中国科学: 信息科学 2021 年 第 51 卷 第 6 期: 927–939
SCIENTIA SINICA Informationis
论文

# 基于多数据融合的 circRNA– 疾病关联关系预测

雷秀娟*, 张文祥, 刘恋

陕西师范大学计算机科学学院, 西安 710000
* 通信作者. E-mail: xjlei@snnu.edu.cn

收稿日期: 2019–07–04; 修回日期: 2019–10–21; 接受日期: 2020–01–05; 网络出版日期: 2021–05–13

国家自然科学基金 (批准号: 61672334, 61972451, 61902230) 和中央高校基本科研业务费专项资金 (批准号: 201901010) 资助项目

**摘要** 环状 RNA (circluar RNA, circRNA) 在基因表达、剪切和转录的过程中扮演着重要角色. 越来越多的证据表明, circRNA 与疾病的产生与发展存在着重要的联系. 本文提出了一种基于多数据融合的非负矩阵分解算法 (EDNMF) 预测 circRNA– 疾病关联关系. 该方法首先对 circRNA– 疾病关联关系进行预处理, 解决了 circRNA– 疾病关联关系过少对算法产生的负面影响的问题. 然后, EDNMF 算法将 circRNA 表达谱和癌症相似性数据转化为约束条件, 基于预处理后的 circRNA– 疾病关联关系采用改进的非负矩阵分解算法得到最终的打分值, 从而预测 circRNA– 疾病关联关系. 五折和十折交叉验证结果表明, EDNMF 算法相比其他算法能更有效地预测 circRNA– 疾病关联关系. 此外, 采用 EDNMF 算法预测新的 circRNA– 结肠直肠癌关联关系打分排名前 10 的结果中, 大部分结果已经得到了佐证, 表明了该算法可以有效地预测未知的 circRNA– 疾病关联关系.

**关键词** circRNA, circRNA 表达谱, circRNA– 疾病关联关系, 非负矩阵分解, 疾病相似性

## 1 引言

环状 RNA (circluar RNA, circRNA) 是一种拥有独特性质和多种细胞功能的特殊内源性非编码 RNA (non-coding RNA, ncRNA) [1,2]. 1976 年, Sanger 等 [3] 在研究植物类病毒的过程中首次通过电子显微镜观察到了 circRNA. 此后, 研究人员在真核生物细胞质 [4]、酵母线粒体 [5] 中发现了 circRNA. 然而, 由于 circRNA 的结构特性、功能未知性和丰度低等原因, 很长一段时间它们被认为是转录噪声或错误拼接的产物 [6,7], 并没有受到过多的关注.

近年来, 随着高通量测序技术的快速发展, 在古生菌、植物和动物中检测到越来越多的 circRNA, 引起了学者的广泛关注 [1,8,9]. 通过对数据进行分析处理, 构建了大量与 circRNA 相关的数据库, 如 PlantcircBase 数据库 [9]、CircBase 数据库 [10]、CircR2Disease 数据库 [11]、CSCD 数据库 [12] 和 circAtlas

**引用格式:** 雷秀娟, 张文祥, 刘恋. 基于多数据融合的 circRNA– 疾病关联关系预测. 中国科学: 信息科学, 2021, 51: 927–939, doi: 10.1360/SSI-2019-0142
Lei X J, Zhang W X, Liu L. Prediction of circRNA-disease associations based on multiple biological data (in Chinese). Sci Sin Inform, 2021, 51: 927–939, doi: 10.1360/SSI-2019-0142

 

数据库 [13] 等等. 随着 circRNA 数据的增加, circRNA 的生物学功能也逐渐被揭示, 如充当 miRNA 分子海绵 [14]、参与转录调控 [2,15]、联结 RNA 结合蛋白 [16] 和发挥翻译功能 [17,18] 等. 此外, circRNA 失调将会导致细胞功能紊乱、表达异常和生长缺陷等.

近年来, 复杂疾病遗传模式复杂、关联基因或 RNA 数量较多, 已经严重威胁了人类的健康 [19~21]. 研究发现, 多种 circRNA 已经被确认与胃癌、结直肠癌、肝癌和神经胶质瘤等复杂疾病的产生与发展存在着重要的联系. circRNAAKT3 作为 miR-198 海绵促进 PIK3R1 表达, 以此增强胃癌患者对顺铂的耐药性 [22]. Li 等 [23] 已经证实 circVAPA 在结直肠癌病人组织和血浆中上调, 通过在结直肠癌中充当 miR-101 海绵发挥其致癌特性. Xu 等 [24] 研究发现 circSETD3 在肝癌组织和细胞中均显著下调, 它可以作为一种新型的肝癌抑制因子. Shi 等 [25] 发现 circ-0014359 在神经胶质瘤细胞中充当 miRNA-153 海绵, 并通过调控 miR-153/PI3K 信号通路影响胶质瘤进展.

以上通过实验来预测 circRNA– 疾病关联关系的方法需要耗费大量的财力与时间 [26]. 机器学习是当前一种有效的方法, 具有很好的预测功能, 可以为实验研究提供前期基础, 以节省大量的时间和财力. 因此, 如何通过机器学习方法来预测 circRNA– 疾病关联关系成为了亟待解决的问题. 早期由于缺少大量已知的 circRNA– 疾病关联关系, 通过机器学习方法预测 circRNA– 疾病关联关系并没有得到广泛的普及 [27]. Fan 等 [11] 在 2018 年通过文献检索的方式提取了 725 个 circRNA– 疾病关联关系, 其中包含 661 个 circRNA 和 100 个疾病, 构建了 CircR2Disease 数据库. CircRNADisease 数据库 [28] 也为计算方法分析 circRNA– 疾病之间的关联关系提供了有效的契机. 随着这些数据库的产生, 一些计算方法相继被提出. Lei 等 [29] 在 circRNA 和疾病组成的异构网络上采用深度优先搜索算法搜索特定 circRNA 到特定疾病的路径, 然后采用路径权重算法识别 circRNA– 疾病关联关系. Fan 等 [26] 通过 circRNA– 疾病关联关系、circRNA 相似性网络和疾病相似性网络构建了异构网络, 并基于构建的异构网络使用 KATZ 算法识别 circRNA– 疾病关联关系. Xiao 等 [30] 使用流行正则化学习框架在 circRNA 和疾病构成的异构网络上预测 circRNA– 疾病关联关系. Yan 等 [27] 基于 circRNA– 疾病关联关系分别计算 circRNA 之间以及疾病之间的相似性, 提出了一种基于克罗内克积 (Kronecker product) 的算法预测 circRNA– 疾病关联关系. Zhao 等 [31] 依据二部图投影和 KATZ 算法来识别 circRNA– 疾病关联关系. Wei 等 [32] 采用了一种改进的非负矩阵分解算法来识别 circRNA– 疾病关联关系.

