基于基因簇判别的人类miRNA功能预测研究

丁 涛 高 洁*

(江南大学理学院, 无锡 214122)

摘要 随着新一代测序技术的不断发展, 面对海量的序列数据, 如果仅靠生物实验的方法来挖掘微RNA(microRNA, miRNA)的基因功能似乎不可能, 因此, 通过判别新miRNA家族归属来预测其相关生物学功能为实际生物实验的研究开辟新的思路。该文运用基因簇判别方法, 基于原始家族信息, 对未确定家族归属或新发现的miRNA进行判别, 确定其基因家族。研究发现, 同一家族的成熟体miRNA成员序列之间存在高度相似性, 并且参与相同的调控通路或作用于相同的靶基因, 具有相似的生物学功能。因此, 通过基因簇判别预测新miRNA家族归属, 对新miRNA的基因表达实验与验证具有十分重要的指导意义。

关键词 miRNA; miRNA家族; 基因簇判别; K-折交叉验证

Prediction of Human miRNA Functions by Gene Cluster Discriminant Analysis

Ding Tao, Gao Jie*

(*School of Science, Jiangnan University, Wuxi 214122, China*)

Abstract With the development of new generation of sequencing technology, it seems impossible to find miRNA functions through biological experiment alone. So identifying the family for new miRNA and predicting its biological functions will provide a new method for experimental research. Based on the original family information, the unclassified or new miRNAs can be classified into a definited family with gene cluster discriminant analysis. The results show that there is a high degree of similarity among mature miRNAs in the same family. Considering that members in the same miRNA family participate in the same pathway or act on the same target genes, the same family miRNAs will have similar biological functions. Thus, the unclassified miRNA can be identified by gene cluster discrimination, which plays a vital role in experimentation and verification for new miRNA functions.

Keywords miRNA; family classification; gene cluster discriminant analysis; K-fold cross-validation

微RNA(microRNA, miRNA)是一类内源性长约19~24个核苷酸的非编码单链RNA, 进化中具有高度保守性, 在转录后水平调控基因的表达, 从而在生物过程中发挥重要的作用。生物信息学分析也表明, 人类全部基因的1/3都受到miRNA的调控。研究发现, miRNA能够特异性和靶mRNA的不精确互补配对而裂解或抑制蛋白质的翻译, 从而进一步抑制蛋白质的合成, 最终采取何种调控机制, 是由靶基因与miRNA的匹配部位和程度来决定的[1]。当然, 任何一个基因都不是孤立的, 而是和其他生命要素相互作用来共同完成某种功能。因此, 一组具有某种相似性的基因往往会和某个共同的疾病都有关系[2]。

收稿日期: 2016-06-13 接受日期: 2016-10-25

国家自然科学基金(批准号: 11271163)和江苏省研究生科研创新计划项目(批准号: KYZZ16_0309)资助的课题

*通讯作者。Tel: 0510-85912033, E-mail: ezhun6669@sina.com.

Received: June 13, 2016 Accepted: October 25, 2016

This work was supported by the National Natural Science Foundation of China (Grant No.11271163) and the Foundation of the Innovation Project of Jiangsu Province (Grant No.KYZZ16_0309)

*Corresponding author. Tel: +86-510-85912033, E-mail: ezhun6669@sina.com

网络出版时间: 2016-12-19 16:12:56 URL: http://www.cnki.net/kcms/detail/31.2035.Q.20161219.1612.012.html

事实上, miRNA在基因组里往往成簇分布, 而一簇或家族中的miRNA常在很大程度上共表达[3], 与共基因相似, 它们很可能具有某种相似的功能, 参与相似的生命过程。例如, 已知一组基因中部分基因和某疾病相关, 那么该一簇或家族中的其他miRNA基因有较大的概率是与该疾病相关的。此外, miRNA无论是核苷酸序列还是二级结构在进化过程中是高度保守的, 因此它们会在进化关系相近的物种之中保守出现[4]。特别地, 隶属于同一家族的miRNA通常具有一致的序列结构和相似的生物功能[5]。依据这个概念, 将新的miRNA按照家族关系快速准确的归类在生物学上是一件非常有意义的工作。

