第七章 小 RNA 分析

内源性非蛋白质编码小 RNA (small non-protein-coding RNA, 12-24nt)广泛存在于高等和低等生物体内，通过对靶标 mRNA 直接切除或抑制其翻译在转录及转录后水平对基因表达起调节作用。已知的小 RNA 主要分为两大类：一类是微小 RNA (miRNA, microRNA)，一类是小干扰 RNA (siRNA, small interfering RNA)。在植物和动物体内，miRNA 与 siRNA 的产生机制和作用形式均有所不同，这里主要介绍植物体内的小 RNA。miRNA 是由具有发夹结构的初级转录本 （pri-miRNA）经过一系列加工过程，包括核酸内切酶 DCL1 加工后生成，而小干扰 RNA 则是通过核酸内切酶 DCL2, DCL3 和 DCL4 对具有较好互补结构的长双链 RNA 前体进行加工形成的 (Vazquez 2006)。目前发现的小干扰 RNA 种类很多，根据前体序列类型和形成机制可分为： ta-siRNAs (*trans* acting siRNAs)，nat-siRNAs (natural antisense transcript-derived siRNAs), hc-siRNA (heterochromatic siRNA), ra-siRNAs(repeat-associated siRNAs)，长茎环结构的 miRNA-like 位点（miRNA-like long hairpin）和 nat-miRNA (natural antisense miRNA)。植物中发现的小 RNA 已有相当的数量，在水稻中至今已鉴定出 451 个 miRNA (miRBase, http://microrna.sanger.ac.uk/sequences/, Release 14.0)、一个 ta-siRNA 家族 （*TAS3*）和一个 mirtron (Zhu et al. 2008)。

由于小 RNA 表达的时空特异性，导致传统的实验方法研究小 RNA 效率很低，成本较高，因此借助计算方法研究小 RNA 是一个很好的补充，大大加速了该领域的研究进程。对保守 miRNA 家族的查找，miRNA 基因簇的发现，基于 miRNA 序列特征预测特异（novel）miRNA，通过高通量测序技术（454 和 SOLEXIA）产生的小 RNA 数据 （往往超过几百或上千万条序列）处理，以及小 RNA 靶位点的预测及其进化分析，这些分析均离不开生物信息学的帮助。随着研究的深入，大量的计算方法，相关软件和小 RNA 数据库不断产生，本章将对相关内容进行介绍。

第一节 miRNA 的主要特征及计算识别

一. miRNA 的主要特征

在植物体内，miRNA基因首先通过Pol Ⅱ酶转录产生一个具茎环结构的miRNA初级转录本 (pri-miRNA) (Lee et al., 2004)，然后在DCL1酶 (Dicer-like enzyme)的作用下切除茎结构的尾巴或loop结构由miRNA前体 (pre-miRNA)得到miRNA:miRNA*双链复合体 (Tang et al., 2003; Kurihara and Watanabe, 2004)。miRNA:miRNA*复合体的两个3'端均有两个碱基的错位，其碱基结合允许一定的错配数，但通常不超过4个，并且没有较大的空位或loop结构。最后双链由解旋酶切开，miRNA*降解，成熟miRNA序列结合到靶基因位点进行调节， 根据与靶位点结合的紧密程度决定了对目标mRNA切割或是抑制其表达 (Bernstein et al., 2001; Papp et al., 2003; Bartel, 2004, 图1)。

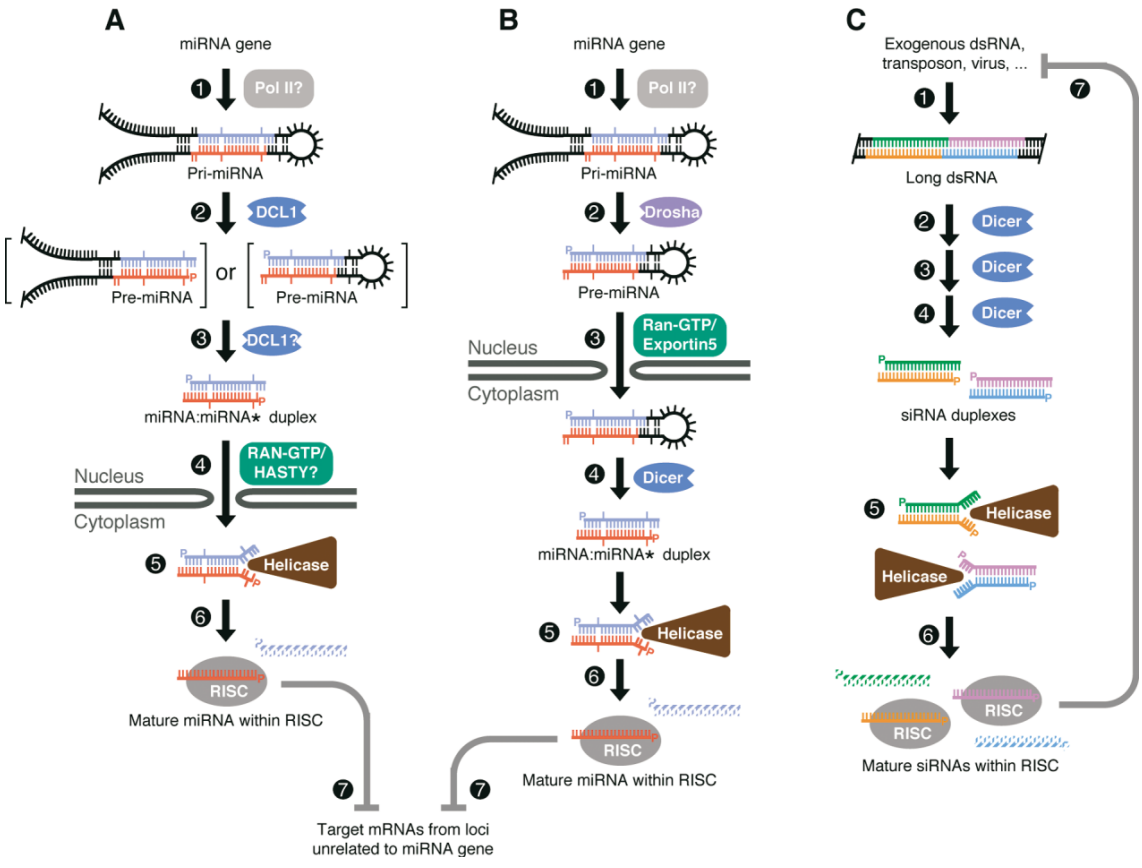


图1 miRNAs和siRNAs的产生途径 (Bartel, 2004)

(A) The biogenesis of a plant miRNA (steps 1–6; see text for details) and its hetero-silencing of loci unrelated to that from which it originated (step 7). The pre-miRNA intermediates (bracketed), thought to be very short-lived, have not been isolated in plants. The miRNA (red) is incorporated into the RISC (step 6), whereas the miRNA* (blue) is degraded (hatched segment). A monophosphate (P) marks the 5_ terminus of each fragment.

(B) The biogenesis of a metazoan miRNA (steps 1–6; see text for details) and its hetero-silencing of loci unrelated to that from which it originated (step 7).
(C) The biogenesis of animal siRNAs (steps 1–6; see text for details) and their auto-silencing of the same (or similar) loci from which they originated (step 7).

miRNA基因长度从几十到几百碱基不等 (Zhang et al., 2006a,b)，但成熟miRNA序列长度一般为20-24个碱基 (Ambros, 2001)，水稻中以21nt和24nt两种长度miRNA含量最丰富，这跟其选择的DCL酶有关。miRNA成簇排列的现象在动物中比较常见，在植物中目前已发现几个miRNA家族像水稻中的miR169，miR395，也在基因组上成簇排列 (Jones-Rhoades and Bartel, 2004; Zhang et al., 2006a)。成簇排列的miRNA类似多顺反子结构，基因表达模式和时期均有同步性 (Bartel, 2004; Altuvia et al., 2005; Baskerville and Bartel, 2005)。

基于miRNA前体的二级结构，一些研究发现miRNA前体有较低的最小折叠自由能 (MFE, minimal folding free energy)，由于MFE跟序列长度相关，Zhang等（2006b）提出了最小折叠自由能指标 (MFEI, minimal folding free energy index)的概念，将序列长度考虑进来，从而为不同长度miRNA前体的MEF比较提供了一个标准，并给出0.85作为miRNA区别于其他类型RNA的MFEI值，不失为一个预测miRNA的较理想指标。

$$\text{MEFI} = \frac{100 \times MEF / L}{(G+C)\%}$$

(L: the length of pre-miRNA)

目前miRBase 14.0 (http://www.mirbase.org/)版本中miRNA的记录已经超过1万条。其中很多miRNA家族均可以在至少2个物种中找到，其中miR159, miR171家族在目前miRBase收录的全部物种中均存在 (Tab. 1)。这种miRNA的保守性对于在新物种中预测保守的miRNA非常有用。尽管miRNA前体在不同物种，或不同成员间的变异非常大，但成熟miRNA序列还是相当保守的，同一miRNA家族不同物种的homologs间往往只有1, 2个碱基的差异。这种便利促使了大量的查找不同物种间保守miRNA的研究 (Llave et al., 2002; Reinhart et al., 2002; Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004; Sunkar and Zhu, 2004; Wang et al., 2004a; Adai et al., 2005; Sunkar et al., 2005; Zhang et al., 2005)。除了保守miRNA外，不同物种中还存在很多物种特异的miRNA (species-specific miRNA)，这类进化上比较“年轻”的miRNA无疑在特定物种的形成和发育过程中扮演着重要的作用。

表1. 植物保守miRNA家族（根据miRBase 14.0和物种多少排序）

miRNA family	No. of species	miRNA family	No. of species	miRNA family	No. of species
miR-159	17	miR-394	6	miR-1510	2
miR-171	17	miR-157	4	miR-1514	2
miR-156	16	miR-2118	4	miR-161	2
miR-166	16	miR-824	4	miR-2111	2
miR-167	16	miR-1507	3	miR-2275	2
miR-396	15	miR-2119	3	miR-413	2
miR-160	14	miR-403	3	miR-414	2
miR-399	14	miR-437	3	miR-415	2
miR-169	13	miR-444	3	miR-416	2
miR-172	13	miR-477	3	miR-417	2
miR-319	13	miR-529	3	miR-418	2
miR-408	12	miR-530	3	miR-419	2
miR-164	11	miR-535	3	miR-420	2
miR-168	11	miR-827	3	miR-426	2
miR-162	10	miR-1122	2	miR-472	2
miR-390	10	miR-1127	2	miR-479	2
miR-393	9	miR-1135	2	miR-783	2
miR-395	9	miR-1139	2	miR-821	2
miR-398	9	miR-1432	2	miR-828	2
miR-397	8	miR-1435	2	miR-845	2
miR-482	7	miR-1509	2		

miRNA通过与靶基因形成互补RNA双链来行使调节功能，这种互补性在进化过程中是保守的 (Rhoades et al., 2002; Jones-Rhoades and Bartel, 2004; Robins et al., 2005a)。互补性的强弱或者说互补碱基的多寡决定了miRNA调节的不同机制。跟靶基因有较好互补的miRNA主要通过对目标mRNA的直接切割调节mRNA的表达，相反，如果miRNA与其靶位点的错配较多，则主要通过转录后抑制的方式干扰mRNA的翻译 (Papp et al., 2003; Bartel, 2004, 图2)。植物miRNA的靶基因一大类都是转录因子 (transcriptional factor)，揭示了miRNA调节通路的复杂性。

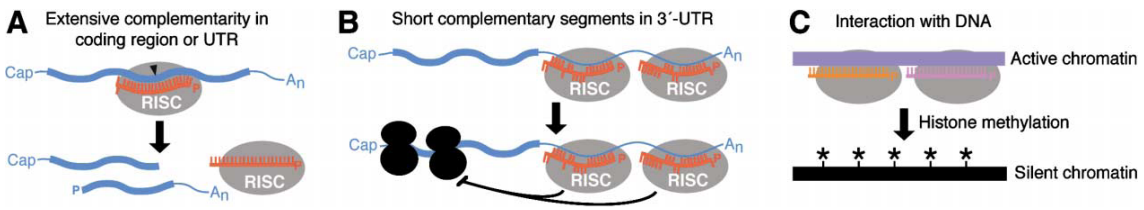


图 2 小RNA调控机制 (Bartel, 2004)

(A) Messenger RNA cleavage specified by a miRNA or siRNA. Black arrowhead indicates site of cleavage.

