DOI: 10.13376/j.cbls/2015131
文章编号：1004-0374(2015)07-0946-07

· 技术与应用 ·

鉴定和预测长非编码RNAs的生物信息学方法

陈思佟[1]，岑　益[2]，柳建发[1]，李　洋[1]，廖　奇[1]*
(1 宁波大学医学院，宁波 315211；2 宁波市公安局鄞州分局，宁波 315100)

摘　要：越来越多的研究表明，长非编码 RNAs(long non-coding RNAs, lncRNAs) 可以调节蛋白质编码基因的表达、稳定性及亚细胞定位，参与众多重要的生物过程。由于 lncRNAs 是一类新发现的非编码 RNAs，挖掘各物种的 lncRNAs 仍然是一个值得研究的问题。其中，利用生物信息学方法挖掘和鉴定 lncRNAs 已经成为当前生物信息学家研究的一个热点。现就基于生物信息学方法对 lncRNAs 的鉴定研究作一综述，主要内容分为两大类：基于测序和基于特征的计算机预测方法。基于测序又包括 EST 测序、cDNA 测序及二代转录组 RNA 测序；而基于特征的计算机预测则主要包含基于序列保守性、基于碱基排列顺序及基于表观遗传修饰特征。通过以上几方面的论述，来阐明目前 lncRNAs 鉴定方法的现状和进展。

关键词：长非编码 RNAs；鉴定；计算机预测，测序
中图分类号：Q522；Q612　　**文献标志码**：A

Bioinformatics methods of identifying and predicting long noncoding RNAs

CHEN Si-Tong[1], CEN Yi[2], LIU Jian-Fa[1], LI Yang[1], LIAO Qi[1]*
(1 School of Medicine, Ningbo University, Ningbo 315211, China; 2 Yinzhou Branch of Ningbo Public Security Bureau, Ningbo 315100, China)

Abstract: More and more researches show that long non-coding RNAs (lncRNAs) play an important role in a number of biological processes through regulating the expression, stability and subcellular location of protein-coding genes. As lncRNAs are a kind of ncRNAs recently found, identification of lncRNAs in each organism is an emergency task. Among them, identification of lncRNAs using bioinformatics methods is a hot topic for bioinformatists. In this review, we mainly summarize the bioinformatics approaches of lncRNAs identification and prediction. The mothods are divided into two major parts: sequencing technology-based method and sequence characters-based computational prediction method. Sequencing technology-based method includes EST sequencing, cDNA sequencing and next-generation RNA-seq, while sequence characters-based computational prediction method includes sequence conservation, nucleotide arrangement and epigenetics modifications. The review aids to clarify the present status and progress of lncRNAs identification method.

Key words: long noncoding RNAs; identification; computational prediction; sequencing

真核生物的基因组由庞大的 DNA 序列构成，其中有 50% 的 DNA 可以转录成 RNA，但约有 5% 的 RNA 负责翻译蛋白质，剩余的大概 95% 则被称为非编码 RNA (non-coding RNA, ncRNA)。ncRNA 即不编码蛋白质的 RNA，其种类繁多，按长度分包含长非编码RNAs (long non-coding RNAs, lncRNAs) 以及 microRNA、siRNA 和 piRNA 等在内的小非编码 RNAs。其中，lncRNAs 是一类长度大于 200 nt，特征与 mRNAs 类似，如可变剪切、可带 Poly-A、可加帽的 ncRNAs。近年来，由于 lncRNAs 在众

收稿日期：2015-01-06；**修回日期**：2015-02-26
基金项目：浙江省自然科学基金项目(LQ13C060002)；国家自然科学基金项目(31301084)；宁波大学王宽诚教育基金
***通信作者**：E-mail: liaoqi@nbu.edu.cn；Tel: 0574-87609602

多生物过程中表现出对蛋白质编码基因的重要调节作用，关于 lncRNAs 的报道越来越多。lncRNAs 通过对关键蛋白质编码基因的多重调控机制，包括调节蛋白质编码基因的表达、稳定性及亚细胞定位，参与一些重要的生物过程，如维持剂量补偿、基因组印记、mRNA 加工、细胞分化和发育、疾病和癌症等。有些与疾病相关的 lncRNAs 甚至可以作为生物标记用于疾病的诊断与预测 [1]，如人们发现有些 lncRNAs 是非常灵敏和特异的肿瘤标记物，如前列腺癌中的 DD3，肿瘤抑制基因 *p15* 的反义链，能通过调控 *p15* 基因甲基化水平促进白血病的发生 [2]。