近年来, 利用直接克隆、正反向遗传学技术等生物实验与方法确定了少数miRNA基因的生理功能, 但是仍有大量的miRNA基因功能尚未确定。值得注意的是, miRNA具有染色体上成簇排列形成同源性基因群的位置特性, 从而相同家族的成员总是有一致的结构和相近的功能。所以, 一个很有生物学意义的研究就是将这些有相近结构和作用的miRNA归类为同一个家族, 根据家族的基因表达谱来预测新miRNA的功能。2003年, 由Griffiths-Jones等[6]实验室构建的Rfam开源数据库主要提供关于这些非编码小分子RNA的家族注释信息, 为每个家族建立一个一致性二级结构和特有的协方差模型。2011年, 丁建栋等[7]基于N-元文本特征提取法(N-Grams)和多分类SVM(multiclass SVM)提出了miRFam, 用于分类miRNA家族。同样2014年, 基于miRNA的家族归属特性, Quan等[8]提出一种分层级联的家族分类预测方法: miRClassify。在这些传统miRNA家族研究中, 大部分处理手段是对miRNA序列及其二级结构数字化, 再将这些高维向量特征提取降维, 建立不同的算法(机器学习、多分类等)来预测miRNA家族信息。这些方法虽然可以较好地预测家族信息, 但随着miRNA数据的爆发式增长, 基因家族分类预测训练又需要考虑到大量的序列和结构信息, 此类方法将耗费相当长的时间。

为了大大减少新miRNA家族预测成本与时间, 本文基于前人miRNA家族研究成果与不断涌现新的miRNA信息, 提出新的miRNA家族预测方法——基因簇判别法。直接利用现有的miRNA家族数据库, 对家族成员高度相似性的序列信息建立判别指标, 确定新miRNA的家族归属。这种方法仅仅从一级序列中提取不同的碱基含量作为特征, 而不考虑复杂的空间结构, 大大提高了实验运行速度, 降低了算法的复杂度。最终, 通过对原始数据进行K-折交叉验证[9](K-fold cross-validation), 实验验证发现, 基因簇判别的确具有较高的精确性与实用性。

1 材料与方法

本文从已知的miRNA家族关系出发, 提出新的基因家族分类方法——基因簇判别法, 为预测新miRNA的生物学功能提供新的思路。miRBase(http://www.mirbase.org/)是基于种子区域的序列相似性的miRNA家族分类[10], 其数据库在过去8年里从Version 8.2(07/2006)更新至Version 21.0 (06/2014), 人血清白蛋白(human serum albumin, Hsa)的miRNA家族数从342个发展到583个, 增长了1.7倍。特别地, 未确定家族miRNA数量较版本8.2增长了7.5倍(具体miRNA及家族增长数见表1), 越来越多的miRNA生物学功能与家族信息亟待我们挖掘。基于这860个成熟体miRNA家族归属问题, 本文将这些miRNA判别到已知583个家族中, 根据原家族成员的共同生物学功能, 为未知miRNA可能具有的功能提供理论依据。

鉴于miRNA在进化过程中的高度保守性, 同一家族的miRNA通常具有相似的序列与二级结构[3]。实验从miRBase(Version 21.0)中提取所有人类成熟体miRNA序列信息, 引用碱基含量将miRNA基因数字特征化。一个成熟体miRNA是由长度为*n*的A、U、

表1 从miRBase版本8.2到版本21.0的miRNA与miRNA家族个数变化

Table 1 Changes of miRNA family numbers from miRBase version 8.2 to version 21.0

版本变化 Version changes	miRNA数 miRNAs	miRNA家族 miRNA families	未确定家族miRNA数 Unclassified miRNAs
Version 8.2	462	342	114
Version 21.0	1 767	583	860
Growth ratio (%)	4.1	1.7	7.5

