



蛋白质结构预测进展*

曹 卫 潘宪明**

(清华大学生命科学院, 蛋白质科学教育部重点实验室, 北京 100084)

摘要 蛋白质的序列决定结构, 结构决定功能。新一代准确的蛋白质结构预测工具为结构生物学、结构生物信息学、药物研发和生命科学等许多领域带来了全新的机遇与挑战, 单链蛋白质结构预测的准确率达到与试验方法相媲美的水平。本综述概述了蛋白质结构预测领域的理论基础、发展历程与最新进展, 讨论了大量预测的蛋白质结构和基于人工智能的方法如何影响实验结构生物学, 最后, 分析了当前蛋白质结构预测领域仍未解决的问题以及未来的研究方向。

关键词 蛋白质结构预测, 深度学习, 能量函数

中图分类号 Q518

DOI: 10.16476/j.pibb.2023.0152

1 蛋白质结构预测的理论基础与意义

蛋白质是生命活动的主要承担者和体现者, 一切生命活动都离不开蛋白质, 了解蛋白质功能的基本方法之一是研究其三维结构。高通量测序技术的发展使蛋白质序列信息呈指数增长, 相对地, 蛋白质结构数据的增长速度远远低于其序列数据的增长速度^[1]。蛋白质结构主要通过X射线晶体学(X-ray crystallography)、核磁共振(nuclear magnetic resonance, NMR)和冷冻电镜(cryo-electron microscopy, cryo-EM)等方法解析, 虽然这些方法可以产生高分辨率和高质量的蛋白质结构, 但是耗时、昂贵且不适用于所有蛋白质。随着计算领域技术的进步, 人工智能在生物学中的应用日益广泛, 因此, 用计算的方法从蛋白质序列出发预测结构是非常必要的。

蛋白质结构非常复杂, 结构化学家从概念上将蛋白质结构分为四个“层次”。一级结构是由氨基酸脱水缩合组成的多肽链; 二级结构描述了局部区域的一般三维(3D)形式, 这些区域与蛋白质的其余部分独立地组织成重复出现的结构片段, 多肽链最主要的局部构象是 α 螺旋(α helices)和 β 片层(β sheets), 不规则形状也是蛋白质结构和功能的重要组成部分, 通常称其为环区(loop); 三级结构是一条多肽链的3D结构, 即每个原子的3D坐

标; 四级结构指其亚基之间是如何定向和排列的, 也就是说四级结构仅适用于多亚基蛋白质^[2]。

2 蛋白质结构预测的进展

2.1 从蛋白质序列出发预测结构

20世纪70年代初期 Anfinsen^[3]进行的经典实验表明, 蛋白质正确折叠所需的所有信息均包含在其氨基酸序列中。近50年来, 研究人员在解决蛋白质折叠问题上做了诸多尝试, 主要可分为两大类, 分别是基于模板(template-based)和无模板(template-free)的预测方法。

基于模板的结构预测方法, 即同源建模方法, 首先选择合适的结构模板, 然后将靶标序列与模板结构比对, 最后使用分子建模等技术补充靶标-模板比对中存在的突变、插入和缺失部分的结构。由于蛋白质折叠类型的总量有限, 且折叠覆盖率随着蛋白质结构数据量的增长而增加^[4], 基于模板的结构预测方法的适用性在不断增长, 根据可用的蛋白质结构, 基于模板的结构预测方法可以预测大约2/3蛋白质家族的结构^[5]。研究表明, 基于模板的方法(同源建模和折叠识别)所预测的模型的准确

* 国家自然科学基金(39625008)资助项目。

** 通讯联系人。

Tel: 010-62792827, E-mail: pan-xm@mail.tsinghua.edu.cn

收稿日期: 2023-03-20, 接受日期: 2023-04-20

率和生物学实用性均要远高于使用无模板的方法(从头预测)^[6]。

无模板的结构预测方法适用于预测在蛋白质结构数据库(protein data bank, PDB)中找不到同源蛋白质的结构, 由于缺少结构模板, 该类方法需要用于生成候选模型的构象采样方法和用于选择类似天然构象的物理能量函数。构象采样的一个显著发展是使用基于片段的组装方法, 它在某些方面连接了基于模板的方法和无模板从头计算的方法, 其中模型是基于已知结构的蛋白质中短连续主链片段(通常长度为3~15个残基)构建的, 并使用蒙特卡罗模拟组装成全长模型。

