蛋白质三级结构预测算法综述

王 超[1),2)] 朱建伟[1),2)] 张海仓[1),2)] 巩海娥[1),2)] 郑伟谋[3)] 卜东波[1)]

[1)](中国科学院计算技术研究所 北京 100190)

[2)](中国科学院大学 北京 100049)

[3)](中国科学院理论物理研究所 北京 100190)

摘 要 了解蛋白质的三维结构对于认识蛋白质的功能有着重要意义.由于蛋白质结构测定的速度远远跟不上蛋白质序列测定的速度,因此使用计算技术依据蛋白质序列预测结构成为结构测定的有力补充.该文首先总结了蛋白质结构预测的 3 类基本方法,包括基于序列-序列联配的同源建模法、基于序列-结构联配的归范法以及基于最小化能量函数的从头预测法,并分析了其中的关键技术;进而总结了有代表性的蛋白质结构预测软件工具,然后通过对蛋白质结构预测 CASP 比赛结果的分析比较了各种方法的性能,并获得了如下结论:当待预测蛋白质与模板蛋白质序列等同度超过 30%时,同源建模法能够产生高质量的预测结果,归范法中的远同源检测以及从头预测法中的能量函数设计是尚待突破的关键点;最后,该文总结分析了未来的发展趋势,并阐释了强序列信号"绑架"蛋白质构象生成的观点,即从整体来说,蛋白质序列与结构的关联关系并不显著,但其中某些特定的局部序列片段具有非常强的结构倾向性,这些强信号区域引导蛋白质的折叠过程,对这些强信号区域的认识将会有助于提升蛋白质结构预测算法的性能.

关键词 蛋白质结构;同源建模;归范法;从头预测法;能量函数;动态规划;线性规划

中图法分类号 TP399 **DOI 号** 10.11897/SP.J.1016.2018.00760

A Survey on Algorithms for Protein Tertiary Structure Prediction

WANG Chao[1),2)] ZHU Jian-Wei[1),2)] ZHANG Hai-Cang[1),2)] GONG Hai-E[1),2)]
ZHENG Wei-Mou[3)] BU Dong-Bo[1)]

[1)](*Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*University of Chinese Academy of Sciences, Beijing* 100049)

[3)](*Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing* 100190)

Abstract Proteins are important macromolecules that consist of one or more long chains of amino acids, and perform a great variety of biological functionalities, including catalyzing, transportation, and responses to stimuli. In natural environment, a protein usually folds into a specific tertiary structural conformation called native structure. The functionalities of a protein are largely determined by its structural conformation, rending the understanding of protein structures fatally important. However, the experimental approaches for the determination of protein structures are commonly labor-intensive and time-consuming, and thus cannot match up the speed of protein sequencing. Therefore, it is invaluable to predict protein structures using

收稿日期:2015-10-08;在线出版日期:2016-03-02.本课题得到国家"九七三"重点基础研究发展规划项目基金(2012CB316502)、国家自然科学基金(31270834,61272318,11175224,11121403)和中国科学院理论物理研究所理论物理国家重点实验室开放工程项目(Y4KF171CJ1)资助.
王 超,男,1987 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为蛋白质结构预测、算法设计.E-mail: 0472452@fudan.edu.cn. **朱建伟**,男,1991 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为蛋白质结构预测、机器学习. **张海仓**,男,1987 年生,博士研究生,中国计算机学会(CCF)会员,主要研究方向为蛋白质结构预测、机器学习. **巩海娥**,女,1991 年生,硕士研究生,中国计算机学会(CCF)会员,主要研究方向为蛋白质结构预测、机器学习. **郑伟谋**,男,1946 年生,研究员,博士生导师,主要研究领域为表面物理、随机过程、非线性动力学、统计物理. **卜东波**,男,1973 年生,研究员,博士生导师,主要研究领域为算法设计与分析、生物信息学.

computational approaches. In this study, we summarized the popular strategies for protein structure prediction, including homology modeling approaches based on sequence-sequence alignment, threading approaches based on sequence-structure alignment, and ab initio approaches based on optimizing energy functions. Furthermore, we listed the popular software packages for protein structure prediction as well as their performance in the CASP (Critical Assessment of protein Structure Prediction) competition. The evaluation of these approaches showed the following three observations: (1) If the sequence identity between a query protein and a certain template protein exceeds 30%, homology modeling approaches usually report accurate prediction results; (2) For threading approaches, how to improve fold recognition for remote homology proteins remains one of the challenges; and (3) For ab initio approaches, the design of an accurate energy function, together with building structural conformation with the assistance of the information of residue-residue contact information still need investigation. Finally, we summarized this study by listing our perspectives on protein structure prediction, especially on the the perspective that protein folding is an elite-driven process. Specifically, the correlation between protein sequence and its native structure is not that strong from the global point of view; however, at certain regions, local sequences carry strong signals of local structural preference. These regions might initialize the protein folding process, and speed up protein folding through significantly reducing search space of possible structural conformations. The understanding of these regions should greatly facilitate the developing of novel approaches for protein structure prediction. In summary, the accurate prediction of protein structures heavily relies on our insights into the relationship between protein sequence and structure, as well as modeling these insights into an efficient statistical models and algorithms. Protein structure is a representative problem that a linear sequence of entities forms specific complex structures under the interactions among these entities; thus, the advances in the field of protein structure prediction will contribute to solving other similar problems.

Keywords　protein structure; homology modeling; threading; ab initio; energy function; dynamic programming; linear programming

《计算机学报》

1　引　言

蛋白质是生命体的重要组成部分.据估计,生命体的细胞中大约15%～20%是蛋白质,是含量最高的有机物.

蛋白质的基本组成单元是氨基酸.常见的氨基酸有20种,由中心碳原子及其相连的氨基(－NH_2)、羧基(－COOH)、氢原子以及区别氨基酸的侧链构成.一个氨基酸的氨基可以和另一个氨基酸的羧基脱水缩合形成肽键.蛋白质即是由氨基酸(特指脱水反应之后的氨基酸)以肽键(见图1)相连而成的一条长链.

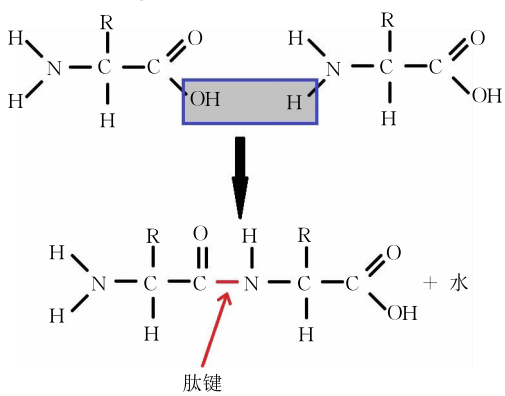


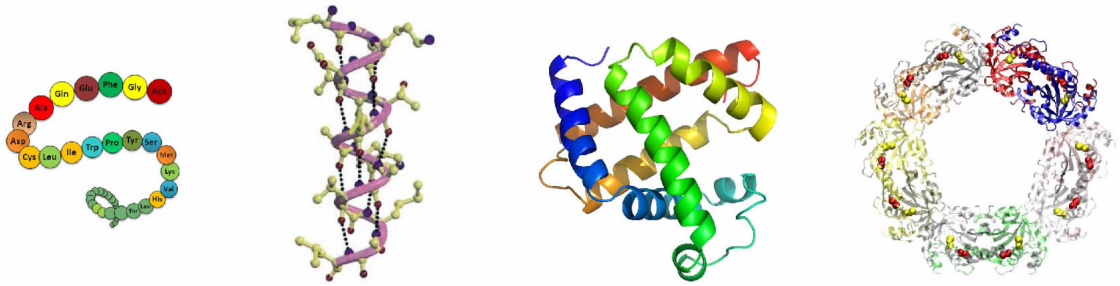
图1　氨基酸脱水后形成肽键,从而连接成长链

在天然环境下,蛋白质在大量的非共价作用之下自发形成特定的结构,以行使其生物学功能.常见的非共价作用有疏水作用、范德华力(van der Waals' force)、离子键、氢键等.

通常来说,蛋白质的结构可分为4个层次(见图2).

(1)一级结构是指构成蛋白质的残基序列;

(2)在蛋白质的局部片段上,常常会观察到主链上的周期性规则构象,比如α螺旋和β片层等,这



一级结构 二级结构 三级结构 四级结构

图 2 蛋白质的 4 个层次结构示意图

种周期性的规则主链构象被称为二级结构；

(3) 蛋白质的三级结构是指由多个二级结构单元拼装而成的具有紧密形态的三维空间结构；

(4) 四级结构描述由不同多肽结构链间相互作用形成具有功能的蛋白质复合物分子[1].

蛋白质只有在折叠成特定的结构之后方能行使特定的生物学功能，因此了解蛋白质结构对于认识其功能有着重要意义[1]. 比如，疯牛病即是由朊蛋白(Prion Protein，PrP)的结构变异引发——在正常情况下，朊蛋白是水溶性的 α 螺旋结构，但在某些诱因下会转变成不溶于水的 β 片层结构，从而沉积在脑组织中并引发神经细胞退行性改变[2].

测定蛋白质三级结构的主流实验方法包括 X-晶体衍射和核磁共振(NMR)等. X-晶体衍射技术能够获得高精度的蛋白质结构，然而，很多蛋白质难以制备用于结构解析的晶体，因此无法通过此方法测定结构. 而核磁共振方法不需要制备蛋白质晶体，但一般仅能测定长度不超过 300 个氨基酸的小蛋白. 值得指出的是：最近冷冻电镜技术发展迅速，其主要优势在于能够测定大蛋白质的结构[3-5].

由于蛋白质结构实验测定的速度远远跟不上序列测定的速度，无法满足大规模蛋白质结构解析的需求，因此使用计算的方法根据蛋白质序列预测结构是非常有意义的工作. 此外，Anfinsen 等人的实验表明：在一般情况下，蛋白质能够自发折叠成特定的结构构象[6]. 换句话说，蛋白质的结构信息蕴含于其序列之中，从而表明从序列出发进行结构预测是可行的.

蛋白质结构预测的直观思路是直接使用分子动力学技术模拟蛋白质的折叠过程. 模拟蛋白质的折叠过程有助于理解蛋白质的折叠机理，以及能量函数设计. 然而，由于需要大量的计算资源，以及缺乏对蛋白质折叠过程的深刻认识，现有的分子动力学模拟技术无法进行快速、准确的结构预测[7]，而基于统计学习和组合优化的预测算法则成为预测蛋白质结构的主要手段.

本篇综述介绍了蛋白质结构预测方法，分别就二级结构预测、三级结构预测的同源建模法(Homology Modeling)、归范法(Threading，旧译穿线法)和从头预测法(Ab initio，又译本初法)进行陈述和分析，并通过对蛋白质结构预测 CASP(The Critical Assessment of protein Structure Prediction)竞赛结果的分析总结了蛋白质结构预测领域的发展动态和趋势. 此外，本文列举了主要的蛋白质数据库以及在线的蛋白质结构预测服务. 本文采用蛋白质结构领域的常用规范，将待预测结构的蛋白质称为目标蛋白质(Target protein)，将已知结构的蛋白质称为模板蛋白质(Template protein).

值得指出的是，蛋白质结构预测问题与计算机科学之间存在着紧密的联系. 一方面，蛋白质结构预测的各类算法中大量使用了现有的计算机科学中的研究成果，比如整数线性规划、马尔可夫随机场、条件随机场等. 蛋白质是“一维序列在相互作用下形成复杂结构”这一现象的典型代表，因此也是检验计算机科学相关成果的良好平台. 另一方面，蛋白质结构预测领域的进展也推进了计算机科学的发展，比如为刻画多序列联配而设计的序列列型隐马尔可夫模型(Profile-HMM)，是生物信息学领域特有的 HMM 模型，也是对 HMM 的一个有益的扩展.

