

k -gram 方法识别 microRNA 前体*

杨良怀^{1,2)} 吕丕明³⁾ 陈立军²⁾ 邓明华^{4,5)**}

(¹⁾浙江工业大学信息工程学院, 杭州 310032; ²⁾北京大学信息科学与技术学院, 北京 100871;

³⁾中国科学院研究生院, 北京 100049; ⁴⁾北京大学数学科学学院, 北京 100871; ⁵⁾北京大学理论生物中心, 北京 100871)

摘要 MicroRNAs (miRNAs) 是动植物中较短的参与调控基因表达的功能性非编码 RNA 序列. 第一个 miRNA 是通过实验手段发现的, 然而通过实验手段识别 miRNA 在技术上仍然具有很大的挑战性和不完整性. 因此, miRNA 基因识别需要寻求计算方法来弥补实验方法的不足. 提出了一个全新的 miRNA 前体的识别方法. 在构造识别模型中, 把初级序列和序列二级结构相结合, 采用 k -gram 方法把序列信息映射到高维特征空间中, 然后通过特征选取方法提取特征, 并用这些特征为 miRNA 前体的识别构造了基于 SVM 的识别模型. 同时, 采用隐马尔可夫模型(HMM)的学习方法进行了比较. 实验结果表明, 该方法是有效的, 可以达到较高的敏感性和特异性.

关键词 microRNA, 基因识别, 支持向量机, 隐马尔可夫模型, microRNA 前体

学科分类号 Q612

1 引言

自从 Watson-Crick 在 1953 年发现 DNA 双螺旋结构以来, 生物学里就有这样一条“中心法则”, 遗传信息的流向 DNA→RNA→蛋白质, RNA 在细胞中的作用仅限于信使 RNA(mRNA), 转运 RNA(tRNA)和核糖 RNA(rRNA). 一直认为那些酶和生物催化剂必定是蛋白质. 然而, 20 世纪 80 年代中期, Zaug 与 Cech^[1]发现 RNA 充当酶的作用, 引发了第一次 RNA 革命. 过去几年中 RNA 干扰(RNAi)和 microRNA 的发现引发了第二次 RNA 革命, 它们可以分割 / 降解 / 抑制与其互补的靶标——信使 RNA(mRNA), 它们调控着复杂的遗传通道. microRNA 在 2002 年被《科学》杂志(Science)列为十大科技突破的榜首. 现在生物学界已认识到没有对 RNAs 功能的了解, 要对细胞进程有一个全面的了解是很困难. 如何识别非编码功能性 RNA 基因, 已成为当前生物学界的一个重大研究课题^[2~4]. 因此, 非编码 RNAs(ncRNAs)的计算方法识别是生物信息学主要研究目标之一. 本文关注的是如何用计算方法来识别 microRNA.

microRNAs(miRNAs) 是动植物中参与调控基因表达的功能性非编码 RNA 短序列. 它们被誉为“生物学中的暗物质, 存在于我们周围却几乎逃脱

了所有的视线.”, 成熟 miRNAs 是从具有长发夹结构的 miRNAs 前体(pre-miRNA)中产生的, 它们包含在发夹结构的双臂中. 所有已报道的成熟 miRNAs 长度在 17 到 29 个核苷(nt)之间, 他们中的大多数是在 21~25 nt 范围. miRNAs 广泛存在于真核生物中, 从拟南芥(*Arabidopsis thaliana*)、线虫(*Caenorhabditis elegans*)、果蝇(*Drosophila melanogaster/pseudoobscura*)到鼠(*Mus musculus*)和人类(*Homo sapien*)^[5]等等. MicroRNAs 通过对信使 RNA 的降解或翻译抑制来实现对真核生物基因表达的调控^[6]. miRNAs 调控功能的范围包括细胞增殖、脂肪新陈代谢, 线虫的神经元结构分化、神经系统疾病, 哺乳动物造血系统分化的调控、发育、细胞的死亡、癌症和植物花叶生长等. miRNAs 通过和靶标 mRNAs 的碱基互补配对降低它们的翻译水平. 在 miRNAs 和它们的靶标 mRNAs 之间的碱基配对通常包括短鼓凸(bulge)及(或)失配.

在动物中, miRNAs 倾向于绑定到靶标 mRNA

*国家自然科学基金(30570425), 国家重点基础研究发展计划(2003CB715903)资助项目.

** 通讯联系人. Tel: 010-62767562, Fax: 010-62751801, E-mail: dengmh@math.pku.edu.cn

收稿日期: 2006-06-20, 接受日期: 2006-08-24

的 3' 非翻译区 (3' UTR) 来抑制翻译. 动物 miRNAs 与其靶标配对区通常包含短的碱基鼓凸或失配. 作为对比, 在所有已知的案例中, 植物 miRNAs 和它们的靶标 mRNAs 绑定时, 引发靶标 mRNA 的降解或翻译抑制, 其碱基失配个数比动物少, 通常少于 3 个.

