第 2-4 章　非编码 RNA 分析

第一节　非编码 RNA 简介

一、非编码 RNA 类型与功能

1. 非编码 RNA 类型

非编码 RNA(non-coding RNA)是指不编码蛋白质的 RNA,因其不编码蛋白质曾被认为是“垃圾 RNA”(junk RNA)。非编码 RNA 主要包括小 RNA 和长链 RNA,在微生物、动植物等许多生物生命活动中发挥着极广泛的调控作用。目前越来越多的科学家开始关注非编码 RNA 的生物学功能及其与重大疾病的关系,人们逐渐意识到,非编码 RNA 对基因调控、基因敲除、农艺性状、病害防治及生物进化探索等都具有重要意义。

内源性非蛋白质编码小 RNA(small non-protein-coding RNA, 18-24nt)广泛存在于高低等生物体内,通过对靶标 mRNA 直接剪切或抑制其翻译,在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 (microRNA,缩写为 miRNA),另一类是小干扰 RNA(small interfering RNA, 缩写为 siRNA)。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 *DCL*1(Dicer-Liker)加工后生成,而小干扰 RNA 则是通过核酸内切酶 *DCL*1、*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(Khraiwesh 等,2012)。目前发现的 siRNA 种类很多,根据前体序列类型和形成机制可分为 ta-siRNA(*trans* acting siRNA)、nat-siRNA(natural antisense transcript-derived siRNA)、hc-siRNA (heterochromatic siRNA)、ra-siRNA(repeat-associated siRNA)和 nat-miRNA (natural antisense miRNA)。迄今为止,在 miRNA 国际数据库 miRBase (www.mirbase.org)中已经有超过 4000 条植物 miRNA 序列记录(Release 21.0)。

长链非编码 RNA(long noncoding RNA,lncRNA),其通常定义为长度大于 200 个核苷酸的非编码 RNA。已有研究表明,lncRNA 对 mRNA 的转录以及转录后都存在着调控作用,并且能够与 DNA 以及蛋白质互作,进一步影响生物体的生命活动。尽管生物体内含有丰富的 lncRNA,但到目前为止人们对于它们在机体内的功能,以及作用机制的了解并不透彻。根据 lncRNA 与相邻编码蛋白基因的关系,我们可以将其分为四类(图 2-4.1):lncRNA 与编码基因有重叠,并且转录方向一致,将其列为同义长非编码 RNA(sense lncRNA);lncRNA 与编码基因有重叠,却在反义链上,将其列为反义长非编码 RNA(antisense lncRNA);lncRNA 由编码基因的内含子中转录产生,将其归类为内含子长非编码 RNA(intronic lncRNA);lncRNA 位于两个编码基因之间非编码区,将其归类为基因间区长非编码 RNA (intergenic lncRNA, lincRNA)。部分 lncRNA 像 mRNA 一样具有 5’帽子和 polyA 尾巴,通过剪接而成熟。生物物理学分析表明,lncRNA 可以折叠形成许多有功能的二级结构。有些 lncRNA 在不同的物

种间相当保守,可能调节不同物种间共有的信号通路,使这些物种具有某些共同的生物学功能。另一方面,有些非保守的 lncRNA 功能具有物种特异性,这可能受限于不同物种的环境选择压力和表型分离相关的进化。

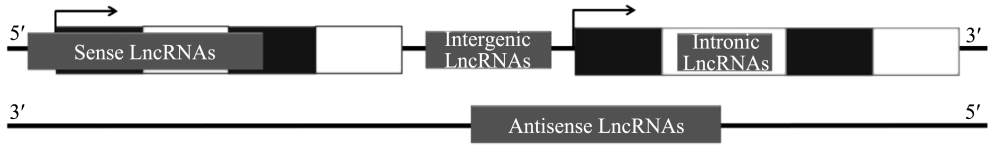


图 2-4.1 lncRNA 的四种类型:与编码蛋白转录本有重叠区域并有一致的转录方向(sense lncRNA);与编码蛋白转录本有重叠区域转录方向相反(antisense lncRNA);转录来自编码基因的内含子区域(intronic lncRNA);由基因间区转录获得(intergenic lncRNA)(引自 Quan 等,2015)

2. 非编码 RNA 功能

非编码 RNA 参与一系列重要调控功能。以下仅以 miRNA 和 lncRNA 为例进行介绍。

miRNA 参与调控许多蛋白质编码基因,特别是转录因子类基因。miRNA 参与调控基因的功能涉及植物生长发育、生殖发育、抗性等等各个方面。图 2-4.2 列出了植物一些保守 miRNA 的靶基因及其参与各种环境胁迫的调控功能。

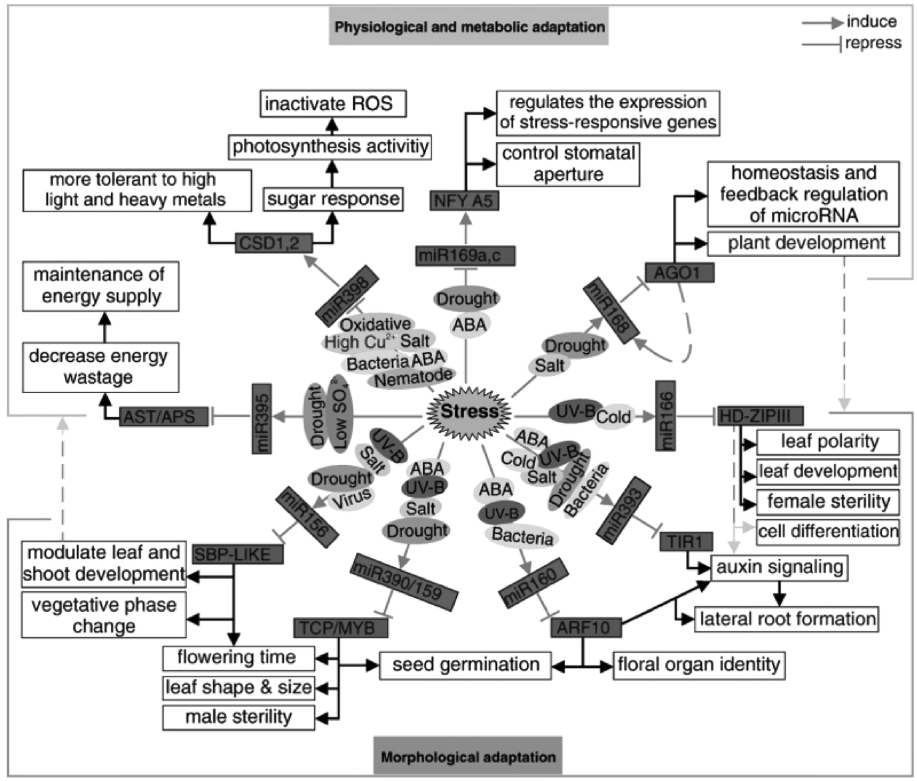


图 2-4.2 部分植物保守 miRNA 的靶基因及其参与各种环境胁迫的调控功能(引自 Khraiwesh 等,2011)

长非编码 RNA(lncRNA)被认为是真核生物基因调控的功能调控元件,很多不同类型的 lncRNA 在真核生物中被发现,并且分别在不同的调控网络中发挥作用(Kim 和 Sung,2012)。例如,有一类参与 miRNA 调控的 lncRNA,称为 eTM(内源诱捕靶标,Endogenous

Target Mimics)。Franco-Zorrilla 等(2007)最早在拟南芥中发现了一个由磷酸盐饥饿诱导 lncRNA 基因 *IPS*1(induced by phosphate starvation 1)——该基因能够与拟南芥中的 miR399 的序列绑定在一起,在 miR399 的剪切位点形成了一个环状凸起结构。因此 *IPS*1 基因无法被 miR399 切割,却能将 miR399 与其真正的靶基因隔离起来。miR399 真正的靶标基因是 *PHO*2(phosphate 2),该基因编码泛素结合酶,对维持细胞内蛋白质生成与降解的相对稳定及细胞的稳态方面起着重要作用(Franco-Zorrilla 等, 2007)。*IPS*1 基因的存在,使得 miR399 靶向 *PHO*2 基因的活性受到抑制,像类似这种具有抑制 miRNA 功能的长非编码 RNA,通常定义为 eTM。本书作者在一项烟草尼古丁合成机制工作中,发现 eTM 还参与植物次生代谢产物合成。在烟草中,我们鉴定到四个烟草特有的 miRNA,其中一个 miRNA(*nta-miRNAX*27)与尼古丁合成途径关键基因 *QPT*2(quinolinate phosphoribosyl transferase 2)基因存在靶向关系。进一步生物信息学分析发现,长非编码基因(*nta-eTMX*27)作为内源 miRNA 诱捕靶标参与 *nta-miRNAX*27 的调控。烟草打顶处理后,*QPT*2 基因表达会显著上调,快速促进尼古丁合成,实验表明,*nta-eTMX*27 和 *QPT*2 基因表达以及尼古丁含量存在显著相关性;进一步转基因功能实验证实,敲除和过量表达 *nta-eTMX*27 显著影响 *QPT*2 基因表达及尼古丁含量。上述结果证明长非编码基因 *nat-eTMX*27 介导调控 *nta-miRX*27-*QPT*2 的功能。我们的实验结果首次揭示了植物中次生代谢产物合成相关的"miRNA-eTM-target"调控模式(Li 等,2015)。

二、非编码 RNA 进化

1. 起源与进化

植物中许多 miRNA 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时间(约 600 百万年前)。目前还没有发现动植物中 miRNA 编码基因或靶基因的起源基因,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 与其靶基因共同进化的一个潜在机制。由于 miR161/163 两个家族都是新产生的 miRNA 基因,而且跟大多数保守 miRNA 家族不同,在基因组上它们均位于其靶基因的附近区域,因此 Allen 等认为 miRNA 家族有可能通过靶基因家族扩增过程中的倒转复制或反向倍增(inverted duplication)产生。如图 2-4.3 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片段复制,从而为形成 miRNA 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)。新形成的创始基因及其相关的家族成员,有可能成为 *DCL* 基因的靶标从而导致 siRNA 的产生,因此 *DCL* 基因会在转录后或染色质水平受到 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* 识别的功能受限制,形成一类特异的 siRNA 家族(图 2-4.3 中步骤 2);而对 *DCL*1 调控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 *DCL*1 识别功能限制下,miRNA 及其互补序列(miRNA*)中只剩下一段与原始序列相似的序列(步骤 4)。miRNA 位点的复制导致了 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 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。

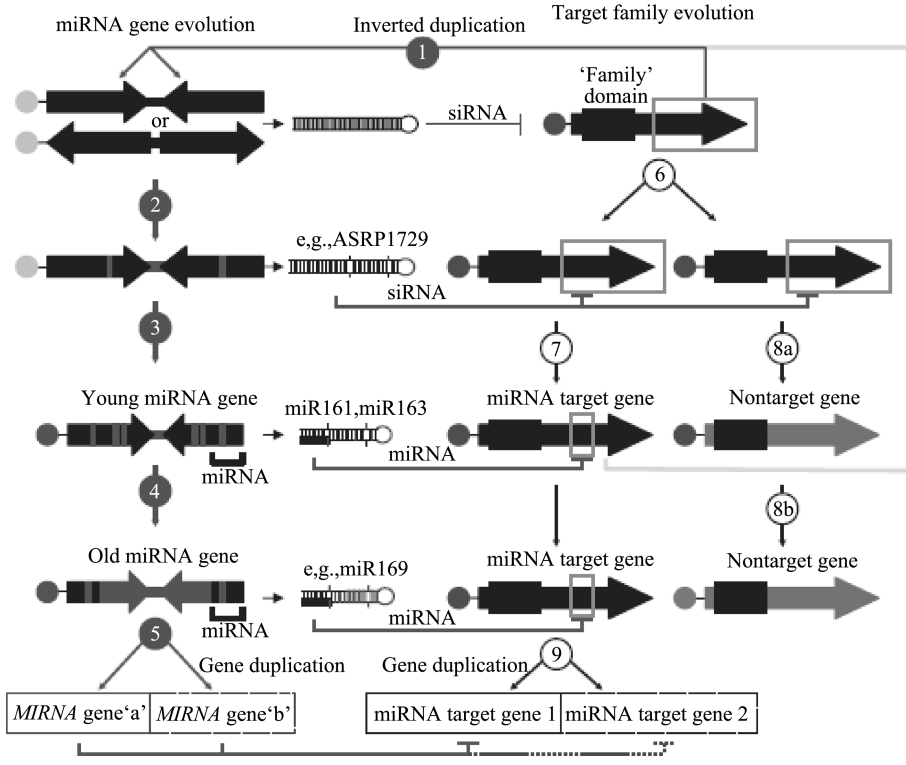


图 2-4.3 植物 miRNA 反向倍增进化模型(引自 Allen 等,2004)

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证据的存在,因此对于保守 miRNA 的解释仍然有待进一步验证。同样,由于动物 miRNA 前体序列较短,也不能为创始基因提供信息。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。与植物 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 等,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 和 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(图 2-4.4):第一,起源于创始基因家族的小 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(见下面有关人工选择部分);第三,可能是最普遍的命运,随

着小 RNA 产生位点的启动子区域、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调控元件的进化提供了很好机会(Chapman 和 Carrington, 2007)。

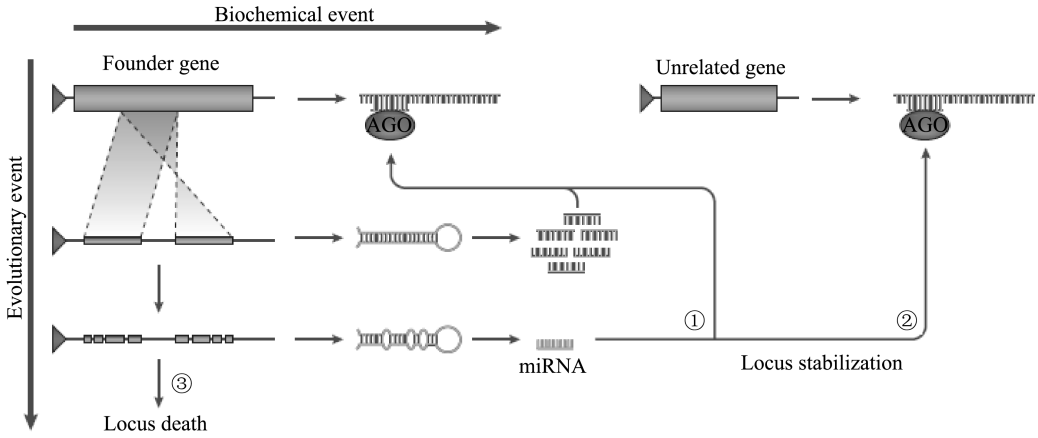
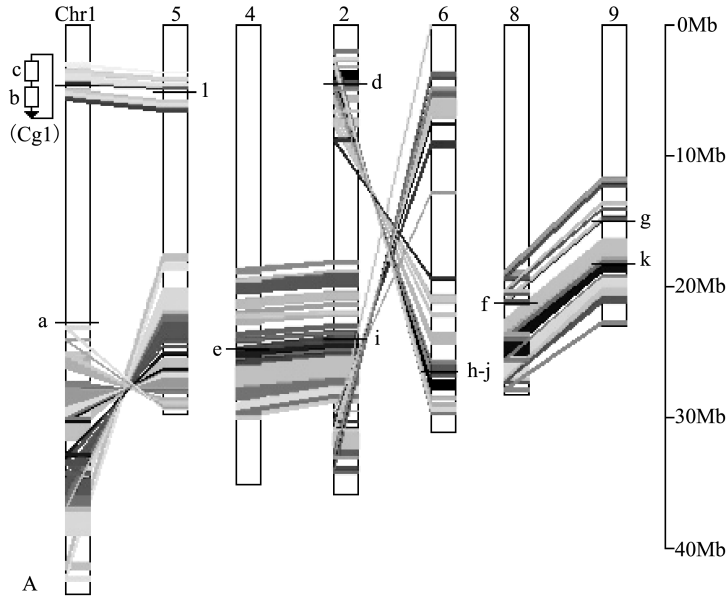


