基于个性化随机游走的基因-表型关联分析

谭好江[1,2],王 峻[2],余国先[1],陈 建[3],郭茂祖[4]
(1. 山东大学软件学院,山东济南 250101;2. 山东大学人工智能国际联合研究院,山东济南 250101;3. 中国农业大学农学院,北京 100083;4. 北京建筑大学电气与信息工程学院,北京 100044)

摘 要: 基因与表型间的关联分析对揭示生物的内在遗传关联具有重要意义. 随机游走算法可以融合多组学数据,聚合一阶或高阶邻居的标签信息,对网络中不同节点间关联信息进行补全,提高关联预测的准确度,进而发现基因和表型间潜在的遗传关联. 但现有随机游走算法通常平等地对待每个节点,忽略了不同节点的重要性,使得非重要节点过度传播,降低了模型性能. 为此,本文提出了一种基于多组学数据融合的个性化随机游走算法(individual Multiple Random Walks, iMRW),在由基因、miRNA及表型节点构建的多组学异质网络上,基于网络拓扑结构,设计个性化多元随机游走策略,为不同重要程度的节点分配不同的游走步长,并结合高斯相互作用属性核相似性与随机游走,对网络不同节点及节点间关联信息进行补全,最终实现多源基因-表型关联矩阵的融合,准确获取基因-表型关联预测矩阵. 在不同实验设置下,与主流算法的对比实验结果均显示iMRW能够取得更优的预测性能. 在玉米光合作用能力和淀粉含量表型的实验分析结果也进一步证实了iMRW在识别潜在的基因-表型关联的实用性与有效性.

关键词: 基因-表型关联;随机游走;异质网络;多组学数据融合;网络拓扑结构
基 金 项 目: 国家自然科学基金项目(No. 62031003, No. 62072380);山东大学中央高校基本业务费(No.2020GN061)
文献标识码: A **文章编号:** 0372-2112(XXXX)XX-0001-14
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.xxxxxxxx *

Individual Random Walks for Gene-Phenotype Association Analysis

TAN Hao-jiang[1,2], WANG Jun[2], YU Guo-xian[1], CHEN Jian[3], GUO Mao-zu[4]
(1. *School of Software*, *Shandong University*, *Jinan*, *Shandong* 250101, *China*;
2. *Joint Centre for Artificial Intelligence Research*, *Shandong University*, *Jinan*, *Shandong* 250101, *China*;
3. *College of Agronomy and Biotechnology*, *China Agricultural University*, *Beijing* 100083, *China*;
4. *College of Electrical and Information Engineering*, *Beijing University of Civil Engineering and Architecture*, *Beijing* 100044, *China*)

Abstract: Association analysis between genes and phenotypes is crucial to reveal the inherent genetic association of organisms. Random walk-based algorithms can fuse multiple omics data, aggregate the label information of first-order or higher-order neighbors, complete the association information between different nodes in the network, improve the accuracy of association prediction and further discover the potential genetic associations between genes and phenotypes. However, existing random walk algorithms usually treat each node equally and ignore the varying importance of different nodes, as such non-important nodes can be excessively propagated and the model performance is compromised. To this end, an individual Multiple Random Walks (iMRW) algorithm based on multi-omics data fusion is proposed. On the heterogeneous genetic network composed with genes, miRNAs and phenotype nodes, we design the individual multiple random walks strategy based on the network topology, assign nodes of different importance with different walking lengths. We then complete the genetic information of different nodes by fusing multi-source association matrix, Gaussian interaction profile kernel similarity and random walk, and accurately obtain the gene-phenotype association prediction matrix. Under different experimental settings, iMRW can achieve the best prediction performance compared with the state-of-the-art algorithms. The case study with respect to maize photosynthetic ability and starch content further confirm the usefulness and effectiveness of iMRW in identifying potential gene-phenotype associations.

Key words: gene-phenotype associations; random walk; heterogeneous network; multi-omics data fusion; network

收稿日期:XXXX-XX-XX;修回日期:XXXX-XX-XX;责任编辑:某某 *投稿时不填写此项*

topology

1 引言

随着高通量测序技术的发展,融合基因组、转录组、蛋白质组、代谢组以及表型组等多组学数据的关联分析受到了广泛关注[1~3],也使得基因-表型关联分析成为了当前遗传学、病因学、育种学等生命科学领域研究的热点[4~7]. 基因组是生物个体全部基因组合的总称,反映了生物体的遗传构成;表型是指具有特定基因型的个体在一定环境条件下所表现出来的性状特征[8~10]. 基因-表型关联分析能够揭示基因与表型之间的内在遗传关联,为生命体性状发育与变异的底层遗传过程的解析提供指导.

传统基因-表型关联是通过生物实验进行验证,这类方法实验周期长、成本高,获取的基因-表型关联数量有限. 随着生物组学数据的增加以及系统生物学的不断发展,研究人员发现结构或功能相近的基因通常会参与相同的生物学通路,进而对表型的产生和变化形成相似的影响[11,12]. 受这类发现激发,研究者将基因-表型关联的研究方向转向了计算科学、统计科学与生命科学相结合的交叉研究[2,13~16],提出了大量的预测方法和模型. 这类基于计算分析的方法能够整合利用分子间遗传关联信息,指导探索导致表型发生的遗传作用机制,对揭示生物体的内在遗传机制具有良好的理论和应用价值.

2 相关工作

生物分子间以及生物分子与表型性状间存在复杂的遗传关联关系,因此,现有面向基因-表型关联分析的计算模型大多是基于遗传分子网络的构建和优化策略.

早期的基因-表型关联分析计算模型主要是基于单一组学数据信息进行设计的. 由于蛋白质互作可以用来刻画基因间相关性,Kohler等人[17]使用测量全局网络距离和随机游走的方法来定义蛋白质-蛋白质相互作用(Protein-Protein Interaction, PPI)网络,进而实现对疾病性状关联的候选基因进行优先级排序. PRINCE[18]使用基于全局网络的方法来推断蛋白质之间的复杂关联并确定与疾病表型性状关联的基因. 上述方法仅利用了PPI刻画的基因相似性网络来进行关联预测,忽略了表型相关性网络. 当已知的与疾病相关的基因较少时,仅依靠基因相似性很难进行大规模基因-表型关联预测. 为了克服以上方法的不足,研究者进一步提出了基于异质网络的方法[13,19,20]. 这类方法首先构建基因相似性网络、表型相关性网络以及基因-表型关联网络,再进行基因-表型关联分析. 然而,上述方法的基因相似性网络是基于PPI构建的,PPI数据普遍存在较多的噪声,为了减少这些噪声对基因-表型关联预测的影响,IDLP[21]将PPI网络的邻接矩阵作为学习变量,通过优化损失函数减少噪声的影响,提升了预测精度.

生物代谢是一个复杂的过程,通常受多个基因或不同层次(基因层、转录组层、蛋白质层、代谢组层等)的生物分子之间的相互作用的影响,上述基于单一组学数据的关联预测方法遗漏了其他组学的生物分子对遗传过程的影响,很难实现对生物系统的复杂调控过程的全面理解和描述[22]. 因此,基于多组学数据的表型相关性研究受到了越来越多的关注. Fu等人[23]对基因、lncRNA、miRNA、药物以及基因本体(Gene Ontology, GO)组成的异构数据源通过矩阵分解的方式来获取潜在的基因、lncRNA与疾病表型的关联;Chen等人[24]整合了lncRNA、miRNA以及疾病层面的数据构建了异质网络,在异质网络上使用标签传播[25]策略来预测miRNA和疾病的关联;Huang等人[26]使用多示例学习和融合多组学(基因组、转录组及表型组)数据来预测可变剪接异构体与疾病的关联. 然而,这些方法更多是从多组学的角度预测人类转录层分子与表型的关联,对于基因-表型的关联关注不足.

与人类基因-表型研究相比,对于植物,尤其是水稻、大豆、玉米等主要农作物,已知的基因-表型关联信息较少,面向作物基因-表型关联的计算遗传分析研究相对缺乏. 现有的作物基因-表型关联分析方法主要集中在基于单一组学(基因组)数据的分析策略. 随着表达谱分析技术和代谢组学的飞速发展,拟南芥、玉米和水稻等植物的多组学数据来源日益丰富[27,28],推动了基于多组学数据指导的作物遗传智能分析和育种研究进一步的发展[29]. Xu等人[30]分别利用基因组、转录组和代谢组学的数据在339个不同的玉米自交系中使用了8种方法预测了6个农艺性状的表型,并比较了不同方法的预测能力. 然而这一方法并没有将不同类型的组学数据进行整合,而是单独进行分析预测,本质仍然是一种单组学分析方法. Jiang等人[29]利用多组学数据分别构建加权网络,并将其整合到一个融合网络中,通过对融合网络的分析,挖掘出了一些在玉米发育过程中起关键作用的孤儿节点,并在此基础上利用基因、mRNA、蛋白质和表达谱数据集构建了相互作用网络,最终研究了与前100个候选基因及已知基因相关的基因本体术语,并分析了这些基因在决定玉米表型中的作用[31].

