ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 1,Jan. 2015,pp. 155－163
DOI:10.16337/j.1004-9037.2015.01.015

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于 LDA 主题模型的功能性 miRNA-mRNA 调控模块识别

张俊鹏[1]　贺建峰[2]

(1.大理学院工程学院,大理,671003;2.昆明理工大学信息工程与自动化学院,昆明,650500)

摘　要:借助 mRNAs 分析 MicroRNAs(miRNAs)的研究已经用于阐述 miRNAs 调控机理,但是它们大部分的准确功能仍然处于未知状态。基于此,本文提出了一种基于 LDA(Latent Dirichlet allocation)主题模型来识别特定生物条件下 miRNAs 和靶标 mRNAs 之间的调控模块。该模型首先利用 Welch's t-检验挖掘具有差异表达的 miRNAs 和 mRNAs,然后采用折叠 Gibbs 抽样法进行参数估计。在上皮细胞-间充质细胞转型(Epithelial to Mesenchymal transition,EMT)数据集中的结果表明,所识别出的功能性 miRNA-mRNA 调控模块(FMRMs)能够构造不同生物条件下 miRNAs 与 mRNAs 之间的调控关系,从而为了解 EMT 生物过程和 miRNA 靶标治疗提供新的视角。与基于 K-means 聚类算法比较,LDA 主题模型比 K-means 聚类在挖掘 FMRMs 上更加有效。

关键词:LDA 主题模型;miRNA;miRNA-mRNA 调控模块;上皮细胞-间充质细胞转型;K-means 聚类
中图分类号:TP391　　**文献标志码**:A

Identifying of Functional miRNA-mRNA Regulatory Modules Based on LDA Topic Model

Zhang Junpeng[1], He Jianfeng[2]

(1. Faculty of Engineering, Dali University, Dali, 671003, China; 2. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China)

Abstract: Although much work has been done to elucidate the regulatory mechanism of miRNAs by associating miRNAs with mRNAs, their precise functions are still largely unknown. Latent dirichlet allocation (LDA) topic model is thus proposed to infer regulatory modules of miRNAs and their targets mRNAs for specific biological conditions. The proposed model firstly uses Welch's *t*-test to mine differentially expressed miRNAs and mRNAs, and then a collapsed Gibbs sampling method is utilized to estimate parameters. The results on epithelial to mesenchymal transition (EMT) data sets show that the inferred functional miRNA-mRNA regulatory modules (FMRMs) can construct regulatory relationships between miRNAs and mRNAs in different biological conditions, and give new insights into EMT biological process and miRNA targets therapy. Compared with K-means clustering algorithm, LDA topic model is more efficient in mining FMRMs.

Key words: LDA topic model; miRNA; miRNA-mRNA regulatory modules; epithelial to mesenchymal transition; K-means clustering

---

基金项目:大理学院青年教师科学研究基金(KYQN201203)资助项目;云南省应用基础研究计划青年(2013FD038)资助项目;国家自然科学基金(20120033)资助项目。
收稿日期:2013-06-06;**修订日期**:2013-09-06

引 言

MicroRNAs(miRNAs)是一类最新发现能够调控基因表达的非编码 RNA,其内源长大约为 21～23 个核苷酸。miRNAs 通过直接剪切靶基因 mRNAs 或者抑制翻译来调控基因表达的水平。一系列研究表明,它们在生长周期、细胞分化、细胞增殖、细胞生长、细胞迁移、细胞凋亡、应激反应等生物过程中发挥重要作用[1-3]。由于 miRNAs 研究的实验方法耗时、耗材和效率低,所以计算机方法成为 miRNAs 研究的替代方法。大部分与 miRNAs 研究相关的计算机方法可以分为三类:miRNAs 基因识别、miRNAs 靶基因预测和 miRNA-mRNA 调控模块的预测。一个公开的问题:miRNAs 如何调控靶基因 mRNAs 来与不同复杂细胞系统相关联。为了理解 miRNAs 在复杂细胞系统中的调控机理,研究识别 miRNAs 与靶基因 mRNAs 之间复杂关系的调控模块变得非常重要。另外,miRNAs 与靶基因 mRNAs 之间复杂关系的调控模块对于深入了解疾病发病机理和基因治疗具有重要作用。