2 蛋白质结构的一维表征及其预测

蛋白质的空间结构是用原子的三维坐标来精确表示的. 此外，蛋白质结构还可以使用一维序列进行表征，这种方法在保持蛋白质结构描述能力的同时，大大简化了描述. 另一方面，对这些一维结构表征的准确预测，能够减小三维结构的搜索空间，因此对三维结构预测起着十分重要的作用.

本节介绍两种蛋白质结构的一维表征，即二级结构和结构码以及相应的预测方法.

2.1　基于主链氢键定义的二级结构

蛋白质二级结构由特定的主链氢键模式所定义. 在蛋白质中，一个氨基酸的酰胺氢原子与另一氨基酸上的羰基氧原子之间形成氢键. 这种主链上的氢键对于维持蛋白质结构起着重要作用. 根据主链氢键模式，研究者定义了规则的蛋白质二级结构，主要为 α 螺旋和 β 片层. 1951 年，Pauling 等人[8]提出蛋白质二级结构字典（Dictionary of Protein Secondary Structure，DSSP），Kabsch 等人[9]将二级结构详细分为 8 种状态，分作 3 类，即螺旋（G、H、I）、片层（E）以及卷曲（T、B、S、C），详见表 1.

表 1　蛋白质二级结构 8 种状态[9]

标注	描述对象
G	3-氨基酸螺旋
H	4-氨基酸螺旋
I	5-氨基酸螺旋
T	氢键转角
E	正平行/反平行延展片层
B	孤立氢键片层
S	大弯折
C	其他卷曲

在蛋白质二级结构预测方面，Jones 等人[10-11]提出一种利用蛋白质进化信息（Information of the evolutionarily related proteins）的预测方法 PSIPRED；Wang 等人[12]提出用条件神经场（Conditional Neural Field，CNF）模型进行 8 态二级结构预测的概率方法；Faraggi 等人[13]提出多步学习（Multi-step learning）策略，并利用其他结构特征如溶液可及表面面积（Solvent accessible surface area）以及主链扭角（Backbone torsion angles）来提高二级结构预测准确率，详细介绍如下.

PSIPRED 方法的基本思想是利用进化上相关蛋白质的信息[10-11]. 整个预测过程分为 3 个阶段：产生序列列型（Sequence profile）、初始二级结构预测以及预测二级结构后处理. 首先，PSIPRED 利用 PSI-BLAST（下一节详述）产生目标蛋白质的序列列型，以表示进化上相关蛋白的序列信息；然后采用神经网络针对每个位置预测局部片段的三态二级结构. 神经网络的输入层是中心残基和相邻左右各 7 个残基的 15 个列型，每个列型是 20 种氨基酸和 1 个空位（Gap）的离散分布，因此共计 315 个特征；输出层是中心残基和相邻左右各 2 个残基长度为 5 的三态二级结构. 最后，用另一个神经网络来对第一次预测的结果进行过滤，输入层依然为长度为 15 的滑动窗口，中间包含了一个 60 个神经元的隐藏层，输出为中心残基的 3 种状态. 对于每个中心残基，PSIPRED 给出二级结构三态的预测打分，并将打分最高的二级结构作为预测结果. PSIPRED 的三态二级结构预测精度在 76.5%与 78.3%之间.

Wang 等人[12]提出一种基于概率图模型条件神经场的二级结构预测方法 RaptorX-SS8. RaptorX-SS8 对三态二级结构标注以及八态二级结构标注都可以进行预测. 该条件神经场的输入层特征由 1 个位置相关特征以及 4 个位置无关特征组成. 位置相关特征是由 PSI-BLAST（下一节详述）产生的位置特异打分矩阵 PSSM（Position Specific Scoring Matrix）；位置无关特征包括是否是二级结构单元终结点的倾向性（Propensity of being endpoints of SSs）、氨基酸物理化学性质（Physico-chemical property）、氨基酸关联接触势能（Correlated contact potential of amino acids）以及氨基酸序列. 条件神经场输出层为相邻氨基酸二级结构构成的二元组. 当神经网络得到输出层之后，通过前向-后向算法计算出每个位置氨基酸标注为三态二级结构或者八态二级结构的概率.

Faraggi 等人[13]开发的 SPINE X 采用迭代策略预测二级结构，其基本模型是仅有输入层和输出层的神经网络. 在每一轮迭代中，输入层不仅包含了序列位置特异打分矩阵以及氨基酸物理化学性质，还包含有前一次迭代所输出的扭角预测值（Torsion angles）以及相对溶液可及面积 RSA（Related solvent accessibility）预测值. 在此基础上，Heffernan 等人[14]利用迭代深度学习（Iterative deep learning）的方法将二级结构、扭角以及相对溶液可及面积一起预测，通过上一轮迭代的预测结果作为神经网络的上一层来进行下一轮预测. 实验结果表明，这种方法对二级结构三态标注的准确度能达到 81.7%.

蛋白质二级结构预测是空间结构预测的基础，其准确度直接影响到空间结构的预测[11]. 现有的三级结构预测软件中，都普遍使用二级结构预测的结果作为特征. 然而，在蛋白质的某些位置上，序列与结构间的关联并不很强[15]，其空间结构是在局部与全局结构竞争过程中形成，因此在这些位置上如何提高二级结构预测准确率是一个难点.

2.2　描述四肽局部结构的结构码

与基于氢键标注的二级结构不同，结构码能够表示四肽的局部结构. 具体地说，结构码采用伪键表示蛋白质结构，即用链中每个氨基酸的中心碳原子 C_α 代表整个氨基酸. 伪键表示的好处在于相邻 C_α

计算机学报

之间距离比较稳定(约 3.8 Å);因此,对于包含 n 个氨基酸的肽段来说,采用角度表示的结构($\theta_2;\tau_3,\theta_3;\cdots;\tau_{n-1},\theta_{n-1}$)和三维空间坐标表示的结构($r_1,r_2,\cdots,r_n$)具有等价性.具体到四肽来说,只使用两个折角一个扭角 θ,τ,θ' 即可表示其三维结构.

对真实蛋白质结构的分析表明:表征四肽的两折角和一扭角(θ,τ,θ')的分布有显著的成簇性;使用高斯混合模型对四肽角度进行聚类,最终获得 17 个簇,每簇对应着一种典型的局部结构;将这些簇按照字母编号,称之为结构码(Conformation LEtter, CLE).图 3 显示了 17 种结构码对应的典型局部结构,其中,H,I,J,K 为典型的螺旋结构,C,D,E,F 为典型的片层结构,而其余 9 种描述了不同的卷曲结构.以蛋白质 1ctfA 为例,图 4 显示了其空间结构、残基序列以及对应的结构码和二级结构.

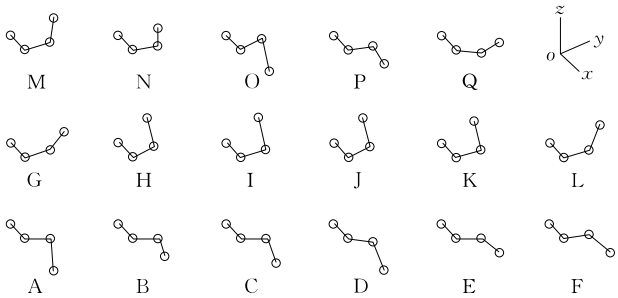
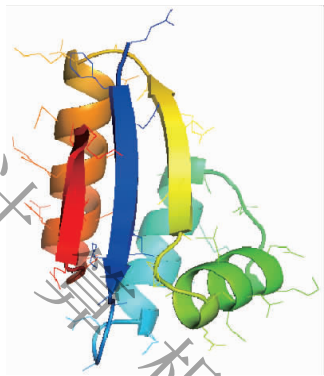


图 3 17 种结构码的形状描述[16]



SEQ: EFDVILKAAGANKVAVIKAVRGATGLGLKEAKDLVESAPAALKEGVSKDDAEALKKALEEAGAEVEVK
CLE: RRDEEBPLBGNMMIHHHHHHHHHIKAGCAIIIHHHHHLFNGBPLENGCAHHHHHHHHHHHHHOLDEEER
SS: CEEEEEEECGGGHHHHHHHHHHHHCCCHHHHHHHHHTCSEEEEEEECHHHHHHHHHHHHHHTCEEEEC

图 4 蛋白质 1ctfA 序列、结构码及二级结构

结构码的典型应用是蛋白质结构联配,即将三维的结构-结构联配问题归结为一维的序列-序列联配问题,此处的序列指结构码序列.与氨基酸序列相似性打分 BLOSUM62[17] 类似,结构码之间也可构造相似性打分 CLESUM[18],以衡量结构码之间的替换概率.使用结构码序列打分可以快速地计算两结构联配和多结构联配.

结构码的另一个应用是对一类相似结构进行标注,比如一类结构模体可以使用不含空格的一串结构码来描述,从而提高结构分类查询的效率.这也是对层次结构分类(例如 SCOP[19])的补充.

在结构码预测方面,Zeng[20] 提出了基于隐马尔可夫模型(Hidden Markov Model, HMM)的预测方法,将序列列型(Profile)、预测的二级结构、预测的相对溶液可及面积作为特征,以结构码作为输出,其预测准确率为 37.3%.笔者的团队应用条件随机场(Conditional Random Field)的方法,利用相同的特征预测结构码,准确率约为 40%.结构码是描述局部结构的重要表征,然而,相邻位置的结构码与氨基酸的关联并不足够显著,这是提高结构码准确率的难点之一.

2.3 小结

蛋白质结构的一维表征及其预测是三级结构预测的基础,常常作为目标蛋白的特征参与到整体结构预测之中,因此其准确度直接影响到结构预测的精度.除二级结构及结构码以外,残基的扭角与溶液可及面积也是蛋白质空间结构在残基层次上的一维表征,也常用于蛋白质空间结构预测.

蛋白质三级结构的预测方法主要包括同源建模法、归范法以及从头预测法,下文分节详述.

3 同源建模法

3.1 基本思想

同源建模法的出发点是认为序列相似的蛋白质其结构也相似[15].基于这种认识,同源建模法计算目标蛋白质序列与模板蛋白质序列之间的相似性,并以序列联配的形式呈现序列相似性计算结果,最

《计算机学报》

后从联配出发构建出目标蛋白质的空间结构（见图 5）. 其关键步骤包括如下两点：(1) 如何表示蛋白质序列信息；(2) 如何衡量并计算序列相似性.

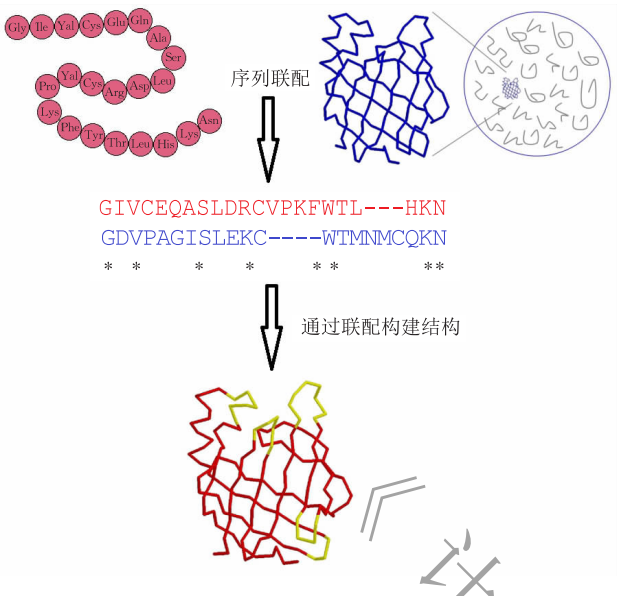


图 5　同源建模法示意图（同源建模法首先计算目标蛋白质序列与模板蛋白质序列之间的序列相似性，然后根据联配以及模板蛋白质的结构构建出目标蛋白质的结构）

3.2　典型的序列-序列联配算法

典型的同源建模法包括 BLAST[21]、PSI-BLAST[22-23]、FFAS[24-25]、ORFeus[26]、SAM-T99[27]、HHpred 等[28-30]. 上述算法的不同之处主要体现在蛋白质序列表示模型不同以及序列相似性计算方式的不同.

