DOI: 10.11992/tis.201806008
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180629.0847.002.html

多特征融合的 lncRNA 识别与其功能预测

常征, 孟军, 施云生, 莫冯然
(大连理工大学 计算机科学与技术学院, 辽宁 大连 116023)

摘 要: 针对传统的基于单一特征的植物 lncRNA 识别的局限性, 提出了融合 RNA 序列的开放阅读框、二级结构以及 k-mers 等多特征方法, 训练高斯朴素贝叶斯、支持向量机和梯度提升决策树 3 种经典的分类模型, 并实现分类结果的集成, 利用交叉验证对模型的性能进行了评估,整体性能优于目前较流行的 CPAT、CNCI 和 PLEK 预测软件, 在拟南芥数据集上总体的准确率达到了 89%。另外, 基于内源性竞争规则以及 RNA 结构信息, 分别对 lncRNA-microRNA 和 microRNA-mRNA 进行靶向预测、筛选, 再通过整合预测数据建立互作网络, 并对网络模块中的 lncRNA 进行功能预测。通过 GO 术语分析, 对与 mRNA 相关的 lncRNA 可能参与的生物调控过程进行预测, 推测它们的相应功能。

关键词: lncRNA; 识别; 特征提取; 多特征融合; 机器学习; 互作关系; 网络构建; 功能预测
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2018)06−0928−07

中文引用格式: 常征, 孟军, 施云生, 等. 多特征融合的 lncRNA 识别与其功能预测[J]. 智能系统学报, 2018, 13(6): 928–934.
英文引用格式: CHANG Zheng, MENG Jun, SHI Yunsheng, et al. LncRNA recognition by fusing multiple features and its function prediction[J]. CAAI transactions on intelligent systems, 2018, 13(6): 928–934.

LncRNA recognition by fusing multiple features and its function prediction

CHANG Zheng, MENG Jun, SHI Yunsheng, MO Fengran
(School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China)

Abstract: Considering the limitations of the traditional plant lncRNA identification based on a single feature, in this paper, a method, in which the open reading frame, secondary structure, and k-mers features of RNA sequences are integrated, is proposed. It involves the training of three classical classification models, Gaussian naive Bayes, support vector machines, and gradient lifting decision tree, and integrating the classification results. The performance of the method was evaluated using cross-validation, and it exhibited superior performance. The accuracy of the proposed method reached 89% when tested with the Arabidopsis thaliana dataset. Using the same dataset, the proposed method outperformed the popular CPAT, CNCI, and PLEK prediction software. In addition, based on the endogenous competition rules and RNA structure information, target prediction and filter rules for lncRNA-microRNA and microRNA-mRNA pairs were executed, and then related tools were used to establish RNA interaction regulatory networks, and the regulatory relationship was analyzed to predict the functions of lncRNAs in modules. Through Gene Ontology term analysis, the possible biological regulation function of lncRNAs can be predicted, and their corresponding functions can be inferred.

Keywords: lncRNA; identification; feature extraction; multiple features fusion; machine learning; interrelationship; network construction; function prediction

近年来, 非编码 RNA(non-coding RNA, ncRNA) 识别的相关研究已成为人们关注的热点。一直以来, 转录本被大家普遍认为只起到翻译蛋白质的作用, 但随着人类基因组注释工作的

收稿日期: 2018−06−04. **网络出版日期**: 2018−06−29.
基金项目: 国家自然科学基金项目 (61472061); 大连理工大学研究生教改基金项目 (Jg2017015); 大连理工大学大学生创新训练项目 (201810141020101 1019).
通信作者: 孟军. E-mail: mengjun@dlut.edu.cn.

