 (score function) 来筛选结构. 这里的打分函数就如力场一样, 可由基于物理或基于知识的方法构建, 可对目标蛋白的任意结构给出评分. 此外, 还有一些方法是基于待选结构的相似性来进行结构筛选, 例如结构聚类[42−44]. 一般认为出现频率越高的结构越有可能接近天然结构 (不论是对于同一预测方法得到的大量结构, 还是不同预测方法得到的结构集合), 这是结构聚类方法的思想基础, 很多著名的结构预测方法都采用了这种方法来筛选结构.

2.4 全原子结构重建

由于目前的结构预测方法普遍都采用蛋白质简化模型进行构象搜索, 通过上述步骤得到的也只是一个或多个简化结构, 接下来需要在简化结构的基础上重建起全原子结构. 对于不同的简化模型, 全原子重建的过程也不一样. 有些结构预测方法采用的是 Cα-原子加“虚拟侧链中心”式的简化模型, 而且其中的“虚拟侧链中心”只用于辅助确定 Cα-原子位置, 构象搜索输出的是仅含 Cα-原子位置的简化结构. 对于这种情况, 往往需要先根据 Cα-原子位置重建起整个主链结构, 很多独立的全原子重建方法均具有这个功能, 如 SABBAC[45], BBQ[46], PULCHRA[47], REMO[48] 等. 这些方法一般都依赖于从已知结构截取的主链片段数据库, 例如, REMO 所使用的主链片段数据库就包含来自 2561 个 PDB 结构的 528798 个长度为四个残基的主链片段. 有了主链结构, 剩下的就是要给每个残基安装上各自的侧链. 与主链的重建类似, 侧链的安装通常也会用到一个数据库——侧链旋转异构体数据库[49], 安装过程中不断搜索数据库, 筛选确定最佳的侧链构象. 侧链的安装也有专门的方法, 如 Scwrl 等[50,51], SCATD[52], RASP[53]. 完成了主链和侧链的重建便可得到全原子的预测结构. 需要多说一句的是, 很多全原子结构重建过程并未考虑结构, 必要时可通过专门的氢原子添加工具进行添加.

2.5 结构优化

虽然通过前面的步骤已经获得了目标蛋白完整的预测结构, 但由于所用力场、构象搜索方法以及全原子重建方法自身的问题和缺陷, 该结构的质量 (尤其是局部结构质量) 往往存在较大的优化空间. 尤其经构象搜索过程输出的简化结构, 很容易出现空间位置冲突、不可能的主链二面角、非常规的氢键网络等问题. 另外, 如果后续的结构筛选采用结构聚类方法, 并以聚类质心 (cluster centroid) 结构为输出, 那么可能还会额外引入一些局部结构偏差[54]. 显然, 简化结构存在的问题会被进一步延伸、放大, 直接影响到在其基础上重建起来的全原子结构. 因此, 有的结构预测方法会在全原子结构重建的同时, 分步骤进行结构优化[55]. 例如, 首先对简化结构自身进行一轮优化, 尽可能修正一些不合理的空间坐标; 接着对重建起来的主链结构进行优化, 从而使后续的侧链安装过程更加顺畅; 最后对重建好的全原子结构进行优化.

和前面的构象搜索过程一样, 结构优化过程也需要一定的力场以及相应的构象搜索方法. 相比而言, 构象搜索过程的主要任务是确定目标蛋白的整体拓扑结构 (或者说主链结构), 因构象变化跨度大, 为保证搜索效率而往往牺牲了结构细节; 结构优化过程的主要目的则是要在整体拓扑结构变动不大的情况下尽可能改善其局部结构细节. 当然, 最好的情况是使目标蛋白的整体拓扑结构以及局部结构细节均得到优化, 目前这仍是一个极具挑战的任务. 由于结构优化过程中涉及的构象变化远远小于构象初始化之后的构象搜索过程, 这让分子动力学模拟得以扬长避短, 充分发挥其构象搜索本领. 过去十余年里发展起了诸多专注于结构优化的工具及方法. FG-MD[56] 就是一个基于分子动力学模拟的结构优化方法, 它同时从不同结构模板和小的结构片段中搜集空间限制信息来指导结构优化. ModRefiner[55] 则是一个基于蒙特卡罗模拟的结构优化方法, 它同时具备全原子结构重建的功能, 其结构重建和结构优化的功能相辅相成. 自第七届 CASP 竞赛起, 还专门设置了模型优化 (model refinement) 项目, 以评估蛋白质结构优化方面的发展情况[57−59].