蛋白质结构预测的下一个重大进展是有效利用共进化信息, 该方法首先用于识别接触的残基对, 并进一步扩展以推导残基距离和二面角分布, 所有这些都用作为无模板从头计算的约束^[7]。基于神经网络的学习方法进一步将多序列比对的使用扩展到端到端的蛋白质结构预测, 2020年CASP14(critical assessment of structure prediction)中, AlphaFold2在有模板预测和无模板预测蛋白质三级结构都有较为出色的表现, 实现了原子水平上的准确, 是一种依赖于PDB中的多序列比对信息和实验结构的深度学习算法。相对的, 如果蛋白质序列在已知数据库中同源性低于30%, 进而产生低质量的多序列比对信息, 最终预测的蛋白质3D结构准确率也会降低, 这种现象在2022年的CASP15中依然存在。

2.2 用于蛋白质结构评估的能量函数

蛋白质结构预测的最终目标之一是在不依赖实验数据的情况下识别高分辨率预测的蛋白质结构, 对这种高精度的预测需要能够区分具有低均方根偏差的非常相似的结构。蛋白质折叠理论的基本假设是蛋白质结构在天然状态下一般具有最低的吉布斯自由能^[3]。因此, 准确的能量函数是解决蛋白质折叠和蛋白质结构预测问题的关键。

科学家们已经做了很多工作来捕捉蛋白质内部的能量相互作用并发展这些力场。这些势能的两个最普遍的类别是基于物理的力场和基于知识的力场。基于物理的力场使用物理的基本定律来结合其原子之间在原子水平上发生的所有类型的相互作用(例如范德华力、氢键、静电相互作用等)。最早由Shneior Lifson团队在1960年代开始, 他们扩展模拟有机小分子的分子力学方法到大分子系统^[8], 开发了一致力场(consistent force field, CFF)能

量函数, 这为现今蛋白质建模中使用的一些最重要的全原子势能方法发展奠定了基础, 包括CHARMM^[9]、Amber^[10]和ECEPP^[11]。然而, 由于未考虑溶剂化效应, 事实证明它们不足以对溶液中稳定的致密蛋白质折叠进行热力学描述, 并且无法区分天然蛋白质和错误折叠的模型^[12]。下一步是在经典势中添加隐式溶剂化项, 由此产生的力场可以识别肽或蛋白质的天然状态, 但是准确性有限^[13]。基于知识的力场需要更少的计算能力, 使用统计方法计算结构特征的频率并将这些频率转换为自由能贡献^[14], 这些势能采用蛋白质的简化粗粒度模型, 同时使用来自PDB实验确定的蛋白质结构信息, 例如TASSER/I-TASSER^[15]、ROSETTA^[16]、DeepAccNet^[17]等。基于知识的力场已成功应用于许多领域, 包括折叠识别、从头算蛋白质结构预测、结构模型评估、蛋白质-蛋白质对接和蛋白质稳定性预测。然而, 蛋白质的巨大构象空间进行采样仍然是一个问题, 基于知识的力场方法也越来越多地依赖蛋白质序列和对应已知蛋白质结构的数据来解决这个问题。

2.3 CASP进展

CASP是一项社区范围的实验, 两年一次, 旨在确定和推进从氨基酸序列建模蛋白质结构的最新技术水平, 参与者将被邀请提交一组尚未公开实验结构的蛋白质模型。CASP最重要的衡量标准是全局距离测试总分(GDT_TS), 较高的值表示模型更接近给定的参考结构。这为研究小组提供了测试其蛋白质结构预测方法的机会, 并向研究团体和软件用户提供了对蛋白质结构建模最新技术水平的独立评估, 结果显示了过去两年取得的进展, 并揭示了未来应重点关注的方向。在2022年的CASP15中, 来自世界的162个小组提交了53 764个模型, 涉及5个预测类别的127个建模目标。

过去几年见证了蛋白质3D结构建模的突破。这一突破可归因于两个主要因素: a. 蛋白质序列数据库的非凡增长, 以及实验确定的结构数据库的多产性较低但也在稳步增长, 这两种类型的数据都可以在公共数据库中获得; b. 逐步将深度学习中的前沿方法引入成熟的蛋白质建模领域。在CASP14中, AlphaFold2预测模型的准确率取得了蛋白质结构预测领域突破性进展, 随后改进版本的RoseTTAFold发布, 使科学界可以广泛使用接近实验精度的蛋白质结构预测工具。CASP15的单体结构预测评估结果中, Yang-server、UM-TBM和

