DOI:10.12113/j.issn.1672-5565.201809002

基于机器学习的 RNA 编辑位点预测方法综述

冷嘉承[1,2],吴凌云[1,2]*
(1.中国科学院数学与系统科学研究院 应用数学研究所,管理、决策与信息系统重点实验室,国家数学与交叉科学中心,北京 100190;
2.中国科学院大学 数学科学学院,北京 100049)

摘 要:RNA 编辑是一个十分重要的生物细胞分子机制。作为转录后修饰的一步,它可以增加蛋白质组学多样性,改变转录产物的稳定性,调节基因表达等。RNA 编辑失调会导致各种疾病,包括神经疾病和癌症。在动物中,腺苷到肌苷(A-to-I)的编辑是最普遍的。高通量测序技术的进步大大提高了在全局范围内检测和量化 RNA 编辑的能力,使得 RNA 编辑的大规模全基因组分析变得可行,产生了一系列基于高通量测序技术的 RNA 编辑位点预测方法。通过对这些方法进行介绍、总结和分析,为 RNA 编辑的进一步研究提供一些思路。

关键词:RNA 编辑;高通量测序;A-to-I;机器学习
中图分类号:Q522+.6 **文献标志码**:A **文章编号**:1672-5565(2019)01-001-08

Review: prediction of RNA editing sites based on machine learning

LENG Jiacheng[1,2], WU Lingyun[1,2]*
(1. *Academy of Mathematics and Systems Science*, *Institute of Applied Mathematics*, *Chinese Academy of Science*, *Key Laboratory of Management*, *Decision and Information Systems*, *National Center for Mathematics and Interdisciplinary Sciences*, *Beijing* 100190, *China*;
2. *School of Mathematics Sciences*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: RNA editing is an important molecular mechanism of biological cells. As a step of post-transcriptional modification, it can increase proteomic diversity, alter the stability of transcription products, regulate gene expression, and so on. RNA editing disorders can lead to a variety of diseases, including neurological diseases and cancer. Among animals, the editing of adenosine to inosine (A-to-I) is the commonest. Advances in high-throughput sequencing technology have greatly improved the ability to detect and quantify RNA editing globally, which can make large-scale genome-wide analysis of RNA editing feasible, thereby resulting in a series of prediction methods of RNA editing sites based on high-throughput sequencing technology. This article will introduce and summarize these methods and provide new perspectives for further research of RNA editing.

Keywords: RNA editing; High-throughput sequencing; A-to-I; Machine learning

RNA 合成、加工、行使功能和降解是细胞生存的关键,并在许多不同的层面进行着调控[1]。RNA 合成是基因表达的第一步,转录因子调控 RNA 聚合酶 II(Pol II)与启动子结合[2],通过一套非常复杂的操作步骤将 DNA 转录为前体 RNA[3]。前体 RNA 随后被加工产生成熟 mRNA、功能性 tRNA 和 rRNA[4]。RNA 的加工包括(1)加帽:将 7-甲基鸟苷酸(m7G)添加到 5'末端[5];(2)聚腺苷酸化:在 3'末端添加 poly-A 尾巴[6];(3)剪接:去除内含子之后拼接外显子[7];(4)RNA 编辑:修改 RNA 分子序列并导致蛋白质多样性[4,8]。而 RNA 编辑作为 RNA 加工中的一环,起着至关重要的作用。

RNA 编辑有很多种改变 RNA 序列的机制,其中涉及了碱基的插入和缺失以及碱基替换,如胞苷(C)到尿苷(U)和腺苷(A)到肌苷(I)的脱氨基化,以及非模板化的核苷酸添加和插入[3]。到目前为

收稿日期:2018-09-06;**修回日期**:2018-11-20.
基金项目:国家自然科学基金(No.11631014; No.91730301;No.11661141019).
作者简介:冷嘉承,男,博士研究生,研究方向:生物信息学.E-mail:amssljc@163.com.
*通信作者:吴凌云,男,研究员,研究方向:生物信息学.E-mail:lywu@amss.ac.cn.

止,人们已经在真核生物、病毒、古细菌以及原核生物中发现了 RNA 编辑。所以,生物学家一直对 RNA 编辑保持着广泛的关注和强烈的兴趣[1-3, 8]。

在哺乳动物中已经发现了两种类型的 RNA 编辑。一种是由催化性类多肽载脂蛋白 B mRNA 编辑酶(APOBEC)催化的 C 至 U 编辑,其频率相当低。另一种是 A-to-I 编辑,其中腺苷通过腺苷脱氨酶(ADAR)的作用脱氨基成肌苷,其频率非常高。因为人类 RNA 编辑中绝大多数的编辑都属于 A-to-I 编辑,故 A-to-I 编辑在生物细胞分子机制中尤为重要。研究表明,A-to-I RNA 编辑同时也与脑功能、病毒感染和人类疾病有关。例如在人体中,RNA 合成和加工中的错误可能引起神经系统疾病,如三核苷酸扩张性疾病强直性肌营养不良和脆性 X 综合征[8]。这些均显现了 RNA 编辑研究的重要性与意义。

