基于双层耦合网的表型-基因关联分析与预测

郁 湧，顾 捷，赵 娜*，骆永军，阚世林

(云南大学软件学院，云南省软件工程重点实验室 昆明 650091)

【摘要】随着基因组测序完成和基因技术不断发展，使得某些疾病的致病基因逐渐得到确认。目前，通过科学实验已经掌握了一部分疾病的致病原因，但是大部分疾病的致病原因，特别是与基因相关的疾病的致病原因还不得而知。该文采用与人类同源相似度高达 85% 的小鼠数据作为研究对象，使用疾病表型数据集、致病基因数据集和已经确认的表型−基因关联关系数据集构成一个双层耦合网络，通过元路径上随机游走的方法进行数据的分析与挖掘，在已经确认的表型−基因关联数据基础上预测未确定的表型−基因关联关系。经验证比较，该文提出的算法所取得的预测效果优于其他算法。

关 键 词 关联关系； 疾病表型； 双层耦合网络； 致病基因

中图分类号 TP391 **文献标志码** A doi:10.12178/1001-0548.2019133

Phenotype-Gene Association Analysis and Prediction Based on Double-Layer Coupled Network

YU Yong, GU Jie, ZHAO Na*, LUO Yong-jun, and KAN Shi-lin

(School of Software, Key Laboratory in Software Engineering of Yunnan Province, Yunnan University Kunming 650091)

Abstract With the completion of genome sequencing and the continuous development of gene technology, the pathogenic genes of some diseases are gradually identified. At present, people have grasped the pathogenic causes of some diseases through scientific experiments, but the pathogenic causes of most diseases, especially those related to genes, are still unknown. In this paper, the mouse data with 85% homology similarity to human is used as the research object. The disease phenotype data set, pathogenic gene data set and confirmed phenotype-gene association data set are constructed into a double-layer coupled network. The data are analyzed and mined by meta-path random walk method, and the uncertainties are predicted on the basis of confirmed phenotype-gene association data. The proposed algorithm achieves better prediction results compared with other algorithms.

Key words correlation; disease phenotype; double-layer coupled network; pathogenic gene

人类第三代测序技术的迅速发展，让生命系统组成元件间的相互作用关系信息得到更加快速的积累。基因数据的不断丰富，表型数据的不断增加，为理解疾病与致病基因之间的关系提供了大量有效的数据。在生物数据大量涌现的前提下，利用相关计算技术和模型对数据进行分析与挖掘，加快了生物学研究前进的步伐，可以深层次挖掘疾病表型与致病基因之间的关系，为了解疾病发病机理、疾病临床诊断和疾病预防与治疗提供了便利。

通过几十年的努力，人类已经发现了一些疾病的致病基因，如 BRCA1 和 BRCA2 基因在乳腺癌的发生中发挥重要的作用[1]，EGFR 在肺癌的发生中发挥重要作用[2]。如果能够知道更多疾病的致病基因，则可以在发病前期进行基因检测预防，在发病过程中进行相应的治疗，后续也可以将发病机理应用到药物设计中，从而有效提高疾病的控制与治愈能力。通过疾病表型和致病基因关系的挖掘，使得疾病发病机理一目了然，在疾病发现过程中能直击疾病发病原因，后续治疗能做到药到病除。

1 疾病基因预测算法研究现状

目前，挖掘疾病表型与致病基因的关联关系是一个极具挑战的课题。如果能够设计出高精度的致病基因预测方法，对于生物学家、临床医师和遗传学家等相关人员来说具有非常重要的意义。这不但有助于提高发现致病基因的准确率，缩短发现致病基

收稿日期：2019－06－03；修回日期：2019－11－07
基金项目：国家自然科学基金 (61462091)；云南省教育厅科研项目 (2019J0008，2019J0010)；云南省数据驱动软件工程创新团队 (2017HC012)
作者简介：郁湧 (1980－)，男，博士，副教授，主要从事网络科学、社交网络与社会媒体分析等方面的研究.
通信作者：赵娜，E-mail：zhaonayx@126.com

因的周期，节省大量的人力物力，同时也为将来的生物医学和基因治疗诊断等技术的发展奠定重要基础。

随着计算机和生物技术的迅猛发展，大量的生物信息数据的产生，疾病和基因知识的可用性大幅度提高，科研人员也相应提出了一系列疾病与基因预测的计算方法。其中，随机游走是疾病与基因关联关系预测中较为常见的办法，主要包括重启随机游走和双向随机游走等几种类型。文献 [3] 在双层耦合网络上提出了重启随机游走，用于推断潜在的 miRNA 与疾病的相关性。文献 [4] 开发了 BiRWHMDA 的计算模型，通过在双层耦合网络上的双向随机游走来预测潜在的微生物与疾病关联。文献 [5] 提出在双层耦合网络上基于多路径的双向随机游走预测微生物与疾病相关性。文献 [6] 结合表型相似网络、基因相似网络和表型基因关联网络构成表型基因双层耦合网络，并在其上采用重启随机游走算法，推出了一种新的预测疾病致病基因的方法。文献 [7] 采用了带重启的随机游走算法和最短路径这两种广泛使用的算法，构造了两种参数化计算方法，即基于 RWR 的方法和基于 SP 的方法，并在此基础上构建了一种新的疾病基因识别的集成方法。