不断推进, 研究结果表明只有大约 1%~2% 的基因参与了编码蛋白的工作[1], 而以往被大家忽略的非编码序列也在整个生命活动中扮演着至关重要的角色。这些非编码序列中, 有一种长度大于 200 nt、无法编码蛋白质的转录本尤其受到关注, 被称为长链非编码 RNA(long non-coding RNA, lncRNA)[2]。近年来发现 lncRNA 具有调节生物体生命活动的重要作用[3-4], 而各种传统的实验方法, 一方面需要花费大量时间和高额费用, 另一方面, 因为 lncRNA 的低表达和低保守性等原因, 在识别 lncRNA 方面受到不同程度的影响。研究人员对人和动物进行了大量的实验, 并且出现了具有良好鲁棒性的 lncRNA 识别软件。

RNAseq 和全基因组阵列分析显示, 植物体内也存在大量的 lncRNA, 它们在植物的开花、雄性不育、营养代谢、生物和非生物胁迫等生物过程中起着调节因子的作用[5]。与哺乳动物相比, 植物 ncRNA 的研究起步比较晚, 且多数集中在短链非编码 RNA 上, 这为植物 lncRNA 识别与分析带来了困难。研究植物 lncRNA 将帮助生命学科的工作者进一步揭示植物内部生命活动, 因此深入研究植物 lncRNA 并预测其功能具有非常重要的意义。

目前, 在计算预测 lncRNA 方面, 许多研究工作都利用机器学习算法建立预测模型, 通过输入各类序列特征、结构特征, 构建识别 lncRNA 的分类器模型。研究表明, 对于 lncRNA 识别, 通过提取开放阅读框、密码子频率偏好性、与已知蛋白质相似度等特征作为输入, 对线性回归、支持向量机以及其他模型进行训练得到的分类器具有良好的分类效果[6]。近年来衍生出的预测软件多采用以上特征。其中, CPC[7]和 CPAT[8]都是通过序列特征来区分编码和非编码 RNA; CNCI[9]能够将训练好的分类器运用到近亲物种的 lncRNA 识别; PLEK[10]可以从高通量测序的转录本中识别 lncRNA。然而, 大多数软件只在动物数据集上得到良好的验证, 专门为植物 lncRNA 识别设计的软件目前还比较稀缺。

随着基因组学研究的不断深入, 产生了大量未被标注的基因序列。由于生物实验方法验证基因功能的代价十分昂贵, 如何通过计算机方法对基因序列功能进行大规模预测成了近年来生物信息学的研究热点之一[11]。

为了进一步提高植物 lncRNA 预测的准确性, 基于机器学习分类算法, 通过对下载的高可信度数据提取开放阅读框、k 核苷酸频率以及二级结构特征等多特征融合[12]作为输入特征, 训练朴素贝叶斯、支持向量机和梯度提升决策树 3 种分类模型, 并采用加权投票的多分类器集成方法, 集成分类结果以得到更好的分类性能。利用标注测序数据测试集对模型进行验证、分析并选择性能最好的将其作为最终分类器。提出的方法通过五折交叉验证, 得到了较好的性能。在功能预测方面, 根据 lncRNA-microRNA、microRNA-mRNA 相互作用关系, 建立调控网络, 利用相关联的 RNA 预测 lncRNA 的功能。

1　基于多特征融合的 lncRNA 预测

1.1　数据集

拟南芥的生物学实验数据和基因注释信息相对比较丰富, 常被广泛用于植物胁迫响应的研究中[13]。本文使用的正集数据为 PNRD[14](http://structuralbiology.cau.edu.cn/PNRD/) 2 565 条具有高可信度的拟南芥 lncRNA 序列。负集数据是从 RefSeq 数据库下载的 48 148 条 mRNA 序列。为了保证正负样本均衡, 从负集原始数据中随机采样出 2 500 条 mRNA 作为最终训练集, 如表 1 所示。

表 1　数据集信息
Table 1　Dataset information

数据集	数目	数据库
lncRNA	2 565	PNRD
mRNA	2 500	RefSeq

1.2　开放阅读框

在分子遗传学中, 开放阅读框 (open reading frame, ORF) 是阅读框的一部分, 具有潜在的翻译能力[15]。研究表明, mRNA 的 ORF 覆盖率明显高于 lncRNA, 且 mRNA 具有更多的完整性开放阅读框[16]。首先利用 TransDecoder 软件包计算得到每个序列的开放阅读框信息, 然后对其分别提取完整性、ORF 覆盖率以及归一化的 ORF 值 3 种特征。将完整性定义为一个布尔变量 bool, 0 代表不存在完整性 ORF, 1 表示存在完整性 ORF。覆盖率 Coverage 等于所有的合法 ORF(本文只考虑正链情况下) 的长度与 RNA 序列长度之比, 归一化 ORF Normalized_ORF 是序列中 ORF 个数 n 与 RNA 序列长度 L 的比值, 分别定义为

