ARTICLE

# A Hybrid method for protein complex prediction in weighted protein-protein interaction networks

Jerome Cary Beltran[1], Catalina Montes[1], John Justine S. Villar*[1,2], Adrian Roy L. Valdez[1]

[1]Scientific Computing Laboratory, Department of Computer Science, University of the Philippines 1101 Diliman, Quezon City, Philippines
[2]Institute of Chemistry, Faculty of Materials Science and Engineering, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

Due to the significance of protein-protein interactions (PPIs) in regulating many significant cellular functions, many studies have focused on detecting protein complexes within PPI networks (PPINs) using computational methods. While a number of these methods are based on graph clustering, experimental studies have revealed that several relevant biological insights about protein complexes are not reflected in these methods. This paper proposes an algorithm that combines an extension of the Markov cluster algorithm (MCL), called the MLR-MCL with balance, and a core-attachment scheme to cluster PPINs. This algorithm was run on the BioGRID and DIP yeast PPI networks, and the output clusters were compared against the CYC2008 protein complex data (Pu et al 2009) by computing F-scores for the predicted complexes. The clustering results showed an improvement in average F-scores between 25.6% to 153.3% with respect to those resulting from clustering done on two datasets, as compared to three other clustering algorithms. Also, the proposed algorithm yielded an improvement of 59.1% for BioGRID and 81.4% for DIP dataset, as compared to original MLR-MCL with balance. These values reflect the positive effect of applying biological information to a pure, graph-theoretic clustering algorithm.

*Corresponding author
Email Address: john_justine.villar@upd.edu.ph
Date Received: 04 July 2016
Date Revised: 01 May 2017
Date Accepted: 02 June 2017

**KEYWORDS**

protein complex prediction, weighted protein-protein interaction network, Markov clustering, core-attachment scheme, graph clustering

## INTRODUCTION

Cellular functions and biochemical events involve complicated interactions among proteins, which is commonly referred to as *protein-protein interactions* (PPIs). These proteins then aggregate to form larger macromolecules, called *complexes*, to regulate specific molecular responses under various physiological conditions. Identifying and characterizing protein complexes from PPI networks (PPINs) are very important in understanding biomolecular processes and organization since it helps reveal the structure-function relationships among complexes.

High-throughput experimental methods have produced a large amount of protein interaction data, which makes it possible to predict complexes from protein-protein interaction networks. However, the relatively small amount of known physical interactions and noisy experimental data may limit complex detection.

Many computational methods that are developed for protein complex detection on PPINs are mainly based on graph clustering (Bader and Hogue 2003, van Dongen 2000, Liu et al. 2009, Nepusz et al. 2012, Chua et al. 2008), which rely solely on the topology of the PPI network. A popular method, called the Markov Clustering Algorithm (MCL), was developed by Van Dongen in 2000 (van Dongen 2000, Enright et al. 2002)

based on simulating a random walk on a PPIN. Due to its robustness, many extensions have been produced from this method. One of its variants is MLR-MCL with balance, which is a multi-level, regularized algorithm that uses a balance parameter to even out the sizes of output clusters (Satuluri et al. 2010). A discussion on this algorithm is presented on Section 2. Despite progress in protein complex detection, it is a great challenge to effectively analyze the massive data for biologically meaningful protein complex detection. More recent studies on protein complex detection focused on applying biological insights into existing graph-theoretic methods. Srihari and others have surveyed a variety of these methods, and discussed that incorporating biological information in network mining improved the performance of the methods (Srihari et al. 2015).
One finding pertains to the inherent organization of the complexes in the network. Many studies assume protein complexes to be dense subgraphs, since most complexes involve multiple interacting proteins to perform specific functions (Pu et al. 2009, Becker et al. 2012). However, studies have revealed that some of these complexes share the same proteins, i.e. have overlapping regions (Palla et al. 2005). These overlapping complexes are not usually captured by some clustering methods, such as MCL (van Dongen 2000), which yields tree clusters, and MCODE (Bader and Hogue 2003), which yield highly dense but disjoint complexes. These methods classify closely related complexes, which share a significant amount of proteins, as a whole when they share many nodes in the network. This problem is called the "soft clustering problem" in graph theory.
In 2006, Gavin and others proposed that there are two components in every complex: a set of *core* proteins and *attachments* (Gavin et al. 2006). Core proteins are central to each complex, with relatively more interactions among themselves, while attachment proteins bind to the core proteins and may appear in several complexes. These attachment proteins assist the complex core to perform secondary functions. This idea of biological organization in the PPI network defines the framework that will be called *core-attachment structure* in the paper.

Several papers used the concept of core-attachment structure in computational protein complex detection (Wu et al. 2009, Srihari et al. 2010, Srihari et al. 2015). The results of these studies show that applying this idea in graph clustering improves its predictive performance, in terms of cluster coverage and sensitivity, as compared to the pure graph-theoretic algorithms.
This study aims to combine the performance of MLR-MCL with balance algorithm with the core-attachment structure to produce a hybrid method to improve clustering efficiency, with the further intent of observing the effects of incorporating biological information to existing graph-theoretic methods.

