doi: 10.12068/j.issn.1005-3026.2018.03.005

基于功能网络信息传播预测疾病-miRNAs的关联

李建华, 雒士源, 张建营, 康 雁
(东北大学 中荷生物医学与信息工程学院, 辽宁 沈阳 110169)

摘 要: 为了快速发现与疾病关联的miRNA, 基于功能网络信息传播,提出PMBP算法用于改进随机游走法,使用留一交叉验证评估了算法性能,最后进行案例分析. 实验结果表明:对于尚未发现关联miRNA的疾病,随机游走法是失效的,而PMBP以疾病相似性作为先验信息,能够有效预测;对于已经关联miRNA的疾病,PMBP提高了预测性能,AUC值为0.866. 对乳腺癌进行案例分析,预测的前50个miRNAs都被证实与乳腺癌相关,体现了PMBP算法的有效性.

关 键 词: 功能网络;疾病网络;网络传播;随机游走;miRNA预测
中图分类号: Q 811.4 **文献标志码**: A **文章编号**: 1005-3026(2018)03-0325-05

Prediction of Disease-Related miRNAs via Functional Network Information Propagation

LI Jian-hua, *LUO Shi-yuan*, *ZHANG Jian-ying*, *KANG Yan*
(School of Sino-Dutch Biomedical & Information Engineering, Northeastern University, Shenyang 110169, China. Corresponding author: KANG Yan, professor, E-mail: kangyan@bmie.neu.edu.cn)

Abstract: In order to quickly find out disease-related miRNAs, PMBP algorithm was proposed for improving random walk based on functional network information propagation. Leave-one-out cross validation was utilized to evaluate the performance of the algorithm and finally a case was analyzed. The results showed that random walk is ineffective for diseases that have not yet been associated with miRNAs, but the miRNA can be effectively predicted by using disease similarities as prior information. For the diseases known to be related with miRNAs, PMPB achieves a better performance and the corresponding AUC value is 0.866. In the case study of breast cancer, the predicted top 50 miRNAs are confirmed to be associated with breast cancer, which indicates the validity of PMBP.

Key words: functional network; disease network; network propagation; random walk; miRNA prediction

miRNA是不编码蛋白质的小分子RNA,它对标靶mRNA的表达进行调控,进而实现对基因的调控. 研究表明,miRNA几乎涉及动物的所有发育和病理过程,人类疾病特别是癌症的发生与miRNA失调密切相关[1]. 采用计算方法对可能与疾病相关的miRNA进行筛选,然后在实验室中验证,可以节省大量实验成本.

目前,多种方法已应用到疾病-miRNA关联预测[2-3],如机器学习方法,包括支持向量机[4-5]、最小二乘法[6]、玻尔兹曼机[7]等,以及基于相似性的方法[8-9]. 随机游走法在基因预测中优势明显[10],在miRNA预测中得到普遍应用[11-13]. 以上方法各有优缺点,机器学习方法易于集成多种生物数据,不足是假设的阴性样本中可能存在阳性样本,引起训练及预测误差. 基于相似性的方法较好地体现了功能相似的miRNA与表型相似的疾病关联,但是相似性度量及多种数据相似性集成需要更深入地研究. 随机游走法易于理解,预测精度较高,但只能对已关联miRNA

收稿日期: 2016-10-20
基金项目: 国家自然科学基金资助项目(61372014).
作者简介: 李建华(1973-),男,河北怀来人,东北大学讲师, 博士; 康 雁(1964-),男,辽宁沈阳人,东北大学教授,博士生导师.

的疾病进行预测.

本文对随机游走方法进行了改进,提出的 PMBP (prioritizing disease miRNA based on PRINCE)算法,既提高传统随机游走方法的预测性能,也能够对尚未发现关联 miRNA 的疾病进行预测. 实验表明,与经典的随机游走法 RWRMDA[12] 及 Chen 的相似性方法[8] 相比,PMBP 性能更好.

