基于改进的PSO算法的关键蛋白质识别方法研究

洪海燕 刘 维
(扬州大学信息工程学院 扬州 225000)

摘 要 关键蛋白质是生物体内维持所有生命活动最重要的物质基础。随着高通量技术的发展,如何从蛋白质相互作用网络中识别出关键蛋白质成为目前蛋白质组学的研究热点。针对大部分现有方法仅仅基于网络拓扑结构信息进行识别以及蛋白质相互作用数据假阳性高的问题,提出了改进的粒子群算法来识别关键蛋白质。通过综合考虑网络拓扑结构特性和多源生物属性信息构建了高质量的加权网络,还考虑使用蛋白质节点间联系的紧密程度来衡量蛋白质的关键性,并扩展局部网络拓扑至二阶邻居,大大提高了预测的准确率。提出了衡量top-p关键蛋白质的整体性指标,降低了计算复杂度。在标准数据集上的实验结果表明,与其他经典算法相比,所提算法更具优势,能够识别出更多的蛋白质,具有较高的准确率。

关键词 关键蛋白质,PSO,蛋白质相互作用网络
中图法分类号 TP39 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.10.007

Research on Essential Protein Identification Method Based on Improved PSO Algorithm

HONG Hai-yan LIU Wei
(College of Information Engineering, Yangzhou University, Yangzhou 225000, China)

Abstract The essential protein is the most important material basis for the maintenance of all life activities in the living body. With the development of high throughput technology, how to identify the essential proteins from the protein interaction network has become a hot research topic in proteomics. For most of the existing methods are only based on the information of network topology for recognition as well as high false positive of protein-protein interaction data, this paper presented the improved particle swarm algorithm to identify the essential proteins. We considered the network topology characteristics and multi-source biological attribute information to construct the high quality of the weighted networks. We also considered node links between protein to measure the essentiality of protein, and expanded the local network topology to the second-order neighbor, improving the accuracy greatly. We proposed a measure of the overall top-p index, which reduces the computational complexity. The experimental results on standard data sets show that our algorithm is superior to other algorithms in comparison with other classical algorithms, which can identify more proteins with higher accuracy.

Keywords Essential protein, PSO, PPI

1 引言

众所周知,蛋白质在生物体的细胞生命活动中发挥着重要的作用,但由于每一种蛋白质在生命体细胞中参与各种不同的生命活动过程,因此不同蛋白质对生命活动的重要程度不尽相同,以此为依据,蛋白质可分为关键蛋白质和非关键蛋白质两大类。关键蛋白质一般是指通过基因剔除式突变将其移除后会造成生物体相关生物功能缺失,并导致生物体生病或无法生存的蛋白质[1-2]。

最初,研究者们大多采用生物实验方法对关键蛋白质进行识别,如单基因敲出[3]、RNA干扰[4]、条件性基因敲除[5]等,虽然生物实验法识别出关键蛋白质的准确率较高,但是该方法不仅费时费力,而且代价极高。因此,为了突破生物实验方法的局限性,利用计算的方法从已有的大量生物实验数据中快速而准确地识别出关键蛋白质,这是蛋白质组学研究的一个重要方向。

Jeong等[6]早在2001年就指出,在蛋白质相互作用(PPI)网络中,与其他蛋白质连接较多的蛋白质比随机选取的蛋白质更有可能是关键蛋白质,即参与相互作用较多的蛋白质对细胞的生存起着更为重要的作用,该现象被称为"中心-致死性"法则。近年来,许多研究者已经证明了蛋白质的关键性和拓扑中心性之间的正相关性[7-10]。基于后者的研究发

到稿日期:2016-09-24 返修日期:2016-12-31 本文受国家自然科学基金(61379066,61379064,61472344,61301220,61402395),江苏省自然科学基金(BK20130452,BK20151314,BK20140492),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,E-mail:1249055464@qq.com;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(通信作者)。

现，大量识别关键蛋白质的中心性方法已经被提出，其中，度中心性[6]（Degree Centrality，DC）、介数中心性[11]（Betweenness Centrality，BC）、接近度中心性[12]（Closeness Centrality，CC）、子图中心性[13]（Subgraph Centrality，SC）、特征向量中心性[14]（Eigenvector Centrality，EC）和信息中心性[15]（Information Centrality，IC）都是最典型的；还有其他的中心性方法，如局部平均连通度[16]（Local Average Connectivity，LAC）、边聚集系数和[17]（Sum of ECC，NC）等都被用于识别关键蛋白质。众所周知，基于网络拓扑方法的性能和 PPI 网络的质量是密切相关的。但是，在 PPI 网络中往往存在着大量的假阳性和假阴性，为了解决这个问题，Li 等[18]通过考虑基因注释来构建相对可靠的带权值的 PPI 网络，在带权值的网络中，基于网络拓扑的方法在很大程度上可以得到优化。此方法证实了拓扑中心性方法对 PPI 网络的噪声是敏感的。为了进一步提高预测准确率，研究者尝试结合生物信息和网络拓扑来识别关键蛋白质[19-21]，为此提出了一些新的方法，例如：PeC 方法考虑了边聚集系数和基因表达数据[22]，CoEWC 整合了聚集系数和基因表达数据[23]；除此之外，还有其他的整合方法，如通过整合边聚集系数和复合物的中心性方法等[24]。