鉴于 lncRNAs 的重要调节作用，lncRNAs 的鉴定也成了科学家们首要研究的问题。从 2003 年和 2005 年 FANTOM (Functional Annotation of the Mouse) 组织在鼠 cDNA 大规模测序中发现大量 lncRNAs 开始 [3-4]，lncRNAs 的鉴定方法随着测序技术的发展也不断更新和改进。目前，借助测序技术已经在人、鼠等哺乳动物中发现了大量的 lncRNAs，不同的研究者也提出了各种鉴定 lncRNAs 的方法和流程。lnRNAs 的鉴定方法主要分为两大类，一类是基于测序技术，它包括早期的 cDNA 文库测序、EST 片段测序以及目前发展盛行的二代转录组 RNA-seq 测序；另一类是基于特征的计算机预测，它基于的特征主要包括序列特征、保守性和表观遗传修饰位点特征。而有些鉴定流程结合测序和计算机预测，结果更加可靠。本文就关于 lncRNAs 的鉴定方法上作一综述。

1　测序方法鉴定lncRNAs

1.1　基于EST鉴定lncRNAs

EST (expression sequence tag) 称为表达序列标签，是将具有 Poly-A 的 RNAs 反转录成 cDNA 并克隆到载体构成 cDNA 文库后，随机选择 cDNA 克隆进行 5′ 端和 3′ 端单一次测序获得的短 cDNA 部分序列。由于之前对 lncRNAs 认识的欠缺，科学家们认为 cDNA 只包含 mRNAs，因此，EST 测序在早期通常用于蛋白质编码基因的鉴定。然而，由于 lncRNAs 序列特征与 mRNAs 类似，可带 Poly-A、可变剪切等，其实 cDNA 中也包含一部分 lncRNAs。因此，EST 也可能是 lncRNAs 的片段，即 EST 测序能够用于 lncRNAs 的鉴定。

早在 2001 年，EST 序列就用于拟南芥 lncRNAs 的鉴定 [5]。2007 年，Wen 等 [6] 同样利用 EST 数据在豆科植物苜蓿中挖掘了不具编码能力的 mRNA-like ncRNAs，类似 mRNAs 的 ncRNAs，也就是 lncRNAs。EST 预测 lncRNAs 的流程如下：(1) 利用 EST2Genome 将 EST 序列与基因组进行比对；(2) 去除与已知蛋白质编码基因 (或包括预测的蛋白编码基因) 重叠较多 (如 10% 以上) 的 EST 序列；(3) 利用 GENEMARK.hmm[7] 或其他软件预测基因；(4) 利用 EMBOSS 软件中的 getORF 或其他 ORF (open reading frame，开放阅读框) 预测软件对转录本序列进行 ORF 预测，去除 ORF 长度较长的转录本；(5) 利用 BLASTX 软件将剩余转录本与 Swiss-Prot、trEMBL 和 GenBank 等其他蛋白质数据库的蛋白质序列进行比对，去除与蛋白质编码基因相似的转录本。

2008 年，Xue 和 Li [8] 提出另一种方法，在人 EST 序列中鉴定出 100 多条 ncRNAs。他们以 50 bp 作为窗口不重叠地扫描整个基因组，统计每个 50 bp 窗口所覆盖的 EST 数目，选择 EST 数目大于 3 并且保守性达到一定程度 (Phastcons 分值大于 0.8)[9] 的窗口作为种子序列，然后对这些种子序列进行电子延伸，得到 contigs，进而除去与 ECgene 软件注释为可变剪切、可选转录起始或可选 Poly(A) 位点相重叠的转录本，除去 ORF 长度较长 (超过 100 aa) 且在基因上下游 2 000 nt 内没被 promoter 2.0 软件预测有启动子的转录本。由于他们当时没有考虑到 lncRNAs，因此，过滤掉了长度大于 1 500 nt 的转录本；如果要预测 lncRNAs，可改变长度的阈值，如设置长度大于 200 nt 即可。

