Omranian and Nikoloski
*Applied Network Science* (2022) 7:71
https://doi.org/10.1007/s41109-022-00508-5

Applied Network Science

RESEARCH Open Access

# CUBCO+: prediction of protein complexes based on min-cut network partitioning into biclique spanned subgraphs

Sara Omranian[1,2,3,4,5] and Zoran Nikoloski[4,5*]

*Correspondence:
Nikoloski@mpimp-golm.mpg.de

[1] Technical University of Munich, Campus Straubing for Biotechnology and Sustainability, Bioinformatics, Petersgasse 18, 94315 Straubing, Germany
[2] Weihenstephan-Triesdorf University of Applied Sciences, Petersgasse 18, 94315 Straubing, Germany
[3] SynBiofoundry@TUM, Technical University of Munich, Schulgasse 22, 94315 Straubing, Germany
[4] Bioinformatics, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany
[5] Systems Biology and Mathematical Modeling, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

## Abstract

High-throughput proteomics approaches have resulted in large-scale protein–protein interaction (PPI) networks that have been employed for the prediction of protein complexes. However, PPI networks contain false-positive as well as false-negative PPIs that affect the protein complex prediction algorithms. To address this issue, here we propose an algorithm called CUBCO+ that: (1) employs GO semantic similarity to retain only biologically relevant interactions with a high similarity score, (2) based on link prediction approaches, scores the false-negative edges, and (3) incorporates the resulting scores to predict protein complexes. Through comprehensive analyses with PPIs from *Escherichia coli*, *Saccharomyces cerevisiae*, and *Homo sapiens*, we show that CUBCO+ performs as well as the approaches that predict protein complexes based on recently introduced graph partitions into biclique spanned subgraphs and outperforms the other state-of-the-art approaches. Moreover, we illustrate that in combination with GO semantic similarity, CUBCO+ enables us to predict more accurate protein complexes in 36% of the cases in comparison to CUBCO as its predecessor.

**Keywords:** Protein complexes, Protein–protein interaction, Network clustering, Species comparison

## Introduction

Proteins are essential components of all living organisms and participate in almost every biological process. However, most proteins do not function as a single entity; instead, they often interact with other proteins to form large macromolecules, i.e. protein complexes, that are involved in different cellular functions. Identifying protein complexes allows assigning functions to proteins of yet unknown roles by using the known function of their interacting partners, following the principle of guilt-by-association (Tian et al. 2008). Moreover, due to the protein structures, proteins are often involved in more than one complex in different subcellular compartments and biological processes. Therefore, studying protein complexes is important to understand the functional principles of the cell system, from signaling to metabolism (Pawson and Nash 2000; Maslov and Sneppen 2002; Reyes-Turcu et al. 2009; Sweetlove and Fernie 2018), and provide a better understanding hierarchy of intra- and inter-cellular activities (Bauer and Kuster 2003).

One way to obtain data on protein–protein interactions (PPIs) relies on high-throughput profiling techniques (Ho et al. 2002; Gavin et al. 2002). The advances in experimental techniques (Fields and Sternglanz 1994; Bauer and Kuster 2003; Fujikawa and Kato 2007; Lin and Lai 2017; McBride et al. 2019) provide us with a plethora of resulting PPI networks from several model organisms (Gavin et al. 2006; Szklarczyk et al. 2014; Babu et al. 2017; McWhite et al. 2020). In a protein–protein interaction (PPI) network, nodes correspond to proteins and edges represent physical interaction between two proteins. However, due to the noisiness of experimental techniques, the resulting PPI networks contain spurious interactions, which result in false-positive and false-negative interactions (Berger et al. 2013).

Computational approaches based on graph clustering algorithms are often used to complement the experimental approaches in the identification of protein complexes. Several studies (Li et al. 2010; Srihari and Leong 2013; Bhowmick and Seah 2016; Wu et al. 2019; Omranian et al. 2022) have categorized the existing computational approaches for protein complex prediction in multiple groups, such as (i) supervised (Qi et al. 2008; Shi et al. 2011) versus unsupervised (Spirin and Mirny 2003; Bader and Hogue 2003), (ii) using only the topological structure of PPI network (Enright 2002; Nepusz et al. 2012) versus integrating additional knowledge or data, such as gene expression (Feng et al. 2011), functional and evolutionary information (King et al. 2004; Sharan et al. 2005; Dost et al. 2008). Further, several protein complex gold standards of different species such as EcoCyc for *Escherichia coli* (Keseler et al. 2016), MIPS, SGD, and CYC2008 for *Saccharomyces cerevisiae* (Mewes 2004; Hong et al. 2007; Pu et al. 2008), and CORUM for *Homo sapiens* (Giurgiu et al. 2018), have been assembled to facilitate the comparison and evaluation of predicted complexes from different approaches.

