



任务进展汇报

2024.4.5

Abstract

- 1、SGNNMD 文献代码部分 理解
- 2、miRNA与疾病关系预测分析的非深度学习
方法（文献）
- 3、TWAS方法开发代码

1、SGNNMD

标签分配

- - **0和1**：分配给中心的miRNA和疾病节点，标记网络的起点和终点。
- - **2和3**：用于直接与疾病节点 v_j 相连的miRNA节点，区分这些节点是上调（2）还是下调（3）。
- - **4和5**：用于直接与miRNA节点 u_i 相连的疾病节点，区分这些节点是上调（4）还是下调（5）。

1-hop内：直接与中心节点相连的节点被标记，反映直接的上调或下调关系。

Input: Enclosing subgraph G_s for miRNA u_i and disease v_j , and hop number $H, l=0$;

Output: subgraph G_s with labeled nodes.

```
1: assign integer 0, 1 to  $u_i$  and  $v_j$ 
2: for unlabeled miRNA nodes linked to  $v_j$  do
3:   if edge is up-regulation then
4:     assign an integer 2 to it.
5:   else
6:     assign an integer 3 to it.
7:   end if
8: end for
9: for unlabeled disease nodes linked to  $u_i$  do
10:  if edge is up-regulation then
11:    assign an integer 4 to it.
12:  else
13:    assign an integer 5 to it.
14:  end if
15: end for
16: all nodes in  $G(1)$  have been labeled.
```

1、SGNNMD

标签分配

- 对于每个新加入的miRNA节点，它们的标签 ($l + 0, l + 1, l + 2$) 根据它们到中心疾病节点的路径中边的类型（全部上调、全部下调或混合）来确定。
- 对于每个新加入的疾病节点，它们的标签 ($l + 3, l + 4, l + 5$) 根据它们到中心miRNA节点的路径中边的类型（全部上调、全部下调或混合）来确定。

h-hop内 ($h > 1$)

```
17: for  $h = 2, 3, \dots, H$  do
18:   for miRNA nodes  $\in G(h) - G(h - 1)$  do
19:     let  $path$  denote the path linking it to target
       disease
20:     if all up-regulation edges in  $path$  then
21:       assign an integer  $l + 0$  to it.
22:     else if all down-regulation edges in  $path$  then
23:       assign an integer  $l + 1$  to it.
24:     else
25:       assign an integer  $l + 2$  to it.
26:     end if
27:   end for
28:   for disease nodes  $\in G(h) - G(h - 1)$  do
29:     let  $path$  denote the path linking it to target
       miRNA
30:     if all up-regulation edges in  $path$  then
31:       assign a integer  $l + 3$  to it.
32:     else if all down-regulation edges in  $path$  then
33:       assign a integer  $l + 4$  to it.
34:     else
35:       assign a integer  $l + 5$  to it.
36:     end if
37:   end for
38:    $l = l + 6$ 
39: end for
40: return subgraph  $G_s$  with labeled nodes
```

1、SGNNMMD

代码不足

- 可解释性不足 无上下调关系来源的具体说明
- 方法疑似对前人的方法整合改进

problems. However, SGNNMMD extracts subgraphs around node pairs from the signed graph to train the prediction model, and it takes lots of training time if we have a number of subgraphs or a large signed graph. In this case, we have to restrict the number of subgraphs and sizes of subgraphs (number of nodes) to reduce the computational complexity.

Key Points

- We study how to predict the deregulation types of miRNA-disease associations, on which little attention has been paid previously. It benefits exploring how genetic variants in miRNA genes affect the expression level of miRNAs and lead to diseases.
- We formulate the original problem as a signed graph link prediction task, and propose a graph neural network-based method SGNNMMD to resolve it. In SGNNMMD, a novel node labeling algorithm is designed for subgraphs from the signed graph, and it can better describe the structural information. SGNNMMD can generalize to miRNAs/diseases unseen in the training set.
- SGNNMMD leverages the structural information learned from subgraphs around miRNA-disease pairs as well as the biological information of miRNAs and diseases, and trains the prediction model in an end-to-end manner. The structural information leads to the high-accuracy prediction model, and the biological information further enhances the performance.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib/article/23/1/bbaa140/6455665> by Huazhong Agricultural University user on 14 November 2023

Data availability

The datasets were derived from the following sources in the public domain, the miRNA-disease association from <https://www.cuilab.cn/hmdd>, the disease MeSH descriptors from <https://meshb.nlm.nih.gov/> and the gene functional interaction network is downloaded from <https://www.inetbio.org/humannet/download.php>. The implementation of SGNNMMD and the preprocessed data is available at <https://github.com/bubblecode/SGNNMMD> and <https://github.com/BioMedicalBigDataMiningLab/SGNNMMD>.

Author contributions statement

W.Z. conceived the project, G.Z. and H.D. conducted the experiment(s), W.Z., G.Z., X.X., M.L. and X.L. analyzed the results and wrote the manuscript.

(62072206, 61772381); Huazhong Agricultural University Scientific & Technological Self-innovation Foundation; Fundamental Research Funds for the Central Universities (2662021JC008).

Acknowledgments

We thank anonymous reviewers for their valuable suggestions.

