计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200200004

基于结构洞的多数据源融合关键蛋白质识别方法

杨 壮 刘培强 费兆杰 刘 畅
山东工商学院计算机科学与技术学院 山东 烟台 264005
山东省高等学校协同创新中心:未来智能计算 山东 烟台 264005
(1083628707@qq.com)

摘 要 关键蛋白质识别是当前计算生物学领域的一个研究热点和难点。通过计算方法识别关键蛋白质的方法主要有DC,BC,LAC,PeC,ION和LIDC等。现有方法的识别准确率还有待进一步提高,主要原因是其仅使用了蛋白质相互作用网络单一数据源,以及蛋白质相互作用网络中存在许多假阳性和假阴性数据等。为了提高识别准确率,提出一种高效识别方法PSHC。首先,PSHC方法首次把结构洞理论引入到关键蛋白质识别方法中;其次,融合了蛋白质相互作用网络和蛋白质复合物两种数据源用于识别关键蛋白质。在真实数据上的实验结果表明,与其他传统方法相比,PSHC方法可以识别更多关键蛋白质,并且敏感度、特异性、准确性、阳性预测值、阴性预测值、F测度等统计指标也明显高于其他方法。

关键词:蛋白质相互作用网络;结构洞;蛋白质复合物;关键蛋白质
中图法分类号 TP301

Essential Protein Identification Method Based on Structural Holes and Fusion of Multiple Data Sources

YANG Zhuang,LIU Pei-qiang,FEI Zhao-jie and LIU Chang
School of Computer Science and Technology,Shandong Technology and Business University,Yantai,Shandong 264005,China
Co-innovation Center of Shandong Colleges and Universities:Future Intelligent Computing,Yantai,Shandong 264005,China

Abstract Essential protein identification is a hot research topic which is difficult in the field of computational biology. The existing methods for identifying essential proteins by computational methods are mainly DC,BC,LAC,PeC,ION,and LIDC,yet the identification accuracy needs to be further improved,mainly because only one data source is used which is protein interaction network,and there are many false positive and false negative data in the network. In order to improve the identification accuracy,an efficient essential protein identification method PSHC is proposed. Firstly,the PSHC method introduced the structure hole theory into the essential protein identification method for the first time. Secondly,the PSHC method combines two data sources of protein interaction network and protein complex to identify the essential proteins. Experimental results on real data show that PSHC can identify more essential proteins than other traditional methods,and statistical indicators such as sensitivity,specificity,accuracy,positive predictive value,negative predictive value,and F-measure are also higher than other methods.

Keywords Protein interaction network,Structural holes,Protein complex,Essential proteins

1 引言

蛋白质是组成生物体细胞、组织的重要成分,如细胞代谢、信号传导等生命活动是通过蛋白质相互作用(Protein-Protein Interaction,PPI)或蛋白质与DNA相互作用得以完成的。将蛋白质作为顶点、蛋白质之间的相互作用作为边,可以将所有相互作用的蛋白质看作一个复杂网络,称之为蛋白质相互作用网络(Protein Interaction Network,PIN)。

根据对生命活动的重要性差异,可以将蛋白质分为两类:关键蛋白质和非关键蛋白质。其中,关键蛋白质是关键基因的产物,在细胞生命中起着决定性作用[1],将其移除后,会造成有关蛋白质功能模块功能的丧失,使细胞无法进行正常生命活动,从而导致生物体无法生存或繁殖[2]。识别关键蛋白质,对于理解细胞存活和发育的最低要求、鉴定人类疾病基因和设计药物具有重要意义[3-4]。

识别关键蛋白质的方法包括生物实验方法和计算方法。生物实验方法有单基因敲除[5]、条件基因敲除[6]和RNA干扰[7]等。生物实验方法虽然得到的结果相对可靠,但存在耗时长、昂贵且低效等不足。酵母双杂交[8]和质谱串联亲和纯化[9-10]等高通量生物实验方法的发展,已积累了海量的PPI数据,使得通过计算方法识别关键蛋白质成为可能。计算方法弥补了实验方法的不足,成为了当前的研究热点。

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:山东省自然科学基金(ZR2017MF049);烟台市重点研发计划项目(2017ZH065)
This work was supported by the Shandong Provincial Natural Science Foundation(ZR2017MF049) and Key Research and Development Program of Yantai City(2017ZH065).
通信作者:刘培强(liupq@126.com)