本文基于 circRNA 表达谱数据和疾病相似性数据提出了 EDNMF 算法 (non-negative matrix factorization algorithm based on circRNA expression profiles data and disease similarity data) 预测 circRNA– 疾病关联关系. 为了解决 circRNA– 疾病关联关系数据量过少而导致的预测不准确的问题, 本文首先通过 circRNA 之间的相似性矩阵、疾病之间的相似性矩阵和原始 circRNA– 疾病关联关系矩阵, 对所有的 circRNA– 疾病关联关系进行预处理. 基于预处理后的 circRNA– 疾病关联矩阵, 使用改进的非负矩阵分解算法预测 circRNA– 疾病关联关系. 在非负矩阵分解过程中, 将 circRNA 表达谱数据和疾病相似性数据转化为约束条件, 使基矩阵偏向于 circRNA 表达谱数据; 系数矩阵的转置与其本身的乘积偏向于疾病相似性数据; 并且通过固定基矩阵和系数矩阵的维度, 解决了传统 NMF 算法中维度不确定的问题.

## 2 数据与方法

本节将详细介绍数据的收集、预处理及 EDNMF 算法.

### 2.1 数据集和预处理

#### 2.1.1 circRNA– 疾病关联关系数据

CircR2Disease 数据库 [11][1)]中收录了 725 个 circRNA– 疾病关联关系数据, 其中包含 661 个 circRNAs 和 100 个疾病. 为了保证数据的准确性, 仅保留存在 CircBase ID 和 gene symbol 的 circRNAs, 最终经过处理之后剩余 427 个 circRNA– 疾病关联关系、372 个 circRNAs 和 77 个疾病.

CircAtlas 2.0 数据库[13][2)]中收录了 1018 条 circRNA– 疾病关联关系数据. 本文采用与 CircR2Disease 数据库同样的处理方法提取 circRNA– 疾病关联关系, 最终处理后的数据包含了由 410 个 circRNAs 和 79 个疾病组成的 465 条 circRNA– 疾病关联关系.

#### 2.1.2 疾病名称标准化

由于同一个疾病可能对应多个疾病名称, 如 “clear cell renal cell carcinoma” 与 “clear cell kidney carcinoma” 本质上代表同一种疾病. 在 Disease ontology 数据库 [33][3)]中对疾病名称进行标准化, 使每一种疾病拥有唯一的标识号 “DOID”. 我们首先合并了 2.1.1 小节中在两个数据库中处理后的数据, 在删除合并的数据中没有 “DOID” 的疾病数据后, 获得了 553 条 circRNA– 疾病关联关系, 这些关联关系由 465 个 circRNAs 和 68 个包含 “DOID” 的疾病组成.

#### 2.1.3 circRNA 表达谱数据

本文从 exoRBase 数据库 [34][4)]中下载了 90 个样本中 58330 个 circRNAs 的表达数据. 将表达数据中的 exor_circ_ID 转换为 circBase_ID 之后, 我们发现 2.1.2 小节中 465 个 circRNAs 有 268 个 circRNAs 拥有在 90 个样本上的表达数据. 因此, 剔除不含 circRNA 表达数据的 circRNAs 后, 最终的数据包含由 268 个 circRNAs 和 62 个疾病组成的 340 条 circRNA– 疾病关联关系.

#### 2.1.4 预处理

本文用矩阵 CD 表示 circRNA– 疾病关联关系矩阵, 其中矩阵 CD 的行表示 circRNA, 列表示疾病. 如果 circRNA $i$ 与疾病 $j$ 存在关联关系, 则矩阵 $\mathrm{CD}(i,j)=1$, 反之为 0. 矩阵 CT 表示 circRNA 表达谱矩阵, 其中 $\mathrm{CT}(i,j)$ 表示 circRNA $i$ 在样本 $j$ 上的表达量.

经过数值统计可知, 矩阵 CT 中表达量取值范围是 $[0, 86451.31]$. 因此, 为了避免较大的值对算法产生影响, 本文对矩阵 CT 进行规范化:

$$T(i,j)=\frac{\mathrm{CT}(i,j)}{\sqrt{\sum_{k=1}^{s}\mathrm{CT}(i,k)\sum_{k=1}^{m}\mathrm{CT}(k,j)}}, \tag{1}$$

其中 $m$ 表示 circRNA 的数量, $s$ 表示在组织表达中样本的数量, $T$ 表示规范化后的 CT 矩阵.

本文采用两种方法来计算疾病之间的相似性. 第 1 种方法是基于高斯 (Gauss) 核相似性方法, 该方法假设相似的疾病有一定概率会与功能上相似甚至相同的 circRNA 存在关联. 具体计算方法如下:

$$D_G(d_i,d_j)=\exp(-\gamma_d\|\mathrm{CD}(\cdot,d_i)-\mathrm{CD}(\cdot,d_j)\|^2), \tag{2}$$

$$\gamma_d=\frac{1}{\frac{1}{n}\sum_{i=1}^{n}\|\mathrm{CD}(\cdot,d_i)\|}, \tag{3}$$

1) http://bioinfo.snnu.edu.cn/.
2) http://circatlas.biols.ac.cn/.
3) http://www.disease-ontology.org/.
4) http://www.exorbase.org/.

其中 $n$ 表示疾病的数量, $\mathrm{CD}(\cdot, d_i)$ 表示 CD 矩阵中第 $d_i$ 列, $D_G$ 表示疾病高斯核相似性矩阵. 此外, 本文通过 R 语言的 DOSE 包利用疾病对应的 DOID 来计算疾病之间的语义相似性 $D_L$ [35]. 采用如下方法来融合两种疾病相似性矩阵:

$$D(i,j) = \max(D_G(i,j), D_L(i,j)). \tag{4}$$

对于 circRNA 之间的相似性, 本文也采用两种相似性计算方法, 分别是高斯核相似性和皮尔森 (Pearson) 相关系数. circRNA 高斯核相似矩阵与求疾病高斯核相似性矩阵方法类似:

$$C_G(c_i, c_j) = \exp(-\gamma_c \|\mathrm{CD}(c_i,\cdot) - \mathrm{CD}(c_j,\cdot)\|^2), \tag{5}$$

$$\gamma_c = \frac{1}{\frac{1}{m}\sum_{i=1}^{m} \|\mathrm{CD}(c_i,\cdot)\|}, \tag{6}$$