图 2-4.4　植物新 miRNA 基因进化模型(引自 Chapman 和 Carrington, 2007)

根据我们在水稻上的研究结果表明,全基因组倍增(whole genome duplication, WGD)和基因水平的倍增事件对水稻 miRNA 家族的形成起到很大的作用。例如水稻 miR156 基因家族,禾本科分化前的那次 WGD 事件使水稻基因组上 miR156 成员几乎增加了一倍(图 2-4.5)。随着进化过程,虽然部分 miR156 成员由于选择压丢失了,但大部分都保留了下来。同时,可以看到,基因水平的复制事件也增加了 miR156 成员数量。另外,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(paralog)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 等, 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 2-4.6 展示了来自 WGD 两个水稻编码基因拷贝,在 miR397 结合位点碱基突变情况,突变导致 miR397 的绑定和调控关系的改变(Wang 等,2010)。



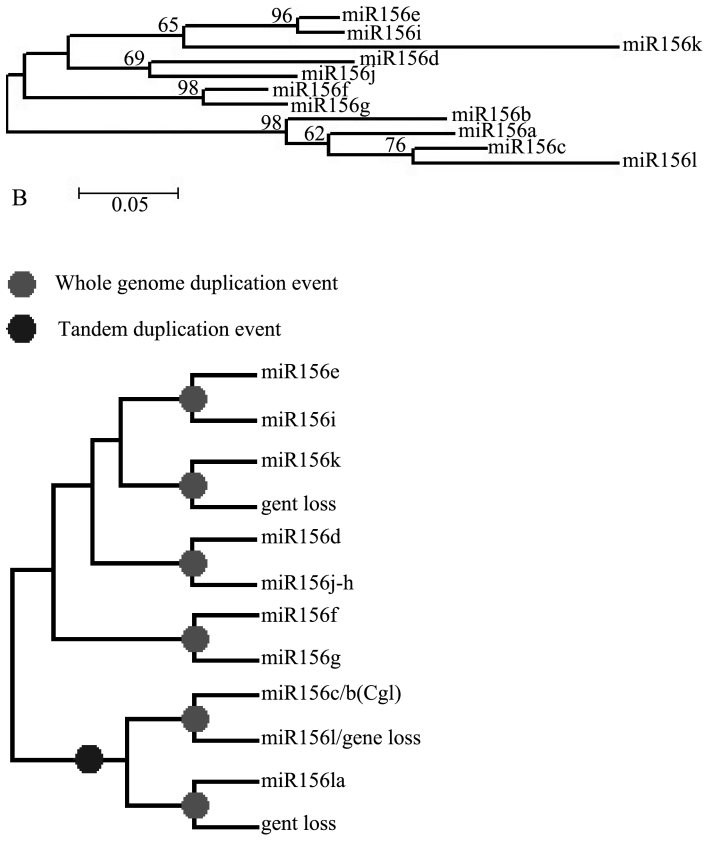
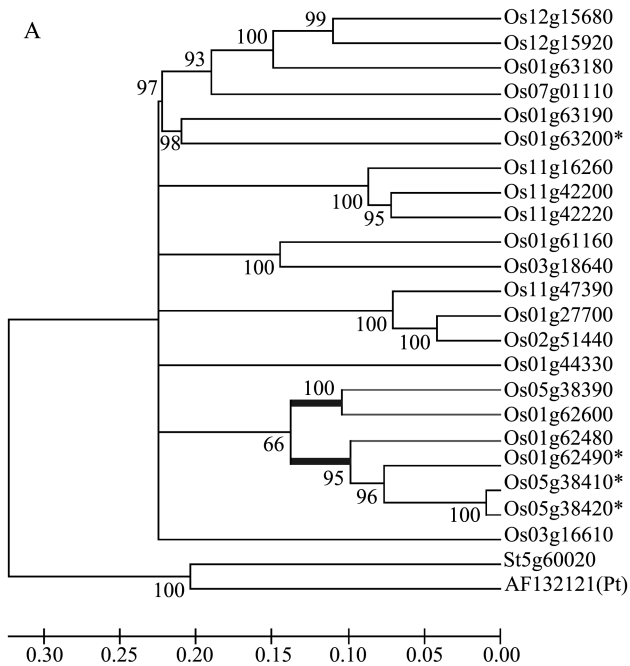


图 2−4.5 水稻 miR156 家族在基因组上的分布和系统进化关系（引自 Wang 等,2007）



B

```
miR397a   5'GUAGUUGCGACGUGAGUUACU 3'
                 |||||||||||||||||| ||
Os05g38420   CAUCAACGCUGCACUCAACGA(1)*
Os01g62480   CAUCAACGCCGCGCUCAACGA(3)
Os05g38390   GAUCAACGCGGCGCUCAACGA(4)
Os01g62600   GAUCAACUCGGCGCUCAACGA(5)
```

图 2-4.6　来自全基因组倍增事件的两个水稻编码基因拷贝在 miR397 结合位点碱基突变情况(引自 Wang 等,2010)

线性 lncRNA,顾名思义,形成 lncRNA 最后的转录本呈线性,是指常规的 lncRNA,用于区分环状 RNA。对于 lncRNA 的分类,之前已经阐述过,主要是依据 lncRNA 在基因组上的转录位置。LncRNA 几乎可以在基因组的任何位置产生。lncRNA 的起源又是一个值得探究的方向(图 2-4.7,图中 A-E 四种机制依次如下):1)第一个可能的起源是在进化过程中,编码蛋白的 RNA 变成为非编码基因。在哺乳动物中,*Xist*(Xinactive specific transcript)作为 lncRNA 是使 X 染色体失活的关键分子。在该 lncRNA 中,有好几个外显子和启动子来源于一个编码基因"*Lnx*3(ligand of numb-protein X 3)"。研究发现 *Lnx*3 在进化过程中,外显子之间发生了框架性损坏(两个相邻外显子合并在一起),从而变成 lncRNA;2)由于染色体的重组事件的发生,两个不能转录的序列区域合并在一起,组成一条含多个外显子的 lncRNA;3)由原来的非编码基因经历逆转录转座事件,形成新的非编码基因;4)由序列上发生的串联重复事件,形成 lncRNA;5)转座子在序列上的随机插入,同样会形成具有功能的 lncRNA。由此可见,lncRNA 起源机制就如同它们的功能一样非常复杂,还有待科学家们继续深入研究。

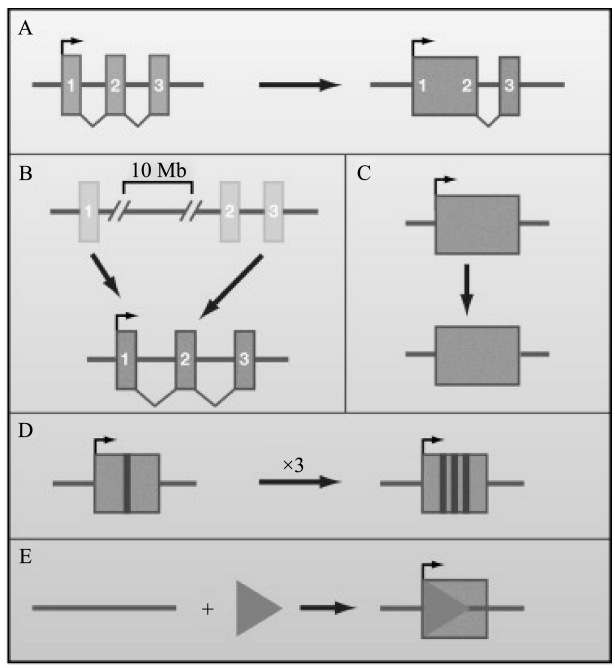


图 2-4.7　长非编码 RNA(lncRNA)几种可能的起源机制(引自 Ponting 等,2009)

2. 序列保守性

相对于编码基因,非编码 RNA 的一个特点是其序列保守性很弱。以小 RNA 为例,miRNA 中只有部分在植物界是保守的。图 2-4.8 和表 2-4.1 给出了目前已知的在植物界保守的 miRNA 家族。

目前国际 miRNA 数据库 miRBase 数据库(www.mirbase.org/)中 miRNA 记录已经接近 3 万条(V21.0 版本),其中很多 miRNA 家族均可以在多个物种中找到,例如 miR156 和 miR166 等家族在许多植物物种中均存在(表 2-4.1)。这种 miRNA 的保守性对于在新物种中预测保守 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 家族不同物种的同源基因间往往只有 1～2 个碱基的差异。这种保守性能够用来查找不同物种间保守 miRNA 的研究。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。

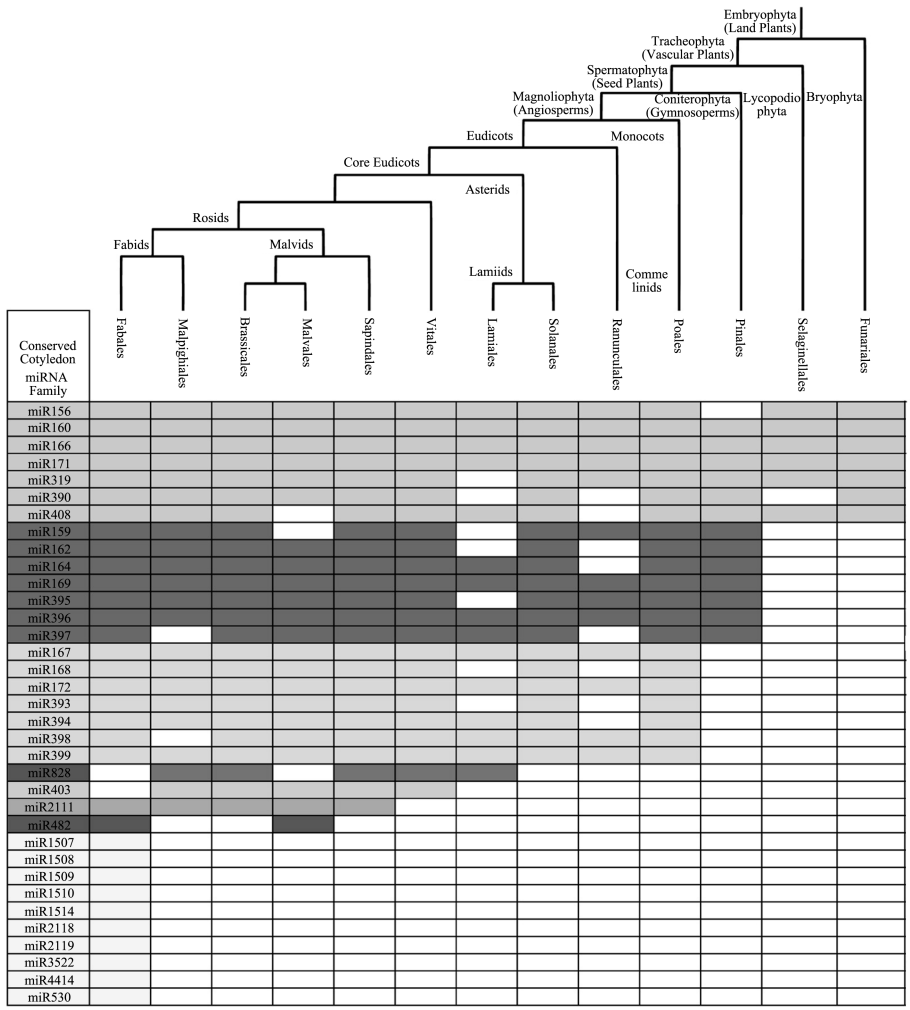


图 2-4.8 植物 miRNA 家族的保守性(改自 Goetell 等,2014)

同样颜色表示相应物种中均发现存在该 miRNA

表 2-4.1　植物保守 miRNA 家族(根据 miRBase 数据库 V21.0,2014)

miRNA 家族	物种数量*	miRNA 家族	物种数量*	miRNA 家族	物种数量*	miRNA 家族	物种数量*	miRNA 家族	物种数量*
miR156	48	miR172	35	miR390	29	miR2111	14	miR1507	5
miR166	42	miR319	34	miR168	28	miR530	13	miR157	5
miR396	41	miR164	33	miR397	27	miR477	12	miR3627	5
miR171	40	miR408	32	miR393	24	miR535	12	miR437	5
miR160	39	miR162	30	miR394	23	miR828	11	miR479	5
miR167	36	miR395	30	miR482	23	miR2118	10	miR528	5
miR159	35	miR398	30	miR403	15	miR529	9	miR824	5
miR169	35	miR399	30	miR827	15	miR444	6	miR858	5

*按照含有特定 miRNA 家族的物种数量多少排序;仅列出物种数量超过 5 个 miRNA 家族

siRNA 与 miRNA 情况类似,其保守的 siRNA 基因不多。一个比较典型的例子是一个 ta-siRNA 基因(*TAS*)——*TAS*3。我们通过 *TAS*3 基因的保守序列片段,克隆测序发现了 51 个来自禾本科的 *TAS*3 基因(Shen 等, 2009)。通过序列比较等方法,我们发现 *TAS*3 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 *TAS*3 基因,在基因组中保持了共线性关系;同时 *TAS*3 在不同禾本科基因组上也存在明显的基因组共线性。

3. 人工选择靶基因

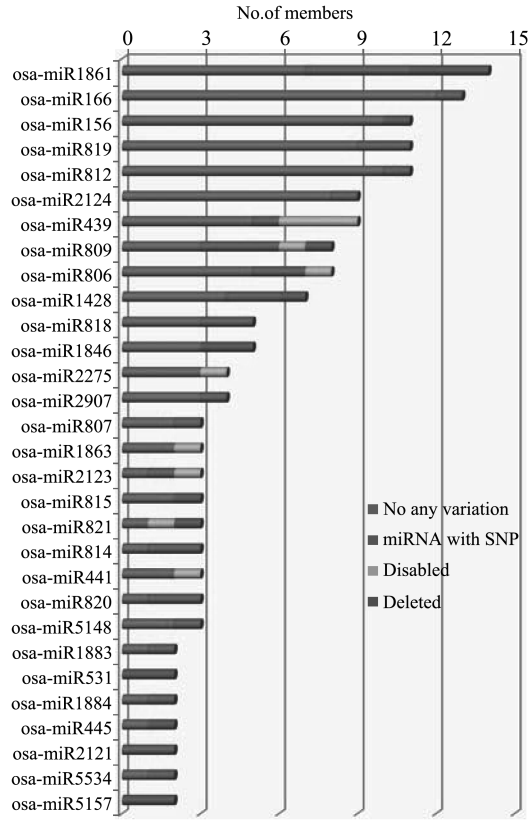
人工选择的靶基因一直以来是作物遗传育种与功能基因研究的一个热点,多年来先后在水稻、玉米等作物上克隆了一批与驯化选择相关基因。比较著名的驯化基因包括玉米中顶端分化和穗型相关基因、番控制茄果实大小相关基因等;在水稻上,最近,国内外科学家发现了一批与人工选择相关的基因,如落粒、分蘖角度、谷色、产量等相关基因。目前发现的这些所谓驯化基因位点均为蛋白质编码基因,而且主要编码的是转录因子类基因。miRNA 主要的调控对象之一为转录调节因子,这说明人工选择的靶基因除了转录因子及其下游基因外,还可能针对转录因子的调控(上游)基因。由此产生一个问题:这些非编码基因位点在我们进行作物驯化和育种过程中,是否同样受到选择?或者说,非编码 miRNA 基因会是否人工选择的直接靶基因?

