

拟南芥基因组中新的 microRNA 预测及分析

金伟波^{1,2}, 孔 栋², 应晓敏¹, 郭蔼光², 李伍举¹

(1. 军事医学科学院基础医学研究所计算生物学中心, 北京 100850; 2. 西北农林科技大学生命科学院, 西安 712100)

摘要: MicroRNA (miRNA) 是一类存在于动植物体内、长度为 21~25 nt 的内源性小 RNA, 对生物体的转录后基因调控起着关键作用, 但一些低丰度的 miRNA 和组织特异性 miRNA 往往很难发现。为了系统识别拟南芥基因组中新的非同源 miRNA, 首先基于已报道的拟南芥 miRNA 的特征, 从全基因组范围中筛选出 453 条可能的 miRNA 前体; 其次, 为了进一步对上述 miRNA 前体进行筛选, 利用人的 miRNA 前体数据构建了支持向量机模型 GenomicSVM, 该模型对人测试集的敏感性和特异性分别为 86.3% 和 98.1% (30 个人 miRNA 前体和 1 000 个阴性 miRNA 前体), 对拟南芥测试集的正确率为 93.6% (78 个 miRNA 前体); 最后, 利用 GenomicSVM 预测上述 453 条 miRNA 前体序列, 得到了 37 条候选的新的拟南芥 miRNA 前体, 为进一步的 miRNA 实验发现研究提供了指导。

关键词: 拟南芥; 基因组; microRNA; 预测

中图分类号: Q74

0 引 言

MicroRNA (miRNA) 是一类存在于动植物体内的内源性的 ncRNA^[1-5], 在生物体内起调节 mRNA 稳定性及翻译作用。最早被发现的 miRNA lin-4 和 let-7 可通过与靶 mRNA 3' 末端形成碱基配对来抑制翻译, 后来很多新发现的 miRNA 也以相似的机制发挥作用^[6]。不过 miRNA 大家族里也有一些成员, 特别是植物 miRNA, 主要是通过 RNAi 途径, 以完全互补或接近完全互补的方式与靶 mRNA 结合, 从而达到降解靶基因目的^[7,8]。

miRNA 在生物体的不同部位和不同的发育阶段对基因的转录后调控都起重要的作用^[1-5], 因此发现各物种的 miRNA 进而揭示其功能具有重要意义。但通过实验手段只能使部分高丰度表达的 miRNA 得到有效克隆, 而大量低丰度表达的 miRNA 却难以得到, 因此, 利用计算生物学方法发现低丰度或组织特异性表达的 miRNA, 进而为实验提供帮助已成为一条切实可行的途径。到目前为止, 已发展了许多算法与软件^[9-15], 如基于比较基因组学方法的软件有 MiRscan^[9,10]、SRNALoop^[11]、miRseeker^[12]和 miRAlign^[13]等, 这些程序都需要通过序列保守性来预测 miRNA 前体, 因此利用它们很难有效地发现某个基因组中非同源的新的 miRNA 基因。另外, 最近机器学习方法也逐渐用于真假 miRNA 前体的区分问题, 如 Nam 等人^[14]于

2005 年开发的基于隐马氏模型 (HMM) 的 ProMIR 来预测人的 pre-miRNA, 正确率为 75%; Xue 等人^[15]利用人 miRNA 前体数据集中的部分数据作为训练数据, 然后采用 SVM 方法构建 miRNA 前体预测模型, 对阳性测试集中的 30 个 miRNA 前体, 预测精度为 93.3% (28/30), 对两个阴性测试集来说, 预测精度分别为 88.1% (881/1000) 和 89.0% (2175/2444); 此外, Loong 等人^[16]也利用 SVM 方法探讨了 miRNA 前体的预测问题, 在他们构建的测试集上, 其敏感性与特异性分别为 84.6% 与 98.0%; 这些预测模型主要是基于已知的数据集来探讨 miRNA 前体的预测问题, 还没有用于基因组水平的 miRNA 前体识别。

基于上述考虑, 我们采用下列策略探讨拟南芥基因组中新的非同源 miRNA 的识别问题, 首先对目前已知的 78 条拟南芥 miRNA 前体进行特征统计, 并以其为基础并结合比较基因组学方法, 构建了拟南芥基因组水平的新的非同源 miRNA 前体预测流程, 获得了 453 条 miRNA 前体; 然后利用人的 miRNA 前体数据集和支持向量机方法构建了

收稿日期: 2007-02-02

基金项目: 国家自然科学基金项目 (30470411, 30500105)

