文章编号:1007-2861(2010)01-0075-06

MiRfilter:一个预测病毒 microRNA 的计算工具

张玉滨, 赵洁苑, 龚云路, 王翼飞

(上海大学 理学院,上海 200444)

摘要:采用生物信息学方法开发了一个预测病毒 miRNA 的计算机程序 MiRfilter,通过提取已知的病毒 miRNA 的生物学特征,规定参数及参数范围,并以此为筛选标准,预测潜在的病毒 miRNA. 用 EBV 和 RLCV 两种病毒的正负样本对 MiRfilter 的预测精度进行检测. MiRfilter 从 39 个 EBV 已知的 miRNA 中正确识别了 30 个,从 116 个负数据中检测到 9 个假阳性数据;从 20 个 RLCV 的 miRNA 中正确识别了 15 个,从 108 个负数据中检测到 2 个假阳性数据. 用取自果蝇和线虫预测区域的 186 个负数据检测 MiRfilter 的特异性,在 186 个负数据中没有预测出假阳性序列.

关键词:病毒 miRNA;MiRfilter;生物基因组;发夹结构;生物学特征

中图分类号:O 224 **文献标志码**:A

A Computational Method for Prediction and Detection of microRNAs

ZHANG Yu-bin, ZHAO Jie-yuan, GONG Yun-lu, WANG Yi-fei

(College of Sciences, Shanghai University, Shanghai 200444, China)

Abstract: This paper develops a computational procedure named MiRfilter to predict the virus miRNAs. MiRfilter selects some important biological features of miRNAs as parameters to predict the potential precursors and miRNAs. When using two samples to test MiRfilter's accuracy, it successfully retrieves 30 EBV-encoded miRNAs from 39 known ones, and predicts 9 false positive miRNAs from 116 negative ones. Then it retrieves 15 RLCV-encoded miRNAs from 20 known ones and predicts 2 false miRNAs from 108 negative ones. Further, when using another species to test MiRfilter's specificity, it predicts none false positive miRNA from 186 negative ones in *D. melanogaster* and *C. elegans*.

Key words: virus miRNA; MiRfilter; biology genome; hairpin structure; biological features

microRNA(miRNA)是一种内源性长约 21 ~ 25 个核苷酸的单链非编码 RNA[1],广泛存在于真核细胞中. miRNA 本身不具有可读框(ORF),但能够调节其他基因的表达活性,与生物的阶段性发育有着密切的联系. miRNA 在生物体中的表达具有时序性、保守性、组织特异性的特点,其前体 pre-miRNA 折叠形成发夹结构或类似发夹的二级结构. 通过对 pre-miRNA 的基因组定位和注释发现,miRNA 主要位于基因间隔区,也有相当一部分来源于内含子. 较大比例的 miRNA 呈现成簇分布,且在相近或多物种间保守. 成熟 miRNA 进入一个类似于 RISC(RNA-induced silencing complex)的核糖蛋白复合体 miRNP,通过与靶基因的 3′UTR 互补配对,指导 miRNP 复合体对靶基因 mRNA 进行切割或者翻译

收稿日期:2008-04-28
基金项目:国家高技术研究发展计划(863 计划)资助项目(2002AA02Z190)
通信作者:王翼飞(1948 ~),男,教授,博士生导师,研究方向为计算分子生物学. E-mail:yifei_wang@staff.shu.edu.cn

抑制. 1993 年,Lee 等在 *C. elegans* 中发现了第一个定时调控胚胎后期发育的基因 *lin*-4;2000 年,Reinhart 等又发现了第二个异时性开关基因 *let*-7[1-2];2008 年 2 月,miRBase(http://microrna.sanger.ac.uk/cgi-bin/sequences/)里通过生物信息学手段和分子克隆方法发现的 miRNA 已经有 5 395 种.

miRNA 的发现主要有两种方法. 最初寻找 miRNA 一般通过 cDNA 克隆方法,然后利用 Northern 杂交技术检测 miRNA 的表达. 由于克隆方法的局限性,近年来通过计算机预测 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. 2003 年出现了第一个计算机预测工具——MiRscan[3],这是 Lim 等人在线虫中寻找 miRNA 基因时总结的方法,适用于线虫和脊椎动物 miRNA 的搜索. 同年,另一个实验室的 Lat 等人在果蝇中寻找 miRNA 基因,开发了一个用于果蝇和昆虫的软件——MiRseeker[4]. 随着 miRNA 研究的深入,预测 miRNA 的计算机方法越来越多,涉及的物种也越来越多. 根据预测方法的本质可以分为 5 种类型[5]:同源片段搜索方法[3]、基于比较基因组学的预测方法[4,6]、基于序列和结构特征打分的预测方法[7]、结合作用靶标的预测方法[8]和基于机器学习的预测方法[9-10].

