2015-02-25 DOI: 10.3724/SP.J.1145.2013.05021 http://www.cibj.com/
应用与环境生物学报 *Chin J Appl Environ Biol* 2015, **21** (1): 68-74

基于全基因组预测莱茵衣藻的新miRNA及其靶基因*

岳建宇 费忠安 郎小强 徐 辉 乔代蓉 曹 毅**
四川大学生命科学学院生物信息与代谢工程共享实验平台 成都 610065

摘 要 莱茵衣藻(*Chlamydomonas reinhardtii*)是一种重要的模式生物,其miRNA的发现相对较晚. 为系统化地预测分析莱茵衣藻的miRNA,采用比较基因组和同源比对相结合的方法,根据miRbase中已知的莱茵衣藻miRNA序列以及前体的特点,并且基于莱茵衣藻的全基因组对其miRNA的前体序列和成熟miRNA进行系统的分析和筛选,使用unigene和JGI的莱茵衣藻相关序列数据库对预测结果进行靶基因预测和功能的分析. 最终发现可能存在的miRNA 36条,其前体结构符合miRNA前体的基本特征且具有高度的同源性,两个数据库所得相匹配靶基因分别为64和32条,其中部分是与莱茵衣藻各项生命活动相关的基因. 本研究表明莱茵衣藻的基因组中具有可能存在的新miRNA家族,并且部分有高度匹配的靶基因,为其后续研究提供了可靠的理论支持. 图1 表3 参19

关键词 莱茵衣藻; miRNA; 全基因组; 比较基因组学; 同源比对; 靶基因; 生物信息学
CLC Q811.4

Genome-wide analysis of miRNAs and target prediction in *Chlamydomonas reinhardtii**

YUE Jianyu, FEI Zhong'an, LANG Xiaoqiang, XU Hui, QIAO Dairong & CAO Yi**
Sichuan Public Experimental Platform of Bioinformatics and Metabolic Engineering, College of Life Sciences, Sichuan University, Chengdu 610065, China

Abstract *Chlamydomonas reinhardtii* is an important model organism, but its miRNA is not fully understood. Therefore, systemic prediction and analysis of *Chlamydomonas reinhardtii* miRNA is of important significance for further researches. This study aimed to set up the parameters including GC content of pre-miRNA, number of loops in pre-miRNA secondary structure, size of loop, folding free energy and minimal folding free energy index, using the genome and homologous alignment method in combination with the characteristics of known 85 miRNA sequences and precursors of *Chlamydomonas reinhardtii* in miRbase. Based on the whole genome, the miRNA precursor sequences and mature miRNA were systemically analyzed and screened. Finally the softwares psRNATarget and psRobot were used based on the sequence data of *Chlamydomonas reinhardtii* from the unigene and JGI genomic databases to conduct target gene prediction and functional analysis of prediction results. All together 36 possible miRNAs were defined, with the two softwares obtaining 64 and 32 matched target genes respectively. The result indicated probable existence of a new miRNA family with some highly matched target genes in *Chlamydomonas reinhardtii* genome .

Keywords *Chlamydomonas reinhardtii*; miRNA; comparative genomics; homologous comparison; genome-wide; target gene; bioinformatics

莱茵衣藻(*Chlamydomonas reinhardtii*)是一种具有多叶绿体的单细胞绿藻,是藻类中一种重要的模式生物,主要用于研究基于叶绿体的光合作用,真核细胞鞭毛的结构、装配以及功能等[1-4]. 近年来,随着技术和需求的发展,也用于生物胁迫、生物能源等方面的研究[5-8]. 对于单细胞生物,miRNA的发现相对晚了很多年,而莱茵衣藻中miRNA于2007年第一次被Tao Zhao等人报道[9],这一发现也是对miRNA只存在于高等动植物种这一个观点的补充. 同年,另一篇关于莱茵衣藻miRNA的文章在《Nature》上发表[10],也丰富了人们对莱茵衣藻miRNA的认识. 2007年由DOE Joint Genome等研究机构承担测序的莱茵衣藻的全基因组基本公布出来[11],这为从整体上分析和预测莱茵衣藻的miRNA提供了非常好的数据基础.

