随机游走技术在网络生物学中的研究进展

李 敏,王晓桐,罗慧敏,孟祥茂,王建新
(中南大学信息科学与工程学院,湖南长沙,410083)

摘 要: 网络生物学是近年来受到国际学术界广泛关注的学术前沿领域,在疾病研究和药物预测等领域有重要应用. 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、物理和社会网络分析中都有广泛应用. 近年来逐渐被应用到网络生物学,并在技术上得到了新的发展. 本文以生物网络为基础,介绍了随机游走技术及其基本理论,并详细阐述了随机游走技术在网络生物学中的应用,具体包括蛋白质功能预测、关键蛋白质识别、疾病基因预测、疾病相关非编码 RNA 预测、药物相关预测等. 最后讨论了随机游走技术在网络生物学研究中存在的问题以及未来的研究方向.

关键词: 随机游走; 生物网络; 网络生物学; 生物信息学; 系统生物学
中图分类号: TP301.6 **文献标识码:** A **文章编号:** 0372-2112 (2018)08-2035-14
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.08.033

Progress on Random Walk and Its Application in Network Biology

LI Min, WANG Xiao-tong, LUO Hui-min, MENG Xiang-mao, WANG Jian-xin
(*School of Information Science and Engineering, Central South University, Changsha, Hunan* 410083, *China*)

Abstract: Network biology, as a hot academic frontier field, has gained increasingly wide attention in international academic circles in recent years, which plays an important role in disease research and drug discovery. Random walk is a mathematical model, which is widely used in financial, physical and social network analysis. Recently, it has gradually been applied in network biology, and the model has been improved constantly. Based on the biological network, this study introduces the technology and basic theory of random walk model firstly. Then, the applications of random walk in network biology are presented in detail, which include predicting protein functions, identifying essential proteins, predicting disease gene, discovering disease related non-coding RNAs, discovering disease related things and so on. Finally, some existing problems and future research directions of random walk in network biology research are discussed in this study.

Key words: random walk; biological network; network biology; bioinformatics; system biology

1 引言

2004 年,Barabasi 和 Oltvai[1] 在其发表在 Nature 上的文章中提出网络生物学(network biology)的概念,给出了对于各种生物系统基于网络的定量描述. 这使得很多生物计算问题,包括蛋白质功能预测、关键蛋白质识别、癌症的驱动基因识别、疾病相关的基因/microRNA/lncRNA/pathway 等预测、药物靶标预测、药物重定位、蛋白质复合物挖掘都可以使用基于网络的方法来展开研究.

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. 随机游走模型在图上应用的基本思想是从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. 每一次的游走过程都是在一定的时间内进行的. 最后,这个概率将会趋于稳定不再改变,得到起始节点到每个节点的概率. 文献[2,3] 深入讨论了随机游走模型,包括它的定义和数学公式.

由于随机游走模型可以迅速在网络中扩散的特

收稿日期:2017-03-23;修回日期:2017-08-10;责任编辑:梅志强
基金项目:国家自然科学基金优秀青年项目(No. 61622213);国家自然科学基金面上项目(No. 61370024);国家自然科学基金重点项目(No. 61232001)

性,以及可以应用在不同拓扑结构的网络中的特点,对于处理各种结构特异的生物网络和基于网络的生物计算问题而言,随机游走技术都是一个极有优势的工具.近年来,随机游走技术不断地被应用到网络生物学中的相关问题研究中,并在技术上得到了新的发展.

2 随机游走模型及方法

随机游走最早是由 Pearson[4] 提出的基于物理中"布朗运动"的微观粒子运动形成的一个模型.随机游走是布朗运动的理想数学状态.它用来表示不规则的变动形式,粒子的运动是无规律的、随机的.1900 年,Bachelier[5] 首次将随机游走模型应用于股票分析上.之后,随机游走模型在股票市场、物理学、化学和生物学等多种方面得到了广泛的应用和发展.

图上的随机游走的基本思想是:给定一个网络图 $G=(V,E)$ 和一个出发点 v_1(其中 $V=\{v_1,v_2,v_3\cdots,v_n\}$ 是图 G 中顶点集合,每一个顶点对应网络中的一个节点;E 是图中边的集合 $E=\{\langle v_i,v_j\rangle \mid v_i,v_j\in V\}$,对应网络中节点的联系),随机粒子以一定的概率随机从出发点 v_1 移动到出发点 v_1 的一个邻居节点,然后将此邻居节点设为出发点,重复上述过程,得到从初始出发点 v_1 到各个节点的稳定概率,即图中各个节点到该节点 v_1 的稳定概率.

接下来,我们将按照单随机游走技术、双随机游走技术和多随机游走技术的顺序详细介绍随机游走的定义、公式.

2.1 单随机游走技术

单随机游走技术是只在单个网络中进行随机游走的技术,单随机游走技术经过改进发展,现常用的几种技术包括:简单随机游走算法,懒散随机游走算法、带限制的随机游走算法和带重启动的随机游走技术.

2.1.1 简单随机游走算法

Lovász[6] 在 1993 年发表的文章中根据随机游走的基本思想提出图上的简单随机游走概念,其计算公式为:

$$P(v_i)^{t+1}=\boldsymbol{W}^{\mathrm{T}}P(v_i)^t \tag{1}$$

其中,v_i 是当前节点,$\boldsymbol{W}$ 是图 G 对应的邻接矩阵矩阵 $\boldsymbol{A}$ 经过处理后的归一化矩阵,矩阵 $\boldsymbol{W}$ 中的值 w_{ij} 的大小表示图 G 中节点 v_i 到 v_j 的相互关系强弱,通常称为概率转移矩阵,为了简便又称为转移矩阵.转移矩阵在下述不同公式中的具体表示不同.$P(v_i)^t$ 是图中其他节点经过 t 步移动到节点 v_i 的概率.$P(v_i)^{t+1}$ 是图中其他节点公式迭代游走到节点 v_i 的概率.经过一定的步数之后,最终的概率将会收敛.简单随机游走是一个可逆的、带转移矩阵的马尔可夫链.简单随机游走是最原始的随机游走方法,其缺点在于它需要经过很长的时间才能到达距离起始节点很远的节点.

2.1.2 懒散随机游走算法

懒散随机游走是 Hodgkinson 等在文章[7] 中提及的一个概念,其计算公式为:

$$P(v_i)^{t+1}=\boldsymbol{L}^{\mathrm{T}}P(v_i)^t \tag{2}$$

其中,$\boldsymbol{L}=(\boldsymbol{W}+\boldsymbol{I})/2$,$\boldsymbol{I}$ 是单位矩阵.懒散随机游走与简单随机游走的不同在于在转移矩阵的基础上增加了一个单位矩阵,使得随机粒子在随机游走过程中有 1/2 的概率留在当前节点和有 1/2 的概率游走到当前节点的邻居节点.

2.1.3 带限制的随机游走算法

简单随机游走和懒散随机游走是无限制的随机游走,但在一些实际网络中需要带限制的游走.因此 Lei 等提出带限制的随机游走(RWS)[8],通过引入两个变量 ε 和 β 来控制随机游走.令 $f(v_i,v_j)^{t+1}$ 为经过 $t+1$ 步从节点 v_i 到节点 v_{ij} 的概率,$f(v_i,v_j)^{t+1}$ 的定义如下:

$$f(v_i,v_j)^{t+1}=\begin{cases}\max(0,P(v_i)^t w_{ij}-\varepsilon, & P(v_i)^t>0\\ \max(0,P(v_i)^t w_{ij}-\varepsilon, & P(v_i)^t=0\ \&\\ & \max(P(v_i)^t w_{ij})\geqslant\beta\\ 0, & \text{otherwise}\end{cases} \tag{3}$$

于是可以得到:

$$P(v_j)^{t+1}=\sum_i f(v_i,v_j)^{t+1}/\sum_{i,j} f(v_i,v_j)^{t+1} \tag{4}$$

式(3)中的参数 ε 用来推断随机游走过程中的概率 $f(v_i,v_j)$,当这个概率小于 0 的时候它会被重新设置为 0.当随机游走进行到一个没有到达过的点的时候,这个点的概率必定会比阈值 β 大.

2.1.4 带重启动的随机游走技术

带重启动的随机游走通过增加一个参数 α 来调节随机粒子留在原节点的概率.最早使用重启动思想的是 Page 等提出的对网页重要性进行排序的 PageRank[9] 算法.

PageRank 算法基于下列两个假设计算网页(节点)的重要性:(1)如果一个页面接收到的其他网页指向它超链接数越多,那么这个页面越重要;(2)指向页面 i 的超链接链重要性不同,重要性高的页面会通过链接向其他页面传递更多的权重.所以越是质量高的页面指向页面 i,则页面 i 越重要.以某个网页(节点)为起始网页(节点),点击起始页面中的超链接会跳转到下一个网页,以网页之间的超链接作为每次随机游走的跳转概率.并且考虑下面一个事实:每次停留在当前网页时,可能会跳转到另一个网页也可能会以一定概率 α 停留在当前网页,即可能会以 1-α 的概率跳转到超链接所链接的下个网页.则 PageRank 算法对应网络中起始节点

跳转到节点 v_i 的概率计算公式为:

$$P(v_i) = \frac{(1-\alpha)}{N} + \alpha \sum_{j \in Ne(i)} \frac{P(v_j)}{|\mathrm{Ne}(v_j)|} \quad (5)$$

式(5)中,N 是网络中的总结点数,v_j 是节点 v_i 邻居节点,$\mathrm{Ne}(v_j)$表示的是节点 v_j 邻居节点数,$P(v_j)$表示起始节点跳转到节点 v_j 的概率,$\alpha \in (0,1]$是一个用来衡量留在原节点概率的参数,$P(v_i)$是节点 v_i 经过 PageRank 算法计算得到的起始节点跳转到它的概率. 在初始时,PageRank 算法赋予每个网页相同的初始值 $1/N$,通过不断的迭代最终的到得到稳定的每个页面的重要性. 可以看出,一个节点的经过 PageRank 算法计算的值是其它节点跳转到该节点的概率,和它留在原点的概率的线性组合.

