生物物理学报 2011 年 1 月 第 27 卷 第 1 期: 28-37
ACTA BIOPHYSICA SINICA Vol.27 No.1 Jan. 2011: 28-37

基于模板的蛋白质结构预测

黄俊峰, 段 鹏, 吴文言

中山大学生命科学学院有害生物防治与资源利用国家重点实验室，广州 510275

收稿日期：2010-06-07；**接受日期**：2010-08-10
基金项目：国家自然科学基金（30772722）、广州市科技计划（2009JI-C541）资助项目
通讯作者：吴文言，电话：(020)84115570，E-mail：lsswwy@mail.sysu.edu.cn

摘要：基于模板的蛋白结构预测和不依赖模板的蛋白结构预测是计算预测蛋白质三维结构的两种方法，前者由于具有快速和较高准确性的优点，而得到了广泛的应用。基于模板的结构预测是通过寻找与目标蛋白序列相似并且有实验测定的结构作为模板，进而构建目标序列的结构模型的方法。文章详细综述了基于模板的结构预测方法的步骤、关键环节，并对影响结构预测精确性的因素进行了分析和讨论。

关键词：基于模板的蛋白结构预测；序列比对；结构精修；结构模型筛选
中图分类号：Q617

引 言

随着人类基因组计划[1]的成功和新一代测序技术[2]的发展，基因序列能够比以往更轻易地获得。而在后基因组时代，基因的功能和调控成为主要的关注点，其中一个不可忽略的问题是，作为基因产物的蛋白，是怎样通过其空间结构发挥组装、催化等作用的。尽管随着现有的蛋白结晶和结构测定方法的改进[3]，更多的蛋白结构被测定，然而，截至 2010 年 2 月，已测序的非冗余蛋白序列超过 10 500 000 条（NR 库，NCBI：ftp://ftp.ncbi.nlm.nih.gov/blast/db/），相比而言，PDB 蛋白结构数据库（http://www.rcsb.org/）的结构数据只有 63 000 多个，相差巨大。为了更充分地理解蛋白是如何执行其功能以及根据蛋白结构设计出更合理的药物，能否获得蛋白的结构甚至预测其结构就显得尤为重要。

随着计算机硬件的快速发展，科学家们期望能利用计算的方法预测蛋白结构。蛋白结构的预测主要基于以下三个前提：1）蛋白的空间结构能够由氨基酸序列唯一确定[4]；2）蛋白的空间结构是稳定的；3）蛋白的天然构象处于自由能最低点[5,6]。因此，理论上，可以利用计算机的快速计算能力，让序列尝试所有的可能空间构型，找出满足上述条件的构象。但是，这种方法需要付出巨大的时间成本。众所周知，在生物体内，通常一个蛋白的折叠会在小于 1 s 内完成。我们不妨作一个简化的设想[7]：对于一段含有 100 个氨基酸的序列，如果折叠中每一个残基都采取三个构象的变化去寻找合适的结构，则整个过程需要 3^{100} 或者

10^{48} 构象的变化。如果从一个构象转变为另一个则需要约 10^{-12} s，那形成一个天然结构则需要 10^{36}s——这是无法想象的漫长时间。显然，每个残基的构象远远不止 3 个，这就说明，在自然情况下，蛋白并不会尝试所有构象。由此可以设想，蛋白的折叠可能是经历一条动力学轨迹达到最低自由能点，而非逐个地进行构象尝试。这提示人们，可以寻找一种满足这一折叠动力学的算法得到蛋白的最终结构，而不必遍历所有构型，蛋白折叠的问题由此可以转化为可实际操作的计算问题。为了能有效地在巨大的采样空间中找到唯一的蛋白结构，各种不同方法相继提出。在过往文献中，把常用的蛋白结构预测方法分为三类：1）针对高相似序列的同源建模（homology modeling）[8]，也称之为比较建模（comparative modeling）；2）针对较低序列相似性的折叠识别方法（fold recognition，又称之为穿线法，threading method）[9]；3）不依赖于模板而利用物理学原理直接进行从头计算（ab initio or de novo modeling）[10]。随着 1994 年开始的两年一度的 CASP 蛋白结构预测比赛（critical assessment of techniques for protein structure prediction）[11]技术的发展，科学家们把前两类方法统一称之为“基于模板的蛋白结构预测”(template-based modeling，TBM)，把第三种方法称为“不依赖模板的蛋白结构预测”(template-free model modeling，FM)。由于基于模板的预测方法取样空间小，计算速度快，所得结构与不依赖于模板的方法相比可靠性高，因此常作为蛋白结构预测的首选方法。

