Computer Science and Application 计算机科学与应用, 2018, 8(8), 1300-1308
Published Online August 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.88140

Hans汉斯

Identifying Protein Complexes Based on Gene Ontology and Core-Attachment Structure

Yang Yu

Software College, Shenyang Normal University, Shenyang Liaoning
Email: yuyangsd1204@126.com

Received: Aug. 5th, 2018; accepted: Aug. 20th, 2018; published: Aug. 28th, 2018

Abstract

Protein complexes are composed of a group of proteins with specific biological functions. Computational methods for protein complexes prediction from biological networks have important practical implications for understanding the mechanisms of biological activity and the pathogenesis of diseases. Some of traditional algorithms are usually based only on network topologies, ignoring the impact of biological information and noise data on complex prediction. Aiming at this problem, we propose a protein complex identification algorithm based on gene ontology and core-attachment structure. Firstly, a weighted graph model is constructed based on semantic similarity by combining protein interaction network with gene ontology information. Secondly, a complex identification algorithm GCA is designed with local subgraph diameter and density as clustering conditions. Finally, GCA is compared with three methods in two real complex data sets. The experimental results indicate that GCA performances significantly better than CFinder, MCode and MCL in terms of recall, f-measure and functional enrichment analysis.

Keywords

Gene Ontology, Core-Attachment Structure, Weighted Network

基于基因本体和核-附属的蛋白质复合物识别算法

于　杨

沈阳师范大学，辽宁　沈阳
Email: yuyangsd1204@126.com

文章引用：于杨. 基于基因本体和核-附属的蛋白质复合物识别算法[J]. 计算机科学与应用, 2018, 8(8): 1300-1308.
DOI: 10.12677/csa.2018.88140

收稿日期：2018年8月5日；录用日期：2018年8月20日；发布日期：2018年8月28日

摘　要

蛋白质复合物由一组具有特定生物功能的蛋白质组成。使用计算方法从生物网络中预测蛋白质复合物对于理解生物活动的机制和疾病的发病机理具有重要的现实意义。传统的复合物识别算法通常仅基于网络拓扑结构，忽略生物特征和噪声数据对复合物识别性能的影响。针对该问题，本文提出一种基因本体和核-附属结构的蛋白质复合物识别算法，首先通过语义相似性融合蛋白质相互作用网络和基因本体信息构建有权图模型；其次，设计以局部子图直径和密度为聚类条件的核-附属结构的复合物识别算法GCA。最后，GCA和三个经典的方法在两个复合物数据集中进行比较和分析。实验结果表明，GCA在召回率、f度量和功能富集分析方面的表现均显著优于CFinder，MCode和MCL。

关键词

基因本体，核-附属结构，有权网络

1. 引言

蛋白质复合物是由一组具有功能相似的蛋白质组成，它是细胞组织形成和功能实现的关键细胞实体。因此，预测蛋白质复合物可以从系统水平上更好地理解细胞的基本组成和组织机制。通过实验方法检测蛋白质复合物通常比较昂贵并耗费大量时间。而且，有些复合物只能在特定的条件下才能被检测。近年来，高通量实验技术和机器学习的方法产生了大量的蛋白质相互作用数据(PPI)。基于复杂网络的聚类技术、机器学习理论和群智能算法[1]的应用为蛋白质复合物识别奠定了理论基础和技术指导。因此，设计有效的聚类方法识别相互作用网络中的复合物是生物网络分析的一个关键问题。

复合物主要特征：1) 复合物内蛋白质具有结构和功能一致性；2) 在拓扑方面复合物具有相对较高的密度。复合物识别任务转化为利用聚类算法从图中识别蛋白质节点集合。整个 PPI 网络可以转化为一个图[2]。根据网络的性质，将其分为有权网络和无权网络，在有权 PPI 网络中，两个节点之间边的权值由这两个节点之间存在的相互作用关系的可能性大小表示。在无权的 PPI 网络中，所有边权值相同，一般使用二进制的邻接矩阵表示蛋白质节点之间的关系，如果两个蛋白质节点之间没有相互作用关系则用 0 表示，否则用 1 表示。研究发现，在有权网络上进行实验更容易识别出蛋白质复合物。

