EasyChair Preprint

№ 1211

Detecting protein complexes based on a combination of topological and biological properties in protein-protein interaction network

Dongwen Zhang, Peiheng Wang and Yunfeng Xu

June 20, 2019

基于蛋白质相互作用网络中拓扑特征和生物学特性的组合检测蛋白质复合物

张冬雯[1] 王佩恒[1] 许云峰[1]

[1]（河北科技大学 信息科学与工程学院 石家庄 050000）
（zdwwtx@163.com）

Detecting protein complexes based on a combination of topological and biological properties in protein-protein interaction network

ZHANG Dongwen[1] , WANG Peiheng[1] , XU Yunfeng [1]

[1](.School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang Hebei 050000, China)

Abstract Protein complexes are aggregates of protein molecules that play important roles in biological processes. The systematic analysis of PPI networks can enable a great understanding of cellular organization, processes and function.Identifying complexes from raw protein protein interactions (PPIs) is an important area of research. Earlier work has been limited mostly to yeast. Such protein complex identification methods, when applied to large human PPIs often give poor performance.We introduce a novel method called CFM to detect protein complexes.Experiments were carried out on the PPI datasets of DIP, Krogan, HPRD and Mouse respectively. MIPS and PCDq were used as standard complexes of Saccharomyces cerevisiae and human.The results show that compared with the six classical PPI algorithms of WCOACH, ClusterONE, SPICi, MCL, MCODE and CFinder,The algorithm is similar to the classical PPI clustering algorithm in Saccharomyces cerevisiae, and the accuracy of protein complex prediction in human PPI network is higher than the other six algorithms, and the number of protein complexes predicted in the mouse PPI network is higher than other algorithms, which improves the accuracy of predicting protein complexes on multi-species PPI networks.In addition, D3 visualization technology was applied to PPI network visualization field, which provided a beneficial reference for the mining and analysis of biological network modules.

Key words protein-protein interaction network; protein complex; clustering;evaluate;

摘要 蛋白质相互作用（PPI）在几乎所有生物过程中都发挥着重要作用。PPI 网络的系统分析可以很好地理解细胞组织，过程和功能。从原始蛋白质蛋白质相互作用（PPI）中鉴定复合物是一个重要的研究领域。目前的研究工作主要限于酵母。当应用于大型人类的 PPI 网络时，蛋白质复合物检测方法通常表现不佳。本文开发了 CFM 模型来检测蛋白质复合物。对 CFM 方法分别在 DIP、Krogan、HPRD 和 Mouse 的 PPI 数据集上进行了实验，MIPS 和 PCDq 分别作为酿酒酵母和人类的标准复合物集。结果表明,相比 WCOACH、ClusterONE、SPICi、MCL、MCODE 和 CFinder 六种经典 PPI 算法,该算法在酿酒酵母预测结果和经典的 PPI 聚类算法相当，在人类 PPI 网络预测蛋白质复合物的准确率优于其他 6 种算法,在小鼠 PPI 网络预测蛋白质复合物的数量多于其他算法,达到了提高在多物种 PPI 网络上预测蛋白质复合物准确率的目的,并且将 D3 可视化技术应用到 PPI 网络可视化领域，为生物网络模块的挖掘和分析提供了有益的参考。

关键词 蛋白质相互作用网络；蛋白质复合物；聚类；评估；

基金项目：国家自然科学基金，国家重点基础研究发展计划(973) 。
The National Natural Science Foundation of China (General Program, Key Program, Major Research Plan)，The National Basic Research Program of China (973 Program)

通信作者：许云峰（386839300@qq.com）

中图法分类号 TP391

在蛋白质组学时代，各种高通量的实验技术和计算方法已经产生了巨大的蛋白质相互作用数据[1]。蛋白质相互作用(Protein-Protein Interaction，PPI)网络是由单独蛋白通过彼此之间的相互作用构成，来参与生物信号传递、基因表达调节、能量和物质代谢及细胞周期调控等生命过程的各个环节[2]。蛋白质复合物是一组蛋白质，它们彼此之间通过物理结合并以连通的方式起作用以执行特定的生物学功能[3]。

