PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2021,48(10):1184~1195
www.pibb.ac.cn

基于随机游走的风险致病基因预测研究进展*

刘丽丽[1,2] 张绍武[1)**]

([1)] 西北工业大学自动化学院，信息融合技术教育部重点实验室，西安 710072；[2)] 陕西师范大学物理学与信息技术学院，西安 710119）

摘要 风险致病基因预测有助于揭示癌症等复杂疾病发生、发展机理，提高现有复杂疾病检测、预防及治疗水平，为药物设计提供靶标. 全基因组关联分析（GWAS）和连锁分析等传统方法通常会产生数百种候选致病基因，采用生物实验方法进一步验证这些候选致病基因往往成本高、费时费力，而通过计算方法预测风险致病基因，并对其进行排序，可有效减少候选致病基因数量，帮助生物学家优化实验验证方案. 鉴于目前随机游走算法在风险致病基因预测方面的卓越表现，本文从单元分子网络、多重分子网络和异构分子网络出发，对基于随机游走预测风险致病基因研究进展进行较全面的综述分析，讨论其所存在的计算问题，展望未来可能的研究方向.

关键词 致病基因，随机游走，单元网络，多重网络，异构网络
中图分类号 R318.04 **DOI**：10.16476/j.pibb.2020.0376

研究发现，基因与疾病的产生密切相关，基因的突变与异常往往是导致疾病产生的重要因素[1]，几乎所有的疾病在某种程度上都受到人类遗传变异的影响[2]. 阿尔茨海默病、自闭症、癌症等复杂疾病通常是由一组基因失调引起的，这些失调基因被称为致病基因或者疾病关联基因[3]. 识别这些基因对于理解疾病机制至关重要，有助于揭示癌症等复杂疾病发生、发展机理，提高现有复杂疾病检测、预防及治疗水平，为药物设计提供靶标[4].

疾病基因预测实质上是一个优选问题，即在众多潜在基因中优选出最有可能与疾病关联的基因. 早期疾病基因预测方法主要包括连锁分析[5-9]和全基因组关联研究[10-14]. 连锁分析主要利用孟德尔性状和具有家族聚集的复杂性状基因定位预测疾病基因[8]，而全基因组关联分析（GWAS）则对全基因组范围内的遗传变异基因进行总体关联分析，能够一次性对疾病进行轮廓性概览[15]. 然而，GWAS和连锁分析通常会产生数百种候选致病基因，尤其GWAS主要依赖统计分析，产生较多的假阳性和假阴性结果，需进行大量的生物学实验验证. 采用生物实验方法进一步验证这些候选致病基因往往成本高、费时费力，因此需要发展新的计算方法对候选致病基因进行排序，减少候选致病基因数量，帮助生物学家进一步优化实验验证方案.

随着高通量生物技术的发展，大量的生物分子作用及组学数据快速涌现，这些分子作用数据通常可表现为复杂网络，其节点代表基因、蛋白质等分子，边表示相应分子间的作用关系[16]. 由于生物数据本身具有复杂性和高维性，从中挖掘疾病风险致病基因面临挑战. 现有的致病基因预测方法可分为机器学习方法、图表示学习方法和网络传播方法三类[17-23]. 其中机器学习方法是从输入的网络数据中提取疾病和基因的各种特征，然后输入传统的机器学习模型用于预测疾病基因的关联. 机器学习方法又分为有监督分类方法[24]和positive-unlabeled（PU）方法[25]，如dgMDL[24]采用node2vec分别从疾病相似性网络和基因相似性网络提取疾病和基因特征、从蛋白质-蛋白质作用（PPI）网络和GO语义相似性网络提取基因特征输入DBN深度学习模型，利用sigmoid激活函数预测致病基因. 由于目前没有实验验证的负样本（即非致病基因）数据，现有的大多数方法一般将已知致病基因之外的

* 国家自然科学基金(61873202)资助项目.
** 通讯联系人.
Tel：029-88431308，E-mail：zhangsw@nwpu.edu.cn
收稿日期：2020-10-22，接受日期：2020-12-14

基因作为非致病基因，从这些基因中随机选取构建负样本集合.但已知致病基因之外的基因并不一定是非致病基因，只是没有发现或实验验证.鉴于此，人们采用PU方法筛选负样本，以提高模型的致病基因预测性能.如PEGPUL[25]采用多层SVM、加权KNN和加权决策树共训练的PU策略，选择可信度较高的正样本和负样本，通过度量学习构建基因相似性网络，在此网络上执行多步行走标签传播预测风险致病基因.机器学习方法虽然能够成功应用于风险致病基因预测，但由于已知的致病基因数量有限，尤其某些疾病的已知致病基因非常少，甚至有些疾病目前没有已知的致病基因，无法构建包含一定数量的致病基因（即正样本）集合，因而该类方法不能有效预测特异疾病（即已知致病基因较少的疾病）的致病基因.另外，目前没有实验验证的非致病基因，因而也面临如何有效构建负样本集的困境.还有如何选择提取特征的问题，这又需要非常深厚的专业领域知识.

近年来，图表示学习方法[26]在生物信息学和生物医学中受到了广泛关注.图表示学习方法通过矩阵分解[27]、图嵌入[28]、图神经网络[29]以及混合方法[30]自动学习疾病和基因的潜在嵌入特征，而不利用人工进行特征选择预测风险疾病基因.如GeneHound[27]采用贝叶斯概率矩阵分解（BPMF）方法求解疾病基因排序问题，该算法将基因-疾病关联作为部分观察数据，并整合多个基因组数据源（包括基于文献的表型和基于文献的基因组信息）.然后，采用贝叶斯模型联合学习基因和疾病的潜在因素及相应的基因和疾病关联矩阵，预测疾病-基因关联.HeteWalk[28]通过引入PPI、miRNA相似度网络、疾病表型相似度网络等数据源，构建加权异构网络，采用元路径选择消除多个实体异构行走产生的潜在冗余和误导性信息，高精度预测风险致病基因.PGCN[29]方法基于图卷积网络，利用异质表型基因网络和节点附加信息（如疾病本体相似性、基因表达数据等），对致病基因进行排序.DW-GCN[30]结合DeepWalk和GCN，在异质疾病基因网络上进行基因-疾病关联预测.