尽管目前测序技术飞速发展，EST 测序已经逐渐退出舞台。然而，以往测序所得的 EST 数据，仍然保存在数据库中，对这些数据再利用和再分析，从中挖掘有意义的信息是生物信息学家们的任务。在没有对样本进行 cDNA 或 RNA 测序的时候，基于 EST 数据预测 lncRNAs 仍然不失为一种较好的方法，目前仍然在采用。如 2012 年，Huang 等 [10] 利用 EST 序列在牛中挖掘出 449 条 lncRNAs。但是，EST 只能检测到部分含 Poly-A 的 lncRNAs，并且 EST 只是基因的部分片段，全长序列需要进行 EST 片段拼接才能获得，有些甚至无法拼接，无法保证 lncRNAs 的完整性。

1.2　cDNA测序方法鉴定lncRNAs

2003 年，FANTOM 组织对 RIKEN 鼠全长 cDNA 进行测序，得到 60 770 条转录本，经过筛选分析，得到 4 280 条 lncRNAs[3]。2004 年，H-Invitational 组织对人的转录组 cDNA 也进行大规模测序，发现

了 21 037 条转录本，其中 1 377 条也被鉴定为 lncRNAs[11]。2005 年，FANTOM 组织又再一次对鼠的 cDNA 进行大规模测序，检测出 102 281 条转录本，其中有 34 030 条序列长度与 mRNA 相当，却没有明显的 ORF，并且也不与其他任何编码蛋白质的 cDNA 序列相似，这类转录也归为 lncRNAs[4]。早期 cDNA 技术测序的目的是为了鉴定编码蛋白质的 mRNAs，然而，对测序片段经过分析筛选后却发现一类不具蛋白质编码基因特性的 lncRNAs。cDNA 测序能够测得 lncRNAs 序列的原因是 lncRNAs 具有与 mRNAs 类似的序列特性，都有 Poly-A。在 cDNA 测序技术中，首先要构建 cDNA 文库，cDNA 文库的构建通常采用 Oligo(dT) 作逆转录引物，且保留长度较长 (如 400 bp 以上) 的 cDNA。具有 Poly-A 且长度较长的 RNAs 过去均被认为是 mRNAs，其实有一部分是 lncRNAs，因此，cDNA 测序可以测得一部分 lncRNAs 的序列。

由于 cDNA 测序可以获得转录本全长序列，因此，cDNA 测序技术鉴定 lncRNAs 的流程相对简单，可以归纳为以下几个步骤：(1) 利用 RepeatMasker 软件剔除具有重复、低复杂性的 cDNA 序列；(2) 利用 BLASTN[12] 或其他比对软件将剩余 cDNA 序列与基因组进行比对，根据一定的阈值 (如相似度 >95%，覆盖率 >90%) 选择与基因组匹配的 cDNA 序列；(3) 如果一个 cDNA 序列对应多个基因组区域，则通过其相似度、覆盖长度、外显子个数等选择一个最好的基因组区域；(4) 将与基因组匹配的 cDNA 序列与 Refseq mRNAs 进行比对，选取与 Refseq mRNAs 不相似的 cDNA 序列；(5) 利用 FASTY[13-14] 和 BLASTX [12] 预测剩余 cDNA 序列的 ORF，除去具有明显 ORF 的 cDNAs。

正如 cDNA 测序技术比 EST 测序技术先进一样，基于 cDNA 测序鉴定 lncRNAs 的准确度和精度也要比基于 EST 的方法高，在二代测序技术问世之前，cDNA 测序技术也是鉴定 lncRNAs 较为可靠的方法之一。

1.3 RNA-seq二代转录组测序技术鉴定lncRNAs

随着测序技术的发展，特别是二代转录组测序 (RNA-seq) 技术的出现，越来越多的 lncRNAs 在人和其他物种的各个组织和细胞系中被发现。二代转录组测序 (RNA-seq) 技术是采用新一代的测序技术，能够快速全面地检测特定物种的某个组织或细胞系的几乎全部的转录本。lncRNAs 的序列特征与 mRNAs 类似，有些具有 Poly-A，有些则无。RNA-seq 在样品提取总 RNA 后，有三种策略：(1) 总 RNA 去除核糖体 RNA，以最大限度保留所有 lncRNAs；(2) 总 RNA 去核糖体 RNA 后再去除含 polyA 的 RNA，以去除大部分编码蛋白质序列；(3) 提取 poly A+ 的 RNAs。由于既有 poly A+ 的 lncRNAs，也有 poly A– 的 lncRNAs，没特殊要求的情况下，第一种方法最好。因此，对这些所提取的 RNAs 进行 RNA-seq，可以检测 lncRNAs 序列以及它们的表达，利用测序片段 (reads) 及表达信息可以从头构建新的转录本，通过对这些转录本进行分析筛选从而获得候选的 lncRNAs。该方法目前被广泛应用于人、鼠、寄生虫、植物等各物种 lncRNAs 的鉴定，已经在各物种的各个组织和细胞系中找到了上万条 lncRNAs[15-18]。因此，RNA-seq 测序技术已经成为当前鉴定 lncRNAs 的主流方法。