本文的主要贡献如下:

1) 我们提出了一种基于拓扑特征和生物学特性组合的聚类方法，通过分析蛋白质相互作用网络来检测重叠的蛋白质复合物。

2) 将基于基因本体论（GO）结构的蛋白质之间的语义相似性度量应用于权衡 PPI 网络，可以减少 PPI 网络噪音数据。

3) 将 D3.js 可视化技术应用到大型的蛋白质相互作用网络上，对 PPI 网络和聚类的结果进行可视化，以更好的方式解释生物现象。

4) 应用不同数据集的实验结果表明我们提出的检测重叠蛋白质复合物算法可以有效发现大型网络中存在的蛋白质复合物。

1 相关工作

近年来，已经从简单地构建 PPI 网络转变为分析现有网络并且发现网络中的蛋白质如何相互作用[4]。研究表明，同一蛋白质复合物中的蛋白质倾向于彼此相互作用，因此 PPI 网络中的密集区域可能是潜在的蛋白质复合物。基于 PPI 网络的蛋白质复合物的检测可以帮助展开细胞生物学的各个方面并识别未表征的蛋白质的生物学功能。聚类技术已被广泛用于使用 PPI 网络检测蛋白质复合物。聚类就是将数据对象分组成多个类或者簇，划分的原则是在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大[5]。

一般来说，将 PPI 网络表示为图，其中顶点是蛋白质，两个顶点之间的边表示这两种蛋白质之间的相互作用[6]。鉴定蛋白质复合物类似于在图中找到簇。这一概念是研究人员通过在 PPI 网络中找到密集连接区域来尝试发现新蛋白质复合物的基础。为了识别蛋白质复合物，许多研究人员采用了不同的方法。例如，MCODE[7]通过识别大型 PPI 网络中的高度连通区域来发现蛋白质复合物。MCL[8]通过图中模拟流体流动识别蛋白质复合物。CFinder[9][10]使用 clique 合并程序来识别复合物。SPICi[11]利用集群拓展方法检测蛋白质复合物，ClusterONE[12]使用内聚力来寻找蛋白质复合物。最近有 COACH 的改进版本 WCOACH[13]。在过去几年中，GO 注释也被用于提高复杂预测的准确性。尽管有许多蛋白质复合物预测方法，但很少有方法在人类数据集上进行验证[14]。对一些现有方法的实证分析表明，这些方法主要用于酵母数据集。因此有必要找到一种能够同样适用于多物种数据集的方法。

本文提出的基于拓扑特征和生物学特性组合的检测蛋白质复合物方法，可以减少消除噪音数据，并且可以在大型 PPI 网络识别功能模块重叠结构。本文主要用 F-measure 标准和 Accury 标准对 CFM 蛋白质复合物识别算法进行详细的评估，并对预测的复合物进行 GO 功能富集分析。

2 PPI 网络中加权骨干度和语义相似度的定义

在本节中，我们主要介绍加权骨干度和语义相似度组合的方法在蛋白质相互作用网络中的应用. 其中，2.1 节给出了模型所需的定义. 2.2 节描述了如何将基于 GO 结构的蛋白质对之间的语义相似性度计算每个相互作用的权重.

2.1 基本定义

给定一个与 PPI 网络对应的图 ***G=（V，E）***，其中 ***V*** 代表一组蛋白质，***E*** 代表一组边缘，任务是找到一组子图，使这些子图与基准复合物紧密对应。

定义 1 让顶点 ***v*** 的标识符为 ***i***，图 ***G*** 中任何顶点的网络权重都可以表示为 ***Vi***。我们可以用 ***NW*** (***v0,G***) 表示网络的重量。

$$\mathrm{NW}(\mathrm{v}_0,\mathrm{G})=\sum_{i=1}^{n} w(v_0,v_i),(v_0,v_i)\in E \qquad (1)$$