基于模板的蛋白结构预测

基于模板的结构预测遵循两个前提假设：1) 两条统计上展现相似性的序列，它们可能采取一致的折叠模式[12]；2) 蛋白的折叠方式是有限的，现有的 PDB 数据库中几乎包含了所有可能的折叠模式[13,14]。这里所谓的“统计上的相似性”，既指在实际的比对中，两条序列同一比对位置的氨基酸残基完全相同，另一方面，也指两条序列同一比对位置上氨基酸残基有相似性的替代，但是性质相近或者不变。前一种情况，主要指两序列的有大于 50%的相似性，原先的“同源建模”主要针对这种情况，后一种情况，主要指两序列相似性在 30%～50%之间，对应原先的“折叠识别”。当序列相似性过低的时候，通常难以找到合适的模板，即使通过各种比对方法勉强能找到可能的模板，但结果往往还不如直接作从头预测[15]。

在基于模板的结构预测的实际操作中，各种不同方法采用的预测步骤不尽相同，总体而言，我们可以将其分为四个步骤：1) 待预测的序列在结构数据库（如 PDB）中比对，搜索可能的模板；2) 通过多序列比对确定保守区和可变区，利用模板构建序列的模型结构；3) 用多种方法对第二步获得的结构进行补全、精修和优化；4) 对所得的结构系综（ensemble）进行估计，选择最可能的结构作为最后结果。在整个结构预测过程中，最重要的是序列比对的正确性和结构的精修。

搜索模板结构及序列比对

基于模板的结构比对，首先的而且相当重要的一个步骤是序列比对。给定一条氨基酸序列，基于模板的结构预测首先需要根据这条序列搜索一个或者多个它的模板结构。在序

列比对中，两个因素显著影响结构预测的质量：敏感性和精确性。

敏感性，指的是在比对中找出该给定序列的远源和同源序列的能力。寻找该序列的模板，可以通过给定序列与 PDB 数据库中已测定结构的蛋白序列比对得到。如果不能得到高序列相似度的 PDB 结构，就需要找其相近的可用的模板。在可接受的序列相似度范围内，一个待预测序列允许拥有几个模板，多模板可以更好地确定结构保守性，为后续的建模提供更准确的结果。在满足敏感性的同时，由于需要把待测序列的每一个氨基酸比对到模板上的唯一位置，这时就要求序列比对有较高的精确性。高敏感性和高精确性的要求还基于这样一个原因：在获得结构模板的同时，还需要知道序列上哪些位置是保守的，哪些位置是不保守的，因此需要通过序列比对搜索非冗余的蛋白数据库，尽可能找出该序列的所有同源序列，并得到正确的比对，以便确定序列的保守性。

现有很多不同的序列比对算法，它们得到的结果不尽相同。每一种比对序列都各自有对结果优劣的判断方法，而且这些方法都是基于一定的统计学结果，有时不同方法比对出的结果不尽相同，但它们均满足各自的判断要求，因此通常很难比较各种算法之间的优劣。在常用的 FASTA[16]和 BLAST[17]启发式比对算法中，它们能够快速地在庞大的数据库中搜索，但是这两种方法往往很难满足前述要求。因此，常用的方法有两种，一种是基于替代矩阵(PSSM)的 PSI-BLAST[18]方法，另一种是基于隐马尔科夫模型 (HMM)[19]的方法。相对而言，基于 HMM 的方法敏感性高，但是速度比 PSI-BLAST 要慢。在兼顾速度的同时为了进一步提高敏感性，一个常用的方法是利用 PSI-BLAST 得到序列间的谱 (profile)，结合动态规划算法进行序列比对，通常可以比单纯利用 PSI-BLAST 的方法多找出 20%～30%的同源序列。在 CASP6 比赛中，SPARKS2 和 SP3 方法就成功地运用该方法找出待预测序列的远源同源序列，从而找到最适合的模板[20]。