然而,现有基于多组学的关联分析方法主要是利用两种组学数据分子间的现有关联构建网络进行统计分析和挖掘,并未对多组学数据,尤其是玉米等作物多

组学数据实现真正意义的融合,对多种组学分子间隐含关联和分子间遗传互补信息缺乏有效挖掘和利用.其次,现有随机游走以及标签传播算法平等地对待网络中的每一个节点,为节点分配相同的随机游走步长,这些方法忽略了不同节点的重要性,使得非重要节点过度传播,限制了模型性能的提升. 为解决上述难题,在现有基因-表型研究工作基础上,以我国主要粮食作物—玉米为范例,本文提出了基于多组学数据融合的多元个性化随机游走算法(individual Multiple Random Walks, iMRW),通过对基因、miRNA及表型数据内部及数据间遗传关联信息的集成融合,实现对基因-表型关联的有效预测. iMRW首先融合分子节点标记信息和表型节点的层次结构信息,设计构建了由基因、miRNA及表型节点组成的异质分子网络,并根据网络的拓扑结构量化了每个节点随机游走的步幅长度;其次,根据每个节点的步幅长度,iMRW在构建的异质分子网络上通过多元个性化随机游走进行链路预测和补全,得到了基因-表型、基因-miRNA和miRNA-表型的关联;最后,iMRW通过关联预测矩阵决策融合策略,充分利用和整合了三种不同网络节点间遗传信息,实现了基因-表型关联预测. 实验结果证明,iMRW能够取得与现有主流算法相比更优的预测效果. 在玉米光合作用能力和淀粉含量相关的表型案例实验分析也进一步验证了iMRW在识别潜在的基因-表型关联方面的有效性和实用性.

3 材料和方法

3.1 异质遗传网络构建

为了综合利用多组学信息指导基因-表型关联预测,本文选择基因、miRNA及现有表型性状作为网络节点,利用同类节点间及不同类别节点间遗传关联性,构建包含 m 个基因,l 个miRNA以及 n 个表型本体(Trait Ontology, TO)的异质遗传网络 $\boldsymbol{W}$,如图1所示. 异质遗传网络通过其加权邻接矩阵 $\boldsymbol{W}\in\mathbb{R}^{(m+n+l)\times(m+n+l)}$ 表示:

$$\boldsymbol{W}=\begin{bmatrix}\boldsymbol{W}_{gg} & \boldsymbol{W}_{gt} & \boldsymbol{W}_{gm}\\ \boldsymbol{W}_{tg} & \boldsymbol{W}_{tt} & \boldsymbol{W}_{tm}\\ \boldsymbol{W}_{mg} & \boldsymbol{W}_{mt} & \boldsymbol{W}_{mm}\end{bmatrix} \tag{1}$$

其中,$\boldsymbol{W}_{gg}\in\mathbb{R}^{m\times m}$、$\boldsymbol{W}_{tt}\in\mathbb{R}^{n\times n}$ 和 $\boldsymbol{W}_{mm}\in\mathbb{R}^{l\times l}$ 分别表示基因、TO和miRNA层面的同质子网络. $\boldsymbol{W}_{gt}\in\mathbb{R}^{m\times n}$、$\boldsymbol{W}_{gm}\in\mathbb{R}^{m\times l}$ 和 $\boldsymbol{W}_{mt}\in\mathbb{R}^{l\times n}$ 对应基于已知基因-表型、基因-miRNA和miRNA-TO间遗传关联构建的异质子网络. $\boldsymbol{W}_{tg}$、$\boldsymbol{W}_{mg}$ 和 $\boldsymbol{W}_{tm}$ 分别是对应关联矩阵的转置矩阵. 下面详细介绍各同质/异质子网络的构建与邻接矩阵计算方法.

3.1.1 基因子网络

基因序列中包含多个功能位点,例如活性结合位点、表型相关的单核苷酸多态性(SNP)位点、单分子肽和基序等,这些功能位点虽然长度较短,但却包含着丰富的遗传信息,可以为基因-表型关联预测提供指导[32]. 因此,本文从MaizeGDB数据库中获取了12098条基因序列(版本号AGPv3.21),并使用BLAST工具计算获取基因序列相似度,再基于基因序列相似度构建了基因子网络 $\boldsymbol{W}_{gg}\in\mathbb{R}^{m\times m}$.

3.1.2 表型子网络

本文研究利用表型本体(Trait Ontology, TO)进行表型子网络构建. TO是描述植物表型性状的控制词汇[33]. 每一个性状都是一株正在发育或成熟的植物的可区分的特征或品质. TO通过有向无环图(Direct Acyclic Graph, DAG)结构化地组织. 在DAG中,每个节点对应一个TO术语,每条边对应一对TO术语之间的关系. TO遵循真路径规则[34]:子术语是其父术语功能的进一步完善,若一个基因被某TO术语进行了注释,那么该基因也被这一TO术语的祖先术语(若存在)进行注释;若一个基因未被某一TO术语进行注释,则该基因不会被这一TO术语的任何后代术语进行注释. 图2给出了玉米基因“GRMZM2G068699”的TO注释示例:“GRMZM2G068699”被“TO:0000696”标注. 根据真路径法则,“GRMZM2G068699”也被“TO:0000696”的祖先节点(“TO:0006007”,“TO:0000489”,“TO:0000291”,“TO:0000281”,“TO:0000277”和“TO:0000387”)标注. 表1给出了各个TO术语的含义.

为了综合利用基因的TO标注信息和TO的层次结构,本文使用以下公式[35]计算了父亲节点 t 和孩子节点 s 之间的过渡概率:

$$p(s|t)=\frac{n_s}{n_t}+\frac{IC(s)}{\sum_{s'\in ch(t)}IC(s')} \tag{2}$$

其中,$ch(t)$ 是节点 t 的孩子节点的集合,n_t 和 n_s 分别表示节点 t 和节点 s 所标注的基因数量. $IC(t)$ 是节点 t 的信息量,它的定义如下:

$$IC(t)=1-\frac{\log(1+|desc(t)|)}{\log n} \tag{3}$$

其中,$desc(t)$ 表示节点 t 以及它的所有后代的集合. 如果 $|desc(t)|$ 越大,那么节点 t 有更多的后代,这些后代比 t 含有更确切的表型信息,那么节点 t 的信息量就越少. 为了使父亲节点到其直接孩子节点的过渡概率和为1,本文使用 $\boldsymbol{W}_{tt}\in\mathbb{R}^{n\times n}$ 来表示TO节点之间的过渡概率:

$$\boldsymbol{W}_{tt}(t,s)=\frac{p(s|t)}{\sum_{v\in ch(t)}p(v|t)} \tag{4}$$

最终构建了TO层网络 $\boldsymbol{W}_{tt}$.

3.1.3 miRNA子网络

miRNA是一类与多种复杂生物过程相关的微小的

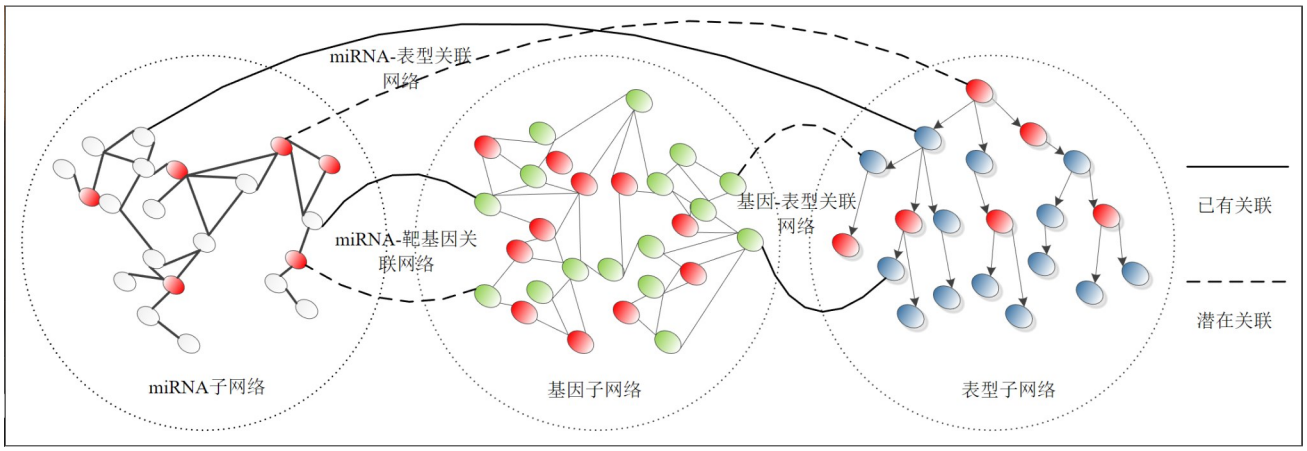


图1 由基因、miRNA 和 TO 节点组成的异质遗传网络

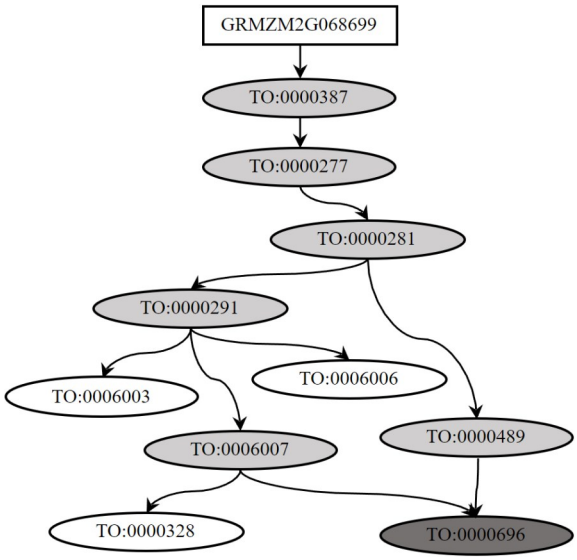


图2 玉米基因"GRMZM2G068699"的 TO 注释示例

表1 典型TO术语的含义结果

TO 术语	含义
TO:0000387	植物性状
TO:0000227	生化特性
TO:0000281	代谢物含量性状
TO:0000291	碳水化合物含量
TO:0006003	低聚糖含量
TO:0006006	单糖含量
TO:0006007	多糖含量
TO:0000489	碳水化合物组成相关性状
TO:0000328	蔗糖含量
TO:0000696	淀粉含量

非编码 RNA，其长度约为 22 个核苷酸[36~38]. 它通过靶向特异性 mRNA 调控基因表达，在多种生物学过程中发挥重要作用[39,40]. 本文从数据库 PmiREN[41,42] 上获取了所有 miRNA 的序列信息. 受 Liang 等人工作[43]的启发，本文根据其序列计算每对 miRNA 的相似性得分，并根据相似度构建了 miRNA 层网络 $\boldsymbol{W}_{mm}$.

