



CUBCO: Prediction of Protein Complexes Based on Min-cut Network Partitioning into Biclique Spanned Subgraphs

Sara Omranian^{1,2} and Zoran Nikoloski^{1,2}(✉)

¹ Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany

Nikoloski@mpimp-golm.mpg.de

² Systems Biology and Mathematical Modeling, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

Abstract. High-throughput approaches have generated large-scale protein-protein interaction (PPI) networks that are used in prediction of protein complexes. Here, we introduce CUBCO—a minimum cut-based algorithm that predicts protein complexes as biclique spanned subgraphs while relying on link prediction approaches to score and incorporate missing interactions. Our comprehensive analyses with PPIs from different organisms show that CUBCO performs on par with the best-performing approaches, that model protein complexes as biclique spanned subgraphs, and outperforms the remaining contenders. We also show that the usage of link prediction approaches in CUBCO improves the prediction of protein complexes on average 34.22% in all comparisons. Finally, CUBCO recovers $\sim 40\%$ and $\sim 11\%$ of known protein complexes from the Pan-Plant and Metazoan PPI networks. Therefore, CUBCO represents an efficient, parameter-free approach for accurate prediction of protein complexes from PPI networks.

Keywords: PPI networks · Protein complex prediction · Biclique spanned graph · Minimum cut

1 Introduction

Biological functions are supported by interactions between cellular components, including proteins. Protein-protein interaction (PPI) networks comprise nodes, corresponding to proteins, and edges, denoting protein interactions. The identification of protein complexes is important to understand the operational principles of cell systems, from signaling to metabolism. Several high-throughput technologies [1] have emerged to study PPIs, resulting in the assembly of large-scale PPI networks for well-studied, model organisms. However, due to the design and steps involved in these technologies, the resulting PPIs contain a sizeable portion of false-positive and false-negative interactions, which affect the identification of protein complexes [2].

Computational approaches are often used in the identification of protein complexes. These computational approaches have been grouped based on several criteria: (i) supervised vs. unsupervised, (ii) using only PPI network or also integrate other data [3].

To facilitate the comparison of these approaches and evaluate the predicted complexes, several gold standards, e.g. EcoCyc for *Escherichia coli* [4], MIPS, SGD, and CYC2008 for yeast [5–7], and CORUM for *H. sapiens* [8], have been assembled.

While there has been considerable improvement in the performance of computational approaches for prediction of protein complexes, there still some key challenges that remain open. The existing approaches rely on the idea that protein complexes correspond to highly connected clusters in PPI networks. As a result, they are often not capable of simultaneously identifying both dense and sparse, as well as, small and large protein complexes [3]. These approaches also depend on multiple parameters, which renders it difficult to interpret the resulting protein complexes. In contrast, the recently proposed PC2P and GCC-v [9, 10] represent parameter-free algorithms that model protein complex as biclique spanned subgraphs [11]. As a result, they can identify sparse as well as dense protein complexes independent of size, since biclique spanned subgraphs include stars, bicliques, and cliques as special graph classes. Nevertheless, if an approach for protein complex prediction uses only a given PPI network as input, its performance will be affected by erroneous and missing interactions in PPI networks. To overcome the issue of missing PPIs, link prediction algorithms have been recently proposed [12]. Therefore, the graph clustering and link prediction algorithms can be used jointly to improve protein complex prediction.

Building on this idea, here we introduce a new approach, referred to as CUBCO, that models protein complexes biclique spanned subgraphs, identified via minimum cut (unlike the local approaches in PC2P and GCC-v), and integrates link prediction to add most probable missing PPIs and investigate their effect on the performance of predicting protein complexes.

