QB期刊 上海交通大学杨旸团队提出了一种准确预测长链非编码RNA亚细胞定位的新方法——Loc4Lnc

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/37885.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！

QB期刊
上
海交
通大学杨
旸团队提出了一种准确预测长链非编码RNA亚细胞定位的新方法——Loc4Lnc。论文标题：Loc4Lnc: Accurate prediction of long noncoding RNA subcellular localization via enhanced RNA sequence representation

期刊：Quantitative Biology

作者：Yujia Cheng, Xiaoyong Pan, Yang Yang

发表时间：21 Apr 2025

DOI：10.1002/qub2.100

微信链接：点击此处阅读微信文章

长链非编码RNA（lncRNA）在基因调控、染色质结构和细胞分化中至关重要，不仅在多种疾病中扮演重要角色，并能作为潜在的生物标志物和治疗的靶点。然而，目前预测lncRNA亚细胞定位的方法在捕捉序列内远距离相互作用方面仍面临挑战。深度学习模型在特征提取上往往难以充分表征这些远距离依赖关系，导致预测准确性受限。尽管目前已发表的许多方法有助于识别局部序列特征，但往往无法捕捉理解lncRNA空间和功能动态的关键远距离相互作用。此外，这些方法大多基于迭代更新了很长时间的数据库早期版本。不同预测工具对亚细胞定位的分类定义尚不明确。各模型预测的亚细胞定位类别数量的不一致性，导致无法对当前的方法进行直接比较和性能评估，这为基于新数据库构建数据集和统一分类标准带来了挑战。

近期，上海交通大学杨旸团队在Quantitative Biology期刊上发表了题为Loc4Lnc: Accurate prediction of long noncoding RNA subcellular localization via enhanced RNA sequence representation的文章，开发了一个名为Loc4Lnc 的深度学习模型，该模型整合了卷积网络和Transformer模块以增强特征提取，使其能更有效地捕捉RNA序列内远端元件之间复杂的远程相互作用。此外，该模型还能准确地预测长链非编码RNA在细胞内的定位，并在独立测试集上与传统特

征提取方法进行比较，能实现最佳的预测性能。

Received: 11 October 2024 | Accepted: 6 January 2025
DOI: 10.1002/qub2.100

METHOD

Quantitative Biology
Open Access

Loc4Lnc: Accurate prediction of long noncoding RNA subcellular localization via enhanced RNA sequence representation

Yujia Cheng[1,2] | Xiaoyong Pan[3] | Yang Yang[1]

全文概要

该团队使用了最新一代数据集RNALocate v2.0构建了一种名为Loc4Lnc的模型。该模型整合了卷积层和Transformer模块以增强特征提取，使其更有效地捕捉RNA序列中远端元素之间的复杂长程相互作用。Loc4Lnc模型有三个主要组件，即数据集构建，特征提取和分类（如图1所示）。