3.1.4 miRNA-靶基因关联网络

miRNA 能够通过切断靶基因转录得到的 mRNA 分子，影响翻译，进而控制基因表达，因此，引入 miRNA 与靶基因之间的关联性能够为基因提供更加丰富的遗传信息，指导基因-表型关联预测. 本文先从玉米公开数据库 MaizeGDB 中得到了 miRNA 及其靶基因，再定义 $\boldsymbol{W}_{gm}$ 存储了已知的 m 个基因和 l 个 miRNA 之间的关联. 如果 miRNAj 对基因 i 有靶向作用，那么 $\boldsymbol{W}_{gm}(i,j)=1$，否则 $\boldsymbol{W}_{gm}(i,j)=0$.

3.1.5 基因-表型关联网络

本文从数据库[44]获取了已知的基因-表型关联信息，使用 $\boldsymbol{W}_{gt}$ 存储了已知的 m 个基因和 n 个 TO 之间的关联. 如果基因 i 和 TOj 有关联，那么 $\boldsymbol{W}_{gt}(i,j)=1$，否则 $\boldsymbol{W}_{gt}(i,j)=0$.

3.1.6 miRNA-表型关联网络

由于 miRNA 对基因具有靶向作用，因此，可以综合利用基因-表型关联网络 $\boldsymbol{W}_{gt}$ 和 miRNA-靶基因关联网络 $\boldsymbol{W}_{gm}$ 得到 miRNA-TO 关联网络：

$$\boldsymbol{W}_{mt}=\boldsymbol{W}_{gm}^{T}\cdot\boldsymbol{W}_{gt} \tag{5}$$

最终，本文基于上述子网络构建了异质遗传网络 $\boldsymbol{W}\in\mathbb{R}^{(m+n+l)\times(m+n+l)}$，表 2 列出了各子网络的相关信息.

3.2 个性化随机游走

基因-表型关联预测是一个多标签分类问题. 一个

表2 子网络相关信息

符号	定义	维度	关联数
$\boldsymbol{W}_{gg}\in\mathbb{R}^{m\times m}$	基因-基因	12,098×12,098	1,664,119
$\boldsymbol{W}_{tt}\in\mathbb{R}^{n\times n}$	TO-TO	155×155	161
$\boldsymbol{W}_{mm}\in\mathbb{R}^{l\times l}$	miRNA-miRNA	447×447	199,803
$\boldsymbol{W}_{gm}\in\mathbb{R}^{m\times l}$	基因-miRNA	12,098×447	2,087
$\boldsymbol{W}_{gt}\in\mathbb{R}^{m\times n}$	基因-表型	12,098×155	120,444
$\boldsymbol{W}_{mt}\in\mathbb{R}^{l\times n}$	miRNA-TO	447×155	11,440

基因可以被多个表型标记,同时,一个表型可以与多个基因进行关联. 基于3.1节构建的异质遗传网络,本文使用多元个性化随机游走算法实现基因-表型关联预测. 多元个性化随机游走算法分为以下三个步骤.

3.2.1 确定个性化随机游走步长

基于网络的关联预测方法通常为所有节点分配相同的步长来进行随机游走以探索网络的拓扑结构[13, 45~47]. 这些方法忽略了不同节点的重要程度,使得所有的节点有相同的游走步长. 这使得影响力较低的节点的信息得以广泛传播,降低了模型性能. 为此,本文个性化地为每个节点确定步长以便更好地探索网络$\boldsymbol{W}$的拓扑结构.

节点的步长通常依赖于它在网络中的影响力[48]. 受Wang等人工作[49]的启发,本文使用改进的Jaccard指标来评估节点的个性化步长. 对于基因-表型关联网络$\boldsymbol{W}_{gt}$,本文使用$N(g_i)$和$N(t_j)$分别表示基因i和表型j的邻居集合,如果基因i和表型j拥有越多的公共邻居,那么它们越有可能相互影响. 然而,在基因-表型二分关联网络$\boldsymbol{W}_{gt}$中,基因i的邻居为表型,表型j的邻居为基因,$N(g_i)\cap N(t_j)$是空集. 因此,本文使用$N'(t_j)=\bigcup_{g_i\in N(t_j)}N(g_i)$为表型$j$的邻居集合. 然后,二分网络的Jaccard指标定义如下:

$$\boldsymbol{JI}_{gt}(g_i,t_j)=\frac{\left|N(g_i)\cap N'(t_j)\right|}{\left|N(g_i)\cup N'(t_j)\right|} \tag{6}$$

$\boldsymbol{JI}_{gt}$表示在基因-表型关联网络$\boldsymbol{W}_{gt}$中基因i和表型j的相互影响. 假设一个影响程度高的节点在随机游走的过程中与其他节点交互的概率更大,该节点的游走长度应该更大. 基于这个假设,本文定义每个节点的游走步长:

$$\begin{aligned}\boldsymbol{L}^{g}_{gt}(g_i)&=\sum_{j=1}^{n}\boldsymbol{JI}_{gt}(g_i,t_j)\\ \boldsymbol{L}^{t}_{gt}{}'(t_j)&=\sum_{i=1}^{m}\boldsymbol{JI}_{gt}(g_i,t_j)\end{aligned} \tag{7}$$

为克服不同类型的节点数量对游走步长的影响,基于TO节点数量n,本文将游走步长进行规范化处理:

$$\begin{aligned}\boldsymbol{L}^{g}_{gt}(g_i)&=\sum_{j=1}^{n}\boldsymbol{JI}_{gt}(g_i,t_j)\\ \boldsymbol{L}^{t}_{gt}(t_j)&=\frac{n}{m}\sum_{i=1}^{m}\boldsymbol{JI}_{gt}(g_i,t_j)\end{aligned} \tag{8}$$

同理,可以得到基于基因-miRNA关联网络$\boldsymbol{W}_{gm}$的基因和miRNA的游走步长:

$$\begin{aligned}\boldsymbol{L}^{g}_{gm}(g_i)&=\frac{n}{l}\sum_{j=1}^{l}\boldsymbol{JI}_{gm}(g_i,m_j)\\ \boldsymbol{L}^{m}_{gm}(m_j)&=\frac{n}{m}\sum_{i=1}^{m}\boldsymbol{JI}_{gm}(g_i,m_j)\end{aligned} \tag{9}$$

基于miRNA-表型关联网络$\boldsymbol{W}_{mt}$的miRNA和表型的游走步长:

$$\begin{aligned}\boldsymbol{L}^{m}_{mt}(m_i)&=\sum_{j=1}^{n}\boldsymbol{JI}_{mt}(m_i,t_j)\\ \boldsymbol{L}^{t}_{mt}(t_j)&=\frac{n}{l}\sum_{i=1}^{l}\boldsymbol{JI}_{mt}(m_i,t_j)\end{aligned} \tag{10}$$

最终,本文融合了在不同关联网络下取得的游走步长:

$$\begin{aligned}\boldsymbol{L}_{g}&=(\boldsymbol{L}^{g}_{gt}+\boldsymbol{L}^{g}_{gm})/2\\ \boldsymbol{L}_{m}&=(\boldsymbol{L}^{m}_{gm}+\boldsymbol{L}^{m}_{mt})/2\\ \boldsymbol{L}_{t}&=(\boldsymbol{L}^{t}_{gt}+\boldsymbol{L}^{t}_{mt})/2\end{aligned} \tag{11}$$

其中,$\boldsymbol{L}_g\in\mathbb{R}^m$、$\boldsymbol{L}_m\in\mathbb{R}^l$和$\boldsymbol{L}_t\in\mathbb{R}^n$分别存储了$m$个基因、$l$个miRNA和$n$个表型的个性化游走步长.

3.2.2 个性化多元随机游走

同一遗传层面生物分子内部和不同层面分子间均存在遗传关联,这些遗传关联信息间存在着传递与互补关系. 为有效利用3.1节中构建的异质遗传网络中包含的同层面和跨层面遗传信息,通过遗传信息间的增量互补,对网络节点间缺失关联信息进行补充,

本文提出了一种基于多组学数据融合的多元个性化随机游走算法(individual Multiple Random Walks, iMRW)来预测基因-表型关联. 随机游走算法将一个分子当作起始节点,并将与它关联的异类节点当作中间的过渡节点,然后跳转到与过渡节点相似的目标节点上,从而实现该分子与目标节点的关联预测. 分子的游走步长控制着跳转次数,从而实现分子节点的个性化随机游走.

iMRW随机游走过程示例如图3所示,重要节点A的随机游走步长为2,非重要节点B的随机游走步长为1. 在第一轮随机游走过程中,基于关联(1)和miRNA子网络,节点A和B都进行随机游走,建立了一阶新建关联(2);由于B的步长为1,第二轮随机游走过程中,B不再进行随机游走,重要节点A基于关联(2)/(3)进行随机游走,最终建立二阶新建关联(4).

