文章编号：2095-2163(2019)01-0047-04 中图分类号：TP301.6 文献标志码：A

一种基于异构网络算法的药物-蛋白关联性研究方法

徐 婷，龚家瑜，宋 晖
（东华大学 计算机科学与技术学院，上海 201620）

摘 要： 药物与蛋白质间关联性的研究，有助于药物的重新定位和发现药物新的使用途径，是网络药理学的重要研究内容。应用现有网络随机游走算法预测药物与蛋白质间新的关联时，一般直接在药物-蛋白质二分图网络内进行随机游走，并且不断重复此过程，这种方法效率很低，还会遗漏药物-药物相似性网络和蛋白-蛋白相似性网络中部分拓扑信息。鉴于此，本文提出一种异构网络异步重启随机游走算法（Drug Restart Walk Random Prediction，DRWRP），构建药物-蛋白质异构网络，深层次挖掘二者间潜在的关联性。该算法分别在药物相似性网络、蛋白质相似性网络以及药物-蛋白质二分图网络中进行随机游走，然后在网络间不停跳转，反复迭代后形成稳态概率向量，最终得到潜在最优关联。仿真实验表明，本文提出的算法可以有效预测药物与蛋白质间新的关联，多数预测结果获得了文献证据支持。

关键词： 药物-蛋白关联性；重启随机游走；异构网络

A drug-protein relevance research method based on isomeristic network algorithm

XU Ting, GONG Jiayu, SONG Hui
（School of Computer Science and Technology, Donghua University, Shanghai 201620, China）

【Abstract】 The study of the relationship between drugs and proteins is helpful for repositioning drugs and discovering new ways of drug use. It is an important research content of network pharmacology. When using the existing network random walk algorithm to predict the new association between drugs and proteins, random walk is generally carried out directly within the drug-protein binary graph network, and this process is constantly repeated. This method is very inefficient. Some topology information in the drug-drug similarity network and protein-protein similarity network is also omitted. In view of this, this paper proposes an asynchronous restart random walk algorithm (DRWRP) for heterogeneous networks to build drug-protein isomerism networks and deeply explore the potential correlation between the two. The algorithm moves randomly in drug similarity networks, protein similarity networks, and drug-protein binary graph networks, and then jumps between networks. After repeated iterations, steady state probability vectors are formed, and the potential optimal correlation is finally obtained. The simulation results show that the proposed algorithm can effectively predict the new association between drug and protein.

【Key words】 pharmaco-protein association; restart random walk; isomeristic networks

0 引 言

在医药行业中，研究一种新的药物所耗费的周期长、投资大、风险高，但成功率却一直偏低。现如今，网络药理学飞速发展，药物重定位被认为是药物研发策略中风险和效益比最好的策略之一。研究表明，导致相同类似药理作用的分子一般在同一个生物模块内，如蛋白质复合体[1]、代谢通路[2]和蛋白质网络[3]。因此，可以利用这种模块性及已知的药物-蛋白作用特性预测潜在的新的关联。目前，基于网络的药物-蛋白质关联性预测方法大致可划分为 2 类，即单源网络方法和多源网络整合方法。单源网络方法多采用蛋白质网络进行药物-蛋白质的关联性预测。多源网络整合方法将多源网络（如蛋白相似性网络、药物相似性网络）信息进行潜在关联性预测。如 Lage[5] 等人利用贝叶斯模型整合蛋白相似性网络和药物相似性网络，实现对药物相关蛋白质复合物的预测；Li 等人[7] 基于药物-蛋白质二元网络，提出二元网络重启随机游走算法来实现关联性预测，该方法可提高药物-蛋白质关联性预测准确率，但构造的状态转移矩阵较为稀疏，这种策略可能遗漏蛋白质网络中的局部拓扑信息，导致预测性降低。本文提出一种异构网络异步重启随机游走算法，将药物网络、蛋白质网络及药物-蛋白质二分图网络三者构建为异构网络，并在内进行随机游走和网络间的跳转，最后验证算法的有效性。

1 材料与方法

蛋白质相似性数据来源于 UniProt 数据库，包含 313 个节点，其邻接矩阵用 $\boldsymbol{A}_P$ 表示，该网络反映蛋

作者简介：徐 婷（1993-），女，硕士研究生，主要研究方向：数据工程。

收稿日期：2018-09-22

白质与蛋白质之间的量化相似性关系。蛋白-蛋白相似性网络中的节点表示靶标蛋白,而蛋白质之间的相似程度则由邻边权重来量化,取值范围为[0,1],越接近1则表示相似程度越高。