Due to the incompleteness and noisiness of interactions data, a variety of computational approaches have been proposed as an alternative to experimental tools to predict protein interactions (Zeng 2016; Kovács et al. 2019; Wang et al. 2020). For instance, link prediction algorithms enable us to overcome some of the disadvantages of experimental approaches by identifying false-negative interactions in the PPI network. Therefore, the link prediction and graph clustering algorithms are jointly used to improve the performance of approaches for the prediction of protein complexes. One can employ a link prediction algorithm as a pre-processing step to tune the PPI network and then predict more accurate protein complexes. Alternatively, one can first employ a graph clustering algorithm to group the proteins that are more likely to interact together, and then apply a variety of local or global structure-based similarity measures to compute the possibility of protein interactions in the same cluster (Hu et al. 2021).

Although the performance of existing computational approaches has gradually increased over time, they still have some notable disadvantages. Overall, the existing computational approaches to solve this problem are based on the idea that protein complexes correspond to highly connected or near-cliques clusters in the PPI network. Therefore, it is most likely that they predict only large and dense protein complexes, while they are incapable of finding sparse and small ones (Srihari and Leong 2013; Wu et al. 2019). If an approach solely depends on the PPI network, as mentioned earlier, its performance is expected to be affected by errors and missing interactions in PPI networks. Although the additional biological information might help in identifying protein complexes, this requires wet-lab

experiments that are time-consuming, labor-intensive, and the functional annotations of a protein might be outdated or unverified (Li et al. 2010).

Moreover, the computational approaches for protein complex prediction are parameter-dependent, which renders it difficult to interpret the resulting protein complexes (Omranian et al. 2021a; b). Finally, the performance of existing computational approaches is mostly evaluated with the PPI network of *S. cerevisiae*, and few of the existing methods conducted experiments to assess their results across other species, such as bacteria, plants, and humans (Sharma et al. 2018; Omranian et al. 2021a, b; Omranian and Nikoloski 2022).

In contrast to the existing computational approaches, PC2P, GCC-v, and CUBCO (Omranian et al. 2021a; b; Omranian and Nikoloski 2022) represent parameter-free algorithms and compare the performance of their results with several state-of-the-art approaches across different species. These approaches detect a protein complex based on partitioning the network into biclique spanned subgraphs, which is also known as coherent network partition (CNP) (Angeleska and Nikoloski 2019; Angeleska et al. 2021). PC2P and GCC-v rely on local properties of the network by finding the minimum cut in complement of the second neighborhood of a node (Omranian et al. 2021a; b) and computing the clustering coefficient for each node to partition the network into biclique spanned subgraphs (Omranian et al. 2021a; b), respectively. Alternatively, CUBCO (Omranian and Nikoloski 2022) is based on the global properties of the network and utilizes global minimum cut to partition the network into biclique spanned subgraphs. Moreover, to overcome the incompleteness of PPI networks, CUBCO integrates link prediction (Kovács et al. 2019) as a pre-processing step to cluster more probable interacting proteins together. The three approaches show consistent performance across different species, in contrast to other approaches that obtain different ranking scores for different combinations of species and the corresponding gold standards.

Here we introduce a new approach, referred to as CUBCO+, that predicts protein complexes based on the same concept as the PC2P, GCC-v, and CUBCO algorithms, i.e. biclique spanned subgraphs. However, CUBCO+ not only considers the effect of false-negative interactions, like CUBCO but also evaluates the false-positive interactions by weighing the interactions with Gene Ontology (GO) semantic similarity. To the best of our knowledge, CUBCO+ is the first algorithm to take the effect of both false-positive and false-negative interactions into account, while providing predictions of protein complexes with improved performance over the contenders. The rest of the paper is organized as follows: "Results" section presents the proposed complex prediction algorithm followed by comparing the performance of CUBCO+ with 17 other state-of-the-art methods based on 12 performance measures; "Method" section contains the related works, the introduction of PPI networks, gold standards, and well-established performance measures that are used in this study; finally, the conclusion of this study with a future scope is presented in "Discussion" section.

## Results

### CUBCO+ algorithm predicts protein complexes by considering both false-negative and false-positive interactions

In a simple graph $G = (V, E, w)$, a set of nodes $V$ corresponds to proteins, a set of edges $E$ denotes PPIs, and $w(e)$ corresponds to the weight of edge $e$ that indicates the reliability

of the interaction based on experimental and computational approaches. The graph denoted by $\overline{G} = (V, \{(u,v)|(u,v) \notin E\})$ is a complement of the simple graph $G$.

