北大与华为云联合发布蛋白质多序列比对开源数据集

时间：2021-09-15 11:19:00 来源： 作者： 访问量：

近日，北京大学化学与分子工程学院、北京大学生物医学前沿创新中心（BIOPIC）、深圳湾实验室高毅勤教授课题组与华为联合推出蛋白质多序列比对（Protein MSA）数据集，希望在标准化的数据集基础上，支撑研究人员开发先进的AI模型，加深对蛋白质结构、功能和进化的认知，并进行蛋白设计与改造。此数据集将发布于华为云AI Gallery平台，相关代码及数据集说明将依托于华为全场景AI计算框架MindSpore进行开源开放、定期扩展与维护，旨在为全世界相关的产、学、研团队提供优质的数据共享解决方案。

本次开源的Protein MSA数据集完全覆盖最新版本（2021年2月发布）的UniRef50数据库中的蛋白质序列，采用学术界的“金标准”搜索方法，对约0.5亿条蛋白序列进行了充分的MSA搜索与比对（MSA平均深度大于1000），是目前世界范围内规模最大、参考数据集最新、覆盖度最广的开源蛋白质MSA数据集（之前最大的开源MSA数据集包含10万个蛋白MSA）【1】。

人类已知的蛋白质序列已经超过4.4亿条，但仅凭这些蛋白质单序列数据库，很难了解蛋白之间的关系。Protein MSA数据库是一个对不同蛋白质序列之间的关系进行了标记的大规模“关系型”数据库，被标记为关联的蛋白质序列之间的相似度、进化关系、突变所在位点的分布等信息对蛋白质结构和功能的预测极为重要。

为了更好地服务于跨领域的研究人员，Protein MSA数据集将被组织成具有多重形态的数据格式。原始数据集（近30T）将以UniRef系列数据库【2】和UniClust数据库【3】的标准文本形式存储，并按照序列长度进行分割与压缩。为了便于AI领域的研究人员直接使用，Protein MSA数据集还会将文本格式的数据集转化为浮点数张量类型压缩存储，并对已有的AI框架如MindSpore进行数据接口的支持。

高毅勤教授表示：“我们鼓励并期待来自生物信息学、数据科学和AI研究等领域的专家和人才充分碰撞与合作，引入、改进或设计全新的AI模型，来充分地挖掘Protein MSA数据集中所隐藏的‘自然的秘密’”。

从科学的角度看，MSA的数量和质量很大程度上影响了目前最先进结构模型的预测速度和精度，而且产生MSA的非参数化算法仍是诸多蛋白预测方法中决定速度的主要步骤之一。因此，Protein MSA数据库本身可以作为这些结构预测模型的预训练材料，用来挖掘序列信息甚至快速生成新的序列特征，这对解决研究、设计蛋白质中所面临的高变异序列和孤儿序列等问题具有巨大的潜在价值。

此次数据库的发布，依托于华为云AI Gallery平台，能够充分保障国内外用户对于数据集的访问和下载，并提供可持续更新与扩充的先进数据维护方案以及下游AI应用与部署的相关支持，融合了产、学、研相结合的研究模式的优势。此外，高毅勤课题组也与华为联合开发并开源了首个国产分子动力学软件MindSponge，希望未来该软件能在材料、生物、医药等领域得到广泛的应用。

消息来源：华为云

附：

数据集开源说明：

https://gitee.com/mindspore/mindscience/tree/master/MindSPONGE/protein_msa

数据集下载地址：

https://marketplace.huaweicloud.com/markets/aihub/datasets/detail/?content_id=5802def2-5fbd-40da-85d8-a4541d1c6f1e

【1】 AlQuraishi, Mohammed. "ProteinNet: a standardized data set for machine learning of protein structure." BMC bioinformatics 20.1 (2019): 1-10.

【2】 Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & UniProt Consortium. (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31(6), 926-932.

【3】Mirdita M.*, von den Driesch L.*, Galiez C., Martin M. J., Söding J.#, and Steinegger M.#, Uniclust databases of clustered and deeply annotated protein sequences and alignments, Nucleic Acids Res. 2016.