利用支持向量机识别 miRNA 成熟链

吴方丽[a,b],金伟波[a],段 敏[a],王保莉[a],曲 东[c]
(西北农林科技大学 a. 生命科学学院,b. 植物保护学院,c. 资源环境学院,陕西杨凌 712100)

[摘 要] 【目的】开发一个用于从 pre-miRNA 上识别其成熟链的 miRNA 预测程序 miR-SVM。【方法】利用支持向量机工具,将 pre-miRNA 的序列结构特征作为支持向量机工具——LibSVM 的输入向量,经 Grid 程序优化参数后,开发出一个可靠的 miRNA 成熟链预测程序 miR-SVM。【结果】检测结果表明,在 miR-SVM 人数据集上得到程序的敏感性和特异性分别为 83.7%和 81.2%。通过杂交验证,获得 ROC 曲线下的面积约为 87.71%,表明研究提出的序列结构特征可以有效地预测 pre-miRNA 成熟链。此外,用人数据训练出来的 miR-SVM 程序对其他 20 个物种的 pre-miRNA 成熟链进行预测,结果正确识别率为 89.2%。【结论】研究开发的 miR-SVM 程序,成功地预测了 pre-miRNA 成熟链,检验结果表明,该程序具有良好的推广性,可用于 miRNA 试验过程中的前期预测分析。

[关键词] miRNA;成熟链;预测分析;支持向量机
[中图分类号] Q811.4 [文献标识码] A [文章编号] 1671-9387(2009)03-0219-04

Idetification of mature microRNA on the precursor using *ab initio* prediction method

WU Fang-li[a,b], JIN Wei-bo[a], DUAN Min[a], WANG Bao-li[a], QU Dong[c]
(*a. College of Life Sciences, b. College of Plant Protection, c. College of Resource and Environment, Northwest A&F University, Yangling, Shannxi* 712100, *China*)

Abstract: 【Objective】 The research predicted mature strand on the miRNA precursor. 【Method】 A program, miR-SVM, was developed based on support vector machine for identifying mature strand on pre-miRNA. To optimize the SVM classifier, the penalty parameter C and the RBF kernel parameter γ were adjusted based on the training set data using the grid search strategy in LibSVM. 【Result】 The miR-SVM had the sensitivity of 83.7% and specificity of 81.2% respectively on human data. The ROC of the model was plotted with the specificity and the sensitivity from the results, and gave an area under the ROC curve of 87.71%. Interestingly, the miR-SVM classifier built on human data can correctly identify up to 89.2% of the real miRNAs from 20 other species. 【Conclusion】 The successful detection of mature strand on the precursors provides a reliable method for predicting mature miRNAs from their precursors.

Key words: miRNA; mature strand; prediction; Support Vector Machine

MicroRNA(miRNA)是一类大小为 21～22 nt、存在于动植物体内的内源性小 RNA,对生物体的转录后基因调控起着关键作用[1-5]。目前可知,动物 miRNA 来源于长的初级转录物(pri-miRNA),然后在 Drosha 酶的作用下,pri-miRNA 被切割成 60～70 nt、具有发夹结构的 miRNA 前体(pre-miRNA)[6-7],最后在 Exportin-5 的作用下,将 pre-miRNA 从细胞核内运至细胞质中,并在 Dicer 酶的作用

* [收稿日期] 2008-05-12
[基金项目] 国家自然科学基金(30771442);陕西省自然科学基金(SJ08-ZT04);浙江省"生物医学工程"重中之重开放基金(SWYX0819);西北农林科技大学人才专项(01140412)
[作者简介] 吴方丽(1979－),女,河南睢县人,讲师,主要从事植物病理与分子生物学研究。
[通信作者] 金伟波(1977－),男,浙江温岭人,博士,主要从事植物分子生物学与生物信息学研究。E-mail:jinweibo@nwsuaf.edu.cn

