AlphaFold 震撼发布 98.5% 的人类蛋白结构预测结果，有哪些重大突破？它们将带来哪些应用？ - 知乎

https://www.zhihu.com/question/474094187/answer/2014201529

孟凡康合成生物学话题下的优秀答主 Sat Jul, 24 22:44

人工智能在生命科学领域的一个新的里程碑！作为合成生物学家十分激动！但是也为结构生物学家捏一把汗（当然不必悲观，具体可以看我整理的一篇文章：孟凡康：AlphaFold能否革命药物设计领域？什么是大的问题？）。

在文章发表之后，我马上写了一篇解读文章，在此分享给大家！

《AlphaFold2 高精度破译几乎所有人类蛋白结构，人工智能驱动的生物学研究时代呈现无限潜力》

关键内容/Highlights

○ 整个人类蛋白质组（98.5%的人类蛋白质）被AlphaFold破译，极大地扩展了蛋白结构覆盖率。

○ 由此产生的数据集包含了58%的残基具有较高置信度，其中一个子集（占所有残基的36%）具有非常高的置信度。

○ 蛋白结构的准确预测带来了高质量的生物学假设，将进一步的激发基础科学、药物研发、合成生物学设计方面的未来发展。

○ Deepmind将通过一个公共数据库（由欧洲生物信息学研究所托管，网址：https://alphafold.ebi.ac.uk/）向社会免费提供所有AlphaFold2的蛋白质预测结果。

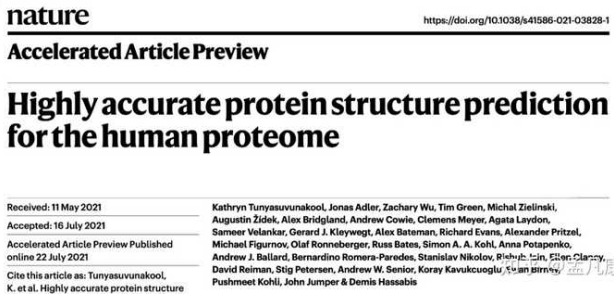
01 蛋白质组的全面结构覆盖仍然是一个突出且巨大的挑战，但蛋白质结构预测可以提供高效解决方案

在全世界科研机构的共同努力下，现在已经有超过50000个人类蛋白质结构被解析，使智人成为迄今为止在蛋白质数据库（PDB）中最具有代表性的物种。

即使如此，仍然只有35%的人类蛋白质被登记到PDB数据库中，而且在许多情况下，结构只包括序列的一个片段。实验性结构测定需要克服许多耗时的障碍：必须生产足够数量的蛋白质、进行纯化、选择适当的样品制备条件并收集高质量的数据。而不同的制备方法、蛋白质的大小、跨膜区域的存在、无序结构的存在或对构象变化的敏感性等进一步的限制结构的解析过程。**因此，蛋白质组的全面结构覆盖仍然是一个突出且巨大的挑战。**

蛋白质结构预测通过快速和大规模地提供可信赖的蛋白质结构，有助于解决上述提到的困境。近年来，结构预测取得了实质性的进展，两年一度的蛋白质结构预测关键评估（CASP）的结果证明了这一点。特别是AlphaFold的最新版本以'AlphaFold2'的团队名称参加了CASP14。AlphaFold2使用了与DeepMind在CASP13参赛时完全不同的模型，并且在提供常规的高精确度方面比以前的方法有了很大的改进，宣称解决「解决生物学50年内最大挑战」。AlphaFold2近期已经正式开源，相关文章以「Highly accurate protein structure prediction with AlphaFold」发表在《自然》期刊。

