applied sciences

MDPI

Article

A Special Structural Based Weighted Network Approach for the Analysis of Protein Complexes

Peter Juma Ochieng [1,2]*, József Dombi [1], Tibor Kalmár [3] and Miklós Krész [4,5,6]*

[1] Institute of Informatics, University of Szeged, 2 Árpád tér, H-6720 Szeged, Hungary; dombi@inf.u-szeged.hu
[2] Bánki Donát Faculty of Mechanical and Safety Engineering, Óbuda University, Népszínház Street 8, H-1081 Budapest, Hungary
[3] Department of Pediatrics and Pediatric Health Center, Albert Szent-Györgyi Health Centre, University of Szeged, H-6725 Szeged, Hungary; kalmar.tibor@med.u-szeged.hu
[4] InnoRenew CoE, Livade 6a, 6310 Izola, Slovenia
[5] Andrej Marušič Institute, University of Primorska, Muzejski trg 2, 6000 Koper, Slovenia
[6] Department of Applied Informatics, University of Szeged, Boldogasszony sgt. 6, H-6725 Szeged, Hungary
* Correspondence: juma@inf.u-szeged.hu (P.J.O.); miklos.kresz@innorenew.eu (M.K.)

Abstract: The detection and analysis of protein complexes is essential for understanding the functional mechanism and cellular integrity. Recently, several techniques for detecting and analysing protein complexes from Protein–Protein Interaction (PPI) dataset have been developed. Most of those techniques are inefficient in terms of detecting, overlapping complexes, exclusion of attachment protein in complex core, inability to detect inherent structures of underlying complexes, have high false-positive rates and an enrichment analysis. To address these limitations, we introduce a special structural-based weighted network approach for the analysis of protein complexes based on a Weighted Edge, Core-Attachment and Local Modularity structures (WECALM). Experimental results indicate that WECALM performs relatively better than existing algorithms in terms of accuracy, computational time, and *p*-value. A functional enrichment analysis also shows that WECALM is able to identify a large number of biologically significant protein complexes. Overall, WECALM outperforms other approaches by striking a better balance of accuracy and efficiency in the detection of protein complexes.

Keywords: protein complexes; core-attachment; local modularity structure; weighted PPI network

Citation: Ochieng, P.J.; Dombi, J.; Kalmár, T.; Krész, M. A Special Structural Based Weighted Network Approach for the Analysis of Protein Complexes. *Appl. Sci.* **2023**, *13*, 6388. https://doi.org/10.3390/app13116388

Academic Editors: Larissa A. Balabanova and Yuri N. Shkryl

Received: 20 April 2023
Revised: 10 May 2023
Accepted: 18 May 2023
Published: 23 May 2023

1. Introduction

The detection of protein complexes in Protein–Protein Interaction (PPI) networks is an essential task in system biology for deciphering the cellular organization and functional mechanism. Protein complexes perform the majority of a cell's functional actions [1–3]. As a result, detecting protein complexes is a critical research topic in systems biology. Understanding biological processes is also important in a variety of cytoplasmic systems and helps in the diagnosis of complex diseases [4–6].

Though there are numerous laboratory techniques for detecting protein complexes, most of them tend to be expensive and time-consuming. This has led to the use of computational methods as an efficient approach to detect protein complexes [7]. Computational methods for protein complex detection are generally classified into two broad classes depending on the information required during the complex detection procedure [8]. The first class is known as a topology-based approach, which just uses PPI network topological information to detect protein complexes. The second class uses both topological and biological data to detect protein complexes such as DPC [9], GMFTP [10], and IPC-BSS [11]. Recently, a number of topology-based approaches have been developed to detect protein complexes. For example, there is k-cliques or cliques-based method such as CMC [12] and CFinder [13]; Sub-network density-based methods such as MCL [14–16], DPClus [17,18],

and SPICi [19]; modularity-based method such as CALM [20] and ClusterONE [21]; core-attachment structure-based methods such as COACH [22]and Core [23] and rank and spoke-based methods such as ProRank+ [24].

Nevertheless, these topological-based methods do not identify the state and structure of protein complexes in a PPI network. For instance, CFinder [13], detects protein complexes based on the clique percolation method (CPM) [25], an approach which is computationally expensive when handling large-scale PPI networks due to the NP-complete problem that requires protein complex to be k-clique [26,27]. Related studies have also applied a sub-network density-based approach such as Markov Clustering (MCL) [15,16], and tend to detect protein complexes based on the interaction of proteins within a sub-network (protein complex) in a random walks fashion [7,8,28,29]. Moreover, a heuristic network clustering technique such as SPICi [19], has shown to be efficient for detecting protein complexes based on the local density and support measure. However, this technique is often unreliable when it comes to the detection of protein complexes with overlapping structures especially with high functional similarity. This has led to the development of DPClus [17] as an efficient method for detecting overlapping protein complexes such as these. However, methods such as ClusterONE that utilize MMR for overlapping complex detection tend to miss some attachment proteins, which could result in false positives for protein complex detection [18,20]. Filtering methods such as ProRank+ [24]and PEWCC [30] have been adopted to increase the reliability of PPI networks. Recently modularity-based clustering techniques such as PCR-FR [31], CALM [20], ClusterONE [21] and EPOF [32] have been proposed for detecting protein complexes in densely and sparsely connected network structures [13,33–38]. Generally, the core of a protein complex is frequently a dense sub-network with attachment proteins that are closely linked to the complex's core proteins which help these proteins perform auxiliary functions [22]. Protein complexes have an inherent organization and a common architecture [39,40]. Several techniques for identifying protein complex cores based on core-attachment structure have been investigated to this point, including COACH [22], and Core [23].

Another popular technique that has been used in the detection of protein complex cores is the co-attachment method which is often based on the network core-network structure [41–43]. Generally, this technique has two steps: namely, the identification of the complex core as a dense sub-network or maximal clique and then the characterization of the core of the protein complex. Although these two steps have been widely adopted in the detection of the protein complexes they tend to be inefficient when attempting to characterize the protein complex core of a dense sub-network [44]. Moreover, the majority of the core-attachment-based methods are based on the selection of proteins whose neighbors interact with more than half of the protein in the complex core in the sparse PPI networks [22]. However, this may result in high false-positive interactions and lead to the inaccurate detection of protein complexes [45–47]. The core-attachment structure is still being investigated; no studies have provided a clear distinction between overlapping proteins, core proteins, and peripheral proteins in terms of the weighted network structure [41]. The majority of studies simply focus on a few structural concepts of these protein complexes [20,45–49].

Recently, method such as CALM has shown to be more efficient in the detection of overlapping protein complexes on large-scale PPI networks. However, this method only focuses on the detection of overlapping protein complexes and tends to ignore local attachment proteins to the complex core, as well as it does not consider the common neighborhood and high-order common neighborhood similarity measures when calculating the initial weight of the PPN. All those factors influence the reliability of the PPN and detection of the protein complexes which may result in false positive prediction. To address these limitations, we propose a special structural-based weighted network approach called the Weighted Edge algorithm, Core-Attachment, and Local Modularity (WECALM) for protein complex analysis. By our WECALM approach, our contributions are: First, we introduce a high-order similarity measure based on the Jaccard measure to compute the edge weights, which ensures the

reliability of the PPI network. Second, we extend protein complex identification by using a weighted connectivity algorithm to discriminate and detect local attachment proteins to complex cores. Third, we extend the detection of protein complexes using the structural similarity measure concept. Fourth, we perform functional enrichment analysis by calculating the *p*-value of the detected complexes to validate their associated functions.

This paper is organized as follows. In Section 2, we provide a preliminary overview of our approach. In Section 3, we give a detailed computational description of our approach to the detection of protein complexes. In Section 4, we describe our experimental PPI datasets and evaluation criteria of our proposed approach. In Section 5, we present the experimental results and discuss them. Lastly, in Section 6, we draw some conclusions and outline our future research plans.

2. Preliminaries

In this section, we introduce some fundamental concepts. Generally, the PPI network can be represented as an undirected unweighted, or weighted graph denoted by $G = (V, E)$ where V are set of nodes denoting the proteins and E are set of edges corresponding to the interaction between pair of proteins. In our approach, we consider the PPI network to be an undirected edge-weighted graph given by $G = (V, E, W)$, where W denotes the weight on the edge representing the confidence score in the range $(0, 1]$ and function $W : E \longrightarrow \Re^+$ quantifies the affinity of the interaction between each pair of nodes or proteins (*i.e.*, edge mapping in E). For node v, $N(v)$ denotes the set of all neighboring nodes of v. The nodes (proteins) of the PPI graph model can be classified into four major classes with respect to protein complexes (groups of two or more proteins that are physically linked together through non-covalent interactions) according to [21,50,51] (see also Figures 1 and 2). The first class is *core nodes*: a node is considered to be a core node in the complex if: it shows a high degree of physical interaction; it has a relatively high weighted degree of direct physical interaction among themselves within the complex and less interaction with nodes outside the complex; the set of core nodes unique in each complex. The second classification is *peripheral node*, a node is considered to be a peripheral node to a complex if: it has a close interaction with the complex core; it is stable and directly interacts with the complex core. The third classification is *overlapping nodes*; a protein is considered to be an overlapping protein to a complex if: it has a higher degree and acts as a betweenness node than the neighborhood nodes; it interacts closely to the complex core; it belongs to more than one complex. The remaining proteins are classified as *interspersed nodes*, which is probably just noise in the PPI network.

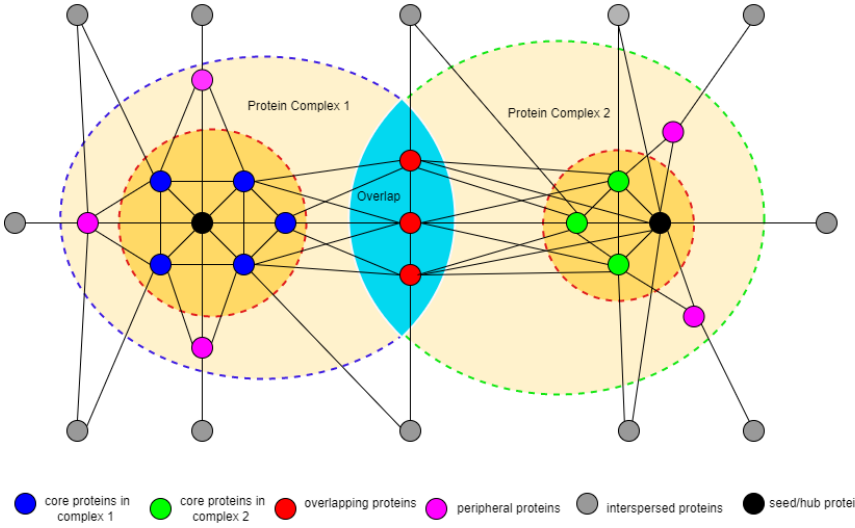

Figure 1. A general structure of a PPI network comprised of two complexes, overlapping proteins, and peripheral and interspersed proteins.

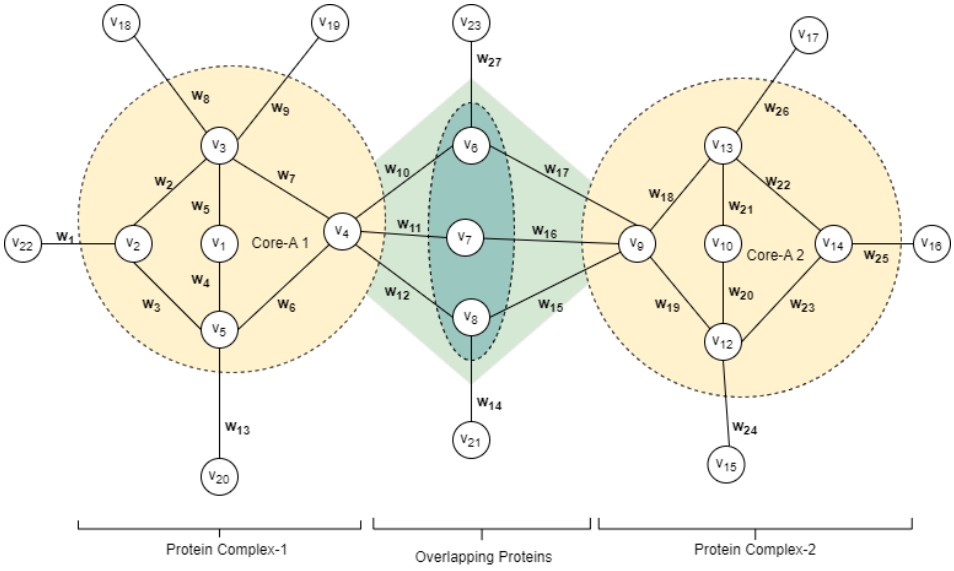

Figure 2. A simple graph representation of the PPI network structure is shown in Figure 1. From the network, v represents the nodes or proteins and w represents the weight of edges or the confidence score.

3. Methods

In this section, we will describe seven main steps of our proposed WECALM algorithm these include: building a weighted PPI network; identifying overlapping structures; identifying seed proteins; identifying local modularity structures; identifying complex core structures; detecting attachment proteins to the complex core and protein core attachment and protein complex formation.

3.1. Building Weighted PPI Network

In general, PPI networks obtained through various experimental techniques are typically noisy and many interactions are presumed to be false positives [52,53]. As a result, we should reduce the rate of false positives. To address this challenge topological properties of PPI networks have been proposed to develop preprocessing strategies for evaluating and eliminating potential false positives [54–58]. According to some experimental findings [59–61], neighbor information-based methods are used to evaluate PPI with high confidence scores and are typically more reliable than other methods. Thus, in this study to build a reliable weighted PPI network, we shall use Jaccard's coefficient similarity (J_s) [62] to compute the proteins interaction scores. Hence, the similarity between two neighboring proteins v and u is defined by

$$J_s(v,u) = \frac{|N(v) \cap N(u)|}{|N(v) \cup (u)|} = \frac{|CN(v,u)|}{|N(v) \cup N(u)|}, \tag{1}$$

where $0 \leq J_s(v,u) \leq 1$, I is the interaction between proteins v and u, $CN(v,u)$ represents the set of common neighbors proteins v and u. $|N(v) \cap N(u)|$ represents number of common neighbors of proteins v and u. $|N(v) \cup N(u)|$ represents the union set of all the different neighboring proteins of v and u. Thus, with Equation (1), we can calculate weight between two neighbouring proteins v and u by

$$w(v,u) = \begin{cases} 1 & if\ |CN(u,v)| \leq 1 \\ 0 & otherwise \end{cases} \tag{2}$$

Based on our computation the similarity of two adjacent proteins will be higher if the two proteins share more common neighbors. On this basis, we propose a high-order

similarity metric based on Jaccard's coefficient between proteins v and u to calculate the connectivity between the adjacent proteins v and u in the common neighbor. Now we will define the common neighbors' support using the formula

$$\rho(v,u) = J_s(v,u) \sum_{u \in CN(v,u)} w(v,u), \tag{3}$$

where ρ is the common neighbor support of the weighted edge(v,u) and w is the weight of the edge between protein v and u stated in the preliminary in Section 2. Thus, with Equations (2) and (3), we can define high-order similarity score by the formula

$$\phi(v,u) = \frac{J_s(v,u) + \rho(v,u)}{1 + \rho(v,u)}, \tag{4}$$

where $\phi(v,u)$ is the high-order similarity score for the common neighbor of two adjacent proteins and it takes the values in the range $[0,1)$. For the rest of the paper, ϕ defines the edge weights W.

