文章编号：2095-2163(2020)03-0375-04 中图分类号：TP391；Q811.4 文献标志码：A

长非编码 RNA 鉴定方法研究

杨 阳
（哈尔滨工业大学 计算机科学与技术学院，哈尔滨 150001）

摘 要：高通量测序技术的出现带来了大量可用的转录组数据，评估进化保守区域的编码潜力成为转录数据分析中的核心任务。对转录本编码潜力的预测可以用来鉴定长非编码 RNA（long noncoding RNA，lncRNA）。lncRNA 是一种长度超过 200 个核苷酸的非编码 RNA，研究表明 lncRNA 在多种生物中都有重要作用，能够在染色质修饰、表观遗传、转录及转录后调控等多种层面发挥重要的调控作用。已经有许多基于机器学习的工具被开发用来区分编码与非编码转录本序列。不同的工具通常是针对不同的情况设计的，因此需要根据特定的情况选择合适的方法。本文分析了几种常用工具各自的特点和适用范围，帮助研究人员选用合适的方法以获得更可靠的结果。

关键词：转录组数据；编码潜力；长非编码 RNA；机器学习

Research on identification methods of long non-coding RNA

YANG Yang
（School of Computer Science and Technology，Harbin Institute of Technology，Harbin 150001，China）

【Abstract】 With the advent of high-throughput sequencing technologies，a large amount of available transcriptome data has been generated，and the evaluation of the coding potential of evolutionarily conserved regions has become a core in the analysis of transcripts. Prediction of the coding potential of transcripts can be used to identify long noncoding RNAs（lncRNAs）. lncRNA is a kind of noncoding RNA with length more than 200 nucleotides，which plays an important role in many organisms. It can play an important regulatory role in various aspects such as chromatin modification，epigenetics，transcription and post-transcriptional regulation. Many machine learning tools have been developed to distinguish between coding and non-coding transcripts. Different tools are designed for different situations，so it is required to choose the suitable method for the specific situation. In this review，several popular tools and their advantages，disadvantages，and application scopes are summarised to assist people in employing a suitable method and obtaining a more reliable result.

【Key words】 transcriptome data；coding potential；long noncoding RNA；machine learning

0 引 言

非编码 RNA（noncoding RNA ，ncRNA）是所有从 DNA 转录但不编码蛋白质的功能性 RNA 的统称。最初，人们将非编码 RNA 基因分类为“垃圾基因”或转录“噪音”，然而在之后的研究中发现，非编码序列在生命体生命活动中具有重要的调控作用[1]。这些非编码序列中，最近研究较多的是长非编码 RNA（long noncoding RNA，lncRNA），lncRNA 是指长度超过 200 个核苷酸且不编码蛋白质的转录物[2]。

为了系统研究 lncRNA 的功能，首要的工作是从基因中识别 lncRNA。高通量测序数据大量涌现为学者们提供了更多有关 lncRNA 的有用信息。与此同时，为了方便后续研究和分析，很多鉴定 lncRNA 的计算机方法被提出。本文对鉴定 lncRNA 的计算方法进行了较为全面的回顾。

1 lncRNA 鉴定工具介绍

lncRNA 鉴定过程中的一个重要问题是区分编码与非编码转录本序列，目前已经有很多生物信息学的方法使用序列的内部特征和结构特点预测非编码 RNA[3]。本文中比较了几种流行的基于机器学习的工具。对此可做阐释分述如下。

1.1 CPAT[4] 介绍

CPAT 是基于逻辑回归模型的蛋白质编码潜力评估工具。使用的特征包括：开放阅读框大小、开放阅读框覆盖率、Fickett 分数和 Hexamer 分数。

Fickett 分数的计算主要是基于序列中每个碱基的位置和碱基的含量。定义如下：

$$N_{content} = \text{核苷酸 } N \text{ 的频率}, \tag{1}$$

$$N_{position} = \frac{\mathrm{MAX}(N_i)}{\mathrm{MIN}(N_i) + 1} \quad i \in \{1,2,3\}, \tag{2}$$

$$N_i = \text{核苷酸 } N \text{ 在第 } i \text{ 个阅读框中的个数}, \tag{3}$$

$$FickettScore = \sum_{i=1}^{8} p_n w_n, \tag{4}$$

$$p_n = N_{content} \text{ 或} N_{position} \text{ 转换的概率值}, \tag{5}$$

作者简介：杨 阳（1996-），女，硕士研究生，主要研究方向：机器学习、生物信息学。

收稿日期：2019-05-08

w_n = 概率值p_n 的权值, (6)