通过计算方法识别关键蛋白质的方法大致分为基于 PIN 拓扑特性的方法和融合生物信息的方法。Jeong 等[11]提出了“中心性-致死性”法则，发现在 PIN 中高度连接的蛋白质节点比低度连接的蛋白质节点更重要。已经有许多基于 PIN 拓扑特性的中心性度量方法来识别关键蛋白质，例如度中心性方法(Degree Centrality，DC)[12]、介数中心性方法(Betweenness Centrality，BC)[13]、子图中心性方法(Subgraph Centrality，SC)[14]、接近度中心性方法(Closeness Centrality，CC)[15]、信息中心性方法(Information Centrality，IC)[16]、特征向量中心性方法(Eigenvector Centrality，EC)[17]、局部连通性方法(Local Average Connectivity，LAC)[18]、邻居中心性方法(Network Centrality，NC)[19]和邻居相互作用密度中心性方法(Local Interaction Density，LID)[20]等。虽然这些方法取得了一定效果，但其准确率仅仅是依靠了 PIN 的拓扑特性和可靠性，还有待于提高。

为提高识别准确率，一种重要方法是将 PIN 和蛋白质生物学信息结合起来。Li 等[21]基于 PIN 和基因表达数据，提出一种名为 PeC(Integration Person Correlationand ECC)的方法；Zhao 等[22]基于亚细胞定位信息和蛋白质同源信息，提出融合多数据源的方法 RWHN(Randomly Walking in the Heterogeneous Network)；Luo 等[23]通过整合蛋白质的局部相互作用密度和复合物信息，提出 LIDC(Local Interaction Density Combined with Protein Complex)方法来识别关键蛋白质；Qin 等[24]基于局部密度、介数中心性和蛋白质复合物信息，提出识别关键蛋白质的方法 LBCC(Local density，Betweenness Centrality and In-degree Centrality of Complex)；Zhang 等[25]整合蛋白质的同源信息、基因表达数据和 PIN，提出一种名为 OGN(Integrating Orthology，Gene Expressions and PPI Networks)的方法；Li 等[26]提出通过整合 PIN 拓扑特性和蛋白质复合物信息来识别关键蛋白质的方法 UC(United Complex Centrality)；Lei 等[27]基于蛋白质复合物参与度和子图密度，提出一种名为 PCSD(Participation Degree in Protein Complex and Subgraph Density)的方法；Li 等[28]结合亚细胞定位信息提出基于子网划分的方法 SPP(Sub-network Partition and Prioritization)；Lei 等[29]基于 GO 语义相似性、基因表达数据、亚细胞定位和蛋白质复合物，提出随机游走的方法 RWEP(Random Walk based method to identify Essential Proteins)。在识别关键蛋白质的精度上，融合生物信息的方法虽然比只使用 PIN 拓扑特性的方法有一定提高，但仍对 PIN 拓扑特性和蛋白质生物信息考虑不够全面，例如对于蛋白质复合物内部的节点，只考虑了该节点的内度，没有考虑蛋白质复合物外部节点的度，即没有考虑节点的邻域信息等。

针对以上问题，通过引入社交网络领域中的结构洞理论，以及融合蛋白质复合物信息，提出基于结构洞的多数据源融合关键蛋白质识别方法 PSHC(essential Protein identification method based on Structural Holes and fusion of protein Complex)。本文主要贡献如下：

(1)首次将结构洞理论应用于关键蛋白质识别方法中，在充分考虑该节点的度和其邻域信息来计算节点间的相似性基础上，提出结合结构洞与局部密度信息的拓扑约束指数；

(2)基于关键蛋白质更倾向于出现在蛋白质复合物中的事实，提出融合拓扑约束指数和蛋白质复合物信息的识别方法，提高了识别准确率；

(3)实现了 PIN 拓扑特性和蛋白质生物信息的多数据融合，提出一个 PSHC 公式来计算蛋白质的关键性。

2　相关概念

定义 1(简单图)　由给定 PIN 对应的无向简单图 $G(V,E)$表示，其中 V 表示节点集合，E 表示边集合，V 中的节点代表蛋白质，E 中的边代表蛋白质的相互作用。$v_i \in V$ 代表 PIN 中第 i 个蛋白质，N_v表示包含节点 v 的所有邻居的节点集。

定义 2(结构洞)　结构洞(Structural Holes，SH)为两个未连接节点之间的间隙。当两个节点通过第三个节点连接时，该间隙被填充，并为桥接节点创造了重要的优势。因此，结构洞(桥接节点)可以访问不同的信息流，并从邻居那里获得更多收益。Burt 给出了网络约束指数[30]，用于测量网络中的结构洞。结构洞将没有直接联系的两个行动者联系起来，拥有信息优势和控制优势，能够为自己提供更多的服务，在网络中的位置更为重要。以图 1 中的图 G_1 为例，YNL306W 和 YNL137C 都有可能是结构洞，其中 YNL306W 在 PIN 中是关键蛋白质。