深度测序技术和生物信息学的发展使得人们可以在全局筛选 A-to-I RNA 编辑位点。但是目前为止,通过高通量测序数据准确识别 RNA 编辑事件仍然是一个巨大的挑战。现在的方法通常将短读段映射到参考基因组或转录组,然后去除相同的读段,过滤低质量读段,调用差异信息并且去除已知的单核苷酸多态性(SNP)[6, 7, 10, 15-17]。但是将大量的短读段映射到参考基因组是非常耗时的,并且只有少数流水线和计算工具可公开用于处理 RNA 编辑[18]。最关键的是,由于成本问题,DNA 测序与 RNA 测序一般不会一起进行,所以很难区分新的 SNP 位点和 RNA 编辑位点。实际上,大量的 RNA 编辑位点已被注释为 dbSNP 中的 SNP[19]。

针对这些现象,尤其是为了区分 SNP 位点和 RNA 编辑位点,陆续产生了许多能够较为准确预测 RNA 编辑位点的方法,本文对一些常见方法进行了总结(见表 1),将在第三节对这些方法进行详细的介绍。

表 1 RNA 编辑位点预测方法概览

Table 1 Overview of RNA editing site prediction methods

名字	主要数据类型	主要方法	正样本来源	负样本来源
SVM[9]	目标位点侧翼 DNA 序列	支持向量机	RADAR DARNED	依碱基频率产生的随机 DNA 序列
Multi-Sampled Method[10]	RNA-seq 数据	统计量过滤	无	无
GIREMI[11]	RNA-seq 数据	互信息	由互信息推导出的 RNA 编辑位点	dbSNP 中的 SNP 位点
RDDpred[12]	RNA-seq 数据	随机森林	RADAR DARNED	MES[13] 方法产生的位点
RED-ML[14]	RNA-seq 数据	逻辑回归	多种过滤方法的重叠位点,加以实验验证	随机选取的一些未能通过过滤标准或实验验证的位点,以及 dbSNP 中的位点

注:正样本指发生 RNA 编辑的位点(包括所有类型的 RNA 编辑,以 A-to-I 编辑为主),负样本指发生碱基变化但没发生 RNA 编辑的位点,其中 SVM 方法只针对 A-to-I 编辑。

1 RNA 编辑数据库

已有文献在研究和评价 RNA 编辑位点预测方法时使用了不同的数据集,目前还没有一个通用的、被广泛接受的基准数据集。除了一些文献中提供的独立数据集,随着越来越多的 RNA 编辑位点被发现,目前已经出现了多个关于 RNA 编辑的数据库。这些数据库为研究基于计算的 RNA 编辑位点预测方法提供了很好的训练数据和研究的基础。

1.1 RADAR 数据库

Ramaswami 和 Li 等人于 2013 年创建了 RADAR 数据库[20],包括了人类、苍蝇、老鼠这三个物种(hg19/mm9/dm3)的 A-to-I RNA 编辑数据。该数据库可以根据用户的要求进行筛选,例如基因名称、是否位于 Alu 区域、与其他物种之间的编辑保守性等,并且提供了一些常用数据的直接下载,如 hg19 中的全部 RNA 编辑位点等。在结果中点击位置信息可以链接到 UCSC 基因组浏览器上浏览更详细的信息,并且都附有数据的文献来源。

1.2 DARNED 数据库

DARNED 数据库[21]是由 Kiran 和 Baranov 创建的。它包括多种数据来源:(1)生物信息学分析 cDNA 序列和基因组序列之间的差异;(2)SNP 的分析数据;(3)miRNA 的分析数据;(4)来自同一组织的 RNA 和 DNA 样品的高通量测序结果[21]。最后将 RNA 编辑事件的位点映射到参考人类基因组。此数据库不仅包括 RADAR 中的三个基因组还包括了 hg18 以及 mm10。目前为止该数据库支持三种 RNA 编辑位点的查询方式:根据区域(染色体、位置、组织

特异性等信息)查询;根据基因名称(如 APOB)查询;根据序列本身查询。在此数据库的查询结果中,通过点击位置信息不仅可以链接到 UCSC 上,还可以链接到 ENSEMBL 上,比较方便。对是否与 SNP 混淆,与哪个 SNP 混淆,基因名称,基因区域都有详细的标注,而且都可以通过点击基因信息链接到 NCBI 上进行查询,十分人性化。

2　RNA 编辑位点预测方法

2.1　基于 DNA 序列数据的有监督学习方法

Sun 等人[9]提出了一个支持向量机(SVM)模型,基于 DNA 测序数据对 A-to-I 类型的 RNA 编辑位点进行预测。其核心思想是,RNA 编辑位点主要是由该位点附近的序列决定的。该方法将序列与序列之间的关系进行相关性分析,得到相似矩阵,然后通过映射将其转化到内核空间得到核矩阵,最终利用 SVM 模型进行训练,得到具有判别能力的 RNA 编辑位点分类器(见图 1)。

