计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2014.20130538
51(10):2178-2186,2014

一种蛋白质复合体模块度函数及其识别算法

郭茂祖　代启国　徐立秋　刘晓燕
(哈尔滨工业大学计算机科学与技术学院　哈尔滨　150001)
(maozuguo@hit.edu.cn)

On Protein Complexes Identifying Algorithm Based on the Novel Modularity Function

Guo Maozu, Dai Qiguo, Xu Liqiu, and Liu Xiaoyan
(*School of Computer Science and Technology, Harbin Institute of Technology, Harbin* 150001)

Abstract　Proteins often interact with each other to form complexes. It is very significant for understanding the activities in cell to carry out their biological functions. In recent years, with the rapid development of new biological experiment technologies, a large amount of protein-protein interaction (PPI) networks are generated. Identifying protein complexes by clustering proteins in PPI networks is hot spot in current bioinformatics research. Many clustering methods, which are mainly based on graph partition or the technologies of community detection in social network, have been proposed to recognize the protein complexes in PPI networks in last decade. However, the performances of most of previous developed detecting methods are not ideal. They cannot identify the overlapping complexes, but according to the biological study found, protein complexes are often overlapping. Therefore, in this paper, a protein complexes modularity function (Q function), namely PQ function, is proposed to identify the overlapping complexes from PPI networks. Based on PQ, a new algorithm for identifying protein complexes BMM (the algorithm based on protein complexes modularity function for merging modules). Firstly, BMM algorithm finds some dense sub-graphs as initial modules. Then, these initial modules are merged by maximizing the modularity function PQ. Finally, several high-quality protein complexes are found. Comparing these protein complexes with two known protein complexes datasets, the results suggest that the performance of BMM is excellent. In addition, compared with other latest algorithms, BMM is more accurate.

Key words　protein complex; protein-protein interaction (PPI); protein complexes modularity function; initial module; based on protein complexes modularity function for merging modules (BMM) algorithm

摘　要　蛋白质复合体对于研究细胞活动具有重要意义.随着新的生物实验技术的不断出现,产生了大量的蛋白质相互作用网络.通过对蛋白质相互作用网络进行聚类识别蛋白质复合体是当前研究热点.然而,目前大多数蛋白质复合体识别算法的性能不够理想.为此,提出了蛋白质复合体模块度函数(PQ),并在此基础上提出了基于蛋白质复合体模块度函数的模块合并(based on protein complexes modularity function for merging modules, BMM)算法.BMM 算法首先识别网络中一些稠密子图作为初始模块,然后依据 PQ 函数对这些初始模块进行合并,最终得到了质量较高的蛋白质复合体.将识别出的复合体分别与 2 种已知的蛋白质复合体数据集进行比对,结果表明 BMM 算法具有很好的识别性能.此外,与其他最新的识别算法相比,BMM 算法的识别准确率较高.

收稿日期:2013-04-07;**修回日期**:2013-10-21
基金项目:国家自然科学基金项目(91335112,61271346,61172098);高等学校博士学科点专项科研基金项目(20112302110040)

关键词 蛋白质复合体;蛋白质相互作用;蛋白质复合体模块度函数;初始模块;BMM 算法

中图法分类号 TP18; TP391; TP3-05

蛋白质作为生物活动的物质基础越来越受关注,近年来蛋白质组学领域的研究成果层出不穷[1].生物学家通过观察细胞活动得出结论:在同一个细胞生命周期中蛋白质很少单独行使功能;蛋白质之间存在着广泛的相互作用,并且这种相互作用在很多生物细胞中都存在[2].

最近几年出现了很多先进的生物实验技术,包括酵母-双杂交法(yeast-two-hybrid)、质谱分析法(mass spectrometry)、蛋白质切片技术(protein chip technologies)等方法[3].生物学家通过这些实验技术,验证了一些蛋白质之间的相互作用关系.大量蛋白质相互作用形成了蛋白质相互作用网络,即PPI(protein-protein interaction)网络.PPI 网络是理解生物系统中蛋白质活动的重要途径,因而是当前蛋白质组学研究的热点.

研究 PPI 网络的目的之一是得到网络中连接紧密的一些模块.这些模块具有 2 种特殊的生物学意义:一种是功能模块,另一种是蛋白质复合体.功能模块是由参与同一个特定分子过程的蛋白质结合而成,但是这种结合在时间和空间位置上不同;蛋白质复合体则是指在相同时间和空间一组相互作用的蛋白质组成的多分子机制[4].然而,通过生物实验验证蛋白质复合体比较困难,主要体现在生物实验成本高,实验受环境条件影响较大,实验结果假阳性较高[5].通过计算机算法识别复合体,可以为生物实验提供指导和帮助.因此,近年来出现了大量的蛋白质复合体识别算法.其中,基于模块度函数[6]的 PPI 网络聚类是一种比较好的方法.因为,模块度函数充分考虑了 PPI 网络的节点分布情况,通过优化该函数,可以将连接紧密的节点聚到相同的模块中,符合蛋白质复合体结构特征.