3.2. Identifying Overlapping Structures

To identify the overlapping structure, let $v \in V, N(v) = \{u | u \in V, (v,u) \in E\}$ be set of neighbour protein v and $deg(v) = |N(v)|$ be the number of neighbours of protein v. Given protein $v \in V$ we can define the neighborhood network $\mathcal{GN}_\mathcal{V} = (V_v, E_v)$ as sub-network of protein v and its direct neighbours interacting in network $\mathcal{G}$. Hence $V_v = \{v\} \cup \{u | u \in V, (v,u) \in E\}$ and $E_v = \{(u_i,u_j) | (u_i,u_j) \in E, u_i, u_j \in V_v\}$. Thus, the weighted degree average of a local neighborhood sub-network $\mathcal{GN}_\mathcal{V}$ is defined by the equation

$$Avg(deg(\mathcal{GN}_\mathcal{V})) = \frac{\sum_{u \in V_v} deg(u)}{|V_v|}, \tag{5}$$

To calculate the global importance of a protein v, we calculate the shortest paths between all protein pairs that pass through the target proteins by defining the betweenness of node v by

$$B(v) = \sum_{\substack{s \neq v, t \neq v \\ s,t,v \in V}} \frac{\delta_{s,t}(v)}{\delta_{s,t}}, \tag{6}$$

where $\delta_{s,t}$ is the number of shortest paths between protein s to t and $\delta_{s,t}(v)$ is the number of shortest paths between protein s to t that pass through the intermediate (bridge) protein v. Thus, using Equation (6) the average betweenness of its local neighborhood sub-network $\mathcal{GN}_v$ is calculated by

$$Avg(B(\mathcal{GN}_v) = \frac{\sum_{u \in V} B(u)}{|V_v|}, \tag{7}$$

where $AvgB(\mathcal{GN}_\mathcal{V})$ is the average of $B(u)$ for all $u \in V_v$ in local neighborhood sub-network $\mathcal{GN}_v$ and $|V_v|$ denotes the total number of nodes in the *PPI network*. Then, using Equations (5) and (7) we defined overlapping protein structure in the PPI network by

$$\mathcal{OVP}(\mathcal{GN}_\mathcal{V}) = \begin{cases} 1 & if\ deg(v) \geq Avg(deg(\mathcal{GN}_\mathcal{V})) \wedge B(v) > Avg(B(\mathcal{GN}_\mathcal{V}))), \\ 0 & otherwise \end{cases} \tag{8}$$

where $\mathcal{GN}_v$ is candidate overlapping protein complex with $\mathcal{OVP}(\mathcal{GN}_\mathcal{V}) = 1$. To measure the degree of overlap between sets of candidate overlapping protein complexes we calculate the overlapping score between the two sets by

$$OS(\mathcal{OVP}_i, \mathcal{OVP}_j) = \frac{|\mathcal{OVP}_i \cap \mathcal{OVP}_j|}{|\mathcal{OVP}_i| + |\mathcal{OVP}_j| - |\mathcal{OVP}_i \cap \mathcal{OVP}_j|}, \tag{9}$$

where $OS(\mathcal{OVP}_i, \mathcal{OVP}_j)$ is overlapping score between $\mathcal{OVP}_i$ and $\mathcal{OVP}_j$ ranging from $[0, 1)$ in which 0 indicates no overlap between the sets and 1 indicates that the sets are identical; $|\mathcal{OVP}_i|$ and $|\mathcal{OVP}_j|$ denote the sizes of sets $\mathcal{OVP}_i$ and $\mathcal{OVP}_j$, respectively. $|\mathcal{OVP}_i \cap \mathcal{OVP}_j|$ denotes the intersection of sets $\mathcal{OVP}_i$ and $\mathcal{OVP}_j$. In this paper, we identify the candidate overlapping protein complex when $OS(\mathcal{OVP}_i, \mathcal{OVP}_j) \geq \pi$, where π is predefined overlap threshold ranging in $(0, 1]$.

3.3. Identifying Seed Proteins

The identification of seed protein for the PPI network is essential for the detection of protein complexes. Here, we introduce the concept of weighted degree and cluster coefficient as a strategy for identifying the seed protein. For this, we defined the weighted node degree by

$$deg_w(v) = \sum_{u \in N(v); (v,u) \in E} w(v, u), \tag{10}$$

where $deg_w(v)$ is the weighted degree of the protein v and w is the edge weight stated in the preliminary in Section 2. To determine the seed protein we consider the small world phenomenon model [63,64] which correspond to local weighted clustering coefficient λ. Then, we define λ_v of protein v as measure of its local connectivity among its immediate neighbors and $\lambda_w(v)$ of protein v as weighted sub-network $\mathcal{GN}_\mathcal{V}$ formed by N_v and their corresponding weighted edges. Thus, we can calculate clustering coefficient of protein v by

$$\lambda_w(v) = \frac{\sum_{u_i \in V_v} \sum_{u_j \in N(u_i) \cap V_v} w(u_i, u_j)}{|N_v| \times (|N_v| - 1)}, \tag{11}$$

where $\lambda_w(v)$ is the clustering coefficient of protein v and $\lambda_w(v) \in (0, 1]$. Using Equation (11), we can calculate the average clustering coefficient of sub-network $\mathcal{GN}_\mathcal{V}$ by

$$Avg(\lambda_w(v)) = \frac{\sum_{u \in V_v} \lambda_w(v)}{|V_v|}, \tag{12}$$

where $\lambda_w(v)$ is the average local weighted clustering coefficient of the protein v, V_v is the number of the protein v and all its local neighbours in a sub-network. With Equations (7) and (12), the seed protein ($\mathcal{S}$) is defined as

$$\mathcal{S}(v) = \begin{cases} 1 & if\ \lambda_w(v) \geq Avg(\lambda_w(v)) \wedge B(v) \leq Avg(B(\mathcal{GN}_\mathcal{V})), \\ 0 & otherwise \end{cases}. \tag{13}$$

where node v is a selected seed protein if $\mathcal{S}(v) = 1$.

3.4. Identifying Local Modularity Structures

To identify local modularity structures, we consider seed proteins calculated by Equation (13) as initial nodes to generate clusters by first computing the support function, followed by the local modularity function. Hence, using Equation (4) first, we calculate the similarity score between a seed protein v and its immediate proteins gradually adding the neighboring proteins with the help of the support function and the local modularity function in order to generate cluster K as sub-network. To prioritize each neighboring

protein u, first, we calculate the support function to measure how close the protein u is to the cluster K using the formula

$$supp(u, K) = \frac{\sum_{u' \in K \cap N(u)} w(u, u')}{\sum_{u' \in N(u)} w(u, u')}, \tag{14}$$

where $supp(u, K)$ is the support function in the range $[0, 1)$, $u \notin K$, and $\sum_{u' \in K \cap N(u)} w(u, u')$ is the summation of the edge weight linking protein u to K, and $\sum_{u' \in N(u)} w(u, u')$ is the total degree weight of protein u. The above-prioritizing approach can be extended iteratively for the neighbors of any initial cluster K. Thus, in each iteration step, according to the priority of the neighbors, the decision to join the cluster is made by the local modularity function.

Given subnetwork K of $\mathcal{G}$, we can define weights in-degree as the sum of the weight of edges linking protein u to other proteins in K denoted by $w_{in}(K)$ and weighted out-degree as the sum of the weight of edges linking protein v to proteins in the rest of $\mathcal{G} - K$ denoted by $w_{out}(K)$. Thus, we can define $w_{in}(K)$ and $w_{out}(K)$ by

$$w_{in}(K) = \sum_{\substack{u,u' \in K \\ w(u,u') \in W}} w(u, u'), \tag{15}$$

and

$$w_{out}(K) = \sum_{\substack{u' \in K, u \notin K \\ w(u,u') \in W}} w(u, u'), \tag{16}$$

where w represents the weight of the edges in sub-network K. To determine the local modularity structure in sub-network K, we defined modular uncertainty correction threshold value η in the interval of $[0, 1)$. Using Equations (15) and (16), we can define the local modularity of sub-network K by

$$\mathcal{Q}(\eta, K) = \frac{w_{in}(K)}{\left(w_{in}(K) + w_{out}(K) + \eta \cdot |V_K|\right)^{\alpha}}, \tag{17}$$

where $\mathcal{Q}(\eta, K)$ takes a value $(0, 1)$; $|V_K|$ is total number of proteins in K, η is predefined modular uncertainty correction parameter in the range of $(0, 1]$, α is the ratio of the internal interaction to the total interaction in the community. We set $\alpha = 1.0$ in order to detect high $w_{in}(K)$ and a low $w_{out}(K)$ which makes it efficient in the detection of local modularity structure. A neighboring node is added to K, if extending K by the given node, the value of the local modularity function would increase.

3.5. Identifying Complex Core Structure

To detect the complex core, let $v \in V, N(v)$ be the set of all immediate neighbor proteins, and the structural neighborhood of protein v is given by $N_s(v) = \{v\} \cup N(v)$, in which $N_s(v)$ entails protein v and its direct neighbors. Now, we can calculate the structural similarity between two neighboring proteins v and w by

$$SS(v, w) = \frac{|N_s(v)| \cap |N_s(w)|}{\sqrt{|N_s(v)|\,|N_s(w)|}}, \tag{18}$$

where $SS(v, w)$ structural similarity is in the range of $(0, 1]$. Here, high $SS(v, w)$ between two proteins indicates that the two proteins shared a similar neighborhood structure. Moreover, the structural similarity is symmetric as $SS(v, w) = S(w, v)$. Based on $SS(v, w)$ we mine a sub-network in the neighborhood network $\mathcal{GN}_{\mathcal{V}}$, which we refer to preliminary complex core. We introduce ω as the default threshold value to compute the optimal structural similarity score between seed protein, v, and each neighbor $w \in N(v)$ from the identified preliminary protein complex $C_p(v)$. Hence, using Equation (18) given the

preliminary complex $C_p(v)$, and structural similarly threshold ω, we can calculate the preliminary complex core of protein v by

$$Core(\omega, C_p(v)) = \{w \in C_p(v) : SS(v, w) \geq \omega\} \tag{19}$$

where $Core(\omega, C_p(v))$ is the preliminary complex core; ω is a default threshold value in ranging from $(0, 1]$; $C_p(v)$ denotes the preliminary complex of protein v. Note that protein v is included in the $Core(\omega, C_P(v))$.

3.6. Detection of Attachment Proteins to Complex Core

Generally, attachment proteins exist in two forms, namely overlapping and peripheral protein attachments [65]. Therefore, to identify protein attachment to the complex core, consider the identified preliminary protein complex denoted by $C_p(v)$, the preliminary complex core as a sub-network represented by $Core(\omega, C_p(v)) = (V_c, E_c)$ and the set $CAP(C_p(v))$ of candidate attachment proteins as a subset of the neighbors of $Core(\omega, C_p(v))$. Here, our two main objectives are: first, to find a subset $CAP(C_p(v)) \subseteq V$ in PPI network in which each protein $p \in CAP(C_p(v))$ is a candidate attachment protein with identified preliminary protein complex $C_P(v)$, and secondly, to predict the category of each protein in $CAP(C_p(v))$.

To achieve the two objectives we set two basic conditions, namely: (1) The attached proteins must interact with the complex cores directly; (2) The attached proteins must be connected to at least two core proteins via complex cores since protein complexes are made of two or more complexes [66]. Therefore, if protein p fulfils the conditions as belonging to the neighborhood of $Core(\omega, C_P(v))$ with $|N(p) \bigcap V_c| \geq 2$, then it is selected for $CAP(C_p(v))$. Below we provide a detailed description of the calculation of the overlapping and peripheral protein attachment to the complex core.

3.6.1. Overlapping Attachment Proteins

To identify overlapping protein attachment let $CAP(C_P(v))$ attached from preliminary complex protein $C_P(v)$ and $\mathcal{OVP}(C_P(v))$ be the set of candidate overlapping proteins attached to the preliminary complex protein $C_P(v)$. We can define weighted candidate protein for a candidate overlapping attachment protein $p \in \mathcal{OVP}(C_P(v))$ interacting with proteins in complex core $Core(\omega, C_P(v))$ by

$$d_w(p, Core(\omega, C_P(v)) = \sum_{t \in V_c} w(p, t), \tag{20}$$

Next, we calculate the average weight of interaction for all candidate core protein p within complex core $Core(\omega, C_P(v))$ by the formula

$$Avg(d_w(OVP(C_P(v)))) = \frac{\sum\limits_{p \in OVP(C_P(v)} d_w(p, Core(\omega, C_P(v))}{|OVP(C_P(v)|} \tag{21}$$

Using Equations (20) and (21), we defined the score of the candidate overlapping protein attachment to the complex core $C_P(v)$ by

$$\mathcal{OVP}(p, Core(C_P(v)) = \begin{cases} 1 & if\ d_w(p, Core(\omega, C_P(v)) \geq Avg(d_w(OVP(C_P(v))))), \\ 0 & otherwise \end{cases} \tag{22}$$

Then, the set $\mathcal{OVP}(Core(C_P(v))$ denotes the set of local overlapping attachment proteins p for which $\mathcal{OVP}(p, Core(C_P(v)) = 1$.

3.6.2. Peripheral Attachment Protein

Here, we consider the set of candidate peripheral proteins $PP(C_P(v))$ obtained by the difference of $CAP(C_P(v)) - \mathcal{OVP}(Core(CP(v))$. Given the weight of the connectiv-

ity of proteins $p \in PP(C_P(v))$ with respect to the complex core as $d_w(p, Core(\omega, C_P(v))$, we define the average weight of interactions of all candidate peripheral proteins with $Core(\omega, C_P(v))$ by

$$Avg(d_w(PP(C_P(v))) = \frac{\sum\limits_{p \in PP(C_P(v)} d_w(p, Core(\omega, C_P(v))}{|PP(C_P(v)|} \tag{23}$$

Hence, using Equation (23), we define the score of peripheral attachment protein by

$$\mathcal{PP}(p, Core(C_P(v)) = \begin{cases} 1 & if\ d_w(p, Core(\omega, C_P(v)) \geq Avg(d_w(PP(C_P(v)))), \\ 0 & otherwise \end{cases} \tag{24}$$

Then the set $\mathcal{PP}(Core(C_P(v))$ denotes the set of local peripheral attachment proteins p for which $\mathcal{PP}(p, Core(C_P(v)) = 1$.

3.7. Protein Core Attachment and Protein Complex Formation

To detect protein complex formation, we first compute the core-attachment proteins by aggregating the overlapped and peripheral protein scores to generate the overall set of attachment proteins in the complex core defined by the formula

$$\mathcal{A}(Core(C_P(v)) = \mathcal{OVP}(Core(C_P(v)) \cup \mathcal{PP}(Core(C_P(v)), \tag{25}$$

where $\mathcal{A}(Core(C_P(v))$ is the overall local attachment proteins to the complex core $Core(C_P(v))$. Next, the protein complex formation is computed by merging sets of preliminary complex cores (see Equation (19)) and the set of detected candidate attachment proteins (see Equation (25)). Hence, using Equations (22) and (25) we defined the score of final protein complex formation by

$$CP(v) = \begin{cases} 1 & if\ |\mathcal{OVP}(Core(C_P(v))| \geq 2 \wedge |Core(\omega, C_P(v))| > 3\,|\mathcal{A}(Core(C_P(v))|\,, \\ 0 & otherwise \end{cases} \tag{26}$$

Therefore, we define the set of distinct protein complexes using the formula

$$\mathcal{CP}(v) = Core(\omega, C_P(v)) \cup \mathcal{A}(Core(C_P(v))), \tag{27}$$

where the protein complexes above are defined only if $CP(v) = 1$.