(B) Translational repression specified by miRNAs or siRNAs.

(C) Transcriptional silencing, thought to be specified by heterochromatic siRNAs.

二. miRNA的计算识别

通过计算方法识别miRNA基因主要基于以上提到的miRNA序列及结构上的特征，以及不同物种间的保守性。可以分为以下几类方法：

1. 同源比对

同源比对的方法主要是通过已知保守miRNA的在不同物种间的序列相似性进行同源序列搜索预测miRNA的方法。以已知miRNA序列为索引，公共DNA序列数据库中的数据作为搜索库，对于全基因组已测序或正在测序的模式生物，如rice，maize等，可利用其全基因组或大规模测序数据；对于基因组序列并未获得的物种来说，小规模的GSS (genome survey sequences)序列和EST (expressed sequence tags)序列也是很好的数据资源。尤其是EST序列，因为其本身就是表达水平的序列，故而预测的结果更加准确可信。搜索程序可以选择BLAST，如果是利用成熟miRNA序列进行搜索，因为序列较短，E值一般要高于1E-2，最小字符长度改为7 (默认13, -W 7)，但利用BLAST比对仍然会因程序本身的原因造成敏感性的降低，笔者在实际数据处理过程中曾发现对于~20nt的miRNA，2个不连续且距离较近的错配会导致错配序列3'端完全略掉联配过程，从而漏掉一个可能的结果，尽管这种情况是极少的。另外，基于轮廓的搜索软件ERPIN (http://tagc.univ-mrs.fr/erpin/)也可以用来搜索数据库中的miRNA同源基因位点。通过提交一组特定RNA的联配序列及二级结构信息，ERPIN可以搜索特定模式的RNA序列，从而获得更加准确特异的结果。同源比对方法还要注意以下几点：1) 数据处理过程中一般先通过BLASTX搜索蛋白质数据库，以排除掉编码蛋白序列，提高检索效率； 2) 往往仅找到已知miRNA的同源序列还远远不够，一般需要对候选miRNA位点周围的序列进行二级结构预测，以确定该段序列是否可能形成stem-loop结构，并需要验证miRNA的位置，及miRNA与miRNA*的互补情况；3) 在确定了可能的miRNA前体序列后，需要计算该段序列的MEF及MEFI值，一般情况下miRNA前体的MEF很小，而MEFI > 0.85，如果所有以上标准均符合，那么该位点即为候选的miRNA基因。

基于同源搜索方法开发了很多软件，包括Wang等（2005b）开发的miRAlign

软件 (http://bioinfo.au.tsinghua.edu.cn/miralign/) (图3), 可以用来预测人miRNA基因的基于概率共同学习模型开发的ProMiR (cbit.snu.ac.kr/~ProMiR2/) (Nam et al., 2005)，以及原理相似，用于植物miRNA预测的microHARVESTER (http://www-ab.informatik.uni-tuebingen.de/brisbane/tb/index.php?view=microharvester) (图4) (Dezulian et al., 2006)。

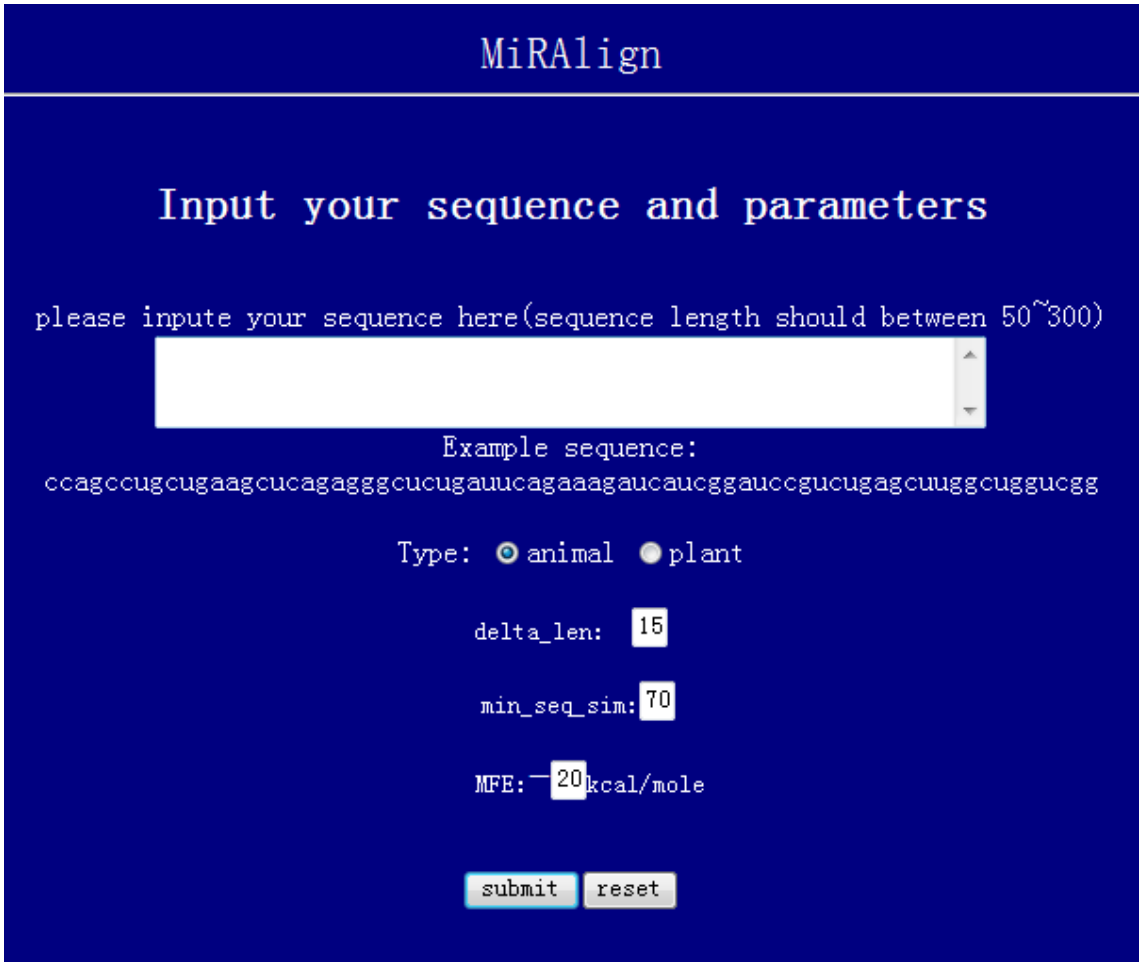


图 3. miRAlign界面 (http://bioinfo.au.tsinghua.edu.cn/miralign/)

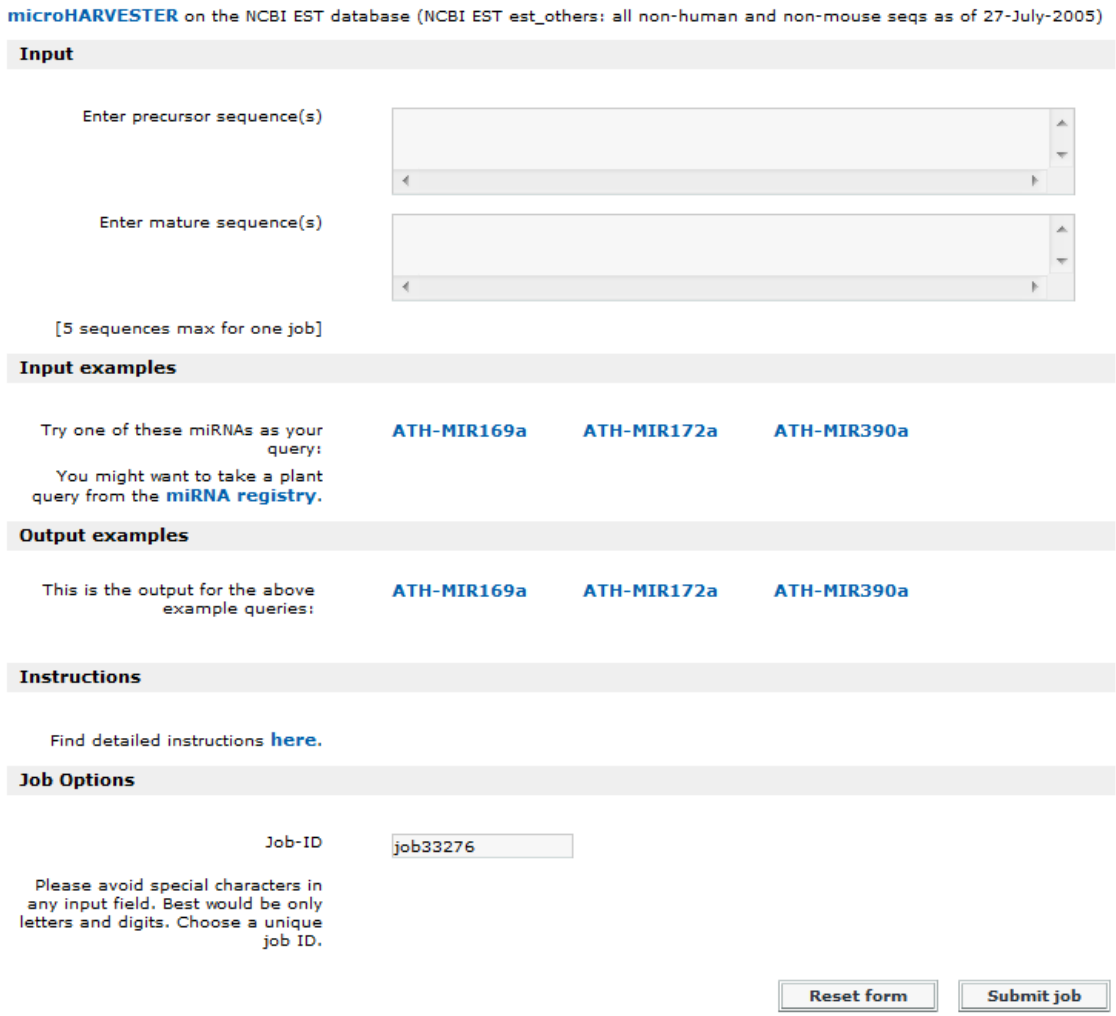


图 4. microHARVESTER界面

(http://www-ab.informatik.uni-tuebingen.de/brisbane/tb/index.php?view=microharvester)

2. 基因查找

基因查找方法可以不考虑miRNA的保守性，对整个基因组进行扫描，但只适用于动物miRNA基因的预测。首先根据不同物种的全基因组联配信息确定保守的非编码区，特别是启动子区及3' UTR区 (Xie et al., 2005a)，然后设定一个窗口大小比如110nt在该区域内滑动，利用二级结构预测软件比如Mfold (http://dinamelt.bioinfo.rpi.edu/download.php)或RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) (图5)对每条110nt长度的序列进行二级结构预测并打分，给出候选的miRNA基因。目前有两个基于该方法的软件成功预测了动物miRNA基因。一个是miRscan (http://genes.mit.edu/mirscan/) (图6)另一个是miRseeker (Lim et al., 2003b)。Lai等（2003）在果蝇基因组中的miRNA基因鉴定

工作表明，以已知的miRNA基因做参照，miRseeker的准确度和灵敏度为75% (18/24)，但是由于两种方法都是基于一定的窗口大小对保守区域进行扫描，因此该方法对于miRNA基因序列长度变化较大的植物miRNA预测来说并不适合。

