DOI:10.3969/j.issn.1005-202X.2020.09.023

医学生物物理

蛋白质结构预测综述

刘子楠,黎河山,宋枭禹
哈尔滨工业大学生命科学与技术学院,黑龙江 哈尔滨 150080

【摘要】蛋白质结构预测对于从分子层面理解蛋白质的生物功能具有重要意义。本研究从同源建模、自由建模等经典方法以及深度学习这几个方面来阐述蛋白质结构预测方面的进展。已知结构蛋白质模板数量的增加、序列比对等算法对信息提取能力的提升以及片段拼接技术的应用使得同源建模在预测蛋白结构的能力大大提升。域分割和片段分割技术及并行计算策略的应用使得自由建模方法在预测远程氨基酸接触能力不断提升。深度学习技术与以上经典方法的结合提升了蛋白结构预测的准确性和速度,但是对于没有同源性蛋白结构的预测,仍然存在巨大的挑战。

【关键词】蛋白质结构预测;深度学习;同源建模;自由建模;综述
【中图分类号】R318;Q71 【文献标志码】A 【文章编号】1005-202X(2020)09-1203-05

Survey on protein structure predication

LIU Zinan, LI Heshan, SONG Xiaoyu
School of Life Science and Technology, Harbin Institute of Technology, Harbin 150080, China

Abstract: The prediction of protein structure has provided important underpinnings for understanding the biological functions of protein at molecular level. Herein the development of protein structure predication is reviewed from the aspects of homology modeling, ab initio prediction and deep learning. Because of the increasing number of protein templates with known structure, the improved performance of sequence alignment algorithm for information extraction and the application of fragment assembly technology, the ability of homology modeling for predicting protein structure is greatly increased. The prediction of long-distance amino-acid contact using ab initio method is more and more precise for the application of regional partition and fragment segmentation and the strategy of parallel computing in ab initio method. Deep learning combined with the above-mentioned classical technologies can promote the accuracy and speed of protein structure prediction, but there is still a great challenge when the protein has no homologous proteins.

Keywords: protein structure prediction; deep learning; homology modeling; ab initio prediction; review

前 言

蛋白质是生命功能的行使者,其构象普遍具有独一无二的生物学活性。蛋白质分子的结构信息蕴含了分子层面的功能信息[1],其结构对于理解生物基本反应是至关重要的。一种蛋白质的生物学功能由其三维结构决定,而蛋白质结构又由其氨基酸单体的一维链进行编码。如何利用氨基酸的一维序列预测蛋白质的三维结构是Anfenshen等[2-3]于1959年提出的问题。距今已经有半个多世纪。蛋白质结构预测已成为多学科交叉领域,包含物理、化学、生物乃至计算机和材料科学领域。

【收稿日期】2020-03-14
【作者简介】刘子楠,硕士研究生,研究方向:生物医学工程,E-mail: lzn_home@qq.com
【通信作者】宋枭禹,E-mail: songxyhit@hit.edu.cn

从目前计算机学的角度来看,设计出一个从一维氨基酸序列出发,最终能够预测蛋白质立体天然结构的算法是极其困难的。在基因组测序计划的巨大成功下,由于高通量测序技术的发展,已知的蛋白序列的数目呈爆发式增加。但是由于主流的获取蛋白质结构的方法都需要繁重的劳动以及其它局限,已知的蛋白结构数目远小于已知蛋白序列的数目。因此,根据一维序列高精度地预测结构的办法是很有研究意义的。

在蛋白质结构预测领域中,创建于1994年的蛋白质结构预测竞赛(Critical Assessment of protein Structure Prediction, CASP)是一个里程碑事件[4-5]。此竞赛每两年一次,竞赛委员会面向全球公布“目标序列”(相应蛋白质的结构已知但尚未公布),每个报名参赛的队伍根据自己设计的算法软件方案来对“目标序列”作出相应的预测结构,最后CASP委员评

估所有的算法。

CASP的历史就是蛋白质结构预测算法二十多年的发展史，CASP代表着蛋白质结构预测的最新进展。本研究结合CASP的发展历史和现状对蛋白质结构预测算法进行阐述，既包括经典方法（自由建模、同源建模），又包括最近在蛋白质领域取得重大成功的深度学习算法，并分析以上方法的优点和局限性。

1 蛋白质结构预测方法

1.1 同源建模预测

同源建模法基于如下的假定：相似的序列会产生相似的结构[6]，以基于一个或者多个已知结构（模板）的同源蛋白为模板，预测未知结构蛋白（靶蛋白）的三维结构[7]。这种方法依赖于现已包含80 000多个结构的PDB（Protein Data Bank）。但是，这80 000多个结构中，很多是一个蛋白的多个变种情况，这些结构中只有大概4 000个结构族[8]。

