Computational and Structural Biotechnology Journal 19 (2021) 5255–5263

ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Efficient and accurate identification of protein complexes from protein-protein interaction networks based on the clustering coefficient

Sara Omranian [a,c], Angela Angeleska [b], Zoran Nikoloski [a,c,*]

[a] Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany
[b] Mathematics Department, University of Tampa, Tampa, FL, USA
[c] Systems Biology and Mathematical Modeling, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

ARTICLE INFO

Article history:
Received 14 May 2021
Received in revised form 13 September 2021
Accepted 13 September 2021
Available online 20 September 2021

Keywords:
Protein complexes
Protein-protein interaction
Network clustering
Species comparison

ABSTRACT

Identification of protein complexes from protein-protein interaction (PPI) networks is a key problem in PPI mining, solved by parameter-dependent approaches that suffer from small recall rates. Here we introduce GCC-v, a family of efficient, parameter-free algorithms to accurately predict protein complexes using the (weighted) clustering coefficient of proteins in PPI networks. Through comparative analyses with gold standards and PPI networks from *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens*, we demonstrate that GCC-v outperforms twelve state-of-the-art approaches for identification of protein complexes with respect to twelve performance measures in at least 85.71% of scenarios. We also show that GCC-v results in the exact recovery of ~35% of protein complexes in a pan-plant PPI network and discover 144 new protein complexes in *Arabidopsis thaliana*, with high support from GO semantic similarity. Our results indicate that findings from GCC-v are robust to network perturbations, which has direct implications to assess the impact of the PPI network quality on the predicted protein complexes.

1. Introduction

Protein-protein interactions (PPIs) govern many key cellular processes, from transcription and translation (e.g. pre-initiation complex and ribosomes) [1] to signaling and metabolism (e.g. protein kinase complexes and enzyme complexes) [2,3]. The collection of all PPIs in a given biological system is represented by a PPI network composed of nodes, denoting proteins, and edges, corresponding to the interactions between protein pairs. Systems biology has focused on assembling networks of PPIs across different organisms by combing computational and experimental approaches [4,5,6,7,8,9,10,11], with different resolution and quality [12,13,14,15,16]. The resulting PPI networks of increasing size and improved quality require the development of efficient algorithms for their mining [17].

A key problem in PPI network mining is that of identification of protein complexes. At the core of the algorithms to solve this problem is the identification of network clusters with particular properties. The existing network-based approaches for identification of protein complexes can be categorized into three groups: cluster-quality-based [18,19], node-affinity-based [20,21,22,23,24], and ensemble clustering methods [25,26]. The performance assessment of these algorithms has been facilitated by the assembly of gold standards of protein complexes (e.g. EcoCyc for *Escherichia coli* [27], MIPS, SGD, and CYC2008 for S. cerevisiae [28,29,30], and CORUM for H. sapiens [31]).

Mounting evidence based on these PPI networks and gold standards has pointed out that the existing methods tend to predict dense and large protein complexes; however, the vast majority of real protein complexes are small and sparse [32]. In addition, comparative analyses have demonstrated that these approaches are not able to predict high-confidence clusters and suffer from small recall [33]. This observation has led to the design of algorithms to identify sparse [34,35] and small complexes [36,37], which have slightly improved the recall of protein complexes. Yet, these algorithms depend on multiple parameters, which render it difficult to gauge the performance in absence of optimal parameter values for different combinations of PPI networks and gold standards. It was recently shown that a parameter-free approach, that models protein complexes as biclique spanned subgraphs, outperforms the existing, seminal approaches [38] and allows for the identification

[*] Corresponding author at: Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany; Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany.

E-mail address: nikoloski@mpimp-golm.mpg.de (Z. Nikoloski).

https://doi.org/10.1016/j.csbj.2021.09.014
2001-0370/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

of both dense and sparse clusters; however, only in unweighted networks of limited size.

Due to the participation of proteins in one or many protein complexes embedded in PPI networks, it can be argued that algorithms for protein complex identification must be able to: (i) predict, with high confidence, both overlapping and non-overlapping clusters of varying density, (ii) consider edge weights that capture the confidence of interactions [39], and (iii) scale with the network size, to facilitate its application on genome-scale PPI networks and verification of the robustness of the resulting predictions. However, the existing approaches are tailored to simultaneously address only some of these requirements, and most importantly, depend on multiple parameters.

Here we propose a class of efficient greedy algorithms, collectively termed GCC-v, to predict protein complexes based on the concept of (weighted) clustering coefficient of nodes in a given PPI network. Our findings demonstrate that GCC-v partitions the network into biclique spanned subgraphs [40], thus allowing the prediction of both sparse and dense protein complexes. The reason why GCC-v partitions the network into biclique spanned subgraphs is that for each of these subgraphs, say H, there exists a node whose first neighborhood contains H; hence, H is necessarily spanned by a star, which is a biclique. Furthermore, GCC-v is parameter-free, easy to implement and is efficient even on large-scale networks. By design, GCC-v can incorporate edge weights, facilitating the prediction of more accurate protein complexes and can detect overlapping clusters. Our comprehensive comparative analysis shows that GCC-v outperforms twelve state-of-the-art approaches based on twelve well-established performance measures on PPI networks from *E. coli*, S. cerevisiae, and H. sapiens, based on the corresponding gold standards. We also show that the predictions of protein complexes resulting from GCC-v are robust against network perturbation. Finally, by using the pan-plant PPI in combination with analysis of domain-domain interactions, we demonstrate that GCC-v accurately predicts existing and new protein complexes in *Arabidopsis thaliana*.

2. Results

2.1. GCC-v algorithm combines clustering coefficient with network transformations

The PPI networks used in this study are weighted, undirected networks. Let $G = (V, E, w)$ be a network with set of nodes V corresponding to proteins, set of edges E denoting PPIs, and $w(e)$ corresponding to the weight (positive real number) of edge e that indicates the reliability of the interaction based on experimental and computational approaches. The set of neighbors consists of all the nodes that are connected to v by an edge. Due to the propensity of edge formation between nodes in the same neighborhood in PPI networks, the number of triangles is larger than in Erdos-Renyi random networks of same density [41,42]. The clustering coefficient [42] quantifies the number of triangles in which a node v participates, normalized by the maximum possible number of triangles in the vicinity of node v:

$$CC(v) = \frac{2t_v}{k_v \times (k_v - 1)}, \tag{1}$$

where t_v denotes the number of triangles that node v is involved, and k_v is a degree of node v. Hence, if none of the neighbors of v are connected, then according to Eq. (1) $CC(v) = 0$, and if all neighbors of v are connected to each other the $CC(v) = 1$ (rendering a clique).

