基于双加权投票的蛋白质功能预测

唐家琪[1] 吴璟莉[1,2,3] 廖元秀[1] 王金艳[1,2,3]

(广西师范大学计算机科学与信息工程学院 广西 桂林 541004)[1]

(广西师范大学广西多源信息挖掘与安全重点实验室 广西 桂林 541004)[2]

(广西区域多源信息集成与智能处理协同创新中心 广西 桂林 541004)[3]

摘 要 蛋白质是完成重要生物活动所必需的分子。准确掌握蛋白质功能,将对生命科学研究及应用起到极大的促进作用。高通量技术的发展产生了海量的蛋白质序列,利用计算技术预测大规模蛋白质功能已成为当今生物信息学的核心任务之一。目前,作为蛋白质功能预测的研究热点,基于蛋白质相互作用网络的预测方法在降低数据噪声影响、充分利用网络拓扑特性及整合多源数据等方面仍不够完善。文中结合带阻力随机游走得到的全局拓扑相似度,及功能术语的语义相似度,设计了一种双加权投票蛋白质功能预测算法BiWV;并在此基础上整合了生物通路信息,提出了带生物通路的双加权投票算法——BiWV-P。在酿酒酵母和人类数据集上,对所提算法与TMC,UBiRW和ProHG 3种算法的预测效果进行对比分析。实验结果显示,算法BiWV和BiWV-P能够有效预测蛋白质功能,并在许多数据集上获得较其他算法更高的微正确率与微F1。

关键词 蛋白质相互作用网络,功能预测,随机游走,语义相似度,生物通路

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.04.035

Prediction of Protein Functions Based on Bi-weighted Vote

TANG Jia-qi[1] WU Jing-li[1,2,3] LIAO Yuan-xiu[1] WANG Jin-yan[1,2,3]

(School of Computer Science & Information Engineering, Guangxi Normal University, Guilin, Guangxi 541004, China)[1]

(Guangxi Key Laboratory of Multi-Source Information Mining & Safety, Guangxi Normal University, Guilin, Guangxi 541004, China)[2]

(Guangxi Regional Multi-Source Information Integration & Intelligent Processing Cooperation Innovation Center, Guilin, Guangxi 541004, China)[3]

Abstract Proteins are the essential molecules to accomplish important biological activities. It will greatly promote the advance of life science research and application to accurately grasp their functions. A tremendous amount of protein sequences has been generated with the development of high-throughput techniques. Thus, prediction of large-scale protein functions with computation technology has become one of the key tasks in bioinformatics today. Currently, the prediction method based on protein-protein interaction network, which is a research hotspot of protein function prediction, still has shortcomings at such aspects as reducing the impact of data noise, making full use of network topology characteristics, integrating multi-source data, and so on. In this paper, the Bi-Weighted Vote(BIWV) algorithm was proposed to predict protein functions, which combines the global topological similarity produced by Random Walk with Resistance (RWS) and the semantic similarity between terms. In addition, the Bi-Weighted Vote algorithm with pathway (BiWV-P) was presented by integrating the information of biological pathway. By using the data sets of saccharomyces cerevisiae and homo sapiens, experiments were performed to compare TMC, UBiRW, ProHG, BiWV and BiWV-P. The experimental results indicate that BiWV algorithm and BiWV-P algorithm can predict protein functions effectively, and achieve higher micro-accuracy and micro-F1 than other algorithms in many data sets.

Keywords Protein-protein interaction network, Function prediction, Random walk, Semantic similarity, Biological pathway

近年来,随着高通量生物技术的发展,研究者们通过酵母 双杂交、串联亲和纯化等实验获取了海量的蛋白质相互作用

到稿日期:2018-03-03 返修日期:2018-05-15 本文受国家自然科学基金项目(61762015,61502111,61662007,61763003),广西自然科学基金项目(2015GXNSFAA139288),"八桂学者"工程专项,广西科技基地和人才专项(AD16380008)资助。

唐家琪(1992—),男,硕士生,主要研究方向为生物信息学、机器学习;**吴璟莉**(1978—),女,博士,教授,CCF会员,主要研究方向为生物信息学、算法设计与分析,E-mail:wjlhappy@mailbox.gxnu.edu.cn(通信作者);**廖元秀**(1963—),女,硕士,副教授,主要研究方向为人工智能、形式化方法、机器人知识表示及推理;**王金艳**(1982—),女,博士,副教授,CCF会员,主要研究方向为数据安全、不确定性理论、自动推理。

(Protein-Protein Interaction，PPI)数据，并将其表示为 PPI 网络来处理蛋白质组学中的一系列问题。基于 PPI 网络的蛋白质功能预测问题正是当前蛋白质组学的一个研究热点，并且已取得一些研究成果。