为了回答这个问题,我们首先选择了一个 miRNA 基因位点(*MIR*156*b*/*c*)进行了研究。*MIR*156*b*/*c* 基因调控的是一个转录因子,在玉米和水稻中,该转录因子的突变将直接导致栽培种多分蘖草状性状的产生,回复到祖先野生种特性。我们的群体遗传学调查表明,该 miRNA 基因位点在栽培群体中的遗传多态性仅保留了野生群体的 8.9%,检测到显著正向选择信号(Tajima $D=-2.01$, $P<0.05$),说明该位点受到强烈的人工选择效应的影响(Wang 等, 2007)。后来进一步功能研究表明,*OsmiR*156 的确调控水稻转录因子 *OsSPL*14(squamosa promoter-bindling-like protein 14, SPL14),进而对株型产生影响(Jiao 等,2010)。这些研究证明,水稻在驯化过程中,通过对 *OsmiR*156 进行直接选择,使栽培稻具有理想株型。

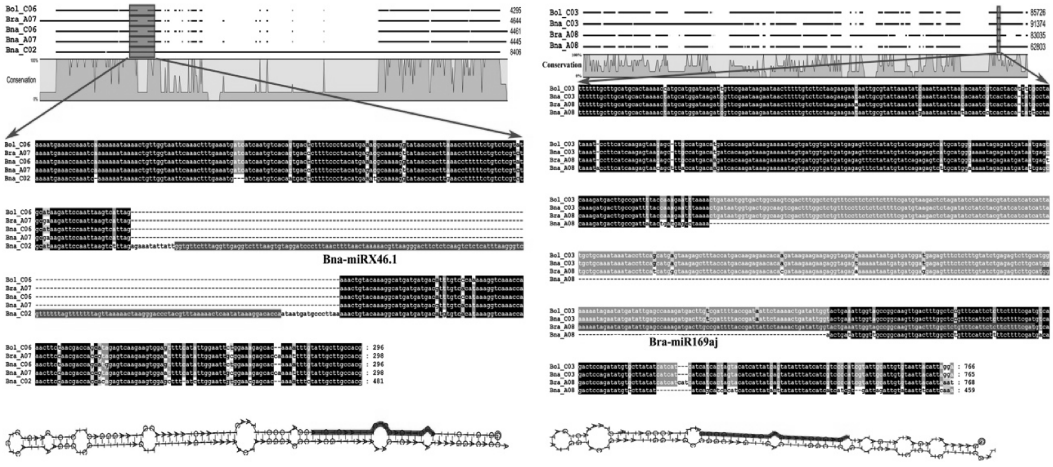
我们进而在水稻基因组上更多位点(120 个水稻 miRNA 和 ta-siRNA 等位点)进行选择效应的群体遗传学调查,确定这些基因在水稻栽培群体和野生群体 (*O. rufipogen*)的基因序

列。然后利用分子进化方法检测这些基因是否受到定向选择(所谓定向选择,即偏离中性进化途径)。与驯化/育种选择相关的小 RNA 位点在栽培群体中受到正向选择,但在野生群体中未受到正向选择。结果发现至少 8 个位点(如 miR164,miR395,*TAS*3)在栽培稻群体中受到显著正向选择效应,但在其祖先野生稻群体中没有检测到相应的信号,表明这些基因位点在驯化/遗传改良过程中,受到了强烈的人工选择。对这些基因序列和芯片相对表达量的比较分析,发现了部分基因在栽培和野生群体之间存在着序列和表达差异(Wang 等,2010)。我们对东乡野生稻小 RNA 群体的大规模调查发现,在水稻驯化过程中,大量 miRNA 基因丢失或去功能化了(图 2-4.9A);同样我们在进行栽培油菜(由甘蓝和白菜两个二倍体近缘种在 7500 年前杂交而成)miRNA 群体调查中发现,其栽培种起源后也伴随着 miRNA 的丢失,但同时我们也发现新的 miRNA 的形成(图 2-4.9B)。这些 miRNA 都可能与上述作物驯化和品种形成直接相关。

基于我们和其他一些研究结果,证实非编码 RNA 为作物人工选择的直接靶基因,其在作物品种形成中发挥重要作用,为作物基因组重要遗传构成。



A



B

图 2-4.9　作物人工选择过程中伴随着大量 miRNA 的丢失,同时也伴随着新 miRNA 的产生

A. 38 个保守 miRNA 在水稻驯化过程中发生了大量遗传变异,导致去功能化甚至丢失(引自 Wang 等,2012);B. miRNA 在杂交形成异源四倍体油菜后的丢失和获得。Bna-miRX46.1 是在油菜基因组 C 序列新获得的位点;Bra-miR169aj 在油菜 A 基因组上缺失(引自 Shen 等,2015)

三、样品采集及其测序方法

由于 RNA 表达的时空特异性,导致传统的实验方法研究 RNA 效率很低,成本较高,因此借助计算方法研究 RNA 是一个很好的补充;同时随着高通量测序技术的迅猛发展,科学界也开始越来越多地应用第二代测序技术来解决生物学问题。例如在转录组水平上进行全转录组测序(whole transcriptome sequencing, RNA-Seq),从而开展可变剪接、编码序列单核苷酸多态性、基因表达情况等研究;小分子 RNA 群体测序,通过分离特定大小的 RNA 分子进行测序,从而发现新的 miRNA 分子;通过去核糖体 RNA 并建立链特异性文库,进而鉴定新的 lncRNA 分子。在转录组水平上,与染色质免疫共沉淀(ChIP)和甲基化 DNA 免疫共沉淀(MeDIP)技术相结合,从而检测与特定转录因子结合的 DNA 区域和基因组上的甲基化位点。利用紫外交联免疫沉淀结合高通量测序(CLIP- Seq),可以在全基因组水平揭示 RNA 分子与 RNA 结合蛋白相互作用。上述技术对植物中样品的采集方式以及涉及的测序方法均有所不同。

样品采集时,为了保证植物材料的代表性,必须运用科学方法采取材料。从大田或实验地、实验器皿中采取的植物材料,称为“原始样品”,再按原始样品的种类(如植物的根、茎、叶、花、果实、种子等)分别选出“平均样品”,然后根据分析的目的、要求和样品种类的特征,采用适当的方法,从“平均样品”中选出供分析用的“分析样品”。取样的地点,一般在距田埂或地边一定距离的株行取样,或在特定的取样区内取样。取样点的四周不应该有缺株的现象。进行 RNA 测序,因此需要用新鲜样品。取样时注意保鲜,取样后应立即进行待测组分提取;也可采用液氮中冷冻保存或冰冻真空干燥法得到干燥的制品。在进行 RNA 提取时,样品的匀浆、研磨一定要在冰浴上或低温室内操作。新鲜样品采后来不及提取或测定的,可放入液氮中速冻,再放入-70 ℃冰箱中保存,避免反复冻融,因为这会导致提取的 RNA 降解和提取量下降。同时,我们也需要注意样品的备份,以应付实验中的意外情况。

目前可以用 Illumina 基因组分析仪、Roche454 基因组测序仪、AB Life Technologies 的

SOLiD 系统等进行小 RNA 测序。使用第二代测序技术对小 RNA 进行检测的基本步骤主要包括:(1)构建 DNA 模板文库,从总 RNA 中分离纯化出 20～30nt 的小 RNA 后,使用 T4 连接酶分别在 miRNA 的 5′端和 3′端连上接头序列,进行 RT-PCR 得到 70～80 bp 的 DNA 片段;(2)将所得单链模板文库,固定在平面或是微球的表面;(3)通过桥式 PCR、微乳滴 PCR 或原位成簇对数据进行扩增;(4)采集并记录 PCR 循环中的光学事件;(5)对产生的阵列图像进行时序分析,获得 DNA 片段的序列。对送样的 RNA 样品也会有一定的要求:(1)样品浓度:总 RNA 浓度≥350ng/μl;(2)样品总量:总 RNA 用量>6ug;(3)样品纯度:OD260/280 为 1.8～2.2,260nm 处有正常峰值;(4)RNA 完整性:总 RNA 28S/18S≥1.5,小 RNA 要求电泳后有单一 5S 条带。

植物中小 RNA 以碱基互补配对的方式靶向 mRNA,导致 mRNA 的降解。为了大规模验证小 RNA 与 mRNA 的互作关系,我们常常会用到降解组测序(degradome sequencing)。降解组测序的原理是,在植物体内绝大多数的 miRNA 是利用剪切作用调控靶基因的表达,且剪切常发生在 miRNA 与 mRNA 互补区域的第十位核苷酸上。靶基因经剪切产生二个片段,5'剪切片段和 3'剪切片段。其中 3'剪切片段,包含有自由的 5'单磷酸和 3'polyA 尾巴,可被 RNA 连接酶连接,连接产物可用于下游高通量测序;而含有 5'帽子结构的完整基因,含有帽子结构的 5'剪切片段或是其他缺少 5'单磷酸基团的 RNA 是无法被 RNA 酶连接,因而无法进入下游的测序实验;对测序数据进行深入地比对分析,可以直观地发现在 mRNA 序列的某个位点会出现一个波峰,而该处正是候选的 miRNA 剪切位点(完整实验流程参见图2-4.10)。利用降解组测序,摆脱了生物信息学预测的限制,真正从实验中找到了 miRNA 的作用靶基因。

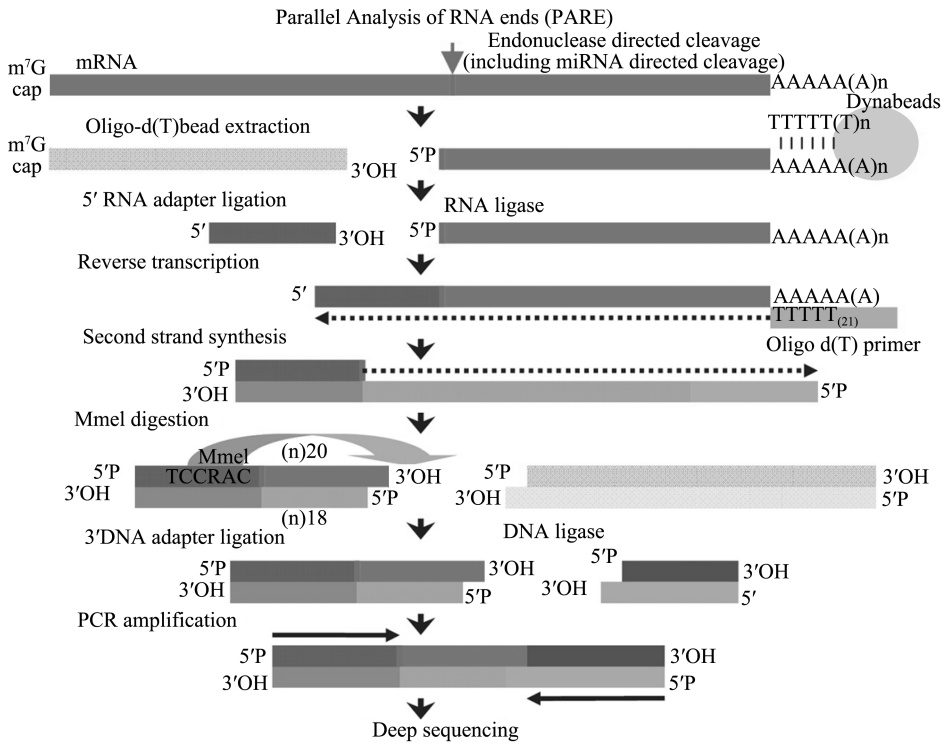


图 2-4.10 降解组测序建库流程

收集靶基因经剪切产生二个片段,5'剪切片段和 3'剪切片段,用 RNA 连接酶连接 3'剪切片段(含有 PolyA 尾巴的片段),进行反转录 PCR 扩增可得到目标片段(引自 Thomson 等,2011)

为了便于 lncRNA 群体的研究,出现了专门针对长非编码 RNA 测序技术。相对 mRNA 来说,lncRNA 表达水平比较低,而数量上来说应该是 mRNA 的几倍甚至到几百倍。在总 RNA 样本中,核糖体 RNA(rRNA)的丰度最高,占到总 RNA 的 80%以上。这些 rRNA 所含的转录组信息很少,浪费了宝贵的测序资源。因此在提取总 RNA 过程中,去除 rRNA 可以最大程度的保留转录组信息,从而达到富集 lncRNA 的效果。之后再进行 cDNA 文库构建,进入测序仪测序。

四、非编码 RNA 主要数据库

1. miRBase 数据库

作为最权威和完整的 miRNA 数据库(http://www.mirbase.org/),截止到目前(2016 年 3 月),miRBase 已经收录了 223 个物种中接近 30 000 条的 miRNA 记录(图 2-4.11)。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,主要收录了动物 miRNA 的靶基因信息。

