一种基于随机游走模型的关键蛋白质预测方法

杨莉萍[1,2] 路松峰[2] 黄 钰[1]

1.华中农业大学信息学院,武汉 430070; 2.华中科技大学计算机学院,武汉 430074

摘要 为了解决目前的关键蛋白质预测方法对生物功能的分析不够深入的情况,利用蛋白质复合物信息,提出1种基于随机游走模型,结合蛋白质相互作用网络中的边聚集系数等数据来预测关键蛋白质的RWP(random walk method for predicting essential proteins)算法。在酿酒酵母(*Saccharomyces cerevisiae*)蛋白质相互作用网络上,以敏感度、特异性、阳性预测值、阴性预测值、准确率等5个统计学指标为评价标准,将RWP与介数中心性、度中心性、信息中心性、CSC算法及LIDC算法等5种用于预测关键蛋白质的方法进行对比实验。结果表明:RWP在关键蛋白质识别率等方面优于这5种测度方法,它具有较好的预测关键蛋白质的性能。

关键词 关键蛋白质; 随机游走模型; 蛋白质互作网络; 蛋白质复合物; 边聚集系数

中图分类号 Q 51; TP 301.6 **文献标识码** A **文章编号** 1000-2421(2016)06-0086-06

关键蛋白质是指在生物体的繁殖和生存中不可或缺的蛋白质。实验证明,若除去生物体中的某些关键蛋白质,会使生物体缺失某些特定的生物功能。关键蛋白质的预测对生命科学的研究意义重大,在药物设计和治疗疾病等方面的应用价值也很大[1]。研究表明,蛋白质的关键性和它们在蛋白质相互作用网络里的中心性关系紧密。Jeong等[2]提出了著名的"中心性-致死性"法则,认为相互作用较多的蛋白质对细胞的生存作用更大,并且在蛋白质网络中的度越高的蛋白质节点越倾向于表现出关键性。于是,研究者先后提出一些基于网络拓扑特性的中心性方法,包括度中心性(degree centrality,DC)[3]、介数中心性(betweenness centrality,BC)[4]、信息中心性(information centrality,IC)[5]等方法用于预测和识别关键蛋白质。但是这些中心性方法也存在一定的局限性,因为大多数中心性方法,例如BC、DC等方法很少去分析蛋白质内在的生物特性,而只是利用它们在网络中的介数、入度、出度等拓扑特性来对蛋白质进行打分,这样会影响关键蛋白质识别的准确性。Hart等[6]研究表明,关键蛋白质常聚集于某些具有特定功能的复合物中,蛋白质的关键性通常与蛋白质复合物的关系密切,不是只取决于个别的蛋白质。2008年,Zotenko等[7]提出了关键复合物模块的概念,指出了具有相同或相近生物功能的高度连通的蛋白质复合物功能模块中具有大量的关键蛋白质。2009年,Hwang等[8]系统地研究了蛋白质互作网络中关键基因和非关键基因的不同的拓扑属性,发现关键基因在互作网络中其拓扑属性上占据更重要的地位。通过统计分析,他们发现了关键基因区别于非关键基因的许多特有的拓扑属性。2013年,Luo等[9]提出了利用蛋白质的内度[10]去识别关键蛋白质的CSC算法。算法定义了结点u的复合物中心度,等于包含u的所有蛋白质复合物中,u为其中的各内度之和;并且CSC算法在结合复合物中心度与ECC中心度[10]的基础上,定义了蛋白质u的综合中心度INC。算法通过计算蛋白质互作网络中各蛋白质的INC值并排序来识别关键蛋白质。2015年,Luo等[11]研究了真正的蛋白质复合物中结点间的交互关系,将蛋白质互作网络中的结点划分成2类:互作结点和孤立结点,并在此基础上提出了识别关键蛋白质的新方法LIDC。该方法分为3部分:第1部分是基于蛋白质复合物中蛋白质的局部拓扑属性的中心性测度方法LID;第2部分是求蛋白质复合物的内度中心性IDC;第3部分是将LID和IDC相结合的策略。CSC算法和LIDC算法在利用蛋白质的复合物信息时,只考虑了复合物内