早期的研究主要利用目标蛋白质在 PPI 网络中邻居[1-3]的功能信息、网络全局特性及模块特性[4-7]等提出功能预测方法，这些方法的共同缺陷在于忽视了蛋白质功能之间的关联性。针对该问题，近年来，研究者们提出利用功能术语之间的关联来提升蛋白质功能预测的效果。Zhang 等[8]利用杰卡德系数(Jaccard Coefficient)来衡量不同功能间的关联性，然后将其整合到一个监督学习框架来预测蛋白质的功能；Wang 等[9]采用一种基于功能关联性的多标签学习方法来预测蛋白质功能；Yu 等[10]利用 PPI 网络和功能关联性，既可预测完全未被注释的蛋白质的功能，还可预测未注释完整的蛋白质的缺失功能；Peng 等[11]提出了不平衡的双随机游走算法 UBiRW，迭代地在 PPI 网络和功能关联网络上分别执行随机游走来建立蛋白质与功能之间的新的映射关系；Yu 等[12]将 PPI 网络和功能关联网络合并，提出了直推式多标签分类器 TMC 来预测蛋白质的功能；在此基础上，Liu 等[13]提出在混合图上随机游走的蛋白质功能预测方法 ProHG，该方法利用功能相似性权重(Function Similarity Weight)对 PPI 网络进行重构，以减轻数据噪声的影响；Prasad 等[14]利用 PPI 网络中最重要的两个领域特性，即亲密度中心性和边聚类系数，设计了一种自底向上的层到层(Level to Level，L2L)的先验算法来预测蛋白质的功能。此外，研究者们还将基因表达[15]、氨基酸序列[16]、进化知识[17]、生物通路[18]、结构域信息[19]等数据整合到 PPI 网络中，以改善预测效果。

上述方法虽然取得了不错的效果，但在降低 PPI 网络中数据噪声的影响、利用更多的网络拓扑特性以及整合多源数据等方面仍然不够完善。为此，本文首先提出双加权投票(Bi-Weighted Voting，BiWV)方法，该方法考虑到 PPI 网络的全局拓扑特性，通过带阻力随机游走(Random Walk with Resistance，RWS)[20]构建蛋白质影响权重矩阵；考虑到 GO 术语之间的关联性，用 Wang 等[21]提出的 GO 术语语义相似度计算方法构建功能影响权重矩阵，并采取加权投票的方式预测蛋白质功能。然后，在 BiWV 方法的基础上，提出带生物通路信息的双加权投票算法 BiWV-P(Bi-Weighted Vote with Pathway)，该算法将生物通路信息整合到 PPI 网络中，提高了数据的可靠性，降低了数据噪声带来的影响。采用酿酒酵母(Saccharomyces Cerevisiae)和人类(Homo Sapiens)数据集进行实验测试，通过整合数据库 DIP[22]，KEGG[23]和 GO[24]中的 PPI 网络、生物通路和功能注释信息得到测试数据集，对算法 TMC[12]，UBiRW[11]，ProHG[13]，BiWV 和 BiWV-P 的预测效果进行对比分析。实验结果表明，算法 BiWV 和 BiWV-P 的排序损失(Ranking Loss)性能较差，但在大部分相对完整的数据集上能获得较其他算法更高的微正确率(Micro-accuracy)与微 F1(Micro-F1)，可作为蛋白质功能预测的一种参考方法。

1　问题与符号的定义

PPI 网络通常表示为无向图 $G(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 为顶点集，$v_i(i=1,2,\cdots,n)$ 表示蛋白质；$E=\{e_{ij}\mid e_{ij}=(v_i,v_j),v_i,v_j\in V\}$ 为边集，e_{ij} 为无向边，表示蛋白质 v_i 与 v_j 之间存在相互作用。令 $distance(v_i,v_j)$ 表示蛋白质 v_i 与 v_j 之间的路径长度。

本文利用基因本体(Gene Ontology，GO)[24]来注释蛋白质功能。GO 包括分子功能(Molecular Functions，MF)、生物过程(Biological Processes，BP)和细胞组件(Cellular Components，CC) 3 个独立的子本体，用有向无环图来表示子本体中术语间的语义关系。令 $G^o(V^o,E^o)$ 表示一个子本体中功能术语的语义网络，其中 $V^o=\{v_1^o,v_2^o,\cdots,v_m^o\}$ 为顶点集，$v_i^o(i=1,2,\cdots,m)$ 表示功能术语(term)；$E^o=\{e_{ij}^o\mid e_{ij}^o=\langle v_i^o,v_j^o\rangle,v_i^o,v_j^o\in V^o\}$ 为边集，e_{ij}^o 为有向边，表示术语 v_i^o 与 v_j^o 之间存在语义关系“is-a”或“part-of”。令 $F=\{f_1,f_2,\cdots,f_w\}\subseteq V^o$ 表示用于注释 PPI 网络中蛋白质的功能术语集，集合 $ancestor(f_i)(i=1,2,\cdots,w)$ 记录术语 f_i 的所有祖先术语，$offspring(f_i)(i=1,2,\cdots,w)$ 记录术语 f_i 的所有后代术语。

生物通路即生物体内完成代谢、膜转运、信号传导、细胞周期等生物学过程所需的一系列生化级联反应[25]，通常表示为化合物图(Compound Graph)和酶图(Enzyme Graph)两种有向简单图。本文采用酶图 $G^e(V^e,E^e)$ 来表示生物通路，其中 $V^e=\{v_1^e,v_2^e,\cdots,v_z^e\}$ 为顶点集，$v_i^e(i=1,2,\cdots,z)$ 表示酶，即参与酶促反应的蛋白质；$E^e=\{e_{ij}^e\mid e_{ij}^e=\langle v_i^e,v_j^e\rangle,v_i^e,v_j^e\in V^e\}$ 为边集，有向边 e_{ij}^e 表示酶 v_i^e 作用的产物是酶 v_j^e 作用的底物。

给定 PPI 网络 $G(V,E)$ 及功能术语集 F，定义功能注释矩阵 $\boldsymbol{Y}_{n\times w}$，矩阵元素 $y_{ij}(i=1,\cdots,n,j=1,\cdots,w)$ 取值如下：

$$y_{ij}=\begin{cases}1, & \text{若蛋白质 } v_i \text{ 被 } f_j \text{ 注释}\\ 0, & \text{否则}\end{cases}\tag{1}$$