虽然基于 PPI 网络拓扑的关键蛋白质识别方法取得了不错的进展，但其识别准确率依旧较低，这主要有两方面原因：1）在 PPI 网络中，数据存在着较高比例的假阳性和假阴性；2）大多数已有的方法仅仅考虑了网络的拓扑结构，没有结合关键蛋白质的内在属性和生物特性加以考虑，即各个特征属性在识别过程中所处的地位和作用不明确，目前缺乏各特征之间的相关性研究。此外，现有的关键蛋白质的识别方法都是首先根据蛋白质的某种重要性指标（如度中心性）排序，然后取前 P 个，虽然这样做可以识别出关键蛋白质，但是这些方法都需要逐一计算顶点的某种指标并排序，无形中大大增加了计算量。针对上述情形，Li 等提出了一种 CEPPK 算法[25]，该算法根据已有的 k 个关键蛋白质，寻找其邻居集合中的最“密切”者，并将其逐个加入关键蛋白质集合中，直至 P 个为止。但是该算法本质上是一种贪心算法，容易陷入局部最优。例如，假设计算出的 k 个关键蛋白质只是分布在 PPI 网络的某一小部分，则结果得到的 P 个关键蛋白质也只能在其所在的邻域。

综上所述，由于大规模蛋白质网络中通常存在一定数量的假阳性数据，而有些网络的数据不够完备，因此仅仅依靠网络的拓扑特性来识别关键蛋白质较依赖于网络的可靠性。考虑到蛋白质网络特有的生物特性，本文提出了一种高效的基于改进的粒子群（PSO）算法用于识别关键蛋白质。该算法通过融合蛋白质网络拓扑特征和生物特性来达到改善由于网络可靠性低造成的识别准确率低的状况。在该方法中只需识别出 P 个关键蛋白质即可，不必根据某种指标逐个计算，大大降低了计算量，因为就指标最高的 P 个蛋白质而言，其关键性有时在 PPI 网络中的整体程度未必最高，可能只是某一局部顶点的代表，特别是一些采用局部链接指标或者是采用链接密切度逐步扩大的算法更容易导致最优解的局部性。因此，我们提出了衡量 top-p 关键蛋白质的整体性指标，而不是评估蛋白质关键性的单个指标。所提算法即 EPPSO 采用选取候选解的方法，每一个候选解含有 P 个蛋白质，整体度量这 P 个蛋白质的关键性，用这些蛋白质与其他蛋白质间联系的紧密度来衡量其关键性，即为所求的适应度函数；然后根据粒子群的算法思想，通过跟踪全局最优值及个体最优值来更新自己。为了评估 EPPSO 算法的性能，在酵母数据集等标准数据集上运行此算法，并与 DC，SC，CC，BC，EC，IC，LAC 和 NC 算法进行各项性能指标的比较，实验结果表明本文提出的 EPPSO 关键蛋白质识别算法明显优于其他识别方法。

2　方法

定义 1（PPI 网络）　PPI 网络通常被认为是一个无向图 $G(V,E)$，V 是顶点的集合，E 是边的集合；图 G 中的顶点表示一个蛋白质，边代表蛋白质之间的相互作用。

2.1　构造适应度函数

首先要提出一种对 P 个关键蛋白质进行整体评估的适应度函数，而 PPI 网络中的数据往往存在大量的噪声，因此需要衡量蛋白质之间的可靠性，即它们之间的相似性度量。本文将综合考虑蛋白质在 PPI 网络中的拓扑特征及在生物属性方面的相似度[26]，得到相互作用的两个蛋白质的相似性的最终度量，即将共同邻居相似度（NTE）、基因表达相似度（GES）、GO 语义相似度（GOS）、域交互程度（DS）、系统发育谱相似度（PPS）进行加权平均，如式（1）所示：

$$w_{ij}=\alpha_1 NTE_{ij}+\alpha_2 GES_{ij}+\alpha_3 GOS_{ij}+\alpha_4 DS_{ij}+\alpha_5 PPS_{ij} \tag{1}$$

其中，w_{ij} 表示蛋白质 i 与 j 的相似度，权重 $\alpha_i(i=1,2,3,4,5)$ 满足 $\alpha_i\in(0,1)$，$\sum_1^5\alpha_i=1$。

由“中心-致死性”法则可知，与其他蛋白质联系紧密的蛋白质是关键蛋白质的可能性较大，因此，研究蛋白质之间联系的紧密程度对于确定蛋白质的关键性非常重要。将蛋白质之间的紧密程度定义如下：

$$p_{ij}=\frac{w_{ij}}{\sum_{(i,k)\in E}w_{ik}} \tag{2}$$

其中，w_{ij} 表示蛋白质之间的相似度；p_{ij} 为蛋白质 i 与蛋白质 j 之间的转移概率或影响力，反映了蛋白质之间联系的紧密程度。

EPPSO 算法将采用选取候选解的方法，即每一个候选解含有 P 个蛋白质，然后通过计算蛋白质之间的紧密程度来整体度量这 P 个蛋白质的关键性。设顶点集合 $U=\{v_1,v_2,\cdots,v_p\}$，其中，v_i 表示一个蛋白质，则 $N_k(U)$ 表示距 U 中顶点的最短距离为 k 的顶点的集合，即：

$$N_k(U)=\{v\mid v\notin u,\exists u\in U,|l_{u,v}|=k\} \tag{3}$$

其中，$|l_{u,v}|$ 为 u 到 v 的最短路径长度，特别地，记 $N_0(U)=U$。最后，定义集合 U 的关键度即适应度函数如下：

$$ESS(U)=\sum_{k=1}^{L}\alpha_k\sum_{i\in N_k(u)}\left[1-\prod_{\substack{(i,j)\in E\\ j\in N_{k-1}(u)}}(1-p_{ij})\right] \tag{4}$$