在 Page 等提出的 PageRank 算法中,其他节点跳转到节点 v_i 的初始概率都是相等的,为 $1/N$. 但是每个页面都有它自己的主题,人们对它的兴趣和关注度是不同的. 考虑到这个问题,Haveliwala 扩展了原始 PageRank 算法的定义并提出了个性化 PageRank 算法[10]. 个性化 PageRank 算法引入了一个包含所有其他节点跳转到某节点的概率向量 $\boldsymbol{x}$,个性化 PageRank 算法的值可以表示为:

$$P(v_i) = (1-\alpha)x(v_i) + \alpha \sum_{j \in \mathrm{Ne}(i)} \frac{P(v_j)}{|\mathrm{Ne}(v_j)|} \quad (6)$$

其中,N、$\mathrm{Ne}(v_j)$、$P(v_j)$、α、$P(v_i)$的含义与公式(5)相同,$x(v_i)$为概率向量 $\boldsymbol{x}$ 中其他节点跳转到 v_i 的概率.

考虑到在现实中的随机网页访问过程中,人们更容易访问到链接更多其它网页的网页,Le 提出了带权重的 PageRank 算法[11],其 $P(v_i)$的计算公式如下:

$$P(v_i) = (1-\alpha)x(v_i) + \alpha \sum_{j \in \mathrm{Ne}(i)} w(v_i,v_j)P(v_j) \quad (7)$$

其中,$w(v_i,v_j)$为列归一的矩阵 $\boldsymbol{W}$ 内节点 w_{ij}的值,它的值取决于网页的权重.

带重启动的随机游走(RWR)算法[12]也是一种改进的 PageRank 算法,它从一个或几个确定的起始节点(种子节点)开始,在每一步游走中,都会以一定的概率 α 跳转到当前节点的邻居节点,或者以 $1-\alpha$ 的概率跳转到种子节点重新开始. 它的公式可以描述为:

$$\boldsymbol{P}^{t+1} = (1-\alpha)\boldsymbol{s} + \alpha \boldsymbol{W}^{\mathrm{T}}\boldsymbol{P}^{t} \quad (8)$$

其中,$\boldsymbol{P}$ 是网络中起始节点跳转到所有节点的概率向量. $\boldsymbol{s}$ 是一个初始向量,$\boldsymbol{s}$ 中除了对应的种子节点数值设置为 1,其余所有的值都设置为 0. 由于向量 $\boldsymbol{s}$ 可以表示整个网络的初始状态,相当于 $\boldsymbol{P}^0$,因此 $\boldsymbol{P}^{t+1}$的计算公式也可以描述为:

$$\boldsymbol{P}^{t+1} = (1-\alpha)\boldsymbol{P}^{0} + \alpha \boldsymbol{W}^{\mathrm{T}}\boldsymbol{P}^{t} \quad (9)$$

PageRank 算法及其改进算法、带重启动的随机游走算法被广泛应用于蛋白质功能预测、疾病基因预测等生物计算问题. 本文在后面章节中讨论双随机游走技术和多随机游走技术的时候,默认其随机游走模型是带重启动的.

2.2　双随机游走技术

当一个网络中的节点不仅和该网络内的其他节点紧密相连,并且和其他类型的网络中的节点有着密不可分的关系时,我们设想可以通过在两个网络中随机游走得到新的结论. 如图 2 所示,给定两个网络 $G_1 = (V_1,E_1)$和 $G_2 = (V_2,E_2)$,其中 V_1、V_2 分别为网络 G_1、G_2 中顶点集合,E_1、E_2 分别为网络 G_1、G_2 中边的集合,$E_{1,2}$为网络 G_1-G_2 中节点的关系. 当我们使用双随机游走技术解决生物问题时,默认是对两种不同类型生物网络数据的关系进行处理,在所有的双随机游走算法中,根据对转移矩阵 $\boldsymbol{W}$ 的处理不同,将双随机游走技术分为基于一个异构网络的双随机游走和基于两个网络的双随机游走,下面将分别讨论这两种双随机游走技术.

2.2.1　基于异构网络中的双随机游走技术

异构网络最早是由 Katz 提出的计算机网络的一个概念[13],本文异构网络是指至少由两种不同类型的相似性网络构成的新网络. 令 $\boldsymbol{A}_1(l_1 \times l_1)$、$\boldsymbol{A}_2(l_2 \times l_2)$为两个网络 G_1 和 G_2 分别构建的对称矩阵. $\boldsymbol{R}(l_1 \times l_2)$为网络 G_1-G_2 的关系构建的关系矩阵. 由 $\boldsymbol{A}_1$、$\boldsymbol{A}_2$、$\boldsymbol{R}$ 及其的转置矩阵 $\boldsymbol{R}^{\mathrm{T}}$ 构成一个新的矩阵为$\begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{R} \\ \boldsymbol{R}^{\mathrm{T}} & \boldsymbol{A}_2 \end{bmatrix}$,这个矩阵就是异构网络对应的邻接矩阵.

基于上述异构网络的构建方法,Li 等提出了基于异构网络的随机游走 RWRH 算法[14]. RWRH 算法从多个来自网络 G_1 或者网络 G_2 的种子节点开始在异构网络上进行随机游走. 在游走过程中,来自网络 G_1 的起始节点的下一步可能会停留在网络 G_1,或者以一定的概率到达网络 G_2. 同理,来自网络 G_2 的起始节点的下一步可能会停留在网络 G_2,也可能会以一定的概率到达网络 G_1,RWRH 算法中随机游走的公式和式(9)相同,但是它的转移矩阵 $\boldsymbol{W}$ 的计算公式为:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{M}_{A_1} & \boldsymbol{M}_{\boldsymbol{R}} \\ \boldsymbol{M}_{\boldsymbol{R}^{\mathrm{T}}} & \boldsymbol{M}_{A_2} \end{bmatrix} \quad (10)$$

虽然 RWRH 算法在随机游走过程中将异构网络看成一个网络,但是由于它的转移矩阵包含了两种不同类型的网络,因此仍将它作为双随机游走类型. 式(14)中,$\boldsymbol{M}_{A_1}$和 $\boldsymbol{M}_{A_2}$分别为矩阵 $\boldsymbol{A}_1$ 和 $\boldsymbol{A}_2$ 的归一化矩阵,表示网络 G_1、G_2 内部节点的相似性比值. $\boldsymbol{M}_{\boldsymbol{R}}$ 为关系矩阵 $\boldsymbol{R}$ 的归一化矩阵,表示网络 G_1 和网络 G_2 之间的节点彼此跳转的概率. $\boldsymbol{M}_{\boldsymbol{R}^{\mathrm{T}}}$是矩阵 $\boldsymbol{M}_{\boldsymbol{R}}$ 的转置矩阵,也是归一化的. $\boldsymbol{M}_{A_1}$、$\boldsymbol{M}_{\boldsymbol{R}}$、的计算公式如下:

$$(\boldsymbol{M}_{\boldsymbol{R}})_{i,j}=\begin{cases}\dfrac{\alpha(\boldsymbol{R})_{i,j}}{\sum\limits_{j}(\boldsymbol{R})_{i,j}}, & \text{if } \sum\limits_{j}(\boldsymbol{R})_{i,j}\neq 0\\ 0, & \text{otherwise.}\end{cases} \tag{11}$$

$$(\boldsymbol{M}_{\boldsymbol{A}_1})_{i,j}=\begin{cases}\dfrac{(\boldsymbol{A}_1)_{i,j}}{\sum\limits_{j}(\boldsymbol{A}_1)_{i,j}}, & \text{if } \sum\limits_{j}(\boldsymbol{R})_{i,j}=0\\ (1-\alpha)\dfrac{(\boldsymbol{A}_1)_{i,j}}{\sum\limits_{j}(\boldsymbol{A}_1)_{i,j}}, & \text{otherwise.}\end{cases} \tag{12}$$

$\boldsymbol{M}_{\boldsymbol{A}_2}$、$\boldsymbol{M}_{\boldsymbol{R}^{\mathrm{T}}}$的计算公式与式(11)、式(12)类似,不再赘述。

$\boldsymbol{P}^0$ 表示两个网络的初始状态,$\boldsymbol{P}^0$ 的计算公式如下:

$$\boldsymbol{P}^0=\begin{bmatrix}\lambda\boldsymbol{u}_0\\(1-\lambda)\boldsymbol{y}_0\end{bmatrix} \tag{13}$$

其中,初始 $\boldsymbol{u}_0$ 和 $\boldsymbol{y}_0$ 分别表示从网络 G_1、G_2 中的种子节点开始的初始概率. 参数 $\lambda\in(0,1]$ 用于控制两个网络的权重.

重复式(13)直到 $\boldsymbol{P}^{t+1}$ 的值趋于稳定.

RWRH 算法中对异构网络中的网络进行同步计算,要求在两个网络中游走的步数相同. 考虑到两个相似性网络的拓扑特性可能不同,在两个不同相似性网络中的步数也可能不同,Zhang 等提出基于异构网络的异步重启动随机游走算法[15]. 异步重启动的转移矩阵 $\boldsymbol{W}$ 公式如下:

$$\boldsymbol{W}=\begin{bmatrix}(1-\alpha)\boldsymbol{M}_{\boldsymbol{A}_1}^{m} & \alpha\boldsymbol{M}_{\boldsymbol{R}}(\boldsymbol{M}_{\boldsymbol{R}^{\mathrm{T}}}\boldsymbol{M}_{\boldsymbol{R}})^{k}\\ \alpha\boldsymbol{M}_{\boldsymbol{R}^{\mathrm{T}}}(\boldsymbol{M}_{\boldsymbol{R}}\boldsymbol{M}_{\boldsymbol{R}^{\mathrm{T}}})^{k} & (1-\alpha)\boldsymbol{M}_{\boldsymbol{A}_2}^{n}\end{bmatrix} \tag{14}$$

其中,参数 m、n、k 为游走步数. α 为随机粒子从相似性网络 G_1 内游走 m 步或网络 G_2 游走 n 步的概率,$1-\alpha$ 为在 G_1-G_2 之间游走 $2k+1$ 步的概率. 由于 m、n、$2k+1$ 三个参数的值可能不相等,实现了在异构网络中的异步重启动.