蛋白质功能预测问题即给定功能注释矩阵 $\boldsymbol{Y}_{n\times w}$，通过分析 PPI 网络 G、生物通路 G^e 和语义网络 G^o，发现不同蛋白质之间的功能联系，从而利用 $\boldsymbol{Y}_{n\times w}$ 中已注释的蛋白质功能来预测其中未注释的蛋白质的功能。

2　双加权投票算法

本节首先提出双加权投票蛋白质功能预测算法 BiWV。在此基础上整合生物通路，提出带生物通路信息的双加权投票算法 BiWV-P。

2.1　BiWV 算法

BiWV 算法的输入为 PPI 网络 $G(V,E)$、语义网络 $G^o(V^o,E^o)$、功能术语集 F 和蛋白质功能注释矩阵 $\boldsymbol{Y}$，输出为蛋白质功能得分矩阵 $\boldsymbol{S}$。BiWV 首先构建蛋白质拓扑相似度矩阵和语义相似度矩阵；然后基于两个矩阵分别构建蛋白质影响权重矩阵和功能影响权重矩阵；最后采用加权投票方式预测蛋白质功能。

2.1.1　构建蛋白质拓扑相似度矩阵

利用 Lei 等[20]提出的 RWS 算法，通过 PPI 网络 $G(V,E)$ 构建蛋白质拓扑相似度矩阵 $\boldsymbol{U}_{n\times n}$。首先，对于给定的图 $G(V,E)$，为顶点 v_i 添加自环，即加入无向边 $e_{ii}=(v_i,v_i)(i=1,2,\cdots,n)$。令 $\boldsymbol{P}_{n\times n}$ 表示概率转移矩阵，其元素 $p_{ij}(i,j=1,2,\cdots,n)$ 表示每次游走时从顶点 v_i 到 v_j 的转移概率，如下所示：

$$p_{ij}=\begin{cases}(d_i+1)^{-1}, & \text{若}\exists e_{ij}\in E\\ 0, & \text{否则}\end{cases} \tag{2}$$

其中，d_i 表示 v_i 的度。假设随机游走的起点为 v_φ（$\varphi=1,2,\cdots,n$），令 $q^t_{\varphi,i}$ 为 t 时刻从 v_φ 游走到 v_i（$i=1,2,\cdots,n$）的标准化概率，即 $\sum_{i=1}^{n}q^t_{\varphi,i}=1$；$c^{t+1}_{\varphi,ij}$ 为 $t+1$ 时刻选中 v_i 到 v_j 间游走路径的概率，如式(3)所示：

$$c^{t+1}_{\varphi,ij}=\begin{cases}\max(0,q^t_{\varphi,i}P_{ij}-\theta_1), & q^t_{\varphi,j}>0\\ \max(0,q^t_{\varphi,i}P_{ij}-\theta_1), & \max\{q^t_{\varphi,k}P_{kj}\mid 1\leqslant k\leqslant n\}\geqslant\theta_2\\ 0, & \text{否则}\end{cases} \tag{3}$$

根据文献[20]，参数 $\theta_1=|V|/|E|^2$，$\theta_2=1/|E|$。$t+1$ 时刻，随机游走到 v_j（$j=1,2,\cdots,n$）的标准化概率 $q^{t+1}_{\varphi,j}$ 的计算方法如下所示：

$$q^{t+1}_{\varphi,j}=\frac{\sum_{i=1}^{n}c^{t+1}_{\varphi,ij}}{\sum_{k=1}^{n}\sum_{i=1}^{n}c^{t+1}_{\varphi,ik}} \tag{4}$$

当 $t\to\infty$ 时，$q^t_{\varphi,i}$ 收敛于某一常数 $q_{\varphi i}$，此时随机游走达到稳态。

然后，将顶点 $v_1,v_2,\cdots,v_n$ 分别作为起点进行上述随机游走，直到达到稳态，则可得到稳态概率矩阵 $\boldsymbol{Q}_{n\times n}$，其元素 $q_{ij}=q^t_{i,j}(t\to\infty,i,j=1,2,\cdots,n)$。接着，对矩阵 $\boldsymbol{Q}_{n\times n}$ 中的元素 q_{ij} 进行归一化处理，减去 q_{ij} 对应列的中位数，即 $q'_{ij}=q_{ij}-\text{median}(\boldsymbol{q}_{-j})(i,j=1,2,\cdots,n)$，得到结果矩阵 $\boldsymbol{Q}'_{n\times n}$。最后，计算 $\boldsymbol{Q}'_{n\times n}$ 中各列间的 Pearson 相关系数，并将其作为蛋白质在 PPI 网络中的拓扑相似度，从而得到拓扑相似度矩阵 $\boldsymbol{U}_{n\times n}$，该矩阵的元素 $u_{ij}=pearson(\boldsymbol{q}'_{-i},\boldsymbol{q}'_{-j})(i,j=1,2,\cdots,n)$。

2.1.2 构建语义相似度矩阵

Wang 等[21]基于 GO 层次结构信息，提出了一种语义相似度计算方法，本文称其为 GS-Wang。利用算法 GS-Wang，根据语义网络 G^o 和功能术语集 F 构建语义相似度矩阵 $\boldsymbol{R}_{w\times w}$，其中每个元素 $r_{ij}(i,j=1,2,\cdots,w)$ 表示 GO 术语 f_i 与 f_j 之间的语义相似度，如式(5)所示：

$$r_{ij}=\frac{\sum_{f_k\in T_i\cap T_j}(S_i(f_k)+S_j(f_k))}{\sum_{f_k\in T_i}S_i(f_k)+\sum_{f_k\in T_j}S_j(f_k)} \tag{5}$$