C和G这4个碱基组成, 我们分别计算序列中的单碱基含量、双碱基含量和三碱基含量:

$$x_i=\frac{n_i}{n}\times 100\%,\ i=m,mn,mnk,\ \text{其中},\ m,n,k\in\{A,U,C,G\}$$

因此, 所有miRNA序列均可转化成$4^1+4^2+4^3=84$维的特征向量x:

$$(x_{A\%},x_{U\%},\cdots,x_{G\%},x_{AA\%},x_{AU\%},\cdots,x_{GG\%},x_{AAA\%},x_{AAU\%},\cdots,x_{GGG\%})'$$

与多元统计中判别分析[11]类似, 本文提出基因簇判别: 设有k个miRNA家族G_1、$G_2\cdots G_k$, 对任一未确定家族的miRNA x, 构造马氏距离线性判别函数$W_{ij}(x)$:

$$W_{ij}(x)=\left(x-\frac{\mu_i+\mu_j}{2}\right)'\Sigma^{-1}(\mu_i-\mu_j),\quad i,j=1,2,\cdots,k \tag{1}$$

其中, μ_i、μ_j为miRNA家族G_i、G_j所有基因特征向量的均值, Σ为协方差矩阵。

值得注意的是, 函数(1)只对该k个家族的协方差阵Σ_i相同时成立, 当方差阵不全相同时, 对函数$W_{ij}(x)$进行修正记为$\widetilde{W}_{ij}(x)$:

$$\widetilde{W}_{ij}(x)=(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)-(x-\mu_j)'\Sigma_j^{-1}(x-\mu_j) \tag{2}$$

当函数值$W_{ij}(x)$或$\widetilde{W}_{ij}(x)>0(\forall i\neq j)$时, 则认为待测miRNA序列$x$属于家族$G_i$; 若当函数值$W_{ij}(x)$或$\widetilde{W}_{ij}(x)=0$时, 则$x$记为待判。

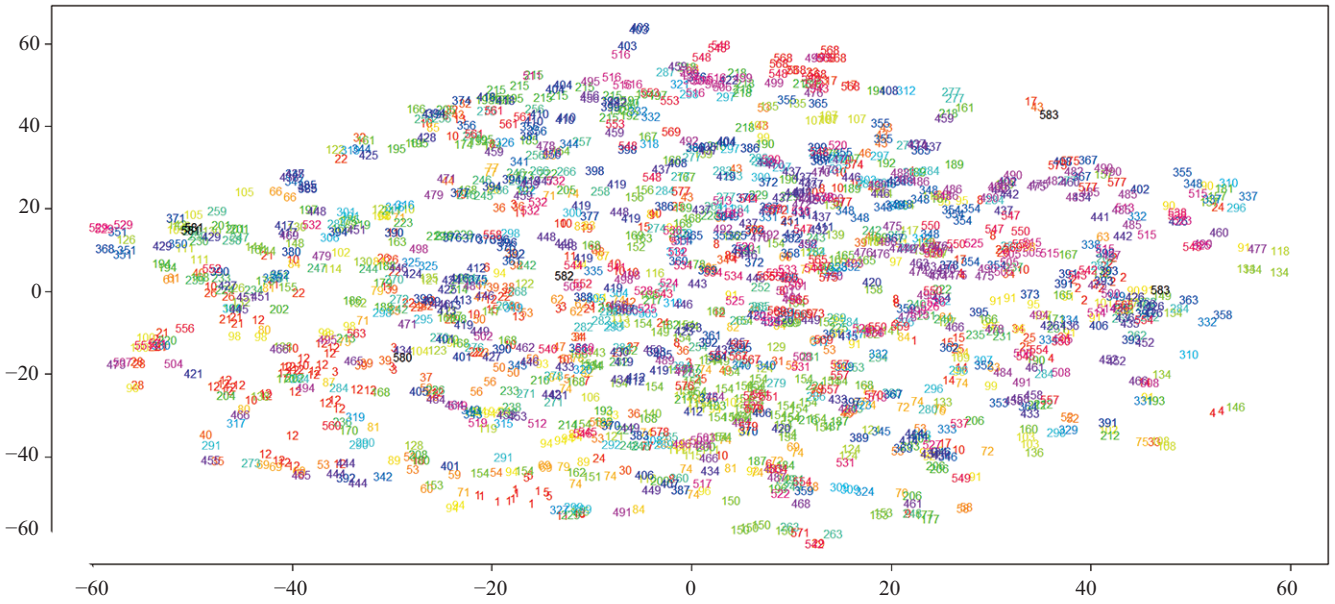
基因簇判别的误差通常用回代误判率E进行估计: 基于上述确定的miRNA家族总体及各家族成员(m_i表示miRNA家族总体的家族成员个数)$G_1:y_{11},y_{12},\cdots,y_{1m_1}$、$G_2:y_{21},y_{22},\cdots,y_{2m_2}$、…、$G_k:y_{k1},y_{k2},\cdots,y_{km_k}$为训练样本, 以全体训练样本$m_1+m_2\ldots+m_k$作为新样品, 逐个代入已建立的基因簇判别准则中判别其家族归属。构造一个$k\times k$阶混淆矩阵N:

$$N=\begin{bmatrix} N_{11} & \cdots & N_{1k} \\ \vdots & \ddots & \vdots \\ N_{k1} & \cdots & N_{kk} \end{bmatrix} \tag{3}$$

其中, $N_{ij}(i\neq j$且$i,j\in k)$表示属于家族G_i的miRNA被误判到属于家族G_j的个数, 混淆矩阵的对角元素即为回代后的结果与回代前相同的miRNA个数。因此得到基因簇判别误判率:

$$E=100\%-\frac{N_{11}+N_{22}+\cdots+N_{kk}}{m_1+m_2+\cdots+m_k}\% \tag{4}$$

2003年, miRNA的靶基因预测算法TargetScan提出了种子区域概念, 它是指miRNA成熟序列中5′端的2~8个碱基[12]。研究发现, 作为miRNA的核心功能区域, 种子区域序列的保守性更高。因此, 本文从miRBase中获取人类1 767个miRNA序列, 最



不同的数字表示不同的miRNA家族。

Different figures represent different miRNA families.

图1　miRNAs平面分布图

Fig.1　The plane distribution map of miRNAs

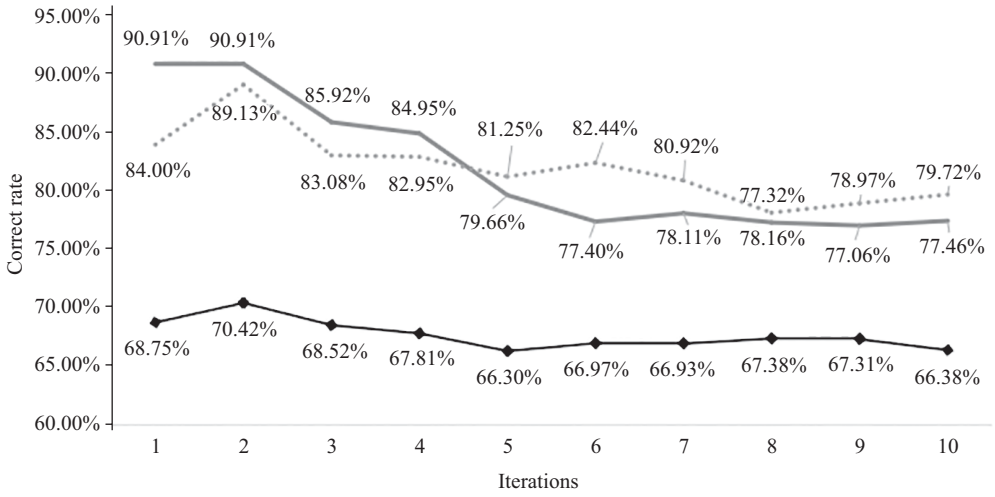


图2 不同数据产生的准确率

Fig.2 Correct rates by different experiments

终去重整合得到1 700个5′端成熟体miRNA, 583个miRNA家族, 实验将未确定家族归属的860个miRNA判别到这583个miRNA家族中, 最终实验得到基因簇判别的误判率为0.98%。统计实验结果发现, 只含有一个成员的家族有227个, 同时存在2个庞大的家族miR-548与miR-515其家族成员分别达到55和38, 而大部分家族成员数均在2~15之间。图2表示由583个数字代表的583个miRNA家族, 共计1 700个miRNA在平面中相对位置分布。