RNA-seq 测序技术鉴定 lncRNAs 的流程可以归纳如下：(1) 首先，利用 cufflink 软件 [19] 从 RNA-seq 的序列数据中构建转录本；(2) 根据转录本的表达情况，去除低表达的转录本 (认为是噪音所致)；(3) 对高重复或低复杂的转录本进行过滤；(4) 去除基因组上与蛋白质编码基因正向重叠的转录本，剩下的转录本包括内含子、基因间的 ncRNAs 以及蛋白质编码基因的反义链；(5) 要求 lncRNAs 长度不小于 200 nt；(6) 利用 ORF Finder[20] 或其他软件寻找转录本的 ORF 区域，要求 lncRNAs 的 ORF 长度不大于 300 nt；(7) 利用 BLASTX[12] 将剩下的转录本与 UniProt-TrEMBL 数据库 [21] 的蛋白质序列进行比对，去除那些具有相似性蛋白质 (匹配长度大于 30aa，E-value 小于 0.01) 的转录本；(8) 利用 ncRNA 预测软件如 CPC(coding potential calculator)[22] 或其他软件对剩下转录本进行过滤，得到更加可信的候选 lncRNAs。

RNA-seq 测序鉴定 lncRNAs 的方法虽然成本较高，并且不同的转录本构建方法会得到不同的转录本，对于低表达的阈值确定同样也是有争议的话题。然而，RNA-seq 检测 lncRNAs 的方法基于表达的信息，对 lncRNAs 起始、终止位点以及可变剪切位点的界定均有较强的表达依据，结果较为可信，是目前最为准确的方法之一。

2 计算机方法预测lncRNAs

目前，关于 lncRNAs 的计算机预测主要基于 lncRNAs 的序列特征，包括保守性、碱基排列以及组蛋白修饰位点。

2.1　基于序列保守性的计算机预测方法

lncRNAs 尽管某些序列特征与 mRNAs 类似，如可被剪切、具有帽子与 Poly-A，长度与 mRNAs 相当等，但仍然具有自己独特的序列特征，其中最重要的一点是，除了外显子及特殊的功能元件外，lncRNAs 序列不具保守性，变异程度较高。科学家们利用这点，通过计算分析已知 mRNAs 和 lncRNAs 的序列保守性，构建数学模型，从而对未知序列进行预测。最经典的是 Lin 等 [23] 提出的密码子替换频率 (codon substitute frequence, CSF) 方法，即利用 CSF 打分对蛋白质编码基因和 ncRNAs 进行鉴定。