近几年,许多相关学者提出了系列计算机方法用于识别 miRNAs 与靶基因 mRNAs 之间的调控网络。Yoon 和 De Micheli[4] 提出了一种基于序列水平的 miRNAs 调控模块(MRMs)计算方法。他们的方法是基于一种观察:当多个结合位点存在于一个靶标时,miRNAs 和靶基因 mRNAs 的结合强度差不多是相似的,而仅仅依靠基因序列信息进行预测将会导致高的错误发现率,更加精确的预测还要包括诸如基因表达谱的信息。Huang 等[5] 采用贝叶斯网络的方法将序列信息与 miRNAs 和 mRNAs 表达谱结合来识别 miRNAs 与 mRNAs 的调控模块,这种结合方法降低了错误识别率并且通过干扰实验证实了 miRNAs 的靶基因。同时,他们发现多个不同组织的 miRNAs 靶基因具有相对较低的错误发现率。Liu 等[6,7] 采用贝叶斯网络方法对 miRNA-mRNA 功能调控模块进行识别,所采用的数据信息包括 miRNA-mRNA 碱基互补配对信息、miRNAs 和 mRNA 表达数据。他们发现,在肿瘤组织模块中 miRNAs 表达下调而 mRNAs 表达上调;相反,在正常组织模块中 miRNAs 表达上调而 mRNAs 表达下调。Joung 等[8] 采用双聚类方法来发现 MRMs。他们发现:miRNA-mRNA 结合强度不仅与 miRNAs-mRNA 碱基互补配对信息有关,而且还与 miRNAs 和 mRNA 表达信息相关,并且证实了利用包括基因序列信息、miRNAs 和 mRNA 表达信息的异构数据源能够达到较好的预测效果。一种基于规则推理法的 MRMs 也被提出[9],这种方法也是利用多种异构数据源:miRNA-mRNA 碱基互补配对信息、miRNAs 和 mRNA 表达信息。他们所发现的 miRNA-mRNA 调控对具有较高的置信度,并且 miRNAs 与 mRNAs 的表达模型高度相关。Lu 等[10] 提出了一种基于 Lasso 回归模型的 miRNA-mRNA 调控网络识别方法,该方法使用 miRNA-mRNA 配对信息、miRNA 协同调控信息、RISC 数据以及 miRNAs 和 mRNAs 表达数据。该方法具有很高的敏感性和特异性,并且从前列腺癌数据中挖掘出许多以 miRNA 为中心的显著性调控模块。目前,Le 等人[11] 提出了一种基于 IDA 因果模型的 miRNA-mRNA 因果调控网络识别方法,该方法只利用了 EMT 数据集中的 miRNA 和 mRNA 表达谱数据。然而,仅仅依靠基因表达谱数据进行预测将会导致高的错误发现率,更加精确的预测还要包括其他信息,诸如靶基因结合信息。

本文所提出的模型是一种半监督模型,该模型首先利用样本信息进行基因差异表达分析,筛选出差异表达的 miRNAs 和 mRNAs,然后基于 LDA[12] 主题模型识别出功能性 miRNA-mRNA 调控模块(FMRMs)。LDA 主题模型已经应用于从多种文本数据源中挖掘隐含的主题,它能够从大量文档和单词异构信息中挖掘一系列主题以及与每个主题最相关的单词。类似地,假设主题对应 mRMs(不同生物条件下的 mRNAs 簇)或 miRMs(不同生物条件下的 miRNAs 簇),LDA 主题模型也能够从多种异构数据源中挖掘与 FMRMs 最相关的 miRNAs 和 mRNAs。使用的多种异构数据源包括 miRNA-mRNA 配对信息、miRNAs 和 mRNAs 基因表达数据。在上皮细胞-间充质细胞转型(Epitheliat-mesenchymal transition,EMT)数据集中的结果表明:提出的模型能够有效地识别特定生物条件下 miRNAs 与 mRNAs 之间的调控关系。与基于 K-means 聚类方法[13] 比较分析得到:基于 LDA 主题模型比基于 K-

means 聚类能够挖掘更多与 EMT 相关的 FMRMs 和与癌症或肿瘤显著相关的靶基因。

1 相关模型与 FMRMs 挖掘

1.1 基于 LDA 主题模型的 mRMs 识别

LDA 主题模型主要用于寻找 mRNAs 与 mRMs(不同生物条件下的 mRNAs 簇)相关以及生物条件与 mRMs 相关。如图 1 所示,在 S 个样本中有 T 个 mRNAs。随机变量 $m_{s,t}$ 代表第 t 个 mRNA 在第 s 个样本中的索引,其中 $s \in \{1,\cdots,S\}$,$t \in \{1,\cdots,T_s\}$。T_s 代表 mRNAs 在第 s 个样本中表达的总次数。随机变量 $z_{s,t}$ 代表第 s 个样本中隐藏的 mRMs,$z_{s,t}$,$m_{s,t}$ 与参数 $\theta_{s,k}$,$\varphi_{k,n}$ 都呈多项式分布。相应地,每个参数与超参数 α 和 β 有一个 Dirichlet 先验分布。参数 $\Theta=\{\theta_{s,k}\}$ 表示样本 s 属于模块 k 的概率,$\Phi=\{\varphi_{k,n}\}$ 表示第 n 个 mRNA 在模块 $Z=\{z_{s,t}\}$ 中表达的概率。因此 mRMs 可以通过估计参数 Θ 和 Φ 来识别。