下最终形成成熟的 miRNA[8-15]。因此,miRNA 的成熟过程主要包括 2 个步骤:(1) pri-miRNA 在 Drosha 的作用下形成 pre-miRNA;(2) pre-miRNA 经 Dicer 作用形成成熟的 miRNA。

miRNA 在生物体的不同部位和不同发育阶段,对基因的转录后调控都起着重要作用[1-5],因此探明各物种的 miRNA,对研究这些物种基因的调控机理有重要意义。但通过实验手段只能克隆到高丰度的 miRNA,而大量表达量较低的低丰度 miRNA 却无法得到,因此利用计算生物学预测低丰度或组织特异性 miRNA 是最有效的方法。到目前为止,已有不少相关的计算机程序被报道,如 MiRscan、SRNAloop、miRseeker 和 miRAlign 等[16-20]。但上述程序都只能用于对 pre-miRNA 的预测,而不能用于对 miRNA 成熟区的预测。因此,本研究拟开发一个高效的 miRNA 成熟区预测程序 miR-SVM,用于在 pre-miRNA 上识别成熟 miRNA 片段,为快速有效地从各物种中识别成熟 miRNA 序列提供保证。

1 材料与方法

1.1 序列资源

本研究所用的 miRNA 序列均从 miRNA 数据库[21](Release 8.0) 中下载,从中去除了无发夹结构的 pre-miRNA 以及发夹结构中两个茎都包含成熟链的 pre-miRNA,最后得到 377 条人 pre-miRNA。

1.2 支持向量机模型

本研究采用的支持向量机——LibSVM 软件包[22]下载于 http://www.csie.ntu.edu.tw/cjlin/libsvm/oldfiles/。此外,将金伟波等[23]提取 pre-miRNA 的序列结构特征作为 LibSVM 的输入向量。

1.3 参数的优化及预测程序的评估

为了提高预测程序的识别效率,利用 Grid 程序优化 LibSVM 的罚分参数 C 和 RBF 的核心参数 γ;此外,本研究通过交叉验证的方法评估预测程序的识别率。具体步骤为:(1)随机抽取一定数量真假序列样本作为训练集,剩余的作为检测数据集;(2)利用 10-fold 杂交验证方法优化程序的参数;(3)利用优化的参数训练模型;(4)利用剩余的检测集检测模型的识别率;(5)重复上述过程 10 次,将所得结果取平均值,然后用 ROC 曲线法评估模型。

2 结果与分析

2.1 真假数据集的获取

用 RNAfold[24]程序预测 377 条人 pre-miRNA 的二级结构,按照金伟波等[23]的报道,提取成熟 miRNA 的 25 nt 特征序列结构信息,将得到的 377 个序列结构数据集作为真集 POS;相反,在人 pre-miRNA 的另一个茎上,以 3 nt 为步长,提取 miRNA 25 nt 的特征序列结构,将最后得到的 2 373 个序列结构数据集作为假集 NEG。

2.2 miR-SVM 模型的训练和检测

为了使 miR-SVM 具有较高的识别率,本研究首先从 POS 和 NEG 中分别抽取 281 和 257 个样品作为训练集 TR;将 POS 中剩余的 96 个样品和 500 个从 NEG 中随机抽取的样本(与 TR 中的没有重复)作为检测集 TE;然后利用 TR 集和 10-fold 交叉验证的方法,执行 Grid 程序以得到优化的模型参数 C 和 γ;接着利用优化的参数训练 miR-SVM 模型,并用 TE 集检测模型的识别率;最后重复上述过程 10 次,结果如表 1 所示。由表 1 可知,模型的平均敏感性和特异性分别为 83.7%和 81.2%。

表 1 利用检测集 TE 评估 miR-SVM 的预测效率

Table 1 Evaluation of miR-SVM on test sets TE

重复次数 Repeat	识别出的真阳性序列 Positive prediction	敏感性/% Sensitivity	识别出的真阴性序列 Negative prediction	特异性/% Specificity
TE1	83	86.5	389	77.8
TE2	80	83.3	405	81.0
TE3	79	82.3	407	81.4
TE4	82	85.4	397	79.4
TE5	78	81.3	411	82.2
TE6	78	81.3	422	84.4
TE7	80	83.3	419	83.8
TE8	83	86.5	398	79.6
TE9	81	84.4	412	82.4
TE10	79	82.3	399	79.8
平均值 Average		83.7		81.2

2.3　miR-SVM 模型的评估

ROC(receiver operating characteristic)曲线法是目前公认的用于评估模型执行效率的可靠方法。本研究采用 LibSVM 工具箱中的 ROC 曲线绘制程序——Plotroc.py,用上述 10 次重复的平均敏感性和特异性绘制 ROC 曲线(图 1)。图 1 显示,ROC 曲线下的面积约为 87.71%,表明本研究采用的序列结构特征可以有效地区分真假 pre-miRNA。