其中，α_k 为系数，$\alpha_1,\alpha_2,\cdots,\alpha_k$ 递减，$\alpha_i\in(0,1)$。因为随着路径的增加，邻居间的紧密程度会降低，所以为不同的紧密程度附上权值，例如，可设 $\alpha_k=\alpha^k(\alpha\in(0,1))$。定义的关键度实际

上是对 U 中所有顶点的 1 至 k 阶邻居顶点的影响力进行综合评价，在实际计算中取 L 为 2 或 3 即可。

2.2 EPPSO 算法

在 PSO 算法中，首先初始化一群随机个体(粒子)，然后通过迭代找到最优解，在每一次迭代中，粒子通过跟踪两个最优值(全局最优值及个体最优值)来更新其速度与位置。设在 D 维目标搜索空间中，由 m 个粒子构成种群，其中第 i 个粒子在第 d 维的位置为 X_{id}，其飞行速度为 V_{id}，该粒子当前搜索到的最优位置为 P_{id}，整个粒子群当前的最优位置为 P_{gd}，则在 t 时刻到 $t+1$ 时刻，粒子的更新公式如下：

$$V_{id}^{(t+1)}=wV_{id}^{(t)}+c_1r_1(P_{id}-X_{id}^{(t)})+c_2r_2(P_{gd}-X_{id}^{(t)}) \quad (5)$$

$$X_{id}^{(t+1)}=X_{id}^{(t)}+\alpha V_{id}^{(t+1)} \quad (6)$$

把粒子群优化算法应用到关键蛋白质的识别问题中，将适应度函数 ESS 作为目标函数进行优化。设种群中每个个体都是由 P 个蛋白质构成的集合，用一个 P 维的向量来表示，例如 $x=(4,5,3,7,6)$ (设 $P=5$)表示第 4,5,3,7,6 的蛋白质被选为关键蛋白质。对于蛋白质而言，利用式(5)、式(6)不便对其进行运算，因此将经典粒子群式(5)、式(6)分别改为如下形式：

$$V_{id}^{(t+1)}=G[wV_{id}^{(t)},c_1F(r_1,P_i,X_{id}^{(t)}),c_2F(r_2,P_g,X_{id}^{(t)})] \quad (7)$$

$$X_{id}^{(t+1)}=X_{id}^{(t)}\oplus\alpha V_{id}^{(t+1)} \quad (8)$$

其中，P_i 为第 i 个粒子的部分最优值，P_g 为全局最优值式中的函数。将 G，F 以及运算⊕分别定义如下。

(1)函数 $F(r,P,X_d)$。其中，r 为随机阈值；P 为一个粒子，表示一个部分(或全局)最优解；X_d 为个体的第 d 个分量。函数 F 的功能是判别 X_d 是否出现在 P 中，如果出现，函数值为 0；否则，以 r 的概率选取 P 中一个元素作为函数值，以 $1-r$ 的概率选取其他的任意元素作为函数值。设 P 中的元素分为 P_y 与 P_n 两部分：$P_y=\{S|S\in P$ 同时 $S\in X\}$ 为 P 中同时属于 X 的蛋白质，$P_n=\{S|S\in P$ 同时 $S\notin X\}$ 为 P 中不属于 X 的蛋白质。

F 的功能可由算法 1 进行描述。

算法 1 $F(r,P,X_d)$

```
Begin
  if Xd∈Py then F=0;
  else 产生随机数 r′∈(0,1);
    if r′<r then
      在 Pn 中随机选取 S;
      F=S;
      Pn=Pn−{S};
    else
      在所有蛋白质中随机选取一个 S∉X;
      F=S;
    end if
  end if
End
```

(2)函数 $G(wV_1,c_1F_1,c_2F_2)$。其中，V_1 为粒子速度，F_1 为局部最优值，F_2 为全局最优值，w，c_1，c_2 是各自所占的权重。函数 G 的功能是综合考虑这 3 个值，以判断粒子飞行的方向。我们用赌轮的方式选取 V_1 或 F_1 或 F_2 作为函数值，设 $q=w+c_1+c_2$，且

$$\lambda_1=\frac{w}{q},\lambda_2=\frac{c_1}{q},\lambda_3=\frac{c_2}{q} \quad (9)$$

则 V_1，F_1，F_2 被选中的概率分别为 λ_1，λ_2，λ_3。

函数 G 可以用算法 2 描述。

算法 2 $G(wV_1,c_1F_1,c_2F_2)$

```
Begin
  按式(9)计算 λ1,λ2,λ3;
  产生随机数 r∈(0,1);
  if r<λ1 then G=V1;
  else if λ1<r≤λ1+λ2 then G=F1;
  else G=F2;
  end if
End
```