$$\text{Coverage} = \frac{\sum_{i=1}^{n} L_i}{L} \tag{1}$$

$$\text{Normalized\_ORF} = \frac{n}{L} \tag{2}$$

式中 L_i 代表序列中第 i 个 ORF 的长度。

整合 3 种特征得到特征向量:

$$\boldsymbol{V}_{\text{ORF}} = [\text{bool}\ \ \text{Coverage}\ \ \text{Normalized\_ORF}] \tag{3}$$

1.3 二级结构

二级结构 (secondary structures, SS) 是单条序列通过碱基配对自身形成茎区和环球，与 RNA 的功能息息相关，可以作为识别 lncRNA 的重要依据。目前，预测二级结构的计算方法分为两类：基于热力学和基于系统发生学。前者认为：生物体在形成高维结构的时候，将使自身达到稳态结构，因此释放的能量应更多。研究表明，二级结构越是稳定 (释放的自由能越多) 其潜在的编码能力越强。同时，二级结构的稳定性与 RNA 序列中配对碱基个数以及核苷酸 C 和 G 的含量有一定的关系。本文使用 ViennaRNA[17]工具包对序列形成二级结构释放的自由能进行计算并得到二级结构的点括号表示形式，然后从中提取出配对碱基的个数以及 C 和 G 碱基的含量。归一化最小自由能由如下：

$$\text{Normalized\_MFE} = \frac{\text{MFE}}{L} \tag{4}$$

式中：MFE 是释放的自由能，L 是 RNA 序列的长度。

整合上述 3 个特征得到如下特征向量：

$$\boldsymbol{V}_{SS} = [\text{Normalized\_MFE}\ \ n_p\ \ \%(C+G)] \tag{5}$$

式中 n_p 为配对碱基的个数。

1.4 k-mers

密码子是遗传物质编码的信息由活细胞转化为蛋白质的一套规则，蛋白质为保证其某些生物功能，在自然选择下会表现出对某些密码子的偏好性。因此 mRNA 在密码子方面表现出一定的保守性，而不编码蛋白的 lncRNA 其保守性较差。所以可以使用密码子频率当作识别 lncRNA 的一个特征。然而，因为无法准确定位 mRNA 编码区域，且 lncRNA 有多个编码区域，直接计算密码子频率存在一定困难。为解决以上问题，使用一个近似的解决方法：k-mers 特征计算。

一个 k-mer 具有 k 个核苷酸，每个核苷酸可以是 A、C、G 或 T，k 取值为 1、2 和 3，则有 4+16+64=84 种模式：4 个 1-mer，16 个 2-mer，64 个 3-mer。使用一个长度为 k 的滑动窗口来匹配上述 k-mer。滑动窗口沿 RNA 序列以步长为 1 核苷酸进行滑动匹配，使用 c_i 表示匹配到的次数 ($i = 1,2,\cdots,84$)，并且为每个 k-mer 分配一个系数 w_k，从而使得每类频率对预测效果的影响一样，具体如下：

$$f_i = w_k \frac{c_i}{s_k}, k = 1,2,3,\ i = 1,2,\cdots,84 \tag{6}$$

$$s_k = L - k + 1,\ k = 1,2,3 \tag{7}$$

$$w_k = \frac{1}{4^{3-k}},\ k = 1,2,3 \tag{8}$$

式中：s_k 是总的匹配次数，L 为 RNA 序列长度，f_i 为每种 k-mer 出现的频率，则得到二级结构特征向量：

$$\boldsymbol{V}_{\text{k-mer}} = [f_1\ \ f_2\ \ f_3\ \ \cdots\ \ f_{84}] \tag{9}$$