以往大部分miRNA的预测方法是用已存入数据库内的miRNA序列为查询序列,而待测物种的DNA序列数据库中的数据作为比对库,全基因组测序完成的模式生物,使用其全基因组作为数据库;对于全基因组尚未完成测序或者尚未公布的物种来说,使用GSS序列或者EST序列来构建比对

收稿日期 Received: 2013-05-13 接受日期 Accepted: 2013-10-07
*国家自然科学基金项目(31272659, 31171447, 30971817)、国家"十二五"科技支撑计划项目(2014BAD02B02, 2013BAD10B01)和国家科技基础条件平台项目(NIMR-2014-8)资助 Supported by the National Natural Science Foundation of China (31272659, 31171447, 30971817), the Sci-tech Pillar Project of the Twelfth Five-year Plan of China (2014BAD02B02, 2013BAD10B01), and the National Sci-tech Infrastructure and Facility Development Program of China (NIMR-2014-8)
**通讯作者 Corresponding author (E-mail: caoyi_01@163.com)

库，然后在通过比对的结果所在的区域进行结构上的分析来判断是否可能为miRNA前体序列[12-15]. 虽然以往的方法也对miRNA序列前后关联部分的结构进行了分析和过滤，但由于其基于同源比对的序列均是已知的miRNA家族的序列，所以无法发现和已知序列非同源的miRNA序列，具有非常大的局限性. 因此，本研究基于莱茵衣藻的全基因组采用基于序列结构打分的方式进行预测，之后同源比对中加入相关物种EST序列进行比对，旨在能够有效地预测到已知miRNA家族以外的新miRNA.

1 材料与方法

1.1 数据来源

本文用于分析的数据主要包括莱茵衣藻的全基因组，莱茵衣藻已公布的miRNA和前体序列，非编码RNA的序列，编码RNA，用于保守性分析的EST序列. 其中莱茵衣藻的全基因组来自JGI（DOE Joint Genome Institute），miRNA及前体序列来自miRBase，莱茵衣藻非编码RNA的序列来自Rfam，而其编码RNA来自Refseq，用于保守性分析的EST序列则来自dbEST.

1.2 实验流程

1.2.1 数据预处理 全基因组共17条染色体加上scaffold的序列总共有88条序列. 按照染色体拆分为子文件并分别使用EMBOSS的einverted工具进行分析，参数设置如下：gap penalty = 40；Minimum score threshold = 40；Match score = 3；Mismatch score = -3；Maximum extent of repeats = 240；得出结果为反向重复序列对，之后用perl脚本将反向重复序列对拼接成符合条件的反向重复序列. Perl脚本主要根据miRNA以及前体的结构特征来对反向重复序列对进行处理，由于miRNA的前体通常为茎环结构，其茎的长度通常大于18 nt，且总长度大于45 nt小于260 nt，所以筛选出每对反向重复序列对中两条序列的长度需要大于或者等于18 nt，并且两条序列所构成的长序列始末位置的长度差大于或者45 nt小于或者等于260 nt，将符合条件的长序列左右两端各自延长10 nt，若长度不够则取序列左右两端的起始或者终止的位置. 经过处理所得到的序列储存为inverted_repeat_seq数据集.

1.2.2 去除待测序列中已知的序列 已知序列包括已知的成熟miRNA、miRNA前体序列、已知的编码序列和已知的非编码序列. 比对工具选择本地BLAST，对BLAST比对的结果用perl脚本进行筛选，踢出*e*值小于或者等于*e*-10或者匹配长度占到二者其中之一0.8以上的序列. 主要过程如下：（1）将inverted_repeat_seq.fasta文件中的序列与莱茵衣藻成熟的miRNA序列以及miRNA前体序列进行比对，按照*e*值与匹配长度的标准将达到标准以上的序列去除.（2）将步骤（1）中得到的结果与莱茵衣藻的编码RNA序列以及非编码序列进行比对，比对之后踢出高匹配的序列.（3）通过上述比对处理之后，将剩余的序列进行自我比对去冗余，最后保留结果为pre_miRNA_ candidate数据集.

1.2.3 二级结构及其自由能分析 miRNA前体的预测主要是对其二级结构进行分析，通过序列的环数、环的大小、自由能的大小，GC含量等信息的分析和处理，筛选出符合条件前体序列. 使用ViennaRNA Package的RNAfold [16]对前体的二级结构进行分析，参数设置为默认. RNAfold能够折叠出序列可能的二级结构，二级结构以RNAfold文本的格式来记录，同时能够生成每条序列的结构图. 该二级结构包含了序列的茎环结构，茎即为前体的双臂. 从二级结构的信息中我们能够得到环的大小、数量，臂的长度，有无错配隆起等信息，并且还能得出该二级结构进行折叠的自由能. 对二级结构进行过滤，参数设置如下：（ⅰ）折叠自由能（Minimal free energies，MFEs）小于或等于-28 kJ/mol；（ⅱ）折叠之后环的数量小于或等于2；（ⅲ）各个环的长度大于或等于4个碱基；（ⅳ）GC含量大于或者等于30%、小于或者等于70%；（ⅴ）最小折叠能指数（Minimal free energy Index，MFEI）大于或者等于0.85. 按照上述标准使用perl脚本对折叠出来的二级结构进行过滤，得到可能的前体序列，将序列存储入pre_miRNA数据集中.