References

1. Llave C, Xie Z, Kasschau KD, et al. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science* 2002; **297**(5589): 2053–6.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004; **116**(2): 281–97.
3. Yue X, Wang Z, Huang J, et al. Graph embedding on biomedical networks: methods, applications and evaluations. *Bioinformatics* 2020; **36**(4): 1241–51.
4. Zhang ZC, Zhang XF, Wu M, et al. A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks. *Bioinformatics* 2020; **36**(11): 3474–81.
5. Shen Z, Zhang YH, Han K, et al. miRNA-Disease Association Prediction with Collaborative Matrix Factorization. *Complexity* 2017; **2017**: 1–9.
6. Chen X, Xie D, Wang L, et al. BNPMDA: bipartite network projection for miRNA-disease association prediction. *Bioinformatics* 2018; **34**(18): 3178–86.
7. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* 2019; **35**(22): 4730–8.
8. Zeng X, Liu L, Lü L, et al. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* 2018; **34**(14): 2425–32.
9. Peng J, Hui W, Li Q, et al. A learning-based framework for miRNA-disease association identification using neural networks. *Bioinformatics* 2019; **35**(21): 4364–71.
10. Zhang W, Li Z, Guo W, et al. A fast linear neighborhood similarity-based network link inference method to predict microRNA-disease associations. *IEEE/ACM Trans Comput Biol Bioinform* 2021; **18**(2): 405–15.
11. Chen X, Wang L, Qu J, et al. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* 2018; **34**(24): 4256–65.
12. Pan X, Shen HB. Inferring disease-associated microRNAs using semi-supervised multi-label graph convolutional networks. *Iscience* 2019; **20**: 265–77.
13. Li J, Zhang S, Liu T, et al. Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. *Bioinformatics* 2020; **36**(8): 2538–46.
14. Huang Z, Shi J, Gao Y, et al. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2019; **47**(D1): D1013–7.
15. Chen X, Clarence Yan C, Zhang X, et al. RBMMMDA: predicting multiple types of disease-microRNA associations. *Sci Rep* 2015; **5**(1): 13877.
16. Huang F, Yue X, Xiong Z, et al. Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations. *Brief Bioinform* 2020; **Bbaa140**.



其他非深度学习方法

2010-2015

2015-2021



Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks

Xiangxiang Zeng, Xuan Zhang and Quan Zou

Corresponding author: Quan Zou, 422F Simingnan Road, Department of Computer Science, Xiamen University, Xiamen, 361005, P.R.China. Tel.: +86-592-2580333; Fax: +86-5922580333; E-mail: zouquan@xmu.edu.cn

Abstract

MicroRNAs (miRNA) play critical roles in regulating gene expressions at the posttranscriptional levels. The prediction of disease-related miRNA is vital to the further investigation of miRNA's involvement in the pathogenesis of disease. In previous years, biological experimentation is the main method used to identify whether miRNA was associated with a given disease. With increasing biological information and the appearance of new miRNAs every year, experimental identification of disease-related miRNAs poses considerable difficulties (e.g. time-consumption and high cost). Because of the limitations of experimental methods in determining the relationship between miRNAs and diseases, computational methods have been proposed. A key to predict potential disease-related miRNA based on networks is the calculation of similarity among diseases and miRNA over the networks. Different strategies lead to different results. In this review, we summarize the existing computational approaches and present the confronted difficulties that help understand the research status. We also discuss the principles, efficiency and differences among these methods. The comprehensive comparison and discussion elucidated in this work provide constructive insights into the matter.

Key words: disease miRNA prediction; network similarity; biological database

Downloaded from <http://bib.oxfordjournals.org/> at University of Man

2015

2021

各类方法

深度学习的方法居多

2010



DOI:10.1093/bib/bbv033

Jiang et al. BMC Systems Biology 2010, 4(Suppl 1):S2
<http://www.biomedcentral.com/1752-0509/4/S1/S2>



RESEARCH

Open Access

Prioritization of disease microRNAs through a human phenome-microRNAome network

Qinghua Jiang^{1†}, Yangyang Hao^{1†}, Guohua Wang¹, Liran Juan¹, Tianjiao Zhang¹, Mingxiang Teng¹, Yunlong Liu², Yadong Wang^{1*}

From The ISIBM International Joint Conferences on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS)
Shanghai, China. 3-8 August 2009

Abstract

Background: The identification of disease-related microRNAs is vital for understanding the pathogenesis of diseases at the molecular level, and is critical for designing specific molecular tools for diagnosis, treatment and prevention. Experimental identification of disease-related microRNAs poses considerable difficulties. Computational analysis of microRNA-disease associations is an important complementary means for prioritizing microRNAs for further experimental examination.

Results: Herein, we devised a computational model to infer potential microRNA-disease associations by prioritizing the entire human microRNAome for diseases of interest. We tested the model on 270 known experimentally verified microRNA-disease associations and achieved an area under the ROC curve of 75.80%. Moreover, we demonstrated that the model is applicable to diseases with which no known microRNAs are associated. The microRNAome-wide prioritization of microRNAs for 1,599 disease phenotypes is publicly released to facilitate future identification of disease-related microRNAs.

Conclusions: We presented a network-based approach that can infer potential microRNA-disease associations and drive testable hypotheses for the experimental efforts to identify the roles of microRNAs in human diseases.

DOI: 10.1186/1752-0509-4-S1-S2

最早的研究miRNA与疾病关系的方法
第一篇文献

2、非深度学习方法

大致归类



2.1、传统机器学习

种类、方法各异

✓ RKNMMDA: Ranking-based KNN for MiRNA-Disease Association prediction

结合了K最近邻 (KNN) 算法和支持向量机 (SVM) 排名模型等技术。这种方法利用了多种生物数据集, 包括疾病相似性网络、miRNA相似性网络...