由两种异构网络中双随机游走的计算公式可知,异构网络的不同游走技术实质上也是对转移矩阵 $\boldsymbol{W}$ 的一种处理形式.

2.2.2 平衡双随机游走算法

基于异构网络的双随机游走的转移矩阵 $\boldsymbol{W}$ 是将两个相似性网络 G_1、G_2 对应的相似性矩阵 $\boldsymbol{A}_1$,$\boldsymbol{A}_2$ 以及两个网络 G_1、G_2 之间的关系对应的矩阵 $\boldsymbol{R}$ 以及 $\boldsymbol{R}$ 的转置矩阵 $\boldsymbol{R}^{\mathrm{T}}$ 共四个矩阵集成在一起构成异构矩阵. 当异构矩阵中数量很多时,实际上增加了很多计算量. Xie 等受到图匹配问题的启发,将全局的两个不同生物相似性网络 G_1、G_2 之间的关系等同于这两个网络 G_1、G_2 之间的配对关系,提出了基于两个网络的双随机游走 BiRW 算法[16,17],包括三种变体:平衡的双随机游走算法 BiRW_bl、不平衡双随机游走算法 BiRW_avg、顺序双随机游走算法 BiRW _seq.

其中平衡的双随机游走算法 BiRW_bl 的计算公式如下:

$$\boldsymbol{W}^{t+1}=(1-\alpha)\boldsymbol{A}_1\times\boldsymbol{W}^{t}\times\boldsymbol{A}_2+\alpha\boldsymbol{R} \tag{15}$$

其中,$\boldsymbol{A}_1$,$\boldsymbol{A}_2$ 分别表示网络 G_1 和网络 G_2 对应的相似性矩阵,t 为迭代次数,参数 $\alpha\in(0,1]$ 为重启动的概率值. 矩阵 $\boldsymbol{R}$ 存储先验知识,即已知的网络 G_1 和 G_2 内节点之间的关系. $\boldsymbol{W}$ 是转移矩阵. 在平衡双随机游走中,两个网络中的节点的关系可以通过式(21)中转移矩阵 $\boldsymbol{W}$ 同时以迭代的方式在两个相似性矩阵 $\boldsymbol{A}_1$、$\boldsymbol{A}_2$ 中扩展,以得到关于网络 G_1、G_2 节点之间的关系.

2.2.3 不平衡的双随机游走算法

由于两个生物网络的拓扑特性不同,在随机游走的过程中游走的步数也应该不同,不加区分的在两个网络中保持相同的步数随机游走显然忽略了这一点,使得随机游走的效果大打折扣. 与基于异构网络的异步重启动随机游走算法相类似,不平衡双随机算法引入两个参数 l_l,l_r 分别控制随机粒子在两个相似性网络 G_1、G_2 中游走的步数,通过迭代更新转移矩阵 $\boldsymbol{W}$ 的值,获得两个网络 G_1 中节点的关系. 不平衡双随机游走算法 BiRW_avg[16] 的具体计算过程如下:

$$\boldsymbol{W}_{\text{left}}^{t+1}=\alpha\boldsymbol{A}_1\times\boldsymbol{W}^{t}+(1-\alpha)\boldsymbol{R},\quad \text{if}(t<l_l) \tag{16}$$

$$\boldsymbol{W}_{\text{right}}^{t+1}=\alpha\boldsymbol{W}^{t}\times\boldsymbol{A}_2+(1-\alpha)\boldsymbol{R},\quad \text{if}(t<l_r) \tag{17}$$

$$\boldsymbol{W}^{t+1}=\frac{\lambda_{\text{left}}\times\boldsymbol{W}_{\text{left}}^{t+1}+\lambda_{\text{right}}\times\boldsymbol{W}_{\text{right}}^{t+1}}{\lambda_{\text{left}}+\lambda_{\text{right}}} \tag{18}$$

由式(16)和式(17)可知,随机粒子在网络 G_1 中游走是通过左乘网络 G_1 对应的相似性矩阵 $\boldsymbol{A}_1$ 完成的,在网络 G_2 中游走是通过右乘网络 G_2 对应的相似性矩阵 $\boldsymbol{A}_2$ 完成的,因此不平衡的随机游走在网络 G_1,G_2 中游走又可以称为左走,右走,下面沿用这个叫法. 式(18)中,λ_{left}和 λ_{right}分别控制随机粒子在网络 G_1、G_2 中游走后的影响占总影响的权重.

Peng 等在进行蛋白质功能预测时提出的 UBiRW 算法[18] 也是不平衡的双随机游走算法,其计算过程与 BiRW_avg 相似.

2.2.4 顺序双随机游走算法

与不平衡双随机游走算法不同,顺序双随机游走算法虽然也引入了变量 l_l 和 l_r 控制左走和右走的步数,但是顺序双随机游走在游走过程中直接将左游走的结果在右游走中迭代. 顺序双随机游走算法 BiRW _seq 的计算公式为:

$$\boldsymbol{W}_{\text{left}}^{t+1}=\alpha\boldsymbol{A}_1\times\boldsymbol{W}^{t}+(1-\alpha)\boldsymbol{R},\quad \text{if}(t<l_l) \tag{19}$$

$$\boldsymbol{W}^{t+1}=\alpha\boldsymbol{W}_{\text{left}}^{t}\times\boldsymbol{A}_2+(1-\alpha)\boldsymbol{R},\quad \text{if}(t<l_r) \tag{20}$$

不同的生物计算问题的类型不同,所采用的随机游走模型也应该不同. 在具体的计算问题上,这三种基于两个网络的随机游走算法各有应用. 其中应用较多

的是不平衡的双随机游走.

2.3　多随机游走技术

由于在很多生物计算问题中,密切相关的生物数据类型可能有多个,多种不同类的生物数据彼此之间存在错综复杂的相互关系. 例如:疾病与基因、miRNA、lncRNA 和环境因子都存在密切关系,而基因与 miRNA 和 lncRNA 也存在很密切的关系,基因、miRNA、lncRNA 都会受到环境的影响. 基于此,随机游走技术经过发展,产生了多随机游走技术. 多随机游走技术当前包括三随机游走技术和四随机游走技术,下面将分别进行叙述这两种多随机游走技术.

三随机游走技术是三个不同但有密切关系的生物网络中节点关系预测的有力工具. 如图 1 所示,$\boldsymbol{A}_1(l_1\times l_1)$、$\boldsymbol{A}_2(l_2\times l_2)$、$\boldsymbol{A}_3(l_3\times l_3)$为三种不同生物相似性网络 G_1、G_2、G_3 构建的相似性矩阵,$\boldsymbol{R}_1(l_1\times l_2)$、$\boldsymbol{R}_2(l_3\times l_2)$、$\boldsymbol{R}_3(l_1\times l_3)$分别为网络 G_1-G_2、G_2-G_3、G_3-G_1 之间的已知关系矩阵. $\boldsymbol{R}_Q(l_1\times l_2)$、$\boldsymbol{R}_D(l_3\times l_2)$为要预测的 G_1-G_2、G_2-G_3 之间的关系矩阵. 目前,应用三随机游走技术主要关注的仍然是其中某两种数据的关系预测,而将另一种数据作为辅助信息添加进来. . 下面,以最新提出来的三随机游走算法 ThrRW[19] 为例,讨论三随机游走算法的主要计算过程. $\boldsymbol{R}_Q$ 的主要计算公式如下:

$$\begin{aligned}&(\boldsymbol{R}_Q)^{t+1}_{\text{left}}=\alpha\boldsymbol{A}_1\times\boldsymbol{R}^t_Q+(1-\alpha)\boldsymbol{R}_1\\&(\boldsymbol{R}_Q)^{t+1}_{\text{mid}}=\boldsymbol{A}_3\times\boldsymbol{R}^t_D\end{aligned},\text{if}(t<l_{l1})\quad(21)$$

$$(\boldsymbol{R}_Q)^{t+1}_{\text{right}}=\alpha\boldsymbol{R}^t_Q\times\boldsymbol{A}_2+(1-\alpha)\boldsymbol{R}_1,\text{if}(t<l_{r1})\quad(22)$$

$$\boldsymbol{R}^{t+1}_Q=\frac{\lambda_{\text{left}}(\boldsymbol{R}_Q)^{t+1}_{\text{left}}+\lambda_{\text{mid}}(\boldsymbol{R}_Q)^{t+1}_{\text{mid}}+\lambda_{\text{right}}(\boldsymbol{R}_Q)^{t+1}_{\text{right}}}{\lambda_{\text{left}}+\lambda_{\text{mid}}+\lambda_{\text{right}}}\quad(23)$$

如式(21)和式(22)所示,在每一步游走过程中,通过左乘网络 G_1 对应的相似性矩阵 $\boldsymbol{A}_1$ 以及右乘网络 G_2 对应的相似性矩阵 $\boldsymbol{A}_2$ 拓展了潜在的网络 G_1 和 G_2 节点之间的关联关系. 矩阵 $\boldsymbol{R}_1$ 存储已知的网络 G_1、G_2 节点之间的已知关系. 由于网络 G_1 与网络 G_2 的拓扑结构的差异性,引入两个参数 l_{l1}、l_{r1} 分别限定在两个网络 G_1、G_2 上的游走的最大迭代数. 式(23)中通过网络 G_1 和 G_3 对应的关系矩阵 $\boldsymbol{R}_D$ 左乘网络 G_3 对应的相似性矩阵 $\boldsymbol{A}_3$ 获得新的 G_3-G_1 关联关系.

矩阵 $\boldsymbol{R}_D$ 也可用同样的方式得到.