其中 $m$ 表示 circRNA 的数量, $\mathrm{CD}(c_i,\cdot)$ 表示 CD 矩阵中第 $c_i$ 行, $C_G$ 表示 circRNA 高斯核相似性矩阵. 此外, 本文也利用 circRNA 表达谱矩阵 CT 通过皮尔森相关系数来计算 circRNA 表达相似性矩阵:

$$C_P(c_i, c_j) = \frac{\sum_{i=1}^{m}(\mathrm{CT}(c_i,i) - \mathrm{avg}(\mathrm{CT}(c_i,\cdot)))(\mathrm{CT}(c_j,i) - \mathrm{avg}(\mathrm{CT}(c_j,\cdot)))}{\sqrt{\sum_{p=1}^{m}(\mathrm{CT}(c_i,p) - \mathrm{avg}(\mathrm{CT}(c_i,\cdot)))^2 \sum_{q=1}^{m}(\mathrm{CT}(c_j,q) - \mathrm{avg}(\mathrm{CT}(c_j,\cdot)))^2}}, \tag{7}$$

其中 $m$ 表示表达谱数据中样本的数量, $\mathrm{CT}(c_i,\cdot)$ 表示 CT 矩阵中第 $c_i$ 行, $\mathrm{avg}(\mathrm{CT}(c_i,\cdot))$ 表示 $\mathrm{CT}(c_i,\cdot)$ 中元素的均值, $C_P$ 表示 circRNA 表达相似性矩阵. 最终, 本文采用了如下方法来融合两种 circRNA 相似性矩阵:

$$C(i,j) = \max(C_G(i,j), C_P(i,j)). \tag{8}$$

算法的表现性能很大程度依赖于大量的已知 circRNA– 疾病关联关系. 然而已知的 circRNA– 疾病关联关系很少, CD 矩阵中存在大量为 0 的元素, 其中必然包含一定的假阴性数据, 因此这必将对预测 circRNA– 疾病关联关系产生负面影响 [30, 32]. 为了减少这种负面影响, 本文通过 circRNA 相似性矩阵 $C$、疾病相似性矩阵 $D$ 和 circRNA– 疾病关联关系矩阵 CD 对 circRNA– 疾病关联关系进行了预处理, 如下:

$$Y_C = C \times \mathrm{CD}, \tag{9}$$

$$Y_D = D \times \mathrm{CD}', \tag{10}$$

$$Y = \max(Y_C, Y_D), \tag{11}$$

其中 $Y_C$ 和 $Y_D$ 分别表示从 circRNA 和疾病的角度对 circRNA– 疾病关联关系进行重定义, $Y$ 表示最终重定义的 circRNA– 疾病关联关系矩阵.

## 2.2 方法

### 2.2.1 两种基本的非负矩阵分解算法

非负矩阵分解 (nonnegative matrix factorization, NMF) 算法由 Lee 等 [36] 在 1999 年正式提出, 经过数十年的发展, 已经被广泛应用于各大领域, 如图像分析、数据挖掘及语音识别等. 其核心思想是将一个原始非负矩阵 $Y \in \mathbb{R}_+^{m\times n}$ 分解成两个非负矩阵相乘的形式, 描述如下:

$$Y \approx WH, \tag{12}$$

其中 $W \in \mathbb{R}_+^{m\times r}$ 称为基矩阵, $H \in \mathbb{R}_+^{r\times n}$ 称为系数矩阵, $r \ll \min\{m,n\}$.

为了能够使分解的矩阵更具有合理性, 可以定义如下的损失函数:

$$\min_{W,H} \|Y - WH\|_F^2 \quad \text{s.t. } W \geqslant 0,\ H \geqslant 0, \tag{13}$$

其中 $\|X\|_F$ 表示矩阵 $X$ 的 Frobenius 范数.

为了平衡最终结果的准确性以及算法的平滑性, Pauca 等 [37] 提出了一种正则化的非负矩阵分解算法, 其损失函数为

$$\min_{W,H} \|Y - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2 \quad \text{s.t. } W \geqslant 0,\ H \geqslant 0, \tag{14}$$

其中 $\alpha \in \mathbb{R}$ 和 $\beta \in \mathbb{R}$ 为正则项参数.

### 2.2.2 EDNMF 算法

本文提出了 EDNMF 算法, 其损失函数定义如下:

$$\min_{W,H} \|Y - WH\|_F^2 + \lambda_1(\|W\|_F^2 + \|H\|_F^2) + \lambda_2\|W - T\|_F^2 + \frac{\lambda_3}{2}\|H'H - D\|_F^2 \quad \text{s.t. } W \geqslant 0,\ H \geqslant 0. \tag{15}$$

EDNMF 算法通过规范化后的 circRNA 表达谱数据 $T$ 和疾病相似性数据 $D$ 对基矩阵和系数矩阵进行了约束. 其中, 约束项 $\|W - T\|_F^2$ 可以确保基矩阵 $W$ 趋向于规范化后 circRNA 表达谱矩阵 $T$, 这样不仅可以使 EDNMF 算法从 circRNA 表达谱角度分析 circRNA– 疾病关联关系, 而且也能够通过固定基矩阵和系数矩阵的维度解决传统 NMF 算法中维度不确定的缺点; 约束项 $\|H'H - D\|_F^2$ 使系数矩阵与系数矩阵转置的内积近似等于疾病相似性矩阵 $D$.

### 2.2.3 优化问题求解

解决式 (15) 最优化方法的主要思路是采用拉格朗日 (Lagrange) 乘子法和 Karush-Kuhn-Tucker (KKT) [38] 互补松弛条件进行非负约束, 式 (15) 的拉格朗日函数可以被定义成如下形式:

$$\begin{aligned} f(W,H) &= \|Y - WH\|_F^2 + \lambda_1(\|W\|_F^2 + \|H\|_F^2) + \lambda_2\|W - T\|_F^2 + \frac{\lambda_3}{2}\|H'H - D\|_F^2 \\ &= \mathrm{tr}(YY') - 2\mathrm{tr}(WHY') + \mathrm{tr}(WHH'W') + \lambda_1\mathrm{tr}(WW') + \lambda_1\mathrm{tr}(HH') \\ &\quad + \lambda_2\mathrm{tr}(WW') - 2\lambda_2\mathrm{tr}(WT') + \lambda_2\mathrm{tr}(TT') \\ &\quad + \frac{\lambda_3}{2}\mathrm{tr}(H'HH'H) - \lambda_3\mathrm{tr}(H'HD') + \frac{\lambda_3}{2}\mathrm{tr}(DD') \\ &\quad + \mathrm{tr}(\Lambda_1 W') + \mathrm{tr}(\Lambda_2 H'), \end{aligned} \tag{16}$$