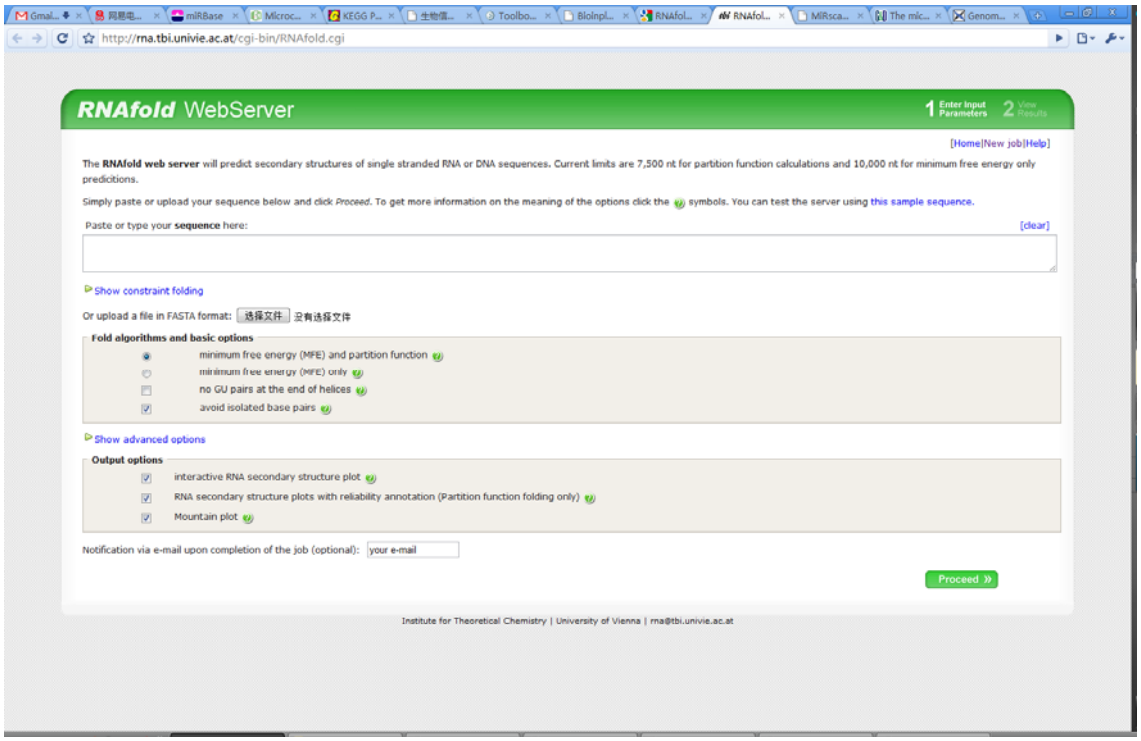


图 5. RNAfold 界面 (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi)

以 RNAfold 为例来说明二级结构预测软件的使用。RNAfold 是 Vienna RNA Package 的一系列用于二级结构预测和计算的工具之一。作为一个开源软件，Vienna RNA package 支持 Unix/Linux/Windows 多平台的版本下载 (http://www.tbi.univie.ac.at/RNA/)，每个软件均有详细的说明文档。上图是 RNAfold 的 web server 界面。将需要做二级结构预测的序列 (RNA/DNA)粘贴到文本框中，或将保存有 fasta 格式的文件提交到 server，选择相应的参数，如果不想在线等待结果，可以提供一个 email，在程序运行完毕后会将结果的链接发到邮箱里。点击 "Proceed"，最佳二级结构结果会在新窗口中显示，包括方便批量处理的“点-括号”格式结果，最小自由能值，以及图形化的结果，可保存为.eps 或.pdf 格式的文件。

图 6. MiRscan 界面 (http://genes.mit.edu/mirscan/)

3. 邻近茎环结构搜索

基于动物miRNA经常成簇存在于基因组上的特点，通过对已知miRNA附近区域进行茎环结构预测来发现成簇存在的miRNA。近期研究表明42%的人类miRNA基因和50%的果蝇miRNA基因都有成簇存在的现象 (Bartel, 2004; Altuvia et al., 2005)。由于植物miRNA成簇存在的现象比较少，只有miR169,miR395等几个家族存在成簇分布，因此该方法在植物miRNA预测方面存在局限性。

4. 基于比较基因组学的算法

基于比较基因组方法代表性研究是Jones-Rhoades和Bartel (2004)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。作者开发了MIRcheck软件 (http://web.wi.mit.edu/bartel/pub/software.html)通过计算一段序列是否存在理想的茎环结构，以及是否有20mers的短序列位于茎的位置上，然后根据其在两个物种中的保守性来查找保守的miRNA基因。Adai等 (2005)开发了findMiRNA (http://sundarlab.ucdavis.edu/mirna/)可以针对单个基因组来查找miRNA，findMiRNA主要依据miRNAs和其靶基因序列互补的保守性，然后利用二级结构预测软件对候选位点进行二级结构预测，找出有理想茎环结构的序列。需要注意的是，因为基

因组中很多类型的序列，如tRNA，逆转座子等元件均能形成发卡结构，因此在前期序列过滤和最终候选结果筛选方面要注意。

5. 基于大规模测序数据的发掘方法

从以上方法可以看出，大部分方法的理论基础都是miRNA的序列保守性，只有基因查找可以从miRNA的结构出发鉴定新的或物种特异的miRNA，但由于它是以一定长度为限制进行扫描，因而该方法对植物miRNA的预测并不适合。随着新一代测序技术如454和solexa技术的成熟和推广，大规模的基因组数据和RNA数据不断产生。针对miRNA的solexa测序每次都可以产生百万级数量的数据。在海量的数据面前仅仅通过前面介绍的传统方法显然不能满足研究的需要，如何有效的从这些海量数据中鉴定出miRNA基因变成了一个迫切而略带挑战的课题。以水稻方面的工作为例，最近发表了几篇大规模鉴定miRNA基因的文章。其中Zhu等 (2008) 以发育的水稻种子为材料最终鉴定了39个新的非保守的miRNA家族；Sunkar等 (2008) 以胁迫处理的水稻幼苗为材料鉴定了23个新的miRNA。虽然采用的计算方法略有不同，但都是基于miRNA序列和结构上的保守性进行预测。

下面以Zhu等 (2008)的工作为例说明一下大规模小RNA测序的数据处理流程。基于Solexa测序的原理，测序得到的原始读序都是一端连接了接头 (adaptor)的同一长度的序列，因此首先需要过滤掉接头和一些低质量的序列，这样得到了一个从十几个碱基到二十几个碱基不等的数据库。对于已有基因组数据的物种，比如水稻、拟南芥等，可以利用序列比对工具如BLAST将测得的小RNA匹配到基因组上 (>18nt)。这样我们就得到了一个全基因组的小RNA的分布图谱。根据全基因组的注释，排除掉匹配到重复序列区域和编码区的小RNA。这样一方面我们可以用上面介绍的方法来搜索保守的miRNA基因，另外，由于已知了小RNA序列和其位置信息，我们就可以利用一些新的标准来识别新的物种特异的miRNA基因。由于miRNA在产生过程中需要形成miRNA:miRNA*复合体，首先，根据小RNA的分布寻找候选的miRNA:miRNA*复合体。标准如下：1) 两条小RNA匹配到同一染色体的同一条链，且相距不超过400nt；2) 不允许有很多其他小RNA匹配到两条序列之间的区域（特别是有另外的小RNA跟其中一条部分配对，形成“拖尾”现象）；3) 每条小RNA在全基因组的匹配位置不能太多 （不超过10处）；4) 两条smallRNA的读序数需要相差5倍以上 （根据miRNA合成原理，miRNA*在与miRNA分开后会很快降解）。两条小RNA的配对也需要符合一定的标准 (Jones-Rhoades et al. 2006)：

1) 总共不超过7个碱基 (更严格的话可以设为4个碱基)的错配；2) 不超过3个碱基的连续错配；3) 不存在一条链上超过两个碱基错配而在另一条链上没有错配碱基的对应。满足以上条件的两条小RNA序列被当做候选的miRNA:miRNA*序列。从基因组上切下包含两条互补小RNA的序列作为候选的miRNA前体序列进行二级结构预测，根据其二级结构及两条序列所处的位置判断是否为候选的miRNA基因。

以上计算方法虽然提供了一种相对方便的鉴定miRNA的手段，而且目前大部分miRNA序列都是通过计算得方法预测出来的，但由于不同的预测方法都存在或多或少的缺陷或者假阳性，所以预测得到的候选miRNA基因仍然需要通过实验方法进行验证，包括直接克隆，Northern, PCR, 5'-RACE (5' rapid amplification of cDNA ends) (Griffiths-Jones, 2004; Griffiths-Jones et al., 2006)。

三. miRNA靶基因的预测

不像动物miRNA结合靶基因的机制那么复杂，植物miRNA主要通过接近完美的互补配对结合到靶位点，从而引发对目标mRNA的直接切割。植物miRNA和靶位点的结合有如下特征：1) 一般不超过3个碱基的错配；2) 5'端前10个碱基结合很紧密，一般只允许1个碱基的错配；3) 5'端第1，11，12个碱基因为剪切功能的关系一般不允许有错配；4) 一般没有连续的错配 (>=3个)出现。动物miRNA靶基因的预测根据结合的不同特点已经开发了很多的软件，从miRanda，TargetScan，Pictar到microTar等，但由于植物miRNA识别靶位点的模式较为简单，所以植物miRNA靶位点的预测软件相对较少，其中miRU (http://bioinfo3.noble.org/miRNA/miRU.htm)是一个网络平台，整合了已知的大部分植物mRNA和gene数据，可提供候选的小RNA，在提供的植物表达数据中预测是否有靶位点（图7）。miRU有几个参数可供设置：一是阀值，即总罚分为3分，根据不同错配类型，罚分不同；二是G:U配对，一般罚0.5分，三是INDEL，一般不超过2个，四是其他类型，即错配，总共不超过3个。然后选择需要预测的靶基因数据库，即Database1，另外还有一个Database2，是预测保守miRNA靶位点提供的参照物种，可以降低预测的假阳性。另外，Zhao et al. 又在miRU的基础上开发了psRNATarget，不仅可以提供小RNA在其植物基因数据库中预测靶位点，还可以提供自己特定的基因数据(< 70Mb)检验是否存在已知的miRNA的靶基因，另外，最灵活的服务是你可以提供特定的小RNA以及特定的植物基因数据，进行完全个性化的靶基因预测，当然你的基因数据大小有一定的限制 (<70Mb)。

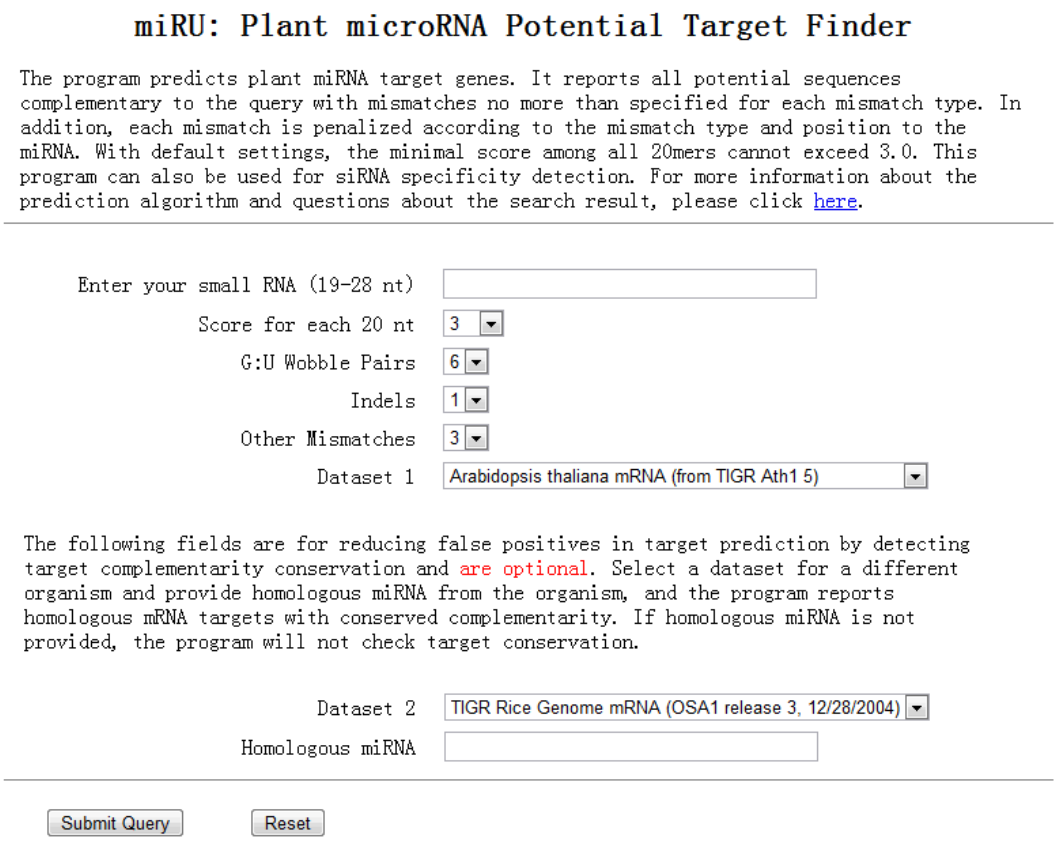


图 7. miRU界面 (http://bioinfo3.noble.org/miRNA/miRU.htm)