完成比对后，在选择搜索结果作为模板时，需要注意的是，合适的模板并不一定意味着序列比对的分值最高的蛋白质序列。我们认为在遵循以下几个原则的前提下可以选择到最优的模板，即：尽量选择与未知蛋白质序列最接近的蛋白质结构，即一致性比较高的蛋白质结构；在这些已知的结构中，尽量选择高分辨率和低 R-free 值的结构；结构越完整越好，尤其是在较大的蛋白质分子中，环状结构部分因其残端无法解析出来而容易被忽略；在有特别用途的时候，合适的辅因子或者其他小分子也是一个需要考量的因素。

构建模型结构和精修

以上述模板为原型，构建待测序列的结构模型。由于蛋白骨架是蛋白结构的主体，基于刚体替代的原则[21]，根据序列的保守区将模板结构的坐标直接拷贝到目标中，从而构建目标蛋白的基本主链骨架。由于蛋白侧链的结构由其相互作用的侧链及水环境决定，因此即使相对位置的氨基酸残基相同，其侧链坐标也不拷贝到模型结构中，而由下一步精修的侧链建模产生。预测序列未能与模板结构比对上的区域通常由环区（loop）组成，这一部分的坐标也不直接拷贝到模型中，而由下述的环区建模构建。因此在这一步中得到的模型只是蛋白保守区的主链骨架。

蛋白结构精修

优化模型骨架

在得到蛋白的骨架结构后，由于待测序列的实际结构与模型结构不可能完全相同（例如二级结构的长短、各氨基酸之间的距离等），因此首先需要优化骨架结构，为后续添加侧链保留足够的位置，同时使骨架结构保持接近天然构象（native fold）。简单而言，就是允许骨架在满足一定拓扑学的约束下进行随机扰动。进行随机扰动的方法有多种，用得比较多的有基于片段的组装法和空间约束法。基于片段的组装法[22]源于一个观察发现：现有PDB 数据库中的蛋白结构均可以通过与其无相关性的蛋白的刚性片段组装得到。因此，该法首先需要对现有已知结构的蛋白进行特定长度（通常为 9 个氨基酸残基）的连续分割，构建一个片段集。随后，随机选择模型蛋白中一个位置上的定长序列，根据片段集与该序列的相似性和二级结构匹配，得到一系列匹配分数由高到低的片段子集，然后依次用子集中片段的骨架扭角替换该模型序列的骨架扭角，再利用蒙特卡罗模拟退火搜索策略，根据相应的打分函数计算此时蛋白的构象能量并决定是否接受新的骨架扭角值。重复此过程，直至模型的构象能量达到某一阈值，由此得到模型骨架的一系列诱导构象（decoy sets），然后对诱导构象进行聚类，最终得到若干候选构象。由于这一方法来源于真实蛋白的取样，所以它得到的结果会更接近实际。ROBETTA[23]、I-TASSER[24]等预测方法主要应用这一策略，近年来基于片段组装的方法在不依赖于模板的结构预测中被广泛运用，成功率高居前列。相对的，著名的 MODELLER[25]软件对骨架模型优化采用了空间约束法，即约束骨架在模板构象附近作采样。由于空间约束采样的速度快，可靠性高，因而被广泛采用。

环区（loop）建模

在初步得到模板骨架结构的时候，由于环区残基的可变性而未被构建，需要在环区建模这步补上。环状结构的模拟相对来讲是比较困难的，因为结构变异区通常出现在环状结构中，而且环区将二级结构连接在一起，同时也起着决定活性中心和结合部位功能专一性的重要作用。目前，主要有两种方法模拟环状结构，分别是：从头预测[26,27]和基于已知环区结构[28]的建模方法，或是将两种方法结合起来使用。尽管最近几年，这两种方法得到了快速的发展，但是在使用的过程中仍然受到诸多限制。从头预测环区的方法是根据物理化学和量子化学的原理预测环区序列最低能量的稳定结构，LOOPY 程序[29]就因其较好的精度而在结构预测中被广泛采用。但是，这种从头预测的方法对于较长的环区和较近的端距计算量会大大增加，另外，需要注意的是，能量最低的结构不一定是真实的结构[30]。第二种方法是通过搜寻 PDB 数据库中已知结构的环区序列，根据序列相关或几何基准，如 C 端和 N 端的距离等，找出所有符合目标序列的环状结构所需的个数和端点距离的构型，再通过过滤和筛选过程，选择最合适的环区构象。该方法在某些环状结构的模拟中既准确又有效率，但是受到环区结构数据库大小的限制。