4. Datasets and Evaluation Criteria

In this section, we will provide a general description of the experimental PPI datasets and evaluation criteria used to validate and compare the performance of our WECALM approach.

4.1. Experimental PPI Datasets

In our study, the three freely accessible PPI networks extracted from *S.cerevisiae* were used for simulation. They were the DIP [67] database that documents experimentally determined Protein–Protein interactions, BioGRID [68] database of physical and genetic interactions, and the Yeast database [17,69]. A brief description of the dataset used for the simulation is given in Table 1. The data from Human [69] was used to build Human PPI networks.

Table 1. The general details of PPI networks used for the simulation.

Datasets	Number of Protein	Number of Edges	Network Density
BioGRID	5640	59,748	3.16×10^{-6}
DIP	4930	17,202	1.42×10^{-3}
Human	15,459	144,687	1.21×10^{-3}
Yeast	6194	74,826	3.90×10^{-3}

For complex simulation data, we used the yeast reference datasets CYC2008 [70] and NewMIPS [71,71] for complex simulation studies. For human complexes, we used data from the CORUM [69] , PINdb [72], and KEGG modules [73] databases. In addition, for functional enrichment analysis, we utilised Aloy [50] and SGD [74] for Gene Ontology. Table 2 lists the details of the benchmark protein complexes employed in this study.

Table 2. The details of the benchmark protein complexes.

Complex Datasets	Number of Protein Complexes	Overlapping Complexes	Non-Overlapping Complexes	Protein Coverage	Average Size
NewMIPS	328	283	45	1171	14.93
CYC2008	236	108	128	1628	4.71
Human complexes	2289	-	-	6206	8.57
Yeast complexes	1045	-	-	2773	8.92

4.2. Evaluation Criteria

We compared the identified protein complexes with the reference complexes to determine how well the algorithms identify protein complexes. To make comprehensive and detailed comparisons, we utilized a wide range of evaluation metrics such as recall, precision, F-measure, coverage rate, and others, as suggested by related studies [20,22,23,75]. In the subsection below, we provide a detailed description of these metrics.

4.2.1. Computation of Recall, Precision and F-Measure

To calculate evaluation metrics, we must first compute the similarity between detected and reference complexes based on neighborhood affinity in order to measure their closeness [53,76,77]. Hence, let $P = \{p_1, p_2, \ldots, p_k\}$ be detected protein complexes $C_P(v)$ and $R = \{r_1, r_2, \ldots, r_l\}$ be the reference protein complexes. Here, we denote the detected and reference proteins complexes by p_i and r_j respectively. Thus, neighborhood affinity between the detected and reference protein complexes is calculated like so:

$$NA(p_i, r_j) = \frac{|N(p_i) \cap N(r_j)|^2}{|N(p_i)|\,|N(r_i)|}, \tag{28}$$

where $NA(p_i, r_j)$ is the neighborhood affinity in the range of $[0, 1)$, $|N(p_i)|$ represents the size of detected complex, $|N(r_j)|$ represents the size of the reference complex, and $|N(p_i) \cap N(r_j)|$ denotes the number of common proteins from the detected and reference complexes. Here, the larger the $NA(p_i, r_j)$, the closer the two complexes are. Given a threshold $\kappa, if NA(p_i, r_j) \geq \kappa$, then p_i is similarly matched with r_j so we set $\kappa = 0.2$ according to [22,53,78]. From Equation (28) we can calculate recall, precision and F-measure. Let $\mathcal{N}_\mathcal{P} = |\{p|p \in P, \exists r \in R, NA(p, r) \geq \kappa\}$ and $\mathcal{N}_\mathcal{R} = |\{r|r \in R, \exists p \in P, NA(r, p) \geq \kappa\}$ be the number of the corrected detected and reference complexes that match at least one real protein and detected complex, respectively. Now, we define recall and precision using the formula

$$Recall = \frac{|\{r|r \in R, \exists p \in P, NA(r,p) \geq \kappa\}|}{|R|} = \frac{\mathcal{N}_{\mathcal{R}}}{|R|}, \tag{29}$$

and

$$Precision = \frac{|\{p|p \in P, \exists r \in R, NA(p,r) \geq \kappa\}|}{|P|} = \frac{\mathcal{N}_{\mathcal{P}}}{|P|}, \tag{30}$$

In general, a smaller protein complex has a higher precision, and a larger protein complex has a higher recall hence the two metrics often have an inverse relationship. Since the F-measure is the harmonic mean of recall and precision using Equations (29) and (30), we can define the F1-measure by

$$F1 - measure = \frac{2 \times Precision \times Recall}{|Precision + Recall|}, \tag{31}$$

4.2.2. Coverage Rate

To evaluate the performance of our proposed WECALM algorithms and peer methods it is necessary to determine the number of potentially covered proteins in the reference complexes by a computation of the coverage rate (CR) [75,77,79]. To calculate the coverage rate, let P and R be the sets of detected and reference protein complexes, respectively. Hence we can represent the matrix of the detected complexes and the reference complexes $|R| \times |P|$ by M, where each component of the matrix $max\{M_{ij}\}$ is the maximum number of proteins sharing a similar function relationship between the i^{th} and j^{th} reference complex and detected complex respectively. Now we defined coverage rate by

$$CR = \frac{\sum_{i=1}^{|R|} \max\{M_{ij}\}}{\sum_{i=1}^{|R|} N_i}, \tag{32}$$

where CR is the coverage rate and N_i denotes the number of proteins in the i^{th} reference complex.

4.2.3. Maximum Matching Ratio

The maximum match ratio, or MMR, is a metric based on the maximum one-to-one mapping between the detected and reference complexes. MMR directly penalizes a reference complex that has been split into two or more parts in the detected set because only one of these parts is permitted to match the correct reference complex. MMR offers a natural, simple method for comparing detected complexes to reference complexes [20,80]. We compute the MMR using a weighted edge between the detected and the reference complexes calculated based on the neighborhood affinity score defined in Equation (28). That is, the maximum match ratio is

$$MMR = \frac{\sum_{i=1}^{|R|} \max_{j=1}^{n} NA\{p_i, r_j\}}{\sum_{i=1}^{|R|} N_i}, \tag{33}$$

where $NA\{p_i, r_j\}$ is the neighborhood affinity score; R is the number of the reference complexes, n is the number of detected complexes; j is a member of the detected complexes; N_i is the number of proteins in the i^{th} reference complex; r_j is the j^{th} reference complex and p_i is the i^{th} detected complex.

4.2.4. Separation and ACC

To avoid the case where proteins of a reference complex are matched with several detected protein complexes we used Separation (Sep) to calculate a one-to-one correspondence between detected protein complexes and reference protein complexes [20]. Here, we defined Separation by

$$Sep_{p_i} = \frac{\sum_{i=1}^{|R|}\sum_{j=1}^{m} Sep_{ij}}{|R|},\ Sep_{r_j} = \frac{\sum_{j=1}^{m}\sum_{i=1}^{|R|} Sep_{ij}}{m},\ Sep = \sqrt{Sep_{p_i} \times Sep_{r_j}}, \tag{34}$$

where $Sep_{ij} = \frac{(t_{ij})^2}{\sum_{i=1}^{|R|} t_{ij} * \sum_{j=1}^{m} t_{ij}}$, $|R|$ is the number of protein complexes in the reference complexes, m is the number of proteins in detected complexes,t_{ij} denotes the degree of intersection between the i^{th} reference complex and the j^{th} detected complex, and N_i is the number of proteins within the i^{th} reference complex. To quantify the quality of detected protein complexes, we compute the geometric means of sensitivity and the positive predictive value (PPV) to obtain the Accuracy ACC [20]. To measure ACC, we used the following formula

$$S_n = \frac{\sum_{i=1}^{|R|} \max_{j=1}^{m}\{t_{ij}\}}{\sum_{i=1}^{|R|} N_i},\ PPV = \frac{\sum_{i=1}^{m} \max_{j=1}^{|R|}\{t_{ij}\}}{\sum_{j=1}^{m}\sum_{i=1}^{|R|}\{t_{ij}\}},\ ACC = \sqrt{S_n \times PPV}, \tag{35}$$

4.2.5. Functional Enrichment Analysis

Even though known protein complexes are often insufficient or incomplete in laboratory-based experiments, it is always necessary to annotate the biological function of the detected complexes by computing the p-value and perform Gene Ontology functional enrichment analysis as a confirmatory test of the biological significance of the detected complexes [9,11,81]. To calculate the significance value of the biological function, we define the p-value by

$$p - value = 1 - \sum_{i=0}^{m-1} \frac{\binom{F}{i}\binom{N-F}{C-i}}{\binom{N}{C}}, \tag{36}$$

where m is the number of observed proteins in the functional group of the detected complex, N is the total number of proteins in a PPI network, C is the size of the detected protein and F represents the size of functional group. Note that in our analysis the p-value is calculated based on the biological processes term descriptions (or ontologies) and the smaller value the more the biological significance that protein complex has. Hence, protein complex with a p-value < 0.01 is deemed to be biologically significant in the PPI network.

5. Results and Discussion

In this section, we will present and discuss the findings of WECALM's performance compared with other algorithms, followed by parametric selection, computational complexity analysis, and validation with function enrichment analysis.

5.1. Performance Comparison of WECALM with Other Algorithm

In our study, it was necessary to compare the performance of our proposed WECALM algorithm with other existing protein complex detection algorithms based on the evaluation criteria stated in Section 4.2. Therefore, we compared our WECALM method with ten recently developed complex detection algorithms namely, CFinder, MCL, COACH, EWCA, Core, CALM, ClusterONE, GMFTP, ProRank+, and CMC. To fairly evaluate the ten algorithms we set the optimal parameters of each algorithm based on the author's recommendations to obtain the results [10,20,82].

5.1.1. Performance on NewMIPS Complexes

We compared the robustness of our proposed WECALM method to other existing methods for detecting protein complexes. We considered the evaluation matrices described in Section 4.2. Based on the NewMIPS complex using the BioGRID dataset (see Figure 3A), we found that WECALM performed better in terms of recall (0.7701), F-measure (0.7252), coverage rate (0.6743), and maximum matching ratio (0.3975). In terms of precision score (0.7131), ProRank+ outperformed all other methods. On the NewMIPS complex using the DIP dataset (see Figure 3B), again we observe that WECALM performed better in terms of recall (0.7141), F-measure (0.5889) and maximum matching ratio (0.3531). ProRank+ (0.6657) and CMC (0.5736) performed best in terms of precision and coverage rate respectively. Though, ProRank+ and CMC in terms of precision and coverage rate WECALM performed best in the overall composite score. Table A1 in Appendix A provides supplementary results for performance comparison of WECALM and another algorithm on the NewMIPS dataset using the BioGRID and DIP complexes.

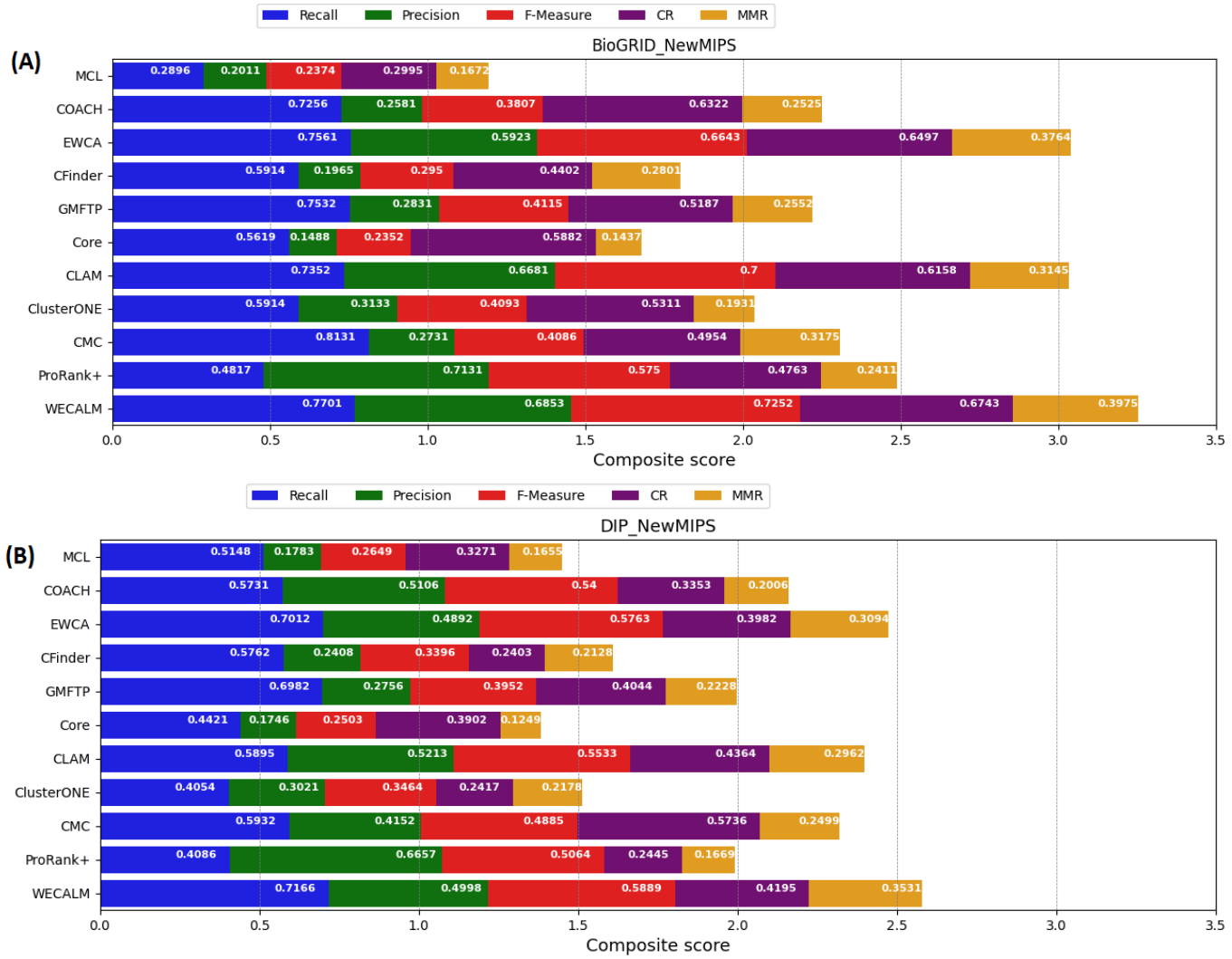

Figure 3. A comparison of the performance of WECALM and other existing algorithms on NewMIPS complexes. (**A**): The BioGRID dataset and (**B**): The DIP dataset. Evaluation matrices include; Recall, Precision, F-measure, coverage rate (CR), and the Maximum Matching Ratio (MMR). The overall composite score is determined by the length of the bar. The longer the bar, the better an algorithm's overall performance is.

