计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100129

基于信息传播的致病基因识别研究

李家文　郭炳晖　杨小博　郑志明
北京航空航天大学大数据与脑机智能高精尖中心　北京 100191
鹏程实验室　广东 深圳 518055
北京航空航天大学数学科学学院教育部数学信息与行为重点实验室　北京 100191
(jiawenli@buaa.edu.cn)

摘　要　基因在生命科学领域的研究中占据着重要地位,而致病基因则是关键重心之一。对致病基因的精准识别可以揭示疾病在分子层面的发病机制,为疾病的预防、诊断及治疗等多个阶段提供强力支撑。准确识别致病基因的关键在于给出基因之间的相似性度量。文中利用复杂网络对生物系统进行建模,并提出了一种带有耗散机制的多源头重启随机游走模型 DRWMR 来度量基因之间的功能相似程度。首先基于 NCBI 等生物数据库构建人类基因相互作用网络,并在 KEGG 的疾病-基因关联数据集上开展实验对已知致病基因进行识别。与 SP,RWR 和 PRINCE 3 种现有模型进行对比,DRWMR 准确预测了 581 种疾病中的 156 种,而其余模型平均正确预测了 121.3 种,DRWMR 的平均预测分数相比其余模型的预测分数均值高出 9.46%。最后使用所提模型预测哮喘、血友病和 PEHO 综合征的潜在致病基因,预测结果均在文献或数据库中找到了理论或实验支持。

关键词:生物信息学;复杂网络;信息传播;基因功能预测
中图法分类号　R319;O29

Disease Genes Recognition Based on Information Propagation

LI Jia-wen,GUO Bing-hui,YANG Xiao-bo and ZHENG Zhi-ming
Beijing Advanced Innovation Center for Big Data and Brain Computing,Beihang University,Beijing 100191,China
Peng Cheng Laboratory,Shenzhen,Guangdong 518055,China
Key Laboratory of Mathematics,Informatics and Behavioral Semantics,School of Mathematical Sciences,Beihang University,Beijing 100191,China

Abstract　Genetic research in the field of life science and medicine occupies an important position,while disease genes are one of its key focuses. Accurate identification of disease-causing genes can reveal the pathogenesis of diseases at the molecular level,and provide strong support for the prevention,diagnosis,treatment and other medical stages of diseases. The key to accurately identifying disease-causing genes is to give a measure of similarity between genes. This paper uses complex networks to model biological systems and proposes a dissipative random walk model with multiple restarts to measure the degree of functional similarity between genes. Firstly,a human gene-gene interaction network is constructed based on the human gene interaction datasets on NCBI. Experiments are then carried out on KEGG's disease-gene association dataset to identify known disease-causing genes. Compared with the three existing models of SP,RWR and PRINCE,DRWMR accurately predicts 156 of 581 diseases while the remaining models predict 121.3 correctly on average. The average prediction score of DRWMR is 9.46% higher. Finally,the potential disease genes of asthma,hemophilia and PEHO syndrome are predicted and the candidate genes are found guilty for the pathologies in the literature or biological database.

Keywords　Bioinformatics,Complex networks,Information propagation,Gene function prediction

到稿日期:2020-11-17　返修日期:2021-04-18　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:科技创新 2030-"新一代人工智能"重大项目(2018AAA0102301);国家自然科学基金项目(11671025);民机项目(MJ-F-2012-04)
This work was supported by the Artificial Intelligence Project(2018AAA0102301),National Natural Science Foundation of China(11671025) and Fundamental Research of Civil Aircraft(MJ-F-2012-04).
通信作者:郭炳晖(guobinghui@buaa.edu.cn)

1 研究背景及意义

在生命科学领域，针对基因开展全方位研究的重要性是不容置疑的，基因是决定生命健康的关键内在因素，其生物意义极其重要和显著。致病基因则是一类特殊的基因，它的突变或缺失会导致生物患上某种疾病。对致病基因的精准识别能在疾病早期阶段提供预警，达到预防的目的，或者是在治疗阶段提供分子视角，为高效的治疗方案提供切入点。