由于模型中参数的准确估计很困难,因此需要采用近似估计算法来对模型参数进行估计。最大期望(Expectation-maximization,EM)算法通过寻找参数后验概率分布来估计隐含变量的参数,它是一种估计参数的标准方法。但是,该算法容易受到局部最大和计算效率低的影响。因此,本文采用一种折叠 Gibbs 抽样法来估计这些参数,该方法利用蒙特卡洛算法[14]对参数的后验概率进行抽样。折叠 Gibbs 抽样法[15]通过对隐含变量 $z_{s,t}$ 和 $m_{s,t}$ 联合抽样,对每个状态样本中的 $\theta_{s,k}$ 和 $\varphi_{k,n}$ 边缘化。对于一个 mRNA 变量 t 的第 s 个样本,$z_{s,t}$ 和 $m_{s,t}$ 的抽样可以表示为一个条件概率

$$P(z_{s,t}=k \mid m_{s,t}=n, Z_{\neg(s,t)}, s) \propto \frac{C_{kn}^{KT}+\beta}{\sum_{n'=1}^{T} C_{kn'}^{KT}+T\beta} \times \frac{C_{sk}^{SK}+\alpha}{\sum_{s'=1}^{S} C_{s'k}^{SK}+S\alpha} \tag{1}$$

式中:$Z_{\neg(s,t)}$ 是除了第 s 个样本类型外其 mRMs 的分配状态。C_{sk}^{SK} 是排除现在状态外,样本 s 分配给第 k 个 mRMs 的次数。C_{kn}^{KT} 是排除现在状态外,表达类型 n 分配给第 k 个 mRMs 的次数。K 是 mRMs 的隐含变量个数,S 是样本类型的总个数。

经过抽样迭代后,参数 $\theta_{s,k}$ 和 $\varphi_{k,n}$ 的估算如下

$$\hat{\theta}_{s,k}=\frac{C'^{SK}_{sk}+\alpha}{\sum_{s'=1}^{S} C'^{SK}_{s'k}+S\alpha},\ \hat{\varphi}_{k,n}=\frac{C'^{KT}_{kn}+\beta}{\sum_{n'=1}^{T} C'^{KT}_{kn'}+T\beta} \tag{2}$$

其中 C'^{SK}_{sk} 和 C'^{KT}_{kn} 来自除了现在状态外由所有数据分配结果计算出来,它与式(1)的抽样过程不同。

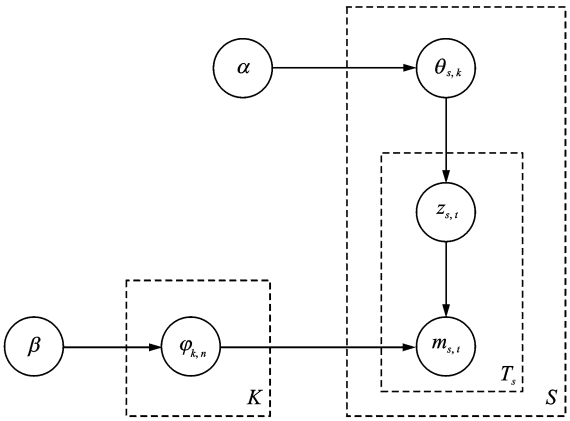


图 1 LDA 主题模型示意图(阴影和非阴影的变量分别代表已知和隐性变量)

Fig. 1 LDA topic model (shaded and unshaded variables indicate observed and latent variables)

1.2 miRNA-mRNA 调控模块与生物条件之间的统计模型

LDA 主题模型中估算的参数在许多层次上为 EMT 数据集提供新的视角，其中 Θ 将样本聚集到模块中，这些模块与特定生物条件相关。构造一个统计模型来连接调控模块和生物条件之间的关系。令数据集的样本数为 S，C 是数据集的生物条件个数，K 是模块个数，c_i 是属于生物条件 i 的样本个数，其中 $\sum_{i=1}^{C} c_i = S$。对于每一个模块，假设前 n 个最有可能属于相同生物条件 i 的样本数是 x，那么随机变量 x 与参数 S，c_i 和 n 服从一个超几何分布，其表达式如下

$$p(x) \sim \text{hypergeometric}(x; S, c_i, n) \tag{3}$$

当 x 处于显著水平（例如 p 值<0.05）时，生物条件 i 分配给模块 k。

1.3 FMRMs 挖掘

LDA 主题模型和统计模型共同挖掘 miRNAs 簇和 mRNAs 簇与生物条件之间的显著性关系，也就是识别 FMRMs，图 2 阐述了基于 LDA 主题模型的 FMRMs 识别流程图。本文中，一个主题在 LDA 主题模型中被定义为与文档最相关的单词。主题对应 mRMs（不同生物条件下的 mRNAs 簇）或 miRMs

