microRNA 识别和鉴定方法

王 玫 李思光* 罗玉萍
(南昌大学生命科学学院, 南昌 330031)

摘要 microRNA 是一类长约 22 nt 的内源非编码小分子 RNA, 在线虫、果蝇、家鼠、人体及拟南芥等生物中普遍存在, 并对其生长发育起着重要的调控作用。目前通过实验和计算机的方法在植物和动物中发现了越来越多的 microRNA。通过对识别和鉴定新 microRNA 的主要方法策略的总结可以为 microRNA 今后的研究和发展提供一些思路和启发。

关键词 microRNA; 非编码 RNA; 计算机分析; 实验鉴定

microRNA是一类非编码小分子RNA, 由于其在植物和动物的生命活动过程中起着重要的调控作用而受到越来越广泛的关注。Lee 等[1]在研究线虫的发育缺陷时发现了第一个能调控胚胎后期发育的小分子 RNA *lin-4*, Rhoades 等[2]在线虫中发现了另一个类似的具有转录后调节功能的小分子 RNA *let-7*。其实 *lin-4* 与 *let-7* 只是一类非编码 RNA 家族中的两个成员, 该家族被称为 microRNA 家族, 其基因数量占目前已经被证实了的多细胞动物基因组基因的 1%[3,4]。microRNA 的前体能形成保守的分子内茎环结构, 而成熟的 microRNA 是长约 21~23 nt 的单链 RNA 分子, 它们需要核糖核酸酶III(Dicer)将其从长的前体序列上剪切下来。microRNA 基因以单拷贝、多拷贝或基因簇等多种形式存在于基因组中, 而且绝大部分位于基因间隔区(intergenic region, IGR), 说明它们的转录独立于其他的基因, 具有自身的转录调控机制。microRNA 在生物体内的作用十分广泛, 主要通过与靶序列互补配对而实现其调控功能[5]。它们在植物和动物中的靶序列通常都是一些重要的调控基因, 而这些调控基因参与了生物体生长发育的各个阶段, 从而表明microRNA在生物体的生理过程中起着重要的调控作用[6]。

最早被人们识别的 *lin-4* 和 *let-7* 是通过遗传学方法发现的。通过对秀丽新小杆线虫细胞周期缺陷的分析表明, *lin-4*的缺失是引起细胞周期缺陷的主要原因[1]。*Let-7* 的发现开启了广阔的 microRNA 领域, 因为相对于 *lin-4* 来说, *let-7* 在一个系统分类的大范围内都是保守的[7,8]。这意味着由 microRNA 介导的基因调控可能比预想的更加古老和广泛。遗传学方法还鉴定出了另外 4 种 microRNA: 果蝇中的 *bantam*、*miR-14*、*miR-278* 以及线虫中的 *lsy-6*。由于早期的遗传学方法对于识别 microRNA 来说效率太低, 所以不能成为发现和鉴别 microRNA 的主要方法[9]。

由于 microRNA 在生物体内重要的调控作用, 人们正尝试着利用各种各样的方法在多种生物体内寻找 microRNA。目前识别和鉴定 microRNA 主要采用实验分析和计算机分析两种方法。

1 实验方法识别 microRNA

1.1 建立富集 microRNA 的 cDNA 文库法

目前主要有两种构建富集microRNA的cDNA文库的方法。第一种方法是将一个组织中的全部 RNA 经变性聚丙烯酰胺凝胶电泳分离, 回收19~25 nt RNA 小片段, 随后在这些小片段 RNA 的 3' 和 5' 端连上接头, 逆转录后用与接头对应的引物进行 PCR 扩增, 随后将这些片段克隆至载体以构建 cDNA 文库, 并对其每一个克隆进行测序[10]。第二种方是利用 15% 变性聚丙烯酰胺凝胶电泳从总 RNA 中分离出 16~28 nt 小片段后, 用 poly(A)聚合酶在小片段 RNA 的 3' 端进行多聚腺苷酸化反应。逆转录得到 cDNA 后, 再在 cDNA 的 3' 端进行多聚鸟苷酸化反应, 随后进行 PCR 扩增, 最后克隆建立 cDNA 文库并对克隆进行测序[11]。随着科技的发展, 测序技术也不断进步, 例如 Lu 等[12]利用大规模平行标记测序(MPSS)技术对小片段 RNA 进行高效测序从而改进目前对单个克隆的测序。利用 MPSS 技术, 一个标签序列可以一次

