PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2024, 51(9): 1989~1999
www.pibb.ac.cn

高通量蛋白质结构生物信息学进展*

祝云箦　陆祖宏**
（东南大学生物科学与医学工程学院，数字医学工程全国重点实验室，南京 211189）

摘要　本文总结了高通量蛋白质结构生物信息学的最新进展，包括结构数据管理、工具软件开发和结构数据挖掘三个主要方面。结构数据管理方面，得益于类AlphaFold系统的发展，蛋白质结构数据量实现爆发式增长，直接促进了压缩技术的升级，也吸引了研究者对结构数据管理的关注。工具软件开发方面，以Foldseek为代表的新算法实现了高速的结构比对，突破了结构分析的通量瓶颈，此外深度学习模型的大量应用从多个方面改进了基于结构的蛋白质功能注释。结构数据挖掘方面，研究者以组学思维处理结构大数据，在持续的探索中提炼分析要素、优化方法，并在新工具的帮助下推动着结构数据挖掘的进阶。随着高通量方法的发展，结构生物信息学有望在生命科学中发挥更重要的作用。

关键词　蛋白质结构生物信息学，高通量，类AlphaFold系统，结构蛋白质组学
中图分类号　Q518　　**DOI**：10.16476/j.pibb.2024.0082

蛋白质是构成生命体的基石，其功能与活性紧密依赖于其复杂的三维结构。解析蛋白质的结构对于深入理解其生物学功能至关重要，这对于疾病机理的探索、新药的开发以及精准医疗的实施具有不可估量的价值[1]。传统的蛋白质结构解析技术，如X射线晶体学、核磁共振光谱学和冷冻电镜等，虽然精确，但往往耗时且成本高昂，与此同时，计算机辅助的蛋白质结构预测技术在过去长期受限于预测精度。上述种种因素导致相关数据积累的速度受限，蛋白质结构生物信息学的发展也因此未能获得应有的关注。

然而，这一局面在2020年底迎来了颠覆性的变革。DeepMind开发的AlphaFold2系统在2020年举办的蛋白质结构预测关键评估（The Critical Assessment of protein Structure Prediction，CASP）赛事（CASP14）中取得了历史性的突破，以其92.4的中位GDT（Global Distance Test）得分遥遥领先，成为首个能够实现大部分蛋白质单体高精度结构预测的系统[2]。AlphaFold2的横空出世，不仅标志着蛋白质结构预测领域的一次飞跃，也为高通量蛋白质生物信息学的研究和应用开辟了新的道路。

在生物信息学的广阔领域中，无论研究的具体方向如何，其核心使命可以归结为三大任务：首先是生物数据的收集与管理，这包括确保数据的高效存储、易于检索和共享，从而为全球的研究人员提供便捷的数据访问，奠定信息分析和数据挖掘的基础；其次是工具软件的开发，这旨在为生物信息学的具体应用提供支持，使研究人员能够更加高效地处理和分析生物数据；最后是数据的挖掘与分析，这一步骤涉及揭示数据间的内在联系，洞察数据的本质，并将这些信息提炼成有意义的生物学知识，以解答科学问题[3]。本文将围绕这三个方面，对高通量蛋白质结构生物信息学的最新进展进行综述。

1　结构数据管理

1.1　数据生产

AlphaFold系统在操作便捷性、成本和生产效率方面相较于传统技术有着显著优势，极大地简化了蛋白质结构预测的过程，从而推动了蛋白质结构数据量的爆炸性增长。2021年7月，在开源

* 国家重点研发计划（2016YFA0501600）资助项目。
** 通讯联系人。
Tel：86-25-83793779，E-mail：zhlu@seu.edu.cn
收稿日期：2024-03-04，接受日期：2024-04-18

AlphaFold2的同时，DeepMind发布了AlphaFold人类蛋白质组，覆盖98.5%的人类蛋白质结构，这是首次尝试应用人工智能系统进行蛋白质组水平结构建模[4]。2021年11月，AlphaFold蛋白质结构数据库（AlphaFold Protein Structure Database，简称AFDB）正式上线[5]。其由欧洲生物信息研究所和DeepMind共同开发、维护，以UniProt数据库为蛋白质序列源（长度范围16~2 700 aa），使用AlphaFold2-monomer进行大规模结构建模。截至2022年7月，该数据库已公开了超过2.14亿个蛋白质结构[6]，这一数字是蛋白质数据库（protein data bank，PDB）数据库50年累积数据量的千倍以上。

AlphaFold2的开源特性激发了众多开发者参与到AlphaFold2系统的改进与类AlphaFold系统的开发。如图1所示，原版AlphaFold2系统与绝大多数类AlphaFold系统遵循“序列-多序列比对-结构”的工作流程，其中“序列-多序列比对”涉及在模板数据集中搜索输入序列，生成多序列比对（multiple sequence alignment，MSA），这一步是CPU密集型计算，“多序列比对-结构”涉及将MSA输入AlphaFold2网络进行结构建模与松弛，是GPU密集型计算。最初级的优化思路以上海交通大学高性能计算中心开发的ParaFold为代表，将两步计算分离至对应的专用节点，以加速计算、节省资源[7]，这在高性能计算集群上尤为适用。许多研究者基于这一策略，对两个步骤分别进行优化，构建出各种各样的类AlphaFold系统。2022年5月发布的ColabFold系统是众多优化版本中的佼佼者。该系统采用MMseqs2替代AlphaFold2中原有的Jackhmmer进行模板搜索，并在GPU计算模块中引入了优化后的模型，使得运行速度提升了40~60倍，同时精度损失极小[8]。