通讯作者: 李伍举, 电话 / 传真: (010)66931324, E-mail:

liwj@nic.bmi.ac.cn; 郭蔼光, 电话 / 传真: (029)87026171,

E-mail: guoai Guang@yahoo.com.cn

miRNA前体预测模型 GenomicSVM, 并基于此模型预测上述 453 条 miRNA 前体, 最终获得了 37 条候选的新的非同源的拟南芥 miRNA 前体, 为进一步的 miRNA 实验发现研究提供了指导, 也为其它物种基因组中新的非同源的 miRNA 识别提供借鉴。

1 材料与方法

1.1 在拟南芥基因组中筛选具有发夹结构的序列

拟南芥基因组从 NCBI (<http://www.ncbi.nlm.nih.gov/>) 库中下载; 根据已公布拟南芥 miRNA 的相关特征, 包括前体发夹结构及自由能 (根据 RNAfold 计算所得)、GC 含量、前体长度、茎区螺旋区长度和发夹环的长度等, 从拟南芥基因组中筛选 pre-miRNA-like 序列。

1.2 真假 pre-miRNA 数据集

人的 pre-miRNA 从 miRNA 数据库^[17]下载, 通过去除那些没有发夹结构的 pre-miRNA, 最后得到 193 条人 pre-miRNA 作为阳性数据集。目前认为, miRNA 基因主要位于基因组的非编码区, 因此可以认为, 对于一些编码蛋白的基因序列, 即使它们具有一些与真正 pre-miRNA 类似的特征, 仍认为是假的 pre-miRNA。基于此认识, 本研究构建了一个 ENCODE 阴性数据集, 它根据 UCSC 的 refGene 列表提取出人类编码基因的序列, 用 RNAfold 程序^[18]预测这些序列的二级结构, 提取所有与已报道人的 pre-miRNA 具有相似特征并具有发夹结构的序列, 最后构建的 ENCODE 包含了 7893 条序列。

1.3 构建 GenomicSVM 的训练集和检测集

为了用支持向量机 (SVM) 方法来判别真假 pre-miRNA, 我们构建了一个训练集 (TR) 和两个测试集 TE-human 和 TE-ath。其中 TR 包括 163 条随机从阳性数据集中抽取的真的 pre-miRNA (阳性样本) 和 1000 条随机从阴性数据集 ENCODE 中抽取的假 pre-miRNA (阴性样本) 的序列; TE-human 包括了剩余的 30 条真 pre-miRNA 和 1 000

条随机从阴性数据集 ENCODE 中抽取的假 pre-miRNA (与 TR 中的数据没有重复)。TE-ath 数据集中包含了从 miRNA 数据库^[17]下载得到的 78 条具有发夹结构的拟南芥 miRNA。

1.4 支持向量机模型及其评估

本研究采用支持向量机软件包 LibSVM^[19], 该软件包由台湾大学林智仁等人开发, 具有操作简单和易于使用等特点, 可以解决分类问题、回归问题以及分布估计等问题, 提供了四种常用核函数 (线性、多项式、径向基和 S 形函数) 供用户选择, 可以有效地解决多类问题、交叉验证选择参数和对不平衡样本加权等, 其中 Grid 方法用于寻找最优的罚分参数 C 和径向基函数 RBF 的核心参数 γ 。该软件包下载于 <http://www.csie.ntu.edu.tw/~cjlin/libsvm/oldfiles/>。

另外, 模型效果评价是评估一个模型在实际中能否应用的关键步骤。对于一个包含阳性样本与阴性样本的测试集来说, 预测结果包括以下四种类型: 正确预测的阳性样本数目 TP 与阴性样本数目 TN , 假阳性样本数目 FP 与假阴性样本数目 FN 。基于这些数值, 可以分别计算出模型的敏感性 (Se)、特异性 (Sp) 和分类精度 (ACC), 具体计算公式如下:

$$Se = TP / (TP + FN) \quad (1)$$

$$Sp = TN / (TN + FP) \quad (2)$$

$$ACC = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

2 结果与分析

2.1 拟南芥 miRNA 特征统计及基因组中 Hairpin-like 序列的搜索

通过对测试集 TE-ath 中的 78 条具有发夹结构的拟南芥 miRNA 前体序列进行分析与特征统计, 得到了如表 1 所示的结果。然后, 以表 1 得到的结果作为参数, 从拟南芥基因组中识别 pre-miRNA-like 序列, 具体流程见图 1, 去除冗余后, 最终得到 453 条可能的前体片段。

