流形排序算法预测 microRNA*

王常武, 刘兵强, 王宝文, 刘文远
(燕山大学 信息科学与工程学院, 河北 秦皇岛 066004)

摘 要: 在已知 microRNA(miRNA)较少的情况下,为了提高算法预测的准确性,提出一种基于流形排序的 miRNA 预测算法。该算法采用加权图模型描述序列,使用置信传播分配排序分数,降低了算法的时间复杂度;算法根据大规模数据内部全局流形结构进行排序,提高了排序结果的准确性。在人类和按蚊全基因组范围内的实验证明,流形排序算法的预测效果优于传统的预测方法,可以作为预测 miRNA 的一个有效工具。

关键词: 微小 RNA; 加权图; 置信传播; 流形排序; 预测; 生物信息学
中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2012)03-0819-04
doi:10.3969/j.issn.1001-3695.2012.03.004

MicroRNA prediction based on manifold ranking

WANG Chang-wu, LIU Bing-qiang, WANG Bao-wen, LIU Wen-yuan
(*College of Information Science & Engineering, Yanshan University, Qinhuangdao Hebei* 066004, *China*)

Abstract: In order to improve the precision of microRNA prediction while the number of known microRNAs is small, this paper proposed a novel microRNA prediction algorithm based on manifold ranking. The algorithm adopted the strategy of modeling microRNA prediction process as belief propagation on a weighted graph, hence reduced the time complexity of the algorithm. The core idea of algorithm was to rank the data with respect to the intrinsic manifold structure collectively revealed by a great amount of data, hence enhanced the accuracy of the ranking results. Experiments on H. sapiens and anopheles gambiae genes show that manifold ranking algorithm is better than the traditional algorithm, and can be worked as an effective tool for predicting novel microRNAs.

Key words: microRNA; weighted graph; belief propagation; manifold ranking; prediction; bioinformatics

0 引言

microRNA 是一类长度约为 20 ~ 24 nt 的内源性非编码调控单链小分子 RNA[1]。早期 miRNA 预测主要通过 cDNA 克隆测序,这类方法很难捕获表达丰度较低的 miRNA。通过生物信息学方法预测 miRNA 不受其表达丰度的影响,可以弥补 cDNA 克隆测序方法的不足。

基于结构特征的预测算法是将最小自由能和二级结构保守性这两个参数相结合进行预测,但这类算法无法找出不具备保守结构的 miRNA[2]。基于序列特征的预测算法需要数量充足的已知 miRNA 作为先验信息,而目前已知 miRNA 的数量相对较少,从而限制了该类算法的使用。基于大规模测序的预测算法依赖于大规模测序技术所提供的转录本分析资源,但得到充足的序列测序数据比较困难,因此具有一定的局限性[3]。

基于机器学习的预测算法是近几年出现的 miRNA 预测算法。文献[4]用 32 个三联体结构—序列特征向量描述二级结构为发卡环的 pre-miRNA 序列和待测序列,通过构建分类器 SVM 预测人类序列的 miRNA,其精度达到了 90%。文献[5]使用 SVM 分析 pre-miRNA 的 18 个特征值来发现新的 miRNA,从而缩小了实验的候选集。文献[6]采用两个级联的分类器 microprocessor SVM 和 miRNA SVM 预测 miRNA。文献[7]采用随机森林(random forest)方法构建区分 pre-miRNA 和非 pre-miRNA 的分类器,增加最小自由能和随机检验值描述样本。文献[8]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。根据已知 pre-miRNA 和未知茎环结构片段估计 HMM 转移概率和发射概率,取得了较好的预测效果。文献[9]在加权图上构建了一个马尔可夫随机游模型预测 miRNA。

传统的基于机器学习方法采用样本间的距离度量样本的相似性,没有整体考虑序列在特征空间中的相关性。基于流形排序的预测算法从标记样本和未标记样本中综合分析数据的结构和分布信息,利用特征空间中的潜在流形分布结构对相应的序列进行排序[10]。算法在保持数据局部结构的同时也保持了数据的整体流形结构。在人类和按蚊中的实验结果证明,在已知 miRNA 较少的情况下,该算法灵敏度相同时的特异性优于传统机器学习方法。