传统基于生物实验或病例统计学的致病基因识别方法普遍具有周期长、耗资不菲等缺点，使研究效率与精度都受到了严重的制约。生物信息学与网络生物学等交叉学科的出现改善了这一局面[1-2]。复杂生命系统的各项功能类似于系统中的元素以及元素之间相互作用展示出的宏观表现，这与复杂网络的思想十分类似，由此诞生了网络生物学这门学科，同时也出现了许多识别致病基因的新方法。最基本的思路是通过已知功能的基因来推断新的基因，其理论基础是“Guilt by association”假设，即功能相近或者具有相似表型的基因通常会存在相互作用关系。因此，如何准确地度量基因之间的相似程度成为了基因功能推断问题的核心。网络生物学的基本思路主要分为两派，一部分学者在生物网络中通过结构模型提取局部结构特征，而另一部分学者则开展传播动力学过程，利用信息传播的方法抽象出全局特征。研究者对结构特征的探究集中在节点的统计特征和网络拓扑结构这两个方面。重要的节点特征通常包括桥节点、Hub 节点等，而拓扑结构通常考虑度分布、聚类系数等指标。网络中的传播动力学过程的思想在于将网络节点处产生的调控或扰动信息抽象为信息流(flow)，通过信息传播模型给定传播动力学规则，流就会沿着节点之间的连边扩散到整个网络，从而以流的分布状态来表示网络中的隐含信息和规律。例如，以突变基因作为源节点在基因或蛋白质的相互作用网络中应用该信息传播模型，就可以模拟该基因的突变对网络中其他基因的影响；或者发现与已知功能基因有较大关联的其他基因，从而预测基因功能；或者是观察信息流在网络中的聚集效应，发掘可能存在的疾病或功能模块。常见的传播模型包括随机游走(Random Walk，RW)、PageRank 算法、扩散模型、传染病“Susceptible-Infectious”模型等[3]。已经有许多学者利用信息传播等网络动力学特征在基因功能识别、药物靶点预测、网络模块发现、疾病亚型分类等方向取得了长足的进展[4]，不论是在效率还是精度上都已经比肩其他方法，为探寻疾病的分子机制提供了新的网络视角。

国内外对该领域的研究，通常以功能已知的致病基因作为信息传播的源节点，通过观察稳定状态时信息流的分布来给出预测的致病基因。Weston 等作为先行者在 2004 年首次将扩散模型的思想加入到蛋白质相似性网络中，提出了 RankProp 算法[5-6]。Qi 等在酵母菌的合成致死相互作用网络上开展了热扩散模型(Diffusion Kernel)对于寻找特定相互作用关系的研究[7]。Vandin 等在其基础上，将方法扩展到一个全基因组规模的基因相互作用网络上，并利用多重假设检验的方法来降低错误发现率[8]。Vanunu 等提出了用于预测致病基因的 PRINCE(PRIoritizatioN and Complex Elucidation)算法，该算法在很长一段时间内成为了致病基因识别的常用模型[9]。Qian 等则在 PRINCE 算法的基础上，加入了对 α 和 T 两个参数选取方式的讨论，预测了克罗恩氏病(Crohn's Disease)的潜在致病基因[10]。Macropol 等在酵母菌的功能网络上开展了自重复随机游走模型(Repeated Random Walk，RRW)，用于不断发现功能相似的基因，并以此来构建功能模块[11]。Wang 在 2019 年提出了一种 PU 归纳矩阵补全方法 PUIMCHIF，通过融合异构信息、RWR 模型和扩散分量分析方法来预测致病基因[12]。Zhao 等提出了一种基于邻层传播的重要节点识别方法 NLD，若一个节点与其他已知重要节点具有大量共同特征，那么它是重要节点的概率也就越大[13]。但通常来说，在复杂生物网络上开展传统的信息传播模型依旧存在着预测精度不高、预测结果难以验证等缺陷，而生物组学大数据也通常具有噪声大、数据缺失等问题，会进一步给模型带来干扰。为解决上述问题，本文提出了一种带有耗散机制的多源头重启随机游走算法，并在真实的数据集上取得了较好的实验效果。

2 DRWMR 算法描述

本文提出的带有耗散机制的多源头重启随机游走算法(Dissipative Random Walk with Multiple Restarts，DRWMR)通过引入重启机制避免了标准随机游走在稳定状态时陷入同一分布的缺陷，建立了稳态分布向量与初始向量的相似程度描述。增加多源头的游走初始节点，用来模拟多基因遗传病中多个基因耦合关联在疾病的发生过程中所发挥的作用。而耗散机制的存在会导致网络更加倾向于留下那些置信程度较高的连边所代表的信息，减少脏数据对模型的影响。本文方法的具体流程如图 1 所示。

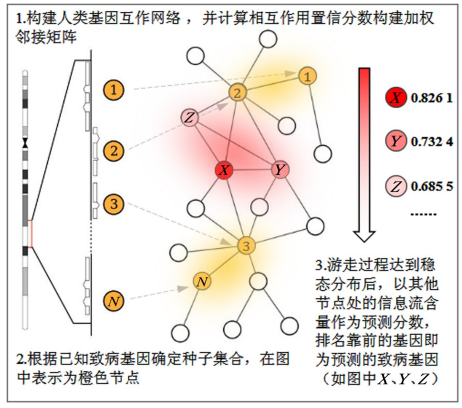


图 1 DRWMR 模型的流程(电子版为彩色)