本文选择融合上述 3 类特征组成含 90 维的特征集作为最终的特征向量：

$$\boldsymbol{V} = [\boldsymbol{V}_{\text{ORF}}\ \ \boldsymbol{V}_{\text{SS}}\ \ \boldsymbol{V}_{\text{k-mer}}] \tag{10}$$

1.5 构建分类模型

朴素贝叶斯方法 (Naive Bayes) 是基于贝叶斯定理的监督学习算法，即简单地假设每对特征之间相互独立。相比于其他更复杂的方法，朴素贝叶斯学习器和分类器非常快，并且有助于解决高维数据问题。支持向量机是一种基于统计学习的分类方法[18]，其模型参数确定会对应到一个凸最优化问题，因此可以保证得到最优解。目前流行的 CPC、CNCI 等软件都使用 SVM 作为分类器。梯度提升决策树 (gradient boost decision tree, GBDT) 是对于任意可微损失函数的提升算法的泛化，它具有强大的预测能力以及在输出空间中对异常点的鲁棒性。

结合训练集高维度、非连续等特征，以及模型自身的鲁棒性，本文选择基于高斯分布的朴素贝叶斯模型、支持向量机以及梯度提升决策树 3 个模型进行训练。然后使用网格搜索法分别调整 3 个分类器的超参数。并且采用加权投票分法来融合上述 3 个分类模型的输出得到最终的预测结果。

1.6 性能评价指标

本文选择使用准确率 (Accuracy, ACC)、精确率 (Precison, P)、召回率 (Recall, R)、F_1 值 (F_1_score) 来评估训练出的分类模型。定义如下：

$$\text{ACC} = \frac{\text{TP}+\text{TN}}{\text{TN}+\text{FP}+\text{TP}+\text{FN}} \tag{11}$$

$$P = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{12}$$

$$R = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{13}$$

$$F_1\text{\_score} = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \tag{14}$$

式中：TP 指将正类预测为正类数，FN 指将正类预测为负类数，FP 指将负类预测为正类数，TN 指将负类预测为负类数。

2 LncRNA 功能预测

2.1 数据集

用于构建互作网络的 microRNA 序列是从 miRBase[19](http://www.mirbase.org/index.shtml) 下载的 427 条成熟拟南芥 microRNA 序列。lncRNA

以及 mRNA 则选用上述下载的具有高可信度的 2 565 条 lncRNA 与 2 500 条 mRNA。

2.2　靶向预测提取互作对

研究证明, 胁迫作用下, 植物的性状将发生改变, 而这个过程是由多个基因相互作用形成的, lncRNA 也参加其中。作为竞争性内源 RNA 或者 microRNA 内源性模拟靶标的 lncRNA, 可以有效抑制 microRNA 的功能, 从而间接作用 mRNA 影响生物形状及蛋白表达[20]。首先使用 RNAhybrid[21] 预测 lncRNA-microRNA 相互作用对。本文综合考虑 microRNA 靶标结合的特征, 设置 RNAhybrid 参数: 最小自由能−25 kcal/mol, 保证种子区域 2~8 位完全配对, *p*-value 值小于 0.05。然后, 在杂交区内结合 RNA 结构信息[22]进行筛选:

1) microRNA 序列 5'端开始的 10~12 位必须有突起点;

2) 中间位置突起点只允许包含 lncRNA 序列 2~4 个核苷酸;

3) 除了中间位置的突起外, microRNA 上的错配和 G:U 配对总数少于 4 并且连续错配小于 2。

利用 psRNAtarget 靶向预测工具预测出 microRNA-mRNA 相互作用对, 并且筛选出有 lncRNA 靶点的 microRNA 靶向 mRNA 的数据。

2.3　构建调控网络与功能预测

融合两类相互作用对, 基于 Cytoscape[23]工具包构建初级的 lncRNA-microRNA-mRNA 互作网络, 然后对该网络进行模块分析, 利用 GO[24]数据库中的术语了解基因特性。这些 GO 术语被划分为 3 类: 细胞成分 (cellular component), 分子功能 (molecular function) 和生物过程 (biological process), 因此可以基于 GO 术语对各个模块进行注释预测 lncRNA 的功能。