该方法在提出的时候主要用于区分果蝇的蛋白质编码基因和 ncRNAs，后来被 Guttman 等 [24] 用于区分小鼠的 lncRNAs 和蛋白质编码基因。CSF 的原理基于 ncRNAs 在人与其他同源物种的密码子替换频率不一样的假设，利用该物种与其他物种的多重比对数据，通过计算 (训练) 已知 mRNAs 和 ncRNAs 的密码子替换频率，得到 mRNAs 和 ncRNAs 的密码子替换矩阵，分别记为 $CSM_{a,b}^{C}$ 和 $CSM_{a,b}^{N}$(a、b 为两个密码子，表示 a 替换为 b)，则 $CSM_{a,b}^{N}/CSM_{a,b}^{C}$ 即为 ncRNAs 与 mRNAs 密码子替换频率的比值。由于多重比对中涉及多个物种，因此，每个物种都能得到人与该物种的替换频率比值矩阵。对于一个序列，首先可以获得该序列与其他物种的多重比对数据，然后考察该序列的第一个 90 bp (30 个密码子) 长的序列，计算 30 个密码子的替换频率比值之和，得到一个 CSF 分值。由于具有多个物种，该 90 bp 序列在每个物种中都对应一个 CSF 值，取最大的 CSF 值作为该段的 CSF 分值。然后，窗口向前移动 3 bp (1 个密码子)，继续计算下一个 90 bp 序列的 CSF 分值，直到算完最后一段 90 bp 的 CSF 分值为止，最终选取最大分值作为该序列的 CSF 值。通过以上的计算方法可以获得已知 mRNAs 和 lncRNAs 的 CSF 分值分布状况，可以发现 mRNAs 与 lncRNAs 的 CSF 分值为两个截然不同的分布，选择一个阈值作为两类 RNA 的界限。对于未知的序列，可以通过计算该序列的 CSF 分值，从而判断该序列是否为 lncRNA。目前，利用 CSF 分值判别 lncRNAs 的算法已开发成软件，名叫 PhyloCSF，可以通过网址 http://compbio.mit.edu/PhyloCSF 进行访问，将源程序下载在本地安装运行 [25]。PhyloCSF 的敏感性为 90%，特异性仅为 63%；并且，PhyloCSF 由于基于多物种序列比对的特征，存在一定的缺陷，如有些物种的序列保守性较差，即使是人类，在 8 195 条 lncRNAs 中，也仅有 993 条在其他物种中具有同源序列 [16]。此外，有些 lncRNAs 在基因组上与蛋白质编码基因相重叠，为蛋白质编码基因的正义链或反义链转录本，这些转录本与其他物种的蛋白质编码能比对上，因此，不能准确地判断为 lncRNAs，并且由于多重比对运行时间较长，因此，PhyloCSF 软件运行的速度也较慢。

2.2　基于碱基序列特征的计算机预测方法

lncRNAs 序列特征除了保守性差以外，还具有其他特有的特征，其中最关键的是 lncRNAs 通常不具有 ORF。因此，最早对 lncRNAs 进行鉴定的方法为确定其序列有无包含较长的 ORF。ORF 的预测软件有很多，如 ORF-finder。此外，BLASTX 通过确定与已知蛋白质编码基因的序列相似性也可以对转录本进行判定。由于一些 lncRNAs 可以包含较短的 ORF，并且有些蛋白质编码基因编码较短序列的蛋白质，因此，仅仅依赖 ORF 不能准确地预测 lncRNAs，但由于 ORF 的重要性，后续的方法中 ORF 仍然作为其中的一个重要特征。

由于 lncRNAs 不编码蛋白质，其碱基排列也与 mRNAs 有所不同，结合多种序列特征，构建分类器，同样也能对 lncRNAs 进行预测，如 CPC (Coding Potential Calculator) 由 Kong 等 [22] 开发，是一个基于转录本序列特征的 SVM 分类器。他们利用 6 个基本的序列特征：由 ORF 预测软件 framefinder 计算出来的 (1)LOG-ODDS SCORE 和 (2) COVERAGE OF THE PREDICTED ORF，这两个特征为 ORF 的指标，值越高，ORF 质量越好；第三个也是关于 ORF 的特征，为 (3)INTEGRITY OF THE PREDICTED ORF，表示 ORF 是否以起始密码子开始，终止密码子结束；第四个为 (4)NUMBER OF HITS，即基于 BLASTX 软件在 UniProt 参考序列数据库中比对寻找到的相似序列 (阈值为 E- 值小于 1e-10) 的数目；此外，比对结果中所有序列的 E 值进行负 Log 化，求其均值，并对 3 种编码形式的序列得到的负 Log(E-value) 平均值再求均值，作为第五个特征，称为 (5)HIT SCORE；最后，将 3 种编码形式的负 Log(E-value) 平均值的方差作为第六个特征，称为 (6)FRAME SCORE，其值越高，越有可能是蛋白质编码基因。基于已知 mRNAs 和 ncRNAs 的这 6 个序列特征，Kong 等 [22] 利用 SVM 机器学习方法构建 mRNAs 以及 ncRNAs 的分类器，预测准确率高达 99%。CPC 同样具有在线的网上服务，研究者可以很方便地从网址 http://cpc.cbi.pku.