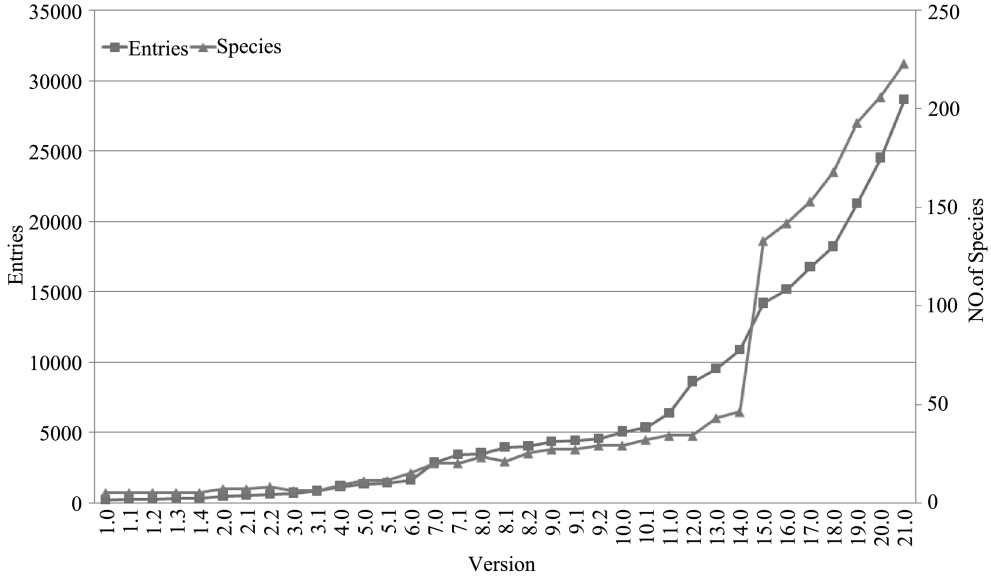


图 2-4.11　miRBase 数据库记录和物种数量增长情况

2. siRNA 数据库

由于 siRNA 种类的多样性,为各种类型 siRNA 建立一个统一的数据库存在很多困难,因此,目前对于 siRNA 数据的组织没有 miRNA 那样规整统一。这里提供两个数据库以供参考,一个是 siRNA 数据库(http://web.mit.edu/sirna/),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。其二是 siRNAdb(http://sirna.sbc.su.se/),搜集了经过实验验证的 siRNA 数据和基于计算预测的靶标基因来自 REFSEQ 数据库的 siRNA。

3. lncRNA 数据库

近些年来由于对 lncRNA 的大量研究,出现了许多与 lncRNA 相关的资源。目前相关数据库质量和覆盖范围不同,因此我们在这介绍一些具有代表性的数据库。这些数据库均在相关刊物中发表,并且可以经过网络访问和搜索。

1) lncRNAdb 数据库

该数据库由 Amaral 等在 2011 年开发,如今已经更新到了 2.0 版本(http://www.lncrnadb.org/)。数据库收录了真核生物的 287 条 lncRNA 注释信息。它也整合了 Illumina 表达图谱数据以及核苷酸序列信息。可以对已知的 lncRNA 序列进行信息检索,同时可以用 BLAST 对未知序列进行搜索,比对是否与已有的 lncRNA 保守。

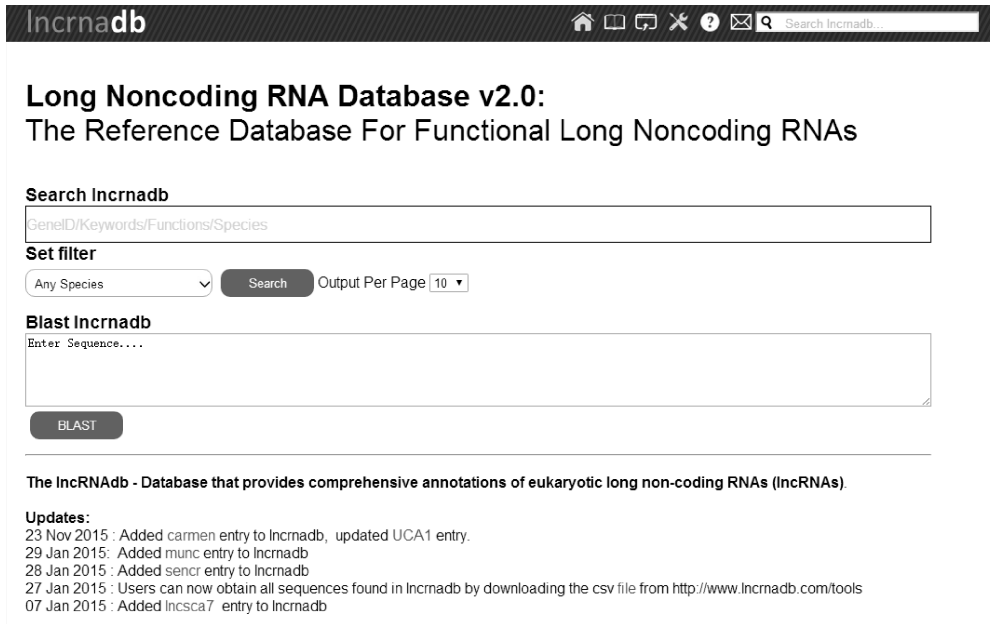


图 2-4.12 lncRNAdb 数据库网络服务主页界面

2) NONCODE 数据库

NONCODE 是一个整合了非编码 RNA 信息的一个大型数据库(http://noncode.org/),它包括 lncRNA, smRNA, tRNA 和 rRNA 等数据。现在已经收录了 16 个物种,主要是关于动物和人类的,植物中只有拟南芥相关信息被收录。数据库的数据来源是通过整合已发表研究结果和其他公共数据库资源。它针对所有非编码 RNA 进行了统一的命名格式,使查阅更加方便。在这里可以对各个物种的非编码基因进行比较,对非编码 RNA 注释比较完善,是目前高收录率和高使用度的非编码数据库之一。

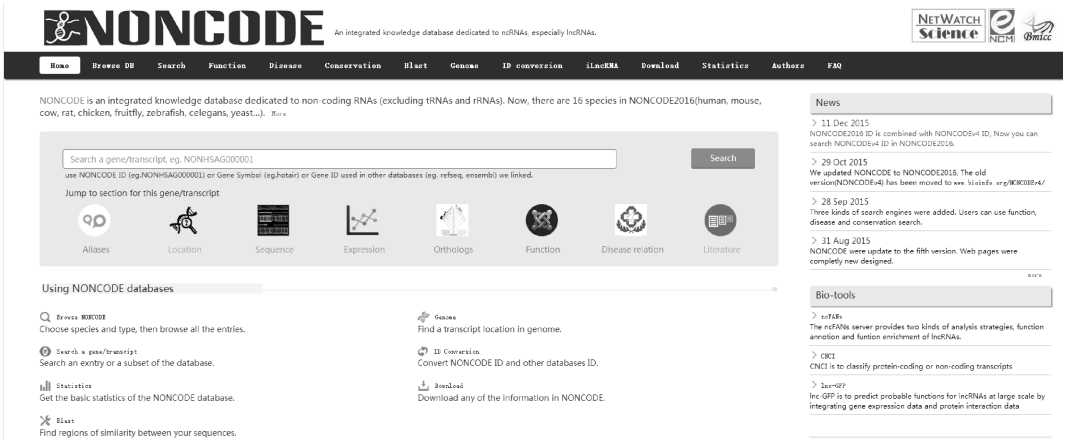


图 2-4.13　Noncode 数据库网络服务访问界面

3）lncRNADisease 数据库

lncRNADisease(cmbi.bjmu.edu.cn/lncrnadisease)收录了与疾病相关的 lncRNA 信息,包括癌症、心血管疾病和神经性疾病等。该数据库不仅整合了经过实验验证的疾病相关 lncRNA,也开发了相应的工具用来预测新类型的 lncRNA 与疾病之间相互关系。

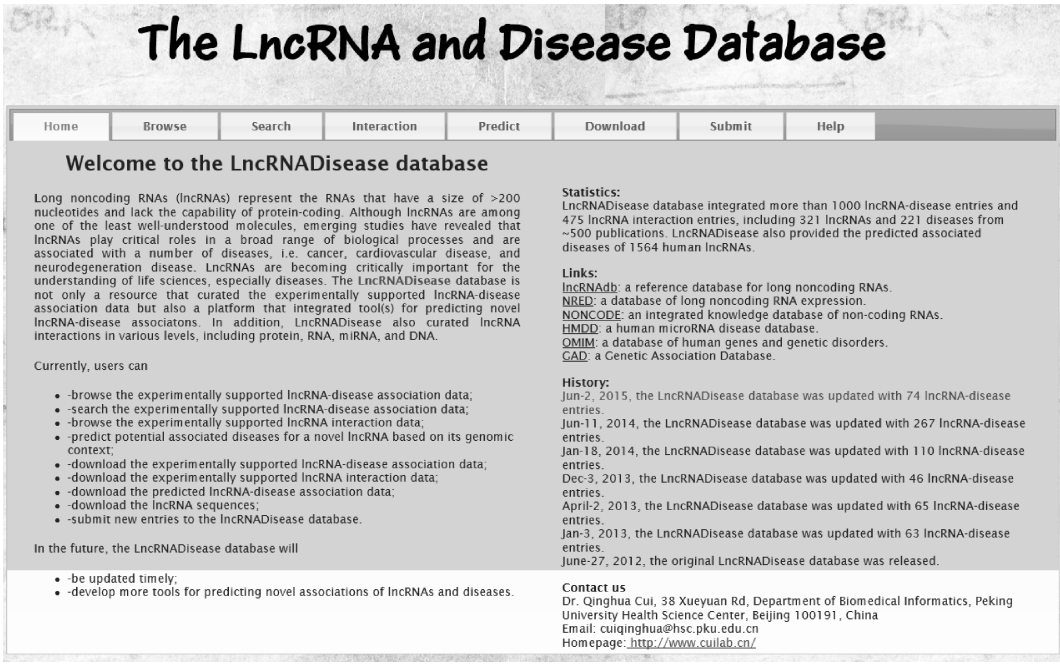


图 2-4.14　lncRNADisease 数据库网络服务访问界面

4）CircNet 环状 RNA 数据库

CircNet(http://circnet.mbc.nctu.edu.tw/)利用已经发表的转录组数据,根据自行开发的流程大规模鉴定环状 RNA 分子。该数据库利用已发表 RNA 数据主要为我们提供了以下资源;circRNAs 的鉴定;整合了与 miRNA 的互作网络;对 circRNA 及其他可变剪切体做了表达分析;对 circRNA 在基因组上做了注释;提供了可变剪切的 circRNA 的序列; 对 circRNA 与

miRNA 的互作、表达、基因组位置及其结合序列深度都做了可视化界面。

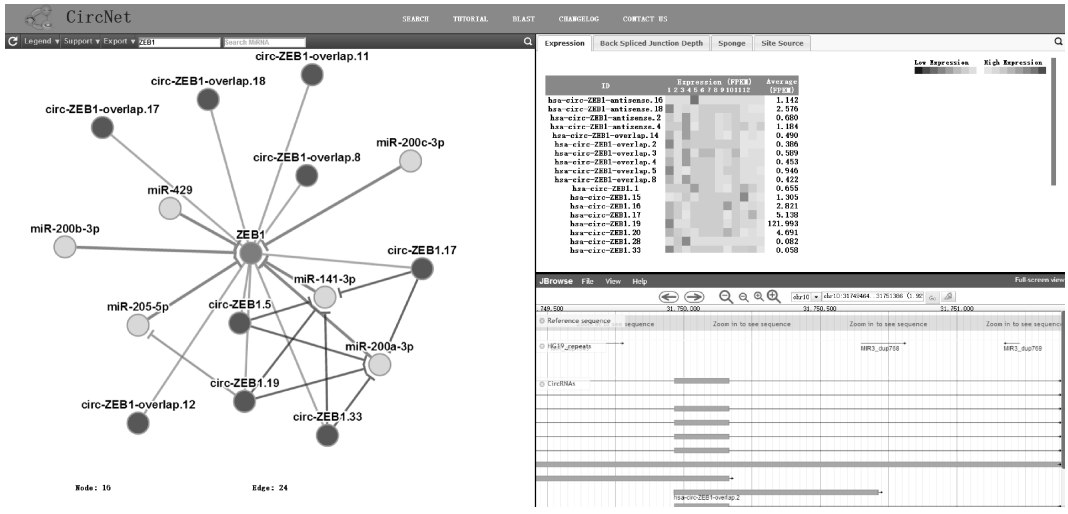


图 2-4.15 CircNet 网络服务访问界面

第二节 小 RNA 计算识别与靶基因预测

一、miRNA 主要特征及计算识别

1. miRNA 主要特征

在植物中,miRNA 的生成起源于一种 miRNA 的初级转录(pri-miRNA),它由 miRNA 基因经 *Pol* Ⅱ(polymerase Ⅱ)转录酶转录并折叠形成具有茎环结构的 miRNA 前体(pre-miRNA)(图 2-4.16)。随后在 *DCL1* 酶(Dicer-like enzyme)、*HYL1*(hyponastic leaves 1)和 *SE*(CZHZ sinc-finger protein SERRATE)共同催化作用下,miRNA 前体茎环结构切割形成 miRNA:miRNA * 的双链复合结构,该 miRNA 复合体结构的 3'端在 *HEN1*(DsRBD protein-like protein 1)酶作用下形成甲基化,并由 *HST1* 间蛋白质输出到细胞质。在细胞质中,miRNA 复合体中的其中一条链与 AGO(argonaute)蛋白结合形成 *RISC* 复合体,该复合体通过碱基互补配对原则作用到靶基因,从而调节目标靶基因在植物体中的表达,而与成熟 miRNA 成互补结构的 miRNA *,通常情况下都会降解并且不具有调控基因表达的功能(Zhang 等,2011)。大部分植物中 miRNA 和靶基因会形成完全或近似完全的匹配,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。

动植物 miRNA 存在着差异。主要差异包括:1)前体序列长度不同。植物 miRNA 前体的茎环结构更大、更复杂,大约是动物中的 3 倍长,预测的折回(fold-back)长度变异(64～303nt)也比动物 miRNA(60～70nt)明显;2)植物的 miRNA 的长度多为 21nt 和 24nt,而动物 miRNA 长度多为 22～23nt,这源于 *Drosha* 与 *Dicer* 切割性能的差异;3)植物 miRNA 5'端更优先选择尿嘧啶(U),热力学分析表明这种末端不稳态是通过 *RISC*(RNA-induced silencing complex)来维持的,另外植物中 miRNA 3'末端 2nt 突出的 3'-OH 存在甲基化,而动物中无甲基化;4)相对于动物 miRNA,植物具有较高的进化保守性(详见上节非编码 RNA 序列保