简要地说，在蛋白质序列表示方面，初始是直接采用残基序列的表示方案，后续发展为采用序列列型以包含同源序列信息，进而扩展到使用隐马尔可夫模型（Hidden Markov Model，HMM）以及马尔可夫随机场（Markov Random Field，MRF）. 蛋白质序列的表示方案不同，直接导致了序列相似性计算方法的不同.

BLAST[21] 直接考虑目标蛋白质序列与模板序列之间的联配. 简要地说，BLAST 通过以将模板序列作为参考，将目标序列标记为 M（匹配）、I（插入）、D（删除）这 3 种状态，目的是获得目标序列与模板序列的最优联配. BLAST 以 M/I/D 状态链的方式表示序列联配，根据残基打分矩阵（如 BLOSUM）衡量联配打分，并采用动态规划技术求出最优联配. PSI-BLAST[22-23] 是 BLAST 的改进版本，它采用迭代方式，根据位置特异打分矩阵计算出序列列型与模板序列的联配.

FFAS[24-25] 采用序列列型方式来表示目标蛋白质序列和模板蛋白质序列，并使用序列列型-序列列型（Profile-profile）比较来衡量两条序列之间的相似性. 和单条序列相比，序列列型的优势在于能够表示在每个残基位点出现各种氨基酸和空位（Gap）的可能性.

和 FFAS 相比，ORFeus 的特色在于使用元序列列型（Meta-profile）表示蛋白质序列信息[26]，即在序列列型中添加了预测的二级结构信息. 简要地说，ORFeus 先预测目标蛋白质中每个残基的二级结构倾向性；然后在序列联配中考虑目标序列的预测二级结构与模板的真实二级结构之间的相似度，从而提高序列联配的准确度.

序列列型的优势在于能够刻画每个位点产生各种变异的可能性. 然而，由于采用了位置无关的空位罚分，导致序列列型联配方法倾向于生成目标序列与模板序列的“成块”（Block）联配. 与序列列型表示方案不同，SAM-T99[27] 使用序列列型隐马尔可夫模型（Profile-HMM）来表示模板蛋白质序列，每个位置具有特异的空位罚分（以 M/I/D 状态转移概率刻画），从而避免了序列列型联配的缺陷. 从另一个角度看，SAM-T99 建立了一个产生式模型，即认为目标蛋白质序列是从模板所对应的隐马尔可夫模型产生的，目标蛋白质与模板之间的最优联配则转化为隐马尔可夫模型的标记问题，可以使用 Viterbi 动态规划算法求解. 在 CAFASP1 比赛中，SAM-T99 表现出比 PSI-BLAST 更好的预测准确度.

HHpred[28] 则做了进一步的推广，即使用序列列型隐马尔可夫模型同时刻画模板蛋白质序列和目标蛋白质序列；模板序列与目标蛋白质序列之间的最优联配则转化成两个隐马尔可夫模型之间的最优联配问题. 为计算两个隐马尔可夫模型之间的联配，HHpred 建立一个包含多组“成对状态”的有限状态自动机（Finite State Automaton，FSA）；“成对状态”中的一个状态来自于目标蛋白质的隐马尔可夫模型，另一个则来自于模板蛋白质的隐马尔可夫模型. 每个成对状态下，会同时发射出两个残基（见图 6）. 两个隐马尔可夫模型之间的相似性，可以通过这两个模型产生相同残基序列的概率来计算.

由于考虑了相邻位置的关联性信息，HHpred 在联配中表现出比 SAM-T99 更优的性能，从而能构建出更高质量的结构.

《计算机学报》

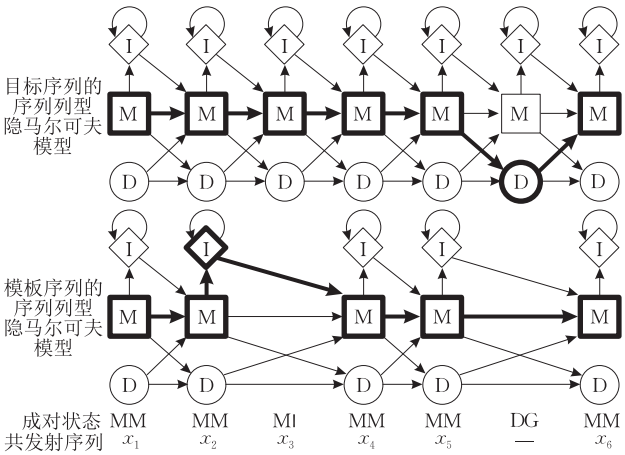


图 6 序列列型隐马尔可夫模型之间的最优联配[28]

3.3 同源建模法小结

同源建模法适用于在模板库中存在同源序列的目标蛋白质.已有的同源建模法的区别体现在如何描述序列信息.事实上,对序列信息表示方案的不断细化贯穿了同源建模法的发展历程.比如最早的BLAST 简单地使用氨基酸序列信息;PSI-BLAST 与 FFAS 进一步刻画了位置特异的序列信息;SAM-T99 采用序列列型隐马尔可夫模型描述模板蛋白质序列信息,允许设置位置特异的空位罚分;HHpred 则对目标蛋白质序列和模板蛋白质序列同时建立序列列型隐马尔可夫模型.需要指出的是,同源建模法考虑的序列特征都是描述残基的局部特征,不考虑序列远程的残基之间的关联,从而使得序列联配可以高效地计算.

尽管同源建模法具有速度快的优势,但是其预测精度则依赖于目标蛋白质与结构模板之间的序列等同度(Sequence identity).当序列等同度比较高(大于 30%)时,同源建模法一般能够以较高的精度预测出蛋白质三级结构;而在序列等同度较低时则往往失效(见图 7).在 CASP7 会议上,Baker 认为

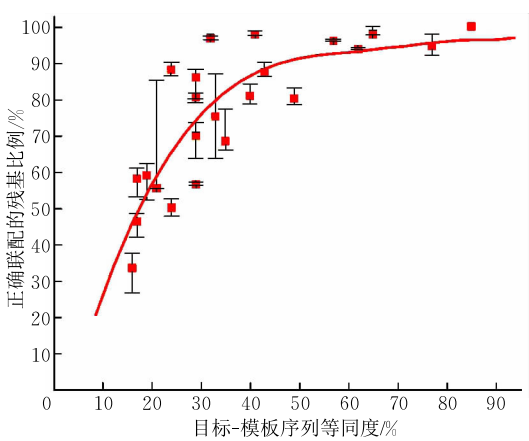


图 7 预测准确度和序列等同度之关系[31](当序列等同度小于 30%时,正确联配的氨基酸比例迅速下降)

"同源建模法事实上已经被终结了",并建议下一步的研究重点应该放在使用归范法和从头预测法对于序列等同度低于 30%的蛋白质进行结构预测.

4 归范法

4.1 基本思想

与同源建模法相同,归范法(Threading,旧译穿线法)也是要计算目标蛋白质与模板蛋白质之间的联配,但不同之处在于同源建模法仅仅使用了模板蛋白质的序列信息,而归范法还利用了模板蛋白质的结构信息,如残基间的接触、残基溶液可及性等.

归范法考虑目标蛋白质序列是否适应于模板蛋白质结构;相应地,归范法的基本手段不再是同源建模法所采用的序列-序列联配,而是序列-结构联配,即假设将目标蛋白质"归入"模板的"模范"结构中,观察目标蛋白质序列是否与模板结构相匹配(见图 8).更详细地说,模板蛋白质中的每个残基都有其局部的"结构微环境",因此在联配时需要衡量目标蛋白质中的残基是否适合某个"结构微环境",这也是"归范法"这一译法的由来.

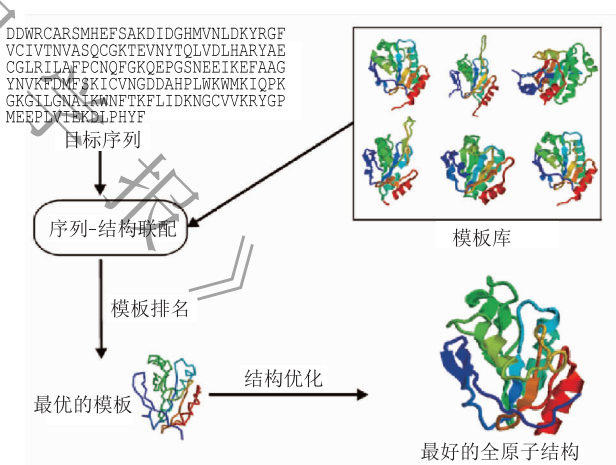


图 8 归范法示意图(归范法首先计算目标蛋白质与模板蛋白质结构的适配程度,然后依据序列-结构联配生成目标蛋白质结构)

归范法的关键步骤包括:(1)如何刻画模板中每个残基的局部"结构微环境";(2)如何衡量并最优化目标蛋白质序列与模板蛋白质结构之间的序列-结构联配;(3)如何生成最终的预测结构.

4.2 典型的序列-结构联配算法

典型的序列-结构联配算法包括 Prospect[32]、SPARKS[33-34]、Raptor 系列[35-37]、MRFalign[38] 等,描述如下.

Prospect[32]使用如下的打分函数衡量序列与模

《计算机学报》

板结构之间的匹配程度：

$$E = \omega_1 E_{\text{mutate}} + \omega_2 E_{\text{single}} + \omega_3 E_{\text{pair}} + \omega_4 E_{\text{gap}},$$

其中变异项 E_{mutate} 表示模板蛋白质上的某个残基变异成目标蛋白质相联配残基的可能性，通常采用 PAM250 打分矩阵；单体项 E_{single} 是指目标蛋白质中某个残基适应于模板上某个位置的结构环境的程度，此处的结构环境包括二级结构和溶液可及性等；二体接触项（Pairwise contact）E_{pair} 用来衡量目标蛋白质中两个残基形成接触（C_β 原子距离小于 7 Å）的可能性；而空位罚分项（Gap penalty）E_{gap} 则用来避免联配中形成过多的空位. Prospect 采用“分而治之”策略来计算使得上述打分函数最优的序列-结构联配. 值得注意的是，由于上述打分函数中包含了二体接触项，因此在将两个子问题的解合并成整个问题的解时，需要枚举所有的连接左右两个部分的二体接触情形，因此算法的复杂度是 $O(2^c)$ 的，其中 c 是模板结构中相互接触的最大覆盖数.

SPARKS 算法的特色表现在模板蛋白质的序列列型生成方式上[33-34]，即对于每一个模板蛋白质，首先通过结构联配获得与其相似的其他模板，然后将联配位点的序列列型合并在一起形成新的序列列型（Structure driven profile）. 在结构微环境方面，SPARKS 考虑了主链扭角和溶液可及程度，相应的打分函数项有：(1) 主链扭角项以衡量目标蛋白质中残基适应模板结构中主链扭角的程度；(2) 溶液可及项以衡量目标蛋白质中残基适应于模板结构中溶液可及情况的程度. 由于上述各打分项都是残基层次的单体项，因此可以利用动态规划算法高效地计算出最优的序列-结构联配.

Raptor 的特色在于将序列-结构联配问题形式化成一个整数线性规划模型[35]. 详细地说，Raptor 首先将模板分成两部分：一部分是由二级结构片段构成的核心（Core）；另一部分是连接核心的卷曲. 然后做了两个简化假设：(1) 在序列-结构联配中，联配空位只会出现在卷曲区域；(2) 残基间的相互作用只出现在核心区域上，即用核心区域间的相互作用代替残基间的相互作用（见图 9）.