基于上述个性化随机游走过程,iMRW分别以基因节点、miRNA节点和表型节点为起始节点,根据量化的个性化游走步长和异构网络拓扑结构,在异质网络上进行随机游走对基因-miRNA,基因-表型与miRNA-表型进行关联预测,具体而言:

针对基因-表型(或miRNA)关联预测,iMRW从基因g_i开始进行随机游走,基于基因g_i与表型t_k(或miRNAm_k)的关联概率,以及表型t_k与表型t_j(或miRNAm_k与miRNAm_j)的相似度,得到基因g_i与表型t_j(或miRNAm_j)的关联概率,计算公式如下:

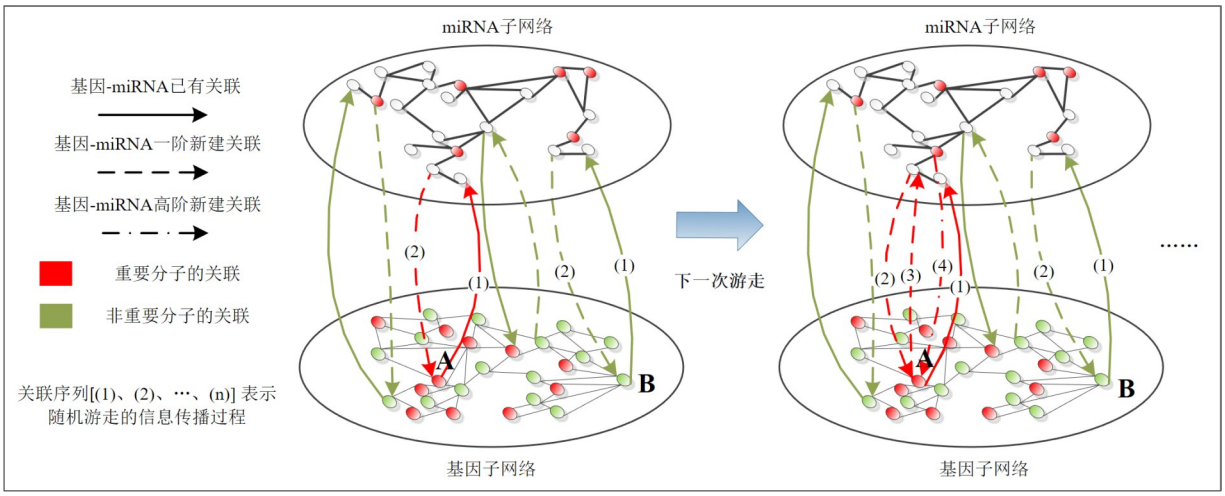


图3 个性化随机游走过程示例

$$\boldsymbol{R}_{gtt}^{t}(g_i,t_j)=\begin{cases}\alpha\sum_{k=1}^{n}\boldsymbol{R}_{gtt}^{t-1}(g_i,t_k)\tilde{\boldsymbol{W}}_{tt}(t_k,t_j)+\\(1-\alpha)\boldsymbol{W}_{gt}(g_i,t_j),\text{if } t\le\lfloor\boldsymbol{L}_g(g_i)\rfloor\\\boldsymbol{R}_{gtt}^{t-1}(g_i,t_j),\text{otherwise}\end{cases}\tag{12}$$

$$\boldsymbol{R}_{gmm}^{t}(g_i,m_j)=\begin{cases}\alpha\sum_{k=1}^{l}\boldsymbol{R}_{gmm}^{t-1}(g_i,m_k)\tilde{\boldsymbol{W}}_{mm}(m_k,m_j)+\\(1-\alpha)\boldsymbol{W}_{gm}(g_i,m_j),\text{if } t\le\lfloor\boldsymbol{L}_g(g_i)\rfloor\\\boldsymbol{R}_{gmm}^{t-1}(g_i,m_j),\text{otherwise}\end{cases}\tag{13}$$

$\boldsymbol{R}_{gtt}^{t}(g_i,t_j)$和$\boldsymbol{R}_{gmm}^{t}(g_i,m_j)$分别是在表型子网络和miRNA子网络上经过$t$次迭代后的基因$i$-表型$j$和基因$i$-miRNA j的关联概率，$\boldsymbol{R}_{gtt}^{0}=\boldsymbol{W}_{gt}$，$\boldsymbol{R}_{gmm}^{0}=\boldsymbol{W}_{gm}$. $\alpha>0$是随机游走的重启概率. $\tilde{\boldsymbol{W}}_{tt}=\boldsymbol{D}_t^{-1/2}*\boldsymbol{W}_{tt}*\boldsymbol{D}_t^{-1/2}$是$\boldsymbol{W}_{tt}$的拉普拉斯标准化矩阵，$\boldsymbol{D}_t$是对角矩阵并且$\boldsymbol{D}_t(t_j,t_j)=\sum_{k=1}^{n}\boldsymbol{W}_{tt}(t_j,t_k)$. $\tilde{\boldsymbol{W}}_{mm}$是$\boldsymbol{W}_{mm}$的拉普拉斯标准化矩阵. 当$t>\boldsymbol{L}_g(g_i)$时，从$g_i$开始的随机游走将会终止.

针对表型-基因(或miRNA)关联预测，iMRW从表型t_j开始进行随机游走，基于表型t_j与基因g_k(或miRNAm_k)的关联概率，以及基因g_k与基因g_i(或miRNAm_k与miRNAm_i)的相似度，得到表型t_j与基因g_i(或miRNA m_i)的关联概率，计算公式如下：

$$\boldsymbol{R}_{gtg}^{t}(g_i,t_j)=\begin{cases}\alpha\sum_{k=1}^{m}\tilde{\boldsymbol{W}}_{gg}(g_i,g_k)\boldsymbol{R}_{gtg}^{t-1}(g_k,t_j)+\\(1-\alpha)\boldsymbol{W}_{gt}(g_i,t_j),\text{if } t\le\lfloor\boldsymbol{L}_t(t_j)\rfloor\\\boldsymbol{R}_{gtg}^{t-1}(g_i,t_j),\text{otherwise}\end{cases}\tag{14}$$

$$\boldsymbol{R}_{mtm}^{t}(m_i,t_j)=\begin{cases}\alpha\sum_{k=1}^{l}\tilde{\boldsymbol{W}}_{mm}(m_i,m_k)\boldsymbol{R}_{mtm}^{t-1}(m_k,t_j)+\\(1-\alpha)\boldsymbol{W}_{mt}(m_i,t_j),\text{if } t\le\lfloor\boldsymbol{L}_t(t_j)\rfloor\\\boldsymbol{R}_{mtm}^{t-1}(m_i,t_j),\text{otherwise}\end{cases}\tag{15}$$

$\boldsymbol{R}_{gtg}^{t}(g_i,t_j)$和$\boldsymbol{R}_{mtm}^{t}(m_i,t_j)$分别是在基因子网络和miRNA子网络上经过$t$次迭代后的基因$i$-表型$j$和miRNA i-表型j的关联概率，$\boldsymbol{R}_{gtg}^{0}=\boldsymbol{W}_{gt}$，$\boldsymbol{R}_{mtm}^{0}=\boldsymbol{W}_{mt}$. $\tilde{\boldsymbol{W}}_{gg}$是$\boldsymbol{W}_{gg}$的拉普拉斯标准化矩阵. 当$t>\boldsymbol{L}_t(t_j)$时，从$t_j$开始的随机游走将会终止.

针对miRNA-基因(或表型)关联预测，iMRW从miRNAm_i开始进行随机行走，基于miRNAm_i与基因g_k(或表型t_k)的关联概率，以及基因g_k与基因g_j(或表型t_k与表型t_j)的相似度，得到miRNAm_i与基因g_j(或表型t_j)的关联概率，计算公式如下：

$$\boldsymbol{R}_{gmg}^{t}(g_j,m_i)=\begin{cases}\alpha\sum_{k=1}^{m}\tilde{\boldsymbol{W}}_{gg}(g_j,g_k)\boldsymbol{R}_{gmg}^{t-1}(g_k,m_i)+\\(1-\alpha)\boldsymbol{W}_{gm}(g_j,m_i),\text{if } t\le\lfloor\boldsymbol{L}_m(m_i)\rfloor\\\boldsymbol{R}_{gmg}^{t-1}(g_j,m_i),\text{otherwise}\end{cases}\tag{16}$$

$$\boldsymbol{R}_{mtt}^{t}(m_i,t_j)=\begin{cases}\alpha\sum_{k=1}^{n}\boldsymbol{R}_{mtt}^{t-1}(m_i,t_k)\tilde{\boldsymbol{W}}_{tt}(t_k,t_j)+\\(1-\alpha)\boldsymbol{W}_{mt}(m_i,t_j),\text{if } t\le\lfloor\boldsymbol{L}_m(m_i)\rfloor\\\boldsymbol{R}_{mtt}^{t-1}(m_i,t_j),\text{otherwise}\end{cases}\tag{17}$$

$\boldsymbol{R}_{gmg}^{t}(g_j,m_i)$和$\boldsymbol{R}_{mtt}^{t}(m_i,t_j)$分别是在基因子网络和表型子网络上经过$t$次迭代后的基因$j$-miRNA i和miRNA i-表型j的关联概率，$\boldsymbol{R}_{gmg}^{0}=\boldsymbol{W}_{gm}$，$\boldsymbol{R}_{mtt}^{0}=\boldsymbol{W}_{mt}$. 当$t>\boldsymbol{L}_m(m_i)$时，从$m_i$开始的随机游走将会终止.

