基于 PPI 网络的关键蛋白质的高效预测算法

洪海燕[1] 刘 维[1,2]
(扬州大学信息工程学院 扬州 225127)[1]
(江苏省动物重要疫病与人畜共患病防控协同创新中心 扬州 225127)[2]

摘 要 关键蛋白质对于细胞生活是不可缺少的,识别关键蛋白质可以帮助了解细胞生活的最小需求,同时对药物设计也有非常大的作用。随着高通量技术的发展,人们可得到越来越多的蛋白质-蛋白质相互作用(PPI)的数据,这就使得可以在网络层次上来研究关键蛋白质。目前,学术界已经提出了一系列的计算方法来识别关键蛋白质,但这些方法并没有完全解决蛋白质相互作用数据的假阳性问题。除此之外,现有方法一般只考虑了网络的拓扑结构,对于生物信息的考虑,目前还是比较欠缺的。蛋白质对于人类细胞的生命活动不仅仅与网络拓扑结构有关,还和蛋白质在网络上的生物信息相关。因此,针对以上问题,提出了一种高效的预测关键蛋白质的新方法 EPP(Essential Proteins Predict),该方法通过计算蛋白质在 PPI 网络中的重要性来进行预测,蛋白质的重要性越高,成为关键蛋白质的可能性就越大。取重要性排名前 $P\%$ 的蛋白质作为关键蛋白质,在进行蛋白质重要性的计算时,综合考虑语义相似度及可信度因素,以综合考虑网络的拓扑结构与蛋白质本身的生物信息。实验结果表明,与其他传统方法相比,提出的新方法复杂度较低,且能够识别出更多的关键蛋白质,并且其统计指标也高于其他的方法。

关键词 关键蛋白质,GO,语义相似度
中图法分类号 TP39 **文献标识码** A

Efficient Prediction Method of Essential Proteins Based on PPI Network

HONG Hai-yan[1] LIU Wei[1,2]
(College of Information Engineering of Yangzhou University, Yangzhou 225127, China)[1]
(Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Diseases and Zoonoses, Yangzhou 225127, China)[2]

Abstract The essential protein is indispensable for cell life, so it is very helpful for us to understand the minimum requirements of cellular life and design the drug through identifying essential protein. With the development of high-throughput technologies, more and more protein- protein interactions (PPI) data has been obtained, which makes it possible to study essential protein from the network level. At present there are already a number of computational methods proposed for essential proteins identification, but these methods do not solve the PPI data false positive issues. In addition, existing methods generally just consider the topology of the network not considering biological information of protein on the network, and is still relatively lacking. Protein for human life activities of cells not only related to the topology of the network, but also related with protein biological information on the network. To solve the above problems, this paper presented an efficient new method to predict essential protein called EPP (Essential Proteins Predict). The algorithm predicts essential proteins through computing the importance score of protein in the PPI network, the higher importance score of protein is, the protein is more likely to be essential. We take the importance of rank $P\%$ of the protein as essential protein. When computing the importance score of essential protein, we synthetically considered the semantic similarity and credibility factors. Our method has low complexity, and considers not only the topology of the network but also the biological meaning of protein itself. Experimental results show that, compared with other conventional methods, our method can identify more essential protein, and its statistical indicators is also higher than other methods.

Keywords Essential protein, GO, Semantic similarity

关键蛋白质是指将其通过基因剔除式突变移除后,会造成有关蛋白质复合体功能丧失,并导致生物体无法生存的蛋白质[1]。对于细胞生活来说,关键蛋白质是不可缺少的,是细胞生命活动中所必需的。在蛋白质-蛋白质相互作用(PPI)网

本文受国家自然科学基金(61379066,61070047,61379064,61472344),江苏省自然科学基金(BK20130452,BK2012672,BK2012128),江苏省高校自然科学基金(12KJB520019,13KJB520026)资助。
洪海燕(1991-),女,硕士生,主要研究方向为生物数据挖掘;刘 维(1982-),女,博士,副教授,主要研究方向为数据挖掘、生物信息学等,E-mail:yzliuwei@126.com。

络中，若移除的是关键蛋白质将导致 PPI 网络的崩溃，使得生命活动不能正常进行。相反，若移除的是非关键蛋白质，则对该网络没有影响，生命活动可以正常进行。因此，能够正确地识别出关键蛋白质对我们研究细胞生活的最小需求具有重要意义。研究表明，致病基因往往表现为关键蛋白质[2]，所以正确地识别出关键蛋白质对药物设计也具有重要的意义。