在此基础上，Raptor 设计了包含残基变异打分、环境适应性、空位罚分、二级结构打分以及二体相互作用打分的序列-结构联配打分函数. 由于打分函数中的二体相互作用需要考虑序列距离较远的一对残基，所以无法使用动态规划技术快速求解. Raptor 使用整数线性规划技术求解最优的序列-结构联配，线性规划的约束主要包括：(1) 要求模板中

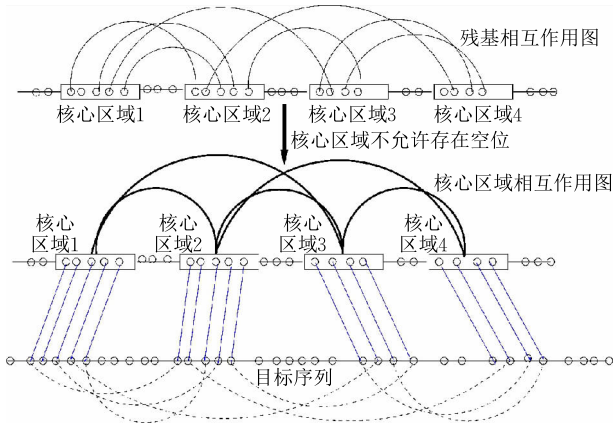


图 9　Raptor 中采用的模板残基相互作用图（在构建序列-结构联配中，仅考虑核心区域之间的相互作用[35]）

的一个核心区域只能联配到目标蛋白质序列中的一个位置；(2) 相邻的两个核心联配的位置不能有交叉. 在 CASP 竞赛中的实验结果表明，Raptor 能够有效地提高序列-联配精度.

在 Raptor 的基础上，RaptorX[36-37] 作了进一步的发展，其主要特色有两点：

(1) RaptorX 不再使用整数线性规划技术求解序列-结构的最优联配，而是采用神经网络（Neural network）技术直接预测目标蛋白质中的残基与模板蛋白质中的残基如何联配. 如图 10 所示，神经网络的输入特征包括目标蛋白质残基的序列信息、以及模板蛋白质残基的结构微环境信息，而输出层则是一条表征联配状态（包括匹配、目标蛋白质插入、模板蛋白质插入）的序列.

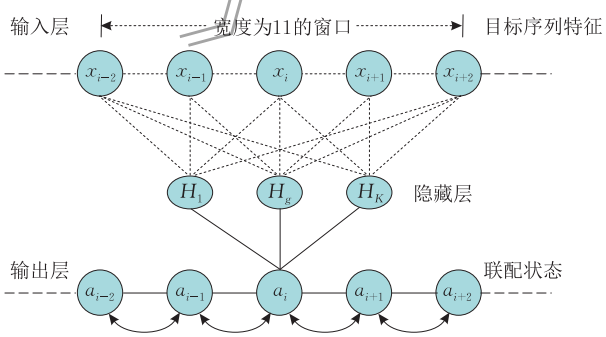


图 10　RaptorX 采用神经网络计算直接预测目标蛋白质上残基的匹配状态[36]

(2) RaptorX 还引入了多模板策略以相互校正，尽量排除单模板的联配错误. 其直观思想是：当计算目标蛋白质和两个模板结构 T_1、T_2 之间的序列-结构联配时，理想的序列-结构联配应该与模板之间的结构联配相一致. 而 T_1 与 T_2 之间的结构联配可以事先计算. RaptorX 建立了一个概率联配模型来尽量提高联配之间的一致性，此方法在 CASP8 和

计算机学报

CASP9 中都表现出优异的性能.

另一个联配算法 MRFalign[38] 可以看作是 HHpred 的一个发展. 如前所述,HHpred 是对目标蛋白质和模板蛋白质分别建立序列列型隐马尔可夫模型,然后对两个隐马尔可夫模型进行联配. 然而序列列型隐马尔可夫模型的缺陷在于无法刻画残基间的远程关联,一种解决方案是使用马尔可夫随机场技术,其中概率图中的连边用于刻画残基间的远程关联. 相应地,目标蛋白质与模板蛋白质之间的联配问题也就转化成两个马尔可夫随机场之间的联配问题. 在 MRFalign 中,衡量联配质量的打分函数由两部分构成,其中单体项刻画序列列型等单残基信息,而两体项则刻画了马尔可夫随机场中有关联的残基对之间的打分(见图 11). 鉴于马尔可夫随机场之间的最优联配计算是 NP-hard 问题,MRFalign 采用交替方向乘子法(Alternating Direction Method of Multipliers,ADMM)以计算次优联配.

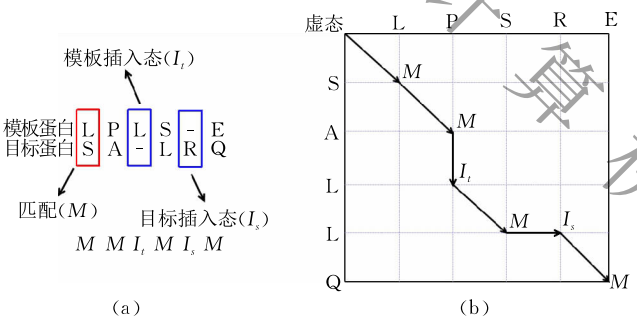


图 11 MRFalign 联配示意图((a)为联配状态的表示;(b)为在联配矩阵中用一条路径来描述特定的联配[38])

4.3 模板选择与折叠型识别

在获得了目标蛋白质与模板结构之间的联配之后,一个有待解决的问题是模板选择,即在多个序列-结构联配中选出最可能的模板. 该问题也称为折叠型识别(Fold recognition)问题,即在模板库中寻找与目标蛋白质具有相同折叠型(Fold)的模板蛋白质.

直观的模板选择方法是直接利用序列-结构联配的原始打分或者依据原始打分计算出的 P 值(P value). 由于原始联配打分通常都和序列-结构联配长度密切相关,因此当模板长度相差很大时不能提供准确的模板选择信息.

目前主流的模板选择方法都基于机器学习技术,比如支持向量机(Supported Vector Machine,SVM)、随机森林(Random forest)、系综分类(Ensemble classification)等,描述如下.

Xu 等人[35] 首次采用 SVM 技术,依据序列-结构联配来判断目标蛋白质与模板蛋白质是否属于同一个折叠型. 此 SVM 模型的输入特征包括序列-结构联配的原始打分、联配中的目标序列和模板的等同残基个数、有联配的模板核心区域与未联配的模板核心区域之间的接触残基数、二级结构打分、空位罚分等. SVM 使用上述特征作为输入,预测目标蛋白质与模板蛋白质的结构联配质量. 实验结果表明:预测出的结构联配质量与真实值之间的相关系数达到 0.95. 这表明,SVM 模型能够有效地进行模板选择.

随机森林技术也被应用于模板选择,其代表性算法是 RF-Fold[39]. 如图 12 所示,RF-Fold 有多棵决策树构成,决策树的输入是序列-结构联配信息,输出是目标蛋白质与模板结构是否属于同一折叠型. RF-Fold 综合多棵决策树的结果以计算目标蛋白质与模板属于同一个折叠型的概率,并最终返回概率最高的折叠类型.

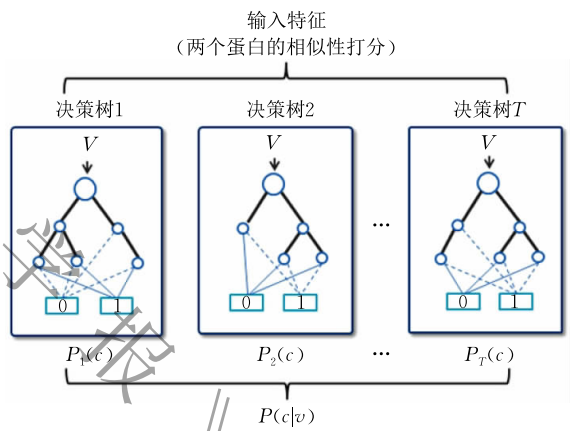


图 12 判别两个蛋白质折叠型是否相同的随机森林算法[39]

Lin 等人[40] 提出了对于不同层次的蛋白质折叠型进行检测的系综分类方法. 具体地说,Lin 等人设计了 188 维的特征向量以刻画蛋白质中的残基出现规律以及物理化学性质,并通过随机森林模型在 SCOP 数据集上训练了系综分类器,以避免单个分类器的缺陷. 在 SCOP 的 7 个蛋白质类(Class)中的实验结果表明:上述方法的折叠型识别率有显著提高.

Wei 等人[41-42] 在此上述方法的基础上加入了预测的二级结构的信息,并开发了软件 PFPA. 在 3 个标准数据集上的实验结果表明 PFPA 方法达到 73.6%的准确率,显著提高了折叠型识别率.

折叠型识别中的一个重要子问题是如何提取蛋白质的序列特征. Zhao 等人[43] 提出了一种综合使用主成分分析、遗传算法和支持向量机技术提取并筛选蛋白质模体特征的方法. 在 6 个超家族的实验结

果表明上述方法具有显著的优势. Huang 等人[44]提出基于残基亲水性的特征提取方法，并利用支持向量机训练蛋白质折叠型的分类器，在视紫红质(Rhdopsin)的家族中达到很高的蛋白质折叠型识别率.

此外，为克服训练集合中的正样本和负样本之间数目的不均衡性，Zhao 等人[45]利用自助法(Bootstrap)生成系综分类器，实验结果显示这种系综分类器的方法能够有效提高蛋白质分类准确度.

模板选择(折叠型识别)对于提高归范法的准确率有着重要的作用. 此问题在本质上是监督学习问题. 如何选择适当的特征以及如何将这些特征加入到机器学习模型中去，是值得进一步研究的关键问题.

4.4　远同源检测

归范法的难点在于远同源关系的识别，即当目标序列与模板序列在进化上远同源时，如何检测出可靠的模板蛋白质.

对于远同源检测问题，Liu 等人[46]利用 PSI-BLAST 生成的多序列联配信息，并组合使用 3 种支持向量机(SVM-Ngram，SVM-pairwise，SVM-LA)进行远同源检测；Liu 等人[47]还借鉴自然语言处理中的排名学习(Learning To Rank，LTR)方法，开发了 ProtDec-LTR 软件，进行远同源检测. 在标准数据集测试的结果显示 ROC 曲线面积达到 0.8442，具有较明显的优势.

笔者的团队开发的 FALCON@home[48]首先从模板蛋白质中抽取出序列和结构共保守区域，然后在进行同源检测时，仅考虑目标序列与模板中的共保守区域的相似性. CASP11 的结果显示这种方法有助于提高预测精度[48].

4.5　预测结构的生成方法

在获得目标蛋白质和模板蛋白质之间的联配之后，最后一个步骤是如何从联配出发计算各个残基的原子的空间坐标，从而生成目标蛋白质的预测结构. 在此步骤，常用的软件和算法包括 MODELLER[49]、MDS[50](Multiple Dimensional Scaling)和 I-TASSER[51-52]，陈述如下.

MODELLER 的基本思想是生成“满足残基间距离和角度约束的结构”[49]，所采用的约束主要包括：(1)残基二面角应该符合统计倾向性；(2)目标蛋白质中的残基之间的距离，应该尽可能与模板蛋白质上对应残基间距离一致. 此外，MODELLER 还考虑了原子间位阻等约束. 上述约束被描述成一个概率密度函数，MODELLER 使用共轭梯度法、模拟退火技术以及分子动力学技术来优化此概率密度函数，以生成最优的预测结构.

MDS 的基本思想是“依据残基间的相互距离计算最优的空间结构”[50]. 简要地说，MDS 从目标蛋白质与模板蛋白质之间的联配以及模板结构信息出发，首先推断目标蛋白质中联配上的残基之间的距离；然后将三维坐标恢复建模成一个优化问题，其目标为最小化各残基空间距离与推断距离之间的差异.