PubMed Central Full Text PDF

PubMed Central Link

✓ Prediction of miRNA-disease associations with a vector space model

使用了分布语义学方法来揭示miRNAs和疾病之间的关联, 通过将它们表示为高维向量并计算向量相似性来定义这种关联

全文

PubMed entry

✓ Predicting Multiple Types of Associations Between miRNAs and Diseases Based on Graph Regularized Weighted Tensor Decomposition

用了一种新颖的计算框架称为WeightTDAIGN来解决miRNA-疾病关联预测的问题。该框架基于加权张量分解, 引入了图拉普拉斯正则化以及L2,1...

PubMed Central Full Text PDF

PubMed Central Link

✓ Predicting miRNA-disease association based on inductive matrix completion

基于相似性计算和矩阵补全的方法来预测miRNA与疾病之间的关联

全文

PubMed entry

✓ PBMDA: A novel and effective path-based computational model for miRNA-disease association prediction

基于图论 利用了已知的miRNA-疾病关联、miRNA功能相似性、疾病语义相似性以及相互作用概要核相似性等信息构建了一个异构图, 并采用深...

全文

PubMed entry

✓ NTSMDA: prediction of miRNA-disease associations by integrating network topological similarity

基于已知的miRNA-疾病网络的拓扑相似性

全文

PubMed entry

✓ Novel Human miRNA-Disease Association Inference Based on Random Forest

机器学习方法中的随机森林算法。

全文

PubMed entry

✓ miRNA-Disease Association Prediction with Collaborative Matrix Factorization

用了矩阵因子分解等传统的机器学习方法, 并结合了 miRNA 的功能相似性、疾病的语义相似性以及已验证的 miRNA-疾病关联信息进行预测

✓ miRNA-Disease Association Prediction with Collaborative Matrix Factorization

基于矩阵分解和协同过滤的计算方法, 用于预测miRNA与疾病的关联

Shen 等 - 2017 - miRNA-Disease Association Prediction with Collabor.pdf

✓ MCMDA: Matrix completion for MiRNA-disease association prediction

使用了矩阵补全算法来更新已知miRNA-disease关联的邻接矩阵, 并预测潜在的关联

PubMed Central Full Text PDF

PubMed Central Link

✓ MCCMF: collaborative matrix factorization based on matrix completion for predicting miRNA-disease associations



矩阵补全、相似性核函数、以及协同矩阵分解等传统机器学习和数据分析方法

✓ LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction

稀疏子空间学习和拉普拉斯正则化, 以及L1范数约束。

全文

PubMed entry

✓ Ensemble of decision tree reveals potential miRNA-disease associations

涉及到特征提取、决策树以及集成学习等技术

Full Text PDF

✓ Adaptive multi-view multi-label learning for identifying disease-associated candidate miRNAs

基于自适应多视图多标签学习 (AMVML)

全文

PubMed entry

✓ Adaptive boosting-based computational model for predicting potential miRNA-disease associations

使用了自适应增强方法 (Adaptive Boosting), 并将决策树作为弱分类器

全文

PubMed entry

✓ A Fast Linear Neighborhood Similarity-Based Network Link Inference Method to Predict MicroRNA-Disease Associations