随着生物信息学的发展，尤其是蛋白质相互作用数据的日趋完善，已有不少学者对关键蛋白质进行了深入的研究，因此近年来研究者们已经提出了很多识别关键蛋白质的方法。早期的方法是生物实验方法，典型的有 RNA 干扰[3]、单基因敲除[4]、条件性基因剔除[5]等。用生物实验方法来识别关键蛋白质虽然准确率高，但是在高通量进行时这些方法很容易出错。此外，对于有些致病的细菌培养是有危险的，所以很多研究者开始使用计算方法来探究蛋白质的关键性。

在计算方法上，大多数研究者都致力于研究 PPI 网络的拓扑结构和关键蛋白质的联系，这些研究已经表明关键蛋白质与 PPI 网络的拓扑结构是有关联的，Jeong 等人[6]在 2001 年提出了“中心·致死”法则(centrality-lethality rule)，此法则证明了关键蛋白质和拓扑结构的紧密关系，而且有高度连接的中心更趋向于是关键蛋白质。所以，一些研究者提出了一些基于节点拓扑中心性测度的方法，这些方法主要有度中心性(DC)[6]、介数中心性(BC)[7]、接近度中心性(CC)[8]、子图中心性(SC)[9]、特征向量中心性(EC)[10]、信息中心性(IC)[11]等。实验结果表明，在发现关键蛋白质时，这些研究方法比随机选取更有意义。除了上述方法外，还有基于网络模块或边的方法，该类方法主要有最大邻居连通度、最大邻居连通密度、局部平均连通度、边聚集系数和、模体中心性等。

但是，仅仅使用 PPI 数据来预测关键蛋白质是不足的，首先蛋白质相互作用的数据存在假阳性和假阴性，其次蛋白质相互作用的数据包含不稳定的相互作用或者发生在不同的时间点，相互作用网络不能代表真实的相互作用网络，其存在相互作用重复的数据。因此，一些研究者开始尝试将生物信息考虑进去，这样预测的关键蛋白质的准确率就大大的提高了，为此，Plaimas 等人[12]提出基于机器学习的方法，融合了多种网络拓扑特性、序列特征以及与关键性相关的共表达特征等，来识别一种细菌和另外一种细菌相关的一些关键基因，实验证明该方法取得了较好的精度。Lu 等[13]将语义相似度考虑进去，作为计算方法的一个参数同时分别将各个中心性的方法作为一个参数进行了实验，发现准确率明显比单纯地使用中心性的方法高出许多。Tang 等[14]将基因表达数据考虑进去，利用基因表达数据来计算两蛋白质节点之间的皮尔森相关系数，以衡量蛋白质之间的可信度，并且也将边聚集系数作为计算的参数，计算相互作用的边权值，这样降低了假阳性数据对实验结果的影响。实验结果表明，该方法比其他单纯的中心性方法准确率高。很明显，融合了生物信息的方法比只考虑拓扑结构的方法能够识别出更多的关键蛋白质。

本文提出了一个识别关键蛋白质的新方法，该算法将 PPI 网络看成是一个不确定的、顶点带有属性的网络，然后在该网络中求最具有重要性的前 $P\%$ 的顶点，该方法首先计算顶点间的相似度，相似度的计算考虑了蛋白质的可信度及语义相似度。首先根据基因本体数据库 GO 中反映的生物特性来计算蛋白质之间的语义相似度；其次，对于 PPI 网络中的每一个顶点(即蛋白质)，计算它的邻域相似度；最后利用上述两个相似度计算顶点的重要性。本文的算法具有复杂度较低、能够综合反映顶点之间的相似度以及它们之间发生相互作用的可信度的优点。实验结果表明，该方法能更准确地识别出更多关键蛋白质。

1 方法

我们可以把蛋白质之间的相互作用网络看成是一个不带权的无向图，在图中每个节点代表一个蛋白质，每条边代表两个蛋白质之间的相互作用，我们称之为蛋白质-蛋白质相互作用(PPI)网络。通过简单的图结构来描述蛋白质-蛋白质相互作用的方法在许多应用中都很常见，包括蛋白质相互作用的预测、关键蛋白质的预测等。但是在 PPI 网络中，由于不同的边被赋予的重要性不同，并且对于生物网络，还要考虑其生物意义，而基因本体数据库 GO 恰好提供了各种生物信息。所以将这两个数据源整合到一起，以便以不同程度的重要性对待 PPI 网络中的相互作用，从而识别出关键蛋白质。