Fickett 分数和 Hexamer 分数是区分编码转录本和非编码转录本的重要特征。Hexamer 是指蛋白质中相邻的氨基酸,Hexamer 分数定义如下:

$$HexamerScore = \frac{1}{m}\sum_{i=1}^{m}\log\left(\frac{F(H_i)}{F'(H_i)}\right),\quad i \in 0,1,2,\cdots,4\ 095. \tag{7}$$

按照定义,Hexamer 有 64 * 64 种, $F(H_i)$ 表示蛋白质编码转录本的 hexamer 频率, $F'(H_i)$ 表示非编码转录本的 hexamer 频率。对于一条转录本序列有 m 个 hexamer,Hexamer 分数为正值倾向于这条转录本序列是蛋白质编码转录本。

1.2 CNCI 介绍

CNCI[5]是基于序列的内在特征评估转录本的支持向量机模型,首先计算相邻核苷酸三联体(ANT)的频率,分析 ANT 在序列编码区和非编码 RNA 序列中的使用频率,构建 2 个 64 * 64 的 ANT 矩阵,最后得到 ANT 评分矩阵。对应的数学公式为:

$$ANTScoreMatrix = \log_2\frac{CDSMatrix}{NoncodingMatrix}, \tag{8}$$

计算单个 ANT 频率的公式如下:

$$X_iN = \sum_{j=1}^{n} S_j(X_i), \tag{9}$$

$$T = \sum_{i=1}^{m} X_iN = \sum_{i=1}^{m}\sum_{j=1}^{n} S_j(X_i), \tag{10}$$

$$m = 64 \times 64; n = 1,\cdots,N,$$

$$X_iF = \frac{X_iN}{T}, \tag{11}$$

其中, X 指一种 ANT; $S_j(X_i)$ 指在序列 S_j 中核苷酸三联体 X_i 的出现频次; T 表示数据集中4 096种 ANT 的总出现频次。因此 X_iF 是核苷酸三联体 X_i 的频率。在得到 ANT 评分矩阵后,利用评分矩阵对序列评分。

使用大小为 150 nt,步长为 3 nt 的滑动窗口扫描每个转录本序列得到 6 个阅读框, 扫描的同时计算每个窗口的 S 得分,每个阅读框对应一个窗口 S 得分的数组。S 得分的定义如下:

$$S\ Score = \sum_{i=1}^{n}\{H_p(X_i)\}, \tag{12}$$

其中, X 表示一种 ANT; H_p 表示 ANT 评分矩阵; n 表示一个窗口中的 ANT 总数。

S 得分可以反映出每个滑动窗口的编码能力,CNCI 从每个阅读框中找到 MLCDS(most-like CDS)。在这六个 MLCDS 中,选择得分最高的 MLCDS 长度和 S 得分作为 CNCI 的 2 个特征。CNCI 的其他三个特征分别是长度百分比、得分距离和密码子偏差。长度百分比的定义如下:

$$LENGTH - Percentage = \frac{M_1}{\sum_{i=0}^{n}(Y_i)},\quad i \in (1,2,3,4,5,6), \tag{13}$$

其中, M_1 是 6 个 MLCDS 中得分最高的 MLCDS 的长度, Y_i 是每个 MLCDS 的长度。

得分距离的定义如下:

$$Score - Distance = \frac{\sum_{j=0}^{n}(S - E_j)}{5},\quad j \in (1,2,3,4,5). \tag{14}$$