网络传播方法[17-23]通过构建疾病相似性网络、基因相似性网络及基因-疾病异构网络，选取某种疾病的已知致病基因为种子节点，利用热扩散、标签传播、重启随机游走等方法，在基因-疾病异构网络中扩散/游走，对基因进行打分排序，进而预测潜在的致病基因.代表性网络传播方法主要包括RWR-MH[19]、InLPCH[17]等.RWR-MH[19]通过蛋白质相互作用网络、基因共表达网络和疾病表型相似性网络，构建蛋白质-基因-疾病三层异构网络，采用重启随机游走算法在三层异构网络游走，预测潜在致病基因.InLPCH[17]利用lncRNAs-疾病关系和lncRNAs-基因关系，构建lncRNAs-基因-疾病三层异构网络，通过信息丢失模型优化异构网络，在优化异构网络执行传播算法预测潜在的致病基因.网络传播方法可以避免有监督预测方法中的负样本（即非致病基因）集构建窘境，可有效预测特异疾病的致病基因.

鉴于随机游走算法在致病基因预测中得到了广泛的应用[31-32]，且预测效果比较突出，本文将从基因/疾病相关数据库、基因/疾病相似性度量、种子基因选取及不同网络下的风险致病基因预测方法等方面，对基于随机游走的风险致病基因预测研究进展进行综述分析.

1　基因/疾病相关数据库

1.1　蛋白质相互作用数据库

目前，已建立多个人类蛋白质相互作用（PPI）数据库，如HPRD数据库[33]、BioGrid数据库[34]、BIND数据库[35]、String数据库[36]、IntAct数据库[37]等，表1汇总了几种常用的PPI数据库.

目前实验证实的PPI数据只占实际存在的相互作用数据量的0.3%[38]，另外高通量实验产生的数据含有大量的假阳性和假阴性.

1.2　通路数据库

通路数据来自于京都基因和基因组数据库（KEGG）.KEGG是一个从分子水平信息，特别是由基因组测序和其他高通量实验技术生成的大规模分子数据集，从该数据集可以了解细胞、生物体和生态系统等生物系统的高级功能和效用，是生物系统的一种计算机表示，由基因和蛋白质的分子构建块（基因组信息）和化学物质（化学信息）组成，这些化学物质与相互作用、反应和关系网络的分子接线图知识（系统信息）集成在一起，且包括疾病和药物等信息[39].

1.3　疾病数据库

疾病相关数据库主要包括OMIM[40]、HPO[41]、DisGeNET[42]、DO[43]数据库、LncRNAdisease、lnc2Cancer、GeneRIF等，这些数据库中的疾病信息可以作为疾病基因预测的先验信息.表2为6个常用的疾病相关数据库.

Table 1 Databases of eight protein-protein interaction（PPI）

表1 蛋白质-蛋白质相互作用数据库

数据库	特点	网址
HPRD	包含基于文献中的实验证据的有关人类蛋白质的注释，包括了翻译后修饰、亚细胞定位、蛋白质结构域结构、组织表达和与人类疾病关联的信息. 除了蛋白质与其他蛋白质的相互作用，HPRD还报道蛋白质与核酸和小分子的相互作用	http://www.hprd.org/
String	收集、评分和整合所有公开的蛋白质-蛋白质相互作用信息来源，并通过计算预测来补充这些信息	https://string-db.org/
BIND	生物分子协会的数据库，分类为3类，二元分子的相互作用，分子复合物和途径	http://bind.ca/
BioGRID	记录、整理包括蛋白质、遗传和化学互作的数据，涵盖人类和所有主要的模式生物	https://thebiogrid.org/
IntAct	描述人类蛋白质的相互作用，实验方法和文献引用，以及从几种其他物种衍生的蛋白质	http://www.ebi.ac.uk/intact/
MINT	数据库中的蛋白相互作用都是由专家审核过的有实验证据支持的，主要是哺乳动物的蛋白质相互作用	https://mint.bio.uniroma2.it/
DIP	存储在DIP中的数据是通过文献检索的人工整理获得的，其中包括直接和复杂的交互作用	http://dip.doe-mbi.ucla.edu/
MIPS	包含从文献中手动整理的哺乳动物交互数据，包括实验类型、交互描述和交互伙伴的结合位点	http://mips.gsf.de/proj/ppi/

Table 2 Databases related with diseases

表2 疾病相关数据库

数据库	特点	网址
DO	对人类常见疾病和罕见病进行归纳整理，并提供了统一的、标准化的疾病分类系统	http://disease-ontology.org/
HPO	提供在人类疾病中遇到的表型异常的标准化词汇，目前包含超过13 000个术语和超过156 000个关于遗传疾病的注释	https://hpo.jax.org/app/
OMIM	人类基因和遗传表型的数据库，包含人类疾病相关的基因、遗传特征等详细描述	https://www.ncbi.nlm.nih.gov/omim/
DisGeNet	收集了大量与人类疾病相关的突变位点和基因数据集	http://www.disgenet.org/
HMDD	收集了疾病与miRNA关联关系	http://www.cuilab.cn/hmdd/
MalaCards	综合了各大数据库，可搜索人类疾病及其注释	https://www.malacards.org/

1.4 基因本体GO数据库

基因本体（GO）数据库是世界上最大的基因功能信息来源，也是生物信息学研究中大规模分子生物学和遗传学实验计算分析的基础[44-45]. 它对不同数据库中关于基因和基因产物的生物学术语进行标准化，对基因和蛋白质的功能进行统一的限定和描述. GO数据从生物过程（BP）、分子功能（MF）和细胞成分（CC）三个方面对基因及其产物进行分类注释.

2 基因/疾病相似性度量

利用随机游走预测致病基因的实质就是在网络中扩散已知疾病基因的影响，对网络中的每个基因进行打分，根据基因打分值对所有基因进行排序，将分值较高的基因视为候选疾病基因. 而相似度度量是建立相似网络的基础，目前相似性度量可分为如下二类：一类为基于术语间层次关系的相似性度量，另一类为结合其他统计信息的相似性度量. 下面介绍一些常用的相似性度量.

2.1 基于语义的相似性度量

通过对基因/疾病本体论数据库中的两个集合中的条目进行比较，分析基因或疾病间的相似性，这种基于本体表型计算的相似性，被称为语义相似性（semantic similarity）. 分析本体条目间语义相似性的度量方法大致可分为以下3类：基于边的方法、基于节点的方法和混合方法. 基于边的方法主要以本体有向无环图（directed acyclic graph，DAG）拓扑结构中的边为数据源，重点分析条目间的关系. 基于节点的方法主要以本体有向无环图拓扑结构中的点为数据源，重点分析条目自身的属性. 混合方法是以上两种方法的结合.

2.1.1 基于节点的度量指标

基于节点的方法依赖于比较所涉及的术语属性，这些属性可能与术语本身、它们的祖先或后代相关. 该方法通常使用信息量（*IC*）描述节点的特殊性和所包含的信息，根据条目c在语料库中出现

的概率度量它所提供的信息量. *IC* 定义为 $IC(c) = -\lg p(c)$，其中 $p(c)$ 是条目 c 所处的子本体全部条目所注释的基因/疾病中被 c 及其后代条目注释的概率. 基于节点的度量指标主要包括 Resnik 指标[46]、Lin 指标[47]、Jiang-Conrath 指标[48]、Relevance 指标[49] 和 Information Coefficient 指标[50].