侧链建模

催化、相互作用、信号传递等功能的行使依赖于蛋白侧链[31~33]，蛋白侧链的多样性决定了即使拥有相同骨架也可能有不同的功能[34,35]，因此蛋白侧链的建模是一个难点，模型的

精确性决定了后续分析的准确性。目前，在侧链建模中，多采用主链依赖的旋转异构体库（backbone dependent rotamer libraries）[36]。从 PDB 数据库所得的旋转异构体库可以减少侧链的空间构象的情况，极大地减少了计算量。并且侧链构象库与骨架结构相关，在加快取样速度的同时增加了精确性。但是，相比蛋白内部的侧链建模，由于蛋白表面受环境影响较大而且缺乏类似于内部的侧链之间的约束，对表面侧链的建模仍是一个难题。另外，由于旋转异构体库是对已有 PDB 结构的取样，其构象空间是离散的，而实际上由于原子键的转动，侧链构象是连续的，因此，实际应用中可能会出现构象缺失的情况。尽管有不足，但现在运用最多的 SCWRL 程序[37]仍是基于骨架的旋转异构体库，该程序结合图论方法判断侧链构象的可接受程度，取得了不错的结果。

结构模型筛选

经过以上过程，我们得到待测序列的一系列满足特定要求（如：最小自由能，回转半径等）的模型。最后，需要根据一些标准，在这一系列模型中选出一个最接近天然构象的作为最终模型。筛选标准主要包括符合能量函数，过滤掉不符合的蛋白，以及对结构进行聚类等。

事实上，从蛋白结构精修开始，便需要使用能量函数去判断添加的侧链或者骨架的构象是否合理。能量函数主要分为两类[38]，从已知蛋白结构的残基分布、溶解可及性、局部原子密度等性质推导出的基于统计的函数（statistical effective energy function，SEEFs），以及从分子力学和物理规律如范德华力、静电相互作用、氢键能量、二面角扭转能量等推导出的基于物理的函数（physical effective energy functions，PEEFs）。基于物理的函数其前提可靠、意义明确，易于被接受。但在区分天然构象和非天然构象上，目前基于物理的函数由于量子力学计算的复杂和难以正确拆分能量表达式而造成精确性不足[39]，并且由于需要计算大量原子间的相互作用力[40]，基于物理的函数耗用计算资源且效率不高，这些问题仍待解决。而基于统计的函数由于来源于已测定结构的蛋白，在判断蛋白是否达到天然构象的效果上优于基于物理的函数，PROCHECK[41]、Verify3D[42]等程序即利用统计函数来判断预测得到的模型是否落在天然构象的区域内，MODELLER 则利用从正确折叠蛋白推导出的 DOPE 函数[43]对结果模型进行判断。虽然近年来基于统计的函数在蛋白结构预测中获得了相当好的结果，但是由于其来源于统计而导致的精确性不足，往往难以区分多个接近同一天然构象的模型。在实际应用中，常根据需要混用两种能量函数对模型结构进行判断。而 Rosetta 则根据不同预测过程的精度要求，分别使用了基于残基水平和基于全原子水平的统计学能量函数和物理学能量函数[44]。更深入的关于能量函数的讨论，可以参见 Boas 的综述[45]。

由于能量函数在结构精修的过程中已被采用，这就在一定程度上削弱了它选择最终模型时的效果，因此，预测过程中往往还采用能量无关的计算方法去过滤不符合要求的结果。这些方法包括蛋白的空间拓扑结构[46]、表面形态[47]、回转半径[48]、氨基酸的接触顺序[49]等等，它们能把符合能量函数要求但明显折叠错误的非天然构象筛选出来。

在经过以上步骤筛选后仍剩余众多蛋白结构模型，这时就需要通过聚类分析获得最后的结构。当有最多的蛋白聚于同一个簇时，便认为这一簇的蛋白是最接近天然构象的。要

做聚类，就必然要求之前对模型的构建有充分大的采样，能够产生足够多的用于聚类的模型，这需要耗用大量的计算资源。在今天硬件资源得到极大发展的情况下，这几乎成为一个标准采样过程，Rosetta 和 TASSER 便在选择最终结构时采用了聚类方法。