## PRELIMINARIES

### A. Graph Theory

The network of interactions between proteins (which is also referred to as a PPI network) can be represented by an undirected graph $G = (V, E)$ excluding self-loops, where $V$ be the set of vertices, $E \subseteq V \times V$ be the set of edges, with each edge $(v_i, v_j)$, $v_i, v_j \in V$, and suppose $|V| = n$. The set $V$ represents the set of distinct proteins and $E$ the set of known interactions between two proteins. In this paper, it is assumed that multiple interactions between the same pair of proteins and transient interactions among the proteins are not considered in the construction of the PPI network.

A protein complex is a set of proteins in $V$ bound together to form a stable structure. In other words, the proteins in the complex, with their interactions, form a subgraph of the PPI network. However, not all subgraphs in the PPI network are relevant complexes.

Define $w: V \times V \to \mathbb{R}_{\geq 0}$ to be the weight function of the graph $G$, where $w(v_i, v_j) > 0$ if $(v_i, v_j) \in E$ and $w(v_i, v_j) = 0$ if $(v_i, v_j) \notin E$. This mapping represents the confidence level of the interaction of a PPI network.

Let $A \in M_n(\mathbb{R})$ be the (weighted) adjacency matrix corresponding to $G$, such that

$$A_{ij} = \begin{cases} w(v_i, v_j) & (v_i, v_j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

Define $M \in M_n(\mathbb{R})$ to be the canonical transition matrix, which is a column-stochastic matrix, i.e., for every column $j$, $\sum_i M_{ij} = 1$. This represents the matrix of transition probabilities of a random walk (or a Markov chain) defined on $G$, with $M_{ij}$ being the transition probability from $v_i$ to $v_j$ (with $M_{ji} = M_{ij}$ since $G$ is undirected).

### B. The Markov Cluster Algorithm and its Variants

The Markov cluster algorithm (MCL) is a clustering method based on a simulation of stochastic flows on the graph (van Dongen 2000). The algorithm simulates random walks on the graph, in which it enhances the flows that tend to gather, and then yields resulting clusters on the graph. This amplification is done based on the idea that there are more paths between two nodes in a cluster than between those in different clusters (van Dongen 2000). Thus, from any given node, there is a higher probability of "walking" to a node within a cluster than to a node in another cluster.

The MCL process consists of two operations on stochastic matrices, referred to as $Expand$ and $Inflate$. The $Expand$ step spreads the flow out of a vertex to potentially new vertices and also enhances the flow to those vertices which are reachable by multiple paths, which has the effect of strengthening intra-cluster flows. The $Inflate$ step introduces a nonlinearity into the process, with the purpose of enhancing intra-cluster flow and weakening inter-cluster flow. Thus, given a canonical flow matrix $M$,

$$Expand(M) = M * M \text{ , and}$$
$$Inflate\ (M, r)[i, j] = \frac{(M_{ij})^r}{\sum_{k=1}^{n} (M_{kj})^r},$$

with $r > 1, r \in \mathbb{R}$. These two operators are applied in alternation iteratively, starting with the canonical flow matrix.

Because of its speed and scalability, many variants of MCL have been developed with the aim of producing more accurate clustering results with respect to experimentally-curated clusters. One variant is the Regularized MCL (R-MCL) algorithm, in which the $Expand$ step is replaced by the operation $Regularize(M) = M \cdot M_G$, where $M_G \in M_n(\mathbb{R})$ is the original matrix of transition probabilities (Satuluri and Parthasarathy 2009), given by

$$M_G[i, j] = \frac{A_{ij}}{\sum_{k=1}^{n} A_{kj}},$$

where $A \in M_n(\mathbb{R})$ is the adjacency matrix of the graph. This approach uses the original topology of the graph to influence the clustering results throughout all iterations, and not just during the first iteration.

An extension of R-MCL, called MLR-MCL, embeds the former within a multi-level framework (Satuluri and Parthasarathy 2009). In this algorithm, the input graph is successively coarsened into a chain of smaller graphs, until a sufficiently small graph is obtained. A few iterations of R-MCL are run on the coarsest graph, and the flow matrix at the end of these few

iterations is used to initialize a few iterations of R-MCL on the next bigger graph, and so on up the chain of graphs until the original graph is reached, and the clusters are returned. It has been observed that MLR-MCL has faster execution time than both R-MCL and MCL, as running the process first on smaller graphs is faster and the flow matrix is sparse by the time it proceeds to the bigger graphs, leading to smaller matrix multiplications.

An improvement of MLR-MCL, which incorporates adjustable balance, has been proposed to address the skewed clustering results produced by the original MCL algorithm (Satuluri et al. 2010). Instead of using $M_G$ in the R-MCL step, a different matrix, $M_R$, is used. The matrix $M_R$ is formed using a penalized $M_G$, which is penalized in a way that nodes will be likely to join a smaller cluster, thus balancing out cluster sizes. The reader is referred to (Satuluri et al. 2010) for the full details of the MLR-MCL algorithm.