我们知道，pageRank 是一个用于分析网页的重要性的著名算法。该算法的出发点是，网页的重要性取决于指向它的所有网页的重要性相加之和，当指向该网页的都是优质网页，则该网页就是优质的，同样，顶点 u 的重要性取决于和它相互作用的顶点的重要性，与 u 连接的顶点 v 的重要性越高，u 与 v 的相似度越高，u 的重要性也越高。因此，在本文中使用 pageRank 算法来预测关键蛋白质，只需要利用两顶点的相似性来提高预测的准确率，两顶点的相似性越大，它们之间有相互作用的可能性就越大。

有关研究表明，在 PPI 网络中蛋白质的关键性也存在着与网络节点的重要性同样的性质，即关键蛋白质之间的联系是非常密切的，它们间会有很大概率产生相互作用；而和其他蛋白质联系较密切的蛋白质成为关键蛋白质的可能性较大[15]。这里，蛋白质之间的联系的密切程度是由它们之间的生物特性的相似度和在 PPI 网络中的拓扑相似度联合决定的。因此，可以根据 pangRank 算法的思想，计算蛋白质在 PPI 网络中的关键性。

为了衡量两顶点之间的相互作用的大小，首先需要计算两个蛋白质间的可信度；其次根据 GO 计算蛋白质间的语义相似度，综合考虑可信度和语义相似度，计算得到两蛋白质间的相似性的大小，即发生相互作用的可能性的大小，基于此计算蛋白质的邻域相似度；最后根据 pangRank 算法，计算该蛋白质在 PPI 网络中的重要性。

1.1 问题定义

PPI 网络可以表示为一个无向不带权的图 $G=(V,E)$，其中 $V=\{v_1,v_2,\cdots,v_n\}$ 是所有蛋白质的集合，v_i 表示其中的一个蛋白质，$E\subseteq V\times V$ 表示所有蛋白质中有相互作用的边的集合，该图即为蛋白质-蛋白质相互作用(PPI)网络。本文提出的方法同时考虑了网络拓扑特征和节点的生物信息，生物信息考虑的是 GO 数据信息，通过它来计算蛋白质的重要性，从而构造出质量更高的 PPI 网络。

1.1.1 蛋白质相互作用可信度

给定表示 PPI 网络的图 $G=(V,E)$，定义其邻接矩阵 $A=$

$[a_{ij}]$，其中 a_{ij} 表示顶点 v_i 与 v_j 发生作用的可信度，对于顶点间的相似性的计算，根据文献[16]，采用共同邻居的方法来计算其相似性，如式(1)所示：

$$a_{uv}=|\Gamma(u)\cap\Gamma(v)| \tag{1}$$

其中，$\Gamma(u)$ 表示顶点 u 的邻居的集合，$\Gamma(v)$ 表示顶点 v 的邻居的集合，a_{uv} 表示顶点 u 和顶点 v 共同邻居的数量，即集合 $\Gamma(u)$ 和集合 $\Gamma(v)$ 中有相同的顶点的数量，这里的邻居指的是直接邻居。

1.1.2 GO 语义相似度

人们通过不断地对生物信息学基因方面的性质进行研究，到目前已有一系列通过基因本体来计算 GO 术语间语义相似性的方法。这些计算方法可以分为以下 3 类：基于信息量的方法、基于本体结构的方法和混合方法。Resnik 方法[17]定义两个术语的语义相似性为最近共同祖先(Lowest Common Ancestor，LCA)节点的信息量。Lin 方法[18]考虑两个 GO 术语的共同祖先的信息量并且还考虑 GO 术语本身的信息量计算短语之间的语义相似性。Xu 等[19]提出以两个术语所对应的节点在本体 DAG 图中的最短路径长度作为它们的语义相似度。

由于生物学家各自的研究角度、研究方法和研究背景的不同，对本质上相同的生物数据的定义并不完全相同，造成了语义的混乱，使研究变得困难。而 GO 的提出正是为了解决这个问题，规范了数据的描述，提供了三层结构的系统定义方式，用来描述基因和基因产物的特征，具体包括：细胞组件(Cellular Component，CC)、分子功能(Molecular Function，MF)和生物过程(Biological Process，BP)。在生物学中某一概念的描述就是通过这 3 个子本体中的术语构成的，把基因产物 g 的术语集合表示为 $Terms(g)=\{t_1,t_2,\cdots,t_n\}$，其中 t_i 表示基因产物的术语，n 表示这个术语集合中的术语数量，需要把数据集里的每个基因产物都用该集合表示，即需要得到每个基因产物的术语集合。