We model a protein complex as a biclique spanned graph, that is a graph $G = (V, E)$, with a node-set that can be partitioned into two subsets, $V_1(G)$ and $V_2(G)$, in which $V_1(G) \cap V_2(G) = \varnothing$ and $V_1(G) \cup V_2(G) = V(G)$. Its edge set contains all possible edges between the two node sets, $V_1(G)$ and $V_2(G)$, as well as additional edges between the nodes in each of the two partitions. A biclique spanned graph has two properties: 1. Its complement, $\overline{G}$, is disconnected (i.e. it contains more than one connected component) (Akiyama and Harary 1981); 2. the distance between any two nodes in a biclique spanned graph is at most two. Hence, these properties provide a natural formation of a network cluster based on connectedness, since the complement of a cluster, intuitively speaking, is disconnected. Moreover, since stars, bicliques, and cliques are special graph classes of biclique spanned subgraph the identified protein complexes will be sparse as well as dense regardless of their size. Thereby, the problem of protein complex prediction is cast to partition graph $G$, $C = \{C_1, C_2, \ldots, C_k\}$, such that each $C_i$ is a biclique spanned subgraph (Angeleska and Nikoloski 2019) by removing a minimum number of edges. It has been shown that the problem of finding an optimal coherent network partition is NP-hard (Angeleska and Nikoloski 2019), while a graph transformation has also been proposed to obtain a $O(\log n)$-approximation algorithm on a bipartite graph (Angeleska et al. 2021). Thus, greedy approximation algorithms, including PC2P, GCC-v, and CUBCO, have been proposed for solving this optimization problem. CUBCO+ is an updated version of CUBCO exploring the whole network at once by using global network properties to define the partition into biclique spanned subgraphs.

Since the complement of the biclique spanned subgraph is disconnected, CUBCO and CUBCO+ employ a global minimum cut algorithm to render the complement of the original graph disconnected. Given an undirected graph with non-negative edge weights, the minimum cut problem (i.e. min-cut) is to partition the node-set into two subsets so that the sum of edge weights between the two subsets is minimized (Stoer and Wagner 1994).

In the following, we present CUBCO+ that iteratively finds the biclique spanned subgraph in a given simple graph $G$ based on identifying the minimum cut in the weighted $\overline{G}$. The procedure of CUBCO+ is similar to CUBCO and predicts protein complexes in four main steps: (i) construct initial PPI network, (ii) determine the complement of a graph $G$, i.e., $\overline{G}$, (iii) assign weights to the edges in $\overline{G}$ based on the degree-normalized number of the path of length three between the endpoint nodes of an edge in original graph $G$; (iv) iteratively find the minimum cut of the edge-weighted graph $\overline{G}$ (Algorithms 1 and 2) until all resulting components are biclique spanned.

In contrast to CUBCO, the updated version, CUBCO+, considers the effect of false-positive edges by computing GO semantic similarity for every interaction in the network as an edge-weight. The steps that are the same between CUBCO and CUBCO+ are marked with * in Algorithms 1 and 2. GO is hierarchical controlled biological vocabularies that estimate the functional similarity of gene products, relating to three categories: (i) Molecular Function (MF), (ii) Biological Process (BP), and (iii) Cellular Component (CC). Hence, GO semantic similarity (Cho et al. 2007) determines the functional similarity of two given proteins. For this purpose, we applied

GOSim R package 4.1 (Fröhlich et al. 2007). To calculate the contribution of all GO semantic similarity domains in the edge weight, first, the geometric average of three categories of GO is calculated. Next, these values are normalized, concerning the maximum value. Last, a set of interactions with weights greater and equal to 0.3 is selected for constructing the initial PPI network. By increasing the threshold value on the edge weight, more edges will be removed from the network, which increases the number of connected components of the network. We opted for the value of 0.3 to retain more topological information while ensuring that biologically relevant interactions are included in the network.

The complement of the graph, $\overline{G}$, contains edges that are not present in the original graph $G$. From a biological perspective, $E(\overline{G})$ corresponds to PPIs that are not included in the original graph $G$ (i.e., false-negative/true-negative interactions). Several studies have predicted the missing edges in PPI networks based on different concepts (Zeng 2016; Kovács et al. 2019; Wang et al. 2020). Among those, (Kovács et al. 2019) proposed a network-based prediction of PPIs that relies on network walks of length 3. This approach has been shown to significantly outperform all existing link prediction approaches. Here, we use the advantage of this approach, but rely on paths of length three, to avoid the effects of direct neighbor consideration, and weigh the edges of $\overline{G}$ based on a normalized number of path length 3, Eq. (1):

$$w(u,v) = \sum_{i,j} \frac{P_{u,i}P_{i,j}P_{j,v}}{\sqrt{k_i k_j}} \tag{1}$$

where $P_{u,i}=1$ if proteins (i.e. nodes) $u$ and $i$ interact, and zero otherwise, and $k_i$ denotes the degree of node $i$.

**Algorithm 1** Preprocessing – Calculating weight for edges in $G$ *and* $\bar{G}$

```
1: procedure Preprocessing(G)
2:     weighted_edge_set ← GOSim(G)
3:     initial_edge_set ← {e|∃e ∈ weighted_edge_set, weight(e) ≥ 0.3}
4:     initial_G ← Graph(initial_edge_set)
*5:    Ḡ ← complement(initial_G )
*6:    for each edge(u, v) in Ḡ do
*7:        w(u, v) ← normalized number of path length 3
*8:        Update edge attribute of (u, v) in Ḡ
9:     return (Ḡ)
```

Next, Stoer-Wagner's algorithm (Stoer and Wagner 1994), which is a deterministic, efficient algorithm that considers positive edge weight in determining min-cuts, is applied to obtain the global min-cut of the graph $\overline{G}$.