1 方法设计

流形排序是一个半监督学习的特例。半监督学习是利用未标记数据学习的主流技术之一,能在不加外界干预的情况

收稿日期: 2011-07-31; **修回日期**: 2011-09-26 **基金项目**: 国家自然科学基金资助项目(60970123)

作者简介: 王常武(1970-),男,黑龙江人,教授,博士,主要研究方向为生物信息计算、智能计算;刘兵强(1983-),男,河北武安人,硕士研究生,主要研究方向为生物信息计算、智能计算(liu_bingqiang@126.com);王宝文(1957-),男,黑龙江齐齐哈尔人,副教授,主要研究方向为生物信息计算、智能计算;刘文远(1968-),男,黑龙江密山人,教授,博士,主要研究方向为智能计算、生物信息计算.

下,自动地利用少量已标记数据和大量未标记数据进行学习[10]。半监督学习通过数据分布与标记之间的联系在所有样本上进行全局优化,对减少标注代价、提高学习机器性能具有重要的实际意义[11]。流形排序算法把数据描述成欧几里德空间中的向量,并根据其内部全局流形结构对大规模数据排序。

本文将预测 miRNA 视为从包含茎环结构的候选序列中检索的问题。已知 miRNA 作为标记样本,在候选集合中依照标记样本检索出新的 miRNA。大规模高维数据集合可视为流形结构数据[11]。本文使用三联体结构—序列特征抽取序列的特征值,得到大量由序列特征值组成的高维向量,并用流形结构数据描述序列的特征向量。流形结构数据能够整体考虑数据集的全局结构,在保持局部特征的同时也保持了全局特征。

在本文中,使用加权图模型描述序列之间的关系;置信传播根据未标记样本和标记样本的关系给每条候选序列分配相应的排序分数;根据分数排序节点,并按照排序结果从未标记样本中检索新的 miRNA。

1.1　提取候选序列的特征向量

预测 miRNA 的原始数据为 miRNA 前体序列,是由 ACGU 四个碱基组成的一维线性序列。例如,hsa-mir-20,其一维结构如图 1 所示。为了提高检索的精确性,使用三联体结构—序列特征描述原始数据的结构特征,把序列的结构特征值用一个高维向量表示[4]。MiRNA 前体序列的二级结构为发卡环,使用 RNAfold 预测它的二级结构和最小自由能。序列的二级结构用点括号图描述。其中‘.’表示未配对碱基;‘(’表示靠近 miRNA 前体序列 5端(miRNA 前体序列的始端)的配对碱基;‘)’表示靠近 miRNA 前体序列 3端(miRNA 前体序列的末端)的配对碱基。

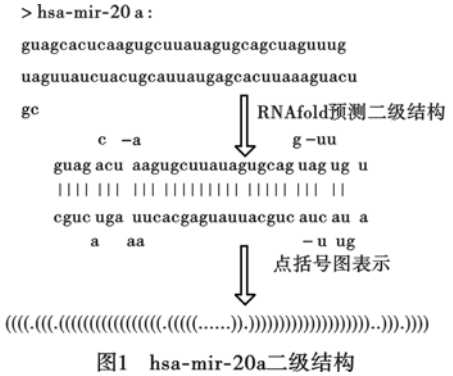

图1　hsa-mir-20a二级结构

由 miRNA 生源论可知,成熟的 miRNA 源自于其前体的双螺旋结构解旋后的一条单链,目前还不清楚参与基因调控单链的位置[1]。所以,三联体结构—序列特征描述原始数据的结构特征只考虑‘(’与‘.’两种状态。由于两个邻接的碱基不能反映序列的结构特征,而选择较多的邻接碱基会在数据规模上呈指数级增长。所以,选择三个相邻碱基描述序列的局部特征。三个‘(’或‘.’符号共有八种组合方式,加上四种不同的碱基(ACGU),共有 32 个结构特征描述每条序列。序列结构特征的数值化可由 $\Psi(x_i)=|(\mu_i^+-\mu_i^-)/(\sigma_i^++\sigma_i^-)|$ 计算得出。序列的每个特征 x_i 可用其对应的 Ψ 值表示。文献[12]计算出了中序列所有局部特征的参数值。表 1 列举了 10 个常用的结构特征参数。