3.2.3 决策融合

通过公式(12)-公式(17)的迭代，可以得到关联预测矩阵$\boldsymbol{R}_{gtt}$、$\boldsymbol{R}_{gmm}$、$\boldsymbol{R}_{gtg}$、$\boldsymbol{R}_{mtm}$、$\boldsymbol{R}_{gmg}$和$\boldsymbol{R}_{mtt}$. 由于以上个性化随机游走算法是独立运行的，miRNA遗传信息不能直接传递给基因-表型关联网络，用于指导基因-表型关联预测. 因此，iMRW首先通过以下公式融合了miRNA相关的关联矩阵，

$$\begin{aligned}\boldsymbol{R}_{gm}&=(\boldsymbol{R}_{gmg}+\boldsymbol{R}_{gmm})/2\\\boldsymbol{R}_{mt}&=(\boldsymbol{R}_{mtt}+\boldsymbol{R}_{mtm})/2\end{aligned}\tag{18}$$

得到基因-miRNA关联预测矩阵$\boldsymbol{R}_{gm}$和miRNA-表型关联预测矩阵$\boldsymbol{R}_{mt}$.

iMRW通过子网络$\boldsymbol{R}_{gm}$(子网络$\boldsymbol{R}_{mt}$)向基因节点(表型节点)投影，获取含有miRNA信息的基因相似度(表

型相似度),再基于这两个相似度矩阵,通过随机游走算法进一步得到基因-表型关联,通过融合基因-miRNA-表型遗传信息,最终实现对基因-表型关联预测过程的指导. 下面详细描述以上过程:

高斯相互作用属性核相似性(Gaussian Interaction Profile Kernel Similarity, GIP kernel similarity)广泛应用于各种半监督学习的预测任务[50~52]. iMRW采用这种相似度度量方法构建包含miRNA信息的基因节点和表型节点的相似度. 基于基因-miRNA关联网络$\boldsymbol{R}_{gm}$,给定两个基因g_i和g_j的高斯相互作用属性核相似性定义如下:

$$\begin{aligned}\boldsymbol{G}_{ggm}(g_i,g_j)&=\exp(-\gamma\left\|\mathbf{IP}(g_i)-\mathbf{IP}(g_j)\right\|^2)\\ \gamma&=\gamma'/(\frac{1}{m}\sum_{i=1}^{m}\left\|\mathbf{IP}(g_i)\right\|^2)\end{aligned}\tag{19}$$

其中,$\boldsymbol{G}_{ggm}\in\mathbb{R}^{m\times m}$表示融合miRNA信息的基因对GIP核相似度,$\mathbf{IP}(g_i)$是矩阵$\boldsymbol{R}_{gm}$的第i行,$\gamma'$是控制核带宽的参数,$m$是基因的数量. 同理,基于miRNA-表型关联网络$\boldsymbol{R}_{mt}$,可以计算融合miRNA信息的表型之间的高斯相互作用属性核相似性$\boldsymbol{G}_{ttm}\in\mathbb{R}^{n\times n}$.

基于获取的含有miRNA信息的基因相似度和表型相似度,本文进一步采取了随机游走算法来计算基因-表型关联:

$$\begin{aligned}\boldsymbol{R}_{gtm}&\leftarrow\alpha\tilde{\boldsymbol{G}}_{ggm}\boldsymbol{W}_{gt}+(1-\alpha)\boldsymbol{W}_{gt}\\ \boldsymbol{R}_{gtm}&\leftarrow\alpha\boldsymbol{R}_{gtm}\tilde{\boldsymbol{G}}_{ttm}+(1-\alpha)\boldsymbol{W}_{gt}\end{aligned}\tag{20}$$

$\boldsymbol{R}_{gtm}\in\mathbb{R}^{m\times n}$是融合miRNA信息的基因-表型关联预测矩阵,$\alpha>0$是随机游走的重启概率. $\tilde{\boldsymbol{G}}_{ggm}$和$\tilde{\boldsymbol{G}}_{ttm}$分别是$\boldsymbol{G}_{ggm}$和$\boldsymbol{G}_{ttm}$的拉普拉斯标准化矩阵.

最终,本文融合了三个基因-表型关联预测子矩阵,构建获取了基因-表型间关联预测矩阵:

$$\boldsymbol{R}_{gt}=(\boldsymbol{R}_{gtg}+\boldsymbol{R}_{gtt}+\boldsymbol{R}_{gtm})/3\tag{21}$$

为了更好的理解异质网络上的多元个性化随机游走算法,本文在算法1中列出了执行步骤.

4 实验分析

4.1 性能评估

为对比分析iMRW的性能,本文选择当下流行的八种算法BiRW[13]、IDLP[21]、NewGOA[35]、ThrRW[53]、tlDLP[14]、TCRW[54]、GCN[55]和GAT[56]作为对比算法. 其中BiRW是双随机游走算法;IDLP是改进的双标签传播算法并重新学习了基因相似度和表型相似度;NewGOA基于过渡概率构建表型相似性网络,在基因和TO组成的混合图上进行双随机游走;ThrRW基于三元随机游走进行基因-表型关联预测;tlDLP利用迁移学习策略重构基因-表型关联;TCRW构建三层异质网络,再通过两层的不平衡双随机游走进行关联预测;GCN和GAT是经典的图深度学习方法,它们都是将邻居节点的特征聚合到中心节点上,GCN是基于拉普拉斯矩阵进行邻居聚合,而GAT是基于注意力系数进行邻居聚合. 在以上对比方法中,ThrRW、tlDLP和TCRW是多组学方法,而BiRW、LDLP、NewGOA、GCN和GAT仅仅考虑了两个组

算法1 iMRW算法.

输入:数据矩阵$\boldsymbol{W}_{gg}$、$\boldsymbol{W}_{tt}$、$\boldsymbol{W}_{mm}$、$\boldsymbol{W}_{gt}$、$\boldsymbol{W}_{gm}$和$\boldsymbol{W}_{mt}$;参数α

输出:基因-表型关联预测矩阵$\boldsymbol{R}_{gt}$.

①根据公式11计算基因、miRNA和表型的个性化游走步长$\boldsymbol{L}_g$、$\boldsymbol{L}_m$和$\boldsymbol{L}_t$

②归一化关联矩阵

$\boldsymbol{W}_{gt}\leftarrow\boldsymbol{W}_{gt}/sum(\boldsymbol{W}_{gt})$

$\boldsymbol{W}_{gm}\leftarrow\boldsymbol{W}_{gm}/sum(\boldsymbol{W}_{gm})$

$\boldsymbol{W}_{mt}\leftarrow\boldsymbol{W}_{mt}/sum(\boldsymbol{W}_{mt})$

③FOR $t=1$ to $\max(\boldsymbol{L}_g)$ DO

FOR $i=1$ to m do

使用公式(12)和公式(13)更新$\boldsymbol{R}^t_{gtt}$和$\boldsymbol{R}^t_{gmm}$

END FOR

END FOR

④FOR $t=1$ to $\max(\boldsymbol{L}_t)$ DO

FOR $i=1$ to n do

使用公式(14)和公式(15)更新$\boldsymbol{R}^t_{gtg}$和$\boldsymbol{R}^t_{mtm}$

END FOR

END FOR

⑤FOR $t=1$ to $\max(\boldsymbol{L}_m)$ DO

FOR $i=1$ to l do

使用公式(16)和公式(17)更新$\boldsymbol{R}^t_{gmg}$和$\boldsymbol{R}^t_{mtt}$

END FOR

END FOR

⑥决策融合

$\boldsymbol{R}_{gm}\leftarrow(\boldsymbol{R}_{gmg}+\boldsymbol{R}_{gmm})/2$

$\boldsymbol{R}_{mt}\leftarrow(\boldsymbol{R}_{mtm}+\boldsymbol{R}_{mtt})/2$

$\boldsymbol{G}_{ggm}\leftarrow GIP(\boldsymbol{R}_{gm},1)$

$\boldsymbol{G}_{ttm}\leftarrow GIP(\boldsymbol{R}_{mt},1)$

使用公式(20)计算$\boldsymbol{R}_{gtm}$

$\boldsymbol{R}_{gt}\leftarrow(\boldsymbol{R}_{gtg}+\boldsymbol{R}_{gtt}+\boldsymbol{R}_{gtm})/3$

⑦返回$\boldsymbol{R}_{gt}$

学的信息. 在iMRW中,本文设置参数$\alpha=0.2$;在其他对比方法中,参数设置为公开代码中的默认值,或在各自文献指定的参考范围内进行优化. 在各种实验配置中,本文对所有方法重复十次实验,记录实验结果的均值和方差作为最终结果,其中,黑体加粗的数据是每列数据成对t检验统计后最好的结果.

为了全面评估模型,本文采用了四种广泛应用的评价指标,即AUROC、AUPRC、F_{max}[57]和S_{min}[58,59]. F_{max}

是在预测的基因-表型关联矩阵 $\boldsymbol{R}_{gt}$ 在给定不同阈值 $\theta\in[0,1]$ 上计算的查准率和查全率的最大调和均值：

$$F_{max}=\max_{\theta\in[0,1]}\frac{2\times P(\theta)\times R(\theta)}{P(\theta)+R(\theta)} \tag{22}$$

其中，$P(\theta)$ 和 $R(\theta)$ 分别表示在阈值为 θ 的条件下的查准率和查全率. S_{min} 是基于TO层次结构在给定不同阈值 θ 上计算得到的预测标签和真实标签之间的最小语义距离[60]：

$$\begin{aligned}&S_{min}=\min\sqrt{ru(\theta)^2+mi(\theta)^2}\\&ru(\theta)=\frac{1}{m}\sum_{i=1}^{m}\sum_{t\in T_i-P_i(\theta)}IC(t)\\&mi(\theta)=\frac{1}{m}\sum_{i=1}^{m}\sum_{t\in P_i(\theta)-T_i}IC(t)\end{aligned} \tag{23}$$

其中，$P_i(\theta)$ 表示第 i 个基因预测概率大于 θ 的标签集合，T_i 表示第 i 个基因的真实标签的集合. $ru(\theta)$ 和 $mi(\theta)$ 分别表示在阈值为 θ 的条件下的剩余不确定性(remaining uncertainty)和误导信息量(misinformation). $IC(t)$ 表示表型标签 t 的信息量，计算方法如公式(3).

4.1.1 五折交叉验证

通过表3的实验结果可以发现iMRW在AUROC、AUPRC、F_{max} 和 S_{min} 上均取得了最优的性能. 在双组学方法中，BiRW是双随机游走算法，IDLP是改进的双随机游走算法，它们仅仅使用了基因子网络和表型子网络，IDLP比BiRW性能更好的原因是IDLP在迭代过程中重新学习了基因相似度和TO相似度，减少了原有数据噪声对模型的影响. NewGOA融合了TO的层次结构和已知的基因标注信息来计算过渡概率，并采用了有向双随机游走算法来预测基因-表型的关联，由于它构建的表型子网络是有向的，比BiRW和IDLP的子网络更加稀疏，含有较少的噪声，所以NewGOA的AUROC比BiRW和IDLP低，AUPRC比BiRW和IDLP高. GCN和GAT是两个基于深度学习的方法，它们仅仅使用了双组学数据，其庞大的参数空间使得它们取得了较好的性能，但这两种方法缺乏对训练参数的可解释性.