5.1.2. Performance on CYC2008 Complexes

We also compared the performance of WECALM with other algorithms based on CYC2008 complexes using both the BioGRID and DIP datasets. Based on the CYC2008 complex using the BioGRID dataset (see Figure 4A), we see that WECALM performed better in terms of recall (0.8291), F-measure (0.6956), CR (0.8831), and MMR (0.4825). In terms of precision score (0.6622), ProRank+ performed better than other methods. On the DIP dataset (see Figure 4B), again we observe that WECALM performed better in terms of recall (0.7315), F-measure (0.6315), and maximum matching ratio (0.3866). ProRank+ (0.6924) and GMFTP (0.6085) performed best in terms of precision and coverage rate respectively. Though, ProRank+ and GMFTP in terms of precision and coverage rate WECALM performed best in the overall composite score. Table A2 in Appendix A provides supplementary results for evaluation of the performance of WECALM and another algorithm on CYC2008 complexes using the BioGRID and DIP datasets.

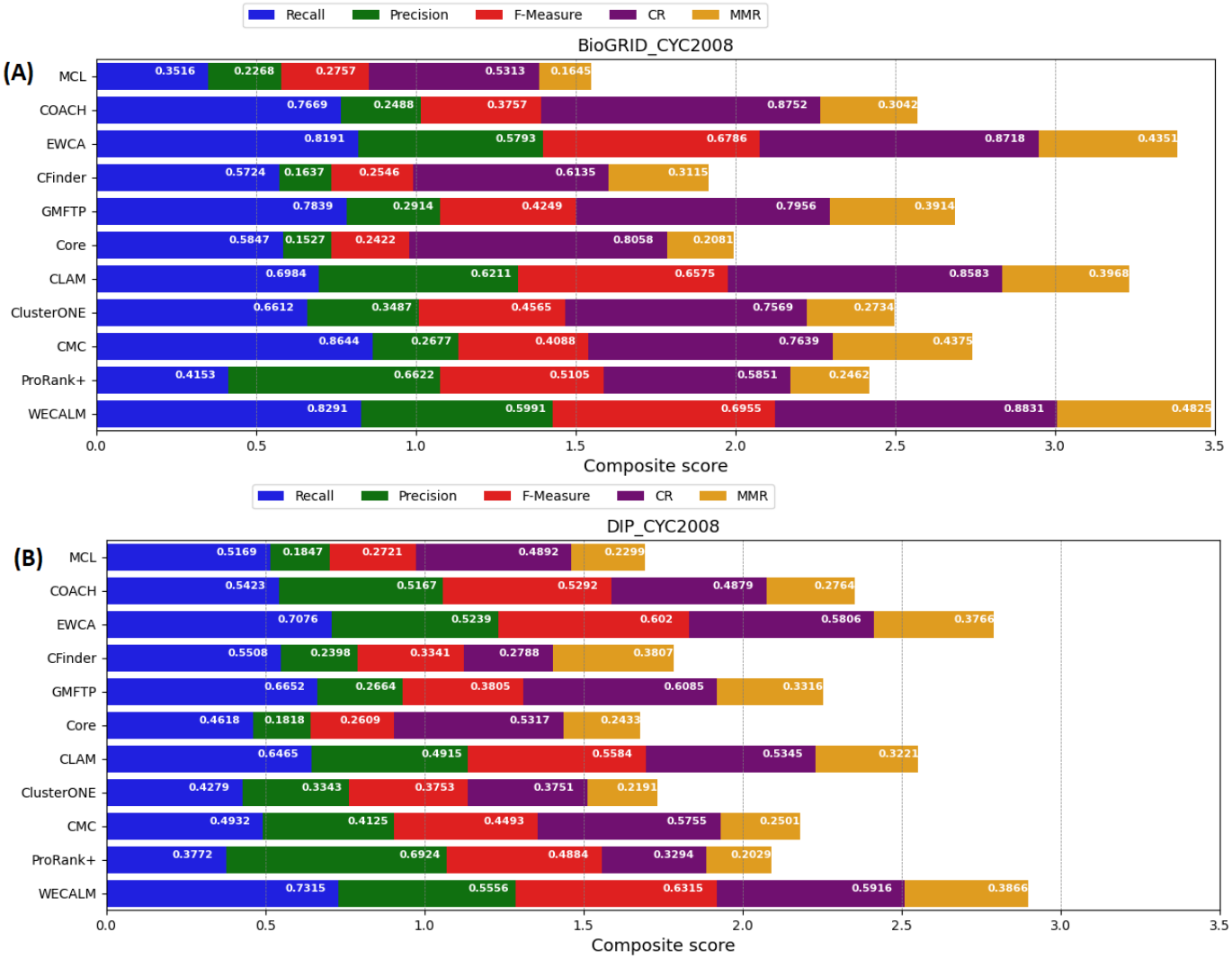

Figure 4. A comparison of the performance of WECALM and other existing algorithms on CYC2008 complexes. (**A**): The BioGRID dataset and (**B**): The DIP dataset. Evaluation matrices include; Recall, Precision, F-measure, coverage rate (CR), and the Maximum Matching Ratio (MMR). The overall composite score is determined by the length of the bar. The longer the bar, the better an algorithm's overall performance is.

WECALM's performance was also evaluated in terms of separation and ACC. According to the results in Tables A1 and A2 in Appendix A, WECALM outperformed all other methods in separation and ACC on both the NewMIPS and CYC2008 complexes both the BioGRID and DIP datasets. A high separation measure indicates that the detected complexes are well separated from one another, indicating good algorithm performance while an ACC score close to 1 indicates perfect performance, meaning that the algorithm detected all the true complexes. A low separation measure indicates that the complexes are overlapping or clumped together, which might mean false positive complexes or inaccurate detection of true complexes while an ACC score of less than 1 indicates that some of the detected complexes were false positives.

5.2. Parametric Selection

Here, we shall evaluate the effects of adjusting the threshold value of π, η, and ω on the overlapping score, local modularity score, and core structural similarity score, respectively on the performance of the WECALM.

5.2.1. Effect of Varying π on the Performance of WECALM

The overlapping score measures the similarity between two protein complexes, and in our simulation, we measure the degree of overlap between sets of candidate overlapping protein complexes using Equation (9). Hence, to assess the effect of π on the performance of the WECALM, we adjust the default threshold value, π from 0.1 to 1.0 with a 0.1 increment, then calculate the composite score based on evaluation metrics including Recall, Precision, F-measure, CR, and MMR. We used the BioGRID and DIP yeast PPI complexes in Table 1 and the NewMIPS and CYC2008 reference protein complexes in Table 2. Figure 5 shows the composite score for WECALM performance at different π values on the BioGRID and the DIP datasets. In Figure 5, we noticed that both the BioGRID (Figure 5A) and DIP (Figure 5B) complexes on the NewMIPS the recall, MMR, and CR scores decrease with increase in π, while the precision and F-measure score is maximum at $\pi = 0.8$ and $\pi = 0.4$, respectively. Figure 5B. We also investigated the effect of π on the performance of WECALM using the CYC2008 reference protein complexes. Again it can be seen that on both the BioGRID (Figure 5C) and DIP (Figure 5D) complexes the recall, and CR scores decrease with an increase in π, whereas the precision and F-measure score is maximum at $\pi = 0.80$ and $\pi = 0.40$, respectively. The MMR is a maximum when $\pi = 0.2$. However, we notice that WECALM has a higher CR score on the BioGRID complex compared to the DIP complex. The overall performance of WECALM is best at $\pi = 0.4$ for both the BioGRID and DIP complexes which provide insights into overlapping structural similarities and differences between different protein complexes. Therefore, by tuning the optimal π value, our WECALM approach can achieve good accuracy and reliable prediction results.

5.2.2. Effect of Varying η on the Performance of WECALM

The local modularity structural similarity score threshold describes the similarity between the local modular structures of protein complexes. In this study, it was also necessary to evaluate the effect of η on the performance of WECALM. To evaluate the WECALM performance, we measured the Recall, Precision, F-measure, CR, and MMR using the BioGRID and DIP yeast PPI complexes (see Table 1) and the NewMIPS and CYC2008 reference protein complexes (see Table 2). To evaluate the performance of WECALM we adjusted the η threshold value in the range of $[0.1, 1.0)$ with a 0.1 increment and set $\eta > 0$ and $\eta \neq 0$. In Figure 6A,B, we noticed that in the BioGRID and DIP complexes on NewMIPS the recall, MMR, and CR scores decrease with an increase in η value, whereas the precision and F-measure score is maximum at $\eta = 0.8$ and $\eta = 0.4$, respectively. Using CYC2008 reference protein complexes we can see that in both BioGRID (Figure 6C) and DIP (Figure 6D) the recall, MMR, and CR scores decrease with an increase in η, whereas the precision and F-measure score is maximum at $\eta = 0.80$ and $\eta = 0.40$, respectively. However, WECALM has a higher CR score on the BioGRID complex compared to the DIP complex.

The overall performance of WECALM is best to perform at $\eta = 0.40$ for both the BioGRID and DIP complexes and on both NewMIPS and CYC2008 reference complexes, which provide an insight into local modularity structural similarities and differences between different modular proteins in the protein complexes. An algorithm with optimal η is able to identify common features or properties of protein complexes, which can give insights into the mechanisms of cellular processes.

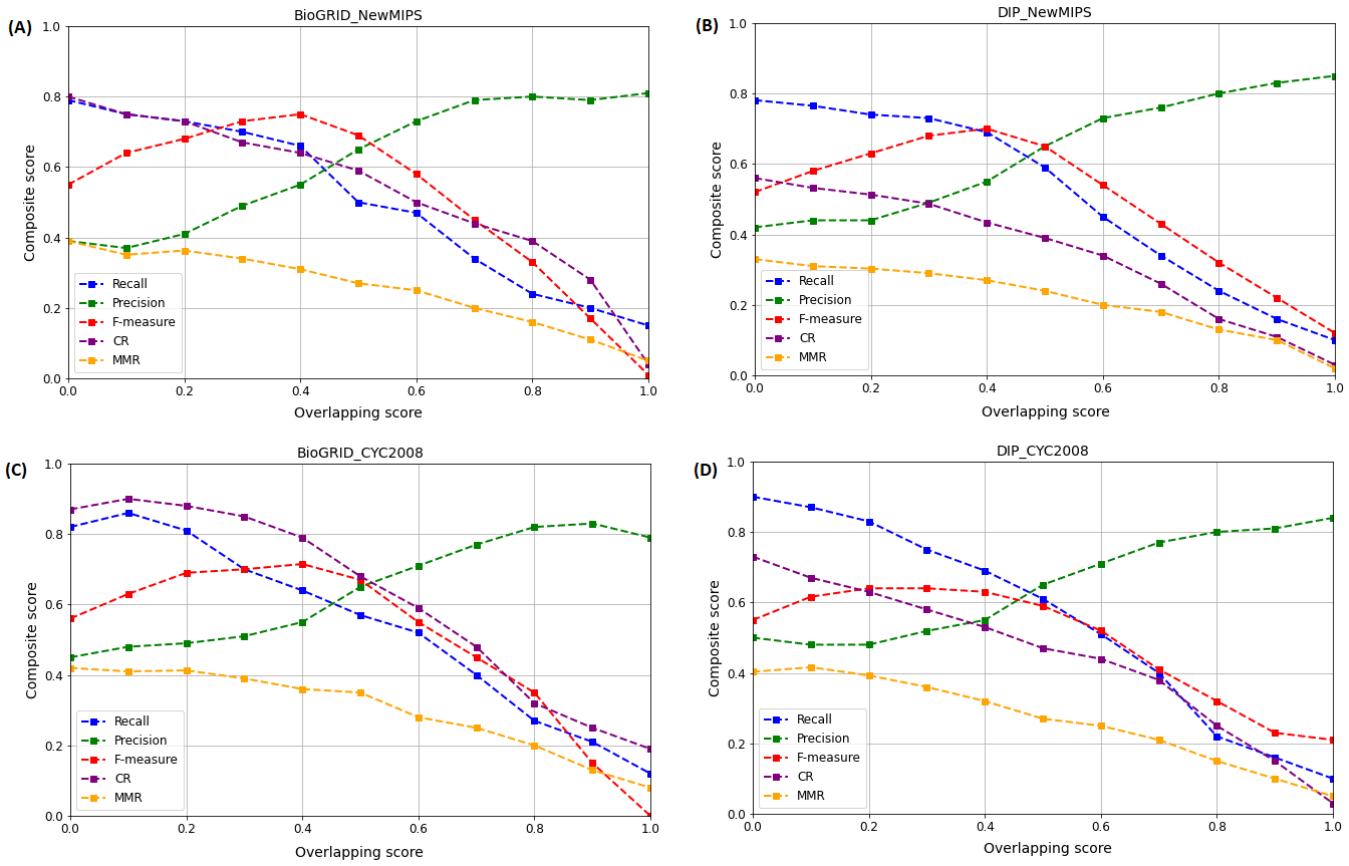

Figure 5. The effect of π on the performance of WECALM based on Recall, Precision, F-measure, CR, and MMR matrices. π is the predefined overlapping threshold. (**A**): The performance on the BioGRID based on NewMIPS. (**B**): The performance on DIP based on NewMIPS. (**C**): The performance on BioGRID based on CYC2008. (**D**): performance on DIP based on CYC2008. The MMR and F-measure are maximum when $\pi = 0.2$ and $\pi = 0.4$, respectively for the BioGRID on both NewMIPS and CYC2008.

5.2.3. Effect of Varying ω on the Performance of WECALM

The core structural similarity score threshold describes the similarity between the core structures of protein complexes. In this study, an evaluation of this parameter was essential to ascertain whether WECALM can correctly detect the core-protein complexes, which plays a key role in the study of the functional relationships between different protein complexes, by comparing their core structures to see if they have similar functions. This can provide insights into the functional roles of protein complexes in cellular processes. Therefore, to evaluate the WECALM performance, we measure the Recall, Precision, F-measure, CR, and MMR using the BioGRID and DIP yeast PPI complexes (see in Table 1) and NewMIPS and CYC2008 reference protein complexes (see Table 2). To evaluate the performance of WECALM we adjusted the ω threshold value in the range of $[0.1, 1.0)$ with a 0.1 increment and set $\omega \geq 0$. In Figure 7, we see that in both BioGRID (Figure 7A) and DIP (Figure 7B) complexes on NewMIPS the recall, MMR, and CR scores decrease with increase in ω, whereas the precision and F-measure score are maximum at $\omega = 0.8$ and

$\omega = 0.4$, respectively. At the same time, we evaluated the effect of ω on the performance of WECALM using the CYC2008 reference protein complexes. Again we notice that in both the BioGRID (Figure 7C) and DIP (Figure 7D) the recall, MMR, and CR scores decrease with increase in ω, while the precision and F-measure score are maximum at $\omega = 0.80$ and $\omega = 0.40$, respectively. However, WECALM has higher CR score on BioGRID complex compared in DIP complex. Overall performance of WECALM is best at $\omega = 0.40$ for both the BioGRID and DIP complexes which provide an insight into core structural similarities and differences between different protein complexes cores.