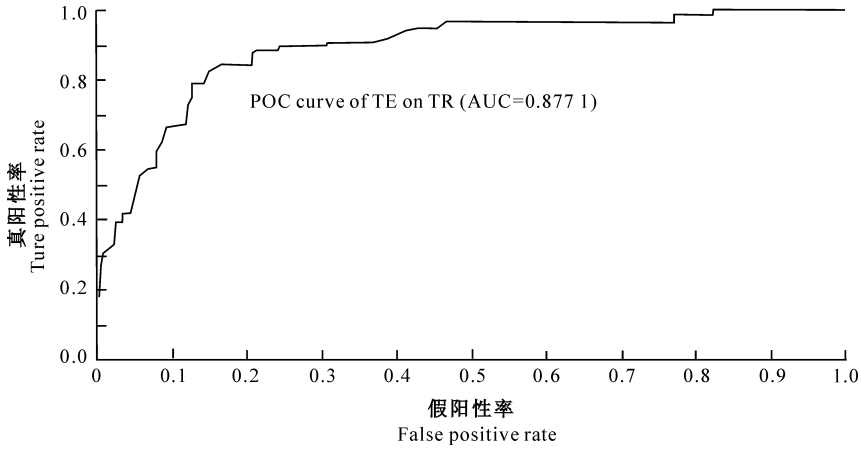

图 1　评估 miR-SVM 执行效率的 ROC 曲线

Fig. 1　The ROC curves of miR-SVM classifier evaluation

2.4　miR-SVM 在其他物种上的验证

由上述结果可知,用人 miRNA 序列数据训练出的 miR-SVM,可以有效地从 pre-miRNA 上识别成熟的 miRNA。为了进一步检测 miR-SVM 是否可以对其他物种的成熟 miRNA 链进行有效预测,又从 miRNA 数据库[21]下载了除人外其他 20 个物种的 pre-miRNAs 序列,同样去除无发夹结构的 pre-miRNAs 序列,以及与人 pre-miRNAs 同源的序列,最后得到包含 1 612 条 pre-miRNAs 的序列。根据金伟波等[23]的方法,提取 25 nt 的序列结构特征作为另一个检测数据集 TE-Other。表 2 显示,用 miR-SVM 对上述 20 个物种的 miRNA 进行预测,平均正确率为 89.2%,表明本研究开发的 miR-SVM 程序对其他物种的 miRNA 预测,同样具有良好的执行效率。

表 2　miR-SVM 在其他物种上的预测正确率

Table 2　Prediction accuracy on test set TE-Other

物种 Species	类型 Type	样本大小 Size	正确率/% Accuracy
黑掌蜘蛛 *Ateles geoffroyi*	真 miRNA 序列(real)	43	97.7
大猩猩 *Gorilla gorilla*	真 miRNA 序列(real)	76	94.7
绒毛猴 *Lagothrix lagotricha*	真 miRNA 序列(real)	44	95.5
狐猴 *Lemur catta*	真 miRNA 序列(real)	14	100.0
豚尾猴 *Macaca nemestrina*	真 miRNA 序列(real)	66	86.4
小家鼠 *Mus musculus*	真 miRNA 序列(real)	298	74.8
猕猴 *Macaca mulatta*	真 miRNA 序列(real)	64	87.5
倭黑猩猩 *Pan paniscus*	真 miRNA 序列(real)	80	92.5
黑猩猩 *Pan troglodytes*	真 miRNA 序列(real)	78	91.0
黄猩猩 *Pongo pygmaeus*	真 miRNA 序列(real)	74	89.2
褐家鼠 *Rattus norvegicus*	真 miRNA 序列(real)	191	81.2
家犬 *Canis familiaris*	真 miRNA 序列(real)	6	100.0
绵羊 *Ovis aries*	真 miRNA 序列(real)	2	100.0
野猪 *Sus scrofa*	真 miRNA 序列(real)	52	90.4
黑腹果蝇 *Drosophila melanogaster*	真 miRNA 序列(real)	64	85.9
果蝇 *Drosophila pseudoobscura*	真 miRNA 序列(real)	70	90.0
线虫 *Caenorhabditis elegans*	真 miRNA 序列(real)	105	78.1
拟南芥 *Arabidopsis thaliana*	真 miRNA 序列(real)	76	79.5
水稻 *Oryza sativa*	真 miRNA 序列(real)	143	83.9
玉米 *Zea mays*	真 miRNA 序列(real)	66	84.8
平均值 Average			89.2

3 小 结

目前,已有较多的生物信息学方法用于预测miRNA基因,然而对于利用计算生物学的方法预测成熟miRNA的报道寥寥无几。本研究利用支持向量机工具,开发了一个用于从pre-miRNA上预测成熟miRNA的程序miR-SVM,并通过检测表明,该程序具有良好的推广性,可以用于miRNAs实验验证前成熟链的识别分析。

[参考文献]

[1] Bartel B, Bartel D P. MicroRNAs: at the root of plant development? [J]. Plant Physiol, 2003, 132: 709-717.