(3) $X_{id}^{(t)}\oplus\alpha V_{id}^{(t+1)}$。若 $V_{id}^{(t+1)}=0$，则该运算结果取值 $X_{id}^{(t)}$，否则以 α 的概率取值 $V_{id}^{(t+1)}$，以 $1-\alpha$ 的概率随机取一蛋白质 $S(S\notin X)$作为结果值。其基本思想是，如果 $V_{id}^{(t+1)}=0$，说明 X_{id} 不需要交换，它在很大程度上与最优解的总数一致；否则，以 α 的概率换成最优解元素或以 $1-\alpha$ 的概率另外选取一个元素替换。

运算⊕可用算法 3 描述。

算法 3 运算⊕

```
输入:X_id^(t)(Xi 的第 d 个分量),α(控制参数),V_id^(t+1)(Xid的速度)
输出:X_id^(t+1)(X_id^(t+1)的更新值)
Begin
  if V_id^(t+1)=0 then X_id^(t+1)=X_id^(t);
  else 产生随机数 r∈(0,1);
    if r<λ then X_id^(t+1)=V_id^(t+1);
    else
      在所有蛋白质中任取一 S(S∉Xi);
      X_id^(t+1)=S;
    end if
  end if
End
```

我们同时考虑了网络拓扑特征和节点生物信息，并基于 GO 数据信息、基因表达数据信息、系统发育谱数据信息等 5 种信息来计算蛋白质间相互作用的可信度，构造质量更高的加权网络，边权表示蛋白质节点间相互作用的强度。除了可信度，根据“中心-致死性”法则，还考虑了蛋白质节点间联系的紧密程度来衡量蛋白质的关键性。在加权网络上考虑节点和边的双重因素，扩展节点局部网络区域至二阶邻居，从而提出了新的整合网络拓扑和生物多源数据的 EPPSO 算法来识别关键蛋白质节点。算法的总体流程图如图 1 所示，具体分为以下几个步骤。

(1)根据 PPI 数据集构建 PPI 网络图 G 的邻接矩阵 A。若蛋白质节点间有边相连，则对应矩阵元素值为 1；若无边直接相连，则对应矩阵元素值为 0。

(2)根据 GO 数据、基因表达数据、域信息、系统发育谱数据计算蛋白质对的可信度 w，构建加权 PPI 网络的可信度矩阵 B。

(3)找出每个顶点 v 的一阶邻居和二阶邻居，即可计算蛋

白质集合的适应度值，即 ESS 的值。

(4)对于初始种群的产生，为了缩小搜索关键蛋白质的范围，取前 1.5×P 个度较大的顶点来产生初始种群，P 为需要识别的蛋白质个数，即每个个体的长度，也可以选取其他指标来缩小范围，如 LAC 等。然后初始化全局最优值和局部最优值，根据式(7)、式(8)来更新每个个体的第 i 分量，每个个体的每个分量更新结束后，计算全局最优值和局部最优值，直到全局最优值收敛，停止迭代。

(5)得到的全局最优值即为所需的关键蛋白质。

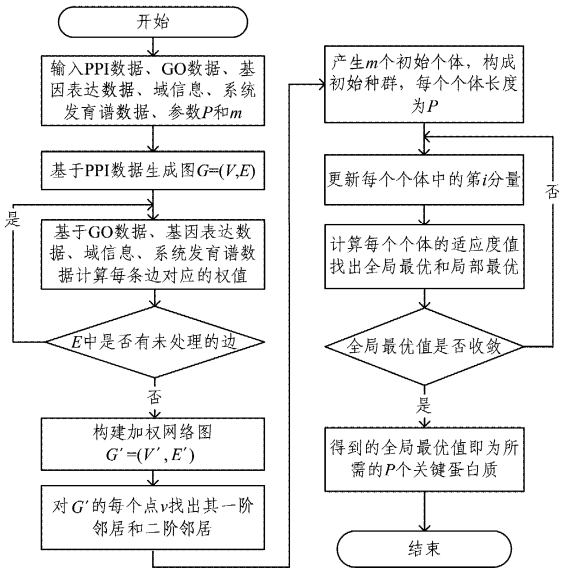


图 1　EPPSO 算法流程图

综上所述，EPPSO 算法的基本框架如算法 4 所示。

算法 4　EPPSO(Essential Protein PSO)算法

输入：V(蛋白质集合)，P(关键蛋白质个数)

输出：u* (P 个关键蛋白质)

Begin

1. n′=1.5×P

2. 保留 n′个度较大的顶点(蛋白质)

3. Initilization(V,W,P,m)；P_g=Ø；

　/* 产生 m 个初始个体，构成初始种群 W，每个个体的长度为 P */

　/* 设 W={X_1,X_2,…,X_m}，X_i=(x_{i1},x_{i2},…,x_{ip}) */

4. for i=1 to m do

　　if ESS(X_i)>ESS(P_g) then P_g=X_i；(全局最优)