patScan是另一个可以方便进行miRNA靶基因预测的软件。patScan提供了Unix/Linux/Windows版本可在 http://iubio.bio.indiana.edu/soft/molbio/pattern/下载。patScan最初的设计是用来查找基因组特定模式的序列，Rhoades et al. (2002)首先将patScan用于miRNA靶基因的预测，并评估了这种预测方式的假阳性 (Rhoades et al., 2002; 图8)。patScan的运行需要调用两个文件，一是指定搜索的pattern文件，由相应的smallRNA序列和匹配模式组成：smallRNA_sequence[4,0,0]；另一个是用来预测的基因序列文件，Fasta格式，标题按照相应的序列类型标示为“>title|CDS ..”或“>title|cDNA ..”等等。smallRNA与靶位点的匹配标准如前所述。另外，前面提到的MIRcheck和findMiRNA软件由于在预测miRNA时需要考虑miRNA和其靶位点的保守性，故而也可用来预测miRNA靶位点。

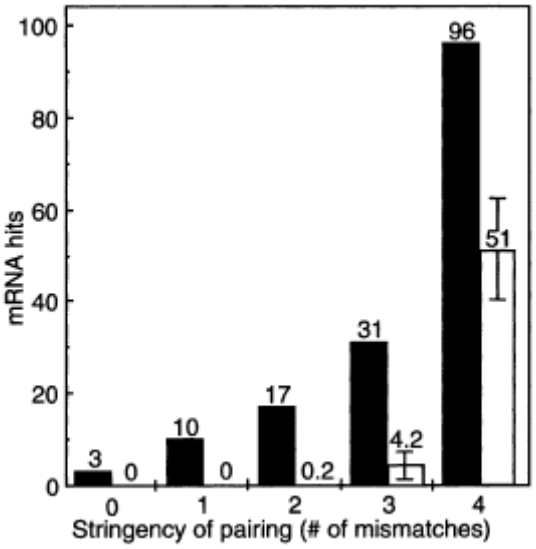


图 8. *Arabidopsis* miRNA与其mRNA的反义匹配情况 (Rhoades et al. 2002)
Annotated Arabidopsis mRNAs were searched for sites complementary to 16 Arabidopsis miRNAs with 0–4 mismatches (solid bars). Identical searches with cohorts of 16 randomized RNAs were also performed (open bars, mean values from ten cohorts; error bars, one standard deviation). Note that two hits by similar miRNAs to the same complementary site within an mRNA were counted as separate hits (Table 1).

第二节 ta-siRNA等的计算识别

一. ta-siRNA的主要特征

与miRNA不同，siRNA主要通过长的双链RNA复合体在DCL酶的切割下产生。植物体演化出几种截然不同的siRNAs，它们在产生机制和调节通路的功能方面都有所不同 (Brodersen and Voinnet, 2006; Vaucheret, 2006)。其中大部分的siRNA类型(24nt)在依赖RNA的RNA聚合酶2 (RDR2)、DCL3、PolⅣ的作用下产生，并通过AGO4引导的DNA甲基化或组蛋白修饰诱导转录沉默 (Zilberman et al., 2003, 2004; Chan et al., 2004; Xie et al., 2004; Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005; Pontier et al., 2005; Tran et al., 2005)。这一代谢通路往往跟转座子、反转座因子等重复序列相关 (Xie et al., 2004; Lu et al., 2006; Rajagopalan et al., 2006; Kasschau et al., 2007)。其他类型的siRNA主要在转录后水平起作用。对病毒RNA和转基因转录本的沉默涉及到依赖RDR6/DCL4的siRNA (21nt)或依赖DCL2的siRNA (22nt)。ta-siRNA就是通过RDR6/DCL4通路产生的。tasiRNA的形成主要是通过miRNA介导的按21nt相位排列的siRNA的剪切 (<=12 phases)。不同的TAS家族受不同的miRNA调节，TAS1和TAS2受miR173的调节，TAS3在拟南芥和水稻中保守，受miR390调节，且有5'端和3'端两个结合位点，TAS4受miR828调节。TAS基因的dsRNA前体在DCL4作用下，由相应的miRNA起始剪切，产生21nt，3'端有两个碱

基错位的双链siRNA复合体 (Dunoyer et al., 2005; Gasciolli et al., 2005; Xie et al., 2005)。不同TAS家族切割产生的siRNA数目不同，其中只有特定的一两个siRNA行使功能。根据以上特征可以通过生物信息学的方法预测tasiRNA。

二. ta-siRNA的计算识别

1. Howell算法

前面提到全基因组序列已测序的物种产生了大量的小RNA的数据，而且这些不同组织或处理下测得的小RNA可以很好的定位到全基因组上。根据一段区域(<300nt)内小RNA是否按照21nt的位移排列这一显著特征，可以找出候选的TAS基因位点。Howell 等(2007) (图9)设计了一套流程用来查找拟南芥中的候选tasiRNA，首先将定位到基因组正反链的小RNA序列合并，将来自不同链的小RNA定位位置抵消掉2个碱基，这样来自一对复合体的正反链小RNA位置可以在计算的时候累加。然后引入P值作为评价步移的参数。P值的计算如下：

$$P = \ln[(1+\sum_{i=1}^{8} k_i)^{n-2}], P > 0,$$

如果一个相位长度设为21nt，n表示在8个相位大小的窗口范围内至少有一个小RNA定位到相位上的相位循环数（即n个相位位置上有小RNA存在）；k表示在调查的这8个相位大小的窗口里面正负链合并过的起点位置刚好位于相位上的小RNA读序总和；由于指数n-2的限定，只有当至少连续三个相位上 (n>=3)都存在至少一个小RNA才能保证P为正值。由公式可以看出，P值受小RNA丰度和所处位置的双面影响。P值的计算按单碱基的步长在基因组上滑动，计算得到的P值分配给该点四个相位距离的位置。因此，可以将小RNA在基因组上的实际分布，如图9 A中READS图所示，转化为P值分布的PHASE图，具有显著高P值的位点被选为候选的phase位点。最后，根据ta-siRNA受相应miRNA调控的现象，在预测到的phase区域两端预测miRNA靶位点，如果可以找到相应的结合位点，那么这段区域可被认为是tasiRNA-like位点。

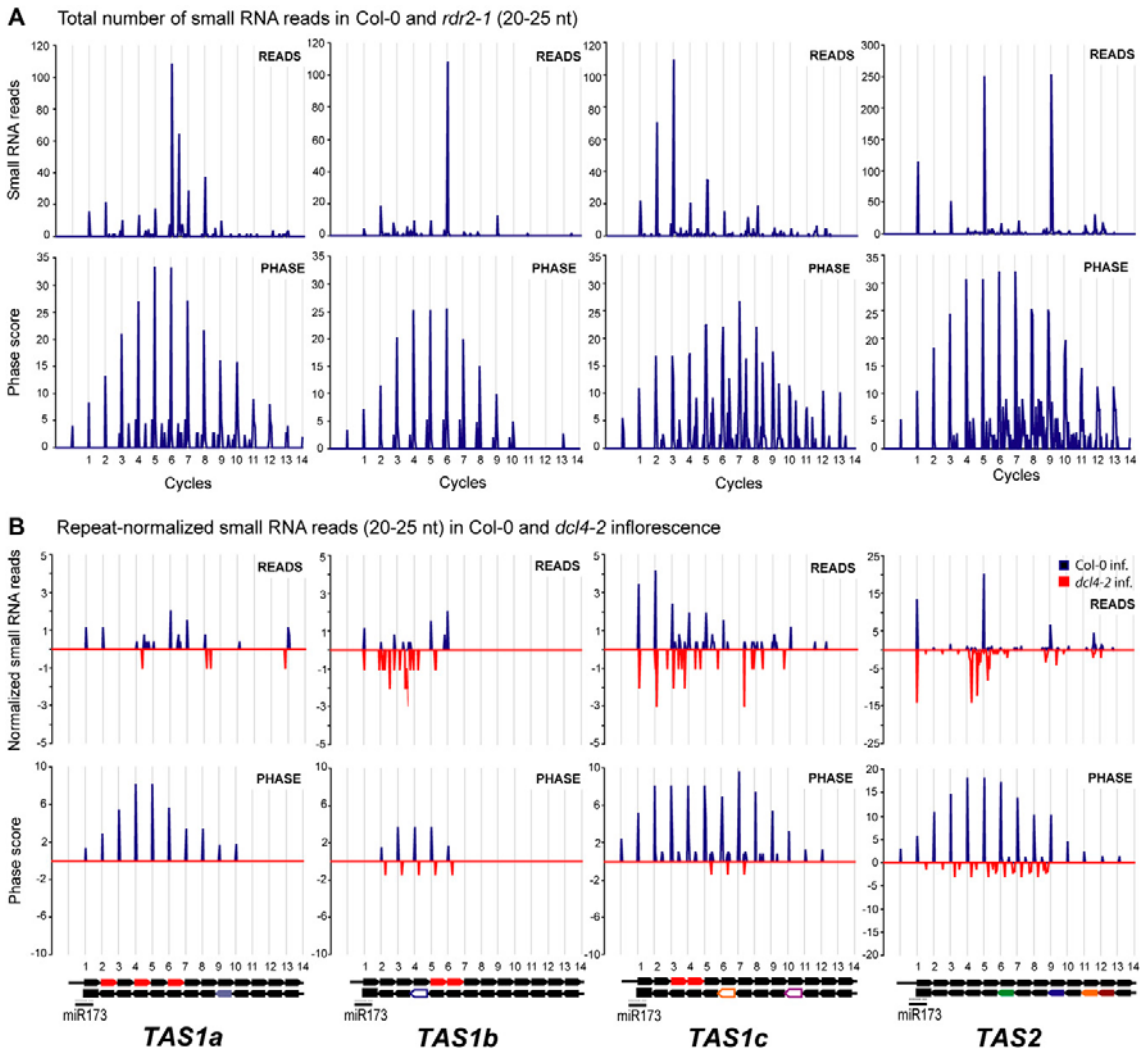


图 9. 拟南芥TAS1a, TAS1b, TAS1c和TAS2位点21nt小RNA分布及数量 (Howell et al. 2007)

2. Chen算法

与Howell的方法类似，Chen等(2007)的方法也是主要考虑tasiRNA的相位分布特征，并构建了一个 P值来查找候选的tasiRNA位点。按照21nt一个相位大小，考虑11个相位长度的一段区域，n表示位于该231bp区间的小RNA读序数；k表示位于该231bp区间相位位置上的小RNA读序数。P 值越大，表示相位（phase）结构越明显。Chen 等提供了相应的perl脚本用于计算P value，可以在其文章附件信息中找到。

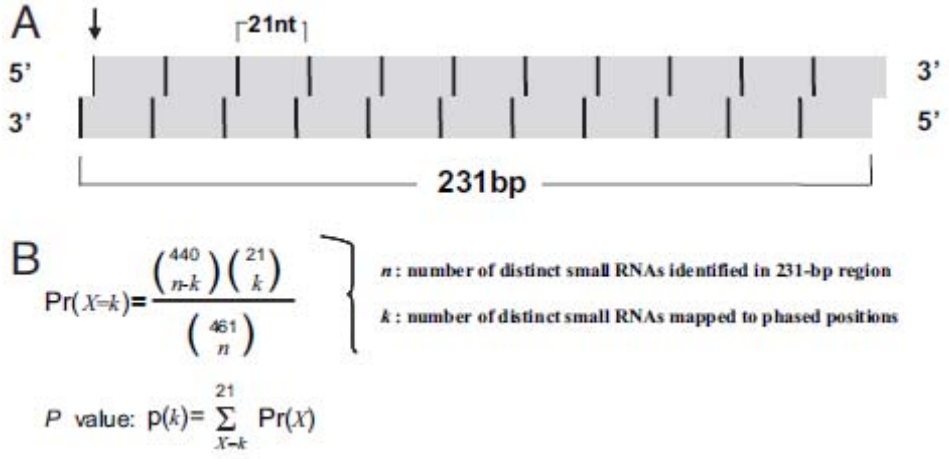


图10 TAS预测算法原理 (Chen et al. 2007)
(A) The vertical arrow indicates the start site for the small RNA used to determine the phased and nonphased positions. 21 phased sites relative to the start site are indicated as black vertical bars. Four hundred forty nonphased sites relative to the start site are indicated as gray. (B) Equation based on hypergeometric distribution for statistically evaluating the presence of phased siRNA in genomic fragment defined in A.