1.1 相关工作

通过实验手段识别 miRNA 在技术上仍然具有很大的挑战性和不完整性. 原因有二: 首先, miRNAs 的表达模式与特定组织和特定时间有很强的关联. 其次, 来自 mRNAs 和其他内生 (endogenous) 非编码 RNAs 的降解物同 miRNAs 共存, 而且它们有时在细胞提取的小分子 RNA 样本中占据优势. 一个有机体可能包含数以百计的 miRNAs, 其中的一些只有在特定阶段、特定组织、特定时间或特定类型的细胞中表达. 这样, 那些仅限在非丰富的细胞类型或特定环境条件下表达的 miRNAs 可能无法克隆. 因此, 研究人员展开了计算方法的研究以弥补实验方法的不足.

许多 miRNAs 已经通过各种计算方法检测出来. 如通过比较基因组学方法, 可以发现完全新的 RNA 族^[7,8]. 到目前为止, 在不同的生物体中已识别出超过 1 600 个 miRNAs^[9]. 多种计算方法被应用到一些动物基因组中, 包括果蝇、线虫和人类^[7, 8, 10]. 他们使用如下策略: a. 利用同源性搜索已知 miRNA 基因的直系同源 (ortholog) 和旁系同源 (paralog)^[11~15]. 采取该策略的根据是注意到如下事实: 一些 miRNA 在很长的演化历史中具有保守性, 从而意味着它们的序列不是任意的. 成熟 miRNA 以及 miRNA 前体所形成的发夹结构上的序列保守性, 有助于进行基因组规模 miRNA 搜索. b. 在已知 miRNA 附近搜索基因簇^[16~19]. 这一策略非常重要, 因为快速演化的 miRNA 基因往往会串联排列, 类似于操纵子族中基因的表现, 一般的方法很难检测这些发散的基因序列. c. 其他不依赖于同源性和 miRNA 基因簇的基因搜索法^[7, 8, 10, 19, 20]. 该方法利用近亲物种中 miRNA 基因序列的保守性、非编码性、以及前体可形成潜在茎环结构等特性来给候选 miRNA 序列计分 (scoring). 此外, 在文献^[21]中提出了采用隐马尔可夫模型的方法对人类基因组进行 miRNA 检测.

上述 miRNA 计算识别方法面临准确率低的问题 (50%~70%), 并且对 miRNA 基因的从头预测 (*de novo* prediction) 很大程度上尚未解决. 如何准确

定位成熟 miRNA 也尚待研究. 我们注意到以上方法的一大缺陷, 是没有任何系统的方法把候选 miRNA 序列的一级结构和二级结构的信息融合起来, 用以捕捉序列数据中可能存在的微弱“信号”. 在已有方法中所采用的一些启发式参数也有待进一步研究, 如麻省理工学院研究人员 Bartel 等在其所提出的 miRscan^[8]方法中, 候选 miRNA 前体经 RNAFold^[22]折叠后的最小折叠自由能阈值下限设为 25 (即 $\Delta G_{\text{folding}} \leq -25$ kcal/mol), 但我们经过对已知的 miRNA 前体进行折叠后发现, 许多最小折叠自由能小于该阈值, 如线虫 miRNA 前体 cel-mir-261 经折叠后其 $\Delta G_{\text{folding}} = -7.08$ kcal/mol. 因此, 如何避免设立这些人为参数成为识别算法的一个重要方面. 还有一个困难在于生物学界对 miRNA 基因的转录机制到目前为止并不清楚. 基于以上各方面的考虑, 机器学习方法成为识别算法的一个很好选择.

本文针对以上问题提出了一个基于支持向量机的 miRNA 基因识别算法, 并和现有的预测算法^[21]进行了比较. 实验结果表明, 我们所提出的方法是有效的, 可以达到较高的敏感性 (sensitivity) 和特异性 (specificity).

1.2 论文的组织

论文其余部分组成如下: 第 2 节阐述 miRNA 前体识别算法, 包括数据编码和特征映射、特征选取方法; 第 3 节讲述数据产生、实验方法和实验结果; 第 4 节给出本文小结.

2 miRNA 前体识别算法

由于面对的数据是字符串序列 (RNA 序列), 为了利用机器学习算法, 一个重要的任务是对数据进行合适的特征映射. 本节探讨如何设计针对 miRNA 前体进行特征映射, 籍此我们构造了基于支持向量机的 miRNA 基因识别模型; 同时作为比较, 我们实现了隐马尔可夫模型.

2.1 miRNA 前体的二级结构编码与输入序列的概念视图

通过计算方法对原核和真核基因组的蛋白质编码基因检测已经取得了相当的成功. 这些计算方法利用了像开放读码框 (ORF) 和密码子等基因特性. 设计了许多讯号检测器来探查各种信号, 如剪切位点、起止密码子、分支点、转录的启动子和终止子、多聚腺苷酸化点、核糖体的结合位点、拓扑异构酶 II 的结合点、拓扑异构酶 I 的剪切点以及各种

转录因子的结合点和 CpG 岛.

