



tRF Prospect: 基于神经网络学习的 tRNA 衍生片段靶标预测算法*

任黛西^{1,2)**} 易健勇^{3)**} 莫勇真¹⁾ 杨 梅¹⁾ 熊 炜¹⁾ 曾朝阳^{1,2)} 石 磊^{2)***}

(¹⁾ 中南大学湘雅基础医学院, 长沙 410078; (²⁾ 中南大学湘雅二医院病理科, 长沙 410011; (³⁾ 复旦大学脑科学研究所, 上海 200032)

摘要 **目的** 转运核糖核酸衍生片段 (transfer RNA derived fragment, tRF) 是一种新型的长度介于 13~50 nt 的非编码 RNA, 最近被发现可能通过靶向结合 mRNA, 促进 mRNA 降解, 抑制蛋白质翻译, 从而在细胞中发挥特定的生物学功能, 但预测其靶基因的生物信息学方法尚且有限。为了研究这一种新型分子在基因网络调控中发挥的作用, 开发一种更全面地预测 tRF 靶标的算法具有重要意义。**方法** 使用多层感知机神经网络深度学习算法, 以确认结合的 38 687 对 tRF : mRNA 为训练数据, 将已知的 tRF 和靶标结合的特征用于模型学习, 并扩大收录的 tRF 和靶标的范围。**结果** 将该基于神经网络学习的 tRF 靶标预测算法命名为 tRF Prospect, 算法 AUC 达 93.4%, 经与过表达 tRF 后的转录组高通量测序数据进行比对, 具有良好的预测性能, 证实该算法具有较高准确度。**结论** tRF Prospect 主要预测 tRF 发挥生物学功能可能作用的靶向 mRNA, 对现有的 tRF 靶标预测平台在收录 tRF 种类、靶标范围和预测模型上进行了补充, 为研究 tRF 在生物中的新作用提供了基础。

关键词 转运核糖核酸衍生片段, 数据库, 神经网络学习

中图分类号 R34, N20

DOI: 10.3724/j.pibb.2025.0059

CSTR: 32369.14.pibb.20250059

转运核糖核酸衍生片段 (transfer RNA derived fragment, tRF) 是一类源自转运核糖核酸 (transfer RNA, tRNA) 的新型短链非编码 RNA, 早期被认为只不过是 tRNA 随机降解的副产物, tRF 的生物学功能几十年来一直被忽视。随着测序技术的发展和生物信息分析技术的完善, 跨物种和数千个样本的研究为 tRF 的重要性提供了强有力的证据^[1-2], 它们在细胞内执行多种重要的生物学功能, 例如促进基因沉默、参与核糖体发生、改变翻译效率和影响表观遗传调控^[3]。同时, tRF 已被证明与多种人类疾病有关, 包括癌症、神经退行性疾病、代谢性疾病和病毒感染等^[2, 4-6]。

tRF 从生物发生上被分为 3 类, tRF-1、tRF-3 和 tRF-5。在 tRNA 加工成熟过程中, 核酸内切酶 Z (RNase Z/ELAC2) 从 pre-tRNA 中去除 3'-trailer 序列, 从而产生 tRF-1。tRF-3 来源于成熟 tRNA TΨC 环的 3'端, 通常被 Dicer 或 ANG 切割形成。tRF-5 在成熟 tRNA 的 D 环或茎区域以 Dicer 依赖性方式切割形成 (图 1)。需要指出的是, 当前 tRF 的研究

仍处于早期探索阶段, 其分子机制尚未完全阐明, 这种基础研究的局限性导致了现有数据库在 tRF 收录标准方面存在差异, 各数据库采用的分类系统缺乏统一性, 且在分子命名规范方面尚未形成广泛认可的标准化体系。tRF 通常长度在 14~50 nt, 与 microRNA (miRNA) 的大小范围相近, 许多研究也已经表明, tRF 发挥和 miRNA 一致的功能, 即靶向结合细胞内 mRNA 导致靶 mRNA 降解和翻译受阻^[7-9]。但目前, tRF 靶标预测往往使用 miRNA 靶标预测平台进行, 这可以提供一些候选靶基因, 但算法还不够准确。最近出现了针对 tRF 的靶标预测工具, 但各有其局限性, 例如, tRFtar 和 tRFtars 不能预测 tRF-1 的靶标, 但 tRF-1 在某些情

* 国家自然科学基金 (82472789), 湖南省自然科学基金 (2025JJ50590), 长沙市自然科学基金 (kq2208329) 和中南大学研究生科研创新项目 (1053320242771) 资助。

** 并列第一作者。

*** 通讯联系人。

Tel: 0731-88660449, E-mail: shilei81@csu.edu.cn

收稿日期: 2025-02-09, 接受日期: 2025-07-14

况下确实会影响靶基因表达^[10-11]。tRFtarget是基于RNA分子间相互作用构建的算法,但排除了可能对tRF靶标预测很重要的几个特征,例如靶位点保护和AU含量,并且只收录了936个tRF^[12]。

tRFforest虽然较为全面地预测了结合特征,但仅收录628个tRF,且只能预测tRFs与mRNA 3'非翻译区(non-translational region, UTR)的结合^[13]。