Several studies [43,44,45] have proposed extensions to the clustering coefficient to consider edge weights. For instance, Onnela et al. defined a weighted clustering coefficient as the geometric average of subgraph edge weights (Eq. (2)) [44]. By this formulation, the contribution of each triangle is given by the product of normalized edge weights $\widetilde{w}_{u,v} = \frac{w_{uv}}{\max(W)}$:

$$\widetilde{CC}(v) = \frac{2}{k_v \times (k_v - 1)} \sum_{u,k} \left(\widetilde{w}_{v,u}\widetilde{w}_{u,k}\widetilde{w}_{k,v}\right)^{1/3}. \tag{2}$$

The clustering coefficient can readily be extended on the level of edges, by applying it to the line graph of a graph G [46]. In a line graph of G, $L(G)$, each node denotes an edge in G and there is an edge between two nodes in $L(G)$if the corresponding edges of G are adjacent (i.e. share a node) [47]. Further and in contrast to the existing approaches that entirely focus on clustering nodes, the edge communities approach predicts overlapping clusters, and helps in revealing the hierarchical organization in PPIs [48]. To obtain overlapping clusters by incorporating edge weights of graph G, one can determine the average score of two vertices (i.e., edges in G) at the endpoints of every edge in the line graph, $L(G)$.

Based on these concepts, we devised four different versions of a greedy algorithm based on the: (i) clustering coefficient (CC), (ii) weighted clustering coefficient (WCC), (iii) overlapping clustering coefficient (OCC), and (iv) overlapping weighted clustering coefficient (OWCC). Given a graph G, the greedy algorithm determines a score for every node based on the clustering coefficient (Eqs. (1) – (2)). Depending on whether the unweighted or weighted clustering coefficient is used to calculate the score for the nodes in the original or the line graph, we obtained the four different variants mentioned above. In the case of OCC and OWCC, the procedure continues by mapping each node (corresponding to an edge in the original graph G) in predicted clusters in a line graph into its corresponding nodes in the original graph G. Moreover, as a preprocessing step in OWCC, the average weight of two adjacent nodes in a line graph (i.e. two adjacent edges in the original graph G) is calculated and assigned to each edge of the line graph. The greedy algorithm selects a node with the highest score and removes its neighbors along with the node itself from the graph. In the next step, it updates the score of the nodes in the first neighborhood of the nodes in the identified cluster. This procedure is repeated as long as there are connected components in G (Fig. 1, Algorithm 1, see Methods).

2.2. GCC-v predicted protein complexes of high GO similarity across PPI networks

We compared the performance of the four versions of our greedy algorithm (GCC-v) with twelve state-of-the-art approaches, including: Markov Clustering (MCL) [24], Molecular Complex Detection (MCODE) [20], CFinder [21], Affinity Propagation (AP) [23], Clustering-based on Maximal Cliques (CMC) [22], Clustering with Overlapping Neighbourhood Extension (ClusterOne) [18], PEWCC [49], Prorank^{+} [50], Discovering Protein Complexes based on Neighbor Affinity and Dynamic Protein Interaction Network (DPC-NADPIN) [51], Core&Peel [19], Inter Module Hub Removal Clustering (IMHRC) [52], and Protein Complexes from Coherent Partition (PC2P) [38]. To facilitate fair comparison, we considered only approaches for which publicly available implementation exists and that do not rely on any additional knowledge (e.g. ontologies or gene expression data) (see Supplementary Table 2). We would like to note that PEWCC and DPC-NADPIN both rely on the concept of the clustering coefficient of a node, but unlike GCC-v, they do not consider a weighted variant.

The critical issue with all of the existing methods (except for PC2P) is that they all rely on at least one parameter. Optimizing the parameters is challenging, since it depends on both the PPI net-

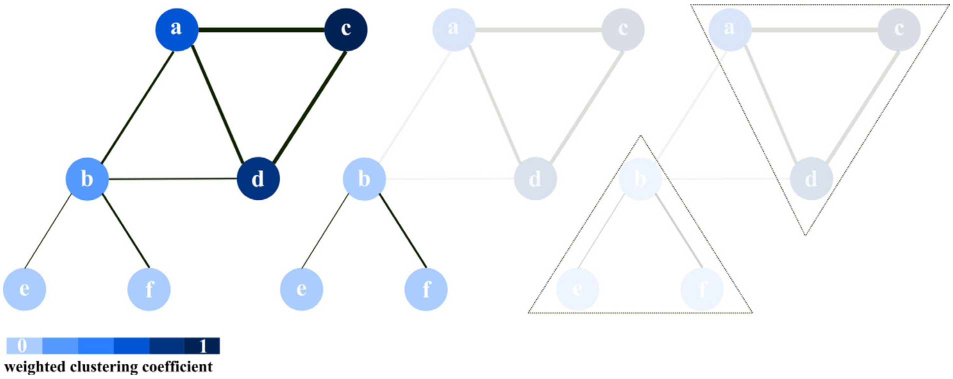

Fig. 1. Illustration of GCC-v. The toy network is composed of six nodes (a – f). The weighted clustering coefficient is calculated for each node based on Eq. (2). The node color indicates the value of the score from 0 to 1 (light to dark blue, see legend). The algorithm first selects node c and its neighbors as a first cluster since it has the largest score. The scores are updated in the next step, and node b together with its neighbors are selected in the second cluster. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

works and gold standards used as well as on the objective to be optimized. Therefore, optimizing each method based on different evaluation metrics yields different sets of predicted protein complexes, rendering it impossible to do meaningful interpretation of the findings. In the comparative analyses, we used the default parameters value for every contending algorithm. In contrast, the proposed greedy variants based on the clustering coefficient are parameter-free and are applicable to large-scale networks.

To compare the performance of the GCC-v with that of the above mentioned contenders, we first determined the GO semantic similarity of the predicted complexes in two PPI networks for *E. coli*, i.e. Babu [5] and Cong [7], four for S. cerevisiae, i.e. Collins [53], Gavin [4], KroganCore, and KroganExt [54], and two for H. sapiens, i.e. STRING [55] and PIPS [6] (see PPI networks and gold standards of protein complexes, Supplementary Table 1). This comparison provides support for the hypothesis that that proteins participating in a protein complex are involved in similar molecular functions and participate in the same cellular component. To this end, we computed the median semantic similarity, based on two different measurements, for every pair of proteins in each predicted complex for every considered method separately, and compared the distribution of these values over all predicted complexes. The first measure is graph-based [56] and the second one is information content-based [57]. We found that all the approaches show comparable distributions of GO semantic similarity across all datasets for the three GO categories, i.e., biological process (BP), cellular component (CC), and molecular function (MF) (Fig. 2 and Supplementary Figs. 1 and 2). Since the two measures lead to similar results, we only illustrate the results of the graph-based measures.