Fig. 1 Workflow of AlphaFold/AlphaFold-like systems

图1 AlphaFold/类AlphaFold系统工作流程简图

开发者的不懈努力显著降低了高精度蛋白质结构预测的门槛，极大提升了结构数据生产通量。2022年11月，东南大学完成超8 000个造礁珊瑚同源蛋白的ColabFold建模[9]；2023年8月，橡树岭国家实验室使用超算特化版AlphaFold[10]完成泥炭藓结构蛋白质组建模[11]；2024年2月，美国国立卫生研究院（NIH）发布淡海栉水母AlphaFold蛋白质组[12]；等等。这些得益于AlphaFold改进版本的大规模建模成果不断拓展着AFDB奠定的蛋白质宇宙疆域。

需要强调的是，高通量的前提是高精度。本小节没有讨论类AlphaFold系统以外的高通量蛋白质结构建模方法，因为它们当中鲜有在实践中得到有力验证者。David Baker团队[13]开发的RoseTTAFold是最早开源的蛋白质结构预测神经网络模型，相比AlphaFold2初始版本有着显著的性能优势，但其精度有限，随着时间推移，该工具的用户群体逐步转向AlphaFold或ColabFold。Meta AI推出的ESMFold基于大语言模型[14]，无需构建MSA，运行速度惊人，但其在2022年CASP15的排名极低，远远落后于担任基准的ColabFold。从CASP15优胜队伍皆或多或少使用AlphaFold的结果[15]看，AlphaFold/类AlphaFold系统仍是当下最稳定可靠的结构预测数据来源，这种状况可能还会持续一段时间（表1）。

Table 1 Medalists of CASP15

表1 CASP15优胜方法一览

分类	优胜级	开发团队	方法	参考文献
蛋白质单体	1	Yang-Server	trRosettaX2、AlphaFold2双系统取最优结果	[16]
	2	UM-TBM	DeepMSA2构建MSA→LOMETS3模板搜索→AlphaFold2初始模型&初始几何约束→I-TASSER副本交换蒙特卡洛模拟	[17]
	3	PEZYFoldings	拓展数据库构建MSA→AlphaFold2建模→深度学习优化、人工干预	[18]
蛋白质复合物	1	Zheng	DeepMSA2构建MSA→AlphaFold-Multimer建模	[17]
	2	Venclovas	ColabFold系统优化	[19]
	3	Wallner	Dropout层激活后的AlphaFold-Multimer 2.1、2.2双版本取最优结果	[19]

1.2　数据存储与共享

PDB格式在结构生物学中扮演着不可或缺的角色，是这一领域里最基础且重要的文件格式。该格式以行为基本数据单元，每行均以特定字符开头，用以标识行的内容和类型。一个标准的PDB文件通常由标题、原子坐标、化学键连接、注释、晶体学参数等多个部分构成。在原子坐标部分，详细列出了原子的空间位置及相关属性，例如元素种类、占有率和温度因子。值得一提的是，由人工智能系统生成的PDB文件往往将温度因子字段用于展示预测模型的评分，如预测局部距离差检验（predicted local distance difference test，pLDDT），这类评分有助于评估模型结构的可靠性，也可服务于下游分析的质量控制。

PDB格式简单易懂，易于手工编辑，被广泛用于存储和共享结构数据，确保可复用性并促进全球研究人员之间的合作。但是，它也有一些缺点，比如数据重复、缺乏对复杂数据类型的支持等。大分子晶体学信息文件（macromolecular crystallographic information file，mmCIF）格式是为了解决PDB格式的局限性而开发的。它是一种基于文本的、自我描述的、可扩展的数据交换格式。该格式不仅可以存储生物大分子的三维结构信息，还可以存储实验条件、结构解析方法、文献引用等多方面的信息。此外，mmCIF格式支持数据分类和嵌套，使得数据的组织结构更加清晰。当然，这也意味着其存储空间占用通常远超PDB格式。随着人工智能技术在结构预测领域的蓬勃发展，大量的预测结构数据不断产生，存储压力的增加逐渐成为一个不容忽视的问题。

AFDB采用GZIP压缩的PDB和mmCIF格式来存储其庞大的数据集，其占用的总存储空间高达23 TiB。随着数据量的不断膨胀，尤其是整个蛋白质组级别数据的下载需求，对网络带宽造成了巨大的压力。这一现象使得之前鲜获关注的PDB压缩技术重新进入了研究者视野。2023年3月，首尔国立大学Martin Steinegger课题组发布了Foldcomp[20]，格式名为FCZ。该工具利用内部坐标和笛卡尔坐标的组合，以及双向NeRF策略来提高压缩比，可将AFDB全数据集压缩至1.2 TiB内。仅仅10天后，密歇根大学Zhang等[21]发布PDC格式处理工具，将蛋白质结构转换为扭转角空间表示，再进行有损压缩，效率与Foldcomp相当。7月，西里西亚技术大学Deorowicz等[22]发布ProteStAr工具，对AlphaFold模型的对齐误差（predicted aligned error，PAE）文件提供了额外的压缩支持。这些工具的接连发布充分体现了学术界对压缩技术的重视。需要注意的是，由于mmCIF格式过于灵活和复杂，当前并没有专门针对该格式的压缩算法，Foldcomp等工具也只能像将其转为PDB格式再压缩，这也意味着它们无法将文件直接解压为mmCIF格式。PDB向mmCIF的格式转换没有也不会有统一标准，因此结构生物信息学软件开发者应在输入处理模块考虑到mmCIF格式的多样性，并在输出环节提供尽量完整的信息。

随着类AlphaFold系统的普及，越来越多的研究者将蛋白质结构预测应用于自己的工作，所得的数据也应以恰当的方式归档与共享。目前，AFDB不收录第三方数据，其维护者建议研究者遵循RDMkit（The ELIXIR Research Data Management Kit[23]）的指导管理自产的AlphaFold模型。可选的公共存储库包括ModelArchive（CC BY-SA 4.0许可证）、PDB-Dev（CC0 1.0许可证）等，其中，截至2024年2月底ModelArchive已收录约55万个模型。研究者也被推荐使用MineProt[24]和3D-Beacons[25]等独立服务应用自建网站开放数据，以提供更丰富的交互功能，但这对他们的硬件条件和运维能力有额外要求。