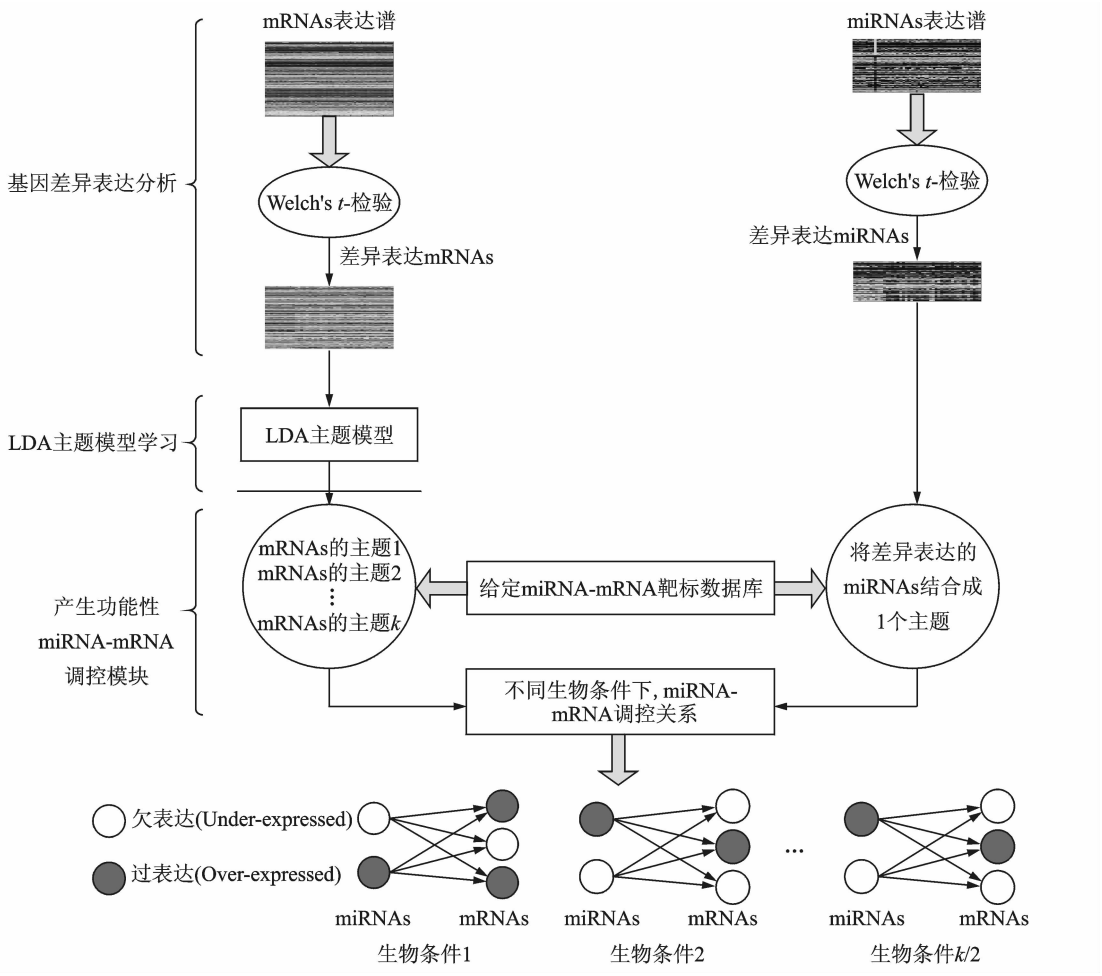


图 2 识别功能性 miRNA-mRNA 调控模块的 LDA 主题模型设计流程图

Fig. 2 Flowchart of LDA topic model for inferring functional miRNA-mRNA regulatory modules

(不同生物条件下的 miRNAs 簇),文档对应样本,单词对应 mRNAs 或 miRNAs。

如图 2 所示,miRNAs 和 mRNAs 表达谱经过归一化处理后,通过参数检验方法 Welch's t-检验[16](显著性 p 值<0.05)得到一系列差异表达的 miRNAs 和 mRNAs。LDA 主题模型用来产生 k 个 mRMs,同时样本属于 mRMs 的概率也被提取出来,差异表达的 miRNAs 合并成 1 个 miRMs。为了控制假阳性率,在产生 FMRMs 的过程中结合了 miRNA-mRNA 靶标数据库信息。因此,功能性 miRNA-mRNA 调控模块在 miRNAs 和 mRNAs 之间的碱基互补配对信息的约束条件下产生。