同源建模法的流程如下：（1）搜索与靶蛋白序列相关的已知蛋白三维结构。将靶蛋白序列与PDB中的所有蛋白序列进行比对，找出序列相似的结果。（2）挑选模板。将得到的相似结果作为潜在的模板，一般来说对于具有同源蛋白的序列，需要选择一个或多个合理的结果作为模板[9-12]。选择的标准绝对不能是孤立的，而应该是综合的、多维度的。

随着CASP的进行，同源建模法对于蛋白质结构预测的准确性也不断提高。在近10年内，该方法的整体准确性（从1994第一届CASP到2004年第六届CASP）的提高是非常明显的，而自CASP6以来，10年中准确性的提高相对缓慢[13]。在最近的几届比赛中，成绩最优秀的队伍具有能构建出优于PDB中最好模板的三维结构模型，此外，在多个模板的取舍问题上也有了长足的进步。预测精度提高的主要原因有：（1）更多的模板。PDB中已知三维结构的蛋白质数量明显增多。（2）更好的序列搜索和比对工具[14]使得探测蛋白质更远的进化关系以及更准确的序列比对成为可能。（3）片段拼接方法的发展[15-18]。若在蛋白结构数据库中无法搜索到相似的序列，片段拼接会发挥出意想不到的性能。片段拼接将所预测的氨基酸序列分成结构已知若干片段以及若干剩余氨基酸残基，然后将片段进行随机组合，如果组合后的结构并没有发生冲突，则会产生新的蛋白质“骨架”，基于这样的新架构再将剩余的氨基酸残基增补即可。片段拼接的思路在一定程度上扩展了以往的经典蛋白质结构预测方法。

如图1所示，预测准确性在从CASP11到CASP12的两年中发生了飞跃，蛋白质主干的预测精度的提高明显高于前10年提高的幅度[19]。除了以上所提的因素外，以下因素也起到决定性的作用：（1）模型的精修能力的提升[20]。（2）从诱饵中选择模型能力的提高。在蛋白质结构预测领域中，将通过自由建模（从头预测法）得到的构象叫诱饵。通过改进模型准确度的评估方法，可以从诸多诱饵中更好地选择模型[20]。

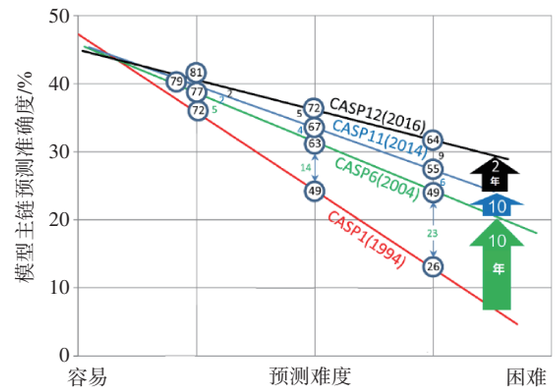

图1 CASP中模型主链预测准确度比较

Fig.1 Comparison of model backbone predication accuracies in CASP

1.2 自由建模

当PDB中没有与所预测蛋白相似的序列时，蛋白质结构的预测变成了极具挑战性的工作[21-22]。此时可利用自由建模对未知结构的序列进行预测。基于设计好的能量函数，自由建模法会通过各种结构变化来进行构象探索，之后就会产生很多的备选结构，从中决定出最佳的三维结构模型。自由建模有3个重要要素：能量函数、构象搜索、模型选择。

1.2.1 能量函数 能量函数分为两类：基于物理的能量函数和基于知识的能量函数。基于物理的能量函数，原子间的彼此作用是基于量子力学和电磁学的[23-24]。但是，由于计算量太大，在蛋白质结构预测算法中，一般使用妥协力场结合分子动力学进行结构预测。但由于计算量仍然很大，因此这类方法主要用于蛋白质结构精修，即从低分辨率的蛋白模型开始，对该模型的部分侧链主链进行精修，最后使这个模型逼近理想的天然结构。

基于知识的能量函数将各种PDB中的结构作为数据进行大规模数据分析，从而推断出一些能够描述结构的相关参量，这些参量能够帮助进行结构预测。

1.2.2 构象搜索 自由建模方法的成功与否取决于构象搜索效率。构象搜索方法是从崎岖的能量环境中搜索到具有全局能量最小的天然结构。常用的两种方法有蒙特卡罗（Monte Carlo, MC）和分子动力学