One factor that affects the distribution of median GO semantic similarity is the number of predicted clusters from different methods. Careful inspection revealed that the methods yielding a smaller number of predicted complexes tend to show, expectedly, narrower distributions for the median GO semantic similarity. However, CC and WCC algorithms showed the highest median GO semantic similarity for Collins and STRING PPI networks, in S. cerevisiae and H. sapiens, respectively, regarding both the BP and CC categories, and in Gavin PPI network of S. cerevisiae for the MF category. Furthermore, CC and WCC exhibited the largest median GO semantic similarity value of 1 regarding CC category in both Cong and Babu PPI network of *E. coli*. The aforementioned methods ranked second in the Collins and Gavin PPI networks S. cerevisiae concerning the median similarity regarding the MF and CC categories, respectively, as well as in the PIPS network H. sapiens for all three GO categories (see Supplementary Fig. 1).

2.3. GCC-v outperformed all contenders across combinations of PPI networks and gold standards

To assess the performance of GCC-v, we employed two *E. coli*, two S. cerevisiae, and one H. sapiens gold standard of protein complexes (see PPI networks and gold standards of protein complexes, Supplementary Table 1). We calculated twelve performance measures, including: 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$^+$ (see Evaluation metrics, Supplementary Files). The larger values for these scores are indicative of better performance. Moreover, to summarize these twelve performance measures, first we calculated a composite score that corresponds to the sum of four metrics, MMR, FMR, ACC, and F-measure [18,58,59,38,38]. Second, we calculated the MMR and F-measure$^+$ over predicted protein complexes with different overlap scores and employed their sum as suggested in [60].

For all combinations of two PPI networks and gold standards of *E. coli*, GCC-v resulted in the highest values for MMR, FRM, and recall$^+$ as well as the largest composite score. More precisely, OCC and OWCC ranked first for the combinations of Cong PPI network with both metabolic and EcoCyc gold standards; CC and WCC ranked first in combinations of Babu PPI network with metabolic and EcoCyc gold standards, respectively (Fig. 3A, and Supplementary Fig. 3, and Supplementary Table 3). The reason that GCC-v obtained a better composite score for Cong than the Babu PPI network with both gold standards is likely due to the smaller density of the latter, since Babu PPI network includes interactions of both high and medium confidence (Supplementary Table 1).

Likewise, for all combinations of protein interaction networks and gold standards in S. cerevisiae, GCC-v exhibited the highest MMR and FRM values as well as the highest recall (Fig. 3B, Supplementary Table 3). Thereby, it is not surprising that the GCC-v exhibited the best composite score and outperformed the other approaches in six out of eight combinations of PPI networks and gold standards for S. cerevisiae. More precisely, in four out of eight combinations, WCC showed the best composite score, and in the two other cases, CC and OWCC exhibited the highest composite score. The only other parameter-free approach, PC2P, also based on modeling of protein complexes as biclique spanned subgraphs, ranked first regarding the composite score for the combinations of KroganExt with CYC2008 and KroganCore with SGD (Supplementary Table 3, and Supplementary Fig. 3). The average composite scores of GCC-v are 5.9% and 2.5% smaller than PC2P in the former

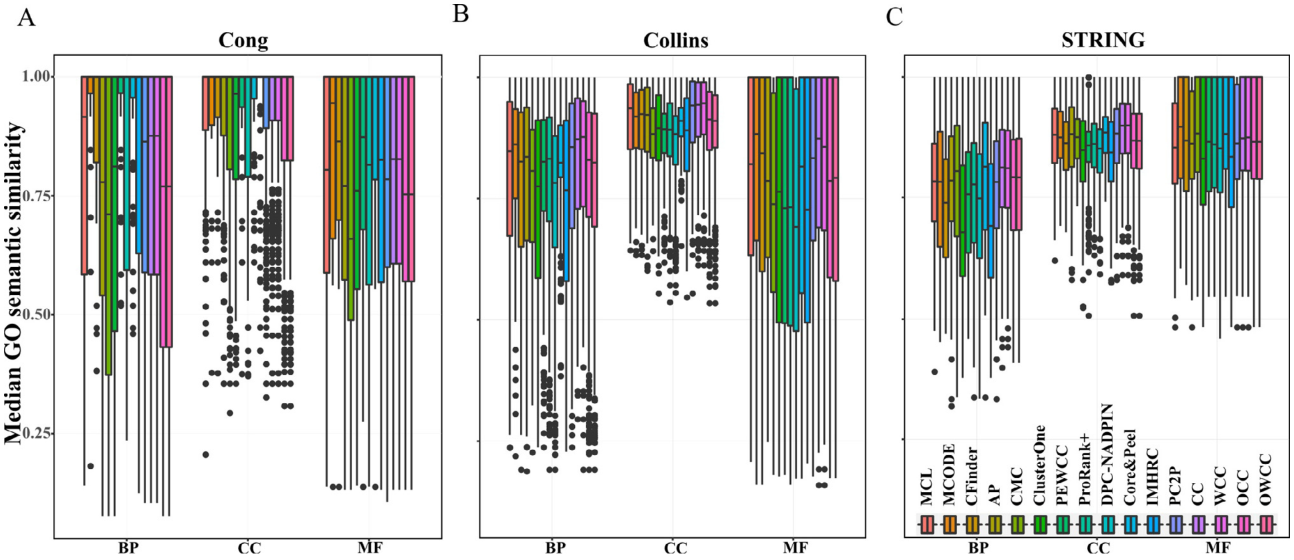

Fig. 2. Comparison of median GO semantic similarity for predicted protein complexes. GCC-v are compared against twelve algorithms (ordered by the year of publication) concerning the distribution of median GO semantic similarity over all clusters, for (A) Cong PPI network of *E. coli*, and (B) Collins PPI network of *S. cerevisiae*, and(C) STRING PPI network of *H. sapiens*. The GO semantic similarity is determined for its three categories: biological process (BP), cellular component (CC), and molecular function (MF).