近年来，复合物识别算法通常基于图的算法聚类 PPI 网络。这些方法主要基于 PPI 中拓扑属性识别复合物，包括：MCODE [3]、CFinder [4]、MCL [5]、FLCD [6]和 CDRWR [7]。MCODE 将 PPI 网络中具有相对高密度的区域定位为蛋白质复合物。CFinder 基于团渗透方法在 PPI 网络中识别重叠蛋白质复合物。文献[5]引入马尔可夫聚类算法识别 PPI 网络中高度联通的蛋白质复合物。文献[8]引入随机游走马尔科夫聚类实现在有权或无权图图中识别蛋白质复合物。FLCD 将 PPI 网络转化为有向无环图，确定子图的搜索空间，能够在加权/无权 PPI 网络挖掘网络中高连通子图。CDRWR 使用相似性计算 PPI 中边的权值，并扩展重要种子节点以确保复合物的完整性，然后通过密度选择外部节点，通过合并策略形成最终的复

合物。这些方法通常只关注单一的网络拓扑特征，其使用的 PPI 数据中含有的噪声影响预测结果。基于这些不足和复合物的特征，本文提出一种基因本体和核-附属结构的蛋白质复合物识别算法，主要工作包括：1) 通过语义相似性融合蛋白质相互作用网络和基因本体信息构建有权图模型；2) 设计以局部子图直径和密度为聚类条件的核-附属结构的复合物识别算法 GCA；3) 比较并分析 GCA 与三个经典的方法，实验结果表明，GCA 在召回率、f 度量和功能富集分析方面的表现均显著优于 CFinder，MCode 和 MCL。

2. 基于基因本体和核-附属的蛋白质复合物识别算法

本文中 PPI 网络用无向有权图 $G=(V,E)$ 表示，其中 V 是网络中蛋白质构成的节点集合，E 是相互作用的蛋白质构成的有权边集合。蛋白质复合物识别工作包括有权网络模型构建和基于核-附属的复合物识别算法。

2.1. 有权网络模型构建

有效的考虑网络权重可以进一步提升蛋白质复合物识别的性能。基因本体(GO)是跨物种的综合资源，描述与生物过程，分子功能和细胞成分相关的基因和基因产物生物学特性，提供了蛋白质对之间的功能关系[9]。本文首先使用语义相似性构建基于基因本体和蛋白质相互作用信息的有权网络模型，通过最佳匹配(BMA)策略[10]计算每个行和列的所有最大相似度的平均值[11]，用于计算图中边的权重，如公式(1)-(3)，其中 A/B 分别表示 DIP 数据中蛋白质对应的基因，$GO(A)=\{GO_{A1},GO_{A2},\cdots,GO_{Am}\}$ 表示由 GO 注释得到关于蛋白质 A 的集合，$GO(B)=\{GO_{B1},GO_{B2},\cdots,GO_{Bm}\}$ 表示由 GO 注释得到关于蛋白质 B 的集合。

$$sim_{MAX}(A,B)=MAX_{t_1\in GO(A),t_2\in GO(B)}\left(sim(t_1,t_2)\right) \tag{1}$$

$$sim_{AVG}(A,B)=AVG_{t_1\in GO(A),t_2\in GO(B)}\left(sim(t_1,t_2)\right) \tag{2}$$

$$sim_{BMA}(A,B)=\frac{AVG_{t_1}\left(MAX_{t_2}sim(t_1,t_2)\right)+AVG_{t_2}\left(MAX_{t_1}sim(t_1,t_2)\right)}{2} \tag{3}$$