1.2.4 成熟miRNA的预测 成熟miRNA预测主要从结构特征、保守性等方面来进行验证. 首先将pre_miRNA数据集中的前体序列数据通过miRCheck [17]中的extract_einverted_20mers.pl模块进行分析，得到了结构特征符合的miRNA序列，经过perl脚本自我去冗余后将数据集命名为20mer_seq. 然后对上述序列集合进行保守性验证，主要通过与其他模式植物的EST序列进行同源比对，最后得出同源性较高的序列则为高保守的成熟miRNA序列，存入conservative_miRNA数据集中. 最后用PsRobot [18]从miRNA的序列出发，分析和比对数据库中的高通量信息，包括莱茵衣藻基因组的特征、Degradome-SeqmRNA（降解组测序数据）等信息，从结构特征和实验数据等方面进行进一步的验证.

通过以上分析步骤所得到的miRNA则为符合前体结构特点且具有高保守性的miRNA序列.

1.2.5 靶基因预测和功能分析 本文预测选择了两款植物和藻类可用的靶基因预测工具，其中psRNATarget [19]数据库选择unigene，为了让结果更为准确，参数设置将Maximum expectation设置为2.5，其余参数为默认. PsRobot则选用JGI数据库来进行预测，条件设置为strict，其中Penalty score threshold设置为1.0，其余为默认值. 预测流程见图1.

2 结果

2.1 miRNA及前体预测的相关结果

本研究通过上述步骤对莱茵衣藻的基因组数据进行处理和分析之后，得出了一些相关数据，并进行了统计和归纳.

2.1.1 实验各个阶段的数据统计 本文使用EMBOSS中的einverted工具分析得出了莱茵衣藻基因组中所有反向重复序列对总共49 634对. 在进行BLAST比对去除已知miRNA和pre-miRNA序列、编码序列以及非编码序列之后，总共得到27 885条符合条件的反向重复序列. 反向重复长序列通过RNAfold处理，得到折叠的二级结构与自由能，在经过perl脚本进行前体结构的过滤，得到的可能的前体序列总共3 292条. miRCheck通过对前体序列的分析，得出成熟miRNA，经过自我去冗余之后总共得到3 539条序列. 对上述miRNA进行同源比对以寻找具有保守性的miRNA，最终得到miRNA序