定义 2（语义相似度）蛋白质对之间的语义相似性(***vi,vj***)由语料库(GO 数据库)中与之关联的概念(GO 术语)的相似性给出：

$$\mathrm{semsim}(v_i,v_j)=sim(GOterms_i,GOterms_j) \qquad (2)$$

定义 3（社区膨胀度）这个指标衡量的是社区外的边缘数量。

$$\mathrm{EX_C}=\frac{|\mathrm{CB_E}|}{\mathrm{C_n}} \qquad (3)$$

定义 4（社区膨胀度差异）这个指标衡量的是社区外的边缘数量。

$$\mathrm{DE(i)}=\mathrm{EX}_{\mathrm{C}\cap\{\mathrm{i}\}}-\mathrm{EX_C} \qquad (4)$$

定义 5（邻里互惠度）给定两个顶点 ***u*** 和 ***v***，让 ***NBu*** 是顶点 ***u*** 的邻域，让 ***NBv*** 是顶点的邻域，让 ***NOuv***

是 ***u*** 和 ***v*** 的邻里互惠度：

$$\mathrm{NOuv}=\frac{\left|\mathrm{NB_v}\cap\mathrm{NB_u}\right|}{\left|\mathrm{NB_v}\cup\mathrm{NB_u}\right|-2} \tag{5}$$

定义 6 设顶点 ***u*** 和顶点 ***v*** 的边的加权骨干度为：

$$\mathrm{D_{uv}}=(\mathrm{NW_u}+\mathrm{N}W_{\mathrm{v}})\times NO_{uv}+\delta \tag{6}$$

Duv 可以测量边的强度和节点相似度。当顶点 ***u*** 和顶点 ***v*** 不相邻的时候 ***NOuv*** 为 0，***Duv=δ***，***δ***是平滑常数的参数，基于经验我们让 ***δ***=0.01。

2.2 基于 GO 结构的蛋白质对之间的语义相似性度

GO 数据库由 GO 术语及其关系组成。GO 术语分为三个域：生物过程（BP），分子功能（MF）和细胞成分（CC）。每个基因或蛋白质注释到一个或多个 GO 术语，并且每个术语与有向无环图（DAG）结构中的一个或多个其他术语相关。根据实验观察，复合物中的蛋白质倾向于发挥共同的生物学功能，因此我们使用 GO 注释来测量每对蛋白质之间相互作用的可靠性。在本文中，蛋白质对之间的语义相似性通过 Lin 方法（GraSM）计算以权衡 PPI 网络。

3 基于 CFM 的聚类过程

在本节中，我们详细介绍如何计算骨干度和膨胀度，以及 CFM 聚类算法的主要过程。

CFM 聚类方法主要过程如下，首先计算蛋白质相互作用网络中每一个骨干的骨干度，并且以降序的方式把这些骨干度保存在骨干度列表里。让初始集群是空的，然后选择骨干度列表中骨干度最大的骨干作为当前集群的初始骨干，形成高质量的集群。在聚类展开过程中，通过引入语义相似度值，进一步增强了节点 v_c 的隶属度。语义相似度阈值为 0.6。通过实验将膨胀度阈值设置为 0.3 较为合理。一旦满足了这两个条件，就可以证明节点 v_c 在拓扑和功能上都是一个很好的选择，可以在分簇中形成一个包含节点 v_a 和 v_b 的复合体。如果加入到当前集群后膨胀度变小，那么不断地把最大的骨干度加入集群，继续寻找同当前集群中有连接的顶点中有着最大加权骨干度的顶点，重复这个过程，直到没有其他节点满足这两个条件，此刻一个新的社区就完全划分出来了。按照上述方案迭代，把剩余的顶点划分到新的集群里。直到骨干度列表中不再有骨干度大于骨干度阈值 f 的骨干存在，或者其余顶点的数目小于参数 ***ω***。这里 ***ω***的值是根据 $/V/$的值得到的。然后 CFM 根据预定义的重叠分数合并高度重叠的局部最佳内聚组对。最后，CFM 输出含有不少于三种蛋白质或其密度大于给定阈值 ***θ***的蛋白质复合体。