2 Results

2.1 CUBCO Predicts Protein Complexes Using Min Cut

Let $G = (V, E)$ be a graph with a set of nodes, V , corresponding to proteins, a set of edges, E , denoting PPIs. The simple graph denoted by $\overline{G} = (V, \{(u, v) | (u, v) \notin E\})$ denotes the complement of G . We formalize the concept of a protein complex by a biclique spanned graph, $G = (V, E)$, whose node set can be partitioned into two subsets, $V_1(G)$ and $V_2(G)$, and its edge set has the edges in the bipartite clique on $V_1(G)$ and $V_2(G)$ as a subset. Intuitively, a biclique spanned subgraph can be seen as a bipartite clique to which additional edges have been added. A biclique spanned graph has two properties: (i) the distance between any two nodes is at most two; (ii) Its complement, \overline{G} , is disconnected [13]. These properties provide a natural formalization of a network cluster based on connectedness, since the complement of a cluster defined this way is disconnected. Thereby, the goal is to partition the graph G , $C = \{C_1, C_2, \dots, C_k\}$, such that each C_i is a biclique spanned subgraph [11]. This can be obtained by utilizing not only local properties (e.g. second neighborhood or clustering coefficient [9, 10]), but also global properties of the graph.

The complement of graph, \overline{G} , contains edges that are not present in the original graph G . From biological perspective, the edges of \overline{G} include false-negative and

true-negative PPIs. Several studies have predicted the missing edges in PPI networks based on different concepts, with the network-based prediction of PPIs that relies on network walks of length three yielding the best results [12]. This approach favors inclusion of edges between nodes that are connected with a higher number of walks of length three in a given graph G . Here, we use the advantage of this approach, but rely on paths—instead of walks—of length three, to avoid effects of direct neighbors. Thereby, we assign weight to the edges of \overline{G} based on normalized number of paths of length three, given by:

$$w(u, v) = \sum_{i,j} \frac{P_{u,i}P_{i,j}P_{j,v}}{\sqrt{d_i d_j}}, \quad (1)$$

where $P_{i,j} = 1$ if nodes i and j are adjacent, and zero otherwise, and d_i denotes the degree of node i . Since one of the features of a biclique spanned subgraph is that its complement is disconnected, here we employ minimum cuts to discover biclique spanned subgraphs [14].

Given a graph G , CUBCO predicts protein complexes in three steps (Fig. 1): (i) determine \overline{G} , (ii) employ ideas from link prediction to weigh the edges in \overline{G} based on the degree-normalized number of path of length three between the end-nodes in G (Eq. (1), Algorithm 1), and (iii) iteratively identify minimum cuts in the edge-weighted graph \overline{G} (Algorithm 2), by using the Stoer-Wagner efficient, deterministic algorithm [15], until all resulting components are biclique spanned.

To render \overline{G} disconnected and obtain the biclique spanned subgraph, C_i , in G , we have to remove nodes, rather than edges (since edges in \overline{G} are not in G). Here, we used the min-cut algorithm that yields a partition of two node subsets, S_1 and S_2 , and a min-cut $E_{cut} = \{(u_i, v_i) | u_i \in S_1 \text{ and } v_i \in S_2, 1 \leq i \leq k\}$ where $k = |E_{cut}|$. Since, in practice, either S_1 or S_2 contain a single node, rather than considering all subsets for node removal, we consider the final biclique spanned subgraph to be either $C_1 = \{(S_1 \cup S_2) / \bigcup_{i=1}^k u_i\}$ or $C_2 = \{(S_1 \cup S_2) / \bigcup_{i=1}^k v_i\}$. The selection of the set C is guided by a score, Eq. (2), that shows the cohesiveness of the subgraph, $G[C]$, induced by C in G :

$$s(C) = \frac{|E_{in}(G[C])|}{|E_{out}(G[C])|}, \quad (2)$$

where $|E_{in}(G[C])|$ is the number of edges inside the subgraph, and $|E_{out}(G[C])|$ is the number of edges connecting the subgraph $G[C]$ to the rest of the network. CUBCO then selects C_i with the largest score and removes it from the graph \overline{G} . The complexity of CUBCO is $O(\frac{n^2}{d+1}(m + n \log n))$, where n is the number of nodes, m is the number of edges, and d is the minimum degree.