edu.cn 中对未知序列进行预测。然而 CPC 的特异性较低，仅为 74%，并且速度较慢。

CPC 软件的运行速度较慢，如 CPC 需要两天的时间计算 Cabili 等 [16] 鉴定的 14 353 个转录本的编码能力。随后，Wang 等 [26] 开发的 CPAT (Coding-Potential Assessment Tool) 软件，不仅能够快速地对 lncRNAs 进行鉴定，而且克服了序列比对造成的缺陷。CPAT 基于 4 个序列特征，采用逻辑回归模型对 lncRNAs 进行鉴定。这 4 个序列特征如下。(1) 最大 ORF 长度。(2) ORF 覆盖比例，即 ORF 的长度比上转录本的整长。(3)Fickett TESTCODE 分数，与核苷酸组成和密码子使用偏倚的组合有关。首先，计算 4 个核苷酸的位置值和组成值，位置值计算为：A_1、A_2、A_3 的最大值与 A_1、A_2、A_3 的最小值加 1 的比值。其中 A_1 为核苷酸 A 在序列 0、3、6….. 的个数；A_2 为核苷酸 A 在序列 1、4、7….. 的个数；A_3 为核苷酸 A 在序列 2、5、8….. 的个数。组成值即为每个核苷酸在序列的组成比例。将这 8 个值转化为编码的概率值 *P*，那么 Fickett 分值为概率值 *P* 与权重 w 乘积的累加和。(4)Hexamer 分值，即邻近氨基酸使用的偏好。首先利用已有的蛋白质编码基因和 lncRNAs 分别计算邻近密码子的使用频率 (分别用 F(Hi) 和 F'(Hi) 代替)；然后，计算两个频率的 log 比值，那么给定一条 DNA 序列，Hexamer 分值为 F(Hi) 和 F'(Hi) 的 log 比值的累加和的平均值。基于以上 4 个特征构建逻辑回归模型，对 lncRNAs 进行预测，敏感性可达 96%，特异性为 97%，并且速度快，比 CPC 和 CSF 快 1 万倍。该软件网址为 http://lilab.research.bcm.edu/cpat/index.php。

另一个基于序列特征的预测软件为 Sun 等 [27] 开发的 CNCI，网址为 http://www.bioinfo.org/software/cnci。CNCI 首先分别将 mRNAs 与 lncRNAs 的邻近密码子 (ANT) 替换频率进行统计，将两两邻近密码子替换的频率之比的 log 值用于构建 ANT 分值矩阵。对于每个转录本，按照 6 种编码框形式进行编码，按照 ANT 分值进行打分，在每条不同编码框序列中选择具有最高分值的区域，而 6 条序列中再进一步选择最高分值的区域，作为最似 CDS 序列 (most-like CDS, MLCDS)，然后选取 MLCDS 的长度、分值作为其中的 2 个特征。此外，由于一条具有编码蛋白质能力的转录本所选取的 MLCDS 会与其他 5 条编码框选择出来的 MLCDS 具有较大的不同，因此，进一步选择长度比例 (所选 MLCDS 的长度与所有 6 条 MLCDS 长度之和的比值)、分值距离 (所选 MLCDS 的分值与其他 5 条分值距离的平均值) 作为另外 2 个特征。最后选择单核苷酸的频率作为最后 1 个特征，共 5 个特征，同样利用 SVM 支持向量机的学习方法，构建了 lncRNAs 的分类器。CNCI 相比其他软件，更适合用于不完整序列及反义链的预测，具有较高的性能。

2014 年，Li 等 [28] 新开发了一个软件，叫做 PLEK (predictor of long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme)。该方法基于改进的 *k*-mer 频次，采用 SVM 算法对 lncRNAs 进行预测。PLEK 比较适用于高插入缺失率的序列，如 454 及 PacBio 测序从头预测所得的转录本。此外，PLEK 的运行速度也较快，比 CNCI 快 8 倍，比 CPC 快 244 倍，比 PhyloCSF 快 1 421 倍。该软件的下载地址为 https://sourceforge.net/projects/plek/files/。

此外，Fan 等 [29] 开发了 lncRNA-MFDL 软件，基于一系列特征包括 ORF、*k*-mer 频次、二级结构和编码蛋白质的功能域，采用深度学习的分类算法，准确率高达 97.1%，并且也同样适用于多物种。该软件的网址为 http://compgenomics.utsa.edu/lncRNA_MDFL/。