序列间的相关性是由字符串距离刻画的:

$$D(a,b) = \frac{w}{L_1} D_{\text{Edit}}(a,b) + \frac{1-w}{L_2} D_{\text{Hamming}}(a,b) \quad (1)$$

其中 a,b 分别代表输入的两个字符串,D_{Edit} 代表的是编辑距离,即 a,b 之间互相转换最少需要插入、删除或者替换多少个字符,而 D_{Hamming} 代表的是汉明距离,与编辑距离不同,只允许替换。而 L_1, L_2 分别代表计算编辑距离和汉明距离所用的字符串长度(见图 1)。w 是 0 到 1 之间取值的权重。D 的值越大,代表两个序列之间的相似性越低。该方法在 LIBSVM[22]中使用字符串核函数将字符串数据转换为向量空间。字符串核函数是对字符串类型数据进行操作的核函数,可写为:

$$K(a,b) = \exp(-gamma \times D(a,b)^2) \quad (2)$$

其中 D 是从等式(1)导出的组合距离。伽玛参数定义单个训练实例影响到达的距离,低值表示“远”,高值表示“近”。

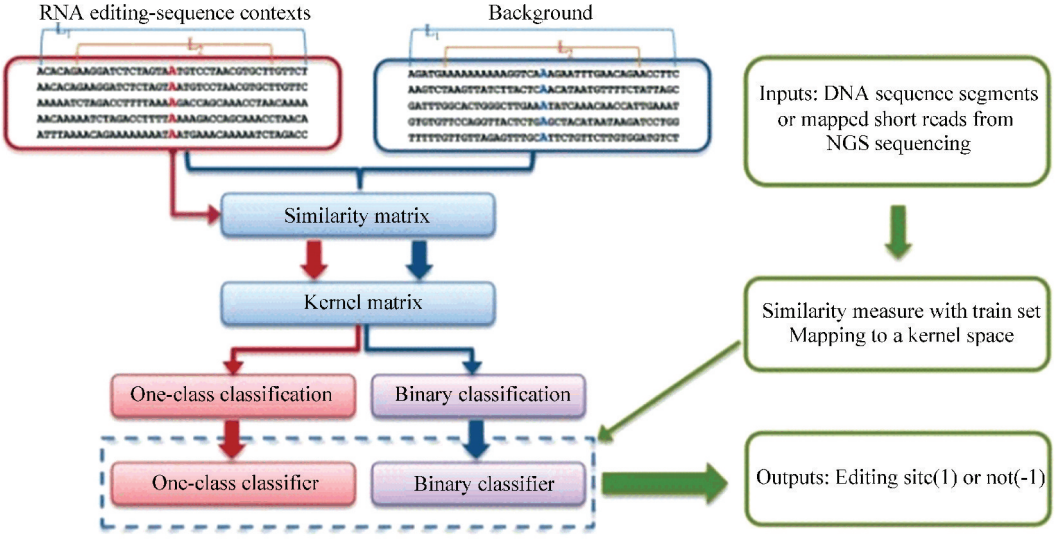

图 1　SVM 方法流程图[9]

Fig.1　Flow chart of SVM method[9]

该论文还尝试了单类 SVM。与传统的支持向量机相比,单类支持向量机尝试学习一个决策边界,实现样本与原点之间的最大分离。他们根据 Schölkopf 等人的研究结果[23],使用二进制值 1 作为编辑事件,-1 作为非编辑事件。该文采用 5 折交叉验证的方式得到,仅用正样本作为训练数据的单类 SVM 的性能要远低于双类 SVM,单类 SVM 在参数 nu=0.5 时的精确度在果蝇、小鼠、人类数据集上分别是 0.489, 0.495, 0.498,而双类 SVM 模型却分别达到了 0.75, 0.85, 0.79 左右。并且单类 SVM 方法通常无法权衡在正负样本上性能的差异。该文还用基于人类样本训练的模型在一个 Sanger 测序集上进行了验证试验,其中 79.3%(46/58)的位点被成功预测。

2.2　基于 RNA-seq 数据的无监督学习方法

仅仅基于 DNA 序列的预测方法需要可靠的 RNA 编辑位点数据作为训练样本,而且这类方法对于训练样本的依赖性非常高。考虑到不同物种和不同类型样本中发生 RNA 编辑概率的差异,以及准确的 RNA 编辑位点数据的缺乏,目前更为常见的是从 RNA-seq 数据中识别出 RNA 编辑位点。按照使用的机器学习方法的不同,基于 RNA-seq 的 RNA 编辑

位点识别方法又可以大致分为两类:基于无监督学习的过滤方法,和基于有监督学习的机器学习方法。第一类方法通过比较 RNA-seq 数据和参考基因组的差异,获得了潜在的 RNA 编辑位点,然后通过与 SNP 数据库的比较以及其他指标进行过滤,去除假阳性位点,最终获得较为可靠的 RNA 编辑位点。而第二种方法则通过 RNA-seq 数据和参考基因组提取出一些特征,建立机器学习模型,利用已知的 RNA 编辑位点数据进行有监督训练,获得 RNA 编辑位点的预测模型。