PEZYFolding 位列前三, 这些工具都将多个预测算法(包括 AlphaFold2) 组合到自己的预测算法中并进行整体优化排序, 这在一定程度上表明, 多方法组合优化可能是提高整体结构预测准确率的有效方向。

值得注意的是, 虽然这些工具不需要模板结构, 但它们确实依赖于多序列比对信息, 低同源性的蛋白质序列多序列比对(multiple sequence alignment, MSA) 质量降低, 从而导致最终预测结构的准确率降低。这些算法提供置信度指标, 例如 pLDDT (predicted local distance difference test) 表示预测模型对其在局部距离差异测试的置信度, 以及 PAE (predicted aligned error) 表示有关残基对之间相对位置的置信度, 为了结构生物学实验工作者可以更有效地使用蛋白质结构预测软件, 针对这些指标的验证与评估工作也陆续从跨膜蛋白、中心体和中心粒蛋白以及整个蛋白质组等各类蛋白质的预测结果中开展^[18]。

2.4 高精度预测蛋白质结构的影响

蛋白质结构预测领域的突破性进展并没有让实验结构生物学过时, 而是提供了前所未有的新机会。通过在分子置换中使用 AlphaFold2 预测的模型可以有助于解析 X 射线或冷冻电镜结构^[19]; 结合实验数据与预测模型有助于确定复杂的核孔复合体结构^[20]; 蛋白质结构预测可为实验工作者筛选可能的翻译后修饰(post-translational modification, PTM) 位点^[21]。

2022 年, AlphaFold PDB 中提供了超过 2.14 亿个预测的蛋白质结构, 涵盖了 UniProt 数据库中的大部分序列^[22], 极大地增加了具有蛋白质结构的蛋白质序列空间的覆盖范围, 并在结构生物学和结构生物信息学领域提出了新的挑战 and 机遇。缺乏蛋白质建模经验的研究人员现在可以使用这些模型来解决具有挑战性的生物学问题, 与此同时, 模型置信度指标对于做出可靠的解释仍然至关重要。这个新的结构数据库和相关的软件工具在计算和实验结构生物学领域引发了一系列研究, 包括优化输入 AlphaFold2 的多序列比对等。此外, 预测结构数据库的出现极大地改变了蛋白质结构域的格局, 数百万个结构域序列可能成为建模良好的结构域结构, 同时数据量的庞大对蛋白质结构域分类工作带来了挑战, 在初始 AlphaFold PDB 的 21 种模型生物中确定了 700 000 个推定的 CATH 结构域, 但其中 49% 的结构域因预测模型质量和无序区域被过滤掉^[23]。

3 挑战与展望

蛋白质结构是研究蛋白质功能的基础, 但结构解析方法耗时、昂贵且不适用于所有蛋白质。因此, 利用计算的方法从蛋白质序列出发预测结构变得非常必要。尽管在蛋白质结构预测领域应用深度学习给结构生物学带来了新的机遇, 但仍然没有从理论上解决蛋白质折叠的问题。此外, 为了识别高分辨率预测的蛋白质结构, 需要使用准确的能量函数来区分具有低均方根偏差的非常相似的结构。

现有的工具缺乏关于蛋白质动态特性的信息, 许多蛋白质采用对其功能至关重要的多种构象状态, 即结合其他蛋白质、核酸和小分子配体或在功能活性和非活性状态之间切换, 这对于理解和模拟蛋白质的功能状态很重要。在这种情况下, 深度学习也提供了一个可行的方向, 首先学习已知结构蛋白质样本的构象状态以及它们之间可能的转换路径, 训练完成的模型用于生成实验上未观察到但类似于天然的蛋白质构象来进一步扩展构象空间的采样^[24], 训练和测试此类方法的关键将是开发单链蛋白质和复合物的生理功能相关结构的数据集。

蛋白质复合物结构预测的方法依赖于蛋白质之间的共同进化信息, 但对于宿主-病原体的相互作用, 由于二者属于不同物质, 进化关系与源自同一生物体两种蛋白质的进化关系非常不同, 在宿主和致病蛋白之间几乎无法找到直系同源序列。因此, 预测这些复合物的预期准确性要低得多。那么完全从序列出发, 不依赖于多序列比对信息的模型开发将会是跨物种蛋白质复合物结构预测的一个可行方案, 另一种选择是利用迁移学习对现有模型进行微调以适应特定任务, 即使在缺乏数据的情况下也可以学习复杂的概念, 这在主要组织相容性复合体(MHC) 受体和肽链相互作用的预测已经有相关应用^[25]。