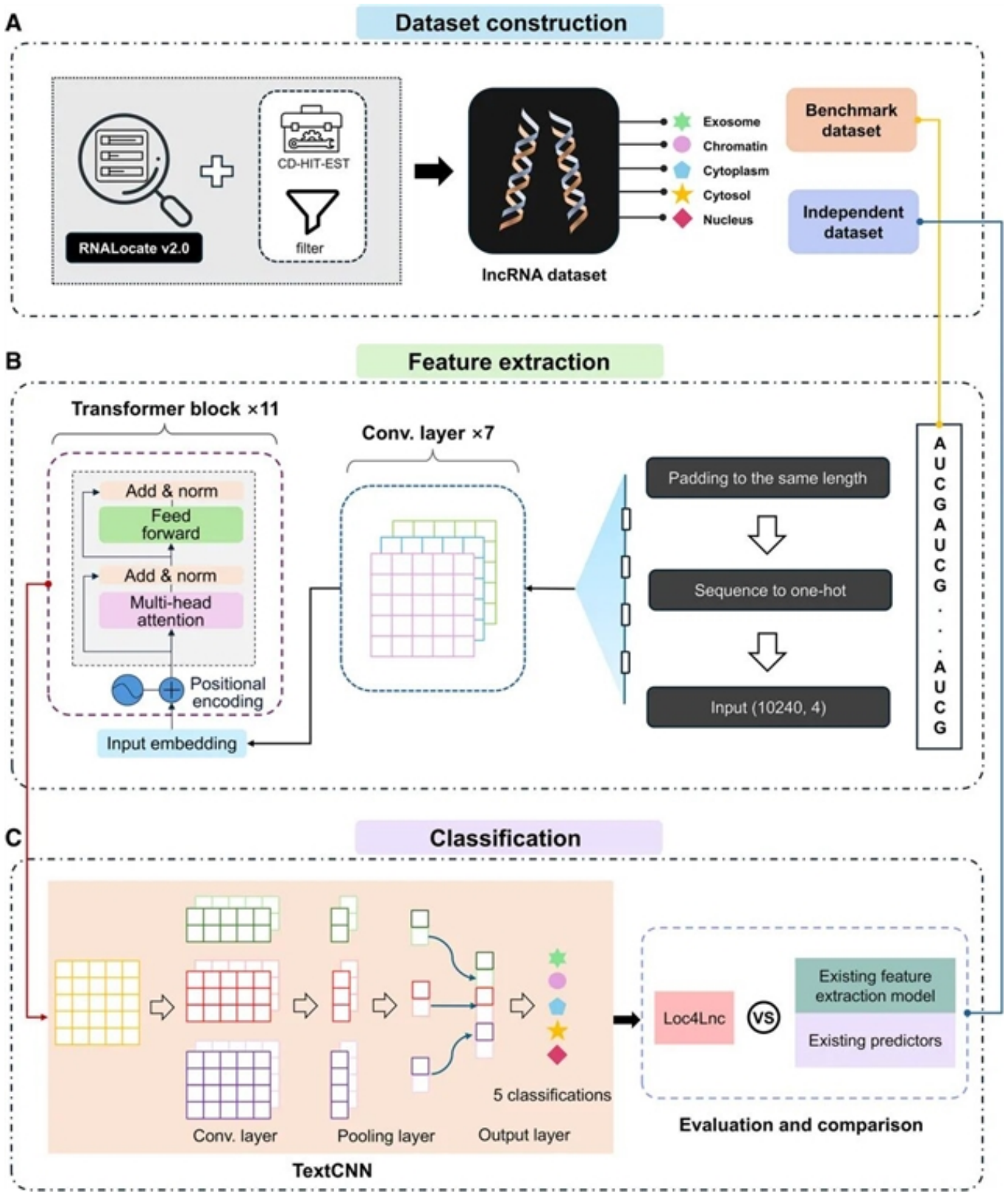


图1：Loc4Lnc框架有三个主要组件：数据集构建，特征提取和分类

一、数据集构建

研究团队首先从 RNALocate v2.0

数据库中获取原始的lncRNA亚细胞定位数据。然后对数据进行过滤和清洗，最终选定 5个待预测的亚细胞位置（细胞质、细胞核、胞质溶胶、染色质、外泌体），并去除冗余序列，构建了一个标准化的基准数据集。

二、 特征提取

受 Enformer 模型的启发，Loc4Lnc框架整体模型架构如图2所示。它由四个主要组成部分组成：茎卷积层、包含6个卷积层的卷积塔、11个变换器块，以及点状层。首先将标准化长度的RNA序列进行独热编码，然后通过一个由大核卷积与注意力池化构成的主干模块进行初步压缩与特征抽象。随后，序列经过一个包含六层卷积与池化模块的卷积塔，以分层方式提取局部至中程的序列模式。最关键的是，模型使用了一个包含11层、并集成了自定义相对位置编码的Transformer块，专门用于有效捕获序列中决定亚细胞定位的长程依赖关系。最终，通过一个逐点模块进行特征裁剪与精炼，输出一个融合了多尺度信息的高维特征嵌入向量，供下游分类器使用。整个流程实现了从原始序列到判别性特征表示的端到端自动学习。

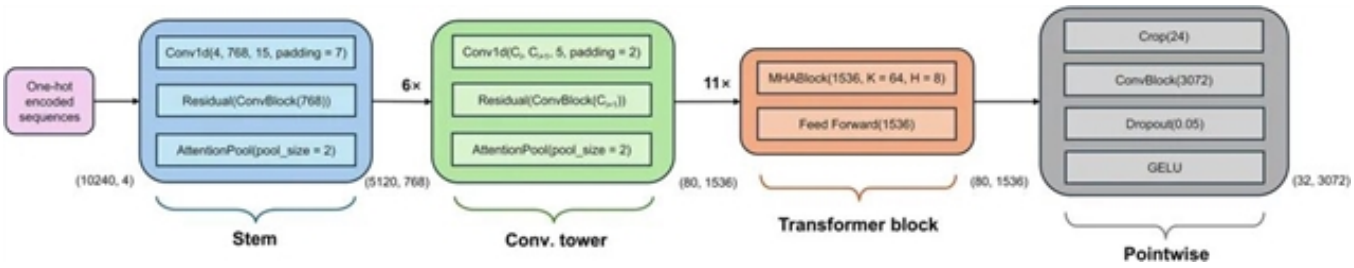


图2. 特征提取过程

三、 分类

利用TextCNN作为分类头，对前端提取的融合特征进行多尺度模式识别，最终通过一个全连接层输出五分类的概率预测。

四、与现有模型的性能对比

为了评估Loc4Lnc模型性能，作者其与结合传统分类器的现有特征提取方法进行了比较。结果如表1所示，Loc4Lnc优于所有基线方法，准确率达到0.636。

TABLE 1 Performance of Loc4Lnc and the existing feature extraction models.