Ramaswami 等人[10]提出了整合多样本 RNA-seq 数据识别 RNA 编辑位点的方法。为了利用大量可公开获得的 RNA-seq 数据集来发现 RNA 编辑位点,该文章提出了两种相关且互补的方法,以使用来自单个物种中的多个个体的 RNA-seq 数据准确鉴定 RNA 编辑位点。在第一种方法(见图 2(a))中,在每个 RNA-seq 样品中分别将测序读段映射到(非匹配的)基因组参考序列后,找出 RNA 的变化,并且将已知常见基因组 SNP 去除,以此将 RNA 编辑位点与其余稀有 SNP 区分开来。这主要是因为相同的编辑位点通常存在于不同的个体中,而罕见的 SNP 很可能不存在。

在第二种方法中(见图 2(b)),将不同样本的 RNA-seq 的读段汇总到一起进行比对,从而提高找出 RNA 变异的灵敏度。然后如第一种方法一样,按接下来的步骤排除掉 SNP,并且找出 RNA 编辑。由于罕见的 SNP 不可能出现在多个个体中,所以在汇聚后的比对中将以非常低的频率存在。

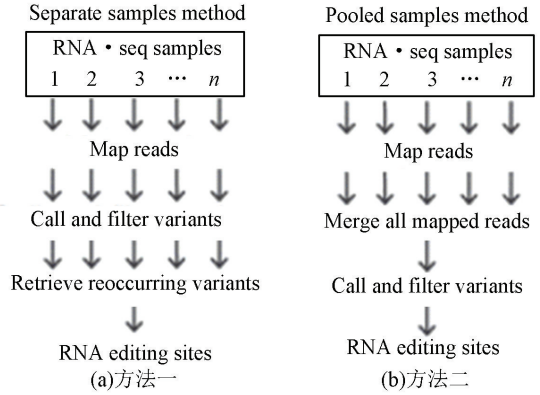

图 2 Multi-Sampled Method 的两种方法[10]

Fig.2 Two methods of Multi-Sampled Method[10]

而 GIREMI 方法[11]则是典型的仅需要单组 RNA-seq 的方法,GIREMI 将 RNA-seq 读段中单核苷酸变异(SNV)对的统计推断模型与机器学习方法结合,以预测 RNA 编辑位点。GIREMI 的输入包括来自 RNA-seq 数据集的 SNV(错配)列表和公共数据库(如 dbSNP)中已知的 SNP,输出是预测的 RNA 编辑位点及其编辑水平。除了公开的 SNP 信息外,GIREMI 仅仅使用感兴趣的 RNA-seq 数据集进行所有分析,而不依赖于任何其他基因组或 RNA-seq 数据集,因此这种方法适用于更广的范围(见图 3)。

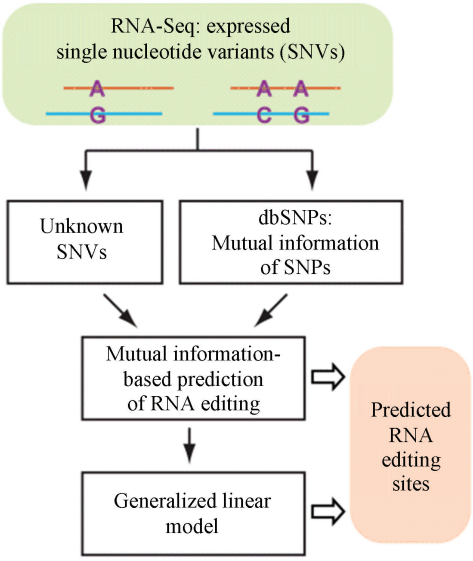

图 3 GIREMI 流程图[11]

Fig.3 Flow chart of GIREMI[11]

GIREMI 分为两个步骤,首先进行 SNV 位点的互信息计算:

对于每个 SNV,我们考虑所有可能的碱基 A,T,C,G 作为变量 s_i 的四种可能状态。对于表示一对碱基 (s_i,s_j) 的联合变量,总共有 16 种可能状态。各种状态 s_i , s_j 或 (s_i,s_j) 的概率可以使用最大似然法进行估计得到。考虑到所有可能的测序错误以及实际数据中的低测序深度,假设在实际数据中未观察到的状态的概率值为 0.01。因此 (s_i,s_j) 的互信息是:

$$I(s_i,s_j)=\sum_{n_i\in N}\sum_{n_j\in N}p(n_i,n_j)\ln\left(\frac{p(n_i,n_j)}{p(n_i)p(n_j)}\right) \quad (3)$$