基于线性邻域相似度和标签传播算法等传统机器学习方法

全文

PubMed entry

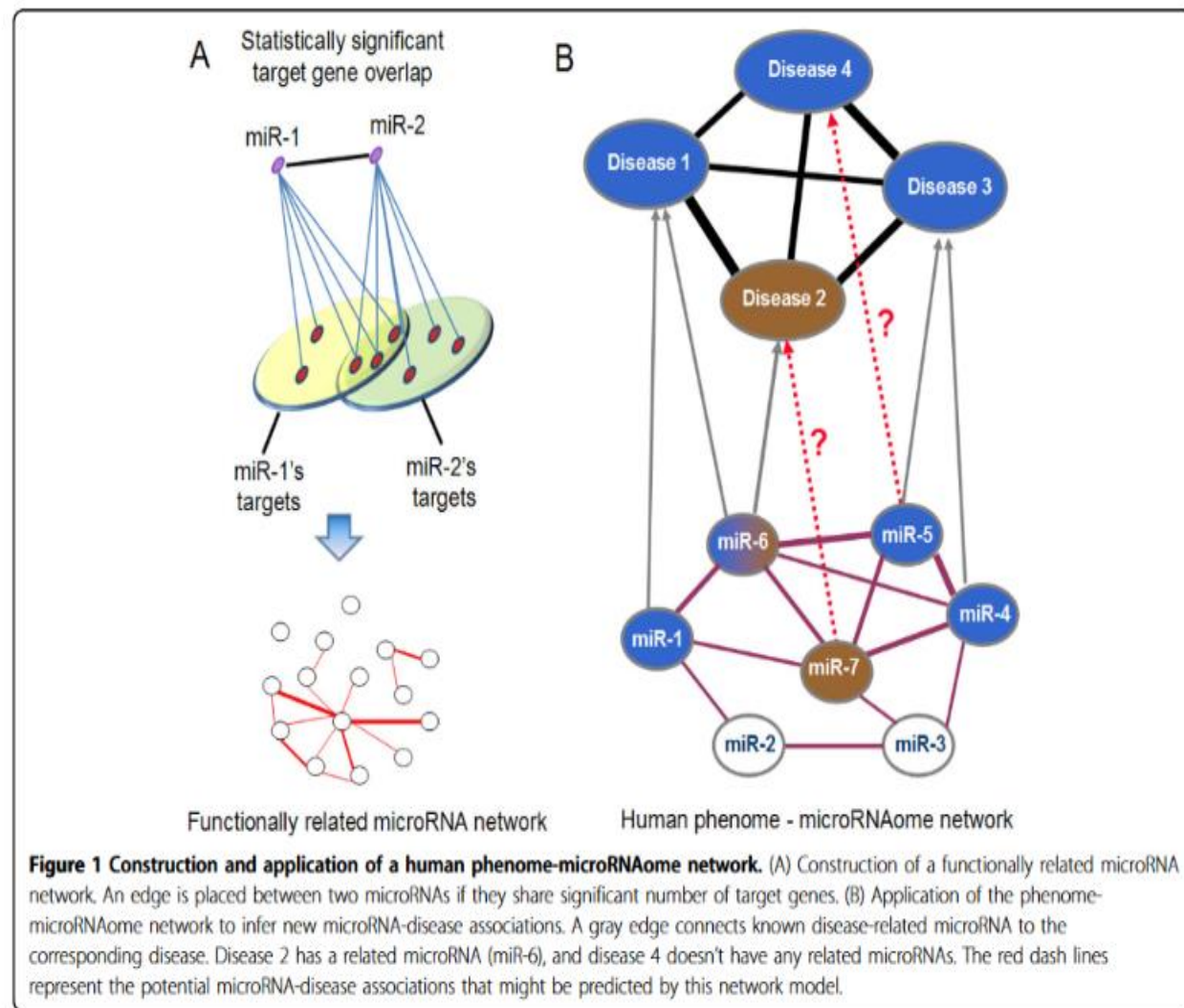
✓ 基于贝叶斯矩阵分解和异构网络算法的microRNA与疾病关联预测研究

两种模型

2.2、基于网络模型

最早的方法

- 将miRNA-疾病关联预测问题看作miRNA-疾病网络上的链路预测问题。
- 提出了第一个使用离散的超几何概率分布来识别miRNA与疾病关联的计算方法



2.3、其他计算方法

- ▼ WBSMDA: Within and Between Score for MiRNA-Disease Association prediction
 - 利用了多种异质生物数据集来预测miRNA与疾病之间的关联
 - 全文
 - PubMed entry
- ▼ Tensor decomposition with relational constraints for predicting multiple types of microRNA-disease associations
 - 基于整合多种异质生物数据集的计算方法来预测miRNA与复杂疾病的关联
 - 已提交版本
 - PubMed entry
- ▼ Predicting miRNA-disease association based on inductive matrix completion
 - 它使用了图正则化非负矩阵分解等方法来分析异质组学数据,
 - 全文
 - PubMed entry
- ▼ HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction
 - 采用了一种结合了多种生物信息学和网络分析方法的计算模型。该模型利用了异质网络的概念,并结合了miRNA功能相似性、疾病语义相似性、高...
 - PubMed Central Full Text PDF
 - PubMed Central Link
- ▼ A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations
 - 整合异质组学数据、构建交互作用评分、图正则非负矩阵分解
 - 全文
 - PubMed entry

基于miRNA功能相似度或疾病相似度
基于miRNA集合富集分析
疾病—miRNA—环境因子

2.4-2.5、数据库与相关综述

- 1. HMDD: 人类miRNA疾病数据库, 提供了大量的miRNA与疾病关联数据。
- 2. miR2Disease: 手工整理的数据库, 收录了miRNA在人类疾病中的失调信息。
- 3. miRCancer: 专注于miRNA在癌症中的作用。
- 4. dbDEMC: 提供了癌症中差异表达miRNA的数据。
- 5. PhenomiR: 提供miRNA表达变化与疾病及生物过程的关系数据。

生理科学进展 2016 年第 47 卷第 3 期

· 203 ·

microRNA 与人类疾病关系研究中的 生物信息学方法和资源

张帆¹ 崔庆华^{2,△}

(¹ 北京大学第三医院老年内科, 北京 100191;

² 北京大学医学部非编码 RNA 医学研究中心、医学信息学系, 北京 100191)

摘要 microRNA (miRNA) 是一类基因调控分子, 其成熟体长度大约为 22 个核苷酸, 其在转录后水平通过碱基互补配对的方式特异性结合靶 mRNA 的 3' 端非翻译区来发挥调控功能, 造成靶 mRNA 的翻译抑制或降解。越来越多的研究表明 miRNA 具有十分重要的分子功能, 几乎在所有的生命过程中均扮演着重要角色。因此, 和 miRNA 有关的功能异常就可能和疾病的发生发展有密切关系。生物信息学旨在为解决生命科学问题提供信息学手段, 目前已经有多种用于研究 miRNA 和人类疾病关系的生物信息学方法和网络资源, 本文将综述这一主题的现状, 并探讨将来的发展趋势。

关键词 miRNA; 人类疾病; 生物信息学

中图分类号 R394

2.4-2.5、数据库与相关综述

介绍相当于小型综述

Hindawi
BioMed Research International
Volume 2021, Article ID 6652948, 16 pages
<https://doi.org/10.1155/2021/6652948>