与蛋白质编码基因不同, 非编码 ncRNAs 在它们的初级序列中缺乏可以可靠地用于识别算法的通用统计信号, 而为 miRNA 寻找类似统计信号并非易事. 通常在 ncRNAs 的基因中只有微弱保守的启动子和终止子信号^[23]. 为使用计算方法预测 miRNAs 前体, 我们需要定义能把已知 miRNA 序列和随机基因序列区别开来的序列和结构属性, 并使用这些属性作为约束在目标基因序列中的基因间隔区或整个基因组中检测候选 miRNAs 前体. 因此, 对于 miRNAs 需要寻找新颖的特征信息.

此外, 对于 miRNAs 计算识别更具挑战性的情形是已知的 miRNAs 基因数量较少. microRNA 序列只能在核苷酸级别上比较, 不能像编码基因那样可在氨基酸级别比较, 而且 miRNA 序列非常短. 如前所述, 对动物来讲成熟 miRNA 大约有 17~25 nt, 它的前体大约 100 nt. 因此, 当搜索 miRNAs 这些弱保守基因时, 和那些随机产生的序列区分的难度要比搜索蛋白质编码基因时困难. 此外, 即便对于那些存在大 RNA 族的情形, 序列保守性也通常在二级结构这一级, 也就是说, 所保守的是碱基配对而不是单个碱基序列. 因此, 单单序列比对将无法识别那些在其初级序列上分化得很远, 但仍保持其碱基配对结构的 miRNAs. 基于此, 我们在特征提取时应当考虑二级结构信息.

EST 搜索表明, 一些人类和鼠类的 miRNAs 是连同它们的上游和下游附近的基因一起被转录

的^[24]. 最近的一项研究表明, microRNA 基因是由 RNA 聚合酶 II 转录的^[25]. 据此, 我们可以设法利用可能存在于 miRNA 前体上下游中的一些信号. 我们把 miRNA 可能的转录方式分为以下两种类型: 第一类, 联合转录 miRNA. miRNA 位于标注宿主基因的内含子中. 对于这种情形, miRNA 与宿主基因共享相同的 $\pm 1\ 000$ nt 上下游序列. 第二类, 独立转录 miRNA. 这些 miRNAs 离那些标注基因不远. 我们进一步把它们分为两类: a. 聚集的 miRNAs. 使用在簇中第一个 miRNA 前体的 -1 000 nt 上游序列和最后一个 miRNA 前体的 +1 000 nt 下游序列; b. 非聚集的 miRNAs. 使用 miRNAs 前体的 $\pm 1\ 000$ nt 上 / 下游序列.

为了捕捉二级结构的信息, 我们首先使用维也纳 RNA 软件包 RNAFold^[22] 预测 miRNA 前体的二级结构. 接下来, 为了便于数据处理, 对碱基对作如下编码: A: U- “1”, C: G- “2”, G: C- “3”, G: U- “4”, U: A- “5”, U: G- “6”, 其他 - “0”. 例如, 线虫的 cel-mir-1 miRNA 前体的二级结构编码见图 1. 在图 1 中我们忽略环和茎部起始失配部份, 因为它们变化大且保守性低.

这样我们得到了如图 2 (上部) 所示的一个输入序列的概念视图. 每个输入有 4 个成分: miRNA 前体上游序列、miRNA 前体的初级序列、miRNA 前体二级结构编码序列以及 miRNA 前体下游序列. 因此, 输入既包含了初级结构的信息又包含了二级结构的信息.