**C. Core-Attachment Structure**

Most of the available protein complex detection algorithms are based on the assumption that densely connected proteins, i.e., dense subgraphs, correspond to complexes in the PPIN. However, these methods fail to consider the inherent organization among protein complexes and the roles of the edges in it.

Core-attachment structure is a framework that many researchers have utilized for designing and improving methods for PPIN clustering, and describes the protein complex organization based on the analysis of experimentally detected protein complexes (Dezso et al. 2003, Gavin et al. 2006). It observes that a protein complex consists of a core, which contains proteins that are highly co-expressed and with strong functional similarity, and attachments to the core, which are other proteins within the complex that help the core proteins carry out their functions. It also implies that protein complexes can share attachment proteins.

The core-attachment structure anchors mainly on three properties (Gavin et al. 2006):

1. the core proteins of a complex constitute a subgraph of the PPI network, with relatively high node degree among themselves,
2. every set of core proteins are disjoint, and
3. if an attachment protein is linked to a subset of core proteins, the attachment protein will be a common neighbor of the subset of core proteins it is connected to in the PPI network.

One of the methods that use core-attachment structure for protein complex detection is MCL-CAw (Srihari et al. 2010), which initially runs MCL on a weighted PPI network and then refines the resulting clusters by identifying the core proteins and attachment proteins within each cluster. For each cluster, core proteins are first identified by considering three measures: *weighted in-connectivity* of a protein, *weighted out-connectivity* of a protein, and *average weighted in-connectivity* of the cluster a protein belongs to. Weighted in-connectivity specifically refers to the sum of a protein's interactions with respect to the other proteins within its cluster, while weighted out-connectivity refers to the sum of a protein's interactions with respect to other proteins outside of its cluster. Moreover, the average weighted in-connectivity refers to the average of the weighted in-connectivities of every protein within the cluster. Given a protein $p$ and a cluster $C_i = (V_i, E_i), V_i \subseteq V, E_i \subseteq E$, the weighted in-connectivity $d_{in}$, weighted out-connectivity $d_{out}$, and average weighted in-connectivity $d_{avg}$, are respectively given by the formulae

$$d_{in}(p, C_i) = \sum_{q \in V_i} w(p, q),$$

$$d_{out}(p, C_i) = \sum_{q \notin V_i} w(p, q),$$

$$d_{avg}(C_i) = \frac{\sum_{q \in V_i}[d_{in}(q, C_i)]}{|C_i|}.$$

For a protein to be considered a core protein, its weighted in-connectivity must be greater than its weighted out-connectivity, and its weighted in-connectivity must be greater than or equal to the average weighted in-connectivity of its cluster.

Non-core proteins are then classified as attachments to the cluster by the following criterion:

$$I_P \geq \alpha \cdot I_C \left(\frac{S_C}{2}\right)^{-\gamma}$$

where $I_P$ is the sum of all interactions of a protein with respect to the core proteins, $I_C$ is the sum of all interactions among the core proteins of the cluster, and $S_C$ is the number of core proteins in the cluster. The user-defined parameters $\alpha$ and $\gamma$ control the effects of the total weight of interactions among the core proteins of a donor cluster and the number of core proteins of a donor cluster, respectively. The final set of complexes is the merged set of cores and attachments, as obtained above.

The pseudocode of the core-attachment structure is presented in Algorithm B1 found in the Supplementary Information. The reader is referred to (Srihari et al. 2010) for the full discussion on the scheme used in this paper.

**D. Functional Similarity as Interaction Reliability Metric**

The reliability of protein interactions in a PPIN can be expressed as an assignment of weights to protein pairs that reflect how likely they will interact with each other, given the topological characteristics of the network and/or other external information. The score is directly proportional to the likelihood of interaction between a pair of proteins.

The functional similarity (FS)-weighting method is based on the observation that the similarity of function between two proteins is highly correlated with the number of interaction partners they have in common (Chua et al. 2008). FS-weighting assumes that proteins can share the same function in two ways: proteins can either interact directly to perform common functions (hence, they have direct functional associations), or indirectly by having many common interaction partners. In other words, if two proteins do not directly interact but do with many other common proteins, it is highly likely that they share similar physical or biochemical characteristics, and therefore share the same functions as well (Chua et al. 2006).

The FS-weight between two proteins $v_i$ and $v_j$ is given by the formula (Chua et al. 2008):

$$w_{FS}(v_i, v_j) = \frac{2\left|N_{v_i} \cap N_{v_j}\right|}{\left|N_{v_i} - N_{v_j}\right| + 2\left|N_{v_i} \cap N_{v_j}\right| + \lambda_{v_i,v_j}} \times \frac{2\left|N_{v_i} \cap N_{v_j}\right|}{\left|N_{v_j} - N_{v_i}\right| + 2\left|N_{v_i} \cap N_{v_j}\right| + \lambda_{v_j,v_i}},$$

where $N_v$ is the set that contains the protein $v$ and the proteins it directly interacts with. Also, $\lambda_{v_i,v_j}$ is given by

$$\lambda_{v_i,v_j} = \max\left[0, n_{avg} - \left(\left|N_{v_i} - N_{v_j}\right| + \left|N_{v_i} \cap N_{v_j}\right|\right)\right],$$

where $n_{avg}$ refers to the average number of neighbors per protein in the network.