3 两类结构预测方法

正如第2节所介绍的, 蛋白质结构预测过程包括了一系列重要步骤, 其中每个步骤几乎都能独自构成一个研究领域或方向, 每个步骤都有大量专门的工具或方法可供选择. 然而除了那些熟悉结构预测或致力于预测方法开发的人, 普通的用户往往对具体预测步骤以及相关技术方法并不感兴趣, 他们只希望能既便捷又可靠地预测蛋白质结构. 经过多年的发展和积淀, 目前有不少综合的结构预测方法脱颖而出, 被世界范围内的蛋白质结构研究者广泛使用[60,61]. 接下来我们就分“基于模板”和“无模板”两类, 举例介绍几种常见的结构预测方法.

3.1 基于模板的结构预测方法

对于大多数目标蛋白来说, 基于序列同源性或结构相似性, 通过序列比对或穿线法总能从PDB数据库中找到合适的模板用于结构预测. 从模板获得的结构信息往往比其他途径得到的更为可靠, 过程也更为便捷. 尤其是目标蛋白与模板具有显著同源性时, 预测结果一般具有极高的可信度. 这使得基于模板的结构预测方法在实际应用中更受欢迎. 下面简单介绍几个具有代表性的基于模板的结构预测方法.

SWISS-MODEL[62,63]是Schwede实验室较早设立的蛋白质结构同源建模在线工具, 其整个预测流程包括模板识别、目标蛋白与模板比对、模型搭建、模型评估等几个步骤. 它使用BLAST[64]和HHblits[65]进行模板搜索比对, 根据比对结果从模板中拷贝原子坐标信息来构建目标蛋白的空间结构, 多个模板时则按一定比重取坐标平均. 对于目标蛋白未比对区域, 需要搜索环区(loop region)结构数据库, 找出匹配最佳的结构片段. 最后是利用QMEAN等[66]打分函数对预测结构进行评估. 根据用户干预程度的不同, SWISS-MODEL提供了自动模式(automated mode)、对比模式(alignment mode)、工程模式(project mode)三种工作模式, 此外还提供了蛋白质多单体结构(四级结构)建模模块. 它功能丰富、操作简单, 且运行稳定、计算快捷, 无论是刚起步的非专业人员还是熟悉结构预测的专家, 都可使用它获得一些同源蛋白的初步信息和结果, 因此SWISS-MODEL也是过去二十余年里使用最为广泛的在线同源建模工具之一. 它的广泛使用也与它开展服务的时间早有关. 尽管如此, 由于其采用的方法简单, 纯粹从模型精度上来看, 它不一定是最具竞争力的. 读者可参考由CAMEO[67]提供的持续更新的蛋白质模型全自动评估结果(http://www.cameo3d.org), 了解更多包括SWISS-MODEL在内的诸多在线建模工具的性能评估信息.

Modeller[68,69]是另一个由Sali实验室早期开发的基于模板的建模工具. 与SWISS-MODEL不同, Modeller是通过满足空间限制信息来进行结构建模, 它可根据目标蛋白与模板的比对自动导出一系列空间约束信息, 并利用这些约束信息指导建模. 除了模板结构, NMR实验数据、荧光光谱数据、位点定向诱变、立体化学限制、残基或原子距离统计势乃至研究者的直觉等都可以成为空间约束信息的来源. 约束信息的形式也多种多样, 可以是距离、角度、二面角等. 正是因为这些特点, Modeller能够非常方便地利用一切可利用的信息为结构预测服务. 最终输出的是五个全原子结构, 每个结构都对应有DOPE等[70]评估函数的打分值. 不同于在线建模工具, Modeller可被下载安装于各种系统的本地计算机上, 用户拥有更大的自由度. 由于Modeller是基于命令行的方式来运行的, 相对而言不太容易上手, 因此近年来出现了EasyModeller等[71]以Modeller为基础的图形界面建模工具.