本文选择如式(2)的计算方法来计算蛋白质间的语义相似度。我们定义蛋白质之间的语义相似度矩阵 $B=[b_{ij}]$，其中 b_{ij} 为 v_i 与 v_j 的语义相似度。设 G_1、G_2 两个基因的语义相似度的计算公式为：

$$Sim_{match}(G_1,G_2)=|Terms(G_1)\cap Terms(G_2)| \tag{2}$$

其中，G_1、G_2 两个基因的语义相似度 $Sim_{match}(G_1,G_2)$ 用两个产物术语集合中有相同术语的数量来衡量。

1.1.3 顶点邻域相似度

根据式(1)和式(2)计算得到顶点间的可信度和语义相似度，首先可以计算出顶点之间的相似度 $Sim(u,v)$，如式(3)所示：

$$Sim(u,v)=\frac{1}{\alpha+\beta}[\alpha\cdot a_{uv}+\beta\cdot b_{uv}] \tag{3}$$

其中，$\alpha,\beta(\alpha,\beta>0)$ 为常数，a_{ij} 为顶点 v_i 与 v_j 发生作用的可信度，b_{ij} 为 v_i 与 v_j 的语义相似度。

然后，对于每个顶点 u，可以定义 u 的邻域为：

$$\Gamma(u)=\{v|(u,v)\in E\} \tag{4}$$

其中，$\Gamma(u)$ 表示和顶点 v 相邻的顶点 u 有链接的顶点的集合。

最后，给出顶点 u 的邻域相似度，定义如下：

$$N(u)=\sum_{v\in\Gamma(u)}Sim(v,u) \tag{5}$$

其中，$Sim(v,u)$ 表示顶点 u 和 v 的相似度，$\Gamma(u)$ 表示顶点 u 的邻域，$N(u)$ 表示与顶点 v 有链接的顶点 u 的相似度之和。

1.1.4 顶点重要性

由于在本文算法中，最后要筛选出重要性排名前 $P\%$ 的顶点，因此需要计算顶点的重要性。设对顶点 u，它的影响力为 $\varphi(u)$，取 $\varphi(u)$ 的初值为 d_u，d_u 为 u 的度数。这里不使用中心性、介数、聚合度等指标，目的是为了节省计算时间，因为仅为初值，后面是要经过修正的，初值的选取与结果无关，仅与迭代的次数有关。对 $\varphi(u)$ 的值使用式(6)进行迭代更新：

$$\varphi(u)=c\cdot\sum_{v\in\Gamma(u)}\frac{\varphi(v)\cdot Sim(u,v)}{N(v)} \tag{6}$$

其中，$c>0$ 为常数，$Sim(v,u)$ 表示的是顶点 u 和 v 的相似度，$N(v)$ 为顶点 v 的邻域相似度，$\Gamma(u)$ 为 u 的领域，$\varphi(u)$ 为蛋白质 u 的重要性。

1.2 算法描述 EPP

算法 1 EPP

输入：G=(V,E)：PPI 网络；
P：关键蛋白质所占百分比；
α，β：调节参数
输出：排名前 P% 的关键蛋白质
Begin
1. /* 初始化矩阵 A 和矩阵 B */
For i=1 to n do/* n=|V| */
For j=1 to n do
{根据式(1)计算 a_{ij}；
根据式(2)计算 b_{ij}
根据式(3)计算 Sim(i,j)；}
2. /* 计算顶点的邻域相似度 */
For i=1 to n do
For j=1 to n do
根据式(5)计算 N[i]；
3. /* 顶点重要性计算 */
While not convergence do
For each u in G do
根据式(6)计算 φ(u)；
4. 根据 φ(u) 值，取前 P% 的蛋白质作为关键蛋白质输出；
End

2 实验结果与分析

2.1 实验环境

实验验证本文所提出的基于蛋白质相互作用网络的关键蛋白质的预测算法 EPP 的有效性，实验程序的运行计算机系统配置为：Intel Core i7 2.5GHZ CPU，内存 8GB，Windows 7 操作系统，MyEclipse 的程序编辑、编译链接环境。所有的算法均采用 Java 语言实现。

2.2 标准数据集的选取

2.2.1 基因本体数据集

基因本体术语(GO terms)数据下载自基因本体数据库[20](2010 年 9 月 12 日的版本)，其中包含 31538 条基因本体术语，分为 3 类："分子组件"术语、"生物过程"术语和"分子功能"术语。这些术语间通过"is-a"和"part-of"等关系联系成 3 个有向无环图 DAG。

GO 的注释数据由参与合作的数据库来完成，它们使用 GO 的定义方法对它们所包含的基因产物进行注释。下载自

UniProt 数据库(2010 年 2 月的版本)[21]。

2.2.2 PPI 数据集

本实验采用酵母蛋白质之间的相互作用数据作为实验的相互作用网络,因为酵母的相互作用数据是所有物种中最为完备的。数据从 DIP[22] 数据库中下载,其中有 1167 个关键蛋白质,3591 个非关键蛋白,其余 335 个关键性未知,数据集中有 24743 组相互作用,去除重复的和自相互作用后还包含 5093 个蛋白质。

