基于关键功能模块挖掘的蛋白质功能预测

赵碧海[1] 李学勇[1] 胡 赛[1] 张 帆[1] 田清龙[1] 杨品红[2] 刘 臻[3]

摘 要 精确注释蛋白质功能是从分子水平理解生物体的关键. 由于内在的困难和昂贵的开销, 实验方法注释蛋白质功能已经很难满足日益增长的序列数据. 为此, 提出了许多基于蛋白质相互作用 (Protein-protein interaction, PPI) 网络的计算方法预测蛋白质功能. 当今蛋白质功能预测的趋势是融合蛋白质相互作用网络和异构生物数据. 本文提出一种基于多关系网络中关键功能模块挖掘的蛋白质功能预测算法. 关键功能模块由一组紧密联系且共享生物功能的蛋白质组成, 它们能与网络中的剩余部分较好地区分开来. 算法通过从多关系网络的每一个简单网络中挖掘高内聚、低耦合的子图形成关键功能模块. 关键功能模块中邻居蛋白质的功能用于注释待预测功能的蛋白质. 每一个简单网络在蛋白质功能预测中的重要性各不相同. 实验结果表明, 提出的方法性能优于现有的蛋白质功能预测方法.

关键词 功能预测, 多关系网络, 蛋白质相互作用, 关键功能模块

引用格式 赵碧海, 李学勇, 胡赛, 张帆, 田清龙, 杨品红, 刘臻. 基于关键功能模块挖掘的蛋白质功能预测. 自动化学报, 2018, **44**(1): 183−192

DOI 10.16383/j.aas.2018.c160592

Prediction of Protein Functions Based on Essential Functional Modules Mining

ZHAO Bi-Hai[1] LI Xue-Yong[1] HU Sai[1] ZHANG Fan[1] TIAN Qing-Long[1] YANG Pin-Hong[2] LIU Zhen[3]

Abstract The accurate annotation of protein functions is a key to understanding living organisms at the molecular level. With its inherent difficulty and expense, experimental characterization of protein functions cannot scale up to accommodate the vast amount of sequence data. As a result, many computational methods based on protein-protein interaction (PPI) networks have been proposed to predict the functions of proteins. Nowadays, the trend in protein functions prediction is to integrate PPI networks and heterogeneous biological data. A novel protein functions prediction algorithm was proposed based on mining essential functional modules from a multi-relational network. An essential functional module is a group of densely connected proteins with shared biological function and can be well-separated from the rest of the network. The proposed algorithm identified subgraph with high cohesion and low coupling on each single network derived from the multi-relational network to form essential functional modules. Functions of neighbor proteins within essential functional modules were used to annotate the testing protein. Each single network has different importance on the prediction of protein functions. Experiment results show that our method outperforms other protein functions prediction methods.

Key words Function prediction, multiple network, protein-protein interaction (PPI), essential functional module

Citation Zhao Bi-Hai, Li Xue-Yong, Hu Sai, Zhang Fan, Tian Qing-Long, Yang Pin-Hong, Liu Zhen. Prediction of protein functions based on essential functional modules mining. *Acta Automatica Sinica*, 2018, **44**(1): 183−192

收稿日期 2016-09-02 录用日期 2017-01-16
Manuscript received September 2, 2016; accepted January 16, 2017
国家自然科学基金 (61772089), 湖南省自然科学基金 (2016JJ3016), 湖南省教育厅项目 (16A020, 16C0137, 17C0133), 水产高效健康生产湖南省协同创新中心资助
Supported by National Natural Science Foundation of China (61772089), Natural Science Foundation of Hunan Province (2016JJ3016), National Scientific Research Foundation of Hunan Province (16A020, 16C0137, 17C0133), and Collaborative Innovation Center for Efficient and Health Production of Fisheries in Hunan Province
本文责任编委 张学工
Recommended by Associate Editor ZHANG Xue-Gong
1. 长沙学院计算机工程与应用数学学院 长沙 410022 2. 湖南文理学院生命科学学院 常德 415000 3. 长沙学院生物与环境工程学院 长沙 410022
1. School of Computer Engineering and Applied Mathematics, Changsha University, Changsha 410022 2. School of Life Science, Hunan University of Arts and Science, Changde 415000 3. School of Biology and Environmental Engineering, Changsha University, Changsha 410022

由于蛋白质在不同生物过程中扮演重要角色, 注释功能未知的蛋白质是后基因时代的重要任务之一. 生物实验确定蛋白质功能存在耗时多和费用高的问题[1]. 因此, 基于计算的功能预测成为非常重要的替代方法. 然而, 这种方法需要准确而可靠的自动功能预测器. 现有的基于计算的功能预测方法都是建立在数据库中已经注释的蛋白质的功能之上. 虽然相互作用数据、序列数据和蛋白质结构数据等都已用于蛋白质功能预测算法, 但是设计一种有效的方法充分利用各种不同的生物信息依然是一个巨大的挑战, 源于这些生物数据的异构性、复杂性和多样性. 根据整合这些不同数据源的方式不同, 这些基于计算的预测方法可以分为四类: 基于多特征向量方法、基于多分类器方法、基于核的方法和基于网络的方法. 网络是一种很好的描述蛋白质之间关系的

途径, 而且大量基于网络的方法为我们提供了有效的工具从网络中挖掘信息, 这也有助于我们理解细胞生命活性物的复杂机制.

大部分基于网络的蛋白质功能预测方法都是从蛋白质相互作用 (Protein-protein interaction, PPI) 网络提取信息. 这些方法都建立在一个发现的基础上: 大约 70 %～80 % 的蛋白质与它们在 PPI 网络的相互作用伙伴至少共享一项功能[2]. 一些方法通过 PPI 网络中的直接或间接邻居节点预测未注释的蛋白质功能. 上述的这些方法独立地为每一个蛋白质预测功能. 还有一些方法将 PPI 网络中的蛋白质分成多个功能模块, 并为相同的模块注释相同的功能[3]. 这类方法聚类形成模块或复合物的方式存在差异. 由于相互作用数据中存在假阳性和假阴性, 一些研究者结合相互作用网络和异构生物数据, 提高功能预测的准确率, 例如基因表达数据[4]、同源数据[5]、蛋白质复合物数据[6]、结构域数据[7] 等.

另一种流行的基于网络并利用生物信息资源的方法是基于 GO term 的功能相似性建立功能关联网络. 蛋白质功能描述为结构化的标准词汇, 并存储在基因本体数据库. GO term 之间的父亲–孩子关系可以表达为有向无环图. 考虑到两个相似的功能共同注释一个共同的蛋白质以及两个相互作用的蛋白质倾向于共享同一功能, 一些研究者结合 PPI 网络和功能相似性, 从而提高功能预测的准确率. 由于 PPI 网络的不完整性, 其他的异构数据也被整合进来. Peng 等[8] 结合 PPI 网络和 Domain 信息, 利用蛋白质的功能相似性, 提出名为 DCS 的蛋白质功能预测方法. 进一步, 加入蛋白质复合物信息, 提出了改进的 DSCP 方法. 大部分整合异构数据的方法基本采取如下思路: 1) 生成各种功能相关网络 (每一个数据源对应一个或多个网络); 2) 这些单独的网络通过加权汇总的方式形成一个复合网络. 这些方法的区别在于单个网络形成复合网络时, 不同方法权值比例和优化方式存在差异.