不同的蛋白质复合体可能包含相同的蛋白质,即蛋白质复合体之间有重叠.然而,传统的模块度 Q 函数无法识别具有重叠特性的蛋白质复合体.所以,本文提出了针对蛋白质复合体识别的模块度函数(PQ),并在此基础上,提出了基于 PQ 函数的模块合并(based on protein complexes modularity function for merging modules, BMM)算法.为了测试 BMM 算法的性能,我们分别利用 BMM 算法、基于重叠邻居扩张的聚类(clustering with overlapping neighborhood expansion, ClusterONE)算法和重叠聚类生成器(overlapping cluster generator, OCG)算法对公开的酵母 PPI 网络进行识别.将它们识别出的蛋白质复合体与 2 种已知的蛋白质复合体数据集进行了比对,结果表明 BMM 算法识别效果优于另外 2 种算法.

1 相关工作

1.1 PPI 网络结构分析

PPI 网络一般由无向图 $G(V, E)$ 表示,其中,V 表示无向图的节点集合,即每个节点表示一个蛋白质;E 表示无向图中边的集合,即蛋白质之间的相互作用关系[7].

对 PPI 网络拓扑结构的研究表明,PPI 网络是典型的无标度网络,也是“小世界网络”的一种,PPI 网络的节点的度分布符合幂律分布的特性[8].即在网络中,度较高的节点数目较少,而度较低的节点数目较多.PPI 网络结构比较稀疏,与其他复杂网络存在结构差异[9].这说明针对其他复杂网络的模块识别算法不适合蛋白质复合体识别.

1.2 蛋白质复合体识别

近几年出现了很多关于蛋白质复合体识别的算法.这些算法主要是基于图分割的方法,如 Markov 聚类算法(MCL)、分子复合物识别算法(MCODE)等[10].这些算法都属于严格的图分割方法,即每一个节点只能属于一个模块,不允许模块间重叠[11].然而,在 PPI 网络中一个蛋白质可能参与多个蛋白质复合体的合成.因此,上述这些方法不能识别重叠蛋白质复合体.

当前,关于识别重叠蛋白质复合体的研究已经有了一些成果.2012 年 Nepusz 等人提出了 ClusterONE 算法.算法的主要思想是:按照节点度由大到小,依次选择种子节点作为聚类核心,然后以“聚类系数”决定其他节点和种子节点是否聚类[12].但是,ClusterONE 算法识别精度较低.

此外,2012 年 Becker 提出了基于层次聚类的 OCG 算法.该算法使用改进的模块度函数作为聚类标准,并以极大团(maximal clique)为核心进行层次聚类[13].然而,OCG 算法不适合识别规模较小的蛋白质复合体,并且该算法的识别精度较低.

1.3 模块度函数

模块度函数是一种用来衡量网络中模块划分质量的函数[14]. 2004 年 Newman 和 Girvan[6] 首次提出了模块度函数(Q 函数). 该函数是基于"随机网络不具有明显的模块结构"的思想,通过比较实际覆盖度(覆盖度是模块内部连接数占总连接数的比例)与随机连接情况下覆盖度的差异来评价模块划分质量的[15]. 该函数定义如下:

$$Q(C)=\frac{1}{2m}\sum_i\sum_j\left(A_{ij}-\frac{k_ik_j}{2m}\right)\delta(c_i,c_j),\quad(1)$$

式(1)中,网络中的全部边数 $m=20$;A_{ij} 表示节点 i 与节点 j 邻接关系;$A_{12}=1$ 表示节点 1 和节点 2 邻接;$A_{15}=0$ 表示节点 1 和 5 不邻接;$k_2=5$ 表示节点 2 的度为 5. 图 1 中标有不同形状的节点分别属于模块 a,b,c,即 $c_1=a$ 表示节点 1 属于模块 a. $\delta(c_1,c_3)=1$ 表示节点 1 与节点 3 属于同一个模块;$\delta(c_1,c_5)=0$ 表示节点 1 和节点 5 不属于一个模块.

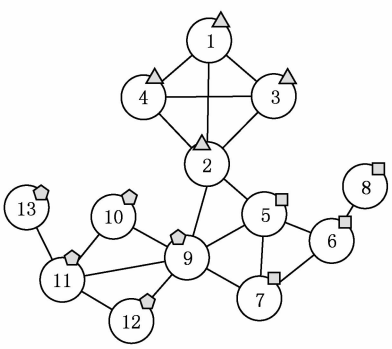

Fig. 1 Non-overlapped modules in a network.

图 1 网络中的非重叠模块

Q 函数具有 2 个重要的性质:1)当所有节点属于一个模块时,Q 值为 0,表明这种模块划分并不合理;2) Q 值越大,表明网络的模块划分质量越好[16]. 通常,Q 值在 0～1 之间,一般以 $Q=0.3$ 作为网络模块结构质量优劣的下限[17].