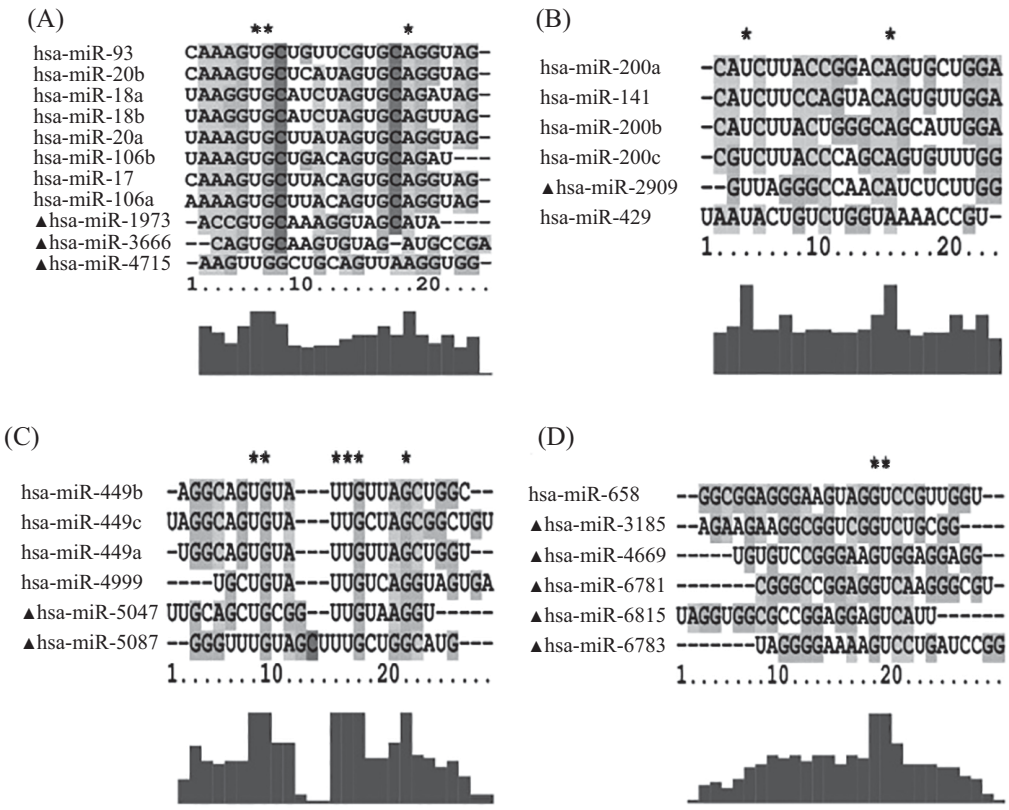
2 结果

2.1 方法验证

为了验证基因簇判别的准确性, 实验对已知家族归属的miRNA进行K-折交叉验证。通过利用大量数据集, 使用不同折数K进行大量的交叉验证结果试验表明, K-折交叉验证是获得最好误差估计的恰当选择[13]。因此, 本文设定K=10, 即将原始miRNA数据集等分为10份, 取9份作为训练集, 1份作为验证集。首先用训练集对分类器进行训练, 再利用验证集来测试训练集, 如此迭代10次, 计算最后的误差来作为预测误差。由于家族成员多的家族提供的序列信息量大, 判别到该家族miRNA的准确率就越高, 因此除了对含有2个及以上家族成员的miRNA进行实验外, 还分别对家族成员为3、4个及以上家族的miRNA进行实验。虽然不同数据集的选择对实验结果的准确率具有一定的影响, 但随着迭代次数的增加实验准确率趋于稳定(图3)。当至少含有2个家族成员的家族作为待测数据时, 实验的准确率为66.38%, 对至少含有3、4个成员的miRNA家族的实验准确率均明显有所提高, 最终实验得到准确率分别为77.46%和79.22%, 这表明基因簇判别对至少含有3个家族成员的家族判别更具有实用性。