其中，$T_l=\{f_l,ancestor(f_l)\}(l=i,j)$ 表示术语 f_l 与其祖先术语组成的集合；$S_l(f_k)$ 表示术语 f_k 对 f_l 的语义贡献：

$$S_l(f_k)=\begin{cases}\max\{w_{k'k}\cdot S_l(f_{k'})\mid\langle f_{k'},f_k\rangle\in E^o\}, & \text{若 } f_k\neq f_l\\ 1, & \text{否则}\end{cases} \tag{6}$$

当 $\langle f_{k'},f_k\rangle$ 为“is-a”关系时，贡献因子 $w_{k'k}$ 取值为 0.8；当 $\langle f_{k'},f_k\rangle$ 为“part-of”关系时，取值为 0.6。

2.1.3 双加权投票预测

根据上述得到的蛋白质拓扑相似度矩阵 $\boldsymbol{U}_{n\times n}$ 和语义相似度矩阵 $\boldsymbol{R}_{w\times w}$，结合蛋白质功能注释矩阵 $\boldsymbol{Y}_{n\times w}$，得到蛋白质功能得分矩阵 $\boldsymbol{S}_{n\times w}$。

首先，由矩阵 $\boldsymbol{U}_{n\times n}$ 得到蛋白质影响权重矩阵 $\boldsymbol{U}'_{n\times n}$，这里仅考虑拓扑相似度 u_{ij} 为正值时蛋白质间的功能影响。矩阵 $\boldsymbol{U}'_{n\times n}$ 的元素 u'_{ij} 反映了蛋白质 v_i 与 v_j 之间的功能相似性。由于蛋白质倾向于与其局部邻居共享功能，因此 v_i 和 v_j 之间的功能影响与其之间的路径长度成反比，如式(7)所示：

$$u'_{ij}=\begin{cases}\left(\frac{u^\alpha_{ij}}{1+distance(v_i,v_j))^\beta}\right., & \text{若 } u_{ij}>0\\ 0, & \text{否则}\end{cases} \tag{7}$$

其中，参数 α 与 β 用于控制不同拓扑相似度与距离对目标蛋白质的影响。

然后，根据语义相似度矩阵 $\boldsymbol{R}_{w\times w}$，构建功能影响权重矩阵 $\boldsymbol{R}'_{w\times w}$。本文仅认为后代术语对祖先术语有较强的影响，即存在术语 f_i 与 f_j 且满足 $f_i\in offspring(f_j)$ 时，若术语 f_i 与 f_j 的语义相似度越大，则 f_i 对 f_j 的影响程度越大（$i,j=1,2,\cdots,w$）。令 $\eta_i=\sum_{k=1}^{n}y_{ki}(i=1,2,\cdots,w)$ 为术语 f_i 注释的蛋白质个数，$M=meadian(\eta_i)(i=1,2,\cdots,w)$，$meadian$ 表示中值函数，则功能 f_i 对 f_j 的影响力 λ_{ij} 如式(8)所示：

$$\lambda_{ij}=\begin{cases}r_{ij}^{\frac{\eta_j}{M}+\gamma}, & \text{若 } i=j \text{ 或 } f_i\in offspring(f_j)\\ 0, & \text{否则}\end{cases} \tag{8}$$

其中，η_i 越小，从 PPI 网络中找到与 f_i 有功能联系的蛋白质的难度就越大，故越需要依赖 GO 术语的语义网络来进行预测；参数 γ 用于控制不同 GO 术语的语义相似度的影响。对 λ_{ij} 进行归一化处理，得到影响权重矩阵 $\boldsymbol{R}'_{w\times w}$ 的元素 $r'_{ij}(i,j=1,2,\cdots,w)$，如式(9)所示：

$$r'_{ij}=\frac{\lambda_{ij}}{\sum_{k=1}^{w}\lambda_{kj}} \tag{9}$$

最后，将矩阵 $\boldsymbol{Y}_{n\times w}$ 分别以左乘矩阵 $\boldsymbol{U}'$、右乘矩阵 $\boldsymbol{R}'$ 的方式进行加权投票，即 $\boldsymbol{U}'\boldsymbol{Y}\boldsymbol{R}'$，从而得到功能得分矩阵 $\boldsymbol{S}$。下面给出算法 BiWV 的形式化描述。

算法 1 BiWV 算法

输入：PPI 网络 G，语义网络 G^o，功能术语集 F，功能注释矩阵 **Y**，参数 α,β,γ

输出：蛋白质功能得分矩阵 **S**

1. $\mathbf{U}_{n\times n}=\text{RWS(G)}$//构建拓扑相似度矩阵
2. $\mathbf{R}_{w\times w}=\text{GS-Wang}(G^o,\text{F})$//构建语义相似度矩阵
3. for each $u_{ij}(i,j=1,2,\cdots,n)$
4. if $u_{ij}>0$ then
5. $u'_{ij}=\frac{u^\alpha_{ij}}{(1+\text{distance}(v_i,v_j))^\beta}$
6. else
7. $u'_{ij}=0$
8. for $i=1,2,\cdots,w$
9. $\eta_i=\sum_{k=1}^{n}y_{ki}$
10. $M=\text{meadian}(\eta_i)(i=1,2,\cdots,w)$
11. for each $r_{ij}(i,j=1,2,\cdots,w)$
12. if $i=j\parallel f_i\in\text{offspring}(f_j)$ then
13. $\lambda_{ij}=r_{ij}^{\frac{\eta_j}{M}+\gamma}$
14. else
15. $\lambda_{ij}=0$
16. for each $\lambda_{ij}(i,j=1,2,\cdots,w)$