Table 1 The characteristics of known 78 pre-miRNAs in *A. thaliana*

Length(nt)	GC content	Stem length	Length of helix	Loop length	Free energy
>70	0.36~0.70	>32 nt	>20 bp	≥5 nt	<-9.6 kcal/mol

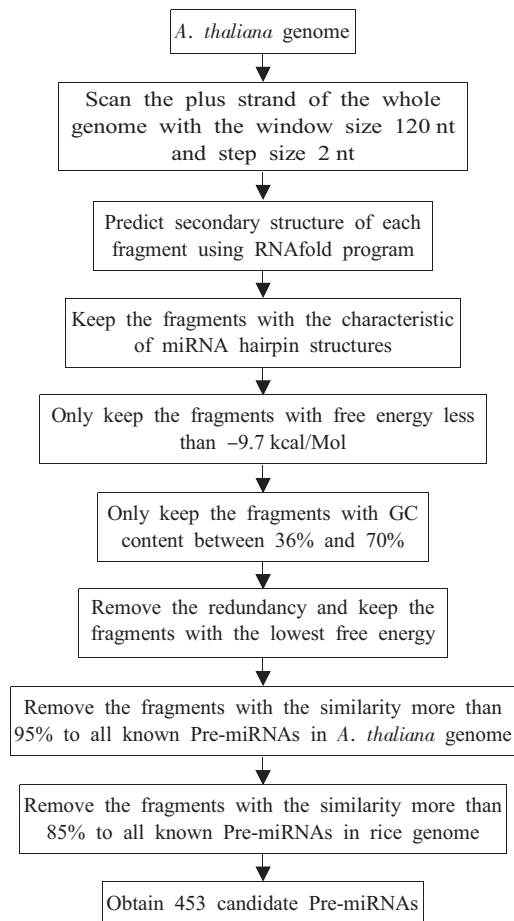


Fig.1 Flowchart for the prediction of pre-miRNA-like sequences in the *A. thaliana* using the characteristics of the known miRNAs and comparative genomics methods

Table 2 Performance of the GenomicSVM model on test sets TE-human and TE-ath

Test set	TP	TN	FP	FN	Se	Sp	ACC
TE-human	25	981	19	5	0.833	0.981	0.977
TE-ath	73	0	0	5	0.936	0.000	0.936

然后将 GenomicSVM 模型用于上述得到的 453 条 pre-miRNA-like 序列的识别, 最后得到了 37 条候选的 pre-miRNA, 其中 20 条位于基因间区, 占 54.1%; 9 条位于基因内含子区, 占 24.3%; 剩余的 8 条位于基因编码区, 占 21.6% (表 3)。但是根据文献^[20], miRNA 在编码区上不应有很高的比例, 我们推测位于编码区上的这些 miRNA 可能是一些反式作用小 RNA。由于本研究在发夹结构序列的筛选过程中去除了与目前已报道拟南芥和水稻 miRNA 的同源序列, 因此, 本研究最后得到的 37 条候选序列是新的拟南芥 miRNA 前体。

2.4 miRNA 的靶标预测

作为 miRNA, 其主要作用是调控靶基因的表

2.2 GenomicSVM 模型的训练和检测

为了进一步从大量 pre-miRNA-like 序列中识别出真正的 pre-miRNA, 我们开发了用于真假 miRNA 前体识别的支持向量机模型 GenomicSVM。GenomicSVM 是根据 pre-miRNA 的二级结构特征, 并利用 SVM 工具来预测 miRNA 前体的一个模型。该模型在训练时选用 *RBF* 作为核函数, 并用 Grid 策略及 10-fold 交叉验证的方法来搜索 *RBF* 的最优参数 γ , 然后用优化的核心参数 ($\gamma=0.125$) 训练 GenomicSVM, 并通过两个测试集来检验 GenomicSVM 的预测效果。如表 2 所示, TE-human 中的 30 条真的 pre-miRNA 有 25 条被正确识别, 得到模型的敏感性 (即从阳性数据集中被正确识别为阳性数据的条数) 为 83.3%; 另外的 1000 条阴性序列有 981 条被正确地识别为假的 pre-miRNA 序列, 得到模型的特异性 (即从阴性数据集中被正确识别为阴性数据的条数) 为 98.1%。GenomicSVM 及所有相关资料都可从网上免费下载得到, 网址为: <http://geneweb.go3.icpcn.com/genomicSVM/>。