2.2.3 关键蛋白质数据集

本文实验中选择的研究对象是酵母的关键蛋白质数据集,数据集来源于 MIPS[23],SGD[24],DEG[25] 和 SGDP[26],总共有 1285 个关键蛋白质。如果一个蛋白质在以上提到的数据库中被两个或两个以上的数据库标记为关键蛋白,那么就认为此标记蛋白质即为关键蛋白质。

2.3 实验结果与分析

比较细胞组件(Cellular Component,CC)、分子功能(Molecular Function,MF)和生物学过程(Biological Process,BP)排序靠前 5%,10%,15%,20%和 25%的蛋白质中能够被正确识别的关键蛋白质数量,结果如表 1 所列。

表 1 基于 BP、CC 和 MF 基因本体 EPP 算法运行效果比较

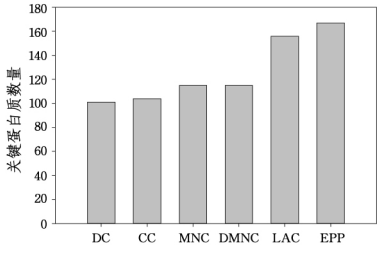
基因本体	5%	10%	15%	20%	25%
BP	158	280	395	487	576
CC	167	319	429	520	598
MF	148	279	382	479	556

从表 1 中可以看出,EPP 算法识别效果最好的是在由“细胞组件”基因本体构建的网络上,因此在下面的实验结果分析中,都采用基于“细胞组件”的 GO 信息来作为实验数据。

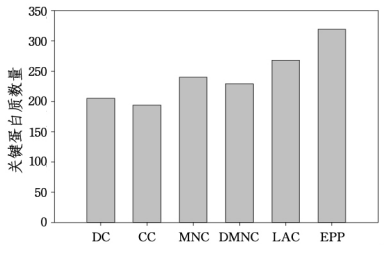
2.3.1 不同比例关键蛋白质预测准确率比较

为验证 EPP 算法的效率,将 EPP 算法与 DC[27],CC[28],MNC[29],DMNC[29],LAC[30] 方法的运行结果进行了对比。

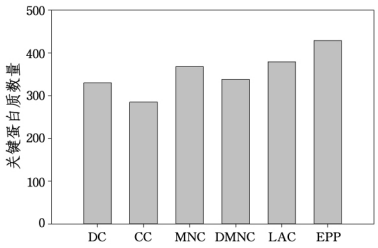
根据上面提出的方法分别计算出各自蛋白质的重要性排名并按递减排列顺序,分别取排序靠前 5%,10%,15%,20% 和 25%的蛋白质作为关键蛋白质的候选集。将每一比例的候选关键蛋白质集与本文所使用的数据库中标记的关键蛋白质逐一比对,从而确定候选集中的蛋白质中有多少是真正的关键蛋白质,计算出每一比例中关键蛋白质的数量。实验结果如图 1 所示。



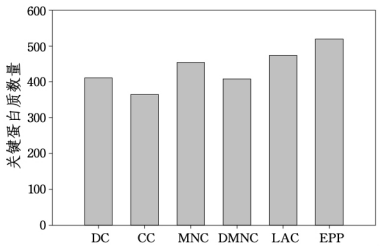
(a)Top5%关键蛋白质数量



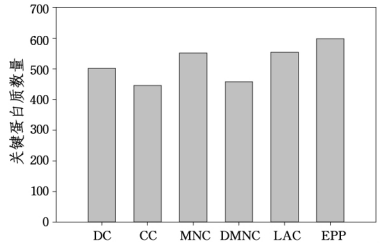
(b)Top10%关键蛋白质数量



(c)Top15%关键蛋白质数量



(d)Top20%关键蛋白质数量



(e)Top25%关键蛋白质数量

图 1 分别取排序靠前 5%,10%,15%,20%和 25%的蛋白质预测为关键蛋白质候选集中的关键蛋白质数量

从图 1 中可以看出,在酵母 PPI 网络中识别关键蛋白质获得的最好效果的是本文提出的 EPP 方法。在候选集取蛋白质总数的前 top5%,top10%,top15%,top20% 和 top25% 时,比较 EPP 与 DC,CC,MNC,DMNC,LAC 方法,可以看到 EPP 算法与这些方法中最好性能的方法还要分别提高 6.6%,16.0%,11.7%,8.8%和 7.4%,可见,GO 信息的加入提高了识别关键蛋白质的水平,同时也反映了关键蛋白质的特征。