2 实验结果与分析

2.1 实现过程

miRBase[17]作为 miRNA-mRNA 关系数据库来去除不合理的 miRNA-mRNA 调控关系,因为它能够提供比实验支持数据库更多的靶标预测信息。模型的输入数据源为 miRNAs 和 mRNAs 表达谱,miRNAs 表达谱可以从文献[18]获取,该数据源来自美国国家癌症研究所(National Cancer Institute,NCI60),NCI60 中的 mRNAs 表达谱可以从 CellMiner[19]中获取。使用的样本被分为三类:上皮细胞(Epithelial,11 个样本)、间充质细胞(Mesenchymal,37 个样本,其中有一个样本不可用)和其他细胞(Non-normal,11 个样本)。

给定一个 237×59 的 miRNAs 表达值矩阵和一个21 225×59 的 mRNAs 表达值矩阵,通过 Welch's t-检验(显著水平 p 值<0.05)基因差异表达分析后,LDA 主题模型的输入数据包含一个 10×59 的 miRNAs 表达值矩阵和一个3 660×59 的 mRNAs 表达值矩阵。在实验中,参数 S 设为 118,这一参数由过表达样本个数(59 个)和欠表达样本个数(59 个)所确定。mRMs 的个数设定为 6,这一参数由 EMT 数据源的细胞类型个数所确定,并且每个 mRM 包含有 40 个(大约是差异表达 mRNAs 个数的 1%)与该 mRM 最相关的 mRNAs。miRMs 的个数设为 1,这一数值由差异表达的 miRNAs 个数所确定。功能性 miRNA-mRNA 调控模块(FMRMs)是通过给定的靶标关系数据来挖掘 miRMs 和 mRMs 之间的关系。超参数 α 和 β 值分别为 0.16 和 0.01,Gibbs 抽样迭代次数设为5 000,这些参数设置都是基于经验实验得来。

2.2 与 EMT 相关的 FMRMs 挖掘

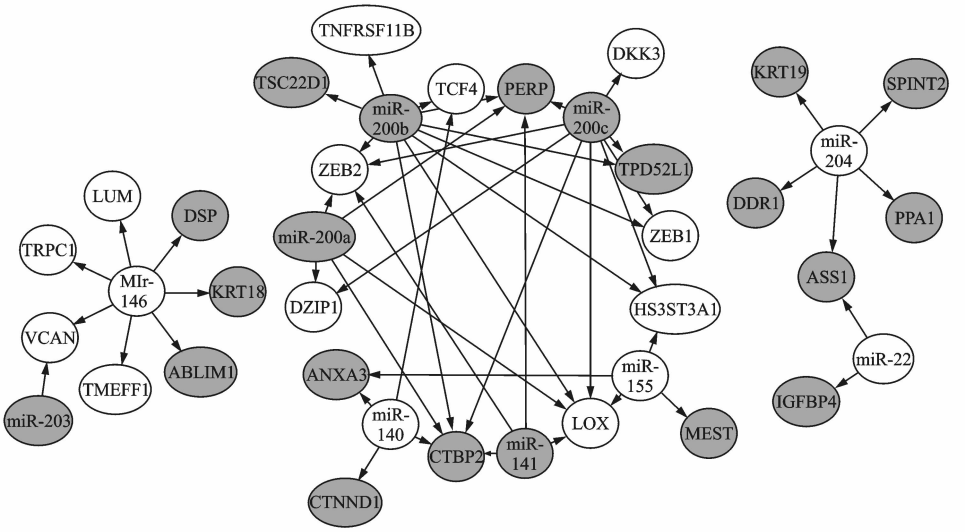
EMT 定义为在细胞行为中产生一个非常复杂变化的总称,它涉及许多基因的差异表达以及许多在细胞内与细胞外的分子功能变化。它产生的结果是:处于连续层的细胞(Epithelial cells,上皮细胞)转变成更加独立和运动状态的细胞(Mesenchymal cells,间充质细胞)。

为了识别 EMT 数据集中特定细胞条件的 FMRMs,miRBase 数据库用来建立 miRMs 中 miRNAs 与 mRMs 中 mRNAs 相互之间的关系。如图 3 所示,特定细胞条件下的 FMRMs 用有向二分图结构表示,其中,灰色节点分别代表 miRNAs 和 mRNAs 过表达,白色节点分别代表 miRNAs 和 mRNAs 欠表达。因此,该网络图不仅能够知道 miRNAs 与 mRNAs 之间的调控关系,而且能够了解不同细胞条件下 miRNAs 和 mRNAs 之间的表达类型。