利用矩阵预测疾病与基因关系也是一个不错的办法。文献 [8] 提出了一种基于归纳式矩阵补全预测潜在 lncRNA 与疾病相关性的方法 (predict lncRNA-disease associations from known data using IMC, SIMCLDA)。文献 [9] 开发了一种利用协同矩阵因子分解预测人类微生物疾病相关性的模型 (collaborative matrix factorization for human microbe-disease association, CMFHMDA)。文献 [10] 提出一种基于 Katz 方法的预估计和基于归纳型矩阵补全方法的精化估计两步骤的 Katz 增强归纳型矩阵补全的基因–疾病关联预测模型。

把高斯相互作用应用于预测之中，文献 [11] 应用高斯相互作用轮廓核相似测度确定微生物相似性和疾病相似性。文献 [12] 建立了用于 miRNAs 与疾病相关性预测的双层耦合网络推理的计算模型，通过整合 miRNAs 功能相似性、疾病语义相似性、高斯相互作用来揭示潜在的 miRNAs 与疾病相关性。

将路径作为预测分数，文献 [13] 引入 PBHMDA (path-based human microbe-disease association)，通过对微生物与疾病之间的所有路径进行评估，得出每个候选微生物与疾病对的预测得分。

研究人员还提出了其他一些疾病与基因关系预测的办法。文献 [14] 提出了一种基于 SimRank 和密度聚类推荐模型的 miRNA 与疾病相关性预测方法 (based on the SimRank and density-based clustering recommender model for miRNA-disease associations prediction, SRMDAP)。文献 [15] 基于 miRNA 与疾病关联预测评分模型 (within and between score for MiRNA-disease association prediction, WBSMDA) 预测与各种复杂疾病关联的 miRNAs。文献 [16] 采用拉普拉斯正则化最小二乘分类器 (Laplacian regularized least squares for human microbe–disease association, LRLSHMDA) 建立预测模型。文献 [17] 将链路预测的思想引入到长非编码 RNA–疾病关联预测中。文献 [18] 提出一种基于密度聚类的二分网络投影算法 (bipartite network projection based on density clustering to predict miRNA-disease associations, BNPDCMDA) 来预测 miRNA–疾病关联。

以随机游走为主导思想的预测方法能够扩大候选基因的范围，可以避免遗漏连接度低和网络边缘的节点，尤其是在多基因疾病的预测中，可以大大提高预测候选致病基因方法的性能；在矩阵预测中，数据的稀疏对预测有很大的影响，PU 问题也是需要面对的另一个问题，加入 Katz 方法也只缓解部分影响；使用高斯相互作用预测将疾病或者基因的相互作用信息作为特征向量，引入高斯核函数，计算疾病或基因间的相似度后在进行疾病和基因之间的相似预测，但是对高斯相互作用相似度参数标准化后，基因或疾病高斯核相互作用相似值就不在依赖于数据集；路径预测利用了生物信息节点之间的拓扑结构，在拓扑结构的基础上预测；其他一些算法都是基于机器学习的一些思想进行关联预测的，然而有监督的机器学习算法，需要假设与疾病相关的基因和不相关的基因是不关联的，但是被证明与疾病相关的基因数量较少，且很少有实验能够证明那些关系是不存在的。

进行多种算法比较研究后，可知基于随机游走的方法相比矩阵预测或聚类的方法存在一定优越性。本文根据疾病表型和疾病基因数据节点属于不同类型节点这一特点，基于疾病表型和疾病基因数据来构成双层耦合网络，提出了在表型–基因的双层耦合网络基础上进行带有元路径的随机游走，从而实现关联关系的预测与分析算法。

2　表型–基因双层耦合网的构建

复杂网络的研究大多局限于单个网络，而事实上单个网络仅仅是更大复杂系统中的一个子集，复杂系统往往是由许多具有不同结构与功能的网络耦合而成的[19]。多层耦合网络由多个子网络构成，网

络中每一层通过一些共享节点而耦合在一起，各层的节点具有不同的属性，并且各层之间的节点存在耦合关系，一般分为相互依赖和相互协作两种关系。例如，在线购物交易平台依赖于因特网，因特网又依赖于电力网；公路网和铁路网组成的双层协作网络，两者相互协作保障了人们出行的方便快捷。作为结果，一个网络中的信息传播可能出现在另一个网络扩散，并最终导致一个信息级联效应。

本文利用小鼠的已知疾病表型之间的关联关系、已知致病基因之间的关联关系和已知疾病表型与致病基因之间的关联关系，构建出表型−基因的双层耦合网络。在表型−基因的双层耦合网络中，上层为表型关联网络，下层为基因关联网络，上下网络之间通过表型与基因的关联关系进行耦合。