三. 起源于NAT的siRNA

Natural Antisense Transcripts (NAT)是指可以跟其他转录本互补形成RNA双链的编码或非编码RNA序列。根据它们在基因组上的相对位置不同，NAT可以分为两类：*cis*-NAT和*trans*-NAT。c*is*-NAT是指来自于跟有义链转录本同一个基因组座位不同染色体链的序列；*trans*-NAT是指跟它的互补序列来自于染色体上的不同位置的转录本。研究表明哺乳动物和植物中大约5%~10%的基因转录本都存在*cis*-NATs。Osato等(2003)从水稻中预测了687组NAT；Wang等(2006)从拟南芥中预测了1320个*trans*-NAT。起源于NATs位点的siRNA称为NAT-siRNA，主要介导转录后沉默。起源于NATs双链RNA复合体的小RNA称作NAT-siRNA，第一个NAT-siRNA是2005年从拟南芥中鉴定出来的，来自P5CDH和SRO5基因转录物形成的dsRNA。目前已有若干大规模鉴定NAT-siRNA的工作在拟南芥和水稻中开展，并发现了许多有意思的结果。其中包括*cis*-NAT-siRNA5'端第一个碱基的偏好性，由于AGO2和AGO4参与该类小RNA的结合，故而第一个碱基常常为腺嘌呤 (A)。对*trans*-NAT的GO分类研究表明，催化活性、信号传感器、转运蛋白活性相关的转录本占很大比例。另外，对NATs结构的功能研究表明植物基因组中的NATs结构可能对逆境胁迫方面起重要作用。

另外还有几种其他类型的siRNA，比如首先从水稻中发现的NAT-miRNA，其长约20nt。前体hpRNA序列两条链分别转录、剪接，反义链RNA产生miRNA，调节正义链mRNA的表达。NAT-miRNA既不同于普通miRNA，因为普通miRNA的前

体hpRNA无需剪接；也不同于NATs-siRNA，后者的序列多来自两条链，而nat-miRNA几乎都是由一条RNA链产生；另外，NATs -siRNA形成需要DCL2，而NAT-miRNA需要DCL1。Zhu等(2008)在水稻中发现了一类miRNA-like long hairpin位点。这类小RNA基因可以像普通miRNA那样形成长的发卡结构，但是有很大的loop环，其茎结构又跟tasi-RNA类似，在双链上有21nt的phase结构。

四. siRNA 靶基因预测

尽管siRNA有着丰富的类型，但其行使功能还是通过与靶基因位点的序列互补来实现 (图1)。因此，miRNA靶基因的预测软件也同样适用于siRNA的靶基因预测。值得注意的是，已有的研究表明，特定类型的siRNA靶基因也有着显著的区别。比如*TAS3*的靶基因是一类大的基因家族，称做激素响应因子 (ARF)。拟南芥中发现的NAT-siRNA被认为与植物的抗逆境代谢有关。

第三节 小 RNA 的进化分析

一. 小 RNA 进化研究概况

作为一类重要的调控小分子，miRNA 在大多数真核生物（Finnegan and Matzke, 2003）甚至是病毒（Sullivan et al. 2005）中通过 RNA 干扰机制调节各种代谢途径。植物中许多编码 miRNA 的基因起源于单双子叶植物分化之前（约 150 百万年前），动物中的 miRNA 编码基因也早于多细胞动物分化的时间（约 600 百万年前）。然而，目前还没有发现动植物中 miRNA 编码基因或靶基因的同源基因。这就提出一个进化上的有趣的问题：这些编码 miRNA 的基因是怎么形成的呢？

Allen 等（2004）通过对两个拟南芥特异 miRNA 家族的研究揭示了 miRNA 编码基因与其靶基因共同进化的一个可能的机制。由于 miR161/163 两个家族都是新产生的年轻 miRNA 编码基因，而且跟大多数保守的 miRNA 家族不同，miR161/163 均位于其靶基因的附近，因此 Allen 等认为 miRNA 家族有可能通过基因家族扩增过程中的倒转复制或反向倍增机制（inverted duplication）产生。如图 11 所示，基因家族在扩增过程中由于倒转复制产生头对头或尾对尾的全部或部分基因复制片段，从而为形成 miRNA 编码的发卡结构提供了可能。倒转复制可能直接从基因组上发生也可能通过逆转录后结合类似假基因序列形成。甚至一个基因家族相近的成员间的结合也可以产生这样的创始基因（founder gene）。新形成的位点转录得

到的具有发卡结构的转录本有可能称为 DCL 的靶标而导致 siRNA 的产生，从而使创始基因及其相关的家族成员在转录后水平或染色质水平受到 RNA 干扰机制的调控。部分创始基因在分化过程中因维持发卡结构以及被 DCL 的识别的功能限制，形成一类特异的 siRNA 家族（步骤 2）。而对 DCL1 调控的代谢途径的适应性进化导致了 miRNA 基因的形成（步骤 3）。由于变异的持续积累，部分基因在发卡结构和 DCL1 识别功能限制下，只剩下 miRNA 及其互补的 miRNA*一段与原始的序列相似（步骤 4）。miRNA 座位的复制导致了 miRNA 家族其他成员的产生（步骤 5），并由于变异的积累导致不同成员拥有了各自特异的 miRNA 靶基因。结合 miRNA 靶基因家族的进化使该模型变得更加完整。大多数 miRNA 的靶基因都是一大类基因家族中的亚类。靶基因家族的复制（步骤 6）为调控的多样化提供了基础。在一个新的 siRNA 或 miRNA 编码基因形成后（步骤 2 或 3），家族成员中小 RNA 结合位点的保留（步骤 7）或丢失（步骤 8a）导致了转录后水平调控的分化。同时也许还伴随着转录调控因子的改变（步骤 8b），导致了进一步的调控机制的差异。miRNA 靶基因随后的复制和分化事件（步骤 9）致使不同 miRNA 家族不同成员间拥有了各自专一的靶位点及调控功能。这样，通过 miRNA 和靶基因之间的复制事件，以及结合位点的保留或丢失而形成了一个新的调控网络。

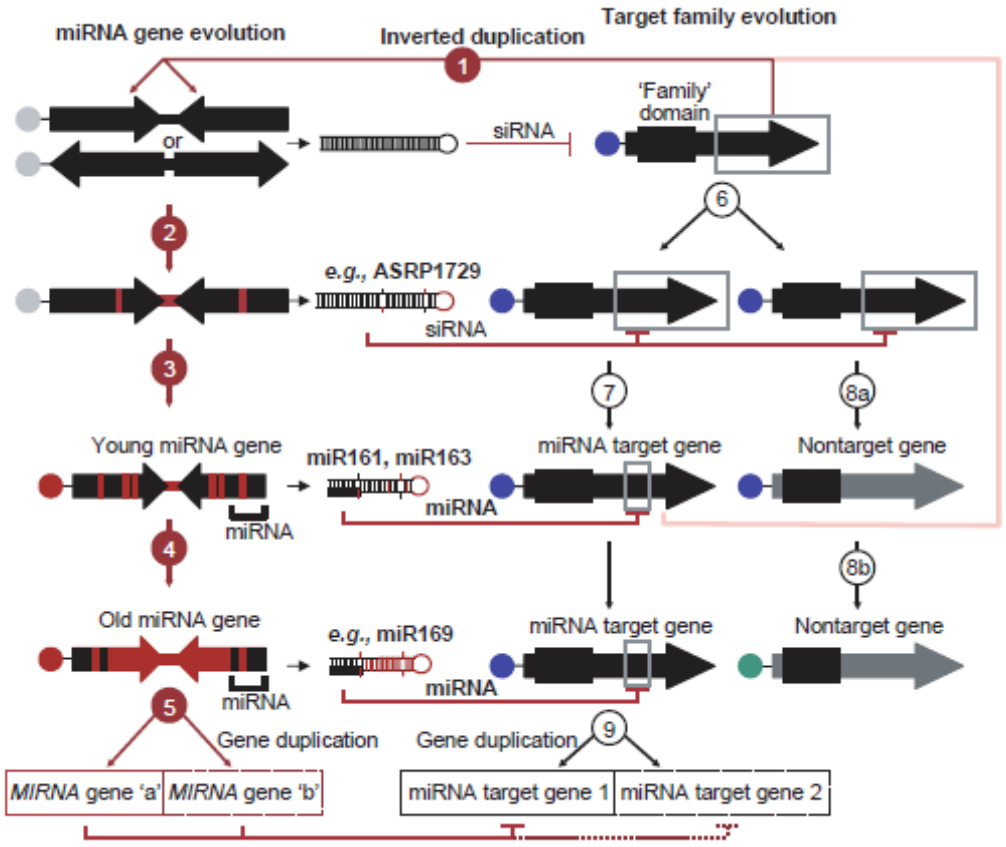


图 11. 植物miRNA反向倍增进化模型(Allen et al. 2004)

然而这样的模型也有很大的局限性。考虑到保守miRNA基因与其靶基因间在结合位点外并没有这种序列相似性的证据存在，对于保守miRNA的解释仍然有待进一步的验证。同样，由于动物miRNA前体序列较短，也不能提供创始基因的信息。一般认为动物miRNA调节机制是通过miRNA和其靶位点间“交互作用获得”事件形成的。跟植物miRNA与靶基因间严格匹配，切割靶基因转录本不同，动物miRNA通过结合到编码基因的3’端干扰其翻译来行使调节作用，并允许其与结合位点间有较多的碱基错配(Bartel et al. 2004)。这一功能模式的不同也表明在动植物miRNA编码基因起源机制上也存在着差异（Li and Mao, 2007）。

对于拟南芥miRNA基因的研究表明，通过上述具有回文结构位点产生的miRNA有几种不同的命运（Fig. 12）：第一，起源于原始基因家族的小RNA保留了调节该基因的能力；第二，小RNA通过遗传漂变获得了特异结合到其他基因或基因家族的能力，很明显，以上两种结果均表明选择作用的存在。第三，也可能是最普遍的命运，随着小RNA产生位点启动子区域、回文结构区域和靶基因结合位点突变的积累而丢失了调节靶基因的能力。因此，植物小RNA的产生机制为研究

特定的调节元件的进化提供了很好的机会（Chapman and Carrington, 2007）。

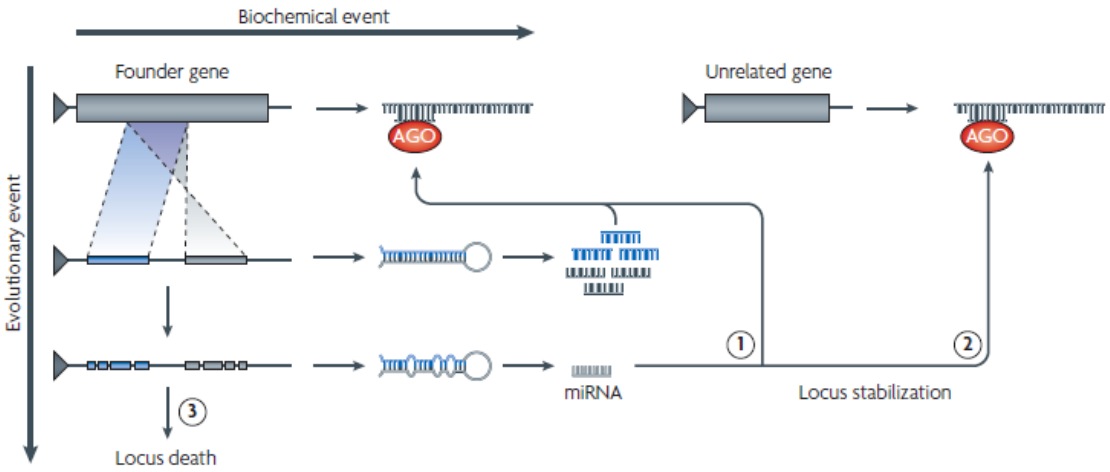


图 12 植物新 miRNA 基因进化模型 (Chapman and Carrington, 2007).