基于模板的蛋白结构预测的应用

基于模板的结构预测可以进行蛋白质的结构与功能关系分析及蛋白质的分子设计包括药物设计等。

各种因素都将引起蛋白质结构和功能的改变，研究人员可以借助于计算机建立的原子水平的分子模型来模拟分子的结构和行为，例如通过替换、添加、删除残基来对多肽进行调整，或用旋转异构体优化侧链构象；或利用不能形成氢键的残基 (如丙氨酸) 取代在天然态中形成氢键的残基等方法，模拟实现蛋白质的 “突变”。并以此分析分子体系的静电势、键强弱、亲疏水性、能量分布等各种物理和化学性质，以及稳定态和折叠过程的动力学性质等，以揭示其结构与功能的关系。如 Jean-Luc Pons 等[50]为解决蛋白 - 配体复合物建模中难以选择模板、流程繁复等问题而构建了 @TOME-2 在线比较建模服务。真核生物中 Ssu72 蛋白由于序列差异大导致功能难以确定，而早前的实验仅显示酵母的 Ssu72 具有磷酸酶活性，却未能得到进一步信息。针对这个问题，研究人员应用 @TOME-2 服务对酵母中的 Ssu72 蛋白进行了建模和对接，成功预测其具有磷酸酶的功能，并指出它可能与四硝基苯磷酸（4-nitro-phenyl phosphate）结合，同时展示了其对接模式和关键位点，为之后的实验验证提供了明确的方向。N- 乙基马来酰亚胺敏感因子（N-ethylmaleimide-sensitive factor，NSF）对细胞膜的融合起关键作用，对其生化机理的研究积累了大量数据，但是由于它包含的活性 ATPase 结构域（D1）的六聚体结构一直无法获得，使得从结构上解释 NSF 的机理难以实现。针对这一情况，Thielmann 等[51]应用 Modeller 软件，结合 3 个不同的模板（中国仓鼠 NSF 蛋白构建 N 结构域、D2 结构域和小鼠 p97/VCP 蛋白构建 D1 结构域）构建了人的全长 NSF 结构，并根据 N、D2 结构域的信息和其他证据，确定其反平行六聚体的组装方式，进一步结合结构上残基的疏水性，利用分子对接的方法确定 NFS 与 GABA 受体连接蛋白（GABAA receptor-associated protein）的结合方式，并通过蛋白体外结合实验（GST pull-down）加以证实。本实验室早前也通过同源模建的方法，删除尿激酶原蛋白的部分肽段并添加导向分子获得了一种新型的溶栓分子 GZ5-sPA 蛋白,酶学分析结果显示该突变体蛋白被激活后，其溶栓活性比标准尿激酶的活性高约 3 倍[66]。

基于模板的蛋白结构预测面临的挑战

尽管在两年一度的 CASP 比赛促进下，蛋白结构预测尤其是基于模板的蛋白结构预测取得了巨大的进展，而且由于其相对较少的计算量和高的可靠性而备受重视，并且在研究中广受应用，但是仍有一些问题亟待解决。

首先，在获得蛋白的模型结构后，如何精修得到接近原子解释度的精确性，关乎能否正确分析蛋白的功能。尽管 D.Baker 小组[52,53]在这方面取得了不俗的成绩，但是，如何使采样得到的模型与天然结构的均方根偏差（root mean square deviation，RMSD）差异减少到

2Å 以内，仍是一个挑战。

在高序列相似性（>50%）下，建模得到的结构一般可以接受，但是在较低（30%～50%）的序列相似度下，多年来仍未能取得突破性进展[54]，CATH 数据库[35]中很多结构相近但是序列高度不同的蛋白便很难通过基于模板的方法建模，而现阶段不基于模板的建模虽然取得不错的进展，但是对模型结构是否真正与天然结构相符仍难以判断[55,56]。

而且，“序列相似的蛋白其结构相似”这一前提假设，在某些情况下也受到质疑。如 Alexander[57,58]的研究指出，一个全 alpha 螺旋的蛋白和一个以 beta 折叠为主的蛋白，原先其序列相似性只有不到 15%，却可以在基本不影响其结构的条件下，通过改变序列上的氨基酸残基，使两蛋白的序列相似性高达 88%，而保持各自的结构不变。如果说这一结果带有很高的人工构建痕迹，那么 Roessler[59,60]发现的一对进化和功能都相关的蛋白，却在较高序列相似性的情况下采取了截然不同的折叠方式，表明了自然中确实存在这种情况。虽然不十分普遍，但这也让人对基于模板的结构预测的结果可靠性提出疑问。需要注意的一个类似情况是，即使是同一个蛋白的序列，在其行使不同功能或者处于不同环境的时候，其结构可能会应需要而作出改变[61]，这就给基于模板的结构预测方法提出了挑战。