In this paper, the weights are bounded by a threshold parameter $\tau$, wherein interactions whose weights fall below the threshold value are removed from the network. This filtering is usually applied to reduce noise related to indirect interactions among proteins. The pseudocode for weighting process used in this paper is detailed in Algorithm A1 found in the Supplementary Information.

### E. The Algorithm

The hybrid algorithm, composed of MLR-MCL with balance (Satuluri et al. 2010) and an adapted core-attachment structure from (Srihari et al. 2010), is presented in Algorithm 1. This consists of three phases: (1) a preprocessing step, where the (weighted) adjacency matrix $A$ is constructed, with respect to the FS-weighting formula in Section 2.3. This is followed by (2) a clustering step, where the (weighted) adjacency matrix is using the MLR-MCL with balance algorithm in (Satuluri et al. 2010). The resulting clusters are further refined in (3), by using the core-attachment structure algorithm in (Srihari et al. 2010), in which the pseudocode is presented in Algorithm B1 found in the Supplementary Information.

**Algorithm 1** Proposed Algorithm

```
procedure PROPOSED_ALGO(A, τ, r, b, α, γ)
    Input: PPI network G = (V, E), weight threshold parameter τ, inflation parameter r,
           balance parameter b, weight parameters α, γ

    //Phase 1: Pre-processing using FS-weighting (Supplementary Information A)
    A ← FS-WEIGHTING(G, τ)

    //Phase 2: Initial clustering using MLR-MCL with balance (Satuluri et al. 2010, Algorithm 6)
    C ← MLR-MCLWITHBALANCE(A, r, b)

    //Phase 3: Refined clustering using the core-attachment scheme (Supplementary Information B)
    C* ← CORE_ATTACHMENT(C, α, γ)

    Interpret C* as a clustering
end procedure
```

## MATERIALS AND METHODS

This section details the properties of the two yeast PPINs used in this study, as well as cluster quality metrics to determine the relative performance of the proposed algorithm, in comparison to other clustering algorithms.

### A. Experimental Setup

The algorithm was run on two *Saccharomyces cerevisiae* PPI networks, namely Biological General Repository for Interaction Datasets (BioGRID) version 3.1.81 (Stark et al 2011) and Database of Interacting Proteins (DIP) version October 27, 2011 (Salwinski et al 2004). The DIP dataset, with 4,995 nodes and 21,875 interactions, contains data from genome-wide yeast two-hybrid screens. Meanwhile, for the BioGRID dataset, which includes 4,364 nodes and 25,464 interactions, only data from low-throughput experiments were considered as these interactions have higher precision (Paccanaro et al 2005). Self-loops are also removed from both datasets.

Note that the two databases apply different rules for capturing the data and often use different systems for cross-referencing genes and proteins across biological databases. For example, the interaction curation in BioGRID mainly follow the "spoke" bait-hit model, where directly pairs bait proteins with associated proteins is applied, with the inclusion of self-interactions, and reciprocal interactions if the bait-hit directionality is clear (Stark et al 2011). Furthermore, DIP includes interactions based on the reliability of individual experimental methods using an expression profile reliability index, analysis of the patterns of interactions between analogous proteins using the paralogous verification method, and domain-domain interaction preferences using the domain pair verification method (Salwinski et al 2004). The reader is directed also to the respective database websites for updated and detailed discussion of the curation guidelines.

The first step of the proposed method is preprocessing through FS-weighting, which modifies the topological properties of the unweighted BioGRID and DIP networks. In this study, FS-weights are bounded by the threshold parameter $\tau$, which is set at 0.2. Table 1 summarizes the profile of both PPI networks before and after the preprocessing step.

**Table 1: Network profile of BioGRID and DIP datasets before and after FS-weighting**

| | Dataset | Vertices | Edges | Average Degree | Avg Clustering Coeff |
|---|---|---|---|---|---|
| BioGRID | before weighting | 4 364 | 25 464 | 11.670 | 0.240 |
| | after weighting | 2 192 | 6 687 | 6.101 | 0.584 |
| DIP | before weighting | 4 995 | 21 875 | 8.759 | 0.123 |
| | after weighting | 1 736 | 6 231 | 7.179 | 0.627 |

Furthermore, in the MLR-MCL with balance phase of the proposed method, the balance parameter $b$ is set to 0.5, and inflation parameter $r$ is fixed to 2.0. The weight parameters for the core-attachment scheme are set at $\alpha = 1.0, \gamma = 0.75$. These set of parameters are chosen, based on the default parameters used in (Satuluri et al. 2010) for the balance and inflation parameters, and in (Srihari et al. 2010) from core-attachment weight parameters.