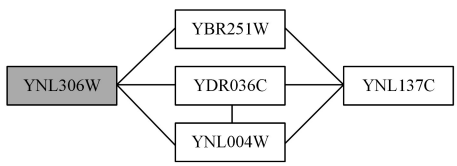


图 1　图 G_1

Fig. 1　Graph G_1

定义 3(复合物度中心性)　在简单图 $G(V,E)$中，节点 $v_i \in V$ 的复合物度中心性(In-degree Centrality of Complex，IDC)表示为：

$$IDC(i)=\sum_{j\in ComplexSet(i)} IN\text{-}Degree\ (i)_j \tag{1}$$

其中，$ComplexSet(i)$表示在真实蛋白质复合物集合中包含有蛋白质 i 的蛋白质复合物子集，$IN\text{-}Degree\ (i)_j$ 表示蛋白质 i 在集合 $ComplexSet(i)$中的第 j 个蛋白质复合物中的内度中心性值。

3　PSHC 方法

首先给出 PPI 网络中节点的拓扑约束指数，然后融合蛋白质复合物信息，给出关键蛋白质识别方法 PSHC。

3.1　拓扑约束指数

根据“中心性-致死性”法则，节点度的高低反映了对应节点在网络中的影响能力，是最常用的中心性测量指标。度越高，其对应蛋白质越倾向于是关键蛋白质。但研究发现，既存在部分度较高的蛋白质不是关键蛋白质，也存在部分度较低的蛋白质是关键蛋白质。此外，无法使用节点的度有效识别重要桥接节点，为此，引入结构洞理论，使用节点的度及其邻域信息来计算节点间的拓扑相似性，以提高识别准确率。

融合结构洞和局部密度，给出节点 v_i的拓扑约束指数，定义如下：

$$SD(i)=\begin{cases}\sum_j C_{ij}{}^2+S_{ij}, & i\neq j, N_i>1\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

$$C_{ij}=P_{ij}+\sum_q P_{iq}\times P_{qj},\quad i\neq q\neq j \tag{3}$$

$$S_{ij}=\frac{|N_i \bigcap N_j|}{\min(N_i,N_j)} \tag{4}$$

$$P_{ij}=\frac{W_{ij}}{\sum_{j}W_{ij}} \tag{5}$$

其中，N_i表示节点i的邻居节点的集合，$j\in N_i$，$q\in N_i$；W_{ij}表示相连的节点i和节点j之间的权重，由于蛋白质相互作用网络是无向无权简单图，且节点j是节点i的邻居节点，因此W_{ij}的值为1；$P_{ij}=1/d_i$，d_i代表节点i的度。

以图2中的图G_2为例，v_7的度为3，其邻居v_8和v_9是叶子节点，度均为1，根据“中心性-致死性”法则，v_7更可能成为关键节点；其邻居v_6的度也为3，而且v_6更靠近网络中心，所以v_6对于v_7的重要性比v_8和v_9更好。对于v_4与v_7、v_5与v_7，它们虽然没有直接相连，但通过v_6间接相连，v_6作为v_4与v_7、v_5与v_7的桥接节点，因此v_6在网络中的拓扑特性相对重要，有可能成为关键节点。

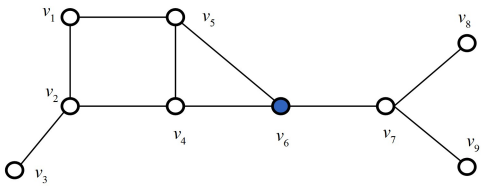


图2 图G_2

Fig. 2 Graph G_2

3.2 关键蛋白质识别方法 PSHC

通过将拓扑约束指数和蛋白质复合物的度中心性进行线性加权组合，提出一种新的方法PSHC。PSHC包含两点：

(1)以节点的度和其邻居信息为参数的拓扑关键性指数SD；

(2)以蛋白质复合物中蛋白质信息为参数的复合物度中心性指数IDC。

蛋白质i的重要性由式(6)给出的$PSHC(i)$进行度量：

$$PSHC(i)=\alpha\times SD(i)+(1-\alpha)\times IDC(i) \tag{6}$$

其中，α是调整SD和IDC贡献的参数，$\alpha\in(0,1)$。当$\alpha=0$时，仅考虑蛋白质复合物信息；当$\alpha=1$时，仅考虑结构洞约束指数。后文实验结果部分将详细讨论α的值。

PSHC描述如算法1所示。

算法1 PSHC

输入：PPI网络G=(V,E)，调节参数α，蛋白质数量K。

输出：得分排名前K的关键蛋白质。

```
1. /* 计算蛋白质复合物信息 */
For i=1 to n do
    根据式(1)计算 IDC(i)；
2. /* 计算顶点的结构洞约束指数 */
  For i=1 to n do /* n=|V| */
    For j=1 to m do /* m=Ni */
      根据式(2)计算 SD(i)；
      根据式(3)计算 Cij；
      根据式(4)计算 Sij；
      根据式(5)计算 Pij；
3. /* 计算蛋白质得分 */
  While not convergence do
    For each i in G do
      根据式(6)计算 PSHC(i)；
4. 根据 PSHC(i)值，取前K的蛋白质作为关键蛋白质输出。
```