Resnik 指标通过两个条目所有最低共同祖先中信息量最大的共同祖先节点（*MICA*）的信息量进行定义，即 $sim_{\text{Resnik}}(c_1,c_2)=IC(c_{MICA})$[46]. 该指标仅考虑条目的共性，忽略了结构信息.

Lin 指标是 Resnik 的改进指标，定义为 $sim_{\text{Lin}}(c_1,c_2)=2\times IC(c_{MICA})/(IC(c_1)+IC(c_2))$，该指标不仅考虑了条目之间的共性，还考虑了差异性[47].

Jiang-Conrath 指标将基于边的距离概念和基于点的信息量相结合，定义为 $sim_{\text{JC}}(c_1,c_2)=1-(IC(c_1)+IC(c_2)-2\times IC(c_{MICA}))$[48].

Relevance 指标在 Resnik 和 Lin 指标基础上发展而来，考虑了两个条目到其最低共同祖先的距离以及最低共同祖先包含的信息量，定义为 $sim_{\text{Rel}}(c_1,c_2)=sim_{\text{Lin}}(c_1,c_2)\times(1-p(c_{MICA}))$[49].

信息指标综合考虑了信息量和结构信息对相似性的影响，定义为 $sim_{\text{IC}}(c_1,c_2)=sim_{\text{Lin}}(c_1,c_2)\times(1-1/(1+IC(c_{MICA})))$[50].

2.1.2　基于边的度量指标

该方法基于两个节点 (A,B) 的最低共同祖先 C 和根节点 *Root* 之间的最长路径长度（即最大共同祖先深度），及每个节点和共同祖先之间的最长路径长度而设计，定义为 $sim_{\text{PS}}(A,B)=\delta(Root,C)/(\delta(A,C)+\delta(B,C)+\delta(Root,C))$，其中 $\delta(A,B)$ 是节点 A 和 B 之间最长路径上的边数，Yu 等[51] 首次将该指标应用于度量 GO 相似性.

2.1.3　混合方法

该方法为一种同时考虑节点信息和边信息的度量方法. 以 GO 数据库为例，给定基因 g_i 可描述为 $DAG(g_i)=(D(g_i),E(g_i))$，其中，$D(g_i)$ 为 *DAG* 中节点 g_i 及其所有父节点对应的基因集合，$E(g_i)$ 为 *DAG* 子节点的边集合. 基因 g_i 和 g_j 间的语义相似性 $S^G(g_i,g_j)$ 定义为[52]：

$$S^G(g_i,g_j)=\frac{\sum_{t\in D(g_i)\cap D(g_j)}(C_{g_i}(t)+C_{g_j}(t))}{C(g_i)+C(g_j)} \quad (1)$$

$$C(g_i)=\sum_{t\in D(g_i)}C_{g_i}(t),$$

$$\begin{cases} C_{g_i}(t)=1, \text{ if } t=g_i \\ C_{g_i}(t)=\max\{w*C_{g_i}(t')|t'\in children\ of\ t\} & t\neq g_i \end{cases} \quad (2)$$

其中，$C_{g_i}(t)$ 表示 $DAG(g_i)$ 中基因 t 对基因 g_i 的贡献度，w 表示贡献度的衰减因子.

利用上述度量方法得到表型语义相似性后，需要进一步计算基因/疾病之间的语义相似性. 由于一个基因/疾病是被多个条目注释的，因此需要获取两个基因/疾病在本体中注释到的多个条目所构成的条目集合，并计算两个条目集合之间的相似性矩阵，其条目集合相似性计算方法主要包括：“Max”方法[53]、“Mean”方法[53]、“funSim Max”方法[49]、“funSim Avg”方法[49] 和 “BMA”方法[52] 等.

2.2　高斯作用谱核相似性度量

从基因（或疾病）与 lncRNA、miRNA 和疾病（或基因）的关联数据出发，利用高斯作用谱可计算出基因/疾病间的相似性. 例如从基因-miRNA 关联数据出发，介绍如何计算基因间的高斯作用谱核相似性. 首先将已知的基因-miRNA 关联数据转换成基因的互作用谱 $IP(g)$，$IP(g)$ 是一个二进制向量，其维数为已知的基因-miRNA 关联所涉及到的 miRNA 个数，如果基因 g_i 与第 k 个 miRNA 存在关联关系，则 $IP(g)$ 的第 k 维为 1，否则为 0；然后根据公式 $K^G(g_i,g_j)=\exp(-\kappa_g\left\|IP(g_i)-IP(g_j)\right\|)$ 计算基因 g_i 和 g_j 间的相似性，其中，κ_g 表示高斯核的带宽，N_g 表示基因的个数. 高斯作用谱核相似性度量能够利用基因/疾病关联数据，获取疾病/基因之间的拓扑相关性或者非线性关系，以及潜在的基因/疾病间关联关系. 该方法最早用于药物-靶点之间相互作用的预测[54]，随后 Chen 等[55-56] 将该方法应用到 miRNA、lncRNA 与疾病关系的预测.

2.3　疾病表型相似性度量

人类表型本体论（HPO）提供了人类疾病中遇到的表型异常标准化词汇. 构建疾病网络时，之前的疾病表型相似性计算方法仅考虑疾病表型信息，而忽略疾病表型之间的结构信息，一些表型间的相关程度大于其他表型，且一般认为共享罕见表型的疾病间的相似性大于共享常见表型的疾病间的

相似性. 于是，Westbury 等[57]综合考虑疾病表型信息和结构信息，提出一种基于信息量的方法度量疾病表型相似性. 即，首先计算HPO数据库中每个表型的相对信息量$IC(i) = -\lg(f_i)$，f_i表示一组HPO疾病中表型i的频率，然后计算二个表型间的相似性$sim(i, j) = \max\limits_{t \in anc(i) \cap anc(j)} IC(t)$，$anc(i)$表示本体图中表型$i$的祖先；疾病$d_k$和疾病间的相似性得分可通过它们共享表型的总$IC$求出，即

$$S^D(d_k,d_l) = \frac{1}{|d_k|}\sum_{i \in d_k}\max_{j \in d_l}(sim(i,j) + \frac{1}{|d_l|}\sum_{j \in d_l}\max_{i \in d_k}(sim(i,j) \quad (3)$$

其中，$|d_k|$表示描述疾病d_k的表型总数目.

