遗传 HEREDITAS (Beijing) 2008 年 6 月, 30(6): 687－696
ISSN 0253-9772 www.chinagene.cn

综 述

DOI: 10.3724/SP.J.1005.2008.00687

microRNA 计算发现方法的研究进展

侯妍妍, 应晓敏, 李伍举

军事医学科学院基础医学研究所计算生物学中心, 北京 100850

摘要: microRNA (miRNA)是近几年发现的一类长度为 ~21 nt 的内源非编码小 RNA, 在植物和动物中发挥着重要而广泛的调控功能。它的发现主要有 cDNA 克隆测序和计算发现两条途径。由于 cDNA 克隆测序方法受 miRNA 表达的时间和组织特异性以及表达水平的影响, 而计算发现可以弥补其不足, 因此 miRNA 的计算发现方法研究受到了广泛的重视。文章对近几年计算发现 miRNA 的研究进展进行了综述, 根据计算发现方法的本质, 将计算发现方法归纳为 5 类, 分别是同源片段搜索方法、基于比较基因组学的预测方法、基于序列和结构特征打分的预测方法、结合作用靶标的预测方法和基于机器学习的预测方法, 并对各类方法的原理、核心思想、优点和局限性进行了分析, 最后探讨了进一步的发展方向。

关键词: microRNA; 计算发现; 同源搜索; 比较基因组学; 作用靶标; 机器学习

Computational approaches to microRNA discovery

HOU Yan-Yan, YING Xiao-Min, LI Wu-Ju

Center of Computational Biology, *Beijing Institute of Basic Medical Sciences*, *Beijing* 100850, *China*

Abstract: microRNAs (miRNAs) are endogenous non-coding RNAs of ~21 nucleotides in length discovered in recent years. They are involved in diverse pathways and play an important role in gene regulation in plants and animals. There are two main groups of approaches to miRNA discovery, which are cDNA cloning and computational identification. Since some miRNAs are expressed at a low level and the expression of many miRNAs has spatio-temporal specificity, it is difficult to find them through cDNA cloning. However, computational approaches can predict the miRNAs specifically expressed or with low abundance, which is complement to cDNA cloning. Computational approaches have hence gained wide attention. In this review, the computational approaches to miRNA discovery were summarized. According to their intrinsic characteristics, computational approaches were categorized into five classes: (1) homology search; (2) prediction based on comparative genomics; (3) scoring candidates using the sequence and structure characteristics; (4) prediction combined with targets; and (5) prediction with machine learning. The principles of each class of the approaches and their advantages and limitations in miRNA discovery were discussed. Finally, the future direction in miRNA discovery was pointed out.

Keywords: microRNA; computational identification; homology search; comparative genomics; target; machine learning

收稿日期: 2007–11–19; 修回日期: 2008–02–01
基金项目: 国家自然科学基金项目(编号: 30500105 和 30470411)资助[Supported by the National Natural Science Foundation of China (No. 30500105 and No. 30470411)]
作者简介: 候妍妍(1983–), 女, 湖南常德人, 硕士研究生, 专业方向：计算生物学。Tel: 010-66932301; E-mail: diana0003@163.com
通讯作者: 李伍举(1966–), 男, 江苏沭阳人, 博士, 研究员, 研究方向：计算生物学。Tel: 010-66931324; E-mail: liwj@bmi.ac.cn
应晓敏(1975–), 女, 江西上饶人, 博士, 副研究员, 研究方向：计算生物学。Tel: 010-66932301; E-mail: yingxm@bmi.ac.cn

miRNA 是近年来发现的一类长度为~21 nt 的内源、单链、非编码小 RNA。目前的研究表明, miRNA 基因由 RNA 聚合酶 II[1,2]或聚合酶 III[3]转录成初级转录物(pri-miRNA), 而后在动物体内经 Drosha 酶剪切形成长度约为 70 nt 的 miRNA 前体(pre-miRNA)[4,5], 在转运蛋白 Exportin-5 的作用下由细胞核内转到细胞质中[6,7], 最后经 Dicer 酶进一步切割产生成熟的 miRNA[8~11]; 在植物体内则由 DCL1(Dicer-like 1 protein)逐步剪切为成熟 miRNA[12], 而后经 HASTY(HST, Exportin-5 的同源蛋白)运输至核外[13,14]。成熟的 miRNA 与 RNA 诱导沉默复合物(RNA-induced silencing complex, RISC)结合, 通过与靶 mRNA 的特定序列结合, 诱导靶 mRNA 剪切或者阻遏其翻译[15]。miRNA 的显著特点是前体折叠形成茎环或类似茎环的二级结构。通过对 pre-miRNA 的基因组定位和注释发现, miRNA 主要位于基因间区或已知转录本的内含子中[16], 较大比例的 miRNA 呈现成簇分布的特点, 且在相近或多物种中保守[17]。尽管目前大部分 miRNA 的确切功能以及其发挥功能的准确调控网络尚在研究之中, 但初步的实验结果表明, miRNA 在生物体内发挥着重要的调控功能, 如调控幼虫发育时序[18,19]、细胞增殖[20]、脂肪代谢[21]、造血系统分化[22]、生殖干细胞自我更新[23]和花的发育[24]等。

miRNA 的发现主要有 cDNA 克隆测序和计算预测两种方法。早期 miRNA 的发现主要通过 cDNA 克隆测序。这种方法直接、可靠, 然而很难克隆出在不同时期表达或只在特定组织或细胞系中表达的 miRNA; 而且由于克隆方法固有的局限性, 也很难捕获表达丰度较低的 miRNA[25,26]。近年来通过计算来预测 miRNA 的方法成为 miRNA 发现的另一条重要途径, 其优点是不受 miRNA 表达的时间和组织特异性以及表达水平的影响, 从而可以弥补 cDNA 克隆测序方法的不足[27]。根据预测方法的本质, 计算预测方法可分为 5 种类型, 分别是同源片段搜索方法、基于比较基因组学的预测方法、基于序列和结构特征打分的预测方法、结合作用靶标的预测方法和基于机器学习的预测方法, 下面分别对各类方法的原理、特点和局限性进行论述。