3.3 PSHC复杂度分析

在包含n个蛋白质的PIN中，有r个蛋白质复合物，每个复合物中有l个蛋白质，计算蛋白质复合物中心性指数(IDC)的时间复杂度为$O(nrl^2)$；每个蛋白质有m个邻居，计算拓扑约束指数(SD)的时间复杂度为$O(nm^2)$；进行线性加权组合的时间复杂度为$O(n)$；排序的时间复杂度为$O(n\log n)$，因此算法1总的时间复杂度为$O(n(rl^2+m^2)+n\log n)$。

4 实验结果与分析

4.1 实验环境

硬件环境为Inter(R) Core(TM)i5-7500 3.41GHz CPU，8GB内存的PC机，软件环境为Windows 10操作系统，开发工具为CODE::BLOCK，实验采用C++语言实现。

4.2 实验数据

为评估PSHC方法的性能，使用酿酒酵母作为实验材料，因为该生物可获得相对可靠和完整的PPI数据。蛋白质相互作用网络来自DIP数据库[31]和Krogan数据库[32]。关键蛋白质数据来自以下4个数据库：MIPS[33]，SGD[34]，DEG[35]和SGDP[36]。其中，DIP数据集包含5093个蛋白质，1167个关键蛋白质；Krogan数据集包含2674个蛋白质，784个关键蛋白质。DIP和Krogan两个数据集的详细信息如表1所列。

表1 DIP和Krogan两个数据集的详细信息

Table 1 Detailed information of two PPI datasets

Dataset	Proteins	Interactions	Density	Essential Proteins
DIP	5093	24743	0.0018	1167
Krogan	2674	7075	0.0020	784

通常使用下述统计指标来衡量各种识别方法的性能：敏感度(Sensitivity，SN)，特异性(Specificity，SP)，阳性预测值(Positive Predictive Value，PPV)，阴性预测值(Negative Predictive Value，NPV)，F-测度(F-measure，F)和准确性(Accuracy，ACC)。

敏感度是指正确选择为关键蛋白的蛋白质与关键蛋白总数的比率：

$$SN=\frac{TP}{TP+FN} \tag{7}$$

特异性是指正确选择为非关键的非关键蛋白质占非关键蛋白质总数的比率：

$$SP=\frac{TN}{TN+FP} \tag{8}$$

阳性预测值是指正确选择为关键蛋白质的比例：

$$PPV=\frac{TP}{TP+FP} \tag{9}$$

阴性预测值是指正确选择为非关键蛋白的比例：

$$NPV=\frac{TN}{TN+FN} \tag{10}$$

F-测度是SN和PPV的谐波平均值：

$$F=\frac{2\times SN\times PPV}{SN+PPV} \tag{11}$$

准确性是指在所有结果中正确选择为关键和非关键蛋白的比例：

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

在上述公式中，真阳性(True Positives，TP)指关键蛋白质候选集中自身是关键蛋白质被识别为关键蛋白质的数目；假阴性(False Negatives，FN)指关键蛋白质候选集中自身是

关键蛋白质被识别为非关键蛋白质的数目；真阴性（True Negatives，TN）指关键蛋白质候选集中自身是非关键蛋白质被识别为非关键蛋白质的数目；假阳性（False Positives，FP）指关键蛋白质候选集中自身是非关键蛋白质被识别为关键蛋白质的数目。

4.3　参数 α 对关键蛋白质识别的影响

在 PSHC 方法中，蛋白质关键性评分由两部分组成：

（1）融合结构洞和局部密度的拓扑约束指数得分；

（2）蛋白质复合物度中心性指数得分。

由参数 α 调整两个部分的比重，其取值范围为[0，1]。当 α 为 0 时，仅考虑蛋白质复合物信息；当 α 为 1 时，仅考虑结构洞约束指数。根据最优化理论求解线性加权组合的专家评判法，将参数 α 设置为从 0.1 到 0.9，不同参数 α 对识别关键蛋白质数量的影响比较如表 2 所列。

表 2　不同参数 α 对识别关键蛋白质数量的影响比较

Table 2　Number of true essential proteins correctly identified by PSHC with different α

Dataset	α	Top100	Top200	Top300	Top400	Top500	Top600
DIP	0.1	79	157	221	277	318	359
	0.2	79	157	221	277	318	362
	0.3	79	157	222	278	319	363
	0.4	79	157	223	277	322	363
	0.5	80	156	224	277	323	367
	0.6	80	155	224	277	329	369
	0.7	77	151	227	278	319	372
	0.8	76	150	220	270	317	365
	0.9	73	141	207	256	298	348
Krogan	0.1	71	143	198	254	292	334
	0.2	71	142	198	254	291	334
	0.3	72	142	198	254	290	334
	0.4	72	141	199	256	290	333
	0.5	72	141	198	256	290	333
	0.6	72	142	198	255	291	331
	0.7	72	142	196	250	288	332
	0.8	73	142	195	245	289	325
	0.9	71	140	191	241	287	330