2.1 信息网络

信息网络[20]是一个带有对象类型的映射函数$\tau:\mathcal{V}\to\mathcal{A}$和链接类型映射函数$\phi:\mathcal{E}\to\mathcal{R}$的图$G=(\mathcal{V},\mathcal{E})$，其中每个对象$v\in\mathcal{V}$属于一个特定的对象类型$\tau(v)\in\mathcal{A}$，每个链接$e\in\varepsilon$属于一个特定的关系$\phi(e)\in\mathcal{R}$，如果两个链接属于同一个关系类型，那么这两个链接具有相同类型的开始对象和结束对象。

2.2 表型关联网络

表型关联网络是一种信息网络，可以定义为$N_P=(P,E_{PP},W_{PP})$，其中$P=\{p_1,p_2,\cdots,p_m\}$表示表型节点的集合，E_{PP}表示表型之间的关联关系，W_{PP}表示关联关系权重值，如果表型i与表型j有关联关系，则权重值为1，否则为0。表示如下：

$$W_{PP}(i,j)=\begin{cases}1 & (p_i,p_j)\in E_{PP}\\ 0 & (p_i,p_j)\notin E_{PP}\end{cases}$$

本文中表型关联网络需要的数据从MGI数据库中获取得到，表型关联网络示意图如图1所示。

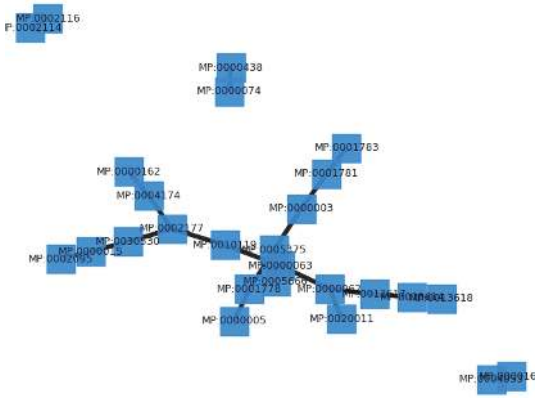

图1 表型关联网络示意图

2.3 基因关联网络

基因关联网络定义为$N_G=(G,E_{GG},W_{GG})$，其中$G=\{g_1,g_2,\cdots,g_n\}$表示基因节点的集合，E_{GG}表示基因之间的关联关系，W_{GG}表示关联关系权重值，基因i与基因j有关联关系则权重值为数据库中所给数值，用α表示，否则为0。表示如下：

$$W_{GG}(i,j)=\begin{cases}\alpha & (g_i,g_j)\in E_{GG}\\ 0 & (g_i,g_j)\notin E_{GG}\end{cases}$$

文中基因关联网络需要的数据从MouseNet下载，基因关联网络示意图如图2所示。

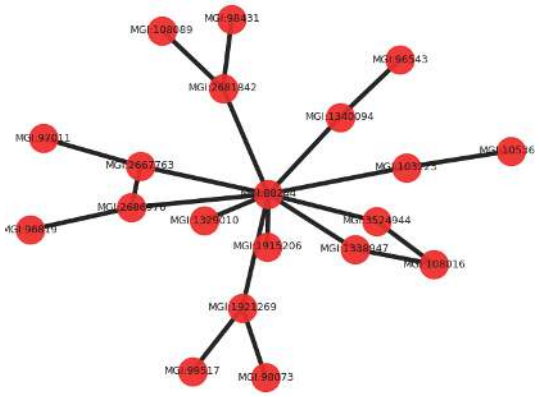

图2 基因关联网络示意图

2.4 表型−基因关联网络

表型−基因网络数据来源于MGI数据库，定义为：$N_{PG}=(P\cup G,E_{PG},W_{PG})$，其中：$P\cup G=\{p_1,p_2,\cdots,p_m,g_1,g_2,\cdots,g_n\}$表示表型和基因节点的集合，$E_{PG}$表示表型与基因之间的关联关系，$W_{PG}$表示关联关系权重值，如果表型$i$与基因$j$有关联关系则权重值为1，否则为0。表示如下：

$$W_{PG}(i,j)=\begin{cases}1 & (p_i,g_j)\in E_{PG}\\ 0 & (p_i,g_j)\notin E_{PG}\end{cases}$$