Research Article

Prediction of miRNA-Disease Association Using Deep Collaborative Filtering

Li Wang¹ and Cheng Zhong²

¹School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China

²School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China

Correspondence should be addressed to Cheng Zhong; chzhong@gxu.edu.cn

Received 1 December 2020; Revised 1 February 2021; Accepted 10 February 2021; Published 24 February 2021

Academic Editor: Stefano Pascarella

Copyright © 2021 Li Wang and Cheng Zhong. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The existing studies have shown that miRNAs are related to human diseases by regulating gene expression. Identifying miRNA association with diseases will contribute to diagnosis, treatment, and prognosis of diseases. The experimental identification of miRNA-disease associations is time-consuming, tremendously expensive, and of high-failure rate. In recent years, many researchers predicted potential associations between miRNAs and diseases by computational approaches. In this paper, we

1. Introduction

miRNAMicroRNAs (miRNAs) are short endogenous non-coding RNAs with about 22 nucleotides. A number of studies have shown that miRNAs play important roles in many biological processes including cell proliferation, development, differentiation, death, apoptosis, metabolism, aging, signal transduction, and viral infection [1–6]. Biological studies have revealed that dysregulation of miRNAs is closely related to the occurrence and development of complex diseases [7–9]. Dysregulation of miR-15 and miR-16 was discovered to be related with B-cell chronic lymphocytic leukemia firstly [10]. So far, it has been verified that many miRNAs are related to cancers. Five members of the miRNA-200 family (miR-200a, miR-200b, miR-200c, miR-141, and miR-429) are downregulated in the development of breast cancer [11]. Epigenetic modulation of the miR-200 family relates to transition to a breast cancer stem cell-like state [12]. Some

studies demonstrate that in human colorectal cancer cells, miR-186, miR-216b, miR-337-3p, and miR-760 could work in synergy to induce cellular senescence by targeting the alpha subunit of protein kinase CKII [13]. By accurately measuring expression levels of miRNAs in the serum of 220 patients with early-stage non-small cell lung cancer and 220 matched controls, researchers found that the expressions of miR-27a, miR-106a, miR-221, miR-146b, miR-155, miR-17-5p, and let-7 were lower than those in controls, while the expression of miR-29c was increased [14].

Identifying the miRNAs associated with diseases will contribute to exploring the pathogenesis, diagnosis, treatment, and prognosis of diseases and help to develop new drugs. Some studies showed that miRNA-23, miRNA-24, and miRNA-27 contained underlying therapeutic factors in ischemic heart and vascular disease [15]. By targeting the BCL6 corepressor such as BCORL1, the migration and invasion of hepatocellular carcinoma (HCC) cells are restrained

by miR-876-5p, which provides a new idea for the treatment of HCC [16]. However, the experimental methods for finding associations between miRNAs and diseases are expensive and time-consuming. The computational methods for predicting potential miRNA-disease associations can provide verifiable hypotheses for further experimental verification, which can reduce biological experiment time and improve the experimental efficiency.

Recently, plenty of computational methods have been proposed to predict potential miRNA-disease associations [17]. Most of the computational methods are based on the assumption that miRNAs with similar functions are more likely to be associated with phenotypically similar diseases and vice versa. These methods are based on different principles to predict miRNA-disease associations, such as similarity-based methods, machine learning-based methods, and matrix factorization-based methods.

The previous similarity-based computational methods were based on miRNA-target interaction network and protein-protein interaction (PPI). For example, Jiang et al. [18] proposed a method to predict potential miRNA-disease associations by applying a scoring system to human phenome-microRNAome network and functionally related miRNA network. Shi et al. [19] developed a computational framework to identify miRNA-disease associations by performing random walk with restart. The method utilized the function connections between miRNA targets and disease genes in protein-protein interaction (PPI) networks. Mrk et al. [20] presented a miRNA-protein-disease association

rhythm, other global network-based methods were proposed. For example, Chen et al. [28] developed the model for miRNA-disease association prediction (WBSMDA) by utilizing within score and between score. The within-score can capture miRNA similarity and disease similarity in known miRNA disease pairs, and the between-score can capture miRNA similarity and disease similarity in unknown miRNA-disease pairs. Next year, Chen et al. [29] proposed a computational model based on super-disease and miRNA for potential miRNA-disease association (SDMMDA) prediction. You et al. [30] proposed a path-based miRNA-disease association (PBMDA) prediction model. PBMDA adopted depth-first search algorithm on a heterogeneous graph. Zeng et al. [31] applied link prediction algorithm named structural perturbation method (SPM) on the miRNA-disease bilayer network to predict potential miRNA-disease associations. Chen et al. [32] proposed a computational model of bipartite network projection for miRNA-disease association (BNPMDA) prediction. The model took advantage of the agglomerative hierarchical clustering and improved the baseline algorithm of bipartite network recommendation based on the constructed bias ratings. In addition, some researcher utilized lncRNA-related other information to predict potential miRNA-disease associations. Chen et al. [33] developed a triple layer heterogeneous network miRNA-disease association (TLHNMDA) prediction model. In the model, the triple layer network was constructed by integrating the known miRNA-disease associations, miRNA-lncRNA interactions, miRNA

2.4-2.5、数据库与相关综述

介绍相当于小型综述

Research Article

miRNA-Disease Association Prediction with Collaborative Matrix Factorization

Zhen Shen,¹ You-Hua Zhang,² Kyungsook Han,³ Asoke K. Nandi,⁴ Barry Honig,⁵ and De-Shuang Huang¹