综上所述, 整合多元生物数据能够有效弥补相互作用网络不完整性和噪声的问题, 提高基于网络的蛋白质功能预测方法的准确率. 但是, 引入其他生物信息后, 使得蛋白质之间的联系更加复杂, 更加多元化. 现有的方法基本都采取合并多种类型的相互作用的处理方式, 这虽然能够一定程度增加正确匹配的功能数量, 但也会同时引入更多的噪声功能, 最终使得整体预测性能提升不大. 上述提及的某些方法先构建多种功能关联网络, 然后再采取加权汇总的方式将多个单独网络构成一个复合网络. 不同网络在加权汇总时的比重各不相同, 而每个网络的比重参数成为影响功能预测方法的重要因素. 参数的设置一般会根据经验值设置. 即便是通过优化的方式获取, 也存在不同数据集有不同设置的问题. 从这些问题出发, 本文在原有研究基础之上, 结合 PPI 网络、蛋白质复合物数据和蛋白质结构域数据建立多关系网络. 考虑到蛋白质功能与模块之间的紧密联系, 提出一种基于多关系网络中关键功能模块挖掘的蛋白质功能预测方法 (Prediction of functions based on essential functional modules mining from a multi-relational network, PEFM). 蛋白质的功能不是由单个蛋白质独立完成, 而是与其他蛋白质相互作用共同执行机体功能, 蛋白质功能与功能模块之间存在紧密联系. 关键功能模块是指相互间紧密联系的蛋白质组成的功能模块或复合物. 移除关键功能模块会使得生物体丧失许多重要分子功能. 因此, 通过挖掘关键功能模块有助于提高蛋白质功能预测算法的准确率. PEFM 方法依次遍历多关系网络分解得到的每一个简单网络, 挖掘高内聚、低耦合的稠密子图形成不同网络层次的关键功能模块集合. 模块中节点的全部功能用于注释测试蛋白质. 多个数据集的实验结果验证了 PEFM 算法的有效性.

1 PEFM 算法

细胞功能不是由单个蛋白质完成, 而是通过多个紧密联系的蛋白质构成模块, 共同执行. 蛋白质功能与模块之间存在紧密联系, 模块划分为蛋白质功能预测提供了途径. 本文通过聚类, 形成高内聚、低耦合的功能模块, 进而实现蛋白质功能预测.

1.1 多关系网络构建

受实验条件限制, 高通量方法获得的蛋白质相互作用数据具有不完整性, 限制了蛋白质功能预测算法的性能. 结合多元的生物信息和蛋白质相互作用网络, 降低相互作用数据的实验错误带来的负面影响, 是当今基于相互作用网络的功能预测算法的发展趋势. 多元异构数据包括基于时间序列的基因表达信息、蛋白质结构域信息、复合物信息、亚细胞定位信息等. 在原有研究基础之上, 本文结合蛋白质相互作用网络的拓扑特性、蛋白质结构域信息和蛋白质复合物信息构建适合功能预测的多关系网络[7]. 相比之前构建的研究基础, 本文在建立多关系网络时, 增加了蛋白质复合物数据. 由于实验方法获得的蛋白质相互作用数据和结构域数据存在假阴性, 存在某些蛋白质执行共同功能, 却没有在前期构建的网络中体现的情况. 通过融入复合物数据, 能够为更多的蛋白质预测功能.

蛋白质结构域是分子的一个特别区域, 具有独立的功能. 有的蛋白质仅仅包含一个结构域, 有的蛋

白质可能包含多个不同类型的结构域. 一个结构域也可能出现在多个不同的蛋白质当中. 蛋白质的新功能常常利用结构域重组完成. 蛋白质执行生物功能离不开结构域, 由此可见, 蛋白质功能与结构域之间存在紧密的联系. 学者们开始尝试利用结构域信息, 提高功能预测算法的准确率.

本文首先针对蛋白质结构域信息与蛋白质功能之间的关联开展统计分析. 本次实验选定的蛋白质相互作用网络包含 5 093 个蛋白质, 其中具有功能注释的蛋白质数量是 2 894, 至少包含 1 个结构域的蛋白质数量是 3 056, 既有功能注释又有包含结构域的蛋白质数量为 1 887 个, 如图 1 所示. 从图 1 不难看出, 具有结构域的蛋白质中, 61.75 % 的蛋白质至少具有 1 项功能; 2 894 个被注释的蛋白质中, 有 65.2 % 的蛋白质包含结构域.

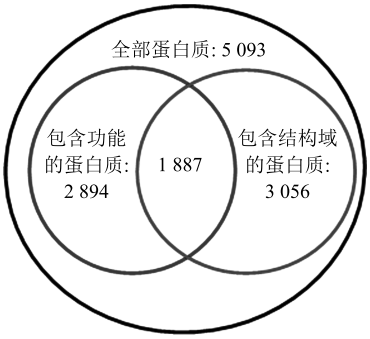

图 1　结构域与蛋白质功能关系综合统计

Fig. 1　Statistics of relationship between domains and protein functions

进一步地, 本文统计分析 2 894 个功能已知的蛋白质之间共享功能和共享结构域的情况, 其中 42 % 的蛋白质与其他蛋白质共享功能的同时还共享相同的结构域. 表 1 详细列出了蛋白质功能数量分布与共享结构域之间的关系.

表 1　蛋白质功能数量统计

Table 1　Quantity statistics of protein functions

功能数量	蛋白质数量	共享结构域的蛋白质数量	所占比例 (%)
1	1 512	523	34.59
2	703	318	45.23
3	317	166	52.37
4	140	77	55.00
5	106	54	50.94
> 5	116	77	66.38

从表 1 不难看出, 1 512 个蛋白质仅有 1 项功能, 其中 34.59 % 的蛋白质与其他蛋白质共享功能的同时还共享结构域. 而当功能数量增多时, 共享结构域的蛋白质比例明显增高. 由此可见, 蛋白质间共享结构域的特性有助于提升蛋白质功能预测性能, 尤其适用于功能数量较多的蛋白质.

本文构建的多关系网络中, 蛋白质之间相互作用的第一种类型为共享结构域. 为提高预测的性能, 我们依据上述的统计分析结论对该种类型的相互作用加权. 统计表明, 两个蛋白质包含相同结构域的比例越高, 它们之间存在联系的可能性越大. 本文提出的 PEFM 算法中, 若两个蛋白质包含共同类型的结构域, 则它们之间存在相互作用. 相互作用的权值通过共同结构域的数量所占比重刻画, 加权计算方式如下:

$$W(v_i, v_j) = \frac{|D_i \cap D_j|^2}{|D_i| \times |D_j|} \tag{1}$$

其中, $W(v_i, v_j)$ 表示蛋白质 v_i 和 v_j 共享结构域的可能性. D_i 和 D_j 分别表示蛋白质 v_i 和 v_j 的不同类型结构域构成的集合, $D_i \cap D_j$ 是两个蛋白质相同结构域类型构成的集合. 若 D_i 或 D_j 为空集, 则权值简单地设置为 0.