Q 函数在蛋白质复合体识别方面存在以下不足:

1) 由于蛋白质复合体具有重叠性,而 Q 函数不能处理模块重叠的情况,所以不适合识别复合体;

2) Q 函数还存在分辨率限制(resolution limit)[18]. 所谓分辨率限制是指:基于 Q 函数的模块识别算法不能识别包含节点较少的模块. 然而,通过对蛋白质复合体的结构分析,发现多数复合体规模较小.

为了弥补 Q 函数存在的不足,本文详细分析了蛋白质复合体的结构特点;并针对分析结果,提出了适合蛋白质复合体识别的模块度函数.

2 蛋白质复合体数据分析

酵母作为模式生物,其蛋白质复合体数据比较丰富. 所以,通常将慕尼黑蛋白质序列信息中心(Munich Information Center for Protein Sequences, MIPS)数据库提供的酵母蛋白质复合体作为评价复合体识别算法性能的"金标准". 然而,随着对蛋白质复合体研究的不断深入,MIPS 数据集已经不能反映现阶段的领域知识[19]. 为了弥补 MIPS 数据集的不足,2008 年 Pu 等人[19] 通过整理文献中的数据,发布了酵母蛋白质复合体数据集 CYC2008. CYC2008 数据集中提供了更多的蛋白质复合体数据.

通过对 CYC2008 数据集进行分析,我们发现,其中包括由 1 627 个蛋白质组成的 408 个蛋白质复合体. 在 408 个复合体中,只有一个复合体包含超过 80 个蛋白质,92%的复合体包含的蛋白质不超过 10 个. 值得注意的是,大约 42%的复合体只包含 2 个蛋白质.

在 PPI 网络中,同一个蛋白质可以属于多个蛋白质复合体. 这些蛋白质通常具有多种生物功能,被称为"多功能蛋白质(multifunction protein)",有重要的生物学意义. 我们发现,在 1 627 个蛋白质中共有 211 个蛋白质是多功能蛋白质,占蛋白质总数的 12.96%. 这些蛋白质中,每个蛋白质最多同时属于 9 个蛋白质复合体.

采用"蛋白质复合体体积(protein complex volume)"表示蛋白质复合体包含的蛋白质个数. 从图 2 可以看出,在 408 个蛋白质复合体中,体积大于 20 的只有不到 10 个. 而半数以上的蛋白质复合体体积不大于 3. 这说明多数蛋白质复合体规模较小.

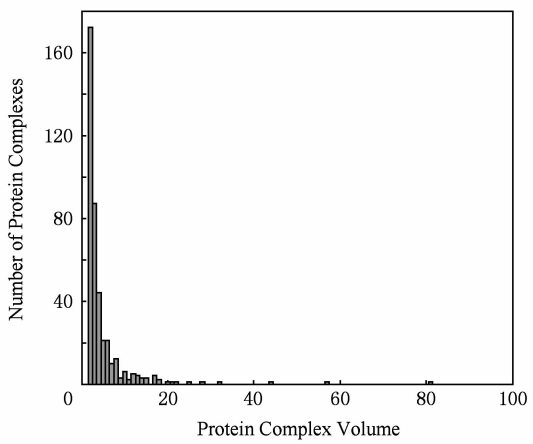

Fig. 2 Volume distribution histogram of protein complex.

图 2 蛋白质复合体体积分布直方图

从图 3 可以看出蛋白质复合体数据集中蛋白质的分布情况.其中共有 164 个蛋白质属于 2 个蛋白质复合体,占多功能蛋白质总数的 77.72%.属于 3 个及以下的复合体的多功能蛋白质有 194 个,占总数的 92%.从而可以得出结论:多数多功能蛋白质只属于少数的复合体.

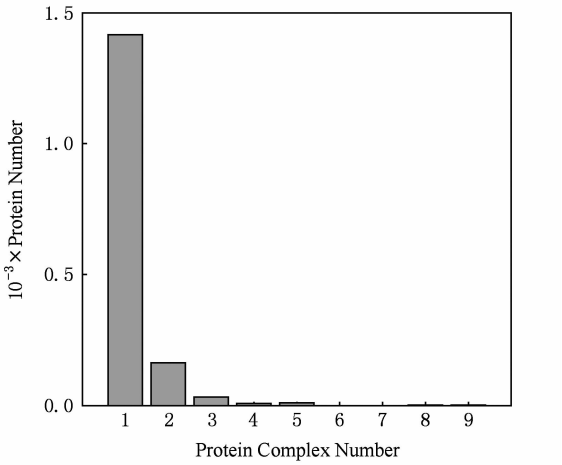

Fig. 3 Histogram of protein distribution.