表 1　三联体结构—序列特征参数表

结构特征	microRNA		待排序列	
	μ^+	σ^+	μ^-	σ^-
A(((	0.121	0.424	0.063	0.032
U(((	0.154	0.048	0.089	0.040
C…	0.006	0.011	0.025	0.030
A…	0.008	0.014	0.025	0.025
U…	0.007	0.011	0.021	0.023
G.((	0.042	0.025	0.063	0.031
C(..	0.009	0.011	0.019	0.017
C((.	0.032	0.022	0.048	0.027
A(..	0.011	0.012	0.020	0.016
G(((	0.151	0.038	0.127	0.040

其中:μ_i^+ 和 σ_i^+ 分别表示已知 miRNA 序列局部特征的均值和标准差,μ_i^- 和 σ_i^- 分别表示待测序列局部特征的均值和标准差。

序列的结构特征用 32 个特征值表示,有效反映了序列局部连续的精细结构和整体布局。从序列的二级结构中抽取出最小自由能(MFE)、环的长度、茎上每条臂的碱基配对数四个全局特征。最终用 36 维向量表示 miRNA 前体的全局和局部特征。通过对基因组中大量原始序列全局特征和局部特征的抽取,这些特征向量形成的数据是一组大规模的高维数据,可用流形结构数据描述。

1.2　构建加权图模型

在一组相同物种的基因序列中,已知 miRNA 组成标记样本,未知序列组成未标记样本。加权图模型能够描述序列以及序列间的关系。在图模型中,每个节点代表一个 miRNA 或者一个候选序列,序列与序列之间的关系用节点间的边表示。边上的权值量化序列间的相似性,权值越大表示边上两条序列的相似性越大。候选序列与已知 miRNA 相似性越大,该序列为新 miRNA 的可能性就越大。通常情况下,权值的大小由序列间的成对距离(pairwise distances)决定,这样就使两个关系密切的序列之间拥有一条权值较大的边。

在加权图模型中,代表序列的节点集合为 $X=\{x_{q1},\cdots,x_{qn},x_{u1},\cdots,x_{um}\}$。其中,节点 x_{q1} 到 x_{qn} 代表标记样本,由已知 miRNA 组成,x_{u1} 到 x_{um} 代表未标记样本,由候选序列组成。未标记样本与标记样本的相似性由权值矩阵 $W_{ij}=\exp\{-d(x_i,x_j)^2/2\sigma^2\}$ 描述。其中,σ 为热核参数(heat kernel parameter),$d(x_i,x_j)$ 表示两个样本间的欧几里德距离。检索新的 miRNA 是根据标记样本排序未标记样本的过程。所以,在初始状态时,为图中每个节点分配初始置信度 F_i,标记样本的初始置信度 $F_i=1$,未标记样本的初始置信度 $F_i=0$。置信传播迭代完毕后,每个节点都更新其置信度。节点 F_i 的值越较大表示该节点所代表的序列与标记样本序列的相似性越高,即该序列为新 miRNA 的概率越大。

1.3　流行排序算法预测 miRNA

置信传播是加权图模型上一个经典的随机过程,用来研究样本的概率以便发现其内在结构。图模型上的每个节点对应

一个样本。置信传播开始时,节点 x_i 到 x_j 的传播概率定义为

$$p_{ij}=\frac{w_{ij}}{d_i} \tag{1}$$

其中:w_{ij}为节点 x_i 与 x_j 边上的权值;$d_i=\sum_j w_{ij}$表示节点 x_i 的度,即节点 x_i 与全部邻居节点边上的权值之和。式(1)用矩阵表示为 $\boldsymbol{P}=\boldsymbol{D}^{-1}\boldsymbol{W}$。其中:$\boldsymbol{D}$ 为一个邻接矩阵,$\boldsymbol{D}$ 中的每个对角元素 $d_{ii}=\sum_j W_{ij}$表示该节点的度;矩阵 $\boldsymbol{P}$ 描述了图模型中各个节点间传播的概率,把矩阵 $\boldsymbol{P}$ 划分为

$$\boldsymbol{P}=\begin{pmatrix} P_{QQ} & P_{QU} \\ P_{UQ} & P_{UU} \end{pmatrix} \tag{2}$$