收稿日期: 2006-12-22 接受日期: 2007-03-23
国家自然科学基金(No.30660042)和江西省自然科学基金(No.0630136)资助项目
* 通讯作者。Tel: 0791-8304099, E-mail: siguangli@163.com

性测出一段序列中的17个核苷酸序列, 测序过程包括标签库的建立、微珠与标签的连接、酶切连接反应等步骤[13]。

1.2 基因芯片法

嵌合基因芯片使用高浓度探针体系, 几乎可以覆盖基因组中的每一个核苷酸[14]。这个转录分析体系能够识别新的转录物, 包括microRNA前体序列。然而嵌合基因芯片法对于microRNA分析不是最佳的方法, 因为大多数的探针可能不能与microRNA或siRNA序列配对。目前许多探针和微阵列都是特别为检测已知的microRNA而设计, 包括将microRNA序列用连接序列连接于微阵列, 而这些连接序列在基因组中是无同源性的[15]。这些microRNA基因芯片可以用于检测特殊的序列, 也可以用于确定microRNA在不同组织和不同物种间的表达情况[14], 以及分析与它们对应的靶序列的表达模式。

2 生物信息学方法识别microRNA

直接克隆法获得的大量microRNA特征信息及多种模式生物基因组测序工作的相继完成使得全基因组搜索microRNA成为可能[16]。目前, 人们依据已知microRNA的特征信息及其对靶分子的作用方式建立了多种microRNA的计算机识别方法。

2.1 利用序列和结构的保守性搜索microRNA

利用microRNA序列和结构的保守性在全基因组范围内搜索microRNA是最常用的搜索方法, 即通过对microRNA及其成熟序列的一级和二级结构保守性分析寻找新的候选分子。一些实验室通过同源性搜索, 即利用microRNA在相关物种中的保守性开发设计软件搜索相关物种中的同源分子, 也有一些实验室通过研究和总结microRNA的二级结构特征设计软件搜索新的microRNA候选分子。

一种典型的利用同源性搜索而设计的软件是srnaloop, 它是由哈佛大学医学院的Grad等[17]利用microRNA的序列保守性和结构相似性设计的, 在线虫基因组中搜索到了214个可能的microRNA基因。Srnaloop是一个类似于BLAST的应用程序, 相比于BLAST, srnaloop支持更短的片段并排列成互补的碱基对(包括GU配对)[17]。从srnaloop网站(http://arep.med.harvard.edu/microRNA/)可获取该软件详细信息并下载该软件。最近, 一些实验室开发了依赖比对在基因组中识别已知microRNA同系物的方法, 它们可以在序列和结构层次与已知microRNA进行比对寻找新的microRNA[18,19]。例如Wang等[18]开发了一种依靠序列和结构比对来寻找动物中的microRNA的计算机方法MiRAlign, 它比起之前的同源性搜索来说有两个主要的特点: 一是它能找到相对较远的同系物; 第二, 该软件考虑到更多的序列保守性特点。另外, Nam等[19]开发的ProMiR是一种microRNA序列和结构的统计学联合软件, 是一般microRNA预测方法的补充, 可以识别亲缘关系近或远的同系物。它可以在人类基因组中搜索无论强或弱的保守的茎环结构。ProMiR成功的检测到了与已知microRNA基因不同的新的microRNA基因, 经过RT-PCR的验证, 人们发现这些新的23个基因中有9个(39%)能够在海拉细胞中表达[19]。Nam等[20]开发了ProMiR的升级版ProMiR II, 它整合了更多的microRNA特征, 如自由能数据库、G/C比例、保守性得分、候选序列熵值等。