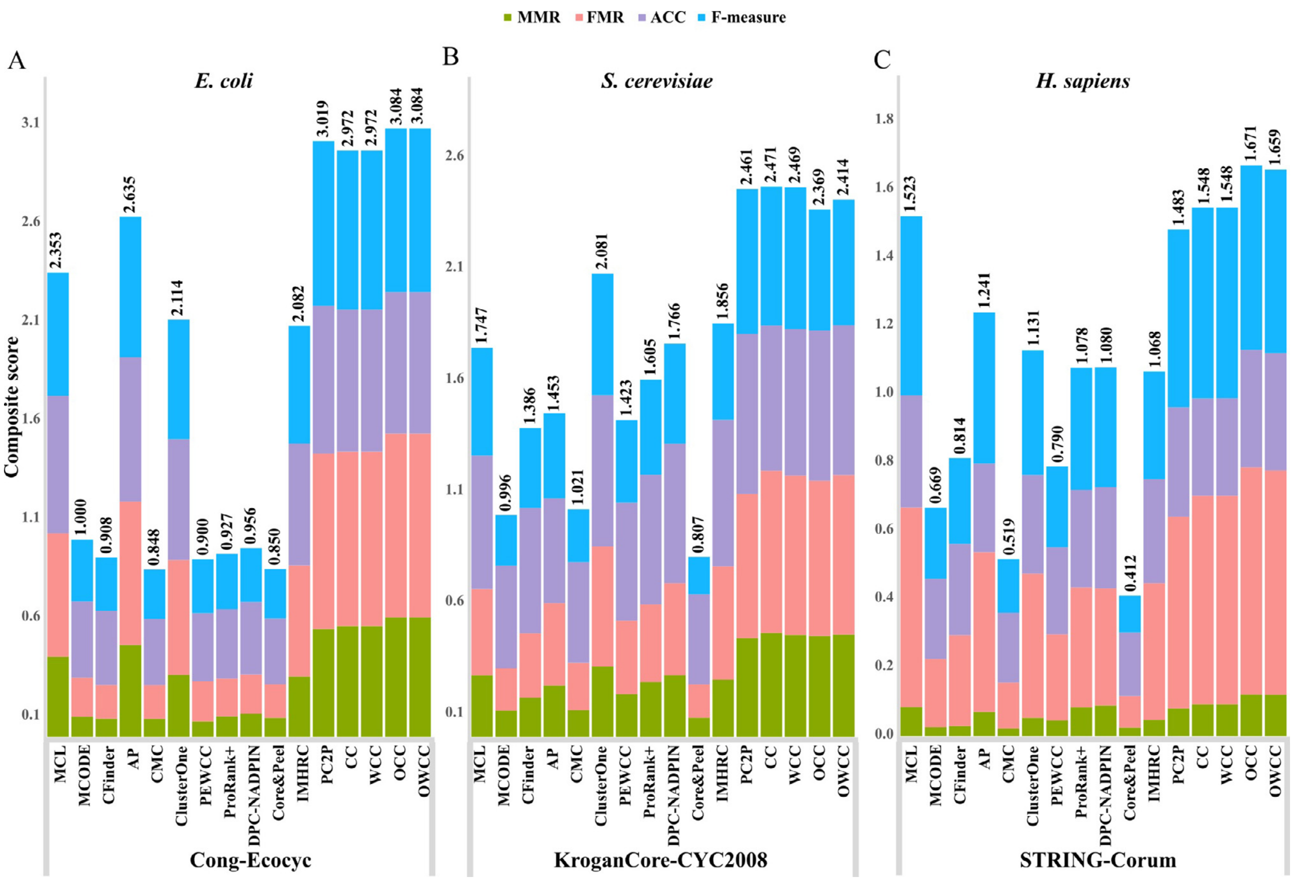

Fig. 3. Comparative analysis of approaches for prediction of protein complexes across PPI networks of different organisms. PPI networks for three organisms are considered (A) *E. coli*, (B) S. cerevisiae, and (C) *H. sapiens*. The comparative analyses are conducted with respect to a composite score that is a sum of MMR, FRM, ACC, and F-measure (see Methods). Sixteen approaches, ordered by the year of publication, are compared on three PPI networks-gold standard combination. The GCC-v outperforms all other approaches based on the composite score.

and latter mentioned cases; however, GCC-v is 6.44% on average more efficient than PC2P. The better performance on KroganCore in comparison to KroganExt can be explained by the smaller density of the former (see PPI networks and gold standards of protein complexes, Supplementary Table 3). The same analysis was carried on two combinations of H. sapiens PPI networks and one gold standard. GCC-v demonstrated the highest MMR, FRM, PPV, recall, f-measure, and f-measure$^+$ and resulted in the top four composite score among all other contenders (Fig. 3C, Supplementary Table 3 and Supplementary Fig. 3). Therefore, these findings demonstrated that predicting protein complexes by partitioning the graph into biclique spanned subgraphs based on simple scoring using the (weighted) clustering coefficients, outperformed other state-of-the-art approaches in 85.71% of the 14 considered scenarios specified by a combination of PPI network and gold standard across the different organisms.

We further evaluated the predicted protein complexes based on another composite score that is the sum of the MMR and F-measure$^+$ calculated over complexes with overlap scores above a given threshold for all combinations of PPI networks and gold standards for *E. coli,* S. cerevisiae, and H. sapiens [52] (see Evaluation metrics, Supplementary Files). Here, too, the GCC-v was among the top five approaches for all combinations of PPI networks and gold standards over all organisms (Supplementary Fig. 4).

Finally, to further analyze the proposed algorithm on PPI networks of even larger size, we used the BioGRID network of *H. sapiens* [61] and the STRING PPI network with two different thresholds value, 700 and 800 (Supplementary Table 1). We selected the top three performers with efficient implementations, including: MCL, ClusterOne, and WCC. The result showed that WCC outperformed the other approaches across all PPI networks. More precisely, for STRING PPI network with two different thresholds, WCC obtained the highest score for all performance measure, except sensitivity for which WCC is ranked second after ClusterOne (Supplementary Table 10).

2.4. Effects of network perturbation on identification of protein complexes

We further examined the reliability of the results acquired from WCC, as the best performing approach according to the composite score, by performing network perturbations and repeating the identification of protein complexes. To this end, we randomly removed 5, 10, 15, and 20 percent of the edges from the PPI networks of *E. coli, S. cerevisiae,* and *H. sapiens*. In doing so, we ensured that the number of connected components of the networks is the same for the cases before and after the edge removal. We repeated the procedure 100 times and calculated the average of each performance measure over all repetitions. The results for all combinations of PPI networks and gold standards showed only a slight decrease in each performance measure, and thereby, in the composite score, when we removed 5 to 20 percent of the edges from the original PPI network. The exception is the PIPS network in *H. sapiens* regarding the CORUM complexes (Supplementary Figure 5), for which edge removal resulted in a composite score of a slightly higher value. Moreover, we compared the composite score obtained by removing edges with the original composite score of WCC across all organisms (Fig. 4). The composite score was generally smaller, except in the cases of Babu (*E. coli*), Gavin, and KroghanExt (*S. cerevisiae*) with the metabolic, CYC2008, and SGD gold standards, respectively. This finding implied that the original PPI network, particularly in the case of KroganExt, contains more false-positive protein interactions than the other PPI networks, under the assumption that the prediction approaches and gold standard applied are of good quality (as shown by our findings).