在多组学方法中，虽然ThrRW、tlDLP和TCRW都引入了miRNA组学数据信息，但是它们没有考虑每个节点的重要性，平等地对待每一个节点，使得影响力小的节点得到广泛传播，不可避免地引入了噪声，降低了模型性能. 通过进一步分析发现，在基于传统的机器学习方法中，多组学方法普遍比双组学方法的性能好，这说明融合多组学数据能够补充遗传信息，指导基因-表型关联预测[22].

4.1.2 新基因(表型)关联预测

本文测试了iMRW预测与新基因(表型)相关的表型(基因)的能力，这些相关的表型(基因)在输入数据中完全不可见. 具体来说，基因被随机划分成五个子集，其中一个基因子集的所有相关表型被移除并且使用其他基因子集中的基因-表型关联进行预测；类似地，表型角度的实验是通过移除相关基因而进一步进行五折交叉验证. 对于一个基因(表型)，

表3 iMRW与对比方法在五折交叉验证实验上的结果

算法	AUROC	AUPRC	F_{max}	$S_{\min}\downarrow$
BiRW	0.9494±0.0002	0.6367±0.0027	0.5889±0.0005	3.7643±0.0058
IDLP	0.9524±0.0004	0.6632±0.0013	0.6241±0.0011	3.1754±0.0086
NewGOA	0.9243±0.0004	0.6714±0.0008	0.6233±0.0009	2.8798±0.0069
ThrRW	0.9628±0.0004	0.6861±0.0023	0.6303±0.0006	3.5157±0.0060
tlDLP	0.9504±0.0003	0.7125±0.0070	0.6703±0.0006	2.8093±0.0067
TCRW	0.9595±0.0004	0.7608±0.0059	0.7044±0.0007	2.7019±0.0049
GCN	0.9403±0.0028	0.7532±0.0101	0.6966±0.0059	2.8632±0.0403
GAT	0.9386±0.0063	0.7660±0.0215	0.6893±0.0243	2.9332±0.1080
iMRW	**0.9687 0.0004**	**0.7948 0.0020**	**0.7133 0.0007**	**2.4703 0.0065**

表4 iMRW与对比方法在预测新基因上的结果

算法	AUROC	AUPRC	F_{max}	$S_{\min}\downarrow$
BiRW	0.8261±0.0003	0.3562±0.0008	0.3589±0.0005	5.3561±0.0031
NewGOA	0.7026±0.0003	0.2653±0.0006	0.3243±0.0007	5.4050±0.0035
ThrRW	0.8433±0.0006	0.3897±0.0012	0.3925±0.0007	5.2478±0.0038
tlDLP	0.6359±0.0004	0.1598±0.0025	0.2494±0.0005	5.6935±0.0040
TCRW	0.8434±0.0000	0.3933±0.0077	0.3839±0.0002	5.3759±0.0005
GCN	0.8472±0.0003	0.4014±0.0012	0.3951±0.0010	5.2581±0.0057
GAT	0.8544±0.0034	0.4072±0.0020	0.4087±0.0037	5.2147±0.0095
iMRW	**0.8614 0.0004**	**0.4150 0.0009**	0.4118±0.0006	**5.1825 0.0036**

表5 iMRW与对比方法在预测新表型上的结果

算法	AUROC	AUPRC	F_{max}	$S_{\min}\downarrow$
BiRW	0.9326±0.0050	0.5056±0.0187	0.5205±0.0180	4.2302±0.1334
NewGOA	0.8177±0.0427	0.6091±0.0487	0.5846±0.0314	2.8543±0.1558
ThrRW	0.9353±0.0067	0.5073±0.0260	0.5891±0.0160	3.9564±0.1718
tlDLP	0.9397±0.0070	0.6859±0.0299	0.6502±0.0316	2.9266±0.1862
TCRW	0.8751±0.0077	0.3650±0.0176	0.4321±0.0158	4.8901±0.1018
GCN	0.9166±0.0033	0.6969±0.0256	0.6080±0.0231	3.1481±0.2325
GAT	0.9331±0.0047	0.7051±0.0276	0.6252±0.0205	2.9652±0.2487
iMRW	**0.9516 0.0038**	0.7307±0.0289	**0. 6786 0.0184**	**2.6012 0.2339**

IDLP依赖它的已知表型(基因)关联进行预测,所以本文没有在此实验中设置IDLP. 表4-5给出了iMRW与对比方法的预测结果,iMRW在所有评价指标下取得了较优的性能.

与3. 1. 1节的五折交叉验证不同,预测与新基因(表型)相关的表型(基因)更具有挑战性,因此,

如表4-5所示的实验结果普遍比五折交叉验证低. 对于新表型,与其相关的基因完全未知,根据网络拓扑结构分析可知,新表型的初始基因-表型关联是基于表型子网络而生成,因此预测新表型相关的基因会依赖表型子网络;同理,预测新基因相关的表型会依赖基因子网络. 由于表型子网络是基于TO层次结构建立,基因子网络是基于基因序列相似度或PPI建立,表型子网络比基因子网络含有更少的噪声,所以预测新表型相关的基因比预测新基因相关的表型普遍取得了更好的效果.

4.2 个性化随机游走步长分析

为了测试个性化随机游走的贡献,本文设置了如下实验:iMRW对所有的节点使用相同的游走步长,步长从1变化到10. 图4展示了不同的随机游走步长在各个评价指标下的结果. 从图中可以看出,当随机游走步长大于2时,AUROC、AUPRC和 不再升高,不再降低,这是因为在相同步长的随机游走过程中,有用的节点信息和噪声节点信息的正负作用相互抵消,使得模型性能趋于饱和. 而个性化随机游走的性能一直优于相同步长随机游走的性能,且不受游走步长的影响.

4.3 消融实验分析

本文设置了三种变种实验iMRW(nM)、iMRW(nMG)和iMRW(nMT)来探索复杂分子网络对模型的影响. iMRW(nM)忽略了与miRNA相关的网络,仅仅使用基因子网络 $\boldsymbol{W}_{gg}$,表型子网络 $\boldsymbol{W}_{tt}$ 和基因-表型关联网络 $\boldsymbol{W}_{gt}$ 进行关联预测;iMRW(nMG)忽略了与miRNA相关的网络和基因子网络,仅仅使用表型子网络 $\boldsymbol{W}_{tt}$ 和基因-表型关联网络 $\boldsymbol{W}_{gt}$ 进行关联预测;iMRW(nMT)忽略了与miRNA相关的网络和表型子网络,仅仅使用基因子网 $\boldsymbol{W}_{gg}$ 和基因-表型关联网络 $\boldsymbol{W}_{gt}$ 进行关联预测.

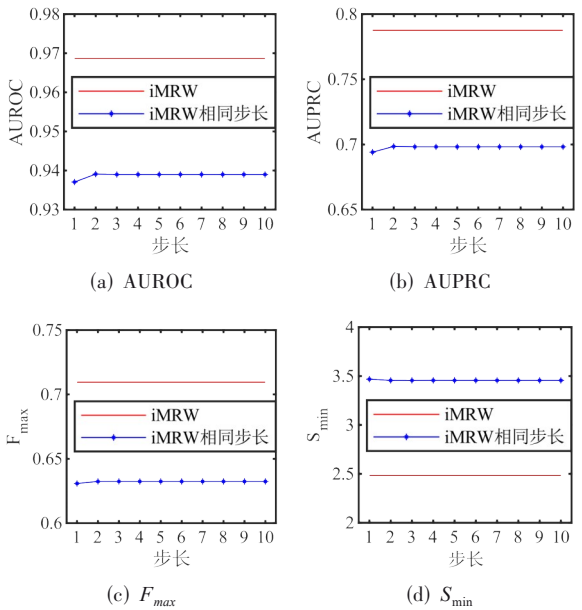


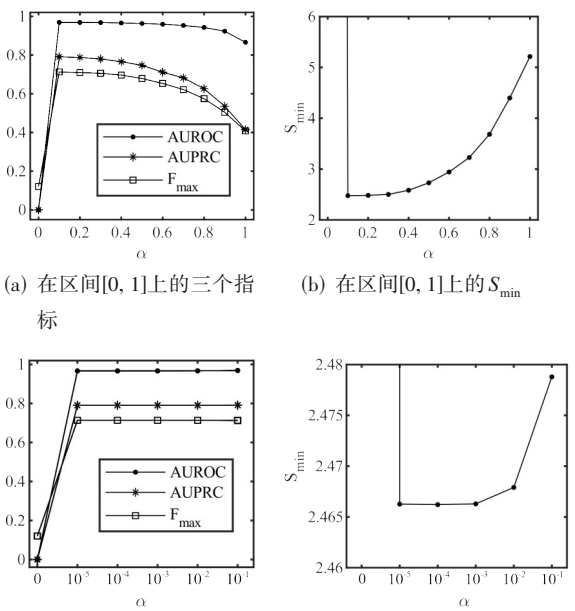
图4 相同的游走步长对随机游走的影响

表6给出了iMRW及其变种实验的预测结果. iMRW在所有指标上比iMRW(nM)取得更优越的性能,说明融合miRNA的组学信息对基因-表型的关联预测起到了指导作用. iMRW(nM)比iMRW(nMG)和iMRW(nMT)效果好说明使用基因子网络和表型子网络能够显著提高模型性能. iMRW(nMG)比iMRW(nMT)取得更好的性能,这是因为表型子网络是根据TO层次结构而建立,含有的噪声较少,而基因子网络通过基因序列相似度而建立,含有噪声较多. 以上结果表明,融合多组学数据对预测基因-表型关联具有重要作用[22].