其次，基于公式(4)~(6)，根据基因本体的分子功能、生物过程和细胞组成三个不同特性，分别构建基于分子功能的有权图模型 M，基于生物过程的有权图模型 O，基于细胞组成的有权图模型 C。最后，根据公式(7)，采用均值方式融合三个模型 M，O 和 C，构建基于基因本体的有权网络模型 G。

$$M=\begin{cases}M_{ij}=sim_{BMA}(i,j), & \text{if } (i,j)\in DIP\\ M_{ij}=0, & \text{else}\end{cases} \tag{4}$$

$$O=\begin{cases}O_{ij}=sim_{BMA}(i,j), & \text{if } (i,j)\in DIP\\ O_{ij}=0, & \text{else}\end{cases} \tag{5}$$

$$C=\begin{cases}C_{ij}=sim_{BMA}(i,j), & \text{if } (i,j)\in DIP\\ C_{ij}=0, & \text{else}\end{cases} \tag{6}$$

$$G=\begin{cases}G_{ij}=0, & \text{if } M_{ij}=0 \text{ and } O_{ij}=0 \text{ and } C_{ij}=0\\ G_{ij}=\dfrac{M_{ij}+O_{ij}+C_{ij}}{3}, & \text{else}\end{cases} \tag{7}$$

2.2. 基于核-附属的蛋白质复合物识别模型

蛋白质网络中相同簇中的蛋白质通常比外部簇中的蛋白质具有更紧密地相互作用；生物网络具有小世界的特性。因此，在复合物识别的聚类过程中，采用子图密度和子图直径为聚类约束条件，其中密度计算如公

式(8)， $\deg(v)$ 是有权图中节点 v 的度， $w(e)$ 是网络中的边 e 的权值， $|N|$ 为有权网络中节点的个数。

$$\deg(v)=\sum_{e=(i,j)\in E} w(e)$$
$$\text{density}(S)=\frac{2*\deg(v)}{|N|*(|N|-1)} \tag{8}$$

算法过程描述如下

基于基因本体和核-附属的蛋白质复合物识别算法
输入：G=（V，E，δ，λ）
输出：识别的复合物集合 CL；
第一步：核心蛋白质复合物识别；
1):准备数据和构建有权图 G；
2):初始化蛋白质复合物集合 CL；
3):计算 G 中节点的度，并将节点度值按降序排列存入数组 D；
4):从数组 D 中选择度最大的节点 D[i]为种子节点，形成初始化簇 C_i；
5):选择簇 C_i 的其他邻接点加入到 C_i 中；
6):更新簇 C_i
7):i=i+1，返回 4)直到 D 空结束。
更新簇 C_i
a) 计算当前簇的所有邻居节点的度，选度最大的节点 V_i。
b)如果 V_i 满足条件(9)，将 V_i 加入 C_i，更新 C_i；否则继续选择 C_i 的其他邻接点尝试加入 C_i 中；
c)如果当前 C_i 不能继续更新，则输出 C_i，并将 C_i 加入集合 CL 中。
第二步：
1)将每个核心蛋白质复合物 C_i 映射到 DIP 蛋白质相互作用网络 G 中得 OC_i；
2)对于每一对 OC_i,如果 OC_i 与 C_i 之间满足 ω<=t(t=0.2);
$complex(CL_i)=core(C_i)$U attachment(OC_i)，输出识别蛋白质簇 CL_i。

算法包括四个步骤：节点度的计算和排序，种子节点的选取，核心簇的形成以及核心-附属复合物的形成。其中核心簇的形成由两个约束条件确定：密度和直径。如果以当前节点为种子节点形成簇的过程中满足条件(9)，则将其加入到当前簇中；否则继续找其他节点，直到不满足条件(9)为止，形成一个核心簇。然后选择新的种子节点，重复这个过程，直到形成所有的核心复合物。根据文献经验得知[12] [13]，通常设置 $\delta=0.7$， $\lambda=2$。核心-附属算法以蛋白质复合物通常具有高度相互作用的蛋白质的密集核心理论的基础。此类模型识别步骤分两步，首先发现网络中高度连接区域，然后通过添加强关联邻居来扩展这些区域。

$$\text{diameter}\le\delta \text{ and density}\ge\lambda \tag{9}$$

3. 实验与结果分析

3.1. 数据集及评价标准

文中实验使用 CYC2008 [14]和 MIPS [15]作为基准复合物数据集评估不同方法预测的复合物的性能。酵母蛋白质相互作用数据 DIP [16]和基因本体论(GO)信息用于构建有权图模型。