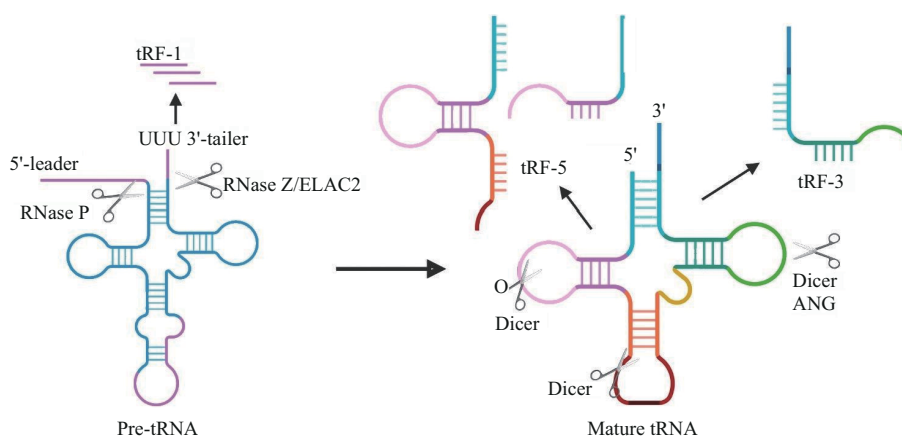


Fig. 1 Biogenesis and classification of tRFs

During the maturation process of tRNA, the endonuclease Z (RNase Z/ELAC2) removes the 3'-trailer sequence from pre-tRNA, thereby producing tRF-1. tRF-3 originates from the 3' end of the mature tRNA TΨC, and is typically formed by cleavage with Dicer or ANG. tRF-5 is formed by a Dicer-dependent cleavage in the D loop or stem region of the mature tRNA.

因此,本文通过多层感知机神经网络深度学习算法构建了一种tRF目标预测算法, tRF Prospect。利用小RNA及其靶标交联生成的交联测序(CLASH)数据作为训练和测试数据集^[14],并从7个特征计算tRF与mRNA间的相互作用。该算法可以对MINT base收录的所有tRF(23 413个)进行靶标预测。经受试者工作曲线(receiver operating characteristic, ROC)和曲线下面积(area under the curve, AUC)检测tRF Prospect的性能,并在过表达tRF后的转录组高通量测序数据中检测了准确性。此外,由于本文采用了基于机器学习的方法,该预测算法可以在未来更新迭代,以更全面的特征迁移到更多的tRF。

1 方 法

1.1 数据和算法源及数据预处理

从既往的研究中获取了tRF的CLASH数据及其相应的基因靶标,该数据鉴定了Helwak等^[7, 14]的CLASH数据中可用的所有tRF:mRNA嵌合读段。该CLASH数据集经过过滤,排除了来自非tRNA来源(包括miRNA)的片段。因此该训练集具有良好的可信度,训练集中共包含5 959对tRF-5:mRNA, 32 293对tRF-3:mRNA和435对

tRF-1:mRNA。

tRF序列是从MINTbase (<https://cm.jefferson.edu/MINTbase/>)获得的^[15]。人类蛋白质编码基因的mRNA序列,以及它们的Ensembl和Refseq IDs,从Ensembl Biomart(版本GRCh38.p13, <http://grch38.ensembl.org/biomart/>)中获得(共16 610个基因)。tRF Prospect的预测特征来自<https://www.trforest.com>^[14]。

1.2 阴性数据生成

CLASH数据仅提供正向的阳性数据来训练模型,因此本文将CLASH数据中发现的目标归类为真实目标。为了使训练模型更有效,必须生成负性数据集。本文使用python 3.7中random模块的shuffle函数对已知结合数据进行打乱重组来生成阴性数据(即不结合的假目标)用于模型训练。

1.3 特征计算

计算CLASH数据集中每个tRF:mRNA相互作用的特征,包括以下7个。

a. AU含量。计算结合位点周围的局部AU含量,分别计算其上下游各30 nt的AU含量然后取两者均值。参考tRFforest模型中的计算公式。公式如下:

$$\sum_{i=1}^{30} \frac{P(i)}{i} = \frac{1}{1} + \frac{0}{2} + \frac{0}{3} + \frac{1}{4} + \dots + \frac{1}{29} = 2.27546$$

$P(i)$ 表示该位点是否为 AU, 是则 $P(i) = 1$, 否则 $P(i) = 0$;

- b. 计算 3'端互补配对位点数量;
- c. 计算最长连续互补配对的位点所在位置;
- d. 计算互补配对的位点数量;
- e. 计算互补配对序列中最长的连续互补配对碱基数;
- f. 计算结合区长度;
- g. 计算种子序列: 看是否存在 7mer-m1 种子序列配对 (种子序列为 tRF 的 1~7 个核苷酸)。

目前已知, tRF 可以使用种子序列以 miRNA 样机制下调靶 mRNA 表达, 但是既往研究中这些相互作用特征大部分来源于 tRF 与 mRNA 3'UTR 的相互作用。由于有文献报道 tRF 也可以与 mRNA 的 5'UTR 和编码序列 (coding sequence, CDS) 区结合, 因此将这些特征在 CDS 区和 5'UTR 也进行了训练, 希望以此推测 tRF 与 mRNA CDS 区和 5'UTR 区结合的作用^[16-19]。

1.4 多层感知机训练和测试

利用 Scikit-learn 的 MLPClassifier 分类器进行模型构建。使用网格搜索方法对模型参数进行优化。模型的训练集和测试集使用 Scikit-learn 中的 train_test_split 函数进行分解。函数参数设置为 70% 样本作为训练集, 30% 样本作为测试集, 并且将样本标签随机打乱, 保证训练集和测试集中的正负样本均衡。并在训练模型时采用十折交叉验证作为防御措施, 防止算法过度拟合, 提高可靠度。

1.5 算法验证

分别训练逻辑回归模型、随机森林模型、决策树模型和多层感知机模型。模型的评估获得以下性能指标: 准确性、ROC 曲线和 AUC 值。选择 AUC 值最高的模型作为 tRF Prospect 的预测算法并进一步对算法进行验证。从 GEO 数据库中分别获取 tRF AS-tDR-007333 过表达 (GSE184690)、tRF-3004b 过表达 (GSE197091) 和 tRF-20-S998LO9D 过表达 (GSE208381) 后的高通量转录组测序数据, 分析测序数据中差异表达的基因作为调控靶基因。分别使用 tRF Prospect、tRFTar、tRForest 和 tRFTarget 对 tRF AS-tDR-007333、tRF-3004b 和 tRF-20-S998LO9D 的靶基因进行预测, 统计比较各算法获得的靶基因数量。