¹Institute of Machine Learning and Systems Biology, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

²School of Information and Computer, Anhui Agricultural University, Changjiang West Road 130, Hefei, Anhui, China

³Department of Computer Science and Engineering, Inha University, Incheon, Republic of Korea

⁴Department of Electronic and Computer Engineering, Brunel University London, Uxbridge UB8 3PH, UK

⁵Center for Computational Biology and Bioinformatics, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA

Correspondence should be addressed to De-Shuang Huang; dshuang@tongji.edu.cn

Received 31 March 2017; Accepted 2 May 2017; Published 28 September 2017

Academic Editor: Fang-Xiang Wu

Copyright © 2017 Zhen Shen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As one of the factors in the noncoding RNA family, microRNAs (miRNAs) are involved in the development and progression of various complex diseases. Experimental identification of miRNA-disease association is expensive and time-consuming. Therefore, it is necessary to design efficient algorithms to identify novel miRNA-disease association. In this paper, we developed the computational method of Collaborative Matrix Factorization for miRNA-Disease Association prediction (CMFMDA) to identify potential miRNA-disease associations by integrating miRNA functional similarity, disease semantic similarity, and experimentally verified miRNA-disease associations. Experiments verified that CMFMDA achieves intended purpose and application values with its short consuming-time and high prediction accuracy. In addition, we used CMFMDA on Esophageal Neoplasms and Kidney Neoplasms to reveal their potential related miRNAs. As a result, 84% and 82% of top 50 predicted miRNA-disease pairs for these two diseases were confirmed by experiment. Not only this, but also CMFMDA could be applied to new diseases and new miRNAs without any known associations, which overcome the defects of many previous computational methods.

1. Introduction

MicroRNAs (miRNAs) are a class of short noncoding RNAs (19~25 nt), which normally regulate gene expression and protein production by targeting messenger RNAs (mRNAs) at the posttranscriptional level [1–9]. Since the first two miRNA lin-4 and let-7 were found in 1993 and 2000 [10, 11], thousands of miRNAs have been detected in eukaryotic organisms ranging from nematodes to humans. The latest version of miRBase contains 26845 entries and more than 2000 miRNAs have been detected in human [12–14]. With the development of bioinformatics and the progress of miRNA-related projects, researches are gradually focused on the function of miRNAs. Existing studies have shown that miRNAs are involved in many important biological processes

[15, 16], like cell differentiation [17], proliferation [18], signal transduction [19], viral infection [20], and so on. Therefore, it is easy to find that miRNAs have close relationship with various human complex diseases [12, 21–26]. For example, researchers found that mir-433 is upregulated in gastric carcinoma by regulating the expression of GRB2, which is a known tumour-associated protein [27]. Mir-126 can not only function as an inhibitor to suppress the growth of colorectal cancer cells by its overexpression, but also can help to differentiate between malignant and normal colorectal tissue [28]. Besides, the change of mir-17~92 miRNA cluster expression has close relationship with kidney cyst growth in polycystic kidney disease [29]. Considering the close relationship between miRNA and disease, we should try all means to excavate all latent associations between miRNA

and disease and to facilitate the diagnose, prevention, and treatment human complex disease [30–33]. However, using experimental methods to identify miRNA-disease association is expensive and time-consuming. As the miRNA-related theories are becoming more and more common, such as the prediction model about miRNA and disease, the function of miRNA in biological processes, and signaling pathways, new therapies are urgently needed for the treatment of complex disease; it is necessary to develop powerful computational methods to reveal potential miRNA-disease associations [12, 15, 20, 34–40].

Previous studies had shown that functionally similar miRNAs always appear in similar diseases; therefore many computational models were proposed to identify novel miRNA-disease associations [13, 41–46]. For example, Jiang et al. [31] analyzed and improved disease-gene prediction model, introduced the principle of hypergeometric distribution and how to use it, and discussed its application in prediction model and its actual effect. In order to realize the prediction function of the improved model, they used different types of dataset including miRNA functional similarity data, disease phenotype similarity data, and the known human disease-miRNA association data. Therefore, the prediction accuracy of this method is greatly impacted by miRNA neighbor information and miRNA-target interaction prediction. Chen et al. [47] reported a new method HGIMDA to identify novel miRNA-disease association by using heterogeneous graph inference. This algorithm can get better prediction accuracy by integrating known miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel

interactions. Random walks theory was applied on miRNA similarity network and EF similarity network. In addition, drug chemical structure similarity, miRNA function similarity, and networked-based similarity were also used in miREFRWR. Based on these biological datasets and efficient calculation method, miREFRWR could be an effective tool in computational biology. What is more, Chen et al. [52] also proposed a computational model RKNNMDA to predict the potential associations between miRNA and disease. Four biological datasets, experimentally verified human miRNA-disease associations, miRNA functional similarity, disease semantic similarity, and Gaussian interaction profile kernel similarity for miRNAs and diseases were integrated into RKNNMDA. It can be found that the prediction accuracy of RKNNMDA is excellent. Moreover, RKNNMDA could be applied for new diseases which do not have any known related miRNA information.