1 同源片段搜索方法

同源片段搜索方法实现简单, 是最早采用的计算发现方法。这类方法的共同点是采用序列或结构比对算法在相同或相近基因组中搜索已知 miRNA 或 pre-miRNA 的同源片段。由于 miRNA 的显著特征是前体折叠形成茎环结构, 因此单纯采用序列比对算法搜索到的同源片段还不足以判断为可能的 miRNA, 还需要根据二级结构特征进行筛选; 而采用序列和结构比对相结合的同源片段搜索方法则可以找出满足 miRNA 结构特征的同源片段, 可初步判断为可能的 miRNA。

由于 pre-miRNA 的序列较长, 且形成茎环结构, 因而目前大部分同源片段搜索方法均是在基因组中搜索 pre-miRNA 的同源片段。Weber 等[28]用 BLAT[29]在人和小鼠基因组中交叉搜索已知小鼠、大鼠和人 pre-miRNA 的同源片段, 而后根据同源片段的二级结构、miRNA 成熟体所在区域的配对数以及成熟体的保守性进行筛选, 最后得到 35 个可能的人 pre-miRNA 和 45 个可能的小鼠 pre-miRNA。Dezulian 等[30]则用 BLAST[31]在 NCBI EST 数据库中搜索已知 pre-miRNA 的同源片段, 而后根据同源片段是否包含已知成熟 miRNA、同源片段的二级结构以及已知成熟 miRNA 是否位于茎区来筛选候选 miRNA。由于 pre-miRNA 的茎区更为保守、环区则容易随着进化距离的增加而趋于不保守[32,33], 因此采用不区分比对区域的同源序列比对算法, 如 BLAST、BLAT, 只能搜索到与已知 pre-miRNA 在各个区域序列同源性均较高的片段, 对于由于进化距离较远、而在环区序列同源性较低的同源片段就难以发现。为了克服这一弱点, Legendre 等[34]采用基于谱的序列比对表示和搜索方法 ERPIN[35]来发现新的 pre-miRNA。他们首先对所有动物和植物的 pre-miRNA 进行了多序列比对, 根据比对情况分出 miRNA 家族, 而后提取同一家族 pre-miRNA 的一致二级结构, 最后采用 ERPIN 表示 miRNA 家族的一致二级结构、并在基因组中搜索同源片段。这种方法在序列同源的基础上增加了结构同源性, 因而可以发现进化距离较远、但结构同源性较高的 pre-miRNA。他们根据 miRNA 家族的一致二级结构、用 ERPIN 在 20 余种动物基因组中搜索同源片段, 发现了 265 个可能的 pre-miRNA, 较用 BLAST 发现的 miRNA 要多出 17%。然而, 采用 ERPIN 表示和搜索同源序列的前提是 miRNA 必须要有较多隶属于同一家族的成员, 对于那些缺乏家族成员的 miRNA 则无法搜索同源序列。Wang 等[36]对这一问题做了进一步的改进。他们首先用 BLAST 搜索已知 pre-miRNA 的同源片段,

而后根据同源片段的最低自由能、miRNA 成熟体在同源片段中的位置进行筛选, 最后计算 pre-miRNA 与同源片段的结构比对, 并给出两条序列二级结构的相似性度量, 根据标准化后的度量进行筛选。利用该方法, Wang 等人在冈比亚按蚊(*Anopheles gambiae*)基因组中发现了 59 个可能的新 miRNA 基因。经比较, Wang 等人提出的这一方法的敏感性要优于采用 BLAST 和 ERPIN 的方法, 特异性则优于采用 BLAST 的方法, 与 ERPIN 方法相当。

与上述方法不同, Li 等[37]通过在基因组中搜索成熟 miRNA 的同源片段来预测新的 miRNA。他们首先在拟南芥和水稻中采用 BLAST 搜索这两物种已知成熟 miRNA 的同源片段, 而后根据同源片段及其侧翼序列的二级结构、与已知 pre-miRNA 的同源性进行筛选, 最后找出 20 个可能的拟南芥 miRNA 和 40 个可能的水稻 miRNA。Qiu 等[38]、Xie 等[39]和 Zhang 等[40]采用了类似的方法, 不同的是, 他们分别在棉花、欧洲油菜和多种植物 EST 序列中搜索可能的 miRNA。

就本质而言, 上述同源序列搜索方法均需要以已知的 miRNA/pre-miRNA 为参照, 搜索与已知 miRNA/pre-miRNA 在序列上和结构上同源的 miRNA/pre-miRNA, 对于不与已知 miRNA/pre-miRNA 同源的 miRNA/pre-miRNA 则无能为力。

2　基于比较基因组学的预测方法

随着对 miRNA 功能研究的深入, 研究者发现部分 miRNA 参与很多基本而重要的生理过程, 这提示其很可能在进化过程中保守。而且, 对 miRNA 进行基因组分析也发现, 较大比例的已知 miRNA 位于基因组中进化保守的区域。因此, 基于比较基因组学搜索在多物种中保守的 miRNA 成为一种可行而有效的方法。