(Molecular Dynamics, MD),无论使用哪一种方法,完整地探测整个构象空间都需要耗费大量的计算资源。

MC 模拟的棘手问题主要在于蛋白结构的能量环境往往不是规则的,而是复杂且高低不平的,也具有很多的凹陷。这类方法中最流行的构象搜索方法是模拟退火[23],该方法能够在很大程度上避免局部陷阱,从而找到全局最优解。

MD模拟在原子运动的每一步骤都会计算运动的牛顿方程,该方法是描述蛋白中原子变化的最直接方法,需要耗费大量的计算资源。如果目前的结构具有低分辨率的模型,且该模型能准确描述天然构象,则该方法也可以用于精修。

1.2.3 模型选择　在大多数情况下,在自由建模的过程中会产生多个候选的结构方案,如何进行结构的挑选也是重要步骤之一。

自由建模思想远在CASP创建之前,然而第一个由自由建模方法准确地预测蛋白模型却出现在几十年以后的CASP4(2008)[18],Rosetta首先在短蛋白质序列结构预测上取得了进展[25],从那时起CASP见证了自由建模方法的预测能力的提升。虽然还没有哪个参赛队伍能够作出对每个蛋白都能进行预测的算法,但是短序列的蛋白,特别是对小于100个氨基酸的蛋白,自由建模已经有了实际性的改善。目前,将自由建模法应用于大型的蛋白复合体仍然是巨大的挑战。

但是,随着生物信息算法信息提取能力的提升,Hhblits等算法能够灵敏而准确地识别远亲同源蛋白算法,仅根据一维的氨基酸序列就可以生成一个二维的接触距离矩阵来预测残基与残基的接触[26-28]。随着计算机计算能力的增长,并且伴随域分割技术和片段组装技术的发展,自由建模技术得到飞跃式发展。以此为基础,首先对蛋白质进行域分割,在各个域内部相互接触的生物力场很强;然后再预测每个域的空间结构,预测各个域的空间结构,主要基于蛋白质主链折叠角度进行预测,进而实现域内氨基酸序列的分割和组装;最终把预测好的结构域进行组装。

在CASP11(2014)上,自由建模方法将以上能量函数构建、构象搜索、模型选择同域分割和片段分割技术融合在一起,并运用并行计算策略,取得了巨大的成功,第一次实现由自由建模方法准确地预测了长度为256的氨基酸序列的蛋白结构[29]。在CASP11和CASP12上蛋白质结构预测领域最明显的进展是预测远程氨基酸接触能力不断提升。在CASP12上,远程氨基酸之间接触预测的平均精度提升了近两倍(从27%提升至47%)。在CASP12上,共有26种方法预测能力超过CASP11最优的预测方法[30]。

随着深度学习技术的发展,自由建模方法的预测能力也得到大大提升。而在CASP13(2018)上,远程氨基酸接触预测的准确性又得到大幅度的提升,准确性高达70%,与CASP12(2016)相比,预测准确性提高了近23%[31]。

1.3 深度学习算法

在2018年CASP上,谷歌deepmind小组开发的AlphaFold取得了优异的成绩。在近几年内,深度学习算法和传统结构预测的结合一直是CASP的主题,而且此次比赛前五名都是深度学习和传统方法的融合方法。本研究以AlphaFold为例来阐述深度学习对蛋白质结构预测领域的影响。

与Rosetta自由建模中所应用的策略进行比较,CASP13(2018)AlphaFold一共提出3种预测方案:第一种方案是典型的自由建模Rosetta策略,不过在打分和片段拼接等环节运用了深度学习技术;第二种方案则是省略了自由建模方法中域内分割和拼接环节;第三种方案则对自由建模方法进一步简化,域间分割环节也被省略。后两者之所以在一定范围内取得成功,得益于深度学习技术的发展[32]。

AlphaFold相对传统的结构预测方法的优势在于其可以通过神经网络来预测蛋白质结构的两种特性:(a)氨基酸对之间的距离;(b)蛋白质主链连接氨基酸化学键之间的角度。基于神经网络,可通过两种不同方法从氨基酸的一维序列信息预测蛋白质的三维结构。

尽管谷歌公司开发的AlphaFold在成绩上远远高于其他小组,但是在结构预测的策略乃至方法上没有本质的提高,AlphaFold的优势在于计算力的运用和信息的表达和提取。策略上,蛋白质结构预测的大多数方法都是遵循了序列匹配、结构预测、打分、诱饵选取这些过程,AlphaFold也是遵循这样的策略。就策略层面,所用的方法和策略与传统蛋白质结构预测方法相比并没有本质的区别。