3　致病种子基因筛选

基于随机游走算法预测疾病的致病基因，需要选取某种疾病的已知疾病基因作为致病种子基因. 对于已知致病基因较少（甚至没有已知致病基因）的某种疾病，将待查疾病作为种子节点，在构建的疾病网络中采用随机游走算法找出得分较高的疾病，作为与待查疾病最相似的疾病，选取得分较高疾病的致病基因作为种子基因，种子基因选取5~30个，预测效果较好. 一般来说，若种子基因太少，则不能提供足够的信息预测潜在疾病基因，选取的数目太多会导致网络预测的生物信息异构化，从而无法正确预测或者反映出实际的疾病信息[58].

4　不同网络结构下的致病基因预测

迄今为止，尽管大多数基于随机游走的致病基因预测方法多集中在单一网络结构上，但整合多个网络可以提高致病基因预测精度[31, 57, 59]. 随机游走不仅放大了基因与表型的弱关联，也放大了不同信息源（如不同分子物种或不同患者）之间的弱相似性. 随着组学数据的不断增长，随机游走算法在不同的网络结构上预测疾病基因，也就是，从早期的单一网络结构（或同构网络）（图1a），到异构网络（图1b），多重网络 （图1c）及多重异构网络（图1d）. 下面我们在单一网络结构、多重网络结构及多重异构网络上分别介绍如何利用随机游走预测疾病的风险治病基因.

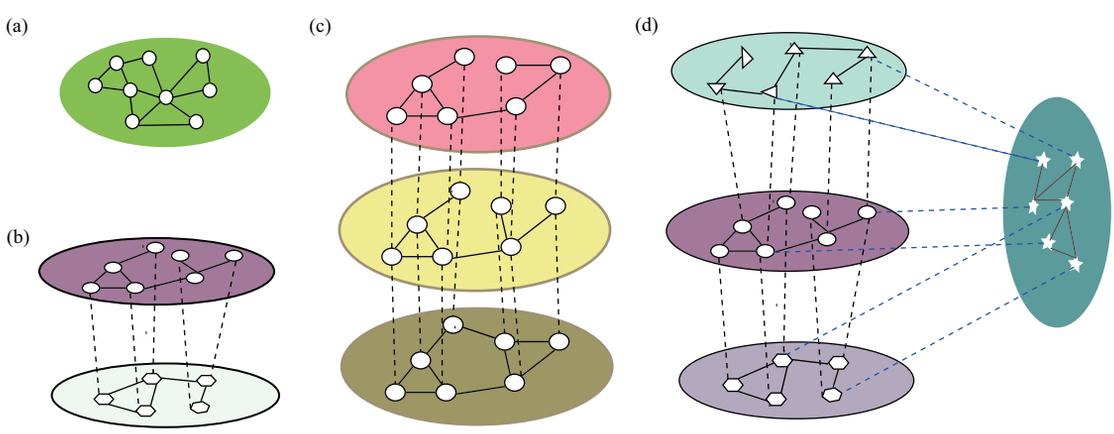

Fig. 1　Schematic diagram of different network structures

图1　网络结构示意图

（a）单一网络.（b）由两个网络组成的异构图，根据两种不同类型节点之间的二部图关联，粒子可以在每层网络中随机游走或跳转到另一层网络.（c）由三层网络组成的多重图，粒子可以在每一层网络中游走或跳转到另一层网络的同一节点上.（d）多重异构网络.

4.1　单一网络结构下的致病基因预测

Kohler 等[2]首先提出将候选基因和已知疾病基因映射到同一个蛋白质相互作用网络中，根据蛋白质相互作用网络求取归一化转移矩阵W，从已知疾病基因开始，通过重启随机游走在蛋白质相互作用网络中迭代更新给候选基因打分，并认为分数越高的候选基因越有可能为潜在疾病基因. Chen 等[60]提出一种K步随机游走算法，通过K步简单

随机游走排序致病基因，预测新的致病microRNA基因. Le等[61]考虑到疾病基因在蛋白质相互作用网络中的模块化结构特性，采取给模块化的疾病基因加权方式，提出一种ORIENT算法预测潜在的疾病基因. 这些方法由于忽略了生物大分子之间的物理和功能关系的丰富多样信息，其预测精度有待进一步改善.

4.2　异构网络结构下的致病基因预测

为弥补单一生物学数据库信息容量小的缺点，研究人员整合蛋白质数据、基因组数据、本体数据、疾病表型数据、功能注释数据等生物学数据，构造异构网络，在此异构网络执行重启随机游走可有效改善疾病致病基因预测精度. 异构网络一般由两层以上的网络构成，每层网络的节点都有各自的属性（即不同的类别）.

Li和Patra[62]提出了一种基于异构网络的RWRH算法预测风险致病基因，其异构网络由PPI网络、疾病表型网络和蛋白质-疾病表型关联二部图组成. 粒子可以在PPI网络上游走，可以在疾病表型相似网络上游走，也可以在蛋白质-疾病表型关联二部图跳跃. RWRH算法从多个来自网络P或者网络G的种子节点开始在异构网络上进行随机游走. RWRH算法异构网络转移矩阵W定义为：

$$W = \begin{bmatrix} W_G & W_{GP} \\ W_{PG} & W_P \end{bmatrix} \tag{4}$$

式（4）中，W_G、W_P分别为疾病表型相似性矩阵G和蛋白质作用网络邻接矩阵P的归一化转移矩阵，W_{PG}和W_{GP}表示疾病-基因关联矩阵的归一化矩阵. 设λ为随机游走者从基因相似性（或蛋白质作用）网络跳到疾病表型相似性网络的概率. 设u_0和v_0分别代表基因相似性（或蛋白质作用）网络基因和疾病表型相似性网络的初始概率向量. 基因相似性（或蛋白质作用）初始概率向量u_0由基因相似性（或蛋白质作用）里的种子节点以等概率构建，也就是u_0里的各个概率之和为1. 类似地可得到疾病表型相似性网络的初始概率. 因此，基因-疾病表型异质网的初始向量可表示为$P^0 = \begin{bmatrix} (1-\eta)u_0 \\ \eta v_0 \end{bmatrix}$，参数$\eta \in (0,1)$用来衡量每个子网的重要性. 如果$\eta = 0.5$，则两个子网的权重相同；如果$\eta > 0.5$，则随机游走者更倾向于重启到疾病表型种子节点，疾病表型相似性网络的权重更大.

RWRH算法采用如下重启游走算法公式预测疾病风险致病基因.

$$P^{t+1} = (1-\gamma)P^t W + \gamma P^0 \tag{5}$$

其中，P^t为t步迭代时的节点概率，参数γ用来调节随机游走概率.

RWRH算法同时为基因和疾病表型节点排序，采用留一交叉检验方法评估算法预测能力. 相对于RWR算法[2]，RWRH算法的AUC值较大，且该算法确定了18个基因-疾病关联关系，多数有文献证据支持.

