DOI:10.3969/j.issn.1672-5565.201709002

基于序列特征的环状 RNA 识别

周 晶[1],谢雪英[2,3*],顾万君[2,3*]
(1.东南大学 学习科学研究中心,南京 210096;2.生物电子学国家重点实验室,东南大学 生物科学与医学工程学院,南京 210096;3.生物医学工程国家级实验教学示范中心(东南大学),南京 210096)

摘 要:环状 RNA 是新发现的一类具有重要生物学功能的 RNA。现有的环状 RNA 识别工具依赖高通量测序数据,因数据本身和识别方式的弊端而普遍存在准确性不足、不同方法间重复性低以及假阳性率/假阴性率高等缺点。为了解决该问题,我们搭建模型来实现不依赖于测序数据而根据序列的内在特征的环状 RNA 从头预测。本文选取了包括剪接位点上下游内含子的长度、A-to-I 密度和 Alu 重复序列等 100 个与 RNA 成环相关的序列特征,建立了机器学习模型,并识别了人类基因组中的环状 RNA,比较了两种机器学习方法随机森林法(RF)和支持向量机(SVM)的分类效果。结果表明,所选序列特征能有效地鉴别 RNA 能否成环,同时,不同序列特征对模型的分类预测能力的贡献也不同。相比于 SVM 方法,RF 分类的效果更好。

关键词:环状 RNA;序列特征;机器学习;随机森林;支持向量机
中图分类号:Q522+.6 **文献标志码**:A **文章编号**:1672-5565(2018)02-113-06

Identification of circular RNAs using genomic sequence features

ZHOU Jing [1], XIE Xueying [2, 3*], GU Wanjun [2, 3*]
(1. *Research Center for Learning Sciences, Southeast University, Nanjing* 210096, *China*; 2. *State Key Laboratory of Bio-electronics, School of Biological Sciences and Medical Engineering, Southeast University, Nanjing* 210096, *China*; 3. *National Demonstration Center for Experimental Biomedical Engineering Education* (*Southeast University*), *Nanjing* 210096, *China*)

Abstract: Circular RNAs (circRNAs) are a class of novel RNAs with important biological functions. Currently, the identification tools of circRNAs are dependent on high-throughput sequencing. However, due to defects in data and their identification mode, low accuracy, low overlapping rate of different methods, high false positive rate, and false negative rate generally exist. To solve this problem, we built a model to identify circRNAs from the very beginning based on the inherent features of the genomic sequence rather than sequencing data. We selected 100 genomic sequence features related to circRNAs including the length of flanking introns, the density of A-to-I RNA editing sites, and the pairing score of Alu elements in the flanking introns, built machine learning model, identified the circRNAs in human genome, compared the classifying results of two machine learning algorithms, random forest (RF) and support vector machine (SVM). The results showed that the selected features could effectively identify circRNAs and different sequence features had different contributions to the identification of circRNAs. In addition, RF model had a better performance than SVM model in identifying RNAs.

Keywords: Circular RNAs; Sequence feature; Machine learning; Random forest; Support vector machines

环状 RNA 是区别于传统线性 RNA 的一类 RNA,具有闭合环状结构,不受 RNA 外切酶影响,不易降解,表达更稳定[1-2]。由于检测方法的限制,环状 RNA 自 1979 年在 Hela 细胞的细胞质中被首次发现以来的近三十年时间内并没有得到研究者的关注。然而,随着近年来高通量测序技术的广泛应用,成千上万的环状 RNA 在不同物种的细胞和组织中被广泛发现[1-2]。研究发现,人类细胞中存在大量

收稿日期:2017-09-10; **修回日期**:2018-01-19.
基金项目:国家自然科学基金(61372164, 61471112 , 61571109);江苏省重点研发计划(BE2016002-3);中央高校基本科研业务费专项资金(2242017K3DN04).
作者简介:周晶,女,硕士研究生,研究方向:生物信息学;E-mail:crys_jing@126.com.
*通信作者:谢雪英,女,副教授,研究方向:生物信息学;E-mail:xxying.rcls@seu.edu.cn.
顾万君,男,研究员,研究方向:生物信息学;E-mail:wanjungu@seu.edu.cn.