3　实验结果与分析

3.1　标注数据测试集验证结果

交叉验证 (cross validation, CV) 是一种模型验证技术, 把给定的数据进行切分, 将切分的数据集组合为训练集与测试集, 用于验证模型的泛化能力, 有效降低模型的过拟合问题。交叉验证方法可分为简单交叉验证法、K 折交叉验证以及留一交叉验证。其中, 应用最多的是 K 折交叉验证。标注数据即从公共数据库采集到的被验证与标注的序列数据。为了减少计算消耗的时间并评估分类模型的泛化性能, 本文直接对分类模型在数据集上进行 5 折交叉验证, 得到的性能效果作为标注数据测试集的测试结果。

为了验证本文提出方法的有效性, 使用目前比较流行的 CPAT、CNCI 与 PLEK 软件在本文采用的数据集上进行分类预测, 将得到的结果进行比较。CPAT 使用逻辑回归模型; CNCI 通过分析序列的内在组成来区分蛋白编码和非蛋白转录本, 使用 ATN 分数矩阵以及序列结构两类特征; PLEK 使用 k-mer 和滑动窗口来分析转录本, 选取 k-mers 频率作为其特征。后两款软件都使用支持向量机作为其分类器, 结果如表 2 所示。可以看出, 本文提出的基于多特征融合的集成方法在精确率上超过 90%, 优于 CPAT、CNCI 与 PLEK; 召回率分别比 CPAT、CNCI、PLEK 高出 6.8%、7.4% 和 8.8%; F_1 得分也优于另外三者。这些结果表明本文提出的方法可以有效地预测植物 lncRNA。

表 2　基于不同方法的分类结果比较

Table 2　Classification results comparison based on different methods

方法	精确率	召回率	F_1 值	准确率
CPAT	0.898	0.810	0.852	0.857
CNCI	0.847	0.804	0.825	0.827
PLEK	0.671	0.790	0.738	0.714
Our	0.914	0.878	0.888	0.890

为了进一步验证本文基于多特征融合所构建的分类模型的有效性, 分别给出单独使用开放阅读框、二级结构、k-mers 作为特征训练分类器得出的预测结果, 交叉验证的结果如表 3 所示。

表 3　基于不同特征的分类结果比较

Table 3　Classification results comparison based on different features

方法	精确率	召回率	F_1 值	准确率
ORF	0.848	0.816	0.816	0.828
SS	0.728	0.716	0.722	0.720
k-mers	0.828	0.794	0.810	0.814
Fusion	0.914	0.878	0.888	0.890

可以看出, 本文提出的方法整体的准确率为 89.0%, 比单独使用开放阅读框、二级结构、k-mers 要分别高出 6.2%、17%、7.6%, 这表明提出的方法对于识别 lncRNA 相较于使用单一类特征是有效的。并且可以看出, 使用 ORF 得到的预测结果要优于其他两类, 这意味着 ORF 在识别 lncRNA 上具有更好的区分度。

3.2　网络构建与功能预测

经过两个靶向预测软件包的预测并且对预测结果按上述规则进行筛选后得到数据如表 4 ~ 表 5 所示。

表 4　筛选后的 microRNA-lncRNA 靶点数据
Table 4　Filtered microRNA-lncRNA target data

靶点数量	microRNA	lncRNA
108	81	70

表 5　筛选后的 microRNA-mRNA 靶点数据
Table 5　Filtered microRNA-mRNA target data

靶点数量	microRNA	mRNA
853	81	421

融合以上两类数据构建的初级调控网络如图 1 所示。

调控网络中的每个模块以 microRNA 为中心，形成 microRNA 同时与 lncRNA、mRNA 相互作用的调控子网络。调控子网络根据 RNA 作用数目和类型的不同可以分为：1) 单 microRNA 作用网络，即单个 microRNA 作为结点与 lncRNA、mRNA 相互作用，但与网络中其他 microRNA 结点没有联系，如图 2 所示；2) 多 microRNA 相互作用网络，不同的 microRNA 通过靶向同一个 mRNA、lncRNA 形成相互作用的模块，如图 3。