2　工具软件开发

2.1　结构比对

比对（alignment）是生物信息学的一个核心概念及基本步骤。从宏观角度看，生物学家普遍认同“序列决定结构，结构决定功能”，结构比对结果对蛋白质功能的阐释理应比序列比对更精准；从微观角度看，结构比对无疑能够更加灵敏地捕获那些埋没在序列差异中的远源同源物（remote homolog，图2）。因此，结构比对问题应得到足够重视。

早在2005年，结构生物信息学的先驱张阳就强调了发展高效结构比对算法的重要性，并推出了TM-align工具[26]。TM-align至今仍是结构生物学领域不可或缺的比对工具，与之伴生的比对指标TM-score[27]亦被广泛采用。与之共同发展的还有DALI[28]、CE[29]等，它们都为蛋白质结构生物信息学的早期发展做出了贡献。然而，正如当年高通量测序数据的爆炸性增长让众多双序列比对算法难以应对，这些当初为双结构比对设计的软件也无法承担面向海量结构数据的搜索任务。对此，2023

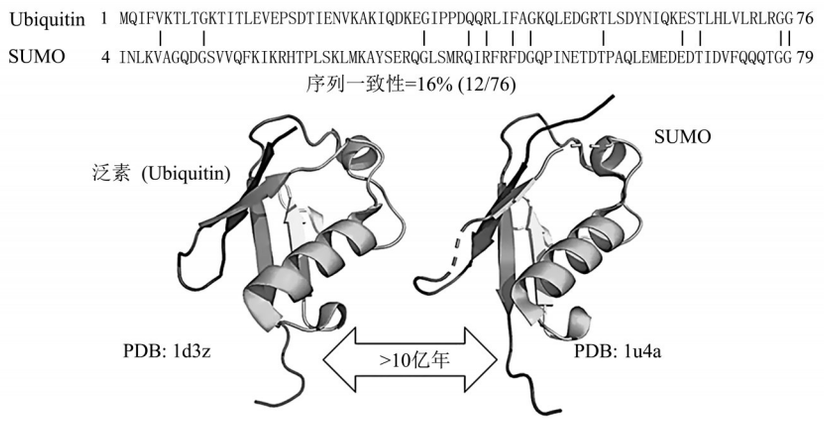


Fig. 2　Demo of remote homolog

图2　远源同源物示例

年5月，Martin Steinegger课题组发布了Foldseek[30]。该算法的主要步骤包括：将蛋白质结构编码为20状态的3Di序列，以描述氨基酸残基之间的三级相互作用；使用MMseqs2在3Di序列数据库中搜索候选结构；使用加速版TM-align对候选结构进行全局比对，打分、排序。Foldseek使用3Di描述残基间的相互作用，提高了信息密度，减少了假阳性，同时显著加快了速度，在SCOPe基准测试中比同类工具快4~5个数量级。结构生物信息学“次世代”的快速搜索工具诞生了[31]，其意义不亚于BLAST之于序列生物信息学，截至2024年2月底，它已收获上百次引用。

当然，结构比对领域的“加速赛道”并非只有Foldseek。2023年9月，纽约大学Richard Bonneau课题组发布TM-Vec、DeepBLAST[32]，利用深度学习直接从序列信息中学习结构特征，实现更快速、更准确的相似结构搜索；2023年11月，上海交通大学沈红斌课题组发布FoldExplorer[33]，该工具采用了序列增强的图嵌入方法来表示蛋白质结构，以向量比对代替序列比对，号称比Foldseek快数十倍；2024年1月，张阳课题组发布TM-search[34]，该工具全库搜索比TM-align快约30倍，且灵敏度极高。这些新工具究竟孰优孰劣还需经实践检验。

此外，加速绝非结构比对算法的唯一赛道。张阳课题组于2022年8月发布的US-align[35]采用通用策略和统一评分函数，实现蛋白质复合物等各类大分子的高效比对，拓展了结构比对的应用范围，也给算法工具开发指出了新方向。尽管在当前的蛋白质复合物结构预测领域中，原版AlphaFold-Multimer[36]预测精度有限，而在CASP15中表现优秀的DMFold[37]、MULTICOM[38]等工具的性能着实需要优化，但可以预期的是，一旦出现兼具高通量与高精度的生产工具，蛋白质复合物结构数据规模将会猛增，进而引发相关软件的开发浪潮。

2.2　基于结构的功能注释

基于结构的蛋白质功能注释是一个涵盖广泛的领域，配体预测、活性位点预测、蛋白质相互作用（PPI）、变异检测等问题皆可纳入。在生物信息学研究中，许多功能注释问题都可以简化为比对问题，但遗憾的是，由于高通量结构搜索的技术瓶颈到2023年才算突破，在此之前基于结构比对的功能注释工具数量不多。荷兰癌症研究所于2022年11月发布的AlphaFill[39]是其中的代表。该算法通过序列比对找到与AlphaFold模型相似的实验确定的蛋白质结构，然后进行结构比对，以确定小分子和离子的位置，最后将它们移植到AlphaFold模型中，从而丰富模型的信息。该团队已对AFDB中的近100万个模型进行移植，产生了1 200万个移植结果，并搭建了在线服务开放数据（alphafill.eu），足见该工具效率合格，能够处理大量结构数据。张阳课题组的COFACTOR也采用了序列比对、结构比对相结合的思路[40]，但该应用尚无公开的本地化版本，因此无法评价其能否满足后AlphaFold时代高通量的需求。