前面两种预测方法都源于同源建模思想, 一般要求目标蛋白与模板具有一定程度的序列相似性. 很多基于穿线法的预测方法则具有更广阔的适用范围. I-TASSER[72,73]是Zhang实验室近年来发展起来的一种基于穿线法的综合性结构预测方法. 它的整个预测流程包括以下几步(见图3): 首先, 利用多重穿线(multiple threading)算法LOMETS[74]从非冗余的蛋白质结构数据库中识别出若干结构模板或超二级结构片段; 第二步, 在这些模板或结构片段的基础上构建起初始构象(非比对区域从头构建), 并展开基于副本交换蒙特卡罗模拟的构象搜索; 第三步, 通过结构聚类方法SPICKER[43]对构象搜索产生的大量构象进行聚类, 计算出最大的几个聚类的质心构象(取聚类中各构象等位残基的平均位置作为质心构象中相应残基的位置); 第四步, 对聚类质心构象再进行一次副本交换蒙特卡罗模拟以尽可能去除因“平均”带

来的原子碰撞现象并优化整体拓扑结构; 第五步, 重建全原子结构并利用FG-MD等[56]方法进行结构优化; 最后I-TASSER将输出五个预测结构以及相应的评估得分. 事实上, I-TASSER在无同源模板的新折叠(new fold)目标蛋白的预测方面也有不错的表现, 其突出的综合性能让它在过去几届的CASP竞赛中一直处于领先地位. 用户可选择通过在线提交预测任务, 然后由Zhang实验室提供高精度的自动在线预测. 其缺点是由于Zhang实验室的计算资源的限制以及全球大量在线用户的使用, 每次每个注册用户只能提交一个蛋白质序列. 另外, Zhang实验室也提供免费的I-TASSER软件套装. 用户也可以下载安装I-TASSER套件, 在本地计算机上进行快速的、批量化的结构预测.

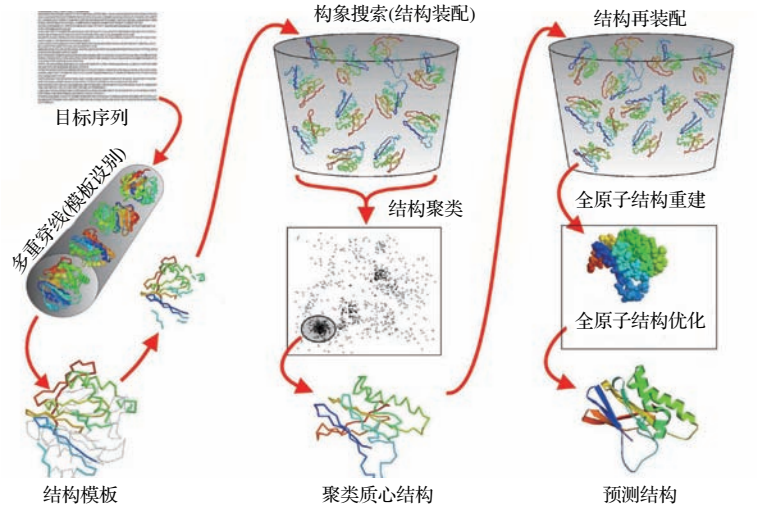

图3 I-TASSER结构预测基本流程[54]

Fig. 3. Flowchart of I-TASSER method for protein structure prediction.

3.2 无模板的结构预测方法

目前绝大部分的结构预测方法都在不同程度上依赖于蛋白质结构数据库提供的信息, 模板的使用是其中最直接的一种形式. 过度地依赖于已知结构显然不利于我们真正去探索和认识蛋白质折叠的本质规律. 无模板结构预测方法的发展不仅受实际应用的驱动(并不是所有目标蛋白都能在结构数据库中找到满意的模板), 更受到蛋白质折叠密码这一基本科学问题的推动. 尽管现有的无模板结构预测方法普遍也使用了已知结构信息(如基于知识的力场、片段装配所用的实验结构片段等), 但相对于基于模板的结构预测方法, 它的发展无疑更能体现结构预测领域的理论和技术水平.