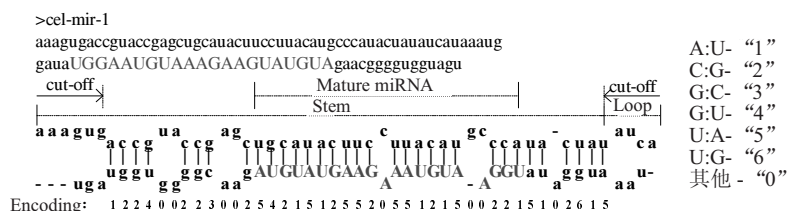


Fig. 1 Encoding the secondary structure

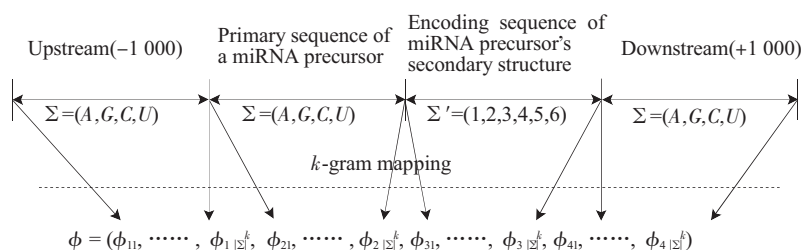


Fig. 2 A conceptual view of the input sequence and its k -gram feature mapping

2.2 特征生成和特征选择

为了使机器学习运算法则能够从已知的 miRNA 序列中进行学习, 需要把输入序列映射到特征空间中的一个特征向量. 在特征生成中, 使用在特征映射中广泛使用的被称为 k -gram 频度^[26]计数方法. 令 Σ 表示一个字母表, 其长度是 $|\Sigma|=L$. 用 X 表示一个来自 Σ 的字母序列. 给定 $1 \leq k < L$, k -gram 是一个长度为 k 的连续子序列. 我们用所有来自 Σ^k 中 k 长度的子序列 α 作为指标的向量来定义特征映射. 特征映射 $\Phi_k: X \rightarrow R^L$ 的形式化定义是: $\Phi_k(X) = (\phi_\alpha(X))_{\alpha \in \Sigma^k}$, 其中 $\phi_\alpha(X)$ 表示出现在 X 中的 α 的频度计数.

如图 2 所示, 在我们的输入数据中, miRNA 前体的上游序列、miRNA 前体的初级序列和 miRNA 前体的下游序列都有相同的字母表 $\Sigma = \{A, C, G, U\}$. 给定 $k=5$, 每个序列经编码形成的特征向量包含 $\sum_{k=1}^5 4^k = 1\,364$ 元素. miRNA 前体的二级结构编码序列的字母表是 $\{1, 2, 3, 4, 5, 6\}$. 这里忽略失配码 “0”. 取 $k=4$, miRNA 前体的二级结构编码序列经映射所形成的向量有 $\sum_{k=1}^4 6^k = 1\,554$ 元素. 因此, 一个输入序列将被映射到一个元素个数达 $3 \times 1\,364 + 1\,554 = 5\,646$ 的特征向量. 我们使用后缀树来加快特征生成. 后缀树中每个深度为 k 的节点都保存一个它所能到达的叶节点的数目.

即便对于较小的 k , 特征的维数也是非常大的. 大部分学习算法都会遇到 “维度灾难”. 因此, 特

征选择对于高维数据的学习是必需的, 我们采用了一个以信息熵为基础的基于相关性的特征选择方法 CBFS^[27].

2.3 识别算法

2.3.1 支持向量机. 支持向量机(SVM)是由 Vapnik 首先提出的一类有指导学习算法. 给定一组带标签的训练向量(正例或反例输入样本), SVM 学习一个线性的决策边界用于判定二个类别, 得到一个能用来分类新测试样本的线性分类规则. SVMs 已经在实践中表现出出色的推广性能, 并且有坚实的统计学习理论基础. 在我们的应用中, 我们将先前经特征选择得到的训练数据作为 SVM 的输入数据进行学习, 以获得一个能区分候选序列是否 miRNA 前体的模型. 模型中采用 RBF 核.

2.3.2 隐马尔可夫模型(HMM). 隐马尔可夫模型在编码 RNA 基因的识别中的应用是相当成功的. 自然人们会问, 它是否也可应用到 miRNA 前体的识别中? 在文献[21]中提出了采用隐马尔可夫模型的方法对人类基因组进行检测, 取得了一定的效果. 其基本思想是把 miRNA 前体的二级结构视为序列比对过程, 把各碱基配对状态根据序列比对的情形分为 4 类: a. 匹配(M), b. 失配(N), c. 删除(D), d. 插入(I). 另外根据这些状态是否包含了成熟 miRNA 碱基又可细分为 ± 2 个状态, 共 8 个状态. 如图 3b 给出了 miRNA 前体的隐马尔可夫训练模型(注: 这里给出的模型是对原文^[21]模型的细化).

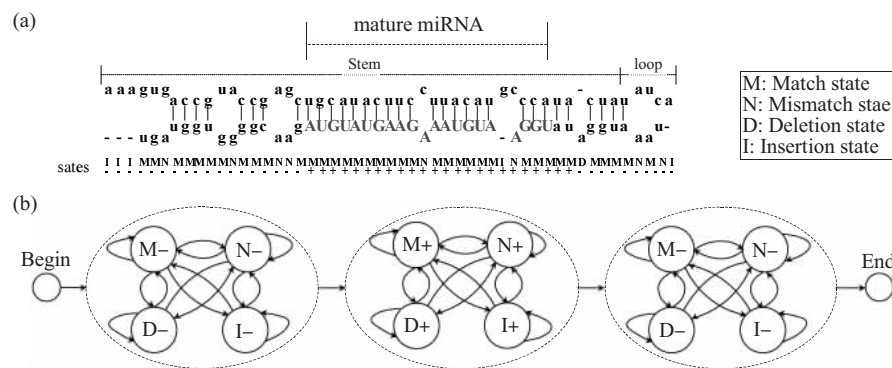


Fig. 3 A hidden markov model for identifying MiRNA precursor

(a)The observation sequence and its corresponding state sequence formed by the folded secondary structure. (b)HMM training model of miRNA precursors.