We repeated the same procedure for the other two best-performing algorithms according to composite score, MCL and ClusterOne. The result suggested that ClusterOne performed similarly to WCC; however, MCL had a huge drop in its composite score. The fluctuation in the composite score is calculated as follows: we first calculated the average of the composite score of 100 times repetition for each network obtained by removing 5 to 20 percentages of edges; next, we computed the overall mean of these composite scores; finally, we compared the average composite score with the one resulted from the original network. On average, MCL has the highest fluctuation in the composite score followed by ClusterOne and then WCC with a fluctuation ratio of 848.57%, 18.88%, and 12.94%, respectively (Supplementary Figure 6–7).

2.5. Application of GCC-v to pan-plant PPI network identifies new, high-confidence protein complexes

Next, we aimed to determine new protein complexes in PPI networks from plant. To this end, we used the recently assembled pan-plant PPI network by using data from co-fractionation mass spectrometry from 13 plant species; the resulting protein interactions are scored based on the likelihood of physical interaction between two proteins [8]. Here, we considered the high-confidence interactions (with scores greater than 0.5), and used the same gold standard of plant protein complexes (see Supplementary Table 1).

We selected the top three best performing approaches, including: ClusterOne, MCL, and WCC, to analyze their performance on the plant PPI network. The result indicates that WCC outperformed the other two contenders and obtained the highest composite score (Supplementary Table 10). With further investigation of predicted protein complexes from the approaches, WCC captured fully 34.75% of known protein complexes in the plant gold standard (Supplementary Table 5), while ClusterOne and MCL captured

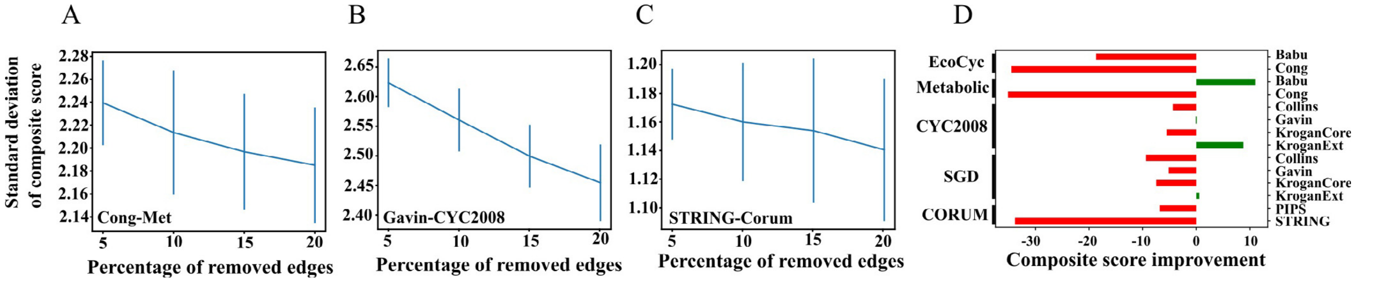

Fig. 4. Standard deviation of the composite score due to network perturbations by edge removal. To investigate the composite scores due to WCC, we removed 5, 10, 15, and 20 percent of the edges from the original PPI networks and then we applied WCC on all combinations of gold standards and PPI networks. The standard deviations of the composite score are calculated over 100 repetitions for (A) Cong-Metabolic complexes, (B) Gavin-CYC2008, and (C) PIPS-CORUM in *E. coli*, *S. cerevisiae*, and *H. sapiens*, respectively. The average standard deviation of the composite score decreases slightly with the increased percentage of removed edges across all PPI networks. (D) The obtained average composite score from removing edges was in turn compared with the original composite score of WCC across all organisms. The composite score for Babu, Gavin, KroganExt PPI networks in combination with Metabolic, CYC2008, and SGD complexes are improved, indicating possible false positive PPIs included in the network.

20.33% and 12.71% of known protein complexes from this gold standard, respectively.

We next determined the PPI network for *A. thaliana* by mapping the each eggNOG ID, used in the pan-plant network, to a TAIR locus ID [8]. By applying WCC and comparing the results clusters against the plant gold standard, we predicted 144 new protein complexes. Interestingly, 39.58% of these new complexes show GO semantic similarity of MF category equal to 1, suggesting coordinated functions of the involved proteins (see Supplementary Table 6). However, it has already been shown that functional modules may perform different molecular functions, but could be involved in the same process and can take place in the same organelle [62,63]. Therefore, we also investigated the new clusters that obtained maximum GO semantic similarity (of 1) for the BP and CC categories, but showed lower GO semantic similarity based on the MF category. Consequently, we considered 8.33% of the new predicted complexes as functional modules (see Supplementary Table 7).

Finally, we evaluated the selected clusters based on the GO semantic similarity by analyzing the domain-domain interactions (DDI) of the proteins comprising these clusters. For $\sim 43\%$ of the protein complexes we identified the domains, based on the Pfam database [64] and relied on the DDI network [65] for further characterization. By using the DDI of high confidence (gold and silver), we found support for eight protein complexes, of which six were also inferred by considering only the Gold DDI category (see Supplementary Table 9).

By inspecting the protein clusters for which there is a support from domain-domain interactions, we found that COL3 (AT2G24790) forms a complex with a peptidase M1 family protein (AT1G63770) and APM1 (AT4G33090), encoding an aminopeptidase. COL3 is a positive regulator of photomorphogenesis that acts downstream of COP1, but can promote lateral root development independently of COP1; it also functions as a daylength-sensitive regulator of shoot branching [66]. Loss-of-function mutants for APM1 show irregular, uncoordinated cell divisions throughout embryogenesis that affect the shape and number of cotyledons [67]. It is therefore expected that this complex may be implicated in control of root traits. In addition, our findings point at the protein complex involving AL3, AL4, and AL6; all AL proteins except AL3 bind to di- or trimethylated histone H3 (H3K4me3/2) [68], and may be involved in controlling root hair elongation under particular nutrient availability [69]. The other protein complexes are composed of two proteins each, related to histones or part of histone superfamily protein, ferredoxin 2 and 3, serine hydroxylmethyltransferases 2 and 3, SERINE HYDROXYMETHYLTRANSFERASE 3, two RNA helicases, and several proteins of unknown function. The largest of the 144 predicted, high-confidence complexes involved the functional interaction between light harvesting proteins of photosystems I and II as well as associated components (particularly subunits of photosystem I), supported by other chloroplast PPI network of *A. thaliana* [70] (see Supplementary Table 9). Therefore, these findings provide a rich resource for generating hypotheses for validation in dedicated experimental studies.

3. Discussion

Assembly of high-quality gold standards has demonstrated that protein complexes exhibit different densities, can overlap in one or more proteins, and show differences in sizes. However, the existing approaches have largely focused on identification of particular class of protein complexes (e.g. dense vs. sparse), affecting the resulting recall values. In addition, the existing approaches rely on multiple parameters, rendering it challenging to compare the predicted complexes in networks of increasing quality and size. Finally, many of the approaches do not scale well for large-scale networks or are difficult to parallelize.