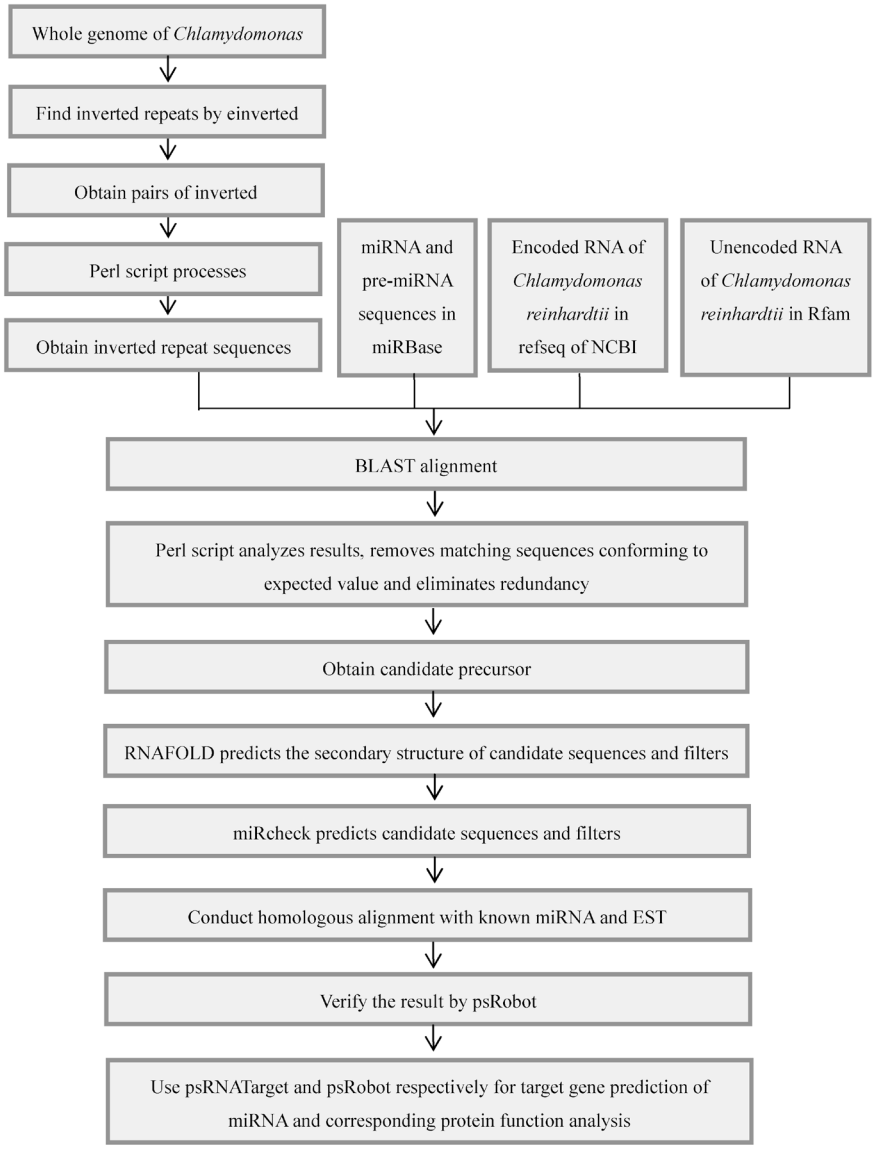

图1 莱茵衣藻的miRNA及靶基因预测流程.

Fig. 1 miRNA of *Chlamydomonas reinhardtii* and target gene prediction process.

列294条.

2.1.2 PsRobot分析得到的结果 通过上述步骤，得到高同源性序列共294条，上述结果中的miRNA序列通过PsRobot进行相关的验证，最终通过验证的miRNA总共为36条，miRNA以及其对应前体信息包括如下几个方面：miRNA编号，miRNA序列来源（在基因组中的位置），是属于-3p还是-5p，前体序列，前体所折叠的自由能，前体的二级结构，以及miRNA（序列为大写字母）在前体中的位置. 序列信息见表1.

2.2 靶基因预测和主要功能分析

本研究总共预测出莱茵衣藻的新miRNA数量为36条，通过psRNATarget预测得到21条miRNA序列65个靶基因（表2）；而通过psRobot进行预测最后得到18条miRNA序列32个靶基因（表3）. 从数量上以及被预测出具有靶基因的miRNA编号上来看，二者结果比较一致，但在各条miRNA靶基因的数量上以及靶基因相关注释内容上，具有一定的差异.

3 讨论

比较基因组学主要是基于基因组图谱或者序列结构，进行相似性的比较分析，用于了解物种之间的进化关系以及基因组内在结构特征. 而其用于miRNA的研究中比较具有代表性的研究应该是Jones-Rhoades和Bartel两人利用拟南芥和水稻全基因组鉴定在两个物种中保守的miRNA序列. 这个研究所发表的工具miRCheck也得到了广泛的应用和认可. 本文虽然是参照miRCheck的算法，但是根据莱茵衣藻miRNA的特点对加入了新的过滤条件，添加这些修改的主要目的是为了降低预测假阳性率，假阳性率在高通量分析中一直都是常见的问题，而随意的设置参数降低假阳性率又有可能遗漏掉