一种基于比较基因组学预测 miRNA 的思路是先在一个物种基因组中根据结构和序列特征找出可能的 pre-miRNA, 而后与其他物种基因组比较, 判断其序列和结构是否保守。Lim 等[41]首先在秀丽线虫(*C. elegans*)基因组中找出可能的 pre-miRNA 片段, 而后与 *C. briggsae* 序列进行比对, 找出同源 pre-miRNA, 最后按照 miRNA 成熟体区域的配对概率总和等 7 个特征对茎环结构对进行打分。利用该方法, Lim 等人预测出 35 个可能的新 miRNA, 其中 16 个得到实验验证。Lim 等[42]还采用相同的方法在人、小鼠及河豚基因组保守的茎环结构对中预测了 188 个可能的 pre-miRNA, 包含 109 个已知 miRNA 中的 81 个, 精度为 74%。Grad 等[43]同样先从线虫基因组的基因间区中找出可能的 pre-miRNA 片段, 而后在果蝇基因组中寻找同源 pre-miRNA, 之后将线虫可能的 pre-miRNA 在 *C. briggsae* 基因组中搜索同源片段, 得到 81 个可能的 pre-miRNA, 其中 6 个为已知的 pre-miRNA; 同时将果蝇 pre-miRNA 在人基因组中寻找同源片段, 要求线虫、果蝇和人 3 物种对应的 pre-miRNA 中可能的成熟 miRNA 在同一端, 得到 40 个可能的 pre-miRNA, 其中 6 个为已知的 pre-miRNA。Wang 等[44]采用类似的方法在拟南芥中预测 miRNA。他们首先在拟南芥基因组中寻找可能的 pre-miRNA, 而后在水稻基因组中寻找高度同源且折叠形成类似茎环结构的片段, 得到 95 个可能的 miRNA, 其中包括 12 个已知的 miRNA, 并通过实验验证了新发现的 83 个 miRNA 中的 25 个。

另一种基于比较基因组学预测 miRNA 的思路是先通过比较两物种的基因组找出保守区域, 而后在保守区域中根据结构和序列特征搜索可能的 miRNA。Bonnet 等[45]首先对拟南芥基因间区序列与水稻基因组进行序列比对, 得到保守短片段, 而后以保守短片段为中心在两物种基因组中寻找可能的 pre-miRNA, 最后得到 91 个可能的 miRNA, 其中 58 个 miRNA 有潜在的靶标。另一种代表性的预测方法是 Lai 等[32]提出的 miRSeeker。他们首先通过比较果蝇 *D. melanogaster* 和 *D. pseudoobscura* 的基因组, 得到保守的内含子和基因间区, 而后在保守序列中寻找可能的 pre-miRNA, 根据奖赏配对、惩罚内部环/膨胀圈等不配对的打分矩阵对茎环结构片段进行打分。Lai 等通过观察 24 对已知 *D. melanogaster* pre-miRNA 与相应 *D. pseudoobscura* 同源片段的序列比对, 发现 pre-miRNA 的茎区受进化压力的影响而更为保守, 环区则更富于变化。利用这一保守模式, Lai 等对保守的茎环结构进行了筛选, 最后得到约 200 个可能的 pre-miRNA。此外, 他们还在进化距离较远的物种(如昆虫、线虫、脊椎动物)中寻找 *D. melanogaster* 保守茎环结构的同源片段。通过实验, 他们验证了 24 个新的 miRNA, 其中包括 20 个在 3 物种中保守的 miRNA 和 4 个果蝇特异的 miRNA。Berezilov 等[33]在 miRNA 保守模式方面做了更为深入的研究, 并基于新发现的保守模式对人 miRNA 进行了预测。他们采用种系发生投影方法

(Phylogenetic shadowing)对 10 个灵长类物种的 122 条 miRNA 序列的侧翼序列进行了多序列比对, 发现 pre-miRNA 的茎区相对保守, 环区更不保守, 且 pre-miRNA 侧翼序列的保守性相对于 pre-miRNA 而言出现骤降。利用这一保守模式, 他们在人/小鼠和人/大鼠的保守谱中(conservation profile)中搜寻满足相应模式的片段, 而后根据折叠形成茎环结构和随机检验 *P* 值[46]进行筛选, 得到 976 个可能的 miRNA。通过实验, 他们验证了其中的 16 个。

从单纯的序列和结构保守发展到通过已知 pre-miRNA 找出保守模式、进而利用该模式搜寻 miRNA, 基于比较基因组学的预测方法使得 miRNA 的计算发现有了很大的进展。相对于同源片段搜索而言, 基于比较基因组学的预测方法能够找到不与已知 miRNA 同源的新 miRNA, 具有更大的优越性。然而, 该类方法由于仅在两个或多个物种基因组的保守序列中预测 miRNA, 限制了其对非保守 miRNA 的发现, 如病毒 miRNA。而且, 由于大部分物种基因组之间的进化距离较远, 通过比较基因组学的方法也难以发现仅出现在某些进化距离很近的物种中的 miRNA。

3 基于序列和结构特征打分的预测方法

随着 miRNA 发现的不断深入, 研究者不仅在高等真核生物基因组中发现了新的 miRNA, 同时还在感染高等真核生物的 DNA 病毒基因组中发现了 miRNA[47~51]。序列分析发现, 病毒 miRNA 之间的序列相似性很低, 仅有 8 例非洲淋巴细胞瘤病毒(EBV)的 miRNA 与猕猴淋巴隐病毒(RLCV)的 miRNA 同源, 其他病毒 miRNA 很少有同源序列[52]; 而且, 对于很多病毒而言, 它们只存在进化距离很远的直系同源成员, 这使得通过同源片段搜索或比较基因组学方法预测病毒 miRNA 变得相当困难, 甚至是不可能。类似的问题也发生在一些高等真核生物上, 如到目前为止, 具有完整基因组序列且与拟南芥进化距离相对最近的物种是水稻, 而水稻与拟南芥基因组早在 2 亿年前就已经分化[53]; 具有完整基因组序列且与人进化距离相对最近的物种是黑猩猩, 而黑猩猩与人基因组也早在 4 百万年前就已经分化[54]。同源片段搜索或基于比较基因组学预测仅能发现一些在进化距离较远的物种基因组中保守的 miRNA, 而难以发现物种特异的 miRNA。根据已知 miRNA 在序列和结构上的特征、对全基因组中可能折叠形成茎环结构的片段进行筛选成为发现这些非同源、物种特异 miRNA 的行之有效的途径。