图 3 蛋白质分布直方图

因此,可以得到如下主要结论:1)在 PPI 网络中存在着大量的蛋白质复合体;2)多数蛋白质复合体的体积较小.这些结论说明了蛋白质复合体规模较小,与其他网络模块结构有差异.根据上述结论,本文提出了基于蛋白质复合体模块度函数的复合体识别算法,旨在更准确地识别出蛋白质复合体.

3 蛋白质复合体模块度函数及识别算法

3.1 蛋白质复合体模块度函数

针对蛋白质复合体具有重叠性和规模较小的特点,本文提出了蛋白质复合体模块度函数(PQ 函数),定义如下:

$$PQ(C)=\frac{1}{2m}\left(1-\frac{1}{|C|}\right)\sum_i\sum_j\frac{|s_i\cap s_j|}{|s_i||s_j|}\left(A_{ij}-\frac{d_id_j}{2m}\right). \tag{2}$$

以图 4 为例,网络中的全部边数 $m=20$,$|C|=3$ 表示当前网络中模块总数目为 3;标有不同形状的节点分别属于模块 a,b,c;$s_2=\{a,b\}$,$s_9=\{b,c\}$ 表示节点 2 和节点 9 都包含在 2 个模块中.$|s_1\cap s_2|=1$ 表示节点 1 和节点 2 共同参与的模块数目为 1.

当整个网络中所有节点属于一个模块时 $PQ=0$,表示模块划分不合理.通常,PQ 值介于 0~1 之间;其值越高,表明 PPI 网络的模块划分质量越好.

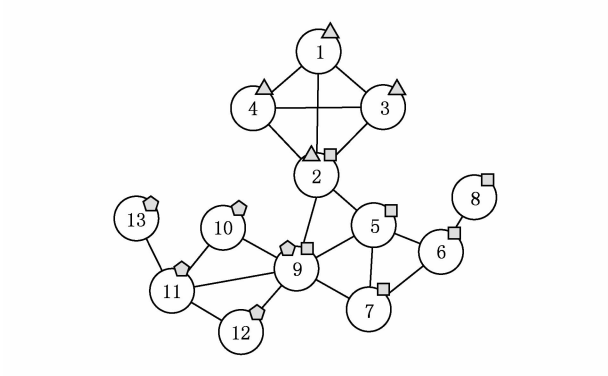

Fig. 4 Overlapped modules partition in a network.

图 4 网络的重叠模块划分

由式(2)中 $\frac{|s_i\cap s_j|}{|s_i||s_j|}$ 一项可知,网络中的节点可以属于多个模块,如图 4 中的节点 2.此外,由 $\left(1-\frac{1}{|C|}\right)$ 一项可知,$|C|$ 与 PQ 值成正比;表明蛋白质复合体数目越多 PQ 值越大,此时蛋白质复合体平均规模也越小;因此,PQ 函数解决了"蛋白质复合体重叠性"和"分辨率限制"问题.

3.2 蛋白质复合体识别算法

以 PQ 函数为基础,我们设计了蛋白质复合体识别算法 BMM.算法的核心思想是:对于给定的 PPI 网络,首先进行初始模块选择,然后迭代地选择能够使得 PQ 值增加的一对模块进行合并;直到模块度不再增加为止,最终将得到的所有模块作为蛋白质复合体.

3.2.1 PPI 网络初始模块选择方法

假设 PPI 网络中节点数目为 n,如果以每一个节点作为一个初始模块,迭代地选择使 PQ 值增加的模块合并,那么算法的时间复杂度为 $O(n^3)$,复杂度过高.为了降低算法的复杂度,我们提出了 PPI 网络初始模块选择方法.主要步骤如下:

步骤 1. 根据节点度降序排列节点.

步骤 2. 按顺序选择一个种子节点,如果该节点不属于任何模块,则把该节点作为种子模块.否则,按顺序选择其他节点.

步骤 3. 按顺序遍历种子节点后面所有节点,如果遍历到的节点和种子模块中一半以上节点邻接,则把遍历到的节点加入种子模块.

步骤 4. 循环执行步骤 2、步骤 3,直到遍历完所有节点.

步骤 5. 去除所有仅包含一个节点的模块,余下的模块作为初始模块.

图 5 是包含 8 个节点的模拟网络. 其中,节点 1～8 按照度降序排列.

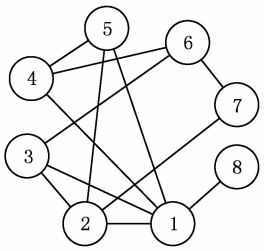

Fig. 5 A simple artificial network.

图 5 一个简单的模拟网络

根据初始模块选择方法,得到图 6 中的 3 个初始模块. 其中,模块 $C_1=\{1,2,3,5\}$, $C_2=\{4,5\}$, $C_3=\{6,7\}$.

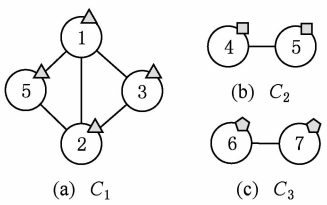

Fig. 6 The selected initial modules.

