microRNA计算识别中的模式识别技术

孙秋凤
(南京师范大学 泰州学院 信息与科学技术系,江苏 泰州 225300)

摘 要:MicroRNAs (miRNAs)是一种大小约21~23个碱基的单链RNA小分子,对多种生物学过程起调控作用,它们主要参与基因转录后水平的调控,能有效地抑制相关蛋白质的合成,与生物体的生长发育和某些疾病的发生密切相关。对microRNAs (miRNAs)的研究正在不断增加,计算识别为分子生物学实验寻找新microRNA提供一组高质量的候选序列。文中从模式识别的角度审视现有的计算识别技术,分析和比较各种方法的特点后发现基于支持向量机的识别方法已经能在识别精度上得到很好的效果,这也是microRNA识别技术将来发展的主要方向。

关键词:microRNA;支持向量机;生物信息学;核函数
中图分类号:TP391.4;Q811.4 **文献标识码**:A **文章编号**:1673-629X(2010)06-0097-04

Pattern Recognition Technology for MicroRNA Identification

SUN Qiu-feng
(Department of Information and Technology, Taizhou College,
Nanjing Normal University, Taizhou 225300, China)

Abstract: MicroRNAs(miRNAs) are ~22nt long non-coding RNAs that are derived from larger hairpin RNA precursors and play important regulatory roles in both animals and plants. The research of miRNAs is continually increasing after the first miRNAs were discovered using experimental methods. Since experimental miRNA identification remains technically challenging and incomplete, this calls for the development for computational approaches to complement experimental approaches to miRNA gene identification. Attempts to look back the existing computational approaches and compare their advantages. Finally find that among all the potential means, the one based on SVM has better precision and that's why this method has become the leading measure for microRNA identification in the future.

Key words: microRNA; support vector machine; bioinformatics; kernel function

0 引 言

miRNA是一些长度约为22nt的非编码调控RNA家族,它有3个显著的特点:

(1) 广泛存在于真核生物中,是一组不编码蛋白质的短序列RNA,本身并不具有开放阅读框;

(2) 通常的长度为20~24nt,但在3'端可以有1~2个碱基的长度变化;

(3) 成熟的miRNA 5'端有一磷酸基团,3'端为羟基。

在线虫、果蝇、小鼠和人等物种中已经发现的数百个miRNAs中的多数具有和其他参与调控基因表达的分子一样的特征,提示miRNAs在高级真核生物体内对基因表达的调控作用可能和转录因子一样重要。

早期传统寻找miRNA的方法主要依赖于分子克隆,此类方法一般都步骤烦琐、周期性长、工作量大,由于目标不明确,效率较低且实验费用昂贵。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,基于机器学习的计算方法已经成为发现新的microRNA的一个重要手段,为实验发现提供候选microRNA基因。

1 miRNA计算识别方法

1.1 基于决策树的计算识别方法

这类识别方法中较为成功软件有:MiRscan[1]和miRseeker[2]。这类方法大致的流程是从某一物种已知miRNA中提取相关特征,建立模型,从大量的数据集中筛选出候选miRNA,然后对其进行打分,若超过某个阈值则认为此序列可归于这一类。其中提取特征并建立模型的操作从模式识别角度来看类似于建立一棵决策树的过程(见图1)。

在文献[3]中考虑到在已识别的miRNA周围可能

收稿日期:2009-09-30;**修回日期**:2009-12-22
基金项目:国家自然科学基金(60275007)
作者简介:孙秋凤(1979-),女,江苏泰州人,硕士,研究方向为模式识别及生物信息技术。

存在新的 miRNA,将待识别的序列长度增加,然后提取其二级结构,使用 SVM 对其进行分类。这个方法的优点是使用了输入序列和二级结构,缺点在于忽略了相关生物信息,可能导致在对哺乳类动物基因进行测试时产生高的假阳性。

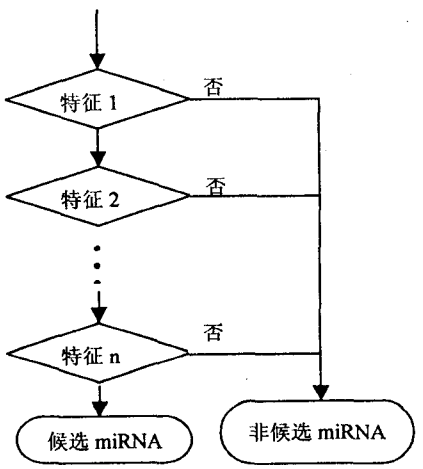

图 1 一个二叉决策树的示例

文献[4]中的方法是基于物种间的保守性而设计的,在特征方面除了使用了众所周知的前体二级结构外,还提出了三个可计算的特征:(1) GC 含量为 38%～70%;(2) 茎环长度在 20～70nt 之间;(3) 与物种 O. sativa 的序列相似性不低于 90%。

