frontiers in Genetics

METHODS
published: 14 December 2021
doi: 10.3389/fgene.2021.794354

An Improved Memetic Algorithm for Detecting Protein Complexes in Protein Interaction Networks

Rongquan Wang[1], *Huimin Ma*[1]* *and Caixia Wang*[2]

[1]School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, China, [2]School of International Economics, China Foreign Affairs University, Beijing, China

Identifying the protein complexes in protein-protein interaction (PPI) networks is essential for understanding cellular organization and biological processes. To address the high false positive/negative rates of PPI networks and detect protein complexes with multiple topological structures, we developed a novel improved memetic algorithm (IMA). IMA first combines the topological and biological properties to obtain a weighted PPI network with reduced noise. Next, it integrates various clustering results to construct the initial populations. Furthermore, a fitness function is designed based on the five topological properties of the protein complexes. Finally, we describe the rest of our IMA method, which primarily consists of four steps: selection operator, recombination operator, local optimization strategy, and updating the population operator. In particular, IMA is a combination of genetic algorithm and a local optimization strategy, which has a strong global search ability, and searches for local optimal solutions effectively. The experimental results demonstrate that IMA performs much better than the base methods and existing state-of-the-art techniques. The source code and datasets of the IMA can be found at https://github.com/RongquanWang/IMA.

Keywords: protein complexes, protein-protein interaction networks, memetic algorithm, fitness function, graph clustering methods

OPEN ACCESS

Edited by:
Liudmila Sergeevna Mainzer, University of Illinois at Urbana-Champaign, United States

Reviewed by:
Jiawei Luo, Hunan University, China
Weihao Ge, University of Illinois at Urbana-Champaign, United States

***Correspondence:**
Huimin Ma
mhmpub@ustb.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *13 October 2021*
Accepted: *22 November 2021*
Published: *14 December 2021*

Citation:
Wang R, Ma H and Wang C (2021) An Improved Memetic Algorithm for Detecting Protein Complexes in Protein Interaction Networks. Front. Genet. 12:794354. doi: 10.3389/fgene.2021.794354

1 INTRODUCTION

Many complex systems in the real world are often modeled with complex networks, such as computer networks, social networks, and biological networks. The detection of community structure is an essential property of complex networks, and it helps us understand the structure and functionally of complex networks. Protein complexes mined from the PPI network are representative of detecting community structure in complex networks. In proteomics, proteins rarely act alone, and often organize together to form protein complexes to perform specific biological functions cooperatively (Spirin and Mirny, 2003). Therefore, accurately identifying protein complexes from PPI networks can contribute to the study of the mechanisms of cellular functions and organization (Gavin et al., 2002) in the post-genomic era. Although some experimental methods such as yeast two-hybrid and tandem affinity purification can detect protein complexes, they have limitations. Specifically, they are expensive and time-consuming. With the development of high-throughput experimental technologies, many PPI networks are now available. The computational methods developed complement the experimental techniques in identifying protein complexes. As a result, many computational methods have been proposed for the identification of protein complexes from PPI networks, which is a type of cluster analysis,

which consists of grouping patterns into clusters based on the similarity, and it is a valuable technology in many areas such as bioinformatics, machine learning, and computer vision.

To date, a variety of computational methods for detecting protein complexes in PPI networks have been proposed. Based on our work, we provide a summary of the related works by classifying the existing methods into three types: methods based on reducing noise, methods based on different topological structures, and methods based on evolutionary algorithms.