的环状 RNA。通过与基因注释信息比对,85%的环状 RNA 能匹配到已知基因上,其中的 84%与编码外显子重合[3]。进一步的研究发现,环状 RNA 通过多种途径在细胞中发挥着重要的生物学功能。一些环状 RNA 作为竞争性内源 RNA,可以调控基因表达[3];环状 RNA 也可以发挥海绵作用,吸附 miRNA,从而阻断后者对靶基因的抑制作用[4-5];环状 RNA 也可以通过与蛋白质结合,募集蛋白质复合体的组分,或调控蛋白质的活性[3,6];有些环状 RNA 甚至能够被翻译成蛋白质[7]。

目前,常见的环状 RNA 的识别工具(如 find_circ[3], CIRI[8], circRNAfinder[9], MapSplice[10], CIRCexplorer[11]等)都是从高通量测序数据中识别反向剪接位点(Back-splicing site)来实现环状 RNA 的识别。相关研究表明,上述 5 种环状 RNA 工具共同预测到的环状 RNA 比例很小,这些工具普遍存在较高的假阳性率和低灵敏度[12]。因此,如何充分利用序列本身的特征来区分环状 RNA 与其他线性 RNA(特别是编码蛋白的信使 RNA),是生物信息学研究中亟待解决的重要问题。近年来,序列特征与机器学习相结合的方法已成功地用于基因调控位点的预测、剪切位点的预测等生物学问题[13]。值得注意的是,前期研究已经表明基于序列特征的机器学习方法也能有效鉴别环状 RNA 与长链非编码 RNA[14]。

本文首先提取一些影响 RNA 成环过程的序列特征,并筛选了一些在环状 RNA 与线性 RNA 之间存在显著差异的序列特征,采用机器学习的分类算法,构建用于仅仅基于基因组序列预测外显子环状 RNA 的预测工具。测试结果表明,我们的方法能很好地区分外显子环状 RNA 与编码蛋白的 mRNA。

1 材料与方法

1.1 数据来源

为了训练和测试模型,我们共选取了 3 组数据集及 1 组独立测试集。每组数据集由正负样本集两部分组成,正样本集代表环状 RNA,负样本集代表线性蛋白质编码 RNA(见表 1)。

3 组数据集中正样本集构成如下:组 1:5 种常见的环状 RNA 预测工具对相同的 RNA 测序数据集进行环状 RNA 预测,选择同时被 5 种工具检出的 848 条环状 RNA[12]作为正样本。组 2:同样来自该项工作,选择 2 种工具及以上检出的 2 463 条环状 RNA。组 3:来源于 circBase 数据库的人类环状 RNA[15]。该数据库收集了不同研究小组的环状 RNA 识别结果,共包含了超过 9 万条环状 RNA 序列及其基因组注释信息。去除长度小于 200 nt 的短序列并且只保留同一基因的最长可变剪接转录本,最后得到 11 472 条环状 RNA。对于每一个正样本集,我们建立对应的负样本集。负样本集来源于 UCSC 人类基因组数据 hg19 版本[16]。首先,选取人类基因组中约 80 000 条编码蛋白的序列。然后,剔除与正样本集中的序列及 circBase 和 circNet[17] 中预测出的 circRNA 相重叠的序列,剩下共 9 976 条序列。最后,随机抽取与每组正样本集数量相近的序列作为该组对应的负样本集。

另外,由于经过实验验证的环状 RNA 数目太少(文献[18]中共收集到 282 个经过 PCR 验证的环状 RNA),不适合作为模型训练的数据集。因此,我们将实验验证的环状 RNA 作为独立数据集,仅用于模型预测能力的验证(见表 1)。