I-TASSER 算法和以上两种算法有较大差异，主要体现在 I-TASSER 对归范法产生的与模板联配中仅考虑可靠的局部结构[51-52](见图 13). 首先，

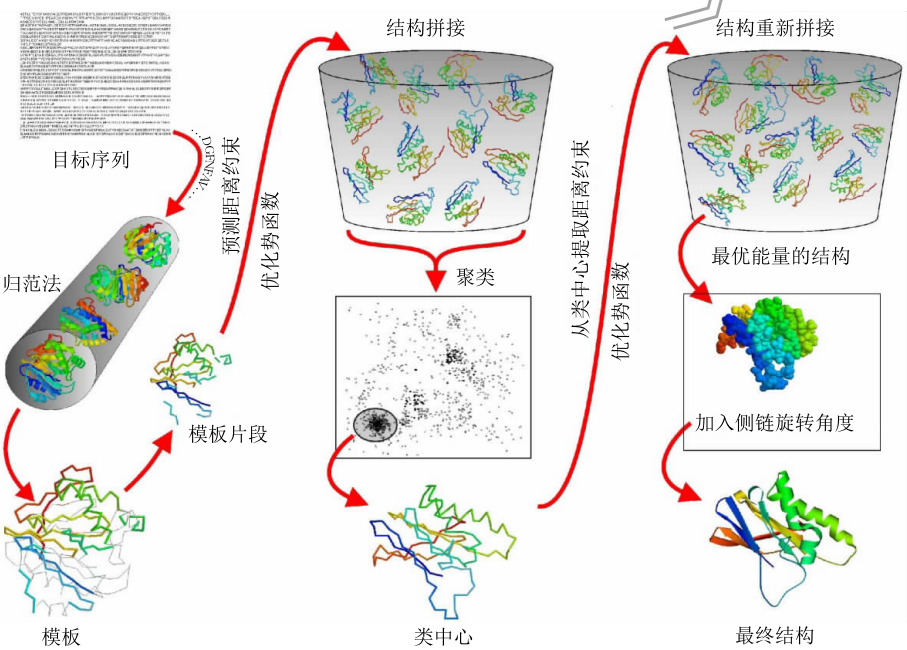


图 13　I-TASSER 算法示意图[52]

I-TASSER 调用 PROSPECT 等归范法软件来获得初始的序列-结构联配；进而从这些初始联配中挑选出最可靠的区域作为局部结构候选，对于卷曲区域内的氨基酸则使用“格点”模型近似其结构.

I-TASSER 的另一个特色在于其能量函数包含了从序列-结构联配中预测出的远程相互作用. 实验结果表明，这样推断出的远程相互作用能够有效地促进最终结构的生成.

4.6 归范法小结

一般来说，归范法能够得到比同源建模法更精确的预测结果，其主要原因在于：归范法能够充分利用模板库中的结构信息，比如氨基酸之间的相互作用、溶液可及性(Solvent accessibility)等，而这是同源建模法所不具备的优势.

蛋白质形成过程充分体现了“序列-结构-进化”三者之间的紧密联系. 在蛋白质的进化过程中，随机突变引发蛋白质序列发生变化[53](被 Marks 等人形象地称为“在序列空间的采样”)；另一方面，选择性压力(Selective pressure)迫使进化过程中蛋白质整体结构保持一定的稳定性，以保证功能的稳定性. 这种选择性压力限制了那些可能破坏结构稳定性的突变. 从这个角度说，同源蛋白质中记录了大量的进化事件，体现了序列变异与结构、功能稳定性之间的平衡.

进一步来讲，近同源蛋白质往往表现出较强的序列相似性，这也是现有的归范法技术成功的原因之一；然而对于远同源(Remote homologue)蛋白质来说，虽然结构相似性依然保持，但是序列相似性不太明显，不足以支撑现有的归范法技术取得较高的精度.

当前常用的归范法多采用隐马尔可夫或者链式条件随机场等统计模型直接进行序列与结构之间的联配，没有显式地说明哪些位置是结构稳定性关键区域，从而在设计联配打分函数时，将关键区域与非关键区域同等看待，每个位置所训练出来的权重是根据极大匹配正确率(Maximum ACcuracy，MAC)得到的，并不能反应在结构稳定性上各个氨基酸或者氨基酸片段的影响.

因此在序列-结构联配中，假如能够对关键区域联配和非关键区域联配分别处理，设计不同的概率模型以反映上述区域进化速率上的差异，有可能显著提高整体联配的精度. 这正是未来归范法的发展趋势.

5 从头预测法

5.1 基本思想

与同源建模法和归范法不同，从头预测法不依赖于已知结构数据库，因此即使在模板数据库中不存在与目标蛋白相似的结构时，从头预测法也能够对目标蛋白质预测. 换句话说，从头预测法具有发现新的结构类型的可能性.

从头预测法一般采用一种粗粒化的片段拼接策略，即利用多个蛋白质结构的片段拼接成目标蛋白结构(见图 14). 其优势在于：直接采用真实结构的片段进行拼接，能够大大降低能量函数设计的难度. 现有的拼接算法大多遵循如下 3 个步骤：

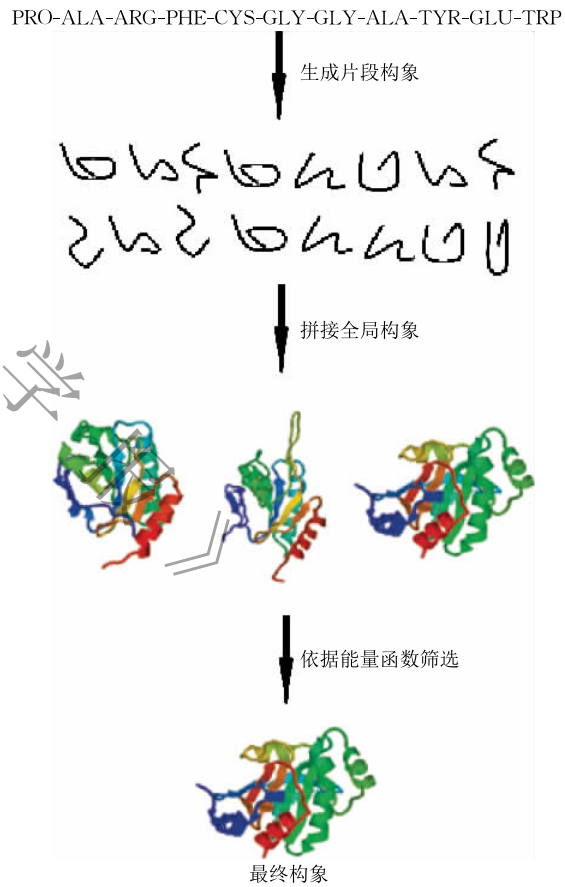


图 14 蛋白质从头预测法示意图

(1) 局部结构预测. 采用滑动窗口等方式将目标蛋白序列分成一系列片段，对每个片段预测其可能的局部结构. 片段一般是定长的，比如 ROSETTA[54-56] 采用 9-mer 和 3-mer 片段，对每个片段预测出 200 个可行的候选结构.

(2) 拼接. 在获得每个片段的候选局部结构之后，下一步就是将局部结构拼接成一个完整的全长

结构，并使得最终产生的完整结构具有最小的能量. 拼接是一个典型的组合优化问题，常常采用 Monte Carlo、模拟退火等技术实现.

(3)构象筛选. 对于目标蛋白质来说，拼接过程常常会产生数千个能量各自不同的全长结构. 如何在这数千个候选中鉴别出和真实结构(Native structure)最接近的结构，是筛选步骤要解决的问题.

采用拼接策略，需要解答如下三个关键问题：(1)如何准确预测局部结构；(2)如何设计能量函数；(3)如何最小化能量函数.

5.2　典型代表

从头预测法的代表有 ROSETTA[54-56]、FALCON[57]以及 Quark[58]等，简要描述如下.

ROSETTA 是当前最为成功的拼接算法，其基本思想是：首先将整条目标序列分割成多个 9-mer 片段，片段间允许相互重叠；对每个片段，从 PSI-BLAST 计算出的相似序列对应的结构中，截取出相应位置的局部结构作为候选；然后使用 Monte Carlo 算法从每个片段的候选局部结构集中挑选出一个进行组合，使得组合成的全长结构能量最小(见图 15).

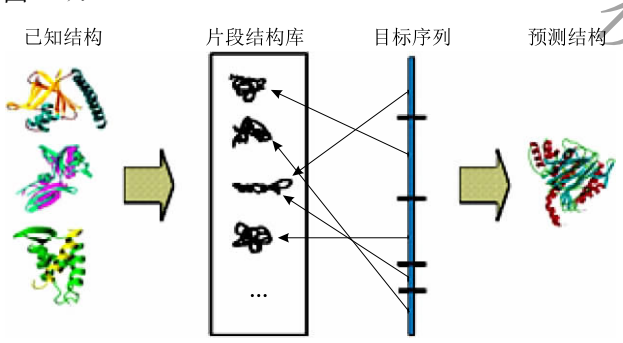


图 15　ROSETTA 方法示意图[54]

ROSETTA 采用的能量函数中考虑了远程相互作用、氢键、范德华力、匹配的 β 折叠和疏水作用等.

ROSETTA 适用于少于 120 个氨基酸的蛋白质. 对 LIVEBENCH7 测试集中的蛋白质来说，在其中大约 50%个蛋白质上，ROSETTA 能够做出准确预测，其均方根偏差(Root-Mean-Square Deviation，RMSD)在 7 Å 以内. ROSETTA 的一个缺点是运行时间较长，比如对于一个长度为 150 氨基酸的蛋白质，ROSETTA 在一个 80 核的 CPU 上运行 6 h 才能产生一个候选构象.

FALCON 使用 von Mises 分布来刻画单个残基的局部结构倾向性；使用位置特异隐马尔可夫模型来刻画相邻残基间的相关性. 对每个残基，FALCON 根据每个残基的 von Mises 分布采样二面角(ϕ,φ)，然后从二面角出发计算出整体结构，并使用能量函数来评价产生结构的优劣(见图 16).

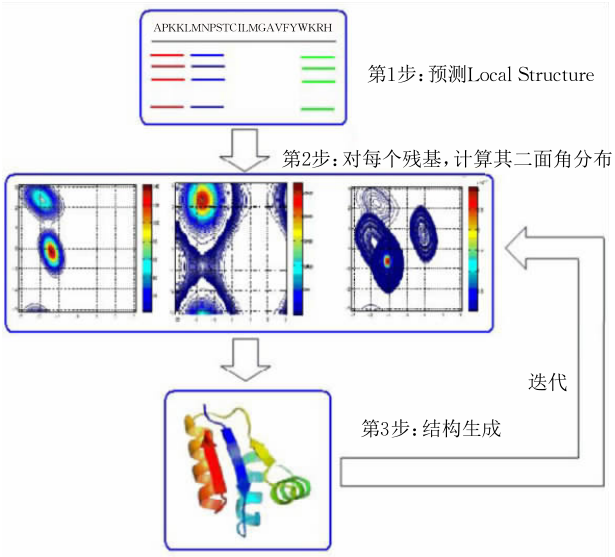


图 16　FALCON 方法示意图[57]

FALCON 的特色在于其采用的迭代策略，即利用所获得的结构中能量较小的结构来产生更加精确的分布，这样不断迭代直至收敛.

相对于 ROSETTA，FALCON 的优势在于：(1)搜索空间大大减小；(2)随着迭代的进行，生成结构的质量逐步提高. 图 17 描述了 FALCON 在蛋白质 1ctfA 上的结果，FALCON 产生的最优的结构与真实结构的 RMSD 仅为 0.557 Å.

QUARK 算法分为 3 步(见图 18)：第 1 步先从目标序列中预测各种特征，例如二级结构、溶液可及性以及残基扭角等，进而依据这些特征推断出各片段的候选结构；第 2 步是拼接过程，利用复制-交换蒙特卡罗技术获得能量较低的全长结构；第 3 步是结构聚类，将得到全长结构按照结构相似性聚类，并且进行原子位置的优化.