理论计算指出，尽管蛋白序列多种多样，但其折叠方式可能只有有限的种类，并且 PDB 数据库中已有的结构几乎涵括了所有的类型[13,14]。但是，Taylor 等人[62]最近对蛋白结构的模拟结果表示，如果不依赖于从已知蛋白推导出的能量函数，仅按照疏水包埋等理想状态的约束预测蛋白的结构，在预测得到的结构中只有十分之一与数据库中已知结构相似。即，还有十分之九可能的结构由于某种原因未被发现，但是目前也没有办法排除其存在的可能性。Taylor 把这十分之九的未被发现的结构称为蛋白折叠空间中的“暗物质”（dark matter）区。如果情况真如 Taylor 所言，还有大量蛋白的折叠模式未被发现，那基于模板的蛋白结构预测是否足够可靠就值得怀疑。

另外，随着对蛋白结构和功能研究的加深，科学家发现很多蛋白是没有十分固定的结构的，他们把这种蛋白称为内在无序蛋白（intrinsic disorder protein，IDP）[63]。研究发现，人体内很多蛋白，特别是一些在信号通路中起重要作用的蛋白是部分或者完全没有固定结构的[64]，这就让结构的预测变得更加困难。2008 年结束的 CASP8 比赛，也着重指出对蛋白中的无序结构的预测，仍然是一个十分困难的问题[65]。

总结与展望

蛋白质结构预测是后基因组时代最具挑战性的问题之一，基于模板的预测方法是相对较为有效的预测方法，目前仍然没有哪种预测方法能够达到没有偏差的理想结果[67]。这些困扰蛋白质结构预测领域发展的“瓶颈”问题的解决，也将会为计算机辅助药物设计、蛋白质设计、药物的高通量筛选等领域提供有力的理论保证和全新的研究手段。

今后该领域的研究将集中在以下几个方面：1) 低序列相似性的蛋白质的结构的预测；2) 序列比较算法的优化；3) 寻找用于模拟支配蛋白折叠和蛋白之间相互作用的动力学研究的力场；4) 在模型的优化的方面，实现从拓扑水平到原子水平方向的发展；5) 建立有效、准确的预测膜蛋白结构的方法。