17. $r'_{ij}=\frac{\lambda_{ij}}{\sum_{k=1}^{w}\lambda_{kj}}$

18. $\mathbf{S}=\mathbf{U}'\mathbf{Y}\mathbf{R}'$

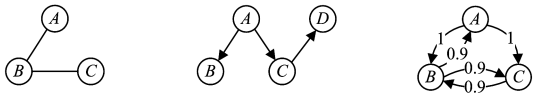
2.2 BiWV-P 算法

由于原始 PPI 网络通常带有一定的数据噪声，且其中一些相互作用在本质上具有方向性[18]，因此可通过整合生物通路构建有向加权的 PPI 网络，从而在考虑蛋白质相互作用的方向性的同时降低数据噪声的影响。基于该研究思路，本文提出带生物通路信息的双加权投票算法 BiWV-P，算法输入为 PPI 网络 $G(V,E)$、生物通路 $G^e(V^e,E^e)$、语义网络 $G^o(V^o,E^o)$、功能术语集 F 和蛋白质功能注释矩阵 $\boldsymbol{Y}$，输出为蛋白质功能得分矩阵 $\boldsymbol{S}$。下面详细介绍 BiWV-P 算法的思想。

首先，构建有向带权网络 $G^d(V^d,E^d,W^d)$，其中，$V^d=V$，$E^d=\{e_{ij}^d\mid e_{ij}^d=\langle v_i^d,v_j^d\rangle,v_i^d,v_j^d\in V^d;v_i^d=v_i,v_j^d=v_j,(v_i,v_j)\in E$ or $v_i^d=v_i^e,v_j^d=v_j^e,\langle v_i^e,v_j^e\rangle\in E^e\}$，有向边 e_{ij}^d 表示蛋白质 v_i^d 对 v_j^d 有作用，$W^d=\{w_{ij}^d\mid w_{ij}^d=weight(e_{ij}^d)\}$，元素 w_{ij}^d 表示 v_i^d 对 v_j^d 作用的可信度，定义如下：

$$weight(e_{ij}^d)=\begin{cases}1, & \text{若 } v_i^d=v_i^e,v_j^d=v_j^e,\langle v_i^e,v_j^e\rangle\in E^e\\ 1-\varepsilon, & \text{否则}\end{cases}\tag{10}$$

参数 ε 控制相互作用方向带来的影响，其值越大表示方向性越强。图 1 给出了有向带权网络 G^d 的构建实例。图 1(a)为包含 3 个节点和 2 条无向边的原始 PPI 网络；图 1(b)为这 3 个顶点和顶点 D 所在的生物通路，包含了 3 条有向边；ε 设置为 0.1；则得到的有向带权的 PPI 网络如图 1(c)所示。



(a)原始 PPI 网络　(b)生物通路　(c)有向带权的 PPI 网络

图 1　有向带权网络的构建实例

Fig. 1　Example of constructing a direct weighted network

算法 BiWV-P 与 BiWV 的主要区别在于前者使用了不同的概率转移矩阵来计算蛋白质间的拓扑相似度。给定有向带权网络 G^d，BiWV-P 利用 RWS 算法在 G^d 上随机游走以获取蛋白质的拓扑相似度，此时概率转移矩阵为 $\boldsymbol{P}_{n\times n}^d$，其元素 p_{ij}^d $(i,j=1,2,\cdots,n)$的计算如下：

$$p_{ij}^d=\begin{cases}(1+\sum_{i=1}^{n}weight(e_{ij}^d))^{-1}, & \text{若 }\exists\, e_{ij}^d\in E^d\\ 0, & \text{否则}\end{cases}\tag{11}$$

算法 BiWV-P 的其余步骤与 BiWV 相同，不再重复介绍。

3　实验与结果分析

3.1　实验数据

本文的实验数据来自于数据库 DIP[22]，KEGG[23] 和 GO[24]。DIP 数据库记录酿酒酵母和人类的 PPI 网络；KEGG 数据库记录酿酒酵母和人类已知的生物通路；GO 数据库记录功能术语及其语义关系。实验时，首先从 DIP 数据库下载 PPI 网络(2017 年高可信版本)，用 UniProtKB/Swiss-Prot[26] 编号对其中的蛋白质进行 ID 转换；然后去除网络中自相互作用、重复相互作用及无法转换的蛋白质；最后利用 R 包 biomaRt[27] 为每个蛋白质的 UniProtKB/Swiss-Prot 编号，获取对应的 GO 术语编号。

为评估功能预测方法的有效性，需要对原始数据进行预处理：1)剔除获取手段为 IEA (Inferred from Electronic Annotation)，NR(Not Recorded)，ND(No biological Data Available)和 IC (Inferred by Curator)的功能注释，以保证功能注释术语的可靠性；2)剔除注释少于 5 个蛋白质的 GO 术语，以避免预测的 GO 术语过于具体；3)分别用 GO 的 3 个子本体 MF，BP 和 CC 上的术语注释 PPI 网络，并从网络中删除未被注释的蛋白质，从而对酿酒酵母和人类分别构建出 3 个不同的 PPI 网络，本文称为 MF，BP 和 CC 网络；4)取 PPI 网络中的极大连通子图作为测试数据，以确保每个蛋白质均有与其相互作用的蛋白质。最终得到的原始 PPI 网络数据集如表 1 所列。