从表 2 可以看出，当 α 的取值为[0.5，0.7]时，*PSHC* 方法的关键蛋白质的识别数目较多；特别地，当 α 的值为 0.6 时，*PSHC* 方法识别的关键蛋白质最佳，因此将 α 的值设置为 0.6。

4.4　与其他方法进行比较

为验证 PSHC 方法的性能，将其与下述方法进行比较：DC，BC，SC，IC，EC，CC，NC，LAC，PeC，LIDC，LBCC 和 UC。首先将 PSHC 方法和以上各种方法分别在 DIP 和 Krogan 两个数据集上进行实验，得到每个蛋白质重要性的评分，其中 PeC，LIDC 和 UC 仅用于 DIP 数据集，CC 仅用于 Krogan 数据集；然后根据每个蛋白质的重要性评分进行降序排序，得到对应的关键蛋白质候选集；最后从每个关键蛋白质候选集中分别选取前 100，200，300，400，500，600 的候选蛋白质，与真实的关键蛋白质集进行比较，得到每种方法识别正确的关键蛋白质数量，如图 3 和图 4 所示，其中横坐标表示每种方法，纵坐标表示该方法识别出的关键蛋白质数目。

对于图 3 所示的 DIP 数据集，在前 100，200，300，400，500，600 个关键蛋白质候选集上，PSHC 方法分别识别了 80，155，224，277，329，369 个关键蛋白质，与其他识别方法相比，PSHC 方法识别的关键蛋白质数量最多。当选择前 100 个作为候选关键蛋白质时，PSHC 识别关键蛋白质的准确率达到了 80%。当选择前 600 个作为候选关键蛋白质时，PSHC 识别关键蛋白的准确率为 61.5%。在融合生物信息特征的方法中，表现较好的是方法 PeC 和方法 LIDC。在 6 个 Top 集上，PSHC 与融合基因表达信息的方法 PeC 相比，关键蛋白质识别准确率分别提高了 8.1%，11.5%，12.0%，9.9%，13.1%，12.5%；与融合蛋白质复合物的方法 LIDC 相比，关键蛋白质识别准确率分别提高了 5.2%，2.0%，7.2%，6.5%，5.1%，4.2%。同时与其他方法相比，基于结构洞的多数据源融合关键蛋白质识别方法也具有明显的优势。