For evaluating the clustering results of the two yeast PPI networks, the CYC2008 protein complex data (Pu et al. 2009) was used as the gold standard for cluster validation. This dataset is a list of 408 experimentally validated protein complexes in *S. cerevisiae*, with 1,920 annotations involving a total of 1,627 proteins (with some proteins having multiple annotations).

### B. Cluster Validation

In order to study the relative performance of different supervised learning algorithms, we need to determine whether a predicted protein complex matches a complex in benchmark set. In (Satuluri et al. 2010), the authors used the precision, recall, and F-score as the criteria, which is defined below.

Let $B = \{B_i\}_{i=1}^{m}$ and $C = \{C_i\}_{j=1}^{n}$ be the sets of benchmark and predicted complexes, respectively. Given a predicted complex $C_j$, the precision and recall value of $C_j$ is calculated based on a benchmark complex in $B$. The precision (sensitivity) $Prec$ and recall (coverage) $Rec$ of $C_j$ is, respectively, defined as

$$Prec(C_j, B_i) = \frac{|C_j \cap B_i|}{|C_j|},$$

$$Rec(C_j, B_i) = \frac{|C_j \cap B_i|}{|B_i|}$$

The quality of predicted complexes is measured using the F-score metric. The F-score $F$ of a given complex $C_j$ is given by

$$F(C_j) = \max_i \frac{2 \cdot Prec(C_j, B_i) \cdot Rec(C_j, B_i)}{Prec(C_j, B_i) + Rec(C_j, B_i)}$$

Note that each predicted complex is matched with the benchmark complex, and then the maximal F-score among all benchmark complexes is identified. For a given clustering result,

the (weighted) average F-score $F_{avg}$ is the weighted average of all maximal F-scores associated per predicted complex, that is:

$$F_{avg}(C) = \frac{\sum_{j=1}^{n} |C_j| F(C_j)}{\sum_{j=1}^{n} |C_j|},$$

where $C_j$ refers to a predicted complex $j$, $j = 1, \ldots, n$, and $F(C_j)$ refers to the maximal F-score matched to the predicted complex $j$. Similarly, the average precision $Prec_{avg}$ and the average recall $Rec_{avg}$ are the weighted averages of the maximal precision and recall of each predicted complex, respectively.

In this paper, the average F-score is computed by comparing the predicted complexes against the CYC2008 catalog (Pu et al. 2009), which is "a comprehensive catalog of manually curated 408 heteromeric protein complexes in *S. cerevisiae* reliably backed by small-scale experiments from the literature".

## RESULTS AND DISCUSSION

This section presents the relative performance of the proposed algorithm, in comparison to four other clustering algorithms over two *Saccharomyces cerevisiae* yeast datasets. The performance of the algorithm was tested across four data preprocessing scenarios, with respect to their average F-scores, as well as their respective precision and recall values.

### A. Comparative Evaluation

For comparison, four other clustering algorithms were also run on the two datasets, namely the (plain) MLR-MCL with balance (Satuluri et al. 2010), MCL-CAw (Srihari et al. 2010), PCP (Chua et al. 2008), and COACH (Wu et al. 2009). The PCP algorithm applies FS-weight scoring scheme to remove unreliable interactions and add indirect interactions, and then merges cliques to produce the final list of complexes. Moreover, COACH also utilizes core-attachment method in complex detection, which starts with characterizing the core proteins from neighborhood graphs and forms protein complexes by including attachments into cores. Table 2 summarizes the different features of the five clustering algorithms.

**Table 2: Features of five clustering methods**

| Method | Weighting Method | Clustering Method | Core-Attachment Structure |
|---|---|---|---|
| COACH | none | Neighborhood affinity | yes |
| PCP | FS-weighting | Clique-merging | no |
| MLR-MCL | none | Markov (regularized) | no |
| MCL-CAw | FS-weighting | Markov | yes |
| Proposed | FS-weighting | Markov (regularized) | yes |

The average F-score of the four algorithms, as well as their average precision and average recall, obtained for the BioGRID and DIP datasets, are presented in Tables 3 and 4, respectively.

**Table 3: Performance of Different Methods for BioGRID Dataset**

| Method | $Prec_{avg}$ | $Rec_{avg}$ | $F_{avg}$ |
|---|---|---|---|
| COACH | 0.250316 | 0.634635 | 0.305980 |
| PCP | 0.262276 | 0.697915 | 0.311961 |
| MLR-MCL | 0.247995 | 0.693017 | 0.289343 |
| MCL-CAw | 0.354374 | 0.543846 | 0.350216 |
| **Proposed** | **0.427992** | **0.654529** | **0.439998** |

**Table 4: Performance of Different Methods for DIP Dataset**

| Method | $Prec_{avg}$ | $Rec_{avg}$ | $F_{avg}$ |
|---|---|---|---|
| COACH | 0.296630 | 0.581313 | 0.327533 |
| PCP | 0.233694 | 0.594110 | 0.270041 |
| MLR-MCL | 0.279371 | 0.548459 | 0.306390 |
| MCL-CAw | 0.345003 | 0.477677 | 0.332294 |
| **Proposed** | **0.525140** | **0.673680** | **0.525485** |

It is evident in Tables 3 and 4 that the proposed method outperforms the other four algorithms, in terms of average F-score, and over two datasets.