表 1 组数据集中正负样本集所对应 RNA 的数目

Table1 Corrernponding number of RNAs in positive and negative sample sets in the dataset

组别	正样本/条	负样本/条
1	848	850
2	2 463	2 500
3	11 472	9 976
独立数据集	282	300

1.2 特征筛选

现有研究结果表明,真核生物中 RNA 序列是否成环与其上下游序列有关。在内含子配对驱动 RNA 环化的过程中,5'端剪切位点存在 GU 重复序列,而对应的 3'剪切位点存在 AG 重复序列,作为上下游剪切信号[19]。

A-to-I RNA 编辑是在哺乳动物中普遍存在的 RNA 编辑方式,通过 ADAR 酶将某些特定位点的腺嘌呤(A)脱氨形成次黄嘌呤(I),在碱基配对中被识别作鸟嘌呤(G),影响碱基配对及位点识别[20]。人类基因组中常伴随双链 RNA 上 Alu 反向互补重复序列[21-22],而 Alu 重复序列的配对结合能在空间上拉近两端剪切位点,从而促进 RNA 环化[2]。对比环状 RNA 与线性 mRNA 上下游内含子中存在的 A-to-I 编辑位点及反向互补重复序列,结果显示在环状 RNA 的上下游内含子中的分布密度显著高于线性 mRNA,尤其体现在剪切位点附近[23]。同时,环状 RNA 上下游的内含子通常较长,且具有较多的 Alu 重复序列。这可能是由于内含子长度越长,在空间上更容易弯曲靠近,使得上下游 Alu 序列更容易结合,也就越容易发生反向剪接促进成环[2]。

另外,RNA 结合蛋白(RBP)与 RNA 成环也存

在着密切的联系。上下游内含子与 RNA 结合蛋白形成共聚体,容易在空间上拉近上下游剪切位点,促进反向剪接[3]。例如,QKI 结合蛋白能调控 RNA 剪切过程,促进反向剪接发生。环状 RNA 与具有相似侧翼内含子长度的线性 mRNA 的 QKI 结合位点分布存在着显著差异[24]。

基于这些先验知识,我们选取序列组分特征(包括内含子长度、G/C 含量、GT、AG 剪切信号分布密度),剪接位点上下游内含子上 A-to-I RNA 编辑分布密度、Alu 反向互补重复序列数量以及人类 94 种不同 RBP[25]结合位点数共 100 种特征,分别计算不同数据集中所有特征的值,构建分类模型的特征值向量。

1.3　随机森林与支持向量机

随机森林算法(RF)是一种用于数据分类和回归的机器学习算法,具有优越的分类性能,在生物信息学领域有着广泛应用[26]。支持向量机(SVM)是由 Vapnik 等提出的基于 VC 维度理论和结构风险最小化原理的机器学习方法,在求解小样本、非线性和高维模式识别方面具有许多特殊的优势[27-28]。

在本工作中,我们还利用 RF 来分析不同特征值对分类预测的重要性。另外,我们采用 5 折交叉验证对 RF 和 SVM 的分类能力进行了评估。

1.4　建立分类模型及评价

利用 R 语言中的 randomForest、ipred、predict、e1071、ROCR、caret 等软件包来实现分类模型。分类特征向量为预先提取的 100 个特征组成的一维向量。首先,我们从每组待测数据集的正负数据集中随机抽取 2/3 作为训练集,剩余 1/3 作测试集,利用 RF 计算多维向量特征值重要性。其次,我们采用 5 折交叉验证,对 3 组待测数据分别进行 RF 和 SVM 算法分类,对预测结果进行评价。分类模型的基本流程见图 1。