2 结 果

2.1 tRF Prospect的建立

从 AGO1 交联、连接和杂交 (CLASH) 数据集^[8] 中获得了经实验验证的可信数据, 并使用第 1 节中描述的计算特征来训练 tRF Prospect。经过下面介绍的几个优化步骤, 就得到了最终的算法 (图 2)。从 CLASH 中获得了 tRF: mRNA 结合数据, 并使用 7 个计算特征来训练 tRF Prospect, 通过十折交叉验证 (10-fold cross validation, 将数据集分成 10 份, 轮流将其中 9 份作为训练集, 1 份作为测试集, 进行试验) 提高模型准确度, 获取模型的表现指标选择准确度最高的模型作为 tRF Prospect 使用的算法, 用户可以通过输入 tRF 序列获得可能的靶基因。

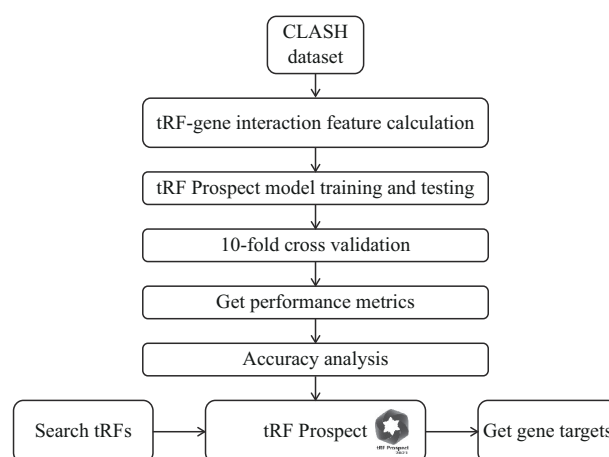


Fig. 2 The procedure for establishing the tRF Prospect

The tRF:mRNA binding data were acquired from CLASH, and 7 computational features were utilized for training the tRF Prospect model. The model's accuracy was enhanced through tenfold cross-validation. A performance index was obtained to select the model with the highest accuracy for tRF Prospect.

2.2 逻辑回归

本研究采用 ROC 曲线的 AUC 作为二元分类任务的性能评价指标, 该指标越趋近于 1 表征模型分类效能越佳。在模型构建阶段, 首先选用传统机器学习中的逻辑回归算法作为基准模型, 该模型因其在二元分类任务中的广泛应用性而被优先采用。从建模原理分析, 逻辑回归通过 Sigmoid 函数将线性回归结果映射为概率输出, 在保持强可解释性的同时突破了传统线性模型对变量关系的限制, 能够有

效表征非线性关联特征^[20-21]。逻辑回归通过非线性映射减小了离分类平面较远的点的权重, 相对提升了与分类最相关的数据点的权重。根据数据的分布情况对逻辑回归模型的参数的进行设置, 参数详情见表1。

针对 mRNA 的 3'UTR、5'UTR 和 CDS 区分别进行模型训练, 并得到了 3 个模型, 对 tRF : mRNA 交互结合的各个特征的权重输出见表2。

Table 1 Parameter selection for the tRF Prospect model

Parameters of the logistic regression model	Application parameter
Penalty	L2
C	0.001
Fit_intercept	True
Solver	sag/saga

然而, 该算法训练得到的 3 个模型 *AUC* 值未达到 0.9, 说明准确度欠佳 (图3)。

Table 2 Weight of each feature in logistic regression model

	The number of 3'end pairs	The position of the longest consecutive pairing	Localization of the binding region	Enumeration of binding sites	The length of the longest consecutive pairing	Binding zone length	Seed sequence
The 3'UTR model feature weight	-0.273 646 61	0.007 582 03	-0.000 741 06	0.002 893 97	0.010 800 48	-0.001 913 81	- 0.038 289 01
The 5'UTR model feature weight	-0.257 242 75	0.015 410 82	-0.006 221	0.028 909 59	-0.023 063 04	-0.001 370 02	- 0.016 954 97
The CDS model feature weight	-0.277 443 16	0.014 703 73	-0.006 679 13	0.023 001 22	0.004 092 62	-0.003 320 68	- 0.005 634 13

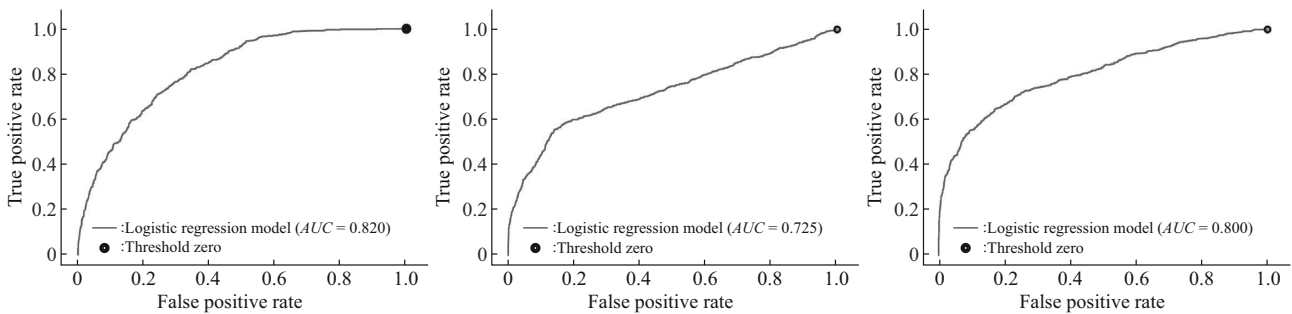


Fig. 3 ROC curve of logistic regression model

Three prediction models for tRF:mRNA target genes (3'UTR, 5'UTR, and CDS) were developed using logistic regression algorithm, with their corresponding ROC curves and *AUC* values.