Model	MaP	MaR	MaF1	ACC
5-Mer + MLP	0.192	0.204	0.459	0.459
word2vec (3-mer) + MLP	0.208	0.236	0.357	0.357
word2vec (5-mer) + MLP	0.259	0.263	0.259	0.538
DNABERT (3-mer) + MLP	0.530	0.559	0.477	0.483
DNABERT (3-mer) + SVM	0.398	0.532	0.372	0.409
DNABERT (5-mer) + MLP	0.504	0.588	0.475	0.547
DNABERT (5-mer) + SVM	0.427	0.617	0.459	0.546
Loc4Lnc	**0.569**	**0.647**	**0.581**	**0.636**

Note: Bold values denote the best results.

Abbreviations: ACC, accuracy; MLP, multi-layer perceptron; SVM, support vector machines.

五、与现有预测工具的比较

本研究还将Loc4Lnc的性能与现有的最先进预测工具进行比较，包括lncLocator (csbio.sjtu.edu.cn/bioinf/lncLocator)、DeepLncLoc (bioinformatics.csu.edu.cn/DeepLncLoc)、GraphLncLoc (csuligroup.com:8000/GraphLncLoc)和LncLocFormer (csuligroup.com:9000/LncLocFormer)，并利用网络服务器进行预测。结果如表2和表3所示，Loc4Lnc可以和lncLocator和DeepLncLoc一样，准确预测五个定位：细胞质、细胞核、细胞质基质、外泌体和染色质。并且Loc4Lnc (ACC=0.662)在预测细胞质、细胞核、细胞质基质和外泌体四个亚细胞定位时（Task 1），其性能显著优于lncLocator (ACC=0.257)和DeepLncLoc (ACC=0.253）；在预测细胞质、细胞核和外泌体定位（Task2）及细胞质、细胞核和染色质定位（Task3）时，其性能也表现出明显的优势。

爱科学
iikx.com

TABLE 2 Comparison of predictable subcellular locations of Loc4Lnc and existing predictors.

Model	Subcellular locations	# of classes
lncLocator	Cytoplasm, nucleus, cytosol, ribosome, and exosome	5
DeepLncLoc	Cytoplasm, nucleus, cytosol, ribosome, and exosome	5
GraphLncLoc	Cytoplasm, nucleus, ribosome, and exosome	4
LncLocFormer	Cytoplasm, nucleus, chromatin, and insoluble cytoplasm	4
Loc4Lnc	Cytoplasm, nucleus, cytosol, exosome, and chromatin	5

TABLE 3 Performance of Loc4Lnc and existing predictors on Tasks 1, 2, and 3.

Task	Model	MaP	MaR	MaF1	ACC
1	lncLocator	0.325	0.260	0.252	0.257
	DeepLncLoc	0.204	0.255	0.226	0.253
	Loc4Lnc	**0.728**	**0.662**	**0.674**	**0.662**
2	GraphLncLoc	0.590	0.406	0.386	0.414
	Loc4Lnc	**0.683**	**0.650**	**0.652**	**0.650**
3	LncLocFormer	0.534	0.517	0.518	0.517
	Loc4Lnc	**0.857**	**0.583**	**0.648**	**0.583**

Note: Bold values denote the best results.
Abbreviation: ACC, accuracy.

五、总结

总结来说，该项研究核心创新在于构建了一个基于卷积神经网络与Transformer相结合的混合模型，有效融合了序列的局部特征与长程依赖信息。在预测准确性上显著超越了现有的主流特征提取方法和预测工具，为解码lncRNA功能与疾病机制研究提供了一个更强大的计算工具。

QB期刊介绍

Quantitative Biology （QB）期刊是由清华大学、北京大学、高教出版社联合创办的全英文学术期刊。QB主要刊登生物信息学、计算生物学、系统生物学、理论生物学和合成生物学的最新研究成果和前沿进展，并为生命科学与计算机、数学、物理等交叉研究领域打造一个学术水平高、可读性强、具有全球影响力的交叉学科期刊品牌。

《前沿》系列英文学术期刊

由教育部主管、高等教育出版社主办的《前沿》（Frontiers）系列英文学术期刊，于2006年正式创刊，以网络版和印刷版向全球发行。系列期刊包括基础科学、生命科学、工程技术和人文社会科学四个主题，是我国覆盖学科最广泛的英文学术期刊群，其中12种被SCI收录，其他也被AHCI

、Ei、MEDLINE或相应学科国际权威检索系统收录，具有一定的国际学术影响力。系列期刊采用在线优先出版方式，保证文章以最快速度发表。

中国学术前沿期刊网

http://journal.hep.com.cn



来源：Quantitative Biology

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