It is worth noting that the F-score for the proposed algorithm increased by 52% for BioGRID dataset and 71.5% for DIP dataset, as compared to the plain MLR-MCL with balance algorithm. This may signify the importance of including interaction reliability in the process of clustering proteins. There is also a raise in F-score of around 25.6% (BioGRID) and 58.1% (DIP), in comparison to MCL-CAw algorithm, which may imply that balance and scalability in clustering affects the quality of the complexes detected. Furthermore, the average F-score increased by 43.8% (BioGRID) and 60.4% (DIP) for COACH, and 41% (BioGRID) and 94.6% (DIP) for PCP.

The proposed method also achieved the highest average precision values over the five algorithms tested and over the two PPINs. For the DIP dataset, the proposed method showed considerable improvement of at least 10.2% in average recall over four other algorithms tested in the study. However, the MLR-MCL and PCP algorithms got higher average recall values, by around 6%, over the BioGRID dataset than the proposed method, which may imply that the latter method is more conservative in extracting out complexes from clusters. Further refinement of the algorithm, by varying the parameters needed by the algorithm, is recommended for future study.

On the other hand, a significant fraction of the proteins in the two PPINs do not have any experimental annotations, with only 27.2% and 28.2% of the proteins are annotated in BioGRID and DIP, respectively. This means that the precision (i.e., the ratio of predicted complexes with annotations over all output clusters) is expected to be low for many output clusters.

### B. Effects of Data Preprocessing in Clustering Results

There are proteins annotated in the CYC2008 that are not present in the BioGRID and DIP datasets, i.e., three are completely absent in the BioGRID data and 288 proteins are completely absent in the DIP data. With these disparities, the average F-scores were computed under two cases. The first case consists of the average F-score computed against the complete CYC2008, while the second case consists of the average F-score computed against a version of the CYC2008 tailored in such a way that this only includes annotations of proteins present in a dataset and discards any complexes having only one protein member as a result of eliminating the annotations of absent proteins. Table 5 summarizes the different scenarios considered in the study, under which the average F-score of the clustering results was computed.

**Table 5: Dataset Scenarios**

| | Include unidentified clusters | Exclude unidentified clusters |
|---|---|---|
| Tailored Benchmark | Scenario A | Scenario B |
| Complete Benchmark | Scenario C | Scenario D |

Figures 1a and 1b details the average F-scores resulting from runs of the five algorithms on the BioGRID and DIP datasets, respectively.

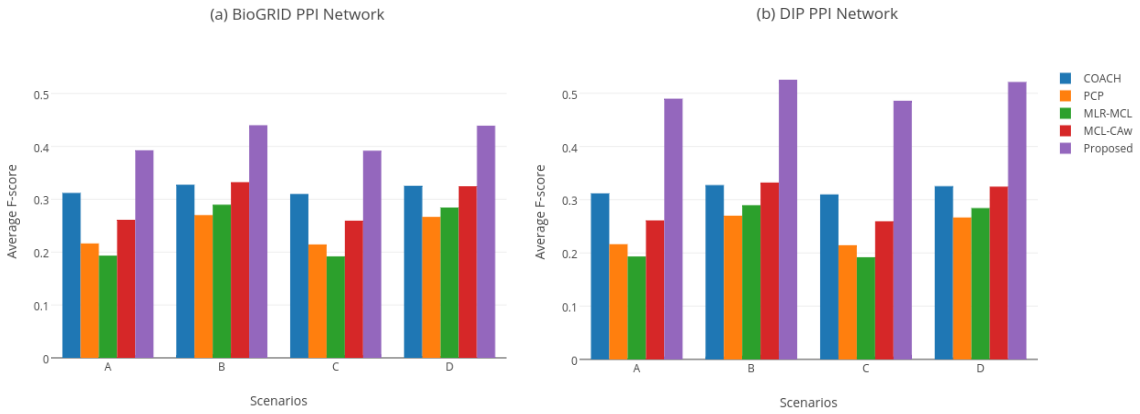


**Figure 1: Average F-scores of the algorithms across the four data preprocessing scenarios over (a) BioGRID dataset (b) DIP dataset**

The improvement in average F-scores ranges from 25.6% to 96.3% across the four data preprocessing scenarios for the BioGRID dataset, and from approximately 58.1% to 153.3% for the DIP dataset.

It is consistent across four scenarios that the core attachment structure has a significant effect in enhancing the cluster quality, as it yielded an increase of 52% – 96.6% for BioGRID dataset and 71.5% – 153.3% for DIP dataset, with respect to MLR-MCL. Furthermore, results suggest that balance and scalability raises the average F-score, given an additional 25.6% – 47.1% (BioGRID) and 58.1% – 87.6% (DIP) in the average F-scores over MCL-CAw.