文献[5]中的 miRAlign 方法在序列信息的基础上加入结构信息来对候选序列打分。在处理待识别序列时,除了使用自由能等常规生物属性来预测序列二级结构外,miRAlign 增加了一个在茎环结构上检测 miRNA 位置的操作,通过比较两个 miRNAs 位置上的差别确认二者是否为同源体。

可以看出,以上方法的策略大致可以分为如下三类:

1)利用同源性搜索已知 miRNA 基因的直系同源(ortholog)和旁系同源(paralog)。

2)在已知 miRNA 附近搜索基因簇。

3)其他不依赖于同源性和 miRNA 基因簇的基因搜索法。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 序列打分。

1.2 基于支持向量机的计算识别方法

1.2.1 支持向量机简介

支持向量机(Support Vector Machine, SVM)是基于统计学习理论的学习方法。它通过构造最优超平面,使得对未知样本的分类误差最小。对于两类线性可分情形,可直接构造最优超平面,使得样本集中的所有向量满足如下条件:

(1) 能被某一超平面正确划分;

(2) 距该超平面最近的异类向量与超平面之间的距离最大,即分类间隔最大。则该超平面为最优超平面。

其中,条件(1)是保证经验风险最小,条件(2)是使 VC 置信度最小,从而使期望风险最小。

这里,最优超平面的构造问题实质上是约束条件下求解一个二次规划问题,以得到一个最优分类函数为:

$$f(x) = \mathrm{sgn}\{\sum_{i=1}^{L} y_i a_i k(x_i, x) + b\}$$

其中 $k(\cdot,\cdot)$ 是一核函数,sgn 是符号函数,L 为训练样本数目。

在该分类函数中,某些 x_i 对应的 a_i 不为零,由于这些具有非零值 a_i 的对应的向量支撑了最优分类面,因此被称为支持向量。

对于线性 SVM,核函数 $k(\cdot,\cdot)$ 就是两向量的点积运算;对于非线性情形,可通过非线性映射把输入向量映射到一个高维特征空间,来构造最优分类面,常用的核函数形式有多项式形式、径向基形式、二层感知器神经网络形式。

1.2.2 一些将序列转化成向量的方法

在基于决策树的方法中已经有部分方法中使用 SVM 来对结果进行分类,但使用仅限于将其作为一个分类工具,并未有任何改进之处。随着研究 miRNA 的人员越来越多,方法也开始呈现多样化,对于一些边缘学科而言,如生物信息学,人们希望将 miRNA 不仅仅看成纯生物的序列,而是将其看成大多数人能接受的某种结构,且这种结构能有利于特征的提取。

文献[6]中将发卡序列中碱基表示成相应的三联组,然后统计 32 个可能的三联组出现的次数,经过归一化后作为 SVM 的输入向量进行训练得到决策面,进而对测试数据进行分类。这个方法的优点在于不需要考虑相关的生物特征,而且测试结果也显示虽然没有考虑那些因素,但所映射后的特征还是反映了 miRNA 二级结构的有关性质。

文献[7]提出了一个依据遵循"特征生成、选择、综合"构造识别模式、基于 SVM 的 de novo 方法来识别前体。为了捕获二级结构中的信息,先通过 RNAfold 将前体折叠,为了便于数据处理,再将碱基对进行编码。

文献[8]的作者认为使用 SVM 识别 miRNA 的方法虽多,但那些方法都没有将一些特征很好地综合起来,因此提出 PSoL(Positive Sample Only Learning Algorithm)算法。

算法的关键在于训练数据集的选择,算法的基本

思想是：

(1) 根据序列统计性，最小自由能及相关基因间的相似性度量将每个序列转换成特征向量。

其中序列统计性包括(A, C, G, T)，dimer(AA, AC...TT)及 trimer(AAA, AAC...TTT)的个数；相似性度量包括 Typhi_CT18, Typhi_Ty2 及 Typhi_LT2。通常特征向量维数过多会降低识别能力，因此对转换后得到的 88 个向量进行筛选。

(2) 将已有的数据分为正类样本及 unlabeled data(即其中既有正类也有负类)。

PSoL 的目的是在 unlabeled data 中预测正类样本，但问题是现有的训练集中没有负类样本，如何产生负类样本是本算法的核心之一。

由于序列的保守性通常在二级结构，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，在特征提取/生成时，与基于决策树类方法将特征的提取重点放在序列信息不同的是，这类方法的特征提取侧重于结构信息。