We have shown that partitioning a PPI network into biclique spanned subgraphs provides the best performing approach to identify protein complexes [38]. However, while this approach is parameter-free and allows to detect both dense and sparse protein complexes, it does not scale well. Therefore, we introduced a new greedy approximation algorithm to efficiently partition a PPI network into biclique spanned subgraphs corresponding to protein complexes. The GCC-v is elegant and fast; it can predict protein complexes in large-scale PPI networks in order of seconds. In addition, GCC-v allows the consideration of edge-weights to prioritize more reliable interactions in the corresponding predicted complexes. Moreover, GCC-v allows the identification of overlapping clusters by employing line graph transformation. Finally, and most importantly, GCC-v is parameter-free and, therefore, can be employed objectively, without user-specified parameters that must be tuned in a case-to-case basis.

Extensive comparative analyses demonstrated that GCC-v outperformed state-of-the-art contending approaches with respect to seminal performance measures in PPI networks from *E. coli*, S. cerevisiae, and H. sapiens, while ensuring that the overall semantic similarity of the predicted proteins is high, in line with biological expectation. Most importantly, GCC-v resulted in the largest recall (and refinements of this measure), MMR, FRM, accuracy as well as F-measure (and its refinement) in majority of examined data sets. Therefore, we demonstrated that GCC-v offers a parameter-free and efficient means to overcome the key shortcoming of the existing approaches. Applications with pan-plant PPI network showed that GCC-v outperforms all other approaches, and resulted in new, high-confidence putative protein complexes of small size involved in variety of biological processes. Therefore, our computational approaches can be readily used to posit testable hypotheses about the role of protein complexes in shaping diverse cellular traits.

4. Conclusion and future work

The proposed family of greedy approaches, called GCC-v, provides a scalable means for accurate prediction of protein complexes as determined by our extensive comparative analyses. Interestingly, our analysis of the effects of network perturbations demonstrate approaches from GCC-v can also point out the quality of the used PPI network, provided a gold standard of protein complexes of high quality, as used in our study. These results point at the intimate relationship between the problem of network clustering based on biclique spanned partitions and that of link (or edge) prediction—a relationship that we aim to explore in future studies.

5. Methods

5.1. PPI networks and gold standards of protein complexes

To evaluate the performance of different network clustering approaches, we used two *E. coli*, four S. cerevisiae, and two H. sapiens PPI networks. All PPI networks are edge-weighted except one from *E. coli*. The S. cerevisiae PPI networks, including Gavin [4], Collins [53], Krogan Core, and Krogan Extended [54], were obtained experimentally, and the weights (in the range between zero and one) denote the reliability of each interaction. In the Collins PPI network, the interaction weights are based on purification enrichment score, while in the Gavin PPI network, the weight indicates the socio-affinity index that measures the affinity between proteins. The socio-affinity index calculates the log-odds of how many times pairs of proteins are observed together as preys, or a bait and a prey in the network. In the Krogan PPI network, each interaction is assigned a probability based on the integration of mass spectrometry scores. This network has two versions, the Krogan core contains highly reliable interactions (probability ≥ 0.273), the Krogan extended network includes more interactions of smaller reliability (probability ≥ 0.101). In addition, we used the up-to-date H. sapiens PPI networks obtained from STRING [55] and PIPS [6]. In the STRING network, the score on each interaction does not indicate the strength but the confidence of an interaction, i.e., given all available evidence, denoting how likely it is that the interaction is real. Two different types of the score are available in the STRING dataset, i.e. combined score and sub-score. In this study, we considered the combined score that is supported by several types of evidence, namely: Conserved neighborhood, Gene fusions, co-expression, phylogenetic co-occurrence, database imports, large-scale experiments, and Literature co-occurrence. The interaction score in the PIPS network corresponds to the posterior odds ratio of interaction computed based on a naïve Bayes network [71]. Intuitively, the score indicates the likelihood of the interaction between pairs of proteins given the evidence. In the prediction of interactions, several features are considered, such as: expression data, protein domains, subcellular localization, and co-occurrence of domains. Therefore, to consider more reliable protein interactions into our study, we set the cut-off score of 999 and 25 for STRING and PIPS PPI networks, respectively. Moreover, two *E. coli* PPI networks that are obtained from [5,7] are used in this study. For simplicity, we named the PPI networks the same as the corresponding first author Babu and Cong throughout the paper. The protein interactions in Babu network inferred experimentally from affinity purification mass spectrometry (APMS). Later, they applied an integrative statistical framework on inferred interactions to obtain a confidence score for each PPI. The protein interactions in Cong network are predicted by utilizing evolutionary signatures in protein sequence and structure.

Different sets of a gold standard are available to assess the predicted protein complexes. Here, the CYC2008 [30], an update to the Munich Information Centre for Protein Sequences (MIPS) catalog [28], and complexes derived from the Saccharomyces Genome Database (SGD) [29] are used as S. cerevisiae protein complex reference sets. Furthermore, we employed CORUM [31] as the gold standard for H. sapiens protein complexes. Finally, we used two gold standards for *E. coli*: metabolic gold standard based on the genome-scale metabolic network of *E. coli* [72] and EcoCyc [27]. The protein complexes in both CYC2008 and CORUM reference sets are verified by small-scale experiments and the protein complexes from EcoCyc are manually curated. The mentioned PPI networks and gold standards differ with respect to the number of proteins and interactions they include. For completeness, Supplementary Table 1 includes these features for the PPI networks, gold standards, and their intersections employed in the analyses.

5.2. Preliminaries

Let $G = (V, E)$ be an undirected graph with set of vertices V and set of edges E. For a node $v \in V$, we define the neighborhood of v by $N(v) = \{u \in V | (u, v) \in E\}$. Regardless of which input graph (original or line graph) and node scoring method (original or weighted clustering coefficient Eqs. (1) – (2)) used, the same Greedy Clustering Coefficient algorithm (GCC) holds. GCC scores all nodes, then identifies the node with highest score and its neighbors in a cluster (ties are broken arbitrarily). Next, it updates the scores and repeats this procedure until there is no node left in the graph (Algorithm 1).