Generally speaking, current prediction model on miRNA-disease association is still demonstrating some shortcomings. For example, unreliable datasets have a great influence on the accuracy of prediction model, such as miRNA-target interactions and disease-genes associations. In addition, for miRNAs and diseases which do not have any known associations, we cannot use some of the existing models to predict its relevant information. In other words, we need to design and develop a new effective computational model. According to the assumption that functionally similar miRNAs always appear in similar diseases, we introduce the model of Collaborative Matrix Factorization for MiRNA-Disease Association prediction (CMFMDA) to reveal novel miRNA-disease association by integrating experimentally



TWAS代码部分

数据读取

```
1 library(data.table)
2
3 # 设置文件路径
4 expr_file_path <- 'F:/TWAS方法开发/expression_and_phenotype/AllStages_Allgenes_FPKM_NCBIID.txt'
5 phen_file_path <- 'F:/TWAS方法开发/expression_and_phenotype/FE_NCBIID_Phe.txt'
6
7 # 读取表达量文件为data.table
8 expr_data <- fread(expr_file_path, na.strings = c("", "NA", "NAN", "nan", "null"))
9
10 # 使用grep函数来获取"ODS"开头的列名
11 col_names_ODS <- names(expr_data)[grep("^ODS", names(expr_data))]
12
13 # 检查是否有提取出数据
14 if (length(col_names_ODS) == 0) {
15   stop("没有找到以 'ODS' 开头的列名，请检查列名是否正确。")
16 }
17
18 # 根据获取到的列名选择数据
19 expr_ODS <- expr_data[, ..col_names_ODS]
20
21 # 提取"ODS"开头的列名
22 col_names_ODS <- grep("^ODS", names(expr_data), value = TRUE)
23
24 # 读取表型文件为data.table
25 phen_data <- fread(phen_file_path, na.strings = c("", "NA", "NAN", "nan", "null"))
26
27 # 提取mean列并转换为数值型
28 phen_data[, mean := as.numeric(mean)]
29
30 # 如果mean列中存在不能转换为数值的数据，会变成NA
31 if (any(is.na(phen_data$mean))) {
32   stop("mean列包含无法转换为数值型的数据。")
33 }
34 # 检查表型数据行数是否为376
35 if (nrow(phen_data) != 376) {
36   stop("表型数据的行数不是376。")
37 }
```





```
# 转置expr_ODS以匹配pheno_data的样本标识符
expr_ODS_transposed <- t(expr_ODS)

# 首先，确保您的phen_data数据框确实有1到376行
n <- nrow(phen_data)
if(n != 376) {
  stop("phen_data does not have 376 rows. It has ", n, " rows.")
}

# 创建从S001到S376的新行名
new_row_names <- sprintf("ODS%03d", 1:376)

# 设置phen_data的新行名
rownames(phen_data) <- new_row_names

# 打印出前几个新行名以确认更改
print(head(rownames(phen_data)))
```

Data	
▶ expr_ODS	70199 obs. of 370 variables 
▶ expr_ODS_transposed	Large matrix (25973630 elements, 207.8 MB) 
▶ expr_data	70199 obs. of 2216 variables 
▶ phen_data	376 obs. of 6 variables 
Values	
col_names_ODS	chr [1:370] "ODS001" "ODS002" "ODS003" "ODS004" "ODS005" ...
expr_file_path	"F:/TWAS方法开发/expression_and_phenotype/AllStages_Allgen...
mean_column	num [1:376] 6.6 6.45 6.93 6.8 6.77 ...
n	376L
new_row_names	chr [1:376] "ODS001" "ODS002" "ODS003" "ODS004" "ODS005" ...
phen_file_path	"F:/TWAS方法开发/expression_and_phenotype/FE_NCBIID_Phe.tx...

数据分析

	SampleID	Rep1	Rep2	Rep3	mean	BLUP
1	S001	6.6	6.5	6.7	6.600000	6.605014
2	S002	6.4	6.5	NaN	6.450000	6.483359
3	S003	7.0	6.8	7.0	6.933333	6.889421
4	S004	6.8	6.7	6.9	6.800000	6.775658
5	S005	6.9	6.7	6.7	6.766667	6.747218
6	S006	6.8	6.8	6.9	6.833333	6.804099
7	S007	6.5	6.6	NaN	6.550000	6.562847
8	S008	6.8	6.7	6.8	6.766667	6.747218
9	S009	6.5	6.7	6.6	6.600000	6.605014
10	S010	NaN	NaN	7.0	7.000000	6.882793
11	S011	6.7	6.8	6.8	6.766667	6.747218
12	S012	6.6	6.5	6.5	6.533333	6.548133
13	S013	NaN	6.7	NaN	6.700000	6.679274
14	S014	6.6	6.6	6.8	6.666667	6.661895
15	S015	6.8	6.9	6.8	6.833333	6.804099

