计算识别水稻基因组中microRNA基因*

金伟波[a,b], 吴方丽[a], 孔 栋[a], 郭蔼光[a,b], 杨淑慎[a]
(西北农林科技大学 a 生命科学学院, b 陕西省农业分子生物学重点实验室, 陕西 杨凌 712100)

[摘 要] 一些低丰度的miRNA和组织特异性miRNA往往很难发现, 利用生物信息学方法, 根据已知水稻miRNA的各种属性, 设计miRNA前体预测程序——PreMiRFind, 从全基因组范围中预测出1 375条miRNA前体, 然后利用已知拟南芥的miRNA对这1 375条前体进行同源检索, 找出166条miRNA前体, 再用水稻已知的miRNA对这166条序列进行检索, 发现其中153条与已知的水稻microRNA完全相同, 对剩余的13条miRNA前体预测序列用mFold作进一步结构分析, 最后得到10条新的候选miRNA基因序列。

[关键词] 水稻基因组; RNA fold; microRNA
[中图分类号] S511; Q522 [文献标识码] A [文章编号] 1671-9387(2006)12-0097-04

MicroRNA (miRNA)是存在于动植物体内的起调节mRNA稳定性及翻译作用的一类ncRNA[1-5]。最早发现的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, 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是通过RNAi途径使靶mRNA降解达到基因阻抑目的, 类似于具有相似长度的siRNA的功能。

随着人们对miRNA认识的加深, 更多的人开始相信, miRNA是早期RNA家族在进化上的遗留物, 其之所以没有被更为先进的蛋白质调控机制所代替, 可能主要是因为miRNA的表达比通常的蛋白质编码基因转录物的表达要迅速得多, 并且miRNA作为调节子不会受到翻译过程的影响, 所以有利于对靶基因进行更加迅速和有效的调控[7]。miRNA的发现, 丰富了人们对蛋白质表达调控的认识, 同时也是对中心法则中RNA作用的重要补充, 将促使人们对生物发育调控进行更深入的探索。

近10年来, 有关miRNA的研究工作已经取得了突飞猛进的进展。到目前为止, 在Rfam中登陆的miRNA已有1 000多条。但Bentwich等[8]认为, 生物体内的miRNA远大于此数, 他们通过计算并结合microarray技术, 从人基因组中又发现了89条新的miRNA, 自此在人类基因组中总共发现321条miRNA, 但他们认为人类基因组编码的miRNA至少有800条。目前在Rfam中登陆的水稻mircoRNA有123条, 而其中通过实验确证的却只有44条。由此推断, 在水稻基因组中也应存在大量尚未被发现的miRNA。本研究采用生物信息方法, 找出了10条新的水稻miRNA, 现将研究结果报道如下。

1 材料与方法

1.1 材 料

本研究所用的水稻基因组和EST数据从NCBI库(http://www.ncbi.nih.nlm.gov/)中下载得到; 133条水稻miRNA和76条拟南芥miRNA均来自于Rfam数据库(http://microrna.sanger.ac.uk/sequences/index.shtml), 用Vienna软件包中的RNAfold程式预测二级结构; RNAstructure为mfold的windows版, 下载于生物软件网(http://www.bio-soft.net/)。

1.2 方 法

1.2.1 miRNA属性的选取 为了区别miRNA与非miRNA, 本研究选取了miRNA的前体结构、平均长度、发夹结构中茎的长度、GC含量、环长及与拟南芥的相似度等几个属性, 用于后面的预测。

1.2.2 水稻miRNA前体的预测 根据上述属性, 对已知的133条水稻miRNA进行统计分析, 以得到的结果为参数, 设计出用于探寻水稻前体miRNA的程序——PreMiRFind, 其工作流程见图1。

* [收稿日期] 2005-11-26
[基金项目] 国家转基因研究与产业化开发专项(JY03A-11-01); 陕西省农业分子生物学重点实验室项目
[作者简介] 金伟波(1977-), 男, 浙江温岭人, 在读博士, 主要从事植物分子生物学研究。
[通讯作者] 郭蔼光(1943-), 女, 陕西西安人, 教授, 博士生导师, 主要从事植物分子生物学研究。E-mail: guoaiguang@yahoo.com.cn