Fig. 1 Process of DRWMR

2.1 随机游走与多源头重启过程

随机游走是一种非常流行的随机过程模型，具有非常丰富的历史。随机游走的概念最早由 Karl Pearson 提出[14]，并在之后扩展到了数学、物理学、计算机科学等多个学科领域。设 $p_i(t)$ 为在 t 时刻访问到节点 i 的概率，即 $p_i(t)=P(X_t=i)$，那么在 $t+1$ 时刻，粒子应该随机访问到节点 i 的某一个邻居节点上，即：

$$P(X_{t+1}=j \mid X_t=i)=\begin{cases}\dfrac{1}{d(i)}, & \text{if } (i,j)\in E\\ 0, & \text{otherwise}\end{cases}$$

其中，$d(i)$为节点 i 的度。那么，易得：

$$p_j(t+1)=\sum_i p_i(t)\cdot\frac{A_{ij}}{d(i)}$$

其中，A_{ij}表示节点 i 和 j 之间是否有连边，如果有，则$A_{ij}=1$，否则$A_{ij}=0$。将上式写成矩阵形式则有：

$$\boldsymbol{p}(t+1)=\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{p}(t)$$

其中，$\boldsymbol{p}$ 为第 i 个位置p_j的向量，$\boldsymbol{p}(t)$称为 t 时刻游走至各顶点的概率分布向量，$\boldsymbol{A}$ 为邻接矩阵，$\boldsymbol{D}$ 是节点的度组成的对角矩阵。随机游走模型的一个特点在于，从任意一个源节点出发，在稳定状态时都会得到同样的全局概率分布。如果用 $\boldsymbol{\pi}$ 来表示 $t\to\infty$时的概率分布，那么从节点 i 转移到节点 j 的概率应该等同于从节点 j 转移到 i 的概率，即：

$$\boldsymbol{\pi}_i\cdot P(X_{t+1}=j\mid X_t=i)=\boldsymbol{\pi}_j\cdot P(X_{t+1}=i\mid X_t=j)$$

即：

$$\frac{\boldsymbol{\pi}_i}{d(i)}=\frac{\boldsymbol{\pi}_j}{d(j)}$$

因此在稳定状态时，随机游走至顶点 i 的概率与顶点的度成正比，而这与源节点无关。

这个特性使得在网络上开展单纯的随机游走算法在一定程度上失去了意义。换句话说，它丢失了与源节点相关的某些网络的全局结构信息。而重启机制的引入可以解决这一问题。重启随机游走的过程可表示为：

$$\boldsymbol{p}(t+1)=\alpha\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{p}(t)+(1-\alpha)\boldsymbol{e}_k$$

其中，$\alpha\in[0,1]$称为重启概率。$\boldsymbol{e}_k\in\mathbb{R}^{n\times1}$是 n 维列向量空间的第 k 个标准正交基，即第 k 个位置的元素为 1，其他均为 0 的列向量，其同时也是随机游走过程在 $t=0$ 时刻的概率分布向量。假设 $t\to\infty$时的概率分布向量为 $\boldsymbol{\tau}$，则有：

$$\boldsymbol{\tau}=\alpha\boldsymbol{A}\boldsymbol{D}^{-1}\boldsymbol{\tau}+(1-\alpha)\boldsymbol{e}_k$$

通过化简可得：

$$(\boldsymbol{I}-\alpha\boldsymbol{A}\boldsymbol{D}^{-1})\boldsymbol{\tau}=(1-\alpha)\boldsymbol{e}_k$$

$$\boldsymbol{\tau}=(1-\alpha)(\boldsymbol{I}-\alpha\boldsymbol{A}\boldsymbol{D}^{-1})^{-1}\boldsymbol{e}_k$$

其中，$\boldsymbol{I}-\alpha\boldsymbol{A}\boldsymbol{D}^{-1}$可以保证其可逆性。$\boldsymbol{A}\boldsymbol{D}^{-1}$的行和至多是 1，从而该非负矩阵的谱半径小于或等于 1，最大特征值也小于或等于 1，而 α 是个小于 1 的正常数，因此 $\boldsymbol{I}-\alpha\boldsymbol{A}\boldsymbol{D}^{-1}$是可逆的。

可以看到重启随机游走的稳态概率分布 $\boldsymbol{\tau}$ 不再仅仅与度有关，从不同的节点出发的游走过程，最终产生的概率分布也是不同的。可以发现，$\boldsymbol{\tau}$ 由 $\boldsymbol{e}_k$ 左乘一个方阵得来，它给出的实际上是 $\boldsymbol{e}_k$ 的某种相似性度量。由此可以看出，相比普通的随机游走，重启随机游走更能捕捉网络的全局特性，并且在很大程度上保留了源节点的有关信息，在生物网络及许多其他学科的复杂网络中都有很大的应用价值。