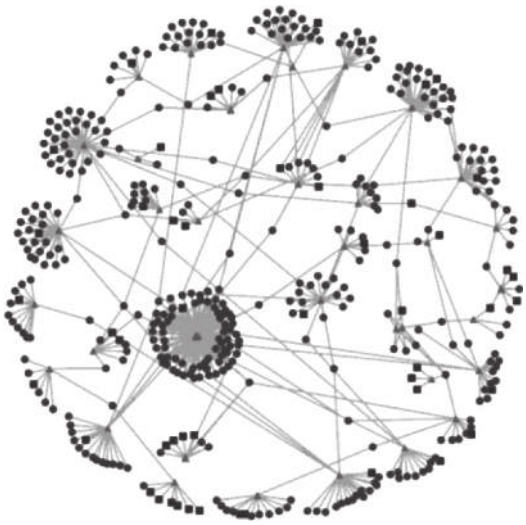
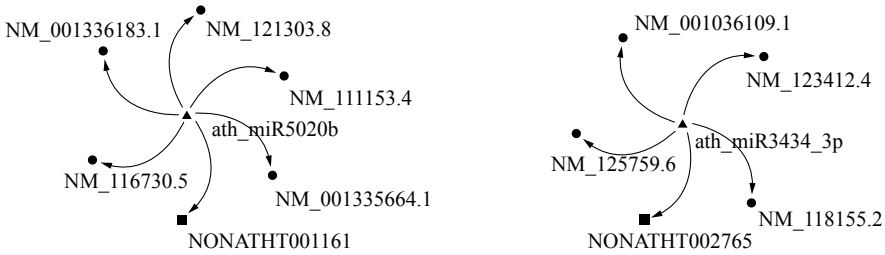


图 1　拟南芥初级调控网络 (三角形代表 microRNA，矩形代表 lncRNA，圆形代表 mRNA)

Fig. 1　Primary regulatory network of Arabidopsis(Triangles represent microRNAs, rectangles represent lncRNAs, and circles represent mRNAs)