1 实验材料和实验方法

1.1 实验数据

本研究基于疾病 – miRNA 关联网络、疾病表型相似网络及 miRNA 功能相似网络进行疾病 – miRNA 关联预测.

文献[8]报道了 242 个实验证实的疾病 – miRNA 关联数据,涉及 51 种疾病和 99 个 miRNA. 其中,35 种疾病与至少 2 个 miRNA 关联. 疾病表型相似网络采用 mimMiner[14] 和 resnikHPO[15],并将后者转化为对称网络,利用 Tanimoto 方法实现数据归一化[16]. miRNA 功能相似网络采用由 271 个 miRNA 构成的网络 MISIM[17].

在疾病致病基因预测中,表型相似数据经过逻辑回归处理[16,18],提高了预测精度,本文也采用相同的处理方法. 逻辑回归方法如下:

$$L(x) = [1 + 10^{(cx+d)}]^{-1}. \quad (1)$$

其中:$d = \lg(9\,999)$;对于 resnikHPO,$c = -17$,对于 mimMiner,$c = -15$.

1.2 基于功能网络传播的预测算法

Vanunu 等提出 PRINCE 算法,成功应用于疾病的基因预测[18]. 借鉴该算法中的网络信息传播,本文提出 PMBP 算法用于疾病 – miRNA 关联预测.

设 miRNA 功能网络为 $G = (V, E, w)$,其中 V 是 miRNA 集合,E 是 miRNA 之间相互关联的边集合,w 是边的权值,表示关联的程度. 已知与疾病 d_i 关联的所有 miRNA 定义为种子集合 s,PMBP 通过网络信息传播,获得 V 中各 miRNA 与 d_i 关联的或然值,移除种子集 s 后,将 V 中剩余的 miRNA 按或然值大小依次排列,从而得到对 d_i 的预测.

上述过程可用迭代公式(2)描述:

$$\boldsymbol{F}^t = \alpha \times W_{\mathrm{NP}} \times \boldsymbol{F}^{t-1} + (1-\alpha) \times Y. \quad (2)$$

其中:Y 是先验信息,即已知的疾病 – miRNA 关联;α 是调整先验信息和上次迭代结果的权重值,范围在(0,1)之间;W_{NP}是按照文献[18]提供的方法变换后的 miRNA 功能网络. 终止条件定义为 $|\boldsymbol{F}^t - \boldsymbol{F}^{t-1}| \leqslant 10^{-9}$或迭代次数 $N \leqslant 100$.

本研究采用留一交叉验证,即如果疾病 d_i 与至少 2 个 miRNA 关联,则将其中 1 个关联移除,通过剩余的种子集合,预测被移除的关联. 如果疾病 d_i 仅与 1 个 miRNA 关联,则将此关联移除后,因没有种子作为先验信息,传统的随机游走算法失效,本文提出利用疾病表型相似性作为先验信息完成预测.

1.3 实验过程

本研究中,miRNA 预测分为三步.

步骤 1 提取待预测疾病 d_i 与 miRNA 关联的先验信息 Y. 如果疾病已与若干 miRNA 关联,Y 中相应分量为 1,而无关 miRNA 的相应位置为 0. 如果 d_i 尚未发现关联的 miRNA,则根据疾病表型相似网络,获得该疾病与其他各疾病的相似向量 $\boldsymbol{S}$,并根据其他疾病与 miRNA 的关联推测 d_i 的先验信息 Y. 当多个疾病与同 1 个 miRNA 关联时,取其中的最大值.

步骤 2 计算疾病 d_i 与各 miRNA 关联的向量 $\boldsymbol{F}$.

在 miRNA 功能网络中通过式(2)实现信息传播,不断迭代直至满足终止条件,此时向量 $\boldsymbol{F}$ 中保存了各个 miRNA 与 d_i 相关联的预测值.

步骤 3 获得疾病 d_i 的预测结果.