给定一组已知的真实蛋白质复合物集合 $R=\{R_1,R_2,\cdots,R_n\}$ 和一组已识别到的蛋白质簇集合 $I=\{I_1,I_2,\cdots,I_n\}$，如果定位到的蛋白质簇 I_i 与真实蛋白质复合物 R_j 满足条件 $\omega\ge t,t=0.2$，则认为 I_i 和 R_j 是匹配的。 $|T|$ 是识别的蛋白质复合物 I_i 和真实蛋白质复合物 R_j 之间交集的大小， $|I_i|$ 和 $|R_j|$ 分别为检测到的复合物集合 I 的大小和真实的复合物集合 R 的的大小。为了评估预测算法的性能，采用 recall，precision 和 f-measure 三个评测指标，如(11)~(13)。

$$\omega=\frac{|T|^2}{|I_i|*|R_j|} \tag{10}$$

$$\text{recall} = \frac{\left|\left\{R_j \middle| R_j \in R \wedge \exists I_i \in I, I_i \text{ matches } R_j\right\}\right|}{|R|} \tag{11}$$

$$\text{precision} = \frac{\left|\left\{I_i \middle| I_i \in I \wedge \exists R_j \in R, R_j \text{ matches } I_i\right\}\right|}{|I|} \tag{12}$$

$$\text{f-measure} = \frac{2 * \text{recall} * \text{precision}}{\text{recall} + \text{precision}} \tag{13}$$

3.2. 实验结果分析

为综合比较算法 GCA 的复合物识别性能，本文将 GCA 与 CFinder，MCode 和 MCL 三个算法就各自算法的准确率，召回率和 f 度量进行了比较。3 种方法的参数都采用文献中提供的默认参数值。如图 1 在数据集 CYC2008 上的表现性能所示，在准确率方面，GCA 弱于 MCode，说明本算法在噪声数据去除方面还有进一步的提升空间。但是综合蛋白质复合物识别性能评测的 3 个指标，GCA 在蛋白质识别的召回率和综合指标 f 度量方面都是最优的。召回率比较说明，本文算法 GCA 识别的复合物能够匹配的真实的复合物的数量最多。在 f 度量方面分别高于 CFinder，MCode 和 MCL 的百分点为 12%，21%和 17%。这说明与图中其他的复合物识别算法相比，本文算法在匹配真实复合物数量，综合表现性能方面具有更好的优势。分析图 2 可以得出类似结果。

3.3. 不同阈值 t 条件下算法性能比较

为了进一步说明 GCA 算法的有效性，图 3 和图 4 列出了四种算法在 9 个不同阈 $t = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ 条件下 f 度量的对比结果。从图 3 可以看出，本文提出的蛋白质复

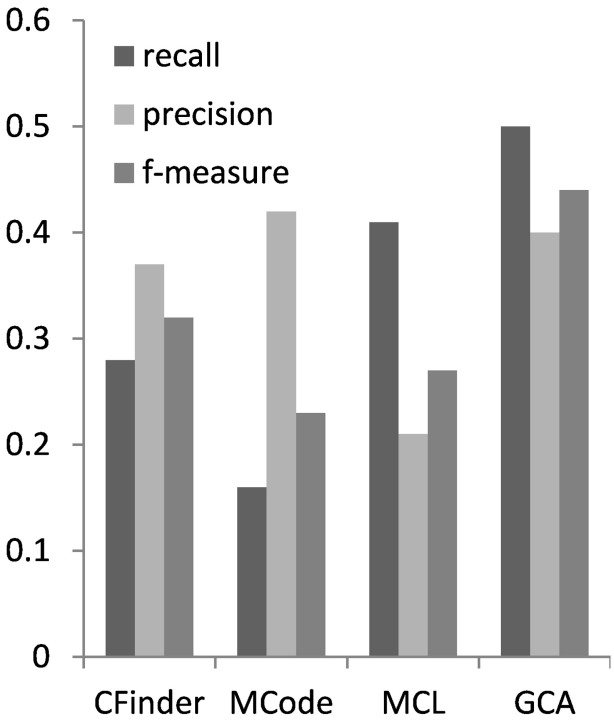

Figure 1. Performance comparison on CYC2008
图 1. 在 CYC2008 上性能比较

合物识别算法的匹配性能要显著优于其他算法。当匹配的阈值越低，算法复合物识别的 f 度量综合性能越高。当阈值越高，算法在复合物网络中 f 度量表现性能逐步下降。特别是在数据集 MIPS 上 GCA 较其