四随机游走技术与三随机游走技术类似,将四个不同但有密切关系的生物网络数据融合在一起,通过扩展更多不同的生物信息,获得需要的潜在的网络节点关系预测. 如图 2 所示,给定 $\boldsymbol{A}_1(l_1\times l_1)$、$\boldsymbol{A}_2(l_2\times l_2)$、$\boldsymbol{A}_3(l_3\times l_3)$、$\boldsymbol{A}_4(l_4\times l_4)$为四种不同网络 G_1、G_2、G_3、G_4 构建的相似性矩阵,$\boldsymbol{R}_1(l_1\times l_2)$、$\boldsymbol{R}_2(l_3\times l_2)$、$\boldsymbol{R}_3(l_3\times l_1)$、$\boldsymbol{R}_4(l_4\times l_2)$、$\boldsymbol{R}_5(l_4\times l_1)$分别为 G_1-G_2、G_2-G_3、G_3-G_1、G_4-G_2、G_4-G_1 四种网络两两之间的已知关系矩阵. $\boldsymbol{R}_{\text{md}}(l_1\times l_2)$、$\boldsymbol{R}_{\text{ed}}(l_3\times l_2)$、$\boldsymbol{R}_{\text{em}}(l_3\times l_1)$、$\boldsymbol{R}_{\text{gm}}(l_4\times l_2)$、$\boldsymbol{R}_{\text{gd}}(l_4\times l_1)$为要更新的网络 G_1-G_2、G_2-G_3、G_3-G_1、G_4-G_2、G_4-G_1 之间的关系矩阵. 与三随机游走技术类似,目前四随机游走技术仍然关注的是某两种不同生物数据之间的关系预测. 最新提出的四随机游走算法 FourRW[20] 预测关系矩阵 $\boldsymbol{R}_{\text{md}}$即网络 G_1-G_2 之间潜在关系的总体上可以分为三个大步骤. 首先,在 G_1-G_2-G_3 网络上游走;然后在 G_1-G_2-G_4 网络上游走;最后在 G_1-G_3 网络上游走.

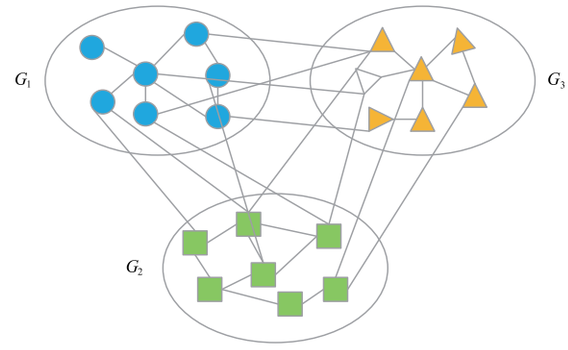


图1　三随机游走网络示意图

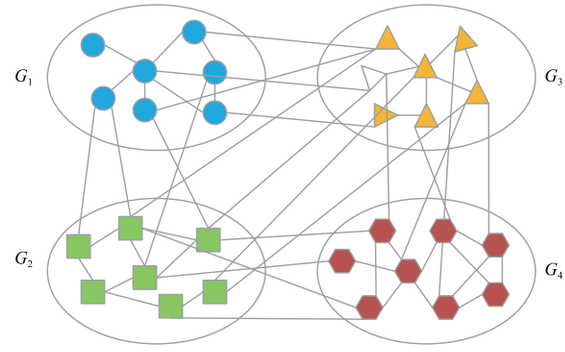


图2　四随机游走网络示意图

步骤 1　在网络 G_1-G_2-G_3 网络中随机游走与三随机游走思想类似,其计算公式为:

$$(\boldsymbol{R}_{\text{ed}})^{t+1}_{e}=\alpha\boldsymbol{A}_3\,\boldsymbol{R}_{\text{ed}}{}^{t}+(1-\alpha)\boldsymbol{R}_2,\text{if}(t\leqslant l_1)\quad(24)$$

$$(\boldsymbol{R}_{\text{ed}})^{t+1}_{\text{right}}=a\,\boldsymbol{R}_{\text{ed}}{}^{t}\boldsymbol{A}_2+(1-\alpha)\boldsymbol{R}_2,\text{if}(t<l_2)\quad(25)$$

$$(\boldsymbol{R}_{\text{ed}})^{t+1}_{\text{mid}}=\boldsymbol{R}_3\times\boldsymbol{R}_{\text{md}}{}^{t}\quad(26)$$

$$(\boldsymbol{R}_{\text{ed}})^{t+1}=\boldsymbol{R}_{\text{em}}\times\boldsymbol{R}_1\quad(27)$$

式(24)、(25)表示通过在网络 G_3 中左走以及在网络 G_2 中右走拓展了潜在的网络 G_2 和 G_3 节点之间的关联关系$(\boldsymbol{R}_{\text{ed}})_{\text{left}}$、$(\boldsymbol{R}_{\text{ed}})_{\text{right}}$. 引入两个参数 l_1、l_2 分别限定在两个网络 G_2、G_3 上的游走的最大迭代数. 通过式(26)左乘 G_1-G_2 网络的已知关系矩阵 $\boldsymbol{R}_3$ 将随机游走扩展到 G_1 网络,获得 G_3-G_1 关联关系$(\boldsymbol{R}_{\text{ed}})_{\text{mid}}$. 式(27)中通过网络 G_1 和 G_3 对应的关系矩阵 $\boldsymbol{R}_{\text{em}}$右乘网络 G_1-G_2 已知关系

矩阵 $\boldsymbol{R}_1$,获得新的 G_3-G_1 关联关系 $\boldsymbol{R}_{ed}$. 再通过加权方法调整矩阵 $(\boldsymbol{R}_{ed})_{left}$、$(\boldsymbol{R}_{ed})_{right}$、$(\boldsymbol{R}_{ed})_{mid}$、$\boldsymbol{R}_{ed}$ 的权重,得到更新的 G_2-G_3 之间的预测关系 $\boldsymbol{R}_{ed}$.

$$(\boldsymbol{R}_{em})_{left}^{t+1}=\alpha\boldsymbol{A}_3\boldsymbol{R}_{em}^{t}+(1-\alpha)\boldsymbol{R}_3,\quad f(t<l_3) \tag{28}$$

$$(\boldsymbol{R}_{em})_{right}^{t+1}=\alpha\boldsymbol{R}_{em}^{t}\boldsymbol{A}_1+(1-\alpha)\boldsymbol{R}_3,\quad f(t<l_4) \tag{29}$$

$$(\boldsymbol{R}_{em})^{t+1}=\boldsymbol{R}_{ed}{}^{t}\times\boldsymbol{R}_1 \tag{30}$$

$$(\boldsymbol{R}_{em})_{mid}^{t+1}=\boldsymbol{R}_3\times\boldsymbol{R}_{md}^{t} \tag{31}$$

式(28)、(29)表示通过在网络 G_3 中左走以及在网络 G_1 中右走拓展了潜在的网络 G_1 和 G_3 节点之间的关联关系 $(\boldsymbol{R}_{em})_{left}$、$(\boldsymbol{R}_{em})_{right}$. 参数 l_3、l_4 分别限定在两个网络 G_3、G_1 上的游走的最大迭代数. 式(30)中通过网络 G_2 和 G_3 对应的关系矩阵 $\boldsymbol{R}_{ed}$ 右乘网络 G_1、G_2 的已知关系矩阵 R_1 获得新的 G_3-G_1 关联关系 $\boldsymbol{R}_{em}$. 式(31)中通过网络 G_2 和 G_3 对应的关系矩阵 $\boldsymbol{R}_{md}$ 左乘 G_3-G_1 已知关系矩阵 $\boldsymbol{R}_3$ 获得新的 G_3-G_1 关联关系 $(\boldsymbol{R}_{em})_{mid}$. 再通过加权方法对 $\boldsymbol{R}_{em}$ 矩阵进行处理,得到更新的 G_1-G_3 之间的预测关系 $\boldsymbol{R}_{em}$.

接着在网络 G_1-G_3 上游走,其计算公式如下:

$$(\boldsymbol{R}_{md})_{left}^{t+1}=\alpha\,\boldsymbol{A}_1\times\boldsymbol{R}_{md}{}^{t}+(1-\alpha)\boldsymbol{R}_1,\quad \text{if}(t<l_{10}) \tag{32}$$

$$(\boldsymbol{R}_{md})_{right}^{t+1}=\alpha\boldsymbol{R}_{md}^{t}\times\boldsymbol{A}_3+(1-\alpha)\boldsymbol{R}_1,\quad \text{if}(t<l_{10}) \tag{33}$$

$$(\boldsymbol{R}_{md})_{em}^{t+1}=\boldsymbol{R}_{em}^{t}\times\boldsymbol{R}_2 \tag{34}$$

$$(\boldsymbol{R}_{md})_{e}^{t+1}=\boldsymbol{R}_3^{\mathrm{T}}\times\boldsymbol{R}_{ed}^{t} \tag{35}$$

$$(\boldsymbol{R}_{md})_{gm}^{t+1}=(\boldsymbol{R}_{gm}{}^{t})^{\mathrm{T}}\times\boldsymbol{R}_4 \tag{36}$$

$$(\boldsymbol{R}_{md})_{right}^{t+1}=\boldsymbol{R}_3\times\boldsymbol{R}_{md}^{t} \tag{37}$$

步骤 2 在网络在 G_1-G_2-G_4 网络上游走与步骤 1 类似,不再赘述。

步骤 3 在网络 G_1-G_2 中进行不平衡的双随机游走获得关系矩阵 $(\boldsymbol{R}_{md})_m$、$(\boldsymbol{R}_{md})_d$,再分别在 G_1-G_3-G_2、G_2-G_3-G_1、G_1-G_4-G_2、G_2-G_4-G_1 中游走探测新的 G_1-G_2 网络的潜在关系矩阵 $(\boldsymbol{R}_{md})_{em}$、$(\boldsymbol{R}_{md})_{e}$、$(\boldsymbol{R}_{md})_{gm}$、$(\boldsymbol{R}_{md})_{d}$,并通过一定的加权算法对加权获得网络 G_1-G_2 的关系矩阵 $\boldsymbol{R}_{md}$.

3 随机游走技术在网络生物学中的应用

自从“网络生物学”这个概念被提出以后,系统生物学的研究有了突破性进展. 生物系统是一种复杂系统,它的一个显著特点为:多个系统组件彼此相互作用. 正是这些相互作用构成一张张复杂的生物网络,进而形成复杂生命活动的生物系统. 随机游走技术作为有效的网络分析技术之一,近年来在很多基于网络的生物计算问题中得到应用. 本节将详细介绍随机游走技术在蛋白质功能预测、关键蛋白质识别、疾病基因预测、疾病相关非编码 RNA 预测、药物相关预测上等方面的应用,讨论随机游走技术是如何有效求解这些生物计算问题的.