表1 预测所得miRNA序列信息

Table 1 Predicted miRNA sequences

miRAN_ID	miRNA_Sequence	Chromosome_NO	Begin	End	MFES	UTR
miR01	ATGCGCTGCTACAGCCTCGC	Chr01	6377931	6378000	-28.5	3'
miR02	TGCTGGGGCTGCACCTGCTG	Chr01	6603660	6603739	-47.1	3'
miR03	GTGGTCGCCATCGTGCTGCC	Chr02	2118683	2118772	-39.2	3'
miR04	CACCCTCCTGCTCACACTCA	Chr02	4324240	4324359	-38.7	3'
miR05	TGTTTGCTGCCGGTAGCCTC	Chr02	8595793	8595882	-39.7	5'
miR06	CGCCGACGGCACCAGCAACA	Chr02	8595803	8595882	-37.6	3'
miR07	TGGCCTGAACGCATGTGGCC	Chr02	8802670	8802749	-33.4	3'
miR08	TTCGGTGGTGCGGCGGCTGA	Chr03	147391	147500	-54.2	3'
miR09	CCGGCGGCGATGCCGTTGAT	Chr03	3024210	3024289	-41.4	3'
miR10	CCGTCGCGACCTTTCTGCTC	Chr03	5583412	5583511	-42.4	5'
miR11	CTGTAGCTACACGGCGGCGT	Chr05	1751056	1751135	-39	3'
miR12	GCGAGGAATCAGGGCCGCTG	Chr06	4853860	4853949	-47.7	5'
miR13	AGCTGCGGCTCCGCTTCCTC	Chr06	4853863	4853952	-47.7	3'
miR14	CGGTGTTATTGGGCGCCGTG	Chr06	7472605	7472704	-44.2	3'
miR15	CTGGAGCTCAGCAGCGTCTC	Chr09	52844	52933	-43.8	5'
miR16	GTAGACGCTGTTGAAGTCAG	Chr09	52851	52930	-37.8	3'
miR17	GTGGCGGTGGCATCAGCGGC	Chr09	1900159	1900228	-30.4	5'
miR18	AGCTTGGTGGCCTTCAAAGG	Chr09	4233801	4233880	-29.9	3'
miR19	GTGAGCGTGAACGGCGTCAT	Chr10	1294051	1294130	-36.7	5'
miR20	CAGCACGCCATTCGCGCTCC	Chr10	1294056	1294135	-36.8	3'
miR21	CGCACCATCAAGATGCTTCA	Chr10	2386987	2387086	-44.2	5'
miR22	ATGAAGCTGCTTGATGGTGA	Chr10	2386991	2387090	-43.3	3'
miR23	CTACTGCTACTGCAGCTGTT	Chr10	6078269	6078358	-43.5	3'
miR24	TGCATGTGTGTCCCCCGTGT	Chr11	142801	142910	-64.6	5'
miR25	TGTGCCCCGCGTGAGGGGCA	Chr11	142805	142904	-63.8	3'
miR26	CATAAGCGGCAGGATGACAT	Chr12	1627598	1627707	-49.5	3'
miR27	TCAAGCTGCTGCCGCCGCAG	Chr12	1731686	1731765	-29.3	5'
miR28	GCCGCAGTAGCTGCAGCAGC	Chr12	8186646	8186775	-59.5	5'
miR29	TTACTGGCCTTCGGGGGCTC	Chr13	605703	605772	-32.9	3'
miR30	CGGTAGGCCCACACAGGACC	Chr14	558419	558538	-51	5'
miR31	GTGTGATGTGGGCGTGGCTG	Chr16	1575924	1575993	-28.8	5'
miR32	GCGTCAGCTGCTGCAGGTGG	Chr16	1821277	1821366	-55.2	5'
miR33	GCGGAGGCGGTCGGAGGCGA	Chr17	3316892	3316971	-41.5	5'
miR34	TTTCGCCTGCTGCCGGCCCC	Chr17	3316898	3316977	-42.2	3'
miR35	GTAAAGCCCGCTGGCGCCGG	Chr17	4711256	4711335	-41.9	5'
miR36	GTCCTTGCGCGTGTACTGGG	Scf19	417464	417543	-33.1	5'

表2 使用psRNATarget所得相匹配的miRNA靶基因

Table 2 Matched miRNA target genes obtained by psRNATarget

miRNA_ID	Target_Acc	Expect	UPE	Target_Description
mir02	TC94895	1.5	22.222	UniRef100_A8HSJ3 Cluster: Predicted protein
	BU655561	1.5	23.988	UniRef100_A8HZH0 Cluster: Predicted protein
	TC82742	2.5	13.432	UniRef100_Q9XEK5 Cluster: Gamete-specific homeodomain protein GSP1
	TC87305	2.5	21.582	UniRef100_A8JEW6 Cluster: Mitogen-activated protein kinase 1
mir03	TC84053	2	23.329	UniRef100_A8JJP0 Cluster: Predicted protein
mir04	TC96965	2	14.409	UniRef100_A8J3S5 Cluster: Predicted protein
	TC91587	2.5	23.955	UniRef100_A8IZJ4 Cluster: Predicted protein
	BG850713	2.5	12.192	Unknown
mir05	TC92308	2.5	22.656	Unknown
mir06	AV388720	2.5	24.97	UniRef100_Q59098 Cluster: Dihydrolipoyllysine-residue acetyltransferase component of pyruvate dehydrogenase complex
mir09	BF863873	2.5	24.622	UniRef100_A1IVY2 Cluster: Cell wall glycoprotein GP3 precursor
	TC93219	2	19.578	UniRef100_A8ITN9 Cluster: Qb-SNARE protein, NPSN-family
	TC87229	2.5	22.103	UniRef100_P81831 Cluster: Phosphoenolpyruvate carboxylase 1
	TC84352	2	14.152	UniRef100_A8ITN9 Cluster: Qb-SNARE protein, NPSN-family