Algorithm 1 Preprocessing – Calculating edge weight in \bar{G}

```

1: procedure Preprocessing( $G$ )
2:    $\bar{G} \leftarrow \text{complement}(G)$ 
3:   for each  $\text{edge}(u, v)$  in  $\bar{G}$  do
4:      $w(u, v) \leftarrow$  normalized number of paths of length 3 in  $G$ 
5:     Update edge attribute of  $(u, v)$  in  $\bar{G}$ 
6:   return  $(\bar{G})$ 

```

Algorithm 2 CUBCO

```

1: procedure CUBCO( $\bar{G}$ )
2:   cluster_set  $\leftarrow []$ 
3:   While there is a node in  $\bar{G}$  do
4:      $(S_1, S_2) \leftarrow$  global min-cut Stoer-Wagner( $\bar{G}$ )
5:      $E_{\text{cut}} \leftarrow \{(u_i, v_i) | u_i \in S_1 \text{ and } v_i \in S_2, 1 \leq i \leq k\}$ 
6:      $C_i = \max(\text{score}((S_1 \cup S_2)/u_i), \text{score}((S_1 \cup S_2)/v_i))$ 
7:     Add  $C_i$  to cluster_set
8:     Remove  $C_i$  from  $\bar{G}$ 
9:   return (cluster_set)

```

2.2 Comparative Performance of CUBCO Without Link Prediction

We used twelve performance measures (Suppl. Information, see GitHub link in Method) to compare the predicted protein complexes from 17 approaches, including CUBCO, with all combinations of PPI networks and protein complexes from two *E. coli*, two *S. cerevisiae*, and one *H. sapiens* gold standards. The performance measures include: maximum matching ratio (MMR), fraction match (FRM), separation (SEP), positive predictive value (PPV), Sensitivity (SN), accuracy (ACC), precision, recall, F-measure, precision⁺, recall⁺, and F-measure⁺, with ranges between 0 and 1 and larger values indicating better performance (see Supplementary Information). We also calculated a composite scores given by the sum of MMR, FRM, ACC, and F-measure, that has been used in comparative analysis of approaches for protein complex prediction [16].

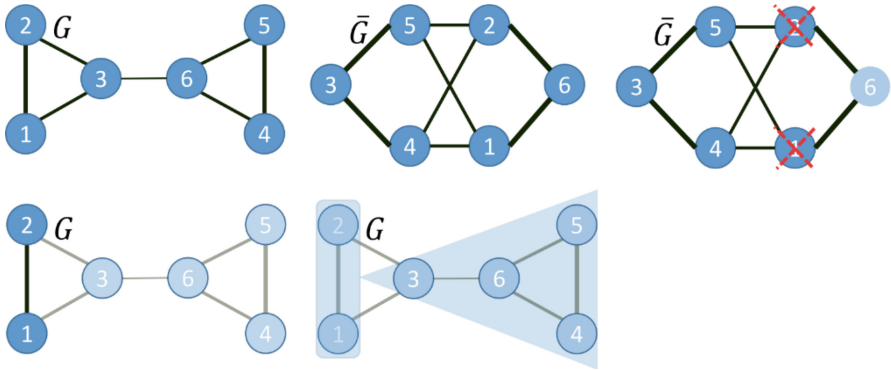


Fig. 1. Illustration of min-cut biclique spanned subgraph identification in CUBCO. The weight for each edge in the complement of the graph G is calculated based on Eq. (1). The global min-cut of the graph \bar{G} is obtained by Stoer-Wagner's algorithm. In this case, the algorithm separates node 6 from node-set $\{3, 4, 5\}$ by removing nodes 2 and 1. Hence, the first cluster $\{3, 4, 5, 6\}$ corresponds to a biclique spanned subgraph in G . After removal of this subgraph from G , the second cluster is given by $\{1, 2\}$.

For the combinations of Babu PPI networks and gold standards of *E. coli*, CUBCO resulted in the highest composite score (Fig. 2A, Tab. S3, Fig. S1, see GitHub link in Method). CUBCO exhibited composite score larger than half of the approaches for the combinations of Kong PPI network and gold standards of *E. coli* (Tab. S3, Fig. S1). For the combination of Gavin PPI network and the SGD gold standards in *S. cerevisiae* CUBCO exhibited the highest composite score in 62.5% of cases, preceded only by PC2P and GCC-v approaches that also represent protein complexes as biclique spanned subgraphs (Fig. 2B). In the remaining combinations of networks and gold standards CUBCO obtained a composite score higher than half of the compared approaches. More precisely, on average, the composite score of CUBCO is only 3.5% smaller than the composite score of the contenders ranked higher than CUBCO (Tab. S3, Fig. S1). In the case of *H. sapiens*, CUBCO exhibited the highest composite score among all other contenders for the PIPS PPI network and CORUM gold standard (Fig. 2C). However, for the combination of the STRING PPI network and CORUM, the composite score of CUBCO was on average 40.4% smaller than the contenders with a higher composite score (Tab. S3, Fig. S1).