Rosetta[18,75]是Baker实验室开发的较早采用片段装配策略的结构预测方法. 所使用的结构片段短至3—9个残基长度, 均来自于已知结构. 与基于模板的方法类似, Rosetta是通过局部序列相似性及预测的二级结构来筛选结构片段. 构象由包含主链原子、Cβ原子以及侧链质心(不计Cβ原子)的简化模型表示, 构象搜索使用蒙特卡罗模拟退火方法, 过程在扭转角空间(torsion space)进行. 之后利用聚类方法从模拟过程产生的大量结构中筛选出最优结构并进行全原子结构重建与优化. QUARK[19]是另一个由Zhang实验室开发的优秀的片段装配预测方法, 它的总体预测流程与Rosetta类似, 而它使用的是1—20个残基不等的结构片段, 并使用副本交换蒙特卡罗模拟进行构象搜索, 各步骤的具体算法也进行了特别的设计. 此外还有SCRATCH等[76], PROFESY[77], FRAGFOLD[78]一系列方法, 都属于片段装配类的结构预测方法. 这些方法与基于模板的预测方法的关键区别在于: 它们并不依赖于任何一个完整的结构模板, 不要求片段所在模板与目标蛋白有任何同源性或结构相似性, 这让它们具有更大的随机性和自由度, 便于模拟已知结构中不存在的全新结构. 尽管如此, 由于计算量巨大、力场精度不够等原因, 目前无模板的结构预测方法还只能应对尺寸相对较小

(<150残基)的目标蛋白.

4 国际蛋白质结构预测技术评估大赛(CASP)

介绍蛋白质结构预测, 不得不介绍国际蛋白质结构预测技术评估大赛(CASP)[13,61], 它在计算生物学领域已久负盛名, 常被誉为蛋白质结构预测的奥林匹克竞赛. CASP是由美国马里兰大学的John Moult于1994年倡导举办, 每两年一届, 目前已成功举办了十余届. 对于参赛者来说, CASP给他们提供了一个能客观评估其预测方法的平台, 且可借此与世界范围内的其他预测方法进行对比, 从而更好地认识自身的优势以及不足. 而对于组织者乃至科学界来说, 通过竞赛可遴选出当前最有效的预测方法, 同时了解整个蛋白质结构预测领域的发展情况, 包括所取得的成绩、存在的困难以及未来的发展方向等.

在竞赛开始前, 必须确保所有参赛者均不知晓目标蛋白信息, 为此, 组织者会选定一些结构暂未经实验测定、或结构已被测定但尚未对外公布的蛋白质作为目标蛋白. 根据预测的难易程度(主要看能否在PDB数据库中找到对目标蛋白有用的结构模板), 所有目标蛋白将被划分成基于模板(TBM)和无模板(FM)两类. 这样做的目的主要是便于后续对相应的两类预测方法进行更合理的评估. 与此同时, 所有参赛方法也被归为人工组和自动组两类, 人工组意味着综合了计算机预测和人工干预, 自动组则纯粹依赖计算机预测. 一般而言, 组织者给人工组的预测时间是三周, 而给自动组的预测时间只有三天. 自动组提交的预测结构会在三天期满后被上传到预测中心的网站上(http://predictioncenter.org), 这些结构接着可被人工组的参赛方法进一步筛选利用. 收集到某个目标蛋白质所有的预测结构后, 组织者便可依据实验测定的结构对其进行综合评估. 除了自动的评估结果, 还会有评估专家对预测结构进行分析和评估, 而评估过程中他们并不知晓每个预测结构来自哪一预测方法或哪个研究组. 评估结果会在CASP大赛当年十二月举行的大会上进行报告, 同时所有的评估数据也将发布于预测中心的网站上, 以便参赛者或其他研究者分析使用. 由评估者、组织者以及优胜的参赛者撰写的相关论文将通过专刊进行发表.