另一方面，现有的信息传播算法通常从单个源节点出发，经过信息传播过程达到稳定状态，通过稳态时信息流的全局分布来给出各个节点与源节点的相似程度。但事实上，只有少部分疾病是单基因遗传病，大部分疾病都由多个基因协同作用而致使发病，因而被称为多基因遗传病。如果依旧从单个源节点开展信息传播过程，将对多基因遗传病的模拟效果大打折扣。因此，本文考虑一种有多个初始源头的信息传播过程。在开始时，随机游走将从多个源节点同时开始，在重启的步骤中，也以等概率回到任意一个源节点中。则重启随机游走基本公式中的 $\boldsymbol{e}_k$将变为 $\boldsymbol{e}_k'$，表示初始时刻多个源头节点上都有一定量信息流的分布向量。根据之前的讨论，多个源头的引入不影响重启随机游走过程中方阵 $\boldsymbol{I}-\alpha\boldsymbol{A}\boldsymbol{D}^{-1}$的可逆性，模型将依旧保证收敛。

2.2 耗散机制

为了减少生物组学大数据的噪声和脏数据干扰，考虑一种基于连边权重的信息耗散机制。如果对网络中的每一条边计算其相互作用的置信程度，将置信分数作为连边的权重，则可得到网络的加权邻接矩阵，记为 $\boldsymbol{W}$，用以替换原模型中的 $\boldsymbol{A}$。即最终 DRWMR 的公式变为：

$$\boldsymbol{p}(t+1)=\alpha\boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{p}(t)+(1-\alpha)\boldsymbol{e}_k'$$

在原来的定义下，根据公式

$$p_j(t+1)=\alpha\left(\sum_i p_i(t)\cdot\frac{A_{ij}}{d(i)}\right)+(1-\alpha)\delta_{jk}$$

和

$$\sum_j\frac{A_{ij}}{d(i)}=1$$

可知，每个节点在 t 时刻的信息流为$p_i(t)$，除去回到源节点的$(1-\alpha)p_i(t)$，剩余部分 $\alpha\,p_i(t)$将以 100% 的比例全部提供给邻居节点，且网络中流的总量保持不变。而在新的定义下，通常有：

$$\sum_j\frac{W_{ij}}{d(i)}<1$$

即信息流沿连边传播到邻居节点时，将强制以一定比例耗散，或者说被遗忘掉。边的权重越大，信息的耗散损失就越少，被记忆的内容就越完整。这也正是我们希望的，即对于置信程度高的相互作用关系，我们希望尽可能地保留其中隐含的特征，而置信程度低的连边则尽可能多地将其遗忘。

综合上述提出的多源头重启和耗散遗忘机制，本文提出了带有耗散机制的多源头重启随机游走模型 DRWMR，算法如下：

(1)构建基因相互作用网络 $G=(V,E)$，获取网络的邻接矩阵 $\boldsymbol{A}$ 及表示各节点的度的对角矩阵 $\boldsymbol{D}$。

(2)从数据库中获取基因相互作用数据，在网络 G 中给连边赋予权重，并构造权重邻接矩阵 W。

(3)对于指定疾病 $\mathcal{D}$，从数据库中获取 $\mathcal{D}$ 的已知致病基因，称为“种子基因集合”，记为 S。根据 S 计算原始状态分布向量 $\boldsymbol{F}_0$。

(4)利用 $\boldsymbol{F}_t$ 表示 t 时刻的状态分布向量，计算 $\boldsymbol{F}_{t+1}=\alpha\boldsymbol{W}\boldsymbol{D}^{-1}\boldsymbol{F}_t+(1-\alpha)\boldsymbol{F}_0$。

(5)如果 $\boldsymbol{F}_t$和 $\boldsymbol{F}_{t+1}$的差距小于给定阈值或者迭代到达最大次数，则终止流程；否则令$\boldsymbol{F}_t=\boldsymbol{F}_{t+1}$，并回到步骤(4)。

(6)对稳态分布向量$\boldsymbol{F}_\infty$的各分量从大到小进行排序，除去 S 中的基因之外，排名靠前的其他基因就是预测出的潜在致病基因。

3 人类基因互作网络的致病基因识别

本节将通过实验验证本文提出方法的性能。首先获取真实的基因相互作用实验数据集，建立人类基因相互作用网络。然后整合多源数据分析相互作用置信程度，计算置信分数并

构建加权网络。在加权网络上开展两部分实验。在第一部分实验中将对每一个疾病都隐藏其中一个致病基因，开展对照实验观察各模型将该基因识别为致病基因的预测准确率，并通过不同准则（如疾病的病理学分类等）对实验数据集进行划分，讨论并分析几种模型在各个疾病子类下的预测效果。而在第二部分实验中将直接利用疾病所有的已知致病基因作为数据集开展实验，预测潜在的致病基因，最终在生物数据库和相关文献中寻找预测结果的理论支持或实验佐证，以验证信息传播模型的有效性，从而为致病基因识别问题提供一种全新思路和高效手段。

3.1 人类基因互作网络加权网络的构建

本文从美国国家生物信息中心（NCBI）中获取了人类基因相互作用数据集。筛选去重之后共有 441 218 个与人类基因相关的基因相互作用二元组，涉及到独立不重复的人类基因共有 18 717 个。这就构成了人类基因相互作用网络的雏形。