蛋白质复合物由多个紧密联系的蛋白质组成, 并共同执行某些生物功能. 很多蛋白质只有聚合成复合物, 并与其他蛋白质相互作用才能体现出某种功能, 由此可见, 蛋白质复合物与功能之间存在紧密联系. 图 2 显示了蛋白质功能数量与共享蛋白质复合物之间的关系.

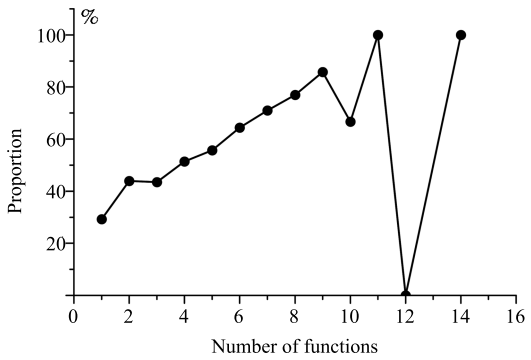

图 2　蛋白质功能与共享复合物统计分析

Fig. 2　Statistics of relationship between protein complexes and functions

从图 2 可以看出, 对于仅包含 1 项功能的蛋白质, 30 % 左右的蛋白质与其他有功能注释的蛋白质包含在相同的复合物中. 随着功能数量的增多, 这个比例明显增高. 包含 12 项功能的蛋白质仅 1 个, 此时比例为 0, 可以认为是偶然事件. 两个蛋白质共享相同的复合物是本文构建的多关系网络中的第二种类型. 这种类型的相互作用加权类似于第一种类型.

对于网络中的两个蛋白质 v_i 和 v_j, C_i 和 C_j 分别表示包含 v_i 和 v_j 的复合物组成的集合, 共享复合物的相互作用加权计算方式如下所示:

$$W(v_i,v_j)=\begin{cases}\dfrac{|C_i\cap C_j|^2}{|C_i|\times|C_j|}, & |C_i|\times|C_j|>0\\ 0, & |C_i|=0 \text{ 或 } |C_j|=0\end{cases} \tag{2}$$

其中, $C_i\cap C_j$ 表示同时包含 v_i 和 v_j 的复合物形成的集合. 若 v_i 或 v_j 没有出现在任何复合物中, 则 $W(v_i,v_j)=0$.

多关系网络中的最后一种类型来源于相互作用网络拓扑特性的分析. 众所周知, 蛋白质相互作用网络具有小世界特性和稀疏性, 且存在假阳性. 如果两个蛋白质都同时与第三个蛋白质发生相互作用, 则这两个蛋白质间相互作用假阳性的可能性比较小, 共同参与模块执行相同功能的可能性比较大. 因此, 一对蛋白质之间相互作用的概率可以通过他们共有的邻居节点数量确定. 本文采用 ECC 计算蛋白质之间连接的权值. 计算公式如下:

$$W(v_i,v_j)=\begin{cases}\dfrac{|N_i\cap N_j|^2}{(|N_i|-1)\times(|N_j|-1)}, & |N_i|>1 \text{ 和 } |N_j|>1\\ 0, & |N_i|=1 \text{ 或 } |N_j|=1\end{cases} \tag{3}$$

其中, N_i 和 N_j 分别表示 v_i 和 v_j 的邻居集合. 图 3 是本文结合 PPI 网络拓扑特性、蛋白质结构域信息和复合物信息构建的多关系网络的可视化展示.

图 3 中, 第一层表示蛋白质间因为隶属同一复合物而发生相互作用, 第二层表示蛋白质间因为包含共同的结构域而相互作用, 第三层则是在相互作用网络的基础上, 通过拓扑特征分析建立. 图中虚线将各层相同的蛋白质相连, 也就是说三层包含相同的蛋白质集合, 不同的是蛋白质间的相互作用.

1.2 关键功能模块挖掘

细胞的功能是由多个紧密联系的蛋白质通过形成功能模块执行. Zotenko 等提出关键复合物生物模块 (Essential complex biological modules, ECO-BIMs)[9], 它是一组紧密联系且共享生物功能的蛋白质组成. Nepusz 等[10] 指出, 子图能够表示为复合物应该满足两点: 1) 子图内包含许多可靠的相互作用; 2) 子图能够与网络的剩余部分很好地区分. 受此启发, 考虑到蛋白质功能与模块之间的紧密联系, 本文通过从多关系网络中挖掘关键功能模块, 实现蛋白质功能预测. 在介绍关键功能模块挖掘算法前, 先简要介绍算法所涉及的几个定义.

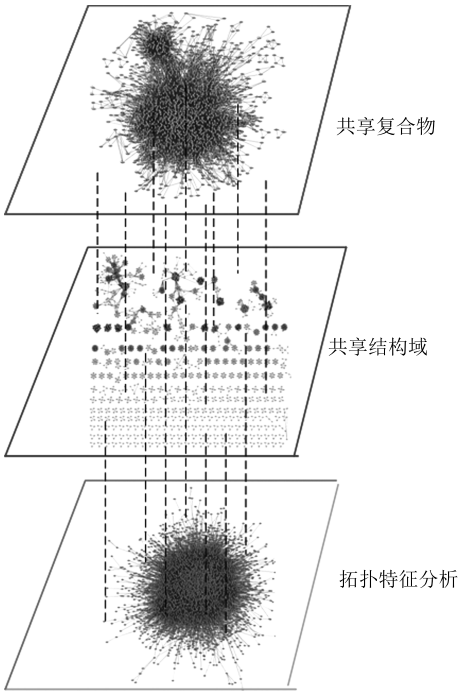

图 3 多关系网络可视化显示

Fig. 3 Visualization of a multi-relationship network

定义 1. 加权度 (Weighted degree, WD). 给定加权网络 $G=(V,E,W)$, 节点 $u\in V$, $V=\{v_1,\ v_2,\ \cdots,\ v_n\}$, $E=\{e_1,\ e_2,\ \cdots,\ e_m\}$, $W=\{w(e_1),\ w(e_2),\cdots,w(e_m)\}$, $w(e_i)$ 表示边 e_i 的权值. $WD(u,G)$ 表示 u 在 G 内的加权度, 定义如下:

$$WD(u,G)=\sum_{i=1}^{n}w(u,v_i),\quad (u,v_i)\in E \tag{4}$$

加权度描述了节点与子图之间的耦合程度. 加权度越大, 节点与子图内节点之间的联系越紧密. 本文采用加权度描述子图与网络剩余部分的区分度.