　　P_i=X_i；(局部最优)

　end for

5. while not convergence do

　　产生随机数 r_1，r_2∈(0,1)；

　　for i=1 to m do

　　for d=1 to p do

　　按式(7)、式(8)更新 x_{id}；

　　end for d

　　if ESS(X_i)>ESS(P_g) then P_g=X_i；(全局最优)

　　if ESS(X_i)>ESS(P_i) then P_i=X_i；(局部最优)

　　end for i

　end while

6. u* =P_g

End

2.3　时间复杂度分析

设算法的迭代过程进行 c 次后结束，在每次迭代过程中，内层循环的复杂度皆为 O(mp)，其中 m 为初始个体，每个个体的长度为 p。在内层循环中，可对每个个体计算适应度值，并按式(4)计算 $ESS(U)$，设图中顶点最大度为 d_{max}，则最多共有$d_{max}+d_{max}^2+\cdots+d_{max}^L=O(d_{max}^{L+1})$个节点，因此复杂度为 $O(d_{max}^{L+1})$，通常取 $L=1$ 或 2。因此，每次迭代的复杂度为 $O(mpd_{max}^{L+1})$，整个算法的复杂度为 $O(cmpd_{max}^{L+1})$。

3　实验结果与分析

3.1　实验数据来源

本文方法主要包括以下几种数据：PPI 数据、基因表达数据、GO 数据、域信息数据、系统发育谱数据和关键蛋白质数据。

文中选择酵母数据进行实验，因为它的 PPI 数据和关键蛋白质数据是最完整和可靠的。PPI 数据从 DIP 数据库下载[27]，在删除自连接和重复连接之后，PPI 网络由 5160 个顶点和 22579 条边构成。酵母的基因表达数据从 GEO[28]数据库搜集，我们选择集合 GSE3431[29]，其是从 3 个连续的代谢周期搜集的，且每个周期包括 12 个时间点，该数据集包括 36 个样本和 6777 个基因。GO[30]数据包括 94181 条注释。域信息[31]数据从 DOMINE 数据库下载。系统发育谱数据[32]从 InParanoid 数据库下载。标准的关键蛋白质数据从以下 4 个数据库搜集：MIPS[33]，SGD[34]，DEG[35]和 SGDP[36]，共搜集了 1011 个关键蛋白质。

3.2　结果分析

EPPSO 算法的候选集是从度中心性指标中选出的，除此之外，也可以选择其他中心性指标，因此，将其与 DC，SC，CC，BC，EC，IC，LAC，NC 等中心性方法进行比较，所提方法名为 EPPSO_DC，若选择了 LAC，则方法名为 EPPSO_LAC。根据选择指标的不同，结果也将不同。

3.2.1　预留蛋白质中关键蛋白质的数量

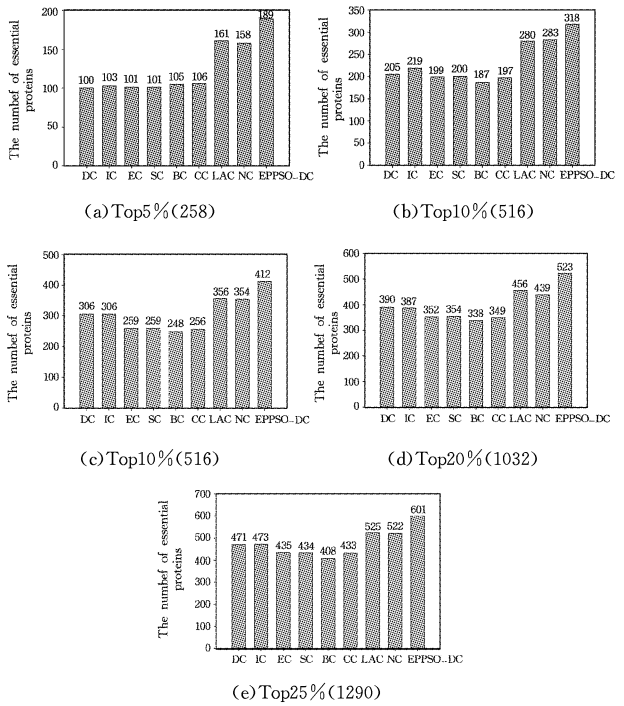
EPPSO_DC 算法选择度中心性指标作为候选集。该算法将计算前 5%，10%，15%，20%，25%的蛋白质中识别的关键蛋白质数量，因此表 1 列出了各比例下预留蛋白质中的关键蛋白质的数量。若选择了其他指标，则预留个数中的关键蛋白质的数量不同。

表 1　预留蛋白质中的关键蛋白质数量

	5%	10%	15%	20%	25%
个数	258	516	774	1032	1290
预留个数(1.5×P)	387	774	1161	1548	1935
关键蛋白质个数	150	306	433	548	645

3.2.2　不同比例的关键蛋白质预测准确率的比较

为了评估 EPPSO 算法的性能，首先在酵母数据集上分别实现了 EPPSO_DC，DC，SC，CC，BC，EC，IC，LAC 和 NC 算法，并分别选择前 5%，10%，15%，20%和 25%的蛋白质作为结果候选集。对于每一比例的候选集，再将其与真实关键蛋白质数据集求交集，得出候选关键蛋白质集中真实关键蛋白质的数量，实验结果如图 2 所示。



(a)Top5%(258) (b)Top10%(516)

(c)Top10%(516) (d)Top20%(1032)

(e)Top25%(1290)