表型−基因关联网络示意图如图3所示。

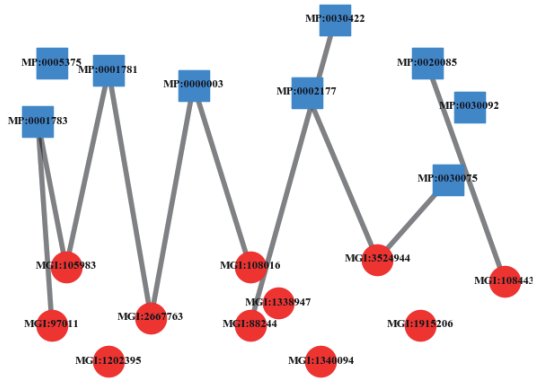

图3 表型-基因关联网络示意图

2.5 表型−基因双层耦合网络

表型−基因双层耦合网络就是在表型关联网络N_P、基因关联网络N_G和表型−基因关联网络N_{PG}基础上，上层为表型网络N_P，下层为基因网络N_G，

表型−基因关联网络N_{PG}节点间的关系作为上下层间的耦合关系而得到，可以定义为：$N_{P-G}=(V=P\cup G, E=E_{PP}\cup E_{PG}\cup E_{GG}, W=W_{PP}\cup W_{PG}\cup W_{GG})$，其中$V=P\cup G$表示包括表型与基因的所有节点，$E=E_{PP}\cup E_{PG}\cup E_{GG}$表示节点间的链接关系，$W=W_{PP}\cup W_{PG}\cup W_{GG}$表示节点链接关系的权重值，表型−基因双层耦合网示意图如图 4 所示。

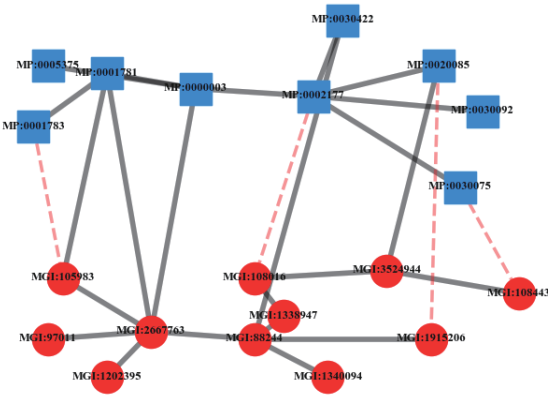

图 4　表型−基因双层耦合网示意图

图 4 中，实线部分为已知存在的关联关系，包括了表型与表型的关联、基因与基因的关联和表型与基因的关联；虚线部分为待预测的表型与基因的关系是否关联。

3　表型−基因双层耦合网上的随机游走

在 2.1 节定义的基础上，如果对象类型$|\mathcal{A}|>1$或者关系类型$|\mathcal{R}|>1$时，该信息网络为异构信息网络。从图 4 中可以看出在表型−基因双层耦合网$N_{P-G}=(V=P\cup G, E=E_{PP}\cup E_{PG}\cup E_{GG}, W=W_{PP}\cup W_{PG}\cup W_{GG})$中，表型关联网络$N_P$和基因关联网络$N_G$的节点分属两个类型，通过表型−基因关联网络$N_{PG}$进行耦合，整体上看表型−基因双层耦合网为一个异构网络。

3.1　表型−基因双层耦合网络上的元路径

元路径 (meta-path)[20] 主要用来描述异构网络中任意两个节点间的不同路径类型，可以定义为：在带有对象类型映射$\tau:\mathcal{V}\to\mathcal{A}$和链接类型映射$\phi:\mathcal{E}\to\mathcal{R}$的异构网络$G=(\mathcal{V},\mathcal{E})$的元模板上的一条路径，其形式为$A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots\to A_{l+1}$。元路径$A_1\xrightarrow{R_1}A_2\xrightarrow{R_2}\cdots\to A_{l+1}$描述了类型$A_1$到类型$A_{l+1}$间的复合关系$R=R_1\circ R_2\circ\cdots\circ R_l$，其中“$\circ$”表示关系上的复合运算。

在表型−基因双层耦合网络N_{P-G}中两个节点之间就存在不同类型不同长度的元路径，以图 4 为例，可以有$P\to P\to G$、$P\to P\to G\to G$、$P\to P\to P\to P\to G$等。对于一个给定的网络，可能存在的元路径数目与路径长度成指数增长[21]。选择不同的元路径，表型与基因之间的关联性也不同，同时，文献 [20] 指出很长的元路径并不是很有意义，反而路径长度越大，关系越弱，预测也越模糊。因此，在表型与基因的关联预测中，本文主要考虑如下 4 条元路径，如表 1 所示。

表 1　元路径表

序号	元路径
MP_1	$P\to P\to G$
MP_2	$P\to G\to G$
MP_3	$P\to P\to G\to G$
MP_4	$P\to G\to P\to G$

3.2　基于元路径的随机游走

随机游走 (random walk) 又称随机游动或随机漫步，是一种数学统计模型，在金融、物理和社交媒体等复杂网络分析中都有广泛应用。随机游走模型是从图上一个或一组节点开始，通过迭代随机的访问图中的每一个节点。每一次移动时，当前节点都以一定的概率移动到他们的邻居节点。因此，图中每个节点都会获得一个经计算得到的当前节点游走到该节点的概率分布值[22]。文献 [23] 提出了基于双层耦合网络的随机游走 RWRH 算法。RWRH 算法在不同的网络中游走，从网络G_1或者网络G_2的某一节点开始进行随机游走，在游走过程中，以一定的概率停留在网络G_1的下一个节点或者网络G_2的一个节点。