算法 1 陈述了骨干度的计算的步骤。

骨干度的计算

输入：无向加权的蛋白质相互作用网络
输出：邻接表
① For all $v_i\in V$ DO
② $NW_i=NW(v_i,G_N)$;//G_{Ni} 为 v_i 的邻居图
③ For all $e_i\in E$ DO
④ $W(v_u,v_v)=D_{uv}$;
⑤ Return Adjacency list

算法 2 陈述了检测蛋白质复合物的步骤。

CFM 算法的实现

数据：无向加权的蛋白质相互作用网络
1 while $BD_b\geq f$ and $nl\geq w$ do
2 if $u\in NL$ and $v\in NL$ then
3 $C_i\Leftarrow\{u,v\}$
4 $E_{C\_PRE}\Leftarrow$ the Expansion degree of C_i .
5 calculate the NB_{Ci} of C_i; $BV_{Ci}\Leftarrow$ null.
6 if $\{NB_C - BV_C\}$= Null
7 if GOsim>n, then
8 add C_i to CF; i ++; goto step2.
9 else
10 find the nearest vertex nv from $\{NB_{Ci} - BV_{Ci}\}$ based on backbone degree, add vertice nv to Ci, calculate the Expansion degree of C_i and note it as E_{C_cur}.
11 if ($E_{C\_cur} - E_{C\_PRE} < 0$), then
12 remove vertice nv from NL and add vertice nv to C_i,
13 goto step11.
14 else delete vertice nv from C_i, add vertice nv to BV_C,
15 if$\{NB_C - BV_C\}$= Null
16 if GOsim>n, then
17 add C_i to CF; i ++, goto step2
18 else
19 goto step11.
20 end if
21 end if
22 end if
23 else
24 goto step2;
25 end if
26 end while
27 Collect all vertices that divided into no community or

several communities.
28 return CF.
29 end

为使得初始膨胀度阈值的取值较为合理，在保证其他参数相同的情况下，在 4 个试验数据集上进行试验，实验结果如下图所示：

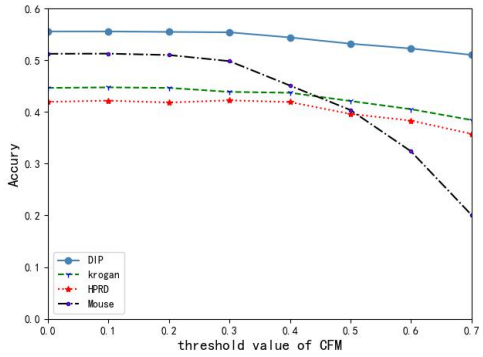

Fig. 1 Change of accuracy under different parameters of CFM.
图 1 CFM 不同参数下准确率的变化

4 基于 D3. js 的 PPI 网络可视化。

D3.js 是一个 JavaScript 库。它的全称是 Data-Driven Documents（数据驱动文档），并且它被称为一个互动和动态的数据可视化库网络。我们将 D3.js 可视化技术应用到蛋白质相互作用网络，使用力导向图布局和树形图布局来可视化最终的蛋白质相互作用网络。

