Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

News & Highlights

人工智能破解生物学 50 年来的一个重大挑战

Sean O'Neill

Senior Technology Writer

2020年11月下旬，谷歌母公司Alphabet旗下子公司DeepMind Technologies（总部位于伦敦，专注于研究人工智能）宣布其AlphaFold系统在仅凭基因序列预测蛋白质的复杂形状方面已达到“无与伦比的精准度”（unparalleled levels of accuracy）[1]。这一壮举遇到生物学50年来的一个重大挑战，即预测蛋白质如何折叠。该挑战的成功破解预计会对药物研发以及蛋白质设计的新兴领域产生重大影响，甚至可能有助于我们应对新冠病毒肺炎疫情[2]，特别是如今迅速出现的多种严重急性呼吸综合征冠状病毒2（SARS-CoV-2）变异株[3]。

DeepMind创始人兼时任首席执行官Demis Hassabis表示：“蛋白质折叠是生物学领域中的一个圣杯问题。我们一直推测人工智能应有助于更快实现这些重大科学突破。”

蛋白质是复杂的大分子，在生物界的各个方面都起着关键作用。蛋白质形状决定了其功能：血红蛋白运输营养物质，酶催化化学反应，胶原蛋白提供结构，胰岛素调节血糖，抗体提供免疫力。这些蛋白质以及其他所有蛋白质均由标准遗传密码中同一组20种氨基酸以长链相连的方式组成。

蛋白质是由生物体或合成过程所产生的氨基酸构成，自然扭曲并折叠在一起，形成复杂形状，呈弯曲结构、螺旋结构和折叠结构。例如，抗体蛋白质为“Y”形，这使其能够锁定且有助于中和引起疾病的细菌或病毒。相反，有害基因突变会导致产生错误折叠的非功能性蛋白质，如囊性纤维化的蛋白质。

产生蛋白质的密码包含在脱氧核糖核酸（DNA）内。不过，尽管DNA测序揭示了给定蛋白质所包含的氨基酸序列，但是并不能说明它们如何折叠成最终形状。蛋白质序列越大，就越难预测其形状。理论上，典型蛋白质分子链可折叠成的构象是一个天文数字，因此使用蛮力去预测其形状几乎是不可能的[4]。

蛋白质折叠问题始于1972年，当时，获得诺贝尔化学奖的美国生物化学家Christian Anfinsen宣称蛋白质氨基酸序列应足以确定其在特定环境中的折叠形状[5]。然而，几十年来，准确确定靶蛋白形状的方法只有核磁共振和X射线晶体分析，以及最近的冷冻电子显微镜等技术，但是这些方法往往价格高昂且费时。此类实验工作可能需要数年时间才能描绘出单个蛋白质的形状，而且无法保证成功。

1994年，为聚集全球科学家共同解决此问题，美国马里兰大学细胞生物学与分子遗传学教授John Moult及其同事开展了一项大型实验，旨在评估生成蛋白质结构的计算方法[6]。这项工作成为两年一次的蛋白质结构预测关键评估（Critical Assessment of Structure Prediction, CASP）活动，Hassabis称之为“蛋白质折叠领域的奥林匹克竞赛”。

CASP竞赛分为三个滚动阶段：①收集约100个靶蛋白，近期实验室工作已揭露其形状，但至关重要的是，尚未发布成果；②向世界各地的研究团队提供这些靶蛋

2095-8099/© 2021 THE AUTHORS. Published by Elsevier LTD on behalf of the Chinese Academy of Engineering and Higher Education Press Limited Company. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
英文原文: Engineering 2021, 7(6): 706–708
引用本文: Sean O'Neill. Artificial Intelligence Cracks a 50-Year-Old Grand Challenge in Biology. *Engineering*, https://doi.org/10.1016/j.eng.2021.04.003

白的基因序列，然后使用软件系统开展工作以预测其形状；③对提交的预测形状进行盲审。CASP主要使用称为“全局距离测试”（global distance test, GDT）的度量标准（范围介于0~100）来判断预测形状的精准度。Moult表示，GDT分数在90分左右，即可视为与人类通过实验方法获取的结果相当。

自1994年以来，研究进展一直稳定但缓慢，直到2018年第13届CASP竞赛，DeepMind团队首次参赛并提出早期版本的AlphaFold系统 [7]。该团队以相当大的优势获胜，在CASP竞赛中一鸣惊人，但AlphaFold系统预测的形状仍与靶蛋白的实际结构相去甚远，其GDT中位数评分为59分（图1）。

然而，在2020年第14届CASP竞赛中，DeepMind团队带来了经过全面改进的AlphaFold系统，这次结果惊人。Moult表示：“简直不可思议。当你看到一个令人惊讶的预测时，你会想，‘这是怎么回事？’。当你拥有三个或四个令人难以置信的准确的结构预测时，你就会意识到发生了非常重要的事情。”