在RWRH算法基础上，许多科研人员后续在构造异质网的数据源和构造方式上对其进行改进[17，19，63-68]，如Luo等[69]解决RWRH算法中的异质网包含一些虚假相互作用的问题，提出一种RWRHN方法；Lei等[17]引入lncRNA数据，构建一个异构三部图，减少预测结果假阳性. Zhao等[66]使用拉普拉斯归一化技术在构造基因-疾病表型异构网之前，对基因相似性矩阵和疾病表型相似性矩阵进行归一化，并对构造的基因-疾病表型异构网络的状态转移矩阵再次进行拉普拉斯归一化处理，消除中心节点的不良影响，以进一步提高预测准确度.

4.3　多重网络结构下的致病基因预测

单一基因组数据来源容易产生偏差、不完整和噪声，忽略了生物大分子之间的物理和功能关系的丰富多样信息. 为实现可靠的疾病基因预测，需要整合不同的基因组数据源，整合来自含噪声的单一数据源的候选疾病基因列表，可能会增加数据中的不确定性，为此，可采用多重（层）网络框架预测风险致病基因. 多重（层）网络指的是共享相同节点但其中的边属于不同类别或表示不同性质的相互作用网络[70-71]，多层网络可表示为L个无向图集合，网络共享同一组n个节点，每层包含不同种类的基因或蛋白质之间的物理和功能相互作用.

设多层网络的层数为L（α=1, ... L），每层邻接矩阵为$A^{[\alpha]}=(A^{[\alpha]}(i,j))_{i,j=1,\dots n}$，如果$\alpha$层中的节点$i$、$j$有连接，则$A^{[\alpha]}(i,j)$=1，否则为0. 若不考虑自动交互关系，即$A^{[\alpha]}(i,i)$=0, $\forall i$=1,...n，则多层网络邻接矩阵为$A=A^{[1]}, ..., A^{[L]}$，即

$$A=\begin{pmatrix}(1-\delta)A^{[1]} & \frac{\delta}{(L-1)}I & \cdots & \frac{\delta}{(L-1)}I\\ \frac{\delta}{(L-1)}I & (1-\delta)A^{[2]} & \cdots & \frac{\delta}{(L-1)}I\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\delta}{(L-1)}I & \frac{\delta}{(L-1)}I & \cdots & (1-\delta)A^{[L]}\end{pmatrix}_{nL\times nL} \quad (6)$$

其中I是单位矩阵，$A^{[\alpha]}$是第α层的邻接矩阵；参数$\delta\in[0, 1]$量化停留在每层或层间跳跃的概率，$\delta = 0$表示非重启后粒子将始终停留在同一层中.

如果用W表示由A得到的列归一化转移矩阵，则多层网络的随机游走公式表示为

$$\overline{P}_{t+1}^{T}=(1-\gamma)W\overline{P}_{t}^{T}+\gamma\overline{P}_{RS}^{T} \quad (7)$$

其中，$\overline{P}_{t}=[P_t^1, \ldots, P_t^L]$和$\overline{P}_{t+1}=[P_{t+1}^1, \ldots, P_{t+1}^L]$为$n\times L$维向量，表示粒子在多层网络中的概率分布，$\overline{P}_{RS}=[P_{RS}^1, \ldots, P_{RS}^L]$表示重启初始概率向量.

基于多重（层）网络重启游走框架，Li等[71]通过构建不同数据源对应独立的基因相似性网络，并将转移概率作为多个网络转移概率的期望值，提出RWRM算法预测风险致病基因. Valdeolivas等[19]构建由PPI网络、基因共表达网络及通路数据网络组成的三层基因网络，在每层上执行RWR算法，成功预测风险致病基因.

4.4 多重异构网络结构下的致病基因预测

Valdeolivas等[19]提出了一种多重异构网络重启随机游走RWR-MH扩展算法，该算法首先构建一个多重网络$G_{MH}=(V_{MH}, E_{MH})$，$V_{MH}=\{V_M\cup U\}$，$E_{MH}=\{U_{\alpha=1, \ldots, L}E_B^{[\alpha]}\cup E_M\cup E_U\}$，其中$V_M=\{v_i^{\alpha}, i=1, \ldots, n, \alpha=1, \ldots L\}$，$U=\{u_1, \ldots u_m\}$，$E_M$，$E_U$，$E_B$为多重网络$G$，网络$U$和二部图的边集合. 该多重网络由$L$层网络组成，每层包含$n$个基因（或蛋白质），其邻接矩阵为$A_{M(nL\times nL)}$（公式6）；然后构建疾病表型相似性网络，其邻接矩阵为$A_{D(m\times m)}$，m表示疾病数目；通过二部图邻接矩阵$B_{(n\times m)}^{1, \ldots L}$把多重网络中，每层基因相似性网络中的基因（或蛋白质）节点与疾病表型相似性网络中的疾病节点链接起来，构建多重异构网络. 每层基因相似性网络与疾病表型相似性网络间的二部图邻接矩阵是相同的，定义为$B_{(n\times m)}$. 因而，多重异构网络中的二部图邻接矩阵可定义为$B_{MH}=\left[B_{(n\times m)}, \cdots, B_{(n\times m)}\right]^T$，多重异构网络全局邻接矩阵可定义为$A=\begin{bmatrix}A_M & B_{MH}\\ B_{MH}^T & A_D\end{bmatrix}$，定义起始概率向量为$P_{RS}=\begin{bmatrix}(1-\eta)u_0\\ \eta v_0\end{bmatrix}$，其中$u_0$为多重基因相似性网络起始概率分布，$v_0$为疾病相似网络初始概率分布. 最后，RWR-MH算法在此多重异构网络执行重启游走高精度预测风险致病基因，且RWR-MH算法鲁棒性较好.

表3对目前不同网络结构下的基于随机游走预测风险致病基因预测方法进行了总结，简述了每种算法原理、优点及局限性.

5 总结与展望

随着高通量技术的进一步发展，各种组学数据急速增长，如何充分利用这些数据预测风险疾病基因仍需进一步开展深入研究. 目前基于随机游走预测风险致病基因可考虑从以下两个方面展开研究：

a. 发展有效的数据融合方法. 目前构建生物网络时，异构数据源融合方法简单，融合后的数据缺乏足够的生物学依据[18]，因此选择合适的策略整合不同数据源，整合过程中如何为不同的数据源赋予合适的权重，以提高数据和背景网络的可信度和覆盖率，获得对疾病发展过程中不同因素影响的全面分析，是一个值得关注的研究方向.

b. 疾病种子基因筛选策略. 随机游走算法中，疾病种子基因的选取对预测精度的提高非常关键，尤其对于已知致病基因较少甚至没有已知致病基因的疾病，如何找到与其相似度较高的疾病，利用该相似度较高的致病基因作为待预测疾病的致病种子基因，也是一个我们需要关注的研究方向.