3 miRNA 前体的分类

在这一节中, 首先描述如何为训练和测试产生所需的数据集, 然后对这些数据进行特征映射和特

征选取, 并利用支持向量机进行训练获得学习模型, 用此模型进行预测. 同时, 我们把预测结果和隐马尔可夫模型的预测结果进行了比较.

3.1 数据的生成

实验中所用的 miRNA 基因和前体(版本 6)来自 miRNA 数据库 (microRNA registry)^[9]. 线虫 (*Caenorhabditis elegans*, 简称为 cel) 和其近亲线虫 (*Caenorhabditis briggsae*, 简称为 cbr) 的基因组序列来自 WormBase^{1(a)}; 所用的果蝇基因组序列 (*Drosophila melanogaster*, 简称为 dme 和 *Drosophila pseudoobscura*, 简称为 dps) 来自 FlyBase (版本 4.1)^{1(b)}; 其余基因组序列及其相应的标注文件 (基因特征文件 GFF) (人类: Homo sapien-hsa, 猿类: Pan troglodytes-ptr, 鼠类: Mus musculus-mmu 和 Rattus norvegicus-rno, 鸡 Gallus gallus-gga) 由 Ensembl^{1(c)} 网站获得 (1 (a)ftp://ftp.wormbase.org; (b) ftp://flybase.net/genomes; (c)http://www.ensembl.org/Download/).

3.1.1 正例样本生成. 根据我们在 2.1 节提出的输入序列的概念视图, 我们需要确定 miRNA 前体的上下游序列. 为此, 给定一个 miRNA 前体, 需要利用其相应基因组的标注文件来确定. 动物 miRNAs 通常是聚集在一起的. 我们称同一 DNA 链上的两个 miRNAs “相邻”, 若一个 miRNA 结束和另一个 miRNA 开始之间的核苷酸数目少于 1 000 nt. 若有多个 miRNAs $mir_1, mir_2, \dots, mir_k$ 满足 $(mir_{r+1}.start - mir_r.end) < 1\ 000\ nt$ ($r = 1, \dots, k-1$), 则称它们形成一个 miRNA 簇. 产生正例样本的步骤如下: a. 对于同一 DNA 序列, 合并相邻的 miRNAs 形成簇. b. 根据 GFF 标注信息, 对于位于编码基因序列 CDS 内含子中的 miRNAs, 获取 CDS 的 $\pm 1\ 000\ nt$ 上下游序列. 对于那些独立转录的 miRNA, 抽取该 miRNA 或 miRNA 簇 $\pm 1\ 000\ nt$ 上下游序列. 连同 miRNA 前体一起形成一个输入序列.

3.1.2 反例样本生成. 在生物信息学中获得反例样本是一个困难的问题. 但我们知道未标注序列中只有非常少的一部分对应于 miRNAs 序列, 这样可以从基因间隔区内生成 miRNA 基因的反例样本进行训练. 据所有已知的 miRNA 前体都形成茎环二级结构, 而且茎的每个臂都可能包含 miRNA, 因此, 我们也要求这些反例样本尽可能和真实 miRNA 前体相似. 特别地, 当产生反例样本时, 两个条件是必须满足的: 第一, 它们形成茎环结构. 这可通过 RNAfold 对候选序列折叠来进行选取. 其次, 茎的匹配部分至少是 15 nt 长 (目前 miRNA 前体茎环二级结构中茎的匹配部份最短的是 17 nt). 生成反例样本的过程如下: a. 经由 GFF 标注文件, 根据相

同的链上 CDS(start, end)位置对 CDS 进行排序, 形成基因间隔区. b. 对于每个基因间隔区, 我们向前滑动序列并使用正态分布 $N(\mu, \sigma)$ 模拟前体的长度, 这里的 μ, σ 都是从已知物种的 miRNA 前体中估计出来的. 例如, 对于线虫 *C. elegans* 来说 $\mu = 98, \sigma = 6.3$. 为避免两个产生的反例样本有过多重叠, 我们以正态分布 $N(50, 20)$ 选择下一个样本的起始位置.

3.2 实验结果

这里我们进行两种识别方法的实验, 即 SVM 和 HMM. SVM 实验采用 LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>) 在训练集上进行训练, 使用 RBF 核. 我们实现了如图 3 所示的 HMM 模型, 并进行了实验. 实验中, 分别用 3 个物种的数据集对 SVM 和 HMM 进行训练. 所用的 3 个物种分别是: 果蝇(dme), 线虫(cbr)和老鼠(mmu). 将获得的训练模型用于预测其他物种的 miRNA 前体. 为了使模型有较好概括性, 在训练阶段, 我们只使用一个物种的正例样本来学习, 而用另一个物种的所有正例样本作为测试集.

