科研人员开发出RNA基础语言模型

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/35820.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！

科研人员开发出RNA基础语言模型
。在真核生物中，RNA转录、剪接、翻译和降解等生物学过程受到顺式调控元件、RNA结构和反式作用因子的调控。解析RNA多层次调控，对研究基因表达分子机制和设计RNA药物具有重要意义。但是，由于调控复杂和数据量不足，构建RNA调控的预测模型面临挑战。

近日，中国科学院上海营养与健康研究所研究员张国庆联合南方科技大学讲席教授王泽峰团队，构建出基于深度神经网络的预训练模型，可微调预测pre-mRNA剪接位点、mRNA翻译效率、mRNA降解率和内部核糖体进入位点（IRES）等RNA调控相关的下游任务，揭示了RNA中调控元件的序列特征，鉴定出新型翻译调控元件，为探讨RNA调控机制和优化RNA生物医学应用提供了新工具和新思路。

研究团队设计并训练了基于多层transformer编码器架构的RNA语言模型LAMAR。研究下载处理约1500万条哺乳动物和病毒的基因和转录本序列，通过掩码学习进行无监督预训练，预先提取RNA序列特征；同时使用含有标签的数据集微调模型，实现RNA调控高效预测。

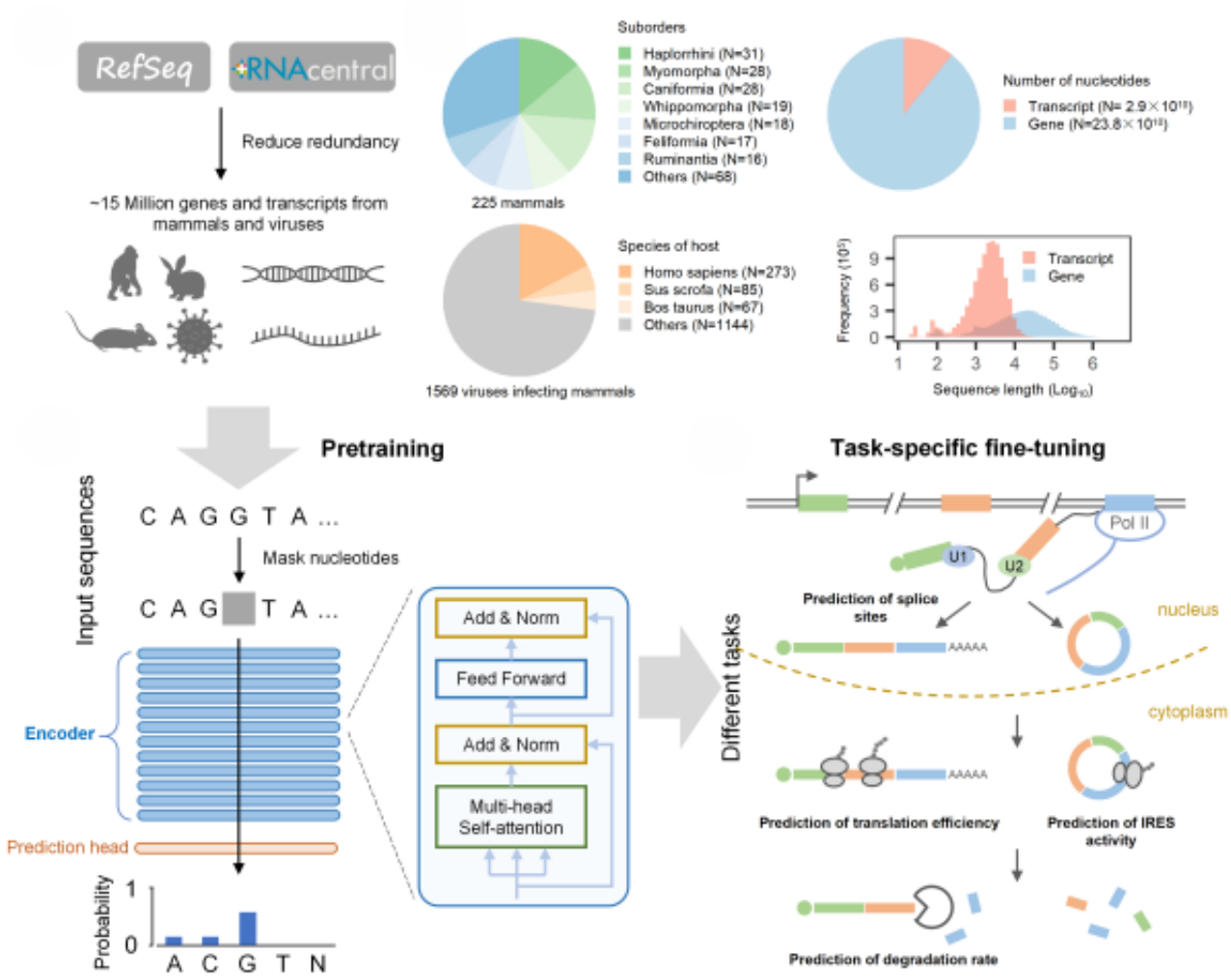
该研究测试了LAMAR模型在多个下游任务中的性能。在mRNA翻译效率和降解率预测任务中，LAMAR模型分别取得0.66和0.65的Spearman相关系数指标，相比最优基线模型提升7%和8%。在剪接位点预测任务中，LAMAR模型取得0.96的PR-AUC指标，与最优基线模型SpliceAI性能相当。

进一步，团队使用公开数据集微调模型预测病毒和真核IRES，取得0.985的AUROC指标。研究预测RNA病毒基因组中潜在的新IRES，并在多个细胞系中测试其中305条序列驱动环形RNA翻译的效率。研究发现序列的预测概率与翻译活性呈正相关，提示模型具有模拟筛选新型调控元件的能力。

9月24日，相关研究成果在线发表在《基因组生物学》（*Genome Biology*）上。研究工作得到国家重点研发计划、国家自然科学基金、中国科学院战略性先导科技专项（B类）等的支持。

目前，LAMAR模型已上传至Github（https://github.com/rnasys/LAMAR）供研究人员使用。

论文链接



LAMAR模型架构及研究流程图

研究团队单位：上海营养与健康研究所

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