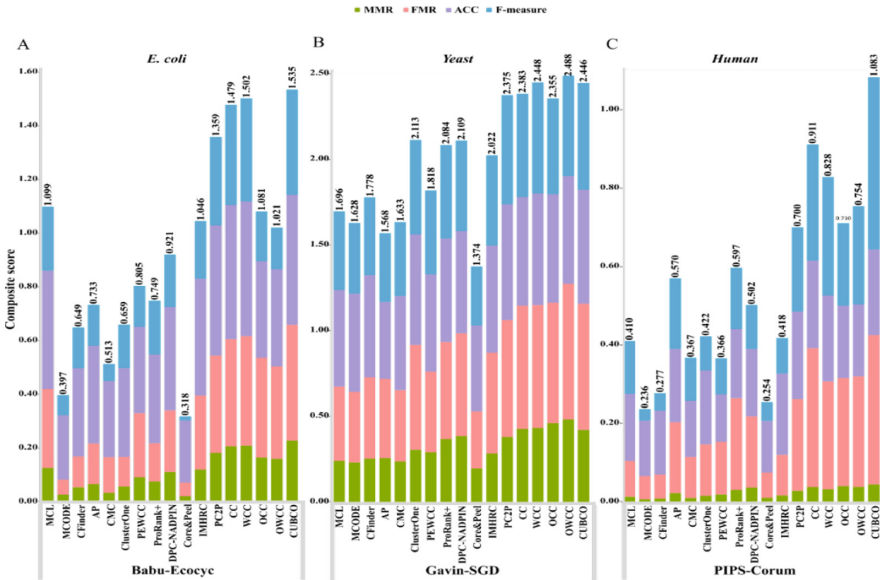


Fig. 2. Comparative analysis of approaches for prediction of protein complexes across PPI networks of different organisms. PPI networks for three organisms are considered (A) *E. coli*, (B) yeast, and (C) human. The comparative analyses are conducted with respect to a composite score that is a sum of MMR, FRM, ACC, and F-measure (see Supplementary file, Evaluation metric). Seventeen approaches, ordered by the year of publication, are compared on three PPI networks-gold standard combination. CUBCO outperforms all other approaches based on the composite score except in yeast, where it ranked third.

2.3 Integrating Link Prediction to Improve Performance of CUBCO

Next, we aimed to integrate a link prediction approach with CUBCO to resolve issues due to missing of PPIs from experimental approaches. To this end, we applied a modification of the L3 approach from [12], and ranked the missing edges by the number of paths of length 3 to avoid effects of immediate neighbors (Eq. (1)). We then selected the first 500, 1000, 1500, 2000, and 2500 of the ranked missing edges, and added them to the original network (as false-negative edges). Finally, we applied CUBCO on the modified network to predict the protein complexes. We found that the performance measures, and thereby, the composite score, increased slightly for all combinations of PPI networks and gold standards. The exception is the combinations of PPI networks of *E. coli* and KroganCore PPI network of yeast with Metabolic and SGD complexes, respectively. After careful investigation of these PPI networks, we realized that before inserting the new edges, they were of larger density and transitivity in comparison to all combinations of PPI networks of *E. coli* and KroganCore with Ecoyc and CYC2008 complexes. The result showed that the highest composite score for half of the combinations of PPI networks and gold standards was obtained by adding the first 500 ranked missing edges. For the combination of KroganExt PPI network of yeast with SGD complexes, the composite score improved gradually by

increasing the number of edges from 500 to 2500 (Fig. 3, Tabs. S4–6, Fig. S2). Overall, the finding implied that integrating of link prediction to graph clustering methods can enhance the predicted protein complexes.