2.3 随机森林

继续训练其他的模型。随机森林算法通过 Bootstrap 聚合策略构建决策树集成, 可以有效克服数据类别不平衡与小样本过拟合问题^[22]。先前已有研究将随机森林算法应用于生物医学, 且表现出良好的效果^[23-25]。基于这一优势, 使用随机森林模型, 分别针对 mRNA 的 5'UTR、3'UTR 和 CDS 区进行训练, 得到的结果见图4。

可见, 随机森林在 5'UTR、3'UTR 和 CDS 区训

练得到的 *ACU* 值均较逻辑回归更高, 接近 0.9, 但并不令人十分满意, 因此继续探索是否有更佳

2.4 决策树

在生物信息学中, 决策树算法因其模型特性具有可解释性而得到应用, 其核心机制在于采用递归二分法策略进行特征空间划分。算法首先将整个数据集作为起点, 通过计算各特征的分类能力进行评分, 选出区分效果最佳的特征作为首个判断节点。

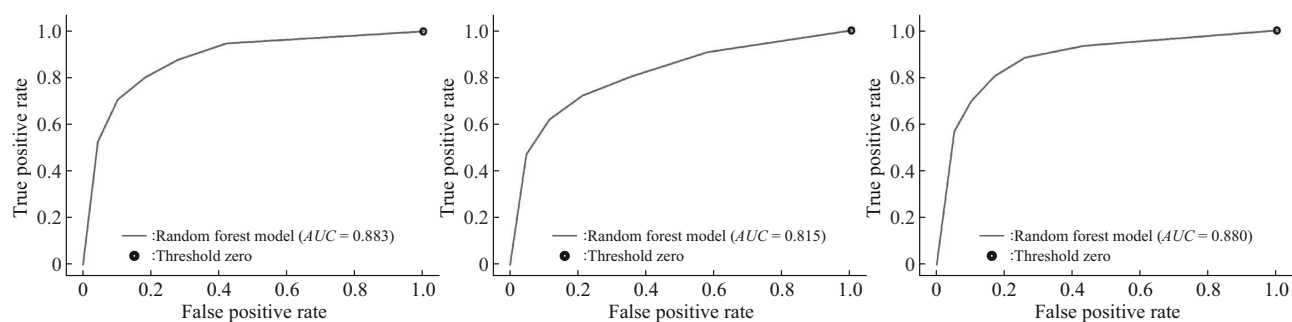


Fig. 4 ROC curve of random forest model

Three prediction models for tRF:mRNA target genes (3'UTR, 5'UTR, and CDS) were developed using random forest algorithm, with their corresponding ROC curves and *AUC* values.

然后根据选定特征的数值范围，将数据集划分成子集。最后重复上述拆分过程，直到剩余特征已无法显著提升分类准确率^[26-28]。相同的，使用该算法

分别对 5'UTR、3'UTR 和 CDS 区域进行训练，得到结果见图 5。

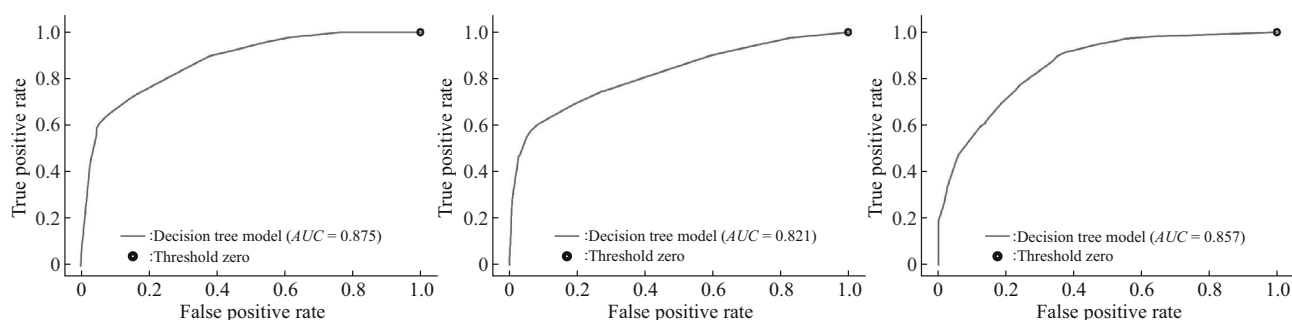


Fig. 5 ROC curve of decision tree model

Three prediction models for tRF:mRNA target genes (3'UTR, 5'UTR, and CDS) were developed using decision tree algorithm, with their corresponding ROC curves and *AUC* values.

结果表明，决策树和随机森林的表现接近，3个模型的 *AUC* 值均在 0.8 以上，但未达到 0.9。

2.5 多层感知机

进一步尝试使用多层感知机模型进行深度学习。多层感知机通过引入非线性激活函数（如 ReLU、Sigmoid）能够突破线性模型的限制，捕捉输入特征间复杂的非线性交互关系，因此在生物医

学研究中具有独特的优势。目前已经有多层感知机应用于生物医学领域的案例，这些尝试都取得了良好的成果^[29-31]。

在训练过程中，使用网格搜索方法对模型参数进行优化，其中主要的优化参数包括 `hidden_layer_sizes` 和 `alpha`。多层感知机模型的各项参数设定如表 3 所示，结果如图 6。

Table 3 Parameter setting of multi-layer perceptron model

Parameters of the multilayer perceptron model	Application parameter
Hidden_layer_sizes	(100, 100)
Activation	relu
Solver	sgd
Alpha	0.001
Batch_size	auto
Random_state	42