为了进一步发挥AlphaFold2的潜力，Deepmind决定将AlphaFold2应用于人类蛋白质组的解析上。2021年7月22日，相关的论文以Highly accurate protein structure prediction for the human proteome为题发表在《自然》期刊上。该工作**利用AlphaFold2破译整个人类蛋白质组结构（98.5%的人类蛋白质），极大地扩展了蛋白结构覆盖率。**同时Deepmind将通过一个公共数据库（由欧洲生物信息学研究所托管，网址：https://alphafold.ebi.ac.uk/）向社会免费提供所有的AlphaFold2蛋白质预测结果（**其中的数据不局限于人类蛋白组，同时也包含部分大肠杆菌、酵母、拟南芥、玉米等在内的超过20中物种的蛋白质结构预测结果**）。。



nature https://doi.org/10.1038/s41586-021-03828-1

Accelerated Article Preview

Highly accurate protein structure prediction for the human proteome

Received: 11 May 2021

Accepted: 16 July 2021

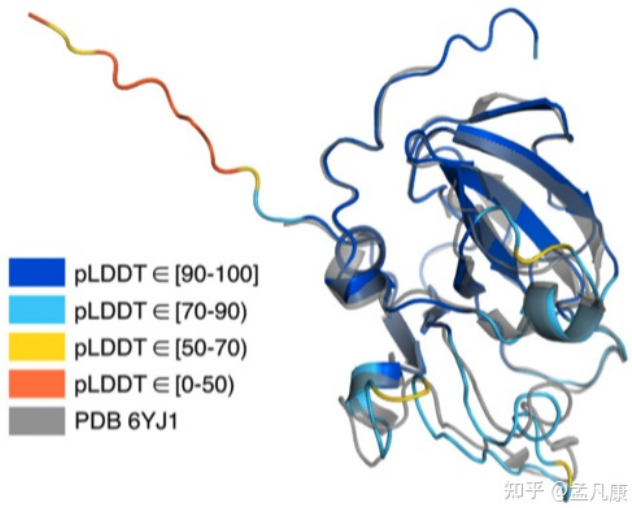
Accelerated Article Preview Published online 22 July 2021

Cite this article as: Tunyasuvunakool, K. et al. Highly accurate protein structure

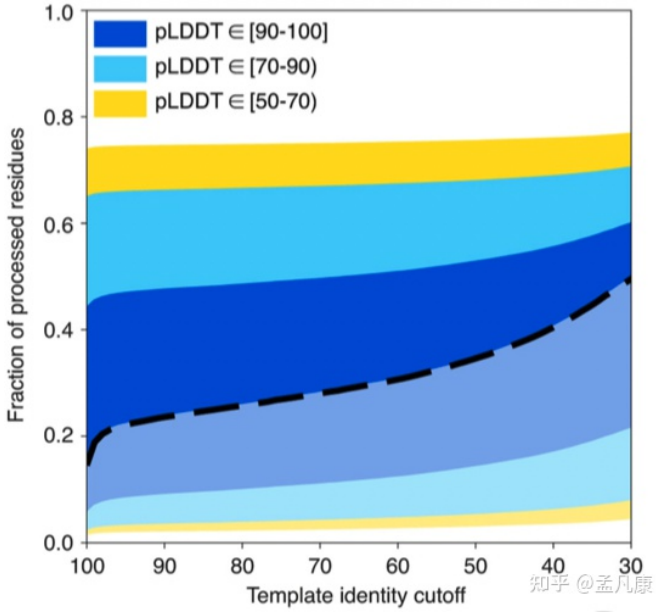
Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper & Demis Hassabis

02 AlphaFold2模型提高了蛋白结构预测的置信度和覆盖率

AlphaFold2用一个范围是0到100的指标pLDDT来衡量单个残基的置信度：将pLDDT>90作为高准确度的分界点，pLDDT>70的较低临界值对应于一个普遍正确的骨架预测。下图显示了AlphaFold2在不同pLDDT范围内对一个示例蛋白质的准确性情况。