将 miRNA 根据 $\boldsymbol{F}$ 中的值从大到小排列. 如果疾病 d_i 有 k 个种子,它们将排在前 k 位,移除这些种子后,最终得到按可能性大小依次排列的 miRNA.

1.4 评估标准

在留一交叉验证中,采用三种指标评估算法,即平均排位率、ROC 曲线下的面积 AUC、Top N 中的真阳率.

平均排位率定义为被预测的 miRNA 在所有候选中的平均排位. 它衡量了算法的总体性能,其值越低表明算法预测性能越高.

AUC 越大,预测性能越好. 在排序后的 miRNA 中,设定一个阈值 τ,如果疾病的 miRNA 排位在 τ 之上(包含 τ),标记其为正确预测的阳性样本,即真阳性 TP;如果被排在 τ 之下,标记为错误预测的阴性样本,即假阴性 FN. 变化 τ 值,可以得到相应的 TPR 和 FPR,从而绘制 ROC 曲线以及计算 AUC.

Top N 中的真阳率是前 N 个中被正确验证的与疾病关联的 miRNA 所占的比例. 它反映了在

不同范围内算法的预测精度，其中 N 分别为 1，5，10，20，30，50，100. 当 $N=1$ 时，真阳率也称为置顶率，它反映了算法的精确预测能力.

2　结果与讨论

2.1　以种子 miRNA 作为先验信息的预测

35 种疾病至少关联 2 个 miRNA，疾病 – miRNA 关联总数是 226，利用 PMBP 和 RWRMDA 分别进行留一交叉验证. 随机游走算法 RWRMDA 的定义为

$$P^t = (1-\gamma) \times \boldsymbol{W}_{\mathrm{RW}} \times P^{t-1} + \gamma \times P^0. \quad (3)$$

其中 $\boldsymbol{W}_{\mathrm{RW}}$ 是对功能矩阵的列向量进行归一化的结果. 因式(2)和式(3)形式类似，为了讨论方便，将 $\beta = 1-\alpha$ 代入式(2)，并将替换后的 β 与式(3)中的 γ 统称为平衡因子. 平衡因子对预测结果具有一定影响，一般大于 0.5 时结果较好. 实验中，当平衡因子为 0.9 时，PMBP 和 RWRMDA 达到最好结果，两种算法预测的 ROC 曲线如图 1 所示，对应的 AUC 分别是 0.866 和 0.848， 平均排位率分别是 13.8% 和 15.2% . Top N 中的真阳率如图 2 所示. 明显地，PMBP 预测结果较 RWRMDA 更好，而在 Top 10 和 Top 50 之间，性能优势更加显著.