Sullivan 等[55]采用根据结构特征打分的方法对猿猴病毒 40(SV40)的 miRNA 进行了预测。他们在 SV40 病毒的基因组中找出形成茎环结构的片段, 而后根据奖赏配对、惩罚膨胀圈和末端环的打分规则对茎环结构片段打分。得到的分值与最低自由能相乘作为每个茎环结构片段的最后分值。采用这一方法他们在 SV40 的基因组中预测出了 2 个可能的 miRNA, 其中 1 个得到实验验证。Grundhoff 等[56]在文献[55]方法的基础上做了改进, 使这一方法能够适用于更大的病毒基因组。他们做的改进主要在两个方面, 一是提高了给配对的奖赏分值和给膨胀圈、末端环的惩罚分值, 另一个改进是增加了对预测成为 pre-miRNA 的片段的分组。采用这种预测方法、结合基因芯片检测, Grundhoff 等发现了 10 个已知的卡波济肉瘤相关疱疹病毒(KSHV)miRNA, 1 个新的 KSHV pre-miRNA 和 18 个新的 EBV pre-miRNA。

与上述打分方法相比, Cui 等[57]采用了一种非常简单的方法对单纯疱疹病毒Ⅰ型(HSV-1)的 miRNA 进行预测。他们采用类似文献[44]的方法取候选片段, 然后根据成熟 miRNA 区域的 GC 含量、复杂度和茎环结构筛选, 得到可能的 pre-miRNA。通过这一方法, 他们预测出了 13 个可能的 HSV-1 pre-miRNA, 编码 24 个可能的成熟 miRNA。通过实验, 他们验证了其中的 1 个。

由于病毒基因组较小且紧致, 形成茎环结构的背景片段数量很小, 因此, 尽管上述基于序列和结构特征打分的预测方法较为简单, 但仍然在病毒 miRNA 预测中取得了较好的效果。然而, 对于高等真核生物而言, 基因组规模在千万至上百亿个碱基对, 形成茎环结构的片段达几万乃至几百万个, 从中通过序列和特征打分挑选出可能的 pre-miRNA 则是一个巨大的挑战。为了能够从大量背景茎环结构片段中挑选出真实的 pre-miRNA, Bentwich 等[58]引入了基于有向图寻找最优分割路径的方法。对每个茎环结构片段, 他们根据序列重复性、最低自由能与随机序列自由能的 Z 值等 11 个序列和结构特征进行量化, 并将茎环片段每个特征的分值都用阈值向量离散化, 这样, 所有茎环结构片段都可以表示为 11 维超空间中按照每个特征阈值向量分割而成的网格中的点。为寻找最优分割路径, Bentwich 等人随机选取了 10 000 非蛋白编码区的茎环结构片段作为背

景片段, 与真实的 pre-miRNA 一起量化, 而后基于有向图搜索最优分类性能的路径。按照这一方法, Bentwich 等在人类全基因组中预测 pre-miRNA, 得到 434 239 个候选片段, 其中包括 86%的真实 pre-miRNA。通过基因芯片检测, 他们发现了 89 个新的 miRNA, 其中 53 个仅在灵长类物种中保守。Li 等[59]则将预测 miRNA 的范围限定在人类 EST 和内含子序列中, 这大大减少了背景茎环结构片段的数量。根据 GC 含量、最小自由能等 4 个序列和结构特征进行筛选, 他们最后预测出 208 个可能的 pre-miRNA, 其中包括 52 个已知的 pre-miRNA, 占已知总 pre-miRNA 的 60% (52/86)。

基于序列和结构特征打分的预测方法由于没有依赖同源序列和多物种中的保守序列, 因而可以找出不与已知 miRNA 同源和物种特异的 miRNA。然而为了从大量背景茎环片段中选出真实的 miRNA、同时降低假阳性, 这类方法往往用异常严格的序列和结构标准筛选候选片段, 因而可能遗漏大量的 miRNA。

4　结合作用靶标的预测方法

miRNA 的作用机制是通过与靶基因的碱基互补配对来发挥调控功能。虽然其序列中的部分碱基可能在进化过程中发生改变, 然而 miRNA 与相应靶序列的碱基互补配对模式从根本上来说却仍然具有严格的保守性, 换言之, miRNA 与其靶序列的相互作用较 miRNA 序列本身更为稳定。部分研究者利用这一特点, 开展了 miRNA 预测的工作, 发现了一些新的 miRNA。

Rhoades 和 Bartel[60]将作用靶标作为筛选候选 miRNA 的最后条件, 在拟南芥和水稻基因组中发现新的 miRNA。他们首先在拟南芥和水稻基因组中分别寻找可能的 miRNA(长度设定为 20 nt, 称为 20 聚体), 而后将两物种的 20 聚体分别在对方 20 聚体集合中搜寻同源片段, 之后在各自物种的基因组中搜寻这些同源 20 聚体的茎环结构前体, 随后分别在对方物种的茎环结构前体中搜索同源片段, 最后, 分别在拟南芥和水稻 mRNA 中搜索 20 聚体的作用靶标。为了提高特异性, 要求靶 mRNA 在另一物种中有同源 mRNA 且结合部位序列在两同源 mRNA 中保守。通过上述方法, Rhoades 和 Bartel 预测了拟南芥中的 24 个 miRNA 家族, 其中 11 个家族为已知的 miRNA 家族, 敏感性达到 85%, 13 个为新发现的 miRNA 家族, 其中 7 个家族中的 23 条 miRNA 经过实验验证。

Xie 等[61]采用了与文献[60]完全不同的思路, 作用靶标不再作为筛选候选 miRNA 的最后条件, 而是一开始就根据 3 ′UTR(Untranslated region)中的保守 8 聚体逆向搜索可能与之作用的 miRNA。他们首先对人、小鼠、大鼠和狗基因组中的 3 ′UTR 序列进行比对, 根据保守性分值, 找到 72 个高度保守的 8 聚体, 随后在 4 物种保守的基因组序列中搜索这些 8 聚体的反向互补片段, 并要求这些片段在 4 个物种中的侧翼序列均折叠形成茎环结构、且最小自由能小于–25 kcal/mol。通过筛选, 他们找到 242 个可能的 miRNA 基因, 其中 113 条编码已知的 miRNA。他们对 129 条候选 miRNA 基因中的 12 条进行了实验验证, 发现了其中的 6 条。