Because the PPI networks are derived from high-throughput experiments, the high false positive and false negative rates in the PPI networks are high (Von Mering et al., 2002; Samanta and Liang, 2003; Liu et al., 2009; Srihari, 2012; Zaki et al., 2013; Lei et al., 2018), which substantially affects the accuracy of protein complex identification. To reduce the influence of noise, some methods utilize various network topological properties for strong interactions and identify protein complexes. These algorithms include PEWCC (Zaki et al., 2013), ProRank+ (Hanna and Zaki, 2014), and EWCA (Wang et al., 2019a). Alternatively, some studies have attempted to integrate the gene expression data (Keretsu and Sarmah, 2016), gene ontology data (Zhang et al., 2013a; Wang et al., 2019b), and subcellular localization data (Wang et al., 2020) to improve the reliability of the interactions. These studies mainly use biological data to weight protein interactions and compensate the PPI network, which better reflects the real protein interactions. DPCT (SabziNezhad and Jalili, 2020) uses TAP and GO data to construct a weighted PPI network and to reduce the noise of PPI, and a memetic algorithm to detect protein complexes. Valdeolivas et al. (2019) extended the Random walk with restart (RWR) algorithm to multiplex and heterogeneous networks. This framework performs better as compared to the aggregation of the different interaction sources and the multiplex framework is more efficient than network aggregations to extract communities from biological networks. Blatti et al. (2020) presented Knowledge Engine for Genomics (KnowEnG), it is a free-to-use computational system for analysis of genomics data sets, it provides the standard clustering pipeline to cluster a collection of samples. However, further reducing the impact of random noise in PPI networks, and improving the performance of protein complex detection methods remain urgent problems to be solved.

In recent years, various computational methods based on different topological structures have been developed to detect protein complexes in PPI networks. Among these methods, different types of topological structures are commonly assumed to be protein complexes. Some methods partition proteins into many non-overlapping clusters by using partition functions or principles. For example, Markov Clustering (MCL) (Van Dongen, 2000) is a simulated random walk method, which mainly uses expansion and inflation operators to manipulate the adjacency matrix and mine protein complexes from the PPI networks. Meanwhile, RNSC (King et al., 2004) first moves the proteins randomly among the clusters to optimize the cost function, and then a post process, based on cluster size, density, and functional homogeneity is carried out. Some methods aim to find cliques. In 2006, CFinder (King et al., 2004) was developed to cluster proteins in the PPI network, and it used the concept of k-clique to discover protein complexes. In 2009, CMC (Liu et al., 2009) tried to enumerate cliques in the PPI network for discovering protein complexes, but it was too strict for most protein complexes. Therefore, several methods based on density have been designed to identify dense subgraphs in the PPI networks, where subgraphs with densities above a predefined threshold were considered as protein complexes. For example, MCODE (Bader and Hogue, 2003) first weighted every node by local neighborhood density and then extended locally dense nodes to detect protein complexes. Later, Li et al. (2008) improved the seed-extended method by modifying the DPClus algorithm based on the diameter and density of the local graph. Furthermore, Gavin et al. (2006) proposed that protein complexes have the core-attachment structure, and some identification methods based on the core-attachment structure. For example, COACH (Wu et al., 2009) and WPNCA (Peng et al., 2014) have been proposed to find protein complexes. They first extracted the protein complex cores from the neighborhood graphs of the proteins, and protein complex cores were further extended to form complete protein complexes. Finally, some protein complexes with a high overlap were merged. There are some variants of network topological features that are used to detect protein complexes; these studies (Nepusz et al., 2012; Giurgiu et al., 2019) have shown that proteins in a protein complex commonly display strong interactions within the core of the protein complex, and weak interactions with the proteins outer surface of the protein complex, ClusterONE (Nepusz et al., 2012) starts from a seed node and inserts neighbors into it to form overlapping protein complexes by using cohesiveness. Subsequently, Wang et al. (2019b) proposed a novel seed-expand algorithm called SE-DMTG to identify protein complexes with a combinatorial function from the weighted PPI networks. Additionally, based on the 3-sigma principle (Wang et al., 2013), MPC-C (Wang et al., 2020) identifies the active points of proteins in a time serial of gene expression data and generates a series of time-sequenced subnetworks to identify static and dynamic protein complexes. In 2021, Liu et al. (2021) proposed a protein complex detection methods based on a semi-supervised model to detect protein complexes with clear module structures. A number of computational methods only consider single topological properties to identify protein complexes, and they recover protein complexes with other types of topological structures.