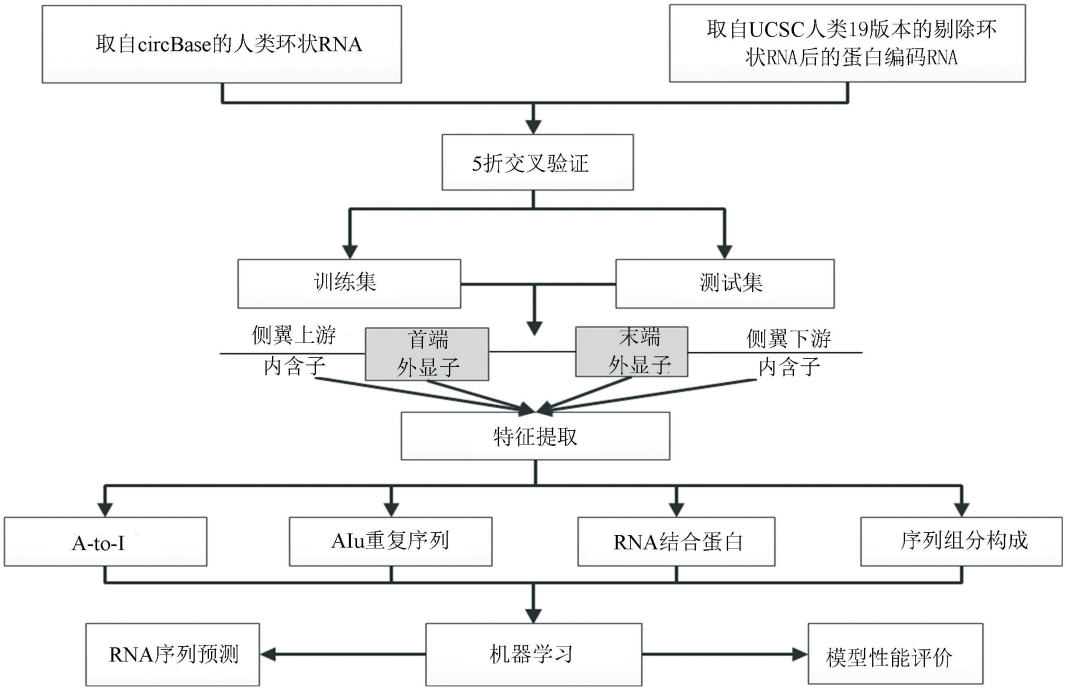

图 1　模型搭建流程图

Fig.1　Flowchart of the model building

我们采用了下列的指标来进行模型评估:灵敏度(Sn)、特异性(Sp)、准确度(Acc)、和马修斯相关系数(MCC),分别定义如下:

$$Sn = \frac{TP}{TP + FN} \tag{1}$$

$$Sp = \frac{TN}{TN + FP} \tag{2}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

2　结果与讨论

2.1　特征重要性排序

为了研究不同的特征在 RNA 成环分类预测中的重要性,我们应用随机森林特征选择对其重要性进

行排序。每个特征值的重要性分数是由森林多次置换前后错误的差异平均值决定的。分值越高表示特征值的重要性越高。图 2 显示了对环状 RNA 分类贡献大的前 40 个特征。可以看出,排在前面的特征有 Alu 反向互补重复序列、A-to-I 位点、多种剪切相关的 RBP、内含子长度和 GT 剪切信号等。这一结果说明,这些特征在区分外显子环状 RNA 及线性 mRNA 中起关键性作用,而排序结果也基本符合我们的预期。