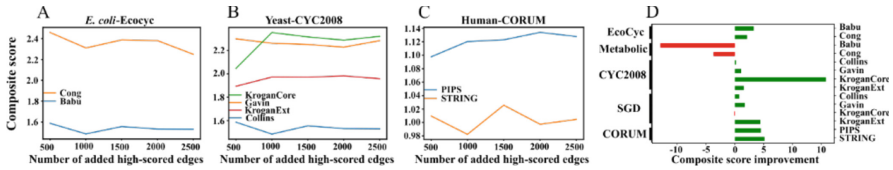


Fig. 3. Composite score due to integration of link prediction by adding the most probable edges. To investigate the composite score of CUBCO with the integration of link prediction, we inserted the first 500, 1000, 1500, 2000, and 2500 edges to the original PPI networks. The composite score is calculated for the combination of PPI networks and the gold standard of (A) *E. coli* and EcoCyc, (B) *S. cerevisiae* and CYC2008, and (C) *H. sapiens* and CORUM, respectively. On average, the composite score increases slightly with the increment of inserted edges across all PPI networks. (D) The original composite score of CUBCO across all combinations of organisms was compared with the obtained average composite score from inserting considered number of edges.

2.4 Prediction of New Protein Complexes in Conserved PPI Networks

We also applied CUBCO to predict new protein complexes in conserved PPI networks across 13 plant species, termed Pan-Plant [17], and across nine animal species, termed Metazoan [18] PPI network. For both PPI networks, we considered the high-confidence interactions with a score greater than 0.5 and used the gold standard, which is included in their corresponding studies (Tab. S1).

By applying CUBCO on the Pan-Plant PPI network, the resulting clusters fully coincided with 47 out of 118 known protein complexes (39.83%), and we predicted 249 new protein complexes that share no protein with the gold standards used (Tab. S7). To support the hypothesis that the proteins participating in a protein complex are involved in similar molecular functions and participate in the same cellular component and biological process, we determined the semantic similarity of the new protein complexes [19]. To this end, we employed the GOSim R package [20] to compute the median semantic similarity for every pair of proteins in each new protein complex. As a result, 77 of the predicted protein complexes (30.92%) showed GO semantic similarity of molecular functions equal to 1, indicating highest similarity of proteins function in the predicted protein complexes (Tab. S8). Further, we evaluated the predicted protein complexes by analyzing the domain-domain interactions (DDI) of proteins. We focused on the PPI network of *Arabidopsis thaliana*, identified the domains for $\sim 46\%$ of proteins based on the Pfam database [21], and considered the network of high-confidence (gold and silver) DDIs [22]. We found DDI support for 27 predicted protein complexes, of which 11 showed GO semantic similarity of 1 (Tab. S8). Finally, we only considered the predicted complexes that share no proteins with complexes in the gold standard, to obtain new predictions. We identified 165 new protein complexes, of which 19 have DDI support (Tab. S9). For instance, we found

that oxidative stress tolerance protein NQR (AT1G49670) forms a complex with AIM1 (AT4G29010) and ECHIA (AT4G16210). They share the same cellular compartment and are located in the peroxisome. One of the major functions of peroxisome is the Beta-oxidation of fatty acid [23], and AIM1 and ECHIA participate in this process, while NQR acts as a response to oxidative stress and co-regulated with Beta-oxidation genes [24]. Therefore, the predicted new complexes are of high quality and can form basis for future experimental confirmation.

Focusing on the Metazoan PPI network, CUBCO fully recovered 105 out of 981 known protein complexes (10.70%) and predicted 221 new protein complexes (Tab. S7). By conducting GO semantic similarity based on molecular function, 46 of the predicted protein complexes (20.81%) showed a value of 1 (Tab. S8). Based on the Pfam database, we then detected the domains for all proteins of the *H. sapiens* PPI network as part of the Metazoan network. We found that 88 of the predicted protein complexes have the DDI support, while only 15 of them also obtained GO semantic similarity of MF category equal to 1 (Tab. S8). Finally, we found only two new protein complexes that share no proteins from the gold standard, of which one is supported with DDI data (Tab. S9). The predicted complex with DDI support is the NXF1:NXT1 complex, which is also known as the TAP:p15 complex. This complex is a general mRNA nuclear export factor, and it is conserved from yeast to humans [25]. Altogether, the results indicate that CUBCO improves the performance of protein complex prediction and can find biologically meaningful protein complexes.