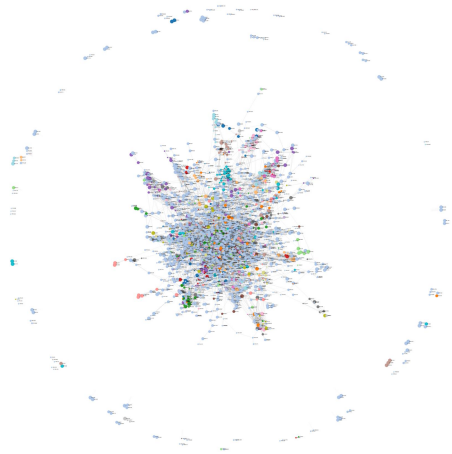

Fig. 2 MIPS labeled krogan PPI network.
图 1 MIPS 标注的 krogan PPI 网络

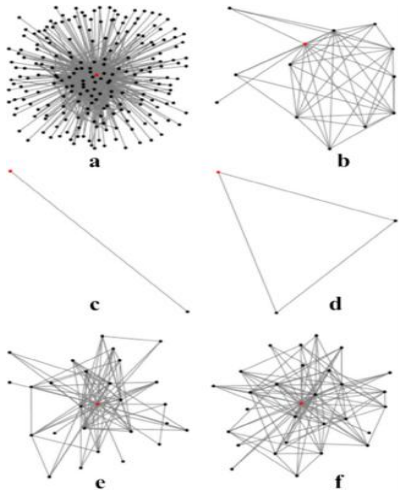

Fig. 3 The visualization of six subgraphs that belong to the protein=protein interaction network.
图 3 蛋白质相互作用网络可视化图中六个子图

5 实验与结果

在本节中，我们使用本文提出的技术构建了一个检测 PPI 网络中的重叠蛋白复合物模型，并且在 4 个 PPI 数据集上测试了我们的方法.

5.1 度量标准

为了评估聚类算法的有效性并验证结果，使用下面的几种评估方法。

5.1.1 基于精确率和召回率的评估

被广泛用于验证识别蛋白质复合体有效性的 *F* 度量（F-measure）[15]，设 *TP*(真阳性)表示与基准配合物相匹配的预测配合物数目，*FN* (假阴性)表示与预测配合物中任何一个不匹配的基准配合物数目，*FP* (假阳性)表示预测配合物减去 *TP* 的数目。精确度（Precision）、召回率（Recall）和 *F* 值的定义如下:

精确度(Precision)是与参考复合体匹配的预测复合体的数量占所有预测复合体数量的比例, Precision 如式:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

召回率(Recall)是与参考复合体匹配的预测复合体的数量占所有参考复合体数量的比例, Recall 如式所示:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

精确度与召回率都只能反映算法单方面的性质，且这 2 个参数往往有相反的变化趋势，因此可以用精确度和查全率的调和平均值 F-measure 来综合评估预测蛋白质复合体的准确性，其 F-measur 如式所示:

$$\text{F-Measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (9)$$

5.1.2 基于聚类灵敏度和阳性预测值的评估

Xie 等人引入的几何精度(Acc)[16]，即灵敏度(Sn)和阳性预测值(PPV)的几何平均值。给定 n 个参考和 m 个预测的复合物，让 t_{ij} 表示在参考复合物 I 和预测的复合物 J 中发现的蛋白质的数目，并且让 Ni 表示参考复合物中的蛋白质的数目。Sn，PPV 和 Acc 定义如下:

$$\text{Sn}=\frac{\sum_{i=1}^{n}\max_{j=1}^{m} t_{ij}}{\sum_{i=1}^{n} n_i} \quad (10)$$

$$\text{PPV}=\frac{\sum_{j=1}^{m}\max_{i=1}^{n} t_{ij}}{\sum_{j=1}^{m}\sum_{i=1}^{n} t_{ij}} \quad (11)$$

$$\text{Acc}=\sqrt{\text{Sn}\times\text{PPV}} \quad (12)$$

5.1.3 预测蛋白质复合物的统计学意义（p 值）

预测蛋白质复合物的统计学意义（p 值）[17]。假设大小为 n 的簇，m 个蛋白质共享特定的生物学注释。还假设数据库中有 N 个蛋白质，其中 M 个已知具有相同的注释。然后使用超几何分布，观察 m 的概率或更多用 n 个蛋白质中相同 GO 术语注释的蛋白质，则 P-value 值定义为：

$$\text{P-value}=\sum_{i=m}^{n}\frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \quad (13)$$

使用在 http://go.princeton.edu/cgi-bin/GOTermFinder 上获得 GO Term Finder 工具计算预测的蛋白质复合物的 P 值。较小的 P 值意味着分组不是随机的，并且在生物学上比具有更高 P 值的分组更重要。使用阈值为 0.01，只要 p 值<0.01，鉴定的蛋白质功能模块被认为具有生物学意义，