病毒编码的 miRNA 是 2004 年以来新发现的一类 miRNA[11],与人类疾病的发生有着千丝万缕的联系. 预测和识别病毒编码的 miRNA,探索其对病毒感染、复制、表达的作用,都将有助于病毒分子生物学的研究,也为研发防治人类疾病的新方法和新途径提供新的思路. 很多病毒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,甚至不可能. 因此需要另辟蹊径,开发针对病毒 miRNA 特点的预测方法和工具. 本研究采用生物信息学方法,开发了一个预测病毒 miRNA 的计算机程序 MiRfilter.

1 材料与方法

1.1 MiRfilter 介绍

MiRfilter 采用 C#语言编程,分为预处理→训练参数→预测三个步骤,能够对预测过程实现自动化. 预处理部分和训练参数部分将对训练集进行处理.

首先,选取并预处理训练集. MiRfilter 是根据已知的 miRNA 的生物学特征预测未知的 miRNA,因此需要选择一个训练集. 如果待测物种已经发现一定数量的 miRNA,则直接用这一物种的已知序列作为训练集;如果待测物种尚没有已知的 miRNA,则可选取相近物种的已知 miRNA 作为训练集,给出各参数的范围. 对所选训练集进行预处理,保留发夹结构中只有一个发夹环的前体和对应的成熟序列.

其次,训练参数. 统计训练集的生物学特征,界定所需的参数及其范围. 这里用到的参数取自前体和成熟体两部分,包括:P1 为前体序列大小(Prelen);P2 为发夹结构自由能(Energy);P3 为发夹环的大小(Hplen);P4 为成熟体到发夹结构茎区末端的距离(Dist2,图 1 椭圆框内碱基的个数);M1 为发夹环到成熟体的距离(Dist1,图 1 方框内碱基个数);M2 为发夹结构茎区内成熟序列中不配对碱基的个数(upNum);M3 为发夹结构茎区内成熟序列中最大内环的大小(Bugle);M4 为成熟序列C + G 含

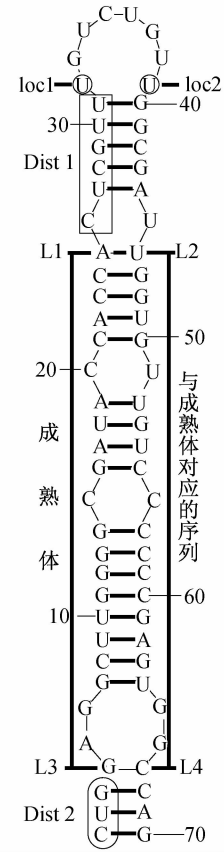

图 1 发夹结构

Fig. 1 Hairpin structure

量(cgCon);M5 为发夹结构茎区内成熟序列中内环

的个数(BugNum);M6 为发夹结构茎区内成熟序列的两端处不配对碱基的个数(terBugNum);M7 为发夹结构茎区内成熟序列中内环的平均大小(AveBugle = (upNum - terBugNum)/BugNum);M8 为发夹结构茎区中与成熟序列对应的序列长度,该参数显示了成熟序列与其对应序列的对称性(Symmetry,见图 1).

另外,统计了一个新的参数 MFEL = Energy/Prelen,即自由能与前体序列长度之比. 这是区分植物 pre-miRNA 一个很有效的参数[12],将其应用于病毒,也能取得很好的效果. 在表示参数的范围时,参数名加后缀 Min 或 Max 表示此参数范围的最小值或最大值. 而参数 Dist2 使用其平均值, BugNum, terBugNum 和 AveBugle 使用最大值.

最后,预测未知 miRNA,实现的流程如图 2 所示,其具体执行过程如下.

(1) 在预测之初对待测物种的基因组序列进行处理.

病毒 miRNA 基因大部分位于基因间隔区,也有相当数量的 miRNA 基因位于内含子内,因此对待测物种的基因组搜索 GenBank 数据库文件中的 CDS 字段信息,提取外显子以外的全部区域作为预测区域. 根据参数 Prelen 的范围,在预测区域中截取连续片段,其中窗口大小设置为[5/4(PrelenMax)] nt (取整),并以[1/4(PrelenMax)] nt 的步长向后移动,这样的取法基本上可以覆盖可能的前体. 需要注意的是,虽然预测区域是由一段段小的区域组成,但它们之间是独立的,在截取片段时应该从每个小区域中分别截取,而长度小于截取片段的小区域则不予考虑.