3.2.1 SVM 实验结果.

SVM 训练方法如下. 关于特征选取, 由于内存的限制, 无法在训练集中包含太多的反例样本来进行特征选择. 在实验中, 我们使用了 4 000 个反例样本用于特征选择. 由于不同的反例样本组合会导致不同的特征选取集, 因此, 为了有较好的测试性能, 我们对不同的组合进行选择. 举例来说, 一个训练集可以由 220 mmu 正例和 4 000 mmu 反例样本组成; 另外一个训练集可以由 220 mmu 正例样本、2 000 mmu 反例样本和 2 000 hsa 反例样本组成. 表 1 中列出了用于特征选择的正反例样本的组合以及经特征选择算法 CBFS 选择后的特征数. 所选取特征集合的一个突出特性是它们主要来自茎环结构的序列以及编码二级结构区域; 对于上下游序列区域, dme、cbr 和 mmu 训练集所选取的特征互不相同, 因此, 这些特征对前体预测精度贡献是否具有普遍的生物学意义还有待进一步研究.

Table 1 Characteristics of the data sets for feature selection

Species	Positives	Negatives	CBFS selected features
mmu	220	4000 mmu	72
dme	78	4000 dme	55
cbr	82	2000cbr+2000cel	55

这里给出果蝇训练实验的说明, 其余类似理解. 用果蝇(dme)的所有已知的正例样本 (78 个)和 4000 个反例样本, 通过 CBFS 方法执行特征选取. 在 55 个所选特征中, 我们经过实验选择前 39 个特征来训练 SVM 模型. 训练时我们采用正反例个数相对平衡的方法, 由于反例数目多, 训练集(果蝇全部正例+部分反例)可有许多不同的组合. 在所得模型之中, 选择一个在 ROC 曲线下有较大面积 (AUC)的模型. 一般而言, 可以得到具有相等 AUC 的许多模型. 在果蝇训练实验, 我们选取了由 {78 个正例样本+120 dme 反例样本+150 dps 反例样本} 训练所得模型, 它具有 86%的敏感性和 93%的特异性, 其 ROC 如图 4 所示. 所得 SVM 模型的最优化参数 $\gamma=0.5$, $C=2$. 表 2 中给出了用一个物种测试

其相关物种的 SVM 实验结果, 第一列的含义是, 所用训练物种(反例样本组合, 所用特征数目, C 值, γ 值), 其余列含义是显见的.

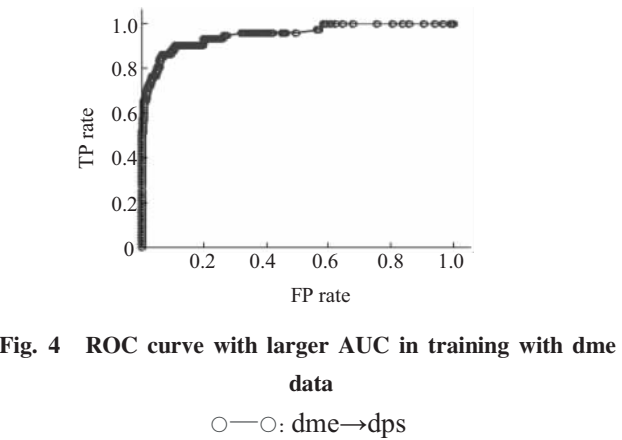


Fig. 4 ROC curve with larger AUC in training with dme data
○—○: dme→dps

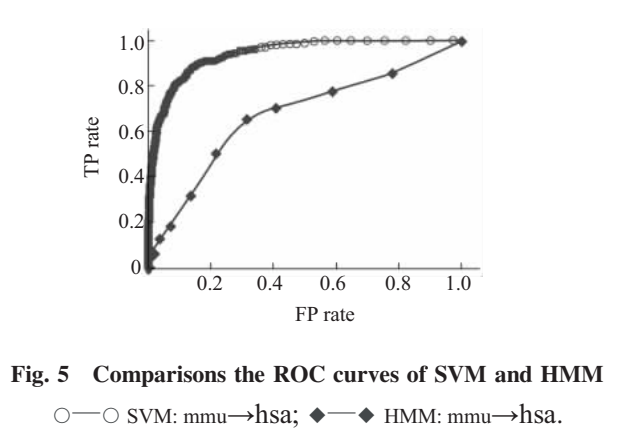
Table 2 Experimental results on SVM

Species in training	Test species	Sensitivity (TP/(TP+FN)%)	Specificity (TN/(TN+FP)%)
dme(120dme150dps, 39, 2, 2 ⁻¹)	dps	62/(62+10)= 86%	39666/(39666+2996)= 93%
cbr(80cbr0cel, 44, 32, 2 ⁻⁹)	cel	88/(88+27)= 76.52%	76661/(76661+3418)= 95.73%
mmu(600mmu150hsa, 62, 8, 2 ⁻³)	rno	172/(172+13)= 92.97%	77370/(77370+4842)= 94.11%
mmu(0mmu350hsa, 62, 512, 2 ⁻⁷)	hsa	258/(258+63)= 80.37%	69792/(69792+6518)= 91.46%
mmu(600mmu450hsa, 62,32, 2 ⁻³)	gga	110/(110+12)= 90.16%	75069/(75069+4338)= 94.54%
mmu(600mmu450hsa, 62,32, 2 ⁻³)	ptr	57/(57+10)= 85.08%	75203/(75203+3451)= 95.61%