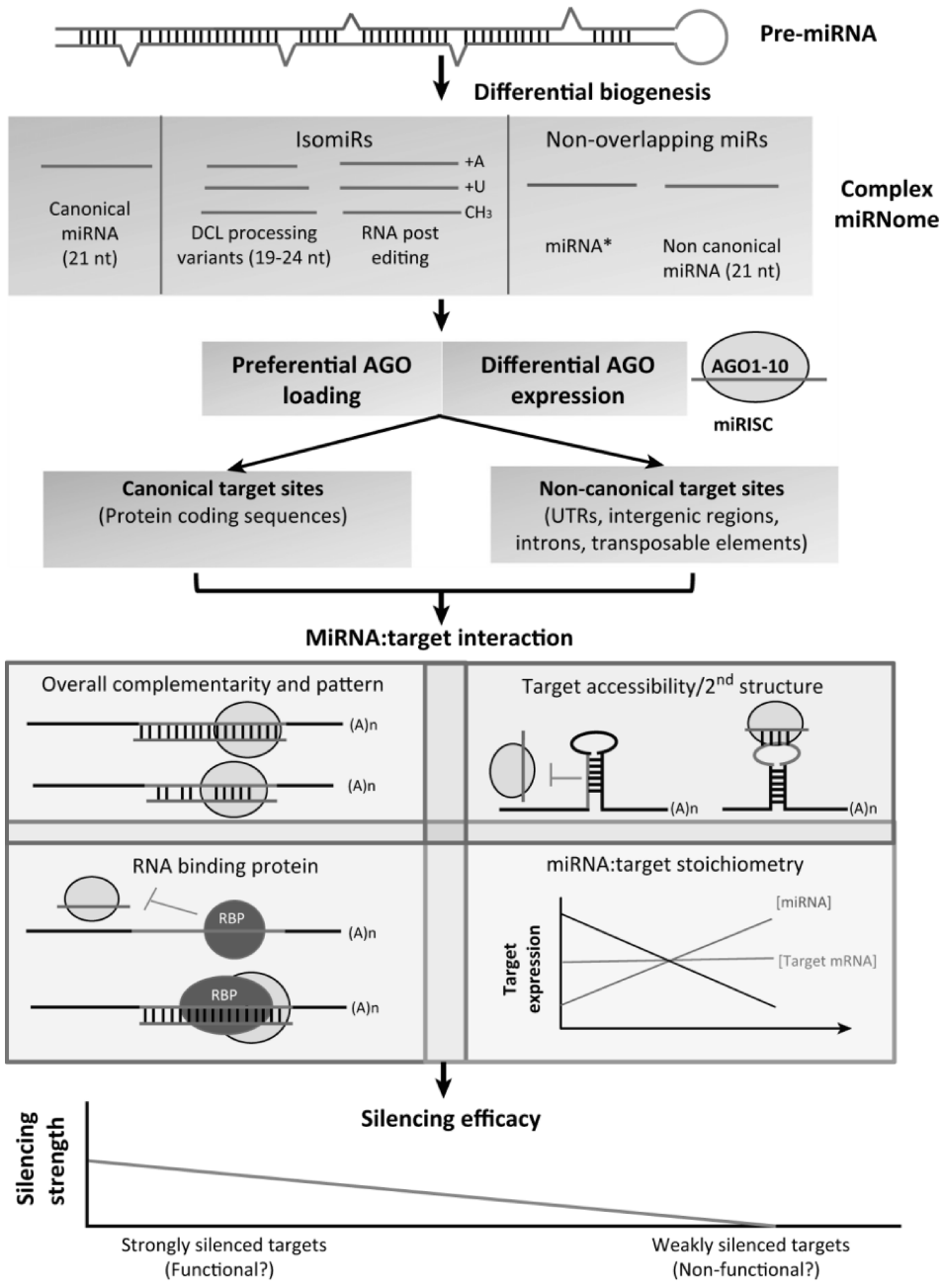


图 2-4.16　植株中 miRNA 产生及 miRNA 调控机制(引自 Li 等,2014)

守性部分),因此,对植物 miRNA 目标的预测要相对简单;5)基因组上的存在位置有差异。动物中 miRNA 广泛存在基因簇现象,即多个 miRNA 由同一个前体加工而来,而植物中 miRNA 多数由单一前体序列加工而来,只有极少数 miRNA,如 miR169 和 miR395 存在基因簇现象。成簇排列的 miRNA 有类似的多顺反子结构,基因表达模式和时期均有同步性;6)加工方式不同,植物中细胞核内编码 miRNA 的基因转录与加工是偶联的,即 miRNA 的形成过程是在细胞核中完成的。成熟的 miRNA 抑或是在细胞核中与类似 *RISC* 的核糖体蛋白

结合形成 *miRNP*，然后被 *Exportin-5* 的同源物——*HASTY* 运送到细胞质中，或者是先被 *HASTY* 运送到细胞质中，再与核糖体蛋白结合形成 miRNP。动物中，细胞核内编码 miRNA 的基因首先在 RNA 聚合酶 II 作用下发生转录，形成长度约为几百个核苷酸的初级转录物，之后在 *Drosha* 酶作用下进一步加工成只含 60～70nt 的 miRNA 前体序列，由转运蛋白 *Exportin-5* 运送到细胞质，之后在 *Dicer* 酶参与下才加工成成熟的 miRNA，形成的成熟 miRNA 与一种类似 *RISC* 的核糖体蛋白结合形成 miRNP 而发挥作用；7）作用机制不同。研究发现，在植物和动物发育过程中，miRNA 与靶 mRNA 结合的程度和部位不同，作用方式也不同。在动物中，多数 miRNA 以不完全互补方式与其靶 mRNA 的 3'端非翻译区（UTR）的识别位点结合，从而阻碍该 mRNA 的翻译来调控基因表达，但不影响 mRNA 的稳定性。植物中的 miRNA 与相应的靶 mRNA 近似完全配对，并且互补区域散布在靶 mRNA 的转录区域内而非仅仅局限于 3'UTR，使得 miRNA 结合到包括编码区域在内的多个位点上去，从而能够直接降解 mRNA，引发基因沉默。

这些区别表明，在生物进化过程中，动植物从最先的共同祖先分化后，各自 miRNA 基因的进化是彼此独立的。miRNA 普遍存在于动植物中，从侧面证明了 miRNA 对于生物个体形成和发展具有重要意义。

同时，我们也发现在编码基因的内含子区域，同样可以转录形成 miRNA，这类 miRNA 叫做 mirtron，该类 miRNA 也能够起到抑制基因表达的作用（Zhu 等，2008）。

基于 miRNA 前体的二级结构，miRNA 前体具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：

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

其中，L 表示前体序列的长度，*MEF* 是最小折叠自由能。

小 RNA 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。互补性的强弱或者说互补碱基的多寡决定了小 RNA 调节的不同机制。与靶基因有较好互补的小 RNA 主要通过对目标 mRNA 的直接切割调节 mRNA 的表达，相反，如果小 RNA 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（图 2-4.17）。植物 miRNA 的靶基因一大类都是转录因子（transcriptional factor，IF），揭示了 miRNA 调节通路的复杂性。

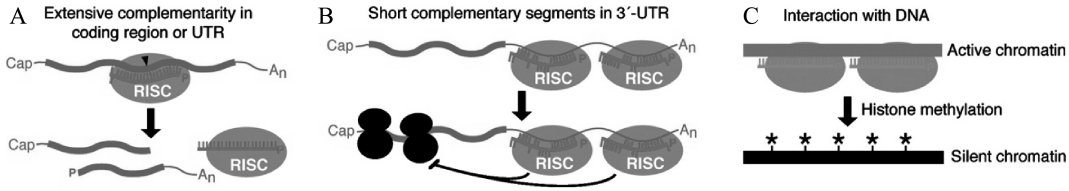


图 2-4.17 小 RNA 调控机制

A，miRNA 和 siRNA 与靶基因互补配对度较高的情况下能够切割 mRNA，黑色箭头是切割位点；B，部分互补配对的情况下，只起到抑制 mRNA 的作用；C，来自异染色质的 siRNA 同样能够抑制 mRNA 的转录（引自 Bartel，2004）

2. 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。

(1) 同源比对

同源比对的方法主要是通过已知保守 miRNA 在不同物种间的序列相似性,进行同源序列搜索预测 miRNA。基于已知 miRNA 序列,搜索公共 DNA 序列数据库,对于全基因组已测序或正在测序的生物,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 sequence)序列和 EST(expressed sequence tag)等表达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,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 值一般设为高于 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'端完全漏掉联配过程,从而漏掉一个可能的结果,尽管这种情况是极少的。另外,基于搜索软件 ERPIN(http://rna.igmors.u-psud.fr/Software/erpin.php)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。同源比对方法应注意以下几点:1) 数据处理过程中一般先通过 BLASTX 搜索蛋白质数据库,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茎环结构,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,那么该位点可为候选的 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/) 以及用于植物 miRNA 预测的 microHARVESTER (http://www - ab. informatik. uni - tuebingen. de/brisbane/tb/index. php) (Dezulian 等,2006)。Artzi 等(2008)开发的 miRNAminer (http://groups.csail.mit.edu/pag/miRNAminer/)可以用 BLAST 进行比对,获取候选序列的一些特征,如二级结构、自由能、保守性等。

(2) 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。研究表明 48%的人类 miRNA 基因和 50%的斑马鱼 miRNA 基因都有成簇存在的现象,一般人类基因组上搜索 miRNA 簇的窗口为 10kb,在斑马鱼中为 3kb。随着高通量测序的成熟,以及植物 miRNA 数据集的完善,在植物中我们也可以用这样的方法寻找 miRNA 簇。不过相对于人和动物,植物中 miRNA 成簇比例要少的多,研究发现在拟南芥中有 25%的 miRNA 成簇,杨树 17%、水稻 22%和高粱 21%的 miRNA 成簇(Zhou 等,2011)。

(3) 基于比较基因组学方法

比较基因组学的基础是相关生物基因组的共线性。如果生物之间存在很近的亲缘关

系,那么它们的基因组就会表现出同线性,即基因序列的部分或全部保守。这样就可以利用基因组之间编码顺序和结构上的同源性,通过已知基因组的信息定位另外基因组中的基因,从而揭示基因潜在的功能、阐明物种进化关系及基因组的内在结构。因此我们可以利用比较基因组学的方法来鉴定非编码 RNA。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在茎的位置上的短序列,然后根据其在两个物种中的保守性来查找保守的 miRNA 基因。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。

(4) 基于高通量小 RNA 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。随着第二代测序技术的成熟和推广,大规模的基因组数据和转录组数据不断产生。经过十几年的发展,根据 miRNA 独有的形成特征以及表达模式,如今利用大规模小 RNA 测序数据从而大规模鉴定 miRNA 的方法已经广为使用。利用高通量测序以及 miRNA 鉴定方法,可以对一个物种的 miRNA 进行从头预测,在数量上和准确性上较传统方法大大提升。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)的水稻 miRNA 研究为例,说明一下大规模小 RNA 测序数据的处理流程。基于 Solexa 测序的原理,测序得到的原始读序都是一端连接着接头(adaptor),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集。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 定位到基因组上(>18nt)。这样我们就得到了一个全基因组的小 RNA 分布图谱。根据全基因组的注释信息,排除掉匹配到重复序列区域和编码区的小 RNA。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。一般标准如下: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 跟其中一条部分读序配对,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) 两条小 RNA 的读序数量需要相差 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 等,2006),如总共不超过 7 个碱基(更严格的话可以设为 4 个碱基)的错配;不超过 3 个碱基的连续错配;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些相对方便的鉴定 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实验验证,包括直接克

隆、Northern、5'-RACE(5' rapid amplification of cDNA ends)等。

二、siRNA 主要特征及计算识别

1. siRNA 和 ta-siRNA 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,并能够激发与之互补的 mRNA 沉默。这个现象在科学界里称为 RNA 干扰现象。RNA 干扰现象是 1990 年由 Jorgensen 研究小组在研究查尔酮合成酶对花青素合成速度的影响时所发现。为得到颜色更深的矮牵牛花而过量表达查尔酮合成酶,结果意外得到了白色和白紫杂色的矮牵牛花,并且过量表达查尔酮合成酶的矮牵牛花中查尔酮合成酶的浓度比正常矮牵牛花中的浓度低 50 倍。Jorgensen 推测外源转入的编码查尔酮合成酶的基因,同时抑制了花中内源查尔酮合成酶基因的表达。1992 年, Romano 和 Macino 也在粗糙链孢霉中发现外源导入基因,可以抑制具有同源序列的内源基因的表达。1995 年,Guo 和 Kemphues 在线虫中也发现了 RNA 干扰现象。1998 年,Andrew Fire 等在秀丽隐杆线虫(*C. elegans*)中进行反义 RNA 抑制实验时发现,作为对照加入的双链 RNA 相比正义或反义 RNA 显示出了更强的抑制效果。从与靶 mRNA 的分子量比分析,加入的双链 RNA 的抑制效果要强于理论上 1∶1配对时的抑制效果,因此推测在双链 RNA 引导的抑制过程中,存在某种扩增效应,并且有某种酶活性参与其中。他们将这种现象命名为 RNA 干扰。

产生 siRNA 的双链复合体,其可有多种来源:主要来源于生物体内存在的反向重复序列,也可能来自自然存在的顺反转录对,由 RNA 聚合酶将单链 RNA 合成双链 RNA,通过病毒 RNA 复制得来的双链 RNA 以及体内存在的大量转录原件(图 2-4.18)。根据其产生机制和功能不同,植物内源 siRNA 被分为四类:异染色质 siRNA(heterochromatic siRNA, hc-siRNA)、反式作用 siRNA(*trans*-acting siRNA, ta-siRNA)、自然反义转录 siRNA(natural antisense transcript-derived siRNA, nat-siRNA)和相位排列 siRNA(phased-siRNA, phasiRNA)。在这里我们以 ta-siRNA 为例进行介绍。

植物基因组演化出几种截然不同的 siRNA,它们在产生机制和功能调节等方面都有所不同,其中大部分的 siRNA 类型(24nt)依赖 RNA 聚合酶 2(*RDR2*)/*DCL3*/*Pol*Ⅳ,并通过 *AGO4* 引导的 DNA 甲基化或组蛋白修饰诱导转录沉默。这一代谢通路往往跟转座子、反转座因子等重复序列相关。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。ta-siRNA 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,比如 *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 等, 2005; Gasciolli 等, 2005; Xie 等, 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 ta-siRNA。