2.3.2 统计指标分析

为了进一步验证 EPP 方法在关键蛋白质预测方面的性能,本文引入敏感度(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)和正确率(Accuracy,ACC) 6 个统计指标来对 EPP 方法与其它 5 种算法进行比较。具体定义如下:

$$(1)SN=\frac{TP}{TP+FN}$$

$$(2)SP=\frac{TN}{FP+TN}$$

$$(3)PPV=\frac{TP}{TP+FP}$$

$$(4)NPV=\frac{TN}{TN+FN}$$

$$(5)F=\frac{2*SN*PPV}{SN+PPV}$$

$$(6)ACC=\frac{TP+TN}{TP+FN+FP+TN}$$

其中,真阳性(True Positives,TP)为关键蛋白质被正确预测

的数目；假阴性(False Negatives，FN)为关键蛋白质被错误预测的数目；真阴性(True Negatives，TN)为非关键蛋白质被正确预测的数目；假阳性(False Positives，FP)为非关键蛋白质被错误预测的数目。以上 6 种指标的值越大，说明算法的识别效果越好，各方法的统计指标的比较如图 2 所示。

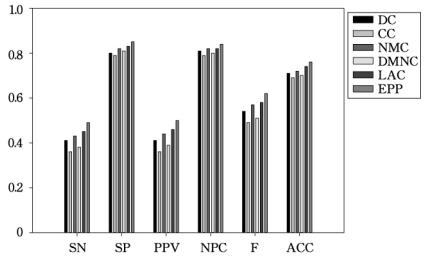


图 2　各方法的统计指标比较

从图 2 可以看出，EEP 算法与 5 种方法中最差的方法 CC 比，各项指标分别高出26.5%、7.1%、28.0%、6.0%、21.0%、9.2%，而与最好的方法 LAC 相比，各项指标分别高出8.2%、2.4%、8.0%、2.4%、6.5%、2.6%。这表明，EPP 算法在识别关键蛋白质上更加有效。

结束语　在后基因组时代，随着实验方法和计算方法的提高，人们获得了越来越多的蛋白质相互作用的数据，这就使得我们可以在网络层次上来识别关键蛋白质。因此，人们提出了一些基于网络拓扑特征的关键蛋白质的识别方法。但这些方法只考虑到了拓扑特性，而由于数据存在着假阳性，因此预测出的关键蛋白质的可靠性和真实性很低。为了衡量相互作用的可靠性，同时考虑可信度和语义相似度的因素来计算顶点之间的相似度，对于语义相似度的计算，使用了 GO 数据库。其次，还根据蛋白质与其邻居之间相似度计算出每个蛋白质的邻域相似度。本文算法在最后为每个蛋白质赋予一个重要性的得分。一个蛋白质的重要性越高，越可能是关键蛋白质。根据 PageRank 的算法，将上述因素综合考虑进去，进行迭代计算，计算出每个蛋白质的重要性，最终完成 EPP 算法。

目前关键蛋白质识别需要解决的关键问题之一是蛋白质间相互作用的可靠性，只要相互作用不可靠，识别出的关键蛋白质也不可靠，因此，本文结合可信度和语义相似度因素，提出了一种新的关键蛋白质识别算法 EPP，实验结果表明，该算法的准确率要优于其他算法。

参 考 文 献

[1] 刘舜民. 基于蛋白质相互作用加权网络的关键蛋白质识别算法研究[D]. 长沙：湖南大学信息科学与工程学院，2013

[2] Li M，Lu Y，Niu Z，et al. United complex centrality for identification of essential proteins from PPI networks[C]// 2015 The Institute of Electrical and Electronics Engineers. Computational Biology and Bioinformatics：IEEE，2015

[3] 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

[4] Wang J，Wei P，Wu F X. Computational approaches to predicting essential proteins：A survey[J]. Proteomics Clinical Applications，2013，7(1/2)：181-192

[5] Roemer T，Jiang B，Davison J，et al. Large-scale essential gene-identification in Candida albicans and applications to antifungal drug discovery[J]. Molecular Microbiology，2003，50(1)：167-181

[6] Jeong H，Mason S P，Barabási A L，et al. Lethality and centrality in protein networks[J]. Nature，2001，411(6833)：41-42

[7] Joy M P，Brock A，Ingber D E，et al. High. betweenness proteins in the yeast protein interaction network[J]. Journal of Biomedicine & Biotechnology，2005，2005(2)：96-103

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

[9] Estrada E，Rodriguez-Veldzquez J A. Subgraph centrality in complex networks[J]. Physical Review E statistical Nonlinear & Soft Matter Physics，2005，71(5)：122-133