There is also a positive change in the average F-scores computed from the results of the proposed algorithm with respect to those from the results of COACH and PCP. For COACH, there is a 35.6% – 43.8% improvement in average F-score over BioGRID dataset and 56.8% – 60.4% improvement over DIP dataset. Moreover, a 41% – 55.7% increase in average F-scores over BioGRID dataset and a 94.6% – 126.5% increase over DIP dataset was observed with respect to PCP algorithm.

## CONCLUSION

In this paper, a hybrid algorithm for protein complex discovery for protein interaction networks has been presented, which integrates an extension of the Markov clustering algorithm with better scalability and balance, with the application of the core-attachment structure to reflect the inherent organization and modularity of protein complexes. Based on performance analysis, the proposed algorithm yielded an improvement of weighted average F-scores, between 25.6% to 153.3% over two datasets and across four data preprocessing scenarios.

The results indicate an improvement in the average F-scores with both the implementation of balance and scalability and core-attachment structure to Markov clustering process, as compared to other algorithms used in the study and with various data preprocessing scenarios. The results show an increase in F-score of 52% – 96.6% (BioGRID) and 71.5% –153.3% (DIP) over MLR-MCL across four data preprocessing scenarios, as well as an improvement of 25.6% – 47.1% (BioGRID) and 58.1% – 87.6% (DIP) over MCL-CAw.

For future work, it is recommended to apply the algorithm to networks with more complex topology (e.g., human PPIN), as well as compare the reliability of the results with other computational algorithms. It is also interesting to look onto the performance of the algorithm with respect to a dynamic PPI network, for example, the successive releases of BioGRID datasets.

## CONFLICT OF INTEREST

The authors declare that they have no competing interests.

## AUTHORS' CONTRIBUTIONS

JCB and CM contributed equally to this work. ARV and JJV designed and supervised the study. JCB and CM constructed the algorithm and the subsequent analysis. All authors have contributed to and approved the final manuscript.

## REFERENCES

Bader G, Hogue C. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 2003; 4(2).

Becker E, Robisson B, Chapple CE, Gúenoche A, Brun C. Multifunctional proteins revealed by overlapping clustering in protein interaction network. Bioinformatics 2012; 28:84–90.

Chua HN, Ning K, Sung W-K, Leong HW, Wong L. Using indirect protein–protein interactions for protein complex prediction. J Bioinform Comput Biol 2008; 6:435–466.

Chua HN, Sung W-K, Wong L. Exploiting indirect neighbours and topological weight to predict protein function from protein–protein interactions. Bioinformatics 2006; 22:1623–1630.

Dezso Z, Oltvai ZN, Barabási A-L. Bioinformatics analysis of experimentally determined protein complexes in the yeast *Saccharomyces cerevisiae*. Genome Res 2003; 13:2450–2454.

Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002; 30:1575–1584.

Gavin A-C, Aloy P, Grandi P, Krause R, Boesche M, Marzioch M, Rau C, Jensen LJ, Bastuck S, Dumpelfeld B, Edelmann A, Heurtier M-A, Hoffman V, Hoefert C, Klein K, Hudak K, Michon A-M, Schelder M, Schirle M, Remor M, Rudi T, HooperS, Bauer A, Bouwmeester T, Casari G, Drewes G, Neubauer G, Rick JM, Kuster B, Bork P, Russell RB, Superti-Furga G. Proteome survey reveals

modularity of the yeast cell machinery. Nature 2006; 440:631–636.

Liu G, Wong L, Chua HN. Complex discovery from weighted PPI networks. Bioinformatics 2009; 25:1891–1897.

Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. Nat Meth 2012; 9:471–472.

Paccanaro A, Trifonov V, Yu H, Gerstein M. Inferring protein–protein interactions using interaction network topologies, In: Proc. IEEE International Joint Conference on Neural Networks 2005; 1:161-166.

Palla G, Derenyi I, Farkas I, Vicsek T. Uncovering the overlapping community structure of complex networks in nature and society. Nature 2005; 435:814–818.

Pu S, Wong J, Turner B, Cho E, Wodak SJ. Up-to-date catalogues of yeast protein complexes. Nucleic Acids Res 2009; 37:825–831.

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 2004; 32: D449–D451.

Satuluri V, Parthasarathy S. Scalable graph clustering using stochastic flows: applications to community discovery. In: Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2009; 737–746.

Satuluri V, Parthasarathy S, Ucar D. Markov clustering of protein interaction networks with improved balance and scalability, In: Proc. First ACM International Conference on Bioinformatics and Computational Biology 2010; 247–256.

Srihari S, Leong HW. A survey of computational methods for protein complex prediction from protein interaction networks. J Bioinf Comput Biol 2013; 11.

Srihari S, Ning K, Leong HW. MCL-CAw: a refinement of MCL for detecting yeast complexes from weighted PPI networks by incorporating core-attachment structure. BMC Bioinformatics 2010; 11(504).

Srihari S, Yong CH, Patil A, Wong L. Methods for protein complex prediction and their contributions towards understanding the organization, function and dynamics of complexes. ArXiv 2015.

Stark C, Breitkreutz B-J, Chatr-aryamonti A, Boucher L, Oughtred R,Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M. The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 2011; 39: D698–D704.