3.2.2 HMM 实验结果. 在隐马尔可夫模型的训练中, 我们采用 10 fold 交叉验证方法来选取概率阈值. 实验中, 我们利用 mmu 正例样本进行训练. 对于正例样本的测试较简单, 由于已知成熟 miRNA 位置, 其 miRNA 前体的状态序列很容易根据 HMM 模型进行计算. 对于反例, 测试过程如下, 对于给定的一个候选 miRNA 前体经 RNAFold 折叠后形成如图 3a 的状态序列, 然后利用滑动窗口 (窗口大小由 19 nt 增加到 25 nt) 滑过整个编码输入序列, 窗口滑过所在位置的状态全部为 ‘+’ 状态, 其余皆为 ‘-’ 状态. 然后计算各种情形的概率, 取最大者和在训练 HMM 时设定的阈值进行比较. 大于阈值, 则判此输入序列为 miRNA 前体. 实验结果见表 3.

Table 3 Experimental results on HMM			
Species in training	Test species	Sensitivity	Specificity
mmu	hsa	35%	85%
	Rno	15%	95%
	Ptr	36%	88%
	Gga	23%	93%

3.2.3 SVM 和 HMM 实验结果比较. 为了比较 SVM 实验结果和 HMM 实验结果的优劣, 我们同时利用老鼠 mmu 数据来对 SVM 和 HMM 模型进行训练并预测人类 hsa 的 microRNA. 它们的 ROC 曲线见图 5. 由图可以得知, SVM 的测试结果明显优于 HMM 的测试结果. 这是因为基于 SVM 的方法, 通过初级序列和序列二级结构相结合, 经过特征映射和选取, 可以捕捉到序列数据中可能存在的微弱 “信号”. 而 HMM 能捕捉的信息相对较少, 从而其



预测效果不及 SVM. 对其他物种的 miRNA 预测也有类似的结果.

4 讨论和结束语

在机器学习的训练数据集中, 由于负样本相对比较充足, 我们对正负样本比例对预测精度的影响也进行了相应的实验. 对于 220 个 mmu 正样本, 我们让 hsa 的负样本数由 50 增加到 600, 观察敏感性和特异性的变化, 结果如图 6 所示. 总的来说, 正负样本比例对预测精度的影响不是十分明显, 而在正负样本比例相当时效果会较好. 当同时使用两个物种的负样本时, 情形有所变化但类似.

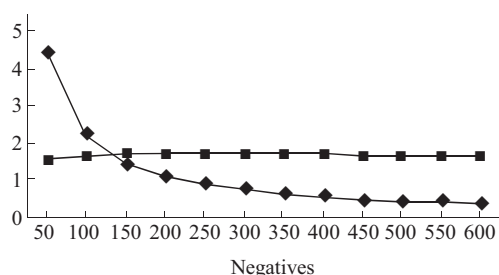


Fig. 6 The effect of pos/neg examples on prediction

◆◆: Positives/Negatives, ■■: Sensitivity+specificity.

在这篇论文中, 我们提出了一个全新的 miRNA 前体的识别方法. 在构造识别模型中, 我们把序列一级结构和二级结构相结合, 采用 k -gram 方法把序列信息映射到高维特征空间中, 然后通过特征选取方法提取特征, 并用这些特征为 miRNA 前体构造了基于 SVM 的识别模型. 同时, 我们和隐马尔可夫模型(HMM)进行了比较. 实验结果表明, 我们所提出的方法是有效的, 基于 SVM 的方法比基于 HMM 方法更有效, 可以达到较高的敏感性和特异性. 其原因是 miRNA 前体较短, 基于 HMM 的方法能捕捉的信息很少. 而基于 SVM 的方法, 通过序列一级结构和二级结构相结合, 经过特征映射和选取, 可以捕捉到序列数据中可能存在的微弱“信号”. 我们在提高 miRNA 前体识别率方面还将做进一步的深入研究.

参考文献

- Zaug A, Cech T. The intervening sequence RNA of Tetrahymena is an enzyme. Science, 1986, **231** (4737): 470~475
- Eddy S R. Non-codingRNA genes and the modern RNA world. Nat Rev Genet, 2001, **2** (12): 919~929
- Storz G. An expanding universe of noncoding RNAs. Science, 2002,