2.3 拟南芥候选 pre-miRNA 的预测

为了检测 GenomicSVM 对拟南芥 pre-miRNA 的识别效率, 我们用 TE-ath 作为测试集来验证。TE-ath 中共有 78 条拟南芥 pre-miRNA, 其中 73 条得到了正确识别 (见表 2), 正确率为 93.6%。

达, 因此, 上节中预测的 37 条候选 miRNA 前体, 如果是真实的 miRNA, 在拟南芥中应该有对应的被调控的靶标, 为此, 我们开展了 miRNA 靶标的预测研究。

目前认为, 植物 miRNA 的作用方式主要通过靶标基因编码区以接近完全互补的方式发生作用 (near-perfect complementarity), 从而降解靶标 mRNA^[7,8,21], 为了进一步检测上述预测的拟南芥 miRNA 的功能, 通过 Blast 程序 (W: 16, S: 2; 表示搜索时 word size 为 16 nt, 仅搜索互补链) 将候选的 37 条成熟 miRNA (分别来自于相应 pre-miRNA 发夹结构的两个茎区) 与 EST 库作互补检索, 如表 3 所示, 发现这些 miRNA 的靶标

mRNA 从几十到几百不等, 然而对于 AthmiR010 却没有找到相应的靶标 EST 序列, 这存在两种可能, 一种是由于目前拟南芥的 EST 序列还不够完

整, 也就是说 AthmiR010 的靶标序列还没有发现; 另一种就是本研究预测的 AthmiR010 可能为假阳性片段。

Table 3 The list of predicted miRNAs in *A. thaliana*

miRNA No.	Chr.	ΔG (kcal/mol)	Tar ¹	Location ²	Sequence
AthmiR001	1	-17.99	33	Intron	UGAAUGAACAGCCUAGACAAAUAAGAAGGCCACAAGUUA AAUGAUUUCUCUAGUAAGAAAAGGAGAAAGAGGAAAUCAG ACCAACUUUGCCAAAAGAAAGAUGUAGCUAGACUCAUAA
AthmiR002	1	-10.22	15	Inter	UGAUAGACAACAGUUCUAAAGAAGGGAAGAUGAGAUUCAGG CAACAAAAACACAUAUAGUUCAAAAAAUAGGUUGGUAG GAUCAAUUAUUCAAAAACAUCACCAAAUAGAUUUUA
AthmiR003	1	-13.00	26	Inter	UAACUAAAGGAUCAAAUUCAAAGUUUUUCUUGUAAACAAAC AUACGUAAAAGGAGGUAAGCUUGGUACAUGAGAAAUAUUUU AAUAUUUAAGUUAGUGGUGAAGAAAAGGGAAGAAGA
AthmiR004	1	-21.51	1	Inter	AAAUUUCUAAAAUUUCAUCUAGUGGGACGGAGGGAGUAU AUGGUUAUUUCAAGAAAAAAACUUAUUUACCAUAU AGUUUCGAUCAUUAACUAAAAACUGAAAAAAGAAAA
AthmiR005	1	-15.80	2	Inter	GAAACCAAAACGAAUAUUUAUAAAAUUCGGAACAGAUUUUA AAUAUUUCUACCGGAAUACACGUACCAAAUAAAAUCCGG AUAUUUAUCUGAAAACCCGAAUAUAAUUUAUUAUAA
AthmiR006	1	-10.80	250	Inter	AGAAGUAGACGAUGAAGAAGAAAAAGAAAAAAGG UACAAAUAGAGAAAGAAUAUAUUAUUGCUUUUUUGAGUA AUUAAAUAUAUAUAAUCUGAUAAUAGCUACAACUCC
AthmiR007	1	-17.40	2	Exon	AAACAGAAAAUUGGCCUGAAUAUAAAAGAAAAUACAUCU UGUAAGUGACAAACAAGCAUCAUGAAAAUGAAUAAGAACA AUACCAUGAACCAUUGCUACCUUGGACUUAUAAGGGAA
AthmiR008	1	-30.60	81	Inter	AAAUAAAAUAAAAUCUCUACAGAAGAACAAGUCUGUCU UUUGCAAAACAUUGUUUGAAGAAAAAGAAAUCAAAUGAUGGAA CAGAGCAGAAGAACAGAACAAGUUCUGUUAGUGCAGAG
AthmiR009	1	-15.72	41	Inter	UAGAAAAAAACAUUCAACUGAAUCAUCAAUGAUCUAUUU AAACAUGAAAAACAAACAACAUUUUCAUCUUAUGAGAU GGUUCAUCCAAUUGGACAAACAACAGGACCUUAAUAG
AthmiR010	1	-18.64	0	Intron	UAAUUUAUACCCUACACAAAACGGACUAUCUCUAAACAAUCU UAAACCUCUUCAGCGAAAAUUGAGGUAAAGGAUAGUGAAAG AUAGUGGUACAUAUGAUUAAACAAAUAACAUAUAA
AthmiR011	1	-16.38	46	Exon	AUCGAUUCAUUACCAGAACCAUUUUCUAAAAUCUGCAAU UGGCAAUGAGAGCAUUAGAAAGAGAGACAGAGGAAAAUGAG AAAGACGAUAAUGAGAAUAAGACAGAGAGAAUAUAAAC
AthmiR012	1	-21.80	3	Intron	AAUGAACAAACAACACUAGAGAACAACAACAUUGGAAUAU CUGUAGCUUAAUUCAUAAAAGGAUACAGAAACUGUAACCUU UUAACUUAAGCUCAACCAUUAGAGAUAAAGUUUAUGAC
AthmiR013	1	-17.45	40	Exon	GAUAAGUCUGGCUUGGCACAACUUGAAUCACACCUGGUUGU UCUGCAACAUCCAAAUCCCAAAAGCAUAUAAGAAAAACAG AACAAAGACAAAAAACAACAUUGAUCUAGUUGAAA
AthmiR014	1	-22.81	21	Exon	AAGAGAUACACAAGAGCAGGAAAAAAGAAGAAGAUAAAGA GAAAAUCACAUCAUUUAUUCUCUUGAGUGAUGAUUAAUAAG UGAAUGAAUCACUCAAGAAGAUUUAAAAAAGUGAUGA