除此之外,AlphaFold通过神经网络构建了打分函数,以此评价所预测的模型是否准确。在AlphaFold中构建两个神经网络用来对预测出来的模型打分。第一个神经网络是内部残基距离神经网络,用来估计氨基酸对之间是否相邻,用已知的蛋白质结构信息训练神经网络,预测在一个蛋白质结构中每一个氨基酸对之间的距离分布,这些距离分布

可以用概率的形式表示，并以此对一个模型的准确度进行打分。第二个神经网络是直接分数神经网络，综合使用所有距离来估计所提出的结构与正确答案的接近程度。通过这些打分函数可以实现在蛋白质的构象空间内更为快速的搜索过程（相对于传统的蛋白质结构预测方法），更为快速和准确地找到符合目标预测的结构。

2 展 望

目前蛋白质结构预测已经取得了非常重要的进展，所预测的最为精确的结构能够充分地在分子尺度上阐明蛋白质作用的生物学机制；并且能够将其应用到药物开发等领域，从而引导生物化学研究。然而，对于大多数的蛋白质结构预测的精度远远不能达到实际应用的程度，仍然存在巨大的挑战。只有10%被预测的结构可以用来指导实际的药物合成等生物化学方面的研究[33]。

尽管PDB中已知结构的蛋白质越来越多，但是基于物理力场的预测技术，而不是依赖于已有结构的蛋白质，对我们理解蛋白质的运动和生物学机制具有重要的意义。对于一些种类的蛋白质如无序蛋白、膜蛋白以及能够折叠的聚合物在一些特定的应用领域如药物诱导结合，我们所获得的已知信息还是远远不够[33]。因此发展基于生物力场的自由建模结构预测方法很有意义，对于没有同源模板的全新的蛋白质结构预测仍然是巨大挑战。

如何确定一个预测程序的预测结果的好坏，目前并没有统一的标准。CASP用实际结构与预测结构的原子距离作为结构预测结果评分标准是不是最佳标准还有待验证[34]。近年来，对于CASP所制定用来评价结构预测评分标准的质疑也是越来越多。因为蛋白质结构并不是刚性结构，它在部分区域是允许有弹性的，用来实现蛋白质的生物学功能。因此CASP制定的标准是不是适合蛋白质动态结构还需要进一步确定。但是目前学术界并没有给出更好的标准。

目前所有的预测程序都没有在过拟合这个问题上有较好的解决办法，此外，目前的诸多方法，包括深度学习，对于不同蛋白的预测效果相差很多。如果有同源性蛋白结构，准确度会极高；若没有同源性蛋白结构，准确度就不能令人满意了[32-35]。

【参考文献】

[1] RIZZUTI B, DAGGETT V. Using simulations to provide the framework for experimental protein folding studies [J]. Arch Biochem Biophys, 2013, 531(1-2): 128.

[2] ANFINSEN C B, HABER E, SELA M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain [C]//Proceedings of the National Academy of Sciences of the United States of America, 1961: 1309-1314.

[3] SCHRÖDER M, KAUFMAN R J. The mammalian unfolded protein response[J]. Annu Rev Biochem, 2005, 74(74): 739.

[4] MOULT J.A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction[J]. Curr Opin Struct Biol, 2005, 15(3): 285-289.

[5] MONASTYRSKYY B, FIDELIS K, MOULT J, et al. Evaluation of disorder predictions in CASP9[J]. Proteins, 2011, 79(Suppl 10): 107-118.

[6] ZHENG W M. Proteins: from sequence to structure[J]. Chinese Phys B, 2014, 23(7): 107-113.

[7] ESWAR N, WEBB B, MARTI-RENOM M A, et al. Comparative protein structure modeling using modeller [J]. Curr Protoc Bioinformatics, 2014, 47(47): 47.

[8] MURZIN A G, BRENNER S E, HUBBARD T J, et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures[J]. J Mol Biol, 1995, 247(4): 536-540.

[9] FERNANDEZ-FUENTES N, MADRID-ALISTE C J, RAI B K, et al. MT: a comparative protein structure modeling server [J]. Nucleic Acids Res, 2007, 35: W363-W368.

[10] FERNANDEZFUENTES N, RAI B K, MADRIDALISTE C J, et al. Comparative protein structure modeling by combining multiple templates and optimizing sequenceto-structure alignments [J]. Bioinformatics, 2007, 23(19): 2558-2565.