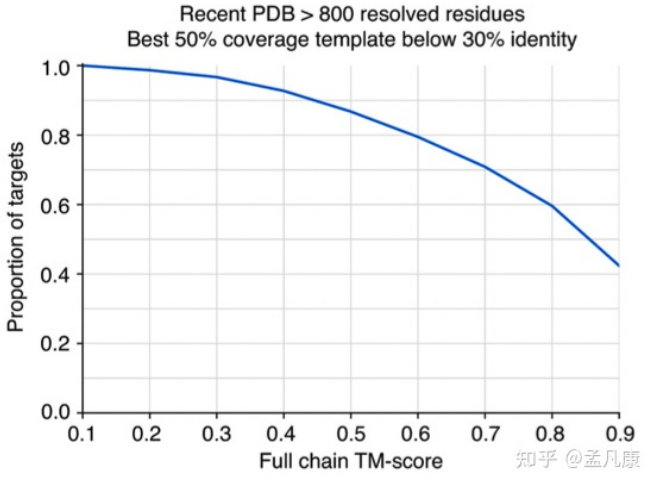


从AlphaFold2的预测结构来看（下图），在人类蛋白质组中，有35.7%的残基落在最高精度带内（相当于38.6%的残基产生了可信的预测结果）。这是现有通过实验所得结构数量的两倍。58.0%的残基被有把握地预测（pLDDT>70），这意味着AlphaFold2也为PDB中没有良好结构的序列增加了大量的覆盖率（原有的结构解析度低于30%）。对于单个蛋白质的预测来说，43.8%的蛋白质中有至少四分之三的序列被有把握地预测。

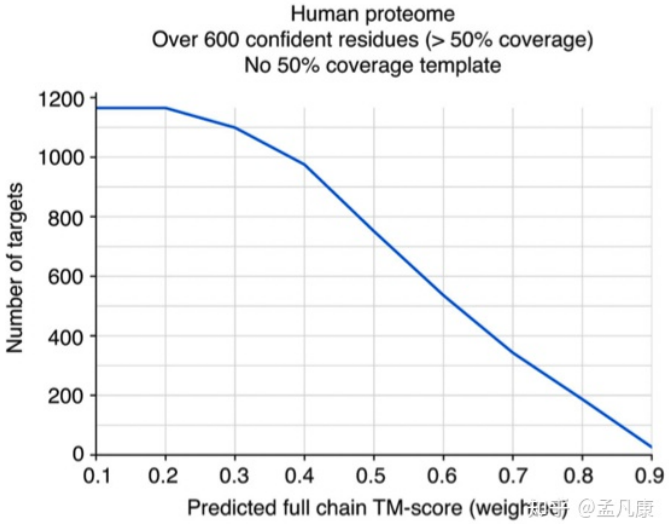


03 AlphaFold2模型在多结构域复合体的预测上同样表现优异

以前的许多大规模结构预测工作都集中具有独立折叠能力的单结构域上。但这会带来一些问题：**1. 将预测局限于预先确定的结构域有可能遗漏尚未被注释的结构区域；2. 它还抛弃了来自序列其他部分的上下文信息，而这些信息在两个或更多的结构域发生实质性相互作用时可能扮演关键的角色。**所以，Deepmind团队尝试利用AlphaFold2对多结构域复合体进行结构预测。为了进一步评估AlphaFold在长的多结构域蛋白上的表现，研究团队编制了一个测试集，只包括大于800个解析残基的蛋白结构。随后研究团队使用新的评估参数—模板建模得分（TM-score）对该测试集的性能进行了评估。TM-score应能更好地反映全局而不是每个单独结构域的准确性。结果表明70%的预测的TM分数大于0.7（下图）。



紧接着团队计算了人类蛋白质组的TM分数，测试集的蛋白结构实验解析覆盖率均小于一半，序列长度至少600个残基。结果中有187个蛋白质的TM>0.8，343个蛋白质的TM>0.7。虽然预计AlphaFold的域间准确度会低于其域内准确度，但这组数据表明AlphaFold2模型在多结构域复合体的预测上同样表现优异