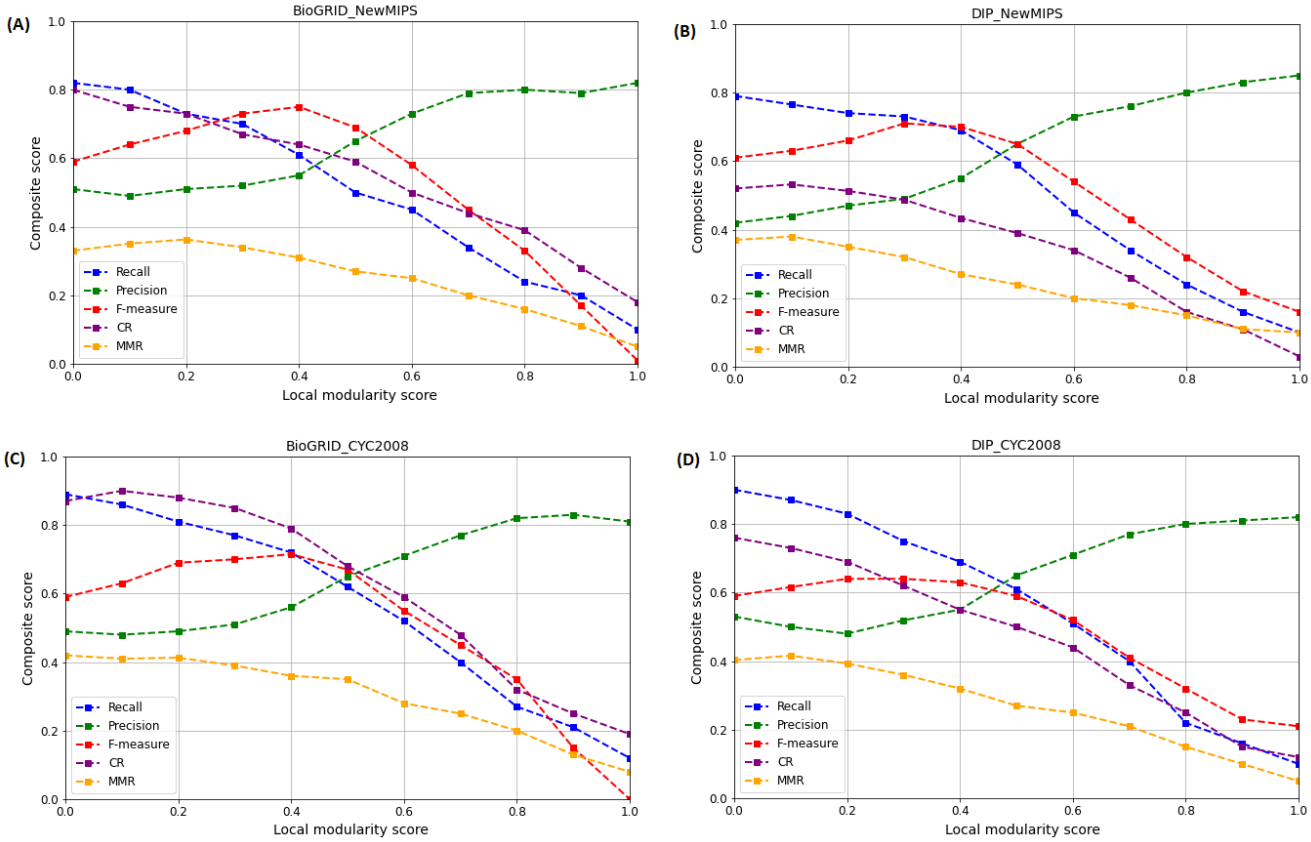

Figure 6. The effect η on the performance of WECALM based on Recall, Precision, F-measure, CR, and MMR matrices. η is a predefined local modularity threshold. (**A**): performance on BioGRID based on NewMIPS. (**B**): performance on DIP based on NewMIPS. (**C**): performance on BioGRID based on CYC2008. (**D**): performance on DIP based on CYC2008. The precision and F-measure is maximum when $\eta = 0.8$ and $\eta = 0.4$ respectively on both the NewMIPS and CYC2008.

5.3. Computational Complexity Analysis

In this study, it was also crucial to perform computational complexity analysis to assess the efficiency of WECALM relative to other algorithms in terms of the time required to detect the total number of protein complexes in standard complexes. To compare the computational complexity of each algorithm for simplicity we ran each program with its default settings. We then compared the time taken to detect the total number of detected protein complexes and matrices including the F-measure, CR, MMR, Sep, and ACC. In this analysis, we used reference Human and Yeast reference complexes (see Table 2).

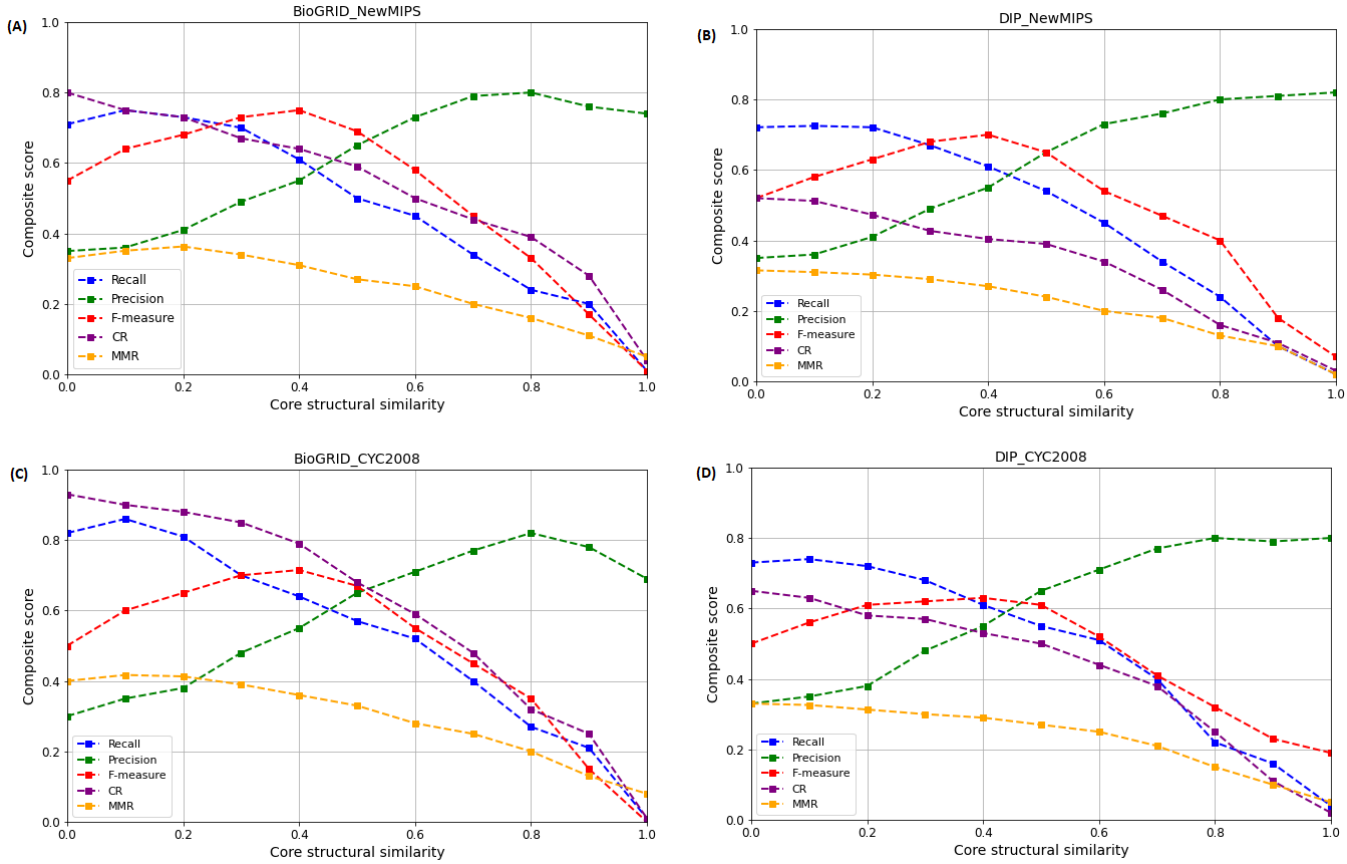

Figure 7. The effect of ω on the performance of WECALM based on Recall, Precision, F-measure, CR, and MMR matrices. is a predefined core structural similarity threshold. (**A**): The performance on BioGRID based on NewMIPS. (**B**): The performance on DIP based on NewMIPS. (**C**): The performance on BioGRID based on CYC2008. (**D**): The performance on DIP based on CYC2008. The precision and F-measure are maximum when $\omega = 0.8$ and $\omega = 0.4$, respectively, on both NewMIPS and CYC2008.

To compare the computational complexity of each algorithm we set the parameters of the other eight algorithms based on the authors' recommendations while for our proposed WECALM we set ω, π, and η at the default values obtained from the experimental results given in Section 5.2. We discovered in Table 3 that WECALM and EWCA had low computational complexity, indicating good efficiency in detecting the total amount of protein complexes in Human standard complexes. Furthermore, when compared to other algorithms, WECALM had the highest MMR, Sep, and ACC scores, demonstrating a balance between accuracy and efficiency. On standard yeast complexes, a similar performance trend was seen, with WECALM and EWCA having the lowest time computational complexity. However, we found that WECALM detected more protein complexes with a better performance efficiency than other algorithms, making it the overall best-performing algorithm for the detection of protein complexes on both Human and Yeast standard complexes.

5.4. Function Enrichment Analysis

We investigated the biological significance of our detected protein complexes to confirm the effectiveness of our WECALM approach because the reference complexes were incomplete. Each identified complex has a p-value calculated by Equation (36) for enrichment analysis. A complex is considered biologically significant if its p-value is less than $p \leq 10^{-2}$ after being detected using a wide range of methods. A complex with a lower p-value has a statistically significantly greater biological significance. Using SGD's GO Term Finder web service [83], we validated the functional relationships and cellular mechanism of the detected complexes based on biological process terms. In this case, the smallest

p-value across all gene ontology terms represents the functional homogeneity of each identification complex. We also evaluated the protein complexes identified by WECALM and calculated the *p*-value of protein complexes identified by MCL, COACH, Core, CALM, EWCA, CFinfer, Core, CALM, GMFTP, ClusterONE, CMC, and ProRank+ whose sizes were ≥ 3 Table 4 shows the *p*-value test results for MCL, COACH, Core, CALM, EWCA, CFinfer, GMFTP, ClusterONE, CMC, ProRank+, and WECALM. To compare the biological significance of protein complexes identified by different algorithms, we computed the number of detected complexes, the total number of detected complexes, and the percentage of detected complexes in different *p*-value ranges. We discovered on the one hand that the majority of algorithms only consider the percentage of detected complexes. The *p*-values of identified protein complexes, on the other hand, are proportional to their size [22,23,84,85].

Table 3. An evaluation of computational complexity and accuracy of WECALM and other algorithms.

Dataset	Algorithm	$C_P(v)$	F-Measure	CR	MMR	Sep	ACC	CPU Run Time (s)
Human	MCL	315	0.1001	0.1759	0.0105	0.1753	0.2167	5906.34
	COACH	4484	o.2455	0.5408	0.0677	0.5216	0.2777	2851.05
	EWCA	1979	0.4048	0.5221	0.0964	0.6081	0.5221	29.37
	CFinder	449	0.1256	0.2834	0.0116	0.3912	0.2511	3896.35
	GMFTP	773	0.2651	0.4193	0.0419	0.4917	0.3852	254.67
	Core	576	0.1621	0.3267	0.1267	0.3573	0.2778	2853.14
	CALM	1108	0.5127	0.5182	0.1394	0.6894	0.5289	198.39
	ClusterONE	375	0.1026	0.3071	0.0207	0.3773	0.2975	4895.78
	CMC	672	0.1251	0.2503	0.0183	0.2975	0.3313	3904.83
	ProRank+	838	0.3651	0.2856	0.0687	0.5526	0.5613	282.66
	WECALM	2367	0.4255	0.5155	0.0981	0.6155	0.6219	28.45
Yeast	MCL	298	0.1104	0.2761	0.0117	0.1625	0.1395	4967.47
	COACH	1551	0.2083	0.5521	0.0466	0.3583	0.3117	3603.31
	EWCA	936	0.4199	0.6182	0.0982	0.5904	0.5879	18.54
	CFinder	351	0.1429	0.2749	0.0281	0.3453	0.4163	3432.07
	GMFTP	675	0.2763	0.3129	0.0309	0.5145	0.4092	229.89
	Core	402	0.2124	0.2968	0.3285	0.1517	0.3218	2543.34
	CALM	732	0.4015	0.6787	0.1433	0.6261	0.6532	154.89
	ClusterONE	317	0.2012	0.2767	0.0285	0.3371	0.3255	3989.92
	CMC	589	0.2115	0.1975	0.0198	0.2934	0.3553	2987.63
	ProRank+	516	0.2712	0.2816	0.0487	0.5471	0.5602	251.54
	WECALM	1891	0.4216	0.6394	0.0487	0.64131	0.6534	17.65

$C_P(v)$: Detected Protein Complex; **CR**: Coverage Rate ; **MMR**: Maximum Match Ratio ; **Sep**: Separation ; **ACC**: Geometrical Mean Accuracy.

When analyzing the function enrichment of identified protein complexes, it is essential to consider both the quantity and the proportion of the identified complexes. On the BioGRID dataset, as shown in Table 4, WECALM detected 97.45% of the significant new protein complexes, slightly less than ProRank+, which recorded the highest significant score (97.59%). The size of the protein complexes identified by WECALM is typically larger than that of other algorithms such as ProRank+, which is most likely why. WECALM detects far more protein complexes than ProRank+. MCL, COACH, Core, CALM, EWCA, CFinfer, GMFTP, ClusterONE, and CMC is 107, 161, 1035, 463, 1341, 269, 449, 210, 832, and 728 protein complexes in the BioGRID dataset, respectively. We also observe that WECALM has detected a maximum of 1376 protein complexes, significantly better than ProRank+. On the DIP dataset, WECALM detected 96.17% of significant protein complexes, compared to ProRank+'s 93.79%, an increase of about 3%. At the same time, WECALM also identified the most protein complexes. In the DIP dataset, there were 113, 144, 315, 603, 869, 272, 350, 235, 146, and 319 protein complexes identified by MCL COACH, Core, CALM, EWCA, CFinfer, GMFTP, and CMC, respectively. In general, as the percentage of detected protein complexes decreases, the proportion of significant protein complexes increases. COACH,

CFinder, and GMFTP discovered far fewer protein complexes than WECALM. Nevertheless, when compared to the WECALM method, they had a lower percentage of significant protein complexes. In terms of the total number of detected protein complexes and the percentage of detected complexes, WECALM outperformed the other methods in terms of functionality and biological significance. According to their *p*-value, these protein complexes detected by WECALM have a higher probability of being actual protein complexes.

Table 4. A function enrichment analysis of protein complexes detected on BioGRID and DIP complexes.