图 2 EPPSO_DC 与其他算法识别出的关键蛋白质数量的比较

从图 2 可以看出,EPPSO_DC 算法要优于其他算法,特别在前 5% 和 10% 情况下,预测准确率达到了 73.3% 和 61.6%;在所有方法中,BC 预测的准确率相对较低,在前 5%,10%,15%,20%和 25%情况下,EPPSO_DC 比 BC 分别提高了 80.0%,70.1%,66.1%,54.7%和 47.3%,即使与最好的结果相比,EPPSO_DC 也分别提高了 19.6%,12.4%,15.7%,14.7%,15.4%。从表 1 和图 2 中还可以看出,EPPSO_DC 算法排名前 5%和 10%预测的关键蛋白质数量要高于预留中(如表 1 所列,即高于 DC 算法识别出的蛋白质数量的 1.5 倍)的关键蛋白质数量,这就说明虽然有的蛋白质相互作用少,但是它与关键蛋白质之间的紧密程度却很高。因此,EPPSO_DC 算法克服了 DC 算法的缺点,能够识别出度较少的关键蛋白质。

3.2.3 统计指标分析

为了进一步评价本文算法的性能,使用以下 6 个统计指标,即敏感度(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),来对 EPPSO 算法与其他 8 种算法进行比较。这些统计指标的定义如下:

(1)$SN=TP/(TP+FN)$,表示关键蛋白质被正确识别的比例。

(2)$SP=TN/(TN+FP)$,表示非关键蛋白质被正确排除的比例。

(3)$PPV=TP/(TP+FP)$,表示候选集中蛋白质被正确预测为关键蛋白质的比例。

(4)$NPV=TN/(TN+FN)$,表示排除的蛋白质中被正确预测为非关键蛋白质的比例。

(5)$F=2*SN*PPV/(SN+PPV)$,表示敏感度和阳性预测值的调和平均值。

(6)$ACC=(TP+TN)/(TP+TN+FP+FN)$,表示所有识别结果中正确结果的比例。

其中,TP(True Positives,真阳性)表示关键蛋白质被正确预测的数目;FN(False Negatives,假阴性)表示关键蛋白质被错误预测的数目;TN(True Negatives,真阴性)为非关键蛋白质被正确预测的数目;FP(False Positives,假阳性)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计算结果如表 2 所列。

表 2 EPPSO_DC 方法与其他 8 种方法的统计指标比较

	SN	SP	PPV	NPV	F	ACC
DC	0.405	0.809	0.410	0.806	0.408	0.664
IC	0.389	0.808	0.391	0.798	0.389	0.651
EC	0.360	0.796	0.370	0.796	0.360	0.645
SC	0.370	0.799	0.379	0.794	0.374	0.650
BC	0.403	0.806	0.398	0.800	0.399	0.655
CC	0.372	0.797	0.373	0.792	0.365	0.649
NC	0.443	0.822	0.446	0.810	0.440	0.681
LAC	0.452	0.823	0.455	0.818	0.455	0.683
EPPSO_DC	0.498	0.842	0.500	0.836	0.503	0.702

从表 2 中可以清楚地看出,在 SN,SP,PPV,NPV,F 和 ACC 这 6 种统计指标上,EPPSO_DC 方法都比其他 8 种方法的精度更高,这说明提出的方法能更准确、有效地识别出关键蛋白质。

3.2.4 识别关键蛋白质的差异性分析

为了进一步验证 EPPSO_DC 算法所具有的优势,分析其与最具代表性的中心性测度方法 DC 在预测小部分蛋白质上的关联和差异(以前 100 的蛋白质为例),结果如表 3 和表 4 所列。通过表 3 和表 4 可以发现:一方面,在 EPPSO_DC 成功识别出的 100 个关键蛋白质中,有 16 个不能够被 DC 算法成功识别,这 16 个蛋白质如表 3 所列,从中可以看出有 8 个蛋白质的度不大于 10,在这 8 个蛋白质中,6 个是关键的,2 个是非关键的,所有 16 个蛋白质中有 75%是关键的。以上说明,与经典算法相比,本文算法对度不高的关键蛋白质的识别具有较强的灵敏性。

表 3 通过 EPPSO_DC 算法识别的 100 个蛋白质中 DC 不包含的集合

Protein name	Essentiality	DC
YJL203W	essential	21
YDL030W	essential	24
YNR016C	essential	9
YHR023W	essential	11
YOR122C	essential	7
YNR007C	non-essential	7
YPR082C	essential	8
YKR037C	essential	6
YKL203C	essential	8
YKL014C	essential	10
YAL047C	non-essential	15
YCR079W	non-essential	26
YGR144W	non-essential	7
YLR268W	essential	39
YBR155W	essential	15
YGR113W	essential	17

表 4　通过 DC 算法识别的 100 个蛋白质中 EPPSO_DC 不包含的集合

Protein name	Essentiality	DC
YBR017C	non-essential	96
YER081W	non-essential	94
YOL055C	non-essential	93
YBR217W	non-essential	91
YLR180W	non-essential	90
YER177W	non-essential	86
YKR026C	non-essential	83
YLR372W	non-essential	79
YLR453C	non-essential	79
YDL047W	non-essential	77
YGR040W	non-essential	76
YHR030C	non-essential	76
YIL035C	non-essential	76
YHR135C	non-essential	75
YDR099W	non-essential	71
YLR423C	non-essential	70
YPL240C	non-essential	70
YGL127C	non-essential	69
YCR034W	non-essential	66
YMR205C	non-essential	65
YDL213C	non-essential	64
YLR044C	non-essential	64
YDR034C	non-essential	63
YOR212W	non-essential	59
YJL088W	non-essential	58
YDR386W	non-essential	56