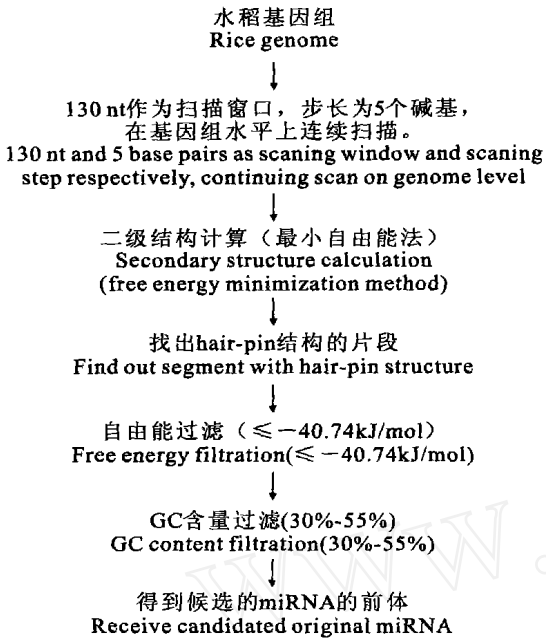

图1 PreMiRFind工作流程图

Fig. 1 Flow chart of PreMiRFind

1.2.3 水稻miRNA的预测 以拟南芥已知的miRNA为quest,对上述获得的前体miRNA进行blast检索,所获片段被认为是最后预测的miRNA,如检索到的序列有重复,只取其前体形成二级结构自由能最小的序列,其他的作为冗余片段丢弃,最后再去除已发表的水稻miRNA,剩下的就是新的miRNA。

2 结果与分析

2.1 水稻miRNA的特征

对已知的133条水稻miRNA进行分析,结果发现水稻前体miRNA的平均长度约为131 nt,所有前体均具有发夹结构,发夹环长度大于4 nt,GC含量为30%～55%;成熟的miRNA由20或21 nt个碱基组成。

2.2 水稻前体miRNA的搜索

以130 nt作为窗口长度,5 nt作为步长,从水稻每条染色体的5′端开始滑动窗口,调用RNA fold预测子片段的二级结构(zuker),判断其是否有发夹结构,没有发夹结构的丢弃,有发夹结构的,若环长>4 nt,茎长≥20 nt,并且GC含量为36%～80%,则保存此片段作为候选前体片段,不符合的丢弃;然后向下滑动5 nt,重复上述工作。经过这一步后,得到约1万条序列,去除冗余后还剩余1 375条候选前体片段。

2.3 成熟水稻miRNA的预测

为了进一步在miRNA前体上预测成熟片段,本研究用已知的拟南芥成熟miRNA作为quest,对上述预测出的1 375条候选片段进行blast分析,要求匹配长度大于17 nt,结果得到166条符合条件的候选片段。然后再用已知的水稻miRNA对这166条片断进行检索,以去除已发表的miRNA,最后得到13条miRNA片断。又用RNA structure对这13条序列进行重折叠,去除3条结构不符片段,最终得到了10条新的候选miRNA前体,其二级结构见图2,序列及其前体在染色体上的位置见表1。再将这10条miRNA递交到水稻EST库进行blast检索验证,其中plus/plus匹配的有miR02,miR05和miR07 3条。

表1 预测出的水稻miRNA的序列及其在基因组上的位置

Table 1 Sequences of new miRNAs and their locations in the rice chromosome

miRNA编号 miRNA number	染色体 Chromosome	前体片段起始位置 position of original sequence	链位置 Chain location	候选的成熟miRNA序列 Candidated maturity miRNA sequence	同源的拟南芥miRNA Originated miRNA of *A rabidopsis thaliana*
miR01	1	38139509～38139603	-	UGCCAAAGGAGAUUUGCAUAC	ath-MIR399
miR02	1	5718080～5718148	+	UGUCAUCUUCAUCAUCAUCUG	ath-MIR414
miR03	2	12956021～12956100	+	UUGGAGAAGAGAGUGAGCACA	ath-MIR156
miR04	3	3034592l～0346041	-	AGCUGCCAGCAUGAUCUAACU	ath-MIR167
miR05	4	32655912～32656028	+	UUGUAUAUAGUGAAGUGUUUG	ath-MIR395
miR06	4	31838759～31838870	-	AGUGUUUGGGGGAACUCUCGA	ath-MIR395
miR07	5	2832917～2833037	+	UCCAAAGGGAUCGCAUUGUCU	ath-MIR393
miR08	7	28469789～28469879	-	GAAGCAGGGCACGUGCAUGCA	ath-MIR164
miR09	8	1701243～1701325	+	CUCCCUGUAUGCCACUCAUCU	ath-MIR160
MiR10	12	25510435～25510546	+	UGCCAGCAUGAUCUAGCUCUG	ath-MIR167

注:+.有意义链;-.无意义链。

Note:+. Sense strand;-. Anti-sense strand.