收稿日期: 2015-12-10
基金项目: 国家自然科学基金项目(61173050);中央高校基本科研业务费专项(2662015QC040)
杨莉萍,博士研究生,讲师. 研究方向:生物信息学、高性能计算. E-mail: teresa@mail.hzau.edu.cn
通信作者: 黄 钰,博士,副教授. 研究方向: 生物信息学. E-mail: yhuang@mail.hzau.edu.cn

部结点的内度，没有考虑复合物外部节点 u 的复合物参与度（即外部结点 u 的邻居结点参与复合物的程度），这样对蛋白质结点的复合物参与程度考虑不够全面。而且这 2 种算法在计算最终值时未使用随机过程中的收敛模型，使得算法的计算复杂度较大，且预测的准确性还不够高。Page 等[12]提出的 PageRank 算法，是基于随机过程中的随机游走模型来解决网络中的节点排序以及计算节点相似性等问题的著名算法。将 PageRank 算法应用到蛋白质功能预测中，可以对蛋白质拥有哪些功能进行排序，从而预测蛋白质的生物功能。2007 年，Freschi 等[13]提出了 ProteinRank 算法，该算法将 PageRank 算法公式中跳转到网络中其他蛋白质节点的跳转概率替换成了表示蛋白质拥有哪些功能的矩阵，算法通过结合全局网络特性来预测蛋白质的功能。但 ProteinRank 算法对于蛋白质网络中参与度小的蛋白质节点，算法的预测性能会比较差。

因此，本研究设计了 1 种基于随机游走模型的关键蛋白质预测方法 RWP（random walk method for predicting essential proteins），该方法结合了蛋白质复合物参与度信息，旨在提升关键蛋白质的预测性能，为关键蛋白质的识别研究提供新的思路。

1　材料与方法

1.1　数据来源

由于酿酒酵母（*Saccharomyces cerevisiae*）的关键蛋白质数据和蛋白质相互作用网络在众多物种中是最可靠和最完整的[14]。因此，本研究的实验数据来源选取的是酿酒酵母蛋白质互作网络。酿酒酵母蛋白质互作数据来源于 DIP 数据库[15]，其中共有 5 093个蛋白质，24 743 条边。酿酒酵母的关键蛋白质数量为 1 285 个，这些关键蛋白质数据通过整合 SGDP、MIPS[16]、SGD[17]等数据库中的蛋白质关键性信息而获得。

1.2　随机游走模型

随机游走模型的思想是：1 个粒子从图 $G=(V,E)$ 上的 1 个或若干个顶点出发开始遍历这张图。记粒子的起始点为 v_0，如果第 t 步它移动到顶点记为 v_t，那么接着粒子移动到与 v_t 相邻顶点的概率为 $1/\mathrm{d}(v_t)$，$\mathrm{d}(v_t)$ 是顶点 v_t 的度。显然，这些点组成的序列构成马尔科夫链[18]。起始点 v_0 可能是 1 个固定的起始顶点，也可能从 v_0 起始是以某个初始概率分布 P_0。这里用 $P_t(i)$ 表示第 t 步时刻走到顶点 i（$v_t=i$）的概率[19]。用 $M=(p_{ij})$，$i,j\in V$ 表示这个马尔科夫链的概率转移矩阵，其中

$$p_{ij}=\begin{cases}1/\mathrm{d}(i), & \text{if } i,j\in E\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

根据以上定义，随机游走的规则可以表示为：

$$P_{t+1}=MP_t \tag{2}$$

$$P_t=M^tP_0 \tag{3}$$

反复迭代这一过程，当 $t\to\infty$（即移动的次数足够多）时，v_t 的分布趋于平稳分布，即 $P_{t+1}=P_t$[20]。

1.3　复合物参与度得分

蛋白质复合物是指在相同的空间和时间通过相互作用组成 1 个多分子机制的 1 组蛋白质[6]。研究表明，蛋白质的关键性和蛋白质复合物之间关系紧密[6]。因此，选择的 RWP 算法在给蛋白质网络中的蛋白质打分时，先给网络中的每个蛋白质 1 个初始得分，即复合物参与度得分，该得分量化了蛋白质参与复合物的程度。