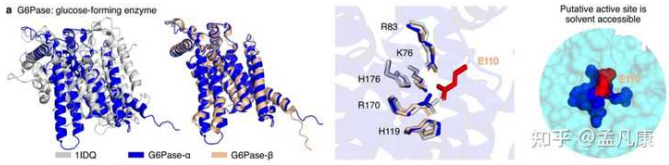


04 蛋白结构的准确预测带来了高质量的生物学假设

蛋白结构的准确预测能够为生物学研究带来丰富的启示。研究团队在文章中提到了三个不同案例。所有的结构预测都是从头开始的，对象均具有少于25%序列同源性或结构解析覆盖少于20%。

○ 葡萄糖-6-磷酸酶

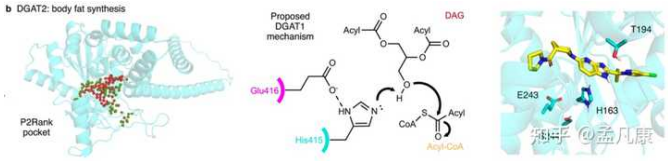
葡萄糖-6-磷酸酶是一种膜结合蛋白，催化葡萄糖合成的最后一步，因此它对维持血糖水平至关重要。但是此蛋白目前还没有实验结构。**AlphaFold2的预测具有很高的置信度（中位数pLDDT为95.5），并给出了一个九螺旋的拓扑结构。**从预测的结构来看，在葡萄糖-6-磷酸酶的结合口袋附近存在一个保守的谷氨酸（Glu110）。谷氨酸可以将结合口袋稳定在一个封闭的构象中，与其他残基形成盐桥。该位点也是推定的活性位点中溶剂暴露最多的残基，表明可能具有门控功能，但是这个残基以前从未被讨论过。此案例说明了可以从高质量的结构预测中获得新的机制性假设。



○ 二酰基甘油 O-酰转移酶2

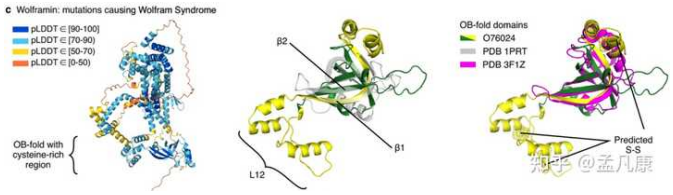
三酰甘油的合成将多余的代谢能量作为脂肪储存在脂肪组织中。DGAT2是两个重要的酰基转移酶之一，在这个途径中催化最后的酰基添加。抑制DGAT2已被证明可以改善肝病小鼠模型的肝功能。

AlphaFold2的预测具有很高的置信度（中位数pLDDT为95.9）。研究团队随后根据预测结构确定了一个已知的抑制剂的结合口袋：分子模拟表明此口袋能够对接抑制剂，并观察了到特定的相互作用。在DGAT2的结合口袋中，研究团队确定了两个对口袋结构关键的残基（Glu243, His163）。以前对DGAT2的实验工作表明，突变His163有更强的负面影响。此外，Glu243和His163在不同的物种中也是保守的，进一步支持了AlphaFold2预测结构的可靠性。