2.3 基于表观遗传修饰特征的计算机预测方法

随着对 lncRNAs 序列特征和功能的进一步研究，科学家们发现 lncRNAs 在组蛋白修饰特征上具有一定的规律性，可用于 lncRNAs 的鉴定。2009 年，Guttman 等 [24] 发现，H3K4me3 和 H3K36me3 这两种组蛋白修饰是基因表达的特征，利用组蛋白修饰图谱鉴定 lncRNAs，他们利用这两种组蛋白修饰特征在鼠中基因间挖掘出 1 000 多处 lncRNAs 区域。如果一个基因表达，那么在其启动子区域会富集 H3K4me3 修饰，而在整个基因转录区域则富集 H3K36me3 修饰，因此，通过在全基因组的基因间区域上挖掘“K4-K36 域”，可以获得可能的转录本。对这些转录本的序列特性进行分析，发现超过 97.5% 的 lncRNAs 与蛋白质编码基因不相似，其外显子的保守性比蛋白质编码基因低，但却比其他基因间区域高，与已知 lncRNAs 的保守性相似，并且包含高保守的元件。此外，其启动子区域的保守性也极高，并富含“CAGE 标签”和 RNA PolII 的结合位点。利用 CPC 非编码 RNAs 预测软件发现这些转录本编码蛋白质的能力极低，因此，推断这些转录本为基因间的 lncRNAs，简称 lincRNAs (large intergenic non-coding RNAs)。随后，利用“K4-K36 域”

特征在人中也找到了上千条 lincRNAs[30]。

基因的表观遗传修饰除了 H3K4me3、H3K36me3 外，还有 H3K9me3、H3K27me3 等组蛋白修饰，以及 DNA 甲基化修饰。lncRNAs 的表观遗传修饰特征是一个巨大的信息资源，其利用价值有待挖掘，可以进一步组合各种修饰特征构建模型，提高预测 lncRNAs 的准确性和精度。现将预测 lncRNAs 的计算机方法介绍如下 (表 1)

表1　预测lncRNAs的计算机方法

预测软件	特点	网上服务或软件下载地址
PhyloCSF	基于多个物种的多重比对数据，准确率较高，适用于多个物种，但计算量大，运行速度慢，并且不适用于不保守的物种及没有参考基因组的物种，并且特异性较低。	http://compbio.mit.edu/PhyloCSF
CPC	基于与已知蛋白质编码基因的相似性，运行速度虽然比PhyloCSF快，但比其他软件慢，准确率高，适用于多物种，但预测特异性较低。	http://cpc.cbi.pku.edu.cn/
CPAT	基于序列特征，不用进行序列比对，因此，运算速度快，准确率也较高。	http://lilab.research.bcm.edu/cpat/index.php
CNCI	基于序列本身的特征，计算速度快，准确率高，对多物种均适用。但不适用于插入缺失率较高的序列，如454或PacBio测序从头组装所得的序列。	http://www.bioinfo.org/software/cnci
PLEK	运行速度快，且对插入缺失的序列不敏感，比较适合于无参考基因组的序列。	https://sourceforge.net/projects/plek/files/.
lncRNA-MFDL	结合二级结构和功能域特征，准确率高，适用于多个物种。	http://compgenomics.utsa.edu/lncRNA_MDFL/

3　问题与展望

测序技术的发展给 lncRNAs 的鉴定带来巨大的推动，计算机预测方法的更新与改进也对 lncRNAs 的研究起到重要的作用。但是，由于 lncRNAs 其特殊的序列属性，与 mRNAs 类似又不完全相同，因此，lncRNAs 的鉴定工作仍然面临着挑战。通常，结合几种鉴定方法，如测序和计算机预测方法来获取更加可信的候选 lncRNAs，比如 Liao 等 [31] 利用疟原虫的 RNA-seq 数据，并结合 ORF 过滤、与蛋白质编码基因的相似性过滤，以及 CPC 软件预测，获得较为可靠的疟原虫 lncRNAs。同时，也可以根据 lncRNAs 的其他特征进行限制和过滤，如表达水平。lncRNAs 的表达水平虽然比蛋白质编码基因低，但仍然有一定的表达。通过序列或组蛋白修饰特征预测出来的 lncRNAs 并不一定表达，因此，可以通过检测其表达水平，筛选过滤获得更加可信的 lncRNAs 集合。比如，利用重注释的芯片平台可以检测到一部分 lncRNAs 的表达水平，从而可以在某个组织和生物过程中鉴定 lncRNAs。Mattick 等 [32] 就利用 Allen Brain Atlas (ABA) 原位杂交数据在小鼠大脑中鉴定了 800 多条 lncRNAs。此外，还利用定制的基因芯片在人和鼠的 CD^{8+} T 细胞中发现上千条 lncRNAs[33]。可见，lncRNAs 的鉴定方法是多样化的，一切有利于 lncRNAs 鉴定的特征均可以作为预测的依据，将来所面临的挑战是寻找最具代表性的特征以及创造最具优化的模型。