其中, S 是 6 个 MLCDS 中得分最高的 MLCDS 的得分, E_j 是其余 MLCDS 的得分。

1.3 PLEK 介绍

PLEK[6]的分类模型为使用了标准径向基函数核的支持向量机模型,是基于序列 k-mer 频率的无对齐工具。使用 k-mer 和具有一个核苷酸步长的滑动窗来分析每个转录物。k-mer 模式是具有 k 个核苷酸的特定字符串,出于对精度和计算时间的考虑,PLEK 选取 k 的范围从 1 到 5。因此,对于一个序列,有 4+16+64+256+1 024=1 364 个模式。用于预测的转录本的特征如下所示:

$$f_i = \frac{c_i}{s_k}w_k, k = 1,2,3,4,5,\ i = 1,2,\cdots,1\ 364, \tag{15}$$

$$s_k = l - k + 1,\qquad k = 1,2,3,4,5, \tag{16}$$

$$w_k = \frac{1}{4^{5-k}},\ k = 1,2,3,4,5. \tag{17}$$

其中, c_i 表示模式 i 在转录本中出现频次, f_i 是模式 i 的频率校准后的值。

1.4 CPC2 介绍

CPC2[7]是 CPC 的升级,仍然使用支持向量机模型,CPC2 能更加快速、准确地评估 RNA 转录本的编码能力。CPC2 中使用了 4 个特征,包括:Fickett 分数、开放阅读框长度、开放阅读框完整性以及预测肽的等电点。开放阅读框的完整性是指开放阅读框以起始密码子开始,以终止密码子结束。等电点可以通过 BioPython 中的 ProtParam 模块计算得到。

1.5 CPPred 介绍

CPPred[8]的实现基于支持向量机分类器和多个序列特征, CPPred 使用开放阅读框长度、开放阅读框覆盖率、Fickett 分数和 Hexamer 分数、开放阅读

框完整性、预测肽的等电点、预测肽的不稳定指数、预测肽的亲水性平均值 Gravy 以及 30 个 CPPred 中提出的 CTD 特征训练分类器。CTD 特征用来描述全局转录本序列，核苷酸组成（特征 C）描述了转录本序列中每个核苷酸的百分比组成；核苷酸转换（特征 T）描述了 4 个核苷酸在相邻位置之间转换的百分比；核苷酸分布（特征 D）计算每个核苷酸在转录物序列的 5 个相对位置（0，25%，50%，75%，100%）来表示每个核苷酸在转录本序列中的分布。

2　lncRNA 鉴定工具比较

本文所涉及的 5 个 lncRNA 鉴定工具包括最常用的 CPAT、CNCI、PLEK，以及 CPC 的最新版本 CPC2 和最新发布的工具 CPPred。其中，CPAT 使用了逻辑回归模型，其余四个工具都使用了支持向量机模型。本文总结每种工具的简要信息和使用细节见表 1。

接着，本文更具体地对 5 种工具所选择的特征进行概述，见表 2。

表 1　有关 lncRNA 鉴定的方法概述

Tab. 1　Overview of the methods concerning lncRNA identification

	时间	模型	训练数据集物种	查询文件格式	网上服务
CPAT	2013	LR	人类；小鼠；果蝇；斑马鱼	FAST；BED	http://lilab.research.bcm.edu/cpat/index.php
CNCI	2013	SVM	人类；植物	FAST；GTF	http://www.bioinfo.org/software/cnci
PLEK	2014	SVM	人类；玉米	FASTA	http://sourceforge.net/projects/plek/files/
CPC2	2017	SVM	人类	FAS；GTF；BED	http://cpc.cbi.pku.edu.cn
CPPred	2019	SVM	人类；小鼠；斑马鱼；果蝇；酿酒酵母；线虫；拟南芥	FASTA	http://www.rnabinding.com/CPPred

表 2　有关 lncRNA 鉴定的方法特征概述

Tab. 2　Summary of the features of each method selected

	开放阅读框	密码子	序列结构	预测肽
CPAT	长度；覆盖率	Hexamer 分数	Fickett 分数	无
CNCI	无	ANT 评分矩阵；密码子偏差	MLCDS	无
PLEK	无	无	改进的 k-mer 方案	无
CPC2	长度；完整性	无	Fickett 分数	等电点
CPPred	长度；覆盖率；完整性	Hexamer 分数	Fickett 分数；CTD 特征	等电点；不稳定指数；Gravy