AlphaFold系统在最困难类别中获得的GDT评分为87分，在所有靶蛋白中的GDT中位数评分为92.4分（图2）[8]。该系统的平均误差约为0.16 nm——大约为一个原子的宽度。为解决这一问题，DeepMind团队开发了一种新型的基于注意力的神经网络系统[9]。在机器学习中，注意力系指模仿人类注意力的设计，即系统识别出数据的关键方面并赋予这些方面更多权重，而对那些它认为不太重要的数据很少关注。有关该深度学习系统的具体技术细节尚待分享，但预计2021年晚些时候会对相关论文进行同行评议。AlphaFold系统（图3）[1]已通过使用蛋白质数据库（PDB）的公开数据进行了训练，该数据库包含大约175 000种蛋白质结构，此外还有包含未知结构蛋白质序列的其他大型数据库。根据DeepMind团队的说法，训练期需要大约16台谷歌TPUv3协处理器（相当于100~200个图形处理器）运行“数周”，

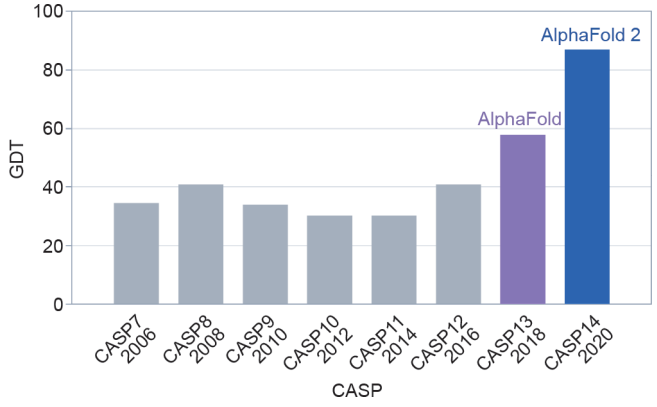

图1. 在两年一度的CASP竞赛中，获胜团队在最困难类别（自由建模类别）中使用GDT预测的中位数精准度。DeepMind团队的AlphaFold系统在2018年和2020年竞赛中均排名第一。图片来源：DeepMind，经许可。

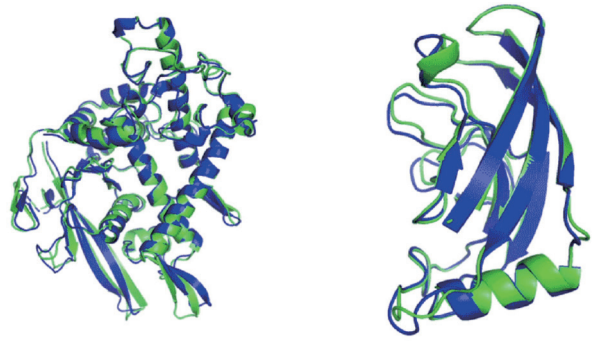

图2. AlphaFold系统在第14届CASP竞赛中预测的几种蛋白质结构（蓝色）与由实验确定的结构（绿色）相重叠。两种预测结果高度匹配。RNA：核糖核酸。图片来源：DeepMind，经许可。

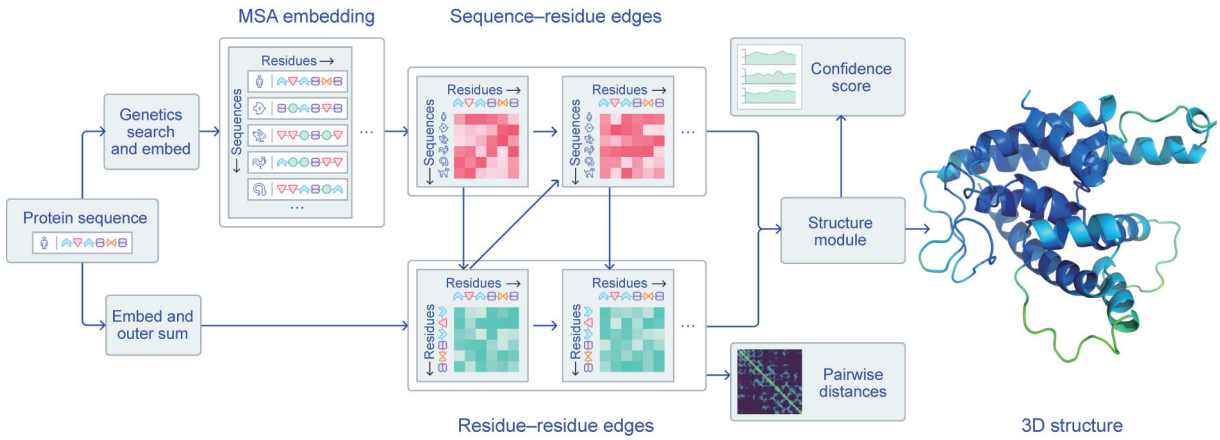

图3. AlphaFold结构概述。DeepMind团队尚未提供其系统的相关细节，但描述了“折叠蛋白质如何被认为是‘空间图’的问题，其中氨基酸残基为节点，并且边缘将残基紧密相连”[1]。MSA：多序列比对；3D：三维。图片来源：DeepMind，经许可。