在表型−基因双层耦合网络N_{P-G}中选定了元路径，随机游走将基于元路径进行游走，但是，游走到元路径中指定类型节点中的哪一个节点是未知的，即规定了下一步游走的节点类型但不固定某个节点。那么，表型−基因双层耦合网络N_{P-G}中节点在既定的元路径$P\to P\to G$、$P\to G\to G$、$P\to P\to G\to G$和$P\to G\to P\to G$下由上一个节点游走到下一个节点的跳转概率有如下 4 种表示：

$$p\left(v^{i+1}|v^i,MP_i\right)=\begin{cases}\dfrac{W_{PP}(i,i+1)}{\sum\limits_{k=1}^{m}W_{PP}(i,k)} & v^i\in P,\quad v^{i+1}\in P\\ \dfrac{W_{PG}(i,i+1)}{\sum\limits_{k=1}^{m}W_{PG}(i,k)} & v^i\in P,\quad v^{i+1}\in G\\ \dfrac{W_{PG}(i,i+1)}{\sum\limits_{k=1}^{n}W^{\mathrm{T}}_{PG(i,k)}} & v^i\in G,\quad v^{i+1}\in P\\ \dfrac{W_{GG}(i,i+1)}{\sum\limits_{k=1}^{n}W_{GG}(i,k)} & v^i\in G,\quad v^{i+1}\in G\end{cases}$$

式中，i表示第i步跳转。

将上式用矩阵形式表示如下：

1) 当$v^i \in P$，$v^{i+1} \in P$，则一步跳转概率矩阵为$\boldsymbol{D}_{PP}^{-1}\boldsymbol{W}_{PP}$；

2) 当$v^i \in P$，$v^{i+1} \in G$，则一步跳转概率矩阵为$\boldsymbol{D}_{PG}^{-1}\boldsymbol{W}_{PG}$；

3) 当$v^i \in G$，$v^{i+1} \in P$，则一步跳转概率矩阵为$\boldsymbol{D}_{GP}^{-1}\boldsymbol{W}_{GP}, \boldsymbol{W}_{GP} = \boldsymbol{W}_{PG}^{\mathrm{T}}$；

4) 当$v^i \in G$，$v^{i+1} \in G$，则一步跳转概率矩阵为$\boldsymbol{D}_{GG}^{-1}\boldsymbol{W}_{GG}$。

其中，$\boldsymbol{D}_{PP}$、$\boldsymbol{D}_{PG}$、$\boldsymbol{D}_{GP}$、$\boldsymbol{D}_{GG}$为对角矩阵，对角线上的值分别为$\boldsymbol{W}_{PP}$、$\boldsymbol{W}_{PG}$、$\boldsymbol{W}_{GP}$、$\boldsymbol{W}_{GG}$中对应行元素之和，即：

$$D_{PP}(i,j)=\begin{cases}\sum_{k=1}^{m} W_{PP}(i,k) & i=j(1\leqslant i,j\leqslant n)\\ 0 & i\neq j\end{cases}$$

$$D_{PG}(i,j)=\begin{cases}\sum_{k=1}^{n} W_{PG}(i,k) & i=j(1\leqslant i,j\leqslant m)\\ 0 & i\neq j\end{cases}$$

$$D_{GP}(i,j)=\begin{cases}\sum_{k=1}^{m} W_{PG}^{\mathrm{T}}(i,k) & i=j(1\leqslant i,j\leqslant n)\\ 0 & i\neq j\end{cases}$$

$$D_{GG}(i,j)=\begin{cases}\sum_{k=1}^{n} W_{GG}(i,k) & i=j(1\leqslant i,j\leqslant m)\\ 0 & i\neq j\end{cases}$$

因此，在表型-基因双层耦合网络$N_{P-G}=(V=P\cup G, E=E_{PP}\cup E_{PG}\cup E_{GG}, W=W_{PP}\cup W_{PG}\cup W_{GG})$中，基于元路径$MP_1: P\to P\to G$的表型$p_i$到基因$g_i$的跳转概率矩阵$\boldsymbol{X}_{PPG}$可表示为：

$$\boldsymbol{X}_{PPG}=\left[\boldsymbol{D}_{PP}^{-1}\boldsymbol{W}_{PP}\right]\left[\boldsymbol{D}_{PG}^{-1}\boldsymbol{W}_{PG}\right]$$

基于元路径$MP_2: P\to G\to G$的表型p_i到基因g_i的跳转概率矩阵$\boldsymbol{X}_{PGG}$可表示为：

$$\boldsymbol{X}_{PGG}=\left[\boldsymbol{D}_{PG}^{-1}\boldsymbol{W}_{PG}\right]\left[\boldsymbol{D}_{GG}^{-1}\boldsymbol{W}_{GG}\right]$$

基于元路径$MP_3: P\to P\to G\to G$的表型p_i到基因g_i的跳转概率矩阵$\boldsymbol{X}_{PPGG}$，可以表示为：

$$\boldsymbol{X}_{PPGG}=\left[\boldsymbol{D}_{PP}^{-1}\boldsymbol{W}_{PP}\right]\left[\boldsymbol{D}_{PG}^{-1}\boldsymbol{W}_{PG}\right]\left[\boldsymbol{D}_{GG}^{-1}\boldsymbol{W}_{GG}\right]$$