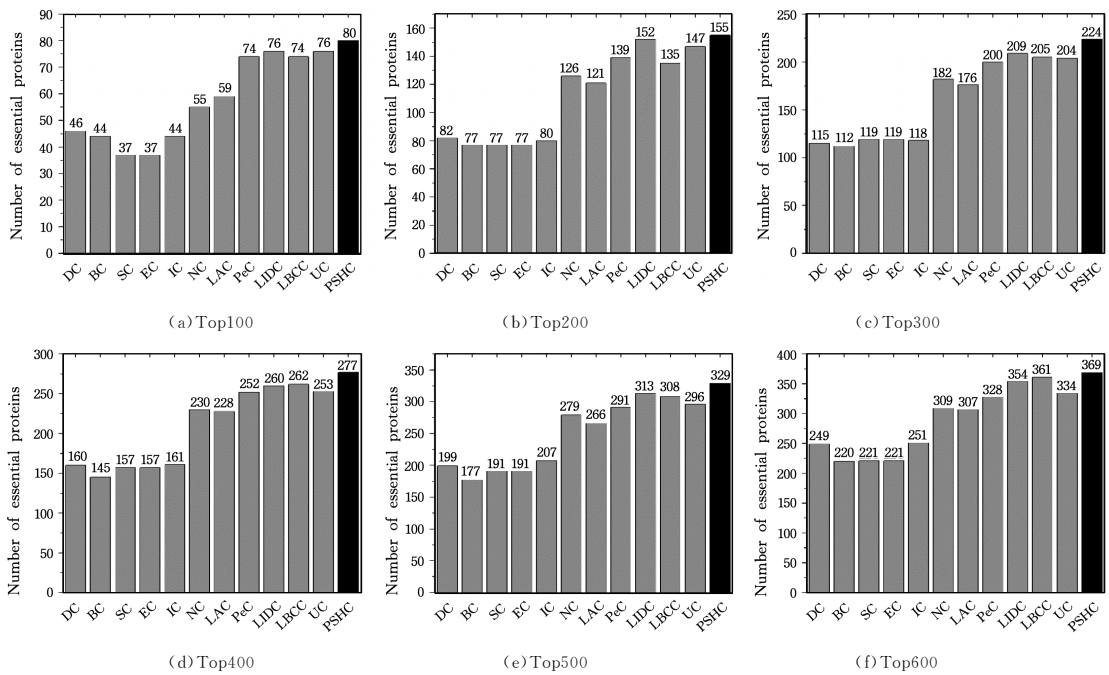
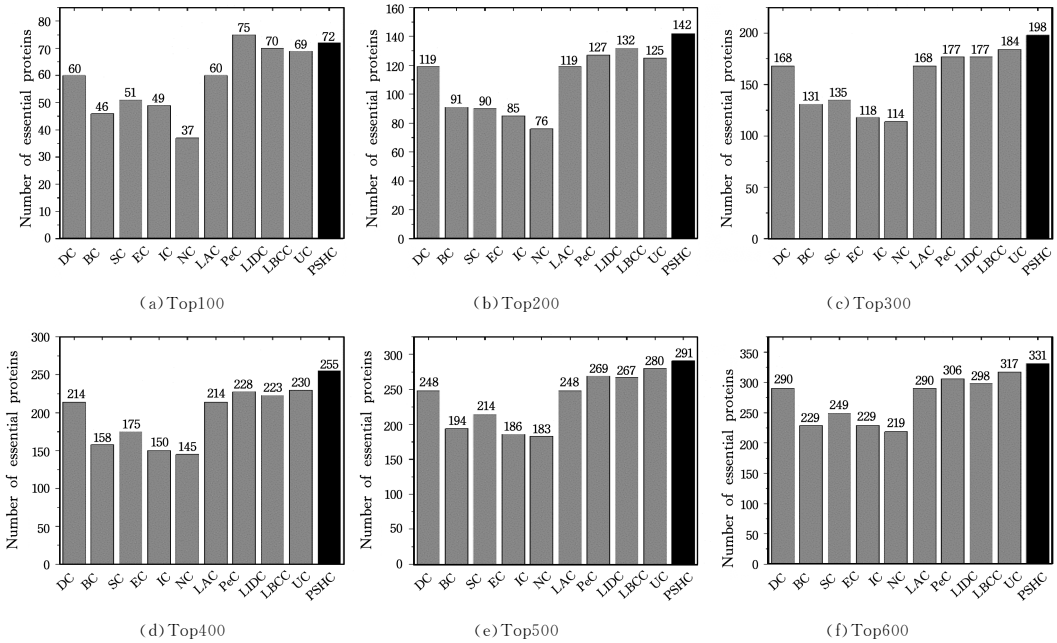


图 3　PSHC 与其他方法在 DIP 数据集上识别正确的关键蛋白质数量

Fig. 3　Number of true essential proteins predicted by PSHC and other several methods on DIP dataset



(a) Top100 (b) Top200 (c) Top300

(d) Top400 (e) Top500 (f) Top600

图 4 PSHC 与其他方法在 Krogan 集上识别正确的关键蛋白质数量

Fig. 4 Number of true essential proteins predicted by PSHC and other several methods on Krogan

对于图 4 所示的 Krogan 数据集，在前 100，200，300，400，500，600 个关键蛋白质候选集上，PSHC 方法分别识别了 72，142，198，255，291，331 个关键蛋白质，当选择前 100 个作为候选关键蛋白质时，NC 获得了最好的识别结果，从前 200 个到前 600 个作为候选关键蛋白质时，PSHC 取得了最好的识别结果。在 6 个 Top 集上，PSHC 与融合蛋白质复合物的方法 LBCC 相比，关键蛋白质识别准确率分别提高了 4.3%，13.6%，7.6%，10.9%，3.9%，4.4%。与其他的方法相比，PSHC 也具有明显的优势。

上述实验结果表明，在大多数情况下，PSHC 可以更有效地识别关键蛋白质。

4.5 基于折刀评价方法的比较实验结果

采用折刀方法(Jackknife)进一步评估 PSHC 和其他各种方法在识别关键蛋白质上的性能。首先，对各种方法得到的每个蛋白质的重要性评分进行降序排序；然后，根据降序后的关键蛋白质候选集，绘制关键蛋白质积累数量的折刀曲线。对比结果如图 5 和图 6 所示。其中，X 轴表示每种方法降序排列得到的候选关键蛋白质；Y 轴表示候选关键蛋白质中真正的关键蛋白质数量。各种方法的曲线与横坐标围成的面积表示对应的关键蛋白质识别准确率，面积越大，则识别关键蛋白质的准确率越高。

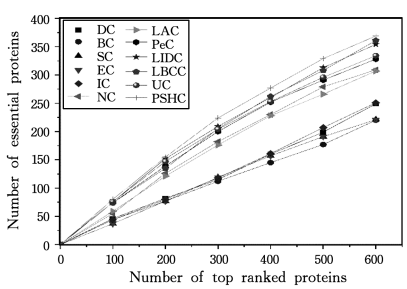


图 5 折刀法曲线图(DIP 数据集)