**Algorithm 2** CUBCO+ algorithm*

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

The min-cut algorithm returns two node-subsets, $S_1$ and $S_2$, such that $S_1 \cup S_2 = V(\overline{G})$ and an edge cut set (i.e. $E_{cut}$) that connects $S_1$ to $S_2$, $E_{cut} = \{(u_i, v_i) | u_i \in S_1 \; and \; v_i \in S_2, \; 1 \leq i \leq k\}$ where $k = |E_{cut}|$. To make $\overline{G}$ disconnected and achieve the biclique spanned subgraph, $C_i$, one of the node sets, $u_i$ or $v_i$, $1 \leq i \leq k$, must be removed from $\overline{G}$; therefore, the final biclique spanned subgraph is either $C_i = \{(S_1 \cup S_2)/\bigcup_{i=1}^{k} u_i\}$ or $C_i = \{(S_1 \cup S_2)/\bigcup_{i=1}^{k} v_i\}$. Finally, a score, which exhibits the cohesiveness of the two node-set in graph $G$, is calculated as follows:

The selection of the set is guided by a score, Eq. (2), that shows the cohesiveness of the induced subgraph of the corresponding node-set in graph $G$:

$$s(C_i) = \frac{|E_{in}|}{|E_{out}|}, \tag{2}$$

where $|E_{in}|$ counts the edges inside the subgraph, and $|E_{out}|$ indicates the number of edges connecting the subgraph to the rest of the network. CUBCO+ then selects a node-set, $C_i$, with the highest score and removes it from the graph $\overline{G}$. The procedure is continued until there is no connected component left in $\overline{G}$.

The complexity of CUBCO+ is the same as the first version, CUBCO, which is $O\big(max\big(O\big(Algorithm1\big), O\big(Algorithm2\big)\big)\big)$. In Algorithm 1, the complexity of finding a complement of a graph on $n$ nodes is in $O(n^2)$. While the complexity of finding all paths of length three between two nodes is in $O(n+m)$, where $m$ is the number of edges in the graph and is calculated for every edge in $E\big(\overline{G}\big)$. Therefore, the complexity of Algorithm 1 is in $O\big(max\big(O\big(n^2\big), O(m(n+m))\big)\big) = O(m(n+m))$. The complexity of Algorithm 2 is that of Stoer-Wagner's algorithm which is in $O(n\big(m + nlogn\big))$. If we assume that in a worst-case scenario, in each iteration, we remove a node with minimum degree ($d$) and its neighbors from the graph, then the procedure will end after $\frac{n}{d+1}$ iterations. Therefore, the complexity of Algorithm 2 is in $O\Big(\frac{n^2}{d+1}\big(m + nlogn\big)\Big)$. Altogether, the complexity of CUBCO+ is dominated by the complexity of Algorithm 2.

The comparative analysis of CUBCO+ and all the other approaches were performed on an Intel(R) Xeon(R) CPU E5-2670 v2 with 2.50 GHz. Moreover, CUBCO+ is freely available on GitHub at https://github.com/SaraOmranian/CUBCOPlus.

#### Comparative analyses of CUBCO+ and other contenders across combinations of PPI networks and gold standards

We compared the predicted protein complexes from 18 contenders including CUBCO+ with the protein complexes from two *E. coli*, two *S. cerevisiae*, and one *H. sapiens* gold standards based on twelve performance measures (see Evaluation metrics, Additional file 1). The range of each of the twelve scores, including Sensitivity (SN), positive predictive value (PPV), accuracy (ACC), recall, precision, F-measure, recall$^+$, precision$^+$, F-measure$^+$, separation (SEP), fraction match (FRM), and maximum matching ratio (MMR), is between 0 and 1. Larger values for a score indicate better performance. We further calculated a composite score to summarize these twelve performance measures. The composite score is a sum of four main performance measures: MMR, FRM, ACC, and F-measure (Nepusz et al. 2012; Cao et al. 2018; Wang et al. 2018; Omranian et al. 2021a, b; Omranian and Nikoloski 2022).

In the case of *E. coli*, CUBCO+ obtains the highest FRM and ranked second after CUBCO in MMR, recall, F-measure for Babu PPI networks, and Ecocyc gold standard. Consequently, it obtained the second best composite score for this combination (Fig. 1A, Additional file 1: Table S3 and Figure S1). For the combinations of the Cong PPI network and the two gold standards of *E. coli*, CUBCO+ obtained a composite score higher than half of the other contenders. (Additional file 1: Table S3 and Figure S1).

The same analysis was carried out on all combinations of PPI networks and gold standards in *S. cerevisiae*. CUBCO+ demonstrated the highest composite score in 62.5% of cases, preceded by GCC-v and PC2P approaches that also modeled a protein complex as a biclique spanned subgraph. While in other cases, the composite score of CUBCO+ is higher than half of the compared approaches. To be precise, the composite score of CUBCO+ is, on average, only ~ 7% smaller than the composite score of the contenders ranked higher than CUBCO+ (Fig. 1B, Additional file 1: Table S3 and Figure S1).