(to be continued)

(continued)

miRNA No.	Chr.	$\Delta G(\text{kcal/mol})$	Tar ¹	Location ²	Sequence
AthmiR015	1	-18.03	123	Exon	UAUACAUAUAUAGAGAGAGAGAAGAGGACAAAGAGUUGAA AGAUGAAGACUCUCAUGUCUUCUAUAGAAACAAGUGAUUAUGU GCGCUAAGAAAAGAAGAAGAAGAAGAAGAAGAAGAAGACA
AthmiR016	1	-20.03	3	Inter	AAAGUAAAUUUAAAUGCAUGGAGAAUAGAAGUAUAAAACU AAAUUUAUUCAAUUCUAUUAUGUAAAAUUUUUAAAGAAGAAG AUUAAAUUUAAAGUAAGACUUUGGUCUCUAAAGAGCAAAUU
AthmiR017	2	-16.00	10	Inter	UUUUGAUUUUGUAACAAAAAUCUUAUGAUACUUAACCGG UCUAACCAAUGCAGACAUUUUAGUAGGAGUAUCAUAACAA AUUUCAAAACAAUAAACAAACAUUAUUAAUUAUACUAAA
AthmiR018	2	-14.62	250	Exon	UCUAAAUCCUAAUUUAUGAGAAGAAAAAGUAGAAUUUUUC ACUAAUCCUAAAAUCAGACAAAACAAAAAGUGAUUUUGUUG AGUGAAAAAAAUCUUUAGAGAGAGAAAAGAAAGAAGAAG
AthmiR019	2	-25.20	26	Inter	CAAAAAAAAAAAAAUUUUAAAAAAAAAGUAUGAGAGAAGGG AGAAAAAGUAGGAGAGAAGGAGAGUUGAGUUUCUCGGAGG AGAAACUUUGAGAAACUAUUCUCAUCCAAUUUGGACAGGU
AthmiR020	2	-21.48	6	Exon	CGAGAAGCAGGAUGAUCUAAAAGAACUGACUUUGUUAUCUU CUGGGGAAAUAUUAAACAAAACUACAGAGGAAAAAGAAACAA AUUUUAGGCUAAACAGAACAAACCUGCAGAGAAACACC
AthmiR021	2	-16.51	10	Intron	AAGAAACAUUACUUAUUCUUAAGAAUUUAAACCCAAAAAAAACA AAUUACUUUUCAAAAACUUAUCUUCUACCCUAUACAAGUAG GAGUGGGCCAAAUUCUAAUAAACAGAGAGAAAAGGUAAA
AthmiR022	2	-14.72	47	Inter	CAUCCAAUUCUCAAAAUCUCCAAUUUCCUAAACAAAUAACAC ACACAGAUCCAUAUUGAACAAAACAAAGAAGACGAAACGU GAUUUUGAAGACUCGUAGAAAAUAGAGAUGAUCAGAAA
AthmiR023	3	-20.20	247	Inter	AGAAGAAGAAAGAAGAAGGAACAAGAAAAAUAUUUUUUUUUU AGAGGAAGGGGGCGAGAGAAAGGAAUAGCAGAAAAUAAUACG CACUGUGAUUUUGGAAGCGUAGGGCUCUCUUUUUUUUUUUU
AthmiR024	3	-19.19	10	Inter	AGAUUAACAAAACAUUCUUCUUCUAUCAAGUAACAAUGUUA UAUAGCAUAAGAGAGAAAAAUGGGCAUGAAUCGAAGAAGAG CUAAUACAUAUUGUUGUAAAGAUGGCACAAGAAGAAAC
AthmiR025	3	-23.06	2	Inter	UAGGUCAAAUUACUCCUAAAUUUAAAGCAGAGUGUUGCUGAG UCAAAAAAAAAACACUAGCCUAAUUUUUACUAAAAAGAAAAAA AAGAGAAAAAGGAGGCUUGAGUUUACUUAAGCAAAUAAA