其中 $\Lambda_1$ 和 $\Lambda_2$ 代表拉格朗日乘子. 根据式 (16), 可以得到 $W$ 和 $H$ 的偏导数:

$$\frac{\partial f(W,H)}{\partial W} = 2((WHH' + (\lambda_1 + \lambda_2)W) - (YH' + \lambda_2 T)) + \Lambda_1, \tag{17}$$

$$\frac{\partial f(W,H)}{\partial H} = 2((W'WH + \lambda_3 HH'H + \lambda_1 H) - (W'Y + \lambda_3 HD)) + \Lambda_2. \tag{18}$$

为了保证 $W \geqslant 0$, $H \geqslant 0$, 由 KKT 条件 [38] 可知 $\Lambda_1 \odot W = 0$ 和 $\Lambda_2 \odot H = 0$ ($\odot$ 代表 Hadamard 积). 进一步结合式 (17) 与 (18), 可得

$$((WHH' + (\lambda_1 + \lambda_2)W) - (YH' + \lambda_2 T)) \odot W = 0, \tag{19}$$

$$((W'WH + \lambda_3 HH'H + \lambda_1 H) - (W'Y + \lambda_3 HD)) \odot H = 0. \tag{20}$$

因此, 可以得到如下的更新规则:

$$W_{ik} \leftarrow \frac{(YH' + \lambda_2 T)_{ik}}{(WH'H + (\lambda_1 + \lambda_2)W)_{ik}} W_{ik}, \tag{21}$$

$$H_{ik} \leftarrow \frac{(W'Y + \lambda_3 HD)_{ik}}{(W'WH + \lambda_1 HH'H + \lambda_3 H)_{ik}} H_{ik}. \tag{22}$$

式 (21) 和 (22) 可以得到每次迭代过程中基矩阵 $W$ 和系数矩阵 $H$ 的更新规则, 并根据最终收敛后的 $Y$ 来预测是否存在 circRNA– 疾病关联关系. 显然, 对于最终收敛的 $Y$ 矩阵, 每一列中得分高的数值所在的行指代的 circRNA 与该列指代的癌症存在关联关系的概率更高.

#### 2.2.4 EDNMF 算法框架

对于 EDNMF 算法伪代码如算法 1 所示.

**Algorithm 1** EDNMF 算法: 计算 $W$, $H$ 使 $Y \approx WH$

**输入:** circRNA– 疾病关联关系矩阵 $Y$, circRNA 表达谱矩阵 $T$, 疾病相似性矩阵 $D$.
**输出:** 收敛后的矩阵 $Y$.

初始化基矩阵 $W$;
初始化系数矩阵 $H$;
$i \Leftarrow 1$;
**while** $i \neq$ MaxIter **do**
  更新基矩阵 $W$ 和系数矩阵 $H$:
    $W_{ik} \leftarrow \frac{(YH'+\lambda_2 T)_{ik}}{(WH'H+(\lambda_1+\lambda_2)W)_{ik}} W_{ik}$;
    $H_{ik} \leftarrow \frac{(W'Y+\lambda_3 HD)_{ik}}{(W'WH+\lambda_1 HH'H+\lambda_3 H)_{ik}} H_{ik}$;
  更新 $Y$ 矩阵;
  **if** $Y$ 矩阵收敛 **then**
    跳出当前 while 循环;
  **end if**
  $i = i + 1$;
**end while**
Return 收敛后的 $Y$ 矩阵.

## 3 实验结果与分析

### 3.1 评价标准

本文采用五折交叉验证和十折交叉验证方法来评估 EDNMF 算法的性能. 在五折交叉验证和十折交叉验证过程中, 我们可以得到对所有的 circRNA– 疾病关联关系的打分值. 这里, 我们将原本在 CD 矩阵中 1 所表示的关联关系作为正样本, 0 所表示的关联关系作为负样本. 然后采用 3 种评价方法进行评价, 即查准率 (precision)、受试者工作特征 (receiver operating characteristic, ROC) 曲线和 ROC 曲线下的面积 (area under ROC curve, AUC).