MiRscan是由Lim等[3,4]依据与已知microRNA相似性开发、设计的。该软件分别在线虫和人类中预测了35个[4]和107个[3]新的microRNA, 其中许多都经过了实验验证。此软件需要先经过两种线虫 *C. elegans* 和 *C. briggsae* 中已经确认的microRNA的驯化, 再利用其他发夹结构与这个驯化体系的相似性寻找这两个物种中其余的microRNA。MiRscan在线地址为http://genes.mit.edu/mirscan/。加利福尼亚大学的Lai等[21]开发的miRseeker不仅利用序列的保守性, 还利用microRNA特殊的保守模式, 例如发夹结构的茎比环状序列更加保守等信息来识别microRNA。它将保守的内含子和基因间隔区序列附加于RNA折叠和评估程序从而识别出保守的含有茎环结构的microRNA前体(图1)。他们利用miRSeeker在果蝇中寻找到了48个microRNA, 其中24个经过了实验的验证。Berezikov等[22]的系统阴影法是对miRSeeker的改进, 他们通过对10个灵长类物种中的122个microRNA进行测序后发现了microRNA基因的保守性特点。这种强保守性表现在microRNA发夹结构的茎上, 而在环上却有很大的变动, 这个特点已被用来跨物种预测新的microRNA。这些软件都是总结了已知microRNA的序列和结构特征后, 将其应用于其他microRNA的预测, 另外Bonnet等[23]证明microRNA的序列具有比tRNAs和rRNAs序列更低的自由能, 说明microRNA二级结构的热力学稳定性也可以用来设计软件搜索microRNA。RNAz软件结合热力学稳定性和二级结构的保守性来预测非编码RNA[24], 此软件成功地应用于在多个物种寻找microRNA分子。

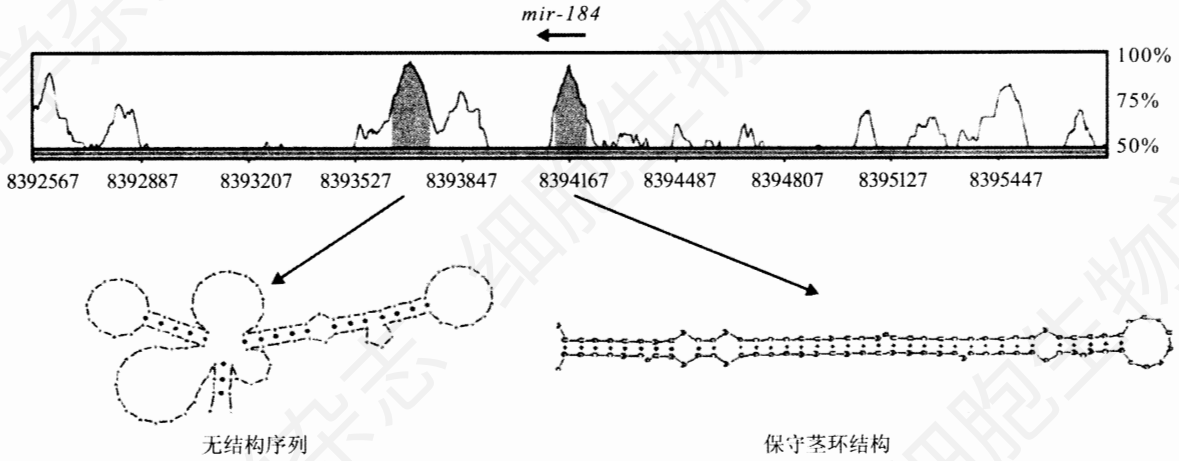

图 1　MiRseeker 预测软件将 microRNA 序列保守与二级结构保守相结合预测 microRNA(部分摘录于文献[21])