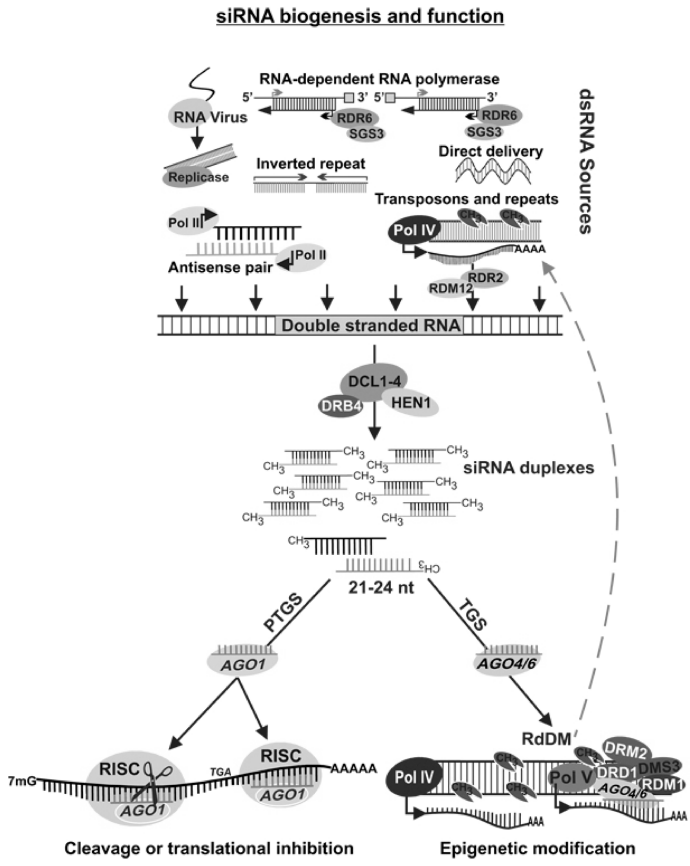


图 2-4.18 siRNA 的来源以及作用方式(引自 Khraiwesh 等,2012)。详见文中说明

2. ta-siRNA 的计算识别

(1) Howell 算法

对于已具备基因组序列的物种,小 RNA 数据(来自不同组织或处理)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 的相位排列这一显著特征,可以找出候选的 *TAS* 基因位点。Howell 等(2007) 设计了一套流程用来查找拟南芥中的候选 ta-siRNA: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:

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

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数(即 n 个相位位置上有小 RNA 存在),k 表示在调查的,8 个相位大小的窗口里面正负链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 2-4.19 中读序图所示,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相位位点。最后,根据 ta-siRNA 受相应 miRNA 调控的现象,在预测的候选相位区域两端预测 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。

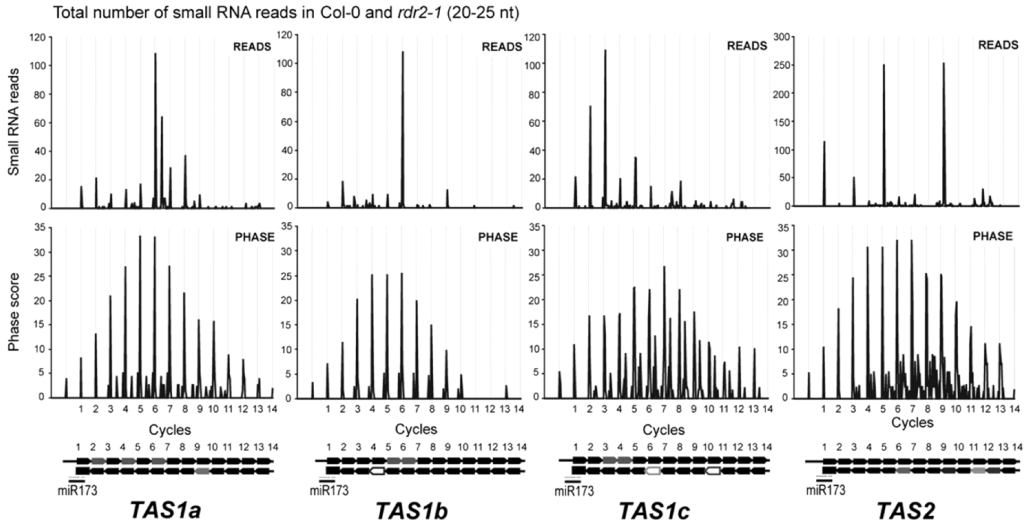


图 2-4.19　四个拟南芥 phasiRNA 基因(*TAS*)位点 21nt 小 RNA 读序分布及相位信号(引自 Howell 等,2007)

(2) Chen 算法

与 Howell 算法类似,Chen 等(2007)也是主要考虑 ta-siRNA 的相位分布特征,并构建了一个统计 *P* 值来查找候选的 ta-siRNA 位点。按照 21nt 一个相位大小,考虑 11 个相位长度的一段区域(11 个相位,总 231bp),*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,表示相位结构越明显(图 2-4.20)。

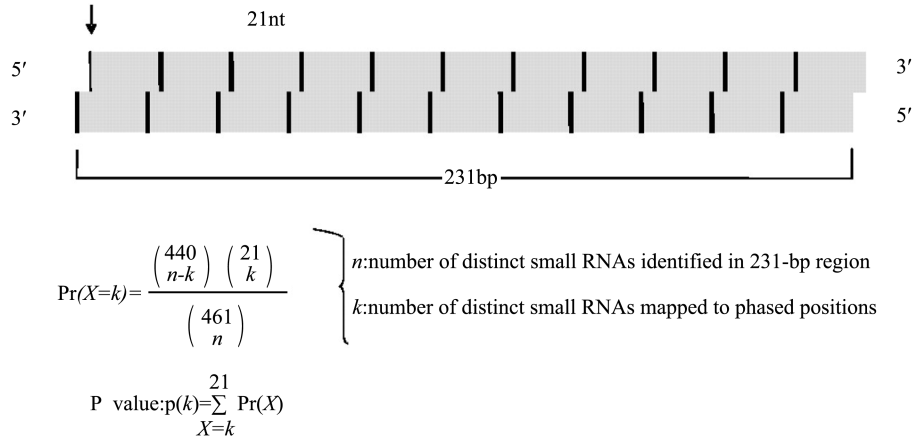


图 2-4.20　相位基因 *TAS* 基因的一个预测算法原理(Chen 等,2007)

垂直箭头表示 phasiRNA 的起始位点,竖线表示跟起始位置相距 21nt 的 siRNA 的相对位置;用来统计分析几何分布 phasiRNA 位点的可靠性公式,其中 *n* 表示在 231 碱基区域内不同小 RNA 的数量,*k* 表示能够比对到相位排列方式的小 RNA 数量

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 等, 2008)。图 2-4.21 给出了部分 *TAS* 位点 21nt 长度读序的 Howell 分布图。

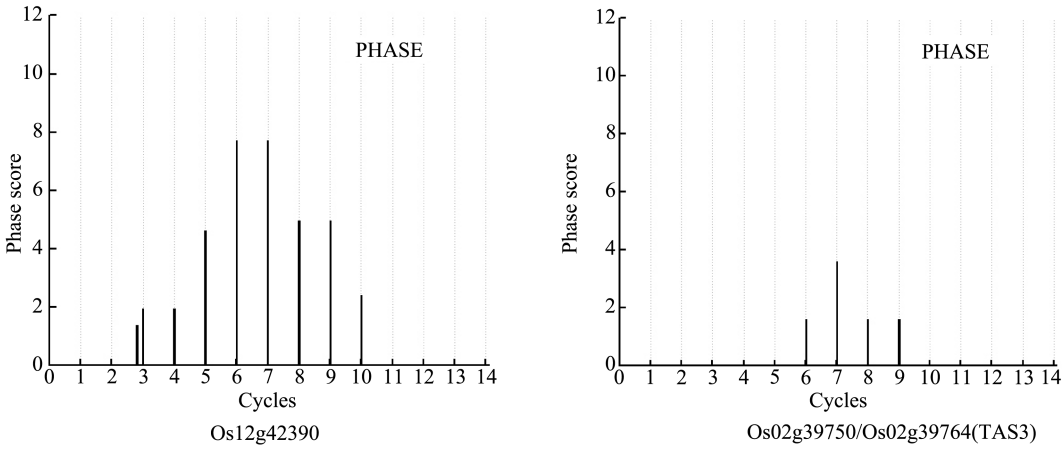


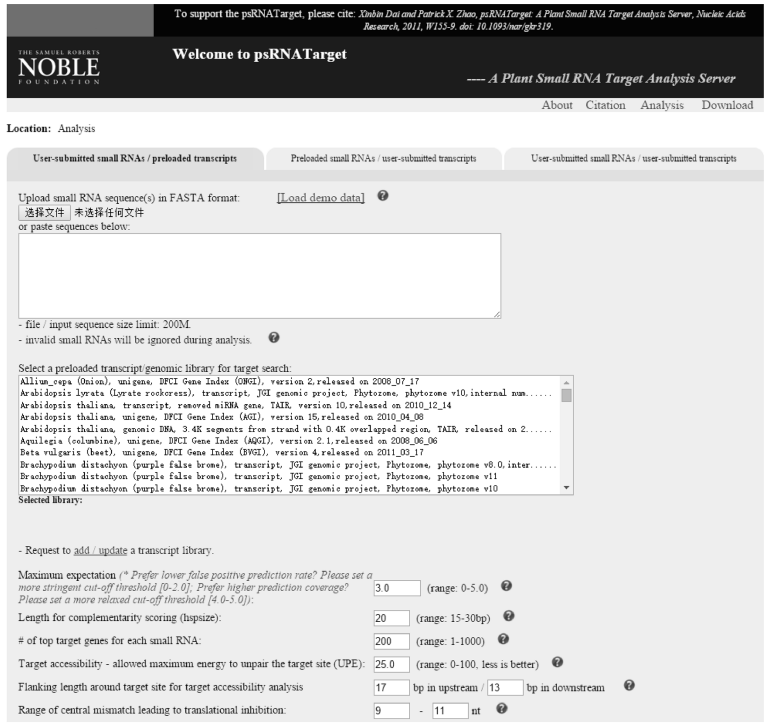
图 2-4.21 水稻 *TAS* 基因 21nt 小 RNA 读序的相位值分布图(来自 Zhu 等,2008)

三、miRNA 和 siRNA 靶基因预测

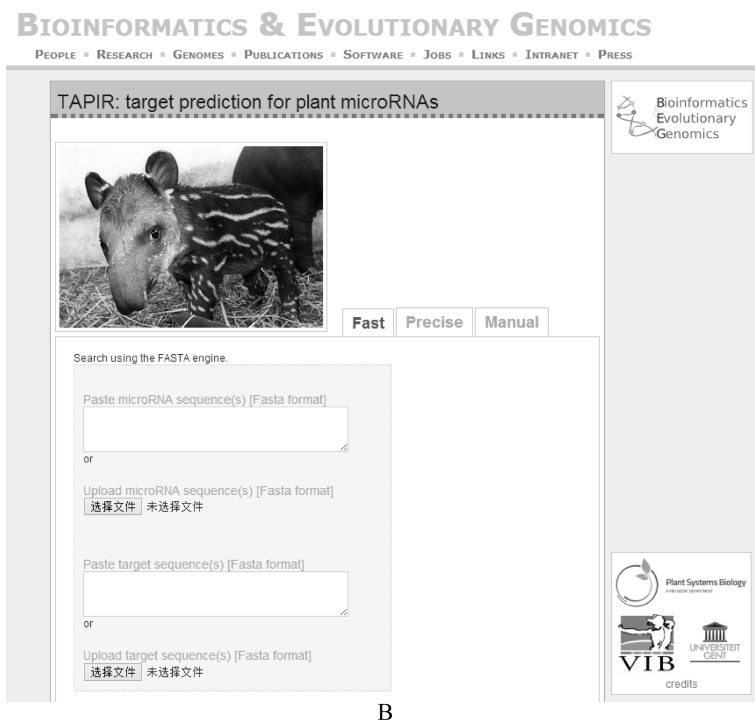
1. miRNA 靶基因预测

动物 miRNA 结合靶基因的机制相对复杂,植物 miRNA 主要通过接近完美的互补配对结合到靶位点,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。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 是一个网络平台,整合了已知的大部分植物 mRNA 和基因数据,可通过提交 miRNA 在植物表达数据中预测是否有靶位点。之后,Zhao 等在 miRU 的基础上开发了 psRNATarget 网络平台(http://plantgrn.noble.org/psRNATarget/)(图 2-4.22A)。它不仅可以预测提供小 RNA 数据(<200Mb)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(< 1000Mb)预测是否可以为已知 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,当然你的基因集大小有一定的限制(<200Mb)。动物 miRNA 靶基因的预测也根据结合位点的不同开发了很多的软件,如 miRecords、PicTar、miRanda、TargetScan、RNAhybrid、microTar、DIANA MicroT Analyzer、MicroInspector 和 TargetBoost 等。

TAPIR(http://bioinformatics.psb.ugent.be/webtools/tapir/)则是另外一个预测植物 miRNA 靶基因的网络平台,有两种预测模式,一种是快速预测模式,还有一种是精确预测模式(图 2-4.22B)。快速预测模式应用的是经典的 FASTA 比对程序,对输入序列反向互补后,与目标 mRNA 比对进行计算。e-value 的阈值设定为 150,搜索的 k-tuple 大小为 1,同时计算了 miRNA-mRNA 复合体的自由能。精确预测模式采用的是 RNAhybrid 搜索引擎,它是对经典 RNA 二级结构预测算法的延伸,采取的是动态规划算法,并对 miRNA 与靶 mRNA 从头到尾可能形成的复合体进行最小自由能计算,严格限制复合体中的凸起或环的长度。利用该搜索引擎,能够精确的计算复合体的自由能,但对 miRNA-mRNA 预测的敏感性会降低,而且从速度上来讲要比之前的一些方法慢得多。



A



B

图 2-4.22　植物 miRNA 靶基因预测工具(A) psRNATarget 和(B) TAPIR 网络服务平台界面

2. 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。因此,miRNA 靶基因的预测软件也同样适用于 siRNA 的靶基因预测。