同样是从作用靶序列逆向搜索 miRNA, Adai 等[62]不要求作用靶序列在多物种中保守, 仅在筛选过程中要求 miRNA 和对应的 miRNA*(pre-miRNA 中与 miRNA 配对的序列片段)在相近物种中保守。他们首先在拟南芥基因间区中搜索与已知转录本匹配的短片段, 按照奖赏配对的原则打分, 根据分值筛选可能的 miRNA; 而后以基因间区中可能的 miRNA 为中心寻找可能的 pre-miRNA, 并在其中搜索相应的 miRNA*, 按照奖赏配对和惩罚间隔的原则打分, 根据分值筛选 miRNA 与相应的 miRNA*; 之后, 根据 miRNA 与转录本中匹配片段的分值、miRNA 与 miRNA*的分值和 miRNA 与相应 miRNA*界定片段的自由能, 对每个转录本对应的所有可能的 miRNA 进行筛选。为减少候选 miRNA 的个数, Adai 等还要求候选 miRNA 在水稻基因组中有完全相同的同源 miRNA, 并且同源 miRNA 和相应的 miRNA*也通过上述方法的筛选。最后得到 236 个包含候选 miRNA 与相应靶序列的簇。对 13 个候选 miRNA 进行了实验验证, 检测到其中的 8 个。

尽管 Xie 等[61]和 Adai 等[62]均根据作用靶标预测 miRNA, 然而为了减少候选 miRNA 的个数, 他们或者要求作用靶序列和候选 miRNA 在多物种中保守, 或者要求 miRNA 和相应的 miRNA*在多物种中保守, 这无疑会使其失去了发现非保守 miRNA 的机会。Lindow 和 Krogh[63]同样从 mRNA 出发, 在拟南芥基因组中搜索与 mRNA 无间隔匹配且长度在 20~27 nt (允许 2 个不匹配)的片段, 但他们不是通过在相近物种中保守来筛选候选 miRNA, 而是通过序列复杂

度、是否位于外显子区域、是否包含重复元件、miRNA 与 mRNA 双链的自由能、在基因组中的拷贝数和前体序列的自由能、环的大小和配对数这 6 条标准严格进行筛选。通过这样的筛选，他们找到 592 个候选 miRNA。他们预测出的候选 miRNA 大部分在其他植物基因组中均未呈现出明显保守性。

尽管理论上结合作用靶标的预测方法既能够预测在多物种中保守的 miRNA，也能够预测非保守的 miRNA，应该具有更好的敏感性和特异性，但在实际应用中，与 mRNA 片段反向互补的基因间区序列数目巨大，使得这类方法也不得不借助在多物种中的保守性来提高特异性，减少预测出的候选 miRNA 的数目，或者采用严格的标准筛选，牺牲了敏感性。而且，由于植物 miRNA 与靶标存在更多的互补配对，而动物 miRNA 与靶标的结合方式存在较大的不确定性，因而这类方法多用于植物 miRNA 的预测。

5 基于机器学习的预测方法

基于机器学习的预测方法是近两年出现的 miRNA 预测方法，与前 4 种方法最大的不同在于，基于机器学习的预测方法不仅需要已知的 miRNA，还需要已知的“非 miRNA”，通过 miRNA(阳性)和非 miRNA(阴性)数据集来构建区分两者的分类器，而后根据学习得到的分类器对未知序列进行预测。

支持向量机(Support vector machines, SVM)方法是目前 miRNA 分类和预测最常采用的机器学习方法。Xue 等[64]根据 163 个已知人 pre-miRNA 和 168 个蛋白编码区(Coding sequences, CDS)中折成 stem-loop 结构的片段、用 32 个三联体结构-序列特征描述样本、构建了分类器 3SVM，该分类器对测试集的敏感性和特异性分别为 93.3%,和 88.1%。Ng 和 Mishra[65]采用了与 Xue 等[64]完全相同的数据源构建分类器 miPred，但他们采用 29 个碱基组成和结构特征描述样本。为了提高分类器的特异性，他们增加了训练集中阴性样本的数量，使阴性与阳性数据集的比例增大到 2∶1，构建了一个有偏的分类器。miPred 对测试集的特异性提高到 97.97%，而相应的敏感性降低到 84.55%。Sewer 等[66]和 Pfeffer 等[67]根据 178 个已知的人 pre-miRNA 和 5 395 个从 tRNA、rRNA、mRNA、人和多种病毒基因组随机选取的序列、采用 37 个结构特征描述样本、构建了分类器 miR-abela。miR-abela 对训练集自身的敏感性为 71%，特异性为 97%。特异性高而敏感性低的主要原因是 miR-abela 所采用的训练集中阴性样本数量远远大于阳性样本数量，这使得分类器倾向于将样本判断为阴性，也就是提高了特异性，但牺牲了敏感性。他们采用该分类器对 8 种以人为宿主的病毒的 miRNA 进行了预测，预测出 32 个可能的 pre-miRNA，其中 13 个得到实验验证；他们还采用相同的方法对人、小鼠和大鼠已知 pre-miRNA 簇上下游 10 kb 的区域进行了预测，分别发现了 89、66 和 105 个可能的 pre-miRNA，其中分别有 20、17 和 6 个可能的 pre-miRNA 能够在小 RNA 克隆库中找到匹配的序列。Hertel 和 Stadler[68]以 295 个动物 pre-miRNA 与其在所有多细胞生物中的直系和旁系同源片段的多序列比对为阳性数据集、294 个随机置乱的 pre-miRNA 与同源序列的多序列比对和 483 个 tRNA 多序列比对为阴性数据集、采用 12 个序列和结构特征描述样本、构建和测试了分类器 RNAmicro。随机选取阳性和阴性数据中的一半为训练集，另一半为测试集，RNAmicro 对测试集的敏感性和特异性分别达到 90%和 99%。特异性高部分地也是源于阴性样本数量大于阳性样本数量。Helvik 等[69]采用两个级联的分类器 Microprocessor SVM 和 miRNA SVM 来预测 miRNA。Microprocessor SVM 是以 327 个人 pre-miRNA 中真实的 Drosha 酶剪切位点为阳性样本、其他位点为阴性样本构建的分类器，用于预测 pre-miRNA 中的 Drosha 酶剪切位点，它的输出、结合 327 个真实的人 pre-miRNA(阳性样本)和 3 000 个从人基因组中随机选取的茎环结构片段(阴性样本)，用于训练分类器 miRNA SVM。他们采用了 686 结构和序列特征描述训练分类器 Microprocessor SVM 的样本，另外还增加了 7 个剪切位点特征描述训练分类器 miRNA SVM 的样本。该级联分类器的敏感性和特异性分别为~90%和~95%。高特异性同样部分地源于训练集中阴性样本数量远大于阳性样本数量。