选择 Bader 等[21]提出的 MCODE 算法来挖掘蛋白质复合物（计算中，参数 haircut 取 True，参数 fluff 取 True，Fluff Density 取值 0.1）。将蛋白质相互作用网络中的所有蛋白质节点分为复合物内部节点和复合物外部节点 2 类。针对复合物内部节点 i，首先定义蛋白质节点 i 的内度，它是指与节点 i 处在同一蛋白质复合物 C_0 内部的 i 的邻居节点的个数[8]。即节点 i 的内度 $k_{in}(i,C_0)$ 用以下公式表示：

$$k_{in}(i,C_0)=|E(i,j)|,i,j\in V(C_0) \tag{4}$$

其中，$E(i,j)$ 表示蛋白质互作网络中两端点分别为节点 i 和 j 的边，$|E(i,j)|$ 表示这样的边的数目，$V(C_0)$ 是蛋白质复合物 C_0 中所有蛋白质节点构成的集合[22]。由于 1 个蛋白质可能属于多个不同的复合物，因此，将复合物内部节点的复合物参与度定义为该节点在其参与所有的蛋白质复合物中的内度的加权和。其中，权值 $w(C_i)$ 取值为每个蛋白质复合物 C_i 根据 MCODE 算法[21]得到的分值，该分值等于复合物 C_i 中蛋白质互作用边数与复合物中蛋白质节点数目之间的比值。针对复合物外部节点，算法主要考察的是该节点的邻居节点中有多少节点直接参与了复合物。即把复合物外部节点的复合物参与度定义为该节点的邻居节点中属于复合物内部节点的节点数目[22]。综上所述，蛋白质节点 i 的复合物参与度 $F_{\mathrm{D}}(i)$ 的计算公式定义如下：

$$F_D(i)=\begin{cases}\sum_{i\in Ci} k_{in}(i,C_i)\times w(C_i), i\in V(|C|)\\ \sum_{j\in V(|C|)} |E(i,j)|, i\notin V(|C|)\end{cases} \quad (5)$$

这里，$V(|C|)$表示蛋白质网络中位于蛋白质复合物内部的所有复合物内部节点组成的集合；C_i代表每个包含蛋白质i的蛋白质复合物[23]；$w(C_i)$是指蛋白质复合物C_i的权值。

1.4 边聚集系数

边聚集系数的概念来源于复杂网络，它刻画了连接边的2个节点与周围节点彼此之间联系的紧密程度。Hart等[6]的研究成果指出，在蛋白质相互作用网络中，蛋白质的关键性与该蛋白质和相邻的蛋白质之间连接的紧密程度非常相关。为此，我们在蛋白质互作网络中增加边聚集系数作为每条边的权重，来识别关键蛋白质。边聚集系数的定义如下：将蛋白质网络看作1个无向图$G=(V,E)$。Z_{ij}是基于边edge(v_i,v_j)能够构成的三角形的数目。k_i和k_j分别表示节点v_i和v_j的度。$m(k_i-1,k_j-1)$表示基于边edge(v_i,v_j)最大可能构成的三角形的数目。于是边edge(v_i,v_j)的边聚集系数为：

$$E_{CC}(i,j)=\frac{Z_{ij}}{m(k_i-1,k_j-1)} \quad (6)$$

1.5 RWP算法

在蛋白质互作网络中用边聚集系数每条边加权后，就可以构建蛋白质网络的关联矩阵来表示网络中蛋白质之间的关联程度。用H表示图$G=(V,E)$的$N\times N$关联矩阵。它的元素$h(i,j)$的值定义如下：

$$h(i,j)=\begin{cases}N_{orm_i}(E_{CC}(i,j))=\frac{E_{CC}(i,j)}{\sum_{w\in N_e(i)} E_{CC}(i,w)}, \text{if} \sum_{w\in N_e(i)} E_{CC}(i,w)>0\\ 0, \text{otherwise}\end{cases} \quad (7)$$