第三节 长非编码 RNA 鉴定与功能分析

一、线性 lncRNA 鉴定

首先,我们需将 RNA-seq 等转录组数据比对到基因组上(可选用软件 TopHat);利用 Cufflink 等转录组拼接软件可以得到新的转录本,用于之后 lncRNA 的甄别。鉴定 lncRNA 最大的难点是确定转录组的非编码性,主要通过过滤编码蛋白质的转录本实现。第一是通过 ORF 长度判别。对于编码蛋白质的 mRNA 来说,其开放阅读框(ORF)长度一般大于 300nt,也就是说编码的蛋白质链长度大于 100 个氨基酸。因此,若 RNA 序列的 ORF 小于 300nt,其编码蛋白质的可能性会非常小,会被判定为 ncRNA。然而这种武断的判断方法会存在一些问题,比如有些 lncRNA 实际上其假定 ORF 长度要大于 300nt,因此在该标准下他们会被错误的划分为 mRNA。类似的,有些 ORF 长度小于此阈值的 mRNA 也会被误判为 lncRNA。因此,可先根据 ORF 保守性,采用比较基因组学的方法进行甄别。mRNA 的 ORF 具有保守性,即可编码蛋白质的转录本序列与已经注释的蛋白质或蛋白质结构域有序列相似性。因此可以采用 BLASTX、Pfam 等方法,将拼接后得到转录组序列放到蛋白质库进行搜索,根据比对得到的序列相似性得分来判别是否可能编码蛋白。不过值得注意的是,有些 mRNA 进化而来的 lncRNA 也会表现出与蛋白质序列类似序列,从而被错误的判断为 mRNA。目前,我们采用综合性方法进行甄别(supervised machine learning),如利用 CPC、CONC、incRNA 等软件,他们可以通过比较肽链长度、氨基酸构成、蛋白质同源性、二级结构、蛋白质比对或表达等多种特征,建立分类模型。图 2-4.23 中给出了我们鉴定 lncRNA 的大致流程,可供大家参考和学习。

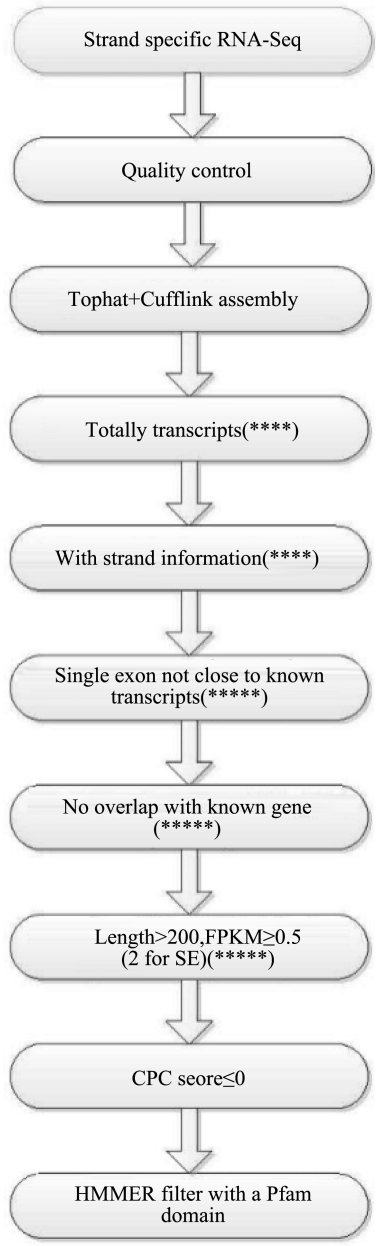


图 2-4.23 植物中鉴定 lncRNA 的流程

二、环化 RNA 鉴定

1. 环状 RNA 定义和特征

环状 RNA 分子是一类由于反向剪切(back splicing)形成的非编码 RNA(图 2-4.24)。尽管生物学家知道环状 RNA 已经有 20 年了(Nigro 等,1991),但是一直认为它们是 RNA 的剪切错误造成的,或者认为它是某些病毒分子如植物类病毒等。直到最近几年,由于高通量测序和生物信息学方法的发展,在动物细胞内发现存在大量的内源性环状 RNA 分子,而这些环状 RNA 根据其在基因组的分布,可以来源于外显子,称为外显子类型的环状 RNA(exonic circRNA)(Zhang 等, 2014),也可以来自于基因的内含子,所谓内含子类型的环状 RNA(intronic circRNA)(Zhang 等,2013)等。

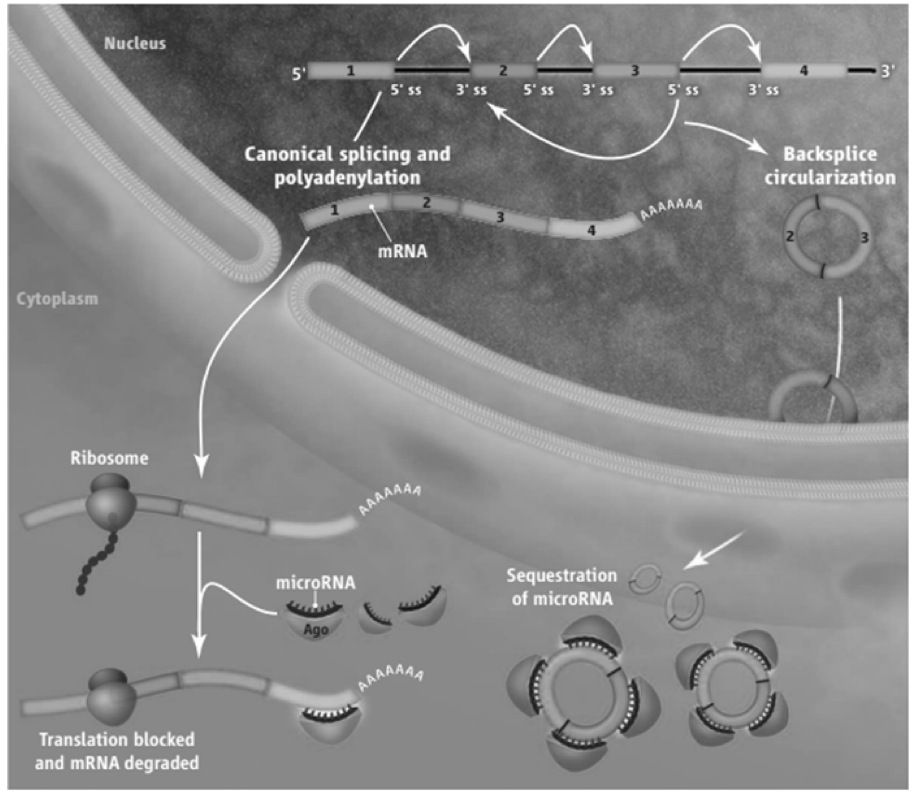


图 2-4.24　环状 RNA 分子的形成(引自 Bolisetty 和 Graveley, 2013)

环状 RNA 具有以下特点:(1)环状 RNA 是一个闭合环状 RNA 分子,存在于生物界的大部分物种之中,包括人类、动物(老鼠,果蝇,线虫等)、植物(拟南芥,水稻等)等等;(2)环状 RNA 是由特殊的可变剪切形成的,通常来源于外显子,存在于细胞质之中;也有来自于内含子的环状 RNA,一般存在细胞核之中;在细胞之中比线性 RNA 分子更稳定,有较长的半衰期,能抵抗 RNAase R 的降解 (Petkovic 和 Muller, 2015);(3)外显子环状 RNA 分子两端经常有较长的内含子,可能与它的形成机制有关系(Jeck 等, 2014);(4)环状 RNA 广泛存在于生物体各个组织和时期,和线性异构体同时存在,一般表达量较低,有时候它的表达量可能超过它的线性转录本,而且有着较强的组织和时期表达特性(Salzman 等, 2013);(5)环状

RNA 也存在高度保守性,有一些则快速进化;(6)有些环状 RNA 能作为竞争性内源分子,富集于 miRNA 的结合位点,起到 miRNA 的海绵作用,从而解除 miRNA 对其靶标的调控作用(Hansen 等, 2013; Memczak 等, 2013);(7)大部分环状 RNA 是非编码分子,不能翻译成蛋白质(Guo 等, 2014)。植物环状 RNA 两端的内含子比线性基因的内含子长,趋势与动物和人类上的研究一致;但与动物和人类上的环状 RNA 相比,其两端的内含子不具有富集的重复序列和反向互补序列。植物上的环状 RNA 表达存在时期和组织特异性,而且一些环状 RNA 的表达与母基因呈现明显的正相关(Ye 等,2015;Lu 等,2015)。

2. 环状 RNA 鉴定

总的来说,预测环状 RNA 的方法可以分为以下三类:(1)候选分子方法;(2)亚读序比对;(3)机器学习类方法。

(1) 候选分子方法。Salzman 等人(2012)首先根据基因组注释信息构建出许多理论上存在的环状 RNA 分子,然后利用 RNA-seq 读序去比对这些假想的环状 RNA 分子,如果读序能刚好比对到反向剪切的切口处,则认为此环状 RNA 分子是存在的。后续研究中,他们改进了算法,根据比对的质量构建了以 FDR 为基础的过滤策略,然而这样的方法需要基因组注释信息,对于没有完全基因组注释信息的基因组则无能为力,而且对于 RNA-seq 的低覆盖度区域也不是很有效(Gao 等, 2015)。

(2) 亚读序比对。这种方法的一般步骤是将不能比对到基因组上的读序(可能来自反向剪切位点)裂开成两段,分别做比对,得到交替比对的情况,最后根据一系列过滤结果得到最终的候选环状 RNA。这里面使用最广泛的工具就是 find_circ (Memczak 等, 2013) (图 2-4.25)。当 RNA-seq 读序比对到基因组上时,线性 RNA 是可以很好的比对回去的,对于环状 RNA 来说,来自成环的剪接位点的读序不能直接比对回基因组上。通过比对筛选出这样的读序,也就是把比对不上的序列都提取出来。然后,将这些比对不上的读序取两头 20bp,变成亚读序。接下来,将亚读序比对到基因组上后,检测这些亚读序是否是环状 RNA 的短序列(anchor)。需要检测的条件如下:GU/AG 在剪接位点的两侧出现;可以检测到清晰的断裂点(breakpoint);只支持 2 个错配(mismatch);断裂点不能在短序列(anchor)2 nt 之外的地方出现;至少有两条读序支持这个反向剪切;比对正确的一个亚读序的位置要比它比对到其他位置的分值高 35 分以上。

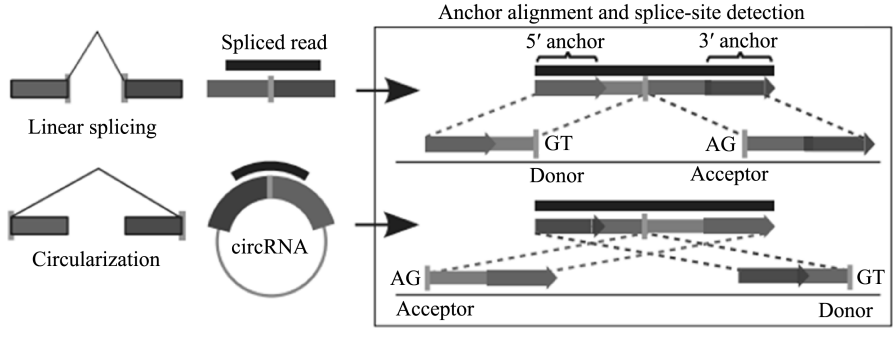


图 2-4.25 环状 RNA 预测工具 find_circ 预测环状 RNA 算法

(3) 机器学习类方法。这种主要是从 *de novo* 拼接的转录本中识别环状 RNA 分子,使用机器学习等方法区分环状 RNA 和线性 lncRNA。首先是提取可以区分的特征,包括保守

信息、序列特征、重复序列、SNP 密度、转录本的开放阅读框(ORF),然后使用机器学习或统计等方法整合这些特征。具体一些方法可参见有关文献 (Pan 和 Xiong, 2015;Szabo 等, 2015)。

环状 RNA 在植物上的研究相对落后于动物和人类,并且目前开发的环状 RNA 预测软件均针对动物或人类基因组,在植物之上预测的结果准确率较低,敏感度不高。利用模式植物水稻和拟南芥,我们和韩斌课题组试图回答环状 RNA 在植物上的存在情况及其特征,并据此开发一套适合植物基因组特点的环状 RNA 预测软件,为植物环状 RNA 的研究奠定基础。针对植物 circRNA 的鉴定,我们提出了改良的算法(Chen 等,2016)。第一步将结合已开发的多种融合 RNA 位点查找软件,如 Tophat-Fusion、STAR-Fusion、MapSplice、segemehl、find_circ 等,希望能找到一个更全面的反向融合 RNA 候选位点,从而最终提高方法的敏感性;第二步设置适宜于水稻等植物基因组特征的参数对第一步获得的位点进行过滤。首先是水稻基因组序列特征,包括基因大小和重复序列(种类、不同大小、所占比例)等,与人类和动物基因组存在明显差异。在过滤步骤中,将考虑这些特征,包括如何避免重复序列产生的错误、水稻 circRNA 的大小范围是多少、水稻剪切信号特别是非 GT-AG 信号、双端测序的读序利用等。

环状 RNA 可视化软件的开发,有利于我们对环状 RNA 的理解和研究。因此我们开发了 circRNA_pocket(Zhang 等,2016),用于对已经鉴定了的环状 RNA 的可视化。通过该软件,可以一目了然环状 RNA 的组成、剪切位置以及可能存在的可变环化情况(图 2-4.26)。

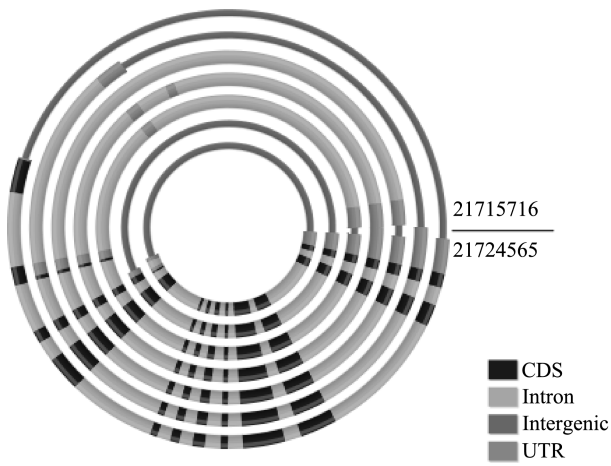


图 2-4.26　环状 RNA 可视化软件 circRNA_pocket 结果示例(Zhang 等,2016)

我们使用去核糖体 RNA-seq 等数据,分别在水稻和拟南芥之中鉴定到 12 037 和 6 012 个环状 RNA,其中水稻的几十个环状 RNA 得到实验验证(Ye 等,2015;Lu 等,2015)。大约 700 个外显子类型环状 RNA 的母基因在水稻和拟南芥之间是同源基因,表明植物环状 RNA 具有一定的保守性(一个例子见图 2-4.27)。拟南芥和水稻中的环状 RNA 表达存在时期和组织特异性,其中 27 个水稻外显子类型环状 RNA 在有磷和无磷营养处理状态下,显著差异表达,而且一些环状 RNA 的表达与母基因呈现明显的正相关。此外,在植物中也存在着可变环化的现象(图 2-4.28),即通过不同的可变剪切方式可以形成不同的环状 RNA 分子。