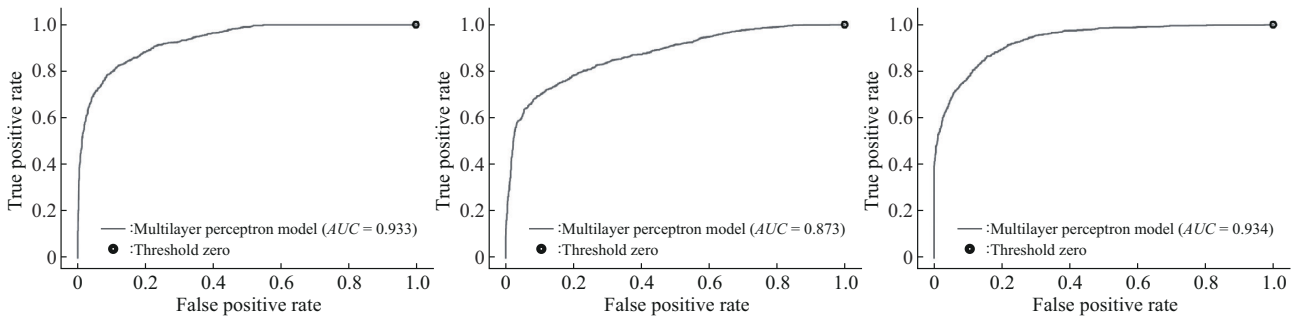


Fig. 6 AUC curve of multi-layer perceptron model

Three prediction models for tRF:mRNA target genes (3'UTR, 5'UTR, and CDS) were developed using multilayer perceptron algorithm, with their corresponding ROC curves and AUC values.

Table 4 AUC values of each model

Model	Logistic regression	Random forest	Decision tree	Multilayer perceptron
3'UTR	0.82	0.883	0.875	0.933
5'UTR	0.725	0.815	0.821	0.873
CDS	0.8	0.880	0.857	0.934

The performance of the 4 models in predicting the binding of tRF to the 3'UTR, 5'UTR, and CDS regions of mRNA is shown.

将通过多层感知机训练得到 3 个模型与上述其他模型的 AUC 值进行比较, 发现其在 3'UTR 和 CDS 区的 AUC 值均超过 0.9, 在 5'UTR 区的 AUC 值也接近 0.9, 优于其他模型 (表 4)。综上所述, 多层感知机在几种模型中表现最好, 因此, tRF Prospect 采用多层感知机训练所得的模型。尽管在 5'UTR 的表现相对不够好, 但这可能是由还未被发现的 tRF 和 mRNA 5'UTR 的特殊作用机制导致的。

3 算法验证

接下来, 使用 tRF 过表达后表达改变的基因作为实验验证的真实靶基因对 tRF Prospect 的准确性

进行验证。从 GEO 数据库获取 tRF AS-tDR-007333 过表达 (GSE184690)^[32], tRF-3004b 过表达 (GSE197091)^[14] 和 tRF-20-S998LO9D 过表达 (GSE208381)^[33] 后的转录组高通量测序数据, 分析测序数据中 tRF 过表达前后差异表达的基因。分别使用 tRF Prospect、tRFTar、tRFforest 和 tRFTarget 对 tRF AS-tDR-007333、tRF-3004b 和 tRF-20-S998LO9D 的靶基因进行预测, 并对鉴定出的靶基因数量进行统计分析。与其他 tRF 靶基因预测算法相比, tRF Prospect 要么可以命中更多的靶基因, 要么可以预测其他平台未收录的靶基因 (表 5)。结果表明, tRF Prospect 在靶基因命中上优于

Table 5 Validation of each algorithm

Experiment	Cell type	Algorithm	# of targets
tRF AS-tDR-007333 overexpression (GSE184690)	A549	tRF Prospect	87
		tRFTar	N/A
		tRFforest	N/A
		tRFTarget	N/A
tRF-3004b overexpression (GSE197091)	HEK293T	tRF Prospect	2 262
		tRFTar	N/A
		tRFforest	111
		tRFTarget	5 000
tRF-20-S998LO9D overexpression (GSE208381)	RL-95-2	tRF Prospect	697
		tRFTar	25
		tRFforest	N/A
		tRFTarget	N/A

tRFTar 和 tRForest。tRFTarget 的性能或较 tRF Prospect 更佳,但仅收录 936 个 tRF,导致其应用的局限性,比如 tRFTarget 不能预测 tRF-20-S998LO9D 和 tRF AS-tDR-007333 的靶基因。

验证结果表明, tRF Prospect 能够有效预测出结合可能性最大的靶基因,且优于一些现有的预测平台。

4 讨 论

tRF 作为新型调控分子,其生物学功能逐渐受到关注。现有研究表明, tRF 可能通过类似于 miRNA 的 Ago 蛋白依赖机制参与基因表达调控^[3, 5]。然而,当前生物信息学工具中,专门针对 tRF 靶基因预测的计算平台仍存在缺口, tRFTar、tRForest、tRFTarget 提供 tRF 靶基因预测服务,但均基于传统机器学习算法^[11, 13-14]。本研究开发了一个更全面的经过验证的 tRF 靶标预测算法。应用多层感知机对 AU 含量、位点是否配对、结合区位置、结合位点点数、互补配对序列中最长的连续互补配对碱基数、结合区长度和种子序列等 7 个特征进行学习,使用 38 687 对 tRF : mRNA 进行训练和测试,得到 tRF Prospect,算法 AUC 达 0.934。经由实验验证也表现出良好的预测性能及更大的效应范围。

本研究构建的 tRF Prospect 算法,整合 7 类关键结合特征,采用多层感知机构建预测模型,基于 38 687 对实验验证的 tRF : mRNA 相互作用数据集训练,并通过转录组高通量测序数据验证。与现有工具的比较分析表明,该模型在效应范围和命中量方面具有显著优势。本文首次将多层感知机应用于 tRF 靶基因预测,基于其非线性建模能力解析特征间的交互作用, tRF Prospect 在相同数据集上展现出更优的性能。但可解释性较传统机器学习算法存在一定差距,后期可以通过增加数据的数量级以及特征数量进一步改善算法的预测效果。

tRF Prospect 在数据可用性和效率方面有几个限制。本研究采用了随机生成的负样本数据来训练模型,这是由于缺乏足够多的通过实验验证过的负样本数据,导致了用于训练的负样本数据存在区分度过高和数据分布失真的问题,未来获取更多的真实负样本数据可提升算法的准确性。目前已知的结合特征并不非常明确,因此 tRF Prospect 目前仅适用于发挥类 miRNA 作用的 tRF。显而易见的是,