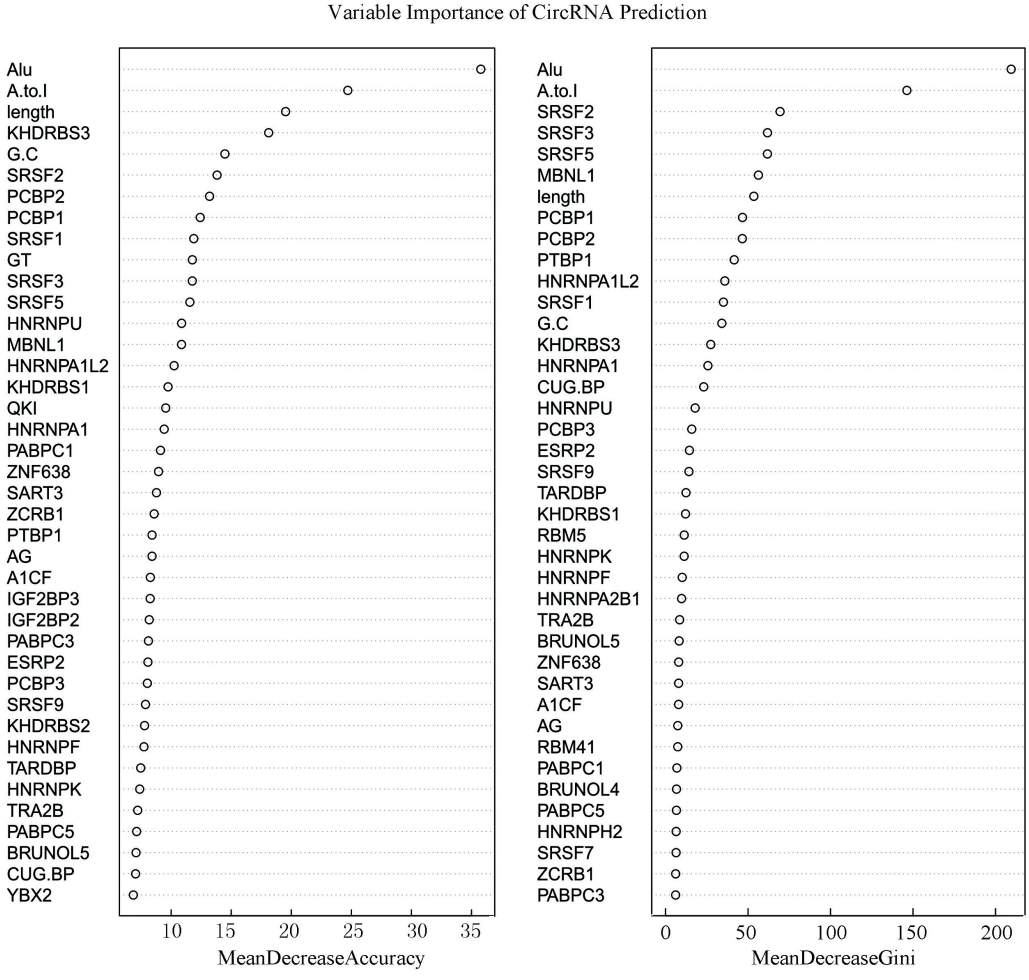

图 2 影响环状 RNA 预测的特征值排序前 40

Fig.2 Top 40 features from Random Forest importance ranking that influence the prediction of circRNAs

2.2 数据测试结果以及 RF 算法和 SVM 算法比较

2.2.1 模型训练

我们采用 5 折交叉检验来评估 RF 和 SVM 在特征向量数据集上的分类情况,结果见表 2。可以看出,对于第 1、2 组数据集,RF 和 SVM 均能有效区别环状 RNA 与线性 RNA,分类性能表现良好,RF 分类表现略优于 SVM。整体来说,这两类算法都能较好的整合不同的特征类型,兼容特征值的异质性,同时也体现出我们提取的特征较好的反映了环状 RNA 与线性 RNA 的特异性。第 3 组数据分类表现不如前两组优秀,但考虑到该组数据集主要来源于 circBase 数据库,该数据库集合了多个高通量测序数据的 find_circ 预测结果。虽然我们对数据进行了序列长度的过滤筛选,但仍有较高的噪声。但由于该数据集样本量大,模型在泛化性和抗过拟合上会有较好的表现。因此,在后续的独立样本预测中,我们选取该组数据训练的模型作为分类器。