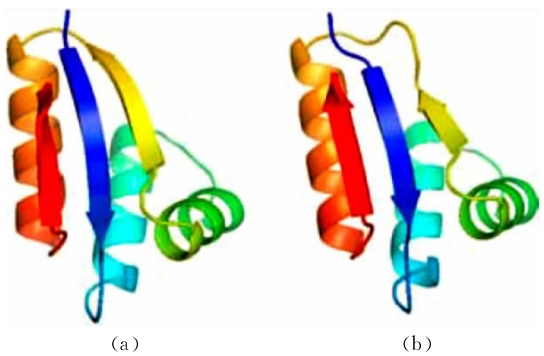


图 17　FALCON 预测 1ctfA 的结果((a)为 1ctfA 的真实结构；(b)为FALCON预测结果. 两者的RMSD仅为0.557 Å[57])

计算机学报

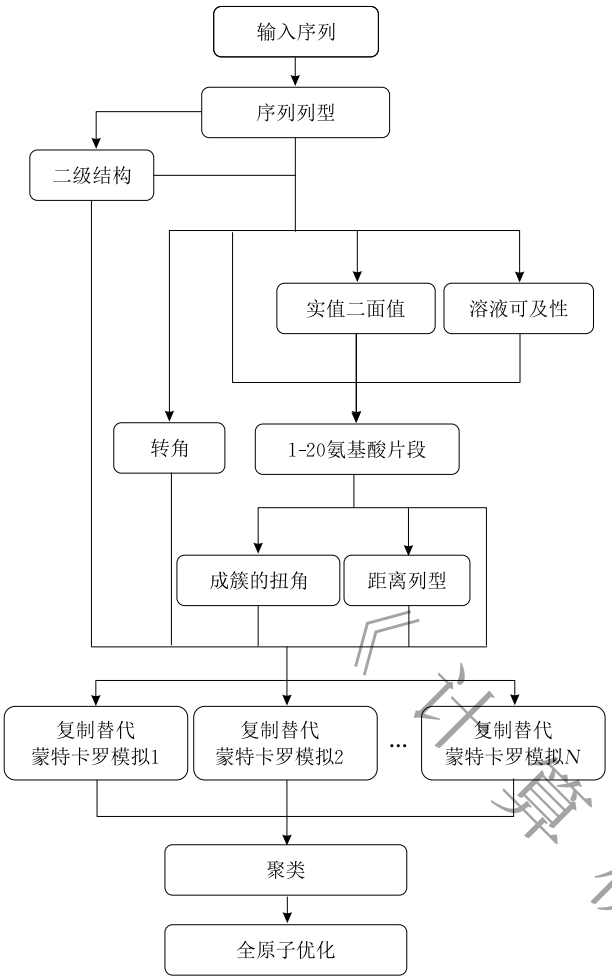


图 18 QUARK 方法流程图[58]

为提高能量函数的准确度以及搜索的效率，QUARK 算法引入了一些新的能量项. 粗略地说，QUARK 的能量项包含 3 类：即原子层面(atomic-level)的、残基层面(residue-level)的以及拓扑层面(topology-level)的能量项，其中原子层面能量项包含主链原子二体项、侧链中心二体项等，残基层面项则考虑氢键、溶液可及性、主链扭角等，拓扑层面项考虑回旋半径、片层-螺旋-片层、螺旋堆叠等信息. 在 CASP10 比赛中，QUARK 表现出优异的性能.

5.3 构象选择

在产生了数千个蛋白质构象后，从头预测法需要对这些构象进行筛选，选出最可能与天然构象相似的预测结构.

常用的策略是对这数千个候选结构进行聚类，最大类的中心往往是和真实结构最相似的结构. 另外能量函数也是其中重要的评估方法，即能量函数最低的构象往往和真实结构最相似.

ROSETTA[54-56]综合使用这两种策略，首先将产生的结构进行聚类，然后计算各个类中心的构象的能量函数，最终选出能量函数最低的构象.

为提高蛋白质构象聚类的效率，Li 等人[59]采取构象辅助组构造、调节 C_α 原子 RMSD 的上界和下界、过滤奇异值构象等策略，开发了软件包 Calibur；在上述工作的基础上，Li 等人[60]进一步采取快速评估类中心以及快速调整不同构象之间的旋转变换阵技术，开发了软件包 ONION，将聚类性能提高了 14 倍，能够在约 12 min 内对 10^5 个蛋白质构象进行聚类.

5.4 从头预测法小结

从头预测法由于没有全序列的模板结构辅助，通常采用片段拼接方法. 片段拼接需要在大量的搜索空间中寻找能量最低的结果，因此不适宜于预测长的、复杂拓扑结构蛋白质(比如 β 类的蛋白质结构). 因此，如何提高对于长的、复杂拓扑结构的蛋白质的预测准确率，是从头预测法下一步的研究重点.

6 蛋白质结构预测 CASP 比赛

CASP 比赛(Critical Assessment of protein Structure Prediction)是全球范围内蛋白质结构预测比赛，从 1994 年开始，每两年一届. 目前，CASP 比赛包括结构预测、接触位点预测、接触位点辅助结构预测、结构优化、结构质量评估 5 个部分，有助于客观评估蛋白质结构预测的质量.

评估结构预测的精度有多种方法，通常采用预测结构与真实结构之间差异的某种度量，比如 GDT_TS、GDT_HA、TM-score[61]等打分，其中 GDT_TS 描述了两个结构在给定最优旋转变换阵的情形下对应的 C_α 原子之间的距离分别在 1 Å、2 Å、4 Å 以及 8 Å 以内的比例的平均值；GDT_HA 是 GDT_TS 的高精度版本，即将 GDT_TS 中的 8 Å 替换为 0.5 Å；TM-score 是在最优旋转变换阵的情形下

$$\frac{1}{L_{\text{full}}}\sum_{i=1}^{L_{\text{full}}}\frac{1}{1+d_i^2/d_0^2}$$

的最大值. 其中 d_i 为预测结构与真实结构中第 i 个残基的距离. 表 2 描述了部分预测服务器在 CASP11 中的结果，其中 Z 值表示用各种打分做加权之后的归一化打分.

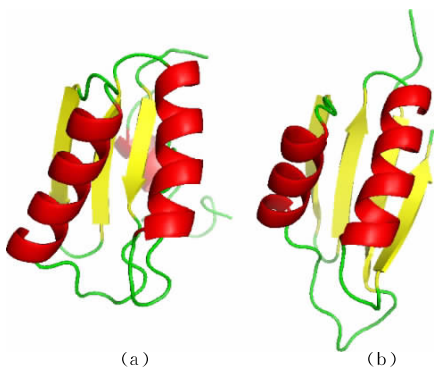
在 2014 年的第 11 届 CASP 比赛中，发布了 78 个目标蛋白质，共包含 126 个结构域，并以结构域为单位进行最终的预测质量评估[62]. 这些结构域又被划分为基于模板建模(Template-Based Modeling，TBM)以及无模板建模(Free Modeling，FM)两类.

《计算机学报》

表 2　部分预测服务器在 CASP11 中的结果(其中 Z 值表示标准化以后的得分之和)

基于模板类(TBM)目标序列		无模板类(FM)目标序列	
预测服务器	Z 值	预测服务器	Z 值
Zhang-Server	71.7964	Zhang-Server	43.8107
nns	69.2103	QUARK	41.4452
BAKER-ROSETTASERVER	67.4999	RBO_Aleph	28.4216
QUARK	64.7060	BAKER-ROSETTASERVER	27.0235
myprotein-me	58.6524	nns	22.6296
RaptorX	42.2683	Seok-server	20.7412
MULTICOM-CONSTRUCT	41.4001	MULTICOM-CLUSTER	18.5134
MULTICOM-CLUSTER	40.9565	myprotein-me	17.2838
FALCON_EnvFold	22.6152	MULTICOM-CONSTRUCT	15.3968
FALCON_MANUAL_X	21.8916	MULTICOM-REFINE	14.6670
FALCON_MANUAL	20.8235	RaptorX	12.4504
FALCON_TOPO	18.6477	FUSION	12.1728
MULTICOM-NOVEL	18.5263	MULTICOM-NOVEL	7.7466
FFAS-3D	13.4991	BhageerathH	6.8712
Seok-server	9.3163	FALCON_MANUAL	1.3082
MULTICOM-REFINE	8.5480	FALCON_TOPO	0.9833
HHPredA	6.7472	Pcons-net	0.3218
raghavagps-tsppred	3.4547	FALCON_EnvFold	−0.6743
HHPredX	2.5288	FALCON_MANUAL_X	−0.8263
IntFOLD3	−1.7179	FFAS-3D	−2.7400

在 TBM 类的评估中，基于 I-TASSER 的 Zhang-Server 预测性能最好，笔者开发的预测算法 FALCON_EnvFold 排名第 9. 图 19 描述了 T0765 目标序列的预测结果，其中图 19(a)为 Zhang-server 组与真实结构的 TM-score 为 0.761，准确度较高. 此外，MULTICOM-CLUSTER 软件[64]充分考虑了多模板之间的关联，通过组合互补与交叠的模板与联配提高了联配准确率和模板识别质量，从而说明了多模板技术的优势.

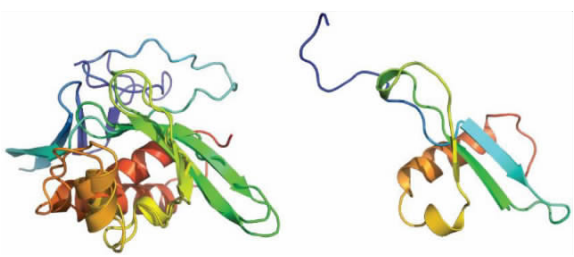


(a)　　(b)

图 19　结构域 T0765 的预测结果((a)为 Zhang-server 组的预测结果；(b)为 T0765 的真实结构)

事实上，目前的蛋白质结构库已经包含了“几乎”所有的球蛋白的结构折叠型. 因此，只要能够找到结构相近的模板并且得到高质量的联配，就能显著提高 TBM 类蛋白的预测准确度，这也是基于模板方法将来的主要努力方向.

在 FM 类的评估中，QUARK 软件表现出优异的性能，这种优异性能部分来源于准确刻画蛋白质的整体特性的能量函数. ROSETTA 也有较好的性能表现，图 20 描述了 FM 类结构域 T0793-D1 的预测结果，在与最好模板联配的 TM-score 仅为 0.48 的条件下，ROSETTA 预测的结构与目标序列的结构的 TM-score 为 0.78. 笔者的团队基于 ROSETTA 能量函数，设计了一种提高优化效率的能量函数加权方案. 结果显示，将这一方案加入到 FALCON 软件中可以显著加速能量优化速度，并且提高预测准确率[65]. 这也是笔者的团队在 CASP11 中的软件 FALCON_TOPO 以及 FALCON_EnvFold 对于无模板序列的实现方法.



(a)　　(b)

图 20　(a)表示结构域 T0793-D1 与预测结构 ROSETTA 的结构联配结果；(b)表示该结构域中最好的模板蛋白 1g6z

另外，在 CASP11 其他比赛项目上得到如下结论.

(1) 对于接触位点预测来说，高精度的接触位点预测的确有助于构建蛋白质三维结构，而基于共进化的接触位点预测方法可以直接用于构建序列-

结构联配，为归范法提供远程相互作用信息.

(2)对于结构优化来说，全局性方法(优化结构的能量函数)和局部性方法(调节结构的局部倾向性)表现出一定的互补性，因此两种方法的结合将有效提高结构优化性能.

(3)对于质量评估来说，通常有两种策略：一种是每个预测构象单独评估；另一种策略是同时考虑多个预测构象之间的关系. CASP 比赛的结果表明第二种策略具有更好的性能.