Dataset	Algorithm	$C_P(v)$	$p \leq 10^{-15}$	$p \leq 10^{-10}$	$p \leq 10^{-5}$	$p \leq 10^{-2}$	Significant Detected $C_P(v)$
BioGRID	MCL	121	41 (33.88%)	28 (23.14%)	26 (21.49%)	12 (9.92%)	107 (88.43%)
	COACH	166	76 (45.78%)	32 (19.28%)	37 (22.29%)	16 (9.64%)	161 (96.98%)
	EWCA	1388	658 (47.41%)	211 (15.20%)	299 (21.54%)	173 (12.46%)	1341 (96.61%)
	CFinder	352	103 (29.26%)	53 (15.10%)	78 (22.16%)	35 (9.94%)	269 (76.42%)
	GMFTP	597	73 (12.23%)	59 (9.88%)	156 (26.13%)	161 (26.97%)	449 (75.21%)
	Core	576	255 (44.27%)	105 (18.23%)	68 (11.81%)	35 (6.08%)	463 (80.38%)
	CALM	1108	587 (52.98%)	236 (21.29%)	116 (10.47%)	96 (8.66%)	1035 (93.41%)
	ClusterONE	294	107 (36.40%)	35 (11.91%)	43 (14.62%)	25 (8.50%)	210 (71.43%)
	CMC	1113	125 (11.23%)	89 (7.99%)	258 (23.18%)	360 (32.34%)	832 (74.75%)
	ProRank+	746	479 (64.21%)	105 (14.08%)	97 (13.00%)	47 (6.30%)	728 (97.59%)
	WECALM	1412	687 (48.65%)	217 (15.37%)	312 (22.09%)	172 (12.18%)	1388 (98.30%)
DIP	MCL	142	41 (28.87%)	29 (20.42%)	17 (11.97%)	26 (18.31%)	113 (79.58%)
	COACH	329	21 (6.38%)	25 (7.59%)	66 (20.06%)	32 (9.73%)	144 (43.77%)
	EWCA	964	188 (19.50%)	126 (13.07%)	319 (33.09%)	236 (24.48%)	869 (90.15%)
	CFinder	352	157 (44.60%)	39 (11.08%)	31 (8.81%)	45 (12.78%)	272 (77.27%)
	GMFTP	548	43 (7.85%)	36 (6.57%)	105 (19.16%)	166 (30.29%)	350 (63.87%)
	Core	412	131 (31.79%)	87 (21.12%)	52 (12.62%)	45 (10.922%)	315 (76.46%)
	CALM	755	256 (33.91%)	127 (16.82%)	112 (14.83%)	108 (14.31%)	603 (80.53%)
	ClusterONE	315	119 (37.78%)	49 (15.56%)	38 (12.06%)	29 (9.21%)	235 (74.60%)
	CMC	303	3 (0.99%)	8 (2.64%)	58 (19.14%)	77 (25.41%)	146 (48.18%)
	ProRank+	338	74 (21.89%)	77 (22.78%)	125 (36.98%)	41 (12.13%)	319 (93.79%)
	WECALM	1018	269 (26.42%)	187 (18.37%)	358 (35.17%)	165 (16.21%)	979 (96.17%)

$C_P(v)$: Protein Complex; p: *p*-value.

The WECALM detected five protein complexes with extremely low *p*-values using the BioGRID and DIP complex datasets, as shown in Appendix B (see Tables A3 and A4), to further validate the biological significance of the identified complexes. The Cluster frequency, Genome frequency, Biological Process *p*-values, False Discovery Rate (FDR), False Positive score value, and Gene Ontology term descriptions were all evaluated in our analysis. Cluster frequency is a metric employed in the evaluation of algorithms designed to detect protein complexes. It represents the number of times the algorithm detects a specific complex across several replicates or runs of a similar test.

In Table A3 in Appendix B we can see that WECALM detected protein complexes with a high cluster frequency on the BioGRID complexes. A high cluster frequency implies that the WECALM detects a protein complex consistently throughout multiple runs, implying a good performance. This also means that most of the detected protein complexes in the BioGRID dataset closely match the gene ontology term and have a functional relationship with high statistical significance. According to results in Table A4, WECALM detected protein complexes with a high cluster frequency on the DIP complexes a clear indication of good performance. In addition, in Table A5, we see that WECALM detected a large number of complexes with a 100% cluster frequency. In the detection of protein complexes, a cluster frequency of 100% means that a particular complex is detected in all runs of the test. This is regarded as a very good indication of the findings' robustness and repeatability, as it implies that the complex is consistently detected by the algorithm across numerous replicates of the same experiment. WECALM had a very low $p < 10^{-10}$, indicating that the detected protein

complexes were biologically significant and meaningful and that they were most likely the true protein complexes, this can be used as a valuable benchmark in future research.

6. Conclusions

Advancements in biological mechanism research have led to the discovery of more disease-associated genes. Analyzing the Protein–Protein interaction (PPI) networks of these genes can help identify new disease-associated genes and clarify their role in specific diseases. This study proposes a new approach called WECALM, which uses a structural-based weighted network analysis of protein complexes using experimentally determined PPI datasets.

WECALM combines different graph mining algorithms based on protein complex structures and local attachment proteins to predict the inherent structure of the protein complexes in the PPI network. The approach introduces a new edge weight calculation method based on the Jaccard similarity measure between interacting proteins in the PPI networks, which improves the reliability of PPI networks for the accurate detection of protein complexes. It also integrates different network structural-based algorithms to detect overlapping structures, local modularity structures, and co-attachment structures in PPI networks, making it more robust in detecting protein complexes with different structures and densities than existing methods [14–16,22,23].

The study demonstrates that WECALM outperforms existing methods in terms of accuracy, computational speed, and the ability to detect biologically significant new protein complexes. Its biggest biological relevance could be that it helps to reduce false positive detection of protein complexes predicted with topological-based only methods. Nevertheless, the accuracy and efficiency of protein complex detection depend on predefined parameter tuning and the size of the PPI network. Additionally, WECALM is an *in silico* method applied to PPI networks, and future research should confirm its efficiency on other biological networks and conduct laboratory tests to validate its findings.

Author Contributions: Conceptualization, P.J.O., J.D., M.K. and T.K.; methodology, P.J.O., J.D. and M.K.; software, P.J.O.; visualization, P.J.O.; formal analysis, P.J.O., J.D., T.K. and M.K.; investigation, P.J.O., J.D., M.K. and T.K.; supervision, J.D. and M.K.; project administration, M.K.; funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

Funding: M.K. gratefully acknowledges the European Commission for funding the InnoRenew CoE project (Grant Agreement no. 739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund). He is also grateful for the support of the Slovenian Research Agency (ARRS) through grants J2-2504, N1-0223 and N2-0171. T.K. was supported by the National Laboratory of Biotechnology through the Hungarian National Research, Development and Innovation Office—NKFIH grant No. 2022-2.1.1-NL-2022-00008.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The code and datasets related to this study is available at https://github.com/peter26jumaochieng (accessed on 20 March 2023).

Acknowledgments: The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program (RRF-2.3.1-21-2022-00004). The research was also funded by the National Research, Development, and Innovation Fund of the Ministry of Innovation and Technology of Hungary under the TKP2021-NVA (Project no. TKP2021-NVA-09) funding scheme.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Performance Comparison of WECALM with the Other Algorithms

Performance Comparison

Tables A1 and A2 compare the performance of WECALM with the other algorithms based on NewMIPS and CYC2008, respectively. We list the performance score of each method in terms of Recall, Precision, F-Measure, CR, MMR, Sep, and ACC scores in both tables.

Table A1. Performance comparison of WECALM with other algorithms on the NewMIPS Complexes.

Dataset	Algorithm	Recall	Precision	F-Measure	CR	MMR	Sep	ACC
BioGRID	MCL	0.2896	0.2011	0.2374	0.2995	0.1672	0.2679	0.2167
	COACH	0.7256	0.2581	0.3808	**0.6322**	0.2525	0.5532	0.2777
	EWCA	**0.7561**	**0.5923**	**0.6643**	**0.6497**	**0.3764**	**0.6149**	0.5221
	CFinder	0.5914	0.1965	0.2950	0.4402	0.2801	0.3912	0.2131
	GMFTP	**0.7532**	0.2831	0.4115	0.5187	0.2552	0.5174	0.4522
	Core	0.5619	0.1488	0.2352	**0.5882**	0.1437	0.4575	0.3456
	CLAM	**0.7352**	**0.6681**	**0.7001**	**0.6158**	**0.3145**	**0.6478**	**0.5576**
	ClusterONE	0.5914	0.3133	0.4096	0.5311	0.1931	0.4851	0.2951
	CMC	0.5131	0.2731	0.3565	0.4954	0.3175	0.3976	**0.5313**
	ProRank+	0.4817	**0.7131**	0.5750	0.4763	0.2411	**0.6276**	0.5119
	WECALM	**0.7701**	**0.6853**	**0.7252**	**0.6743**	**0.3975**	**0.7765**	**0.6015**
DIP	MCL	0.5148	0.1783	0.2649	0.3271	0.1655	0.2927	0.1655
	COACH	0.5731	0.5106	**0.5400**	0.3353	0.2006	0.3833	0.2917
	EWCA	**0.7012**	0.4892	**0.5763**	**0.3982**	**0.3094**	0.6198	**0.5994**
	CFinder	0.5762	0.2408	0.3397	0.2403	**0.2128**	0.3543	0.3613
	GMFTP	**0.6982**	0.2756	0.3952	**0.4044**	0.2228	0.5548	0.4229
	Core	0.4421	0.1746	0.2503	0.3902	0.1249	0.5439	0.4374
	CLAM	0.5895	**0.5213**	**0.5533**	**0.4364**	**0.2962**	**0.6625**	0.5456
	ClusterONE	0.4054	0.3021	0.3462	0.2417	0.2178	0.3041	0.3185
	CMC	0.5932	0.4152	0.4885	**0.5736**	0.2499	0.3793	0.3151
	ProRank+	0.4086	**0.6657**	0.5064	0.2445	0.1669	**0.6451**	**0.5567**
	WECALM	**0.7166**	**0.4998**	**0.5889**	**0.4195**	**0.3531**	**0.8195**	**0.6317**

CR: Coverage Rate ; **MMR**: Maximum Match Ratio; **Sep**: Separation ; **ACC**: Geometrical Mean Accuracy; bold value indicates the best score.

Table A2. Performance comparison of WECALM with other algorithms on the CYC2008 complexes.

Dataset	Algorithm	Recall	Precision	F-Measure	CR	MMR	Sep	ACC
BioGRID	MCL	0.3516	0.2268	0.2757	0.5313	0.1645	0.3831	0.2549
	COACH	0.7669	0.2488	0.3757	**0.8752**	0.3042	0.5375	0.4117
	EWCA	**0.8191**	**0.5793**	**0.6786**	**0.8718**	**0.4351**	**0.6578**	**0.6035**
	CFinder	0.5724	0.1637	0.2546	0.6135	0.3115	0.5634	0.4215
	GMFTP	**0.7839**	0.2914	0.4249	0.7956	0.3914	0.6192	0.4591
	Core	0.5847	0.1527	0.2422	**0.8058**	0.2081	0.4126	0.2742
	CLAM	**0.6984**	**0.6211**	**0.6575**	**0.8583**	**0.3968**	**0.7293**	**0.7153**
	ClusterONE	0.6612	0.3487	0.4566	0.7569	0.2734	0.5162	0.3574
	CMC	0.4644	0.2677	0.3396	0.7639	0.3375	0.4611	0.4375
	ProRank+	0.4153	**0.6622**	0.5105	0.5851	0.2462	0.6105	0.5979
	WECALM	**0.8291**	**0.5991**	**0.6956**	**0.8831**	**0.4825**	**0.7825**	**0.6983**
DIP	MCL	0.5169	0.1847	0.2722	0.4892	0.2299	0.3519	0.3125
	COACH	0.5423	0.5167	**0.5292**	0.4879	0.2764	0.4272	0.3967
	EWCA	**0.7076**	0.5239	**0.6020**	**0.5806**	**0.3766**	**0.6436**	0.5527
	CFinder	0.5508	0.2398	0.3341	0.2788	**0.3807**	0.3758	0.4187
	GMFTP	**0.6652**	0.2664	0.3804	**0.6085**	0.3316	0.6235	0.4136
	Core	0.4618	0.1818	0.2609	0.5317	0.2433	0.3351	0.3519
	CLAM	0.6465	**0.4915**	**0.5584**	0.5345	**0.3221**	**0.6833**	**0.6721**
	ClusterONE	0.4279	0.3343	0.3754	0.3751	0.2191	0.3957	0.3519
	CMC	0.4932	0.4125	0.4493	**0.5755**	0.2501	0.4576	0.3251
	ProRank+	0.3772	**0.6924**	0.4884	0.3294	0.2029	0.5929	**0.5817**
	WECALM	**0.7315**	**0.5556**	**0.6315**	**0.5916**	**0.3866**	**0.7596**	**0.6569**

CR: Coverage Rate; **MMR**: Maximum Match Ratio; **Sep**: Separation ; **ACC**: Geometrical Mean Accuracy; bold value indicates the best score.

Appendix B. A Function Enrichment Analysis

Here, we provide supplementary results for a functional enrichment analysis to confirm the biological significance of detected complexes by WECALM on the BioGRID and DIP complexes listed in Tables A3 and A4, respectively. Tables A3 and A4 provide a list of the complex ID, Cluster frequency, Genome frequency, *p*-value (Biological process), False Discovery Rate (FDR), False Positive score and Gene Ontology term description. Our evaluation the selection of significant Gene Ontology term is purely based on the e Cluster frequency, *p*-values and FDR values.

Appendix B.1. A Function Enrichment Analysis on BioGRID Complex

Table A3 lists significant GO Ontology terms shared by proteins in the BioGRID complexes dataset. From the results we notice that the majority of detected protein complexes match the Gene ontology term well. In addition, it can be seen that the *p*-value of detected complexes is very low, which implies that the detected protein complexes have a high statistical significance.

Table A4 lists significant GO Ontology terms shared by proteins in the DIP complexes dataset. Similar to BioGRID complex dataset, we see in the DIP complex that most of the detected protein complex match the Gene ontology term well. We notice that the *p*-value of detected complexes is very low, which implies that the detected protein complexes have a high statistical significance.

Table A3. Top five protein complexes with significant low *p*-value detected by WECALM on BioGRID complex.