(2) 利用 RNAstructrue 4.2 程序对截取的片段作二级结构模拟,并从中提取合格的发夹结构.

MiRfilter 采用 RNAstructure 4.2 程序对这些片段模拟二级结构. 当使用 RNAstructure 4.2 程序的默认参数时,可以得到最多 20 个可能的二级结构结果,这些结果以最小自由能顺序排列. MiRfilter 只对第一个结果,也就是自由能最小的结构进行分析,尽管它不一定为真实的 RNA 结构. 从二级结构中提取形成发夹结构的序列,判断合格发夹结构的标准如下:① 序列的二级结构由发夹环和能够相互匹配的两条茎组成,如果茎上有单链区,那么这些单链区只能是内环或突环,不允许有多分支环;② 发夹结构中只能包含一个发夹环;③ 整个发夹结构的长度在 PrelenMin ~ PrelenMax 之间;④ 两条茎上配对碱基的个数在 18 nt 以上;⑤ 发夹环的大小介于 HplenMin ~ HplenMax 之间. 保留满足条件的发夹结构,同时记录发夹环的起始位置[loc1,loc2](见图 1).

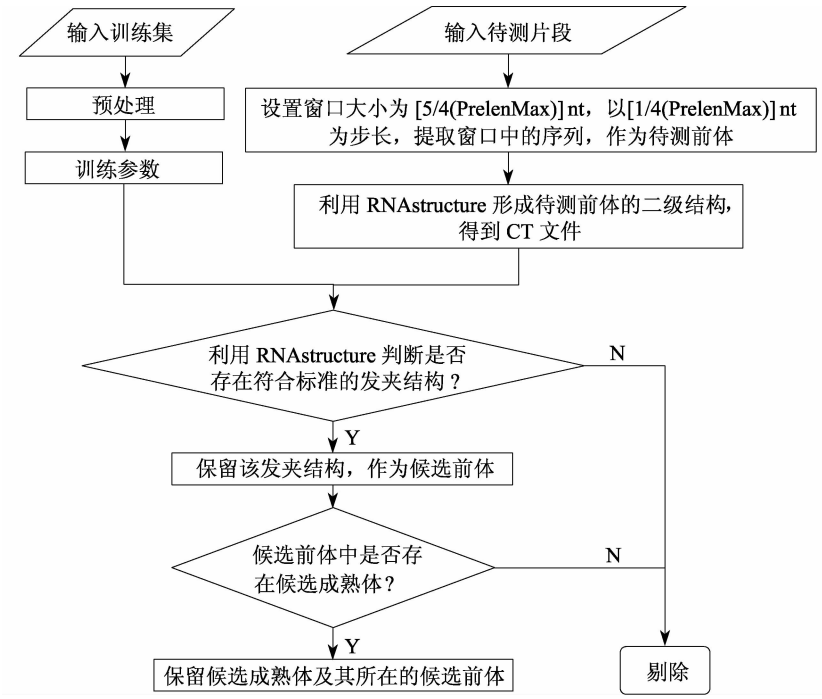

图 2　预测过程流程图

Fig. 2　Flowchart of whole prediction

(3) 根据参数从发夹结构中筛选候选 pre-miRNA 和 miRNA 序列.

对于病毒来说,一个前体两条臂上的两个成熟 miRNA 可能都会保留下来. 因此,在预测病毒的 miRNA 时,对前体的两条臂均做筛选. 筛选步骤如下:

步骤一 从 loc1 向前移动 dst = (Dist1Min + 1) nt,记录下此时的位置 L1,与 L1 对应的位置记为 L2(见图 1). 从 L1 开始继续向前移动 22 nt,终止位置记为 L3,与 L3 对应的记为 L4. 得到的序列 L1-L3 被视为可能的 miRNA, 与之对应的序列为 L2-L4.

步骤二 沿 L3 向前再取 Dist2 nt(如果有的话),同时保证右臂的碱基与左臂对应,这样就得到一个新的前体. 利用 RNAstructure 4.2 程序重新模拟这个前体的二级结构,如果它的自由能不在训练参数范围内,则剔除;否则,判断序列 L1-L3 和对应序列 L2-L4 是否满足参数 M2 ~ M8 的范围.