3.1 蛋白质功能预测

蛋白质是生物体内承担着重要功能的大分子,是生命体的重要组成部分,正确的注释蛋白质功能有助于从分子水平上理解生命活动的运转机制,揭开生命活动的面纱. 早期的蛋白质功能预测的方法大都基于序列相似性或者同源性方法[21,22]等等. 实验表明,基于序列相似性或者同源性的方法结果并不准确,因为相同序列的蛋白质在形成不同结构时执行的功能并不相同,相同功能的蛋白质也不一定拥有相同的序列,蛋白质功能不仅与其序列密切相关,并与其折叠结构息息相关.

随着大量蛋白质相互作用数据的产生,基于蛋白质相互作用网络和已知的蛋白质功能信息来预测未知的蛋白质功能的预测方法获得了研究者的关注. 有研究者指出两个相互作用的蛋白质倾向于共享一个功能,两个相似的蛋白质功能也通常共同注释同一个蛋白质[23],目前已提出了一系列基于网络的蛋白质功能预测方法[18,19,24~29]. 其中,随机游走技术是基于网络的蛋白质功能预测的最主要方法之一.

Freschi 等提出了 ProteinRank 算法[24],通过结合全局网络特性将 PageRank 算法公式中跳转到网络中其他蛋白质节点的跳转概率替换成了表示蛋白质拥有哪些功能的矩阵,来预测蛋白质的功能. 但 ProteinRank 算法对于蛋白质节点参与度小的蛋白质网络,预测性能会比较差. Yu 等[25]利用蛋白质相互作用网络、功能相似性网络和已知的蛋白质-功能关系网络构建异构网络,由基于异构网络的双随机游走技术结合核函数构造一个多分类器,在异构网络中进行随机游走并将随机游走的结果进行分类,达到了分层预测蛋白质功能的目的. Hu 等[26]则从功能的依赖性方面考虑,提出使用已知蛋白质-功能关系作为初始值分别对蛋白质相互作用网络和功能相似性网络进行平衡的双随机游走探测蛋白质功能. Peng 等[18]在平衡的双随机游走基础上进一步考虑蛋白质相互作用网络和功能相似性网络的拓扑结构差异,根据不同层级的蛋白质可能具有不同的功能相似性以及不同层次的蛋白质功能注释可以标注的蛋白质功能不同,通过控制随机游走在两个网络中游走的步数,利用不平衡的双随机游走预测蛋白质功能. 另外 Ma[27]和 Deng[28]结合 K 近邻技术和随机游走技术以分类的方式对蛋白质功能进行预测.

除了上述提到的 ProteinRank 算法、平衡的双随机游走技术以及非平衡的双随机游走技术,三随机游走技术也被用来进行蛋白质功能预测. 结构域(domian)是蛋白质中独立的功能结构,不同的蛋白质所拥有的

结构域通常因为蛋白质功能的不同存在着差异. 2009 年,Zhang 等人[29]首次提出结合蛋白质相互作用网络数据与蛋白质结构域数据进行蛋白质功能预测的方法. Peng 等提出的通过不平衡的三随机游走模型 ThrRW 算法[19]将蛋白质相互作用网络数据、蛋白质结构域数据、功能注释数据相融合共同进行功能预测. 其使用的不平衡的三随机游走算法中采用了边聚集系数对蛋白质相互作用网络加权,从一定程度上减少了蛋白质相互作用网络数据的假阳性和假阴性,减少了蛋白质相互作用网络信息中噪声对蛋白质功能预测造成的影响. 其实验结果也佐证了蛋白质结构域信息在蛋白质功能预测过程中起到的正向作用.

3.2　关键蛋白质识别

不同的蛋白质对生命活动的作用是不一样的,关键蛋白质是指那些在细胞生命过程中起关键作用的蛋白质. 关键蛋白质对生物来说必不可少,它的去除或者突变会导致生物体死亡或停止生长、发育. 识别关键蛋白质有助于从系统水平上理解生命活动的生物机理,且在药物设计等方面也具有重要作用. 研究表明,蛋白质的关键性与它们在蛋白质相互作用网络里的"地位"密切相关. 研究者通常采用各种中心性计算方法来衡量每个蛋白质在网络中的"地位". 例如,Jeong 等发表在 nature 上的文献[30]指出相互作用较多的蛋白质对细胞的生存作用更大,即蛋白质相互作用网络中蛋白质的度越高,越倾向于是关键的(即著名的"中心性-致死性"法则). 之后很多研究者从网络拓扑角度来研究关键蛋白质. 近年来,随机游走技术也被有效地用来分析生物网络中节点的重要性,并用于关键基因或关键蛋白质的预测[31~33].

Del Rio 等[32]通过使用原始的 PageRank 算法在代谢网络中给基因排序、选取排序分数最高的几个作为候选关键基因. Yang 等结合蛋白质复合物参与度信息提出了一种基于带重启动的随机游走的关键蛋白质预测算法 RWP[33]来识别关键蛋白质.

根据关键蛋白质的保守特性和关键蛋白质在相互作用网络中趋向于彼此之间紧密连接成簇(cluster)的假设,Peng 等结合蛋白质的同源信息和其拓扑特性在带权重的 PageRank 算法的基础上提出了新的预测关键蛋白质的方法 ION[34]. 方法 ION 通过在加权的蛋白质相互作用网络中不断随机游走迭代更新蛋白质的 PageRank 值来预测关键蛋白质. 其使用的蛋白质相互作用网络加权方法来自文献[35]中所使用的边聚集系数.

3.3　疾病基因预测

研究发现,大多数疾病可以从基因层面反映出来. 很多复杂疾病都属于多基因疾病,而非孟德尔式的单基因疾病. 这些复杂疾病往往是多个基因和其他因素共同作用的结果. 基于生物网络预测疾病基因、研究复杂疾病的发病机理已经展开了很长一段时间. 基于生物网络预测疾病基因的问题模型主要有两大类:一类是给定一个要查询的特定疾病 Q 和一些已知的与该疾病存在密切关系的疾病及对应的基因信息,预测查询的特定疾病潜在的疾病基因;另一类是给定部分已知疾病和基因之间的关系预测潜在的新的疾病-基因关系.

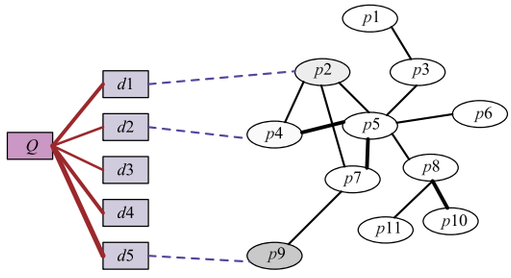
针对第一种疾病基因预测问题模型,Vanunu 等提出了基于随机游走技术的 PRINCE 算法[36]. 如图 3 所示,对于一个给定的疾病 Q,$d1 \sim d5$ 表示和它具有不同程度表型相似的疾病,并作为先验知识初始化带重启动的单随机游走. 将蛋白质相互作用网络经过一定的处理(相似性计算、归一化)之后作为转移矩阵 $\boldsymbol{W}$ 参与到随机游走过程中. PRINCE 算法将疾病-基因关系在蛋白质相互作用网络构成的转移矩阵 $\boldsymbol{W}$ 中不断的迭代更新直到稳定. 依据最后游走得到的结果对潜在的基因进行打分,选出得分最高的基因,并将这个基因作为最有可能和给定疾病 Q 相关的基因. 我们在 PRINCE 算法的基础上通过进一步改进先验知识的计算方法和转移矩阵 $\boldsymbol{W}$ 的计算,提出了随机游走技术的疾病基因预测方法[37].

针对第二种疾病基因预测问题模型,Chen 等[38]提出将已知基因作为个性化的 PageRank 算法中的先验知识,通过在蛋白质相互作用网络上随机游走,预测疾病基因. Kohler 等[12]以全局的方式采用带重启动的单随机游走技术对相互作用网络中的蛋白质进行排序来预测疾病基因. Kohler 等提出的算法首先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,并通过处理将蛋白质相互作用网络转换成随机游走公式中的归一化转移矩阵 $\boldsymbol{W}$. 初始概率向量 $\boldsymbol{P}^0$ 存储已知疾病基因信息,所有已知疾病基因的初始概率是相等的,为总已知疾病基因数的倒数. 该算法从已知疾病基因开始,通过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可能是潜在的疾病基因.

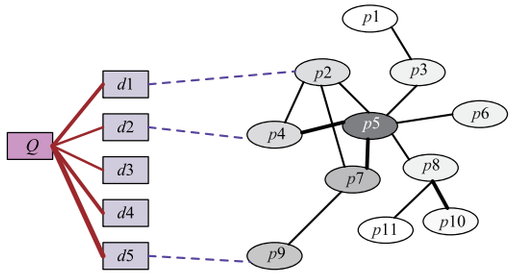
后来,考虑到疾病基因在相互作用网络中的模块化结构特性[12],Le 等采取给模块化的疾病基因加权的方式提出 ORIENT 算法[11]. 该算法采用邻居加权的方式进行预处理,并采用带重启动的单随机游走技术在加权的蛋白质相互作用网络中游走迭代更新获得潜在的疾病基因. 此外,Chen 等[39]提出了一个 K 步随机游走算法,通过 K 步简单随机游走来排序疾病基因.

随着双随机游走技术的发展,研究者通过构建由蛋白质相互作用网络、表型相似性网络和疾病-基因之

间的关系网络组成的异构网络,提出了新的疾病基因预测方法. 例如,Li 等[14]通过构建这样的异构网络提出了基于异构网络的随机游走算法来处理疾病基因预测问题,获得了很好的效果. Guo 等[40]使用平衡的双随机游走算法通过在构建的蛋白质相互作用网络和疾病相似性网络上同时独立的游走预测疾病基因关系. Xie 等[16]在 Guo 的基础上提出一种不平衡的双随机游走方法通过在蛋白质相互作用网络和疾病相似性网络分别游走不同的步数来获取疾病-基因关系.