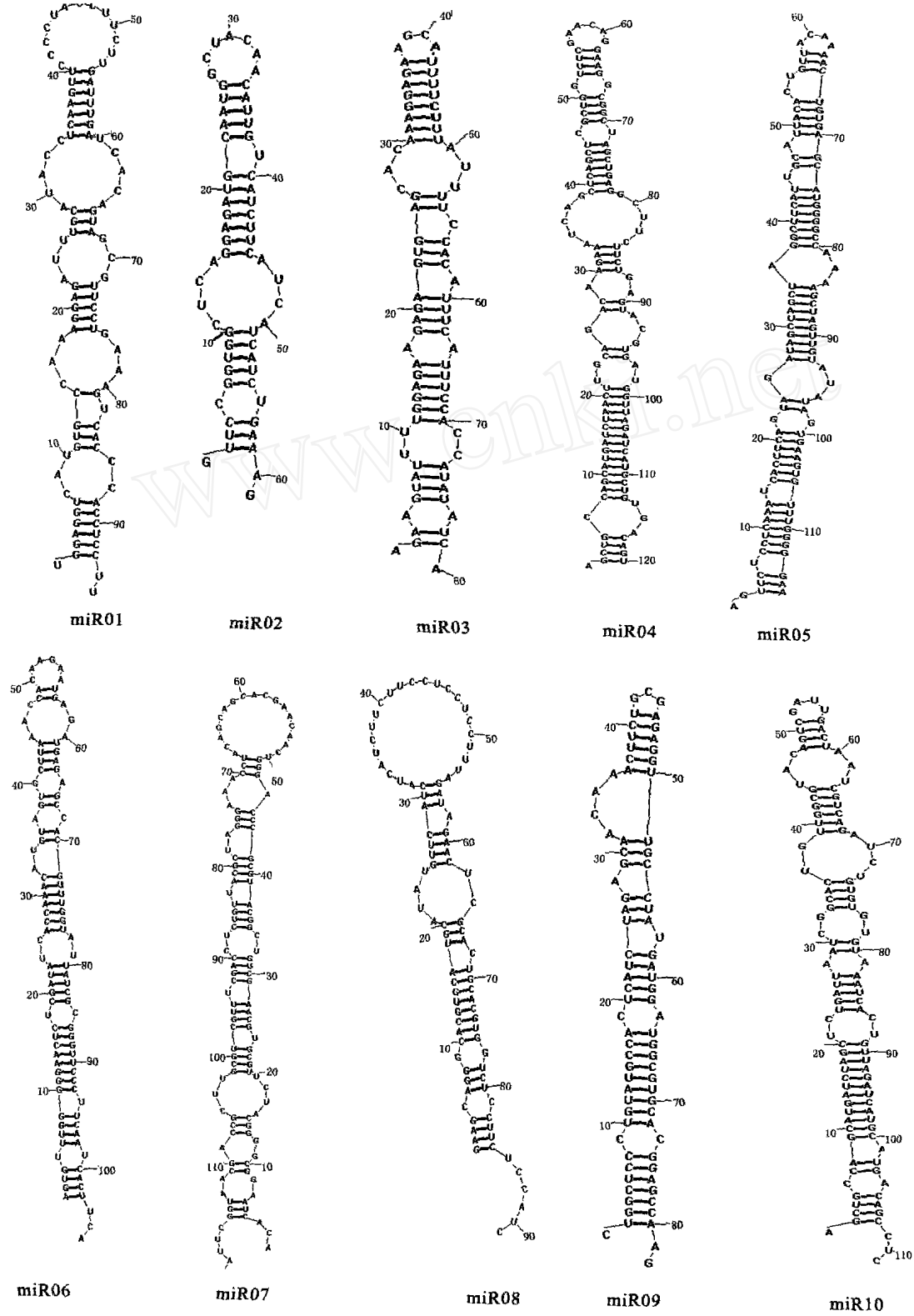

图2　预测出的水稻m iRNA 前体的二级结构

Fig. 2　Putative secondary structures of selected m iRNA precursors

3　讨　论

m iRNA 在基因调控中非常重要[9]。L ew is 等[10]研究认为, 人类有 1/3 的基因由m iRNA 所调控。因此, 尽快找出所有的m iRNA 并研究其功能, 对进一步了解基因表达调控具有重要意义。但由于其片段较短小, 在实验水平快速识别m iRNA 具有很大困