Showing 1 to 16 of 376 entries, 6 total columns

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
ODS001	0.0617	3.1652	0.0261	0.8334	8.4962	0.0000	3.5864	0.0535	0.0000	45.9388	0.6941	1.0596	27.3389	0.2702	
ODS002	0.1223	5.9424	0.0000	0.9746	9.0187	0.0000	4.4120	0.0000	0.0000	53.8060	0.3785	0.5654	31.2462	0.2416	
ODS003	0.2396	5.0082	0.0292	0.8719	8.0230	0.0000	3.5904	0.0000	0.0851	48.1290	0.3924	1.2925	22.5972	0.0000	
ODS004	0.3506	7.1412	0.0000	0.7270	5.2719	0.0248	4.6014	0.0000	0.0605	46.1642	0.7882	1.1359	27.9237	0.0000	
ODS005	0.0907	3.5829	0.0000	0.7646	8.6244	0.0000	4.7863	0.0000	0.0437	50.3054	0.6021	0.6724	23.2548	0.1033	
ODS006	0.2560	5.1769	0.0000	0.3621	11.3256	0.0000	4.1799	0.0000	0.0000	43.2988	0.7867	1.1009	25.4382	0.0000	
ODS007	0.3164	4.3684	0.0000	0.6385	6.6494	0.0000	4.6704	0.0804	0.1226	55.4951	0.3538	0.4479	23.8417	0.0000	
ODS008	0.3553	3.8395	0.0233	0.7301	9.7629	0.0000	4.2024	0.0000	0.0810	51.1850	0.4529	0.5226	26.7844	0.0000	
ODS009	0.1229	6.7203	0.0000	0.7170	8.0486	0.0000	3.8095	0.0505	0.0722	43.5789	0.5150	0.6775	27.5284	0.0000	
ODS010	0.3762	4.1115	0.0266	0.7616	9.4191	0.0000	6.1682	0.0000	0.0000	50.5262	0.7856	1.2906	22.1719	0.2321	
ODS011	0.5387	5.7613	0.0000	0.9424	6.5250	0.0000	3.9631	0.0000	0.0000	50.2526	0.0704	0.6895	25.8500	0.0000	
ODS012	0.1134	4.0235	0.0000	0.9272	6.6549	0.0000	4.2123	0.0000	0.1305	49.6275	0.6473	1.1171	30.0516	0.0000	
ODS013	0.2941	2.4235	0.0000	0.8817	7.1848	0.0000	5.8073	0.0644	0.0000	53.9929	1.0042	0.6164	26.2747	0.0000	
ODS014	0.2336	3.5648	0.0000	0.8981	9.3168	0.1253	3.4415	0.0000	0.0000	59.2294	0.6273	0.7597	25.1280	0.2366	
ODS015	0.1595	4.1920	0.0000	0.9978	7.3755	0.0000	4.7298	0.0502	0.0000	48.9715	0.7284	1.1963	26.7848	0.0000	

Showing 1 to 15 of 370 entries, 70199 total columns

获取两个数据集的行名

```
rownames_mean_column <- rownames(phen_data) # mean_column的行名应该与phen_data一致
rownames_expr_ODS_transposed <- rownames(expr_ODS_transposed)
```

找出两者共同的行名

```
common_rownames <- intersect(rownames_mean_column, rownames_expr_ODS_transposed)
```

根据共有的行名提取每个数据集中的相应行

```
mean_column_aligned <- mean_column[common_rownames]
expr_ODS_transposed_aligned <- expr_ODS_transposed[common_rownames, ]
```

现在mean_column_aligned和expr_ODS_transposed_aligned中的行是对应的，可以进行分析

计算相关性

```
correlation_results <- cor(mean_column_aligned, expr_ODS_transposed_aligned)
```

查看相关性结果

```
print(correlation_results)
```

数据分析

```
# 获取两个数据集的行名
```

```
rownames_mean_column <- rownames(phen_data) # mean_column的行名应该与phen_data一致
rownames_expr_ODS_transposed <- rownames(expr_ODS_transposed)
```

```
# 找出两者共同的行名
```

```
common_rownames <- intersect(rownames_mean_column, rownames_expr_ODS_transposed)
```

```
# 根据共有的行名提取每个数据集中的相应行
```

```
mean_column_aligned <- mean_column[common_rownames]
expr_ODS_transposed_aligned <- expr_ODS_transposed[common_rownames, ]
```

```
# 现在mean_column_aligned和expr_ODS_transposed_aligned中的行是对应的, 可以进行分析
```

```
# 计算相关性
```

```
correlation_results <- cor(mean_column_aligned, expr_ODS_transposed_aligned)
```

```
# 查看相关性结果
```

```
print(correlation_results)
```

The screenshot displays the RStudio environment. The top pane shows a data table with columns V1 through V15 and rows ODS001 through ODS015. The bottom pane shows the R console with the following code:

```
R 4.3.3 ~ />
rownames_mean_column <- rownames(phen_data) # mean_column的行名应该与phen_data一致
rownames_expr_ODS_transposed <- rownames(expr_ODS_transposed)


# 找出两者共同的行名
common_rownames <- intersect(rownames_mean_column, rownames_expr_ODS_transposed)

# 根据共有的行名提取每个数据集中的相应行
mean_column_aligned <- mean_column[common_rownames]
expr_ODS_transposed_aligned <- expr_ODS_transposed[common_rownames, ]

# 现在mean_column_aligned和expr_ODS_transposed_aligned中的行是对应的, 可以进行分析

# 计算相关性
correlation_results <- cor(mean_column_aligned, expr_ODS_transposed_aligned)

# 查看相关性结果
print(correlation_results)
```

- 
- **1. 非深度学习方法宏观上大致都了解了，如果有需要可在看具体的文献**
 - **2. TWAS方法开发的代码遇到瓶颈。**
 - 如何处理，处理什么数据间的关系
 - 代码的输出结果应该是什么样
 - 数据分析各类方法的区别和代码调用

总结

完毕



[https://github.com/yimengzhiyan/
bioinformatics_group_lbz](https://github.com/yimengzhiyan/bioinformatics_group_lbz)