2.2.2 独立样本表现

为进一步验证模型的分类特性,使用基于第 3 组数据训练得到的模型对独立数据集中的 RNA 进行分类,分类结果见表 3。结果说明,RF 模型和 SVM 模型均具有较高的环状 RNA 识别正确率。除了识别敏感性(Sn)外,RF 均要优于 SVM。从分类 ROC 曲线(见图 3)可以看出,RF 模型的总体分类水平(AUC = 0.947)要略优于 SVM(AUC = 0.937)。以上结果表明,不管采用 RF 和 SVM,我们提取的 100 个序列特征均能很好地区分外显子区间的环状 RNA 和 mRNA。结果证实,仅仅通过序列本身进行环状 RNA 预测是完全可行的。

表 2　模型训练分类性能打分

Table 2　Performance in model training　%

DataSet	Algorithm	Sn	Sp	ACC	MCC
group1	RF	97.45	97.96	97.72	95.43
	SVM	94.18	97.62	95.96	91.94
group2	RF	98.79	98.77	98.78	97.49
	SVM	98.10	97.30	97.78	95.37
group3	RF	97.45	92.11	94.92	89.89
	SVM	81.59	95.36	88.11	77.21

表 3　独立样本测试集表现打分

Table 3　Performance of independent testing dataset　%

Algorithm	Sn	Sp	ACC	MCC
RF	80.87	96.53	88.62	82.46
SVM	88.54	86.99	87.67	77.96

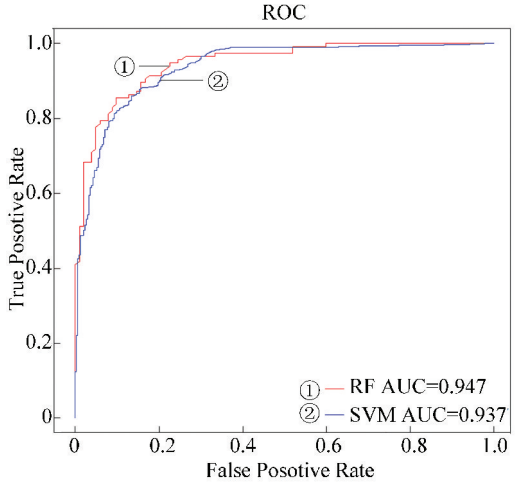

图 3　独立样本测试集分类结果 ROC 曲线

Fig.3　The ROC curves of prediction result of independent testing dataset

2.3　和其他特征提取方法的比较

另外,2016 年发表的 PredicircRNATool 是基于侧翼内含子上双核苷酸构象特征和机器学习算法来识别环状 RNA 和组成型外显子[29]。和我们的模型不同,PredicircRNATool 选取了完全不同的一组特征来实现环状 RNA 的识别。他们基于 RNA 构象及热力学特征,利用 DiProDB 数据库进行特征筛选,共得到了 125 个特征[29]。为了和该方法进行比较,我们计算了表 1 三组数据集中 PredictRNATool 所提取的 125 个特征值,同样利用 5 折交叉验证和 RF 以及 SVM 机器学习方法进行分类测试,结果如表 4。比较表 2 和表 4,我们可以看出尽管基于这两组不同特征所建立的模型的分类表现中,基于本研究的特征建立的模型具有更好的识别能力。对第 3 组数据的分类表现相似,但对于环状 RNA 的识别,我们的模型在 RF 分类方法下具有更好的灵敏度。另外,从特征提取的复杂性来看,我们的特征提取过程不依赖于特定的构象数据库,不需要实验数据的支持,提取过程更简单。同时,我们提取的每一类特征都具有明确的生物学解释,能较好地描述影响 RNA 成环的因素。

表 4　构象及热力学特征模型表现打分

Table 4　Performance of model based on conformational and thermodynamic features　%

DataSet	Algorithm	Sn	Sp	ACC	MCC
group 1	RF	86.64	91.83	89.07	78.31
	SVM	91.98	82.50	87.38	74.96
group 2	RF	91.18	96.71	93.84	87.85
	SVM	96.26	91.47	93.86	87.83
group 3	RF	90.23	93.72	92.06	84.09
	SVM	90.66	92.63	91.67	83.32