基于 AI 的结构预测算法的其他应用可能包括模拟翻译后修饰的结构效应、突变和变异的构象后果以及蛋白质设计领域的应用, 目前, 基于深度学习的方法仍然无法解释氨基酸单点突变对整体蛋白质结构的影响^[26]。开发更准确的基于先进人工智能技术的可靠分子对接算法来模拟蛋白质和小分子之间的相互作用, 可以促进基于结构的药物研发并加速医学研究。

随着端到端机器学习方法得到改进并被更广泛

的结构生物学界所掌握, 对很大一部分蛋白质及复合物仅从其氨基酸序列进行建模将成为常规, 同样, 将这些新方法扩展到预测核酸结构, 特别是RNA, 以及它们与蛋白质形成的复合物结构是未来结构预测领域重要的前进方向。要实现这些拓展, 一个主要挑战是整理足够的实验数据来训练和验证机器学习方法。未来计算和实验方法将更紧密结合, 从冷冻电镜数据中提取大分子复合物结构异质性信息, 基于AI的结构预测算法和分子模拟技术相结合, CASP在检验相关领域的算法预测结果中依然会发挥重要的参考作用。

参 考 文 献

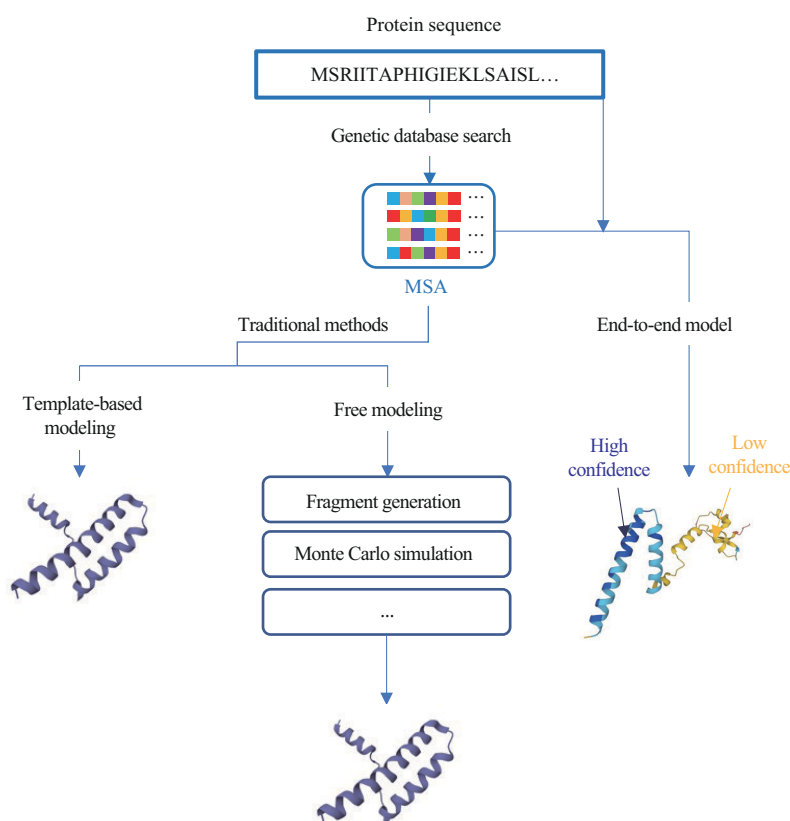
- [1] Ghouzam Y, Postic G, Guerin P E, *et al.* ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Sci Rep*, 2016, **6**: 28268
- [2] Rangwala H, Karypis G. *Introduction to Protein Structure Prediction*. New York: Wiley, 2010
- [3] Anfinsen C. B. Principles that govern the folding of protein chains. *Science*, 1973, **181**(4096): 223-230
- [4] Berman H M, Westbrook J, Feng Z, *et al.* The protein data bank. *Nucleic Acids Res*, 1999, **28**(Pt 6 No 1): 235-242
- [5] Ovchinnikov S, Park H, Varghese N, *et al.* Protein structure determination using metagenome sequence data. *Science*, 2017, **355**(6322): 294-298
- [6] Roy A, Yang Z. *Protein Structure Prediction*. New York: Wiley, 2012
- [7] Pearce R, Zhang Y. Deep learning techniques have significantly impacted protein structure prediction and protein design. *Curr Opin Struct Biol*, 2021, **68**: 194-207
- [8] Hagler A T, Huler E, Lifson S. Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals. *J Am Chem Soc*, 1974, **96**(17): 5319-5327
- [9] Brooks B R, Brooks C L, Mackerell A D, *et al.* CHARMM: the biomolecular simulation program. *J Comput Chem*, 2009, **30**(10): 1545-1614
- [10] Wang J, Wolf R M, Caldwell J W, *et al.* Development and testing of a general AMBER force field. *J Comput Chem*, 2004, **25**(9): 1157-1174
- [11] Isogai Y, Némethy G, Scheraga H A. Enkephalin: conformational analysis by means of empirical energy calculations. *Proc Natl Acad Sci USA*, 1977, **74**(2): 414-418
- [12] Novotný J, Bruccoleri R, Karplus M. An analysis of incorrectly folded protein models. Implications for structure predictions. *J Mol Biol*, 1984, **177**(4): 787-818
- [13] Zhou R. Free energy landscape of protein folding in water: explicit vs. implicit solvent. *Proteins*, 2003, **53**(2): 148-161
- [14] Tanaka S, Scheraga H A. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, 1976, **9**(6): 945-950
- [15] Zhang Y. I-TASSER: fully automated protein structure prediction in CASP8. *Proteins*, 2009, **77**(Suppl 9): 100-113
- [16] Simons K T, Kooperberg C, Huang E, *et al.* Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J Mol Biol*, 1997, **268**(1): 209-225
- [17] Hiranuma N, Park H, Baek M, *et al.* Improved protein structure refinement guided by deep learning based accuracy estimation. *Nat Commun*, 2021, **12**(1): 1340