单个蛋白质结构预计“在几天内”即可完成[1]。

Moult曾听说，神经网络是被美化的模式识别，然而他表示：“AlphaFold系统能够从其训练中获得原子级认知的水平是惊人的。其达到的抽象层次意义深远。仿佛这台机器已经学会了物理学。在任何涉及蛋白质结构的情况中，其可在原子层面得到正确结果。然而，仅通过识别训练数据中的一组模式无法实现这一点。”

该项突破为整个生物学领域带来了机遇，但其最直接的影响可能是药物发现。大多数药物通过与体内蛋白质相结合而起效，从而触发其功能变化。采用诸如AlphaFold这样的机器学习系统，能够迅速算出靶蛋白的形状，然后设计药物（或重新利用现有药物）以有效结合这些蛋白质。

例如，随着2020年年初新冠病毒肺炎疫情规模扩大，以及后来在第14届CASP竞赛中，DeepMind团队提取了构成SARS-CoV-2的几种蛋白质的基因序列，并提供了结构预测，这些预测后来基本都通过实验得到证实[10]。此类工作有可能加快可阻遏这种疾病的药物设计。实际上，蛋白质设计是形状预测的另一方面：一旦机器对支撑蛋白质折叠的原子过程具有深刻了解，那么设计能够折叠成所需形状的蛋白质就变得更加容易。

美国华盛顿大学的蛋白质设计研究所所长David Baker表示：“我们一直使用现有蛋白质设计方法来开发看起来非常具有前景且已进行或即将进行临床试验的新冠病毒肺炎治疗剂、疫苗和检测装置。通过改进的蛋白质设计，我们应该能够做得更好、更快。”David Baker领导的团队在第14届CASP竞赛上名次仅次于DeepMind团队[11]。

诸如AlphaFold系统之类的技术还可用于探索分解工业废物或旧塑料的蛋白质和酶，如有效吸收大气中的碳。马里兰大学生物化学教授及第14届CASP竞赛的蛋白质结构贡献者Osnat Herzberg表示：“对结构生物学领域的直接影响是巨大的。这些方法会产生重要医学应用，并带来我们目前无法想象的技术进步。”

伦敦大学学院生物信息学教授兼生物信息学团队负责人David Jones的看法则更为谨慎。Jones表示：“这样的结果使人们意识到，机器学习可在机器视觉和自然语言处理的领域之外产生巨大影响。但我并不相信仅仅因为我们现在可以比以往任何时候能更精确地对蛋白质结构进行建模，我们就会有新的疾病治疗方法。重要的是，在能够确定其能力或局限性之前，我们需要在许多不同条件下对诸如这样复杂的系统进行测试。”

References

[1] AlphaFold: a solution to a 50-year-old grand challenge in biology [Internet]. London: DeepMind; 2020 Nov 30 [cited 2021 Feb 4]. Available from: https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology.

[2] Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature 2020;588:203–4.

[3] About variants of the virus that causes COVID-19 [Internet]. Atlanta: Centers for Disease Control and Prevention; [updated 2021 Feb 12; cited 2021 Feb 26]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/transmission/variant.html.

[4] Dill KA, Ozkan SB, Shell MS, Weikl TR. The protein folding problem. Ann Rev Biophys 2008;37:289–316.

[5] Protein folding and the thermodynamic hypothesis, 1950–1962. [Internet]. Washington, DC: US National Library of Medicine; [cited 2021 Feb 18]. Available from: https://profiles.nlm.nih.gov/spotlight/kk/feature/protein.

[6] Moult J, Pedersen JT, Judson R, Fidelis KA. A large-scale experiment to assess protein structure prediction methods. Proteins Struct Funct Bioinform 1995;23(3):ii–v.

[7] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577:706–10.

[8] 14th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction [Internet]. Davis: University of California, Davis; c2007–2020 [cited 2021 Feb 18]. Available from: https://predictioncenter.org/casp14/results.cgi?groups_id=205&submit=Submit.

[9] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Tunyasuvunakool K, et al. High accuracy protein structure prediction using deep learning. In: Critical Assessment of Techniques for Protein Structure Prediction (CASP14), abstract book; 2020 Nov 30–Dec 4; online conference. 2020.

[10] Computational predictions of protein structures associated with COVID-19 [Internet]. London: DeepMind; 2020 Aug 4 [cited 2021 Feb 18]. Available from: https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.

[11] Cao L, Goreshnik I, Coventry B, Case JB, Miller L, Kozodoy L, et al. *De novo* design of picomolar SARS-CoV-2 miniprotein inhibitors. Science 2020;370(6515):426–31.