表 1　酿酒酵母与人类数据集

Table 1　Data sets of saccharomyces cerevisiae and homo sapiens

GO	酿酒酵母			人类		
	$\|V\|$	$\|E\|$	w	$\|V\|$	$\|E\|$	w
MF	749	1140	108	2229	3919	402
BP	1297	2798	368	2190	3853	1325
CC	1277	2738	167	2253	3944	344

实验中，利用 R 包 KEGGRESET[28] 提供的 KEGG 数据库接口下载 116 个酿酒酵母生物通路和 321 个人类生物通路数据，分别将这些生物通路合并，得到针对酿酒酵母或人类的酶图 $G^e(V^e,E^e)$，进而与 PPI 网络 G 整合，得到有向带权的 PPI 网络 $G^d(V^d,E^d,W^d)$，其有向边的数目如表 2 所列。

表 2　酿酒酵母与人类数据集中有向边 E^d 的数目

Table 2　Number of directed edges E^d of saccharomyces cerevisiae and homo sapiens data sets

GO	酿酒酵母	人类
MF	3053	14652
BP	6169	14485
CC	6025	14764

3.2　评价指标

蛋白质功能预测可以看成一个多标签分类问题，每个蛋白质为样本，每个功能为标签。本文采用基于类别的评价指标即微正确率、微 F1 和基于排序的评价指标排序损失来综合评估多标签分类的结果[29-30]。

将求解的功能得分矩阵 $\boldsymbol{S}$ 转化为二值矩阵 $\boldsymbol{Y}_{n\times w}^b$，即从行 $\boldsymbol{s}_{i-}$ $(i=1,2,\cdots,n)$中选择功能得分最高的 K 个术语作为蛋白质 v_i 的功能，本文 K 取各蛋白质功能数的均值。统计 $\boldsymbol{Y}_{n\times w}^b$ 中每列 $y^b{}_j$ $(j=1,2,\cdots,w)$的真正例(True Positive，TP)、假正例 (False Positive，FP)、真反例 (True Negative，TN)、假反例(False Negative，FN)，分别记 TP_j,FP_j,TN_j,FN_j，得到 w 列的均值 $\overline{TP},\overline{FP},\overline{TN},\overline{FN}$。由此计算微正确率与微 F1：

$$\text{微正确率}=\frac{\overline{TP}+\overline{TN}}{\overline{TP}+\overline{TN}+\overline{FP}+\overline{FN}}\tag{12}$$

$$\text{微 F1}=\frac{2\times\text{微查准率}\times\text{微查全率}}{\text{微查准率}+\text{微查全率}}\tag{13}$$

其中，微查准率(Micro-precision)和微查全率(Micro-recall)的定义如式(14)－式(16)所示：

$$\text{微查准率}=\frac{\overline{TP}}{\overline{TP}+\overline{FP}} \tag{14}$$

$$\text{微查全率}=\frac{\overline{TP}}{\overline{TP}+\overline{FN}} \tag{15}$$

排序损失定义为无关标签的得分排在相关标签的得分之前的概率均值，该值越小表示分类效果越好。其定义如下：

$$\text{排序损失}=\frac{1}{n}\sum_{i=1}^{n}\frac{|(f',f'')|h(v_i,f')\leqslant h(v_i,f'')|}{|F_i||\overline{F_i}|} \tag{16}$$

其中，F_i 表示蛋白质 v_i 的功能标签集合，$\overline{F_i}$ 为 F_i 的补集，$f'\in F_i$，$f''\in\overline{F_i}$，$(f',f'')\in F_i\times\overline{F_i}$，函数 $h(v_i,f)$ 表示蛋白质 v_i 在标签 f 上的功能得分。

3.3 性能评价

采用 10 折交叉验证来测试算法 BiWV 及 BiWV-P 的预测效果。本文将 TMC，UBiRW，ProHG 3 种近年提出的随机游走预测算法和本文提出的 BiWV 及 BiWV-P 进行比较。本文算法的参数设置如下：$\alpha=5$，$\beta=8$，$\gamma=5$，$\varepsilon=0.1$；对比算法的参数设置与原文献相同。表 3－表 5 分别给出了这 5 种预测算法在酿酒酵母和数据上的实验结果，粗体表示在数据集上的最佳评价指标值。

表 3 不同方法的微正确率比较

Table 3 Micro-accuracy comparison of different methods

方法	酿酒酵母			人类		
	MF	BP	CC	MF	BP	CC
TMC	0.9752	0.9887	0.9841	0.9847	0.9897	0.9805
UBiRW	0.9759	0.9891	0.9854	0.9616	0.9786	0.9740
ProHG	**0.9769**	0.9889	0.9839	0.9865	0.9903	0.9827
BiWV	0.9758	0.9890	0.9855	0.9868	0.9902	0.9832
BiWV-P	0.9764	**0.9893**	**0.9859**	**0.9872**	**0.9905**	**0.9836**

表 3 基于微正确率指标对算法进行比较，从表中数据可以看出，算法 BiWV 与 ProHG 和 UBiRW 的微准确率接近，并高于 TMC。BiWV-P 在大多数数据集上的微准确率高于其他算法，而在酿酒酵母 MF 网络上低于 ProHG。