4.4 参数分析

iMRW使用了一个参数 α 来控制随机游走的重启概率. 本文研究了参数 α 对iMRW预测性能的影响. 首先对 α 设置搜索空间为0~1,从0开始以0. 1为步长递增进行分析,获取的各个指标变化如图5(a)、(b)所示. 当重启概率 $\alpha=0$ 时,AUROC=AUPRC=0, $F_{max}=0.1207$, $S_{min}=122.7993$,效果最低,这是因为iMRW没有进行随机游走;当 $\alpha\geq 0.1$ 时,随着重启概率的增大,

iMRW的性能降低，这是因为已知关联信息所占权重$(1-\alpha)$的减小所带来的负向影响大于重启概率增大所带来的正向影响. 由于$\alpha\in[0,0.1]$时，模型性能有明显的上升趋势，为进一步探索该区间内α对预测性能的影响，本文从10^{-5}开始以10倍为步长递增进行分析，获取的各个指标变化如图5(c)、(d)所示. 结果显示，当$\alpha\in[0,10^{-5}]$时，模型性能呈现明显上升趋势，这是因为即使α较小，iMRW也进行了个性化随机游走，补全了基因-表型之间的关联，且已知的关联信息得到充分保留. 当$\alpha>10^{-5}$时，iMRW模型性能增长趋于平稳，并在$\alpha=0.1$时达到最大值. 根据参数分析结果，本文选取了$\alpha=0.1$作为模型参数.



(a) 在区间[0, 1]上的三个指标　(b) 在区间[0, 1]上的$S_{\min}$

(c) 在区间[0, 0.1]上的三个指标　(d) 在区间[0, 0.1]上的$S_{\min}$

图5 参数 对iMRW预测性能的影响

4.5 案例研究

本文实施了案例研究，以测试iMRW识别与重要表型有关的潜在基因的有效性. 在已知基因-表型关联和异质遗传网络的基础上，运行iMRW得到基因-表型关联矩阵$\boldsymbol{R}_{gt}$. 然后剔除已知的基因-表型关联，

在表7-8中分别列出了可能与光合作用能力和淀粉含量相关的前10个基因. 基于基因本体(Gene Ontology，GO)注释[61]，可以直接(或间接)验证预测结果.

基因“GRMZM2G036996”、“GRMZM2G072513”分别被“GO：0007165”和“GO：0009579”标注. 其中，“GO:0007165”与细胞间的信号(包括光信号等)传导有关，“GO:0009579”与在植物中含有光合色素的膜状细胞结构有关. 基因“GRMZM2G169967”与“AC198371.3_FG011”都被“GO：0009416”标注，“GO：0009416”与对光刺激，波长为红外、可见光或紫外光的电磁辐射的反应有关. 以上GO证据表明，这些基因直接参与了与光合作用相关的生物过程，它们与光合作用能力直接相关. 光合作用是一种氧化还原反应，它将光能转化为化学能[62]，基因“GRMZM2G170017”被“GO：0016616”(氧化还原反应的催化)标注. 光合作用与氮素[63]密切相关. 基因“GRMZM2G169994”注释为“GO：0071705”(含氮化合物的反应)，可能会影响光合作用的效率. 基因“GRMZM2G095209”、“GRMZM2G015578”、“GRMZM2G447847”、“GRMZM2G395114”和“GRMZM2G169962”都被“GO：1901576”标记，“GO：1901576”反应了导致有机物形成的化学反应和途径，并与任何含碳有机物的生物合成过程有关. 基因“GRMZM2G170016”、“GRMZM2G484344”和“GRMZM2G170017”被“GO：0071704”标记，“GO：0071704”涉及有机物、任何含有碳的分子实体的化学反应和途径. 其他基因-表型关联也已被湿实验或生物文献验证.

这些案例结果再次证明了iMRW在识别与重要性状相关的潜在基因方面的有效性.

5 结论

针对已有的随机游走算法忽略了不同节点的重要性，使得非重要节点过度传播，降低了模型性能的问题，本文提出了一种基于多组学数据融合的个性化随机游走算法(individual Multiple Random Walks，iMRW). iMRW基于网络拓扑结构为不同重要程度的节点分配不同的游走步长，并通过高斯相互作用属性核相似性结合随机游走，最终获取基因-表型关联预测矩阵. 在不同实验设置下，与主流算法的对比实验结果均显示iMRW能够取得了较优的预测性能. 在玉米光合作用能力和淀粉含量表型的实验分析结果也进一步证实了iMRW在识别基因-表型关联分析中的实用性与

表6 消融实验分析

算法	AUROC	AUPRC	F_{max}	$S_{\min}\downarrow$
iMRW	**0.9687 0.0004**	**0.7948 0.0020**	**0.7133 0.0007**	**2.4703 0.0065**
iMRW(nM)	0.9599±0.0011	0.7531±0.0017	0.6726±0.0009	2.6981±0.0078
iMRW(nMG)	0.8447±0.0004	0.6398±0.0027	0.6043±0.0011	2.7946±0.0108
iMRW(nMT)	0.8515±0.0011	0.3921±0.0014	0.3929±0.0003	5.2487±0.0027

表7 与"TO:0000316"(光合作用能力)相关的潜在基因

等级	基因	是否被证实	证据
1	GRMZM2G036996	√	GO:0007165
2	GRMZM2G462986	√	GO:0071704
3	GRMZM2G072513	√	GO:0009579
4	GRMZM2G169967	√	GO:0009416
5	GRMZM2G169994	√	GO:0071705
6	GRMZM2G170017	√	GO:0016616
7	GRMZM2G170016	√	GO:0071704
8	GRMZM2G169962	?	
9	AC198371.3_FG011	√	GO:0009416
10	GRMZM2G104310	√	GO:0071704

表8 与"TO:0000696"(淀粉含量)相关的潜在基因

等级	基因	是否被证实	证据
1	GRMZM2G095308	√	GO:0008152
2	GRMZM2G095209	√	GO:1901576
3	GRMZM2G170016	√	GO:0071704
4	GRMZM2G015578	√	GO:1901576
5	GRMZM2G095305	?	
6	GRMZM2G447847	√	GO:1901576
7	GRMZM2G395114	√	GO:1901576
8	GRMZM2G484344	√	GO:0071704
9	GRMZM2G170017	√	GO:0071704
10	GRMZM2G169962	√	GO:1901576

有效性. 理论分析和实验结果均表明iMRW能够充分利用多组学数据,实现对基因-表型关联更准确的预测;同时,iMRW作为通用的异质信息网络融合框架,也可以推广应用到药物靶点预测、商品推荐等领域.

参 考 文 献

[1] LI Y F, WU F X, ALIOUNE N. A review on machine learning principles for multi-view biological data integration[J]. Briefings in Bioinformatics, 2018, 19(2): 325-340.

[2] PAN Y, LEI X J, ZHANG Y C. Association predictions of genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, factor environment, and networks disease: a comprehensive approach[J]. Medicinal research reviews, 2022, 42(1): 441-461.

[3] DING Y L, LEI X J, LIAO B, et al. Machine learning approaches for predicting biomolecule disease associations [J]. Briefings in Functional Genomics, 2021, 20(4): 273-287.

[4] PIERUSCHKA R, POORTER H. Phenotyping plants: genes, phenes and machines[J]. Functional Plant Biology, 2012, 39(11): 813-820.

[5] YANG W N, DUAN L F, CHEN G X, et al. Plant phenomics and high-throughput phenotyping: accelerating rice functional genomics using multidisciplinary technologies [J]. Current opinion in plant biology, 2013, 16(2): 180-187.

[6] DHONDT S, WUYTS N, INZE D. Cell to whole-plant phenotyping: the best is yet to come[J]. Trends in plant science, 2013, 18(8): 428-439.

[7] PENG C, LI A, WANG M H. Discovery of bladder cancer-related genes using integrative heterogeneous network modeling of multi-omics data[J]. Scientific reports, 2017, 7 (1): 1-11.

[8] DAVIS B D. The isolation of biochemically deficient mutants of bacteria by means of penicillin[J]. Proceedings of the National Academy of Sciences of the United States of America, 1949, 35(1): 1-10.

[9] SOULE M. Phenetics of natural populations i. phenetic relationships of insular populations of the side-blotched lizard[J]. Evolution, 1967, 21(3): 584-591.

[10] SCHORK N J. Genetics of complex disease: approaches, problems, and solutions[J]. American journal of respiratory and critical care medicine, 1997, 156(4): S103-S109.

[11] GANDHI T K B, ZHONG J, MATHIVANAN S, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets[J]. Nature genetics, 2006, 38(3): 285-293.

[12] OTI M, BRUNNER H G. The modular nature of genetic diseases[J]. Clinical genetics, 2007, 71(1): 1-11.

[13] XIE M Q, XU Y J, ZHANG Y G, et al. Network-based phenome-genome association prediction by bi-random walk[J]. PLoS one, 2015, 10(5): 1-18.

[14] PETEGROSSO R, PARK S, HWANG T H, et al. Transfer learning across ontologies for phenome genome association prediction[J]. Bioinformatics, 2017, 33(4): 529-536.

[15] 傅广垣, 余国先, 王峻, 等. 基于正负样例的蛋白质功能预测[J]. 计算机研究与发展, 2016, 53(8): 1753-1765.
FU G Y, YU G X, WANG J, et al. Protein function prediction using positive and negative examples[J]. Journal of Computer Research and Development, 2016, 53(8): 1753-1765 (in Chinese).

[16] 李敏, 王晓桐, 罗慧敏, 等. 随机游走技术在网络生物学中的研究进展[J]. 电子学报, 2018, 46(8): 2035-2048.
LI M, WANG X T, LUO H M, et al. Progress on random walk and its application in network biology[J]. Acta Electronica Sinica, 2018, 46(8): 2035-2048 (in Chinese).

[17] KOHLER S, BAUR S, HORN D, et al. Walking the inter-

actome for prioritization of candidate disease genes[J]. The American journal of human genetics, 2008, 82(4): 949-958.