目前研究表明:miR-200 家族成员在 EMT 生物过程中扮演重要角色,因此通过分析 miR-200 家族成员与靶标 mRNAs 之间的调控关系可以验证模型的有效性。如图 3 所示,结果显示:在 FMRMs 中 ZEB1 和 FN1 是 miR-200a,miR-200c 和 miR-141 共同的靶标。miRNAs 和 mRNAs 的表达类型表明:miR-200 家族成员负调控靶标 ZEB1 和 ZEB2。文献[20-23]已经证明:miR-200 家族成员靠直接靶标 ZEB1 和 ZEB2 来抑制转移和 EMT 过程的初始阶段。miR-200 家族成员的表达和抑制直接决定 ZEB1 和 ZEB2 的上调控和下调控,实验结果与验证结果一致。LOX 在 FMRMs 中被所有的 miR-200 家族成

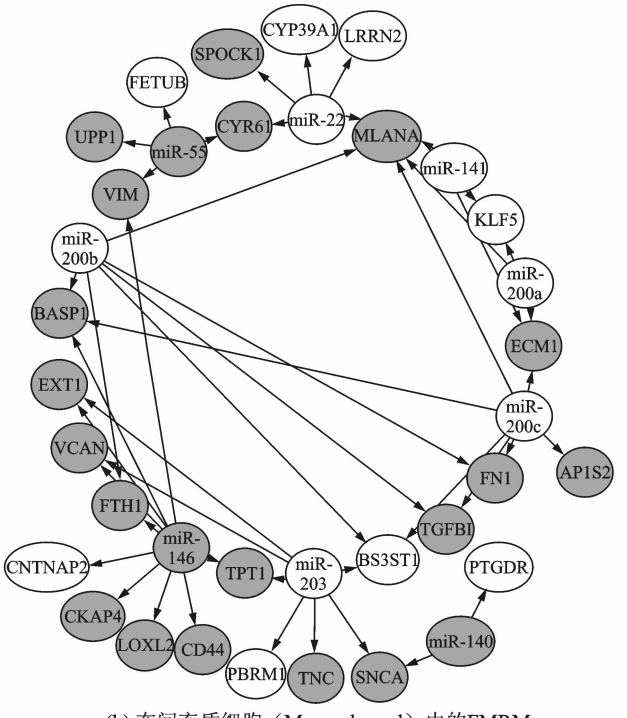
员负调控,这一结果与已知的发现(LOX 调控 EMT[24])一致。这一结果同时可以推断:在 EMT 中,LOX 与 miR-200 家族成员具有广泛的关联。

图 3(a)和(b)同时也显示 miR-200(*a*, *b*, *c*)和 miR-141 在上皮细胞类型中过表达,而在间充质细胞类型中欠表达。这一结果表明 miRNAs 的表达类型与上皮细胞-间充质细胞转型密切相关。



