生物学领域最大规模AI模型发布，可按需编写DNA

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/31919.html

生物学领域最大规模AI模型发布，可按需编写DNA

。2月19日，美国弧形研究所、美国芯片制造商英伟达公司和美国斯坦福大学等机构的研究人员共同开发的人工智能（AI）生物学模型Evo 2正式发布。目前，该模型已开放给全球科研人员，他们可通过网页使用该模型，还可免费该模型的源代码、训练数据及参数。

美国弧形研究所在其官网发布公报称，在前一代模型Evo 1的基础上，Evo 2已发展成为目前生物学领域规模最大的AI模型。Evo 1基于8万个细菌、古菌基因组及病毒等序列进行训练，Evo 2则基于超过12.8万个基因组数据的9.3万亿个核苷酸进行训练。这些模型使机器能够“用核苷酸语言来读、写和思考”。

用于训练Evo 2模型的酵母等真核生物基因组图片。图片来源：NCMIR/Science Photo Library

?

据《自然》报道，在过去几年里，科学家开发了日益强大的“ 蛋白质语言模型 ”，如美国互联网公司Meta开发的ESM-3模型。这类模型通过训练数百万蛋白质序列，已被用于预测蛋白质结构和设计包括基因编辑工具、荧光分子在内的全新蛋白质。

与这些模型不同，Evo 2的训练数据既包含指导蛋白质合成的“ 编码序列 ”，也包含可调控基因活动时空特征的非编码DNA。

相较于原核生物，真核基因组通常更长、更复杂——基因由编码区与非编码区交替构成，非编码调控DNA可能远离其调控的基因。为处理这种复杂性，Evo
2被设计成能学习百万碱基范围内的DNA序列模式。

为验证该模型解析复杂基因组的能力，美国弧形研究所的生物工程师Patrick Hsu团队使用Evo
2预测乳腺癌相关基因BRCA1中已知突变的影响。在相关测试中，Evo
2在预测哪些突变是良性突变、哪些是潜在致病突变方面均达到90%以上的准确率。

“ 在判断编码区变异是否致病方面，其表现接近最佳生物AI模型，已达到顶尖水平。 ”Hsu表示

，Evo 2有助于识别患者基因组中难以解读的变异。

此外，该模型还可用于设计新的生物工具或治疗方法，且有助于节省大量用于细胞或动物实验的时间和研究资金，通过找到人类疾病的遗传原因来加速新药研发。

美国生物模型开发公司Tatta Bio的计算生物学家Yunha Wang认为，Evo 2或擅长将细菌和古菌基因组的规律应用于人类新蛋白质设计。

“蛋白质语言模型等AI工具已引发生物设计革命。”斯坦福大学的计算生物学家Brian Hie及同事希望能用AI建模整个细胞。他们期待Evo-2等基因组模型可以帮助他们取得更大突破。

公报强调称，考虑到潜在的伦理和安全风险，研究人员在Evo 2的基础数据集中已排除了感染人类和其他复杂生物的病原体，并确保该模型不会对这些病原体的相关查询返回有效答案。

作者：李惠钰 来源：中国科学报