(a) ath_miR5020b 作用网络　　(b) ath_miR3434_3p 作用网络

图 2　单 microRNA 作用网络

Fig. 2　Single microRNA interaction network

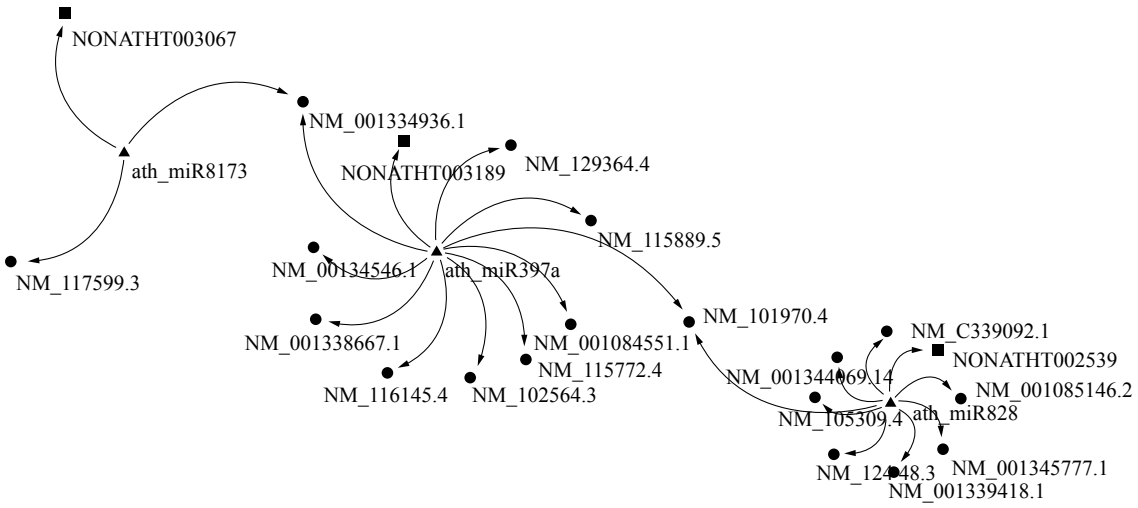


图 3　多 microRNA 作用网络

Fig. 3　Multiple microRNA interaction network

在构建调控网络并进行模块分析后, 使用 GO 术语检查模块中的 mRNA 的功能注释, 并对和 mRNA 相关的 lncRNA 可能参与的生物调控过程进行预测, 部分结果如表 6。可以看到根据相关联的 RNA, 本文预测的 lncRNA 所具有的生物调控功能。例如 NONATHT002539 参与到氮化合物代谢、分解代谢以及生物合成过程; NONATHT000372 促进蛋白质磷酸化; NONATHT002765 和 NONATHT002470、NONATHT002469 都会影响细胞转化的过程等。

表 6　lncRNA 功能预测

Table 6　lncRNA function prediction

microRNA	基因	生物过程	lncRNA
ath_miR3434_3p	AT5G63640; AT1G54220; AT4G20360	lysosomal transport; protein targeting; cellular process	NONATHT002765
ath_miR844_5p	AT5G24940	Protein phosphorylation	NONATHT000372
ath_miR399c_5p	AT3G14460; AT3G08960	anion transport; protein localization; cellular process	NONATHT002470; NONATHT002469
ath_miR8173	AT1G79920; AT4G15100	protein complex assembly; proteolysis	NONATHT003067
ath_miR397a	AT1G79920 AT1G21160	protein complex assembly; biosynthetic process	NONATHT003189
ath_miR828	AT1G21160; AT1G66380AT5G54670; AT3G43210	biosynthetic process; catabolic process; nitrogen compound metabolic process	NONATHT002539

4　结束语

本文基于植物 RNA 序列, 提取开放阅读框、二级结构和 k-mers 3 类特征, 并将它们融合成一个 90 维的特征向量作为输入, 训练朴素贝叶斯、支持向量机、梯度提升决策树 3 种机器学习模型, 并采用加权投票分法来集成分类结果。通过与现有的识别软件 CNCI 和 PLEK 相比, 本文提出方法取得了较好的性能, 可以有效地识别预测植物 lncRNA。基于内源性竞争规则, 筛选 lncRNA-microRNA、microRNA-mRNA 相互作用数据, 并整合两类数据构建调控网络, 基于互作网络利用 GO 术语对各个模块的 mRNA 注释, 进而通过 mRNA 预测 lncRNA 功能。未来将结合深度学习技术, 进一步改善预测的准确率。

参考文献:

[1] COSTA F F. Non-coding RNAs: meet thy masters[J]. Bioassays, 2010, 32(7): 599–608.

[2] PALAZZO A F, LEE E S. Non-coding RNA: what is functional and what is junk?[J]. Frontiers in genetics, 2015, 6: Article No.2.

[3] SCHMITZ S U, GROTE P, HERRMANN B G. Mechanisms of long noncoding RNA function in development and disease[J]. Cellular and molecular life sciences, 2016, 73(13): 2491–2509.

[4] O'LEARY V B, OVSEPIAN S V, CARRASCOSA L G, et al. PARTICLE, a triplex-forming long ncRNA, regulates locus-specific methylation in response to low-dose irradiation[J]. Cell reports, 2015, 11(3): 474–485.

[5] CUI Jun, LUAN Yushi, JIANG Ning, et al. Comparative transcriptome analysis between resistant and susceptible tomato allows the identification of lncRNA16397 conferring resistance to Phytophthora infestans by co-expressing glutaredoxin[J]. The plant journal, 2017, 89(3): 577–589.

[6] HAN Siyu, LIANG Yanchun, LI Ying, et al. Long noncoding RNA identification: comparing machine learning based tools for long noncoding transcripts discrimination[J]. BioMed research international, 2016, 2016: Article No. 8496165.

[7] KONG Lei, ZHANG Yong, YE Zhiqiang, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine[J]. Nucleic acids research, 2007, 36(S2): W345–W349.

[8] WANG Liguo, PARK H J, DASARI S, et al. CPAT: coding-potential assessment tool using an alignment-free logistic regression model[J]. Nucleic acids research, 2013, 41(6): Article No.e74.

