人工智能革命席卷蛋白质测序

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/32701.html

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

人工智能革命席卷蛋白质测序
。人工智能（AI）已经颠覆了蛋白质的研究工作，比如它能帮助研究人员轻松预测蛋白质的3D结构。相关成就获得了2024年的诺贝尔化学奖。

现在，AI深入蛋白质测序，通过组成蛋白质的氨基酸序列识别蛋白质。与传统方法相比，AI更快，还能帮助研究人员对以前从未见过的蛋白质进行测序。对未知蛋白进行测序，一直是医学诊断、环境研究和考古学中的常见挑战。

近日，一项发表于《自然-机器智能》的研究中，名为InstaNova的蛋白质测序AI，可以识别伤口中的致病蛋白和海水样本中微生物产生的未知蛋白。

事实上，InstaNova并非个例，在过去4年里，研究人员已经推出了20多种蛋白质测序AI。

“很明显，这是该领域的发展方向。”美国华盛顿大学蛋白质组学AI开发人员William Noble说。

蛋白质远比DNA和RNA要复杂得多。人类基因组包含约2万个基因，但这些基因会产生1000万种不同的蛋白质。

传统上，生物学家通过将蛋白质分解成肽这样的短片段来识别蛋白质。每个肽由5到20个氨基酸组成。研究人员利用质谱仪中测量这些短片段，将其重量与数十个数据库中已知肽的重量进行匹配，以确定其身份，然后在将这些片段拼成完整的分子。

但这种传统方法存在一些问题。比如，质谱法发现的高达70%的肽并不存在于现有的数据库中。

“传统蛋白质组学有点像利用搜索引擎进行搜索。如果它不在数据库中，你就搜索不到它。”丹麦技术大学蛋白质组学专家Timothy Patrick
Jenkins说，特别是随着肽数据库的不断扩充，发现匹配项所需的计算机时间变得越来越长。

而AI则不会费力寻找匹配的已知肽选线。它们计算了所有可能由给定长度肽化学修饰产生的潜在肽片段的重量。如果产生了与实际样本中的肽片段相匹配的片段，它会尝试将其组装成全长蛋白质。

为了提高准确性，蛋白质测序AI在数百万个已知肽及其如何组装成已知蛋白质的基础上进行训练。这使得AI能够习得氨基酸链结合的最常见方式。

Jenkins说，这种方法类似于大型语言模型，就像ChatGPT在大量文本上训练，以学习语法规则一样，蛋白质组学AI习得了一种蛋白质“语法”，为给定的一组肽提供了最可能的序列。

2021年，Noble和同事推出了Casanovo。这是第一个使用深度神经网络的蛋白质测序AI。在2024年发表于《自然-通讯》的论文中，Noble团队报告说，AI被证明擅长识别训练数据中没有的新肽序列。额外的实验表明，Casanovo擅长识别免疫系统攻击癌症时靶向的细胞表面肽，以及海水样本中的未知蛋白质。

而Jenkins和同事们开发的InstaNova在使用深度神经网络的基础上，引入了扩散模型。AlphaFold等蛋白质结构预测模型也采用了这种策略。

在与Casanovo的面对面测试中，InstaNova与升级款InstanNova+结合，在9种生物的实验室蛋白质混合物中鉴定出42%的肽。

当Jenkins团队将InstaNova应用于现实中的蛋白质组学测试时，除其他结果外，它从感染的腿部伤口中鉴定出1225种人血白蛋白特有的肽，是传统方法检索结果的10倍。其中254种是数据库中没有的新肽。

其他领域的研究人员也在用蛋白质测序AI。英国剑桥大学蛋白质组学研究员Matthew Collins最近就在测试几种蛋白质测序AI工具分析考古样本的能力。

Collins指出，大多数情况下，样本中的蛋白质在地下经过漫长岁月后发生了化学变化，或者它们来自早已灭绝的动植物，因此不太可能存在于传统蛋白质和肽数据库中，而这些AI模型尤其适用于混乱环境中蛋白质的检测。

利用AI工具，Collins团队已经在尼安德特人遗址中发现兔子蛋白质的特征，并在古代巴西的盆中发现鱼类肌肉蛋白质特征。

作者：许悦 来源：中国科学报

更多 科学进展 请访问 https://www.iikx.com/news/progress/

本文版权归原作者所有，请勿用于商业用途，爱科学iikx.com转发