矩阵 $\boldsymbol{P}$ 的划分包括四部分,分别为标记样本间传播的概率 P_{QQ};标记样本向未标记样本传播的概率 P_{QU};未标记样本向标记样本传播的概率 P_{UQ};未标记样本间传播的概率 P_{UU}。相应地,权值矩阵 $\boldsymbol{W}$ 和邻接矩阵 $\boldsymbol{D}$ 也划分为

$$\boldsymbol{W}=\begin{pmatrix} W_{QQ} & W_{QU} \\ W_{UQ} & W_{UU} \end{pmatrix},\boldsymbol{D}=\begin{pmatrix} D_{QQ} & \boldsymbol{O} \\ \boldsymbol{O} & D_{UU} \end{pmatrix} \tag{3}$$

其中,$\boldsymbol{O}$ 表示元素全为 0 的矩阵。

图模型上的节点通过置信传播更新和传播置信度。当节点 x_i 到 x_j 的置信传播发生时,由 x_i 传播其置信度给 x_j,x_j 根据传播规则更新自己的置信度。这是一个动态传播过程:节点把置信度传播给邻居节点,同时也接收邻居节点的置信度。在传播结束后,节点根据邻居节点传播的置信度更新自身节点的置信度。因为标记样本作为置信度最高的节点在传播过程中起范例的作用,所以标记样本只传播置信度,不接收邻居节点的置信度。图模型中置信传播规则为

$$f_i^{(k+1)}=\alpha\sum_{x_j\in U}p_{ij}f_j^{(k)}+\sum_{x_j\in Q}p_{ij}f_j \tag{4}$$

其中,k 为置信传播的迭代次数,$\alpha\in[0,1)$设定未标记样本在置信传播中比例的大小。上式的矩阵表达式为

$$f_U^{(k+1)}=\alpha P_{UU}f_U^{(K)}+P_{UQ}f_Q \tag{5}$$

由式(1)(3)可知,式(5)可表示为

$$f_U^{(k+1)}=\alpha D_{UU}^{-1}W_{UU}f_U^{(k)}+D_{UU}^{-1}W_{UQ}f_Q \tag{6}$$

在本文中,使用置信集 $L=\{1,\cdots,c\}$ 标记集 $X=\{x_{q1},\cdots,x_{qn},x_{u1},\cdots,x_{um}\}$ 中每条序列的置信度。把集合 X 中前 n 个已知 miRNA 序列标记为 $y_i\in L$ 的标记样本,将未知序列标记为未标记样本。为了讨论方便,定义 Φ 为一个 $(n+m)\times c$ 阶矩阵集。矩阵 $\boldsymbol{F}=[F_1,\cdots,F_n]^{\mathrm{T}}\in\Phi$ 作为分类器给序列 x_i 分配一个置信度 $y_i=\arg\max_{j\leqslant c}F_{ij}$。分类器也可以写成一个矢量函数 $\boldsymbol{F}:X\to\mathbb{R}^c$,这个函数为每条序列分配置信度 F_i。使用矩阵 $\boldsymbol{Y}\in\Phi$ 中元素 Y_{ij}描述集合 X 中节点的标记情况,$Y_{ij}=1$ 表示节点 x_i 被标记为 y_i,否则 $Y_{ij}=0$。在初始状态时,对于已知 miRNA 序列,$Y_{ij}=1$;对于未知序列,$Y_{ij}=0$。流形排序算法的步骤可描述为:

a)通过建立矩阵 $\boldsymbol{S}=\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$归一化矩阵 $\boldsymbol{W}$。对 $\boldsymbol{W}$ 进行归一化处理可以使传播规则收敛,从而保证置信传播对称地执行。

b)迭代计算 $F(t+1)=\alpha SF(t)+(1-\alpha)Y,\alpha\in[0,1)$。其中,函数 $\boldsymbol{F}:X\to\mathbb{R}$ 根据传播规则为集合 X 分配置信度。参数 α 指定邻居节点与前一循环结束时节点的排序分数对本次排序分数的影响系数。

c)当集合 X 中每个节点的置信度不再更新时,迭代完毕。记 F^* 为序列$\{F(t)\}$的极限,F^* 为集合 X 中每条序列 x_i 分配置信度 $y_i=\arg\max_{j\leqslant c}F_{ij}^*$。文献[11]证明了 $F(t+1)=\alpha SF(t)+(1-\alpha)Y$ 的收敛性。