由于单个集群的 p 值在统计上不具有代表性，定义了一个集群得分函数来量化整个集群，如下所示[18]：

$$\text{Clustering score}=1-\frac{\sum_{i=1}^{n_S}\min(p_i)+(n_I * cutoff)}{(n_S+n_I)*cutoff} \quad (14)$$

其中 n_S 是显著的簇数，n_I 是不显著的簇数，参数 $cutoff$ 是截止的 P 值，用来区分重要群集和无意义群集。本文设置 $cutoff$ 为 0.01，如果 P 值大于 $cutoff$ 的集群，则它是无关紧要的集群。$min（p\,i）$表示有效聚类 i 的最小 p 值，较大的聚类得分(CS)通常意味着更好的聚类结果。

5.2 数据集

本文使用了四个数据集。酵母 PPI 网络 DIP 数据集[19]，krogan 数据集[20]，一个人类 PPI 网络 HPRD 数据集[21]，另一个小鼠 PPI 网络 Mouse 数据集[22]。使用 WCOCHA 算法的作者提供的 DIP 数据集，由 17201 个相互作用和 4606 种蛋白质组成的酵母数据集。krogan 包括 2675 种蛋白质和 7084 个相互作用。HPRD 是包含 39209 个相互作用和 10080 种蛋白质的人类 PPI 数据集。Mouse 是包含 22488 个蛋白质相互作用和 7590 种蛋白质的的小鼠 PPI 数据集。MIPS[23]作为酿酒酵母标准复合物集，PCDq[24]作为人类标准复合物集来评估算法预测的蛋白质复合物。对于基准复合物组，过滤掉不参与 PPI 数据的蛋白质。此外，参考 Nepusz 等人的方法，只考虑至少包含三种蛋白质的复合物。因此，预处理后的 MIPS 含有 200 种复合物，1059 种蛋白质。PCDq 含有 501 种复合物，1800 种蛋白质。

5.3 实验和结果

在本节中，我们进行 4 个实验来验证 CFM 算法检测蛋白质复合物的可行性．对于 ClusterONE，MCODE，MCL 算法，使用 Cytoscape 中已有的插件进行计算。对于 SPICi，CFinder，WCOACH 算法，下载的初始作者开发的开源软件。将 CFM 聚类算法的性能与上述算法进行对比，表 1 和表 2 显示了 4 种 PPI 网络上所有算法的对比结果：

Table 1 Comparison of prediction results of clustering algorithms

表 1 各聚类算法预测结果对比

数据集	算法	复合物个数	F-measure	Accuracy	CS
DIP	CFm	762	0.40	0.46	0.92
	WCOACH	648	0.41	0.40	0.89
	ClusterONE	895	0.36	0.45	0.79

	SPICi	219	0.41	0.38	0.90
	MCODE	26	0.23	0.21	0.99
	MCL	238	0.23	0.30	0.60
	CFinder	243	0.29	0.36	0.59
Krogan	CFM	452	0.42	0.43	0.86
	WCOACH	638	0.36	0.36	0.79
	ClusterONE	523	0.33	0.44	0.69
	SPICi	133	0.41	0.40	0.82
	MCODE	91	0.34	0.32	0.93
	MCL	376	0.33	0.44	0.60
	CFinder	115	0.33	0.37	0.75
HPRD	CFM	1860	0.29	0.29	0.96
	WCOACH	1674	0.17	0.21	0.87
	ClusterONE	1643	0.23	0.23	0.82
	SPICi	853	0.28	0.24	0.76
	MCODE	26	0.16	0.17	0.98
	MCL	739	0.21	0.22	0.90
	CFinder	414	0.18	0.28	0.69

对于小鼠，由于无法获得基准复合物，因此只对比识别的复合物个数。

Table 2 Comparison of prediction results of clustering algorithms

表 2 各聚类算法预测结果对比

数据集	算法	复合物个数
Mouse	CFM	1459
	WCOACH	956
	ClusterONE	811
	SPICi	361
	MCODE	86
	MCL	430
	CFinder	414