- 296 (5571): 1260~1263
- Mattick J S. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays, 2003, **25** (10): 930~939
- Bartel D P. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell, 2004, **116** (2): 281~297
- Abrahante J E, Daul A L, Li M, et al. The *Caenorhabditis elegans* hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. Dev Cell, 2003, **4** (5): 625~637
- Lai E C, Tomancak P, Williams R W, et al. Computational identification of *Drosophila* microRNA genes. Genome Biol, 2003, **4** (7): R42
- Lim L P, Lau N C, Weinstein E G, et al. The microRNAs of *Caenorhabditis elegans*. Genes Dev, 2003, **17** (8): 991~1008
- Griffiths-Jones S. The microRNA registry. Nucleic Acids Res, 2004, **32** (Database issue): D109~D111
- Grad Y, Aach J, Hayes G D, et al. Computational and experimental identification of *C. elegans* microRNAs. Mol Cell, 2003, **11** (5): 1253~1263
- Legendre M, Lambert A, Gautheret D. Profile-based detection of microRNA precursors in animal genomes. Bioinformatics, 2005, **21** (7): 841~845
- Pasquinelli A E, Reinhart B J, Slack F, et al. Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide let-7 heterochronic regulatory RNA. Nature, 2000, **408** (6808): 86~89
- Lagos-Quintana M, Rauhut R, Meyer J, et al. New microRNAs from mouse and human. RNA, 2003, **9** (2): 175~179
- Lau N C, Lim L P, Weinstein E G, et al. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science, 2001, **294** (5543): 858~862
- Lee R C, Ambros V. An extensive class of small RNAs in *Caenorhabditis elegans*. Science, 2001, **294** (5543): 862~864
- Ohler U, Yekta S, Lim L P, et al. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. RNA, 2004, **10** (9): 1309~1322
- Aravin A A, Lagos-Quintana M, Yalcin A, et al. The small RNA profile during *Drosophila melanogaster* development. Dev Cell, 2003, **5** (2): 337~350
- Seitz H, Youngson N, Lin S P, et al. Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nature Genetics, 2003, **34** (3): 261~262
- Ambros V, Lee R C, Lavanway A, et al. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr Biol, 2003, **13** (10): 807~818
- Wang X J, Reyes J L, Chua N H, et al. Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. Genome Biol, 2004, **5** (9): R65
- Nam J W, Shin K R, Han J, et al. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Research, 2005, **33** (11): 3570~3581
- Hofacker I L, Fontana W, Stadler P F, et al. Fast folding and

- comparison of RNA secondary structures. Monatshefte für Chemie, 1994, **125** (2): 167~188
- 23 Barciszewski J, Erdmann V A. Noncoding RNAs: Molecular Biology and Molecular Medicine. New York: Kluwer Academic, 2004. 33~48
- 24 Smalheiser N R. EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues. Genome Biol, 2003, **4** (7): 403
- 25 Lee Y, Kim M, Han J, *et al.* MicroRNA genes are transcribed by RNA polymerase II. EMBO J, 2004, **23** (20): 4051~4060
- 26 Liu H, Wong L. Data mining tools for biological sequences. J Bioinform. Comput Biol, 2003, **1** (1): 139~167
- 27 Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. The Twentieth International Conference on Machine Learning, Washington D C. 2003. 856~863

A *k*-gram Approach for Identifying MicroRNA Precursors*

YANG Liang-Huai^{1,2}, LÜ Pi-Ming³, CHEN Li-Jun², DENG Ming-Hua^{4,5}**

⁽¹⁾College of Information Engineering, Zhejiang University of Technology, Hangzhou 310032, China;

⁽²⁾School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China;

⁽³⁾Graduate School of The Chinese Academy of Sciences, Beijing 100049, China;

⁽⁴⁾School of Mathematics Sciences, Peking University, Beijing 100871, China;

⁽⁵⁾Center for Theoretical Biology, Peking University, Beijing 100871, China)

Abstract MicroRNAs(miRNAs) are short non-coding RNAs that play important regulatory roles in both animals and plants. While the first miRNAs were discovered using experimental methods, experimental miRNA identification remains technically challenging and incomplete. Hence, computational approaches are a natural choice to complement experimental approaches to miRNA gene identification. A *de novo* miRNA precursor prediction method was proposed. In constructing the recognition model, both primary sequence and secondary structure were combined into an input sequence through encoding, and the input space was mapped into a feature space *via k*-gram method. After applying feature selection, those selected features were used to construct SVM-based models for the recognition of miRNA precursors. In the mean time, the method was compared with the HMM learning method. Experimental results show that the method outperforms HMM. The reason is that microRNAs are so short that it is not easy for HMM model to capture the signals for differentiating the genuine microRNAs from those pseudo-microRNA genes. From features selected, it was found that they are mostly come from the primary and secondary structure of microRNAs. This phenomenon may tell us to put more efforts in the microRNAs themselves in designing computational method before we fully understand the transcription mechanism of microRNA biologically.

Key words microRNA, gene identification, support vector machine, hidden Markov model, microRNA precursor

*This work was supported by grants from The National Natural Science Foundation of China (30570425) and National Basic Research Program of China (2003CB715903).

**Corresponding author. Tel: 86-10-62767562, Fax: 86-10-62751801, E-mail: dengmh@math.pku.edu.cn

Received: June 20, 2006 Accepted: August 24, 2006