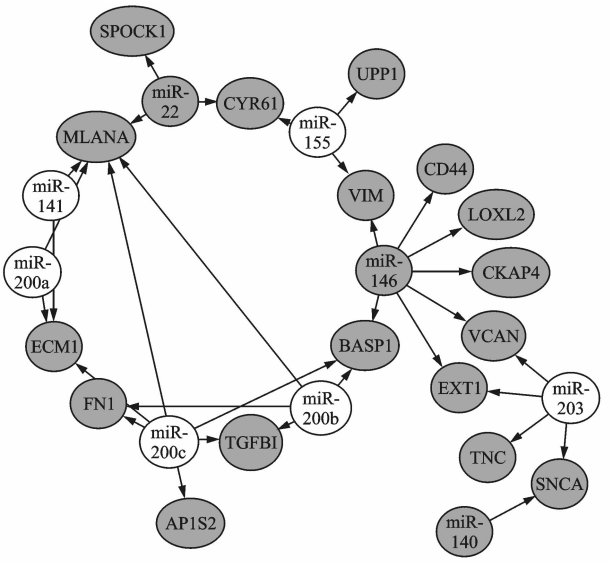
(a) 在上皮细胞（Epithelial）内的$FMRM_1$

(a) $FMRM_1$ in Epithelial cells



(b) 在间充质细胞（Mesenchymal）内的$FMRM_2$

(b) $FMRM_2$ in Mesenchymal cells



(c) 在其他细胞（Non-normal）内的$FMRM_3$

(c) $FMRM_3$ in Non-normal cells

图 3 在上皮细胞(Epithelial)、间充质细胞(Mesenchymal)和其他细胞(Non-normal)内的 FMRMs

Fig. 3 FMRMs in Epithelial, Mesenchymal, and Non－normal cells

2.3 结果比较

LDA 主题模型是一种基于狄利克雷分布的产生式概率模型,每个主题相当于一个聚类。而在聚类算法中,K-means 算法是一种经典聚类算法,在处理大数据集时,该算法具有相对可伸缩性和高效率的特点。与基于 LDA 主题模型的方法类似,miRNAs 和 mRNAs 表达谱首先经过归一化处理,然后通过基因差异表达分析得到一系列差异表达的 miRNAs 和 mRNAs。在本文中,一个聚类在 K-means 聚类中定义为与该聚类最相关的观察变量的集合。观察变量对应 mRNAs 或 miRNAs,聚类对应 mRMs(不同生物条件下的 mRNAs 簇)或 miRMs(不同生物条件下的 miRNAs 簇)。mRMs 的个数设置为 6,并且每个 mRM 包含有 40 个(大约是差异表达 mRNAs 个数的 1%)与该 mRM 最相关的 mRNAs,miRMs 的个数设置为 1。

2.3.1 功能性 miRNA-mRNA 调控对数比较

基于 LDA 主题模型和 K-means 聚类都能够产生与 EMT 细胞类型相关的 miRNA-mRNA 调控对。如表 1 所示,基于 LDA 主题模型在上皮细胞和间充质细胞中生成的 miRNA-mRNA 调控对数(分别是 51 和 48 对)比基于 K-means 聚类(分别是 40 和 36 对)要多,然而在其他细胞中,基于 K-means 聚类挖掘的 miRNA-mRNA 调控对数(38 对)要比基于 LDA 主题模型(32 对)多 6 对。这一结果表明:LDA 主题模型更适合挖掘特异性细胞类型(上皮细胞和间充质细胞)中的 miRNA-mRNA 调控关系,而 K-means 更适合挖掘由多种细胞类型(其他细胞)中的 miRNA-mRNA 调控关系。总体来说,基于 LDA 主题模型(131 对)比基于 K-means 聚类(114 对)在挖掘与 EMT 相关的 FMRMs 上更加有效,详细的 miRNA-mRNA 调控对数比较情况见表 1。

表 1 与 EMT 细胞类型相关的 miRNA-mRNA 调控对数

Table 1 Number of miRNA-mRNA pairs associated with EMT cells

模型	上皮细胞	间充质细胞	其他细胞	总数
LDA	51	48	32	131
K-means	40	36	38	114

2.3.2 靶基因 mRNAs 的显著性功能比较

IPA(Ingenuity pathway analysis, http://www.ingenuity.com/)软件能够有效地分析所识别的靶基因是否与特定生物功能显著性相关。为了挖掘与靶基因 mRNAs 显著相关的生物功能,IPA 软件的显著性 p 值设为 0.05。所挖掘的显著性生物功能中,细胞运动的子类别:细胞迁移、细胞入侵和细胞扩散作为 EMT 的体外功能指标,对研究 EMT 功能具有重要作用。如表 2 所示,基于 LDA 主题模型所挖掘的 EMT 体外功能指标对应的靶基因 mRNAs 个数(28 个,其中细胞迁移 18 个、细胞入侵 8 个和细胞扩散 2 个)比基于 K-means 聚类(18 个,其中细胞迁移 12 个、细胞入侵 6 个和细胞扩散 0 个)要多 10 个。与此同时,基于 LDA 主题模型所挖掘的与癌症或肿瘤显著相关的靶基因 mRNAs(39 个)比基于 K-means 聚类(20 个)大约多 1 倍,并且其显著性水平(6.46E−09～1.16E−02)比基于 K-means 聚类(1.28E−03～3.68E−02)要高。靶基因 mRNAs 的显著性功能比较的详细情况见表 2。

表 2 与 EMT 体外功能指标和癌症或肿瘤显著相关的靶基因 mRNAs 个数

Table 2 Number of target mRNAs significantly associated with the vitro functional markers of EMT and cancer or tumor

模型	细胞迁移	细胞入侵	细胞扩散	癌症或肿瘤
LDA	18	8	2	39
K-means	12	6	0	20

3 结束语

本文基于 LDA 主题模型并且结合基因差异表达分析方法(Welch's t-检验)来对 EMT 数据集进行功能性 miRNA-mRNA 调控模块识别,该调控模块由 miRNAs,mRNAs 以及它们在每个生物条件下的表达类型组成。所提出的模型能够建立生物关系链"miRNA → mRNA → 生物条件",其中 miRNA 是调控子。实验结果表明:在挖掘 FMRMs 上,基于 LDA 主题模型比基于 K-means 聚类方法更加有效。LDA 主题模型的实现软件平台是 Matlab 7.8.0,在 CPU 主频为 2.33 GHz 的计算机平台上,Gibbs 抽样迭代5 000次所花费的时间约为 50 min。因此,为了能够使 LDA 主题模型能够适用于更大的数据集,将来的研究还需要设计一种并行计算方法来提高 LDA 主题模型的效率。

参考文献:

[1] Ambros V. The functions of animal microRNAs [J]. Nature, 2004, 431(7006):350-355.