7 蛋白质数据库及结构预测工具

常见的蛋白质数据库包含两个大类：序列数据库以及结构数据库. 蛋白质序列数据库主要由基因预测以及蛋白质质谱技术得到，而蛋白质结构数据主要由 X-晶体衍射和核磁共振等方法得到.

最常用的蛋白质序列库是 NR(Non-Redundant)库[66]，由美国生物技术信息中心(National Center for Biotechnology Information，NCBI)创建. NR 主要应用于蛋白质同源序列的搜索，例如 BLAST、PSI-BLAST 等工具. 另外 3 个主要的蛋白质序列库见表 3.

最常用的蛋白质结构库是 PDB(Protein Database Bank)数据库[67]. PDB 数据库所提供的信息是最重要的蛋白质结构资料，主要包含以下信息：重原子的空间坐标、引用文献、形成 α 螺旋和 β 片层的氨基酸序列、二硫键连结模式、与蛋白质结构的溶液环境、参与生物化学功能的残基等. 截至 2015 年 9 月 22 日，PDB 收录了 104196 条蛋白的结构，并以每周约 200 条蛋白的速度递增. 表 4 列举了另外 3 个主要的蛋白质结构数据库的主要信息.

表 5 列举了蛋白质结构预测的在线服务，这些在线服务能够高通量地预测蛋白质的结构以及残基间相互作用.

表 3 常见的蛋白质序列数据库

名称	网址	描述
UniProt	http://www.uniprot.org/	来自文献的蛋白质的生物功能的信息
PIR	http://pir.georgetown.edu/	乔治城大学医学中心建立的蛋白质信息源(Protein Information Resource)
Pfam	http://pfam.xfam.org/	包含结构标注的用隐马尔可夫模型建立的多序列联配

表 4 常见的蛋白质结构数据库

名称	网址	描述
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop/	蛋白质层次分类数据库，将蛋白质分为类、折叠型、超家族、家族 4 层
CATH	http://www.cathdb.info/	蛋白质层次分类数据库，在 SCOP 的折叠型与超家族之间加入拓扑层次
Protein	http://www.ncbi.nlm.nih.gov/protein/	美国卫生研究院蛋白质数据库，包含基因组数据库标注的编码区域的翻译

表 5 常见的蛋白质预测在线工具

在线服务工具名称	对应的 CASP 服务器名称	地址
I-TASSER	Zhang-Server	http://zhanglab.ccmb.med.umich.edu/I-TASSER/
RaptorX	RaptorX	http://raptorx.uchicago.edu/StructurePrediction/predict/
ROBETTA	BAKER-ROSETTASERVER	http://robetta.bakerlab.org/
HHpred	HHPredX	http://toolkit.lmb.uni-muenchen.de/hhpred/
MULTICOM	MULTICOM-CLUSTER	http://sysbio.rnet.missouri.edu/multicom_cluster/
FALCON@home	FALCON_TOPO	http://protein.ict.ac.cn/TreeThreader/

8 总结与发展趋势分析

经过多年的发展，蛋白质结构预测技术虽然取得了较大的进展，但是依然存在预测精度和速度等方面的不足. 蛋白质结构预测的性能提高需要在如下几个方面开展研究.

(1)蛋白质序列中结构信号的识别

组成蛋白质的氨基酸序列极端偏离随机序列. 对于蛋白质序列与结构的关联来说，整条序列与结构的关联并不显著，但其中某些特定的局部序列片段具有非常强的结构信号，其结构倾向性非常明显. 这些区域的序列在序列列型上表现出高保守性，并且与之强关联的局部结构构象与其天然态的结构构象相一致. 其余的序列片段，虽然有自身的结构倾向性，但是信号相对模糊. 不同层次的强信号在引导形成稳定结构的过程中发挥不同的作用. 局部片段的强结构倾向性序列可以看作是底层的强信号，例如 α 螺旋中的某些单圈、β 发卡型结构中的转角等等. 在此之上，高层的强信号，出现在相连或者远程二级

结构单元之间，主导蛋白质形成了稳定的“拓扑”结构. 该稳定的结构需要符合强信号结构空间的约束. 因此，蛋白质在强信号的引导下，其折叠过程极端偏离各态历经.

强信号“绑架”了蛋白质形成某些构形，而偏离该构形的结构空间并非在蛋白质折叠搜索的范围内. 不同层次的强信号，可逐级缩减构象空间的搜索范围，最终形成天然构象.

(2) 保守的蛋白质核心区域拓扑框架识别

蛋白质序列不同层次的强信号引导蛋白质形成稳定的核心区域拓扑框架，这些稳定的拓扑框架的正确识别能够有效地提高蛋白质折叠型识别的性能.

在蛋白质结构预测的归范法中，识别出强信号所在的拓扑框架能够显著促进序列-结构联配性能的提升. 最近，笔者的团队[68]提出了拓扑框架(Common structural framework)的概念，描述同源蛋白质的拓扑结构，将折叠型识别问题转化为目标序列与模板拓扑结构之间的联配问题，在 CASP11 中取得了较好的预测性能.

蛋白质中的强信号具有较高的序列保守性，以保证蛋白质形成稳定的拓扑结构. 这些强信号区域在稳定蛋白质拓扑结构后，弱信号区域通过局部调节，形成最终的三级结构. 序列不同区域的强弱信号保守性的差异也体现了在进化中不同区域突变被接收程度的不同.

传统的进化模型，例如 Kleinberg 等人[69]使用的马尔可夫链蒙特卡罗(Markov Chain Monte-Carlo)技术估计了蛋白质结构的进化空间，Marks 等人[53]将序列-结构关系放在进化与功能这个空间中建立概率模型，都没有体现出不同区域中进化速率的差异.

笔者的团队[68]在识别蛋白质拓扑结构的基础上，针对强信号保守区域以及可调节非保守区域不同的进化速率，建立不同的概率模型分别刻画. 强信号区域稳定蛋白质结构和功能，往往进化速率较慢；而弱信号区域由于只要能实现特定的局部构象，往往会表现出较大的序列差异性，体现出进化速率较快的特点. 针对不同进化速率的位置分别建立序列进化模型，对提高联配质量和折叠型识别率会有重要的帮助.

(3) 能量函数设计以及高效的结构构象搜索算法

蛋白质序列中的强弱信号搭配使得在蛋白质形成过程中，不同的相互作用在不同阶段所处的角色不同. 简要地说，底层强信号一般出现在序列中局部片段，近程相互作用在这里起主导作用；而高层强信号，包括跨二级结构单元的超二级结构模体(Super-secondary structure motifs)，也包括出现在不连续的局部片段之间，远程相互作用在这个阶段占据主导. 而弱信号，在序列列型上表现出相对不保守的特点，对于实现最终结构并不起主导作用，其稳定的结构靠局部调节形成. 在从头预测法的能量函数设计中，应当对于不同层次的能量函数在搜索过程中分别考虑. ROSETTA[54-56]和 FALCON[57]在片段拼接的搜索过程中，对于不同的阶段采用不同的能量函数，在 CASP 中取得了好的效果. 目前的能量函数的设计，无论物理化学势函数，以及统计势函数，仅考虑了在特定粗拓扑框架下的结构构象调节. 对于强信号形成拓扑结构的能量函数的设计，是下一步工作的重点.

(4) 高精度的残基接触预测

CASP 比赛的经验显示，有接触位点信息的蛋白质结构预测比无此信息的预测具有更高的准确度. Cheng 等人[70]利用接触位点信息是否与产生的结构模型匹配来对结构模型进行打分，并据此识别折叠型. 从强信号的观点看，接触位点有利于建立从底层强信号到高层强信号的桥梁，通过底层强信号之间的接触位点信息刻画蛋白质的拓扑结构. 高质量的拓扑结构预测对于蛋白质折叠型的识别以及归范法的准确率具有极高的帮助作用.

(5) 高通量结构预测

高通量的蛋白质结构预测需求越来越显著. 为了达到高通量的要求，需要大量的计算资源. Baker 团队的 ROSETTA@home 利用 BOINC 分布式计算平台高通量地预测蛋白质结构、蛋白质-蛋白质相互作用以及设计蛋白质[71]. 笔者的团队开发的 FALCON@home 将蛋白质结构预测归范法软件 TreeThreader[72]应用到中国科学院志愿计算平台 CAS@home 上，有效地提高了远同源蛋白质的折叠型识别率.

(6) 肽段模拟对结构预测的补充

肽段的分子动力学模拟为蛋白质结构预测提供了折叠中间态的有力补充. 目前的蛋白质结构预测考虑的序列-结构空间是对于整条序列而言的，对于局部片段与局部片段之间接触模体的粗拓扑识别，尤其是粗拓扑的能量函数设计还很欠缺. 利用分布式计算的软件，无论是对于局部片段的底层强信号，

《计算机学报》

还是超二级结构或是局部片段之间的高层强信号，可以有效地对肽段进行模拟，从而得到可靠的粗拓扑结构. 由于这些强信号区域所在的序列具有高保守性，在蛋白质结构预测中大幅度减小搜索空间.

蛋白质结构预测是使用计算技术从一级序列预测三级结构的典型代表. 计算技术的使用，是协助揭示蛋白质结构等复杂生物学现象内在规律的重要手段. 在蛋白质结构的研究过程中人们积累了足够丰富的观察和知识，开发了高效的算法.

蛋白质结构预测是蛋白质组计划中的重要组成部分. 首先，通过蛋白质结构预测的研究，可以有效理解蛋白质折叠的机理. 例如，对于蛋白质序列列型的分析可以判别哪些氨基酸是蛋白质折叠的启动位点[15]. 其次，结构预测可以帮助预测蛋白质的功能，并指导相对应的生物学实验. 另外，蛋白质结构预测的研究也是蛋白质设计等问题的基础，对于生物医学制药等问题提供了有力的理论支撑.

参 考 文 献

[1] Branden C I. Introduction to Protein Structure. United States: Garland Science, 1999

[2] Khan M Q, Sweeting B, Mulligan V K, et al. Prion disease susceptibility is affected by β-structure folding propensity and local side-chain interactions in PrP. Proceedings of the National Academy of Sciences, 2010, 107(46): 19808-19813

[3] Hattne J, Reyes F E, Nannenga B L, et al. MicroED data collection and processing. Acta Crystallographica Section A: Foundations and Advances, 2015, 71(4): 353-360

[4] Eibauer M, Pellanda M, Turgay Y, et al. Structure and gating of the nuclear pore complex. Nature Communications, 2015, 6(6): 7532

[5] Gold V A M, Ieva R, Walter A, et al. Visualizing active membrane protein complexes by electron cryotomography. Nature Communications, 2014, 5(5): 4129

[6] Anfinsen C B, Haber E, Sela M, et al. Kinetics of formation of native ribonuclease during oxidation of reduced polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America, 1961, 47(9): 1309-1314

[7] Ho B K, Dill K A. Folding very short peptides using molecular dynamics. PLoS Computational Biology, 2006, 2(4): e27

[8] Pauling L, Corey R B. Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets. Proceedings of the National Academy of Sciences of the United States of America, 1951, 37(11): 729

[9] Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 1983, 22(12): 2577-2637

[10] Heringa J. Computational methods for protein secondary structure prediction using multiple sequence alignments. Current Protein and Peptide Science, 2000, 1(3): 273-301

[11] Jones D T. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology, 1999, 292(2): 195-202

[12] Wang Z, Zhao F, Peng J, et al. Protein 8-class secondary structure prediction using conditional neural fields. Proteomics, 2011, 11(19): 3786-3792

[13] Faraggi E, Zhang T, Yang Y, et al. SPINE X: Improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. Journal of Computational Chemistry, 2012, 33(3): 259-267

[14] 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. Scientific Reports, 2015, 5(5): 11476

[15] Wei-Mou Z. Proteins: From sequence to structure. Chinese Physics B, 2014, 23(7): 8705