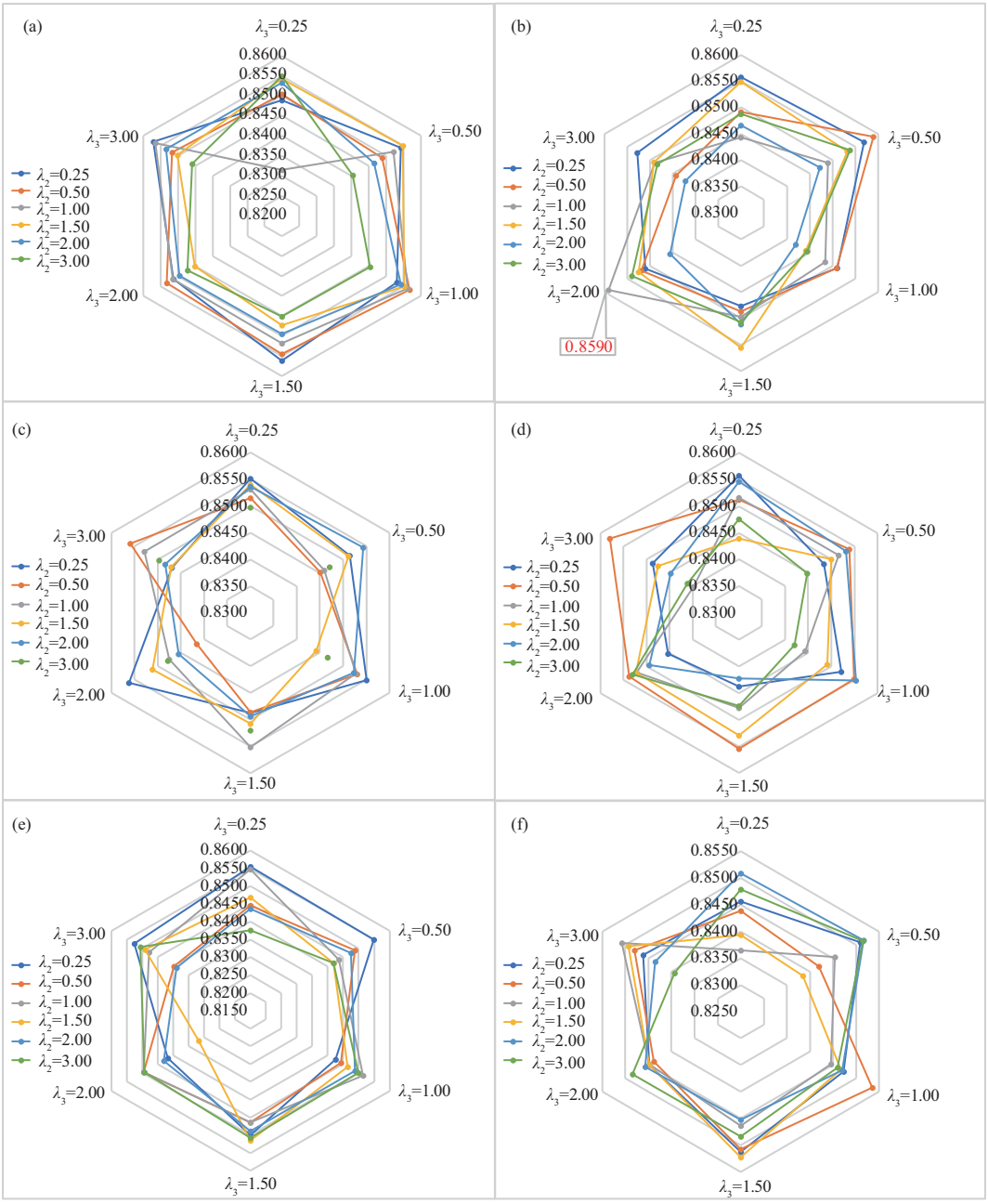


**图 1 (网络版彩图) 十折交叉验证过程中参数对 EDNMF 算法表现性能的影响**

**Figure 1** (Color online) The effect of different parameter values for EDNMF algorithm in 10-fold cross-validation. (a) $\lambda_1 = 0.25$; (b) $\lambda_1 = 0.5$; (c) $\lambda_1 = 1$; (d) $\lambda_1 = 1.5$; (e) $\lambda_1 = 2$; (f) $\lambda_1 = 3$

### 3.2 参数分析

EDNMF 算法中涉及到 3 个参数, 为了分析参数对预测结果的影响, 本文在十折交叉验证过程中分别令 $\lambda_1 \in \{0.25, 0.5, 1, 1.5, 2, 3\}$, $\lambda_2 \in \{0.25, 0.5, 1, 1.5, 2, 3\}$ 和 $\lambda_3 \in \{0.25, 0.5, 1, 1.5, 2, 3\}$, 并将参数进行组合, 不同参数值组合情况下的结果如图 1 所示. 由图 1 可知, 在所有的组合情况下, 当 $\lambda_1 = 0.5$,

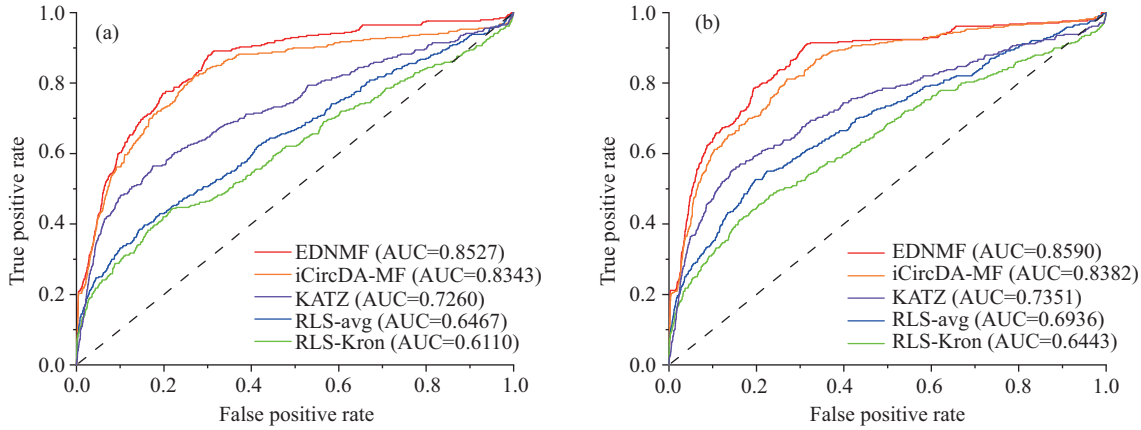


**图 2 (网络版彩图) 不同算法之间在 ROC 曲线和 AUC 值上的比较. (a) 表示在五折交叉验证上的表现性能; (b) 表示在十折交叉验证上的表现性能**

**Figure 2** (Color online) Compared with different algorithm on ROC curve and AUC value. (a) The performance of 5-fold cross-validation; (b) the performance of 10-fold cross-validation

$\lambda_2 = 1$ 和 $\lambda_3 = 2$ 时, AUC 取得最大值 0.8590.

### 3.3 与其他方法的比较

为了分析 EDNMF 算法 ($\lambda_1 = 0.5$, $\lambda_2 = 1$ 和 $\lambda_3 = 2$) 在预测 circRNA– 疾病关联关系上的表现性能, 将 EDNMF 分别与现有最新的 4 种算法进行比较, 分别是 iCircDA-MF 算法 [32]、KATZ 算法 [26,31]、RLS-Kron 算法 [39] 和 RLS-avg 算法 [39]. 图 2 展示了五折交叉验证和十折交叉验证中 EDNMF 算法与其他 4 种算法的 ROC 曲线以及 AUC 值; 图 3 表示在五折交叉验证和十折交叉验证中 EDNMF 算法与其他 4 种算法的 precision 曲线; 图 4 表示在十折交叉验证中 EDNMF 算法与其他 4 种算法在 6 种常见疾病上的表现性能. 显然, 由图 2~4 可以看出, EDNMF 算法在五折交叉验证和十折交叉验证上获得了令人满意的表现性能; 此外, EDNMF 算法在五折和十折交叉验证上的 AUC 分别为 0.8527 和 0.8590, 显然, 它们差别很小, 这体现了 EDNMF 算法的稳定性.

### 3.4 预测新 circRNA– 癌症关联关系

为了展现 EDNMF 算法在预测新的关联关系上的表现性能, 本研究选择结肠直肠癌 (colorectal cancer, DOID:9256) 进行实验验证. 在我们的数据集中, 它与 22 个 circRNAs 存在关联关系. 在实验过程中, 所有的 circRNA– 疾病关联关系不变, 然后执行 EDNMF 算法为所有未知的 circRNA 进行打分. 得到打分值之后, 我们将对未知 circRNA 打分值进行降序排序, 然后取排名前 10 的 circRNAs 进行交互网络分析.