[18] VANUNU O, MAGGER O, RUPPIN E, et al. Associating genes and protein complexes with disease via network propagation[J]. PLoS computational biology, 2010, 6(1): 1-9.

[19] CHEN Y, JIANG T, JIANG R. Uncover disease genes by maximizing information flow in the phenome interactome network[J]. Bioinformatics, 2011, 27(13): i167-i176.

[20] 谢雨洋, 冯栩, 喻文健, 等. 基于随机化矩阵分解的网络嵌入方法[J]. 计算机学报, 2021, 44(3): 447-461.
XIE Y Y, FENG X, YU W J, et al. Learning network embedding with randomized matrix factorization[J]. Chinese Journal Computers, 2021, 44(3): 447-461 (in Chinese).

[21] ZHANG Y G, LIU J H, LIU X H, et al. Prioritizing disease genes with an improved dual label propagation framework[J]. BMC bioinformatics, 2018, 19(1): 1-12.

[22] RITCHIE M D, HOLZINGER E R, LI R, et al. Methods of integrating data to uncover genotype phenotype interactions[J]. Nature reviews genetics, 2015, 16(2): 85-97.

[23] FU G Y, WANG J, DOMENICONI C, et al. Matrix factorization-based data fusion for the prediction of lncrna disease associations[J]. Bioinformatics, 2018, 34(9): 1529-1537.

[24] CHEN X, ZHANG D H, YOU Z H. A heterogeneous label propagation approach to explore the potential associations between mirna and disease[J]. Journal of translational medicine, 2018, 16(1): 1-14.

[25] 马慧芳, 贾美惠子, 张迪, 等. 融合标签关联关系与用户社交关系的微博推荐方法[J]. 电子学报, 2017, 45(1): 112-118.
MA H F, JIA M H Z, ZANG D, et al. Microblog recommendation based on tag correlation and user social relation. Acta Electronica Sinica, 2017, 45(1): 112-118 (in Chinese).

[26] HUANG Q Y, WANG J, ZHANG X L, et al. Isoform-disease association prediction by data fusion[C]//International Symposiumon Bioinformatics Research and Applications. [S.l.]: Springer, 2020: 44-55.

[27] GARTNER T, STEINFATH M, ANDORF S, et al. Improved heterosis prediction by combining information on dna-and metabolic markers[J]. PLoS one, 2009, 4(4): 1-12.

[28] RIEDELSHEIMER C, TECHNOW F, MELCHINGER A E. Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines[J]. BMC genomics, 2012, 13 (1): 1-9.

[29] JINAG J, XING F, WANG C Y, et al. Investigation and development of maize fused network analysis with multi-omics[J]. Plant Physiology and Biochemistry, 2019, 141 (1): 380-387.

[30] XU Y, XU C, XU S. Prediction and association mapping of agronomic traits in maize using multiple omic data[J]. Heredity, 2017, 119(3): 174-184.

[31] JIANG J, XING F, ZENG X X, et al. Investigating maize yield-related genes in multiple omics interaction network data[J]. IEEE Transactions on Nanobioscience, 2019, 19 (1): 142-151.

[32] YU G X, YANG Y Q, YAN Y Y, et al. Deepida: predicting isoform-disease associations by data fusion and deep neural networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, PP(99): 1-1.

[33] COOPER L, MEIER A, LAPORTE M A, et al. The planteome database: an integrated resource for reference ontologies, plant genomics and phenomics[J]. Nucleic acids research, 2018, 46(D1): D1168-D1180.

[34] VALENTINI G. True path rule hierarchical ensembles for genome-wide gene function prediction[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010, 8(3): 832-847.

[35] YU G X, FU G Y, WANG J, et al. Newgoa: predicting new go annotations of proteins by bi-random walks on a hybrid graph[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 15(4): 1390-1402.

[36] AMBROS V. microrna: tiny regulators with great potential[J]. Cell, 2001, 107(7): 823-826.

[37] AMBROS V. The functions of animal microrna[J]. Nature, 2004, 431(7006): 350-355.

[38] 王磊, 徐涛, 宋传东, 等. 基于深度学习的miRNA与疾病相关性预测算法[J]. 电子学报, 2020, 48(5): 870-877.
WANG L, XU T, SONG C D, et al. Prediction algorithm of association between miRNAs and diseases based on deep learning[J]. Acta Electronica Sinica, 2020, 48(5): 870-877 (in Chinese).

[39] MISKA E A. How microrna control cell division, differentiation and death[J]. Current opinion in genetics & development, 2005, 15(5): 563-568.

[40] BARTEL D P. Microrna: target recognition and regulatory functions[J]. Cell, 2009, 136(2): 215-233.

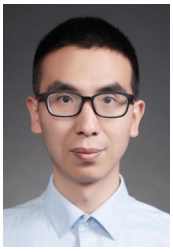
[41] GUO Z L, KUANG Z, WANG Y, et al. Pmiren: a comprehensive encyclopedia of plant mirnas[J]. Nucleic acids research, 2020, 48(D1): D1114-D1121.
[42] KUANG Z, WANG Y, LI L, et al. mirdeep-p2: accurate and fast analysis of the microrna transcriptome in plants [J]. Bioinformatics, 2019, 35(14): 2521-2522.
[43] LIANG C, YU S P, LUO J W. Adaptive multi-view multi-label learning for identifying disease-associated candidate mirnas[J]. PLoS computational biology, 2019, 15(4): 1-18.
[44] PAN Q C, WEI J F, GUO F, et al. Trait ontology analysis based on association mapping studies bridges the gap between crop genomics and phenomics[J]. BMC genomics, 2019, 20(1): 1-13.
[45] HU J L, GAO Y Q, LI J, et al. A novel algorithm based on bi-random walks to identify disease-related lncrnas[J]. BMC bioinformatics, 2019, 20(18): 1-11.
[46] YU G X, WANG K Y, DOMENICONI C, et al. Isoform function prediction based on bi-random walks on a heterogeneous network[J]. Bioinformatics, 2020, 36(1): 303-310.
[47] XIE G B, WU C H, GU G S, et al. Haubrw: Hybrid algorithm and unbalanced bi-random walk for predicting lncrna-disease associations[J]. Genomics, 2020, 112(6): 4777-4787.
[48] ZHAO Y W, WANG J, GUO M Z, et al. Cross-species protein function prediction with asynchronous-random walk[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2019, 18(4): 1439-1450.
[49] WANG Y H, GUO M Z, REN Y Z, et al. Drug repositioning based on individual bi-random walks on a heterogeneous network[J]. BMC bioinformatics, 2019, 20(15): 1-13.
[50] PAN Z X, ZHANG H X, LIANG C, et al. Self-weighted multi-kernel multi-label learning for potential mirna-disease association prediction[J]. Molecular Therapy-Nucleic Acids, 2019, 17(1): 414-423.
[51] YIN M M, CUI Z, GAO M M, et al. Lwpcmf: logistic weighted profile based collaborative matrix factorization for predicting mirna-disease associations[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021, 18(3): 1122-1129.
[52] TAN H J, SUN Q M, LI G H, et al. Multiview consensus graph learning for lncrna disease association prediction [J]. Frontiers in Genetics, 2020, 11(89): 1-10.
[53] PENG W, LI M, CHEN L, et al. Predicting protein functions by using unbalanced random walk algorithm on three biological networks[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(2): 360-369.
[54] YU L M, SHEN X J, ZHONG D, et al. Three-layer heterogeneous network combined with unbalanced random walk for mirna-disease association prediction[J]. Frontiers in Genetics, 2020, 10(1316): 1-10.
[55] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[C]. International Conference on Learning Representations, 2017, 1 14.
[56] VELICKOVIC P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]. International Conference on Learning Representations, 2018, 1-12.
[57] RADIVOJAC P, CLARK W T, ORON T R, et al. A large-scale evaluation of computational protein function prediction[J]. Nature Methods, 2013, 10(3): 221-227.
[58] CLARK W T, RADIVOJAC P. Information-theoretic evaluation of predicted ontological annotations[J]. Bioinformatics, 2013, 29(13): i53-i61.
[59] JIANG Y X, ORON T R, CLARK W T, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy[J]. Genome biology, 2016, 17(1): 1-19.
[60] ZHOU G J, WANG J, ZHANG X L, et al. Predicting functions of maize proteins using graph convolutional network[J]. BMC bioinformatics, 2020, 21(16): 1-16.
[61] CONSORTIUM G O. The gene ontology resource: 20 years and still going strong[J]. Nucleic acids research, 2019, 47(D1): D330-D338.
[62] ALLEN J F, ALEXCIEV K, HAKANSSON G. Photosynthesis: Regulation by redox signalling[J]. Current Biology, 1995, 5(8): 869-872.
[63] ECANS J R. Photosynthesis and nitrogen relationships in leaves of c3 plants[J]. Oecologia, 1989, 78(1): 9-19.

作者简介

谭好江 男，1998 年出生，山东潍坊人．现为山东大学硕士研究生．主要研究方向为机器学习和生物信息学． E-mail：tanhaojiang@mail. sdu. edu. cn

王 峻 （通讯作者）女，1983年出生，重庆人．现为山东大学人工智能国际联合研究院教授．主要研究方向为数据挖掘和生物信息学． E-mail：kingjun@sdu. edu. cn

余国先 男，1985年出生，湖北孝感人．现为山东大学软件学院教授．主要研究方向为机器学习、数据挖掘和生物信息学． E-mail：gxyu@sdu. edu. cn



陈 建 男，1988年出生，浙江衢州人．现为中国农业大学农学院副教授．主要研究领域为玉米基因组学和生物信息学． E-mail：jianchen@cau. edu. cn

郭茂祖 男，1966年出生，山东德州人．现为北京建筑大学电气与信息工程学院教授．主要研究方向为机器学习、数据挖掘和生物信息学． E-mail：guomaozu@bucea. edu. cn