其中，$N_e(i)$是节点v_i的邻居集合。$N_{orm_i}(E_{CC}(i,j))$是标准化后的每条边的边聚集系数数值[24-26]。令$r(i)$表示蛋白质节点v_i的排序得分。$\sum_{j\in Ne(i)} h(i,j)r(j)$表示它邻居节点导出的得分。利用随机游走模型，对于蛋白质相互作用网络中的每个蛋白质，它的排序得分可计算如下：

$$r(i)=(1-\alpha)F_D(i)+\alpha\sum_{j\in Ne(i)} h(i,j)r(j) \quad (8)$$

蛋白质的排序得分可以看作是它的邻居相关节点的得分和该蛋白质的复合物参与度得分之间的线性组合。参数α ($0\leqslant\alpha<1$)用来调节2个得分的比重。由于在蛋白质互作网络中蛋白质的数量众多，计算复杂度非常大。因此，为了简化计算，将公式(8)改写成向量计算的形式，并利用Jacobi迭代的方法写成如下求解公式：

$$r^{t+1}=(1-\alpha)F_D+\alpha H\times r^t \quad (9)$$

其中，向量$r=(r(1),\cdots,r(N))$，向量$F_D=(F_D(1),\cdots,F_D(N))$。$t=0,1,2,\cdots$，表示的是迭代次数。

综上，RWP算法步骤描述如下：①依照公式(5)计算蛋白质互作网络中每个蛋白质的复合物参与度$F_D(i)$。②依照公式(6)算出每条边的边聚集系数。③用公式(7)构建矩阵H。④初始化r值为$r^0=F_D$。⑤用公式(9)计算r^t，迭代$t=t+1$。⑥重复第⑤步直到：$\|r^{t+1}-r^t\|1\leqslant\varepsilon$($\varepsilon$为结束误差)。⑦将蛋白质按照它们的$r$值降序排序。⑧输出排在前面$p\%$的蛋白质，作为预测的关键蛋白质。

2 结果与分析

2.1 实验结果

为了检测RWP算法的性能，本研究对本文材料与方法“1.1”节所述实验数据，分别使用度中心性法(DC)、信息中心性法(IC)、介数中心性法(BC)、CSC算法、LIDC算法和RWP算法进行计算。对计算结果先排序后筛选，选择排列在前1%、5%、10%、15%的蛋白质作为识别的关键蛋白质，再与已知的关键蛋白质数据集进行比对。实验结果如图1所示。从图1可知，RWP识别的关键蛋白质数量与经典的DC、IC、BC等3种中心性测度方法相比明显更多。无论是在前1%、前5%、前10%，还是前15%的样本水平上，RWP都比BC测度参数的预测

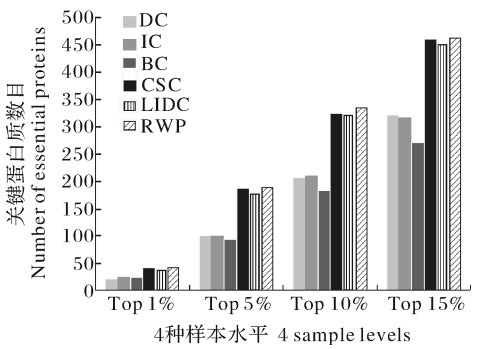

图1 RWP与其他5种预测方法比较

Fig.1 Comparing RWP with other five prediction measures

命中率高 25%以上。DC 是一种使用较为广泛的关键蛋白质预测方法，而与 DC 相比，RWP 方法也具有较明显的优势。与较新的 CSC 和 LIDC 方法相比，RWP 的预测准确率也略高些。总体来说，RWP 算法具有较好的预测性能。

2.2　评价指标

为了进一步检验 RWP 在关键蛋白质预测方面的性能，使用几个经典的检验指标来对比评价各种预测方法，包括特异性(specificity，SP)，即被正确排除掉的非关键蛋白质所占的比例；敏感度(sensitivity，SN)，即被正确识别的关键蛋白质所占的比例；准确率(accuracy，ACC)，即所有预测结果中正确结果的比例；阳性预测值(positive predictive value，PPV)，即甄选出的蛋白质其中被正确预测为关键蛋白质的占总数的比例；阴性预测值(negative predictive value，NPV)，即排除掉的蛋白质其中被正确预测为非关键蛋白质占总数的比例[14]。它们的计算公式分别为 SP＝TN/(TN＋TP)，SN＝TP/(TP＋FN)，ACC＝(TP＋TN)/(P＋N)，PPV＝TP/(TP＋FP)，NPV＝TN/(TN＋FN)。其中，TP(true positive)表示预测所得的关键蛋白质中被正确预测为关键蛋白质的个数，FN (false negative)表示预测所得结果中被错误预测为非关键蛋白质的个数，TN (true negatives)表示预测所得的非关键蛋白质中被正确预测为非关键蛋白质的个数，FP (false positives)表示预测所得结果中被错误预测为关键蛋白质的个数[14]。