从DrugBank数据库获取药物数据,构建药物相似性网络,该网络反映药物与药物之间的量化相似性关系。从获取的信息看,药物-药物相似性网络中含有773个药物节点,邻接矩阵用 A_D 表示。从Chembl数据库中查询获取已知药物-蛋白质相互作用数据,构建药物-蛋白质网络,该二分图网络表示药物与蛋白质的对应关系,包含773个药物节点,313个蛋白质节点。经过分析网络得到药物和蛋白质之间直接相连的边有215条,用B表示其邻接矩阵,为了算法运行效率提高和计算简便,邻接矩阵的元素只能为0或者1,其中1表示蛋白质与药物间有关系,0表示蛋白质与药物无关系。药物-蛋白质异构网络的邻接矩阵用 $\boldsymbol{A}=\begin{bmatrix} A_P & B \\ B^T & A_D \end{bmatrix}$ 表示,邻接矩阵 $\boldsymbol{A}_P$、$\boldsymbol{A}_D$ 和 $\boldsymbol{B}$ 的维数分别为 $s \times s$、$w \times w$ 和 $s \times w(s=313, w=773)$。图1为药物-药物、药物-蛋白、蛋白-蛋白组成的异构网络。

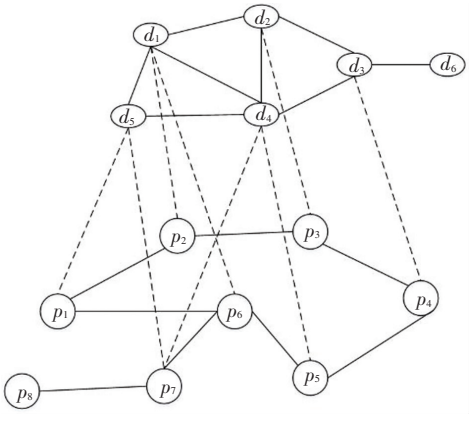

图1 药物-蛋白质异构网络模型

Fig. 1 Model of drug-protein isomerism network

DRWRP算法基本思想:在建立的多个网络中,从某一节点出发,按照一定概率向相邻的节点跳转,下一个节点即为下一个状态,重复初始状态行为。网络中的所有节点都可进行跳转。具体过程类似于数学中的马尔科夫链。算法的数学表示如下:

$$P^{t+1}=(1-\lambda)MP^t+\lambda P^0 \tag{1}$$

其中,P^0 表示初始游走概率向量;P^t 表示为 t 时刻的状态,这个状态和初始状态相似;λ 是经验参数[$\lambda \in (0,1)$],用来确定重启概率;$\boldsymbol{M}$ 是转移矩阵,反映网络的拓扑特性。本文认为当 $t+1$ 时刻的状态 P^{t+1} 与前一时刻的状态 P^t 的范数收敛到某个很小的值 ε 的时候,游走不再进行,整个网络处于静止状态,在实验中,将 ε 设为 10^{-6}。对静止的网络进行得分计算,对于某一个节点来讲,可以计算出下一步跳转到任一节点的概率,根据把下一节点按照概率由大到小进行排序,从而进行推荐。本文采用的异构网络存在2种游走,第一种为同源节点网络内的游走,即药物-药物相似性网络和蛋白-蛋白相似性网络,第二种为异源节点网络间的游走,即药物-蛋白质相似性网络中游走。DRWRP算法状态转移概率矩阵定义为:

$$M=\begin{bmatrix} (1-\alpha)M_P & \alpha M_{PD}(M_{DP}M_{PD}) \\ \alpha M_{DP}(M_{PD}M_{DP}) & (1-\alpha)M_D \end{bmatrix} \tag{2}$$

其中,$(1-\alpha)$ 为节点随机游走后停止的概率;$\boldsymbol{M}_P$ 为蛋白相似性网络状态转移矩阵;$\boldsymbol{M}_D$ 为药物相似性网络的状态转移矩阵;$\boldsymbol{M}_{PD}$ 为从蛋白相似性网络到药物相似性网络的状态转移概率矩阵;M_{DP} 为从药物性相似性网络到蛋白相似性网络的状态转移概率矩阵。

蛋白 p_i 跳转到蛋白 p_j 的转移概率定义为:

$$(M_P)_{ij}=P(p_i \mid p_j)=(A_P)_{ij}/\sum_j (A_P)_{ij} \tag{3}$$

药物 d_i 跳转到药物 d_j 的转移概率定义为:

$$(M_D)_{ij}=P(d_i \mid d_j)=(A_D)_{ij}/\sum_j (A_D)_{ij} \tag{4}$$

蛋白 p_i 跳转到药物 d_j 的转移概率定义为:

$$(M_{PD})_{ij}=P(d_j \mid p_i)=\begin{cases}(B)_{ij}/\sum_i (B)_{ij} & \sum_i (B)_{ij}\neq 0 \\ 0\end{cases} \tag{5}$$

药物 d_i 跳转到蛋白 p_j 的转移概率定义为:

$$(M_{DP})_{ij}=P(p_j \mid d_i)=\begin{cases}(B)_{ji}/\sum_j (B)_{ji} & \sum_j (B)_{ji}\neq 0 \\ 0\end{cases} \tag{6}$$

初始游走概率向量定义为:

$$P^0=\begin{bmatrix}\mu^0 \\ \nu^0\end{bmatrix} \tag{7}$$

其中,μ^0 和 ν^0 分别为蛋白相似性网络和药物相似性网络游走初始概率向量,指定药物对应节点的游走初始概率为1,其它药物节点的游走初始概率为0。设已知指定药物对应的靶标蛋白为 h 个,如果蛋白 p_i 为靶标蛋白,则 $\mu_i^0=\frac{1}{h}$,否则 $\mu_i^0=0$,$\sum_i \mu_i^0=1$。