3 Method

3.1 Contending Approaches for Prediction of Protein Complexes

We compared the performance of CUBCO with sixteen other state-of-the-art methods, including: Markov Clustering (MCL), Molecular Complex Detection (MCODE), CFinder, Affinity Propagation (AP), Clustering-based on Maximal Cliques (CMC), Clustering with Overlapping Neighbourhood Extension (ClusterOne), PEWCC, Pro-rank⁺, Discovering Protein Complexes based on Neighbor Affinity and Dynamic Protein Interaction Network (DPC-NADPIN), Core&Peel, Inter Module Hub Removal Clustering (IMHRC), Protein Complexes from Coherent Partition (PC2P), and GCC-v (for references and software used, see Tab. S2). To facilitate fair comparison, the approaches are selected based on the public availability of their implementations and independence of any additional knowledge or data. Where applicable, we have used the default value of parameters.

3.2 PPI Networks and Gold Standards of Protein Complexes

We carried out all the experiments on PPI networks and gold standards of three model organisms: *E. coli.*, *S. cerevisiae.*, and *H. sapiens.* All the PPI networks are edge-weighted except one from *E. coli.* We used the two PPI networks of *E. coli.* from [26, 27], named by the first authors' names (Babu and Cong). The two gold standards were given by manually curated protein complexes from Ecocyc [4] and protein

complexes based on the genome-scale metabolic network of *E. coli* [28]. For yeast, we used: Collins [29], Krogan core (edge-weight ≥ 0.273), Krogan extended (edge-weight ≥ 0.101) [30], and Gavin [31] PPI networks. The two gold standards were retrieved from CYC2008 [7] and complexes derived from the Saccharomyces Genome Database (SGD) [6]. For *H. sapiens*, the two PPI networks were obtained from STRING (edge-weight ≥ 999) [32] and PIPS (edge-weight ≥ 25) [33]. In addition, we employed CORUM as the gold standard for human protein complexes [8] (Tab. S1). The comparative analysis of CUBCO and all the other approaches was performed on an Intel(R) Xeon(R) CPU E5-2670 v2 with 2.50GHz. CUBCO is freely available on GitHub at <https://github.com/SaraOmranian/CUBCO>, which also includes the supplementary figures and tables.

4 Conclusion

We proposed a new approaches, called CUBCO, to predict protein complexes from PPI networks. Since we have shown that partitioning a PPI network into biclique spanned subgraphs shows the best prediction performance of protein complexes [9, 10], CUBCO adopted the same concept. However, unlike the existing approaches, it relies on min-cuts, as a global network property, to determine the partition into biclique spanned subgraphs. CUBCO also employs paths of length three to rank false-negative interactions and include them in the networks to boost prediction performance. Future work will inspect the effect on coupling other approach for link prediction with approaches for prediction of protein complexes. In addition, we plan to consider weighted node cuts, rendering CUBCO applicable with large-scale proteomics and gene expression data.

References

1. Peng, X., Wang, J., Peng, W., Wu, F.-X., Pan, Y.: Protein–protein interactions: detection, reliability assessment and applications. *Brief. Bioinform.* **18**(5), 798–819 (2016). p. bbw066
2. Berger, B., Peng, J., Singh, M.: Computational solutions for omics data. *Nat. Rev. Genet.* **14**, 333–346 (2013)
3. Wu, Z., Liao, Q., Liu, B.: A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. *Brief. Bioinform.* **21**, 1531–1548 (2019)
4. Keseler, I.M., et al.: The EcoCyc database: reflecting new knowledge about *Escherichia coli*K-12. *Nucleic Acids Res.* **45**, D543–D550 (2016)
5. Mewes, H.W.: MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **32**, 41D – 44 (2004)
6. Hong, E.L., et al.: Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.* **36**, D577–D581 (2007)
7. Pu, S., Wong, J., Turner, B., Cho, E., Wodak, S.J.: Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **37**, 825–831 (2008)
8. Giurgiu, M., et al.: CORUM: the comprehensive resource of mammalian protein complexes —2019. *Nucleic Acids Res.* **47**, D559–D563 (2018)