3　结　论

本文主要考虑 RNA 内在序列特征对 RNA 能否成环的影响,包括对成环有显著促进作用的 Alu 重复序列、A-to-I 编辑位点、内含子上多种调控 RNA 剪切的 RNA 蛋白结合位点以及序列组分构成特征(包括内含子长度、G/C 含量、GT、AG 剪切信号分布密度)等,这些特征的选取综合了近年来环状 RNA 领域的诸多工作,有较好的广泛性和代表性。

模型的测试和验证结果表明:

1)本文构建的基于序列特征的机器学习分类模型能有效地鉴别环状 RNA 与线性 mRNA,准确度高达 85%以上。

2)对于机器学习分类方法,相比于 SVM 方法,RF 分类的效果更好。

3)不同特征对分类结果贡献度不同,其中以 Alu 重复序列及 A-to-I RNA 编辑位点对分类结果影响最显著。

4)与已发表的相关工作比较也体现出本文特征选取方法能够更好地提升环状 RNA 的识别效率。

针对现有基于测序数据的环状 RNA 识别工具的低一致性、高假阳性率和高假阴性率问题,本论文提出的基于序列特征的机器学习模型能在现有工具基础上,进一步有效地区分环状 RNA 和线性 mRNA,为环状 RNA 的后续功能研究提供更可靠的数据。

参考文献(References)

[1]SALZMAN J, GAWAD C, WANG P L, et al. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types[J]. Plos One, 2012,