另一种思路是利用蛋白质结构功能注释库，例如由PDBe和UniProtKB共同维护、建立PDB结构与生物数据交叉引用的SIFTS[41-42]，由张阳课题组

维护、专注蛋白质-配体相互作用的BioLiP系列[43-44]等，来训练深度学习模型。这些数据库的筹建具有一定的前瞻性，因为在前AlphaFold时代，蛋白质数据集多序列而寡结构，绝大多数基于人工智能的功能注释模型都是从序列中训练。DeepFRI[45]是少有的可以直接接受结构输入的模型，它结合了提取序列特征的长短期记忆网络（long short-term memory，LSTM）语言模型与学习蛋白质结构特征的图卷积网络，利用梯度加权类激活映射方法，实现蛋白质结构上的功能位点预测，同时具有强大的去噪能力，可以精准高效地对AlphaFold模型进行基因本体（gene ontology，GO）注释。该工具被广泛应用于结构蛋白质组学分析中。不过，DeepFRI的成功并不意味着序列依赖的深度学习模型会逐步没落，随着由结构转化、表征氨基酸互作的3Di序列的出现，它们有机会获得进一步的发展。2023年11月发布的TT3D（Topsy-Turvy 3D）[46]便是一个极好的例子。该工具是PPI预测模型D-SCRIPT[47]的拓展，输入层添加了3Di序列的编码，在跨物种PPI预测任务中取得了显著提升的效果。序列模型的本质又使其足够轻量化，能够应对全蛋白质组级别的预测任务。由此可见，即使进入高通量结构蛋白质组时代，基于序列的人工智能模型仍将在蛋白质功能注释方面发挥巨大作用。

分子对接在计算生物学中的地位举足轻重，也常被用于蛋白质功能预测。然而，传统的分子对接软件，如AutoDock[48]、SwissDock[49]、ClusPro[50]、HDOCK[51]、ZDOCK[52]、HADDOCK[53]等，普遍存在操作繁琐、计算资源消耗大等问题[1]，这导致当下许多用户转向使用AlphaFold-Multimer之类的人工智能系统[54]。后AlphaFold时代的海量蛋白质结构数据更是暴露了许多分子对接工具本地化工作不足的缺陷，对算法性能提出了进一步的挑战。一些研究者将深度学习与分子对接结合起来，提升精度的同时简化操作、降低计算成本，代表性成果有gnina[55]、KarmaDock[56]、DeepRMSD-Vina[57]等，但总体而言，分子对接算法开发者对通量问题的关注程度仍有待提高。

3　结构数据挖掘

AlphaFold和AFDB无疑给广大生物学者的工作带来了巨大影响[58]，在生物医学领域已有数不清的应用[59]。其中受影响最大的无疑是结构生物学工作者，AlphaFold模型与传统实验技术的结合提高了大尺寸蛋白质的结构解析速度和精度，改变了他们的工作方式。AlphaFold模型本身就是优质的模板，而剑桥大学Oeffner等[60]在phenix.process_predicted_model工具的开发中进一步探究了基于PAE的结构域分割、基于pLDDT的距离约束等，以发挥AlphaFold模型各类置信度指标的作用，将模型更充分地集成至结构解析流程中，服务于后续的分子置换或冷冻电镜重构。后AlphaFold时代的许多结构解析工作，如VP8*结构域[61]、解旋酶-引发酶D5[62]、IL-27信号复合体[63]、嗜热毛壳菌（*Chaetomium thermophilum*）代谢子（metabolon）[64]等，都含有这些策略的影子，到2024年，一些可用的流水线已经成型，例如结合OpenFold（AlphaFold的PyTorch移植版[65]）与冷冻电镜的CryoFEM[66]，综合AlphaFold结构预测与深度学习优化策略以从冷冻电镜密度图中构建结构的EMBuild[67]、DEMO-EM系列[68-69]、DeepMainmast[70]，结合AlphaFold与X射线晶体学的IPCAS 3.0[71]等。尽管受“木桶效应”影响，此类工作的数据生产缓慢，很难达到高通量标准，但它们无疑提高了蛋白质结构数据质量，有助于蛋白质功能数据库的完善，还为下游生物信息学分析的质量控制环节提供了许多经验。

对生物信息学工作者而言，“AlphaFold革命”的实质在于通量提升造成的质变，正如当年二代测序（next-generation sequencing）技术的突破带来了基因组学的全面繁荣。全蛋白质组建模（whole-proteome structuring，WPS）如今已经可行，那么WPS的下游分析管道也应开始构建。在2023年结构搜索算法“大提速”之前，已有研究者开始思考如何将AlphaFold的预测结果上升至组学高度加以分析。最早的AlphaFold人类蛋白质组[4]虽然是单调的建模报告，只包含少数关键蛋白质的讨论而无蛋白质组水平的系统分析，但是报告中基于pLDDT、pTM的质量控制为之后所有WPS工作提供了宝贵的参考；2022年东南大学团队针对珊瑚结构蛋白质组[9]的分析环节加入了对公共结构数据集的比对，但这一分析主要停留在统计层面，并未实现全面深入的解析；2023年橡树岭国家实验室团队在泥炭藓结构蛋白质组[11]的解析中，将结构比对方法应用于酶筛选，尽管自动化程度不高、分析的蛋白质数量有限，但已经展现了基于结构的

功能分析管道的雏形。也有研究者从公共数据入手，对庞大的AFDB展开挖掘。2022年5月，德国马普所Bludau等[72]利用AlphaFold模型研究不同类型翻译后修饰（post-translational modification，PTM）在蛋白质结构中的位置；2022年9月，日本理化学研究所Tang等[73]对AFDB中不同生物体的组成蛋白质进行了比较分析，主要关注了蛋白质的回转半径、螺旋分数和振动频率等特征；2023年1月，芬兰赫尔辛基大学团队使用DALI[28]在AFDB中发现了100个新的远源同源关系，其中约一半涉及未知功能的蛋白质家族[74]。还有研究者通过手工整理或计算机辅助，在AFDB基础上建立各种子数据集以服务不同领域的需求，代表性工作有关注跨膜蛋白的AFTM[75]、使用升级版GalaxySite标注结合位点的HProteome-BSite[76]等。在缺乏高通量结构搜索工具的时间段，这些工作所投入的计算资源和辛勤劳动是无法估量的，早期研究者的努力值得尊重。