[10] Bonacich P. Power and centrality：A family of measures[J]. American Journal of Sociology，1987，92(5)：1170-1182

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

[12] Plaimas K，Eils R，König R. Identifying essential genes in bacterial metabolic networks with machine learning methods[J]. BMC systems biology，2010，4(1)：56-72

[13] Lu Yu，Li Min，Li Qi，et al. A new method for predicting essential proteins based on topology potential[J]. IEEE International Conference on Bioinformatics and Biomedicine，2013：109-114

[14] Tang Xi-wei，Wang Jian-xin，Zhong Jian-cheng，et al. Predicting Essential Proteins Based on Weighted Degree Centrality[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics，2014，11(2)：407-418

[15] Li Min，Zheng Rui-qing，Zhang Han-hui，et al. Effective identification of essential proteins based on prbri knowledge，network topology and gen expressions[J]. Methods，2014，67(3)：325-333

[16] 吕琳媛. 复杂网络链路预测[J]. 电子科技大学学报，2010，39(5)：651-661

[17] Resnik P. Semantic similarity in a taxonomy：an information-based measure and its application to problems of ambiguity in natural language[J]. Semantic Ambiguity & Underspecification，1995，4196(1)：159-201

[18] Lin D K. An information-theoretic definition of similarity[C]// Proceedings of the 15th International Conference Machine Learning. New York：Association for Computing Machinery，1998，1：296-304

[19] Xu C L，Liang Y C，Guan Y L，et al. Turbo product codes for mobile multimedia broadcasting with partial-time jamming[J]. IEEE Transactions on Broadcasting，2007，53(1)：256-262

[20] Ashburner M，Ball C A，Blake J L，et al. The Gene Ontology Consortium[J]. Nat Genet，2010，25：25-29

[21] Consortium U. The Universal Protein Resource (UniProt) in 2010[J]. Nucleic Acids Research，2010，38(1)：142-148

[22] Xenarios L，Salwinski X J，Duan P，et al. DIP，the Database of Interacting Proteins：a research tool for studying cellular networks of protein interactions[J]. Nucleic Acids Res，2002，30(1)：303-305

[23] Mewes H W，Frishman D，Mayer K F X，et al. MIPS：analysis and annotation of proteins from whole genomes in 2005[J]. Nucleic Acids Res，2006，34(1)：169-172

[24] Cherry J M，Adler C，Ball C，et al. SGD：Saccharomyces Genome Database[J]. Nucleic Acids Res，1998，26(1)：73-79

(下转第 25 页)

数取 1.8，输入神经元个数为 9，隐藏层节点数为 6，输出神经元数为 5，构成 9-6-5 的网络结构。文献[14]中提出的粗糙集-BP 神经网络的扩展常数为 1.8，输入神经元个数为 9，隐藏层节点数为 7，输出神经元个数为 6，构成 9-7-6 的网络结构。本文的 IRSRBF 神经网络的扩展常数取 1.8，输入神经元个数为 5，隐藏层节点数为 5，输出层神经元个数为 5，构成 5-5-5 的网络结构。网络的故障判断阈值取 0.5，网络输出如果高于0.5，则判定出现故障。

将节点故障样本作为测试样本，比较 3 种诊断方法的效果，诊断结果如表 8 所列。从表中可以看到，3 种方法都可以得到正确的诊断结果。但是 3 种方法相比较，本文提出的 IRSRBF 神经网络经过改进的归纳属性约简后，只选取对故障诊断起重要作用的条件属性，使得神经网络的结构明显比其它两种方法的网络结构简单，训练时间更短，诊断结果的准确性更高。

表 8　常规 RBF，文献[14]方法和 IRSRBF 诊断结果对比表

故障类型	常规 RBF 神经网络的诊断结果	文献[14]方法的诊断结果	IRSRBF 神经网络的诊断结果
1	0.1255	0.1048	0.1000
2	0.1916	0.1455	0.1255
3	0.1000	0.1002	0.1000
4	0.6810	0.9120	0.9005
5	0.3120	0.2110	0.1005

图 4 是 IRSRBF 诊断算法与文献[14]中的同类诊断算法和常规 RBF 诊断算法的诊断结果比较。从图中可以看出 IRSRBF 的正确率在不同数据可靠性下，均比其它两种算法的正确率要高。并且当数据精度下降时，正确率的差值进一步扩大即表示数据可靠性越低，IRSRBF 算法越能保证正确率。分析其原因，故障诊断的实质是一个搜索属性的过程，由于 WSN 数据很难得到 100%的可靠性保证，常规 RBF 神经网络和同类方法在诊断时需要依赖的属性较多，其可靠性被分散在众多的数据上，导致可靠性下降。本文提出的 IRSRBF 算法通过数据属性约简减小搜索的空间，有效地避免了众多冗余属性对故障诊断正确性的干扰；同时利用 RBF 神经网络对噪声的不敏感性，抗干扰能力强的特征，进一步提高了诊断的正确率。