AthmiR026	3	-15.99	8	Intron	AUAAUAAAGAAAAAAUAAACAAGAUAAAAAAGGAUUAUUUG UUAUCGCAUGUAUUUCAAAAAAAAAUAUUAUACAAAGGAUUA UUACGUAAUUACAUGUGUUCUCCAACAUAUUUCCGACAA
AthmiR027	3	-9.70	23	Intron	UUUAUAUGUUAUUCUAAUUUUAAGAUUAUUAGUUGAAAAUAA UCGUGACAAAAAUAUUAGAGGAGAGAGGAAAAAUGAAAA ACAACAUAUAGCACAAAUAUUUAGGACGUAGAAAAUAAA
AthmiR028	4	-24.12	250	Inter	UUCUUCUUCUUCUUCACCAUCGAAAAGAGAUAAUGAACCAA GAAGAAGAAAAACAGAGAACAAAAGGAUCAACGAGAUCCA UGAAGACGAAGAAGAAGAGUUGGAGAACAAGAAGAUGG
AthmiR029	4	-17.51	72	Inter	AAUCAUCAUCAGUCUGCAUAGAAGAAUCAAGAAGCUAAAGA AUCUUAAAAACGAAAAUAAUAAUAAAAUCAAGAAACAUAG AUUCUUGAGGAAUGUGAAGUUACCAAGUCUGAUUGAUU

(to be continued)

(continued)

miRNA No.	Chr.	$\Delta G(\text{kcal/mol})$	Tar ¹	Location ²	Sequence
AthmiR030	4	-25.40	9	Inter	AUCAUUCAGAUGCAUCAUCCAAAUGGAUCAUGUAAAUGAAU CAUUUGGAUGUAAAUGCUAAAUGAUGAAACAAACAGGACCU AAAUAUAUAACACAAAAAUAAACAAAUAUACUAUAAU
AthmiR031	4	-21.24	253	Intron	AGAGCAGCAGAAGAAGAUGAAGACCCACGUUGGUGCUG CUAAUCUCAGAUACAAACAUGGGUUCUUAUAAACCAGAGAA UCUAAAAAAGAUUUGAAAAGAAGCAUCAAAAAUAAUAA
AthmiR032	4	-18.49	9	Intron	UUCACUCAAAACGAAAAUAUCUAAUGGCUAAACCACUAGUCU AGACACUUUAAAAGAAUAAUUGAAAAUGAUUUUUGUAAAAA AAAAGAAAAGUGAGACUGUGAGAAAAGCCAUGCCCAUUA
AthmiR033	5	-19.90	3	Inter	AGAAAAUCAAAACAUAAACACAACAUUAGAUGGUUAGUCUCU CCCCCAAACUUAUUUCACACCCGUCUCGGUGUAAAGAUAAUU CCGAAAAAAGACUAACGAAAAACAAAGAGAAAAUGAA
AthmiR034	5	-21.20	250	Inter	AUUAAAUAUCAAAUGAUUAAACAAACACUAGAAACAUCAUU CAAUUGCAUCAUUUAAAUGAAUCAUGUAAAUGAAUCAUGUA AAUGAAAAUGCUGUAAAUGAUGAAACAAACAGGACCUAAC
AthmiR035	5	-17.95	4	Inter	AACAAACCACCAUCGUUGCAAUUUCUUGAUGUUGAUCUGG CAAAAACCGAAGAUGAUCAGCGAAACAUUAAAAAAAAAAAAA CAGAUCAAUCAAGAAAACCAUAAAUCUGUAAAGAACAUG
AthmiR036	5	-20.05	3	Inter	AGCUCCAAGAUGUGUAAGAGUGCCUAAAUGACUCAAAACA UACGAAAAGAUUAGAAAGAGUCUAAAAACAACUCUGAAACU AUGUUUGAAAUCAGUAAAACUAGGACAUUCAAAAAG
AthmiR037	5	-16.43	31	Intron	ACAAGUGCAGAAAGAUCCUUCAUAAUUUGAGCAAGAUCGUU UACUGAUUCUACAACCUGGACAAUGACACAGACAACGAAGA AGAAAAAAGAUUUGAUUGAUUGCAAAAAGAAAAA