Jiang 等[70]尝试采用随机森林(Random forest)方法构建区分 pre-miRNA 和非 pre-miRNA 的分类器 MiPred。他们采用的训练数据集与 Xue 等[64]完全相同，同时也用了 Xue 等[64]文中所用的 32 个三联体结构-序列特征，他们还另外增加了最小自由能和自由能的随机检验 P 值[46]两个特征来描述样本。MiPred 对测试集的敏感性和特异性分别为 89.35% 和 93.21%，远高于 3SVM[64]在相同测试集上的性能。

Nam 等[71]采用隐马尔可夫模型(Hidden markov

model, HMM)描述真实 pre-miRNA 和非 pre-miRNA 的二级结构, 根据 136 个已知人 pre-miRNA 和 1 000 个从人基因组中随机选取的茎环结构片段估计 HMM 的转移概率和发射概率, 构建了分类器 ProMiR。当取阈值 0.033 时, 该分类器的 5 折交叉检验特异性高达 96%, 而敏感性只有 73%。他们利用该分类器、结合同源 EST 搜索、自由能随机检验 *P* 值[46]、在脊椎动物基因组中的保守性模式[33]等条件对人 16、17、18 和 19 号染色体进行了预测, 最后预测出 23 个可能的 miRNA。通过实验, 他们验证了其中的 9 个。

Yousef 等[72]则采用 Naive Bayes 分类器构建了区分 pre-miRNA 和非 pre-miRNA 的分类器 BayesMiRNAfind。由于增加训练样本的数量有助于提高分类器的性能, 因此 Yousef 等将多种病毒、植物和动物共计 1 420 个 pre-miRNA 作为阳性训练集, 将 30 000 个多物种保守序列中折成茎环结构的片段作为阴性训练集来训练分类器。他们采用几千个序列和结构特征描述样本。利用该分类器, 结合片段的长度等结构特征和与人以及河豚基因组的保守性, 他们在小鼠基因组的正链中预测出 533 个可能的 pre-miRNA(阈值为 0.99), 其中包括 135 条已知 miRNA 中的 53 条, 敏感性为 39%。

理论上只要阳性和阴性训练样本选取合理, 描述样本的特征能够很好地反映两类样本的差异, 并采用适当的机器学习方法, 完全可以高效地预测 miRNA。然而, 在实际应用中, 由于难以选取到足以描述整个阴性样本空间的代表样本, 也难以找到足以区分 miRNA 和非 miRNA 的特征, 使得基于机器学习预测 miRNA 的效果不尽如人意。即使有的分类器对训练集和测试集均表现出很好的性能, 然而在对基因组进行预测时, 仍然会预测出大量的候选 miRNA。尽管不排除各物种 miRNA 的真实数量可能远超过现在估计的几百个, 但其中仍然包含了大量的假阳性。如何降低假阳性、进一步提高敏感性, 是基于机器学习的预测方法需要进一步探索和解决的问题。

6　结论与展望

自 1993 年第一个 miRNA——lin-4 发现以来, 到目前为止已有 5 000 多个 miRNA 被陆续发现[73], 其中较大比例的 miRNA 是通过计算方法预测、而后经实验验证的。而且, 越来越多的研究显示, 部分 miRNA 是机体或细胞在逆境胁迫时表达或表达量增加[74~76], 这使得 cDNA 克隆方法更加难以捕捉到这些逆境 miRNA。因此, 尽管一开始计算发现方法是作为 cDNA 克隆方法的补充出现, 但发展到现在, 计算发现方法已经在 miRNA 的发现中发挥着举足轻重的作用。

到目前为止, 通过计算预测 miRNA 的方法已经有几十种, 尽管根据算法本质可分为 5 大类, 但无论算法本质是同源搜索, 还是机器学习, 这些方法都存在共同的问题, 就是根据少量的已知 miRNA 或 pre-miRNA 总结规律, 去发现大量的新 miRNA。这个问题导致计算预测方法的精度还不能令人满意。由于已知 miRNA 的数量较少, 因而从中总结的规律不足以代表整个 miRNA 家族, 使得计算预测存在大量的假阳性和假阴性, 尤其是当对全基因组进行预测时, 往往预测出几十万个可能的 miRNA, 其中包含大量的假阳性, 而漏检的比例也很高。这个问题的解决途径一方面可以借助新的大规模并行测序技术发现更多的 miRNA, 以利于计算发现方法总结出更为细致、准确的规律, 提高计算发现方法的敏感性和特异性; 另一方面也有待于探索新的计算发现方法或将现有的预测方法进行有效整合, 以便在现有知识的情况下, 尽可能的在提高特异性的同时, 也提高敏感性。

最近, 研究者在小鼠、大鼠和人的睾丸组织中发现了一类长度在 26~31 nt 的小 RNA——piRNA (Piwi-interacting RNA)[77~81]。尽管目前尚未发现这类小 RNA 具有特殊的结构, 但 miRNA 计算发现方法中的很多思路在 piRNA 等其他类型的小 RNA 的发现方面仍然具有借鉴意义。

参考文献(References):

[1] Cai X, Hagedorn CH, Cullen BR. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 2004, 10(12): 1957–1966.

[2] Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH, Kim VN. MicroRNA genes are transcribed by RNA polymerase Ⅱ. *EMBO Journal*, 2004, 23: 4051–4060.

[3] Borchert GM, Lanier W, Davidson BL. RNA polymerase Ⅲtranscribes human microRNAs. *Nat Struct Mol Biol*, 2006, 13(12): 1097–1101.