步骤三 令步骤一中 dst = dst + 1,重复步骤一和二,直到 dst = (Dist1Max + 1) nt.

步骤四 以同样的手段处理右臂,从 loc2 向后移动 dst = (Dist1Min + 1) nt,记下此时的位置 L2,与之对应的位置记为 L1. 从 L2 开始向后继续移动 22 nt,终止位置记为 L4,与 L4 对应的位置记为 L3. 得到的序列 L2-L4 被视为可能的 miRNA,与之对应的序列为 L1-L3.

步骤五 沿 L4 向后再取 Dist2 nt,同时保证左臂的碱基与右臂对应,这样也得到一个新的前体. 模拟其二级结构,判断二级结构的自由能以及序列 L2-L4,L1-L3 是否满足条件.

步骤六 令步骤五中 dst = dst + 1,重复步骤五和六,直到 dst = (Dist1Max + 1) nt,终止.

最后再从筛选出的成熟体中取出重叠部分作为候选的 miRNA.

1.2 数据集

本研究用两组正负样本检测 MiRfilter 的预测精度. miRBase 上已经公布了野生型 EB 病毒(Epstein Barr virus,EBV)23 个前体上的 39 个成熟体(某些前体的两条臂上均有一个成熟体),且这 23 个前体的二级结构均只含有一个发夹环. 从 NCBI 网站上下载野生型 EBV 的基因组序列,从其外显子部分取了 58 条能形成发夹结构的序列. 这些序列对应了 116 个负数据,将它们作为第一组正负样本. 恒河猴淋巴隐秘病毒(Rhesus lymphocryptovirus,RLCV)已经被鉴定了 16 个前体的 22 个成熟体,经预处理将发夹结构中含有两个发夹环的 1 个前体去掉,保留剩下的 20 个成熟体. 从 RLCV 的基因组序列的外显子部分取了 54 条具有发夹结构的序列,对应了 108 个负数据,将它们作为第二组正负样本. 另外,为了检测 MiRfilter 的特异性,在线虫和果蝇的预测区域中取了 93 条能形成发夹结构的序列,它们对应了 186 个负数据.

1.3 评价标准

对于每一个测试样本,只可能属于以下 4 种类型之一:正确识别的正样本 T_P、正确识别的负样本 T_N、本来是负样本却被识别为正样本(假阳性样本) F_P、本来是正样本却被识别为负样本(假阴性样本) F_N. 用 n 表示样本总数,Q 表示总预测精度,Q_P 表示正样本的预测精度,Q_N 表示负样本的预测精度,R_{FP} 表示假阳性预测率,R_{FN} 表示假阴性预测率,M_{CC} 表示 Matthew 相关系数,分别定义如下[22]:

$$Q = \frac{T_P + T_N}{n}, Q_P = \frac{T_P}{T_P + F_N}, Q_N = \frac{T_N}{T_N + F_P},$$

$$R_{FP} = \frac{F_P}{F_P + T_N}, R_{FN} = \frac{F_N}{F_N + T_P},$$

$$M_{CC} = \frac{T_P \times T_N - F_P \times F_N}{\sqrt{(T_P + F_N)(T_P + F_P)(T_N + F_P)(T_N + F_N)}}.$$

2 结果与讨论

搜索 EBV 的基因组序列,得到 90 个小的预测区域. 利用 MiRfilter 在这 90 段预测区域中搜索,总共找到了 31 个前体和 58 个成熟体,正确识别了 30 个已知成熟体,正样本的预测精度达到 76.92%. 在没有识别出来的 9 个 miRNA 中,mir-BHRF1-1 和 ebv-mir-BART16 没有预测出来;而 ebv-mir-BART13 和 ebv-mir-BART6-3p 是由于预测出来的序列比已知 miRNA 少 2 ~ 3 个碱基而被排除,其中 ebv-mir-BART5 缺少开始位置的 3 个碱基(CAA),ebv-mir-BART13 缺少末尾的 3 个碱基(UGA),ebv-mir-BART6-3p 缺少末尾的 2 个碱基(GA);还有 4 个成熟体是由于预测出来的序列比它们本身少 1 个碱基而被排除,它们是 ebv-mir-BART10(末尾缺少 U)、ebv-miR-BART11-5p(起始缺少 U)、ebv-mir-BART15(起始缺少 G)、ebv-mir-BART18-3p(末尾缺少 C). EBV 负样本的 116 个负数据中被预测为 miRNA 的假阳性数据有 9 个,假阳性率为 7.76%,被识别为 miRNA 的 9 个假阳性数据由表 1 给出.