二. 水稻小 RNA 的进化分析

遗传学方面近几年的一个重要的研究进展是在动植物基因组中发现了大量小 RNA 等非蛋白质编码基因，这些小基因（一般 100-200bp）在生理生化等代谢过程中起到重要作用。由此产生一个有待回答的问题：对于水稻等作物中发现的编码小 RNA 的这些基因位点在我们人类进行作物驯化和育种过程中是否同样受到选择（参见第八章）？我们目前在研究作物骨干亲本遗传成因中是否和如何考虑这些基因对骨干亲本形成的影响？目前发现的人工选择（育种）的基因位点主要编码转录调节因子和其他蛋白质编码基因，我们的研究发现非蛋白质编码基因在人工驯化过程中同样受到人工选择效应的影响。我们利用水稻为模式作物，发现小 RNA 之一，miRNA 基因 *MIR156b/c* 基因位点可能受到强烈的自然和人工选择效应的影响，说明人工选择的对象除了转录因子及其下游基因外，还可能针对转录因子调控（上游）基因（Wang et al, 2007）。

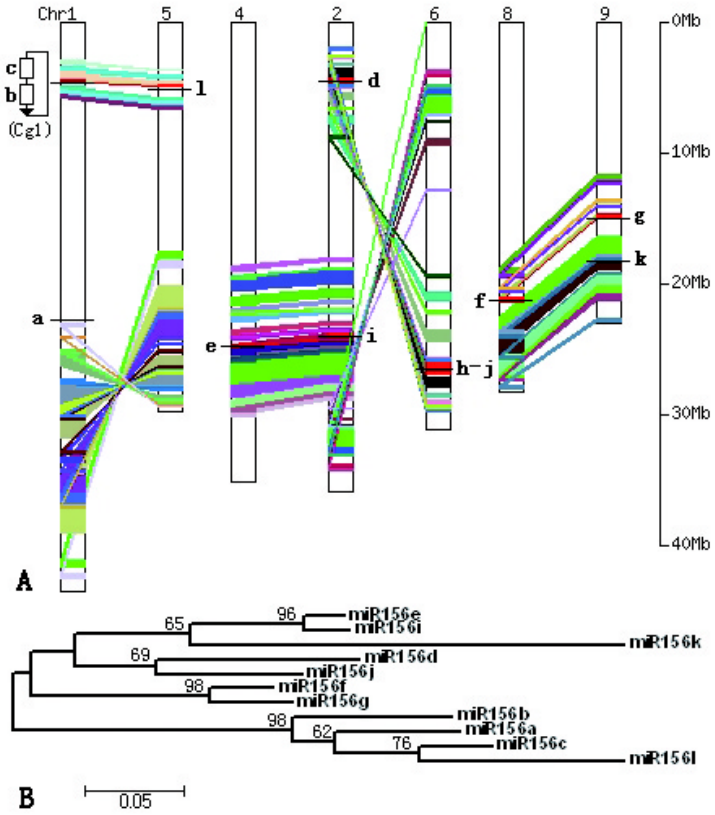


图 13. 水稻 miR156 家族在基因组上的分布和系统进化关系 (Wang et al. 2007)

通过水稻 miRNA 及其靶基因结合位点序列变异的调查和直系同源基因（Paralogs）分析，发现水稻 miRNA 基因在不断地捕获新的结合位点（靶基因），同时也不断丢失对靶基因的调控功能（Guo et al, 2008b）。这种动态的进化过程主要通过 miRNA 序列突变来实现，同时插入和删除也发挥一定作用。图 14 展示了水稻 miR397 靶基因在全基因组前后的突变进化情况，有些靶基因位点由于序列突变而脱离了 miR397 的绑定和调控。

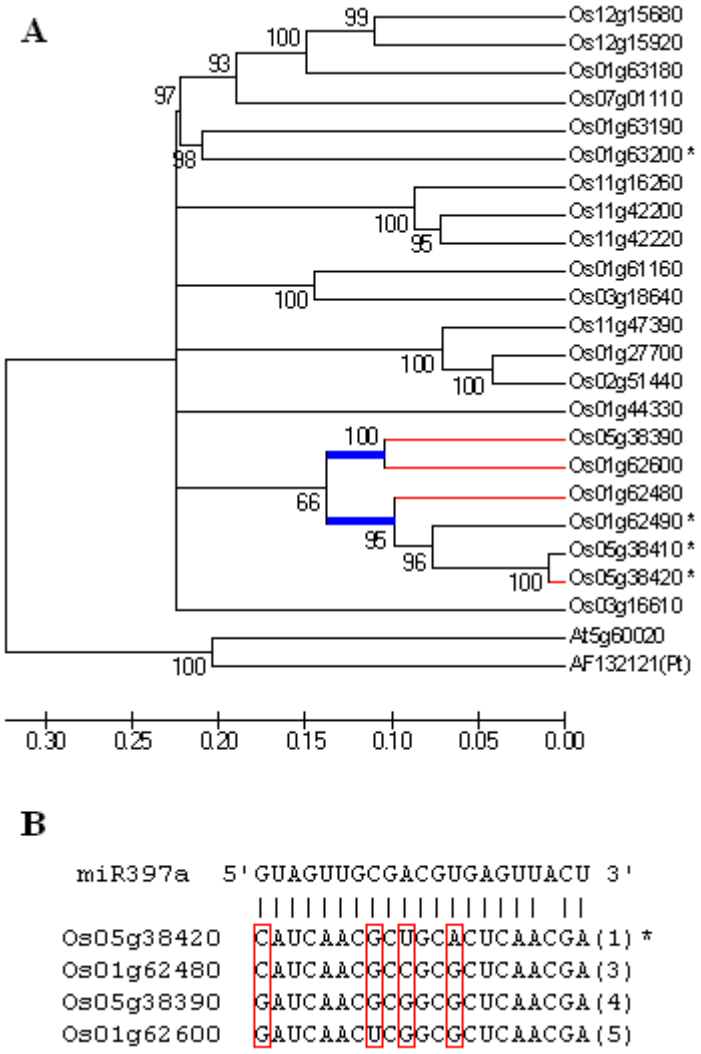


图 14. 水稻 miR397 靶基因进化（A）及其结合位点的序列突变情况（B）(Guo et al. 2008b)

ta-siRNA (trans acting siRNAs)是植物中发现的一类 siRNA 基因(*TAS*)，其在 miR390 等的辅助下，调控生长素相关基因 ARF(auxin response factor)，在植物生长发育过程中发挥重要调控功能。目前已在拟南芥中发现四个亚家族（*TAS1-4*）,其中 *TAS3* 在植物界是保守的。在水稻上，我们通过 Howell 等（2007）和 Chen 等（2007）方面找到了 4 个 *TAS3* 基因（Zhu et al. 2008）。其中部分 21nt 长度读序的 Howell 分布图见图 15。

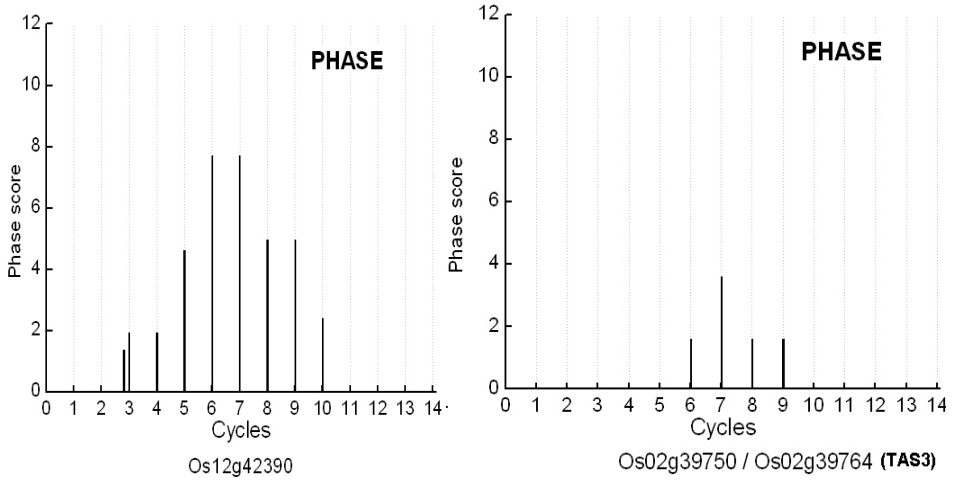
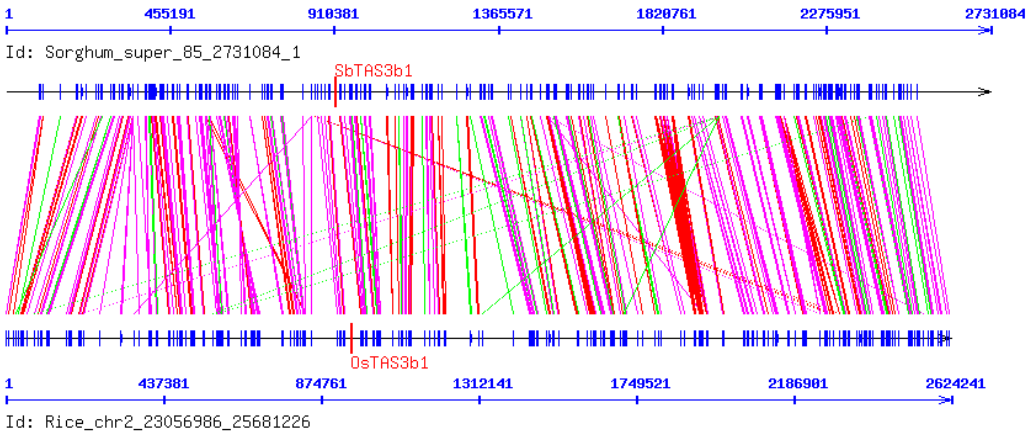


图 15 水稻 *TAS3* 基因 21nt 小 RNA 读序的相位值分布图（Zhu et al. 2008）

A



B

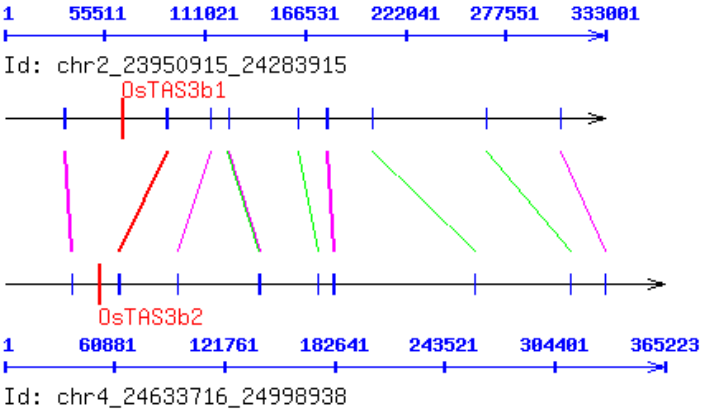


图 16 水稻 *TAS3* 基因倍增及其与高粱同源基因的比较基因组学分析 (Shen et al. 2009)