由于网络的分支情况对于信息传播过程具有十分重要的作用，因此对上述网络的连通情况进行分析可以发现，网络中有一个巨型分支和 13 个小分支，其余部分则是一个网络的最大连通分支（Largest Connected Component，LCC）。我们删去小分支，得到了最终的人类基因互作网络。该网络共 18 703 个节点，441 205 条边。

网络科学的重要学者 Albert-László Barabási 曾经指出，复杂生物网络中同样会涌现出无标度特性[2]。为了探究人类基因互作网络是否满足幂律分布，画出网络的度分布图，并在双对数坐标系下进行观察。网络的度分布如图 2 所示。

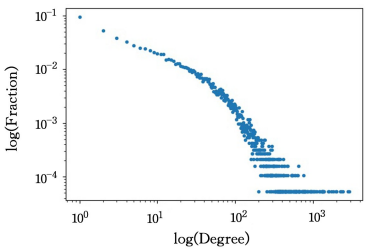


图 2 人类基因互作网络的度分布

Fig. 2 Degree distribution of humangene-gene interaction network

由图 2 可以看到，双对数坐标系下人类基因网络节点的度分布大致处于一条直线上，大致呈幂律分布。

在此无权网络的基础上，我们整合了 BioGRID，MIPS 和 HIPPIE 3 个数据集的基因相互作用置信程度数据，用于给连边赋权值[15]。融合上述 3 个多源异构数据，最终的置信分数由下式可知：

$$C_{ij}=1-\prod_{r}(1-C_{ij}^{r})$$

其中，C_{ij}^{r} 是基因 i-基因 j 相互作用关系在第 r 个数据集中的置信分数。

3.2 现有致病基因识别

基于信息传播方法的致病基因识别，其理论基础在于通过已知致病基因推断未知基因，因此需要获取疾病及其相关基因的关联数据。从京都基因与基因组百科全书（KEGG）数据库中获取了疾病-基因相关关系数据共 1 676 个，涉及到 4 635 个疾病-基因关联。每个疾病至少有一个已知的致病基因，最多可有 59 个致病基因，具体如表 1 所列。在该数据集的基础上，针对每一种感兴趣的疾病可以构建相应的种子基因集合（seed genes），而剩余的基因则成为候选基因（candidate genes）。通过描述候选基因到种子基因集合的距离来对候选基因进行排序，排序靠前的基因，即与现有种子基因相似度高的基因，则被认为是潜在的未知致病基因。

表 1 1 676 种疾病中种子基因数量的分布情况

Table 1 Distribution of seed gene numbers among 1 676 diseases

疾病的种子基因数量	对应的疾病数量
1	1 095
2	169
3	94
4	77
5	32
6	42
7	25
8	19
9～16	90
17～32	27
33～64	6

假设疾病 $\mathcal{D}$ 已经存在有 L 个致病基因，那么从中随机剔除任一基因，一个好的致病基因识别算法应该能够重现这个疾病-基因关联关系。如果以剩下的 $L-1$ 个致病基因作为信息传播过程的源头，那么在稳定状态时，缺失的第 L 个致病基因应该留存有非常高的流量，在排序时应该排在靠前的位置。考虑到在网络生物学中常用的 k-fold 富集分数（k-fold enrichment）定义[16]，可以类似地定义一种度量手段，来衡量未知基因预测结果的准确性。其公式为：

$$f(k)=\begin{cases}1, & \text{if } k\leqslant L\\ \dfrac{L}{k}, & \text{if } k>L\end{cases}$$

可以看到，假如将缺失的基因成功排到了前 L 个位置，说明算法将缺失的疾病-基因关联完全复现了出来，这是最理想的情况，此时称模型对疾病 $\mathcal{D}$ 预测正确；否则称预测不正确，预测分数将有一个严格小于 1 的值。

对于一个有 L 个已知致病基因的疾病 $\mathcal{D}$ 来说，每次随机隐藏一个基因将有 L 种选择，因此实验考虑对这些种子基因进行遍历，一共得到 L 个预测分数，再从其中选取最大的预测分数作为该疾病预测效果的最终评估分数。遍历 KEGG 数据集的每一个疾病，得到模型的总体评判。最短路径方法（Shortest Path，SP）是一种基于网络静态结构特征来度量节点接近程度的方法，而重启随机游走的基本模型和 Vanunu 等提出的致病基因识别方法 PRINCE 则是基于网络动态传播特征的节点相似性度量模型[9]。因此，我们在同样的数据集上应用 SP，RWR 和 PRINCE 这 3 种模型，对已缺失致病基因的还原准确度进行对比，从而验证本文模型的优越性。