表 1　EBV 负样本预测出的假阳性数据

Table 1　False positive miRNAs from the EBV negative sample

序号	假阳性数据	假阳性数据所在的发夹结构
1	GUGGAGAUGGGCAGGCAGUUUGGCU	GCGUGGAGAUGGGCAGGCAGUUUGGCUUCGAGGUGCAUAGAAGC CGGCCCUCCUUCCGUCAGUUCCAGGC
2	ACGUACUGCGCGGGGUCCAUUAUGG	CGGGUCCAGCCCACGUACUGCGCGGGGUCCAUUAUGGCCGCCCUG GGGACGACGAAGCGGUCGACGUAGGCCAGGAUGU
3	UGAGCUAAUGCCAUAGUCACGGA	GGGACAGACCCUGGGCCUUGGCUAUGGUCAUGGCCAGCUUUGAGC UAAUGCCAUAGUCACGGAUGCUGCAGAGGUUC
4	GACAUACUCGGGCUGCUCAUGCC	GGAAGCUGUAGACAUACUCGGGCUGCUCAUGCCCGUGGGCCUCCA CGAAGCUGUCCGCCUCGAGCGUGUCCAUAAAGUCC
5	CCCGGCGUCUAGGUUGUCACUUCGC	GUAGAAUAUCCGCCCGGCGUCUAGGUUGUCACU
6	AGCGCAAGUCCAAGUCUGGUGCUGG	UCGCUCGGCCGGCCAGAAGAGCGCAAGUCCAAGUCUGGUGCUGGG GCCGAUGUGCAG
7	UGUGCCCGCAGUUGUAGACUGUCAU	GUGCAGCGGUUUGUGCCCGCAGUUGUAGACUGUCAUUUUUAUGG GCGAGUGGGCGGUCCACACGCGCGGGCGCAGCACCCAUUGGUCG
8	GAGCGUGCACCUGGAAGAUGCAGC	UGGGAGCGUGCACCUGGAAGAUGCAGCUGGGG
9	UUUACAUCUUUACAGGGCGCAGCGG	UAGAUCUUUACAUCUUUACAGGGCGCAGCGGCCG

MiRfilter 搜索了 RLCV 的 48 段预测区域,总共预测出 42 个前体上 66 个成熟体,其中正确识别了 75%(15/20)的已知成熟体. 而在没有识别出来的 5 个 miRNA 中,均是由于预测出来的序列比已知 miRNA 少 1 ~4 个碱基而丢失. rlcv-miR-rL1-5-3p 和 rlcv-miR-rL1-9 都是缺少末尾的一个碱基(U),rlcv-miR-rL1-8 缺少起始位置的 2 个碱基(UA),rlcv-miR-rL1-12-3p 缺少末尾的 2 个碱基(GA),而 rlcv-miR-rL1-11 则在末尾缺少 4 个碱基(GGGG). RLCV 负样本的 108 个负数据只检测到 2 个假阳性数据,假阳性率仅为 1.85%,2 个被识别为 miRNA 的假阳性数据由表 2 给出. 第一组和第二组正负样本的预测精度见表 3.

表 2　RLCV 负样本预测出来的假阳性数据

Table 2　False positive miRNAs from the RLCV negative sample

序号	假阳性数据	假阳性数据所在的发夹结构
1	CGGGGUGAGGACACUGAGACGUGGU	GGCAGACCACCCCGGGGUGAGGACACUGAGACGUGGUUAGAACG UAACCUGUCUCAGGUCGCCGGGCUUCGGUUGGC
2	GUCUGUUCUUCACCGACGGGGUG	AACUUUGUGGCCCGCAAGUACGUGGUGAAGGAGACGGCGUUCAC CGUCAGUCUGUUCUUCACCGACGGGGUGGGGGCCAACCUGGC

表 3　EBV 和 RLCV 两组正负样本的预测精度

Table 3　Prediction accuracy by EBV and RLCV positive and negative samples

物种	T_P	F_P	T_N	F_N	Q_P/%	Q_N/%	Q/%	R_{FP}/%	R_{FN}/%	M_{CC}
EBV	30	9	107	9	76.92	92.24	88.39	7.76	23.08	0.68
RLCV	15	2	106	5	75.00	98.15	94.53	1.85	25.00	0.78

由两种病毒的正负样本检测发现,MiRfilter 具有较高的预测精度,同时假阳性率控制得也较好. 而取自果蝇和线虫预测区域的负样本,186 个负数据中并没有检测出假阳性数据,假阳性率为 0%,说明 MiRfilter 有很好的特异性.