Likewise, the composite score is calculated for the two combinations of *H. sapiens* PPI networks and one gold standard. CUBCO+ obtains composite scores on average ~ 22% smaller than the other contenders achieving higher scores (Fig. 1C, Additional file 1: Table 3 and Figure S1).

Altogether, these findings demonstrated that CUBCO+ obtains a higher composite score in ~ 36% of the cases in comparison with CUBCO, and it ranked exactly after CUBCO in other cases across all combinations of PPI networks and gold standards of the three species. Therefore, CUBCO+ illustrates almost similar performance with CUBCO. In addition, we concluded that partitioning the graph into biclique spanned subgraphs based on local properties, such as PC2P and GCC-v, resulting in better performance across all species in comparison to CUBCO and CUBCO+ that incorporates global properties in predicting protein complexes.

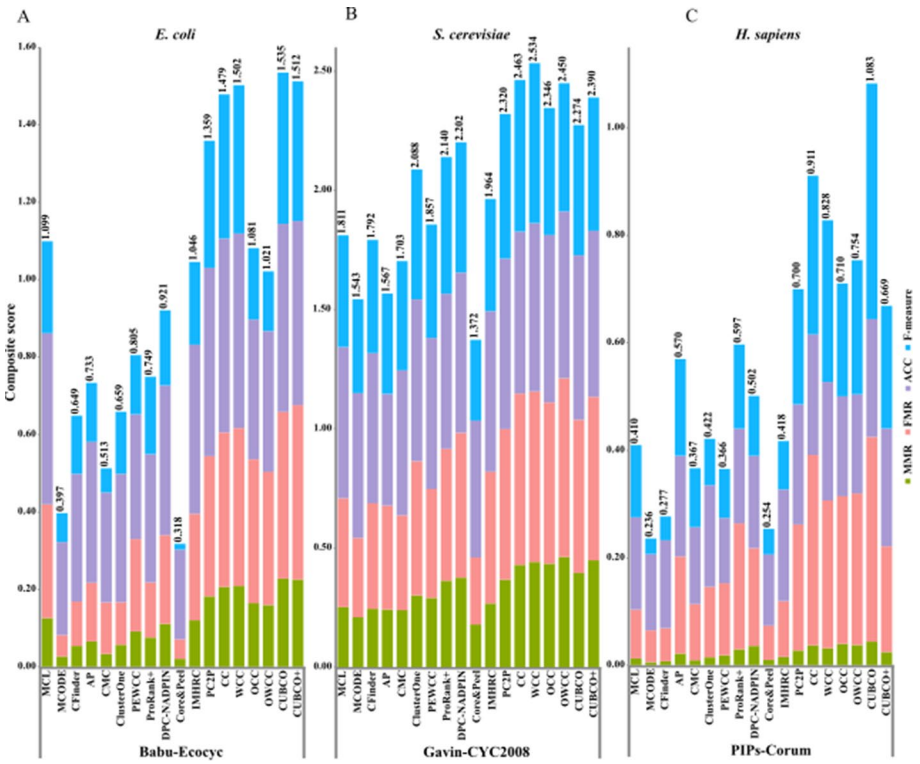


**Fig. 1** Comparative analysis of predicted protein complexes from different approaches across PPI networks of different organisms. PPI networks of **A** *E. coli*, **B** *S. cerevisiae*, and **C** *H. sapiens* are considered. The comparative analyses are based on a composite score that is a sum of ACC, F-measure, FRM, and MMR (see Evaluation metrics). Eighteen approaches are compared on three combinations of PPI networks and the gold standard. CUBCO+ performs on par with the other best-performing approaches that model a protein complex as biclique spanned subgraph, PC2P, GCC-v (CC, WCC, OCC, and OWCC), and outperforms all other approaches based on the composite score

## Method

### Contending protein complex prediction approaches

To compare the performance of CUBCO+ with other algorithms in this field, we selected seventeen state-of-the-art approaches, including minimum CUt to detect Biclique spanned subgraphs as protein COmplexes (CUBCO) (Omranian and Nikoloski 2022), a family of greedy algorithms based on clustering coefficient (GCC-v) (Omranian et al. 2021a, b), Protein Complexes from Coherent Partition (PC2P) (Omranian et al. 2021a, b), Inter Module Hub Removal Clustering (IMHRC) (Maddi and Eslahchi 2017), Core&Peel (Pellegrini et al. 2016), Discovering Protein Complexes based on Neighbor Affinity and Dynamic Protein Interaction Network (DPC-NADPIN) (Shen et al. 2016), Prorank$^+$ (Hanna and Zaki 2014), PEWCC (Zaki et al. 2013), Clustering with Overlapping Neighbourhood Extension (ClusterOne) (Nepusz et al. 2012), Clustering-based on Maximal Cliques (CMC) (Liu et al. 2009), CFinder (Adamcsek et al. 2006), Molecular Complex Detection (MCODE) (Bader and Hogue 2003), and Markov Clustering (MCL) (Enright 2002). To conduct a fair comparison, we selected the approaches that first, their implementation is publicly available, and second, do not rely on any additional data and/or knowledge such as ontologies or gene expression data (Additional file 1: Table S2).