Complex ID	Cluster Frequency	Genome Frequency	*p*-Value (BP)	FDR	FALSE Positive	Gene Ontology Term
1	9 of 12 genes,75.0%	44 of 7166 genes, 0.6%	1.93×10^{-16}	0.0000	0.0000	positive regulation of transcription elongation by RNA polymerase II
	9 of 12 genes, 75.0%	47 of 7166 genes, 0.7%	3.72×10^{-16}	0.0000	0.0000	regulation of transcription elongation by RNA polymerase II
	9 of 12 genes,75.0%	52 of 7166 genes, 0.7%	1.00×10^{-15}	0.0000	0.0000	positive regulation of DNA-templated transcription, elongation
	9 of 12 genes,75.0%	55 of 7166 genes, 0.8%	1.73×10^{-15}	0.0000	0.0000	regulation of DNA-templated transcription elongation
	9 of 12 genes, 75.0%	96 of 7166 genes, 1.3%	3.47×10^{-13}	0.0000	0.0000	transcription elongation by RNA polymerase II
2	12 of 13 genes, 92.3%	936 of 7166 genes, 13.1%	4.40×10^{-4}	0.0000	0.0000	amide metabolic process
	12 of 13 genes,92.3%	1348 of 7166 genes,18.8%	8.30×10^{-4}	0.0000	0.0000	organonitrogen compound biosynthetic process
	12 of 13 genes, 92.3%	1816 of 7166 genes, 25.3%	1.18×10^{-3}	0.0000	0.0000	cellular nitrogen compound biosynthetic process
	12 of 13 genes, 92.3%	2109 of 7166 genes, 29.4%	5.58×10^{-3}	0.0000	0.0000	cellular biosynthetic process
	12 of 13 genes, 92.3%	2725 of 7166 genes, 38.0%	7.22×10^{-3}	0.0000	0.0000	cellular nitrogen compound metabolic process
3	11 of 12 genes, 91.7%	367 of 7166 genes, 5.1%	4.78×10^{-10}	0.0000	0.0000	rRNA processing
	11 of 12 genes, 91.7%	423 of 7166 genes, 5.9%	1.98×10^{-9}	0.0000	0.0000	rRNA metabolic process
	11 of 12 genes, 91.7%	482 of 7166 genes, 6.7%	7.29×10^{-9}	0.0000	0.0000	ribosome biogenesis
	11 of 12 genes, 91.7%	492 of 7166 genes, 6.9%	8.94×10^{-9}	0.0000	0.0000	ncRNA processing
	11 of 12 genes, 91.7%	2159 of 7166 genes, 30.1%	1.14×10^{-3}	0.0000	0.0000	gene expression
4	13 of 14 genes, 92.9%	204 of 7166 genes, 2.8%	3.84×10^{-18}	0.0000	0.0000	cytoplasmic translation
	13 of 14 genes, 92.9%	820 of 7166 genes, 11.4%	3.38×10^{-10}	0.0000	0.0000	translation
	13 of 14 genes, 92.9%	824 of 7166 genes, 11.5%	3.60×10^{-10}	0.0000	0.0000	peptide biosynthetic process
	13 of 14 genes, 92.9%	841 of 7166 genes, 11.7%	4.70×10^{-10}	0.0000	0.0000	peptide metabolic process
	13 of 14 genes, 92.9%	879 of 7166 genes, 12.3%	8.34×10^{-10}	0.0000	0.0000	amide biosynthetic process
5	12 of 13 genes, 92.3%	204 of 7166 genes, 2.8%	1.32×10^{-16}	0.0000	0.0000	ribosomal large subunit biogenesis
	12 of 13 genes, 92.3%	820 of 7166 genes, 11.4%	2.78×10^{-9}	0.0000	0.0000	biosynthetic process
	12 of 13 genes, 92.3%	824 of 7166 genes, 11.5%	2.95×10^{-9}	0.0000	0.0000	peptide biosynthetic process
	12 of 13 genes, 92.3%	841 of 7166 genes, 11.7%	3.77×10^{-9}	0.0000	0.0000	ribonucleoprotein complex biogenesis
	12 of 13 genes, 92.3%	879 of 7166 genes, 12.3%	6.39×10^{-9}	0.0000	0.0000	cellular biosynthetic process

FDR: False Discovery Rate ; **BP**: Biological Process; BP is significant at *p*-value $< 10^{-2}$.

Appendix B.2. A Function Enrichment Analysis on DIP Complex

Table A4. Top five protein complexes with significant low *p*-value detected by WECALM on DIP complex.

Complex ID	Cluster Frequency	Genome Frequency	*p*-Value	FDR	FALSE Positive	Gene Ontology Term
1	11 of 12 genes, 91.7%	125 of 7166 genes, 1.7%	9.76×10^{-15}	0.0000	0.0000	ribosomal large subunit biogenesis
	11 of 12 genes, 91.7%	482 of 7166 genes, 6.7%	1.08×10^{-10}	0.0000	0.0000	ribosome biogenesis
	11 of 12 genes, 91.7%	576 of 7166 genes, 8.0%	7.74×10^{-10}	0.0000	0.0000	ribonucleoprotein complex biogenesis
	11 of 12 genes, 91.7%	1272 of 7166 genes, 17.8%	4.49×10^{-6}	0.0000	0.0000	cellular component biogenesis
	11 of 12 genes, 91.7%	2424 of 7166 genes, 33.8%	4.55×10^{-3}	0.0008	0.0000	cellular component organization or biogenesis
2	3 of 4 genes, 75.0%	56 of 7166 genes, 0.8%	1.10×10^{-4}	0.0000	0.0000	purine ribonucleoside triphosphate metabolic process
	3 of 4 genes, 75.0%	58 of 7166 genes, 0.8%	1.30×10^{-4}	0.0000	0.0000	purine nucleoside triphosphate metabolic process
	3 of 4 genes, 75.0%	119 of 7166 genes, 1.7%	1.14×10^{-3}	0.0000	0.0000	nucleotide biosynthetic process
	3 of 4 genes, 75.0%	121 of 7166 genes, 1.7%	1.20×10^{-3}	0.0000	0.0000	nucleoside phosphate biosynthetic process
	3 of 4 genes, 75.0%	125 of 7166 genes, 1.7%	1.33×10^{-3}	0.0000	0.0000	ribonucleotide metabolic process
3	9 of 10 genes, 90.0%	20 of 7166 genes, 0.3%	9.69×10^{-22}	0.0000	0.0000	ATP biosynthetic process
	9 of 10 genes, 90.0%	20 of 7166 genes, 0.3%	9.69×10^{-22}	0.0000	0.0000	proton motive force-driven ATP synthesis
	9 of 10 genes, 90.0%	24 of 7166 genes, 0.3%	7.54×10^{-21}	0.0000	0.0000	purine nucleoside triphosphate biosynthetic process
	9 of 10 genes, 90.0%	24 of 7166 genes, 0.3%	7.54×10^{-21}	0.0000	0.0000	purine ribonucleoside triphosphate biosynthetic process
	9 of 10 genes, 90.0%	30 of 7166 genes, 0.4%	8.25×10^{-20}	0.0000	0.0000	ribonucleoside triphosphate biosynthetic process
4	10 of 11 genes, 90.9%	20 of 7166 genes, 0.3%	1.59×10^{-24}	0.0000	0.0000	proton transmembrane transport
	10 of 11 genes, 90.9%	20 of 7166 genes, 0.3%	1.59×10^{-24}	0.0000	0.0000	purine ribonucleotide metabolic process
	10 of 11 genes, 90.9%	24 of 7166 genes, 0.3%	1.69×10^{-23}	0.0000	0.0000	nucleotide biosynthetic process
	10 of 11 genes, 90.9%	24 of 7166 genes, 0.3%	1.69×10^{-23}	0.0000	0.0000	nucleoside phosphate biosynthetic process
	10 of 11 genes, 90.9%	30 of 7166 genes, 0.4%	2.59×10^{-22}	0.0000	0.0000	ribonucleotide metabolic process
5	9 of 10 genes, 90.0%	444 of 7166 genes, 6.2%	1.08×10^{-8}	0.0000	0.0000	intracellular protein transport
	9 of 10 genes, 90.0%	449 of 7166 genes, 6.3%	1.19×10^{-8}	0.0000	0.0000	vesicle-mediated transport
	9 of 10 genes, 90.0%	630 of 7166 genes, 8.8%	2.52×10^{-7}	0.0000	0.0000	protein transport
	9 of 10 genes, 90.0%	651 of 7166 genes, 9.1%	3.38×10^{-7}	0.0000	0.0000	establishment of protein localization
	9 of 10 genes, 90.0%	742 of 7166 genes, 10.4%	1.09×10^{-6}	0.0000	0.0000	intracellular transport

FDR: False Discovery Rate ; **BP**: Biological Process; BP is significant at *p*-value $< 10^{-2}$

Appendix B.3. Detected Protein Complexes with 100% Cluster Frequency

Table A5 lists the top 10 complexes with 100% cluster frequency detected by WECALM on the BioGRID and DIP complexes.

Table A5. Top 10 protein complexes with 100% cluster frequency detected by WECALM on the BioGRID and DIP complex datasets.

Dataset	Complex ID	Cluster Frequency	Genome Frequency	*p*-Value (BP)	FDR	FALSE Positive	Gene Ontology Term
BioGRID	1	12 of 12 genes, 100.0%	122 of 7166 genes, 1.7%	9.16×10^{-15}	0.0000	0.0000	mRNA splicing, via spliceosome
	2	42 of 42 genes, 100.0%	123 of 7166 genes, 1.7%	1.27×10^{-10}	0.0000	0.0000	RNA splicing,
	3	10 of 10 genes, 100.0%	10 of 7166 genes, 0.1%	3.64×10^{-10}	0.0000	0.0000	spliceosomal tri-snRNP complex assembly
	4	19 of 19 genes, 100.0%	132 of 7166 genes, 1.8%	1.49×10^{-7}	0.0000	0.0000	RNA splicing, via transesterification reactions
	5	36 of 36 genes, 100.0%	157 of 7166 genes, 2.2%	4.55×10^{-3}	0.0000	0.0000	RNA splicing
	6	11 of 11 genes, 100.0%	20 of 7166 genes, 0.3%	9.29×10^{-21}	0.0000	0.0000	spliceosomal snRNP assembly
	7	10 of 10 genes, 100.0%	229 of 7166 genes, 3.2%	9.08×10^{-16}	0.0000	0.0000	mRNA processing
	8	17 of 17 genes, 100.0%	350 of 7166 genes, 4.9%	1.08×10^{-15}	0.0000	0.0200	mRNA metabolic process
	9	42 of 42 genes, 100.0%	347 of 7166 genes, 4.8%	1.51×10^{-15}	0.0000	0.0000	DNA-directed 5′-3′ RNA polymerase activity
	10	23 of 23 genes, 100.0%	34 of 7166 genes, 0.5%	6.25×10^{-20}	0.0000	0.0000	5′-3′ RNA polymerase activity
DIP	1	19 of 19 genes, 100.0%	62 of 7166 genes, 0.9%	4.49×10^{-19}	0.0000	0.0000	nucleotide-excision repair
	2	39 of 39 genes, 100.0%	234 of 7166 genes, 3.3%	7.76×10^{-19}	0.0000	0.0000	ubiquitin-dependent protein catabolic process
	3	36 of 36 genes, 100.0%	240 of 7166 genes, 3.3%	9.25×10^{-19}	0.0000	0.0000	modification-dependent protein catabolic process
	4	10 of 10 genes, 100.0%	262 of 7166 genes, 3.7%	3.40×10^{-18}	0.0000	0.0000	modification-dependent macromolecule catabolic process
	5	14 of 14 genes, 100.0%	264 of 7166 genes, 3.7%	8.19×10^{-18}	0.0000	0.0000	proteolysis involved in protein catabolic process
	6	13 of 13 genes, 100.0%	293 of 7166 genes, 4.1%	1.83×10^{-17}	0.0000	0.0000	protein catabolic process
	7	15 of 15 genes, 100.0%	309 of 7166 genes, 4.3%	3.03×10^{-17}	0.0000	0.0000	DNA repair
	8	12 of 12 genes, 100.0%	350 of 7166 genes, 4.9%	5.50×10^{-17}	0.0000	0.0000	cellular response to DNA damage stimulus
	9	31 of 31 genes, 100.0%	407 of 7166 genes, 5.7%	1.07×10^{-16}	0.0000	0.0000	organonitrogen compound catabolic process
	10	17 of 17 genes, 100.0%	416 of 7166 genes, 5.8%	3.64×10^{-16}	0.0000	0.0000	proteolysis

FDR: False Discovery Rate ; **BP**: Biological Process; BP is significant at *p*-value $< 10^{-2}$.

References

1. Almeida, R.M.; Dell'Acqua, S.; Krippahl, L.; Moura, J.J.; Pauleta, S.R. Predicting Protein–Protein interactions using bigger: Case studies. *Molecules* **2016**, *21*, 1037. [CrossRef]
2. Bustamam, A.; Siswantining, T.; Kaloka, T.P.; Swasti, O. Application of bimax, pols, and lcm-mbc to find bicluster on interactions protein between hiv-1 and human. *Austrian J. Stat.* **2020**, *49*, 1–18. [CrossRef]
3. Tripathi, S.; Moutari, S.; Dehmer, M.; Emmert-Streib, F. Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules. *BMC Bioinform.* **2016**, *17*, 129. [CrossRef] [PubMed]
4. Li, X.L.; Ng, S.K. *Biological Data Mining in Protein Interaction Networks*; IGI Global: Hershey, PA, USA, 2009.
5. Wu, D.; Hu, X. Topological analysis and sub-network mining of Protein–Protein interactions. In *Research and Trends in Data Mining Technologies and Applications*; IGI Global: Hershey, PA, USA, 2007; pp. 209–240.
6. Larsen, P.E.; Collart, F.; Dai, Y. Incorporating network topology improves prediction of protein interaction networks from transcriptomic data. *Int. J. Knowl. Discov. Bioinform. (IJKDB)* **2010**, *1*, 1–19. [CrossRef]
7. Ahnert, S.E.; Marsh, J.A.; Hernández, H.; Robinson, C.V.; Teichmann, S.A. Principles of assembly reveal a periodic table of protein complexes. *Science* **2015**, *350*, aaa2245. [CrossRef] [PubMed]
8. Tong, A.H.Y.; Drees, B.; Nardelli, G.; Bader, G.D.; Brannetti, B.; Castagnoli, L.; Evangelista, M.; Ferracuti, S.; Nelson, B.; Paoluzi, S.; et al. A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules. *Science* **2002**, *295*, 321–324. [CrossRef] [PubMed]
9. Shen, X.; Yi, L.; Jiang, X.; Zhao, Y.; Hu, X.; He, T.; Yang, J. Neighbor affinity based algorithm for discovering temporal protein complex from dynamic PPI network. *Methods* **2016**, *110*, 90–96. [CrossRef] [PubMed]
10. Zhang, X.F.; Dai, D.Q.; Ou-Yang, L.; Yan, H. Detecting overlapping protein complexes based on a generative model with functional and topological properties. *BMC Bioinform.* **2014**, *15*, 186. [CrossRef]
11. Shen, X.; Zhou, J.; Yi, L.; Hu, X.; He, T.; Yang, J. Identifying protein complexes based on brainstorming strategy. *Methods* **2016**, *110*, 44–53. [CrossRef]
12. Liu, G.; Wong, L.; Chua, H.N. Complex discovery from weighted PPI networks. *Bioinformatics* **2009**, *25*, 1891–1897. [CrossRef]
13. Adamcsek, B.; Palla, G.; Farkas, I.J.; Derényi, I.; Vicsek, T. CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics* **2006**, *22*, 1021–1023. [CrossRef] [PubMed]
14. Van Dongen, S.M. Graph clustering by Flow Simulation. Ph.D. Thesis, University of Utrecht, Utrecht, The Netherlands, 2000.
15. Vlasblom, J.; Wodak, S.J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinform.* **2009**, *10*, 99. [CrossRef] [PubMed]
16. Ochieng, P.J.; Kusuma, W.; Haryanto, T. Detection of protein complex from Protein–Protein interaction network using Markov clustering. In *Proceedings of the Journal of Physics: Conference Series*; IOP Publishing: Bristol, UK, 2017; Volume 835, p. 012001.
17. Wang, R.; Liu, G.; Wang, C. Identifying protein complexes based on an edge weight algorithm and core-attachment structure. *BMC Bioinform.* **2019**, *20*, 471. [CrossRef] [PubMed]
18. Xie, D.; Yi, Y.; Zhou, J.; Li, X.; Wu, H. A novel temporal protein complexes identification framework based on density–Distance and heuristic algorithm. *Neural Comput. Appl.* **2019**, *31*, 4693–4701. [CrossRef]
19. Jiang, P.; Singh, M. SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics* **2010**, *26*, 1105–1111. [CrossRef]
20. Nepusz, T.; Yu, H.; Paccanaro, A. Detecting overlapping protein complexes in Protein–Protein interaction networks. *Nat. Methods* **2012**, *9*, 471–472. [CrossRef]
21. Wang, R.; Liu, G.; Wang, C.; Su, L.; Sun, L. Predicting overlapping protein complexes based on core-attachment and a local modularity structure. *BMC Bioinform.* **2018**, *19*, 305. [CrossRef]
22. Wu, M.; Li, X.; Kwoh, C.K.; Ng, S.K. A core-attachment based method to detect protein complexes in PPI networks. *BMC Bioinform.* **2009**, *10*, 169 . [CrossRef] [PubMed]
23. Leung, H.C.; Xiang, Q.; Yiu, S.M.; Chin, F.Y. Predicting protein complexes from PPI data: A core-attachment approach. *J. Comput. Biol.* **2009**, *16*, 133–144. [CrossRef] [PubMed]
24. Hanna, E.M.; Zaki, N. Detecting protein complexes in protein interaction networks using a ranking algorithm with a refined merging procedure. *BMC Bioinform.* **2014**, *15*, 204. [CrossRef] [PubMed]
25. Palla, G.; Derényi, I.; Farkas, I.; Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **2005**, *435*, 814–818. [CrossRef] [PubMed]
26. Karp, R.M. Reducibility among combinatorial problems. In Proceedings of the Complexity of Computer Computations: Proceedings of a symposium on the Complexity of Computer Computations, New York, NY, USA, 20–22 March 1972; Springer: Berlin/Heidelberg, Germany, 1972; pp. 85–103.
27. Gens, G.V.; Levner, E.V. Computational complexity of approximation algorithms for combinatorial problems. In Proceedings of the Mathematical Foundations of Computer Science 1979: Proceedings, 8th Symposium, Olomouc, Czechoslovakia, 3–7 September 1979; Springer: Berlin/Heidelberg, Germany, 1979; pp. 292–300.
28. Spirin, V.; Mirny, L.A. Protein complexes and functional modules in molecular networks. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12123–12128. [CrossRef]
29. Bader, S.; Kühner, S.; Gavin, A.C. Interaction networks for systems biology. *FEBS Lett.* **2008**, *582*, 1220–1224. [CrossRef]
30. Zaki, N.; Efimov, D.; Berengueres, J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinform.* **2013**, *14*, 163. [CrossRef]