Intensive studies on evolutionary algorithms have also been conducted. In recent years, some researchers have provided new ideas for solving protein complex identification problems by using optimized algorithms, by employing the characteristics of highly adaptive and good optimization abilities. Some successful methods have been applied to tackle the problems of identifying protein complexes and efficiency. In 2015, Ramadan et al. (2016) introduced a genetic algorithm to detect protein complexes. Subsequently, in 2016, Lei et al. (2016) presented F-MCL based on Markov clustering and the firefly method, which automatically determines the parameters by using the firefly method. In the same year, a novel fruit fly optimization

TABLE 1 | Statistics of used four PPI networks in the study.

Dataset	Number of node	Number of edge	Density
Krogan	2674	7075	0.0019796849348
DIP	4930	17201	0.0014157219124
combined6	3869	17327	0.0023156247135
WI-PHI	5955	49604	0.0027980540426

clustering method was designed to detect dynamic protein complexes (Lei et al., 2016). In 2017, Zhang et al. (2017) proposed a new firefly clustering algorithm for transforming the protein complex detection problem into an optimization problem. Zhao et al. (2017) proposed a novel improved cuckoo search clustering method for discovering protein complexes in dynamic weighted PPI networks. In 2019, Lei et al. (2019b) used a nature-inspired optimization method to detect protein complexes. In 2019, a moth-flame-optimization-based protein complex detection method was presented (Lei et al., 2019a). In 2020, an evolutionary algorithm based on a heuristic biological operator was introduced to detect protein complexes (Abduljabbar et al., 2020). These evolutionary methods have a strong global search ability, but they have difficulty in locating the local optima efficiently.

To solve these issues, we present a novel algorithm, named IMA, which uses an improved memetic algorithm we designed to detect protein complexes from the PPI network. First, we constructed a weighted PPI network by using the common neighbor, gene expression data, GO-slim data, and subcellular location data to reduce the impact of noise on our IMA. Second, many high-quality initial individuals, including protein complexes with different topological structures, are generated using EWCA (Wang et al., 2019a), SE-DMTG (Wang et al., 2019b), and MPC-C (Wang et al., 2020). We propose a fitness function to identify protein complexes with various topological properties. Third, a new improved memetic algorithm is proposed to mine the protein complexes by optimizing this fitness function in the weighted PPI network. Remarkably, its selection, recombination, and updating population operators are used for the global search of the best individual, and a local optimization strategy is designed to locate the local optima individually. Finally, our IMA was applied to four different yeast PPI networks and compared with the 12 existing excellent methods. The experimental results illustrate that the IMA achieves state-of-the-art performance of computational metrics and biological relevance metrics in identifying protein complexes.

In the materials and methods section, we introduce the datasets and standard protein complexes used in the evaluation of IMA and define all phases of IMA separately. In the experiments and results section, we evaluated the proposed method and compare it with the state-of-the-art methods. The case study and discussion section shows some examples of protein complexes detected by IMA and we conclude this paper in conclusion section.

2 MATERIALS AND METHODS

2.1 Datasets

In this study, we used four PPI networks including Krogan (Krogan et al., 2006), DIP (Xenarios et al., 2002), combined6 (Liu et al., 2009), and WI-PHI (Kiemer et al., 2007). Their information is presented in **Table 1**.

Furthermore, two sets of standard protein complexes from the literature (Wang et al., 2020) were used to evaluate the performance of the protein complex detection methods, and their information is shown in **Table 2**. Here, standard protein complexes 1 consists of the known protein complexes from MIPS (Mewes et al., 2004), SGD (Hong et al., 2007), TAP06 (Gavin et al., 2006), ALOY (Aloy et al., 2004), CYC 2008 (Pu et al., 2009) and NEWMIPS (Friedel et al., 2009). Meanwhile, standard protein complexes 2 also is a combined protein complexes dataset (Ma et al., 2017). It consists of the Wodak database (Pu et al., 2009), PINdb and GO complexes (Ma et al., 2017). Additionally, we also used CYC2008 protein complexes (Pu et al., 2009) and MIPS protein complexes (Mewes et al., 2004), and they come from other people's work, and they are shown in https://github.com/RongquanWang/IMA/Additional file 3.