All the contending algorithms depend on multiple parameters, except PC2P, GCC-v, and CUBCO. To optimize the parameters and obtain the best performance for each contender based on different performance measures and combinations of PPI networks and gold standards is challenging. Finding the best parameters, by optimization of various performance measures, yields divergent predicted protein complexes. Therefore, it is impossible to do meaningful interpretation and combination of the findings. Hence, the default value of parameters is used as suggested in corresponding studies.

#### PPI networks and gold standards of protein complexes

All experiments are carried out on eight PPI networks and five gold standards of three model organisms: *E. coli*., *S. cerevisiae*, and *H. sapiens*. All the PPI networks, excluding one from *E. coli,* are edge-weighted. Babu and Cong (Babu et al. 2017; Cong et al. 2019) are the two PPI networks of *E. coli*. that are used in this study, and for simplicity, we named the PPI networks the same as the corresponding first author throughout the paper. The two gold standards are generated by manually curated protein complexes from Ecocyc (Keseler et al. 2016) and protein complexes based on the genome-scale metabolic network of *E. coli* (King et al. 2015).

We used four PPI networks of *S. cerevisiae*, including Gavin (Gavin et al. 2006), Krogan core (edge-weight $\geq 0.273$), Krogan extended (edge-weight $\geq 0.101$) (Krogan et al. 2006), and Collins (Collins et al. 2007). The two gold standards were retrieved from complexes derived from the Saccharomyces Genome Database (SGD) (Hong et al. 2007) and CYC2008 (Pu et al. 2008).

For *H. sapiens*, we selected two PPI networks, STRING (edge-weight $\geq 999$) (Szklarczyk et al. 2014) and PIPs (edge-weight $\geq 25$) (McDowall et al. 2009). Moreover, we employed CORUM as the gold standard for human protein complexes (Giurgiu et al. 2018). Additional file 1: Table S1 summarizes all the properties of the used PPI networks and gold standards such as the number of proteins as well as interactions, and their intersections employed in the analyses.

#### Evaluation metrics

Twelve metrics are commonly used to evaluate the predicted protein complexes from the contending algorithms, including maximum matching ratio, fraction match (Nepusz et al. 2012), sensitivity, positive predictive value, accuracy, and separation from (Brohée and van Helden 2006), precision, recall, and F-measure from (Liu et al. 2009), and precision$^+$, recall$^+$, and F-measure$^+$ from (Maddi et al. 2019). Therefore, the predicted protein complexes are compared with complexes from gold standards across all organisms based on mentioned twelve metrics. Moreover, these metrics were selected since they have been employed in seminal studies (i.e. prediction of protein complexes) (Adamcsek et al. 2006; Liu et al. 2009; Nepusz et al. 2012; Wang et al. 2018). The twelve metrics are summarized into a composite score, which is the sum over MMR, FRM, ACC, and F-measure (Nepusz et al. 2012; Cao et al. 2018; Wang et al. 2018; Omranian et al. 2021a, b; Omranian and Nikoloski 2022). The definition and notations of evaluation metrics are comprehensively explained in the Additional file 1.

## Discussion

In this study, we propose a new method called CUBCO+ to identify protein complexes from PPI networks. The available large-scaled PPI networks, as well as gold standards of several model organisms, are obtained from different high-throughput technologies. However, these PPI networks have low quality and contain false-positive along with false-negative protein interactions. CUBCO+ disputes the noisiness and incompleteness of available PPI networks by first assigning weight to the protein interactions by utilizing GO semantic similarity to remove less biologically relevant interactions (i.e. false-positive), and second, employing network properties to rank the false-negative interactions, which are independent of biological data but capture the biological principles. To the best of our knowledge, CUBCO+ is the first algorithm that considers the effect of both false-positive and false-negative interactions.

Since we have shown that partitioning a PPI network into biclique spanned subgraphs provides the best performing approach to identify protein complexes to date (Omranian et al. 2021a, b; Omranian and Nikoloski 2022), CUBCO+ also adopted the same concept to define a protein complex. Therefore, CUBCO+ can identify protein complexes from sparse to dense graphs, since the class of biclique spanned graphs includes stars, bicliques as well as cliques as special subclasses. Hence, this feature leads to improvement of recall over the existing solutions.

We conducted thorough analyses with PPI networks from three model organisms, namely, *E. coli*, *S. cerevisiae*, and *H. sapiens*, while previous studies only evaluated their methods on at least one or two model organisms. As a result, based on twelve performance measures, CUBCO+ outperformed other approaches that are not based on biclique-spanned partitioning in ~ 64% of the cases, while in the other instances with a slight decrease in its composite score, it ranked after PC2P, GCC-v, and CUBCO. Furthermore, CUBCO+ is a parameter-free algorithm and showed consistency in its performance across all organisms.