对于交互网络方法, 本文首先从 DISEASE 数据库 [40] 中提取了结肠直肠癌疾病相关基因. DISEASE 数据库 [40] 中收录了基因 – 疾病关联关系, 并且依据一定证据对每一条基因 – 疾病关联关系进行打分, 以此来评价基因 – 疾病关联关系的可靠性. 因此, 为了保证数据的准确性, 我们仅仅提取打分大于 2 的基因 – 结肠直肠癌关联关系. 此外, 我们也从 Valdeolivas 等 [41] 文章中分别获取了蛋白质 – 蛋白质交互作用 (protein-protein interaction, PPI) 数据和 Pathway 交互作用数据. 然后, 我们寻找与结肠癌相关的 circRNAs 的宿主基因, 以及预测的结肠癌相关的 circRNAs 的宿主基因, 接下来判断它们是否在 PPI 数据和 Pathway 交互作用数据中与结肠直肠癌基因有交互作用. 具体结果如图 5 所示.

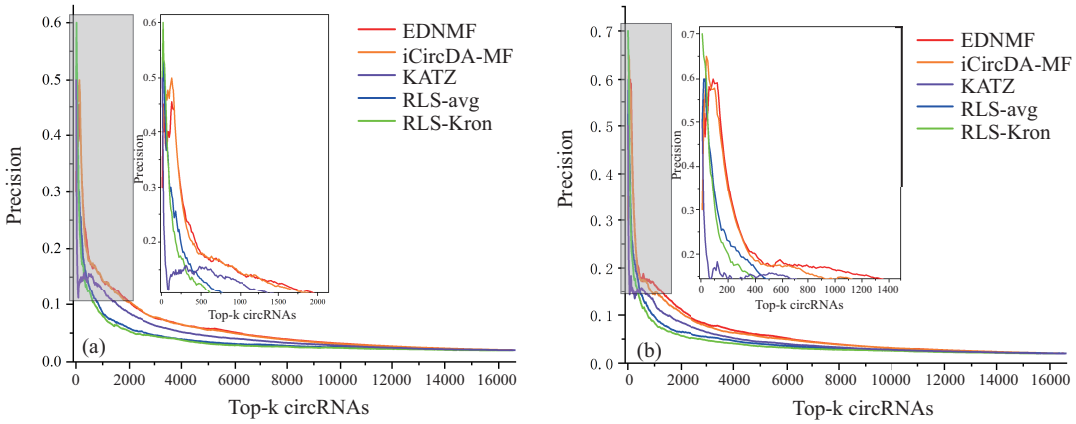


**图 3** (网络版彩图) 不同算法之间在 precision 上的比较. (a) 在五折交叉验证上的表现性能; (b) 在十折交叉验证上的表现性能

**Figure 3** (Color online) Comparison of different algorithms on precision. (a) The performance of 5-fold cross-validation; (b) the performance of 10-fold cross-validation

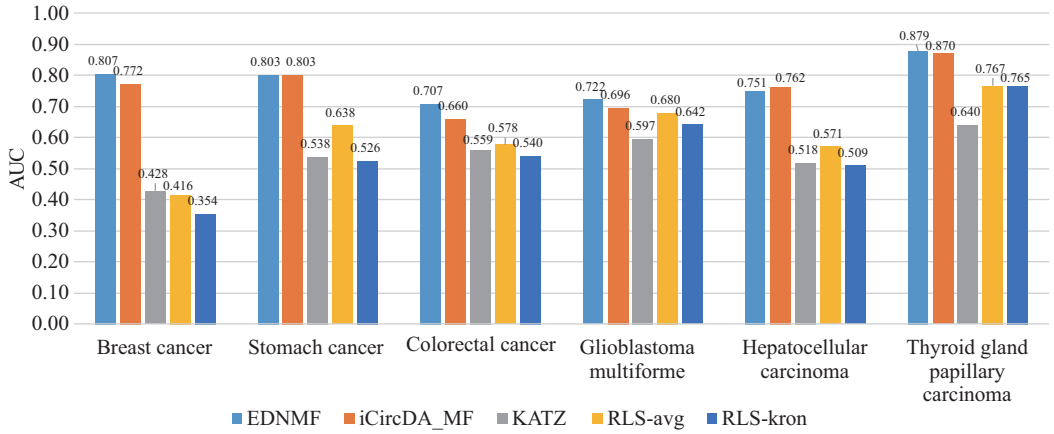


**图 4** (网络版彩图) 不同算法在十折交叉验证过程中在 6 中疾病上的比较

**Figure 4** (Color online) Comparison of different algorithms on 6 diseases in the process of 10-fold cross-validation

由图 5 可以看出, 在预测的前 10 个 circRNAs 中有 6 个 circRNAs 所对应的基因与结肠直肠癌相关的基因存在交互作用; 预测的 circRNA hsa_circ_0004680 与已知的与结肠直肠癌相关的 circRNA hsa_circ_0014717 指向了相同的宿主基因 CCT3; 此外, 预测的 circRNA hsa_circ_0004680, hsa_circ_0007707, hsa_circ_0018168, hsa_circ_0001017 与已知的 circRNA hsa_circ_0001649, hsa_circ_0014717, hsa_circ_0008494, hsa_circ_0024169 的宿主基因存在间接的交互作用. 以上实验结果表明, 预测出来前 10 的 circRNA 在一定程度上很有可能与结肠直肠癌有关.

综上所述, EDNMF 算法可以很好地预测未知的 circRNA– 疾病关联关系.

## 4 总结

本文将 circRNA 表达谱数据和疾病相似性数据转化为非负矩阵分解过程中的约束条件, 提出了一种基于 circRNA 表达谱数据和疾病相似性数据的非负矩阵分解算法预测与疾病相关联的 circRNA.