定义 2. 加权稠密度 (Weighted density degree, WDD). 给定加权子网络 $G=(V,E,W)$, $V=\{v_1,v_2,\cdots,v_n\}$, $E=\{e_1,e_2,\cdots,e_m\}$, $W=\{w(e_1),\ w(e_2),\cdots,w(e_m)\}$, $w(e_i)$ 表示边 e_i 的权值. $WDD(G)$ 表示子网 G 内的加权稠密度, 定义如下:

$$WDD(G)=\frac{2\times\sum_{i=1}^{m}w(e_i)}{n\times(n-1)} \tag{5}$$

加权稠密度用以描述子图内部节点之间的连接紧密程度. 本文通过加权稠密度衡量子图能否表示为高内聚的功能模块.

在 PEFM 方法中, 若子图的加权稠密度超过给定阈值, 且内部节点与子图的加权度大于节点与邻

居子图的加权度, 则该子图可以表示为一个高内聚、低耦合的关键功能模块. 邻居子图由子图内部节点的邻居节点组成, 并且这些邻居节点不出现在子图内.

关键功能模块挖掘的基本思路是: 对于待注释功能的蛋白质 v, PEFM 算法每次遍历同种类型的相互作用, 从而得到不同类型相互作用对应的关键功能模块. 本文中, 从 v 出发, 通过 3 次遍历, 最多可以得到 3 个关键生物模块. 每次遍历时, v 的邻居节点根据与 v 的连接紧密程度从大到小的顺序进入队列. 初始的关键功能模块集合 $S=\{v\}$, 算法依次从队列中取出一个邻居节点并尝试加入集合 S, 若加入邻居节点后, S 对应的子图加权稠密度超过设定的阈值 T, 则保留该节点, 否则将邻居节点从 S 中移除, 得到一个高内聚的稠密功能模块. 考虑到模块中某些节点可能与外部子图存在更加紧密的联系, 需要对子图 S 做进一步的筛选. NS 是由 S 中所有节点的邻居节点形成的子图, 若 S 中某一节点 u 在 NS 中加权度超过其在 S 中的加权度, 则从 S 中移除 u. 若 S 的尺寸超过 2 个, 则形成一个高内聚、低耦合的关键功能模块. 我们通过一个实例描述算法在某一网络中关键功能模块的挖掘过程. 如图 4 所示, A 节点为待注释功能的测试蛋白质, 加权稠密度阈值 $T=0.2$. 首先将 A 的邻居节点根据连接紧密程度依次放入队列 $Q=\{C,B,D,E\}$, 初始关键功能模块集合 $S=\{A\}$. 依次从队列 Q 中取出节点尝试放入 S 中, 并计算 S 的加权稠密度. 依次将 C, B, D, E 放入 S 中后, 得到的关键功能模块集合分别是 $\{A,C\}$, $\{A,C,B\}$, $\{A,C,B,D\}$ 和 $\{A,C,B,D,E\}$, 对应的加权稠密度分别是 0.5, 0.42, 0.24, 0.16. 由于加入邻居节点 E 后, 模块的加权稠密度低于设定的阈值, 因此从模块中移除节点 E, 形成高内聚的关键功能模块集合 $S=\{A,C,B,D\}$. C, B 和 D 的邻居节点形成邻居子图 $NS=\{H,F,G,K\}$. 由于 D 在 NS 中的加权度为 0.7, 大于其在 S 中的加权度 0.2, 从 S 中移除 D. 最终得到关键功能模块 $S=\{A,C,B\}$.

以下是关键功能模块挖掘伪代码描述:

输入. 多关系网络 $MG=(V,E,M)$; 阈值 T
输出. SM: 关键功能模块集合;

For $M=1$ to 3 **Do**
  For 节点 $v\in V$
    v 放入 S; //候选关键功能模块
    $Q=\{v_i|v_i\in V\wedge \mathrm{dis}(v,v_i)=1\}$
    For 元素 $q\in Q$
      将 q 放入 S 中;
      If $WDD(S)<T$
        从 S 中移除 q
    $NS=\{v_i|(v_i,v_j)\in E\ \wedge\ v_j\in S\ \wedge\ v_i\notin S\}$
    For 节点 $vs\in V$
      If $WD(vc,S)\le WD(vc,NS)$
        从 S 中移除 vc
      将 S 插入 SM, 标记为 (v,M).

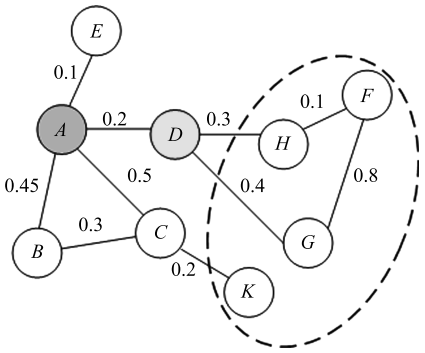

图 4　关键功能模块挖掘实例

Fig. 4　Example of an essential functional module mining

1.3　功能预测

算法的最后一个阶段是根据挖掘的关键功能模块形成候选功能列表, 并注释测试蛋白质. 在上一阶段, 已经产生每一种类型联系对应的关键功能模块. 然而, 不同类型的联系对于蛋白质功能预测的重要性各不相同. 为此, 我们为不同类型的联系设置不同的重要性系数. 重要性系数的计算如下:

$$IC(i)=2^{p(i)},\quad i=1,2,3 \tag{6}$$

其中, $P(i)$ 表示第 i 种类型联系的优先级. 优先级的设置源于统计分析的结果. 本文分别在每种类型联系构成的简单网络上运行经典的功能预测算法–邻居计数法 (Neighbour counting, NC), 预测蛋白质功能, 并计算每种情形下 NC 法的预测性能, 包括敏感性、特异性和 F-measure (相关定义见第 2.2 节), 实验结果如图 5 所示. 当 NC 方法运行在仅包含共享复合物类型的网络时, 能获得最高的敏感性和综合性能指标 F-measure. 共享结构域类型的性能次之, PPI 拓扑特征类型的性能最低. 因此, 共享复合物类型的优先级设置为 1, 共享结构域类型的优先级设置为 2, 而 PPI 拓扑特征类型的优先级为 3.

对于功能未知的测试蛋白质 u, 假设挖掘的关键功能模块集合 $FM=\{fm_1,fm_2,fm_3\}$, $fm_i=\{p_{i1},\ p_{i2},\cdots,p_{in}\}$ $(i\in[1,3])$ 表示第 fm_i 个模块包含的蛋白质集合, $WDD=\{wdd1,\ wdd2,\ wdd3\}$ 表示关键功能模块的加权稠密度, $F=\{f_1,f_2,\cdots,f_m\}$ 是三个关键功能模块中所有蛋白质的全部已知功能形成的集合. 对于 f_i 中某一功能, 可根据下式计算其排名得分:

$$Score(f_k) = \sum_{i=1}^{3} IC(i) \times \sum_{j=1}^{|fm_i|} t_{ijk} \times \frac{w(u, p_{ij})}{wdd(i)} \tag{7}$$

其中, $w(u, p_{ij})$ 表示蛋白质 u 和 p_{ij} 通过第 i 种联系时的权值. 若关键功能模块 fm_i 内的蛋白质 p_{ij} 包含功能 f_k, 则 $t_{ijk} = 1$, 否则 $t_{ijk} = 0$.