[2] Bartel D P. MicroRNAs: Genomics, biogenesis, mechanism, and function [J]. Cell, 2004, 116: 281-297.

[3] Mallory A C, Vaucheret H. MicroRNAs: something important between the genes [J]. Curr Opin Plant Biol, 2004, 7: 120-125.

[4] Carrington J C, Ambros V. Role of microRNAs in plant and animal development [J]. Science, 2003, 301: 336-338.

[5] Hunter C, Poethig R S. Missing links: miRNAs and plant development [J]. Curr Opin Genet Dev, 2003, 13: 372-378.

[6] Lee Y, Jeon K, Lee J T, et al. MicroRNA maturation: stepwise processing and subcellular localization [J]. Embo J, 2002, 21: 4663-4670.

[7] Lee Y, Ahn C, Han J, et al. The nuclear RNase Ⅲ Drosha initiates microRNA processing [J]. Nature, 2003, 425: 415-419.

[8] Lund E, Guttinger S, Calado A, et al. Nuclear export of microRNA precursors [J]. Science, 2004, 303: 95-98.

[9] Yi R, Qin Y, Macara I G, et al. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs [J]. Genes Dev, 2003, 17: 3011-3016.

[10] Bohnsack M T, Czaplinski K, Gorlich D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs [J]. RNA, 2004, 10: 185-191.

[11] Bernstein E, Caudy A A, Hammond S M, et al. Role for a bidentate ribonuclease in the initiation step of RNA interference [J]. Nature, 2001, 409: 363-366.

[12] Grishok A, Pasquinelli A E, Conte D, et al. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing [J]. Cell, 2001, 106: 23-34.

[13] Hutvagner G, McLachlan J, Pasquinelli A E, et al. Acellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA [J]. Science, 2001, 293: 834-838.

[14] Ketting R F, Fischer S E, Bernstein E, et al. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans* [J]. Genes Dev, 2001, 15: 2654-2659.

[15] Knight S W, Bass B L. Arole for the RNase Ⅲ enzyme DCL-1 in RNA interference and germ line development in *Caenorhabditis elegans* [J]. Science, 2001, 293: 2269-2271.

[16] Lim L P, Lau N C, Weinstein E G, et al. The microRNAs of *Caenorhabditis elegans* [J]. Genes Dev, 2003, 17(8): 991-1008.

[17] Lim L P, Glasner M E, Yekta S, et al. Vertebrate microRNA genes [J]. Science, 2003, 299(5612): 1540.

[18] Grad Y, Aach J, Hayes G D, et al. Computational and experimental identification of *C. elegans* microRNAs [J]. Mol Cell, 2003, 11: 1253-1263.

[19] Lai E C, Tomancak P, Williams R W, et al. Computational identification of Drosophila microRNA genes [J]. Genome Biol, 2003, 4: R42.

[20] Wang X W, Zhang J, Li F, et al. MicroRNA identification based on sequence and structure alignment [J]. Bioinformatics, 2005, 21(18): 3610-3614.

[21] Griffiths-Jones S. The microRNA registry [J]. Nucleic Acids Res, 2004, 32: D109-111.

[22] Chang C C, Lin C J. LIBSVM: a library for support vector machines [DB/OL]. (2006-09-17) [2007-12-03]. http://www.csie.ntu.edu.tw/cjlin/~libsvm/.

[23] 金伟波,吴方丽,孔 栋,等. 水稻microRNA预测及实验验证[J]. 生物化学与分子生物学报, 2007, 23(9): 743-750.
Jin W B, Wu F L, Kong D, et al. Prediction and validation of microRNAs from rice genome using mature-SVM [J]. Chinese Journal of Biochemistry and Molecular Biology, 2007, 23(9): 743-750. (in Chinese)

[24] Hofacker I L, Fontana W, Stadler P F, et al. Fast folding and comparison of RNA secondary structures [J]. Monatshefte füur Chemie, 1994, 125: 167-188.