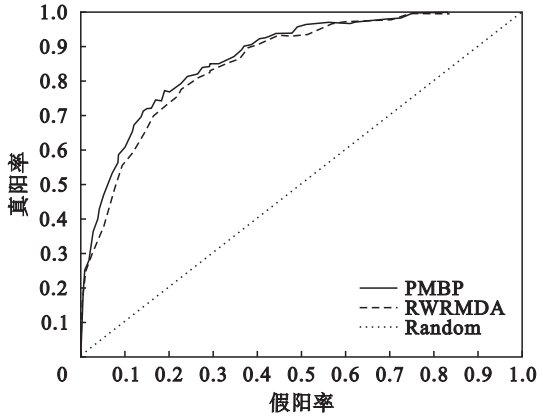

图 1　PMBP 和 RWRMDA 的 ROC 曲线
Fig. 1　ROC curves for PMBP and RWRMDA

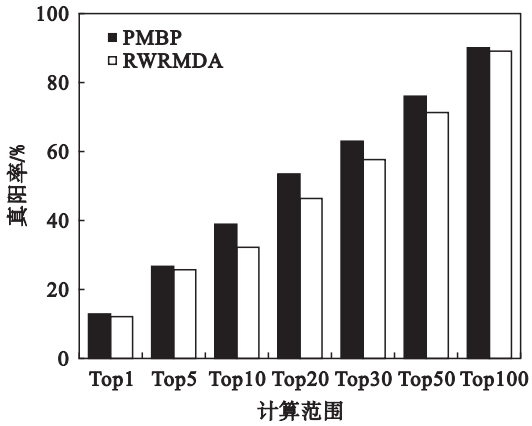

图 2　PMBP 和 RWRMDA 预测 Top N 中的真阳率
Fig. 2　TPR in the Top N for PMBP and RWRMDA

2.2　以疾病表型相似性作为先验信息的预测

RWRMDA 不能对只与 1 个 miRNA 关联的疾病进行留一交叉验证. PMBP 可利用 mimMiner 或 resnikHPO 提供的疾病相似性预测 miRNA.

当采用 mimMiner 时，从中提取 51 种疾病构成相似网络，对所有 16 个具有单一 miRNA 的疾病进行验证，当平衡因子为 0.6 时，得到最低平均排序率 23.3% 以及最大 AUC 值 0.769. Top N 中的真阳率如表 1 所示.

当采用 resnikHPO 时，只能对 15 个具有单一 miRNA 的疾病进行验证，平衡因子取 0.6，平均排序率为 28.4% ，AUC 值是 0.718. 尽管上述指标不及 mimMiner，但是 resnikHPO 在 Top 20 和 Top 30 中预测的真阳率分别是 26.7% 和 33.3% ，优于表 1 中相应的结果. 换言之，与 mimMiner 相比，resnikHPO 将更多的正确预测排在前 30 位.

表 1　PMBP 基于疾病表型相似性在不同范围内预测的真阳率
Table 1　TPR in Top N predicted by PMBP based on disease phenotype similarity

计算范围	Top 1	Top 5	Top 10	Top 20	Top 30	Top 50	Top 100
真阳率/%	12.5	12.5	12.5	18.8	31.3	50.0	68.8

2.3　PMBP 与基于相似性的算法比较

Chen 等[8]基于全局网络相似性和疾病表型相似性报道了三种预测算法 PBSI，MBSI 和 NetCBI，应用这三种预测算法对本文中所有 51 种疾病进行验证. 表 2 列出了 PMBP 与三种算法的比较，其中在与多个 miRNA 关联的疾病验证中平衡因子取 0.9，在与 1 个 miRNA 关联的疾病验证中平衡因子取 0.6. 明显地，PMBP 优于三个算法中最好的 NetCBI.

表 2　PMBP 与 Chen 的三种方法的比较
Table 2　Comparison of PMBP with Chen's three methods

预测方法	PMBP	NetCBI	MBSI	PBSI
AUC	0.859	0.807	0.748	0.540

2.4　案例分析

为了展示 PMBP 预测的有效性，对乳腺癌进

行预测分析. 在本数据集中,已发现 27 种 miRNA 与乳腺癌相关联. 以这些 miRNA 作为种子,利用 PMBP 预测,在预测结果中选择前 50 个 miRNA 进行分析. 首先,在三个权威数据库 miR2Disease, PhenomiR 和 HMDD 中进行检索验证. 如果未能检索出,则在 Pubmed 文献数据库中进行验证. 表 3 列出所有 50 个 miRNA,证实它们都与乳腺癌相关,其中证据部分列出了支持的数据库以及相关文献的 Pubmed 编号.