Algorithm 1. Greedy Clustering Coefficient algorithm

```
procedure GCC ($G$)
  clusters ← []
  connected_components ← connected component of$G$
  while there is a component in connected_components do
    nodes_score ← score(component,V[component])
    $v$ ← argmax(nodes_score)
    cluster← $[v + N(v)]$
    append cluster to clusters
    remove cluster from graph$G$
    connected_components ← connected component of$G$
    Update score(N(cluster))
  return (clusters)
```

The GCC-v and all the other contenders were carried on the same machine, an Intel(R) Core™ i7-8650U with 1.90 GHz-2.11 GHz, except PC2P that was performed on an Intel(R) Xeon(R) CPU E5-2670 v2 with 2.50 GHz. We investigated the empirical running time on the four PPI networks of S. cerevisiae, two PPI networks of H. sapiens, and two PPI networks of *E. coli*, for WCC, PC2P, and MCL approaches (Supplementary Table 4). We observed that the WCC has the lowest running time in comparison with the other two approaches. The WCC running time is always in the order of magnitude of seconds for all datasets however the MCL and PC2P are in the order of magnitude of minutes and in some cases for PC2P is hours or a day. GCC-v is available on GitHub at https://github.com/SaraOmranian/GCC-v.

5.3. Evaluation metrics

Here, we adopt twelve well-established metrics to assess the quality of predicted protein complexes, including sensitivity, positive predictive value, accuracy and separation from [73], fraction match and maximum matching ratio from [18], precision, recall, and F-measure from [22], and precision$^+$, recall$^+$, and F-measure$^+$ from [60].

These evaluation metrics are selected based on their usage in seminal studies about the prediction of protein complexes [18,22,21,19,60]. We further use the composite score which is the sum of the values of MMR, FMR, ACC, and F-measure [58,18,59], as well as the sum of MMR and F-measure$^+$ over different threshold values ($0 \leq \theta \leq 1$) [60] as other metrics to illustrate the overall performance. The definition and notations of evaluation metrics are all well explained in the Supplementary Files.

The functional similarity between two proteins can be assessed by semantic similarity of their respective GO annotation terms [74]. Thereby, we employed the GOSim R package [75] to determine the similarity between protein pairs in a given predicted complex. The final semantic similarity of each predicted complex is summarized by the median of the semantic similarity of the protein pairs in the corresponding complex.

6. Data availability

The data underlying this article are publicly available and their corresponding references are provided within the article.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Z.N. thanks the European Union's Horizon 2020 research and innovation programme, project PlantaSYST (SGA-CSA No. 739582 under FPA No. 664620).

Author contributions

Conceived and designed the study: S.O. and Z.N. Developed the model: S.O., A.A., and Z.N. Implemented: S.O. Wrote the manuscript: S. O. and Z.N.. Made comments and approved the final version submitted: S.O., A.A. and Z.N.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2021.09.014.

References

[1] Martinez E. Multi-protein complexes in eukaryotic gene transcription. Plant Mol Biol 2002;50:925–47.
[2] Sweetlove LJ, Fernie AR. The role of dynamic enzyme assemblies and substrate channelling in metabolic regulation. Nat Commun 2018;9.
[3] Pawson T, Nash P. Protein-protein interaction define specificity in signal transduction. Genes Dev 2000;14:1027–47.
[4] Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, et al. Proteome survey reveals modularity of the yeast cell machinery. Nature 2006;440:631–6.
[5] Babu M, Bundalovic-Torma C, Calmettes C, Phanse S, Zhang Q, Jiang Y, et al. Global landscape of cell envelope protein complexes in Escherichia coli. Nat Biotechnol 2017;36:103–12.
[6] McDowall MD, Scott MS, Barton GJ. PIPs: human protein-protein interaction prediction database. Nucleic Acids Res 2009;37:D651–6.
[7] Cong Q, Anishchenko I, Ovchinnikov S, Baker D. Protein interaction networks revealed by proteome coevolution. Science 2019;365:185–9.
[8] McWhite CD, Papoulas O, Drew K, Cox RM, June V, Dong OX, et al. A Pan-plant Protein Complex Map Reveals Deep Conservation and Novel Assemblies. Cell 2020;181:460–474.e14.
[9] Cui L, Acharya S, Mishra S, Pan Y, Huang JZ. MMCo-Clus — An Evolutionary Co-clustering Algorithm for Gene Selection. IEEE Trans Knowl Data Eng 2020. p. 1–1.
[10] Acharya S, Cui L, Pan Y. Multi-view feature selection for identifying gene markers: a diversified biological data driven approach. BMC Bioinf 2020;21.
[11] Acharya S, Cui L, Pan Y. A refined 3-in-1 fused protein similarity measure: application in threshold-free hub detection. IEEE/ACM Trans Comput Biol Bioinf 2020. p. 1–1.
[12] Fields S, Sternglanz R. The two-hybrid system: an assay for protein-protein interactions. Trends Genet 1994;10:286–92.
[13] Lin J-S, Lai E-M. Protein–Protein Interactions: Co-Immunoprecipitation. In: *Methods in Molecular Biology*. New York: Springer; 2017. p. 211–9.
[14] Bauer A, Kuster B. Affinity purification-mass spectrometry. Eur J Biochem 2003;270:570–8.
[15] Fujikawa Y, Kato N. TECHNICAL ADVANCE: Split luciferase complementation assay to study protein-protein interactions in Arabidopsis protoplasts. Plant J 2007;52:185–95.
[16] McBride Z, Chen D, Lee Y, Aryal UK, Xie J, Szymanski DB. A label-free mass spectrometry method to predict endogenous protein complex composition. Mol Cell Proteomics 2019;18:1588–606.
[17] Silverbush D, Sharan R. A systematic approach to orient the human protein–protein interaction network. Nat Commun 2019;10.
[18] Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods 2012;9:471–2.
[19] Pellegrini M, Baglioni M, Geraci F. Protein complex prediction for large protein protein interaction networks with the Core&Peel method. BMC Bioinf 2016;17.
[20] Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinf 2003;4:2.
[21] Adamcsek B, Palla G, Farkas IJ, Derényi I, Vicsek T. CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 2006;22:1021–3.
[22] Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics 2009;25:1891–7.
[23] Frey BJ, Dueck D. Clustering by Passing Messages Between Data Points. Science 2007;315:972–6.
[24] Enright AJ. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002;30:1575–84.
[25] Ou-Yang L, Wu M, Zhang X-F, Dai D-Q, Li X-L, Yan H. A two-layer integration framework for protein complex detection. BMC Bioinf 2016;17.
[26] Wu M, Ou-Yang L, Li X-L. Protein Complex Detection via Effective Integration of Base Clustering Solutions and Co-Complex Affinity Scores. IEEE/ACM Trans Comput Biol Bioinf 2017;14:733–9.
[27] Keseler IM, Mackie A, Santos-Zavaleta A, Billington R, Bonavides-Martɨnez C, Caspi R, et al. The EcoCyc database: reflecting new knowledge aboutEscherichia coliK-12. Nucleic Acids Res 2016;45:D543–50.
[28] Mewes HW. MIPS: analysis and annotation of proteins from whole genomes. Nucleic Acids Res 2004;32:41D–4D.
[29] Hong EL, Balakrishnan R, Dong Q, Christie KR, Park J, Binkley G, et al. Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res 2007;36:D577–81.
[30] Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 2008;37:825–31.
[31] Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, et al. CORUM: the comprehensive resource of mammalian protein complexes—2019. Nucleic Acids Res 2018;47:D559–63.
[32] Wu Z, Liao Q, Liu B. A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. Briefings Bioinf 2019;21:1531–48.
[33] S. R. I. G. A. N. E. S. H. SRIHARI and H. W. A. I. LEONG, "A survey of computational methods for protein complex prediction from protein interaction networks," Journal of Bioinformatics and Computational Biology, vol. 11, p. 1230002, 2013.
[34] Srihari S, Leong HW. Employing functional interactions for characterisation and detection of sparse complexes from yeast PPI networks. Int J Bioinf Res Appl 2012;8:286.
[35] Yong CH, Liu G, Chua HN, Wong L. Supervised maximum-likelihood weighting of composite protein networks for complex prediction. BMC Syst Biol 2012;6:S13.
[36] Ruan P, Hayashida M, Akutsu T, Vert J-P. Improving prediction of heterodimeric protein complexes using combination with pairwise kernel. BMC Bioinf 2018;19.
[37] Yong C, Maruyama O, Wong L. Discovery of small protein complexes from PPI networks with size-specific supervised weighting. BMC Syst Biol 2014;8:S3.
[38] Omranian S, Angeleska A, Nikoloski Z. PC2P: parameter-free network-based prediction of protein complexes. Bioinformatics 2021;0–0.
[39] Bhowmick SS, Seah BS. Clustering and Summarizing Protein-Protein Interaction Networks: A Survey. IEEE Trans Knowl Data Eng 2016;28:638–58.
[40] Angeleska A, Nikoloski Z. Coherent network partitions. Discrete Appl Math 2019;266:283–90.
[41] Holland PW, Leinhardt S. Transitivity in Structural Models of Small Groups. Comparative Group Studies 1971;2:107–24.
[42] Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature 1998;393:440–2.
[43] Barrat A, Barthelemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. Proc Natl Acad Sci 2004;101:3747–52.
[44] Onnela J-P, Saramäki J, Kertész J, Kaski K. Intensity and coherence of motifs in weighted complex networks. Phys Rev E 2005;71.
[45] Holme P, Park SM, Kim BJ, Edling CR. Korean university life in a network perspective: Dynamics of a large affiliation network. Physica A 2007;373:821–30.
[46] Evans TS, Lambiotte R. Line graphs, link partitions, and overlapping communities. Phys Rev E 2009;80.
[47] Harary F, Norman RZ. Some properties of line digraphs. Rendiconti del Circolo Matematico di Palermo 1960;9:161–8.
[48] Ahn Y-Y, Bagrow JP, Lehmann S. Link communities reveal multiscale complexity in networks. Nature 2010;466:761–4.
[49] Zaki N, Efimov D, Berengueres J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinf 2013;14.
[50] Hanna EM, Zaki N. Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. BMC Bioinf 2014;15.
[51] Shen X, Yi L, Jiang X, Zhao Y, Hu X, He T, et al. Neighbor affinity based algorithm for discovering temporal protein complex from dynamic PPI network. Methods 2016;110:90–6.
[52] Maddi AMA, Eslahchi C. Discovering overlapped protein complexes from weighted PPI networks by removing inter-module hubs. Sci Rep 2017;7.
[53] Collins SR, Kemmeren P, Zhao X-C, Greenblatt JF, Spencer F, Holstege FCP, et al. Toward a Comprehensive Atlas of the Physical Interactome ofSaccharomyces cerevisiae. Mol Cell Proteomics 2007;6:439–50.
[54] Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 2006;440:637–43.
[55] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2014;43:D447–52.
[56] Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics 2007;23:1274–81.
[57] Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinf 2006;7.
[58] Cao B, Deng S, Qin H, Ding P, Chen S, Li G. Detection of Protein Complexes Based on Penalized Matrix Decomposition in a Sparse Protein-Protein Interaction Network. Molecules 2018;23:1460.