[2] Du T, Zamore P D. Beginning to understand microRNA function [J]. Cell Research, 2007,17(8):661-663.

[3] Bushati N, Cohen S M. MicroRNA functions [J]. Annu Rev Cell Dev Biol, 2007, 23:175-205.

[4] Yoon S, De Micheli G. Prediction of regulatory modules comprising microRNAs and target genes [J]. Bioinformatics, 2005, 21(2):93-100.

[5] Huang J C, Morris Q D, Frey B J. Detecting MicroRNA targets by linking sequence, MicroRNA and gene expression data [J]. Research in Computational Molecular Biology, 2006, 3909:114-129.

[6] Liu B, Li J, Tsykin A. Discovery of functional miRNA-mRNA regulatory modules with computational methods [J]. Journal

of Biomedical Informatics, 2009, 42(4): 685-691.
[7] Liu B, Li J, Tsykin A, et al. Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy [J]. BMC Bioinformatics, 2009, 10:408.
[8] Joung J G, Hwang K B, Nam J W, et al. Discovery of microRNA-mRNA modules via population-based probabilistic learning [J]. Bioinformatics, 2007,23(9): 1141-1147.
[9] Tran D H, Satou K, Ho T B. Finding MicroRNA regulatory modules in human genome using rule induction [J]. BMC Bioinformatics, 2008, 9((S12):S5.
[10] Lu Y, Zhou Y, Qu W, et al. A Lasso regression model for the construction of microRNA-target regulatory networks [J]. Bioinformatics, 2011, 27(17):2406-2413.
[11] Le T D, Liu L, Tsykin A, et al. Inferring microRNA-mRNA causal regulatory relationships from expression data [J]. Bioinformatics, 2013,29(6):765-771.
[12] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(4/5):993-1022.
[13] MacQueen J. Some methods for classification and analysis of multivariate observations [C] // Proc Fifth Berkeley Symp on Math Statist and Prob. Calif:Univ of Calif Press, 1967:281-297.
[14] 王尔申,蔡明,庞涛. MCMC 粒子滤波的 GPS 定位数据处理算法[J]. 数据采集与处理,2013, 28(2):213-218.
Wang Ershen, Cai Ming, Pang Tao. GPS positioning data processing algorithm based on MCMC particle filter [J]. Journal of Data Acquisition and Processing, 2013, 28(2):213-218.
[15] Steyvers M, Griffiths T. Probabilistic topic models [R]. Lawrence Erlbaum Associates, Inc, 2007.
[16] Welch B L. The generalization of student's problem when several different population variances are involved [J]. Biometrika, 1947, 34(1-2):28-35.
[17] Griffiths-Jones S, Saini H K, van Dongen S, et al. miRBase:Tools for microRNA genomics [J]. Nucleic Acids Res, 2008, 36:D154-158.
[18] Park S M, Gaur A B, Lengyel E, et al. The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2 [J]. Genes & Dev, 2008, 22(7):894-907.
[19] Shankavaram U T, Varma S, Kane D, et al. CellMiner:A relational database and query tool for the NCI-60 cancer cell lines [J]. BMC Genomics, 2009, 10: 277.
[20] Korpal M, Lee E S, Hu G, et al. The miR-200 family inhibits epithelial-mesenchymal transition and cancer cell migration by direct targeting of E-cadherin transcriptional repressors ZEB1 and ZEB2 [J]. The Journal of Biological Chemistry, 2008, 283(22):14910-14914.
[21] Gregory P A, Bracken C P, Bert A G, et al. MicroRNAs as regulators of epithelial-mesenchymal transition [J]. Cell Cycle, 2008, 7(20):3112-3118.
[22] Burk U, Schubert J, Wellner U, et al. A reciprocal repression between ZEB1 and members of the miR-200 family promotes EMT and invasion in cancer cells [J]. EMBO Rep, 2008, 9(6):582-589.
[23] Gregory P A, Bert A G, Paterson E L, et al. The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1 [J]. Nature Cell Biology, 2008, 10(5):593-601.
[24] Higgins D F, Kimura K, Bernhardt W M, et al. Hypoxia promotes fibrogenesis in vivo via HIF-1 stimulation of epithelial-tomesenchymal transition [J]. Journal of Clinical Investigation, 2007, 117(12):3810-3820.

作者简介:**张俊鹏**(1987-),男,助教,研究方向:生物信息学,E-mail:zhangjunpeng_411@yahoo.com;**贺建峰**(1965-),男,教授,研究方向:生物医学图像与生物信息学,E-mail:jfenghe@kmust.edu.cn。