图 6 选择出的初始模块

3.2.2 基于 PQ 函数的模块合并步骤

假设根据 PPI 网络初始模块选择方法,得到初始模块集合 $C=\{C_1,C_2,\cdots,C_p\}$(初始模块数目为 p). 定义:如果模块 C_i 中至少存在一个节点和模块 C_j 中至少一个节点邻接,则称模块 C_i 和模块 C_j"相互邻接". 模块合并步骤如下:

1) 根据式(2)计算 PQ 函数值.

2) 依次遍历所有模块,如果相邻 2 个模块相互邻接,并且合并后 PQ 函数值增加,则合并 2 个模块,用合并后的模块替换原来 2 个模块,并更新 PQ 函数值.

3) 迭代执行步骤 2),直到 PQ 函数值不再增加为止.

4) 将 C 中的所有模块作为蛋白质复合体.

由于算法首先按照节点度降序排列全部节点,所以,所有的初始模块也是按照种子节点度降序排列的. 因此,集合 $C=\{C_1,C_2,\cdots,C_p\}$ 中相邻的初始模块相互邻接的可能性较大;此外,度较高的节点通常存在规模较大的模块中,度较小的节点通常存在于规模较小的模块中. 为了能够识别出更多的小规模蛋白质,我们希望合并的模块规模相近.

利用上述模块合并算法,将图 6 中的模块合并为图 7 中的 2 个模块,分别为 $C_1=\{1,2,3,4,5\}$, $C_2=\{6,7\}$.

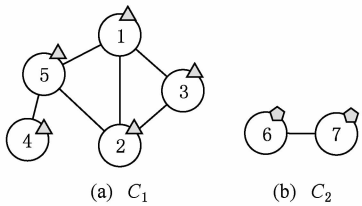

Fig. 7 Identified modules.

图 7 识别得到的模块

3.2.3 BMM 算法描述

算法 1. BMM 算法.

输入:PPI 网络(蛋白质相互作用关系列表);

输出:蛋白质复合体集合 C.

① 预处理 PPI 网络,对节点按度从大到小排序,得到节点序列 $V=\{v_1,v_2,\cdots,v_n\}$;
② 令 $l=0$, $C=\varnothing$;
③ for $i=1$ to n
④ if v_i 不属于任何模块
⑤ $l=l+1$;
⑥ 以 v_i 为核心组成新模块 $C_l=\{v_i\}$, $C=C\cup\{C_l\}$;
⑦ for $j=i+1$ to n
⑧ if v_j 和 C_l 中的一半以上节点邻接
⑨ 加入 $C_l=C_l\cup\{v_j\}$;
⑩ end if
⑪ end for
⑫ if $|C_l|\leqslant 1$
⑬ 从 C 中删除 C_l;
⑭ end if
⑮ end if
⑯ end for
⑰ 根据式(2)计算 C 的模块度 $PQ(C)$;
⑱ $flag=$true;
⑲ while $flag$
⑳ $flag=$false;
㉑ for $i=1$ to $|C|-1$
㉒ 计算 C_i 与 C_{i+1} 合并后 C' 的模块度 $PQ(C')$;
㉓ if $PQ(C')>PQ(C)$
㉔ $C_i=C_i\cup C_{i+1}$;
㉕ 删除 C_{i+1};
㉖ $PQ(C)=PQ(C')$;
㉗ $flag=$true;

㉘ end if

㉙ end for

㉚ end while

㉛ return C.

3.2.4 算法复杂度分析

由于网络中节点数目为 n,则初始模块选择步骤的时间复杂度为 $O(n^2)$. 选择出初始模块数目为 p,则模块合并步骤的时间复杂度为 $O(p^2)$. 所以,算法总时间复杂度为 $O(n^2)+O(p^2)$. 由于 $p<n$,所以 BMM 算法时间复杂度近似为 $O(n^2)$.

此外,Newman 等人[6]的识别社团结构的同类快速算法是基于模块度 Q 函数,其时间复杂度为 $O((m+n)n)$. 由于 $m>>n$,因而 BMM 算法时间复杂度较低.

4 实验结果分析

本文使用 Java 语言实现了 BMM 算法,算法实现源码可在 http://nclab.hit.edu.cn/bmm/下载. 为了验证 BMM 算法,我们将其与 ClusterONE 和 OCG 等算法进行了比较. 分别使用根据 ClusterONE 算法开发的软件工具① 和根据 OCG 算法开发的软件工具② 进行实验. 实验测试是在采用 Windows XP 操作系统的 PC 机(AMD 2.71 GHz, 2 GB 主存)上进行的.