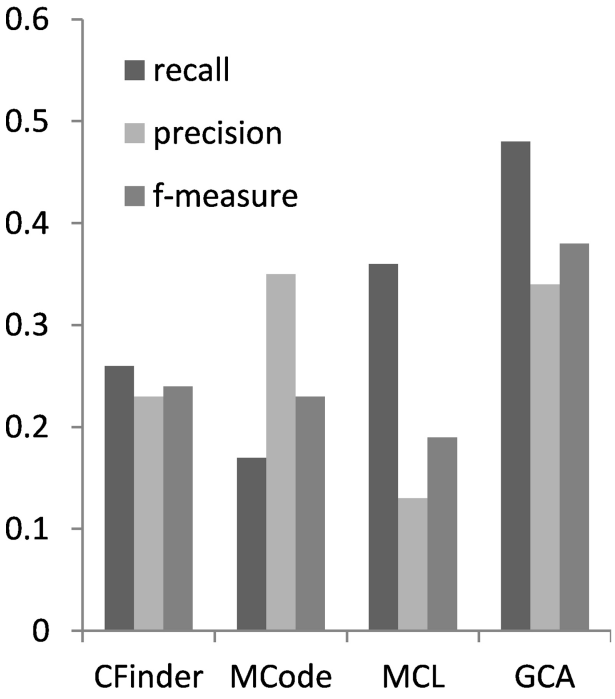

Figure 2. Performance comparison on MIPS
图 2. 在 MIPS 上性能比较

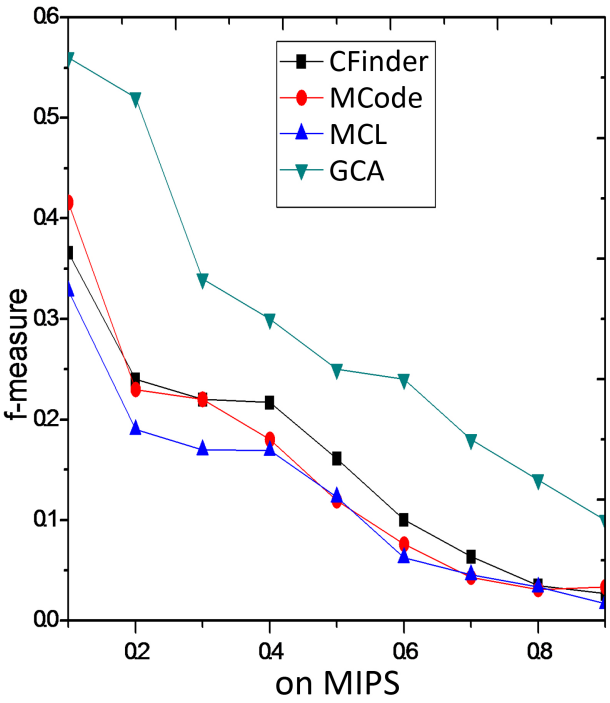

Figure 3. F-measure comparison on MIPS
图 3. 在 MIPS 上 F-measure 性能比较

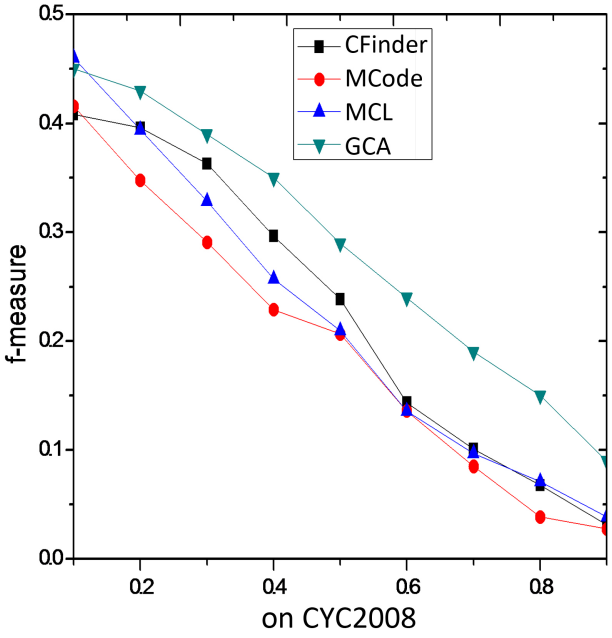

Figure 4. F-measure comparison on CYC2008
图 4. 在 CYC2008 上 F-measure 性能比较

他算法优势更加明显。图 4 在 CYC2008 上的比较结果与图 3 类似。综合分析，GCA 算法在酵母菌的网络 DIP 上的 f 度量表现整体性能要优于其他三类算法，具有较强的蛋白质复合物识别的能力，进一步说明引入基因本体的核-附属算法的有效性。

3.4. 功能富集分析

实验通过计算每个识别出蛋白质复合物的 P_{value} 值可以分析识别到的复合物所具备的生物意义。P_{value} 值反映的是蛋白质复合物所包含的某一特定生物功能的富集程度，可以注释出该复合物所具备的主要功能。P_{value} 值的计算公式如(14)：

$$P_{\text{value}} = 1 - \sum_{i=0}^{l-1} \frac{\binom{|F|}{i}\binom{|V|-|F|}{|C|-i}}{\binom{|V|}{|C|}} \tag{14}$$

其中，$|C|$ 表示识别的蛋白质复合物中蛋白质的个数，l表示复合物C中具有某一功能x的蛋白质个数。$|V|$ 表示蛋白质网络中蛋白质的个数，其中具有功能x的蛋白质的个数为 $|F|$。P_{value} 体现了识别复合物中蛋白质功能富集的概率。如果复合物被注释成好几个功能，则只取 P_{value} 最小时对应的功能。

本实验对从酵母 PPI 网络中识别蛋白质复合物进行功能富集分析进一步验证 GCA 算法的有效性，具体 P_{value} 值的分析和比较如表 1。每个复合物的 P_{value} 的范围从小到大划分成四个区间：<E-15，[E-15, E-10]，[E-10, E-5]，[E-5, 0.001]。P_{value} 值大于 0.001，通常认为复合物的功能被随机指派的可能性很大，基本没有生物意义。