Foldseek作为“次世代”结构搜索工具，有望奠定高通量结构蛋白质组学分析方法的基石。2023年9月，欧洲生物信息研究所等多家单位联合发表了AFDB的结构聚类分析工作[77]。在Foldseek的辅助下，AFDB的2.14亿个蛋白质结构被聚为230万簇，相比AFDB全库，该非冗余数据集降低了存储需求，提高了搜索效率，极大增强了可用性，使AFDB的本地化操作变得可行。在结构生物信息学领域，这一成就的重要性堪比NCBI非冗余蛋白质序列数据库（NCBI non-redundant protein sequences，NR）之于序列生物信息学，结构蛋白质组学分析管道上游自此拥有了宝贵的“参考组”，为蛋白质功能注释和进化分析提供了新线索[78]。同日，巴塞尔大学Durairaj等[79]发表了在AFDB中鉴定的290个潜在新蛋白质家族、1个新蛋白质折叠类型β-flower以及1个新的毒素-抗毒素超家族TumE-TumA，在这项工作中，Foldseek同样发挥了关键作用。由此可见，不断发展的高通量结构搜索技术会不断推动结构数据挖掘的深入与WPS下游分析方法的完善，并促进高级软件工具的开发。

如今，结构数据挖掘仍处于初级阶段，需要进一步明确分析要素及评价指标、发展高通量分析方法、打磨分析工具链，从而完善WPS下游分析管道，以服务于更多科学问题的解答。

4 总结与展望

以AlphaFold2为代表的人工智能系统使得蛋白质结构预测达到了前所未有的精度和通量，推动了结构数据量的爆炸性增长，将结构蛋白质组学与蛋白质结构生物信息学推入了高通量时代。本文系统梳理了高通量蛋白质结构生物信息学的最新进展，涵盖结构数据管理、工具软件开发、结构数据挖掘三个主要方面。结构数据规模的增长促成了数据管理方案的改进、激励了相关工具软件的开发，而前两者的升级又推动着结构数据挖掘的进阶。然而，由于高通量时代初来乍到，这一领域目前存在两大问题：

第一，生物信息学工具的“寡极格局”。如表2所示，在符合“高通量”的前提下，蛋白质结构生物信息学数据生产以AlphaFold/类AlphaFold系统独大，数据搜索以Foldseek独快，目前尚无其他工具能明显挑战它们的地位，下游分析的可用工具数量较少，许多子领域还鲜有研究者涉足，由于开发时间较短，许多软件存在安装困难、用户体验不友好等问题。相信随着时间发展，在开发者的共同努力下，“寡极格局”终将变为“一超多强”乃至“百花齐放”，蛋白质结构生物信息学软件生态的繁荣能够为更多研究工作提供助力。

第二，蛋白质复合物结构预测的通量瓶颈仍未突破，这导致“缺乏复合物处理功能”成了当下高通量蛋白质结构生物信息工具的普遍问题（表2）。如今，学术界的目光已不再局限于蛋白质单体结构，高精度结构预测的曙光已照向核酸、蛋白质复合物、含PTM的生物分子等更多物质，CASP15赛场也涌现了一批优秀的方法模型。在解决精度问题的基础上，研究人员应更加关注通量问题，借鉴ColabFold开发能本地化、全自动化运行的高通量复合物结构预测工具，或者建设类似AFDB的大规模复合物结构数据库。只有突破了通量瓶颈，蛋白质复合物结构预测才能对生物信息学产生显著的影响，这一方面亟需更多研究者的参与。

总之，无数的挑战与机遇将不断推动蛋白质结构生物信息学的发展。高通量技术的发展有望全方位抬升结构生物信息学在生命科学中的地位，期待着在这一领域出现更多颠覆性的创新成果。

Table 2　High-throughput toolkit for protein structural bioinformatics

表2　高通量蛋白质结构生物信息学工具箱一览

生物信息学问题	解决方案	已知缺点
数据生产	AlphaFold/类AlphaFold系统	复合物预测通量不足、处理大蛋白质易受GPU显存限制、部分特殊蛋白预测效果不佳等
结构压缩	Foldcomp、PDC等	有损压缩；不支持复合物
结构比对	一对多、多对多：Foldseek	复合物搜索功能有待完善
	一对一：TM-align、US-align等	/
结构策展	3D-Beacons Client、MineProt等	对硬件、运维有较高要求；复合物支持有待完善
功能注释	GO注释：DeepFRI	GPU显存利用不足；不支持复合物
	配体注释：AlphaFill	安装困难
	PPI预测：TT3D	内存占用过大

参　考　文　献

[1] Paiva V A, Gomes I S, Monteiro C R, *et al*. Protein structural bioinformatics: an overview. Comput Biol Med, 2022, **147**: 105695

[2] Jumper J, Evans R, Pritzel A, *et al*. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, **596**: 583-589

[3] 孙啸, 陆祖宏, 谢建明. 生物信息学基础. 北京: 清华大学出版社, 2005: 4-5
Sun X, Lu Z H, Xie J M. Fundamentals of Bioinformatics. Beijing: Tsinghua University Press, 2005: 4-5

[4] Tunyasuvunakool K, Adler J, Wu Z, *et al*. Highly accurate protein structure prediction for the human proteome. Nature, 2021, **596**(7873): 590-596

[5] 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

[6] Varadi M, Bertoni D, Magana P, *et al*. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. Nucleic Acids Res, 2024, **52**(D1): D368-D375

[7] Zhong B, Su X, Wen M, *et al*. ParaFold: paralleling AlphaFold for large-scale predictions//International Conference on High Performance Computing in Asia-Pacific Region Workshops. New York: Association for Computing Machinery, 2022: 1-9

[8] Mirdita M, Schütze K, Moriwaki Y, *et al*. ColabFold: making protein folding accessible to all. Nat Meth, 2022, **19**: 679-682