续表2

Table 2 Continued

miRNA_ID	Target_Acc	Expect	UPE	Target_Description
	BI718714	2.5	22.535	UniRef100_A8JAG1 Cluster: Predicted protein
mir10	BG846790	1.5	22.721	UniRef100_Q6K2M9 Cluster: Cell wall protein-like
	BG847776	2.5	7.197	UniRef100_Q9VPS3 Cluster: CG2839-PA
	BG852463	2.5	21.683	UniRef100_Q5ZD27 Cluster: Splicing coactivator subunit-like protein
mir12	BE121876	2.5	10.776	Unknown
mir13	TC86639	2.5	15.851	UniRef100_A8JA42 Cluster: Predicted protein
	TC87600	2	24.968	UniRef100_A8HME1 Cluster: Predicted protein
mir14	TC94304	0	18.77	UniRef100_A7CZJ0 Cluster: Regulatory protein LacI
mir15	TC89585	2.5	23.219	UniRef100_A8I2D8 Cluster: Predicted protein
mir17	TC84825	1.5	18.958	UniRef100_A4RD35 Cluster: Protein transport protein SEC31
	TC97817	2	16.882	UniRef100_UPI0000250939 Cluster: myeloid/lymphoid or mixed lineage-leukemia transloction to 6 homolog (Drosophila)
	TC85341	2.5	21.613	UniRef100_A8J1A2 Cluster: Serine/threonine-protein phosphatase PP2A-3 catalytic subunit
	TC84027	2.5	21.221	UniRef100_Q0DUA1 Cluster: Os03g0197700 protein
	TC87417	2	21.252	UniRef100_A8IVL6 Cluster: Predicted protein
	TC91727	2.5	22.653	UniRef100_A8IIQ5 Cluster: Predicted protein
	TC86094	2.5	20.08	UniRef100_A8J0A7 Cluster: Early light-inducible protein
	TC93610	2.5	24.362	UniRef100_Q9PF60 Cluster: Endo-1,4-beta-glucanase
mir18	TC84865	2.5	22.307	UniRef100_A8IRG8 Cluster: Predicted protein
	TC83328	2.5	23.891	UniRef100_A8I3F6 Cluster: Predicted protein
mir19	BG858972	2.5	13.276	Unknown
	BP092845	2.5	22.654	UniRef100_A8IKQ0 Cluster: Fructose-1,6-bisphosphatase
mir21	BG854605	2	17.667	UniRef100_Q8LKK4 Cluster: R1B72 protein
mir22	TC90245	2.5	21.188	UniRef100_A8IRX9 Cluster: Low-CO2-inducible protein
	FC108953	2.5	24.594	UniRef100_Q21KQ5 Cluster: Invasion gene expression up-regulator, SirB
mir23	TC82874	1.5	20.34	UniRef100_A8I635 Cluster: Predicted protein
	TC85436	1.5	24.551	UniRef100_Q9VJT7 Cluster: CG17341-PA
	FC105936	1.5	24.551	UniRef100_Q5IAC3 Cluster: PFL activating enzyme
	TC86556	1.5	21.709	UniRef100_Q9BIU8 Cluster: Flagelliform silk protein
	BF863866	1.5	21.5	UniRef100_Q5VPS5 Cluster: Immediate early protein-like
	AV627612	1.5	23.341	UniRef100_Q67WR0 Cluster: Fibroin heavy chain-like
	AV619159	1.5	21.709	UniRef100_Q84IE8 Cluster: PopA
	BU646639	1.5	17.949	UniRef100_Q0J5F7 Cluster: Os08g0438400 protein
	NP964402	2	21.253	GB|AY450930.1|AAS07044.1 plus agglutinin
	TC94047	2	23.329	UniRef100_A8JG80 Cluster: Phosphoglycerate mutase
	TC93002	2	20.689	UniRef100_A8IL90 Cluster: Predicted protein
	TC93971	2	24.469	UniRef100_A8JEK1 Cluster: Predicted protein
	FC101448	2	24.13	UniRef100_A8JEH6 Cluster:Predicted protein
	TC90644	2	23.614	UniRef100_A8J221 Cluster: Predicted protein
	BQ811441	2	22.703	UniRef100_A8J221 Cluster: Predicted protein
mir24	TC95885	2	24.379	Unknown
	BE212130	2.5	24.693	UniRef100_Q091H2 Cluster: Chemotaxis coupling protein CheW
	BU649619	2.5	17.997	UniRef100_A8IWA8 Cluster: Predicted protein
mir28	FC095522	2	18.941	UniRef100_Q9BIT6 Cluster: Major ampullate spidroin 1
	TC90767	2.5	19.888	UniRef100_A7UCH9 Cluster: Carbonic anhydrase
mir32	TC82741	2.5	19.65	UniRef100_Q5XPS1 Cluster: Axonemal inner arm I1 intermediate chain dynein IC138
	TC85204	2.5	24.898	UniRef100_A8JGD0 Cluster: Predicted protein
mir33	TC97654	0	22.322	UniRef100_A4QP49 Cluster: Zgc:103532 protein
	TC91619	2	11.43	UniRef100_A8I854 Cluster: Predicted protein
	BG854399	2.5	6.377	UniRef100_Q7BR61 Cluster: Topoisomerase IV subunit A
mir36	TC95006	2.5	19.051	UniRef100_A8J287 Cluster: Chloropyll a-b binding protein of LHCII type I, chloroplast

