机器学习辅助定向进化新方法

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/23914.html

机器学习辅助定向进化新方法。定向进化是模拟自然进化机制，利用现代分子生物学方法创造大量的突变基因文库，采用灵敏的定向筛选策略，创造出自然界不存在的或改良特性的蛋白质等生物分子的一种方法。定向进化已广泛应用于蛋白质的分子改造和优化，被认为是生产具有改良或全新特性的蛋白质的高效方法，对于酶工程、多肽和大分子药物设计都具有重要意义。传统的定向进化实验流程包括筛选测试大量突变序列的功能，将得到的最优序列作为亲本序列进行下一轮的突变和筛选，实行多轮突变筛选以得到功能优化的蛋白序列。然而，传统的定向进化方式容易陷入局部最优，且实验所得的突变序列空间受限。

近年来，机器学习辅助定向进化得到越来越多的关注，通过计算机模型模拟实验筛选过程，可以显著减少实验筛选负担、提高筛选效率。机器学习方法最重要的是建立模型学习目标蛋白的序列突变体-功能的函数映射关系。这种映射关系被称为蛋白质适应度图景(protein fitness landscape)，其中适应度(fitness)是一个抽象概念，可定量刻画特定蛋白质序列具有的某种生物学功能(如蛋白的热稳定性、与其他蛋白质的相互作用强弱、催化特定酶促反应的效率等)。由于蛋白质功能不同，适应度图景本身的内涵各不相同。此外，蛋白质突变效应数据难以获得、实验费时费力、蛋白质适应度图景较为复杂。因此，如何使用有限的实验数据学习蛋白质适应度图景以指导定向进化实验成为难题之一。

中国科学院上海药物研究所郑明月课题组、廖苍松课题组，提出了新的深度神经网络模型GVP-MSA。该模型基于已有的不同类型的蛋白质适应度图景，通过迁移学习的方式构建新的目标蛋白的适应度图景。8月16日，相关研究成果以Learning protein fitness landscapes with deep mutational scanning data from multiple sources为题，在线发表在《细胞系统》(Cell Systems)上。

研究从蛋白质热稳定性、上位性效应和序列保守性等多个方面，探讨了适应度图景的共同机制。蛋白质发挥功能的基础是能够折叠和维持稳定的三维结构。研究对不同蛋白的计算结果发现，突变导致适应度的变化和热稳定性变化的数值上有相关性。上位性效应在不同蛋白的适应度图景中也隐含有相似机制。上位性效应表示残基在蛋白质中存在相互作用，导致多点突变效应并不等于其组成的单点突变效应的加和。研究发现，在不同蛋白适应度图景中，具有正向上位效应的双点突变的两个氨基酸的位置在三维结构上更加接近。此外，突变效应与同源序列的隐含分布之间的关系具有共性。这些共性是适应度图景迁移学习的基础(图1)。

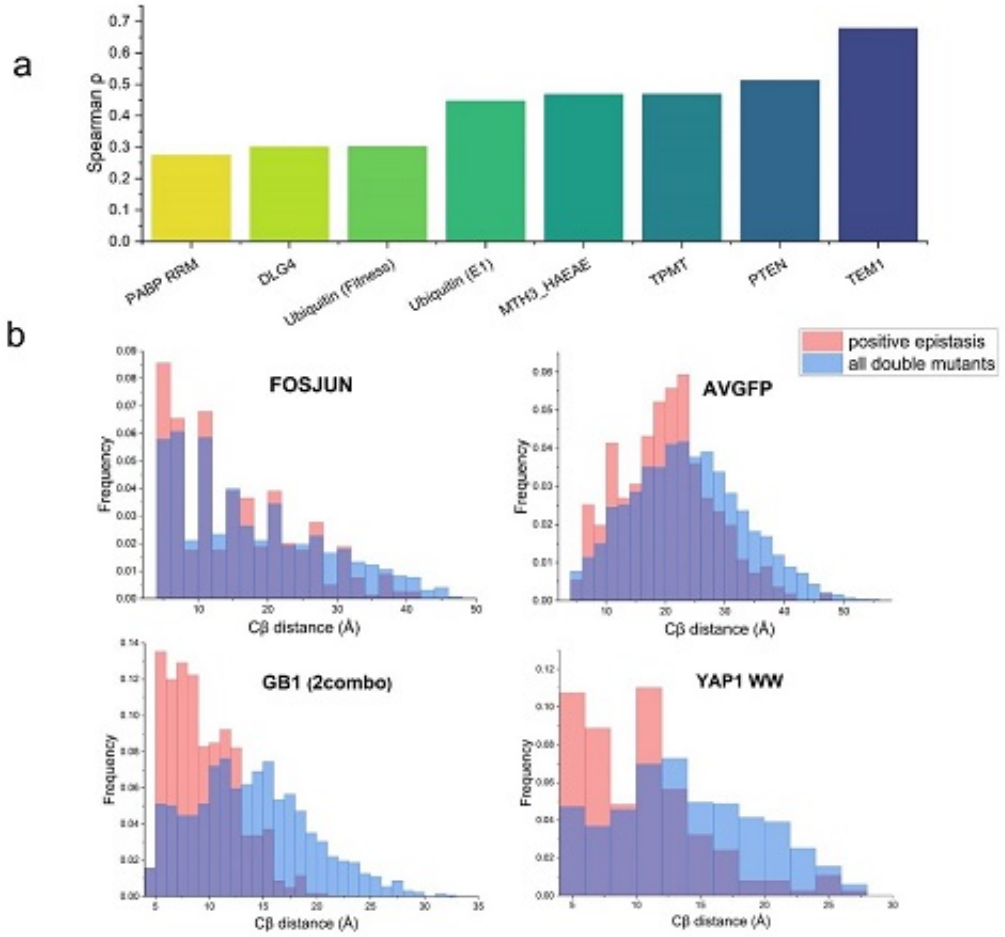

图1. 蛋白质适应度图景迁移学习的动机和基础。a、在不同蛋白的深度突变扫描实验中，突变导致的热稳定性变化与适应度变化相关。柱状图显示了由Rosetta计算的热稳定性和适应度变化之间的Spearman相关性。b、具有正上位效应的双点突变的残基在三维结构上更加接近。粉色直方图表示具有正向上位效应的双点突变的残基间距离，蓝色直方图表示所有双点突变的残基间距离。