[4] Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN. The nuclear RNaseⅢ Drosha initiates microRNA processing. *Nature*, 2003,

425(6956): 415–419.

[5] Zeng Y, Yi R, Cullen BR. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO Journal*, 2005, 24(1): 138–148.

[6] Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 2003, 17(24): 3011–3016.

[7] Bohnsack MT, Czaplinski K, Gorlich D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA*, 2004, 10(2): 185–191.

[8] Ketting RF, Fischer SE, Bernstein E, Sijen T, Hannon GJ, Plasterk RH. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev*, 2001, 15(20): 2654–2659.

[9] Jiang F, Ye X, Liu X, Fincher L, McKearin D, Liu Q. Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. *Genes Dev*, 2005, 19(14): 1674–1679.

[10] Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, Carthew RW. Distinct roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, 2004, 117(1): 69–81.

[11] SHENG Xi-Hui, DU Li-Xin. Progress on the research of microRNAs and its function in humans and animals. *Hereditas* (*Beijing*), 2007, 29(6): 651–658.
盛熙晖, 杜立新. MicroRNA 及其在人和动物上的研究进展. 遗传, 2007, 29(6): 651–658.

[12] Kurihara Y, Watanabe Y. *Arabidopsis* micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci USA*, 2004, 101(34): 12753–12758.

[13] Park MY, Wu G, Gonzalez-Sulser A, Vaucheret H, Poethig RS. Nuclear processing and export of microRNAs in *Arabidopsis*. *Proc Nat Acad Sci USA*, 2005, 102(10): 3691–3696.

[14] LI Pei-Wang, LU Xiang-Yang, LI Chang-Zhu, FANG Jun, TIAN Yun. Advances in the study of plant microRNAs. *Hereditas* (*Beijing*), 2007, 29(3): 283–288.
李培旺, 卢向阳, 李昌珠, 方俊, 田云. 植物 microRNAs 研究进展. 遗传, 2007, 29(3): 283–288.

[15] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 2004, 116: 281–297.

[16] Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. Identification of mammalian microrna host genes and transcription units. *Genome Res*, 2004, 14(10A): 1902–1910.

[17] Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H. Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 2005, 33(8): 2697–2706.

[18] Moss EG, Lee RC, Ambros V. The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the lin-4 RNA. *Cell*, 1997, 88: 637–646.

[19] Reinhart BJ, Slack FJ, Basson M, Bettinger JC, Pasquinelli AE, Rougvie AE, Horvitz HR, Ruvkun G. The 21 nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 2000, 403: 901–906.

[20] Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM. *Bantam* encodes a developmentally regulated microrna that controls cell proliferation and regulates the proapoptotic gene *hid* in *Drosophila*. *Cell*, 2003, 113: 25–36.

[21] Xu P, Vernooy SY, Guo M, Hay BA. The *Drosophila* microRNA mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol*, 2003, 13(9): 790–795.

[22] Chen CZ, Li L, Lodish HF, Bartel DP. MicroRNAs modulate hematopoietic lineage differentiation. *Science*, 2004, 303(5654): 83–86.

[23] Park JK, Liu X, Strauss TJ, McKearin DM, Liu Q. The miRNA pathway intrinsically controls self-renewal of drosophila germline stem cells. *Curr Biol*, 2007, 17(6): 533–538.

[24] Chen X. A microRNA as a translational repressor of APETALA2 in *Arabidopsis* flower development. *Science*, 2004, 303: 2022–2025.

[25] Berezikov E, Cuppen E, Plasterk RHA. Approaches to microRNA discovery. *Nat Genet*, 2006, 38(Suppl.): S2–S7.

[26] Bentwich I. Prediction and validation of microRNAs and their targets. *FEBS Letter*, 2005, 579(26): 5904–5910.

[27] Kim VN, Nam JW. Genomics of microRNA. *Trends Genet*, 2006, 22(3): 165–173.

[28] Weber MJ. New human and mouse microRNA genes found by homology search. *FEBS Journal*, 2005, 272(1): 59–73.

[29] Kent WJ. BLAT —— The BLAST-Like alignment tool. *Genome Res*, 2002, 12(4): 656–664.

[30] Dezulian T, Remmert M, Palatnik JF, Weigel D, Huson DH. identification of plant microRNA homologs. *Bioinformatics*, 2005, 22(3): 359–360.

[31] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*, 1990, 215(3): 403–410.

[32] Lai EC, Tomancak P, Williams RW, Rubin GM. Computational Identification of *Drosophila* microRNA genes. *Genome Biol*, 2003, 4(7): R42.

[33] Berezikov E, Guryev V, van de BJ, Wienholds E, Plasterk RH, Cuppen E. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 2005, 120(1): 21–24.

[34] Legendre M, Lambert A, Gautheret D. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 2005, 21(7): 841–845.

[35] Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol*, 2001, 313(5):

1003−1011.

[36] Wang X, Zhang J, Li F, Gu J, He T, Zhang X, Li Y. MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 2005, 21(18): 3610−3614.

[37] Li Y, Li W, Jin YX. Computational identification of novel family members of microRNA genes in *Arabidopsis thaliana* and *Oryza sativa*. *Acta Biochimica Biophysica Sinica* (*Shanghai*), 2005, 37(2): 75−87.

[38] Qiu CX, Xie FL, Zhu YY, Guo K, Huang SQ, Nie L, Yang ZM. Computational identification of microRNAs and their targets in *Gossypium hirsutum* expressed sequence tags. *Gene*, 2007, 395(1−2): 49−61.

[39] Xie FL, Huang SQ, Guo K, Xiang AL, Zhu YY, Nie L, Yang ZM. Computational identification of novel microRNAs and targets in *Brassica napus*. *FEBS Letter*, 2007, 581(7): 1464−1474.

[40] Zhang BH, Pan XP, Wang QL, Cobb GP, Anderson TA. Identification and characterization of new plant microRNAs using EST analysis. *Cell Research*, 2005, 15(5): 336−360.