在特异性等 5 个统计学指标上，对 RWP 和其他 5 种预测方法的预测结果进行了比较，对比结果如表 1 所示。从表 1 可以看出，RWP 算法的各个指标都比其他 5 种方法的指标值要高，说明 RWP 算法识别关键蛋白质的性能较好。

表 1　RWP 和其他 5 种方法在 SP、SN、ACC、PPV、NPV 指标值上的比较

Table 1　Compare RWP with other five prediction measures on SP,SN,ACC,PPV,NPV

预测方法 Prediction methods	特异性 Specificity	敏感度 Sensitivity	准确率 Accuracy	阳性预测值 Positive predictive value	阴性预测值 Negative predictive value
DC	0.807	0.438	0.724	0.429	0.822
IC	0.821	0.429	0.732	0.428	0.826
BC	0.826	0.379	0.725	0.387	0.812
CSC	0.827	0.456	0.736	0.462	0.825
LIDC	0.868	0.441	0.770	0.501	0.840
RWP	0.868	0.457	0.776	0.507	0.841

在式(8)中，对参数 α 的取值采用了经验值 0.5。随后又研究参数 α 的不同取值对 RWP 预测关键蛋白质准确性的影响。这里设置不同的 α 值，从 0 到 0.99，研究不同 α 值下 RWP 在识别关键蛋白质性能上的不同。实验结果如表 2 所示。其中，参数 p 的取值分别为前 1%、前 5%、前 10%、前 15%；预测准确性是指预测所得关键蛋白质中被正确预测为关键蛋白质所占的比例。表 2 中，当 $\alpha=0$ 时，识别关键蛋白质时仅仅只是考虑了蛋白质的复合物参与度得分；而 $\alpha=0.99$ 时，识别关键蛋白质几乎只是考虑了邻居相关节点的信息。并且当 $\alpha=0$ 或者 $\alpha=0.99$时，RWP 的预测准确性明显比 α 值介于 0 到 0.99之间时 RWP 的性能更差，这说明，结合蛋白质的复合物参与度属性和它的邻居节点属性会比只考虑其中任何一种属性更能准确地识别关键蛋白质。实验结果也证明，α 为 0.5 时 RWP 预测的准确性总

表 2　参数 α 对 RWP 预测准确性的影响

Table 2　Influence of the parameter α on RWP's prediction accuracy

参数 p Parameter p	参数 α Parameter α										
	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.99
前 1% Top 1%	0.728	0.823	0.781	0.783	0.790	0.812	0.803	0.786	0.804	0.751	0.702
前 5% Top 5%	0.662	0.688	0.721	0.724	0.725	0.751	0.743	0.727	0.732	0.713	0.678
前 10% Top 10%	0.613	0.638	0.648	0.662	0.653	0.661	0.647	0.636	0.632	0.628	0.622
前 15% Top 15%	0.572	0.578	0.582	0.590	0.576	0.581	0.579	0.573	0.567	0.562	0.528

体上最高。

3 讨 论

本研究引入了蛋白质复合物信息，提出了1种基于随机游走模型、结合蛋白质相互作用网络中边聚集系数等数据来预测关键蛋白质的RWP方法。与其他机器学习方法相比，本算法不需要事先输入部分已知的关键蛋白质数据。与现有的中心性方法相比，RWP方法预测关键蛋白质凭借的不仅是蛋白质的复合物参与度属性而且还包括它们邻居的属性，这样能够较好地克服蛋白质互作网络信息不可靠的弊端。实验结果证明，RWP方法识别的关键蛋白质的数量在任何一种样本水平上均多于中心性方法预测的数量。同时，RWP方法也为研究者结合蛋白质互作信息和其他生物信息来识别关键蛋白质提供了一个新的思路。

参 考 文 献

[1] WINZELER E A, SHOEMAKER D D, ASTROMOFF A, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis [J]. Science, 1999, 285 (5429): 901-906.

[2] JEONG H, MASON S P, BARABSI A L, et al. Lethality and centrality in protein networks [J]. Nature, 2001, 411 (6833): 41-42.