4.1 实验数据

由于酵母是所有物种中相互作用数据最完备的,所以选择酵母相互作用数据为实验数据[20]. 我们采用 DIP 数据库提供的 Scere20120818 数据集,其中包含 22 466 对相互作用. 在数据预处理阶段,去除数据集中的 318 对自相互作用数据. 形成由 5 078 个蛋白质组成的,包含 22 148 对相互作用的 PPI 网络. 最终,BMM 算法、ClusterONE 算法和 OCG 算法分别识别出 1 366,1 326 和 470 个蛋白质复合体.

为了评价识别出的复合体的准确性,我们使用 2 种已知复合体数据集作为参照. 分别是 CYC2008 复合体③ 和 MIPS 数据库中的复合体数据(2006-05-18 版本). CYC2008 数据集中包含 408 个蛋白质复合体,MIPS 数据集中包含 203 个蛋白质复合体.

① http://www.paccanarolab.org/cluster-one/

② http://tagc.univ-mrs.fr/tagc/index.php/software

③ http://wodaklab.org/cyc2008/

4.2 性能评价

我们将 BMM 算法与 ClusterONE 和 OCG 等算法在 Scere20120818 网络上运行,并比较了它们的识别结果. 实验发现 ClusterONE 算法在输入参数为 0.5 时识别效果最佳.

评价蛋白质复合体识别算法性能的方法主要有算法识别出的复合体和已知复合体匹配程度,以及灵敏度和特异性分析等[21]. 下面,从这 2 个方面分析 3 种算法的性能.

4.2.1 匹配统计

我们用 Ic 表示算法识别出的蛋白质复合体(identified complexes)集合,用 Kc 表示已知的蛋白质复合体(known complexes)集合. 对于蛋白质复合体 $A\in Ic$ 和 $B\in Kc$,我们使用 $OS(A, B)$[22] 评价 A 和 B 的匹配程度:

$$OS(A,B)=\frac{|A\cap B|^2}{|A|\times|B|}, \tag{3}$$

其中,$|A\cap B|$ 表示复合体 A 和 B 包含的相同蛋白质的数目,$|A|$ 表示复合体 A 包含的蛋白质数目.

根据设定的阈值 $os\text{-}value$,我们可以判断复合体 A 和复合体 B 的匹配程度. 若 $OS(A, B)\geqslant os\text{-}value$,则表示 A 与 B 匹配,否则为失配. 若 $OS(A, B)=1$,则表示 A 与 B 完全匹配(perfect matching),即复合体 A 与 B 所包含的蛋白质完全相同. 完全匹配的复合体数目越多说明算法的识别能力越强.

图 8 给出了 3 种算法完全匹配 2 种已知复合体的数目对比. 对于 2 种已知复合体,BMM 算法均表现出明显的优势. 特别地,BMM 算法完全匹配 CYC2008 复合体数目明显高于另外 2 种算法,这说明 BMM 算法更适合识别蛋白质复合体.

图 9 描述了不同阈值下 3 种算法匹配 CYC2008 复合体的能力. 如果阈值越高则说明匹配效果越好,所以我们选取的阈值范围是 $0.6\leqslant os\text{-}value\leqslant 1$. 从图 9 可以看出,BMM 算法识别出的复合体匹配 CYC2008 复合体数目明显高于另外 2 种算法,表现出较大的优势.

图 10 比较了不同阈值下的 3 种算法匹配 MIPS 复合体的能力. 从图 10 可以看出,BMM 算法匹配已知复合体数目明显高于 ClusterONE 算法和 OCG 算法(OCG 算法匹配已知蛋白质复合体的数目为 0),证明了 BMM 算法的识别效果更好.

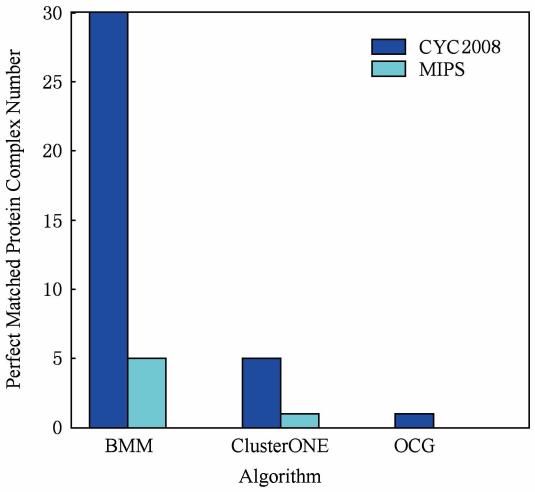

Fig. 8 Comparison of known complex numbers perfect matched by three algorithms.

图 8 3 种算法完全匹配已知复合体数目比较

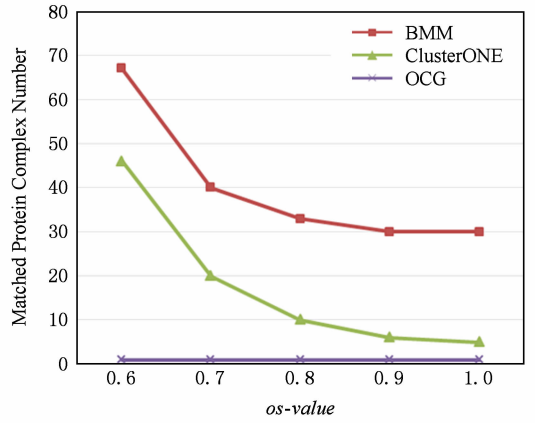

Fig. 9 Comparison of three algorithms in identifying CYC2008.