致病基因数量大于 2 的疾病共 581 种，4 种模型在该数据集上预测正确的疾病数量和平均预测分数如表 2 所列。同样给出 4 种模型预测分数在[0，1]之间的分布情况，如图 3 和图 4 所示，图中的灰色虚线表示随机选择的实验效果。一个好的模型，其评估分数应该倾向于集中在图像的右侧，即更靠近 $x=1$ 的位置。

表 2 4 种模型在 581 种疾病上的表现

Table 2 Performance of 4 models on 581 diseases

模型	预测成功的疾病数量	平均预测效果
PRINCE	113	0.3597
SP	112	0.4189
RWR	139	0.4194
DRWMR	156	0.4371

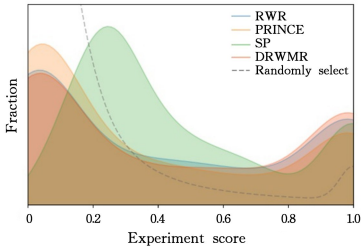


图 3 平均预测分数在[0,1]之间的分布情况(电子版为彩色)

Fig. 3 Average prediction score distribution in [0,1]

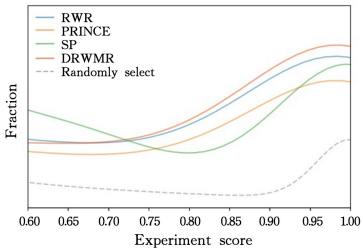


图 4 平均预测分数在[0.6,1]之间的局部情况(电子版为彩色)

Fig. 4 Average prediction score distribution in [0.6,1]

从图 3、图 4 可以看到,DRWMR 算法在重现疾病-基因关联时能够取得最佳效果,整体的分布更加偏向 $x=1$,在[0.6,1]的局部区域可以更清楚地观察到这一现象。

KEGG 数据集对疾病采取“三级分类”的方案,在一级分类下共有 15 种类型,上述疾病均可以划分在该 15 种疾病大类下。此外,表 1 实际上也给出了一种划分方式。因此,可以检验 4 种模型在这两种分类思路下对 581 种疾病的预测效果。

在图 5 可以很明显地看到,SP 算法和其他 3 种基于传播动力学特征的方法在预测分数上有着明显不同的趋势,随着种子基因数量的增加,预测的效果也在持续降低。这是因为 SP 算法通过两个节点之间的测地距离来度量接近程度,当且仅当基因与已知的 $L-1$ 个致病基因均为邻居时,才能对该疾病成功的进行预测。这样的条件在已知致病基因越少时越容易达成。而在其余 3 种方法中,随着已知致病基因数量的增加,预测的效果都在不断提高。这是由于这 3 种方法的普遍思路都是通过已知致病基因来推断未知基因,种子集合实际上充当了先验知识的角色,更大的种子基因集合意味着更充分、更完备的先验知识,因此推断出的结果会更加准确。对比这 3 种模型的预测效果可以看到,DRWMR 准确预测的疾病数量是最多的,即使在某些分组(如 $L=6$)上正确预测的疾病数量比其他模型少,但是从图 5 可以很清晰地看出,在种子基因数量 $L>3$ 的几乎每一个分段,DRWMR 的平均预测分数都是最高的。

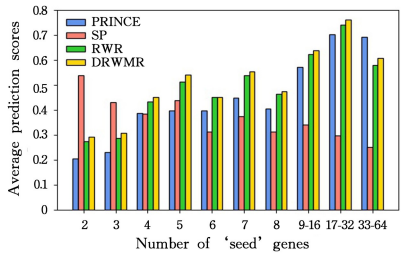
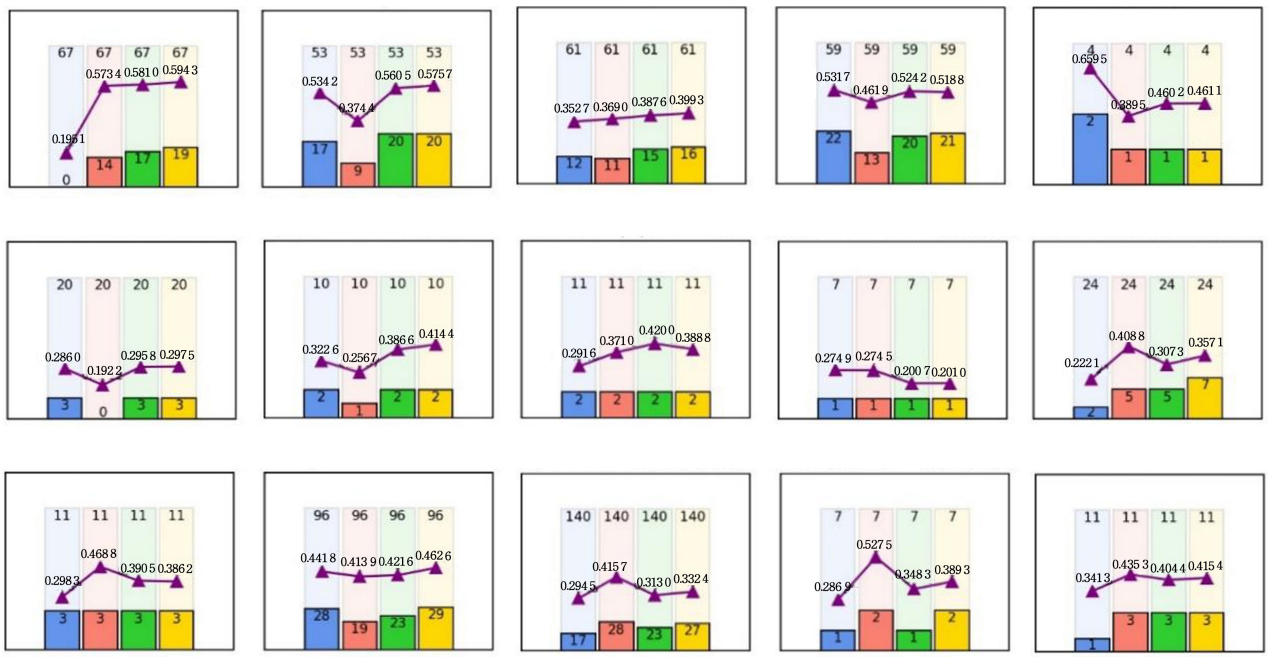