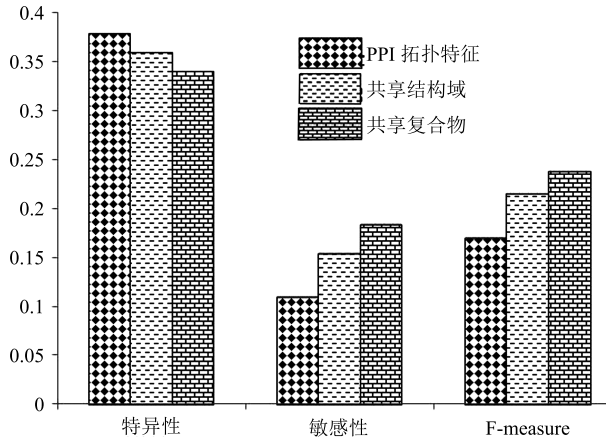

图 5 不同类型联系对预测的影响

Fig. 5 Impact of different types of connection

由于预测得到的候选功能比较多, 有的功能是噪声, 不宜注释测试蛋白质. 为此, PEFM 算法将所有候选功能按照得分降序排列, 然后从中选取前 N 项功能作为 u 的预测功能. N 是关键功能模块中与测试蛋白质 u 联系最为紧密的蛋白质的功能数量. 联系的紧密程度可以用相互作用的权值表示.

2 实验结果和分析

2.1 实验数据

本次实验将采用酵母蛋白质相互作用网络. 因为该物种的相互作用数据和功能数据较为完整, 并被用于现有的功能预测算法实验分析. 我们将详细介绍和分析 DIP[11] 数据集的结果, 也将简要分析 BioGrid[12] 数据集、Gavin[13] 数据集和 Krogan[14] 数据集. DIP 数据包含 5 093 个蛋白质和 24 743 组相互作用, Krogan 数据集则包括了 3 672 个蛋白质和 14 317 个相互作用, Gavin 数据库由 1 855 蛋白质及 7 669 蛋白质间相互作用组成, BioGrid 包括 5 616 个蛋白质和 52 833 组相互作用. 蛋白质功能数据为最新版本, 从 GO 官方网站获取[15]. 本次实验去除了注释蛋白质数量小于 10 个或者大于 200 的功能条目, 旨在提高算法的公平性. 处理完毕后, 注释文件包含 267 个不同的 GO 条目. 下载的 GO 文件进行了格式转换, 原始的 GO 文件为 UniProtKB[16] 格式, 转换后的格式为 Ensemble Genomes Protein. 用于构建多关系网络的 Domain 数据从 Pfam[17] 数据库获取. Domain 文件包含 1 107 种不同类型的结构域, 覆盖相互作用网络的 3 056 个蛋白质. 另一种异构数据, 蛋白质复合物数据采用基准集 CYC2008[18]. CYC2008 通过高通量的生物实验获得, 由 408 个 Benchmark 复合物组成. 为了检验 PEFM 算法的有效性和预测准确率, 我们选取了 FPM[7], Zhang[19], D-PIN[4], DCS[8], NC[2], PON[20] 作为对比算法. 本文将从多方面对比 PEFM 算法和竞争算法的性能.

2.2 评价指标

在测试蛋白质功能预测算法性能时, 通常采用交叉验证法. 蛋白质集合被划分为测试集和训练集. 训练集中的蛋白质用于帮助功能预测算法实现对未知功能的蛋白质注释. 测试集中蛋白质的功能被人为剥离, 利用预测算法得到其预测功能. 预测结束后, 对比预测的功能与真实的蛋白质功能的匹配情况, 从而计算功能预测算法的预测准确率. 交叉验证进一步可以划分为留一法验证和留部分法验证. 留一法验证是指每一轮预测时, 仅保留一个功能已知的蛋白质在测试集中, 剩余的蛋白质全部进入训练集. 留部分法验证是指随机地选取一定比例的蛋白质放入测试集, 例如 10 %, 20 %, 50 %. 剩下的功能已知的蛋白质放入训练集. 然后根据预测算法设定的功能选取策略选取一定数量的功能. 算法的预测准确率由预测的功能与实际功能之间的匹配率决定.

在计算功能预测算法的预测准确率时, 一般采用 Specificity (特异性) 和 Sensitivity (敏感性) 两种评价指标. Specificity 主要针对预测功能集, 指预测集合中被真实功能匹配的功能所占比例. Sensitivity 主要针对标准集, 指标准集中被预测的功能匹配的功能所占比例. Specificity 和 Sensitivity 的形式化定义如下:

$$\text{Specificity} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

其中, TP (True positive) 指预测集合中被标准集合中匹配的功能数量. FP (False positive) 指预测集合中没有被任何真实功能匹配的数量. 换句话说, FP 等于预测的功能数量减去 TP. FN (False negative) 指标准集合中没有被任何预测功能匹配的真实功能数量. 由于真实功能数量是固定的, 在预测蛋白质功能时, 提高候选功能数量, 可以提高 TP 值, 从而提高 Sensitivity 值. 同时导致 FP 增长更快, 导致 Specificity 明显下降. F-measure 是一项综合衡量

预测算法性能的指标, 是 Specificity 和 Sensitivity 的调和平均值.

2.3　参数分析

PEFM 算法中, 为评估子图加权稠密度, 我们引入自定义参数 T. 本节将分析 T 对算法性能的影响, 并确定 T 的合适取值. 根据定义 2 可知, T 的取值范围在区间 [0, 1]. 图 6 显示了在四个数据集 (DIP, Krogan, Gavin 和 BioGrid) 上, PEFM 算法的 F-measure 值随着 T 值变化的情况.

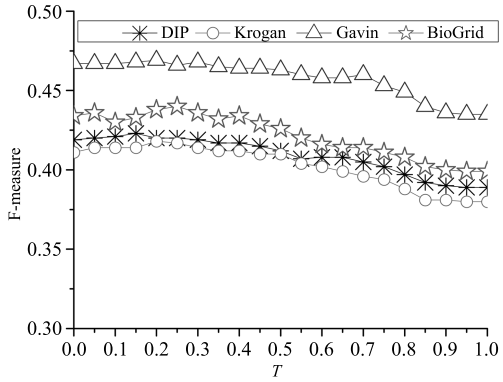

图 6　参数 T 的影响

Fig. 6　The effect of threshold T

从图 6 可以看出, 在 DIP 数据集上, 参数 T 取值 0.15 时, PEFM 算法获得最高的 F-measure 值 0.423. 对于 Krogan 和 Gavin 数据集, T 取值 0.2 时, 综合性能指标 F-measure 最大, 分别是 0.418 和 0.469. 对于 BioGrid 数据集, T = 0.25 时, F-measure 达到最大值 0.44.