表 4 不同方法的微 F1 比较

Table 4 Micro-F1 comparison of different methods

方法	酿酒酵母			人类		
	MF	BP	CC	MF	BP	CC
TMC	0.2473	0.3210	0.3756	0.2543	0.1686	0.1356
UBiRW	0.2669	0.3440	0.4303	0.2553	0.2327	0.4031
ProHG	**0.2997**	0.3344	0.3679	0.3426	0.2168	0.4125
BiWV	0.2713	0.3384	0.4311	0.3591	0.2064	0.4272
BiWV-P	0.2848	**0.3548**	**0.4487**	**0.3767**	**0.2328**	**0.4427**

表 4 给出了各算法的微 F1。可以看出，算法 BiWV 和 BiWV-P 在大多数数据集上的微 F1 高于其他方法，ProHG 依然在酿酒酵母 MF 网络上具有优势。

表 5 不同方法的排序损失比较

Table 5 Ranking loss comparison of different methods

方法	酿酒酵母			人类		
	MF	BP	CC	MF	BP	CC
TMC	0.1487	0.1091	0.0544	0.0906	0.1371	0.0550
UBiRW	0.1802	0.1324	0.0715	0.1803	0.1309	0.0704
ProHG	**0.1292**	**0.0997**	**0.0528**	**0.0821**	**0.1283**	**0.0504**
BiWV	0.1540	0.1202	0.0600	0.0998	0.1678	0.0660
BiWV-P	0.1514	0.1139	0.0550	0.0909	0.1469	0.0587

表 5 中对不同方法的排序损失进行了比较。从表中数据可以看出，算法 BiWV 和 BiWV-P 在大多数数据集上能获得较算法 UBiRW 更小的排序损失，但相对于其他算法没有优势。

由上述实验结果可知，TMC 算法的微正确率与微 F1 指标均劣于本文算法，但排序损失优于本文算法；UBiRW 算法的微正确率和微 F1 与 BiWV 算法相近，但劣于 BiWV-P 算法，且其排序损失均劣于本文算法；ProHG 算法除在酿酒酵母 MF 网络上取得了较本文算法更高的微准确率和微 F1 之外，在其他数据集上的微准确率和微 F1 均劣于本文算法；此外，ProHG 算法在所有数据集上的排序损失均优于其他方法。

由式(11)－式(14)可知，微正确率主要取决于$\overline{TP}$和$\overline{TN}$，微 F1 值主要取决于$\overline{TP}$。由于 $\boldsymbol{Y}_{n\times w}$ 和 $\boldsymbol{Y}^{b}_{n\times w}$ 通常为稀疏矩阵(即正例远少于反例)，蛋白质功能数的差异对$\overline{TP}$的影响大于对$\overline{TN}$的影响。而功能较多(即正例较多)的蛋白质对$\overline{TP}$的影响较大，因此微正确率和微 F1 结果的优劣取决于具有较多功能的蛋白质的预测情况。在 PPI 网络中，通常，功能较多的蛋白质的顶点度(degree)较高[31]，对于顶点度较低的蛋白质而言，其更容易利用网络拓扑结构来预测功能。本文算法正是利用了这个特点，借助网络拓扑结构来提高具有较高顶点度的蛋白质的预测精度，从而获得较其他算法更高的微正确率和微 F1。排序损失表示基于功能得分矩阵 $\boldsymbol{S}$，每个蛋白质无关标签的得分排在相关标签的得分之前的概率均值，因此不同功能数的蛋白质的预测精度对该指标的影响权重相同。如上所述，本文算法对具有较高顶点度的蛋白质的预测情况较好，而对于顶点度较低的蛋白质的预测效果劣于算法 TMC 和 ProHG，因此排序损失指标较大。由上述分析可知，本文提出的算法 BiWV 和 BiWV-P 对 PPI 网络中顶点度较高(即功能数较多)的蛋白质具有较好的预测效果。同时，由于其对网络拓扑具有较大的依赖性，网络数据的缺失将对其预测效果具有一定的负面影响，因此在缺失较多顶点和边的酿酒酵母 MF 网络上其预测效果较差。

综上所述，算法 BiWV 和 BiWV-P 能够有效预测蛋白质的功能，虽然排序损失较大，但在大部分相对完整的 PPI 网络数据集上能够获得较同类方法更高的准确率与微 F1。同时，BiWV-P 在所有指标和数据集上均优于 BiWV，因此整合生物通路对于提升蛋白质功能预测性能是有效的。

结束语 基于 PPI 网络的蛋白质功能预测是近年来生物信息学的研究热点之一。本文提出了一种基于双加权投票的蛋白质功能预测算法 BiWV，其分别从蛋白质和功能术语角度加权投票来进行预测；并在此基础上提出了 BiWV-P 算法，该方法基于生物通路构建有向加权 PPI 网络以提升性能。为评估预测效果，整合数据库 DIP，KEGG 和 GO 中的 PPI 网络、生物通路和功能注释信息，得到酿酒酵母和人类的测试数据集。实验结果表明，BiWV 和 BiWV-P 算法能有效预测蛋白质的功能，并且在大部分相对完整的数据集上的微正确率与微 F1 高于 TMC，UBiRW，ProHG 这 3 种同类方法的性能，但排序损失较大。基于 PPI 网络的蛋白质功能预测的未来发展可以从以下几个方面入手：1)构建动态的 PPI 网络对蛋白

质相互作用进行建模；2）提高 PPI 网络的数据质量，降低数据噪声对预测结果的干扰；3）研究合理利用 GO 术语关联性的方法，从术语的语义网络中挖掘出更多的信息用于蛋白质功能预测。

参 考 文 献

[1] SCHWIKOWSKI B，UETZ P，FIELDS S. A network of protein-protein interactions in yeast[J]. Nature Biotechnology，2000，18(12)：1257-1261.