2.2　利用靶序列的保守性识别 microRNA

利用同源性搜索 microRNA 主要是在相近物种间搜索同源的 microRNA。如果想要找出未曾发现的新 microRNA 就必须采用其他搜索策略, 例如利用靶序列的保守性搜索 microRNA。在生物体内, 多个 microRNA 可能作用于同一个 mRNA 靶分子, 另一方面, 同一个 microRNA 也可能调控多个靶分子的表达。目前, 一些实验室利用靶序列潜在的保守性作为识别 microRNA 的一个重要的依据[25]。在成熟 microRNA 中, 5' 区域 2~8 位置的 7 个核苷酸被称为种子序列, 它们在与靶 mRNA 结合中起着重要的作用[6]。通过基因组系统比较分析法, Xie 等[26]在 mRNA 的 3'UTR 区发现了 106 个保守的基序, 其中 72 个模体与大约一半的已知 microRNA 的 5' 端相结合组成 6~8 bp 的种子双螺旋。根据这一结果, 他们使用 mRNA 上完整的保守模体在人类中预测到了 129 个新的 microRNA。同样的靶序列预测 microRNA 方法也应用到了拟南芥、果蝇和线虫中。

findMicroRNA 是 Adai 等[27]利用单基因组方法设计的在拟南芥中搜索 microRNA 的软件, 它主要依赖植物 microRNA 与其靶序列紧密的互补性来识别候选 microRNA。随后这个软件要进一步对这些候选 microRNA 成熟序列的相邻序列的互补性进行分析, 使得一个 RNA 内的茎环结构的组成和已知 microRNA 前体结构保持一致。拟南芥基因组中有 29 399 个转录本可以和 27 987 个特定基因座相对应, 用此软件对其分析可以在基因间隔区内识别潜在的 microRNA 前体序列。

实验方法和计算机方法识别 microRNA 各有优缺点。就实验法检测 microRNA 来看, 当 microRNA 在生物体内表达量低或在特定阶段表达时, 就很难检测到 microRNA。一些 microRNA 由于其自身的特性, 包括序列组成或转录后编辑或甲基化修饰而难以克隆[28]。由于大多数内源 siRNA 和 microRNA 的低水平表达产生的低信号使得利用基因芯片技术鉴定 microRNA 也困难重重[13]。近年来, 随着人类基因组和许多模式生物基因组测序工作的完成, 利用计算机搜索 microRNA 的方法逐渐普及, 但是利用生物信息学搜索出的新的 microRNA 中有很多并非是真正的 microRNA, 所以计算机搜索出的 microRNA 还需经过实验验证以最终确定其是否为真正的 microRNA。另外, 对于一些序列上不是十分保守或二级结构特征不明显的新的 microRNA, 使用计算机方法也难以检测到。综上所述, 只有将实验方法和计算机方法相结合起来, 才能在各种物种中准确找到尽可能多的新 microRNA。

3　microRNA 的实验验证

采用计算机方法预测或实验方法获得的 microRNA 还需要进一步进行实验验证, 只有能够在生物体内稳定存在的才能最终确定为 microRNA。而 microRNA 的验证是 microRNA 鉴别工作中的一个瓶颈, 因为 microRNA 表达量较低, 而目前的验证方法又不够灵敏[13]。一般来说, 计算机搜索获得的 microRNA 如果被证明存在大约 22 nt 的成熟序列表达就可被认为是真正的 microRNA。一般验证 microRNA 的方法主要可以分为两类: 能够确定成熟 microRNA 精确末端的方法和能证明其表达但不能识别其精确末端的方法[9]。