其中 $N=\{A,T,C,G\}$, n_i 和 n_j 分别表示 s_i 和 s_j 的状态。而 SNP s_i 的信息值被定义为:

$$I(s_i)=\frac{\sum_{s_j\in S}I(s_i,s_j)}{T} \quad (4)$$

其中 S 代表带有 s_i 的 SNP 对的集合, T 代表集合 S 中 SNP 对的个数。

这样,每个 RNA-seq 数据样本均产生一个基于互信息的 SNP 信息值分布 $I(s_i)$ (见图 4)。同样,对这个样本的每个 SNV 位点以同样的方式(SNV 对)求一个信息值,取 95%的置信度,如果该位点的信息值落在 SNP 信息分布的置信区间外,则判定该位点是 RNA 编辑位点,否则为 SNP 位点。

其次,为了提高 RNA 编辑位点判定的精确度,GIREMI 用第一步中识别出的 RNA 编辑位点作为正样本来训练广义线性模型。该模型采用了两个特征,一个是读段中杂合 SNP 等位基因比率与 SNV 等位基因比率的差值的绝对值 d,另一个则是基于序列本身特征的复合序列得分 c(通过位置权重矩阵中+1 和−1 位置计算)。其回归模型为:

$$Y = g^{-1}(\beta_0 + \beta_1 d + \beta_2 c) = \frac{1}{1 + e^{-(\beta_0+\beta_1 d+\beta_2 c)}} \quad (5)$$

其中,β_0,β_1,β_2 分别为需要学习的系数,g 为逻辑回归函数。

该文只对 GM12878 数据集进行了性能测试,并且达到了 99.4%的准确度,但是其准确度的定义为 100%−被预测为 RNA 编辑位点中 SNP 位点的百分比。

2.3　基于 RNA-seq 数据的有监督学习方法

不同于 GIREMI 基于自主产生的正样本来训练广义线性模型,RED-ML和RDDpred是基于已知样本进行有监督学习训练的 RNA 编辑位点检测模型(见图 5)。

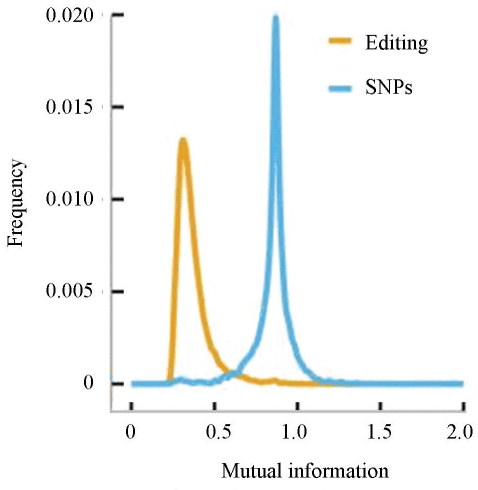

图 4　SNP 与 RNA 编辑位点的信息分布[11]

Fig.4　Information distribution of SNP and RNA editing sites[11]

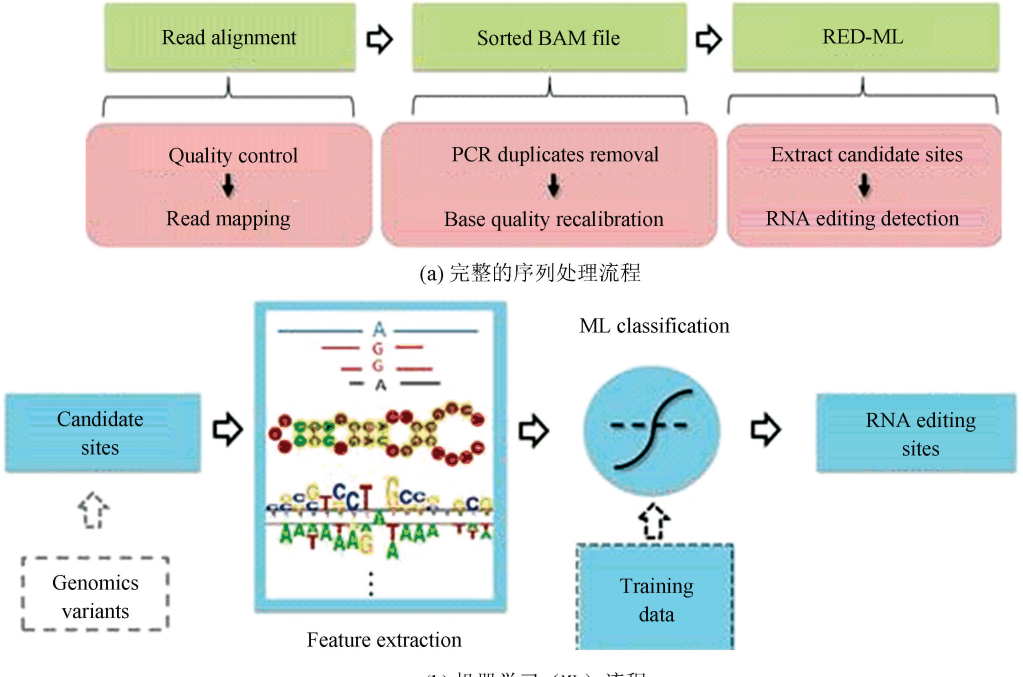

图 5　RED-ML 流程图[12]