(*a*) 应用随机游走之前



(*b*) 应用随机游走之后

图3 PRINCE算法流程[35].*Q*表示要查询的疾病, *d*1~*d*5表示与*Q*有关联关系的疾病, *Q*与*d*1~*d*5的连线粗细分别代表了*Q*与*d*1~*d*5的关联关系的强弱.*p*1~*p*11表示蛋白质, *p*1~*p*11之间的连线代表他们的相互作用关系.*d*1~*p*2、*d*2~*p*4、*d*5~*p*9之间的虚线表示已知的疾病-基因关系.

3.4 疾病相关非编码 RNA 预测

人类基因组的测序结果显示,人类染色体中大概只有大约2%的基因序列编码蛋白质,其余非编码区通常被认为是无用的. 然而,越来越多的实验证据表明,大部分的非编码区在生物体的多个生命进程中也扮演着十分重要的角色. 这类不编码蛋白质的 RNA 被称为非编码 RNA(non-coding RNA,ncRNA). 非编码蛋白质根据其长度,又分为 miRNA(microRNA)和 lncRNA(long nocoding RNA),这两种非编码蛋白质都和疾病有着息息相关的联系.

3.4.1 miRNA-疾病关系预测

miRNA 是长度为 ~22nt 的一类非编码调控 RNA,它通过结合 3'端配对的方式与信使 RNA 进行绑定,进而抑制靶标 mRNA 的表达. 近年来,越来越多的研究表明 miRNA 与多种疾病的发生发展密切相关. 由于功能相近的 miRNA 通常和表型相似的疾病相关[41],因此 miRNA 和疾病之间的关系可以通过构建 miRNA 功能相似性网络和疾病表型网络来预测[42]. Chen 等[43]认为采用全局网络相似度量方式预测 miRNA 和疾病之间的关系要比传统的基于局部信息的方法更适合. 因此 Chen 等通过构建 miRNA 功能相似性网络,采用带重启动的单随机游走算法在 miRNA 功能相似性网络中进行游走来预测疾病相关的 miRNA. Chen 等[44]基于 OMIM 数据库中疾病相似性数据,采用带重启动的单随机游走算法在疾病相似性网络中预测潜在的 miRNA. 此方法由于实验验证过的 miRNA 调控 mRNA 的数量不足导致得到的结果存在很高的假阳性和假阴性. 为了减少样本的误差,Xuan 等[45]基于 miRNA 功能相似性网络,有标签样本和无标本样本加权方式构建转移矩阵,采用带重启动的随机游走方法预测 miRNA-疾病关系.

考虑到环境因子和 miRNA 及疾病表型都具有很密切的关系,Peng 等[46]提出基于三随机游走技术的 miRNA-疾病关系预测方法 ThrRWMDE. 建立 miRNA 功能相似性网络、疾病表型网络、环境因子相似性网络这三种相似性网络,并通过在上述三个网络和它们之间的关联网络里进行不平衡的三随机游走获得 miRNA-疾病之间的预测结果. 由于 Peng 等的方法涉及六个网络,六个网络中的拓扑特性和结构特性不同,因此使用随机游走可以通过调节重启动和各个网络游走的步数来消除网络中拓扑特性和结构特性带来的偏差.

考虑到基因和 miRNA、疾病、环境因子有着不可忽视的关系,在文献[47]的基础上,peng 等又通过将 miRNA、disease、环境因子、基因四种生物数据结合起来,提出了四随机游走 FourRW 算法[20]. 四随机游走算法具体步骤在本文第二节多随机游走技术中已经讨论过,不再赘述.

3.4.2 lncRNA-疾病关系预测

越来越多的证据表明 lncRNA 在许多人类疾病中发挥重要的作用. 预测 lncRNA 与疾病的相关性不仅有助于在 lncRNA 水平上了解疾病的致病机制并且可以加速识别疾病诊断中的生物标志物. 由于功能相近的 lncRNA 通常和相似表型的疾病相关,因此 lncRNA 和疾病之间的关系可以通过构建 lncRNA 功能相似性网络和疾病表型网络来预测. Sun 等[48]提出通过在构建的 lncRNA 功能相似性网络上进行带重启动的单随机游走预测疾病的候选 lncRNA. 这种方法的缺点在于只能应用于有已知和疾病相关 的 lncRNA 上,限制了算法的应用. zhou 等[49]通过集成 lncRNA 相似性网络、疾病相似性网络和已知 lncRNA-疾病关系构建异构网络并在异构网络上进行随机游走,预测疾病的候选 lncRNA. 由于 zhou 等构建的 lncRNA 相似性网络是基于功能相似的

lncRNA 往往和有明显相互作用的 miRNA 相连的假设,所以无法应用到没有已知 lncRNA-miRNA 关系的问题中. 考虑到已知 lncRNA-疾病关系仍然是十分稀少的,Chen 等[50]通过融合 lncRNA 表达相似性和疾病语义相似性网络设置随机游走的初始向量 $\boldsymbol{P}^0$. 因此可以应用于新的(没有和任何疾病相关联)lncRNA 与疾病关系预测.

3.5　药物相关预测

传统的新药研发存在周期长、耗资大、风险高、成功率低的问题. 如今在药物研发方面的投入不断增长,但是实际产出却停滞不前,因此如何有效提高药物研发的效率是制药企业所面临的挑战性问题. 针对这个问题,药物-靶标关联预测以及药物重定位技术等,正在成为药物研发的重要策略. 随着随机游走技术被广泛应用于基于网络的生物计算问题,这种技术也逐渐被应用于药物靶标预测与药物重定位等问题中.

目前药物靶标的预测基本上都是基于相似的药物易于靶向相似靶标的假设. Chen 等[51]提出一种扩展随机游走算法到药物-靶标异构网络上的方法,预测潜在的药物-靶标关联的方法. 首先,基于药物的化学结构信息和已知的药物-靶标关联计算药物相似性,构建药物网络;基于靶标蛋白的氨基酸序列信息计算靶标相似性,构建靶标网络;然后集成药物相似性网络、靶标相似网络和已知药物-靶标关联网络,创建药物-靶标异构网络;最后利用随机游走技术在所构建的异构网络上进行游走,基于网络的全局信息预测未知的、潜在的药物-靶标关联. 和传统的监督学习方法相比,该方法能有效的融合不同的网络来预测药物-靶标关系.

Emig 等[52]提出一种集成局部算法、随机游走算法和网络扩散算法的网络模型来预测药物靶标关系. 该方法融合了基因表达数据,蛋白质相互作用数据和已知的药物靶标关系,采用不同局部的和全局的网络方法来预测药物靶标关系该方法优势在于能够有效利用网络的全局信息和局部信息来预测潜在药物-靶标关系. 此外该方法也能够有效的预测新的药物对应的靶标.

Chen 等[53]提出结合 A* 启发式搜索算法和随机游走模型的药物-靶标预测方法. 该方法通过融合蛋白质相互作用数据和基因表达数据对相互作用网络进行加权. 然后利用 A* 启发式搜索算法识别药物治疗相互作用网络中包含显著基因集的动态子网,然后基于该子网采用随机游走方法预测药物相关的靶标.

由于药物的杂泛性,即通常会与多个靶标发生相互作用,这种药物靶向的非特异性会产生药物的副作用. 因此,通过计算方法预测药物的副作用对用药安全方面有重大意义. Rahmani 等[54]提出一种扩张随机游走算法来预测药物的潜在副作用. 首先构建反映药物之间关联的药物网络,然后基于药物副作用集合,扩张药物网络到包含药物和副作用的集成网络,最后在该网络上进行随机游走预测药物-副作用之间的关联度.

药物重定位技术,即挖掘已有药物的新适应症. 由于疾病种类和已知药物的数量繁多,完全通过实验筛选已知药物的新用途仍然具有很高的成本. 随着组学数据和药物信息学数据的积累,药物重定位进入到了理性设计和实验筛选相结合的阶段. 通过计算方法发现药物的新适应症,即基于计算方法的药物重定位,已经成为计算生物学和系统生物学的重要研究方向.

Luo 等[55]提出一种基于集成的相似性度量方法和双向随机游走的药物重定位方法,基于相似药物易于关联相似疾病的假设,预测当前存在药物的新适应症. 首先利用集成相似性度量方法计算药物相似性、疾病相似性,所计算的相似性值能更好的反应药物间的相似度和疾病间的相似性度;在此基础上,构建了药物-疾病异构网络;采用双向随机游走算法在异构网络上进行游走,有效利用药物相似性信息、疾病相似性信息和已知的药物-疾病关联信息,为给定药物的所有候选疾病进行打分,分数越高,表示给定药物能用于治疗该候选疾病的可能性越大.

3.6　其他网络生物问题

随机游走技术除在上述介绍的蛋白质功能预测、关键蛋白质识别、疾病基因预测、疾病相关非编码 RNA 预测、药物相关预测等问题上得到了有效应用,在其他一些生物计算问题上也得到了一定的应用. 例如,生物网络模块挖掘、疾病相似性计算、生物网络中节点距离度量、生物子网络查询、肿瘤影像分割等.

生物网络模块挖掘　有助于识别执行特定功能的蛋白质复合物,理解细胞功能的运行机制。Enright 等人[56]使用 MCL 算法划分蛋白质相互作用网络,来挖掘网络中的蛋白质复合物。Peng 等认为随机游走过程中当前节点的邻居节点权重值是不同的,当前节点倾向于跳转到更可能和它一起形成模块的邻居节点,于是提出 WPNCA[57]算法用来挖掘蛋白质复合物。WPNCA 算法本质是使用一个加权 PageRank-Nibble 算法筛选出初始的蛋白质复合物,利用核-附属结构进一步筛选获得最终的蛋白质复合物。

疾病相似度计算　是很多生物计算问题(包括疾病基因、疾病相关非编码 RNA 预测等)的基础[58]. 疾病相似度计算问题本身也可以通过采用随机游走技术来提升预测效果. Li 等[59]通过使用单随机游走算法来预计算疾病和基因之间的相关性,再进一步利用预测信息来进行疾病的相似性计算. 当然,其他相似性计算问题(例如药物相似性、功能相似性、miRNA 相似性、ln-

cRNA 相似性等)也可以采用相似的计算策略.