二十余年来, CASP竞赛全方位见证了蛋白质结构预测领域的发展. 在最初几届竞赛中, 基于模板的预测方法得到的预测结构往往比模板本身更加偏离目标蛋白天然结构, 而在最近数届竞赛中, 许多目标蛋白的最佳预测结构都显著优于模板结构, 这主要得益于多模板综合建模的发展以及(基于知识的)力场精度的提高. 无模板的结构预测进步同样也很大, 目前对于100个左右残基长度的无模板类目标蛋白, 不少预测方法能够给出拓扑结构基本正确(RMSD约4—10 Å)的预测结果, 少数例子中甚至能获得RMSD接近于2 Å的预测结构. 图4显示了对CASP无模板目标蛋白进行预测的两个成功的例子. 第一例是I-TASSER利用基于人工智能网络得到的氨基酸配对接触信息对T0604_D1结构进行了成功预测. 第二例是QUARK对T0837的预测结果, 其中也利用了从片断库中提取氨基酸配对接触数据. 可以说, 小蛋白长程氨基酸接触信息对两个例子的成功起到了关键作用. 尽管如此, 较大尺寸的蛋白质(> 150残基)仍是无模板结构预测的最大挑战, 目前为止还很难预测得到具有较高应用价值的结构. 促进无模板结构预测的发展将是CASP竞赛今后最重要的目标和着力方向. 值得一提的是, 由于蛋白质结构数据库所涵盖的折叠构象日趋丰富, 能归入无模板一类的目标蛋白质也愈发有限, 为此, 自2011年起, 组织者又引入了一种新的称为CASP ROLL的竞赛[79], 即在CASP大赛时间之外继续收集无模板目标蛋白质, 以充分利用可能的资源给无模板的预测方法提供更多评估机会.

除了蛋白质三维结构预测, CASP大赛还包括对蛋白质结构其他方面的一些预测方法的评估, 如残基接触预测[80]、无序区域预测[81]、结构质量评估[38]、结构优化等[58]. 历届CASP竞赛的预测结构及评估结果都公开在预测中心的网站上, 以供世界范围内的研究者参考、使用. 有兴趣的读者也可参看历届CASP会议文集(http://www. predictioncenter. org/index. cgi?page=proceedings).

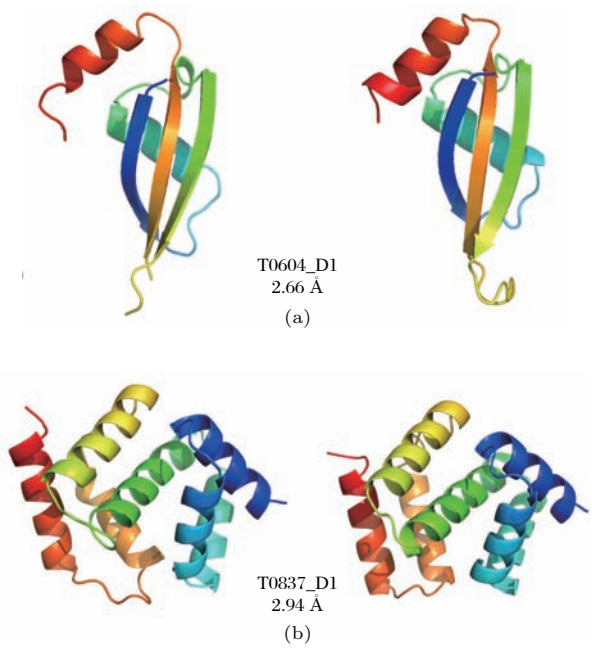

图 4 CASP 大赛中两个成功无模板结构预测的例子 (左右分别为实验和预测结构) (a) I-TASSER 在 CASP9 中对 T0604_D1 预测的第一个模型, RMSD 2.66 Å, 长度 79 个氨基酸, 类别为无模板 β 蛋白; (b) QUARK 在 CASP11 中对 T0837_D1 预测的第一个模型, RMSD 2.94 Å, 长度 128 氨基酸, 类别为无模板 α 蛋白

Fig. 4. Two examples of ab initio modeling in CASP (the left and right panels are X-ray structures and predicted models respectively). (a) The first model of T0604_D1 in CASP9 by I-TASSER, RMSD 2.66 Å, length, 79 classification FM target; (b) the first model of T0837_D1 in CASP11 by QUARK, RMSD 2.94 Å, length 128 Classification, FM target.