[59] Wang R, Liu G, Wang C, Su L, Sun L. Predicting overlapping protein complexes based on core-attachment and a local modularity structure. BMC Bioinf 2018;19.

[60] Maddi AMA, Moughari FA, Balouchi MM, Eslahchi C. CDAP: An Online Package for Evaluation of Complex Detection Methods. Sci Rep 2019;9.

[61] Stark C. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9.

[62] Lu H, Shi B, Wu G, Zhang Y, Zhu X, Zhang Z, et al. Integrated analysis of multiple data sources reveals modular structure of biological networks. Biochem Biophys Res Commun 2006;345:302–9.

[63] Chen B, Fan W, Liu J, Wu F-X. Identifying protein complexes and functional modules–from static PPI networks to dynamic PPI networks. Briefings Bioinf 2013;15:177–94.

[64] Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res 2020;49: D412–9.

[65] Ziaeddine AS, Amina A-N, Hiba N, Ritchie DW, Marie-Dominique D. ‘‘PPIDomainMiner : Inferring domain-domain interactions from multiple sources of protein-protein interactions,” 3 2021.

[66] Datta S, Hettiarachchi GHCM, Deng X-W, Holm M. Arabidopsis CONSTANS-LIKE3 Is a Positive Regulator of Red Light Signaling and Root Growth. Plant Cell 2005;18:70–84.

[67] Peer WA, Hosein FN, Bandyopadhyay A, Makam SN, Otegui MS, Lee G-J, et al. Mutation of the Membrane-Associated M1 Protease APM1 Results in Distinct Embryonic and Seedling Developmental Defects in Arabidopsis. Plant Cell 2009;21:1693–721.

[68] Liang X, Lei M, Li F, Yang X, Zhou M, Li B, et al. Family-Wide Characterization of Histone Binding Abilities of PHD Domains of AL Proteins in Arabidopsis thaliana. Protein J 2018;37:531–8.

[69] Chandrika NNP, Sundaravelpandian K, Yu S-M, Schmidt W. ALFIN-LIKE 6 is involved in root hair elongation during phosphate deficiency in Arabidopsis. New Phytol 2013;198:709–20.

[70] Yu Q-B, Li G, Wang G, Sun J-C, Wang P-C, Wang C, et al. Construction of a chloroplast protein interaction network and functional mining of photosynthetic proteins in Arabidopsis thaliana. Cell Res 2008;18:1007–19.

[71] Scott MS, Barton GJ. Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinf 2007;8.

[72] King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, et al. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res 2015;44:D515–22.

[73] Brohée S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinf 2006;7.

[74] Cho Y-R, Hwang W, Ramanathan M, Zhang A. Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinf 2007;8.

[75] 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 Bioinf 2007;8.