DRWRP算法具体描述如下:

输入：状态转移矩阵 M，初始游走概率向量 P^0 和重启的概率 λ。

输出：蛋白关联性得分 μ^{∞}。

步骤：

(1)初始化 P 值为 P^0；

(2) 初始化 P^t 的值为 P；

(3) 对 P^{t+1} 进行迭代；

(4) 重复步骤(3) 直至 $\| P^{t+1} - P^t \|^1 \leqslant 10^{-6}$；

(5)将蛋白按照关联性得分 μ^{∞} 的值按照降序排列；

(6)输出排在前 p% 的蛋白作为算法识别的关联性蛋白。

2　实验结果与分析

本文以药物 DB00619 为例，根据 DrugBank 数据库的记录，该药物的靶标蛋白有 9 个，在上述筛选的数据库中有 8 个对应的靶标蛋白。本文选取准确率作为评价指标，对计算结果先排序后筛选，选择排列在前 1%、5%、10%、15%的蛋白质作为识别的关联性蛋白，再与已知的靶标蛋白数据集进行比对。实验结果如图 2 所示。

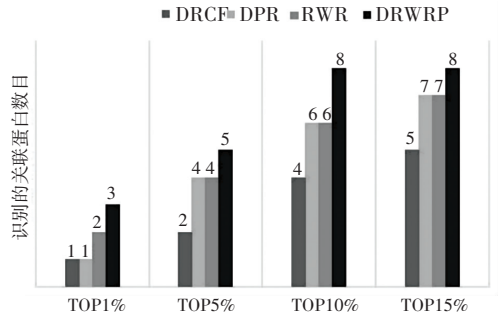

图 2　DRWRP 与其它关联研究方法比较

Fig. 2　Comparison of DRWRP and other correlation research methods

从图 2 可知，本文提出的 DRWRP 算法识别的关联蛋白质数量与采用 DRCF 算法、DPR 算法和 RWR 算法识别的关联蛋白质数量相比明显更多。无论是在前 1%、前 5%、前 10%，还是前 15%的样本水平上，DRWRP 算法的预测命中率都比其它算法高 15%以上。总体来说，DRWRP 算法具有较好的预测性能。

在上式中，对参数 α 和参数 λ 的取值都采用了经验值 0.5。为了研究这 2 个参数对 DRWRP 算法预测性能的影响，先固定其中一个为 0.5，然后调整另一个参数。实验结果如图 3 和图 4 所示。结果表明，当 $\alpha = 0.5$，$\lambda = 0.5$ 时，DRWRP 算法的性能总体上最高。

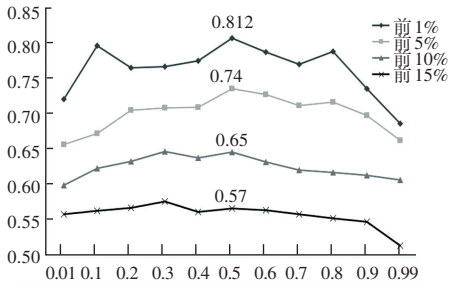

图 3　参数 λ 对 DRWRP 预测准确性的影响

Fig. 3　Influence of λ on the accuracy of DRWRP

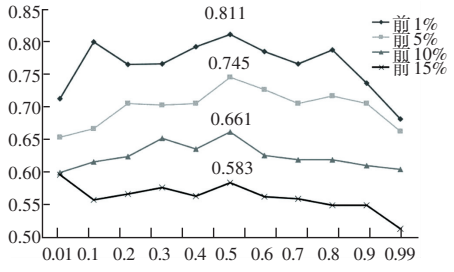

图 4 参数 α 对 DRWRP 预测准确性的影响

Fig. 4　Influence of α on the accuracy of DRWRP

3　结束语

本文提出了一种异构网络异步重启游走算法，构建了药物-蛋白质异构网络，深层次挖掘药物与蛋白质之间的潜在关联性。该算法分别在药物-药物相似性网络、蛋白质-蛋白质相似性网络以及药物-蛋白质二分图网络中进行随机游走，然后在网络间不停跳转，反复迭代后形成稳态概率向量，最终得到药物与蛋白质间的潜在最优关联。对已知药物靶标蛋白的验证结果表明，与现有的随机游走算法和推荐算法相比，DRWRP 算法体现出更好的预测性能。

参考文献

[1] 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，2015，6(3)：73-78.

[2] LIM J, HAO T, SHAW C, et al. A protein-protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration[J]. Cell, 2006, 125(4): 801-814.

[3] VAN DRIEL M A, BRUGGEMAN J, VRIEND G, et al. A text-mining analysis of the human phenome[J]. European Journal of Human Genetics, 2006, 14(5): 535-542.

[4] KÖHLER 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.

(下转第 52 页)