图 5 4 种模型在种子基因数量 L 不同时的平均预测分数

Fig. 5 Average prediction scores of 4 models in circumstances of different number of seed genes



注:从左至右分别是 PRINCE,SP,RWR 和 DRWMR

图 6 4 种模型在 15 类疾病上的预测表现

Fig. 6 Performance of 4 models on 15 diseases families

在 KEGG 一级分类下的 15 种疾病大类中也有类似的现象。从图 6 可以看到,PRINCE 算法在心血管疾病、呼吸道疾病和生殖系统疾病这 3 类疾病上表现最好,SP 算法在皮肤病等 4 类疾病上有较好的预测效果,标准的 RWR 算法在泌尿系统疾病上的预测分数最高,而对于剩余的癌症、免疫系统疾病等 7 类疾病,DRWMR 都给出了最优解。在所有的疾病中,DRWMR 均取得了最好的效果,准确预测出了 156 种疾病的缺失致病基因,而其余 3 种模型(PRINCE,SP,RWR)分别只预测出了 113 种、112 种和 139 种疾病。由前文的讨论可知,SP 算法在已知致病基因数较少的疾病上预测缺失致病基因的效果较好,那么通过观察皮肤病、先天畸形、其他先天性疾病和其他疾病这 4 类疾病的致病基因数量可以发现,$L\leqslant 3$ 的疾病在这些疾病中的数量都超过了 50%,皮肤病甚至有三分之二的疾病都仅有少于 3 个已知致病基因。因此 SP 算法在这 4 类疾病中有较好的预测效果。而对疾病研究的深入程度也同样会影响到模型的预测效果。例如,癌症是投入研究力度最大的疾病类型,关于癌症的文献与实验数据相对来说最丰富,这意味着其拥有最全面的先验知识。而 DRWMR 的多源头机制能够使模型更充分地利用这些内容,因此能够获得最好的效果。

3.3 潜在致病基因的预测

本节利用带耗散机制的多源头重启随机游走算法来对潜在的致病基因进行预测。区别于上述过程随机隐藏一个致病基因,此时对于每一种疾病应取其已知的全部致病基因作为种子集合,达到"通过已知来推断未知"的目的。在哮喘、血友病和 PEHO 综合征 3 种疾病上展开了实验。3 种疾病的已知致病基因数量分别为 10,4 和 1,代表了现阶段研究深入程度各不相同的多种疾病。相应的预测结果都分别在文献或数据库中找到了理论或实验的佐证。

哮喘是一种常见的以气道慢性炎症为特征的疾病,在 KEGG 数据集中哮喘被归类于免疫系统疾病。在预测结果的 5 个基因中,HLA-DQA1,IL13RA2 和 IL2RG 均是与已知致病基因的同族基因,PIK3R1 的常染色体显性突变会导致活化磷酸肌醇 3-激酶综合征,与哮喘有直接关联,最后一个 A2M 被发现该基因的编码蛋白质可以抑制炎症细胞因子,从而阻断炎症级联反应[17-18]。血友病是由于体内的 F8 或 F9 基因缺陷或者异常而导致的,预测出的致病基因中除了同样为凝血因子家族成员的 F2 和 F10 基因,还有 HSPA5,因表达产物 70-kDa 蛋白 5,其也被验证与该疾病有相关关系[19]。PEHO 综合征指的是一种伴有水肿、心律失常和视神经萎缩的进行性脑病,目前仅已知其有一个致病基因 ZNHIT3。预测结果中的基因 NUFIP1 编码一种核 FMR1 相互作用蛋白 1,可能与 ZNHIT3 的表达产物共同参与 pre-rRNA 的加工,因此参与 PEHO 综合征的发病过程[20]。