[2] HISHIGAKI H，NAKAI K，ONO T，et al. Assessment of prediction accuracy of protein function from protein-protein interaction data[J]. Yeast，2001，18(6)：523-531.

[3] CHUA H N，SUNG W K，WONG L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions[J]. Bioinformatics，2006，22(13)：1623-1630.

[4] CHRISTINE B，FRANÇOIS C，DAVID M，et al. Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network[J]. Genome Biology，2003，5(1)：6-18.

[5] NABIEVA E，JIM K，AGARWAL A，et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps[J]. Bioinformatics，2005，21(1)：302-310.

[6] DENG M，TU Z，SUN F，et al. Mapping Gene Ontology to proteins based on protein-protein interaction data.[J]. Bioinformatics，2004，20(6)：895-902.

[7] VAZQUEZ A，FLAMMINI A，MARITAN A，et al. Global protein function prediction from protein-protein interaction，networks[J]. Nature Biotechnology，2003，21(6)：697-700.

[8] ZHANG X F，DAI D Q. A Framework for Incorporating Functional Interrelationships into Protein Function Prediction Algorithms[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics，2012，9(3)：740-753.

[9] WANG H，HUANG H，DING C. Function-Function Correlated Multi-Label Protein Function Prediction over Interaction Networks[C]// International Conference on Research in Computational Molecular Biology. Berlin：Springer，2012：302-313.

[10] YU G，ZHU H，DOMENICONI C. Predicting protein functions using incomplete hierarchical labels[J]. BMC Bioinformatics，2015，16(1)：1-12.

[11] 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 & Peptide Science，2014，15(6)：529-539.

[12] YU G，RANGWALA H，DOMENICONI C，et al. Protein Function Prediction using Multi-label Ensemble Classification[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics，2013，10(4)：1045-1057.

[13] LIU J，WANG J，YU G. Protein Function Prediction by Random Walks on a Hybrid Graph[J]. Current Proteomics，2016，13(2)：130-142.

[14] PRASAD A，SAHA S，CHATTERJEE P，et al. Protein Function Prediction from Protein Interaction Network Using Bottom-up L2L Apriori Algorithm[C]// International Conference on Computational Intelligence，Communications，and Business Analytics. Singapore：Springer，2017：3-16.

[15] LICHTENBERG U D，JENSEN L J，BRUNAK S，et al. Dynamic Complex Formation During the Yeast Cell Cycle[J]. Science，2005，307(5710)：724-727.

[16] XIONG W，LIU H，GUAN J，et al. Protein function prediction by collective classification with explicit and implicit edges in protein-protein interaction networks[J]. BMC Bioinformatics，2013，14(Suppl 12)：4-16.

[17] COZZETTO D，BUCHAN D W，BRYSON K，et al. Protein function prediction by massive integration of evolutionary analyses and multiple data sources[J]. BMC Bioinformatics，2013，14(Suppl 3)：1-11.

[18] CAO M，PIETRAS C M，FENG X，et al. New directions for diffusion-based network prediction of protein function：incorporating pathways with confidence[J]. Bioinformatics，2014，30(12)：219-227.

[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] LEI C，RUAN J. A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity[J]. Bioinformatics，2013，29(3)：355-364.

[21] WANG J Z，DU Z，PAYATTAKOOL R，et al. A new method to measure the semantic similarity of GO terms[J]. Bioinformatics，2007，23(10)：1274-1281.

[22] XENARIOS I，RICE D W，SALWINSKI L，et al. DIP：the database of interacting proteins.[J]. Nucleic Acids Research，2000，32(1)：289-291.

[23] OGATA H，GOTO S，SATO K，et al. KEGG：Kyoto Encyclopedia of Genes and Genomes.[J]. Nucleic Acids Research，2000，27(1)：29-34.

[24] ASHBURNER M，BALL C J，BOTSTEIN D，et al. Gene ontology：tool for the unification of biology. The Gene Ontology Consortium[J]. Nature Genetics，2000，25(1)：25-29.

[25] CARY M P，BADER G D，SANDER C. Pathway information for systems biology[J]. FEBS Letters，2005，579(8)：1815-1820.

[26] CONSORTIUM U P. The Universal Protein Resource (UniProt) in 2010[J]. Nucleic Acids Research，2010，38(Database issue)：142-148.

[27] BIRNEY E. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt[J]. Nature Protocols，2009，4(8)：1184-1191.

[28] TENENBAUM D. Client-side REST access to KEGG[EB/OL]. http：//rpackages. ianhowson. com/bioc/KEGGREST.

[29] ZHANG M L，ZHOU Z H. A Review on Multi-Label Learning Algorithms[J]. IEEE Transactions on Knowledge & Data Engineering，2014，26(8)：1819-1837.

[30] 周志华. 机器学习[M]. 北京：清华大学出版社，2016：23-33.

[31] GILLIS J，PAVLIDIS P. The Impact of Multifunctional Genes on "Guilt by Association" Analysis[J/OL]. http：//www. oalib. com/paper/134869@. W-vO7ywYxAs.