5 总结与展望

本文简要介绍了蛋白质结构预测的发展背景、基本的研究步骤及研究内容, 并例举了若干代表性的预测方法; 另外对国际蛋白质结构预测技术评估大赛进行了扼要介绍. 期望这些内容能帮助读者对蛋白质结构预测有一个综合的了解. 可以说, 蛋白质结构预测已成为计算生物学及生物信息学领域最具代表性和影响力的研究方向之一. 它不是一个单一的、孤立的科学问题, 它是一个与不同学科、不同课题相互交融的综合体系. 从生物序列比对、生物数据库构建、生物信息数据挖掘, 到分子力场描述、大分子结构模拟等, 蛋白质结构预测涉及生物信息学、生物物理研究的诸多方面. 它的发展也在很大程度上推动着整个结构生物信息学领域的发展. 随着预测精度以及效率的不断提高, 很多结构预测方法已经从理论走向应用, 成为细胞和分子生物学家的得力工具[82]. 例如, 仅仅 I-TASSER 在线服务器在过去几年时间内就已经为来自世界 120 多个国家和地区的近 7 万名科研工作者提供了蛋白结构预测服务.

尽管如此, 蛋白质结构预测所存在的问题和挑战依然很多. 例如, 目前的序列比对和穿线算法对目标蛋白远程同源模板的识别能力有限, 构象搜索力场的设计远不够精确, 整体拓扑结构与局部结构细节难以同时获得优化, 实验难以测定的膜蛋白结构同样也是结构预测的软肋, 蛋白质复合体结构 (四级结构) 的预测仍困难重重等. 相信随着研究的深入, 这些问题都会逐步得到解决. 而正如 CASP 竞赛组织者所指出的, 目前绝大部分的预测方法都过分依赖于已知的蛋白质结构信息, 即便是无模板的结构预测方法, 几乎也无一幸免. 减少对已知结构的依赖, 强化对本质规律的研究, 这是未来蛋白质结构预测寻求理论突破的必由之径. 我们期待蛋白质结构预测的发展最终能将第二遗传密码揭开.

参考文献

[1] Kolata G 1986 *Science* **233** 1037
[2] Consortium U 2015 *Nucleic Acids Res.* **43** D204
[3] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N, Bourne P E 2000 *Nucleic Acids Res.* **28** 235
[4] Anfinsen C B 1973 *Science* **181** 223
[5] Bowie J U, Luthy R, Eisenberg D 1991 *Science* **253** 164
[6] Jones D, Thornton J 1993 *J. Comput. Aided Mol. Des.* **7** 439
[7] Jones D T, Taylor W R, Thornton J M 1992 *Nature* **358** 86
[8] Jones D T 1999 *J. Mol. Biol.* **287** 797
[9] Chothia C 1992 *Nature* **357** 543
[10] Zhang Y, Skolnick J 2005 *Nucleic Acids Res.* **33** 2302
[11] Huang Y J P, Mao B C, Aramini J M, Montelione G T 2014 *Proteins* **82** 43
[12] Tai C H, Bai H J, Taylor T J, Lee B 2014 *Proteins* **82** 57
[13] Moult J 2005 *Curr. Opin. Struct. Biol.* **15** 285
[14] Kryshtafovych A, Fidelis K, Moult J 2010 *Introduction to Protein Structure Prediction: Methods and Algorithms* (Hoboken: John Wiley & Sons, Inc.) pp15–32
[15] Needleman S B, Wunsch C D 1970 *J. Mol. Biol.* **48** 443
[16] Smith T F, Waterman M S 1981 *J. Mol. Biol.* **147** 195
[17] Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W, Lipman D J 1997 *Nucleic Acids Res.* **25** 3389
[18] Rohl C A, Strauss C E, Misura K M, Baker D 2004 *Methods Enzymol.* **383** 66
[19] Xu D, Zhang Y 2012 *Proteins* **80** 1715
[20] Dill K A, MacCallum J L 2012 *Science* **338** 1042