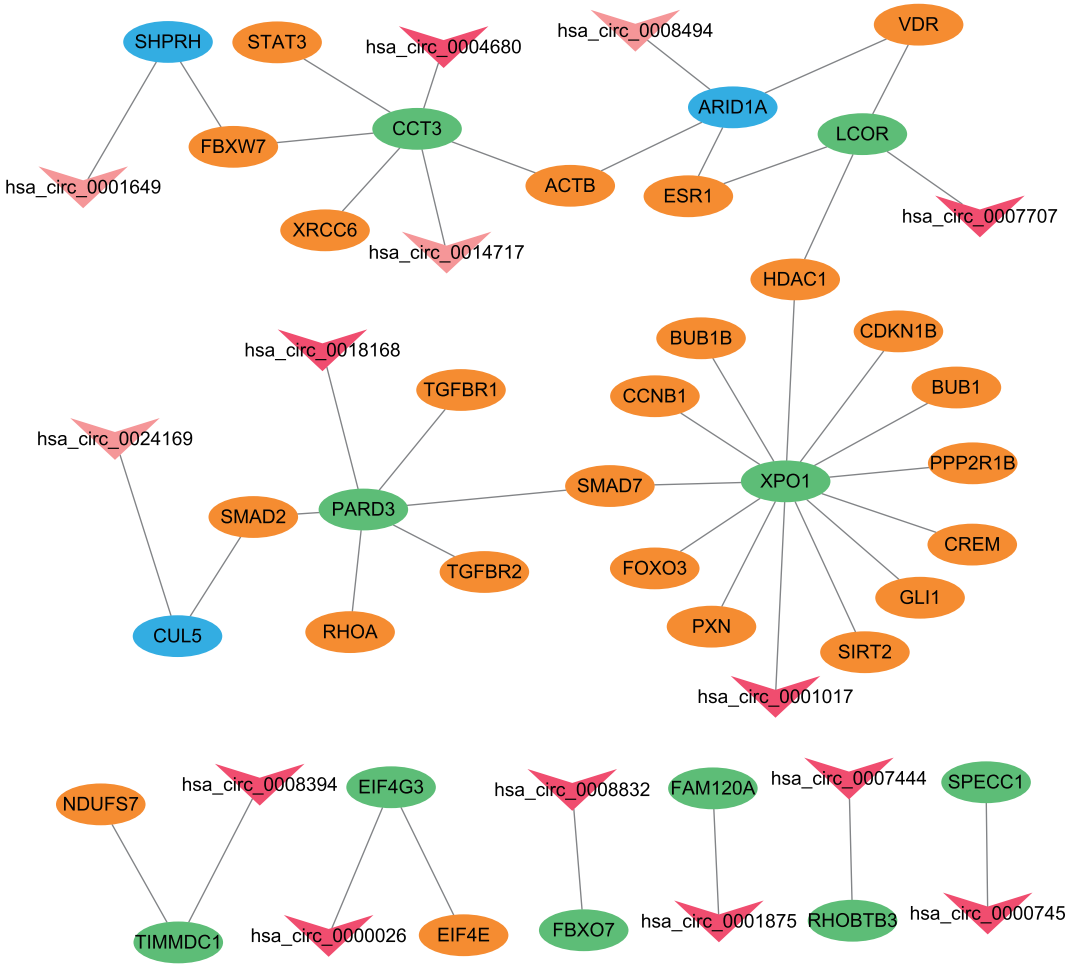


**图 5** (网络版彩图) 交互网络方法验证分析 **circRNA–** 结肠直肠癌关联关系中排名在前 **10** 的结果. 深红色表示预测与结肠直肠癌相关的 **circRNAs;** 浅红色代表已知的与结肠直肠癌相关的 **circRNAs;** 蓝色表示已知与结肠直肠癌相关的 **circRNAs** 所对应的基因; 绿色表示预测的与结肠直肠癌相关的 **circRNAs** 所对应的基因; 橙色表示结肠直肠癌相关的基因

**Figure 5** (Color online) The interactive network method validate the top 10 results in predicting the associations between circRNAs and colorectal cancer. Crimson color represents the top 10 circRNAs; light red color represents the known circRNAs associated with colorectal cancer; blue represents gene corresponding to the known circRNAs associated with colorectal cancer; green color represents gene corresponding to the top 10 circRNAs; orange represents genes associated with colorectal cancer

该算法将 circRNA 表达谱数据和疾病相似性数据转化为约束条件, 在一定程度上使分解出来的基矩阵和系数矩阵符合实际情况, 并且整合了多种类型的生物数据, 从多个角度综合分析 circRNA– 疾病关联关系. 实验结果表明, 该算法可以很好地预测 circRNA– 疾病关联关系.

虽然 EDNMF 算法能够很好地预测 circRNA– 疾病之间的关联关系, 但复杂疾病的产生受到多种因素的影响, 算法仍然存在一定缺陷. EDNMF 算法的参数相对较多, 需要进行大量的实验来寻找最优参数. 在大规模数据分析中, circRNA 的数量可能在数千到数万之间, EDNMF 算法仅仅利用极少量的 circRNA 构建模型. 主要的原因有两点: (1) circRNAs 的命名并不统一 [42], 这必将导致在不同数据库中收录的 circRNA– 疾病关联关系很难进行数据合并; (2) circRNAs 的研究为目前生物信息学领域

新兴的热点问题, 因此 circRNA– 疾病关联关系数据较少且不标准. 这些问题我们后续将进一步深入研究.

## 参考文献


1 Memczak S, Jens M, Elefsinioti A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 2013, 495: 333–338
2 Zhang Y, Zhang X O, Chen T, et al. Circular intronic long noncoding RNAs. Mol Cell, 2013, 51: 792–806
3 Sanger H L, Klotz G, Riesner D, et al. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. Proc Natl Acad Sci USA, 1976, 73: 3852–3856
4 Hsu M T, Coca-prados M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. Nature, 1979, 280: 339–340
5 Matsumoto Y, Fishel R, Wickner R B. Circular single-stranded RNA replicon in saccharomyces cerevisiae. Proc Natl Acad Sci USA, 1990, 87: 7628–7632
6 Lasda E, Parker R. Circular RNAs: diversity of form and function. RNA, 2014, 20: 1829–1842
7 Nigro J M, Cho K R, Fearon E R, et al. Scrambled exons. Cell, 1991, 64: 607–613
8 Danan M, Schwartz S, Edelheit S, et al. Transcriptome-wide discovery of circular RNAs in archaea. Nucleic Acids Res, 2012, 40: 3131–3142
9 Chu Q, Zhang X, Zhu X, et al. PlantcircBase: a database for plant circular RNAs. Mol Plant, 2017, 10: 1126–1128
10 Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA, 2014, 20: 1666–1670
11 Fan C, Lei X, Fang Z, et al. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. J Biol Databases Curation, 2018, 2018: 44
12 Xia S, Feng J, Chen K, et al. CSCD: a database for cancer-specific circular RNAs. Nucleic Acids Res, 2017, 46: 925–929
13 Ji P, Wu W, Chen S, et al. Expanded expression landscape and prioritization of circular RNAs in mammals. Cell Rep, 2019, 26: 3444–3460
14 Hansen T B, Jensen T I, Clausen B H, et al. Natural RNA circles function as efficient microRNA sponges. Nature, 2013, 495: 384–388
15 Chao C W, Chan D C, Kuo A, et al. The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis. Mol Med, 1998, 4: 614–628
16 Du W W, Yang W N, Chen Y, et al. Foxo3 circular RNA promotes cardiac senescence by modulating multiple factors associated with stress and senescence responses. Eur Heart J, 2017, 38: 1402–1412
17 Legnini I, Di Timoteo G, Rossi F, et al. Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. Mol Cell, 2017, 66: 22–37
18 Yang Y B, Gao X Y, Zhang M L, et al. Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. J Natl Cancer Inst, 2018, 110: 304–315
19 Guo M Z, Wang S M, Liu X Y, et al. Algorithm for predicting the associations between MiRNAs and diseases. J Softw, 2017, 28: 3094–3102 [郭茂祖, 王诗鸣, 刘晓燕, 等. miRNA 与疾病关联关系预测算法. 软件学报 2017, 28: 3094–3102]
20 Guo M Z, Wu X J, Zhao N, et al. A method for mining core modules of cancer based on multi-omics biological network. Sci Sin Inform, 2017, 47: 1510–1522 [郭茂祖, 武雪剑, 赵宁, 等. 一种基于多组学生物网络的癌症关键模块挖掘方法. 中国科学: 信息科学 2017, 47: 1510–1522]
21 Deng Y, Gao L, Guo X L, et al. Integrating phenotypic features and tissue-specific information to prioritize disease genes. Sci China Inf Sci, 2016, 59: 070101
22 Huang X, Li Z, Zhang Q, et al. Circular RNA AKT3 upregulates PIK3R1 to enhance cisplatin resistance in gastric cancer via miR-198 suppression. Mol Cancer, 2019, 18: 71
23 Li X N, Wang Z J, Ye C X, et al. Circular RNA circVAPA is up-regulated and exerts oncogenic properties by sponging miR-101 in colorectal cancer. Biomed Pharmacother, 2019, 112: 108611
24 Xu L L, Feng X F, Hao X Y, et al. CircSETD3 (Hsa_circ_0000567) acts as a sponge for microRNA-421 inhibiting hepato- cellular carcinoma growth. J Exp Clin Cancer Res, 2019, 38: 15