表3 使用psRobot所得相匹配的miRNA靶基因

Table 3 Matched miRNA target genes obtained by psRobot

miRNA_ID	Target_ID	Expect	Target_Description
mir02	Cre03.g201700.t1.2	0.5	Transcription factors
	Cre01.g026300.t1.1	1	HCP-like superfamily protein
	Cre10.g425000.t1.2	1	Protein kinase superfamily protein
mir05	Cre02.g137950.t1.2	0	Unknown
mir06	Cre02.g137950.t1.2	0	Unknown
mir08	Cre03.g148900.t1.2	0	Flavin-dependent monooxygenase 1
mir09	Cre03.g171950.t1.2	0	Phosphoenolpyruvate carboxylase 4
mir13	Cre03.g154150.t1.2	1	ARF GTPase-activating protein
mir14	Cre06.g310100.t1.1	0	Mitogen activated protein kinase kinase kinase-related
mir15	Cre09.g386600.t1.2	0	Transcription factors
	Cre09.g411350.t1.2	0.5	Transcription factors
mir16	Cre09.g386600.t1.2	0	Unknown
mir17	Cre14.g629100.t1.2	1	RWP-RK domain-containing protein
mir21	Cre10.g435700.t1.1	0	Unknown
mir22	Cre10.g435700.t1.1	0	Unknown
mir23	Cre01.g070000.t1.2	0.5	Dicer-like 1
	Cre02.g100850.t1.2	1	Organic cation/carnitine transporter4
mir28	Cre14.g608150.t1.2	0.5	Histidine kinase-, DNA gyrase B-, and HSP90-like ATPase family protein
	Cre19.g755900.t1.2	1	PAS domain-containing protein tyrosine kinase family protein
	Cre07.g314950.t1.1	0.5	DNA binding;DNA-directed RNA polymerases
	Cre01.g057256.t1.1	1	Integrase-type DNA-binding superfamily protein
	Cre09.g409250.t1.1	1	Nucleotide-diphospho-sugar transferases superfamily protein
	Cre10.g455950.t1.2	1	Flavodoxin-like quinone reductase 1
	Cre12.g552700.t1.2	0	Arabidopsis phospholipase-like protein (PEARLI 4) family
	Cre03.g188100.t1.1	1	F-box family protein
	Cre01.g008850.t1.2	1	RGPR-related
	Cre04.g227000.t1.2	1	DNA mismatch repair protein, putative
mir31	Cre16.g660050.t1.2	0	Unknown
mir32	Cre16.g661900.t1.2	0	P-loop containing nucleoside triphosphate hydrolases superfamily protein
	Cre07.g356700.t1.2	0.5	Unknown
mir35	Cre17.g733150.t1.2	0	Histidine kinase 3
mir36	Cre19.g754500.t1.2	0	Leucine-rich receptor-like protein kinase family protein

真正的miRNA分子. 本文通过对莱茵衣藻已知miRNA序列分析发现, 分别满足上述条件的miRNA序列的前体序列分别占到了85%-92%, 综合计算得出满足上述所有条件的miRNA占到了82.7%以上, 说明加入这些条件能够有效降低假阳性率, 同时也不会因大幅度降低预测的覆盖率而遗漏掉可能的miRNA.

同源性比对一直是miRNA高通量预测的重要手段之一. miRBase数据库目前更新到19.0版本, 成熟的miRNA总量为25 141条, 其中有非常多的miRNA家族至少在2个物种中出现, 植物当中特别明显, 比如miR159和miR171在数据库收录的大部分植物物种中都能找到. 正是由于这种保守性的存在, 同源性比对被广泛运用于预测各种新物种中的miRNA. 而本研究为了提高预测的准确性, 将符合条件的miRNA序列进行同源性比对.

本文通过一系列的预测和处理得到了一批具有高同源性, 且其前体具有典型符合miRNA前体标准的发夹结构的miRNA序列, 序列数为36条. 其他未通过同源比对的序列为3 243条, 以及未通过psRobot验证的序列258条, 这些序列的前体序列同样满足miRNA前体序列的特征, 由于条件的局限性本实验并未作进一步的验证. 而这3 501条序列当中是否含有新的miRNA序列, 还有待进一步的验证. 综上所述, 本研究对今后莱茵衣藻miRNA的研究工作具有参考作用, 并为miRNA的生物信息学分析提供了新的思路.