[9] Zhu Y, Liao X, Han T, *et al*. Utilizing an artificial intelligence system to build the digital structural proteome of reef-building corals. Gigascience, 2022, **11**: giac117

[10] Gao M, Coletti M, Davidson R B, *et al*. Proteome-scale Deployment of Protein Structure Prediction Workflows on the Summit Supercomputer//2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). Washington: IEEE Computer Society, 2022: 206-215

[11] Davidson R B, Coletti M, Gao M, *et al*. Predicted structural proteome of Sphagnum divinum and proteome-scale annotation. Bioinformatics, 2023, **39**(8): btad511

[12] Moreland R T, Zhang S, Barreira S N, *et al*. An AI-generated proteome-scale dataset of predicted protein structures for the ctenophore *Mnemiopsis leidyi*. Proteomics, 2024: e2300397

[13] Baek M, DiMaio F, Anishchenko I, *et al*. Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021, **373**(6557): 871-876

[14] Lin Z, Akin H, Rao R, *et al*. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 2023, **379**(6637): 1123-1130

[15] Callaway E. After AlphaFold: protein-folding contest seeks next big breakthrough. Nature, 2023, **613**(7942): 13-14

[16] Peng Z, Wang W, Wei H, *et al*. Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. Proteins, 2023, **91**(12): 1704-1711

[17] Zheng W, Wuyun Q, Freddolino P L, *et al*. Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. Proteins, 2023, **91**(12): 1684-1703

[18] Oda T. Improving protein structure prediction with extended sequence similarity searches and deep-learning-based refinement in CASP15. Proteins, 2023, **91**(12): 1712-1723

[19] Ozden B, Kryshtafovych A, Karaca E. The impact of AI-based modeling on the accuracy of protein assembly prediction: insights from CASP15. Proteins, 2023, **91**(12): 1636-1657

[20] Kim H, Mirdita M, Steinegger M. Foldcomp: a library and format for compressing and indexing large protein structure sets. Bioinformatics, 2023, **39**(4): btad153

[21] Zhang C, Pyle A M. PDC: a highly compact file format to store protein 3D coordinates. Database: Oxford, 2023, **2023**: baad018

[22] Deorowicz S, Gudys A. Efficient protein structure archiving using ProteStAr. bioRxiv, 2023. DOI: 10.1101/2023.07. 20.549913

[23] Harrow J, Drysdale R, Smith A, *et al*. ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics, 2021, **37**(16): 2506-2511

[24] Zhu Y, Tong C, Zhao Z, *et al*. MineProt: a stand-alone server forstructural proteome curation. Database, 2023, **2023**: 0

[25] Varadi M, Nair S, Sillitoe I, *et al*. 3D-Beacons: decreasing the gap

between protein sequences and structures through a federated network of protein structure data resources. Gigascience, 2022, **11**: giac118

[26] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res, 2005, **33**(7): 2302-2309

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

[28] Holm L, Sander C. Dali: a network tool for protein structure comparison. Trends Biochem Sci, 1995, **20**(11): 478-480

[29] Shindyalov I N, Bourne P E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng Des Sel, 1998, **11**(9): 739-747

[30] van Kempen M, Kim S S, Tumescheit C, *et al*. Fast and accurate protein structure search with Foldseek. Nat Biotechnol, 2024, **42**: 243-246

[31] Hutson M. Foldseek gives AlphaFold protein database a rapid search tool. Nature, 2023. DOI: 10.1038/d41586-023-02205-4

[32] Hamamsy T, Morton J T, Blackwell R, *et al*. Protein remote homology detection and structural alignment using deep learning. Nat Biotechnol, 2024, **42**(6): 975-985

[33] Liu Y, Shen H. FoldExplorer: fast and accurate protein structure search with sequence-enhanced graph embedding. arXiv, 2023. DOI: 10.48550/arXiv.2311.18219

[34] Liu Z, Zhang C, Zhang Q, *et al*. TM-search: an efficient and effective tool for protein structure database search. J Chem Inf Model, 2024, **64**(3): 1043-1049

[35] Zhang C, Shine M, Pyle A M, *et al*. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. Nat Meth, 2022, **19**: 1109-1115

[36] Evans R, O'Neill M, Pritzel A, *et al*. Protein complex prediction with AlphaFold-Multimer. bioRxiv, 2021. DOI: 10.1101/2021.10.04.463034

[37] Zheng W, Wuyun Q, Li Y, *et al*. Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. Nat Meth, 2024, **21**: 279-289

[38] Liu J, Guo Z, Wu T, *et al*. Enhancing alphafold-multimer-based protein complex structure prediction with MULTICOM in CASP15. Commun Biol, 2023, **6**(1): 1140

[39] Hekkelman M L, de Vries I, Joosten R P, *et al*. AlphaFill: enriching AlphaFold models with ligands and cofactors. Nat Meth, 2023, **20**: 205-213

[40] Zhang C, Freddolino P L, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Res, 2017, **45**(W1): W291-W299

[41] Dana J M, Gutmanas A, Tyagi N, *et al*. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res, 2019, **47**(D1): D482-D489

[42] Choudhary P, Anyango S, Berrisford J, *et al*. Unified access to up-to-date residue-level annotations from UniProtKB and other biological databases for PDB data. Sci Data, 2023, **10**(1): 204

[43] Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res, 2013, **41**(D1): D1096-D1103

[44] Zhang C, Zhang X, Freddolino P L, *et al*. BioLiP2: an updated structure database for biologically relevant ligand-protein interactions. Nucleic Acids Res, 2024, **52**(D1): D404-D412

[45] Gligorijević V, Renfrew P D, Kosciolek T, *et al*. Structure-based protein function prediction using graph convolutional networks. Nat Commun, 2021, **12**(1): 3168

[46] Sledzieski S, Devkota K, Singh R, *et al*. TT3D: leveraging precomputed protein 3D sequence models to predict protein-protein interactions. Bioinformatics, 2023, **39**(11): btad663