○ Wolframin

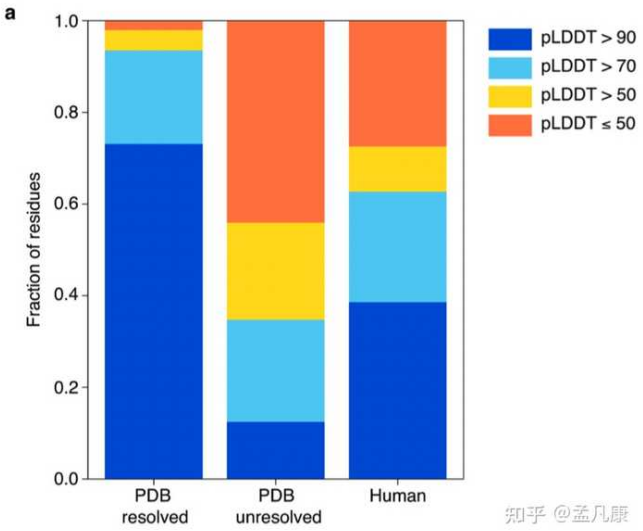
Wolframin是一种定位在内质网中的跨膜蛋白。WFS1基因的突变与Wolfram综合征-1有关。Wolfram综合征-1是一种神经退行性疾病，其特征是早发的糖尿病，逐渐的视觉和听觉丧失，以及早期死亡。**虽然AlphaFold2对Wolframin的预测置信度较低（中位数pLDDT为81.7），但是预测得到的结构信息与之前的理论分析高度吻合。**通过结构分析，研究团队推测一个富含半胱氨酸的结构域可以用来招募其他蛋白，所以此部分的结构信息将为未来研究其招募的蛋白提供重要的参考。



05 低置信度的预测结果可能代表着蛋白结构的无序状态

无序结构在真核生物的蛋白质组中很常见。之前的一项工作估计人类蛋白质组中无序残基的比例为37-50%。因此，当AlphaFold2对整个蛋白质组的全面预测时，我们应该期望有相当比例的残基在溶液中总是或有时处于无序的区域。此处的无序包括本质无序的结构域和仅在复合体时

具有稳定折叠结构的结构域。此外，研究团队还观察到PDB序列中已解决和未解决的残基之间的pLDDT分布有很大差异（下图）。同时研究团队发现pLDDT是非常好的蛋白无序状态预测器。**无序预测结果表明，相当大比例的低置信度残基可能是由某种形式的无序状态导致的。**



总之，研究团队目前对AlphaFold2对部分结构域表现出低置信度解释是：**这些结构域有很大可能是孤立条件下的无序化。目前，AlphaFold2把pLDDT<50的长结构域呈现为带状外观，应该被解释为对无序状态的预测，不应该被解释为结构信息。**

06 人工智能驱动生物学研究时代呈现无限潜力

在这项工作中，Deepmind团队利用AlphaFold2对人类蛋白质组进行了最全面的结构预测。由此产生的数据集为蛋白质组的结构覆盖度提升做出了巨大的贡献；**通过提供可扩展的结构预测和前所未有的准确度，AlphaFold在结构生物信息学上进展令人震撼，而这将进一步极大的拓展生物蛋白质的可研究空间。**

当然，未来的Alphafold仍有关键问题需要解决。

○**人类蛋白质组中仍然没有可靠预测的部分代表了未来研究的方向。**其中有一部分的预测是失败的，即存在一个固定的结构，但当前版本的AlphaFold并不能预测它。在其他许多情况下，序列是孤立的、非结构化的。要**解析这些结构域的话，开发基于生物学原理的新预测方法至关重要，例如预测该结构在复合体中的折叠方式或预测复杂细胞环境中可能的折叠状态的分布。**

○ 同时我们也应当意识到，**Alphafold只是提供了强大的工具，但是对生物学问题的解决需要全世界科学家在不同领域的不断探索。**各领域包括蛋白结构、药物设计、合成生物学元件开发或者蛋白质设计领域的复杂性不仅仅是蛋白质结构预测问题，还有很多更加复杂，需要我们正视的难题。

但是，**人工智能驱动的生物学研究时代呈现的无限潜力是如此的激动人心，更多的生物蛋白质组也值得去探索和挖掘。**人类蛋白质组因为其在健康和医学上重要性而被深入研究，但是其他生物体在现有的结构数据则相差甚远，包括很多具有生物学意义、医学意义或经济意义的物种。精确的结构信息可能会开辟出全新的研究途径，对这些生物的研究产生更深远的影响。**同时精确的结构信息也将对合成生物学领域中的蛋白质元件设计、定量生物分析，定向进化等等提供强大的分析工具，进一步提升人类对于生物系统的工程改造能力。**

\ END \