Fig. 5 Jackknife curves of PSHC and other several methods for DIP dataset

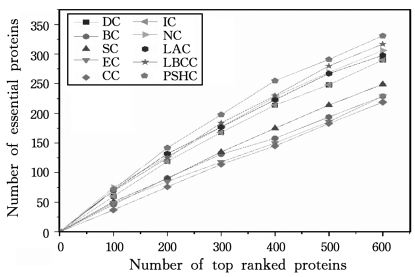


图 6 折刀法曲线图(Krogan 数据集)

Fig. 6 Jackknife curves of PSHC and other several methods for Krogan dataset

从图 5 和图 6 可以看出，随着候选关键蛋白质数量的变化，PSHC 方法所对应的曲线一直处于最高的位置，与横坐标围成的面积最大。因此与其他各种方法相比，PSHC 方法有着更高的关键蛋白质识别准确率，从而说明 PSHC 具有比其他方法更好的性能。

4.6 统计指标分析

为进一步分析 PSHC 方法的性能，使用敏感性(SN)、特异性(SP)、阳性预测值(PPV)、阴性预测值(NPV)、F 量度(F)和准确性(ACC)这 6 个指标对各种识别方法进行比较实验。对于 DIP 数据集，从所有方法的排名结果中选择前 1 167 个蛋白质作为关键蛋白质集，而其余蛋白质被认为是非关键蛋白集；对于 Krogan 数据集，从所有方法的排名结果中选择前 669 个蛋白质作为关键蛋白质集，而其余蛋白质被认为是非关键蛋白集。各种方法的统计指标比较结果如表 3 所列。从表 3 可以看出，对于 DIP 数据集，与其他各种方法相比，PSHC 方法的 6 种统计指标值最高，分别为 0.478，0.845，0.478，0.845，0.478，0.761。除 PSHC 方法外，其余各种方法的 6 种统计指标值最高的是 UC 方法，分别是 0.475，0.838，0.475，0.838，0.475，0.75。PSHC 方法与 UC 方法相比，各指标值分别高出了 0.003，0.007，0.003，0.007，0.003，0.011。对于 Krogan 数据集，PSHC 方法的 6 种统计指标值也最高，

分别为 0.457，0.835，0.535，0.788，0.493，0.724。除 PSHC 方法外，其余各种方法的 6 种统计指标值最高的是 LBCC 方法，分别是 0.441，0.829，0.517，0.782，0.476，0.715。PSHC 方法与 LBCC 方法相比，各指标值分别高出了 0.016，0.006，0.016，0.006，0.017，0.009。综上所述，PSHC 方法的 6 种统计指标的值均高于其他方法的值，表明 PSHC 方法可以更准确地识别关键蛋白质。

表 3　PSHC 方法与其他各种方法的统计指标比较

Table 3　Comparison of SN，SP，PPV，NPV，F and ACC between PSHC and other several methods

Dataset	Methods	*SN*	*SP*	*PPV*	*NPV*	*F*	*ACC*
DIP	DC	0.401	0.822	0.401	0.822	0.401	0.726
	BC	0.35	0.807	0.35	0.807	0.35	0.702
	SC	0.368	0.812	0.368	0.812	0.368	0.71
	IC	0.401	0.822	0.401	0.822	0.401	0.726
	EC	0.368	0.812	0.368	0.812	0.368	0.71
	NC	0.435	0.832	0.435	0.832	0.435	0.741
	LAC	0.451	0.837	0.451	0.837	0.451	0.748
	PeC	0.436	0.832	0.436	0.832	0.436	0.742
	LIDC	0.466	0.841	0.466	0.841	0.466	0.755
	LBCC	0.466	0.841	0.466	0.841	0.466	0.755
	UC	0.475	0.838	0.475	0.838	0.475	0.75
	PSHC	**0.478**	**0.845**	**0.478**	**0.845**	**0.478**	**0.761**
Krogan	DC	0.406	0.814	0.475	0.768	0.438	0.695
	BC	0.316	0.777	0.371	0.733	0.341	0.642
	SC	0.347	0.79	0.407	0.745	0.374	0.66
	EC	0.323	0.78	0.378	0.735	0.348	0.646
	CC	0.305	0.772	0.357	0.728	0.329	0.635
	IC	0.4	0.812	0.469	0.766	0.432	0.691
	NC	0.412	0.817	0.483	0.77	0.445	0.698
	LAC	0.415	0.818	0.486	0.771	0.447	0.7
	LBCC	0.441	0.829	0.517	0.782	0.476	0.715
	PSHC	**0.457**	**0.835**	**0.535**	**0.788**	**0.493**	**0.724**