另一方面，表 4 列出了 DC 方法成功识别的 100 个关键蛋白质中，被 EPPSO_DC 成功过滤掉的 26 个非关键蛋白。从表中可以看出，这些非关键蛋白质的度非常高，但通过 DC 方法很难发现这些蛋白质是非关键的，一般来说，DC 的值越高，EPPSO_DC 越容易识别出是关键的，反之亦然。但正如表 4 所列，对于一些假阳性数据，通过 EPPSO_DC 方法可以成功过滤，但是 DC 却无法做到。

根据上述分析可知，关键蛋白质之间联系的紧密程度不仅仅体现在拓扑结构上，还与它们之间的生物属性信息密切相关。通过表 3 和表 4 可以发现，度少的节点不一定是非关键蛋白；反之，度高的节点也不一定必是关键蛋白。因此，我们通过综合网络拓扑结构和生物属性信息的识别方法，大大提高了预测准确率，同时较好地将粒子群算法应用到该问题中，避免了对每个蛋白质的某个指标值进行逐一计算，节省了计算时间，提高了效率。

结束语　关键蛋白质的识别不仅对帮助了解细胞新陈代谢、生长发育等生命活动过程非常重要，而且在疾病研究和药物设计等方面具有重大的应用价值。本文在综合考虑网络拓扑特性和生物属性信息的基础上，提出了一个高效的基于改进 PSO 的关键蛋白质的识别新方法 EPPSO。在该方法中，首先，融合了如 GO 数据、基因表达数据等多源生物信息数据来计算蛋白质间相互作用的可信度，构造了质量更高的加权网络；除了可信度，根据“中心-致死性”法则，还考虑了蛋白质节点间联系的紧密程度，以衡量蛋白质的关键性，并扩展局部网络拓扑至二阶邻居，因此大大提高了预测的准确率；另外，还将 PSO 算法应用到识别问题中对其进行优化，所提算法可以整体度量 P 个蛋白质的关键性而不必逐一计算某个蛋白质的某个指标，大大减少了计算量，提高了算法性能。在 DIP 和 GEO 等国际公认的标准数据上进行实验，结果表明，与其他经典的中心性方法相比，无论是在预测准确率还是统计指标分析方面，本文算法 EPPSO 都具有更好的性能；同时，本文方法在识别度少的关键蛋白与过滤度高的非关键蛋白时也具有一定的优势。

参 考 文 献

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

[2] CLATWORTHY A E, PIERSON E, HUNG D T. Targeting virulence: a new paradigm for antimicrobial therapy[J]. China Animal Husbandry & Veterinary Medicine, 2007, 3(9): 541-548.

[3] GIAEVER G, CHU A M, NI L, et al. Functional profiling of the Saccharomyces cerevisiae genome[J]. Nature, 2002, 418(6896): 387-391.

[4] CULLEN LM, ARNDT G M. Genome-wide screening for gene function using RNAi in mammalian cells[J]. Immunology & Cell Biology, 2005, 83(3): 217-223.

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

[8] BATADA N N, HURST L D, TYERS M. Evolutionary and physiological importance of hub proteins[J]. Plos Computational Biology, 2006, 2(7): e88.

[9] VALLABHAJOSYULA R R, CHAKRAVARTI D, LUTFEALI S, et al. Identifying hubs in protein interaction networks[J]. PLoS One, 2009, 4(4): e5344.

[10] ESTRADE E. Virtual identification of essential proteins within the protein interaction network of yeast[J]. Proteomics, 2006, 6(1): 35-40.

[11] FREEMAN L C. A set of measures of centrality based on betweenness[J]. Sociometry, 1977, 40(1): 35-41.

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

[13] ESTRAD E, RODRÍGUEZ-VELÁZQUEZ J A. Subgraph centrality in complex networks[J]. Phys. Rev. E Stat. Nonlin. & Soft. Matter. Phys., 2005, 71(2): 056103.

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

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

[16] LI M, WANG J, 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.

[17] WANG J, LI M, WANG H, et al. Identification of essential proteins based on edge clustering coefficient [J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2012, 9(4): 1070.

[18] LI M, WANG J, WANG H, et al. Essential proteins discovery from weighted protein interaction networks[J]. Bioinformatics Research & Applications, 2010, 6053(3): 89-100.