基于元路径$MP_4: P\to G\to P\to G$的表型p_i到基因g_i的跳转概率矩阵$\boldsymbol{X}_{PPGG}$可表示为：

$$\boldsymbol{X}_{PGPG}=\left[\boldsymbol{D}_{PG}^{-1}\boldsymbol{W}_{PG}\right]\left[\boldsymbol{D}_{GP}^{-1}\boldsymbol{W}_{GP}\right]\left[\boldsymbol{D}_{PG}^{-1}\boldsymbol{W}_{PG}\right]$$

4 表型–基因双层耦合网中节点的关联预测

对于表型–基因双层耦合网络$N_{P-G}=(V=P\cup G, E=E_{PP}\cup E_{PG}\cup E_{GG}, W=W_{PP}\cup W_{PG}\cup W_{GG})$中的任意的表型$p_i$和基因$g_j(i=1,2,\cdots,m; j=1,2,\cdots,n)$，如果$(p_i,\mathbf{g}_j)\notin E_{PG}$或者$W_{PG}(i,j)=0$，则需要对其关联性进行预测。

在综合不同元路径的情况下，按不同元路径所占权重进行累加，可以得到表型p_i到基因g_j之间的跳转概率矩阵$\boldsymbol{X}$：

$$\boldsymbol{X}=\alpha_{PPG}\boldsymbol{X}_{PPG}+\alpha_{PGG}\boldsymbol{X}_{PGG}+\alpha_{PPGG}\boldsymbol{X}_{PPGG}+\alpha_{PGPG}\boldsymbol{X}_{PGPG}$$

在得到的跳转概率矩阵$\boldsymbol{X}$中，其对应的取值就是表型p_i到基因g_i的关联值大小，值越大，关联越紧密；反之亦然。

5 实验结果与分析

5.1 数据说明

MGI 是实验室小鼠的国际数据库资源，包含：小鼠基因组数据库 (MGD)、基因表达数据库 (GXD)、小鼠肿瘤生物学 (MTB) 数据库、基因本体 (GO) 项目等。本文用到的表型数据和表型-基因数据集从 MGI 数据库资源下载获得。其中，表型数据集包含了 12 838 个疾病表型，构成了 16 108 对表型与表型关联对；表型-基因数据集共有表型与基因的关联数据对 37 246 对。

MouseNet V2 是许多生物医学研究选择的一种改进的实验小鼠功能基因网络。MouseNet V2 为 2008 年 MouseNet 的改进版本，加入了大量来自不同生物的新微阵列数据。MouseNet V2 现在覆盖 88% 的编码基因组，具有更高的准确性。本文使用基因数据即从 MouseNet V2 中获得，共有 17 710 个基因，构成了关联基因对 788 081 对。

5.2 实验步骤

在 4 条元路径$MP_1: P\to P\to G$、$MP_2: P\to G\to G$、$MP_3: P\to P\to G\to G$和$MP_4: P\to G\to P\to G$中进行随机游走得到了表型在 4 条元路径下游走到基因的跳转概率矩阵，即$\boldsymbol{X}_{PPG}$、$\boldsymbol{X}_{PGG}$、$\boldsymbol{X}_{PPGG}$和$\boldsymbol{X}_{PGPG}$。在所得到的$\boldsymbol{X}_{PPG}$、$\boldsymbol{X}_{PGG}$、$\boldsymbol{X}_{PPGG}$和$\boldsymbol{X}_{PGPG}$数据中，找出 4 个数据都同时存在的表型到基因的概率，在此前提下使用主成分分析的办法，即通过变量变换的方法把相关的变量变为若干不相关的综合指标变量，从而实现对数据集的降维，在

过程中求出综合评价函数而得到不同元路径下的权重值，即是$\boldsymbol{X}=\alpha_{PPG}\boldsymbol{X}_{PPG}+\alpha_{PGG}\boldsymbol{X}_{PGG}+\alpha_{PPGG}\boldsymbol{X}_{PPGG}+\alpha_{PGPG}\boldsymbol{X}_{PGPG}$中$\alpha_{PPG}$、$\alpha_{PGG}$、$\alpha_{PPGG}$和$\alpha_{PGPG}$的值。最后进行表型到基因在元路径下按权重累加，并选出前k名为最终结果，作为表型与基因关联关系的预测值。

5.3　算法验证

为了评价本文算法预测表型与基因关联关系的性能，采用留一交叉验证法 (leave-one-out cross validation, LOO) 实验。在数据的 N 个样本中，每次实验将一个样本作为测试集，剩下的 N–1 个样本作为训练集，直到所有的样本都被作为测试集，即得到 N 个模型，在此过程中利用接收者操作特征 (ROC) 曲线[24] 对预测性能进行评价，绘制截止时的真阳性率 (TPR、敏感性或召回) 与假阳性率 (FPR、1-特异性) 的关系曲线。