Van Dongen S, Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht, 2000.

Wu M, Li X, Kwoh C-K, Ng S-K. A core-attachment based method to detect protein complexes in PPI networks. BMC Bioinformatics 2009; 10(169).

Zhang A. Protein Interaction Networks: Computational Analysis, 1st ed. New York: Cambridge University Press, 2009.

## SUPPLEMENTARY INFORMATION

### A. FS-Weighting Algorithm

**Algorithm A1** FS-Weighting Scheme (Chua et al. 2008)

**function** FS-WEIGHTING($G, \tau$)
  **Input:** PPI network $G = (V, E)$, weight threshold parameter $\tau$

  $A \leftarrow \mathbf{0}$
  **for** $(v_i, v_j) \in E$ **do**
    $\lambda_{v_i,v_j} \leftarrow \max[0, n_{avg} - (|N_{v_i} - N_{v_j}| + |N_{v_i} \cap N_{v_j}|)]$ ▷ $n_{avg}$ is the mean neighbor count per protein
    $w_{ij} \leftarrow \dfrac{2|N_{v_i} \cap N_{v_j}|}{|N_{v_i} - N_{v_j}| + 2|N_{v_i} \cap N_{v_j}| + \lambda_{v_i,v_j}} \cdot \dfrac{2|N_{v_i} \cap N_{v_j}|}{|N_{v_j} - N_{v_i}| + 2|N_{v_i} \cap N_{v_j}| + \lambda_{v_j,v_i}}$
    **if** $w_{ij} < \tau$ **then**
      Delete $(v_i, v_j)$ from $E$
    **end if**
  **end for**
  **for** $(v_i, v_j) \in E$ **do**
    $A_{ij} \leftarrow w_{ij}$
  **end for**
  **return** $A$
**end function**

**B. Core-Attachment Sturcture Algorithm**

**Algorithm B1** Core-Attachment Scheme (Srihari et al. 2010)

**function** CoreAttachment($C, \alpha, \gamma$)
  **Input:** preliminary clusters $C = \{C_i = (V_i, E_i)\},\ i = 1 \ldots k$; weight parameters $\alpha, \gamma$

  //**Phase 1:** Finding the set of preliminary cores
  **for** $i = 1 \rightarrow k$ **do**
    **for** $p \in V_i$ **do**
      $d_{in}(p, C_i) \leftarrow \sum w(p, q) : q \in V_i$
      $d_{out}(p, C_i) \leftarrow \sum w(p, q) : q \notin V_i$
      $d_{avg}(C_i) \leftarrow |C_i|^{-1} \sum d_{in}(q, C_i) : q \in V_i$
      **if** $[(d_{in}(p, C_i)) \geq d_{avg}(C_i)] \wedge [(d_{in}(p, C_i)) > d_{out}(p, C_i)]$ **then**
        $PCore(C_i) \leftarrow p$
      **end if**
    **end for**
  **end for**

  //**Phase 2:** Finding the set of extended cores
  **for** $i = 1 \rightarrow k$ **do**
    **for** $p \in V_i$ **do**
      $d_{in}(p, PCore(C_i)) \leftarrow \sum w(p, q) : q \in PCore(C_i)$
      $d_{out}(p, PCore(C_i) \leftarrow \sum w(p, r) : r \notin PCore(C_i), r \in C_i$
      $d_{avg}(PCore(C_i)) \leftarrow (|C_i| - |PCore(C_i)|)^{-1} \sum d_{in}(q, C_i) : q \in PCore(C_i)$
      **if** $[(d_{in}(p, PCore(C_i))) \geq d_{avg}(PCore(C_i)] \wedge [(d_{in}(p, PCore(C_i)) > d_{out}(p, PCore(C_i))]$ **then**
        $ECore(C_i) \leftarrow p$
      **end if**
    **end for**
  **end for**

  // Combine sets of preliminary and extended clusters to form the final set of core proteins
  $Core(C_i) \leftarrow PCore(Ci) \cup ECore(C_i)$

  //**Phase 3:** Finding the set of attachment proteins
  **for** $i = 1 \rightarrow k$ **do**
    $I(p, Core(C_i)) \leftarrow \sum w(p, q) : q \in Core(C_i)$
    $I(Core(C_i)) \leftarrow \frac{1}{2} \sum w(q, r) : q, r \in Core(C_i)$
    $S_C \leftarrow |Core(C_i)|$ ▷ normalized to yield 1 for core sets of size 2
    **if** $I(p, Core(C_i) \geq \alpha I(Core(C_i)) \left(\frac{1}{2} S_C\right)^{-\gamma}$ **then**
      $Attachment(C_i) \leftarrow p$
    **end if**
  **end for**

  // Combine sets of core and attachment proteins to form the set of complexes
  **for** $i = 1 \rightarrow k$ **do**
    $Complex(C_i) \leftarrow Core(C_i) \cup Attachment(C_i)$
  **end for**

  **return** $\{Complex(C_i)\}, i = 1, \ldots k$
**end function**