[19] HE X, ZHANG J. Why do hubs tend to be essential in protein

networks?[J]. PLoS Genet,2006,2(6):826-834.
[20] ZOTENKO E,MESTRE J,O'LEARY D P,et al. Why do hubs in the yeast protein interaction network tend to be essential:re-examining the connection between the network topology and essentiality[J]. Plos Computational Biology,2008,4(8):e1000140.
[21] CHUA H N,TEW K L,LI X L,et al. A unified scoring scheme for detecting essential proteins in protein interaction networks [C]//Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence(ICTAI'08). USA:IEEE Computer Society,2008:66-73.
[22] 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.
[23] ZHANG X,XU J,XIAO W X. A new method for the discovery of essential proteins[J]. PloS One,2013,8(3):e58763.
[24] LUO J,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.
[25] LI M,ZHENG R,ZHANG H,et al. Effective identification of essential proteins based on priori knowledge,network topology and gene expressions[J]. Methods,2014,67(3):325-333.
[26] JIANG Y,WANG Y,PANG W,et al. Essential protein identification based on essential protein-protein interaction prediction by Integrated Edge Weights[J]. Methods,2015,83:51-62.
[27] http://dip.doe-mbi.ucla.edu/dip/Download.cgi?SM=7.
[28] http://www.ncbi.nlm.nih.gov/geo.
[29] TU B P,KUDLICKI A,ROWICKA M,et al. Logic of the Yeast Metabolic Cycle:Temporal Compartmentalization of Cellular Processes[J]. Science,2005,310:1152-1158.
[30] ASHBURNER M,BALL C A,BLAKE J A,et al. Gene ontology:tool for the unification of biology. The Gene Ontology Consortium[J]. Nat. Genet.,2000,25(1):25-29.
[31] YELLABOINA S,TASNEEM A,ZAYKIN D V,et al. DOMINE:a comprehensive collection of known and predicted domain-domain interactions[J]. Nucleic Acids Res.,2011,39(supp_1):730-735.
[32] O'BRIEN K P,REMM M,SPMMHAMMER E L. Inparanoid:a comprehensive database of eukaryotic orthologs[J]. Nucleic Acids Res.,2005,33:D476-480.
[33] MEWES H W,AMID C,ARNOLD R,et al. MIPS:analysis and annotation of proteins from whole genomes[J]. Nucleic Acids Res.,2004,32(Suppl 1):41-44.
[34] CHERRY J M,ADLER C,BALL C,et al. SGD:Saccharomyces genome database[J]. Nucleic Acids Res.,1998,26(1):73-79.
[35] ZHANG R,LIN Y. DEG 5.0,a database of essential genes in both prokaryotes and eukaryotes[J]. Nucleic Acids Res.,2009,37(Suppl. 1):D455-D458.
[36] http://www.sequence.stanford.edu/group.

(上接第 18 页)

[2] ZHAO Y,LI Y,RAICU I,et al. Enabling scalable scientific workflow management in the Cloud[J]. Future Generation Computer Systems,2015,46(c):3-16.
[3] RAMNARAYAN J,MOZAFARI B,WALE S,et al. SnappyData:A Hybrid Transactional Analytical Store Built on Spark [C]//International Conference on Management of Data. ACM,2016:2153-2156.
[4] YUAN H,LI C,DU M. Optimal Virtual Machine Resources Scheduling Based on Improved Particle Swarm Optimization in Cloud Computing[J]. Journal of Software,2014,9(3):705.
[5] HUANG Q,SHUANG K,XU P,et al. Prediction-based Dynamic Resource Scheduling for Virtualized Cloud Systems[J]. Journal of Networks,2014,9(2):375-383.
[6] ACETO G,BOTTA A,DONATO W D,et al. Cloud monitoring:A survey[J]. Computer Networks,2013,57(9):2093-2115.
[7] CS S,S B M S. A Comparative Analysis of Scheduling Policies in Cloud Computing Environment[J]. International Journal of Computer Applications,2013,67(20):16-24.
[8] ZHANG Q,CHEN H,SHEN Y,et al. Optimization of virtual resource management for cloud applications to cope with traffic burst[J]. Future Generation Computer Systems,2016,58:42-55.
[9] ZHENG Q,LI R,LI X,et al. Virtual machine consolidated placement based on multi-objective biogeography-based optimization[J]. Future Generation Computer Systems,2016,54(C):95-122.
[10] LIU Z,ZHOU H,FU S,et al. Algorithm Optimization of Resources Scheduling Based on Cloud Computing[J]. Journal of Multimedia,2014,9(7):1451-1456.
[11] SHAO Y. Virtual Resource Allocation based on Improved Particle Swarm Optimization in Cloud Computing Environment[J]. International Journal of Grid & Distributed Computing,2015,8(1):228-233.
[12] HASSAN M M,ALAMRI A. Virtual Machine Resource Allocation for Multimedia Cloud:A Nash Bargaining Approach[J]. Procedia Computer Science,2014,34:571-576.
[13] SINGH S,CHANA I. Q-aware:Quality of service based cloud resource provisioning[J]. Computers & Electrical Engineering,2015,47:138-160.
[14] SALAH K,ELBADAWI K,BOUTABA R. An Analytical Model for Estimating Cloud Resources of Elastic Services[J]. Journal of Network & Systems Management,2016,24(2):285-308.
[15] SHYAM G K,MANVI S S. Virtual resource prediction in cloud environment:A Bayesian approach[J]. Journal of Network & Computer Applications,2016,65(C):144-154.
[16] HASSANI H,SILVA E S. Forecasting with Big Data:A Review [J]. Annals of Data Science,2015,2(1):5-19.
[17] ALTINTAS N,TRICK M. A data mining approach to forecast behavior[J]. Annals of Operations Research,2014,216(1):3-22.
[18] HANSUN S. A new approach of moving average method in time series analysis[C]//New Media Studies. IEEE,2013:1-4.
[19] https://en.wikipedia.org/wiki/Moving_average.
[20] https://en.wikipedia.org/wiki/Polynomial_regression.
[21] LI J,SHEN L,TONG Y. Prediction of Network Flow Based on Wavelet Analysis and ARIMA Model[C]// International Conference on Wireless Networks and Information Systems. IEEE Computer Society,2009:217-220.
[22] https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average.