25 Shi F, Shi Z H, Zhao Y D, et al. CircRNA hsa_circ_0014359 promotes glioma progression by regulating miR-153/PI3K signaling. Biochem Biophys Res Commun, 2019, 510: 614–620
26 Fan C Y, Lei X J, Wu F X. Prediction of circRNA-disease associations using KATZ model based on heterogeneous networks. Int J Biol Sci, 2018, 14: 1950–1959
27 Yan C, Wang J X, Wu F X. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. BMC Bioinf, 2018, 19: 520
28 Zhao Z, Wang K, Wu F, et al. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. Cell Death Dis, 2018, 9: 475
29 Lei X J, Fang Z Q, Chen L N, et al. PWCDA: path weighted method for predicting circRNA-disease associations. Int J Molecular Sci, 2018, 19: 3410
30 Xiao Q, Luo J, Dai J. Computational prediction of human disease- associated circRNAs based on manifold regularization learning framework. IEEE J Biomed Health Inform, 2019, 23: 2661–2669
31 Zhao Q, Yang Y, Ren G, et al. Integrating bipartite network projection and KATZ measure to identify novel circRNA-Disease Associations. IEEE Transon Nanobiosci, 2019, 18: 578–584
32 Wei H, Liu B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. Briefings Bioinf, 2019, : 57
33 Kibbe W A, Arze C, Felix V, et al. Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. Nucleic Acids Res, 2015, 43: 1071–1078
34 Li S, Li Y, Chen B, et al. exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. Nucleic Acids Res, 2018, 46: 106–112
35 Yu G, Wang L G, Yan G R, et al. DOSE: an r/bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics, 2015, 31: 608–609
36 Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401: 788–791
37 Pauca V P, Piper J, Plemmons R J. Nonnegative matrix factorization for spectral data analysis. Linear Algebra Its Appl, 2006, 416: 29–47
38 Facchinei F, Kanzow C, Sagratella S. Solving quasi-variational inequalities via their KKT conditions. Math Program, 2014, 144: 369–412
39 van Laarhoven T, Nabuurs S B, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics, 2011, 27: 3036–3043
40 Pletscher-Frankild S, Pallejá A, Tsafou K, et al. DISEASES: text mining and data integration of disease-gene associations. Methods, 2015, 74: 83–89
41 Valdeolivas A, Tichit L, Navarro C, et al. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics, 2019, 35: 497–505
42 Xu S, Zhou L Y, Ponnusamy M, et al. A comprehensive review of circRNA: from purification and identification to disease marker potential. Peerj, 2018, 6: e5503

# Prediction of circRNA-disease associations based on multiple biological data

Xiujuan LEI*, Wenxiang ZHANG & Lian LIU

*College of Computer Science, Shaanxi Normal University, Xi'an* 710000*, China*
* Corresponding author. E-mail: xjlei@snnu.edu.cn

**Abstract** Circular RNA (circRNA) plays a significant role in gene expression, splicing, and transcription. More and more evidence indicates that circRNA is related to the pathogenesis and development of diseases. In this paper, a non-negative matrix factorization algorithm based on circRNA expression profiles data and disease similarity data (EDNMF) is proposed to predict circRNA-disease associations. The EDNMF algorithm firstly preprocesses the circRNA-disease associations to solve the impact of too little the number of known circRNA-disease associations. Then, the EDNMF algorithm converts circRNA expression profile and cancer similarity data into constraints. Finally, we can obtain the final scores for circRNA-disease associations by improved NMF algorithm based on pre-processed circRNA-disease associations. The performance results of 5-fold and 10-fold cross-validation indicate that the EDNMF algorithm achieves satisfactory performance comparing with other algorithms. Besides, the case study shows that EDNMF can mine new circRNA-disease associations very well, which can provide a reference for studying circRNA-disease associations.

**Keywords** circRNA, circRNA expression profiles, circRNA-disease associations, non-negative matrix factorization, disease similarity

**Xiujuan LEI** received her Ph.D. from Northwestern Polytechnical University, Xi'an, China, in 2005. She is currently a professor at School of Computer Science, Shaanxi Normal University, Xi'an, China. Her current research interests include bioinformatics, intelligent computing, pattern recognition and data mining.

**Wenxiang ZHANG** was born in Zaozhuang, Shandong, China in 1994. He received his B.S. degree in computer science and technology from Qufu Normal University, Ri'zhao, China, in 2017. He is currently a master at School of Computer Science, Shaanxi Normal University, Xi'an, China. His research interests include bioinformatics, intelligent computing and machine learning, etc.

**Lian LIU** received her Ph.D. from Northwestern Polytechnical University, Xi' an, China, in 2018. She is now a postdoctoral student at Shaanxi Normal University. She is currently an assistant research fellow at School of Computer Science, Shaanxi Normal University, Xi' an, China. Her current research interests include bioinformatics, pattern recognition and machine learning.