[21] Pearlman D A, Case D A, Caldwell J W, Ross W S, Iii T E C, Debolt S, Ferguson D, Seibel G, Kollman P 1995 *Comput. Phys. Commun.* **91** 1

[22] Brooks B R, Bruccoleri R E, Olafson B D, States D J, Swaminathan S, Karplus M 1983 *J. Comput. Chem.* **4** 187

[23] Tanaka S, Scheraga H A 1976 *Macromolecules* **9** 945

[24] Miyazawa S, Jernigan R L 1984 *Macromolecules* **18** 534

[25] Sippl M J 1990 *J. Mol. Biol.* **213** 859

[26] Samudrala R, Moult J 1998 *J. Mol. Biol.* **275** 895

[27] Lu H, Skolnick J 2001 *Proteins* **44** 223

[28] Zhou H, Zhou Y 2002 *Protein Sci.* **11** 2714

[29] Rykunov D, Fiser A 2010 *BMC Bioinformatics* **11** 1

[30] Deng H, Jia Y, Wei Y, Zhang Y 2012 *Proteins* **80** 2311

[31] Van Gunsteren W F, Bakowies D, Baron R, Chandrasekhar I, Christen M, Daura X, Gee P, Geerke D P, Glätli A, Hünenberger P H 2006 *Angew. Chem. Int. Edit* **45** 4064

[32] Sugita Y, Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141

[33] Hansmann U H E, Okamoto Y 1999 *Curr. Opin. Struct. Biol.* **9** 177

[34] Li Z, Scheraga H A 1987 *Proc. Natl. Acad. Sci.* **84** 6611

[35] Kirkpatrick S C, Gelatt C D, Vecchi M P 1983 *Science* **220** 671

[36] Swendsen R H, Wang J S 1986 *Phys. Rev. Lett.* **57** 2607

[37] Kihara D, Lu H, Kolinski A, Skolnick J 2001 *Proc. Natl. Acad. Sci.* **98** 10125

[38] Kryshtafovych A, Barbato A, Fidelis K, Monastyrskyy B, Schwede T, Tramontano A 2014 *Proteins* **82** 112

[39] Samudrala R, Levitt M 2000 *Protein Sci.* **9** 1399

[40] Tsai J, Bonneau R, Morozov A V, Kuhlman B, Rohl C A, Baker D 2003 *Proteins* **53** 76

[41] Deng H, Jia Y, Zhang Y 2016 *Bioinformatics* **32** 378

[42] Shortle D, Simons K T, Baker D 1998 *Proc. Natl. Acad. Sci.* **95** 11158

[43] Zhang Y, Skolnick J 2004 *J. Comput. Chem.* **25** 865

[44] Kozakov D, Clodfelter K H, Vajda S, Camacho C J 2005 *Biophys. J.* **89** 867

[45] Maupetit J, Gautier R, Tuffery P 2006 *Nucleic Acids Res.* **34** W147

[46] Gront D, Kmiecik S, Kolinski A 2007 *J. Comput. Chem.* **28** 1593

[47] Rotkiewicz P, Skolnick J 2008 *J. Comput. Chem.* **29** 1460

[48] Li Y Q, Zhang Y 2009 *Proteins* **76** 665

[49] Dunbrack R L, Karplus M 1993 *J. Mol. Biol.* **230** 543

[50] Krivov G G, Shapovalov M V, Dunbrack R L 2009 *Proteins.* **77** 778

[51] Canutescu A A, Shelenkov A A, Dunbrack R L 2003 *Protein Sci.* **12** 2001

[52] Xu J 2005 *Research in Computational Molecular Biology* Cambridge May 14–18, 2005 p423

[53] Miao Z, Cao Y, Jiang T 2011 *Bioinformatics* **27** 3117

[54] Wu S, Skolnick J, Zhang Y 2007 *BMC Biol.* **5** 17

[55] Xu D, Zhang Y 2011 *Biophys. J.* **101** 2525

[56] Zhang J, Liang Y, Zhang Y 2011 *Structure* **19** 1784

[57] MacCallum J L, Pérez A, Schnieders M J, Hua L, Jacobson M P, Dill K A 2011 *Proteins* **79** 74