1.2.3 一些使用特殊核函数的方法

尽管 SVM 在生物学应用比较成功，但通常都要涉及到将结构化的生物数据转化成特征向量。这导致即使一个复杂的结构也会被转化成简化的数值，这会损失一些生物信息。为了避免这种信息的损失，基于链以及图的一些核函数被应用到支持向量机中。

文献[9]提出一种基于链的 Spectrum kernel，所使用的特征是长度为 k 的氨基酸的所有可能子序列的集合(见图 2)，若两个蛋白序列含有许多相同的 k 长度的子序列，则 k－spectrum kernel 下的内积就会越大，即两个序列的相似度也会越大。

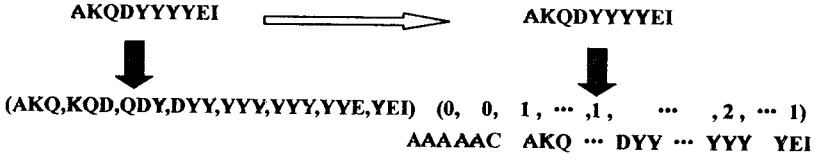

图 2　k－spectrum kernel 特征映射示例

文献[10]是在 spectrum kernel 基础上提出一种新的用于 SVM 的核－mismatch kernel。mismatch kernel 的计算是基于(k, m)－patterns 出现的次数，而(k, m)－patterns 是由所有与一个 k 长度子串最多有 m 个不匹配的 k 长度子串的集合生成的，特征映射见图 3。

文献[11]则在 spectrum kernel 和 mismatch kernel 基础上提出一种 generalized string kernel(GSK)。GSK 是所有(ki, mi)－mismatch kernel 的总和：

$$K_{(k_1,m_1,\cdots,k_s,m_s)}(x,y)=\sum_i\langle\phi_{ki,mi}(x),\phi_{ki,mi}(y)\rangle=\sum_i K_{(ki,mi)}(x,y)$$

GSK 的基本思想是将序列映射(见图 4)成 1－GSK, 2－GSK 和 3－GSK，通过训练找出权值较高的映射子串并把它们作为特征输入 SVM 对数据进行测试。

Sequence : x = ABBA　$|\Sigma|$ = {A,B,C}　2-mers in x: AB, BB, BA

$\Phi^{Mismatch}_{(2,1)}(\alpha)$		α: AA	AB	AC	BA	BB	BC	CA	CB	CC	$\Phi^{Mismatch}_{(2,1)}(x)$ x = ABBA
β	AA	1	1	1	1			1			2
	AB	1	1	1		1			1		2
	AC	1	1	1			1			1	1
	BA	1			1	1	1	1			2
	BB		1		1	1	1		1		3
	BC			1	1	1	1			1	2
	CA	1			1			1	1	1	1
	CB		1			1		1	1	1	2
	CC			1			1	1	1	1	0

图 3　(k, m)－mismatch kernel 特征映射示例

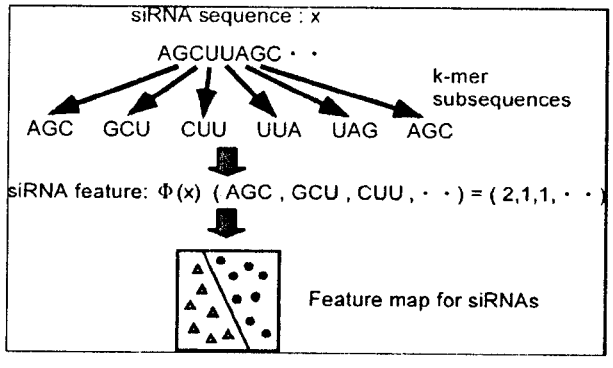

图 4　GSK 映射示例

文献[12]又提出另一种类似的核－stem kernel。与上述核的不同之处在于参与计算的子串不再是连续的，而是允许不连续的子串参与映射向量。

虽然上述的识别方法大多使用二级结构作为输入数据，但在一定程度上忽略了其本身含有的拓扑信息，文献[13]提出将二级结构表示成图结构，这样可以直接比较两个图之间的相似性，避免训练数据、调整参数等。Labeled Dual Graphs(LDG)即是用来表示二级结构的图，两个 LDGs 之间的相似性则是用 Marginalized Kernels 来进行计算。

2　结束语

目前，早期 miRNA 计算识别方法面临较低准确率

的问题，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。且此类方法的一大缺陷是没有一个系统的方法把候选 miRNA 序列的初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：MiRscan 方法中，候选 miRNA 前体经 RNAFold 折叠后的最小折叠自由能量阈值下限设为 25(即 $\Delta G_{folding} \leqslant -25$ kcal/mole)，但经过对已知的 miRNA 前体进行折叠后发现，许多最小折叠自由能量小于该阈值，如线虫 miRNA 前体 cel-mir-261 经折叠后其 $\Delta G_{folding} = -7.08$ kcal/mole。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。

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。

目前研究的热点在于如何将序列和结构信息综合起来，以达到更好的分类效果。已经提出的基于 string kernel 和 graph kernel 的算法是个很好的尝试。

参考文献：

[1] Lim L P, Lau N C, Weinstein E G, et al. The microRNAs of Caenorhabditis elegans[J]. Genes Dev., 2003, 17: 991-1008.