Fig.5　Flow chart of RED-ML[12]

Xiong 等人[12]建立了一个基于机器学习的 RNA 编辑位点预测工具 RED-ML,并选择了逻辑回归(Logistic regression)进行模型的训练。RED-ML 使用的特征有三大类。第一类是基本读段特征,包括候选位点的支持读段数量和计算出的编辑频率。第二类特征与可能的测序失误和错位有关,包括支持读段的图谱质量、候选位点在定位读段中的相对位置、链偏差的指示、候选位点是否落入简单重复区域等。第三类是基于 RNA 编辑的已知属性,如编辑类型(是否为 A-to-I),候选位点是否在 Alu 区域以及它的序列上下文。需要注意的是,与前两类特征不同,第三类特征不能直接用来过滤非 RNA 编辑位点。然而,它们仍然可以提供有价值的信息,通过机器学习方法,将不同来源的信息结合起来做出分类决策。

作为一款 RNA 编辑检测软件工具,RED-ML 的输入是一个 BAM 文件,也可以利用相应的基因组差异信息。RED-ML 将提取候选 RNA 编辑位点及其相应的特征,然后应用逻辑回归分类器以相应的置信度检测真正的 RNA 编辑位点。RED-ML 可以仅基于人类 RNA-seq 数据执行全基因组 RNA 编辑检测,也可以利用匹配的 DNA-Seq 数据,并与其他常见的 RNA-seq 数据分析步骤很好地结合。

该文用从其他文章中已发表的数据自己筛选出正负样本进行训练,其中随机选取了 80%作为训练集,20%作为测试集。在 ROC 曲线上的 AUC 达到了0.98,在 PR 曲线上的 AUC 达到了 0.94。并且该文在 CH24T、CH62T 和 HeLa 样本上做了 RNA-seq 验证实验,取阈值为 0.5 时成功验证了 90%的 RNA 编辑位点。

尽管 RNA-seq 数据可以用于 RNA 编辑位点检测,但目前用 RNA-seq 进行 RNA 编辑位点检测的算法也具有相当大的假阳性(False positive)风险,这是用 RNA-seq 检测 RNA 编辑位点的最大挑战之一。由于短读段误对齐而产生的假阳性,本质上是由以下几种因素导致:(1)基因组序列固有的重复片段;(2)模糊的剪切连接;(3)个体之间普遍的多态性;(4)测序读段的短缺。RDDpred (RNA/DNA Differences prediction)[14]是一种基于随机森林算法的 RNA 编辑位点预测方法,能够大大减少样本中的假阳性数据,从而提升 RNA 编辑位点的预测准确率。

RDDpred 首先对输入的测序数据进行初始比对,产生编辑位点候选者,然后从中选择满足特定条件的样本作为训练数据(见图 6)。该方法使用 RNA 编辑数据库 RADAR 和 DARNED 中的 RNA 编辑位点作为正样本。而负样本则通过 MES (mapping error set)方法来收集,这种方法可以在比对时计算基因组内的容易导致错误的区域[13]。从 RDD 候选者中收集正/负样本后,所有剩余的样本被视为预测目标。然后 RDDpred 建立了一个包含 15 个特征的随机森林预测器来预测 RNA 编辑位点。

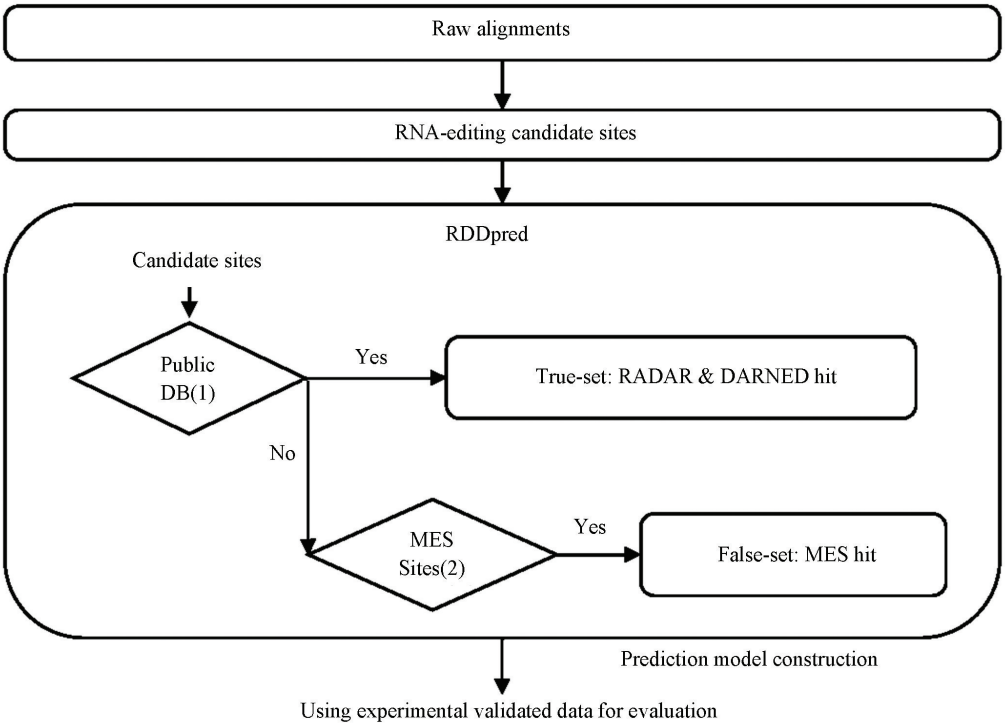

图 6 RDDpred 整体流程图[14]