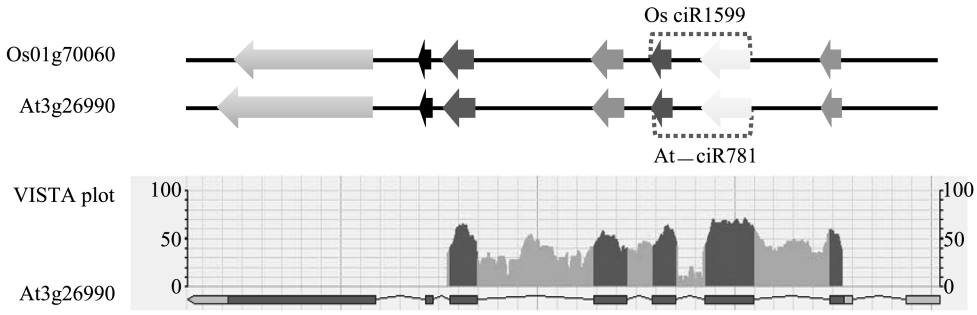


图 2-4.27 一个植物中保守的外显子类型环状 RNA

上图表示环状 RNA 来自于水稻和拟南芥同源基因的相同位置。下图表示以拟南芥 At3g26990 基因做参考的水稻 Os01g70060 VISTA 图,竖线代表同源基因序列的相似程度(引自 Ye 等,2015)

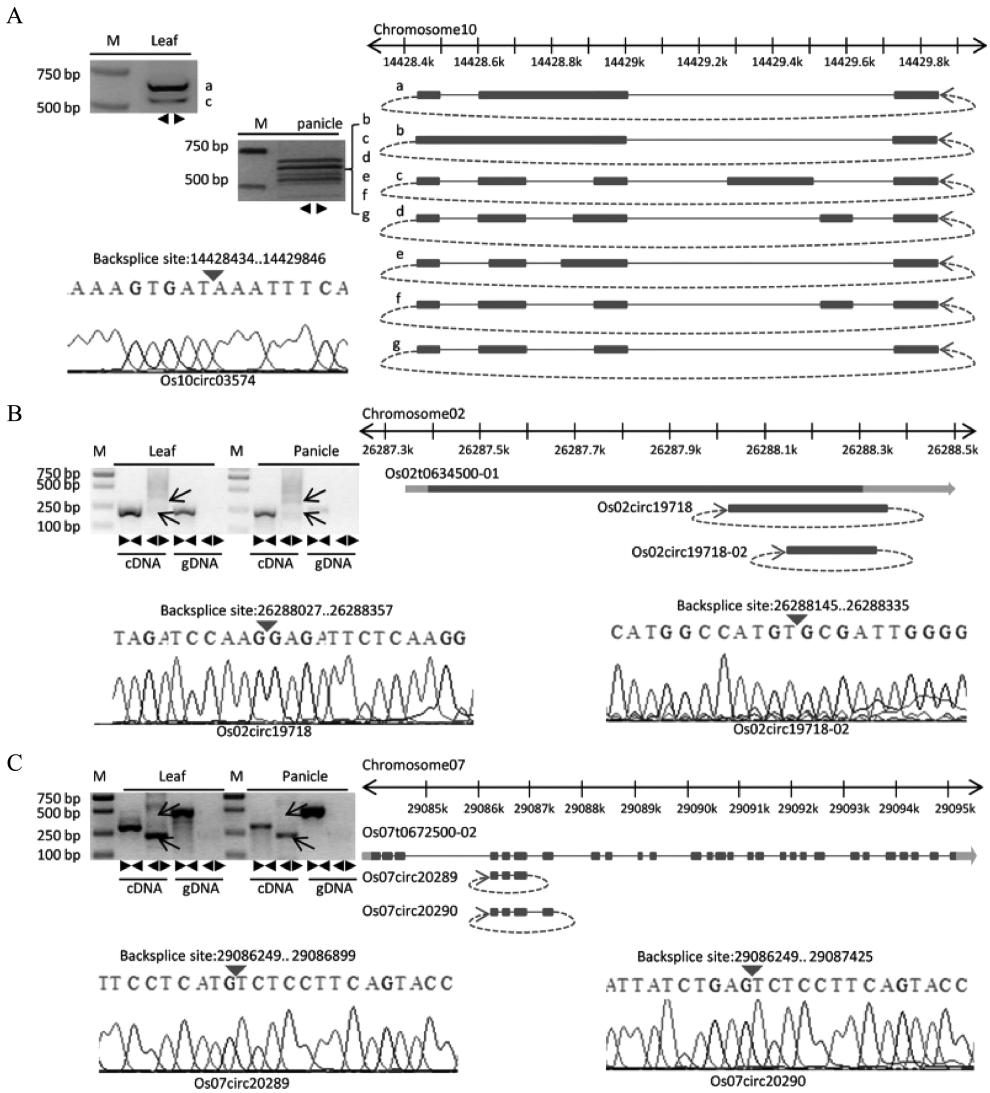


图 2-4.28 水稻中三个可变环化 RNA 例举(引自 Lu 等,2015)

(A)左上图部分是设计反向引物对“Os10circ03574”的 PCR 结果,左下图是利用桑格测序验证反向剪接位点,红色箭头表示的是剪接信号位点;右图是一个环化 RNA 位点 7 个可变环化结果的可视化。长方形表示外显子,实线表示内含子。另外二个例子(B,C)图表示的意思与(A)相同。

三、lncRNA 功能预测

1. lncRNA 与 RNA 分子互作预测

（1）lncRNA 作为 miRNA 的诱捕靶标

miRNA 在动物和植物的生长发育中起着重要的调控作用。为了实现这些调控功能，miRNA 先与 AGO 蛋白形成复合物，再通过碱基互补配对来绑定特定的信使 RNA 序列，导致信使 RNA 的翻译受阻或在特定位点剪切。Franco-Zorrilla 等（2007）在拟南芥中发现了一个由磷酸盐饥饿诱导不编码蛋白的基因 *IPS1*。该基因能够与拟南芥中的 miR399 的序列绑定在一起，但是在 miR399 的剪切位点形成了一个环状凸起结构。因此 *IPS1* 基因无法被切割，却能将 miR399 隔绝起来。而 miR399 真正的靶向基因是 *PHO2*，该基因编码泛素结合酶，对维持细胞内蛋白质的产生和降解的平衡及维持细胞的稳态起着重要作用。*IPS* 基因的存在，使得 miR399 靶向 *PHO2* 基因的活性受到抑制，像类似这种具有抑制 miRNA 功能的长非编码 RNA，定义为 eTM（endogenous target mimics）。在动物中，常常将这类 lncRNA 命名为 miRNA 海绵体或者 miRNA 诱饵（‘miRNA sponge’ or ‘miRNA decoy’），其与植物中的区别主要还是在于 miRNA 与靶标序列结合的方式不同。

利用人工靶向基因模拟序列可以研究特定 miRNA 的功能，目前科学家已经对 miRNA 与其 eTM 的绑定规则有了了解，这为利用生物信息学方法大规模鉴定 eTM 建立了基础。基于已有研究（Meng 等，2012；Bank 等，2012；Wu 等，2013），目前预测 eTM 主要用了如下方法（Ye 等，2014）：首先，用 FASTA3 程序包中的搜索引擎首先获得与 miRNA 反向互补的 cDNA 序列，在搜索过程中允许互补位点有一个较大的凸起；其次，对获得的序列进一步筛选，遵循如下规则：1）在与相应 miRNA 互补配对的中间区域，必须存在一个 3 至 5 个核苷酸的凸起；2）除了中间的凸起区域，所有的错配数要小于 4，且不允许产生连续两个错配；3）除中间区域外，其他地方不允许产生凸起。与 miRNA 预测方法的发展过程一样，随着试验验证结果的不断积累，eTM 预测方法也在不断地被完善和改进。例如新的研究建议遵循以下规则：1）只允许在 miRNA 5’端序列上的第 9 到第 12 个位点出现凸起；2）eTM 中的凸起部分由三个核苷酸组成；3）在 miRNA 5’端第 2 到第 8 个位点要与 eTM 完全配对，但允许 G/U 错配；4）除了中间凸起部分，其余错配数要 $\leqslant 3$，eTM 的长度要大于 200 个核苷酸（Wu 等，2013）。这些改进为今后全基因组上大规模鉴定 eTM 和验证其功能提供了更加完善的方法。

（2）lncRNA 与其他 RNA 分子互作预测

计算预测 RNA-RNA 的互作是基于分子间的作用能，它是通过对两个 RNA 分子的分子内和分子间碱基配对的结合能进行估算。

Busch 等（2008）通过结合分子杂化自由能以及互作时所需自由能，开发了 IntaRNA 算法。如果给出两条潜在互作序列 S^1 和 S^2，其长度分别为 n 和 m。他们之间的靶位点是一对可定义的坐标[x, y]，x 表示靶位点中的起始位点，而 y 表示最后的坐标位点。因此在 RNA 与 RNA 互作过程中，第一条序列 S^1 中的靶位点可以表示为$[i, k]$，S^2 可以表示为$[j, l]$，它们的杂化能可以表示成 $E^{\text{hybrid}}(i, j, k, l)$。随后他们提出递推公式来计算杂化能：

$$H(i,j)=\begin{cases}\min\limits_{p,q}\{E^{loop}(i,j,p,q)+H(p,q)\}, & \text{如果 } S_i^1, S_j^2 \text{ 能成对}\\ \infty\end{cases}$$

其中 $E^{loop}(i,j,p,q)$ 代表成环的 RNA 对中碱基对 (i,j) 和 (p,q) 的自由能。而最终的杂化能可以由 $min_{i,j}\{H(i,j)\}$ 计算获得。

第二点考虑的是 RNA 互作对中靶位点的亲和性，可由以下公式计算：

$$Z_S = \sum_{Q \in S} e^{-\frac{E(Q)}{RT}}, E^{ens}(S) = -RT\ln(Z_s), ED(i,k) = E^{ens}(S_{i,k}^{unpaired} - E^{ens}(s)$$

其中 Z_S 是序列 S 的配分函数，E(Q) 是序列 S 折叠成二级结构时的自由能。$E^{ens}(s)$ 是指折叠成特定结构的 S 序列所拥有的集元能量。$S_{i,k}^{unpaired}$ 是序列 S 中所有未能配对的结构集合。因此 $ED(i,k) \geqslant 0$。结合这两个主要参数，得到最后推导出的公式：

$$H(i,j,k,l) = \begin{cases} \min\limits_{p,q}\{E^{loop}(i,j,p,q) + H(p,q,kl)\}, & \text{如果 } S_i^1, S_j^2 \text{ 能成对} \\ \infty \end{cases}$$

由于计算 ED 值时，需要在两条 RNA 中同时计算从第一个作用碱基到最后一个作用碱基的配对情况。因此，计算杂化能的时候需要扩充到四维矩阵 $H(i,j,k,l)$。这些公式的提出，为以后研究 RNA 与 RNA 之间的互作，提供了重要基础。Wright 等(2013)在 IntaRNA 的基础上，开发了 CopraRNA，其主要改进是结合了比较基因组学方法。最近，Terai 等(2015)提出了一个计算流程(图 2-4.29)，该流程结合了一系列软件(Raccess，TanTan，LAST，IntaRNA 和 RactIP)用于预测人类中 lncRNA-RNA 的互作关系。不过在植物中，像这类预测方法和软件还有待深入研究。

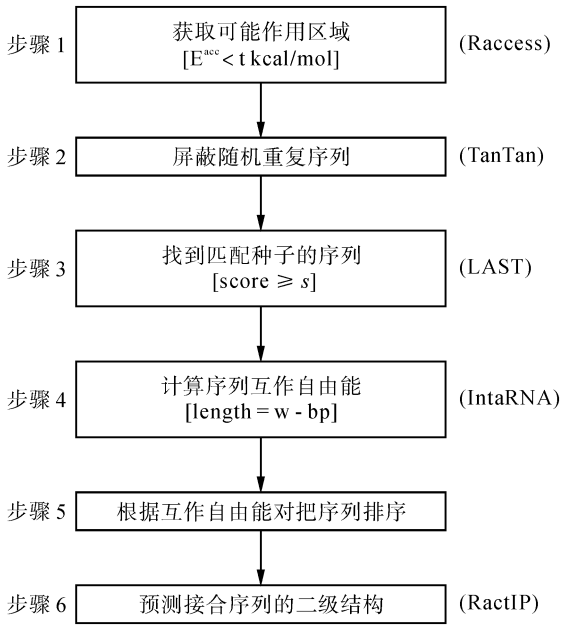


图 2-4.29 预测 lncRNA 与 RNA 互作综合算法流程(引自 Terai 等，2015)

2. lncRNA 与蛋白质分子互作预测

已经有多个算法可以用来预测 RNA 与蛋白质的互作关系。在这些方法中，机器学习方法，比如 Fisher 线性判别分析、支持向量模型以及随机森林等可以用来判断 RNA 与蛋白质是否互作。根据输入数据的不同格式，这些方法可以划分为三类，一类是基于序列的方法，第二类是基于序列和结构的方法，第三类是基于实验数据的方法。

RPI-seq, catRAPID 和 lncPRO 都是基于序列的方法,他们只需要 RNA 以及蛋白质序列作为输入数据。在三个方法之中,catRAPID 和 lncPRO 利用氨基酸和核苷酸的物理化学特征来预测蛋白质和 RNA 的二级结构,作为判别 RNA 与蛋白是否互作的证据。另一种方法 RPI-Pred,在预测 RNA 与蛋白质互作时,不仅用到了它们的序列信息,还需要用到 RNA 与蛋白质的三维结构作为输入数据。通过三维结构可以得到蛋白质结构域以及 RNA 的二级结构,这些可以作为判别 RNA 与蛋白质是否作用成对的特征。相对来说,RPI-Pred 方法准确性更高,然而难度更大,因为目前只有为数不多的 lncRNA 具有结构注释信息。Pancaldi 和 Bahler(2011)开发了基于多种实验数据的预测方法,这些实验数据包括蛋白质定位、RNA 半衰期、核糖体分析以及帕尔斯分析。对于该方法来说,由于可应用的数据更有限,所以难度就来得更大了。可见机器学习方法可能更适应以后的发展趋势。为了评估预测的准确度,还需要实验或其他方法交叉验证。在这方面来看,用生物信息学的方法预测 lncRNA 与蛋白质的互作关系还是一个挑战,值得我们深入思考。

习　题

1. 简述非编码 RNA 类型。
2. 简述基于高通量测序数据的 miRNA 生物信息学预测方法。
3. 什么是长非编码 RNA? 举 2 个例子说明其功能。
4. 什么是相位 siRNA(phasiRNA)基因? 如何用生物信息学方法鉴定?
5. 非编码 RNA 与作物育种有何关系? 举例说明。