[2] Lai E C, Tomancak P, Williams R W, et al. Computational identification of Drosophila microRNA genes[J]. Genome Biol, 2003(4): 1-20.

[3] Sewer A, Paul M, Landgraf P, et al. Identification of clustered microRNAs using an ab initio prediction method[J]. Bioinformatics, 2005(6): 267-281.

[4] Wang X J, Reyes J L, Chua Nam-Hai, et al. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets[J]. Genome Biol., 2004(5): 1-15.

[5] Wang X W, Zhang J, Li F, et al. MicroRNA Identification Based on Sequence and Structure Alignment[J]. Bioinformatics 2005, 21(18): 3610-3614.

[6] Xue Chenghai, Li Fei, He Tao, et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine[J]. Bioinformatics 2005 (6): 310-317.

[7] Yang Liang Huai, Hsu Wynne, Lee Mong Li, et al. Identification of MicroRNA Prediction via SVM[C]//Proceeding of the 4th Asia-Pacific Bioinformatics Conference. Taipei, Taiwan: [s.n.], 2006: 267-276.

[8] Kim Sung-Kyu, Nam Jin-Wu, Rhee Je-Keun, et al. miTarget: microRNA target gene prediction using a support vector machine[J]. Bioinformatics, 2006(7): 411-422.

[9] Leslie C S, Eskin E, Noble W S. The spectrum kernel: a string kernel for SVM protein classification[C]//Proc. Pac. Biocomput. Symp. [s.l.]: [s.n.], 2002: 1441-1448.

[10] Leslie C S, Eskin E, Cohen A, et al. Mismatch string kernels for discriminative protein classification[J]. Bioinformatics 2004, 20(4): 467-476.

[11] Teramoto R, Aoki M, Kimura T, et al. Prediction of siRNA functionality using generalized string kernel and support vector machine[J]. FEBS Lett, 2005, 579(13): 2878-2882.

[12] Yasubumi, Sakakibara. Kernel Functions for RNA sequence analyses[C]//2nd Taiwan-Japan Bilateral Symposium on Bioinformatics. [s.l.]: [s.n.], 2006.

[13] Karklin Y, Meraz R F, Holbrook S R. Classification of Non-Coding RNA Using Graph Representations of Secondary Structure[C]//Pacific Symposium on Biocomputing. [s.l.]: [s.n.], 2005.

(上接第 96 页)

参考文献：

[1] Pawlak Z. Rough Sets[J]. International Journal of Computer and Information Sciences, 1982, 11(5): 341-356.

[2] ZHANG Wen-xiu, MI Ju-sheng, WU Wei-zhi. Approaches to knowledge reductions in inconsistent systems[J]. International Journal of Intelligent Systems, 2003, 18(9): 989-1000.

[3] 赵荣利，崔志明，陈建明. 一种改进的基于差别矩阵的属性约简方法[J]. 计算机技术与发展，2006，16(11)：32-33.

[4] 汪小燕，杨思春. 一种基于分辨矩阵的新的属性约简算法[J]. 计算机技术与发展，2008，18(2)：77-78.

[5] 陈鑫影，邱占芝. 基于可分辨重要度的属性约简算法[J]. 大连交通大学学报，2008，29(4)：83-84.

[6] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法[M]. 北京：科学出版社，2001.

[7] 田卫东，周创德，胡学刚，等. 基于简化分辨矩阵的粗糙集属性约简算法[J]. 计算机科学，2008，35(3)：210-211.

[8] 葛 浩，杨传健，李龙澍. 一种改进的基于二进制可分辨矩阵属性约简算法[J]. 计算机技术与发展，2008，18(8)：13-14.

[9] HU X H, Cercone N. Learning in relational databases: A rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-337.

[10] 胡可云. 基于概念格核粗糙集的数据挖掘方法研究[D]. 北京：清华大学，2001.

[11] 李 侃，刘玉树. 一种粗糙集属性约简算法[J]. 计算机工程与应用，2002，38(5)：15-19.