7(2):e30733.DOI: 10.1371/journal.pone.0030733.
[2]JECK W R, SORRENTINO J A, WANG K, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats[J]. Rna-a Publication of the Rna Society, 2013, 19(2):141-157.DOI: 10.1261/rna.035667.112.
[3]MEMCZAK S, JENS M, ELEFSINIOTI A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency[J]. Nature, 2013, 495(7441): 333-338.DOI: 10.1038/nature11928.
[4]HANSEN T B, JENSEN T I, CLAUSEN B H, et al. Natural RNA circles function as efficient microRNA sponges [J]. Nature, 2013, 495(7441):384-388.DOI: 10.1038/nature11993.
[5]VALDMANIS P N, KAY M A. The expanding repertoire of circular RNAs[J]. Molecular Therapy, 2013, 21(6): 1112-1114.DOI: 10.1038/mt.2013.101.
[6]HENTZE M W, PREISS T. CIRCULAR RNAs: splicing´s enigma variations[J]. Embo Journal, 2013, 32(7):923-925.DOI: 10.1038/emboj.2013.53.
[7]WANG Y, WANG Z. Efficient backsplicing produces translatable circular mRNAs[J]. Rna-a Publication of the Rna Society, 2015, 21(2): 172-179. DOI: 10.1261/rna.048272.114.
[8]GAO Y, WANG J, ZHAO F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification[J]. Genome Biology, 2015, 16(1):4.DOI: 10.1186/s13059-014-0571-3.
[9]FU X, LIU R. CircRNAFinder: a tool for identifying circular RNAs using RNA-Seq data[C]. Proceedings of the 6th International Conference on Bioinformatics and Computational Biology, BICOB. 2014.
[10]WANG K, SINGH D, ZENG Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery[J]. Nucleic Acids Research, 2010, 38(18):e178. DOI: 10.1093/nar/gkq622.
[11]ZHANG X O, WANG H B, ZHANG Y, et al. Complementary sequence-mediated exon circularization[J]. Cell, 2014, 159(1): 134-147.DOI: 10.1016/j.cell.2014.09.001.
[12]HANSEN T B, VENØ M T, DAMGAARD C K, et al. Comparison of circular RNA prediction tools[J]. Nucleic Acids Research, 2016, 44(6):e58. DOI: 10.1093/nar/gkv1458.
[13]XIONG H Y, ALIPANAHI B, LEE L J, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease[J]. Science, 2015, 347(6218):124-125.DOI: 10.1126/science.1254806.
[14]PAN X, XIONG K. PredcircRNA: computational classification of circular RNA from other long non-coding RNA using hybrid features[J]. Molecular BioSystems, 2015, 11(8): 2219-2226.DOI: 10.1039/c5mb00214a.
[15]GLAŽAR P, PAPAVASILEIOU P, RAJEWSKY N. CircBase: a database for circular RNAs[J]. Rna, 2014, 20(11): 1666-1670.DOI: 10.1261/rna.043687.113.
[16]KENT W J, SUGNET C W, FUREY T S, et al. The Human Genome Browser at UCSC[J]. Genome Research, 2002, 12(6):996-1006.DOI: 10.1101/gr.229102. Article published online before print in May 2002.
[17]LIU Y C, LI J R, SUN C H, et al. CircNet: a database of circular RNAs derived from transcriptome sequencing data [J]. Nucleic Acids Research, 2016, 44(D1): D209-D215.DOI: 10.1093/nar/gkv940.
[18]ZENG X, LIN W, GUO M, et al. A comprehensive overview and evaluation of circular RNA detection tools[J]. Plos Computational Biology, 2017, 13(6): e1005420. DOI: 10.1371/journal.pcbi.1005420.
[19]GUO J U, AGARWAL V, GUO H, et al. Expanded identification and characterization of mammalian circular RNAs [J]. Genome Biology, 2014, 15(7):409.DOI: 10.1371/journal.pcbi.1005420.
[20]WASHBURN M C, HUNDLEY H A. Controlling the editor: the many roles of RNA-binding proteins in regulating A-to-I RNA editing[M]//RNA Processing. Springer International Publishing, 2016: 189-213.DOI: 10.1007/978-3-319-29073-7_8.
[21]NISHIKURA K. Functions and regulation of RNA editing by ADAR deaminases[J]. Annual Review of Biochemistry, 2010, 79(1):321-349.DOI: 10.1146/annurev-biochem-060208-105251.
[22]RAMASWAMI G, WEI L, PISKOL R, et al. Accurate identification of human Alu and non-Alu RNA editing sites [J]. Nature Methods, 2012, 9(6):579-581.DOI: 10.1038/nmeth.1982.
[23]IVANOV A, MEMCZAK S, WYLER E, et al. Analysis of intron sequences reveals hallmarks of circular rna biogenesis in animals[J]. Cell Reports, 2015, 10(2):170-177. DOI: 10.1016/j.celrep.2014.12.019.
[24]CONN S J, PILLMAN K A, TOUBIA J, et al. The RNA binding protein quaking regulates formation of circRNAs [J]. Cell, 2015, 160(6): 1125-1134.DOI: 10.1016/j.cell.2015.02.014.
[25]PAZ I, KOSTI I, ARES J R M, et al. RBPmap: a web server for mapping binding sites of RNA-binding proteins [J]. Nucleic Acids Research, 2014, 42(W1): 361-367. DOI: 10.1093/nar/gku406.
[26]BREIMAN L. Random forests [J]. Machine Learning, 2001, 45(1):5-32.DOI: 10.4236/abb.2010.14040.
[27]VAPNIK V. The nature of statistical learning theory[J]. Conference on Artificial Intelligence, 1995 , 10 (5): 988-999.
[28]WANG Y, LI Y, WANG Q, et al. Computational identification of human long intergenic non-coding RNAs using a GA - SVM algorithm[J]. Gene, 2014, 533(1): 94-99. DOI: 10.1016/j.gene.2013.09.118.
[29]LIU Z, HAN J, LV H, et al. Computational identification of circular RNAs based on conformational and thermodynamic properties in the flanking introns[J]. Computational Biology and Chemistry, 2016, 61: 221-225. DOI: 10.1016/j.compbiolchem.2016.02.003.