2.4　留一法验证

本次实验选定的 PPI 网络中, 共有 5 093 个蛋白质, 其中 2 894 个蛋白质有功能注释. 我们首先分析 PEFM 和其他六种方法对这 2 894 个蛋白质预测功能的整体性能. 图 7 显示了各种方法的特异性、敏感性和 F-measure 的平均值. 2 894 个蛋白质中被 PEFM, D-PIN, FPM, Zhang, DCS, NC 和 PON 至少正确预测一个功能的蛋白质数量分别为 1 546, 1 506, 1 407, 801, 1 118, 1 626 和 566. PEFM 覆盖蛋白质数量比 D-PIN, FPM, Zhang, DCS 和 PON 分别提高 2.67 %, 9.88 %, 93.01 %, 38.28 % 和 173.14 %.

从图 7 可以看出, PEFM 具有最高的特异性 (Specificity), 这意味着 PEFM 算法预测的功能中错误 (噪声) 功能所占比例最少. 敏感性 (Sensitivity) 方面, PEFM 比 FPM、Zhang、DCS 和 PON 分别提高了 15.37 %, 95.63 %, 37.03 % 和 206.7 %. 这说明, 相比这四种功能预测算法, PEFM 算法在不增加噪声功能比例的前提下能够注释更多的蛋白质. PEFM 算法的敏感性明显低于 NC. 这是因为 PEFM 算法只选择了排名靠前的部分功能用于注释功能未知的蛋白质, 而 NC 方法是将邻居的所有功能全部赋予测试的蛋白质. 但是这种策略导致 NC 方法预测的功能中包含大量的噪声功能, 使得特异性急剧下降. 本次实验中, 虽然 NC 方法的敏感性比 PEFM 提高了 12.93 %, 但是特异性却比 PEFM 下降了 236.3 %. 因此, 就综合性能而言, PEFM 方法的 F-measure 值分别比 D-PIN, FPM, Zhang, DCS, NC 和 PON 提高 1.71 %, 20.72 %, 90.43 %, 35.28 %, 114.53 % 和 192.33 %. 由此可见, PEFM 方法具有最高综合性能.

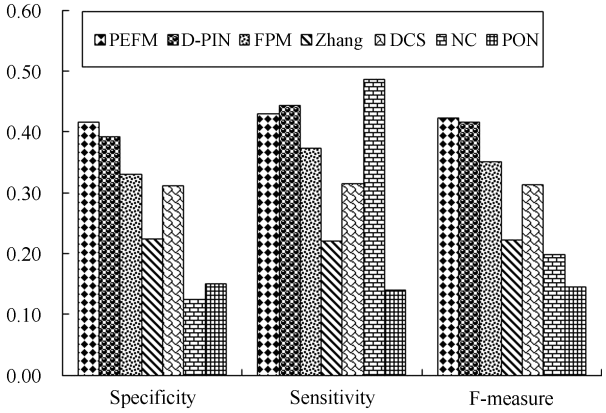

图 7　各种算法综合性能对比

Fig. 7　Overall performance comparison of various algorithms

为了更加全面、客观地对比分析各种方法的性能, 我们将尽可能地为各种方法选择相同的功能数量选取策略, 对每一个蛋白质, 分别选取各种方法预测的前 K 项功能进行预测. 针对 Zhang 方法和 DCS 方法, 选取前 M ($M \le K$) 个最相似的蛋白质, 从这 M 个蛋白质的功能列表中选取前 K 项功能作为预测的功能. 功能根据蛋白质的相似值的最大值降序排列 (例如, 有多个蛋白质具有某项功能 F_i, 则取这些蛋白质中与待预测的蛋白质最相似的蛋白质的相似值作为功能 F_i 的排序得分); 对于 D-PIN, FPM, PEFM, NC 和 PON 方法, 我们分别选取各自方法预测的前 K 个 GO Term 对功能未知的蛋白质进行功能注释. K 的取值从 1 ~ 50, 对于不同的 K 值, 分别计算各种方法的平均 F-measure 值, 对比结果如图 8 所示.

图 8 清晰地显示, 当 K 从 1 增长到 50 时, PEFM 始终具有最高的平均 F-measure 值. 随着 K 值的增长, PEFM 方法的 F-Measure 值虽然略微有所波动, 但基本能维持在 0.33 左右, Zhang 方法和

DCS 方法的 F-Measure 值则下降非常明显, 这说明 K 的选取对于 PEFM 算法的影响不大.

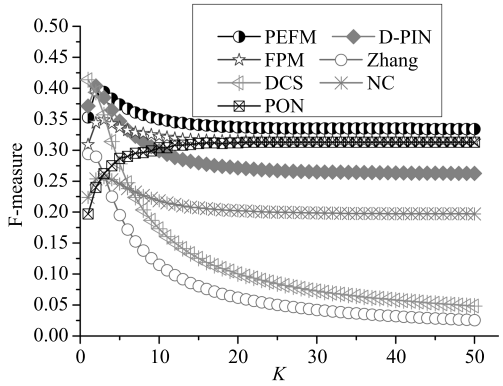

图 8 不同 K 值时各种算法的 F-measure 对比

Fig. 8 Comparison of average F measure of various algorithms under different K values

2.5 留部分法验证

我们已经采用留一法测试了 PEFM 算法的性能, 实验结果表明, PEFM 方法确实在现有方法的基础上提高了预测准确率. 实际应用中, 很多蛋白质的功能是缺失的. 本节将采用留部分法测试 PEFM 方法是否能在部分蛋白质功能缺失的情形下依然保持较高的准确率. 图 9 是留部分法实验结果.

我们随机移除 10 %、20 %、50 % 和 80 % 蛋白质的功能信息, 这部分蛋白质作为测试集, 剩余蛋白质为训练集, 用于对这部分蛋白质进行功能注释. 为尽量降低随机性对实验结果造成的误差, 我们对每个方法运行 1 000 次, 取平均值作最终结果.

从图 9 不难发现, 即便是移除 10 %, 20 %, 50 % 和 80 % 的蛋白质后, PEFM 方法依然获得最高的 F-Measure 值, 且优势比较明显. 即便部分蛋白质的功能信息缺失, 该方法依然能够取得优于现有功能预测方法的性能.

2.6 其他数据集结果

为了全面对比各种功能预测算法, 我们还采用留一法在其他三个不同的酵母相互作用网络 (Krogan 数据集、Gavin 数据集和 BioGrid 数据集) 测试了 PEFM 方法和其他六种对比方法. 表 2 列出了不同方法在三个网络上预测功能的实验结果.