参考文献 [References]

1 Grossman AR. Chlamydomonas reinhardtii and photosynthesis: genetics to genomics [J]. *Curr Opin Plant Biol,* 2000, **3** (2): 132-137

2 Goldschmidt-Clermont M, Rahire M. Sequence, evolution and differential expression of the two genes encoding variant small subunits of ribulose bisphosphate carboxylase/oxygenase in *Chlamydomonas reinhardtii* [J]. *J Mol Biol*, 1986, **191** (3): 421-432

3 Cole DG. The intraflagellar transport machinery of Chlamydomonas reinhardtii [J]. *Traffic*, 2003, **4** (7): 435-442

4 Funke RP, Kovar JL, Weeks DP. Intracellular carbonic anhydrase is essential to photosynthesis in *Chlamydomonas reinhardtii* at atmospheric levels of CO_2 (demonstration via genomic complementation of the high-CO_2-requiring mutant ca-1) [J]. *Plant Physiol*, 1997, **114** (1): 237-244

5 Elbaz A, Wei YY, Meng Q, Zheng Q, Yang ZM. Mercury-induced oxidative stress and impact on antioxidant enzymes in *Chlamydomonas reinhardtii* [J]. *Ecotoxicology*, 2010, **19** (7): 1285-1293

6 Ledford HK, Chin BL, Niyogi KK. Acclimation to singlet oxygen stress in *Chlamydomonas reinhardtii* [J]. *Eukaryotic Cell*, 2007, **6** (6): 919-930

7 Wang H, Alvarez S, Hicks LM. Comprehensive comparison of iTRAQ

and label-free LC-based quantitative proteomics approaches using two *Chlamydomonas reinhardtii* strains of interest for biofuels engineering [J]. *J Proteome Res*, 2011, **11** (1): 487-501

8 Schenk PM, Thomas-Hall SR, Stephens E, Marx UC, Mussgnug JH, Posten C, Kruse O, Hankamer B. Second generation biofuels: high-efficiency microalgae for biodiesel production [J]. *Bioenergy Res*, 2008, **1** (1): 20-43

9 Zhao T, Li G, Mi S, Li S, Hannon GJ, Wang XJ, Qi Y. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii* [J]. *Gene Dev*, 2007, **21** (10): 1190-1203

10 Molnár A, Schwach F, Studholme DJ, Thuenemann EC, Baulcombe DC. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii* [J]. *Nature*, 2007, **447** (7148): 1126-1129

11 Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernández E, Fukuzawa H, González-Ballester D, González-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riaño-Pachón DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martínez D, Ngau WC, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS, Grossman AR. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions [J]. *Science*, 2007, **318** (5848): 245-250

12 郭强, 项安玲, 杨清, 邱承祥, 杨志敏. 利用EST及生物信息学方法挖掘马铃薯中miRNA及其靶基因[J]. 科学通报, 2007, **52** (14): 1656-1664

13 Zhang B, Pan X, Anderson TA. Identification of 188 conserved maize microRNAs and their targets [J]. *Febs Lett*, 2006, **580** (15): 3753-3762

14 Ritchie W, Rajasekhar M, Flamant S, Rasko JE. Conserved expression patterns predict microRNA targets [J]. *PLoS Comput Biol*, 2009, **5** (9): e1000513

15 叶可勇, 陈瑶, 李瑞梅, 符少萍, 郭建春. 小果野蕉microRNAs及其靶基因的生物信息学预测[J]. 热带生物学报, 2012, **3** (3): 222-227 [Ye KY, Chen Y, Li RM, Fu SP, Guo JC. Bioinformatic prediction of conserved microRNAs and their target genes in *Musa acuminata*. *J Trop Org*, 2012, **3** (3): 222-227]

16 Benman RB. Using RNAFOLD to predict the activity of small catalytic RNAs [J]. *Biotechniques*, 1993, **15** (6): 1090-1095

17 Jones-Rhoades MW, Bartel DP. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA [J]. *Mol Cell*, 2004, **14** (6): 787-799

18 Wu HJ, Ma YK, Chen T, Wang M, Wang XJ. PsRobot: a web-based plant small RNA meta-analysis toolbox [J]. *Nucleic Acids Res*, 2012, **40** (W1): W22-W28

19 Dai X, Zhao PX. psRNATarget: a plant small RNA target analysis server [J]. *Nucleic AcidsRes*, 2011, **39** (suppl 2): W155-W159