Fig.6 Overall flow chart of RDDpred[14]

RDDpred 用来自 Bahn 和 Peng 的小组进行的独立研究的两个数据集进行了评估[13, 15]。在 Bahn 的研究中,RNA-seq 产生了 115 132 348 个读段,RDDpred 检测到6 856 440个初始 RDD 并预测了 105 564个 RNA 编辑位点。在 Peng 的研究,RNA 序列产生了 583 640 030 个读段,RDDpred 检测到 58 666 976个初始 RDD 并预测3 076 908个 RNA 编辑位点。虽然这两项研究都使用人体组织,但它们产生了不同数量的 RNA 编辑位点(105 564 与 3 076 908),这表明 RNA 编辑事件的表达模式在两种环境下可能不同。而同时,该方法用自己的模型验证了 Bahn 和 Peng 方法发现的编辑位点,分别成功验

证了 95.32%(3 947/4 141) 和 90.37%(20 504/22 688),并且都大幅度减少了错误的编辑位点,NPV 分别达到了 84.21%和 75.86%。

3　讨　论

本文介绍了五种 RNA 编辑位点预测方法。第一种方法仅仅需要 DNA 序列。基于 DNA 序列的方法主要利用了序列间的相似性进行预测,然而其缺点就是 DNA 序列中包含的信息是有限的,这使得其性能没有达到较高的水准(精确度大概在 0.7 左右)。此外,这种方法无法用于研究 RNA 编辑在不同条件(例如疾病)、不同个体、不同组织中的差异。

后四类方法则基于 RNA-seq 测序数据和机器学习模型,它们的共同特点就是高度依赖于高通量测序的质量与深度。与此同时,还有另外一个因素也限制了这些方法的效果,那就是 RNA 编辑水平,如果该位点处于一个比较低的编辑水平,那么预测难度将会大大提升。

RED-ML 方法提出了大量可能与 RNA 编辑事件有关的特征,通过逻辑回归模型进行整合。它的一个缺陷就是模型对训练数据的依赖性很高,例如用人类数据得到的模型,在其他动物上的预测效果并不理想。而且这种方法对于比对工具有很强的依赖性(目前版本只优化了 BWA 和 TopHat2),这导致用户在选择不同的比对软件时会产生截然不同的结果阻碍研究的进行。

如何准确地区分 SNP 和 RNA 编辑,这是 RNA 编辑位点识别的一个核心问题。GIREMI 方法通过计算不同 SNV 位点之间的互信息能够更准确地区分 SNP 和 RNA 编辑。但是其代价就是计算量的增加,如果考虑到时间成本的话,可能会对效率有所影响。并且要求对测序读段的长度不能太短,否则无法覆盖两个感兴趣位点。

RDDpred 方法通过 RNA 编辑数据库和 MSE 对数据进行了筛选,然后用随机森林模型进行训练。该文章主要提出了组织特异性导致错误正样本的问题,因为在 RADAR 和 DARNED 数据库中有 97%的编辑位点都是只存在于一个组织中,如果不清楚地将其筛选出来将会导致“预测危机”,因为这将从根本上(样本上)导致训练失败,从而降低预测性能。如果将其剔除假阳性的方法引用到其他模型中,或许会产生更好的效果,值得让人期待。

4　总　结

RNA 编辑事件的识别对于理解转录后调控是非常重要的。本文首先介绍了 RNA 编辑的概念和意义,然后介绍了两个现有的 RNA 编辑数据库(DARNED、RADAR)。而随着机器学习的发展,其在 RNA 编辑的相关研究中也起到了重要的作用,故本文对已有的 RNA 编辑位点预测方法进行了概述与讨论,得到以下结论:(1)对于 RNA 编辑位点,我们更关注其在样本中的表达;(2)对于 RNA 编辑来说,仍没有一套像人类参考基因组一样较为完备的标准;(3)虽然已经有了一些基于机器学习的预测 RNA 编辑位点的方法,但是并没揭示 RNA 编辑的本质,即提取到判别 RNA 编辑位点的本质特征。所以 RNA 编辑领域的研究还有很多亟待解决的问题和现象,希望以后能够通过更深层次的模型去解释 RNA 编辑,从而促进相关疾病的研究以及精准医疗的发展。

参考文献(References)

[1] ROSENTHAL J J, SEEBURG P H. A-to-I RNA editing: effects on proteins key to neural excitability[J]. Neuron, 2012, 74(3): 432-439. DOI: 10.1016/j.neuron.2012.04.010.