图 9 3 种算法匹配 CYC2008 复合体能力比较

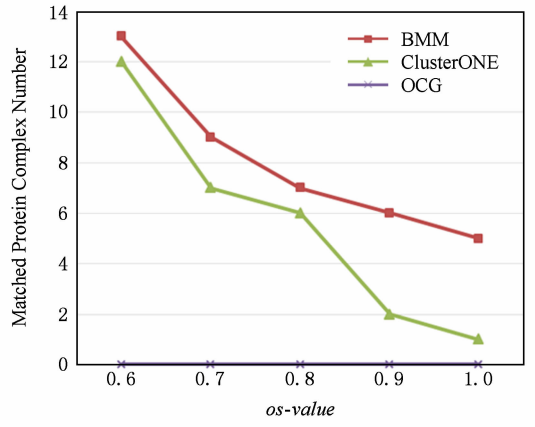

Fig. 10 Comparison of three algorithms in identifying MIPS.

图 10 3 种算法匹配 MIPS 复合体能力比较

4.2.2 灵敏度和特异性分析

灵敏度(sensitivity)S_e 和特异性(specificity)S_p 是衡量识别算法性能的 2 个重要指标. 灵敏度是指已知蛋白质复合体中被算法识别出的比例;特异性则是指算法识别出的复合体中与已知复合体匹配的比例. 灵敏度和特异性的计算如下:

$$S_e=\frac{TP}{TP+FN}, \tag{4}$$

$$S_p=\frac{TP}{TP+FP}, \tag{5}$$

其中,TP 被称为"真阳样本",表示识别出的复合体和已知复合体匹配的数目;FN 被称为"假阴样本",表示已知复合体中没有被算法识别出来的数目;FP 被称为"假阳样本",表示识别出的复合体中没有匹配已知复合体的数目,等于识别的全部复合体数目减去 TP 的值.

为了更好地衡量算法的性能,有的文献综合灵敏度和特异性提出了综合评价指标 *F-measure* 值,该值越高说明算法性能越好[23]. 计算如下:

$$F\text{-}measure=\frac{2\times S_e\times S_p}{S_e+S_p}. \tag{6}$$

表 1 列出了 BMM 算法和另外 2 种算法在 $os\text{-}value=0.6$ 的条件下针对 2 种已知复合体的灵敏度、特异性和综合评价 *F-measure* 的值. 从表 1 可以看出,BMM 算法的各项指标均高于另外 2 种算法对应值. 这说明在蛋白质复合体识别方面,BMM 算法比 ClusterONE 算法和 OCG 算法性能优越.

Table 1 Comparison of BMM and Other Algorithms under Sensitivity, Specificity, and *F-measure*

表 1 BMM 与其他算法的敏感度、特异性及 *f* 的比较

Algorithm	Real Complex Set	S_e	S_p	*F-measure*
BMM	CYC2008	0.15	0.049	0.074
	MIPS	0.064	0.009	0.017
ClusterONE	CYC2008	0.109	0.035	0.053
	MIPS	0.059	0.009	0.016
OCG	CYC2008	0.002	0.002	0.002
	MIPS	0	0	0

另外,从表 1 可见,不同算法在 CYC2008 数据集上的识别结果均显著好于 MIPS 算法. 这种差异是由于 CYC2008 收录的复合体质量较高、数量更多.

5 结 论

针对 PPI 网络中的蛋白质复合体识别问题,本

文分析了蛋白质复合体的结构特点,提出了蛋白质复合体模块度函数 PQ 以及 PPI 网络初始模块选择方法;在此基础上,提出了 BMM 算法.实验部分采用真实的酵母相互作用数据,分别使用 BMM 算法、ClusterONE 算法和 OCG 算法进行识别;并使用 2 种已知的复合体数据集,分别匹配 3 种算法的识别结果.结果表明 BMM 算法匹配效果最好,说明 BMM 算法适合解决蛋白质复合体识别问题.