本文使用特异性（*Specificity*）、敏感度（*Sensitivity*）、准确率（*Accuracy*）来评估 5 个分类工具。定义如下：

$$Specificity = \frac{TN}{TN + FP}, \tag{18}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{19}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}. \tag{20}$$

5 种鉴定工具都是不包含比对过程的，适用于对未充分研究的物种的转录物分析。其中，CNCI 和 PLEK 都可以用于有测序错误的数据集，PLEK 在这类数据上表现更好。与 CPAT、CPC2 和 CPPred 相比，PLEK 在除人类以外的其他物种中表现不佳。5 种鉴定工具在不同测试集上的表现见表 3。

由于不同物种 lncRNA、不同测序数据之间存在一定的差异性，不同的 lncRNA 鉴定工具设计上存在一定的针对性。CPAT 和 CPPred 为小鼠转录本的鉴定提供了专门的模型。在分析其他物种时，CPAT 还提供了果蝇和斑马鱼的模型；CNCI 和 PLEK 可以预测脊椎动物和植物的序列；CPC2 还可以预测果蝇、斑马鱼、拟南芥、蠕虫；CPPred 提供了适用于斑马鱼、果蝇、酿酒酵母、线虫和拟南芥的模型。不同工具在不同条件下的适用性见表 4。

表 3 鉴定工具在不同测试集上的表现概述

Tab. 3 Overview of each tool's performance on different testing datasets

测试集	测试指标	CPAT	CNCI	PLEK	CPC2	CPPred
Human MCF-7 (PacBio)	特异性		91.80	94.70		
	敏感度		78.70	95.80		
	准确率		91.30	94.70		
Human HelaS3 (454)	特异性		93.90	95.50		
	敏感度		81.10	92.50		
	准确率		93.70	95.40		
Human (from GENCODE)	特异性	94.07		98.10	95.30	97.04
	敏感度	94.58		95.42	90.92	95.44
	准确率	94.33		96.73	93.07	6.23
Mouse (from GENCODE)	特异性	96.65		93.43	95.86	97.70
	敏感度	96.10		87.61	95.86	95.57
	准确率	96.32		89.88	95.61	96.40

表 4 不同工具在不同条件下的优先权

Tab. 4 Priority of employing different methods on different situations

	CPAT	CNCI	PLEK	CPC2	CPPred
编码潜力评估	√			√	
人类 lncRNA 识别	√	√	√	√	√
小鼠 lncRNA 识别	√			√	√
其他物种 lncRNA 识别	√	√	√	√	√
数据包含测序错误		√	√	√	
小 ORF 数据					√
大规模数据	√		√	√	
允许用户训练模型	√		√		
网站服务	√			√	

3 结束语

lncRNA 的鉴定一直以来都是生物信息学研究的一个挑战,在 2010 年之前,以 CPC 软件为代表的 lncRNA 鉴定工具会依赖比对过程,此后,大部分软件通过提取序列的内在特征来进行分类。这篇综述中,集中探讨了常用的和最新的 lncRNA 鉴定工具,总结了其相应的适用范围,帮助研究人员来选择使用适合的工具,同时获得令人信服的结果。未来 lncRNA 鉴定工具的趋势是针对不同类型的序列,开放不同的工具来解决各种特定情况下的问题。

参考文献

[1] PALAZZO A F, LEE E S. Non-coding RNA: What is functional and what is junk? [J]. Frontiers in Genetics, 2015, 6: 2.

[2] 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.

[3] 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: 8496165.

[4] 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): e74.

[5] 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): e166.

[6] 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: 311.

[7] KANG Yujian, YANG Dechang, KONG Leilei, et al. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features [J]. Nucleic acids research, 2017, 45(W1): W12.

[8] TONG Xiaoxue, LIU Shiyong. CPPred: Coding potential prediction based on the global description of RNA sequence [J]. Nucleic Acids Research, 2019, gkz087:1