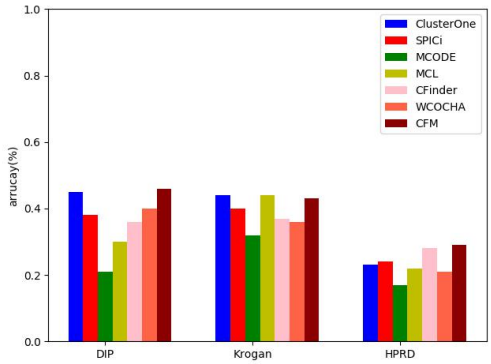

Fig. 4 Evaluation of accuracy

图 4 Accuracy 评估

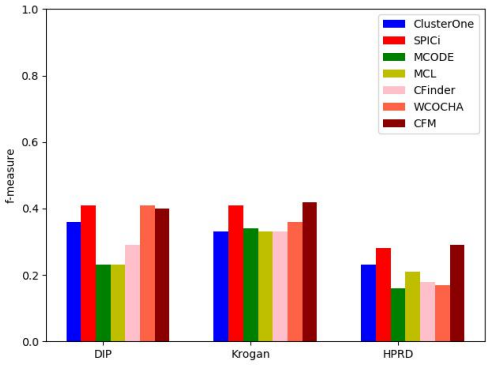

Fig. 5 Evaluation of F-measure
图 5 F-measure 评估

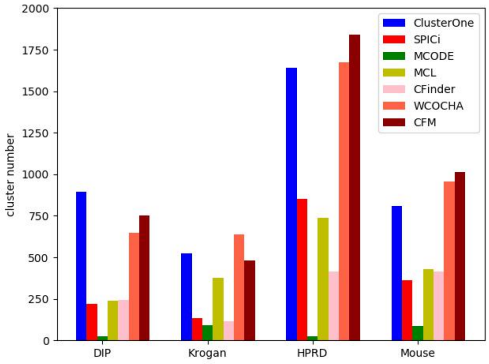

Fig. 6 Comparison of the cluster number
图 6 cluster 数量对比

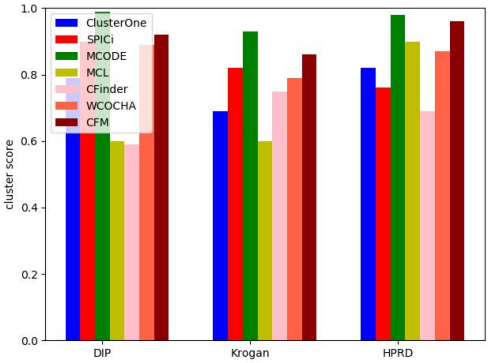

Fig. 7 Evaluation of Clustering score
图 7 聚类得分评估

6 总结

本文提出了一种基于蛋白质相互作用网络中拓扑特征和基因表达数据的的组合的 CFM 算法是一种有效的检测蛋白质复合物方法，有助于生物学家更好地了解蛋白质复合物。实验结果表明，本文提出的 CFM 方法针对大型生物网络在准确度和生物学意义方面优于许多已有的蛋白质复合物检测方法．因此，本文的方法是可行且高效的．

参 考 文 献

[1] OU-YANG L, WU M, ZHANG X F, et al. A two-layer integration framework for protein complex detection[J]. BMC bioinformatics, 2016, 17(1): 100.

[2] SHEN X, YI L, JIANG X, et al. Mining temporal protein complex based on the dynamic pin weighted with connected affinity and gene co-expression[J]. PloS one, 2016, 11(4): e0153967.

[3] PIZZUTTI C, ROMBO S E. Algorithms and tools for protein‐protein interaction networks clustering, with a special focus on population-based stochastic methods[J]. Bioinformatics, 2014, 30(10): 1343-1352.

[4] SHARMA A, ALI H H. Analysis of clustering algorithms in biological networks[C]// IEEE International Conference on Bioinformatics and Biomedicine. IEEE Computer Society, 2017:2303-2305.

[5] SUSYMARY J, LAWRANCE R. Graph theory analysis of protein-protein interaction graphs through clustering method[C]//2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS). IEEE, 2017: 1-5.