生物网络中节点距离度量 是处理很多生物计算问题的有效工具和基础,例如蛋白质相互作用预测、蛋白质功能预测等. Erten 等[60]提出采用带重启动的随机游走算法来替代带限制的随机游走算法来计算从不同节点开始的网络中所有节点的平稳概率. 该方法的基本思想是:当选择某个节点作为起始节点,经过带重启动的随机游走计算得到的平稳概率表示这个节点和其他节点之间的密切关系,也就是显示了起始节点的拓扑结构. 对每个起始节点,都有一个向量存储所有节点的平稳概率,通过计算不同起始节点的平稳概率向量之间的皮尔逊相关系数可以得到这些起始节点的相似性. Chipman 等[61]通过使用带重启动的随机游走方法测量两个节点之间的拓扑相似性,并预测遗传相互作用.

生物子网络查询 随着高通量技术的发展,产生了越来越多的大型生物网络. 如何从这些大型网络中找到具有特定结构或功能的子网络,即生物子网络查询,是一个非常重要的生物计算问题[62 63]. Sahraeian 等[64]使用随机游走技术分别对查询网络和目标网络进行处理,找到查询网络中与目标网络相似的子网络.

肿瘤影像分割 文献[65,66]认为可以通过将医学图像(肿瘤 PET、CT 图片)划分为小区域,将小区域作为网络中的节点通过不同区域之间的关系大小以一定的方式构建网络. 据此,Ju[65]通过将随机游走和图分割两种算法相结合,同时分割 PET 和 CT 图像识别肺肿瘤得到一个初始的肿瘤轮廓. Hu[66]将随机游走算法应用于脑部胶质瘤 MR 图像构成的网络中,寻找 MR 图像中脑部胶质瘤位置.

4 研究难点和发展趋势

在当前的网络生物学分析方法中,随机游走技术经过不断地发展已经成为其中很重要的一类计算方法. 此外,包括机器学习、矩阵分解等的其他方法技术也得到了长足发展. 从目前已有的一些研究成果看,随机游走技术在很多具体问题上仍然具有一定的优势. 当使用随机游走技术时,可根据具体问题使用不同的预处理方法、或者改进随机游走算法使其更适用具体问题,达到更好的效果. 此外,由于生物问题本身很复杂,不同生物网络也各有其生物特性,加之目前的生物网络数据还不完善,且存在较大噪声,如何在选择适用的现有随机游走技术、进一步发展随机游走技术来处理基于生物网络的生物计算问题仍需进一步研究. 主要的技术难点及研究方向如下:

(1)有效的预处理技术. 由于不同生物网络的拓扑特性和生物特性不同,高通量技术测得的生物数据有噪声,为更好的应用随机游走技术处理基于网络的生物计算问题,应采用一定的预处理技术对生物网络数据进行处理. 在生物计算问题中,常用的和随机游走技术结合使用的预处理技术包括对转移矩阵的处理(拉普拉斯归一化[67]、邻居加权[11]等)技术、初始概率向量的处理技术以及对参与随机游走的网络本身的降噪处理技术. 如何针对每种生物计算问题进一步开发适合的预处理技术是进一步研究的方向.

(2)利用生物特征的随机游走技术. 为适应不同的生物计算问题,随机游走技术发展了不同的模型,例如:带重启动的单随机游走、基于异构网络的双随机游走、平衡双随机游走、非平衡双随机游走、三随机游走、四随机游走等等. 但是,由于随机游走技术并不是为解决生物计算问题而研发的,所以并没有考虑生物数据本身的特征. 而生物计算问题除了复杂外还具有较强的生物特征,并且不同的生物数据不是孤立存在的,彼此之间存在复杂的关联关系,如何通过充分分析生物数据之间的复杂关联关系和有效挖掘生物数据背后隐藏的特征来发展新的面向生物特征和融合多元数据的随机游走技术是未来该方向的最主要的挑战之一.

(3)基于多网络的随机游走模型. 虽然,目前已经提出了三随机游走和四随机游走技术,但这仍然是对双随机游走技术的扩充,并且预测的目标也仍然是多种关联关系中的一种,并不能同时预测多种关联关系. 正如前面本文中讨论的,一种生物网络可能与多个其他类型的生物网络存在密切关系. 如图 8 所示,当人们研究疾病时,它不仅与基因的突变等相关,可能与 miRNA、lncRNA 以及环境等都存在密切关系. 而这些与疾病密切相关的不同因素彼此之间也都存在密切关系. 因此,如何开发基于多关系网络的随机游走技术以及可以同时预测多种关联关系的随机游走技术是未来需要重点关注的方向.

(4)在动态网络中应用随机游走模型. 本文第三章讨论的随机游走应用在基于网络的生物计算问题,都假设其使用的生物网络是静态的. 而生物系统本质上是随着时间动态演化的[68~71],基因的调控关系、蛋白质之间的相互作用等都是随着时间动态变化的,其构成的生物网络也是动态变化的. 由于高通量技术的发展,使得获取大量随时间变化的表达数据成为可能. 目前已有一些构建时间序列下的或组织特异性基因表达谱数据的动态子网计算方法被提出[72]. 因此,在今后的工作中,可以探讨如何将随机游走技术应用到基于动态网络的生物计算问题上,或者面向动态网络进一步发展新的随机游走技术.

(5)随机游走技术和其他技术结合. 随机游走技术在一些生物计算问题上单独应用取得了不错的效果,但也有一些生物计算问题模型仅仅使用随机游走技术

不能获得很好的效果. 目前,已有一些文章将随机游走技术与矩阵分解、多标签学习方法、深度学习、A * 启发式搜索算法等方法结合在一起取得良好的效果,如何寻找合适的技术与随机游走技术结合,是未来关注的重点.

5　总结

生物网络在生物问题处理方面有着重要的地位,随机游走则是研究和处理各种基于网络的问题的有效工具. 本文从介绍随机游走模型的基本定义和基本方法开始,以基于网络的生物问题为基点,详细阐述了随机游走技术是如何有效地应用到基于网络的生物计算问题包括关键蛋白质识别、蛋白质功能预测、疾病相关的基因/非编码 RNA 预测、药物靶标预测、药物重定位等问题中. 最后,本文讨论了随机游走技术在网络生物学研究中存在的问题以及未来的研究方向. 总的来说,虽然随机游走技术在网络生物学中的应用已经发展一段时间,但未来在生物计算问题中仍然大有可为.

参考文献

[1] Barabasi A L, Oltvai Z N. Network biology: understanding the cell's functional organization[J]. Nature reviews genetics, 2004, 5(2): 101 – 113.

[2] BERKHIN P. A survey on pagerank computing[J]. Internet Mathematics, 2005, 2(1): 73 – 120.

[3] CODLING E A, Plank M J, Benhamou S. Random walk models in biology[J]. Journal of the Royal Society Interface, 2008, 5(25): 813 – 834.

[4] PEARSON K. The problem of the random walk[J]. Nature, 1905, 72(1865): 294.

[5] BACHELIER L. Théorie De La Spéculation[M]. Paris: Gauthier-Villars, 1900.

[6] Lovász L. Random walks on graphs: A survey[J]. Combinatorics, Paul Erdos is Eighty, 1993, 2(2): 353 – 397.

[7] HODGKINSON L, KARP R M. Algorithms to detect multiprotein modularity conserved during evolution[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics(TCBB), 2012, 9(4): 1046 – 1058.

[8] LEI C, RUAN J. A random walk based approach for improving protein-protein interaction network and protein complex prediction[A]. Bioinformatics and Biomedicine (BIBM)[C]. Philadelphia: IEEE, 2012. 1 – 6.

[9] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: Bringing order to the web[R]. California: Stanford InfoLab, 1998.

[10] HAVELIWALA T H. Topic-sensitive pagerank[A]. Proceedings of the 11th International Conference on World Wide Web[C]. Honolulu: ACM, 2002. 517 – 526.

[11] LE D H, KWON Y K. Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization[J]. Computational Biology and Chemistry, 2013, 44(1): 1 – 8.

[12] K? HLER S, BAUER S, HORN D, et al. Walking the interactome for prioritization of candidate disease genes[J]. The American Journal of Human Genetics, 2008, 82(4): 949 – 958.

[13] KATZ R H, BREWER E A. The Case for Wireless Pverlay Networks[M]. New York: Springer US, 1996. 621 – 650.

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

[15] 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[J]. 生物物理学报, 2015(1): 33 – 44.
ZHANG S Y, ZHANG S W. Prediction of the risk of lung cancer risk in lung cancer based on the two element network with asynchronous reset random walk algorithm[J]. Acta Biophysica Sinica, 2015(1): 33 – 44. (in Chinese)

[16] XIE M, HWANG T, KUANG R. Prioritizing disease genes by bi-random walk[J]. Advances in Knowledge Discovery and Data Mining, 2012, 7302: 292 – 303.

[17] XIE M, HWANG T H, KUANG R. Reconstructing disease phenome-genome association by bi-random walk[J]. Bioinformatics, 2012, 1(02): 1 – 8.

[18] PENG W, WANG J, CHEN L, et al. Predicting protein functions by using unbalanced bi-random walk algorithm on protein-protein interaction network and functional interrelationship network[J]. Current Protein and Peptide Science, 2014, 15(6): 529 – 539.

[19] 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.

[20] PENG W, LAN Wei, WANG Jianxin, et al. Predicting microRNA-disease associations by walking on four biological networks[A]. Bioinformatics and Biomedicine (BIBM)[C]. Shenzhen: BIBM, 2016. 299 – 302.

[21] TATUSOV R L, KOONIN E V, LIPMAN D J. A genomic perspective on protein families[J]. Science, 1997, 278(5338): 631 – 637.

[22] MARCOTTE E M, PELLEGRINI M, THOMPSON M J, et al. A combined algorithm for genome-wide prediction of protein function[J]. Nature, 1999, 402(6757): 83 – 86.

[23] ZHANG X F, DAI D Q. A framework for incorporating functional interrelationships into protein function predic-

tion algorithms[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics(TCBB),2012,9(3):740 -753.