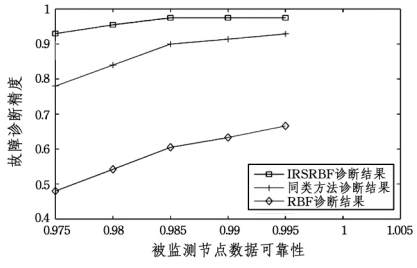


图 4　IRSRBF 与同类方法和 RBF 诊断结果比较

结束语　本文提出了改进的归纳属性约简算法 IIARA，并集成 RBF 神经网络，提出 WSN 故障诊断算法 IRSRBF。IRSRBF 算法能够在 WSN 节点信息不完整或者部分信息有误的情况下进行有效地故障诊断，特别是当数据的可靠性下降时，该算法在诊断速度和诊断精度上优势更大，提高了诊断系统的可靠性和 WSN 的实用性。

参 考 文 献

[1] 朱大奇，史慧. 人工神经网络原理及应用[M]. 北京：科学出版社，2006：33-56

[2] 韩力群. 人工神经网络教程[M]. 北京：北京邮电学院出版社，2006

[3] Zhu Liang，He Fei，Tong Yi-fei，et al. Fault detection and diagnosis of belt weigher using improved DBSCAN and Bayesian regularized neural network[J]. Mechanika，2015，21(1)：70

[4] Tayyab W，Mustafa D，Cemal K. Fault Diagnosis on Bevel Gearbox with Neural Networks and Feature Extraction[J]. Electronics & Electrical Engineering，2015，21(5)：69-74

[5] Xiong Guo-jiang，Shi Dong-yuan，Chen Jin-fu，et al. Divisional fault diagnosis of large-scale power systems based on radial basis function neural network and fuzzy integral[J]. Electric Power Systems Research，2013，105：9-19

[6] 熊国江，石东源，朱林，等. 基于径向基函数神经网络的电网模糊元细胞故障诊断[J]. 电力系统自动化，2014，38(5)：59-65

[7] Yan Jun，Jiang Xun-lin，Cheng Hai-song，et al. A hybrid training method for ANNs and its application in multi faults diagnosis of rolling bearing[J]. Journal of Vibroengineering，2015，17(7)：3538-3550

[8] Pawlak Z. Rough sets [J]. International Journal of Computer and InformationScience，1982，11(5)：341-355

[9] Akyildiz I F，Su W，Sankarasubramaniam Y. Wireless sensor networks：a survey[J]. Computer Networks，2002，38(4)：393-422

[10] 孙利民，李建中，陈渝，等. 无线传感器网络[M]. 北京：清华大学出版社，2008

[11] Noury N，Herve T，Riale V，et al. Monitoringbehavior in home using a smart fail sensor[C]// IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology. Lyon：IEEE Computer Society，2000：607-610

[12] Basile C，Gupta M，et al. An approach for detecting and distinguishing errors versus attacks in sensor networks [C]// Proceedings of the 2006 International Conference on Dependable Systems and Networks. 2006：474-484

[13] 王国胤. 粗糙集理论与知识获取[M]. 西安：西安交通大学出版社，2001：1-155

[14] 尚志信，周宇. 基于粗糙集和 BP 神经网络算法的网络故障诊断模型研究[J]. 宁波大学学报(理工版)，2013，26(2)：45-48

(上接第 20 页)

[25] Zhang R，Lin Y. DEG 5.0，a database of essential genes in both prokaryotes and eukaryotes[J]. Nucleic Acids Res，2009，37(1)：455-458

[26] Saccharomyces Genome Deletion Project[OL]. http://www.sequence.stanford.edu/group

[27] Jeong H，Mason S P，Barabasi A L，et al. Lethality and centrality in protein networks[J]. Nature，2001，411(6833)：41-42

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

[29] Lin C Y，Chin C H，Wu H H，et al. Hubba：hub objects Analyzer-a framework of interactome hubs identification for network biology[J]. Nucleic Acids Res，2008，36(2)：438-443

[30] Li M，Wang J X，Chen X，et al. A local average connectivity-based method for identifying essential proteins from the network level[J]. Computational Biology & Chemistry，2011，35(3)：143-150