针对 5'UTR 的训练结果在各个模型中均是较差的一个,这说明 tRF 和 mRNA 5'UTR 的结合作用机制可能存在特殊性,目前已经发表的几篇研究也报道了这一事实^[34-35]。进一步明确 tRF 和 mRNA 的作用机制有助于完善算法。一些 tRF 仅能获得 5'UTR/3'UTR/CDS 可结合的基因,还有一些 tRF 无法被预测到可能的靶基因,这可能是由于 Ensembl Biomart (版本 GRCh38.p13) 中收录的 mRNA 不够全面,也可能是由于部分 tRF 的作用机制并不符合目前的认知。

tRF Prospect 与 tRFTar 和 tRForest 相比,在命中靶基因的数量方面存在更显著的优势。原因是 tRF Prospect 采用了多层感知机算法,相比 tRFTar 和 tRForest 采用的支持向量机算法和随机森林算法具有更强的识别弱结合靶点的能力。多层感知机算法通过深层非线性网络拟合了 tRF 和 mRNA 结合的连续渐变特性,而支持向量机和随机森林更依赖离散规则划分,导致对处于特征决策边界的靶点(比如弱结合和非经典种子匹配)的捕捉能力较差。然而这一特性也使 tRF Prospect 与 tRFTar 和 tRForest 相比,精确率可能有所下降,以换取显著的召回率提升。但是这也使得 tRF Prospect 可用于探索更多潜在靶点。在 GSE197091 数据集中, tRFTarget 相比 tRF Prospect 命中了更多的靶基因,这可能是由于 tRFTarget 是利用 RNAhybrid 和 IntaRNA 进行预测后,将两者的结果收录进数据库,扩大了收录的靶基因数量,但是 tRFTarget 仅收录了 936 个 tRF 的预测结果,所以在预测范围上不具备 tRF Prospect 的优势。需要说明的是, tRFTar、tRForest 和 tRFTarget 提供了交互式网页,为用户提供除了靶基因预测以外的更多功能。比如 tRFTar 还提供基因组可视化查询、tRF 功能富集分析; tRForest 提供 tRF:基因双链体的结合自由能查询、tRF 在基因上结合的位置及可视化互作图像和功能富集分析; tRFTarget 则提供 8 个物种的 tRF 靶基因预测^[11, 13-14]。但这些平台均是将计算好的结果存储后供用户查询, tRF Prospect 则可以即时针对 tRF 的序列计算靶基因。

总的来说,本研究提供了一种新的基于多层感知机的深度学习模型来预测 tRFs 靶基因的算法。tRF Prospect 扩大了研究人员可以预测的 tRF 范围,为进一步探索 tRF 的生物学功能提供新的方法基础。

5 结 论

本研究针对tRF靶基因预测领域的研究缺口, 开发了基于深度学习的算法tRF Prospect。该算法首次将多层感知机(MLP)引入tRF靶基因预测, 通过整合AU含量、位点配对长度、结合区位置等7类关键特征, 利用38 687对实验验证的tRF:mRNA相互作用数据进行训练与测试。模型表现出优秀的预测性能, AUC值高达0.934, 并通过转录组测序结果检验了其在效应范围和命中量上的优势, 为tRFs的调控机制研究提供了更全面的计算工具。

相较于现有工具, tRF Prospect实现了更多的靶基因命中数量和更广的tRF预测范围, 支持即时靶基因计算, 而非依赖预存结果。但存在局限性: 当前模型主要适用于类miRNA作用机制的tRFs, 对5'UTR区域的预测效果欠佳; 预测范围受限于Ensembl Biomart数据库的mRNA覆盖度及现有tRFs作用机制的认知边界; 未纳入组织或疾病特异性表达数据。此外, 深度学习的“黑箱”特性导致模型可解释性弱于传统算法。

未来的工作需结合更多实验数据完善特征体系, 尤其需深化对tRF结合mRNA 5'UTR结合机制的解析, 并通过扩大训练数据集规模和使用高质量负样本数据提升预测准确度。尽管存在优化空间, tRF Prospect通过深度学习, 显著拓展了可预测tRF的范围, 为揭示tRF的生物学功能提供了新方法论基础。

数据可用性声明 本论文的预测算法代码已上传, 可公开获取 (https://github.com/YiJianying/tRFTargets_Prediction/tree/main)。