目前有多种方法可以用于验证 microRNA, 在这

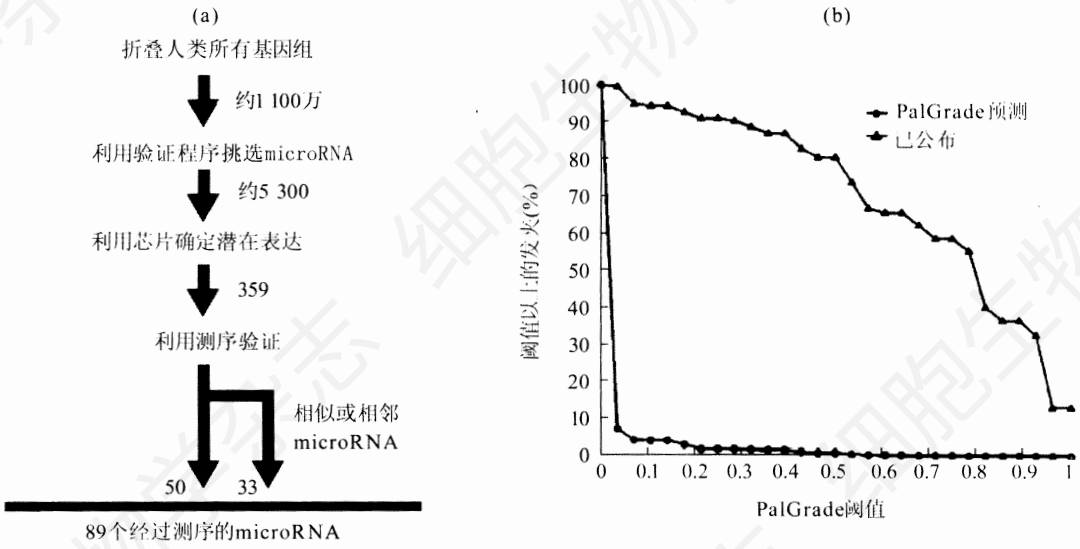

图2 利用PalGrade软件对预测人类新的microRNA(部分摘录于文献[32])

里主要介绍两类验证 microRNA 的实验方法。(1)依赖杂交的实验方法。这类方法首先需要根据预测的 microRNA 的成熟序列设计探针，这些特殊标记的探针可用于 Northern 杂交分析、引物延伸、基因芯片分析和原位杂交等方法验证 microRNA 的表达。Northern 杂交法是目前验证 microRNA 表达应用最广泛且有效的的方法，它能够提供所预测microRNA的有关大小和表达信息。引物延伸分析法能够确定 microRNA 的5'末端，可作为 Northern 杂交的补充[29]。这种方法所用的引物比所预测的microRNA少几个核苷酸，它被杂交至 RNA 样品上，并以 RNA 为模板经逆转录酶延伸，最后通过凝胶电泳检测延伸产物[30]，但是通过这种方法只能识别 microRNA 的5'端。这种引物延伸法的相反的策略叫做RAKE微阵列法，在这种方法中，RNA 在微阵列上与探针杂交，并作为引物经 Klenow 酶延伸，引物能延伸的 RNA 即为表达的候选 microRNA[31]。RAKE 首先是为了研究已知 microRNA 的表达模式所设计的，但是它可以高产量的确定所预测 microRNA 的3'端。(2)依赖克隆的实验方法。例如依赖 PCR 的定向克隆法，该方法使用一对引物，其中一个引物是能与5'接头互补的通用引物，另一个引物与 microRNA 的3'区域相同，这样有利于从小 RNA 文库中扩增出特定的 cDNA 克隆[4]。这种方法只能测定出microRNA的一个末端(5'端)，它具有较高灵敏度但是当不知道 microRNA 成熟序列时却很难操作。另一种定向克隆的方法是根据候选 microRNA 的序列设计探针，这些经生物素标记的探针与RNA样品杂交后经筛选出阳性克隆构建文库并进行测序。这个方法的优点是可以推测出成熟 microRNA 的完整序列[32]。

一些实验室正在研究开发 microRNA 识别和鉴定的综合软件，将生物信息学预测、实验鉴定和表达验证相结合，更为有效和准确的预测microRNA，例如 PalGrade 软件就是集合了生物信息学预测、基因芯片分析以及直接序列克隆为一体的综合的 microRNA 预测软件。PalGrade 由 Bentwich 等[32]开发，具体步骤如下：(1)在整个人类基因组中通过计算机方法搜索发夹结构；(2)对保守的、重复的和蛋白质编码区域的发夹结构进行注释；(3)利用热力学稳定性和结构特征对每一个发夹结构打分，选出高分的已知的 microRNA 发夹和相对低分的基因组发夹；(4)通过高产microRNA基因芯片在不同组织(胎盘，睾丸，胸腺，大脑和前列腺)中测定计算机预测的microRNA的表达；(5)对于在基因芯片中显示出强信号的microRNA基因使用一种新的定向序列克隆和测序的方法进行验证(图2)。他们应用这种软件克隆和测序出89个新的人类的 microRNA。其中53个在灵长类动物中是不保守的。