[24] FRESCHI V. Protein function prediction from interaction networks using a random walk ranking algorithm[A]. IEEE 7th International Symposium on BioInformatics and BioEngineering[C]. Silicon valley, USA: BIBE, 2007. 42 -48.

[25] YU G, RANGWALA H, DOMENICONI C, et al. Protein function prediction using multilabel ensemble classification [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(4): 1045 -1057.

[26] HU P, JIANG H, EMILI A. Predicting protein functions by relaxation labelling protein interaction network[J]. BMC Bioinformatics, 2010, 11(S1): S64.

[27] 马吉权,贾翠翠,张军杰. 基于随机游走的蛋白质功能预测算法设计与实现[J]. 黑龙江大学工程学报,2015(03):73 -78.
MA Gequan, JIA Cuicui, ZHANG Junjie. Prediction of protein based on random walk algorithm design and implementation[J]. Engineering Journal of Heilongjiang University, 2015(03): 73 -78. (in Chinese)

[28] 邓小龙. 基于随机游走的蛋白质功能预测方法的研究[D]. 长春:吉林大学,2012.
DENG X L. Study on protein function prediction based on random walk[D]. Changchun: Jilin University, 2012. (in Chinese)

[29] ZHANG S, CHEN H, LIU K, et al. Inferring protein function by domain context similarities in protein-protein interaction networks [J]. BMC Bioinformatics, 2009, 10(1): 395.

[30] JEONG H, MASON S P, BARABáSI A L, et al. Lethality and centrality in protein networks[J]. Nature, 2001, 411(6833): 41 -42.

[31] PENG W, WANG J, CHENG Y, et al. UDoNC: an algorithm for identifying essential proteins based on protein domains and protein-protein interaction networks [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics(TCBB), 2015, 12(2): 276 -288.

[32] DEL RIO G, KOSCHüTZKI D, COELLO G. How to identify essential genes from molecular networks? [J]. BMC Systems Biology, 2009, 3(1): 102.

[33] 杨莉萍,路松峰,黄钰. 一种基于随机游走模型的关键蛋白质预测方法[J]. 华中农业大学学报,2016,35(6):86 -91.
YANG Liping, LU Songfeng, HUANG Yu. A method for predicting essential proteins based on random walk mode [J]. Journal of Huazhong Agricultural University, 2016, 35(6): 86 -91. (in Chinese)

[34] PENG W, WANG J, WANG W, et al. Iteration method for predicting essential proteins based on orthology and protein-protein interaction networks[J]. BMC Systems Biology, 2012, 6(1): 87.

[35] WANG J, LI M, CHEN J, et al. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(3): 607 -620.

[36] VANUNU O, MAGGER O, RUPPIN E, et al. Associating genes and protein complexes with disease via network propagation[J]. PLoS Comput Biol, 2010, 6(1): e1000641.

[37] LI M, ZHENG R, LI Q, et al. Prioritizing disease genes by using search engine algorithm[J]. Current Bioinformatics, 2016, 11(2): 195 -202.

[38] CHEN J, ARONOW B J, JEGGA A G. Disease candidate gene identification and prioritization using protein interaction networks[J]. BMC Bioinformatics, 2009, 10(1): 73.

[39] CHEN J, BARDES E E, ARONOW B J, et al. Toppgene suite for gene list enrichment analysis and candidate gene prioritization[J]. Nucleic Acids Research, 2009, 37(suppl 2): W305 -W311.

[40] GUO X, GAO L, WEI C, et al. A computational method based on the integration of heterogeneous networks for predicting disease-gene associations[J]. PloS One, 2011, 6(9): e24171.

[41] OMER A, SINGH P, YADAV N K, et al. MicroRNAs: role in leukemia and their computational perspective[J]. Wiley Interdisciplinary Reviews: RNA, 2015, 6(1): 65 -78.

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

[43] CHEN X, LIU M X, YAN G Y. RWRMDA: predicting novel human micro RNA-disease associations[J]. Molecular BioSystems, 2012, 8(10): 2792 -2798.

[44] CHEN H, ZHANG Z. Prediction of associations between OMIM diseases and MicroRNAs by random walk on OMIM disease similarity network [J]. The Scientific World Journal, 2013, 2013: 204658.

[45] XUAN P, HAN K, Guo Y, et al. Prediction of potential disease-associated microRNAs based on random walk[J]. Bioinformatics, 2015, 31(11): 1805 -1815.

[46] PENG W, LAN W, YU Z, et al. Predicting microrna-disease associations by random walking on multiple networks [A]. International Symposium on Bioinformatics Research and Applications. Springer International Publishing [C]. Minsk, Belerus: ISBRA, 2016. 127 -135.

[47] GUPTA R A, SHAH N, WANG K C, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote

cancer metastasis[J]. Nature, 2010, 464(7291): 1071 – 1076.

[48] SUN J, SHI H, WANG Z, et al. Inferring novel lncRNA – disease associations based on a random walk model of a lncRNA functional similarity network[J]. Molecular BioSystems, 2014, 10(8): 2074 – 2081.

[49] ZHOU M, WANG X, LI J, et al. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network[J]. Molecular BioSystems, 2015, 11(3): 760 – 769.

[50] CHEN X, YOU Z, YAN G, et al. IRWRLDA: improved random walk with restart for lncRNA-disease association prediction[J]. Oncotarget, 2016, 7(36): 57919 – 57931.

[51] CHEN X, LIU M X, YAN G Y. Drug-target interaction prediction by random walk on the heterogeneous network[J]. Molecular BioSystems, 2012, 8(7): 1970 – 1978.

[52] EMIG D, IVLIEV A, PUSTOVALOVA O, et al. Drug target prediction and repositioning using an integrated network-based approach[J]. PLoS One, 2013, 8(4): e60618.

[53] CHEN L C, YEH H Y, YEH C Y, et al. Identifying co-targets to fight drug resistance based on a random walk model[J]. BMC Systems Biology, 2012, 6(1): 5.

[54] RAHMANI H, WEISS G, MéNDEZ-LUCIO O, et al. ARWAR: A network approach for predicting adverse drug reactions[J]. Computers in Biology and Medicine, 2016, 68: 101 – 108.

[55] LUO H, WANG J, LI M, et al. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm[J]. Bioinformatics, 2016, 32(17): 2664 – 2671

[56] ENRIGHT A J, VAN DONGEN S, OUZOUNIS C A. An efficient algorithm for large-scale detection of protein families[J]. Nucleic Acids Research, 2002, 30(7): 1575 – 1584.

[57] PENG W, WANG J, ZHAO B, et al. Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2015, 12(1): 179 – 192.

[58] SUTHRAM S, DUDLEY J T, CHIANG A P, et al. Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets[J]. PLoS Comput Biol, 2010, 6(2): e1000662.

[59] LI P, NIE Y, YU J. Fusing literature and full network data improves disease similarity computation[J]. BMC Bioinformatics, 2016, 17(1): 326.

[60] ERTEN S, BEBEK G, KOYUTüRK M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks[J]. Journal of Computational Biology, 2011, 18(11): 1561 – 1574.

[61] CHIPMAN K C, SINGH A K. Predicting genetic interactions with random walks on biological networks[J]. BMC Bioinformatics, 2009, 10(1): 17

[62] YOON B J, QIAN X, SAHRAEIAN S M E. Comparative analysis of biological networks: Hidden markov model and markov chain-based approach[J]. IEEE Signal Processing Magazine, 2012, 1(29): 22 – 34.

[63] 赵建邦,董安国,高琳. 一种用于生物网络数据的频繁模式挖掘算法[J]. 电子学报,2010,38(8):1803 – 1807.
ZHAO J, DONG A, GAO L. An algorithm for frequent pattern mining in biological networks[J]. Acta Electronica Sinica, 2010, 38(8): 1803 – 1807. (in Chinese)

[64] SAHRAEIAN S M E, YOON B J. RESQUE: Network reduction using semi-Markov random walk scores for efficient querying of biological networks[J]. Bioinformatics, 2012, 28(16): 2129 – 2136.

[65] 鞠薇. 基于随机游走和图割算法的 PET-CT 肺肿瘤分割[D]. 苏州:苏州大学,2015.
JU W. Random walk and graph cut for co-segmentation of lung tomor on PET-CT images[D]. Suzhou: Suzhou University, 2015. (in Chinese)

[66] 胡洁. 基于图论的医学图像分割随机游走算法研究[D]. 广州:南方医科大学,2013.
HU J. Research on random walk algotithm based graph theory on the application of medical image segmentation[D]. Guangzhou: Southern Medical University, 2013. (in Chinese)

[67] ZHAO Z Q, HAN G S, YU Z G, et al. Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization[J]. Computational Biology and Chemistry, 2015, 57C: 21 – 28.

[68] WANG J, PENG X, PENG W, et al. Dynamic protein interaction network construction and applications[J]. Proteomics, 2014, 14(4 – 5): 338 – 352.

[69] LI M, WU X, WANG J, et al. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data[J]. BMC Bioinformatics, 2012, 13(1): 109.

[70] 李敏,孟祥茂. 动态蛋白质网络的构建、分析及应用研究进展[J]. 计算机研究与发展,2017,54(6):1281 – 1299.
LI M, MENG X. The construction, analysis, and applications of dynamic protein-protein interaction networks[J]. Journal of Computer Research and Development, 2017, 54(6): 1281 – 1299. (in Chinese)

[71] LI M, MENG X, ZHENG R, et al. Identification of protein complexes by using a spatial and temporal active protein interaction network[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, DOI:

10.1109/TCBB.2017.2749571.
[72] SUN S Y,LIU Z P,ZENG T,et al. Spatio-temporal analysis of type 2 diabetes mellitus based on differential expression networks [J]. Scientific Reports, 2013, 2013, 3(2268):1 – 13.

作者简介

李 敏 女,1978 年出生,辽宁人. 教授、博士生导师. 中国计算机学会会员. 主要研究方向为生物信息学和数据挖掘.
E-mail: limin@mail.csu.edu.cn

王晓桐 女,1994 年出生,河南人. 硕士研究生. 主要研究方向为生物信息学和随机游走技术.
E-mail:kathrynwalls@foxmail.com