[58] Nugent T, Cozzetto D, Jones D T 2014 *Proteins* **82** 98

[59] Modi V, Xu Q, Sam A, Roland L, Dunbrack J 2016 *Proteins* **0** 00

[60] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A 2014 *Proteins* **82** 1

[61] Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A 2016 *Proteins* 0

[62] Guex N, Peitsch M C 1997 *Electrophoresis* **18** 2714

[63] Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino T G, Bertoni M, Bordoli L 2014 *Nucleic Acids Res.* **42** 252

[64] Altschul S F, Madden T L, Schäffer A A, Zhang J, Zhang Z, Miller W, Lipman D J 1997 *Nucleic Acids Res.* **25** 3389

[65] Remmert M, Biegert A, Hauser A, Söding J 2011 *Nature Methods* **9** 173

[66] Benkert P, Künzli M, Schwede T 2009 *Nucleic Acids Res.* **37** W510

[67] Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T 2013 *Databsae Oxford* **2013** bat031

[68] Sali A, Blundell T L 1993 *J. Mol. Biol.* **234** 779

[69] Fiser A, Do R K, Sali A 2000 *Protein Sci.* **9** 1753

[70] Shen M y, Sali A 2006 *Protein Sci.* **15** 2507

[71] Kuntal B K, Aparoy P, Reddanna P 2009 *BMC Res. Notes* **3** 1

[72] Roy A, Kucukural A, Zhang Y 2010 *Nat. Protoc.* **5** 725

[73] Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y 2014 *Nature Methods* **12** 127

[74] Wu S, Zhang Y 2007 *Nucleic Acids Res.* **35** 3375

[75] Simons K T, Kooperberg C, Huang E, Baker D 1997 *J. Mol. Biol.* **268** 209

[76] Cheng J, Randall A Z, Sweredoski M J, Baldi P 2005 *Nucleic Acids Res.* **33** 72

[77] Lee J, Kim S Y, Joo K, Kim I, Lee J 2004 *Proteins* **56** 704

[78] Jones, David T 2001 *Proteins* **5** (Suppl.) 127

[79] Kryshtafovych A, Monastyrskyy B, Fidelis K 2014 *Proteins* **82** 7

[80] Monastyrskyy B, D' Andrea D, Fidelis K, Tramontano A, Kryshtafovych A 2014 *Proteins* **82** 138

[81] Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K 2014 *Proteins* **82** 127

[82] Zhang Y 2009 *Curr. Opin. Struct. Biol.* **19** 145

SPECIAL TOPIC — Progress in Soft Matter Research

Protein structure prediction*

Deng Hai-You[1)] Jia Ya[2)] Zhang Yang[3)†]

1) (*College of Science, Huazhong Agricultural University, Wuhan 430070, China*)

2) (*College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China*)

3) (*Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 45108, USA*)

(Received 22 June 2016; revised manuscript received 21 July 2016)

Abstract

Predicting 3D structure of proteins from the amino acid sequences is one of the most important unsolved problems in computational biology and biophysics. This review article attempts to introduce the most recent effort and progress on this problem. After a brief introduction of the background and basic concepts involved in protein structure prediction, we went through the specific steps that have been taken by most typical structural modeling approaches, including fold recognition, model initialization, conformational search, model selection, and atomic-level structure refinement. Several representative structure prediction methods were introduced in detail, including those from both template-based modeling and ab initio folding approaches. Finally, we overview the results shown in the community-wide Critical Assessment of protein Structure Prediction (CASP) experiments that have been developed for benchmarking the state of the art of the field.

Keywords: protein structure prediction, homology modeling, *ab initio* prediction, structure refinement

PACS: 87.14.E–, 82.30.–b **DOI:** 10.7498/aps.65.178701

* Project supported by the National Natural Science Foundation of China (Grant Nos. 11547255, 11474117), the Fundamental Research Funds for the Central Universities, China (Grant No. 2662015BQ045) and the National Institute of General Medical Sciences (GM083107, GM116960).

† Corresponding author. E-mail: zhng@umich.edu