参 考 文 献

- [1] Krishna S, Raghavan S, DasGupta R, *et al.* tRNA-derived fragments (tRFs): establishing their turf in post-transcriptional gene regulation. *Cell Mol Life Sci*, 2021, **78**(6): 2607-2619
- [2] Fu M, Gu J, Wang M, *et al.* Emerging roles of tRNA-derived fragments in cancer. *Mol Cancer*, 2023, **22**(1): 30
- [3] Chen Q, Zhang X, Shi J, *et al.* Origins and evolving functionalities of tRNA-derived small RNAs. *Trends Biochem Sci*, 2021, **46**(10): 790-804
- [4] Mathew B A, Katta M, Ludhiadch A, *et al.* Role of tRNA-derived fragments in neurological disorders: a review. *Mol Neurobiol*, 2023, **60**(2): 655-671
- [5] Yang M, Mo Y, Ren D, *et al.* Transfer RNA-derived small RNAs in tumor microenvironment. *Mol Cancer*, 2023, **22**(1): 32
- [6] Deng J, Ptashkin R N, Chen Y, *et al.* Respiratory syncytial virus utilizes a tRNA fragment to suppress antiviral responses through a novel targeting mechanism. *Mol Ther*, 2015, **23**(10): 1622-1629
- [7] Kumar P, Anaya J, Mudunuri S B, *et al.* Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol*, 2014, **12**: 78
- [8] Kuscu C, Kumar P, Kiran M, *et al.* tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA*, 2018, **24**(8): 1093-1105
- [9] Światowy W, Jagodziński P P. Molecules derived from tRNA and snoRNA: entering the degradome pool. *Biomed Pharmacother*, 2018, **108**: 36-42
- [10] Zhou Y, Peng H, Cui Q, *et al.* tRFTar: prediction of tRF-target gene interactions via systemic re-analysis of Argonaute CLIP-seq datasets. *Methods*, 2021, **187**: 57-67
- [11] Xiao Q, Gao P, Huang X, *et al.* tRFTars: predicting the targets of tRNA-derived fragments. *J Transl Med*, 2021, **19**(1): 88
- [12] Li N, Shan N, Lu L, *et al.* tRFtarget: a database for transfer RNA-derived fragment targets. *Nucleic Acids Res*, 2021, **49**(D1): D254-D260
- [13] Parikh R, Wilson B, Marrah L, *et al.* tRForest: a novel random forest-based algorithm for tRNA-derived fragment target prediction. *NAR Genom Bioinform*, 2022, **4**(2): lqac037
- [14] Helwak A, Kudla G, Dudnakova T, *et al.* Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 2013, **153**(3): 654-665
- [15] Pliatsika V, Loher P, Magee R, *et al.* MINTbase v2.0: a comprehensive database for tRNA-derived fragments that includes nuclear and mitochondrial fragments from all The Cancer Genome Atlas projects. *Nucleic Acids Res*, 2018, **46**(D1): D152-D159
- [16] Liu X, Mei W, Padmanaban V, *et al.* A pro-metastatic tRNA fragment drives Nucleolin oligomerization and stabilization of its bound metabolic mRNAs. *Mol Cell*, 2022, **82**(14): 2604-2617.e8
- [17] Kim H K, Xu J, Chu K, *et al.* A tRNA-derived small RNA regulates ribosomal protein S28 protein levels after translation initiation in humans and mice. *Cell Rep*, 2019, **29**(12): 3816-3824.e4
- [18] Goodarzi H, Liu X, Nguyen H C B, *et al.* Endogenous tRNA-derived fragments suppress breast cancer progression via YBX1 displacement. *Cell*, 2015, **161**(4): 790-802
- [19] Yamasaki S, Ivanov P, Hu G F, *et al.* Angiogenin cleaves tRNA and promotes stress-induced translational repression. *J Cell Biol*, 2009, **185**(1): 35-42
- [20] Schober P, Vetter T R. Logistic regression in medical research. *Anesth Analg*, 2021, **132**(2): 365-366
- [21] Pate A, Riley R D, Collins G S, *et al.* Minimum sample size for developing a multivariable prediction model using multinomial logistic regression. *Stat Methods Med Res*, 2023, **32**(3): 555-571
- [22] Hu J, Szymczak S. A review on longitudinal data analysis with random forest. *Brief Bioinform*, 2023, **24**(2): bbad002

- [23] Jin Y, Lan A, Dai Y, *et al.* Development and testing of a random forest-based machine learning model for predicting events among breast cancer patients with a poor response to neoadjuvant chemotherapy. *Eur J Med Res*, 2023, **28**(1): 394
- [24] Wallace M L, Mentch L, Wheeler B J, *et al.* Use and misuse of random forest variable importance metrics in medicine: demonstrations through incident stroke prediction. *BMC Med Res Methodol*, 2023, **23**(1): 144
- [25] Feng H, Wang F, Li N, *et al.* A random forest model for peptide classification based on virtual docking data. *Int J Mol Sci*, 2023, **24**(14): 11409
- [26] Hammann F, Drewe J. Decision tree models for data mining in hit discovery. *Expert Opin Drug Discov*, 2012, **7**(4): 341-352
- [27] Wilson A. CORR synthesis: can decision tree learning advance orthopaedic surgery research. *Clin Orthop Relat Res*, 2023, **481**(12): 2337-2342
- [28] Botha D, Steyn M. The use of decision tree analysis for improving age estimation standards from the acetabulum. *Forensic Sci Int*, 2022, **341**: 111514
- [29] Riina N, Harris A, Siesky B A, *et al.* Using multi-layer perceptron driven diagnosis to compare biomarkers for primary open angle glaucoma. *Invest Ophthalmol Vis Sci*, 2024, **65**(11): 16
- [30] Fan Y, Wang W. Using multi-layer perceptron to identify origins of replication in eukaryotes *via* informative features. *BMC Bioinformatics*, 2021, **22**(1): 516
- [31] Yang X, Zheng Y, Xing X, *et al.* Immune subtype identification and multi-layer perceptron classifier construction for breast cancer. *Front Oncol*, 2022, **12**: 943874
- [32] Yang W, Gao K, Qian Y, *et al.* A novel tRNA-derived fragment AS-tDR-007333 promotes the malignancy of NSCLC *via* the HSPB1/MED29 and ELK4/MED29 axes. *J Hematol Oncol*, 2022, **15**(1): 53
- [33] Qian T, Yu X, Xu A, *et al.* tRF-20-S998LO9D inhibits endometrial carcinoma by upregulating SESN2. *Epigenomics*, 2022, **14**(24): 1563-1577
- [34] Guzzi N, Muthukumar S, Cieřla M, *et al.* Pseudouridine-modified tRNA fragments repress aberrant protein synthesis and predict leukaemic progression in myelodysplastic syndrome. *Nat Cell Biol*, 2022, **24**(3): 299-306
- [35] Guzzi N, Cieřla M, Ngoc P C T, *et al.* Pseudouridylation of tRNA-derived fragments steers translational control in stem cells. *Cell*, 2018, **173**(5): 1204-1216.e26

tRF Prospect: tRNA-derived Fragment Target Prediction Based on Neural Network Learning*