在 ROC 曲线绘制和 AUC 面积的计算时，使用到如下的定义：

$$\mathrm{TPR}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},\quad \mathrm{TNR}=\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}},$$

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}},\quad \mathrm{PPV}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})}}$$

$$F_1=2\frac{\mathrm{PPV}\times\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}}=\frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{EP}+\mathrm{EN}}$$

其中，条件正 (P)：数据中实际正案例数；条件负 (N)：数据中的实际负案例数；TP 和 TN 代表正确预测的真正和真负数量；FP 和 FN 代表错误预测的假阳性和假阴性。

将本文算法与其他 3 种相关预测算法 RWR[25]、LPIHN[26] 和 PRINCE[27] 进行测试比较。RWR 算法从已知的致病基因以相同的概率出发，随机走向邻居节点，当前后两次游走的概率向量相同或者前后两次游走的概率差值小于某个阈值时，认为游走达到平衡，然后将概率值从大到小排序，排名靠前的说明基因与疾病的相关性较大，认为该基因是该疾病的致病基因。LPIHN 是一种在异构网络上实现随机游走的方法。PRINCE 是一种基于对优先级函数的约束的全局方法，从某个查询疾病表型出发游走至整个网络，通过计算在基因节点邻居中与查询疾病关联的基因的优先次序后，合并相似性信息中分数高的基因作为致病基因。RWR 方法中的重启概率r经过多次试验，对试验结果影响不大，所以设置$r=0.5$；LPIHN 的参数根据[26] 文中提及参数值特设置如下：$\gamma=0.5$，$\beta=0.5$，$\delta=0.3$；PRINCE 的参数根据[27] 文中提及数值而设置如下：$\alpha=0.5$，$c=-15$，$d=\lg(9\,999)$，传播迭代次数为 10。所得结果如图 5 所示，其中 THIS 代表本文提出的算法。

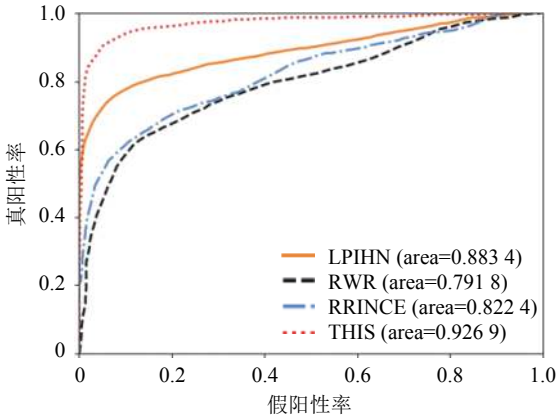

图 5　不同算法测试 ROC 曲线

结果表明，在所给数据实验中，本文提出的算法的AUC 得分为93%，高于RWR、LPIHN 和PRINCE 的 AUC 值，分别为 79%、88% 和 82%。

6　结 束 语

随着基因数据和表型数据的不断增加，为理解疾病与致病基因之间的关系提供了大量有效的数据，也为利用数据分析与挖掘的手段找出疾病表型与致病基因之间的关系提供了便利。为此，旨在设计一种算法来找到表型节点与基因节点的更多关联关系。本文在经典的随机游走方法上加入了元路径的概念，充分利用先验知识及网络中包含的生物关系来预测发现表型与基因的关联关系。从实验结果可以看出，本文算法的正确率高于 RWR、LPIHN 和 PRINCE 等算法，能够得到较好的预测效果。

在后续的工作中，有如下几方面可以做进一步研究：1) 整合更可靠的生物网络数据。生物信息知识的缺乏和实验数据的假阳性都会对实验的预测结果造成误差，整合其他有用的生物数据将会提高生物网络数据的可靠性。2) 整合多重生物网络数据。如将序列相似性、功能注释、微阵列表达、蛋白质域、通路成员等数据库整合为一个完整数据进行相应的预测。3) 改变生物网络的拓扑特征。可以适当改变网络的拓扑特征，如介数中心性、紧密中心性、聚类系数等，再进行关联预测。

参 考 文 献

[1] REBBECK T R, FRIEBEL T M, MITRA N, et al. Inheritance of deleterious mutations at both BRCA1 and BRCA2 in an international sample of 32, 295 women[J]. Breast Cancer Research, 2016, 18(1): 112.

[2] CHABON J J, SIMMONS A, LOVEJOY A F, et al. Circulating tumour DNA profiling reveals heterogeneity of EGFR inhibitor resistance mechanisms in lung cancer patients[J]. Nature Communications, 2016, 7(1): 11815-11815.

[3] LIU Y, ZENG X, HE Z, et al. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(4): 905-915.

[4] ZOU S, ZHANG J, ZHANG Z, et al. A novel approach for predicting microbe-disease associations by bi-random walk on the heterogeneous network[J]. PLOS ONE, 2017, DOI: 10.1371/joural.pone.0184394.

[5] SHEN X, ZHU H, JIANG X, et al. A novel approach based on bi-random walk to predict microbe-disease associations[M]//Intelligent Computing Methodologies. [S.l.]: Springer, 2018: 746-752.