[参　考　文　献]

[1] Lin R, Maeda S, Liu C, et al. A large noncoding RNA is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. Oncogene, 2007, 26(6): 851-8

[2] Yu W, Gius D, Onyango P, et al. Epigenetic silencing of tumour suppressor gene p15 by its antisense RNA. Nature, 2008, 451(7175): 202-6

[3] Numata K, Kanai A, Saito R, et al. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. Genome Res, 2003, 13(6B): 1301-6

[4] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome. Science, 2005, 309(5740): 1559-63

[5] MacIntosh GC, Wilkerson C, Green PJ. Identification and analysis of Arabidopsis expressed sequence tags characteristic of non-coding RNAs. Plant Physiol, 2001, 127(3): 765-76

[6] Wen J, Parker BJ, Weiller GF. *In Silico* identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula. In Silico Biol, 2007, 7(4-5): 485-505

[7] Borodovsky M, Lomsadze A, Ivanov N, et al. Eukaryotic

gene prediction using GeneMark.hmm. Curr Protoc Bioinformatics, 2003, Chapter 4: Unit4 6

[8] Xue C, Li F. Finding noncoding RNA transcripts from low abundance expressed sequence tags. Cell Res, 2008, 18(6): 695-700

[9] Nakaya HI, Amaral PP, Louro R, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol, 2007, 8(3): R43

[10] Huang W, Long N, Khatib H. Genome-wide identification and initial characterization of bovine long non-coding RNAs from EST data. Anim Genet, 2012, 43(6): 674-82

[11] Imanishi T, Itoh T, Suzuki Y, et al. Integrative annotation of 21,037 human genes validated by full-length cDNA clones. PLoS Biol, 2004, 2(6): e162

[12] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol, 1990, 215(3): 403-10

[13] Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol, 2000, 132: 185-219

[14] Mackey AJ, Haystead TA, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics, 2002, 1(2): 139-47

[15] Chen G, Yin K, Shi L, et al. Comparative analysis of human protein-coding and noncoding RNAs between brain and 10 mixed cell lines by RNA-Seq. PLoS One, 2011, 6(11): e28318

[16] Cabili MN, Trapnell C, Goff L, et al. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev, 2011, 25(18): 1915-27

[17] Pauli A, Valen E, Lin MF, et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res, 2012, 22(3): 577-91

[18] Prensner JR, Iyer MK, Balbin OA, et al. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol, 2011, 29(8): 742-9

[19] Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol, 2010, 28(5): 511-5

[20] Rombel IT, Sykes KF, Rayner S, et al. ORF-FINDER: a vector for high-throughput gene identification. Gene, 2002, 282(1-2): 33-41

[21] UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res, 2014, 42(Database issue): D191-8

[22] Kong L, Zhang Y, Ye ZQ, et al. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res, 2007, 35(Web Server issue): W345-9

[23] Lin MF, Carlson JW, Crosby MA, et al. Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res, 2007, 17(12): 1823-36

[24] Guttman M, Amit I, Garber M, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature, 2009, 458(7235): 223-7

[25] Lin MF, Jungreis I, Kellis M. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. Bioinformatics, 2011, 27(13): i275-82

[26] Wang L, Park HJ, Dasari S, et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res, 2013, 41(6): e74

[27] Sun L, Luo H, Bu D, et al. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res, 2013, 41(17): e166

[28] Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme. BMC Bioinformatics, 2014, 15: 311

[29] Fan XN, Zhang SW. lncRNA-MFDL: identification of human long non-coding RNAs by fusing multiple features and using deep learning. Mol Biosyst, 2015, 11(3): 892-7

[30] Khalil AM, Guttman M, Huarte M, et al. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci USA, 2009, 106(28): 11667-72

[31] Liao Q, Shen J, Liu J, et al. Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data. Parasitol Res, 2014, 113(4): 1269-81

[32] Mercer TR, Dinger ME, Sunkin SM, et al. Specific expression of long noncoding RNAs in the mouse brain. Proc Natl Acad Sci USA, 2008, 105(2): 716-21

[33] Pang KC, Dinger ME, Mercer TR, et al. Genome-wide identification of long noncoding RNAs in CD8+ T cells. J Immunol, 2009, 182(12): 7738-48