2.2 结果分析

为了更详细地说明基因簇判别对miRNA家族成员的判别能力, 我们采用Clustal X2[14]对判别结果进行后续处理。随机选取4个miRNA家族: miR-17家族、miR-200家族、miR-449家族、miR-658家族。为了区分未知家族的miRNA和已知的miRNA, 我们在未知的miRNA序列名称前加了三角形“▲”, 如图3所示, 当家族所有成员在某位点碱基相同时, 该碱基位点上方用“*”表示, 图形下方的直方图表示各位点众数最大的碱基在该位点的碱基比重。由上面验证结果可知, 基因簇判别对新miRNA判别到至少含有3个家族成员的miRNA家族中的准确性较高。因此, miR-17家族(成员数11)、miR-200家族(成员数6)和miR-449家族(成员数6)在图中能清晰地说明各家族内确定家族的miRNA与新加入的miRNA成熟体序列之间具有较高的相似性, 即家族中成员基因序列具有良好的保守性, 这极大地说明了基因簇判别对新miRNA的家族归属判别质量非常好。除此, 我们还选取了只含1个确定家族信息的miR-658家族, 不难发现新判别的5个新miRNA序列与miRNA-658序列也具有一定的相似性。



图中A、B、C、D 分别表示miR-17家族、miR-200家族、miR-449家族和miR-658家族。在每个家族中, miRNA名称前的“▲”表示该miRNA为新判别进来的成员。当新家族所有成员在某位点碱基相同时, 该碱基位点上方用“*”表示, 图形下方的直方图表示各位点众数最大的碱基在该位点的碱基比重。

In the figure, A, B, C and D represent the family of miR-17, miR-200, miR-449 and miR-658, respectively. At the each of families, we add an asterisk “▲” in front of the new family members. When all the members of the family have the same base at a certain position, the top of the base site is also indicated by “*”. The histogram below the graph shows the percentage of the largest number of bases at each site.

图3　4个家族经过Clustal X2处理后的结果

Fig.3　The results of 4 miRNA families by Clustal X2

通过人类miRNA家族预测研究, 确定新的家族信息对miRNA家族表达谱以及单个miRNA功能研究提供重要的帮助。具体各家族成员分别居于何种调节机制, 还需要对靶基因结合点和通路进行进一步的研究。但接智慧等[15]阐述了miR-17家族基因在宫颈癌、子宫内膜癌与卵巢癌多种细胞系中的作用机制, 而最近Li等[16]发现, 本文预测的miR-17家族新成员miR-3666能够增强宫颈癌细胞的转移, 这再一次证明我们实验结果的实用性。相似地, miR-200家族调控消化系统肿瘤的发生[17]等, 这些工作均为我们预测新发现的miRNA基因功能提供了有利的信息。

3　讨论

miRNA在参与个体发育、细胞分化增殖、细胞凋亡、激素分泌和脂质代谢等多个生理过程, 与肿瘤、代谢性疾病、应激性疾病、心血管疾病和自身免疫性疾病的发生、发展密切相关[18]。随着miRNA生物信息学方向预测的发展与实际实验技术的提升, 大量的miRNA被发掘与验证, 研究miRNA功能相关工作也在不断深入。新miRNA的发现为基因表达调控的多样性和复杂性开辟了新的研究视角, 因此, 利用已有的miRBase数据库中miRNA家族信息, 借助于生物信息学的分析, 将基因序列信息转化成数字特征信息。同多元统计中的判别分析类似, 将确定家族的基因序列数字特征作为训练样本, 未确定家族的基因作为待判样本进行基因簇判别, 分析出目前已经发现的所有人类miRNA家族信息。依赖于miRNA家族的正确分类与同一家族成员生物学功能的相似性, 通过基因簇判别确定新miRNA的家族归属, 为miRNA生物学功能预测提供新的思路。虽然更确切的结果还需