[9] SUN Liang, LUO Haitao, BU Dechao, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts[J]. Nucleic acids research, 2013, 41(17): Article No.e166.

[10] LI Aimin, ZHANG Junying, ZHOU Zhongyin. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme[J]. BMC bioinformatics, 2014, 15: Article No.311.

[11] 郭杏莉, 高琳, 刘永轩, 等. 长非编码 RNA 生物特征研

究与分析[J]. 科学通报, 2013, 58(27): 2779–2786.
GUO Xingli, GAO Lin, LIU Yongxuan, et al. Research and analysis of biocharacteristics of long non-coding RNAs[J]. Chinese science bulletin, 2013, 58(27): 2779–2786.

[12] 李同宇, 李卫军, 覃鸿. 基于特征融合的人脸图像性别识别[J]. 智能系统学报, 2013, 8(6): 505–511.
LI Tongyu, LI Weijun, QIN Hong. Facial image gender recognition method based on feature fusion[J]. CAAI transactions on intelligent systems, 2013, 8(6): 505–511.

[13] KARIM S. Exploring plant tolerance to biotic and abiotic stresses[D]. Uppsala, Sweden: Swedish University of Agricultural Sciences, 2007: 18–23.

[14] YI Xin, ZHANG Zhenhai, LING Yi, et al. PNRD: a plant non-coding RNA database[J]. Nucleic acids research, 2015, 43(D1): D982–D989.

[15] DINGER M E, PANG K C, MERCER T R, et al. Differentiating protein-coding and noncoding RNA: challenges and ambiguities[J]. PLoS computational biology, 2008, 4(11): Article No.e1000176.

[16] FRITH M C, BAILEY T L, KASUKAWA T, et al. Discrimination of non-protein-coding transcripts from protein-coding mRNA[J]. RNA biology, 2006, 3(1): 40–48.

[17] LORENZ R, BERNHART S H, HÖNER ZU SIEDERDISSEN C, et al. ViennaRNA package 2.0[J]. Algorithms for molecular biology, 2011, 6: Article No.26.

[18] 王振武, 孙佳骏, 尹成峰. 改进粒子群算法优化的支持向量机及其应用[J]. 哈尔滨工程大学学报, 2016, 37(12): 1728–1733.
WANG Zhenwu, SUN Jiajun, YIN Chengfeng. A support vector machine based on an improved particle swarm optimization algorithm and its application[J]. Journal of Harbin engineering university, 2016, 37(12): 1728–1733.

[19] GRIFFITHS-JONES S, GROCOCK R J, VAN DONGEN S, et al. miRBase: microRNA sequences, targets and gene nomenclature[J]. Nucleic acids research, 2006, 34(S1): D140–D144.

[20] CESANA M, CACCHIARELLI D, LEGNINI I, et al. A long noncoding RNA controls muscle differentiation by functioning as a Competing Endogenous RNA[J]. Cell, 2011, 147(2): 358–369.

[21] KRÜGER J, REHMSMEIER M. RNAhybrid: microRNA target prediction easy, fast and flexible[J]. Nucleic acids research, 2006, 34(S2): W451–W454.

[22] WU Huajun, WANG Zhimin, WANG Meng, et al. Widespread long noncoding RNAs as endogenous target mimics for microRNAs in plants[J]. Plant physiology, 2013, 161(4): 1875–1884.

[23] SHANNON P, MARKIEL A, OZIER O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks[J]. Genome research, 2003, 13(11): 2498–2504.

[24] ASHBURNER M, BALL C A, BLAKE J A, et al. Gene ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25–29.

作者简介：

常征，男，1995 年生，硕士研究生，主要研究方向为机器学习、数据挖掘和生物信息。

孟军，女，1964 年生，教授，博士生导师，博士，主要研究方向为机器学习、数据挖掘和大数据处理。主持参与国家自然科学基金、国家重大专项、教育部专项和省自然基金等项目。在国际 SCI 收录和国内核心期刊发表学术论文 70 余篇。

施云生，男，1994 年生，硕士研究生，主要研究方向为机器学习、数据挖掘和生物信息。