结束语　关键蛋白在生物体的生存和繁殖中起着至关重要的作用，而且关键蛋白的鉴定有助于促进疾病研究和药物设计。目前，由于高通量技术的快速发展，PPI 数据急剧增加，已经有许多识别关键蛋白的方法。然而，蛋白质相互作用网络中存在噪声数据的影响，为准确地识别关键蛋白质带来了挑战。为提高关键蛋白质识别的准确率，结合结构洞理论，提出一种融合了蛋白质相互作用网络和蛋白质复合物两种数据源的关键蛋白质高效识别方法 PSHC。将 PSHC 在 DIP 和 Krogan 数据集上进行实验，并与其他各种方法(DC，BC，SC，IC，EC，CC，NC，LAC，PeC，LIDC，LBCC，UC)进行比较，实验结果表明，PSHC 在识别关键蛋白质方面明显优于其他方法。因此，PSHC 是有效的关键蛋白质识别方法。在未来的研究中，将融合其他生物信息，例如基因表达数据，更有效且更准确地识别关键蛋白质。

参 考 文 献

[1] PÁL C，PAPP B. Genomic function：Rate of evolution and gene dispensability[J]. Nature，2003，421(6922)：496-497.

[2] CLATWORTHY A E，PIERSON E，HUNG D T. Targeting virulence：a new paradigm for antimicrobial therapy[J]. Nat Chem Biol，2007，3(9)：541-548.

[3] LAMICHHANE G，ZIGNOL M，BLADES N J. A postgenomic method for predicting essential genes at subsaturation levels of mutagenesis：Application to Mycobacterium tuberculosis[J]. PNAS，2003，100(12)：7213-7218.

[4] STEINMETZ L M，SCHARFE C，DEUTSCHBAUER A M，et al. Systematic screen for human disease genes in yeast[J]. Nat Genet，2002，31(4)：400-404.

[5] GIAEVER G，CHU A M，LI N. Functional profiling of the Saccharomyces cerevisiae genome[J]. Nature，2002，418(6896)：387.

[6] ROEMER T，JIANG B，DAVISON J，et al. Large-scale essential gene identification in Candida albicans and applications to antifungal drug discovery[J]. Mol Microbiol，2003，50(1)：167-181.

[7] CULLEN L M，ARNDT G M. Genome-wide screening for gene function using RNAi in mammalian cells[J]. Immunol Cell Biol，2005，83(3)：217-223.

[8] ITO T，CHIBA T，OZAWA R，et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome[J]. Proceedings of the National academy of Sciences of the United States of America，2001，98(8)：4569-4574.

[9] AEBERSOLD R，MANN M. Mass spectrometry-based proteomics[J]. Nature，2003，422(6928)：198-207.

[10] HO Y，GRUHLER A，BADER G D，et al. Systematic identi © cation of protein complexes in Saccharomyces cerevisiae by mass spectrometry[J]. Nature，2002，415(6868)：180-183.

[11] H J，SP M，AL B. Lethality and centrality in protein networks[J]. Nature，2001，411(6833)：41-42.

[12] HAHN M W，KERN A D. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks[J]. Mol Biol Evol，2005，22(4)：803-806.

[13] JOY M P，BROCK A，INGBER D E，et al. High-betweenness proteins in the yeast protein interaction network[J]. J Biomed Biotechnol，2005，2005(2)：96-103.

[14] ESTRADA E，RODRIGUEZ-VELAZQUEZ J A. Subgraph centrality in complex networks[J]. Phys Rev E Stat Nonlin Soft Matter Phys，2005，71(5 Pt 2)：056103.

[15] WUCHTY S，STADLER P F. Centers of complex networks[J]. Journal of Theoretical Biology，2003，223(1)：45-53.

[16] STEPHENSON K，ZELEN M. Rethinking centrality：Methods and examples[J]. Social Networks，1989，11(1)：1-37.

[17] BONACICH P. Power and Centrality：A Family of Measures[J]. American Journal of Sociology，1987，92(5)：1170-1182.

[18] LI M，WANG J，CHEN X，et al. A local average connectivity-based method for identifying essential proteins from the network level[J]. Comput Biol Chem，2011，35(3)：143-150.

[19] WANG J，LI M，WANG H，et al. Identification of essential proteins based on edge clustering coefficient[J]. IEEE/ACM Trans Comput Biol Bioinform，2012，9(4)：1070-1080.

[20] QI Y，LUO J. Prediction of Essential Proteins Based on Local Interaction Density[J]. IEEE/ACM Trans Comput Biol Bioinform，2016，13(6)：1170-1182.

[21] LI M，ZHANG H，WANG J X，et al. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data[J]. Bmc Systems Biology，2012，6(1)：15.

[22] ZHAO B，ZHAO Y，ZHANG X，et al. An iteration method for identifying yeast essential proteins from heterogeneous network[J]. BMC Bioinformatics，2019，20(1)：355-368.

[23] LUO J，QI Y. Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Complexes[J]. PLoS One，2015，10(6)：e0131418.

(下转第 72 页)