- [18] Tunyasuvunakool K, Adler J, Wu Z, *et al.* Highly accurate protein structure prediction for the human proteome. *Nature*, 2021, **596**(7873): 590-596
- [19] Kryshchuk A, Moulton J, Albrecht R, *et al.* Computational models in the service of X-ray and cryo-electron microscopy structure determination. *Proteins*, 2021, **89**(12): 1633-1646
- [20] Mosalaganti S, Obarska-Kosinska A, Siggel M, *et al.* AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, 2022, **376**(6598): eabm9506
- [21] Bludau I, Willems S, Zeng W F, *et al.* The structural context of posttranslational modifications at a proteome-wide scale. *PLoS Biol*, 2022, **20**(5): e3001636
- [22] Varadi M, Anyango S, Deshpande M, *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res*, 2022, **50**(D1): D439-D444
- [23] Bordin N, Sillitoe I, Nallapareddy V, *et al.* AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Commun Biol*, 2023, **6**(1): 160
- [24] Ramaswamy V K, Musson S C, Willcocks C G, *et al.* Deep learning protein conformational space with convolutions and latent interpolations. *Phys Rev X*, 2021, **11**(1): 011052
- [25] Hekkelman M L, De Vries I, Joosten R P, *et al.* AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods*, 2023, **20**(2): 205-213
- [26] Buel G R, Walters K J. Can AlphaFold2 predict the impact of missense mutations on structure?. *Nat Struct Mol Biol*, 2022, **29**(1): 1-2

Advances in Protein Structure Prediction*

CAO Wei, PAN Xian-Ming**

(Key Laboratory of Ministry of Education for Protein Science, School of Life Sciences, Tsinghua University, Beijing 100084, China)

Graphical abstract



Abstract The structure of protein is determined by the sequence, and the function of protein is determined by its structure. The advent of accurate protein structure prediction tools has created new opportunities and challenges in the fields of structural biology, structural bioinformatics, drug discovery and many other fields of life sciences. The accuracy of single-chain protein structure prediction has reached a level comparable to that of experimental methods. In this review, we provide an overview of the theoretical basis, development history, and recent advances in the field of protein structure prediction. Additionally, we discuss how the large number of predicted protein structures and artificial intelligence-based methods affect experimental structural biology. Open questions and future research directions in the field of protein structure prediction are analyzed.

Key words protein structure prediction, deep learning, energy function

DOI: 10.16476/j.pibb.2023.0152

* This work was supported by a grant from The National Natural Science Foundation of China (39625008).

** Corresponding author.

Tel: 86-10-62792827, E-mail: pan-xm@mail.tsinghua.edu.cn

Received: March 20, 2023 Accepted: April 20, 2023