我们又通过 TAS3 基因的保守序列片段，克隆测序和生物信息学方法发现了 51 个来自禾本科的 *TAS3* 基因（Shen et al, 2009）。通过序列比较等，发现 *TAS3* 基因通过基因组和单基因倍增，在禾本科基因组中至少有 2 个拷贝，多的可达到近 10 个。水稻基因组倍增而来的 TAS3 基因在基因组保持了其共线性关系；同时 TAS3 在不同禾本科基因组上也存在明显的基因组共线性（图 16）。

三. 水稻 miRNA 位点遗传多样性与驯化选择研究

Ehrenreich 和 Purugganan（2008）对拟南芥 miRNA 编码基因及其靶基因的序列变异情况作了大规模调查。通过对 16 个 miRNA 家族 66 个成员及其对应的 52 个靶基因位点的群体数据的分析，表明成熟 miRNA 位点相对于其上下游序列有更高的保守性，并通过中性检验检测到了可能经受选择压力的 miRNA 位点（MiR166f, miR167d, and miR395c）。

为了调查模式作物—水稻中 miRNA 是否经受人工选择即驯化的影响。我们对水稻 miRNA 进行了大规模的群体调查。对 40 个 miRNA 家族的 97 个成员位点进行了重测序，包括了 30 个水稻籼粳亚种的材料。结果表明，与拟南芥的群体调查结果一致，在 miRNA 成熟位点其核苷酸多态性明显低于两端序列，暗示了 miRNA 通过序列互补结合靶基因功能限制的存在。同时，对于保守的 miRNA 家族，其整体的 DNA 多态性相较水稻特异的 miRNA 来说要低一倍，由于保守 miRNA 一般参与基础的代谢网络的调控，因而有可能遭受更强的净化选择而保持序列的保守性(Wang et al. 2010)。

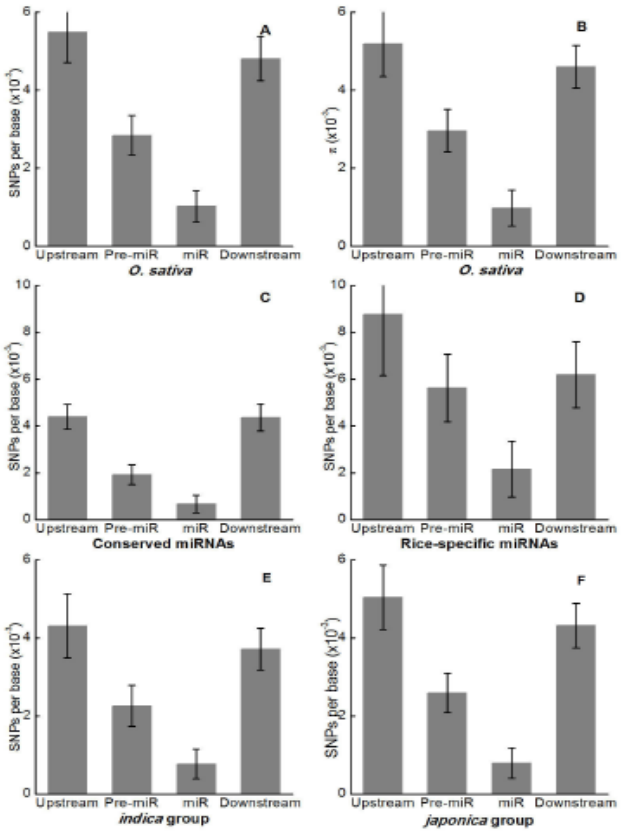


图 16. 水稻 miRNA 位点的序列多态性（Wang et al. 2010）。

另外，我们还对 Tajima's *D* 检验显著的 miRNA 位点进行了进一步的正向选择信号的调查。对相应的 miRNA 位点普通野生稻群体（*O.rufipogon*）进行重测序用于中性检验等分析，结合 Tajima's *D* 检验、HKA 检验的结果，我们找到了几个 miRNA 位点在驯化过程中可能经历了正向选择作用。以 miR390 为例，其调控基因为另一类小 RNA，*TAS3*，中性检验的信号表明，miR390 可能由于选择作用的影响而维持了其特异的调控作用。

表 2. 驯化选择候选 miRNA 位点的 DNA 多态性和中性检验结果（Wang et al. 2010）。

Small RNA	*O. sativa*												*O. rufipogon*						Candidate Status	Target
	indica subgroup						*japonica* subgroup													
	N	*L*	*S*	π	*D*	HKA	*N*	*L*	*S*	π	*D*	HKA	*N*	*L*	*S*	π	*D*	HKA		
MIR390	25	723	2	0. 32	-1.2137	0.1064	20	741	1	0. 36	-0.0861	0.0138	11	741	7	2.25	-1.2177	0.2568	**Selected**	TAS3
MIR395a-b	27	418	11	3.59	-1.558	NA	21	646	13	2.56	-1.9396*	NA	11	594	8	2.93	-1.4933	NA	Putative	APS/AST
Os12g42390	25	645	2	0. 91	0.2607	0.0034	18	644	8	1.77	-1.7663	0.0546	11	645	34	10.06	-2.0541*	0.2498	Putative	Unknown
TAS3a2	28	765	33	5.49	-1.8578*	0.0288	24	763	21	5.66	-0.8475	0.158	10	764	12	3.89	-1.3459	0.5026	Putative	ARF
MIR166f	29	515	9	2.99	-1.0247	0.1610	23	515	10	3.99	-0.814	0.3416	10	515	24	15.79	-0.1961	0.4192	Not	HD-ZIP Ⅲ
MIR395i-k	25	651	8	4.26	0.9811	0.1058	18	687	6	2.14	-0.5168	0.2478	9	687	8	3.71	-0.5981	0.2446	Not	APS/AST
MIR440	25	619	4	1.73	0.0357	0.0966	20	621	4	1.45	-0.5705	0.137	14	621	19	5.96	-1.5827	0.3386	Not	Unknown
MIR443	29	531	9	4.51	0.1466	0.1482	21	530	20	6.10	-1.5683	0.2736	9	529	12	7.66	-0.3859	0.3006	Not	Unknown

第四节 小 RNA 数据库

一. miRBase 数据库

作为目前最权威和完整的 miRNA 数据库 (http://mirdb.org/miRDB/)，截止到目前 (2009 年 11 月)，miRBase 已经收录了一百余个物种中超过 10000 条的 miRNA 记录 (图 17)。其中来自植物体的 miRNA 序列有 1834 条。数据库主要由 3 部分组成：miRBase:Registry，主要是用于提交新的 miRNA 序列；miRBase:Database，用来搜索、比对、下载所有已知 miRNA 相关信息的数据库，包括成熟序列、前体序列、前体二级结构、基因组位置、相关文献等等，并可进行 BLAST 搜索、FTP 下载。miRBase:Targets，存放了所有 miRNA 靶基因的信息。目前已经移至 EBI，并更名为 microCosm。但主要收录了动物 miRNA 的靶基因信息。

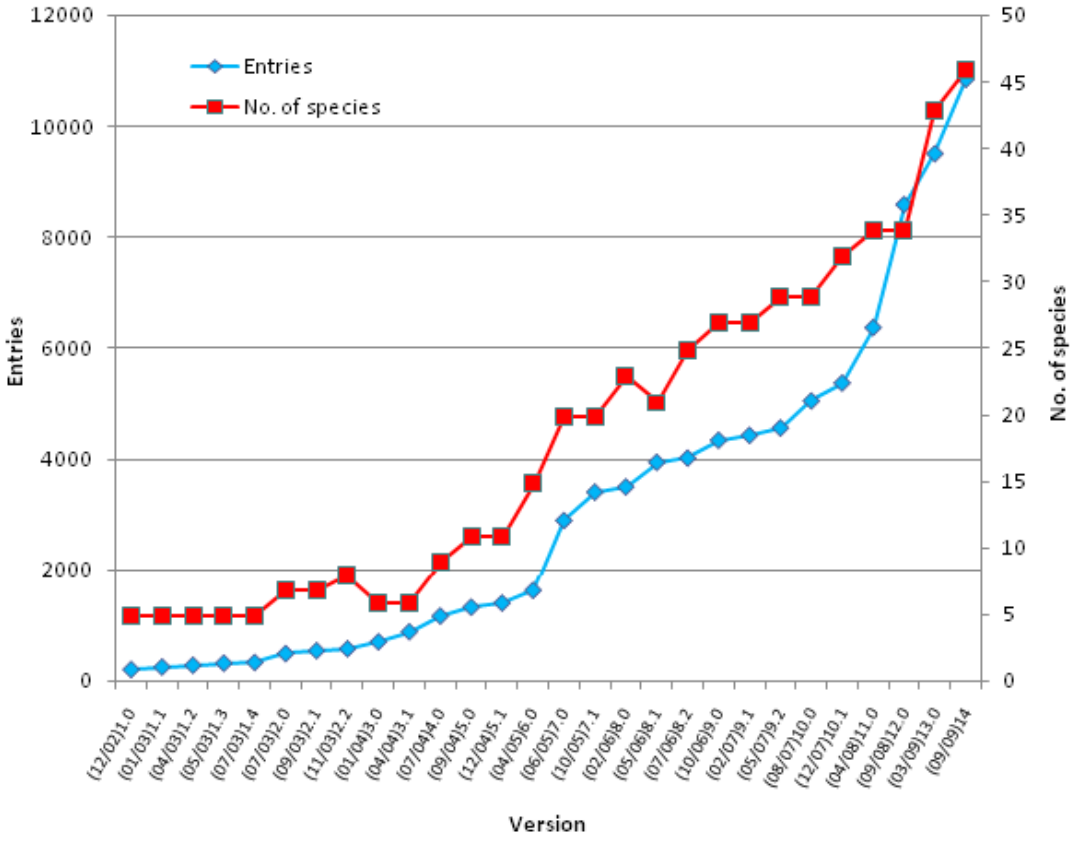


图 17 miRBase 记录和物种数量增长情况

二. siRNA 数据库

由于 siRNA 种类的多样性，为各种类型的 siRNA 建立一个统一的数据库存在很多困难，因此，目前 siRNA 数据的组织没有 miRNA 那样整齐。这里提供两个数据库以供参考，一个是 siRNA Database (http://www.rnainterference.org/)，数据库包括了来自人、大鼠、小鼠的 siRNA 以及 RNAi 等方面的一些资源。另一个是 siRNAdb (http://sirna.sbc.su.se/)，搜集了一千多条经过实验验证的 siRNA 数据和基于计算预测的靶标基因来自 REFSEQ 数据库的 siRNA。

三. CSRDB 和 ASRP

CSRDB (Cereal small RNAs Database, http://sundarlab.ucdavis.edu/smrnas/) 作为专门研究玉米和水稻小 RNA 的数据库，利用 454 测序技术产生了数十万条小 RNA 的数据。可以通过 Genome browser 查看在基因组上的位置信息，并提供了相应的利用 FASTH 软件预测的靶基因数据库 Small RNA target pair (SRTP) dataset。

相应地，ASRP (http://asrp.cgrb.oregonstate.edu/)记录了拟南芥主要生态型和不同组织的小 RNA 数据，包括已知的 miRNA 和 tasiRNA。并提供 BLAST 搜索、Genome Browser 查看、和数据下载。

四. Gene Expression Omnibus (GEO)

Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/gds)作为收录基因表达数据的一个平台，存储了许多原始的表达数据，其中也包括大规模测序的小 RNA 数据。大量原始数据的获取，对于从中挖掘小 RNA 研究相关的信息提供了很大的方便。