风险致病基因的预测和发现是生物医学和生物信息学的基本问题，在大量候选基因中识别出最有可能的风险疾病基因，为生物学家、药物研发及临床治疗提供支持和帮助. 本文综述分析了致病基因的预测方法和发展现状，重点阐述了随机游走算法在不同网络结构下的风险致病基因预测原理及其研究进展，并对基于随机游走风险疾病基因预测研究的未来研究方向进行了展望.

Table 3　Algorithms for predicting the risk disease genes with random Walk on different network architectures

表3　不同网络结构下基于随机游走的风险致病基因预测算法总结

网络结构	算法	原理	优点	局限性	参考文献
单一网络	Kohler *et al*.算法	从Entrez Gene和STRING下载数据构建PPI网络，将候选基因和已知疾病基因映射到同一个PPI网络，根据PPI网络求取归一化转移矩阵，从已知疾病基因开始，通过重启随机游走在蛋白质相互作用网络中迭代更新给候选基因打分，分数越高的候选基因其潜在疾病基因可能越高	考虑了交互网络的全局结构，在确定候选疾病基因的优先级方面具有明显的性能优势	数据源单一，忽略部分作用信息	［2］
	ORIENT	构建权重网络和PPI网络，并对已知疾病基因的邻居基因进行加权强化权重，通过重启随机游走在加权的PPI网络中迭代更新打分，获得潜在的疾病基因	预测性能好，可以适用于任何类型的网络	算法质量依赖于数据库	［61］
	RWRMDA	构建miRNA功能相似度网络，从文献得到种子miRNA，将候选miRNA和种子miRNA映射到miRNA功能相似度网络，执行RWR推断潜在的miRNA疾病相互作用	较高预测能力，得到实验验证的miRNA疾病关联关系	整合更多的数据源衡量两个miRNA之间的功能相似性，以构建可靠的相似性网络	［64］
异构网络	RWRH	利用HPRD数据构建PPI网络，利用MimMiner计算表型相似性构建表型相似性网络，从OMIM数据库得到基因-表型关系，形成异构网络，在基因和表型网络上同时执行RWR，获得潜在的基因和表型关系	较强的鲁棒性	依赖异构网络的拓扑结构，低质量网络可能限制其性能	［62］
	Jiang *et al*.算法	基因本体（GO）的生物过程域构建基因语义相似度网络，然后使用疾病表型相似网络推断与查询疾病相关的基因	比PPI网络有更高的基因覆盖率，覆盖率能够提高预测准确性	仅利用了GO中生物过程域条目，应该综合HPRD数据考虑相似性	［72］
	lnLPCH	利用lncRNAs-疾病关系和lncRNAs-基因关系，构建lncRNAs-基因-疾病三层异构网络，通过信息丢失模型优化异构网络，在此优化异构网络执行传播算法预测潜在的致病基因	有效地减少了二元网络预测疾病基因时的误报率，提高了预测精度	应融合多源生物信息数据，进一步丰富网络信息	［17］
	RWRHN	利用RWS和皮尔逊相关性重构PPI网络，重构网络融入拓扑相似性；融合重构的PPI网络、表型相似性网络以及已知疾病与基因之间的二部关联，构建了一个可靠的异构网络，执行RWR获得候选基因的打分排序	高的预测率，稳定性能好	计算方法高度依赖蛋白质相互作用数据的质量	［69］
	Liu & Luo算法	构建基于数据集成和拉普拉斯归一化的疾病-蛋白异构网络；在该网络上应用RWRH算法来预测蛋白质与查询疾病之间的关联；将成员蛋白的分数相加，得到每个候选蛋白复合物的总分，根据分数对所有候选蛋白复合物进行排序	将数据集成与拉普拉斯归一化技术相结合，增强种子节点之间的权值；预测性能良好合理	集成不同数据源的权重分配需手动设置	［63］
	Lin & Yang算法	利用HumanNet数据库建立PPI网络，将表型相似度信息映射到PPI网络中，构建表型特异性网络，在原始PPI网络权重上添加相应的表型相似度评分，建立相同表型基因之间的链接；利用RWR计算网络中候选基因到种子基因的概率，作为拓扑距离	较高的预测准确率	利用的表型相似性信息有限	［73］
	LapRWRH	在构建异质网络之前，利用拉普拉斯归一化思想对基因相互作用矩阵和表型相似矩阵进行归一化，并利用拉普拉斯归一化思想对转移概率矩阵进行归一化	重现已知的基因-表型关系方面性能明显较RWRH有了明显提升	有些已知的疾病基因关系没有被预测出来	［66］

续表3

网络结构	算法	原理	优点	局限性	参考文献
多重网络	Li & Li算法	利用HPRD数据库建立PPI网络，GO数据库建立三个功能相似性网络，然后将这四个网络融合成单一网络，从疾病的基准数据集选取种子基因，执行RWR获得候选基因的打分排序	精度高	没有给不同数据源分配适当的权重，仅根据当前网络的拓扑结构计算了转移概率	［71］
	Valdeolivas *et al*.算法	构建PPI网络、共表达网络、通路网络三层基因网络，利用已知疾病基因作为种子节点，然后在每层网络上执行RWR获得候选基因排序	较高的预测准确率	三个网络的重启权重分配应考虑各自网络的质量	［19］
	Liu *et al*.算法	构建PPI网络，获得基因的拓扑相似度；用HRSS算法获得GO数据的计算语义相似度；在PPI网络执行RWR算法时初始种子基因的概率由语义相似性得分确定，进而获得候选基因打分	结合基因语义相似度值设定RWR的初始概率；算法鲁棒性强	识别的排名靠前基因在不同的方法之间差异很大，这种现象不能根据额外的实验结果进行分析	［1］
多重异构网络	RWR-MH	构建PPI网络、共表达网络、通路网络三层基因网络形成多层网络，然后与疾病网络通过疾病-基因二部图网络关联形成多重异构网络，把已知疾病及基因作为种子在节点，执行RWR，对候选基因排序	预测精度高	重构网络的无法完全消除干扰数据，数据的偏好等对性能有影响	［19］
	Lin *et al*.算法	构建基因、化学物质（包括药物）和疾病之间所有六种边缘的多模异构网络；运用三种扩散算法确定基因-疾病关联关系	提高预测的覆盖率，特异性和敏感度未显著降低	增加低质量的网络信息将增加噪音和降低性能， 如何降低集成数据的噪声	［18］

参 考 文 献

[1] Liu B, Jin M, Zeng P. Prioritization of candidate disease genes by combining topological similarity and semantic similarity. J Biomed Inform, 2015, **57**:1-5