9. Omranian, S., Angeleska, A., Nikoloski, Z.: PC2P: parameter-free network-based prediction of protein complexes. *Bioinformatics* **37**, 73–81 (2021)
10. Omranian, S., Angeleska, A., Nikoloski, Z.: Efficient and accurate identification of protein complexes from protein-protein interaction networks based on the clustering coefficient. *Comput. Struct. Biotechnol. J.* **19**, 5255–5263 (2021)
11. Angeleska, A., Nikoloski, Z.: Coherent network partitions. *Discret. Appl. Math.* **266**, 283–290 (2019)
12. Kovács, I.A., et al.: Network-based prediction of protein interactions. *Nature Commun.* **10**, 3 (2019)
13. Akiyama, J., Harary, F.: A graph and its complement with specified properties. IV. Counting self-complementary blocks. *J. Graph Theory* **5**, 103–107 (1981)
14. Dantzig, G.B., Fulkerson, D.R.: 12. On the max-flow min-cut theorem of networks. In: *Linear Inequalities and Related Systems*. (AM-38), pp. 215–222. Princeton University Press (1957)
15. Stoer, M., Wagner, F.: A simple min cut algorithm. In: van Leeuwen, J. (ed.) *ESA 1994*. LNCS, vol. 855, pp. 141–147. Springer, Heidelberg (1994). <https://doi.org/10.1007/BFb0049404>
16. Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9**, 471–472 (2012)
17. McWhite, C.D., et al.: A pan-plant protein complex map reveals deep conservation and novel assemblies. *Cell* **181**, 460–474.e14 (2020)
18. Wan, C., et al.: Panorama of ancient metazoan macromolecular complexes. *Nature* **525**, 339–344 (2015)
19. Cho, Y.-R., Hwang, W., Ramanathan, M., Zhang, A.: Semantic integration to identify overlapping functional modules in protein interaction networks. *BMC Bioinform.* **8**, 7 (2007)
20. Fröhlich, H., Speer, N., Poustka, A., Beißbarth, T.: GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinform.* **8**, 5 (2007)
21. Mistry, J., et al.: Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2020)
22. Ziaeddine, A.S., Amina, A.-N., Hiba, N., Ritchie, D.W., Marie-Dominique, D.: PPIDomainMiner: inferring domain-domain interactions from multiple sources of protein-protein interactions, March 2021
23. Shani, N., Jimenez-Sanchez, G., Steel, G., Dean, M., Valle, D.: Identification of a fourth half ABC transporter in the human peroxisomal membrane. *Hum. Mol. Genet.* **6**, 1925–1931 (1997)
24. Wiszniewski, A.A.G., Zhou, W., Smith, S.M., Bussell, J.D.: Identification of two Arabidopsis genes encoding a peroxisomal oxidoreductase-like protein and an acyl-CoA synthetase-like protein that are required for responses to pro-auxins. *Plant Mol. Biol.* **69**, 503–515 (2008)
25. Aibara, S., Katahira, J., Valkov, E., Stewart, M.: The principal mRNA nuclear export factor NXF1:NXT1 forms a symmetric binding platform that facilitates export of retroviral CTE-RNA. *Nucleic Acids Res.* **43**, 1883–1893 (2015)
26. Babu, M., et al.: Global landscape of cell envelope protein complexes in Escherichia coli. *Nat. Biotechnol.* **36**, 103–112 (2017)
27. Cong, Q., Anishchenko, I., Ovchinnikov, S., Baker, D.: Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019)
28. King, Z.A., et al.: BiGG models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* **44**, D515–D522 (2015)

29. Collins, S.R., et al.: Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteom.* **6**, 439–450 (2007)
30. Krogan, N.J., et al.: Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643 (2006)
31. Gavin, A.-C., et al.: Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006)
32. Szklarczyk, D., et al.: STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2014)
33. McDowall, M.D., Scott, M.S., Barton, G.J.: PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res.* **37**, D651–D656 (2009)