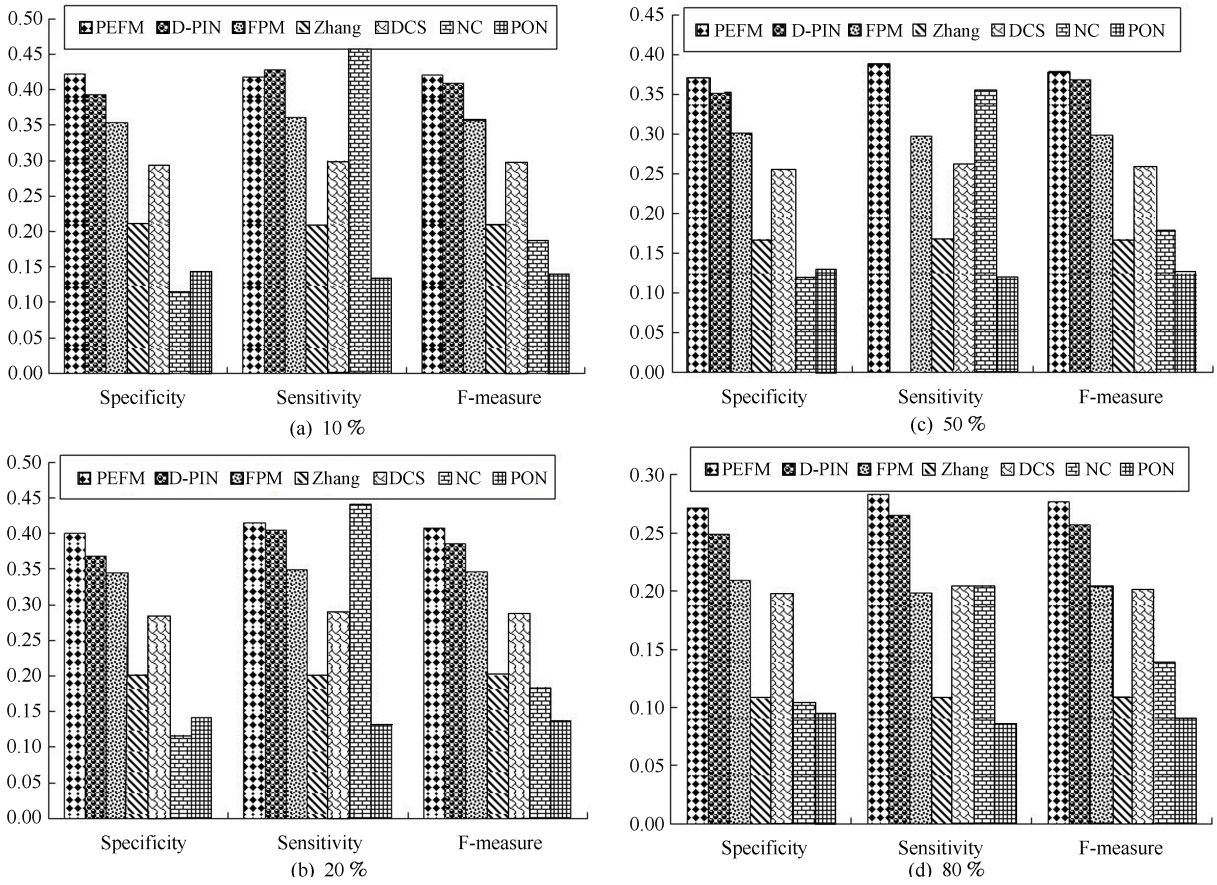

图 9 留部分法实验结果

Fig. 9 Results of leave-percent-out cross validation

表 2 Krogan, Gavin 和 BioGrid 运行结果
Table 2 Results of methods on Krogan, Gavin and BioGrid

Dataset	Method	Specificity	Sensitivity	F-Measure
Krogan	PEFM	0.412	0.423	0.418
Krogan	D-PIN	0.367	0.405	0.385
Krogan	FPM	0.317	0.342	0.329
Krogan	Zhang	0.231	0.221	0.226
Krogan	DCS	0.316	0.321	0.318
Krogan	NC	0.076	0.583	0.135
Krogan	PON	0.170	0.161	0.165
Gavin	PEFM	0.466	0.472	0.469
Gavin	D-PIN	0.443	0.486	0.463
Gavin	FPM	0.401	0.404	0.403
Gavin	Zhang	0.197	0.190	0.194
Gavin	DCS	0.381	0.393	0.387
Gavin	NC	0.210	0.603	0.311
Gavin	PON	0.155	0.145	0.150
BioGrid	PEFM	0.433	0.447	0.440
BioGrid	D-PIN	0.392	0.445	0.417
BioGrid	FPM	0.393	0.415	0.403
BioGrid	Zhang	0.236	0.233	0.235
BioGrid	DCS	0.370	0.375	0.372
BioGrid	NC	0.076	0.583	0.135
BioGrid	PON	0.170	0.161	0.165

从表 2 可以看出, 采用留一法在三个网络进行功能预测时, PEFM 依然取得最高特异性和 F-measure 值. 在不同数据集上的测试结果也证明了特异性算法的有效性. 综合上述分析, 相比其他几种功能预测算法, PEFM 算法具有最高的预测准确率.

3 结论

现有的蛋白质功能预测方法整合 PPI 网络和多元生物信息数据, 从而提高功能预测性能. 而融入多元信息后, 蛋白质之间的相互作用变得多样化. 不同类型的相互作用在功能预测中的作用各不相同. 将两个蛋白质间的多种相互作用进行简单合并, 虽然能有效地降低假阴性的影响, 增加预测的功能数量, 但同时也增加了假阳性功能的数量, 使得功能预测的整体性能提高不大. 本文利用网络拓扑特性、结构域信息和复合物信息构造多关系的蛋白质相互作用网络. 鉴于蛋白质功能与模块之间的紧密联系, 本文从多关系网络中挖掘关键功能模块, 利用关键功能模块的功能对蛋白质进行功能注释. 四个酵母的 PPI 网络上的实验结果验证了方法的有效性.

References

1 Zhao B H, Wang J X, Li M, Li X Y, Li Y H, Wu F X, Pan Y. A new method for predicting protein functions from dynamic weighted interactome networks. *IEEE Transactions on NanoBioscience*, 2016, **15**(2): 131−139

2 Schwikowski B, Uetz P, Fields S. A network of protein-protein interactions in yeast. *Nature Biotechnology*, 2000, **18**(12): 1257−1261

3 Dutkowski J, Ideker T. Protein networks as logic functions in development and cancer. *PLoS Computational Biology*, 2011, **7**(9): e1002180

4 Hu Sai, Xiong Hui-Jun, Zhao Bi-Hai, Li Xue-Yong, Wang Jing. Construction of dynamic-weighted protein interactome network and its application. *Acta Automatica Sinica*, 2015, **41**(11): 1893−1900
(胡赛, 熊慧军, 赵碧海, 李学勇, 王晶. 动态加权蛋白质相互作用网络构建及其应用研究. 自动化学报, 2015, **41**(11): 1893−1900)

5 Zhao B H, Wang J X, Li X Y, Wu F X. Essential protein discovery based on a combination of modularity and conservatism. *Methods*, 2016, **110**: 54−63