[41] Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP. The microRNAs of *Caenorhabditis elegans*. *Genes Dev*, 2003, 17(8): 991−1008.

[42] Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. Vertebrate microRNA genes. *Science*, 2003, 299(5612): 1540−1540.

[43] Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J. Computational and experimental identification of *C. elegans* microRNAs. *Molecular Cell*, 2003, 11: 1253−1263.

[44] Wang XJ, Reyes JL, Chua NH, Gaasterland T. Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biol*, 2004, 5(9): R65.

[45] Bonnet E, Wuyts J, Rouze P, Van de Peer Y. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc Nat Acad Sci USA*, 2004, 101(31): 11511−11516.

[46] Bonnet E, Wuyts J, Rouze P, Van de Peer Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 2004, 20(17): 2911−2917.

[47] Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C, Tuschl T. Identification of virus-encoded microRNAs. *Science*, 2004, 304(5671): 734−736.

[48] Cai X, Schafer A, Lu S, Bilello JP, Desrosiers RC, Edwards R, Raab-Traub N, Cullen BR. Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed. *PLoS Pathogens*, 2006, 2(3): e23.

[49] Cai X, Lu S, Zhang Z, Gonzalez CM, Damania B, Cullen BR. Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Nati Acad Sci USA*, 2005, 102(15): 5570−5575.

[50] Omoto S, Ito M, Tsutsumi Y, Ichikawa Y, Okuyama H, Brisibe EA, Saksena NK, Fujii YR. HIV-1 nef suppression by virally encoded microRNA. *Retrovirology*, 2004, 1(1): 44.

[51] Burnside J, Bernberg E, Anderson A, Lu C, Meyers BC, Green PJ, Jain N, Isaacs G, Morgan RW. Marek's disease virus encodes MicroRNAs that map to meq and the latency-associated transcript. *J Virol*, 2006, 80(17): 8778−8786.

[52] Cullen BR. Viruses and microRNAs. *Nat Genet*, 2006, 38: S25−S30.

[53] Nelson DR, Schuler MA, Paquette SM, Werck-Reichhart D, Bak S. Comparative genomics of rice and *Arabidopsis*. Analysis of 727 cytochrome P450 genes and pseudogenes from a monocot and a dicot. *Plant Physiology*, 2004, 135(2): 756−772.

[54] Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genetics*, 2007, 3(2): e7.

[55] Sullivan CS, Grundhoff AT, Tevethia S, Pipas JM, Ganem D. SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature*, 2005, 435(7042): 682−686.

[56] Grundhoff A, Sullivan CS, Ganem D. A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA*, 2006, 12(5): 733−750.

[57] Cui C, Griffiths A, Li G, Silva LM, Kramer MF, Gaasterland T, Wang XJ, Coen DM. Prediction and identification of herpes simplex virus 1-encoded microRNAs. *J Virol*, 2006, 80(11): 5499−5508.

[58] Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, Sharon E, Spector Y, Bentwich Z. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, 2005, 37(7): 766−770.

[59] Li SC, Pan CY, Lin WC. Bioinformatic discovery of microRNA precursors from human ESTs and introns. *BMC Genomics*, 2006, 7: 164−164.

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

[61] Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. Systematic discovery of regulatory motifs in human promoters and 3′UTRs by comparison of several mammals. *Nature*, 2005, 434(7031): 338−345.

[62] Adai A, Johnson C, Mlotshwa S, Archer-Evans S, Manocha V, Vance V, Sundaresan V. Computational prediction

of miRNAs in *Arabidopsis thaliana*. *Genome Res*, 2005, 15(1): 78–91.

[63] Lindow M, Krogh A. Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, 2005, 6(1): 119.

[64] Xue C, Li F, He T, Liu GP, Li Y, Zhang X. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 2005, 6(1): 310.

[65] Ng KLS, Mishra SK. *De novo* SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures. *Bioinformatics*, 2007, 23(11): 1321–1330.

[66] Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein M, Tuschl T, van Nimwegen E, Zavolan M. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics*, 2005, 6(1): 267.

[67] Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grasser FA, van Dyk LF, Ho CK, Shuman S, Chien M, Russo JJ, Ju J, Randall G, Lindenbach BD, Rice CM, Simon V, Ho DD, Zavolan M, Tuschl T. Identification of microRNAs of the herpesvirus family. *Nat Methods*, 2005, 2(4): 269–276.

[68] Hertel J, Stadler PF. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 2006, 22(14): e197–e202.

[69] Helvik SA, Snove O Jr, Saetrom P. Reliable prediction of Drosha processing sites improves microRNA gene prediction. *Bioinformatics*, 2007, 23(2): 142–149.

[70] Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Research*, 2007, 35(Web Server issue): W339–W344.

[71] Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT. Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 2005, 33(11): 3570–3581.

[72] Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC, Showe MK. Combining multi-species genomic data for microRNA identification using a Naive Bayes classifier. *Bioinformatics*, 2006, 22(11): 1325–1334.

[73] Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 2006, 34(Suppl_1): D140–D144.

[74] Zhao B, Liang R, Ge L, Li W, Xiao H, Lin H, Ruan K, Jin Y. Identification of drought-induced microRNAs in rice. *Biochem Biophys Res Commun*, 2007, 354(2): 585–590.

[75] Sunkar R, Zhu JK. Novel and stress-regulated microRNAs and other small RNAs from *Arabidopsis*. *Plant Cell*, 2004, 16(8): 2001–2019.

[76] Leung AK, Sharp PA. microRNAs: a safeguard against turmoil? *Cell*, 2007, 130(4): 581–585.

[77] Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 2006, 442(7099): 199–202.

[78] Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, Chien M, Russo JJ, Ju J, Sheridan R, Sander C, Zavolan M, Tuschl T. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 2006, 442(7099): 203–207.

[79] Grivna ST, Beyret E, Wang Z, Lin H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev*, 2006, 20(13): 1709–1714.

[80] Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, Minami N, Imai H. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev*, 2006, 20(13): 1732–1743.

[81] Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. Characterization of the piRNA complex from rat testes. *Science*, 2006, 313(5785): 363–367.