结束语 基因在生命科学与医学领域的研究中占据着重要地位,而致病基因则是基因研究的关键重心之一。对致病基因进行准确快速的识别可以揭示疾病-基因关联,从而探寻疾病在分子层面的发病机理,为疾病的早期预防、精准智能诊断以及基因靶向治疗提供突破口。而度量两个基因之间的相似性程度,成为了通过已知基因来预测潜在致病基因的关键。

本文基于信息传播模型提出了一种基因相似性度量的新型网络方法,并在大规模真实生物网络上进行了实验验证,并对 3 种实际疾病进行了致病基因的预测,最后在文献与数据库中找到了这 3 种疾病与基因关联关系的理论支持。DRWMR 为探寻疾病的分子机制提供了新的网络视角,有望在未来持续推进生命科学与医学领域的长期探索,并最终助力建设智慧化、全面化的医疗体系。

参 考 文 献

[1] MASYS D R. New directions in bioinformatics[J]. Journal of research of the National Institute of Standards and Technology, 1989,94(1):59.

[2] BARABASI A L,OLTVAI Z N. Network biology:understanding the cell's functional organization[J]. Nature Reviews Genetics,2004,5(2):101.

[3] KERMACK W O,MCKENDRICK A G. A contribution to the mathematical theory of epidemics[J]. Proceedings of the Royal Society of London. Series A,Containing Papers of a Mathematical and Physical Character,1927,115(772):700-721.

[4] COWEN L,IDEKER T,RAPHAEL B J,et al. Network propagation:a universal amplifier of genetic associations[J]. Nature Reviews Genetics,2017,18(9):551.

[5] WESTON J,KUANG R,LESLIE C,et al. Protein ranking by semi-supervised network propagation[J]. BMC Bioinformatics, 2006,7(1):S10.

[6] WESTON J,ELISSEEFF A,ZHOU D,et al. Protein ranking: from local to global structure in the protein similarity network [J]. Proceedings of the National Academy of Sciences,2004, 101(17):6559-6563.

[7] QI Y,SUHAIL Y,LIN Y,et al. Finding friends and enemies in an enemies-only network:a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions[J]. Genome Research,2008,18(12): 1991-2004.

[8] VANDIN F,UPFAL E,RAPHAEL B J. Algorithms for detecting significantly mutated pathways in cancer[J]. Journal of Computational Biology,2011,18(3):507-522.

[9] 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):e1000641.

[10] QIAN Y,BESENBACHER S,MAILUND T,et al. Identifying disease associated genes by network propagation[C]// BMC Systems Biology. BioMed Central,2014,8(1):S6.

[11] MACROPOL K,CAN T,SINGH A K. RRW:repeated random walks on genome-scale protein networks for local cluster discovery[J]. BMC Bioinformatics,2009,10(1):283.

[12] WANG X P. Research on disease-causing gene prediction algorithm based on heterogeneous information fusion[D]. Harbin Institute of Technology, 2019.

[13] ZHAO N, LI J, WANG J, et al. Relatively important node mining method based on adjacent layer propagation[J]. Journal of University of Electronic Science and Technology of China, 2021, 50(1): 121-126.

[14] PEARSON K. The problem of the random walk [J]. Nature, 1904, 72(1867): 342.

[15] ALANIS-LOBATO G, ANDRADE-NAVARRO M A, SCHAEFER M H. HIPPIE v2.0: enhancing meaningfulness and reliability of protein-protein interaction networks[J]. Nucleic Acids Research, 2016, 45(1): D408-D414.

[16] KÖHLER 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.

[17] ELKAIM E, NEVEN B, BRUNEAU J, et al. Clinical and immunologic phenotype associated with activated phosphoinositide 3-kinase δ syndrome 2: a cohort study[J]. Journal of Allergy and Clinical Immunology, 2016, 138(1): 210-218. e9.

[18] A2M alpha-2-macroglobulin "Summary" [EB/OL]. https://www.ncbi.nlm.nih.gov/gene/2.

[19] SAMELSON-JONES B J, ARRUDA V R. Protein-engineered coagulation factors for hemophilia gene therapy[J]. Molecular Therapy-Methods & Clinical Development, 2019, 12: 184-201.

[20] SABAIE H, AHANGAR N K, GHAFOURI-FARD S, et al. Clinical and genetic features of PEHO and PEHO-Like syndromes: A scoping review[J]. Biomedicine & Pharmacotherapy, 2020, 131: 110793.

LI Jia-wen, born in 1996, postgraduate, is a member of China Computer Federation. His main research interests include complex networks and bioinformatics.

GUO Bing-hui, born in 1982, associate professor, is a professional member of China Computer Federation. His main research interests include data science and complex intelligent system.

(责任编辑:喻藜)