[47] Sledzieski S, Singh R, Cowen L, *et al*. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. Cell Syst, 2021, **12**(10): 969-982.e6

[48] Trott O, Olson A J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem, 2010, **31**(2): 455-461

[49] Grosdidier A, Zoete V, Michielin O. SwissDock, a protein-small molecule docking web service based on EADock DSS. Nucleic Acids Res, 2011, **39**(suppl_2): W270-W277

[50] Kozakov D, Hall D R, Xia B, *et al*. The ClusPro web server for protein-protein docking. Nat Protoc, 2017, **12**(2): 255-278

[51] Yan Y, Tao H, He J, *et al*. The HDOCK server for integrated protein-protein docking. Nat Protoc, 2020, **15**(5): 1829-1852

[52] Pierce B G, Wiehe K, Hwang H, *et al*. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics, 2014, **30**(12): 1771-1773

[53] van Zundert G C P, Rodrigues J P G L M, Trellet M, *et al*. The $HADDOCK_{2.2}$ web server: user-friendly integrative modeling of biomolecular complexes. J Mol Biol, 2016, **428**(4): 720-725

[54] Tsuchiya Y, Yamamori Y, Tomii K. Protein-protein interaction prediction methods: from docking-based to AI-based approaches. Biophys Rev, 2022, **14**(6): 1341-1348

[55] McNutt A T, Francoeur P, Aggarwal R, *et al*. GNINA 1.0: molecular docking with deep learning. J Cheminform, 2021, **13**(1): 43

[56] Zhang X, Zhang O, Shen C, *et al*. Efficient and accurate large library ligand docking with KarmaDock. Nat Comput Sci, 2023, **3**(9): 789-804

[57] Wang Z, Zheng L, Wang S, *et al*. A fully differentiable ligand pose optimization framework guided by deep learning and a traditional scoring function. Brief Bioinform, 2023, **24**(1): bbac520

[58] Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. Proteomics, 2023, **23**(17): e2200128

[59] Yang Z, Zeng X, Zhao Y, *et al*. AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther, 2023, **8**(1): 115

[60] Oeffner R D, Croll T I, Millán C, *et al*. Putting AlphaFold models to work with phenix. process_predicted_model and ISOLDE. Acta Crystallogr D Struct Biol, 2022, **78**(pt 11): 1303-1314

[61] Hu L, Salmen W, Sankaran B, *et al*. Novel fold of rotavirus glycan-binding domain predicted by AlphaFold2 and determined by X-ray crystallography. Commun Biol, 2022, **5**(1): 419

[62] Li Y, Zhu J, Guo Y, *et al*. Structural insight into the assembly and working mechanism of helicase-primase D5 from Mpox virus. Nat Struct Mol Biol, 2024, **31**(1): 68-81

[63] Jin Y, Fyfe P K, Gardner S, *et al*. Structural insights into the assembly and activation of the IL-27 signaling complex. EMBO Rep, 2022, **23**(10): e55450

[64] Skalidis I, Kyrilis F L, Tüting C, *et al*. Cryo-EM and artificial intelligence visualize endogenous protein community members. Structure, 2022, **30**(4): 575-589.e6

[65] Ahdritz G, Bouatta N, Kadyan S, *et al*. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. bioRxiv, 2022. DOI: 10.1101/2022.11.20.517210

[66] Dai X, Wu L, Yoo S, *et al*. Integrating AlphaFold and deep learning for atomistic interpretation of cryo-EM maps. Brief Bioinform, 2023, **24**(6): bbad405

[67] He J, Lin P, Chen J, *et al*. Model building of protein complexes from intermediate-resolution cryo-EM maps with deep learning-guided automatic assembly. Nat Commun, 2022, **13**(1): 4066

[68] Zhou X, Li Y, Zhang C, *et al*. Progressive assembly of multi-domain protein structures from cryo-EM density maps. Nat Comput Sci, 2022, **2**(4): 265-275

[69] Zhang Z, Cai Y, Zhang B, *et al*. DEMO-EM2: assembling protein complex structures from cryo-EM maps through intertwined chain and domain fitting. Brief Bioinform, 2024, **25**(2): bbae113

[70] Terashi G, Wang X, Prasad D, *et al*. DeepMainmast: integrated protocol of protein structure modeling for cryo-EM with deep learning and structure prediction. Nat Methods, 2024, **21**(1): 122-131

[71] Li Z, Fan H, Ding W. Solving protein structures by combining structure prediction, molecular replacement and direct-methods-aided model completion. IUCrJ, 2024, **11**(pt 2): 152-167

[72] 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

[73] Tang Q Y, Ren W, Wang J, *et al*. The statistical trends of protein evolution: a lesson from AlphaFold database. Mol Biol Evol, 2022, **39**(10): msac197

[74] Holm L, Laiho A, Törönen P, *et al*. DALI shines a light on remote homologs: One hundred discoveries. Protein Sci, 2023, **32**(1): e4519

[75] Pei J, Cong Q. AFTM: a database of transmembrane regions in the human proteome predicted by AlphaFold. Database (Oxford), 2023, **2023**: baad008

[76] Sim J, Kwon S, Seok C. HProteome-BSite: predicted binding sites and ligands in human 3D proteome. Nucleic Acids Res, 2023, **51**(D1): D403-D408

[77] Barrio-Hernandez I, Yeo J, Jänes J, *et al*. Clustering predicted structures at the scale of the known protein universe. Nature, 2023, **622**: 637-645

[78] Bordin N, Lau A M, Orengo C. Large-scale clustering of AlphaFold2 3D models shines light on the structure and function of proteins. Mol Cell, 2023, **83**(22): 3950-3952