¹Tar: Number of Targets; ²Loc: Locations of miRNA (include Inter, intron and Exon)

3 讨 论

miRNA 在基因调控中扮演着重要的角色^[22]。Lewis 等^[23]研究认为, 人类有 1/3 的基因由 miRNA 调控。因此尽快找出所有的 miRNA 并研究其功能, 对进一步理解基因表达调控具有重要意义。但由于其片段较短, 利用实验方法快速识别 miRNA 具有很大困难^[21]。因为在实验水平上, 目前检测 miRNA 主要是分离 18~28 nt 的小片段 RNA, 然后再通过克隆和测序的手段来获得。由于实验本身的问题, 使得研究者们克隆到的 miRNA 仅仅是表达丰度较高的少数 miRNA, 而大批的低丰度 miRNA 却很难通过实验手段分离到。因此, 利用计算生物学方法来预测 miRNA 具有重要意义。通过计算的方法来预测 miRNA 能在短时间内识别出大量的 miRNA, 但同时也会产生大量的假阳性序列, 因此如何提高识别的准确率是在 miRNA 预测

中亟待解决的问题, 也是生物信息学领域中普遍存在的问题。

目前, 对于拟南芥 miRNA 的预测一般仅从基因间区入手, 这将直接导致部分由基因区编码的 miRNA 被遗漏。此外, 目前在拟南芥的 miRNA 预测中, 对于大量具有发夹结构的 pre-miRNA-like 序列的筛选主要是根据 miRNA 的种间保守性, 利用比较基因组学方法来进行^[18,24]。比较基因组学方法虽然可以对 miRNA 进行有效的识别, 但却很难发现非同源的 miRNA。为了克服上述缺陷, 从拟南芥基因组中找出新的 miRNA, 本研究首先从基因组入手, 在拟南芥全基因组范围预测 miRNA, 从而克服了目前方法对基因内编码 miRNA 的遗漏的问题; 然后本研究发展了一个 SVM 模型 GenomicSVM, 它基于机器学习算法, 对 miRNA 前体的筛选, 无需比较基因组方法, 也可以从大量 pre-miRNA-like 序列中识别出真正的 pre-miRNA

基因。检测表明该模型的敏感性为 83.3%，特异性为 98.1%。通过应用 GenomicSVM 模型，最后从 453 条可能的 miRNA 前体中预测出 37 条新型的拟南芥候选 miRNA。

发夹结构被认为是 pre-miRNA 的一个重要特征，在 miRNA 前体识别中是必不可少的环节。然而，在基因组中，往往存在着大量的具有类似结构的序列，因此如何从这些具有类似发夹结构的片段中找出少数真正的 pre-miRNA 是目前计算生物学预测 miRNA 的核心问题。本研究开发的 GenomicSVM 模型具有 98.1% 的预测特异性，因此应用该模型能够有效地从大量具有类似发夹结构的序列中筛选出少数真正的 pre-miRNA，然而，令人遗憾的是该模型还存在 16.7% 的假阳性识别。因此，如何进一步优化该模型，以提高敏感性是今后工作的重点。

参考文献:

- [1] Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. *Science*, 2001,294(5543):853~858
- [2] Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 2001,294(5543):858~862
- [3] Lee RC. Ambros an extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, 2001,294(5543):862~864
- [4] Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP. MicroRNAs in plants. *Genes Dev*, 2002,16(13):1616~1626
- [5] Llave C, Kasschau KD, Rector MA, Carrington JC. Endogenous and silencing associated small RNAs in plants. *Plant Cell*, 2002,14(7):1605~1619
- [6] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 2004,116(2):281~297
- [7] Sunkar R, Zhu JK. Novel and stress regulated micro-RNAs and other small RNAs from *Arabidopsis*. *Plant Cell*, 2004, 16(8):2001~2019
- [8] Wang XJ, Reyes JL, Chua NH. Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol*, 2004,5(9):R65
- [9] Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of *Caenorhabditis elegans*. *Genes Dev*, 2003,17(8):991~1008
- [10] Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. Vertebrate microRNA genes. *Science*, 2003,299(5612):1540~1546
- [11] Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J. Computational and experimental identification of *C. elegans* microRNAs. *Mol Cell*, 2003,11(5): 1253~1263
- [12] Lai EC, Tomancak P, Williams RW, Rubin GM. Computational identification of *Drosophila* microRNA genes. *Genome Biol*, 2003,4(7):R42
- [13] Wang XW, Zhang J, Li F, Gu G, He T, Zhang XG, Li YD. MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 2005,21(18):3610~3614
- [14] Nam JW, Shin KR, Han JJ, Lee Y. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic acids research*, 2005,33(11): 3570~3581
- [15] Griffiths-Jones S. The microRNA registry. *Nucleic Acids Res*, 2004,32(Database issue):D109~D111
- [16] Loong SK, Mishra SK. De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 2007, 10.1093/bioinformatics/btm026
- [17] Xue CH, Li F, He T, Liu GP, Li YD, Zhang XG. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 2005,6:310
- [18] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. *Monatshefte f Chemie*, 1994,125(2): 167~188
- [19] Chang CC, Lin CJ. LIBSVM: a library for support vector machines. 2001
- [20] Allen E, Xie Z, Gustafson AM, Sung GH, Spatafora JW, Carrington JC. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat Genet*, 2004,36(12):1282~1290
- [21] Allen E, Xie Z, Gustafson AM, Carrington JC. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, 2005,121(2):207~221
- [22] Bonnet E, Wuyts J, Rouze P, Van de Peer Y. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc Natl Acad Sci USA*, 2004(101):11511~11516
- [23] Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 2005,120(1):15~20
- [24] Alex A, Cameron J, Sizolwenkosi M, Sarah AE, Varun M, Vicki V, Venkatesan S. Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Research*, 2005,15(1):78~91

PREDICTION AND ANALYSIS OF NOVEL miRNA IN *Arabidopsis thaliana*JIN Wei-bo^{1,2}, KONG Dong², YING Xiao-min¹, GUO Ai-guang², LI Wu-ju¹

(1. Institute of Basic Medical Sciences, Academy of Military Medical Sciences, Beijing 100850; 2. College of Life Science, Northwest A&F University Yangling, Xi'an 712100, China)

Abstract: MicroRNAs (miRNAs), ranging in size from 20~25 nt, are a growing family of noncoding RNAs. They play an important role in the regulation of gene expression. The low abundance of some miRNAs and their time- and tissue-specific expression patterns make them difficult to be identified. To identify the novel miRNA systematically in *A. thaliana*, the authors firstly found 453 pre-miRNA candidates from the genome using the characteristics of the known *A. thaliana* miRNAs and comparative genomics methods. Then, in order to reduce the number of putative pre-miRNA candidates, the authors developed a SVM (support vector machine) model, GenomicSVM, using the human miRNA dataset as the training dataset. The model had the sensitivity 86.3% and specificity 98.1% respectively on the human test dataset, which contained 30 positive human pre-miRNAs and 1000 negative pre-miRNAs. For the 78 positive pre-miRNAs in *A. thaliana*, the model could pick up 73 pre-miRNAs and therefore the correct rate was 93.6%. Finally, the GenomicSVM was used to discriminate whether each 453 pre-miRNA-like sequence was pre-miRNA or not. The results indicated that there were 37 novel miRNA candidates. Therefore, the study in this report provides bioinformatics help for the experimental identification of miRNAs in *A. thaliana*.

Key Words: *A. thaliana*; Genome; microRNA; Prediction

This work was supported by a grant from The National Natural Sciences Foundation of China (30470411, 30500105)

Received: Feb 2, 2007

Corresponding author: LI Wu-ju, Tel/Fax: +86(10)66931324, E-mail: liwj@nic.bmi.ac.cn;

GUO Ai-guang, Tel/Fax: +86(29)87026171, E-mail: guoaiguang@yahoo.com.cn