小结

本章介绍了作为内源性非编码的小 RNA 分子，小 RNA 在最近几年研究的进展。尽管各种新类型的小 RNA 仍在不断地被发现，但依据小 RNA 产生的前体主要分为两类：miRNA 和 siRNA，miRNA 前体可以形成发卡结构，在茎结构处产生成熟的 miRNA，siRNA 主要形成长的双链 RNA，通过各种酶的切割和加工产生成熟序列。植物小 RNA 通过剪切降解靶标 mRNA 分子或在转录后水平干扰翻译来行使调节功能。小 RNA 靶基因一大类是转录因子，miRNA 可以启始 tasiRNA 的剪切。siRNA 类型非常丰富，其中重复序列相关 siRNA 占了很大部分。不同类型小 RNA 的功能研究已经发现了一些结果，但很多疑问还需要深入调查。

生物信息学在计算和数据分析方面的优势决定了其在小 RNA 研究领域所起的重要作用。小 RNA 在序列和结构上存在很多明显的特征，这导致计算方法在不同类型小 RNA 预测，靶位点查找和功能分析方面都取得了卓越的成就。如何利用现有的数据和工具，并开发更加有效更加强大的分析工具是生物信息学人员需要考虑的课题。综合利用不同的数据和方法对提高计算结果的可靠性有重要意义。

可以说作为一个非常重要而且在飞速发展的研究领域，小 RNA 方面的形成机制跟作用机理还有很多的谜团等待着进一步的挖掘。小 RNA 在表达层次表现的功能及复杂性也许正是高等生物进化过程中获得的一个重要的调控机制。小 RNA 序列“身材”上的小巧和通过序列互补调控的机制在生物进化的经济高效方面得到完美体现，并且其中的翻译抑制调节机制是一个可逆的过程，对于生物不断适应变化的生境有着很强的调节机动性。因此，随着研究的深入，不断发现的小 RNA 的新功能和新类型也会将这类 RNA 序列在生物体高效复杂的调控网络中所起的“四两拨千斤”的作用展示得更加令人惊叹！

（王煜，樊龙江）

主要参考文献

[1] Adai A., Johnson C., Mlotshwa S., Archer-Evans S., Manocha V., Vance V..Sundaresan V. (2005) Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15(1): 78-91
[2] Allen E., Xie Z., Gustafson A. M., Sung G. H., Spatafora J. W..Carrington J. C. (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nat Genet*, 36(12): 1282-1290
[3] Altuvia Y., Landgraf P., Lithwick G., Elefant N., Pfeffer S., Aravin A., Brownstein M. J., Tuschl T..Margalit H. (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 33(8): 2697-2706
[4] Ambros V. (2001) microRNAs: tiny regulators with great potential. *Cell*, 107(7): 823-826
[5] Bartel D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2): 281-297
[6] Baskerville S..Bartel D. P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3): 241-247
[7] Bernstein E., Caudy A. A., Hammond S. M..Hannon G. J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818): 363-366
[8] Bonnet E., Wuyts J., Rouze P..Van de Peer Y. (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci USA*, 101(31): 11511-11516
[9] Brodersen P..Voinnet O. (2006) The diversity of RNA silencing pathways in plants. *Trends Genet*, 22(5): 268-280
[10] Chan S. W., Zilberman D., Xie Z., Johansen L. K., Carrington J. C..Jacobsen S. E. (2004) RNA silencing genes control de novo DNA methylation. *Science*, 303(5662): 1336
[11] Chapman E. J..Carrington J. C. (2007) Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet*, 8(11): 884-896
[12] Chen H. M., Li Y. H..Wu S. H. (2007) Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc Natl Acad Sci USA*, 104(9): 3318-3323
[13] Dezulian T., Schaefer M., Wiese R., Weigel D..Huson D. H. (2006) CrossLink: visualization and exploration of sequence relationships between (micro) RNAs. *Nucleic Acids Res*, 34(Web Server issue): W400-404
[14] Dunoyer P., Himber C..Voinnet O. (2005) DICER-LIKE 4 is required for RNA interference and produces the 21-nucleotide small interfering RNA component of the plant cell-to-cell silencing signal. *Nat Genet*, 37(12): 1356-1360
[15] Finnegan E. J..Matzke M. A. (2003) The small RNA world. *J Cell Sci*, 116(Pt 23): 4689-4693
[16] Gasciolli V., Mallory A. C., Bartel D. P..Vaucheret H. (2005) Partially redundant functions of Arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Curr Biol*, 15(16): 1494-1500
[17] Griffiths-Jones S. (2004) The microRNA Registry. *Nucleic Acids Res*, 32(Database issue): D109-111
[18] Griffiths-Jones S. (2006) miRBase: the microRNA sequence database. *Methods Mol Biol*, 342: 129-138
[19] Guo X., Gui Y., Wang Y., Zhu Q. H., Helliwell C..Fan L. (2008) Selection and mutation on microRNA target sequences during rice evolution. *BMC Genomics*, 9:

454
[20] Herr A. J., Jensen M. B., Dalmay T..Baulcombe D. C. (2005) RNA polymerase IV directs silencing of endogenous DNA. *Science*, 308(5718): 118-120
[21] Hofacker I. L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13): 3429-3431
[22] Howell M. D., Fahlgren N., Chapman E. J., Cumbie J. S., Sullivan C. M., Givan S. A., Kasschau K. D..Carrington J. C. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell*, 19(3): 926-942
[23] Jones-Rhoades M. W..Bartel D. P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell*, 14(6): 787-799
[24] Jones-Rhoades M. W., Bartel D. P..Bartel B. (2006) MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol*, 57: 19-53
[25] Kanno T., Huettel B., Mette M. F., Aufsatz W., Jaligot E., Daxinger L., Kreil D. P., Matzke M..Matzke A. J. (2005) Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat Genet*, 37(7): 761-765
[26] Kasschau K. D., Fahlgren N., Chapman E. J., Sullivan C. M., Cumbie J. S., Givan S. A..Carrington J. C. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol*, 5(3): e57
[27] Kurihara Y..Watanabe Y. (2004) Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci USA*, 101(34): 12753-12758
[28] Lee Y., Kim M., Han J., Yeom K. H., Lee S., Baek S. H..Kim V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20): 4051-4060
[29] Li A..Mao L. (2007) Evolution of plant microRNA gene families. *Cell Res*, 17(3): 212-218
[30] Lim L. P., Lau N. C., Weinstein E. G., Abdelhakim A., Yekta S., Rhoades M. W., Burge C. B..Bartel D. P. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8): 991-1008
[31] Llave C., Kasschau K. D., Rector M. A..Carrington J. C. (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, 14(7): 1605-1619
[32] Lu C., Kulkarni K., Souret F. F., MuthuValliappan R., Tej S. S., Poethig R. S., Henderson I. R., Jacobsen S. E., Wang W., Green P. J..Meyers B. C. (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res*, 16(10): 1276-1288
[33] Nam J. W., Shin K. R., Han J., Lee Y., Kim V. N..Zhang B. T. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11): 3570-3581
[34] Onodera Y., Haag J. R., Ream T., Nunes P. C., Pontes O..Pikaard C. S. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell*, 120(5): 613-622
[35] Osato N., Yamada H., Satoh K., Ooka H., Yamamoto M., Suzuki K., Kawai J., Carninci P., Ohtomo Y., Murakami K., Matsubara K., Kikuchi S..Hayashizaki Y. (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol*, 5(1): R5
[36] Papp I., Mette M. F., Aufsatz W., Daxinger L., Schauer S. E., Ray A., van der Winden J., Matzke M..Matzke A. J. (2003) Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol*, 132(3): 1382-1390
[37] Pontier D., Yahubyan G., Vega D., Bulski A., Saez-Vasquez J., Hakimi M. A.,

Lerbs-Mache S., Colot V..Lagrange T. (2005) Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev*, 19(17): 2030-2040
[38] Rajagopalan R., Vaucheret H., Trejo J..Bartel D. P. (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20(24): 3407-3425
[39] Reinhart B. J., Weinstein E. G., Rhoades M. W., Bartel B..Bartel D. P. (2002) MicroRNAs in plants. *Genes Dev*, 16(13): 1616-1626
[40] Rhoades M. W., Reinhart B. J., Lim L. P., Burge C. B., Bartel B..Bartel D. P. (2002) Prediction of plant microRNA targets. *Cell*, 110(4): 513-520
[41] Robins H., Li Y..Padgett R. W. (2005) Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci USA*, 102(11): 4006-4009
[42] Shen D., Wang S., Chen H., Zhu Q. H., Helliwell C..Fan L. J. (2009) Molecular phylogeny of miR390-guided trans-acting siRNA genes (TAS3) in the grass family. *Plant Systematics and Evolution*, 283(1-2): 125-132
[43] Sullivan C. S..Ganem D. (2005) MicroRNAs and viral infection. *Mol Cell*, 20(1): 3-7
[44] Sunkar R., Girke T..Zhu J. K. (2005) Identification and characterization of endogenous small interfering RNAs from rice. *Nucleic Acids Res*, 33(14): 4443-4454
[45] Sunkar R..Jagadeeswaran G. (2008) In silico identification of conserved microRNAs in large number of diverse plant species. *BMC Plant Biol*, 8: 37
[46] Sunkar R..Zhu J. K. (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, 16(8): 2001-2019
[47] Tang G., Reinhart B. J., Bartel D. P..Zamore P. D. (2003) A biochemical framework for RNA silencing in plants. *Genes Dev*, 17(1): 49-63
[48] Tran R. K., Zilberman D., de Bustos C., Ditt R. F., Henikoff J. G., Lindroth A. M., Delrow J., Boyle T., Kwong S., Bryson T. D., Jacobsen S. E..Henikoff S. (2005) Chromatin and siRNA pathways cooperate to maintain DNA methylation of small transposable elements in Arabidopsis. *Genome Biol*, 6(11): R90
[49] Vaucheret H. (2006) Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev*, 20(7): 759-771
[50] Vazquez F. (2006) Arabidopsis endogenous small RNAs: highways and byways. *Trends Plant Sci*, 11(9): 460-468
[51] Wang H., Chua N. H..Wang X. J. (2006) Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol*, 7(10): R92
[52] Wang J. F., Zhou H., Chen Y. Q., Luo Q. J..Qu L. H. (2004) Identification of 20 microRNAs from Oryza sativa. *Nucleic Acids Res*, 32(5): 1688-1695
[53] Wang S., Zhu Q. H., Guo X., Gui Y., Bao J., Helliwell C..Fan L. (2007) Molecular evolution and selection of a gene encoding two tandem microRNAs in rice. *FEBS Lett*, 581(24): 4789-4793
[54] Wang X., Zhang J., Li F., Gu J., He T., Zhang X..Li Y. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18): 3610-3614
[55] Wang Y., Shen D., Bo S. P., Zheng J., Zhu Q. H., Cai D. G., Helliwell C., Fan L. J. (2010) Sequence variation and selection of small RNAs in domesticated rice. *BMC Evol Biol*, 10: 119
[56] Xie Z., Allen E., Wilken A..Carrington J. C. (2005) DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 102(36): 12984-12989
[57] Xie Z., Johansen L. K., Gustafson A. M., Kasschau K. D., Lellis A. D., Zilberman

D., Jacobsen S. E..Carrington J. C. (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol*, 2(5): E104

[58] Zhang B., Pan X., Cannon C. H., Cobb G. P..Anderson T. A. (2006a) Conservation and divergence of plant microRNA genes. *Plant J*, 46(2): 243-259

[59] Zhang B. H., Pan X. P., Cox S. B., Cobb G. P..Anderson T. A. (2006b) Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*, 63(2): 246-254

[60] Zhang B. H., Pan X. P., Wang Q. L., Cobb G. P..Anderson T. A. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res*, 15(5): 336-360

[61] Zhang Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res*, 33(Web Server issue): W701-704

[62] Zhu Q. H., Spriggs A., Matthew L., Fan L., Kennedy G., Gubler F..Helliwell C. (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res*, 18(9): 1456-1465

[63] Zilberman D., Cao X..Jacobsen S. E. (2003) ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science*, 299(5607): 716-719

[64] Zilberman D., Cao X., Johansen L. K., Xie Z., Carrington J. C..Jacobsen S. E. (2004) Role of Arabidopsis ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr Biol*, 14(13): 1214-1220