[6] SIKANDAR A , ANWAR W , BAJWA U I , et al. Decision Tree Based Approaches for Detecting Protein Complex in Protein Protein Interaction Network (PPI) via Link and Sequence Analysis[J]. IEEE Access, 2018:1-1.

[7] BADER G D, HOGUE C W V. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC bioinformatics, 2003, 4(1): 2-0.

[8] ENRIGHT A J, VAN DONGEN S, OUZOUNIS C A. An efficient algorithm for large-scale detection of protein families[J]. Nucleic acids research, 2002, 30(7): 1575-1584.

[9] PALLA G, DERE N I, FARKAS I, et al. Uncovering the overlapping community structure of complex networks in nature and society[J]. Nature, 2005, 435(7043): 814-818.

[10] ADAMCSEK B, PALLA G, FARKAS I J, et al.CFinder: locating cliques and overlapping modules in biological networks[J]. Bioinformatics (Oxford, England),2006, 22(8):1021-1023.

[11] JIANG P, SINGH M. SPICi: a fast clustering algorithm for large biological networks[J]. Bioinformatics, 2010, 26(8): 1105-1111.

[12] NEPUSZ T, YU H, PACCANARO A. Detecting overlapping protein complexes in protein-protein interaction networks[J]. Nature methods, 2012, 9(5): 471.

[13] KOUHSAR M, ZARE-MIRAKABAD F, JAMALI Y. WCOACH: Protein complex prediction in weighted PPI networks[J]. Genes & genetic systems, 2016: 15-00032.

[14] SHAMMA P, BHATTACHARYYA D K, KALITA J K. Detecting protein complexes based on a combination of topological and biological properties in protein-protein interaction network[J]. Journal of Genetic Engineering and Biotechnology, 2018, 16(1): 217-226.

[15] RADIVOJAC P, CLARK W T, ORON T R, et al.A large-scale evaluation of computational protein function prediction[J].Nature Methods, 2013, 10 (3) :221-227.

[16] ZHANG, Y, L H, YANG Z, WANG J ,Li Y, Xu B. Protein complex prediction in large ontology attributed protein-protein interaction networks. IEEE/ACM Trans. Comput. Biol. Bioinform. 2013, 10, 729–741.

[17] OU-YANG L , YAN H , ZHANG X F . A multi-network clustering method for detecting protein complexes from multiple heterogeneous networks[J]. BMC Bioinformatics, 2017, 18(S13):463.

[18] HAQUE M, SARMAH R, BHATTACHARYYA D K. A common neighbor based technique to detect protein complexes in PPI networks[J]. Journal of Genetic Engineering and Biotechnology, 2018, 16(1): 227-238.

[19] XENARIOS I, SALWINSKI L, DUAN X J, et al. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions[J]. Nucleic acids research, 2002, 30(1): 303-305.

[20] KROGAN N J, CAGNEY G, Yu H, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae[J]. Nature, 2006, 440(7084):637-643.

[21] PERI S, NAVARRO J D, AMANCHY R, et al. Development of human protein reference database as an initial platform for approaching systems biology in humans[J]. Genome research, 2003, 13(10): 2363-2371.

[22] CALDERONE A, CASTAGNOLI L, CESARENI G. Mentha: a resource for browsing integrated protein-interaction networks[J]. Nature methods, 2013, 10(8): 690.

[23] MEWES H W, AMID C, AMOLD R, et al. MIPS: analysis and annotation of proteins from whole genomes.[J]. Nucleic Acids Research, 2006, 34(Database issue):169-72.

[24] KIKUGAWA S, NISHIKATA K, MURAKAMI K, et al. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from h-invitational protein-protein interactions integrative dataset[C]//BMC systems biology. BioMed Central, 2012, 6(2): S7.

Zhang Dongwen, born in 1964. PhD,professor.Member of China Computer Federation. Her main research interests include Web mining, information retrieval.

Wang Peiheng, born in 1995. MS.His main research interests include Web mining, machine learning.

Xu Yunfeng, born in 1980. MS, associate professor. Member of China Computer Federation. His main research interests include complex network module analysis, cloud computing, data mining.