表 3 PMBP 预测的前 50 个 miRNA 与乳腺癌相关联

Table 3 Top 50 breast cancer-related miRNAs predicted by PMBP

排序	miRNA	证据	排序	miRNA	证据
1	hsa – mir – 18a	miR2Disease, PhenomiR	26	hsa – mir – 146b	miR2Disease, HMDD
2	hsa – mir – 19a	PhenomiR, HMDD	27	hsa – mir – 151	PhenomiR
3	hsa – mir – 19b	PMID: 27602768, 27630665	28	hsa – mir – 223	PhenomiR, HMDD
4	hsa – mir – 145	miR2Disease, PhenomiR	29	hsa – mir – 92b	PMID: 25047087, 26878388
5	hsa – mir – 127	miR2Disease, PhenomiR	30	hsa – mir – 132	PhenomiR, HMDD
6	hsa – mir – 34a	HMDD	31	hsa – mir – 34b	PhenomiR, HMDD
7	hsa – let – 7d	miR2Disease, PhenomiR	32	hsa – mir – 135b	PhenomiR, HMDD
8	hsa – mir – 9	PMID: 25086633, 23617747	33	hsa – mir – 103	PMID: 24088786
9	hsa – mir – 25	PhenomiR, HMDD	34	hsa – mir – 181b	miR2Disease, HMDD
10	hsa – mir – 194	PMID: 22829924, 27221739	35	hsa – mir – 339	PhenomiR, HMDD
11	hsa – let – 7f	miR2Disease	36	hsa – mir – 101	miR2Disease
12	hsa – mir – 106b	PhenomiR	37	hsa – let – 7g	PhenomiR, HMDD
13	hsa – let – 7b	PhenomiR, HMDD	38	hsa – mir – 18b	HMDD
14	hsa – mir – 199a	PMID: 23504322, 25515522	39	hsa – mir – 191	PhenomiR, HMDD
15	hsa – mir – 125a	miR2Disease, HMDD	40	hsa – mir – 214	PhenomiR, HMDD
16	hsa – mir – 16	PMID: 26031775, 27157613	41	hsa – mir – 153	PMID: 27012032, 26392359
17	hsa – mir – 93	PhenomiR, HMDD	42	hsa – mir – 143	miR2Disease, PhenomiR
18	hsa – mir – 34c	HMDD	43	hsa – mir – 30d	PhenomiR, HMDD
19	hsa – let – 7c	PhenomiR, HMDD	44	hsa – mir – 15a	PhenomiR, HMDD
20	hsa – mir – 29b	miR2Disease	45	hsa – mir – 30c	miR2Disease
21	hsa – mir – 20b	HMDD	46	hsa – mir – 302b	PhenomiR
22	hsa – let – 7i	miR2Disease, PhenomiR	47	hsa – mir – 218	PMID: 25900794, 25394901
23	hsa – mir – 29a	PhenomiR, HMDD	48	hsa – mir – 1	PMID: 26275461, 26676637
24	hsa – let – 7e	miR2Disease	49	hsa – mir – 133a	PMID: 26107945, 25051376
25	hsa – mir – 92a	HMDD	50	hsa – mir – 219	PMID: 23813567

3 结 论

1) 基于功能网络传播的方法应用于疾病 miRNA 预测是可行的,本文提出的 PMBP 算法性能优于文献报道的 RWRMDA 和 NetCBI.

2) 对尚未发现关联 miRNA 的疾病,本文首次提出从疾病表型网络 mimMiner 和 resnikHPO 提取先验信息进行预测,前者的 AUC 优于后者,而后者倾向于将更多的正确预测排在前 30 位.

3) PMBP 算法的改进方向. 首先,由于数据集规模较小,未能系统比较两种疾病表型网络在预测中的异同,将来可通过提取可靠的、更大规模的数据集进行分析. 其次,如果进一步融合与 miRNA 相关的生物数据,如疾病 – 基因关联、miRNA – 基因关联,将有助于提高预测精度.

参考文献:

[1] Minju H, Narry K. Regulation of microRNA biogenesis [J]. *Nature Reviews Molecular Cell Biology*, 2014, 15(8): 509 – 524.

[2] Zeng X X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks [J]. *Briefings in Bioinformatics*, 2016, 17(2): 193 – 203.

[3] 张帆,崔庆华. MicroRNA 与人类疾病关系研究中的生物信息学方法和资源[J]. 生理科学进展, 2016, 47(3): 203 – 209.
(Zhang Fan, Cui Qing-hua. Bioinformatics methods and resources for the research on the relationship between MicroRNAs and human diseases [J]. *Progress in Physiological Sciences*, 2016, 47(3): 203 – 209.)

(下转第 344 页)