参考文献：

1. Klug A. The Human Genome Project. *IUBMB Life*, 2001, 51(1): 1~4
2. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*, 2008, 26(10): 1135~1145
3. Hura GL, Menon AL, Hammel M, Rambo RP, Poole FL, 2nd, Tsutakawa SE, Jenney FE Jr, Classen S, Frankel KA, Hopkins RC and others. Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). *Nat Methods*, 2009, 6(8): 606~612
4. Anfinsen. Principles that govern the folding of protein chains. *Science*, 1973, 181: 223~230
5. Bryngelson JD, Wolynes PG. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA*, 1987, 84(21): 7524~7528
6. Wolynes PG. Energy landscapes and solved protein-folding problems. *Philos Transact A Math Phys Eng Sci*, 2005, 363(1827): 453~464; discussion 464~467
7. Feng HQ, Zhou Z, Bai YW. A protein folding pathway with multiple folding intermediates at atomic resolution. *Proc Natl Acad Sci USA*, 2005, 102(14): 5026~5031
8. Xiang ZX. Advances in homology protein structure modeling. *Curr Protein Pept Sc*, 2006, 7(3): 217~227
9. Jones D, Thornton J. Protein fold recognition. *J Comput Aided Mol Des*, 1993, 7(4): 439~456
10. Osguthorpe DJ. Ab initio protein folding. *Curr Opin Struct Biol*, 2000, 10(2): 146~152
11. Moult J, Pedersen JT, Judson R, Fidelis K. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 1995, 23(3): ii~v
12. Ginalski K, Grishin NV, Godzik A, Rychlewski L. Practical lessons from protein structure prediction. *Nucleic Acids Res*, 2005, 33(6): 1874~1891
13. Du PC, Andrec M, Levy RM. Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. *Protein Engin*, 2003, 16 (6): 407~414
14. Chothia C. Proteins. One thousand families for the molecular biologist. *Nature*, 1992, 357(6379): 543~544
15. Lee J, Liwo A, Ripoll DR, Pillardy J, Scheraga HA. Calculation of protein conformation by global optimization of a potential energy function. *Proteins*, 1999, Suppl 3: 204~208
16. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA*, 1988, 85(8): 2444~2448
17. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 1990, 215(3): 403~410
18. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res*, 1997, 25(17): 3389~3402
19. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, 1998, 14(10): 846~856
20. Zhou H, Zhou Y. SPARKS 2 and SP3 servers in CASP6. *Proteins*, 2005, 61(Suppl 7): 152~156
21. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. *Protein Sci*, 2005, 14(5): 1315~1327
22. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. *J Comput Aid Mol Des*, 2003, 17(11): 725~738
23. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res*, 2004, 32: 526~531
24. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*, 2007, 69 (Suppl 8): 108-117
25. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A. Comparative protein structure modeling using Modeller. *Curr Protoc Bioinformatics*, 2006, Chapter 5: Unit 5~6
26. Bruccoleri RE, Karplus M. Prediction of the folding of short polypeptide segments by uniform conformational sampling. *Biopolymers*, 1987, 26(1): 137~168
27. Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA. A hierarchical approach to all-atom protein loop prediction. *Proteins*, 2004, 55(2): 351~367
28. Lee DS, Seok C, Lee J. Protein loop modeling using fragment assembly. *J Korean Phys Soc*, 2008, 52 (4): 1137~1142
29. Xiang ZX, Soto CS, Honig B. Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction. *Proc Natl Acad Sci USA*, 2002, 99(11): 7432~7437
30. Felts AK, Gallicchio E, Chekmarev D, Paris KA, Friesner RA, Levy RM. Prediction of protein loop conformations using the AGBNP implicit solvent model and torsion angle sampling. *J Chem Theory Comput*, 2008, 4(5): 855~868
31. Bowers KE, Fierke CA. Positively charged side chains in protein farnesyltransferase enhance catalysis by stabilizing the formation of the diphosphate leaving group. *Biochemistry*, 2004, 43(18): 5256~5265
32. Jackson RM. Comparison of protein-protein interactions in serine protease-inhibitor and antibody-antigen complexes: implications for the protein docking problem. *Protein Sci*, 1999, 8(3): 603~613
33. SmockRG, Gierasch LM. Sending signals dynamically. *Science*, 2009, 324(5924): 198~203
34. Schluter KD. PTH and PTHrP: Similar Structures but Different Functions. *News Physiol Sci*, 1999, 14: 243~249
35. Orengo CA, Pearl FM, Bray JE, Todd AE, Martin AC, Lo Conte L, Thornton JM. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res*, 1999, 27(1): 275~279

36. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool. *J Mol Biol*, 1997, 267(5): 1268~1282
37. Canutescu AA, Shelenkov AA, Dunbrack RL Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 2003, 12(9): 2001~2014
38. Lazaridis T, Karplus M. Effective energy functions for protein structure prediction. *Curr Opin Struct Biol*, 2000, 10(2): 139~145
39. Tsai J, Bonneau R, Morozov AV, Kuhlman B, Rohl CA, Baker D. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins*, 2003, 53(1): 76~87
40. Misura KM, Baker D. Progress and challenges in high-resolution refinement of protein structure models. *Proteins*, 2005, 59(1): 15~29
41. Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, 1996, 8(4): 477~486
42. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. *Methods Enzymol*, 1997, 277: 396~404
43. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. *Protein Sci*, 2006, 15(11): 2507~2524
44. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using rosetta. *Methods Enzymol*, 2004, 383: 66~93
45. Boas FE, Harbury PB. Potential energy functions for protein design. *Curr Opin Struct Biol*, 2007, 17(2): 199~204
46. Mucherino A, Costantini S, di Serafino D, D'Apuzzo M, Facchiano A, Colonna G. Understanding the role of the topology in protein folding by computational inverse folding experiments. *Comput Biol Chem*, 2008, 32(4): 233~239
47. Kuhn LA, Siani MA, Pique ME, Fisher CL, Getzoff ED, Tainer JA. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *J Mol Biol*, 1992, 228(1): 13~22
48. Lobanov MY, Bogatyreva NS, Galzitskaya OV. Radius of gyration as an indicator of protein structure compactness. *Molecular Biology*, 2008, 42(4): 623~628
49. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, 1998, 277(4): 985~994
50. Pons JL, Labesse G. @TOME-2: A new pipeline for comparative modeling of protein-ligand complexes. *Nucleic Acids Res*, 2009; 37(Web Server issue): W485~91
51. Thielmann Y, Weiergraber OH, Ma P, Schwarten M, Mohrluder J, Willbold D. Comparative modeling of human NSF reveals a possible binding mode of GABARAP and GATE-16. *Proteins*, 2009, 77(3): 637~646
52. Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D. High-resolution structure prediction and the crystallographic phase problem. *Nature*, 2007, 450(7167): 259~264
53. Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei JM, Kim D, Kellogg E, DiMaio F, Lange O and others. Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins-Struct Funct Bioinform*, 2009, 77: 89~99
54. Zhang Y. Protein structure prediction: When is it useful? *Curr Opin Struc Biol*, 2009, 19(2): 145~155
55. Ben-David M, Noivirt-Brik O, Paz A, Prilusky J, Sussman JL, Levy Y. Assessment of CASP8 structure predictions for template free targets. *Proteins-Struct Funct Bioinform*, 2009,77: 50~65
56. Zhang Y. I-TASSER: Fully automated protein structure prediction in CASP8. *Proteins-Struct Funct Bioinform*, 2009, 77: 100~113
57. Alexander PA, He Y, Chen Y, Orban J, Bryan PN. The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci USA*, 2007, 104(29): 11963~11968
58. He Y, Chen Y, Alexander P, Bryan PN, Orban J. NMR structures of two designed proteins with high sequence identity but different fold and function. *Proc Natl Acad Sci USA*, 2008, 105(38): 144112~144127
59. Roessler CG, Hall BM, Anderson WJ, Ingram WM, Roberts SA, Montfort WR, Cordes MH. Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity butdifferent folds. *Proc Natl Acad Sci USA*, 2008, 105(7): 2343~2348
60. Davidson AR. A folding space odyssey. *Proc Natl Acad Sci USA*, 2008, 105(8): 2759~2760
61. Takano K, Katagiri Y, Mukaiyama A, Chon H, Matsumura H, Koga Y, Kanaya S. Conformational contagion in a protein: Structural properties of a chameleon sequence. *Proteins*, 2007, 68(3): 617~625
62. Taylor WR, Chelliah V, Hollup SM, MacDonald JT, Jonassen I. Probing the "dark matter" of protein fold space. *Structure*, 2009, 17(9): 1244~1252
63. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: Intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit*, 2005, 18(5): 343~384
64. Shimizu K, Toh H. Interaction between intrinsically disordered proteins frequently occurs in a human protein-protein interaction network. *J Mol Biol*, 2009, 392(5): 1253~1265
65. Noivirt-Brik O, Prilusky J, Sussman JL. Assessment of disorder predictions in CASP8. *Proteins*, 2009, 77 (Suppl 9): 210~216
66. Lin J, Yang XY, Deng RQ, Yu BG, Lai HJ, Sun WL, Wu WY. Soluble expression of a strong thrombolytic pro-urokinase mutant in *Escherichia coli*. *Protein Expres Purif*, 2006, 48(1): 69~73
67. Liu YL, Tao L. Discussion of the mothods of protein structure prediction. *China J Bioinform*, 2007, 5 (4): 185~187

Template-Based Protein Structure Prediction

HUANG Junfeng, DUAN Peng, WU Wenyan

State Key Laboratory of Biocontrol, School of Life Science, Sun Yat-Sen University, Guangzhou 510275, China

This work was supported by grants from The National Natural Science Foundation of China (30772722) and The Science and Thchnology Foundation of Guangzhou (2009JI-C541)
Received: Jun 7, 2010 **Accepted:** Aug 10, 2010
Corresponding author: WU Wenyan, Tel: +86(20)84115570, E-mail: lsswwy@mail.sysu.edu.cn

Abstract: Methods for predicting three-dimensional structure of protein molecules can be classified into two categories: template-based modeling and template free modeling. Template-based prediction is widely used in application due to its speed and relatively high accuracy. This technique predicts the three-dimensional structure of a given protein sequence based primarily on its alignment to one or more proteins of known structure. In this paper, the progress in the methodology for template-based modeling is reviewed. The critical steps and factors that influence the prediction accuracy are analysed and discussed.
Key Words: Template-based modeling; Sequence alignment; Model refinement; Model selection