该研究建立了一种新型的深度神经网络模型GVP-MSA，利用预训练的蛋白质语言模型处理目标蛋白的同源序列比对(MSA，multiple sequence alignment)信息，利用E-(3)等变图神经网络提取蛋白质三维结构信息，使用多任务学习的方式有效地学习整合不同维度、不同功能的蛋白质数据，从而泛化到新的目标蛋白体系。

此外，该团队设计了多种测试场景：单点突变效应的随机和按位置外推、对新蛋白质突变效应的零样本预测以及由单点突变效应预测多点突变效应(图2)。这些场景模拟了在定向进化实验中不同阶段的实际需求。GVP-MSA在这三种测试情景中均有良好的表现，验证了适应度图景迁移学习的有效性。该工作为机器学习辅助定向进化提供了新思路，有助于更加高效地探索蛋白质序列突变空间、快速设计具有改良或全新特性的蛋白质序列。

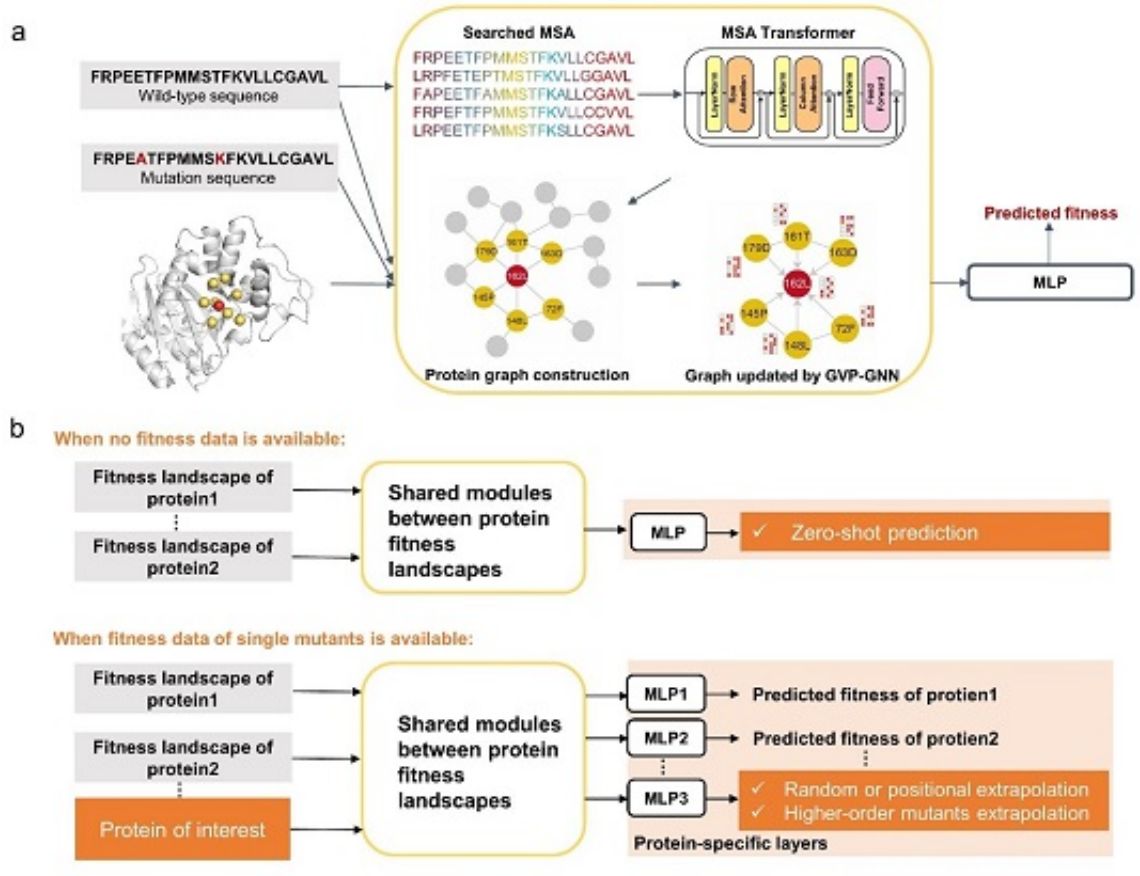

图2. GVP-MSA模型架构和应用场景需求概览。a、GVP-MSA的模型架构。b、蛋白质定向进化的应用场景需求：(1)没有目标蛋白质的适应度数据时，对新蛋白质的零样本预测能力;(2)已有少量目标蛋白的适应度数据时，模型的随机和按位置外推能力;(3)只有单点突变的适应度数据时，模型对多点突变效应的预测能力。

研究工作得到国家自然科学基金、临港实验室、国家重点研发计划、中国科学院青年创新促进会、上海市自然科学基金以及上海药物所与上海中医药大学中医药创新团队联合研究项目的支持。(来源：中国科学院上海药物研究所)

相关论文信息：https://doi.org/10.1016/j.cels.2023.07.003

作者：郑明月等 来源：《细胞系统》