进一步实验验证, 且原始基因家族的分类、靶基因集合的质量, 是直接影响接下来生物富集分析成功与否最重要的基础之一, 但是理论预测为实验验证提供了方向和可能, 节省了更多的资源。

参考文献 (References)

1 Czech B, Hannon GJ. Small RNA sorting: Matchmaking for argonautes. Nat Rev Genet 2011; 12(1): 19-31.

2 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2012; 102(43): 15545-50.

3 Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA 2005; 11(3): 241-7.

4 Cuperus JT, Fahlgren N, Carrington JC. Evolution and functional diversification of MIRNA genes. Plant Cell 2011; 23(2): 431-42.

5 Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, Gorodkin J. Structural profiles of human miRNA families from pairwise clustering. Bioinformatics 2009; 25(3): 291-4.

6 Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: An RNA family database. Nucleic Acids Res 2003; 31(1): 439-41.

7 Ding J, Zhou S, Guan J. miRFam: An effective automatic miRNA classification method based on n-grams and a multiclass SVM. BMC Bioinformatics 2011; 12(1): 216.

8 Quan Z, Mao Y, Hu L, Wu Y, Ji Z. miRClassify: An advanced web server for miRNA family classification and annotation. Comput Biol Med 2014; 45(2): 157-60.

9 Rodríguez JD, Pérez A, Lozano JA. Sensitivity analysis of k-fold cross validation in prediction error estimation. IEEE TPAMI 2010; 32(3): 569-75.

10 Griffiths-Jones S, Saini HK, Van DS, Enright AJ. miRBase: Tools for microRNA genomics. Nucleic Acids Res 2008; 36(Database issue): 154-8.

11 何晓群. 应用多元统计分析. 北京: 中国统计出版社(He Xiaoqun. Applied multivariate statistical analysis. Beijing: China Statistics Press) 2010, 89-112.

12 Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell 2004; 115(7): 787-98.

13 李艳芳, 王 钰, 李济洪. 几种交叉验证检验的可重复性. 太原师范学院学报: 自然科学版(Li Yanfang, Wang Yu, Li Jihong. The replicability of several cross-validated tests. Jouranl of Taiyuan Normal University, Natural Science Edition) 2013; 12(4): 46-9.

14 Larkin MA, Blackshields G, Brown NP, Chenna RM, Mcgettigan PA, Mcwilliam H, *et al*. Clustal W and Clustal X version 2.0. Bioinformatics 2007; 23(21): 2947-8.

15 接智慧, 吴建磊, 陶 陶, 史春雪, 王 敏. miR-17~92基因簇在宫颈癌、子宫内膜癌与卵巢癌多种细胞系中的表达及意义. 中国医科大学学报(Jie Zhihui, Wu Jianlei, Tao Tao, Shi Chunxue, Wang Min. Expression and significance of miR-17-92 gene clusters in cervical cancer, endometrial carcinoma and ovarian cancer cell lines. Journal of China Medical University) 2013; 42(3): 209-13.

16 Lin L, Han LY, Ming Y, Qi Z, Xu JC, Ping L. Pituitary tumor-transforming gene 1 enhances metastases of cervical cancer cells through miR-3666-regulated ZEB1. Tumor Biol 2015: doi: 10.1007/s13277-015-4047-1.

17 唐峰波, 杨 铭. MiR-200家族与消化道肿瘤的相关研究. 中华临床医师杂志: 电子版(Tang Fengbo, Yang Ming. Allied study between miR-200 family and digestive system neoplasms. Chinese Journal of Clinicians, Electronic Edition) 2016; 10(1): 110-5.

18 Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. Cell 2012; 148(6): 1172-87.