Sullivan 等人在 2005 年采用对结构特征打分的方法,给出了预测病毒 miRNA 的方法. Grundhoff 等人于 2006 年对此方法作了改进,提高了给配对奖赏分值和给膨胀圈、末端环的惩罚分值. MiRfilter 没有采用他们的方法,而是充分考虑了 miRNA 前体及成熟体的生物学特征,使得预测参数既具有待测物种 miRNA 共有的特点,又能很好区别于其他物种,因此能够在保证灵敏性和特异性的同时较好地控制假阳性率. MiRfilter 的预测思想易于理解,程序界面友

好且易于使用.

不同物种的 miRNA 有着相异的生物学特征,这反映在计算机预测时需要根据不同的物种设计不同的筛选方法,确定不同的识别参数. MiRfilter 的设计仅依赖于已知 miRNA 前体和成熟体的生物学特征. 这一设计思想使得 MiRfilter 能够根据不同物种的特点训练不同的参数范围,使之适用于不同物种 miRNA 的预测,且在寻找缺乏同源性 miRNA 序列方面也具有优越性.

由于 miRNA 的时空和组织的特异性表达以及克隆方法的局限性,运用生物信息学方法来探测 miRNA 基因就显得很重要. 用生物信息学方法预测 miRNA,加快了对 miRNA 的研究步伐. 而研究 miRNA 的前体和成熟序列,总结它们的生物学特征,也将促进 miRNA 的计算机预测方法的发展和完善. 各种寻找 miRNA 基因的方法和策略都有了成功的例子,但单独应用某一种方法寻找出所有物种的 miRNA 基因是不现实的. 目前已鉴定的 miRNA 基因还远未达到饱和,要进一步鉴定出未知的 miRNA 基因,这有赖于各种方法和策略的相互配合,以及新技术和新方法的发展.

对预测的 miRNA 基因,还须进一步实验验证. 通常利用 Northern 杂交实验等作进一步确认,而最具有说服力的是功能实验,即通过体内、体外实验,证明这些小 RNA 通过抑制靶基因的表达而参与基因调控. 虽然用生物信息学方法识别出具有前体发夹结构特征的小 RNA 只是判断其是否为 miRNA 的初步工作,但是发现这些小 RNA 也是非常有意义的,它为进一步的实验研究提供了极其宝贵的信息,可大大节省实验工作的时间和经费开销.

参考文献:

[1] 何晨,谭军,陈薇,等. MicroRNA 研究进展[J]. 生物技术通报,2006(1):18-21;25.

[2] 金由辛. 核糖核酸与核糖核酸组学[M]. 北京:科学出版社,2005:106-130.

[3] LIM P L, LAU N C, WEINSTEIN E G, et al. The microRNAs of *Caenorhabditis elegans* [J]. Genes & Dev, 2003, 17:991-1008.

[4] LAI E C, TOMANCAK P, WILLIAMS R W, et al. Computational identification of *Drosophila* microRNA genes [J]. Genome Biology, 2003, 4(7):R42.

[5] 侯妍妍,应晓敏,李伍举. microRNA 计算方法的研究进展[J]. 遗传,2008,30(6):687-696.

[6] 张冬宁,刘阳,王翼飞. MiRdetector:一个预测与搜寻 miRNA 基因的计算工具[J]. 上海大学学报:自然科学版,2006,12(4):376-382.

[7] GRUNDHOFF A, SULLIVAN C S, GANEM D. A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses [J]. RNA, 2006, 12:733-750.

[8] JONES-RHOADES M W, BARTEL D P. Computational identification of plant microRNA and their targets, including a stress-induced miRNA [J]. Molecular Cell, 2004, 14(6):787-799.

[9] XUE C H, LI F, HE T, et al. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine [J]. BMC Bioinformatics, 2005, 6:310.

[10] NAM J W, SHIN K R, HAN J, et al. Human micro-RNA prediction through a probabilistic co-learning model of sequence and structure [J]. Nucleic Acids Research, 2005, 33(11):3570-3581.

[11] 贾万忠,李志,伦照荣. 病毒微小 RNA 的发现及其功能[J]. 科学通报,2007,52(23):2705-2714.

[12] 陈薇,谭军,何晨. 植物 miRNAs 前体的生物信息分析[J]. 重庆邮电学院学报:自然科学版,2006,18(6):803-808.

(编辑:刘志强)