6 Li X Y, Wang J X, Zhao B H, Wu F X, Pan Y. Identification of protein complexes from multi-relationship protein interaction networks. *Human Genomics*, 2016, **10**(S2): 17

7 Hu Sai, Xiong Hui-Jun, Li Xue-Yong, Zhao Bi-Hai, Ni Wen-Yin, Yang Pin-Hong, Liu Zhen. Construction of multi-relation protein networks and its application. *Acta Automatica Sinica*, 2015, **41**(12): 2155−2163
(胡赛, 熊慧军, 李学勇, 赵碧海, 倪问尹, 杨品红, 刘臻. 多关系蛋白质网络构建及其应用研究. 自动化学报, 2015, **41**(12): 2155−2163)

8 Peng W, Wang J X, Cai J, Chen L, Li M, Wu F X. Improving protein function prediction using domain and protein complexes in PPI networks. *BMC Systems Biology*, 2014, **8**(1): 35

9 Zotenko E, Mestre J, O′Leary D P, Przytycka T M. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality. *PLoS Computational Biology*, 2008, **4**(8): e1000140

10 Nepusz T, Yu H Y, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 2012, **9**(5): 471−472

11 Xenarios I, Rice D W, Salwinski L, Baron M K, Marcotte E M, Eisenberg D. DIP: the database of interacting proteins. *Nucleic Acids Research*, 2000, **28**(1): 289−291

12 Stark C, Breitkreutz B J, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone M S, Nixon J, Van Auken K, Wang X D, Shi X Q, Reguly T, Rust J M, Winter A, Dolinski K, Tyers M. The BioGRID interaction database: 2011 update. *Nucleic Acids Research*, 2011, **39**(S1): D698−D704

13 Gavin A C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen L J, Bastuck S, Dümpelfeld B, Edelmann A, Heurtier M A, Hoffman V, Hoefert C, Klein K, Hudak M, Michon A M, Schelder M, Schirle M, Remor M, Rudi T, Hooper S, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick J M, Kuster B, Bork P, Russell R B, Superti-Furga G. Proteome survey reveals modularity of the yeast cell machinery. *Nature*, 2006, **440**(7084): 631−636

14 Krogan N J, Cagney G, Yu H Y, Zhong G Q, Guo X H, Ignatchenko A, Li J, Pu S Y, Datta N, Tikuisis A P, Punna T, Peregrín-Alvarez J M, Shales M, Zhang X, Davey M, Robinson M D, Paccanaro A, Bray J E, Sheung A, Beattie B, Richards D P, Canadien V, Lalev A, Mena F, Wong P, Starostine A, Canete M M, Vlasblom J, Wu S, Orsi C, Collins S R, Chandran S, Haw R, Rilstone J J, Gandi K, Thompson N J, Musso G, Onge P S, Ghanny S, Lam M H Y, Butland G, Altaf-Ul A M, Kanaya S, Shilatifard A, O′Shea E, Weissman J S, Ingles C J, Hughes T R, Parkinson J, Gerstein M, Wodak S J, Emili A, Greenblatt J F. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, 2006, **440**(7084): 637−643

15 Martin D M A, Berriman M, Barton G J. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 2004, **5**(1): 178

16 Lima T, Auchincloss A H, Coudert E, Keller G, Michoud K, Rivoire C, Bulliard V, de Castro E, Lachaize C, Baratin D, Phan I, Bougueleret L, Bairoch A. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Research*, 2009, **37**(S1): D471−D478

17 Hawkins T, Chitale M, Luban S, Kihara D. PFP: automated prediction of gene ontology functional annotations with confidence scores using protein sequence data. *Proteins: Structure, Function, and Bioinformatics*, 2009, **74**(3): 566−582

18 Pu S Y, Wong J, Turner B, Cho E, Wodak S J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, 2009, **37**(3): D825−D831

19 Zhang S, Chen H, Liu K, Sun Z R. Inferring protein function by domain context similarities in protein-protein interaction networks. *BMC Bioinformatics*, 2009, **10**(1): 395

20 Liang S D, Zheng D D, Standley D M, Guo H R, Zhang C. A novel function prediction approach using protein overlap networks. *BMC Systems Biology*, 2013, **7**(1): 61

赵碧海 博士, 长沙学院计算机工程与应用数学学院副教授. 2014 年获得中南大学信息学院博士学位. 主要研究方向为生物信息学, 数据挖掘.
E-mail: bihaizhao@163.com
(**ZHAO Bi-Hai** Ph. D., associate professor at the School of Computer Engineering and Applied Mathematics, Changsha University. He received his Ph. D. degree from Central South University in 2014. His research interest covers bioinformatics and data mining.)

李学勇 长沙学院计算机工程与应用数学学院教授. 2016 年获得中南大学信息学院博士学位. 主要研究方向为生物信息学. E-mail: xueyongli@163.com
(**LI Xue-Yong** Professor at the School of Computer Engineering and Applied Mathematics, Changsha University. He received his Ph. D. degree from Central South University in 2016. His main research interest is bioinformatics.)

胡 赛 长沙学院计算机工程与应用数学学院副教授. 2003 年获得湖南大学数学与计量经济学院硕士学位. 主要研究方向为生物信息学, 统计学. 本文通信作者. E-mail: husaiccsu@163.com
(**HU Sai** Associate professor at the School of Computer Engineering and Applied Mathematics, Changsha University. She received her master degree from Hunan University in 2003. Her research interest covers bioinformatics and statistics. Corresponding author of this paper.)

张 帆 长沙学院计算机工程与应用数学学院讲师. 2014 年获北京航空航天大学计算机学院博士学位. 主要研究方向为生物信息学.
E-mail: zf_ccsu@163.com
(**ZHANG Fan** Lecturer at the School of Computer Engineering and Applied Mathematics, Changsha University. She received her Ph. D. degree from Beihang University in 2014. Her main research interest is bioinformatics.)

田清龙 长沙学院数学与计算机科学系讲师. 2012 年获湖南大学信息科学与工程学院硕士学位. 主要研究方向为生物信息学, 机器学习.
E-mail: chinatql@126.com
(**TIAN Qing-Long** Lecturer at the School of Computer Engineering and Applied Mathematics, Changsha University. He received his master degree from Hunan University in 2012. His research interest covers bioinformatics and machine learning.)

杨品红 博士, 湖南文理学院生命科学学院教授. 1999 年获博士学位. 主要研究方向为水生生物资源与利用.
E-mail: yph098@163.com
(**YANG Pin-Hong** Professor at the School of Life Science, Hunan University of Arts and Science. He received his Ph. D. degree in 1999. His research interest covers aquatic biological resources and utilization.)

刘 臻 博士, 长沙学院生物与环境工程学院教授. 2010 年获博士学位. 主要研究方向为分子营养与调控研究.
E-mail: zhenliuccsu@163.com
(**LIU Zhen** Professor at the School of Biology and Environmental Engineering, Changsha University. He received his Ph. D. degree in 2010. His research interest covers molecular nutrition and regulation.)