难[11]。因为目前在实验水平上检测microRNA,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,但由于实验本身的问题,研究者们克隆到的miRNA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,利用生物信息学的方法来预测miRNA意义深远。通过计算的方法预测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,因此如何提高识别准确率是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。

本研究采用全基因组扫描结合同源检索来识别miRNA的方法是一种高效、准确的方法。在本研究中总共识别出166条候选片段,而在用已知的水稻miRNA进行筛选后只剩下13条,说明其中包含了大量的已知水稻miRNA,这也从另一个侧面反映了本研究方法的识别率较高。最后将得到的10条microRNA对水稻EST库进行blast检索,发现有3条是Plus/Plus匹配,说明这3条microRNA在水稻生长阶段肯定能转录出来。但其他7条miRNA是否为真正的miRNA,还需要进一步的研究。

[参考文献]

[1] Lagos-Quintana M, Rauhut R, Lendeckel W, et al. Identification of novel genes coding for small expressed RNAs[J]. Science, 2001, 294(5543): 853-858.

[2] Lau N C, Lim L P, Weinstein E G, et al. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*[J]. Science, 2001, 294(5543): 858-862.

[3] Lee R C. Ambros an extensive class of small1 RNAs in *Caenorhabditis elegans*[J]. Science, 2001, 294(5543): 862-864.

[4] Reinhart B J, weinstein E G, Rhoades M Bartel B, et al. MicroRNAs in plants[J]. Genes Dev, 2002, 16(13): 1616-1626.

[5] Llave C, Kasschau K D, Rector M A, et al. Endogenous and silencing associated small RNAs in plants[J]. Plant Cell, 2002, 14(7): 1605-1619.

[6] Bartel D P. MicroRNAs: genomics, biogenesis, mechanism, and function[J]. Cell, 2004(116): 281-297.

[7] Ruvkun G. Glimpses of a tiny RNA world[J]. Science, 2001, 294(5543): 797-799.

[8] Isaac Bentwich, Amir Avniel, Yael Karov, et al. Identification of hundreds of conserved and nonconserved human microRNAs[J]. Nature Genetics, 2005(37): 766-770.

[9] Bonnet E, Wuyts J, Rouze P, et al. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes[J]. Proc Natl Acad Sci USA, 2004(101): 11511-11516.

[10] Lewis B P, Burge C B, Bartel D P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets[J]. Cell, 2005, 120: 15-20.

[11] Wang X J, Reyes J L, Chua N H, et al. Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets[J]. Genome Biol, 2004, 5: 65.

Computational identification of novel family members of microRNA genes in *Oryza sativa* genome

JIN Wei-bo[a,b], WU Fang-li[a], KONG Dong[a], GUO Ai-guang[a,b], YANG Shu-shen[a]

(*a College of Life Science, b Key Laboratory of Agriculture Molecular Biology in Shannxi, Northwest A & F University, Yangling, Shaanxi* 712100, *China*)

Abstract: MicroRNAs (miRNA) are about 21 nt long non-coding RNAs derived from larger hairpin RNA precursors and play important regulatory roles in both animals and plants. The low abundance of some miRNAs and their time- and tissue-specific expression patterns make experimental miRNA identification difficult. Here a program- PreMiRFind for genome-wide prediction of *Oryza* microRNAs was presented. This method used characteristic features of known *Oryza* miRNAs as criteria to search for miRNAs. Our prediction identified 1 375 pre-miRNA candidates. The 166 pre-miRNAs were reservation as criteria to search for miRNAs conserved between *Arabidopsis* and *Oryza sativa*. 153 pre-miRNAs were the same as known miRNAs and then the structures of the 13 pre-miRNAs were analyzed further. Finally 10 new miRNAs were obtained.

Key words: rice genome; RNA fold; microRNA