31. Cao, B.; Luo, J.; Liang, C.; Wang, S.; Ding, P. Pce-fr: A novel method for identifying overlapping protein complexes in weighted Protein–Protein interaction networks using pseudo-clique extension based on fuzzy relation. *IEEE Trans. Nanobiosci.* **2016**, *15*, 728–738. [CrossRef] [PubMed]
32. Wang, J.; Chen, G.; Liu, B.; Li, M.; Pan, Y. Identifying protein complexes from interactome based on essential proteins and local fitness method. *IEEE Trans. Nanobiosci.* **2012**, *11*, 324–335. [CrossRef]
33. Kreimer, A.; Borenstein, E.; Gophna, U.; Ruppin, E. The evolution of modularity in bacterial metabolic networks. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 6976–6981. [CrossRef]
34. Luo, F.; Yang, Y.; Chen, C.F.; Chang, R.; Zhou, J.; Scheuermann, R.H. Modular organization of protein interaction networks. *Bioinformatics* **2007**, *23*, 207–214. [CrossRef] [PubMed]
35. Poyatos, J.F.; Hurst, L.D. How biologically relevant are interaction-based modules in protein networks? *Genome Biol.* **2004**, *5*, R93. [CrossRef]
36. Ren, J.; Wang, J.; Li, M.; Wang, L. Identifying protein complexes based on density and modularity in Protein–Protein interaction network. *BMC Syst. Biol.* **2013**, *7*, S12. [CrossRef]
37. Bóta, A.; Csizmadia, L.; Pluhár, A. Community detection and its use in Real Graphs. In Proceedings of the 2010 Mini-Conference on Applied Theoretical Computer Science Koper, Slovenia, 13–14 October 2010.
38. Gera, I.; London, A.; Pluhár, A. Greedy algorithm for edge-based nested community detection. In Proceedings of the 2022 IEEE 2nd Conference on Information Technology and Data Science (CITDS), Debrecen, Hungary, 16–18 May 2022; pp. 86–91.
39. Dezső, Z.; Oltvai, Z.N.; Barabási, A.L. Bioinformatics analysis of experimentally determined protein complexes in the yeast Saccharomyces cerevisiae. *Genome Res.* **2003**, *13*, 2450–2454. [CrossRef]
40. Pu, S.; Vlasblom, J.; Emili, A.; Greenblatt, J.; Wodak, S.J. Identifying functional modules in the physical interactome of Saccharomyces cerevisiae. *Proteomics* **2007**, *7*, 944–960. [CrossRef] [PubMed]
41. Gavin, A.C.; Aloy, P.; Grandi, P.; Krause, R.; Boesche, M.; Marzioch, M.; Rau, C.; Jensen, L.J.; Bastuck, S.; Dümpelfeld, B.; et al. Proteome survey reveals modularity of the yeast cell machinery. *Nature* **2006**, *440*, 631–636. [CrossRef]
42. Bruckner, S.; Hüffner, F.; Komusiewicz, C. A graph modification approach for finding core–periphery structures in protein interaction networks. *Algorithms Mol. Biol.* **2015**, *10*, 16. [CrossRef] [PubMed]
43. Meng, X.; Li, W.; Peng, X.; Li, Y.; Li, M. Protein interaction networks: Centrality, modularity, dynamics, and applications. *Front. Comput. Sci.* **2021**, *15*, 156902. [CrossRef]
44. Ma, X.; Gao, L. Predicting protein complexes in protein interaction networks using a core-attachment algorithm based on graph communicability. *Inf. Sci.* **2012**, *189*, 233–254. [CrossRef]
45. Mete, M.; Tang, F.; Xu, X.; Yuruk, N. A structural approach for finding functional modules from large biological networks. *BMC Bioinform.* **2008**, *9*, S19. [CrossRef] [PubMed]
46. Yang, J.; Leskovec, J. Overlapping communities explain core–Periphery organization of networks. *Proc. IEEE* **2014**, *102*, 1892–1902. [CrossRef]
47. Vieira, V.d.F.; Xavier, C.R.; Evsukoff, A.G. A comparative study of overlapping community detection methods from the perspective of the structural properties. *Appl. Netw. Sci.* **2020**, *5*, 51. [CrossRef]
48. Gu, L.; Han, Y.; Wang, C.; Chen, W.; Jiao, J.; Yuan, X. Module overlapping structure detection in PPI using an improved link similarity-based Markov clustering algorithm. *Neural Comput. Appl.* **2019**, *31*, 1481–1490. [CrossRef]
49. Wang, Y.; Qian, X. Functional module identification in protein interaction networks by interaction patterns. *Bioinformatics* **2014**, *30*, 81–93. [CrossRef]
50. Aloy, P.; Bottcher, B.; Ceulemans, H.; Leutwein, C.; Mellwig, C.; Fischer, S.; Gavin, A.C.; Bork, P.; Superti-Furga, G.; Serrano, L.; et al. Structure-based assembly of protein complexes in yeast. *Science* **2004**, *303*, 2026–2029. [CrossRef]
51. Luo, F.; Li, B.; Wan, X.F.; Scheuermann, R.H. Core and periphery structures in protein interaction networks. *BMC Bioinform.* **2009**, *10*, S8. [CrossRef] [PubMed]
52. Bader, G.D.; Hogue, C.W. Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.* **2002**, *20*, 991–997. [CrossRef]
53. Bader, G.D.; Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 2. [CrossRef] [PubMed]
54. Kourtellis, N.; Alahakoon, T.; Simha, R.; Iamnitchi, A.; Tripathi, R. Identifying high betweenness centrality nodes in large social networks. *Soc. Netw. Anal. Min.* **2013**, *3*, 899–914. [CrossRef]
55. Barabasi, A.L.; Oltvai, Z.N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113. [CrossRef]
56. Gosak, M.; Markovič, R.; Dolenšek, J.; Rupnik, M.S.; Marhl, M.; Stožer, A.; Perc, M. Network science of biological systems at different scales: A review. *Phys. Life Rev.* **2018**, *24*, 118–135. [CrossRef]
57. Han, J.D.J. Understanding biological functions through molecular networks. *Cell Res.* **2008**, *18*, 224–237. [CrossRef]
58. Del Sol, A.; O'Meara, P. Small-world network approach to identify key residues in protein–protein interaction. *Proteins Struct. Funct. Bioinform.* **2005**, *58*, 672–682. [CrossRef] [PubMed]
59. Del Sol, A.; Fujihashi, H.; O'Meara, P. Topology of small-world networks of protein–protein complex structures. *Bioinformatics* **2005**, *21*, 1311–1315. [CrossRef] [PubMed]

60. Wang, X.; Li, L.; Cheng, Y. An overlapping module identification method in Protein–Protein interaction networks. *BMC Bioinform.* **2012**, *13*, S4. [CrossRef] [PubMed]
61. Liu, C.; Li, J.; Zhao, Y. Exploring hierarchical and overlapping modular structure in the yeast protein interaction network. *BMC Genom.* **2010**, *11*, S17. [CrossRef]
62. Jaccard, P. The distribution of the flora in the alpine zone. 1. *New Phytol.* **1912**, *11*, 37–50. [CrossRef]
63. Goodrich, M.T.; Ozel, E. Modeling the small-world phenomenon with road networks. In Proceedings of the 30th International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 1–4 November 2022; pp. 1–10.
64. Menezes, M.B.; Kim, S.; Huang, R. Constructing a Watts-Strogatz network from a small-world network with symmetric degree distribution. *PLoS ONE* **2017**, *12*, e0179120. [CrossRef] [PubMed]
65. Zahiri, J.; Emamjomeh, A.; Bagheri, S.; Ivazeh, A.; Mahdevar, G.; Tehrani, H.S.; Mirzaie, M.; Fakheri, B.A.; Mohammad-Noori, M. Protein complex prediction: A survey. *Genomics* **2020**, *112*, 174–183. [CrossRef]
66. Lensink, M.F.; Velankar, S.; Wodak, S.J. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. *Proteins Struct. Funct. Bioinform.* **2017**, *85*, 359–377. [CrossRef]
67. Xenarios, I.; Salwinski, L.; Duan, X.J.; Higney, P.; Kim, S.M.; Eisenberg, D. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **2002**, *30*, 303–305. [CrossRef]
68. Stark, C.; Breitkreutz, B.J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34*, D535–D539. [CrossRef] [PubMed]
69. Ma, C.Y.; Chen, Y.P.P.; Berger, B.; Liao, C.S. Identification of protein complexes by integrating multiple alignment of protein interaction networks. *Bioinformatics* **2017**, *33*, 1681–1688. [CrossRef] [PubMed]
70. Pu, S.; Wong, J.; Turner, B.; Cho, E.; Wodak, S.J. Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.* **2009**, *37*, 825–831. [CrossRef] [PubMed]
71. Mewes, H.W.; Frishman, D.; Mayer, K.F.; Münsterkötter, M.; Noubibou, O.; Pagel, P.; Rattei, T.; Oesterheld, M.; Ruepp, A.; Stümpflen, V. MIPS: Analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.* **2006**, *34*, D169–D172. [CrossRef] [PubMed]
72. Luc, P.V.; Tempst, P. PINdb: A database of nuclear protein complexes from human and yeast. *Bioinformatics* **2004**, *20*, 1413–1415. [CrossRef]
73. Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40*, D109–D114. [CrossRef]
74. Dwight, S.S.; Harris, M.A.; Dolinski, K.; Ball, C.A.; Binkley, G.; Christie, K.R.; Fisk, D.G.; Issel-Tarver, L.; Schroeder, M.; Sherlock, G.; et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* **2002**, *30*, 69–72. [CrossRef]
75. Li, X.; Wu, M.; Kwoh, C.K.; Ng, S.K. Computational approaches for detecting protein complexes from protein interaction networks: A survey. *BMC Genom.* **2010**, *11*, S3. [CrossRef]
76. Li, M.; Chen, J.e.; Wang, J.x.; Hu, B.; Chen, G. Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinform.* **2008**, *9*, 398. [CrossRef]
77. Brohee, S.; Van Helden, J. Evaluation of clustering algorithms for Protein–Protein interaction networks. *BMC Bioinform.* **2006**, *7*, 488. [CrossRef]
78. Li, X.L.; Foo, C.S.; Ng, S.K. Discovering protein complexes in dense reliable neighborhoods of protein interaction networks. In *Computational Systems Bioinformatics: (Volume 6)*; World Scientific: Singapore, 2007; pp. 157–168.
79. Friedel, C.C.; Krumsiek, J.; Zimmer, R. Bootstrapping the interactome: Unsupervised identification of protein complexes in yeast. *J. Comput. Biol.* **2009**, *16*, 971–987. [CrossRef] [PubMed]
80. Maulik, U.; Mukhopadhyay, A.; Bhattacharyya, M.; Kaderali, L.; Brors, B.; Bandyopadhyay, S.; Eils, R. Mining quasi-bicliques from HIV-1-human protein interaction network: A multiobjective biclustering approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2012**, *10*, 423–435. [CrossRef]
81. Cao, B.; Luo, J.; Liang, C.; Wang, S. Identifying protein complexes by combining network topology and biological characteristics. *J. Comput. Theor. Nanosci.* **2016**, *13*, 7666–7675. [CrossRef]
82. Wu, Z.; Liao, Q.; Liu, B. idenPC-MIIP: Identify protein complexes from weighted PPI networks using mutual important interacting partner relation. *Briefings Bioinform.* **2021**, *22*, 1972–1983. [CrossRef] [PubMed]
83. Cherry, J.M.; Adler, C.; Ball, C.; Chervitz, S.A.; Dwight, S.S.; Hester, E.T.; Jia, Y.; Juvik, G.; Roe, T.; Schroeder, M.; et al. SGD: Saccharomyces genome database. *Nucleic Acids Res.* **1998**, *26*, 73–79. [CrossRef] [PubMed]
84. Li, B.; Liao, B. Protein complexes prediction method based on core—Attachment structure and functional annotations. *Int. J. Mol. Sci.* **2017**, *18*, 1910. [CrossRef] [PubMed]
85. Xiao, Q.; Luo, P.; Li, M.; Wang, J.; Wu, F.X. A Novel Core-Attachment–Based Method to Identify Dynamic Protein Complexes Based on Gene Expression Profiles and PPI Networks. *Proteomics* **2019**, *19*, 1800129. [CrossRef] [PubMed]