置信传播的规则是通过多次重复运算得到欲求答案的计算过程,一次比一次接近精确欲求的答案。置信传播结束后,未标记样本的排序分数收敛于 F^*。在置信传播开始前,集合 X 中已知 miRNA 序列的置信度设为 1;未知序列的置信度设为 0,即 $F(0)=Y$。由 $F(t+1)=\alpha SF(t)+(1-\alpha)Y$ 可知:

$$F(t)=(\alpha S)^{t-1}Y+(1-\alpha)\sum_{i=0}^{t-1}(\alpha S)^iY \tag{7}$$

由 $0<\alpha<1$ 且 S 的特征值位于闭区间$[-1,1]$可知:

$$\lim_{t\to\infty}(\alpha S)^{t-1}=0,\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha S)^i=(I-\alpha S)^{-1}$$

因此,

$$F^*=\lim_{t\to\infty}F(t)=(1-\alpha)(I-\alpha S)^{-1}Y \tag{8}$$

因为参数$(1-\alpha)$对置信度的排序结果没有影响,所以式(8)可简写为

$$F^*=(I-\alpha S)^{-1}Y \tag{9}$$

当置信传播迭代完毕时,节点的置信度不再改变且收敛于 F^*。通过迭代计算式(9),集合 X 中每个节点都得到自己的置信度 F^*。节点的 F^* 越高,该节点所代表序列为新 miRNA 的概率越大。对节点的 F^* 排序,检索出 F^* 较大节点,这些节点所代表的序列即为流行排序算法预测的结果。

2　实验及结果

2.1　实验数据和检验标准

从 miRNA 前体序列数据库 miRBase(http://www.mirbase.org/)中下载人类和按蚊的已知 miRNA 前体序列,从 UCSC(http://genome.ucsc.edu/)中下载具有茎环结构的 RNA 序列。

在 miRNA 的预测中,对预测准确率的评价使用敏感度和特异性。敏感度 $X=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$是指所有的已知 miRNA 对中被正确预测到的百分比,特异性 $Y=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$是指在所有预测到的结果中正确预测的百分比。其中,TP(true positive)表示正确预测 miRNA 的个数;FN(false negative)表示真实存在但没有被正确预测出的 miRNA 个数;FP(false positive)表示不存在但被错误预测到的个数。

2.2　实验结果

a)提取原始数据的全局和局部特征。四个全局特征分别是序列的最小自由能、环的长度、茎上每条臂的碱基配对参数。本文采用三联体结构—序列特征描述原始数据的局部特征。把每条序列的序列特征抽取为一个 32 维向量,经过对每条序列特征值的抽取,使用 36 个特征值组成的向量表示序列的全局和局部特征。

b)根据序列间的关系建立加权图模型,并利用在图模型上的置信传播给每个节点分配排序分数,根据分数大小进行相似性排序预测新 miRNA。

为了更准确地评估算法的预测效果,比较了基于 SVM 的分类算法和基于流形数据排序算法在人类基因和按蚊基因组中的预测效果。基于 SVM 的分类算法需要大量的训练样本作为正集,使用全部样本使得算法复杂性很高,核函数随训练集的增大而增加。预测的结果不具有统计解释性,而且由于计算过程牵涉复杂的优化过程而增加计算复杂性。为了直观地分析算法在已知 miRNA 较少数据集中的预测效果,参考了文献[9]的策略设计出四个正集数目分别为 1、10、20 和 50 的分类模型并与之比较。表 2 给出了在获取相同敏感度的情况下,流形排序算法在人类基因数据中的预测结果以及与 3SVM[4]、MiRFinder[5]、DSVM(microprocessor SVM + microRNA SVM)[6]、random forest[7]、ProMiR[8]、MiRank[9]方法的比较。从表 2 中可以看出,在已知 miRNA 较少的情况下,基于流形排序算法的预测效果明显优于其他方法。当 $N=1$ 时,特异性达到了 43.3%;当已知 miRNA 较多的情况下,该算法略优于传统预测算法,或者与传统预测算法预测效果相当。当 $N=50$ 时,特异性达到了 90%。

表 2　MicroRNA 预测方法在人类基因中的实验结果比较

方法	$N=1$	$N=10$	$N=20$	$N=50$
3SVM	0.218	0.628	0.705	0.810
MiRFinder	0.201	0.602	0.751	0.811
DSVM	0.291	0.611	0.714	0.801
随机森林	0.287	0.641	0.762	0.789
ProMiR	0.274	0.624	0.714	0.814
MiRank	0.292	0.695	0.753	0.868
流形排序	0.433	0.753	0.823	0.900