[16] Zheng W M. The Use of a Conformational Alphabet for Fast Alignment of Protein Structures. Berlin Heidelberg, Germany: Springer Press, 2008

[17] Henikoff S, Henikoff J G. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences, 1992, 89(22): 10915-10919

[18] Wang S. Hydropathy conformational letter and its substitution matrix HP-CLESUM: An application to protein structural alignment. Ithaca, NY 14850, United States: Connell University, arXiv preprint:1001.2879, 2010

[19] Murzin A G, Brenner S E, Hubbard T, et al. SCOP: A structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 1995, 247(4): 536-540

[20] Zeng Hui. From Protein Sequence to Structure. Institute of Theoretical Physics, Beijing, 2012(in Chinese)
(曾辉. 从蛋白质序列到结构[博士学位论文]. 中国科学院理论物理研究所，北京，2012)

[21] Altschul S F, Gish W, Miller W, et al. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403-410

[22] Altschul S F, Madden T L, Schäffer A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Research, 1997, 25(17): 3389-3402

[23] Altschul S F, Koonin E V. Iterated profile searches with PSI-BLAST—A tool for discovery in protein databases. Trends in Biochemical Sciences, 1998, 23(11): 444-447

[24] Rychlewski L, Li W, Jaroszewski L, et al. Comparison of sequence profiles: Strategies for structural predictions using sequence information. Protein Science, 2000, 9(2): 232-241

[25] Jaroszewski L, Rychlewski L, Li Z, et al. FFAS03: A server for profile—profile sequence alignments. Nucleic Acids Research, 2005, 33(Suppl. 2): W284-W288

《计算机学报》

[26] Ginalski K, Pas J, Wyrwicz L S, et al. ORFeus: Detection of distant homology using sequence profiles and predicted secondary structure. Nucleic Acids Research, 2003, 31(13): 3804-3807

[27] Karplus K, Hu B. Evaluation of protein multiple alignments by SAM-T99 using the BAliBASE multiple alignment test set. Bioinformatics, 2001, 17(8): 713-720

[28] Söding J. Protein homology detection by HMM-HMM comparison. Bioinformatics, 2005, 21(7): 951-960

[29] Bujnicki J M, Fischer D. Meta Approaches to Protein Structure Prediction. Practical Bioinformatics. Berlin Heidelberg, Germany: Springer Press, 2004

[30] Ginalski K, Grishin N V, Godzik A, et al. Practical lessons from protein structure prediction. Nucleic Acids Research, 2005, 33(6): 1874-1891

[31] Vitkup D, Melamud E, Moult J, et al. Completeness in structural genomics. Nature Structural & Molecular Biology, 2001, 8(6): 559-566

[32] Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. Proteins: Structure, Function, and Bioinformatics, 2000, 40(3): 343-354

[33] Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins: Structure, Function, and Bioinformatics, 2004, 55(4): 1005-1013

[34] Zhou H, Zhou Y. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins: Structure, Function, and Bioinformatics, 2005, 58(2): 321-328

[35] Xu J, Li M, Kim D, et al. RAPTOR: Optimal protein threading by linear programming. Journal of Bioinformatics and Computational Biology, 2003, 1(1): 95-117

[36] Peng J, Xu J. RaptorX: Exploiting structure information for protein alignment by statistical inference. Proteins: Structure, Function, and Bioinformatics, 2011, 79(S10): 161-171

[37] Peng J, Xu J. A multiple-template approach to protein threading. Proteins: Structure, Function, and Bioinformatics, 2011, 79(6): 1930-1939

[38] Ma J, Wang S, Wang Z, Xu J. MRFalign: Protein homology detection through alignment of Markov random fields. PLoS Computational Biology, 2014, 10(3): e1003500

[39] Jo T, Cheng J. Improving protein fold recognition by random forest. BMC Bioinformatics, 2014, 15(Suppl. 11): S14

[40] Lin C, Zou Y, Qin J, et al. Hierarchical classification of protein folds using a novel ensemble classifier. PLoS ONE, 2013, 8(2): e56499

[41] Wei L, Liao M, Gao X, et al. An improved protein structural prediction method by incorporating both sequence and structure information. IEEE Transactions on Nanobioscience, 2014, 6(14): 649

[42] Wei L, Liao M, Gao X, et al. Enhanced protein fold prediction method through a novel feature extraction technique. IEEE Transactions on NanoBioscience, 2015, 14(6): 649-659

[43] Zhao X M, Cheung Y M, Huang D S. A novel approach to extracting features from motif content and protein composition for protein sequence classification. Neural Networks, 2005, 18(8): 1019-1028

[44] Huang D S, Zhao X M, Huang G B, et al. Classifying protein sequences using hydropathy blocks. Pattern Recognition, 2006, 39(12): 2293-2300

[45] Zhao X M, Li X, Chen L, et al. Protein classification with imbalanced data. Proteins: Structure, Function, and Bioinformatics, 2008, 70(4): 1125-1132

[46] Liu B, Zhang D, Xu R, et al. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. Bioinformatics, 2014, 30(4): 472-479

[47] Liu B, Chen J, Wang X. Application of learning to rank to protein remote homology detection. Bioinformatics, 2015, 31(21): 3492-3498

[48] Wang C, Zhang H, Zheng W M, et al. FALCON@home: A high-throughput protein structure prediction server based on remote homologue recognition. Bioinformatics, 2016, 32(3): 462-464

[49] Webb B, Sali A. Comparative protein structure modeling using MODELLER. Current Protocols in Bioinformatics, 2014, 5(6): 1-30

[50] Chen Y, Shang Y, Xu D. Multi-dimensional scaling and MODELLER-based evolutionary algorithms for protein model refinement//Proceedings of the Evolutionary Computation (CEC). Beijing, China, 2014: 1038-1045

[51] Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(20): 7594-7599

[52] Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. Nature Protocols, 2010, 5(4): 725-738

[53] Marks D S, Colwell L J, Sheridan R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE, 2011, 6(12): e28766

[54] Baker D, Sali A. Protein structure prediction and structural genomics. Science, 2001, 294(5540): 93-96

[55] Simons K T, Strauss C, Baker D. Prospects for ab initio protein structural genomics. Journal of Molecular Biology, 2001, 306(5): 1191-1199

[56] Bradley P, Malmström L, Qian B, et al. Free modeling with Rosetta in CASP6. Proteins: Structure, Function, and Bioinformatics, 2005, 61(S7): 128-134

[57] Li S C, Bu D, Xu J, et al. Fragment-HMM: A new approach to protein structure prediction. Protein Science, 2008, 17(11): 1925-1934

《计算机学报》

[58] Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins: Structure, Function, and Bioinformatics, 2012, 80(7): 1715-1735

[59] Li S C, Ng Y K. Calibur: A tool for clustering large numbers of protein decoys. BMC Bioinformatics, 2010, 11(1): 25

[60] Li S C, Bu D, Li M. Clustering 100,000 protein structure decoys in minutes. Computational Biology and Bioinformatics, 2012, 9(3): 765-773

[61] Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins: Structure, Function, and Bioinformatics, 2004, 57(4): 702-710

[62] Kryshtafovych A, Barbato A, Monastyrskyy B, et al. Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins: Structure, Function, and Bioinformatics, 2015, 9(9): 24919

[63] Yang J, Yan R, Roy A, et al. The I-TASSER Suite: Protein structure and function prediction. Nature Methods, 2015, 12(1): 7-8

[64] Wang Z, Eickholt J, Cheng J. MULTICOM: A multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics, 2010, 26(7): 882-888

[65] Wang C, Wei Y, Liu J, et al. Optimizing weights of protein energy function to improve ab initio protein structure prediction. Ithaca, NY 14850, United States: Connell University, arXiv preprint: 1312.6960, 2013

[66] Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Research, 2007, 35(Suppl. 1): D61-D65

[67] Laskowski R A, Hutchinson E G, Michie A D, et al. PDBsum: A Web-based database of summaries and analyses of all PDB structures. Trends Biochemistry Science, 1997, 22(12): 488-490

[68] Zhu J, Zhang H, Wang C, et al. TOPO: Improving remote homologue recognition via identifying common protein structure framework. Ithaca, NY 14850, United States: Connell University, arXiv preprint: 1507.03197, 2015

[69] Meyerguz L, Kempe D, Kleinberg J, et al. The evolutionary capacity of protein structures//Proceedings of the 8th Annual International Conference on Resaerch in Computational Molecular Biology. San Diego, USA, 2004: 290-297

[70] Adhikari B, Deng X, Li J, Cheng J. A contact-assisted approach to protein structure prediction and its assessment in CASP10//Proceedings of the Workshops at the 27th AAAI Conference on Artificial Intelligence. Bellevue, USA, 2013: 6455

[71] Das R, Qian B, Raman S, et al. Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins: Structure, Function, and Bioinformatics, 2007, 69(S8): 118-128

[72] Wu W, Chen G, Kan W, et al. Harness public computing resources for protein structure prediction computing//Proceedings of the International Symposium on Grids and Clouds (ISGC). Taipei, China, 2013: 37

计算机学报

WANG Chao, born in 1987, Ph. D. candidate. His research interests include protein structure prediction, algorithm design.

ZHU Jian-Wei, born in 1991, Ph. D. candidate. His research interests include protein structure prediction, machine learning.

ZHANG Hai-Cang, born in 1987, Ph. D. candidate. His research interests include protein structure prediction, machine learning.

GONG Hai-E, born in 1991, M. S. candidate. Her research interests include protein structure prediction, machine learning.

ZHENG Wei-Mou, born in 1946, professor, Ph. D. supervisor. His research interests include surface physics, stochastic processes, nonliear dynamics and statistical physics.

BU Dong-Bo, born in 1973, professor, Ph. D. supervisor. His research interests include algorithm design and analysis, and bioinformatics.

Background

Protein structure prediction, inferring protein tertiary structure based on the information of protein sequence, remains a great challenge in the field of computational biology. Protein functions are mainly determined by protein structures, making protein structure prediction an important task. On the other hand, protein structure prediction is also feasible since a protein spontaneously folds into its specific tertiary structure in the natural environment.

Protein structure prediction is a representative problem that a sequence of entities form a complex structure under interactions among the entities. The existing prediction approaches can be categorized into three families: (1) the

homology modeling approaches to align the query sequence against the sequences of proteins with known structure; (2) the threading approaches to thread the query sequence into a known structure, and evaluate the fitness of the sequence and structures; (3) the ab initio approaches that attempt to identify the structural conformation with the minimal free energy. The existing approaches applied a great variety of algorithmic techniques, including dynamic programming, linear programming, Markov random field, conditional random field, etc. More importantly, the applications also promote the development of new algorithms, called profile hidden Markov model (HMM), and the comparison of two profile HMMs, etc.

In the study, we summarized the existing approaches to protein structure prediction. We also presented the evaluation of the existing approaches by using the performance in the Critical Assessments of Protein Structure Prediction (CASP) competitions. The current status can be summarized as: (1) A protein structure can be accurately predicted if it has a sequence identity of over 30% with a template protein; thus, the homology modeling problem has been closed. (2) For threading approaches, the fold recognition accuracy is around 60%, and more researches are needed for the remote homologue proteins. (3) How to assemble the structures with complex topology remains one of the challenges for the ab initio approaches.

In the field of protein structure prediction, we have developed a fragment HMM model called FALCON, which was ranked 3rd in the CASP-8 competition. We also developed a high-throughput prediction server called FALCON@home. Currently FALCON@home has over 20 000 volunteer CPUs from all around the world, and can process over 2130 proteins per day.

The study was funded by the National Basic Research Program of China (973 Program) under Grant No. 2012CB316502, the National Natural Science Foundation of China under Grant Nos. 31270834, 61272318, 11175224 and 11121403, and the Open Project Program of State Key Laboratory of Theoretical Physics, Institute of Theoretical Physics, Chinese Academy of Sciences, China (No. Y4KF171CJ1).

《计算机学报》