[6] TIAN Z, GUO M, WANG C, et al. Constructing an integrated gene similarity network for the identification of disease genes[J]. Journal of Biomedical Semantics, 2017, 8(1): 27-41.

[7] CHEN L, YANG J, XING Z, et al. An integrated method for the identification of novel genes related to oral cancer[J]. PLOS ONE, 2017, DOI: 10.1371/joural.pone.0175185.

[8] LU C, YANG M, LUO F, et al. Prediction of lncRNA-disease associations based on inductive matrix completion[J]. Bioinformatics, 2018, 34(19): 3357-3364.

[9] SHEN Z, JIANG Z, BAO W. CMFHMDA: Collaborative matrix factorization for human microbe-disease association prediction[C]//International Conference on Intelligent Computing. [S.l.]: Springer, 2017: 261-269.

[10] 浦建宇, 陈蕾, 邵楷. 基于 Katz 增强归纳型矩阵补全的基因-疾病关联关系预测[J]. 计算机科学与探索, 2019(7): 1154-1164.
PU Jian-yu, CHEN Lei, SHAO Kai. Exploiting Katz method to boost inductive matrix completion for predicting gene-disease associations[J]. Journal of Frontiers of Computer Science and Technology, 2019(7): 1154-1164.

[11] CHEN X, YAN G. Novel human lncRNA-disease association inference based on lncRNA expression profiles[J]. Bioinformatics, 2013, 29(20): 2617-2624.

[12] CHEN X, YAN C C, ZHANG X, et al. HGIMDA: Heterogeneous graph inference for miRNA-disease association prediction[J]. Oncotarget, 2016, 7(40): 65257-65269.

[13] HUANG Z, CHEN X, ZHU Z, et al. PBHMDA: Path-based human microbe-disease association prediction[J]. Frontiers in Microbiology, 2017, 8(2): 233.

[14] LI X, LIN Y, GU C, et al. SRMDAP: SimRank and density-based clustering recommender model for mirna-disease association prediction[J]. BioMed Research International, 2018, DOI: 10.1155/2018/5747489.

[15] CHEN X, YAN C C, ZHANG X A, et al. WBSMDA: Within and between score for mirna-disease association prediction[J]. Scientific Reports, 2016, 6(1): 21106.

[16] WANG F, HUANG Z A, CHEN X, et al. LRLSHMDA: Laplacian regularized least squares for human microbe-disease association prediction[J]. Scientific Reports, 2017, 7(1): 7601.

[17] 郑经龙. 基于链路预测的长非编码 RNA-疾病关联预测方法[D]. 西安: 西安电子科技大学, 2015.
ZHENG Jing-long. Method for prediction of LncRNA-disease associations based on link prediction[D]. Xi'an: Xidian University, 2015

[18] 郭茂祖, 王诗鸣, 刘晓燕, 等. MiRNA 与疾病关联关系预测算法[J]. 软件学报, 2017, 28(11): 3094-3102.
GUO Mao-Zu, WANG Shi-ming, LIU Xiao-yan, et al. Algorithm for predicting the associations between MiRNAs and diseases[J]. Journal of Software, 2017, 28(11): 3094-3102.

[19] 唐明, 崔爱香, 龚凯. 关注耦合网络及其传播动力学研究[J]. 复杂系统与复杂性科学, 2011(2): 87-91.
TANG Ming, CUI Ai-xiang, GONG Kai. On spreading dynamics on coupled networks[J]. Complex Systems and Complexity Science, 2011(2): 87-91.

[20] SUN Y, HAN J, YAN X, et al. PathSim: Meta path-based Top-K similarity search in heterogeneous information networks[J]. Very Large Data Bases, 2011, 4(11): 992-1003.

[21] LAO N, COHEN W W. Relational retrieval using a combination of path-constrained random walks[J]. European Conference on Machine Learning, 2010, 81(1): 53-67.

[22] 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 Min, WANG Xiao-tong, LUO Hui-min, et al. Progress on random walk and its application in network biology[J]. Acta Electronica Sinica, 2018, 46(8): 2035-2048.

[23] LI Y, PATRA J C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network[J]. Bioinformatics, 2010, 26(9): 1219-1224.

[24] METZ C E. Basic principles of ROC analysis[J]. Seminars in Nuclear Medicine, 1978, 8(4): 283-298.

[25] KOHLER S, BAUER S, HORN D, et al. Walking the interactome for prioritization of candidate disease genes[J]. American Journal of Human Genetics, 2008, 82(4): 949-958.

[26] LI A, GE M, ZHANG Y, et al. Predicting long noncoding RNA and protein interactions using heterogeneous network model[J]. BioMed Research International, 2015, DOI: 10.1155/2015/671950.

[27] VANUNU O, MAGGER O, RUPPIN E, et al. Associating genes and protein complexes with disease via network propagation[J]. PLOS Computational Biology, 2010, DOI: 10.1371/journal.pcbi.1000641.

编 辑 税 红