参 考 文 献

[1] Gavin A C, Superti F G. Protein complexes and proteome organization from yeast to man [J]. Current Opinion in Chemical Biology, 2003, 7(1): 21-27

[2] Von M C, Krause R, Sne B, et al. Comparative assessment of large-scale data sets of protein-protein interactions [J]. Nature, 2002, 417(6887): 399-403

[3] Harwell L H, Hopfield J, Leibler S, et al. From molecular to modular cell biology [J]. Nature, 1999, 402(6761): C47-C52

[4] Victor S, Leonid A. Protein complexes and functional modules in molecular networks [J]. Proc of the National Academy of Sciences of the United States of America, 2003, 100(21): 12123-12128

[5] Luo Feng, Yang Yunfeng, Chen Chinfu, et al. Modular organization of protein interaction networks [J]. Bioinformatics, 2007, 23(2): 207-214

[6] Newman M E J, Girvan M. Finding and evaluating community structure in networks [J/OL]. Physical Review E, 2004, 69(2): 026133. [2014-04-01]. http://journals.aps.org/pre/abstract/10.1103/PhysRevE.69.026113

[7] Wang Jianxin, Li Min, Deng Youping, et al. Recent advances in clustering methods for protein interaction networks [J]. BMC Genomics, 2010, 11(Suppl 3): S10

[8] Brohee S, Helden J. Evaluation of clustering algorithms for protein-protein interaction networks [J]. BMC Bioinformatics, 2006, 7: 488

[9] Przulj N, Wigle D A, Jurisica I. Functional topology in a network of protein interactions [J]. Bioinformatics, 2004, 20(3): 340-348

[10] Li X, Wu M, Kwoh C, et al. Computational approaches for detecting protein complexes from protein interaction networks: A survey [J]. BMC Genomics, 2010, 11 (Suppl): S3

[11] Li Min, Wang Jianxin, Liu Binbin, et al. An algorithm for identifying protein complexes based on maximal clique extension [J]. Journal of Central South University: Science and Technology, 2010, 41(2): 560-565(in Chinese)
(李敏,王建新,刘彬彬,等. 基于极大团扩展的蛋白质复合物识别算法[J]. 中南大学学报:自然科学版,2010, 41(2): 560-565)

[12] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks [J]. Nature Methods, 2012, 9(5): 471-475

[13] Becker E, Robisson B, Chapple C E, et al. Multifunctional proteins revealed by overlapping clustering in protein interaction network [J]. Bioinformatics, 2012, 28(1): 84-90

[14] Newman M. Communities, modules and large-scale structure in networks [J]. Nature Physics, 2012, 8(1): 25-31

[15] Zhang Cong, Shen Huizhang. Modularity function for community structure based on natural density of networks [J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 185-191(in Chinese)
(张聪,沈惠璋. 网络自然密度社团结构模块度函数[J]. 电子科技大学学报,2012, 41(2): 185-191)

[16] Nicosia V, Mangioni G, Carchiolo V, et al. Extending the definition of modularity to directed graphs with overlapping communities [J/OL]. Journal of Statistical Mechanics: Theory and Experiment, 2009, 2009 (03): P03024. [2014-03-16]. http://iopscience.iop.org/1742-5468/2009/03/P03024

[17] Wang Xiaofan, Liu Yabing. Overview of algorithms for detecting community structure in complex networks [J]. Journal of University of Electronic Science and Technology of China, 2009, 38(5): 537-543(in Chinese)
(汪小帆,刘亚冰. 复杂网络中的社团结构算法综述[J]. 电子科技大学学报,2009, 38(5): 537-543)

[18] Santo F, Marc B. Resolution limit in community detection [J]. Proc of the National Academy of Sciences of the United States of America, 2007, 104(1): 36-41

[19] Pu S, Wong J, Turner B, et al. Up-to-date catalogues of yeast protein complexes [J]. Nucleic Acids Research, 2009, 37(3): 825-831

[20] Chen Jingchun, Yuan Bo. Detecting functional modules in the yeast protein-protein interaction network [J]. Bioinformatics, 2006, 22(18): 2283-2290

[21] Wang Jianxin, Li Min, Chen Jianer, et al. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks [J]. IEEE/ACM Trans on Computational Biology and Bioinformatics, 2011, 8(3): 607-620

[22] Tang Xiwei, Wang Jianxin, Hu Qiuling. Analysis and compare of methods predicting protein complex [J]. Application Research of Computers, 2011, 28(10): 3611-3614 (in Chinese)
(汤希玮,王建新,胡秋玲. 蛋白质复合物预测方法分析与比较[J]. 计算机应用研究,2011, 28(10): 3611-3614)

[23] Li X, Tan S H, Foo C S, et al. Interaction graph mining for protein complexes using local clique merging [J]. Genome Informatics, 2005, 16(2): 260-269

Guo Maozu, born in 1966. Professor and PhD supervisor in Harbin Institute of Technology. Member of China Computer Federation. His research interests include computational biology, machine learning and image understanding.

Dai Qiguo, born in 1985. PhD candidate in the School of Computer Science and Technology, Harbin Institute of Technology. His research interests include bioinformatics and artificial intelligence.

Xu Liqiu, born in 1989. Master in the School of Computer Science and Technology. Harbin Institute of Technology. His main research interest includes bioinformatics.

Liu Xiaoyan, born in 1963. PhD and associate professor in Harbin Institute of Technology. Her research interests include bioinformatics, engineering database and artificial intelligence, etc.