In the future, for further improvement of the CUBCO+ algorithm, we will focus on integrating gene expression and protein abundance data into PPI networks to bring time points and dynamics to PPI networks to predict both protein complexes and functional modules.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-022-00508-5.

**Additional file 1.** Explanation of evaluation metrics; Supplementary figures for Comparative analysis of approaches for prediction of protein complexes.

**Acknowledgements**
S.O. and Z.N. would like to acknowledge the support from the Max Planck Society.

**Author contributions**
Conceived and designed the study: SO and ZN. Developed the model: SO and ZN. Implemented: SO. Wrote the manuscript: SO and ZN. Made comments and approved the final version submitted: SO and ZN. Both author reads and approved the final manuscript.

**Funding**
Open Access funding enabled and organized by Projekt DEAL.

**Availability of data and materials**
The implementation of CUBCO+ and data we used are publicy available on GitHub at https://github.com/SaraOmranian/CUBCOPlus.

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

Received: 10 March 2022 Accepted: 19 September 2022
Published online: 11 October 2022

**References**

Adamcsek B et al (2006) CFinder: locating cliques and overlapping modules in biological networks. Bioinformatics 22:1021–1023
Akiyama J, Harary F (1981) A graph and its complement with specified properties. IV. Counting self-complementary blocks. J Graph Theory 5:103–107
Angeleska A, Nikoloski Z (2019) Coherent network partitions. Discret Appl Math 266:283–290
Angeleska A, Omranian S, Nikoloski Z (2021) Coherent network partitions: characterizations with cographs and prime graphs. Theor Comput Sci 894:3–11
Babu M et al (2017) Global landscape of cell envelope protein complexes in Escherichia coli. Nat Biotechnol 36:103–112
Bader GD, Hogue CWV (2003) An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinform 4:2
Bauer A, Kuster B (2003) Affinity purification-mass spectrometry. Eur J Biochem 270:570–578
Berger B, Peng J, Singh M (2013) Computational solutions for omics data. Nat Rev Genet 14:333–346
Bhowmick SS, Seah BS (2016) Clustering and summarizing protein-protein interaction networks: a survey. IEEE Trans Knowl Data Eng 28:638–658
Brohée S, van Helden J (2006) Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinform 7:1
Cao B et al (2018) Detection of protein complexes based on penalized matrix decomposition in a sparse protein-protein interaction network. Molecules 23:1460
Cho Y-R, Hwang W, Ramanathan M, Zhang A (2007) Semantic integration to identify overlapping functional modules in protein interaction networks. BMC Bioinform 8:1
Collins SR et al (2007) Toward a Comprehensive atlas of the physical interactome of saccharomyces cerevisiae. Mol Cell Proteom 6:439–450
Cong Q, Anishchenko I, Ovchinnikov S, Baker D (2019) Protein interaction networks revealed by proteome coevolution. Science 365:185–189
Dost B et al (2008) QNet: a tool for querying protein interaction networks. J Comput Biol 15:913–925
Enright AJ (2002) An efficient algorithm for large-scale detection of protein families. Nucl Acids Res 30:1575–1584
Feng J, Jiang R, Jiang T (2011) A max-flow-based approach to the identification of protein complexes using protein interaction and microarray data. IEEE/ACM Trans Comput Biol Bioinf 8:621–634
Fields S, Sternglanz R (1994) The two-hybrid system: an assay for protein-protein interactions. Trends Gene 10:286–292
Fröhlich H, Speer N, Poustka A, Beißbarth T (2007) GOSim – an R-package for computation of information theoretic GO similarities between terms and gene products. BMC Bioinform 8:1
Fujikawa Y, Kato N (2007) TECHNICAL ADVANCE: split luciferase complementation assay to study protein-protein interactions in Arabidopsis protoplasts. Plant J 52:185–195
Gavin A-C et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415:141–147
Gavin A-C et al (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440:631–636
Giurgiu M et al (2018) CORUM: the comprehensive resource of mammalian protein complexes—2019. Nucl Acids Res 47:D559–D563
Hanna EM, Zaki N (2014) Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. BMC Bioinform 15:1
Ho Y et al (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415:180–183
Hong EL et al (2007) Gene Ontology annotations at SGD: new data sources and annotation methods. Nucl Acids Res 36:D577–D581
Hu L et al (2021) A survey on computational models for predicting protein–protein interactions. Brief Bioinform 22:p. bbab036
Keseler IM et al (2016) The EcoCyc database: reflecting new knowledge aboutEscherichia coliK-12. Nucl Acids Res 45:D543–D550
King AD, Przulj N, Jurisica I (2004) Protein complex prediction via cost-based clustering. Bioinformatics 20:3013–3020
King ZA et al (2015) BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. Nucl Acids Res 44:D515–D522
Kovács IA et al (2019) Network-based prediction of protein interactions. Nat Commun 10:1
Krogan NJ et al (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440:637–643