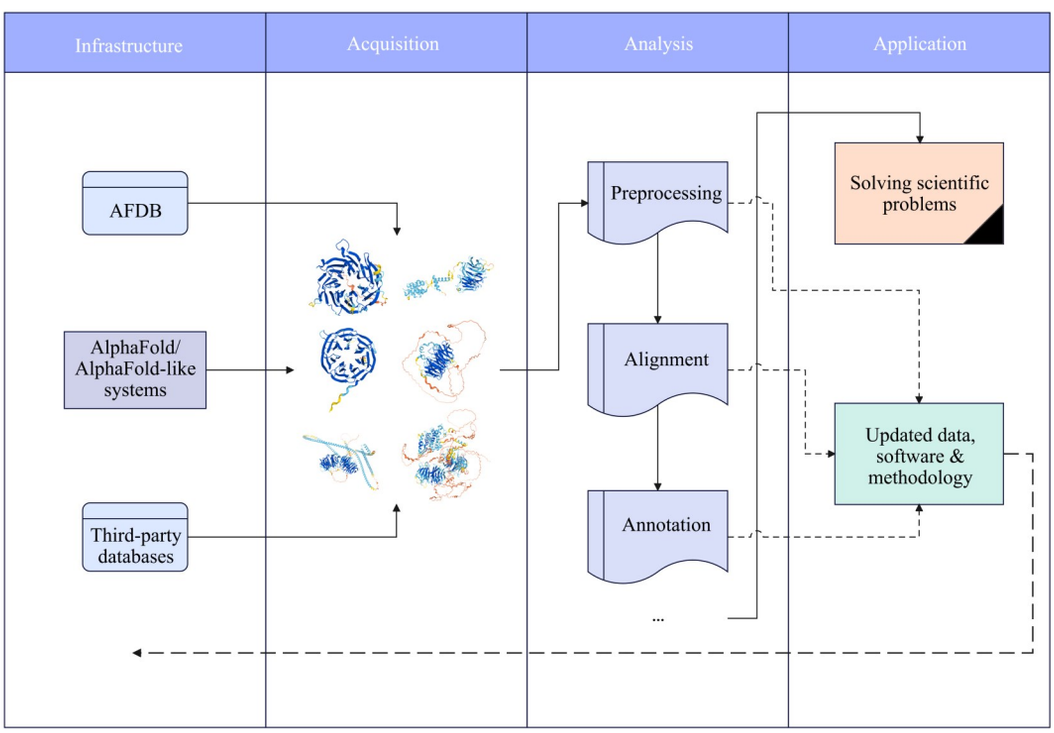
[79] Durairaj J, Waterhouse A M, Mets T, *et al*. Uncovering new families and folds in the natural protein universe. Nature, 2023, **622**: 646-653

Advances in High-throughput Protein Structural Bioinformatics*

ZHU Yun-Chi, LU Zu-Hong**

(*State Key Laboratory of Digital Medical Engineering, School of Biological Science and Medical Engineering, Southeast University, Nanjing* 211189, China)

Graphical abstract



Abstract　This review provides a comprehensive summary of the latest advancements in high-throughput protein structural bioinformatics, a field that has undergone a revolutionary transformation with the advent of deep learning-based protein structure prediction systems like AlphaFold2. These systems have significantly increased the accuracy, speed, and scale of protein structure prediction, resulting in an exponential growth in the number of protein structures available for analysis. Notably, the AlphaFold Protein Structure Database (AFDB) has amassed over 214 million protein structures, surpassing the PDB's 50-year cumulative data by over 1 000-fold within several months. Big data is driving the comprehensive upgrade of protein structural bioinformatics. This review focuses on three main areas: structure data management, tool development, and structure data mining. In the realm of structure data management, the review spotlights the optimization strategy of AlphaFold-like systems, which

* This work was supported by a grant from National Key Research and Development Program of China (2016YFA0501600).

** Corresponding author.

Tel：86-25-83793779，E-mail：zhlu@seu.edu.cn

Received：March 4，2024　Accepted：April 18，2024

significantly reduces the resource requirements for protein folding, enabling more researchers to make custom structure predictions and further enlarging the data scale. The resulting “data explosion” has exerted increased pressure on storage and bandwidth, prompting the development of cutting-edge tools such as Foldcomp, PDC, and ProteStAr for compressing PDB files. Moreover, the review underscores the critical role of public repositories like ModelArchive and PDB-Dev in archiving and sharing third-party AlphaFold models. It also highlights the utilization of independent services like MineProt and 3D-Beacons to create more interactive and accessible data portals. In terms of tool development, the review spotlights recent breakthroughs in structure alignment algorithms, represented by Foldseek, which enable ultra-fast searching of large protein structure databases. It also covers tools for functional annotation of proteins based on their structures, including AlphaFill for ligand annotation, DeepFRI for Gene Ontology (GO) annotation, TT3D for protein-protein interaction (PPI) prediction, among others. It is proposed that 3Di sequences born concurrently with Foldseek can enhance many sequence-based deep learning models developed in the pre-AlphaFold era, enabling them to be applied to structure-based function prediction. The challenges on traditional molecular docking methods in the high-throughput era are mentioned at last, in a gesture to arouse the attention of researchers. Finally, the review explores the burgeoning field of structure data mining. Whole proteome structuring has become feasible in recent years, and scientists are processing large structure datasets from an omics viewpoint, continuously identifying analyzable elements and optimizing methodologies, as well as utilizing newly developed tools to push the boundaries. Notable examples include the identification of new protein families, the development of protein structure clustering, and the integration of AlphaFold with conventional experimental techniques to solve large structures. These advancements are paving the way for a deeper understanding of protein structure and function and have the potential to unlock new discoveries in the life sciences. However, the review also acknowledges the challenges and limitations that persist in the field, including the lack of diversity in high-throughput software for protein structural bioinformatics and the existing bottleneck in rapidly predicting protein complex structures. Overall, structural bioinformatics is expected to play an even more crucial role in the life sciences with the development of high-throughput methodology.

Key words protein structural bioinformatics, high-throughput, AlphaFold-like system, structural proteomics
DOI：10.16476/j.pibb.2024.0082