[2] GARRETT S, ROSENTHAL J J. RNA editing underlies temperature adaptation in K+ channels from polar octopuses [J]. Science, 2012, 335(6070): 848-851. DOI: 10.1126/science.1212795.

[3] CHEN L, LI Y, LIN C H, et al. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma[J]. Nature Medicine, 2013, 19(2): 209-216. DOI: 10.1038/nm.3043.

[4] LI M, WANG I X, LI Y, et al. Widespread RNA and DNA sequence differences in the human transcriptome[J]. Science, 2011, 333(6038): 53-58. DOI: 10.1126/science.1207018.

[5] SCHRIDER D R, GOUT J F, HAHN M W. Very few RNA and DNA sequence differences in the human transcriptome [J]. PLoS One, 2011, 6(10): e25842. DOI: 10.1371/journal.pone.0025842.

[6] RAMASWAMI G, LIN W, 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.

[7] PICARDI E, D'ANTONIO M, CARRABINO D, et al. ExpEdit: a webserver to explore human RNA editing in RNA-seq experiments[J]. Bioinformatics, 2011, 27(9): 1311-1312. DOI: 10.1093/bioinformatics/btr117.

[8] LIANG G, KITAMURA K, WANG Z, et al. RNA editing of hepatitis B virus transcripts by activation-induced cytidine deaminase[J]. Proceedings of the National Academy of

Science, 2013, 110(6): 2246-2251. DOI:10.1073/pnas.1221921110.

[9] SUN J M, YANG D M, OSMARK P, et al. Discriminative prediction of A-to-I RNA editing events from DNA sequence [J]. PLoS One, 2016, 11(10): e0164962. DOI:DOI:10.1371/journal.pone.0164962.

[10] RAMASWAMI G, ZHANG R, PISKOL R, et al. Identifying RNA editing sites using RNA sequencing data alone [J]. Nature Methods, 2013, 10(2): 128-132.

[11] ZHANG Q, XIAO X. Genome sequence-independent identification of RNA editing sites[J]. Nature Methods, 2015, 12(4): 347-350. DOI: 10.1038/nmeth.3314.

[12] XIONG Heng, LIU Dongbing, LI Qiye, et al. RED-ML: a novel, effective RNA editing detection method based on machine learning[J]. Gigascience, 2017, 6(5): 1-8. DOI:10.1093/gigascience/gix012.

[13] PENG Zhiyu, CHENG Yanbing, TAN B C, et al. Comprehensive analysis of RNA-seq data reveals extensive RNA editing in a human transcriptome[J]. Nature Biotechnology, 2012, 30(3): 253-260. DOI:10.1038/nbt.2122.

[14] KIM M S, HUR B, KIM S. RDDpred: a condition-specific RNA-editing prediction model from RNA-seq data [J]. BMC Genomics, 2016, 17 (Suppl 1): 5. DOI:10.1186/s12864-015-2301-y.

[15] BAHN J H, LEE J H, LI G, et al. Accurate identification of A-to-I RNA editing in human by transcriptome sequencing[J]. Genome Research, 2012, 22(1): 142-150. DOI:10.1101/gr.124107.111.

[16] PARK E, WILLIAMS B, WOLD B J, et al. RNA editing in the human ENCODE RNA-seq data [J]. Genome Research, 2012, 22(9): 1626-1633.

[17] ZHU Shanshan, XIANG Jianfeng, CHEN Tian, et al. Prediction of constitutive A-to-I editing sites from human transcriptomes in the absence of genomic sequences[J]. BMC Genomics, 2013, 14(1): 206. DOI:10.1186/1471-2164-14-206.

[18] PICARDI E, PESOLE G. REDItools: high-throughput RNA editing detection made easy [J]. Bioinformatics, 2013, 29(14): 1813-1814. DOI: 10.1093/bioinformatics/btt287.

[19] EISENBERG E, ADAMSKY K, COHEN L, et al. Identification of RNA editing sites in the SNP database [J]. Nucleic Acids Research, 2005, 33(14): 4612-4617. DOI: 10.1093/nar/gki771.

[20] RAMASWAMI G, LI J B. RADAR: a rigorously annotated database of A-to-I RNA editing [J]. Nucleic Acids Research, 2014, 42(Database issue): D109-113. DOI: 10.1093/nar/gkt996.

[21] KIRAN A, BARANOV P V. DARNED: a Database of RNA Editing in humans [J]. Bioinformatics, 2010, 26(14): 1772-1776. DOI: 10.1093/bioinformatics/btq285.

[22] CHANG C C, LIN C J. LIBSVM: a library for support vector machines[J]. DNA Research, 2011, 2(3): 1-27.

[23] SCHOLKOPF B, PLATT J C, SHAWE-TAYLOR J, et al. Estimating the support of a high-dimensional distribution [J]. Neural Computation, 2001, 13(7): 1443-1471. DOI: 10.1162/089976601750264965.