Li X, Wu M, Kwoh C-K, Ng S-K (2010) Computational approaches for detecting protein complexes from protein interaction networks: a survey. BMC Genom 11:S3
Lin J-S, Lai E-M (2017) Protein-Protein interactions: co-immunoprecipitation. Methods in molecular biology. Springer New York, pp 211–219
Liu G, Wong L, Chua HN (2009) Complex discovery from weighted PPI networks. Bioinform 25:1891–1897
Maddi AMA, Eslahchi C (2017) Discovering overlapped protein complexes from weighted PPI networks by removing inter-module hubs. Sci Rep 7:1
Maddi AMA, Moughari FA, Balouchi MM, Eslahchi C (2019) CDAP: an online package for evaluation of complex detection methods. Sci Rep 9:1
Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296:910–913
McBride Z et al (2019) A label-free mass spectrometry method to predict endogenous protein complex composition∗. Mol Cel Proteom 18:1588–1606
McDowall MD, Scott MS, Barton GJ (2009) PIPs: human protein-protein interaction prediction database. Nucl Acids Res. 37:D651–D656
McWhite CD et al (2020) A pan-plant protein complex map reveals deep conservation and novel assemblies. Cell 181:460-474.e14
Mewes HW (2004) MIPS: analysis and annotation of proteins from whole genomes. Nucl Acids Res 32:41D – 44
Nepusz T, Yu H, Paccanaro A (2012) Detecting overlapping protein complexes in protein-protein interaction networks. Nat Methods 9:471–472
Omranian S, Nikoloski Z (2022) CUBCO: prediction of protein complexes based on min-cut network partitioning into biclique spanned subgraphs. Complex networks & their applications X. Springer International Publishing, pp 605–615
Omranian S, Angeleska A, Nikoloski Z (2021a) Efficient and accurate identification of protein complexes from protein-protein interaction networks based on the clustering coefficient. Comput Struct Biotechnol J 19:5255–5263
Omranian S, Angeleska A, Nikoloski Z (2021b) PC2P: parameter-free network-based prediction of protein complexes. Bioinformatics 37:73–81
Omranian S, Nikoloski Z, Grimm DG (2022) Computational identification of protein complexes from network interactions: present state, challenges, and the way forward. Comput Struct Biotechnol J 20:2699–2712
Pawson T, Nash P (2000) Protein–protein interactions define specificity in signal transduction. Genes Dev 14:1027–1047
Pellegrini M, Baglioni M, Geraci F (2016) Protein complex prediction for large protein protein interaction networks with the Core&Peel method. BMC Bioinform 17:37
Pu S et al (2008) Up-to-date catalogues of yeast protein complexes. Nucl Acids Res 37:825–831
Qi Y et al (2008) Protein complex identification by supervised graph local clustering. Bioinformatics 24:i250–i268
Reyes-Turcu FE, Ventii KH, Wilkinson KD (2009) Regulation and cellular roles of ubiquitin-specific deubiquitinating enzymes. Annu Rev Biochem 78:363–397
Sharan R et al (2005) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. J Comput Biol 12:835–846
Sharma P, Bhattacharyya DK, Kalita JK (2018) Detecting protein complexes based on a combination of topological and biological properties in protein-protein interaction network. J Genet Eng Biotechnol 16:217–226
Shen X et al (2016) Neighbor affinity based algorithm for discovering temporal protein complex from dynamic PPI network. Methods 110:90–96
Shi L, Lei X, Zhang A (2011) Protein complex detection with semi-supervised learning in protein interaction networks. Proteom Sci 9:S5
Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Nat Acad Sci 100:12123–12128
Srihari S, Leong HW (2013) A survey of computational methods for protein complex prediction from protein interaction network. J Bioinform Comput Biol 11:1230002
Stoer M, Wagner F (1994) A simple min cut algorithm. Algorithms — ESA\textquotesingle94. Springer Berlin Heidelberg, pp 141–147
Sweetlove LJ, Fernie AR (2018) The role of dynamic enzyme assemblies and substrate channelling in metabolic regulation. Nat Commun 9:1
Szklarczyk D et al (2014) STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucl Acids Res 43:D447–D452
Tian W et al (2008) Combining guilt-by-association and guilt-by-profiling to predict Saccharomyces cerevisiae gene function. Genome Biol 9:S7
Wang R et al (2018) Predicting overlapping protein complexes based on core-attachment and a local modularity structure. BMC Bioinform 19:1
Wang X, Hu P, Hu L (2020) A novel stochastic block model for network-based prediction of protein-protein interactions. Intelligent computing theories and application. Springer International Publishing, pp 621–632
Wu Z, Liao Q, Liu B (2019) A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. Brief Bioinform 21:1531–1548
Zaki N, Efimov D, Berengueres J (2013) Protein complex detection using interaction reliability assessment and weighted clustering coefficient. BMC Bioinform. https://doi.org/10.1186/1471-2105-14-163
Zeng S (2016) Link prediction based on local information considering preferential attachment. Phys A Stat Mech Appl 443:537–542

## Publisher's Note