[2] Kohler S, Bauer S, Horn D, *et al*. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet, 2008, **82**(4):949-958

[3] Piro R M, Di Cunto F. Computational approaches to disease-gene prediction: rationale, classification and successes. FEBS J, 2012, **279**(5):678-696

[4] Luo P, Tian L P, Ruan J, *et al*. Disease gene prediction by integrating PPI networks, clinical RNA-seq data and OMIM data. IEEE/ACM Trans Comput Biol Bioinform, 2019, **16**(1):222-232

[5] Altshuler D, Daly M J, Lander E S. Genetic mapping in human disease. Science, 2008, **322**(5903):881-888

[6] Bailey-Wilson J E, Wilson A F. Linkage analysis in the next-generation sequencing era. Hum Hered, 2011, **72**(4):228-236

[7] Lander E S, Schork N J. Genetic dissection of complex traits. Science, 1994, **265**(5181):2037-2048

[8] Ott J, Wang J, Leal S M. Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet, 2015, **16**(5):275-284

[9] Pulst S M. Genetic linkage analysis. JAMA Neurol, 1999, **56**(6): 667-672

[10] Cantor R M, Lange K, Sinsheimer J S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. Am J Hum Genet, 2010, **86**(1):6-22

[11] Lee I, Blom U M, Wang P I, *et al*. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res, 2011, **21**(7):1109-1121

[12] Pers T H, Hansen N T, Lage K, *et al*. Meta-analysis of heterogeneous data sources for genome-scale identification of risk genes in complex phenotypes. Genet Epidemiol, 2011, **35**(5): 318-332

[13] Santos-Cortez R L, Lee K, Giese A P, *et al*. Adenylate cyclase 1 (ADCY1) mutations cause recessive hearing impairment in humans and defects in hair cell function and hearing in zebrafish. Hum Mol Genet, 2014, **23**(12):3289-3298

[14] van Vliet-Ostaptchouk J V, Onland-Moret N C, Van Haeften T W, *et al*. HHEX gene polymorphisms are associated with type 2 diabetes in the Dutch Breda cohort. Eur J Hum Genet, 2008, **16**(5): 652-656

[15] Yan J, Takahashi T, Ohura T, *et al*. Combined linkage analysis and exome sequencing identifies novel genes for familial goiter. J Hum Genet, 2013, **58**(6):366-377

[16] Cowen L, Ideker T, Raphael B J, *et al*. Network propagation: a universal amplifier of genetic associations. Nat Rev Genet, 2017, **18**(9):551-562

[17] Lei X, Zhang Y. Predicting disease-genes based on network information loss and protein complexes in heterogeneous network.

Inform Sciences, 2019, **479**:386-400

[18] Lin C H, Konecki D M, Liu M, *et al*. Multimodal network diffusion predicts future disease-gene-chemical associations. Bioinformatics, 2019, **35**(9):1536-1543

[19] Valdeolivas A, Tichit L, Navarro C, *et al*. Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics, 2019, **35**(3):497-505

[20] Vanunu O, Magger O, Ruppin E, *et al*. Associating genes and protein complexes with disease *via* network propagation. PLoS Comput Biol, 2010, **6**(1):e1000641

[21] Wang B, Mezlini A M, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods, 2014, **11**(3):333-337

[22] 王一斌, 程咏梅, 张绍武 . 基于节点-模块置信度及局部模块度双重约束挖掘前列腺癌候选疾病模块 . 生物化学与生物物理进展, 2015, **42**(4):375-389
Wang Y B, Cheng Y M, Zhang S W. Prog Biochem Biophys. 2015, **42**(4):375-389

[23] 王一斌, 程咏梅, 张绍武 . 基于扩展起始节点和加权融合策略预测肺癌风险致病基因 . 生物化学与生物物理进展, 2016, **43**(2):176-186
Wang Y B, Cheng Y M, Zhang S W. Prog Biochem Biophys. 2016, **43**(2):176-186

[24] Luo P, Li Y, Tian L P, *et al*. Enhancing the prediction of disease-gene associations with multimodal deep learning. Bioinformatics, 2019, **35**(19):3735-3742

[25] Jowkar G, Mansoori E G. Perceptron ensemble of graph-based positive-unlabeled learning for disease gene identification. Comput Biol Chem, 2016, **64**:263-270

[26] Cai H, Zheng V W, Chang K C. A comprehensive survey of graph embedding: problems, techniques, and applications. IEEE Trans Knowl Data Eng, 2018, **30**(9):1616-1637

[27] Zakeri P, Simm J, Arany A, *et al*. Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information. Bioinformatics, 2018, **34**(13): 447-456

[28] Xiong Y, Guo M, Ruan L, *et al*. Heterogeneous network embedding enabling accurate disease association predictions. BMC Med Genomics, 2019, **12**(Suppl 10):186

[29] Li Y, Kuwahara H, Yang P, *et al*. PGCN: disease gene prioritization by disease and gene embedding through graph convolutional neural networks. bioRxiv, 2019:532226

[30] Zhu L, Hong Z, Zheng H. Predicting gene-disease associations *via* graph embedding and graph convolutional networks. 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE, 2019:382-389

[31] Lee B, Zhang S, Poleksic A, *et al*. Heterogeneous multi-layered network model for omics data integration and analysis. Front Genet, 2019, **10**:1381

[32] Seyyedrazzagi E, Navimipour N J. Disease genes prioritizing mechanisms: a comprehensive and systematic literature review. Network Modeling Analysis in Health Informatics and Bioinformatics, 2017, **6**(1):13

[33] Prasad T S K, Goel R, Kandasamy K, *et al*. Human protein reference database--2009 update. Nucleic Acids Res, 2009, **37**(Database issue):D767-D772

[34] Chatr-Aryamontri A, Breitkreutz B J, Oughtred R, *et al*. The BioGRID interaction database: 2015 update. Nucleic Acids Res, 2015, **43**(Database issue): D470-D478

[35] Bader G D, Betel D, Hogue C W. BIND: the biomolecular interaction network database. Nucleic Acids Res, 2003, **31**(1): 248-250

[36] Szklarczyk D, Gable A L, Lyon D, *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res, 2019, **47**(D1):D 607-D613

[37] Kerrien S, Aranda B, Breuza L, *et al*. The intAct molecular interaction database in 2012. Nucleic Acids Res, 2012, **40**(Database issue):D841-D846

[38] Amaral L A. A truer measure of our ignorance. Proc Natl Acad Sci USA, 2008, **105**(19):6795-6796

[39] Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res, 2006, **34**(Database issue):D354-D357

[40] Hamosh A, Scott A F, Amberger J S, *et al*. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res, 2005, **33**(Database issue): D514-D517