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

[4] JOY M P, BROCK A, INGBER D E, et al. High-betweenness proteins in the yeast protein interaction network [J]. Journal of biomedicine and biotechnology, 2005 (2): 96-103.

[5] STEVENSON K, ZELEN M. Rethinking centrality: methods and examples [J]. Social networks, 1989, 11 (1): 1-37.

[6] HART G T, LEE I, MARCOTTE E M. A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality [J]. BMC bioinformatics, 2007, 8 (1): 236.

[7] ZOTENKO E, MESTRE J, O'LEARY D P, et al. Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality [J]. PLoS computational biology, 2008, 4 (8): e1000140.

[8] HWANG Y C, LIN C C, CHANG J Y, et al. Predicting essential genes based on network and sequence analysis [J]. Molecular biosystems, 2009, 5 (12): 1672-1678.

[9] LUO J W, MA L. A new integration-centric algorithm of identifying essential proteins based on topology structure of protein- protein interaction network and complex information [J]. Current bioinformatics, 2013, 8 (3): 380-385.

[10] PADICCHI F, CASTELLANO C, CECCONI F, et al. Defining and identifying communities in networks [J]. PNAS, 2004, 101: 2658-2663.

[11] LUO J W, QI Y. Identification of essential proteins based on a new combination of local interaction density and protein complexes [J]. Plos One, 2015, 10 (6): 23-25.

[12] PAGE L, BRIN S, MOTWANI R, et al. The PageRank citation ranking: bringing order to the web [J]. Computer network and ISDN systems, 1998, 30 (1): 107-117.

[13] FRESCHI V. Protein function prediction from interaction networks using a random walk ranking algorithm [R]. [S.l.]: Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering, 2007: 42-48.

[14] 王峘. 基于蛋白质网络的关键蛋白质识别方法研究 [D]. 长沙：中南大学，2011.

[15] XENARIOS I, SALWINSKI L, DUAN Q J, et al. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions [J]. Nucleic Acids Res, 2002, 30 (2): 303-305.

[16] 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.

[17] CHERRY J M. SGD: saccharomyces genome database [J]. Nucleic Acids Res, 1998, 26 (1): 73-79.

[18] 徐晓华. 图上的随机游走学习 [D]. 南京：南京航空航天大学，2008.

[19] 邓小龙. 基于随机游走的蛋白质功能预测方法的研究 [D]. 长春：吉林大学，2012.

[20] 张春英. 基于属性图的社交网络建模与态势分析理论研究 [D]. 秦皇岛：燕山大学，2013.

[21] BADER G D, HOGUE C W. An automated method for finding molecular complexes in large protein interaction network [J]. BMC bioinformatics, 2003, 4 (2): 1-44.

[22] 李美满，邱炳城，王峘，等. 一种基于复合物参与度的关键蛋白质预测方法 [J]. 湘潭大学自然科学学报，2014, 36 (4): 84-87.

[23] 彭玮. 基于随机游走模型的蛋白质网络研究 [D]. 长沙：中南大学，2013.

[24] 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.

[25] 贾翠翠. 基于随机游走的蛋白质功能预测算法设计与实现 [D]. 哈尔滨：黑龙江大学，2014.

[26] BRIN S, PAGE L. The anatomy of a large-scale hypertextual web search engine [J]. Computer network and ISDN systems, 1998, 30 (1): 107-117.

A method for predicting essential proteins based on random walk model

YANG Liping[1,2]　LU Songfeng[2]　HUANG Yu[1]

1.*College of Informatics*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*;
2.*College of Computer Science and Technology*, *Huazhong University of Science and Technology*, *Wuhan* 430074, *China*

Abstract　The method for predicting essential proteins based on protein-protein interaction network is not deep enough to discover biological functions. In order to solve this problem, we utilize the protein complex information and propose an algorithm named RWP based on random walk model combining with the edge clustering coefficients in PPI network to recognize essential proteins. In protein-protein interaction network of *Saccharomyces cerevisiae*, 5 criteria of statistics evaluation criteria such as SN etc were taken to experiment with RWP and five centrality measure methods (DC, etc.) contrastively. The results showed that the number of essential proteins predicted by RWP was more than that predicted by other five centrality measure methods.

Keywords　essential protein; random walk model; protein-protein interaction network; protein complex; edge clustering coefficient

（责任编辑：陆文昌）