上海科技报
(/index.html)
www.shkjb.com

可精准预测蛋白质结构的人工智能模型AlphaFold2开源

《自然》曾称"它将改变一切！"

2021年08月04日 版面：A3

作者：陈怡

继去年在第14届"蛋白质结构预测关键评估大赛（CASP14）"中揭晓了其新开发的人工智能模型AlphaFold，并承诺会分享该方法，为科学共同体提供广泛、免费的获取途径之后，位于英国伦敦、由人工智能程序师兼神经科学家Demis Hassabis等人联合创立的人工智能前沿企业——DeepMind公司的研究人员近日在《自然》上发表论文，将AlphaFold2开源。AlphaFold2能以就计算机方法而言前所未有的准确度，根据蛋白质的氨基酸序列预测其三维结构。这意味着，普通研究人员曾需要花费几年时间才能破解的蛋白质结构，现在用AlphaFold2只要几个小时就能解析出来。这被认为破解了由克里斯蒂安·安芬森于1972年提出、困扰了学界长达50年之久的"蛋白质折叠"难题。

"意义超过人类基因组测序"

蛋白质是由氨基酸链组成的，折叠成三维结构的氨基酸链决定了细胞内蛋白质的功能。确定蛋白质的结构有助于确定蛋白质的功能，了解各种突变的作用，为理解生物学过程提供宝贵信息，并有望指导药物研发。数十年时间里，研究人员一直在用X射线晶体学和冷冻电镜这类实验技术解析蛋白质结构。但是，这类方法存在费时耗钱的问题，对一些蛋白也不适用。截至目前，约有10万个蛋白质的结构已由实验方法得到了解析，但这在已经测序的数以10亿计的蛋白质中只占了很小一部分。在50多年的时间里，研究人员一直尝试根据蛋白质的氨基酸序列预测其折叠而成的三维结构。然而，当前使用的计算方法准确度有限，实验方法对人力和时间的要求也非常高。

DeepMind的研究人员John Jumper、Demis Hassabis和同事在论文中描述了基于神经网络的新模型AlphaFold2，其预测的蛋白质结构能达到原子水平的精度。去年5—7月举办的CASP14大赛要求参赛团队根据蛋白质的氨基酸序列解析它们的结构。AlphaFold在大赛中预测的准确性达40分左右（满分为100）。去年12月，DeepMind介绍了Alphafold2，将这一准确性直接拔高到了92.4/100，和蛋白质真实结构之间只差一个原子的宽度，真正解决了蛋白质折叠的问题。AlphaFold2的神经网络能在几分钟内预测出一个典型蛋白质的结构，也能预测较大蛋白质（比如

一个含有2180个氨基酸、无同源结构的蛋白质）的结构，还能根据每个氨基酸对其预测可靠性进行精确预估，方便研究人员使用其预测结果。使用Alphafold2预测，准确性与实验方法不相上下，且远超解析新蛋白质结构的其他方法。

研究人员强调，这是一个完全不同于AlphaFold的新模型。AlphaFold使用的神经网络是类似ResNet的残差卷积网络， AlphaFold2则借鉴了AI研究中最近新兴起的Transformer架构。

Alphafold2于2020年入选“Science年度十大突破”，被称作结构生物学“革命性的突破”“蛋白质研究领域的里程碑”。它的出现，可以让蛋白质结构解析技术跟上基因组革命的发展步伐，使蛋白质与分子结合的概率能得到更好的预判，从而极大地加速新药研发的效率。《自然》杂志在去年11月底发表的文章中称：“它将改变一切！”但这也使当时的生物学界哀嚎一片。曾带领团队开发出了RoseTTaFold的华盛顿大学生物化学家David Baker说：“如果有人解决了你正在研究的问题，但不肯告诉你解决方法，你还怎么研究下去呢？”结构生物学家Petr Leiman则感叹：“我用价值1000万美元的电镜努力地解了好几年，Alphafold2竟然一下就算出来了。”

如今Alphafold2的开源，使生物学界和AI界再次沸腾。有研究人员称：“这是生命科学史上最重要的事件之一，比奥运会更激动人心，其意义超过人类基因组测序。”

大规模的准确结构预测将成为一种重要工具

也是在近期，《自然》还发表了一篇由Kathryn Tunyasuvunakool、John Jumper和Demis Hassabis为通信作者的论文。文章描述了AlphaFold对人类蛋白质组（人类基因组编码的所有蛋白质的集合）的准确结构预测，由此得到的数据集涵盖了人类蛋白质组近60%氨基酸的结构位置预测，且预测结果具有可信度。据悉，预测信息将通过欧洲生物信息研究所托管的公用数据库免费向公众开放。

考虑到理解人类蛋白质组对健康和医药的重要性，科研人员曾付出大量努力来确定这些蛋白质的结构。虽然开展了数十年的攻关研究，但通过实验方法确定的结构只覆盖了人类蛋白质组17%的氨基酸。氨基酸是连接起来形成蛋白质的亚单位。利用实验方法解析蛋白质结构需要跨越诸多十分耗时的障碍，因此，扩大蛋白质组覆盖面仍是一项艰巨挑战。

位于伦敦的Kathryn Tunyasuvunakool、John Jumper、Demis Hassabis和同事利用前沿机器学习方法AlphaFold确定了覆盖几乎整个人类蛋白质组（所有人类蛋白中的98.5%）的蛋白质结构。他们发现，AlphaFold能对人类蛋白质组58%氨基酸的结构位置给出可信预测。其中，对35.7%的结构位置的预测达到了很高的置信度，是实验方法覆盖的结构数量的两倍。在蛋白水平上，AlphaFold对43.8%的蛋白的至少3/4的氨基酸序列给出了可信预测。

论文作者认为，大规模的准确结构预测将成为一种重要工具，让人们能从结构的角度解答新的科学问题。

编辑：chunchun 审核：刘纯