表 1 中括号内的百分数表示 P_{value} 落在某个区间的复合物数与落在所有区间的复合物数的比值。例如 GCA 总共预测到 219 个复合物，其中 P_{value} 落在区间<E-15 的复合物数占百分比为 21.75%。从表 1 中可以发现，在 P_{value} 最小的区间(即<E-15)，GCA 预测的复合物无论数量还是百分比都远远大于其他算法，如前所述，P_{value} 越小表示复合物的生物意义越显著，因此，这些复合物都具有最显著的生物意

Table 1. Comparison of functional enrichment
表 1. 功能富集性比较

算法	数量	有效性	<E-15	E-15-E-5	E-5-0.001
CFinder	112	84.7%	9.25%	35.24%	40.21%
MCode	40	91%	13.5%	61%	16.5%
MCL	265	63%	7.75%	28.45%	26.8%
GCA	219	95.85%	21.75%	62.47%	11.63%

义。GO 功能富集分析都说明 GCA 算法性能优于比其他算法。

4. 结论

在 PPI 网络中检测蛋白质复合物是生物医学领域的重要任务。因此，随着技术的进步，PPI 网络的增长速度比以往任何时候都快，这使得任务变得非常重要。在本文中，我们提出了一种基于基因本体和核-附属的有权图聚类方法，通过融合基因本体和蛋白质相互作用数据构建有权图模型，设计基于核-附属的算法进行复合物识别，可以有效地从复杂的网络中找出密集连接的子图。与 CFinder，MCode 和 MCL 相比，本文中提出的算法可以挖掘更多的蛋白质复合物，提高算法的识别性能。在今后进一步的研究工作中，我们将考虑有差异的利用不用的数据源信息构建有权图模型，并设计基于分类型的聚类识别方法。

基金项目

辽宁省教育厅科技项目(L201605)。

参考文献

[1] Lei, X., Ding, Y., Fujita, H. and Zhang, A. (2016) Identification of Dynamic Protein Complexes Based on Fruit Fly Optimization Algorithm. *Knowledge-Based Systems*, **105**, 270-277. https://doi.org/10.1016/j.knosys.2016.05.019