表 3 给出了在获取相同敏感度的情况下,在按蚊基因组中使用流形排序算法的预测结果及与其他已有方法的比较。可以看出,当 $N=1$ 时,特异性达到了 37.4%;其他情况下略好于文献[4~9]中的方法,当 $N=50$ 时,特异性达到了 96.2%。

表 3　MicroRNA 预测方法在按蚊基因中的实验结果比较

方法	$N=1$	$N=10$	$N=20$	$N=50$
3SVM	0.172	0.551	0.642	0.781
MiRFinder	0.189	0.513	0.689	0.795
DSVM	0.192	0.573	0.652	0.824
随机森林	0.235	0.604	0.728	0.831
ProMiR	0.212	0.592	0.752	0.815
MiRank	0.324	0.741	0.853	0.938
流形排序	0.374	0.773	0.871	0.962

3　结束语

使用流形数据排序算法预测 microRNA 是一种新的尝试。用流形结构描述数据有效避免了空间距离因素对节点相似性分配的影响。结合三联体结构—序列特征提取序列特征值,利用流形排序方法改善了检索排序的结果。在保持数据分布局部一致性的基础上,充分考虑了数据分布的全局一致性,成功实现了从候选序列中预测 MicroRNA。该算法有以下优点:a)不需要序列的注解,可以从大量没有详细注解的序列中预测 miRNA;b)对序列的保守性没有要求,可从不同物种的序列中预测 miRNA;c)对标记样本的数量没有要求,可从已知 miRNA 较少的物种序列中预测 miRNA。该算法新颖简洁、意义明确、运算速度快、容易实现,不像 SVM 方法那样模型复杂,而且计算量大,需要专门的软件。实验结果证实了该算法的有效性。实验预测出的大量候选 miRNA 可为进一步 miRNA 的研究提供指导。

参考文献:

[1] CAI Yi-mei, YU Xiao-min, HU Song-nian, *et al*. A brief review on the mechanisms of miRNA regulation[J]. Genomics, Proteomics & Bioinformatics,2009,7(4):147-154.

[2] 侯妍妍,应晓敏,李伍举. MicroRNA 计算发现方法的研究进展[J]. 遗传,2008,30(6):687-696.

[3] 郑凌伶,屈良鹤. 计算 RNA 组学:非编码 RNA 结构识别与功能预测[J]. 中国科学,2010,40(4):294-310.

[4] XUE Ceng-hai, LI Fei, HE Tao, *et al*. Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine[J]. BMC Bioinformatics,2005,5(6):310-317.

[5] HUANG Ting-hua, FAN Bin, MAX F, *et al*. MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans[J]. BMC Bioinformatics,2007,7(8):341-349.

[6] HELVIK S, SNOVE O, SAETROM P. Reliable prediction of Drosha processing sites improves microRNA gene prediction[J]. Bioinformatics,2007,23(2):142-149.

[7] JIANG Peng, WU Hao-nan, WANG Wen-kai, *et al*. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features[J]. Nucleic Acids Research,2007,35(4): 339-344.

[8] NAM J, SHIN K, HAN Jin-ju, *et al*. Human microRNA prediction through a probabilistic co-learning model of sequence and structure [J]. Nucleic Acids Research,2005,33(7):3570-3581.

[9] XU Yun-peng, ZHOU Xue-feng, ZHANG Wei-xiong. miRNA prediction with a novel ranking algorithm based on random walks[J]. Bioinformatics, 2008,24(13):50-58.

[10] ZHOU Deng-yong, JASON W, GRETTON A, *et al*. Ranking on data manifold[C]//THRUN S, SAUL L, SCHLKOPF B, *et al*. Advances in Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology,2004:169-176.

[11] ZHOU Deng-yong, BOUSQUENT O, LAL T, *et al*. Learning with local and global consistency [C]//THRUN S, SAUL L, SCHLKOPF B, *et al*. Advances in Neural Information Processing Systems. Cambridge: Massachusetts Institute of Technology,2004:321-328.

[12] DROR G, SOREK R, SHAMIR R. Accurate identification of alternatively spliced exons using support vector machine[J]. Bioinformatics,2005,21(7):897-901.