[11] SANCHEZ R, SALI A. Evaluation of comparative protein structure modeling by MODELLER-3[J]. Proteins, 1997, 29(Suppl 1): 50-58.

[12] VENCLOVAS C, MARGELEVICIUS M. Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment[J]. Proteins, 2005, 61 (Suppl7): 99.

[13] KRYSHTAFOVYCH A, FIDELIS K, MOULT J. CASP9 results compared to those of previous casp experiments[J]. Proteins, 2011, 79(Suppl10): 196-207.

[14] ALTSCHUL S F. Gapped BLAST and PSI-BLAST: a new generation of protein detabase search programs[J]. Nucleic Acids Res, 1997, 25: 2389-3402.

[15] JONES D T. Predicting novel protein folds by using FRAGFOLD[J]. Proteins, 2001, 45(Suppl 5): 127-132.

[16] JONES D T, WARD J J. Prediction of disordered regions in proteins from position specific score matrices[J]. Proteins, 2003, 53(Suppl 6): 573-578.

[17] SIMONS K T, BONNEAU R, RUCZINSKI I, et al. Ab initio protein structure prediction of CASP III targets using ROSETTA[J]. Proteins, 1999, 37(Suppl 3): 171-176.

[18] BONNEAU R, TSAI J, RUCZINSKI I, et al. Rosetta in CASP4: progress in ab initio protein structure prediction[J]. Proteins, 2001, 45(Suppl 5): 119-126.

[19] KRYSHTAFOVYCH A, MONASTYRSKYY B, FIDELIS K, et al. Evaluation of the template-based modeling in CASP12[J]. Proteins, 2018, 86(Suppl 1): 321-334.

[20] READ R J, SAMMITO M D, KRYSHTAFOVYCH A, et al. Evaluation of model refinement in CASP13 [J]. Proteins, 2019. DOI: 10.1002/prot.25794.

[21] PILLARDY J, CZAPLEWSKI C, LIWO A, et al. Recent improvements in prediction of protein structure by global optimization of a potential energy function[J]. Proc Nat Acad Sci USA, 2001, 98(5): 2329-33.

[22] KALYAANAMOORTHY S, CHEN Y P. Modelling and enhanced molecular dynamics to steer structure-based drug discovery[J]. Prog Biophys Mol Biol, 2014, 114(3): 123-136.
[23] CORNELL W D, CIEPLAK P, BAYLY C I, et al. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules[J]. J Am Chem Soc, 2015, 117(117): 5179-5197.
[24] JORGENSEN W L, MAXWELL D S, TIRADO-RIVES J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids[J]. J Am Chem Soc, 2017, 118(45): 11225- 11236.
[25] ROHL C A, STRAUSS C E, MISURA K M, et al. Protein structure prediction using Rosetta[J]. Method Enzymol, 2003, 383(383): 66.
[26] MORCOS F, PAGNANI A, LUNT B, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families[J]. Proc Nat Acad Sci USA, 2011, 108(49): E1293-E1301.
[27] MARKS D S, COLWELL L J, SHERIDAN R, et al. Protein 3D structure computed from evolutionary sequence variation[J]. PLoS One, 2011, 6(12): e28766.
[28] JONES D T, BUCHAN D W, COZZETTO D, et al. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments[J]. Bioinformatics, 2011, 28(2): 184-190.
[29] KRYSHTAFOVYCH A, MONASTYRSKYY B, FIDELIS K. CASP11 statistics and the prediction center evaluation system[J]. Proteins, 2016, 84(Suppl 1): 15-19.
[30] SCHAARSCHMIDT J, MONASTYRSKYY B, KRYSHTAFOVYCH A, et al. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age[J]. Proteins, 2017, 86(Suppl 1): 51-66.
[31] HOU J, WU T, CAO R, et al. Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13[J]. Protein, 2019. DOI: 10.1101/552422.
[32] ALQURAISHI M. AlphaFold at CASP13[J]. Bioinformatics, 2019, 35(22): 4862-4865.
[33] DILL K A, MACCALLUM J L. The protein-folding problem, 50 years on[J]. Science, 2012, 338(6110): 1042-1046.
[34] SINGH A, KAUSHIK R, JAYARAM B. Quality assessment of protein tertiary structures: past, present, and future[M]//Bioinformatics: Sequences, Structures, Phylogeny. Springer Nature Singapore Pte Ltd, 2018: 271-288.
[35] DRORI I, THAKER D, SRIVATSA A, et al. Accurate protein structure prediction by embeddings and deep learning representations[J]. arXiv Preprint.arXiv: 1911.05531, 2019.

（编辑：谭斯允）