[41] Köhler S, Carmody L, Vasilevsky N, *et al*. Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res, 2018, **47**(D1):D1018-D1027

[42] Pinero J, Ramirez-Anguita J M, Sauch-Pitarch J, *et al*. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res, 2020, **48**(D1):D845-D855

[43] Bello S M, Shimoyama M, Mitraka E, *et al*. Disease ontology: improving and unifying disease annotations across species. Dis Model Mech, 2018, **11**(3):dmm032839

[44] Ashburner M, Ball C A, Blake J A, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000, **25**(1):25-29

[45] The gene ontology consortium. The gene ontology resource: 20 years and still going strong. Nucleic Acids Res, 2018, **47**(D1): D330-D338

[46] Resnik P. Using information content to evaluate semantic similarity in a taxonomy // Mellish C S. Proceedings of the 14th International Joint Conference on Artificial Intelligence. Montreal: Morgan Kaufmann, 1995: 448-453

[47] Lin D. An information-theoretic definition of similarity. Proceedings of the ICML, 1998, **98**:296-304

[48] Jiang J J, Conrath D W. Semantic similarity based on corpus statistics and lexical taxonomy // Chen K J, Huang C R, Sproat R. Proceedings of the 10th Research on Computational Linguistics International Conference, Taiwan: ACLCLP, 1997: 19-33

[49] Schlicker A, Domingues F S, Rahnenfuhrer J, *et al*. A new measure for functional similarity of gene products based on gene ontology.

BMC Bioinformatics, 2006, **7**:302

[50] Li B, Wang J Z, Feltus F A, *et al*. Effectively integrating information content and structural relationship to improve the GO-based similarity measure between proteins // Hamid R, Arabnia, Quoc-Nam T. International Conference on Bioinformatics & Computational Biology (BIOCOMP), Las Vegas: CSREA Press, 2010: 166-172

[51] Yu H, Gao L, Tu K, *et al*. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene, 2005, **352**:75-81

[52] Wang J Z, Du Z, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. Bioinformatics, 2007, **23**(10): 1274-1281

[53] Lord P W, Stevens R D, Brass A, *et al*. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics, 2003, **19**(10): 1275-1283

[54] van Laarhoven T, Nabuurs S B, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics, 2011, **27**(21):3036-3043

[55] Chen X, Wang L, Qu J, *et al*. Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics, 2018, **34**(24):4256-4265

[56] Chen X, Yan C C, Zhang X, *et al*. Long non-coding RNAs and complex diseases: from experimental results to computational models. Brief Bioinform, 2016, **18**(4):558-576

[57] Westbury S K, Turro E, Greene D, *et al*. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. Genome Med, 2015, **7**(1):36

[58] Moreau Y, Tranchevent L C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet, 2012, **13**(8):523-536

[59] Salehi M, Sharma R, Marzolla M, *et al*. Spreading processes in multilayer networks. 2015, **2**(2):65-83

[60] Chen J, Bardes E E, Aronow B J, *et al*. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res, 2009, **37**(Web Server issue): W305-W311

[61] Le D H, Kwon Y K. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization. Comput Biol Chem, 2013, **44**:1-8

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

[63] Liu Z, Luo J. Genome-wide predicting disease-related protein complexes by walking on the heterogeneous network based on data integration and laplacian normalization. Comput Biol Chem, 2017, **69**:41-47

[64] Chen X, Liu M X, Yan G Y. RWRMDA: predicting novel human microRNA-disease associations. Mol Biosyst, 2012, **8**(10): 2792-2798

[65] Zhang Z, Wu F X, Wang J, *et al*. Prioritizing disease genes by using search engine algorithm. Curr Bioinform, 2016, **11**(2):195-202

[66] Zhao Z Q, Han G S, Yu Z G, *et al*. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization. Comput Biol Chem, 2015, **57**:21-28

[67] Zhu J, Qin Y, Liu T, *et al*. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. BMC Bioinformatics, 2013, **14**(Suppl 5):S5

[68] 张松瑶, 张绍武. 基于二元网络异步重启随机游走算法预测肺癌风险致病基因. 生物物理学报, 2015, **31**(1):33-44
Zhang S Y, Zhang S W. Acta Biophysica Sinica, 2015, **31**(1):33-44

[69] Luo J, Liang S. Prioritization of potential candidate disease genes by topological similarity of protein-protein interaction network and phenotype data. J Biomed Inform, 2015, **53**:229-236

[70] Battiston F, Nicosia V, Latora V. Structural measures for multiplex networks. Phys Rev E Stat Nonlin Soft Matter Phys, 2014, **89**(3): 032804

[71] Li Y, Li J. Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data. BMC Genomics, 2012, **13**(Suppl 7):S27

[72] Jiang R, Gan M, He P. Constructing a gene semantic similarity network for the inference of disease genes. BMC Syst Biol, 2011, **5**(Suppl 2):S2

[73] Lin L, Yang T, Fang L, *et al*. Gene gravity-like algorithm for disease gene prediction based on phenotype-specific network. BMC Syst Biol, 2017, **11**(1):121

Advances in Predicting The Risk Pathogenic Genes With Random Walk*

LIU Li-Li[1,2], ZHANG Shao-Wu[1]**

([1])*Key Laboratory of Information Fusion Technology of Ministry of Education, School of Automation, Northwestern Polytechnical University, Xi'an* 710072, China;

[2])*School of Physics & Information Technology, Shaanxi Normal University, Xi'an* 710119, China)

Abstract　Risk pathogenic genes prediction is important for uncovering the occurrence and development mechanism of complex diseases (*i.e.*, cancer), improving the disease detection, prevention and treatment, and providing the targets for drug design. Traditional gene-mapping approaches, such as linkage analysis and genome-wide association studies (GWAS), often predict hundreds of candidate genes. But it is costly, time-consuming and laborious to further validate these candidate genes with biological experiments. However, the number of candidate pathogenic genes can be effectively reduced by computational and prioritization methods. Considering the excellent performance of random walk with restart (RWR) in predicting the risk pathogenic genes, in this work, we comprehensively discuss the recent progresses of predicting the risk pathogenic genes with RWR from databases related with genes and diseases, the metrics of measuring the similarity between genes/diseases, the strategies of choosing the seed genes of specific disease, and the different genes/diseases network structures. We also point out the computational problems and challenges faced in the process of pathogenic genes prediction.

Key words　risk pathogenic genes, random walk, monoplex network, multiplex network, heterogeneous network

DOI：10.16476/j.pibb.2020.0376

---

* This work was supported by a grant from The National Natural Science Foundation of China（61873202）.

** Corresponding author.

Tel：86-29-88431308，E-mail：zhangsw@nwpu.edu.cn

Received：October 22，2020　Accepted：December 14，2020