目前，有许多新的技术例如MPSS等方法逐渐应用于 microRNA 的搜索中，大规模的 microRNA 识别鉴定已指日可待。值得注意的是，单凭一种方法鉴定 microRNA 通常是不准确的，因此需要将几种实验方法或计算机方法相结合起来，才能更全面、更准确地寻找和鉴定出新的 microRNA。

参考文献(References)

[1] Lee RC *et al. Cell*, 1993, **75**: 843
[2] Rhoades MW *et al. Cell*, 2002, **110**: 513
[3] Lim LP *et al. Science*, 2003, **299**: 1540
[4] Lim LP *et al. Genes Dev*, 2003, **17**: 991
[5] 王丽丽等。*细胞生物学杂志*, 2006, **28**: 646
[6] Lewis BP *et al. Cell*, 2003, **115**: 787
[7] Pasquinelli AE *et al. Nature*, 2000, **408**: 86
[8] Slack FJ *et al. Mol Cell*, 2000, **5**: 659
[9] Berezikov E *et al. Nat Genet*, 2006, **38**: S2
[10] Ambros V *et al. Methods Mol Biol*, 2004, **265**: 131
[11] Wang JF *et al. Nucleic Acids Res*, 2004, **32**: 1688
[12] Lu C *et al. Science*, 2005, **309**: 1567
[13] Meyers BC *et al. Curr Opin Biotechnol*, 2006, **17**: 139
[14] Kapranov P *et al. Science*, 2002, **296**: 916
[15] Barad O *et al. Genome Res*, 2004, **14**: 2486
[16] Berezikov E *et al. Hum Mol Genet*, 2005, **14**: R183
[17] Grad Y *et al. Mol Cell*, 2003, **11**: 1253
[18] Wang X *et al. Bioinformatics*, 2005, **21**: 3610
[19] Nam JW *et al. Nucleic Acids Res*, 2005, **33**: 3570
[20] Nam JW *et al. Nucleic Acids Res*, 2006, **34**: W455
[21] Lai EC *et al. Genome Biol*, 2003, **4**: R42
[22] Berezikov E *et al. Cell*, 2005, **120**: 21
[23] Bonnet E *et al. Bioinformatics*, 2004, **20**: 2911
[24] Washietl S *et al. Proc Natl Acad Sci USA*, 2005, **102**: 2454
[25] Park W *et al. Curr Biol*, 2002, **12**: 1484
[26] Xie X *et al. Nature*, 2005, **434**: 338
[27] Adai A *et al. Genome Res*, 2005, **15**: 78
[28] Yang W *et al. Nat Struct Mol Biol*, 2006, **13**: 13
[29] Kim VN *et al. Trends Genet*, 2006, **22**: 165
[30] Seitz H *et al. Genome Res*, 2004, **14**: 1741
[31] Nelson PT *et al. Nat Methods*, 2004, **1**: 155
[32] Bentwich I *et al. Nat Genet*, 2005, **37**: 766

The Methods in Identifying and Predicting Novel MicroRNAs

Mei Wang, Si-Guang Li*, Yu-Ping Luo

(College of Life Sciences, Nanchang University, Nanchang 330031, China)

Abstract MicroRNAs are a class of endogenous small non-coding RNAs about 22 nt long that present extensively in many species such as *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus*, *Homo sapiens* and *Arabidopsis thaliana* etc, which play an important role in the development of various organisms. Several hundred microRNAs have been being experimentally identified and predicted by computational methods both in animals and plants. The summary of the methods in identifying and predicting novel microRNAs can provide some new clues and elicitations for further researching on microRNAs.

Key words microRNAs; ncRNA; computational analysis; experimental identification

Received: December 22, 2006 Accepted: March 23, 2007

This work was supported by the National Natural Sciences Foundation of China (No.30660042) and the Natural Science Foundation of Jiangxi Province (No.0630136)

*Corresponding author. Tel: 86-791-8304099, E-mail: siguangli@163.com