[2] Srihari, S., Yong, C.H., Patil, A. and Wong, L. (2015) Methods for Protein Complex Prediction and Their Contributions towards Understanding the Organization, Function and Dynamics of Complexes. *FEBS Letters*, **589**, 2590-2602. https://doi.org/10.1016/j.febslet.2015.04.026

[3] Bader, G.D. and Hogue, C.W. (2003) An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC Bioinformatics*, **4**, 2. https://doi.org/10.1186/1471-2105-4-2

[4] Adamcsek, B., Palla, G., Farkas, I.J., Derényi, I. and Vicsek, T. (2006) Cfinder: Locating Cliques and Overlapping Modules in Biological Networks. *Bioinformatics*, **22**, 1021-1023. https://doi.org/10.1093/bioinformatics/btl039

[5] Ochieng, P.J., Kusuma, W.A. and Haryanto, T. (2017) Detection of Protein Complex from Protein-Protein Interaction Network Using Markov Clustering. *International Symposia on Bioinformatics*, *Chemometrics and Metabolomics*, 1-13. https://doi.org/10.1088/1742-6596/835/1/012001

[6] Wang, Y. and Qian, X. (2017) Finding Low-Conductance Sets with Dense Interactions (flcd) for Better Protein Complex Prediction. *BMC Systems Biology*, **11**, 537-538. https://doi.org/10.1186/s12918-017-0405-5

[7] Jiang, J.W., Luo, C., Liang, J.H. and Chen, Q.F. (2017) Protein Complex Detection by Seed-Expansion Method Based on Random Walk with Restart.

[8] Van Dongen, S. (2000) Graph Clustering by Flow Simulation. Phd Thesis University of Utrecht.

[9] Zhang, S.B. and Tang, Q.R. (2016) Protein-Protein Interaction Inference Based on Semantic Similarity of Gene Ontology Terms. *Journal of Theoretical Biology*, **401**, 30-37. https://doi.org/10.1016/j.jtbi.2016.04.020

[10] Lin, D. (1998) In An Information-Theoretic Definition of Similarity. *International Conference on Machine Learning*, 296-304.

[11] Pesquita, C., Faria, D., Bastos, H., Ferreira, A.E., Falcão, A.O. and Couto, F.M. (2008) Metrics for Go Based Protein Semantic Similarity: A Systematic Evaluation. *BMC Bioinformatics*, **9**, S4. https://doi.org/10.1186/1471-2105-9-S5-S4

[12] Li, X., Wu, M., Kwoh, C.K. and Ng, S.K. (2010) Computational Approaches for Detecting Protein Complexes from

Protein Interaction Networks: A Survey. *BMC Genomics*, **11**, 1-19. https://doi.org/10.1186/1471-2164-11-S1-S3

[13] Yang, Y., Liu, J., Feng, N., Song, B. and Zheng, Z. (2017) Combining Sequence and Gene Ontology for Protein Module Detection in the Weighted Network. *Journal of Theoretical Biology*, **412**, 107-112. https://doi.org/10.1016/j.jtbi.2016.10.010

[14] Pu, S., Wong, J., Turner, B., Cho, E. and Wodak, S.J. (2009) Up-to-DATE catalogues of Yeast Protein Complexes. *Nucleic Acids Research*, **37**, 825-831. https://doi.org/10.1093/nar/gkn1005

[15] Mewes, H.W., Dietmann, S., Frishman, D., Gregory, R., Mannhaupt, G., Mayer, K.F., Münsterkötter, M., Ruepp, A., Spannagl, M. and Mips, S.V. (2008) Analysis and Annotation of Genome Information in 2007. *Nucleic Acids Research*, **36**, 196-201. https://doi.org/10.1093/nar/gkm980

[16] Xenarios, I., Salwínski, Ł., Duan, X.J., Higney, P., Kim, S.M. and Eisenberg, D. (2002) Dip, the Database of Interacting Proteins: A Research Tool for Studying Cellular Networks of Protein Interactions. *Nucleic Acids Research*, **30**, 303-305. https://doi.org/10.1093/nar/30.1.303