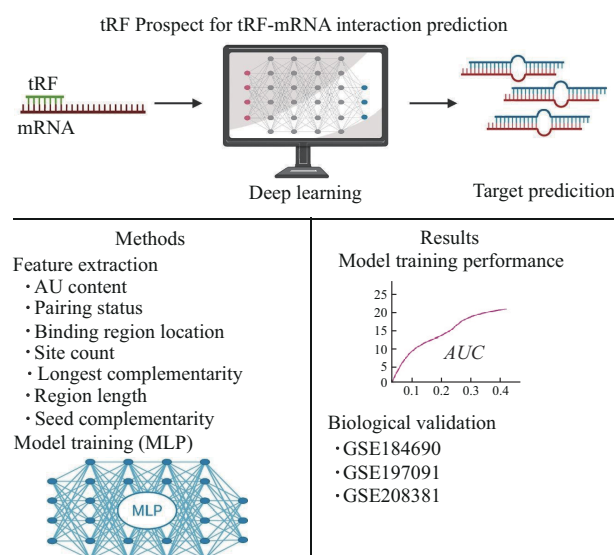
REN Dai-Xi^{1,2)}**, YI Jian-Yong³⁾**, MO Yong-Zhen¹⁾, YANG Mei¹⁾, XIONG Wei¹⁾,
ZENG Zhao-Yang^{1,2)}, SHI Lei²⁾***

¹⁾Xiangya School of Basic Medical Sciences, Central South University, Changsha 410078, China;

²⁾Department of Pathology, the Second Xiangya Hospital, Central South University, Changsha 410011, China;

³⁾Institutes of Brain Science, Fudan University, Shanghai 200032, China)

Graphical abstract



Abstract Objective Transfer RNA-derived fragments (tRFs) are a recently characterized and rapidly expanding class of small non-coding RNAs, typically ranging from 13 to 50 nucleotides in length. They are derived from mature or precursor tRNA molecules through specific cleavage events and have been implicated in a wide range of cellular processes. Increasing evidence indicates that tRFs play important regulatory roles in gene expression, primarily by interacting with target messenger RNAs (mRNAs) to induce transcript degradation, in a manner partially analogous to microRNAs (miRNAs). However, despite their emerging biological relevance and potential roles in disease mechanisms, there remains a significant lack of computational tools capable of systematically predicting the interaction landscape between tRFs and their target mRNAs. Existing databases

* This work was supported by grants from The National Natural Science Foundation of China (82472789), Natural Science Foundation of Hunan Province (2025JJ50590), Natural Science Foundation of Changsha (kq2208329), and Graduate Student Scientific Research Innovation Project of Central South University (1053320242771).

** These authors contributed equally in this work.

*** Corresponding author.

Tel: 86-731-88660449, E-mail: shilei81@csu.edu.cn

Received: February 9, 2025 Accepted: July 14, 2025

often rely on limited interaction features and lack the flexibility to accommodate novel or user-defined tRF sequences. The primary goal of this study was to develop a machine learning based prediction algorithm that enables high-throughput, accurate identification of tRF:mRNA binding events, thereby facilitating the functional analysis of tRF regulatory networks. **Methods** We began by assembling a manually curated dataset of 38 687 experimentally verified tRF:mRNA interaction pairs and extracting seven biologically informed features for each pair: (1) AU content of the binding site, (2) site pairing status, (3) binding region location, (4) number of binding sites per mRNA, (5) length of the longest consecutive complementary stretch, (6) total binding region length, and (7) seed sequence complementarity. Using this dataset and feature set, we trained 4 distinct machine learning classifiers—logistic regression, random forest, decision tree, and a multilayer perceptron (MLP)—to compare their ability to discriminate true interactions from non-interactions. Each model's performance was evaluated using overall accuracy, receiver operating characteristic (ROC) curves, and the corresponding area under the ROC curve (AUC). The MLP consistently achieved the highest *AUC* among the four, and was therefore selected as the backbone of our prediction framework, which we named tRF Prospect. For biological validation, we retrieved 3 high-throughput RNA-seq datasets from the gene expression omnibus (GEO) in which individual tRFs were overexpressed: AS-tDR-007333 (GSE184690), tRF-3004b (GSE197091), and tRF-20-S998LO9D (GSE208381). Differential expression analysis of each dataset identified genes downregulated upon tRF overexpression, which we designated as putative targets. We then compared the predictions generated by tRF Prospect against those from three established tools—tRFTar, tRFforest, and tRFtarget—by quantifying the number of predicted targets for each tRF and assessing concordance with the experimentally derived gene sets. **Results** The proposed algorithm achieved high predictive accuracy, with an *AUC* of 0.934. Functional validation was conducted using transcriptome-wide RNA-seq datasets from cells overexpressing specific tRFs, confirming the model's ability to accurately predict biologically relevant downregulation of mRNA targets. When benchmarked against established tools such as tRFTar, tRFforest, and tRFtarget, tRF Prospect consistently demonstrated superior performance, both in terms of predictive precision and sensitivity, as well as in identifying a higher number of true-positive interactions. Moreover, unlike static databases that are limited to precomputed results, tRF Prospect supports real-time prediction for any user-defined tRF sequence, enhancing its applicability in exploratory and hypothesis-driven research. **Conclusion** This study introduces tRF Prospect as a powerful and flexible computational tool for investigating tRF:mRNA interactions. By leveraging the predictive strength of deep learning and incorporating a broad spectrum of interaction-relevant features, it addresses key limitations of existing platforms. Specifically, tRF Prospect: (1) expands the range of detectable tRF and target types; (2) improves prediction accuracy through multilayer perceptron model; and (3) allows for dynamic, user-driven analysis beyond database constraints. Although the current version emphasizes miRNA-like repression mechanisms and faces challenges in accurately capturing 5'UTR-associated binding events, it nonetheless provides a critical foundation for future studies aiming to unravel the complex roles of tRFs in gene regulation, cellular function, and disease pathogenesis.

Key words tRF, database, neural network learnin

DOI: 10.3724/j.pibb.2025.0059

CSTR: 32369.14.pibb.20250059