The GO-slim data can explain the biological function of proteins, and it can be downloaded from https://downloads.yeastgenome.org/curation/literature/go_slim_mapping.tab. The gene expression data were obtained from https://www.ncbi.nlm.nih.gov/sites/GDSbrowser. Additionally, the subcellular localization data set for yeast proteins was obtained from https://compartments.jensenlab.org/Downloads.

2.2 Methods

2.2.1 Preliminaries

Since PPI networks are defined using graph-theoretic concepts, we first provide some of the terminologies used in our paper and then describe the IMA method in detail.

A PPI network can be represented as an undirected graph $G(V, E)$, where V is the set of vertices (individual proteins) and E is the set of edges (protein interactions) between the vertices. The neighbors of v in G, denoted by $N(v)$, are the set of vertices adjacent to v.

Biologically, protein complexes are groups of proteins that interact with each other at the same time and place, forming a single multi-molecular machine. However, due to the inherent topological structures of protein complexes (Nepusz et al., 2012) in PPI networks, protein complexes are usually assumed to be the subgraphs of PPI networks. Let $C = (V_C, E_C, W_C)$ be a subgraph of G. The neighbors of C are defined by **Eq.1**:

$$N(C) = \{v | (v, u) \in E, u \in V_C, v \in V - V_C\}, \quad (1)$$

As a result, the task of identifying protein complexes can be formulated as mining connected clusters that are densely connected inside and well separated from the rest of the PPI

TABLE 2 | Statistics of used standard protein complexes.

Datasets	Number	Protein coverage	Avg size
Standard protein complexes 1	812	2773	8.92
Standard protein complexes 2	1045	2778	8.97
CYC2008 protein complexes	193	1371	8.33
MIPS protein complexes	212	1202	15.61

networks, and the clusters correspond to the protein complexes. The protein complex detection method obtains a set of clusters, $P = (P_1, P_2, \ldots, P_t)$.

2.2.2 Methods

2.2.2.1 Constructing a Weighted PPI Network

Many studies (Lei et al., 2018, 2019b; Wang et al., 2020) have shown that the performance of protein complex identification methods can be improved by integrating multiple data sources into a single weighted PPI network, which enhances the confidence of interactions in PPI networks. Thus, we integrated the topological structures and multiple biological data to weight the interactions in the PPI networks. Our goal was to weight the edges of the PPI network to reflect the reliability of the corresponding interactions. Graph clustering algorithms can use these weights to reduce the influence of noisy edges and yield meaningful clusters.

2.2.2.1.1 Protein Common Neighbor similarity. The larger the number of common neighbors between the two proteins, the stronger the interaction of the two proteins, and they are more likely to participate in the same protein complex. Some common neighbor similarity measures (Zaki et al., 2013; Wang et al., 2019a; Wang et al., 2019b) were used to calculate the similarity of protein pairs. This paper defines a higher-order common neighbor (HCN) similarity measure to estimate the reliability of the interaction between two proteins, v and u. The HCN is defined by **Eq. 2**.

$$HCN(v,u) = \sqrt{\frac{|NCN(v,u)|^2}{|N(v)|*|N(u)|*|N(v) \cup N(v)|}}, \quad (2)$$

where $NCN(v, u) = N(v) \cap N(u)$ is the number of common neighbors between proteins v and u. $N(v)$ and $N(u)$ represent the number of neighbors that proteins v and u are connected with, respectively. $HCN(v, u)$ can further balance the comprehensive connectivity of the two interacting proteins, which may consist of the same protein complex.

2.2.2.1.2 Protein Co-Expression similarity. Next, the gene expression data describes proteins under various conditions in a biological process (Zhang et al., 2013b; Wang et al., 2014). The gene expression vector of each protein comprised of a series of expression values within the period. If two proteins have a similar degree of expression at the same time interval, they have a high co-expression value, and then they are more likely to form a protein complex. The gene expression profiles of a pair of proteins v and u in a PPI network, their gene expression profiles are $v = \{v_1, v_2, \ldots, v_n\}$ and $u = \{u_1, u_2, \ldots, u_n\}$, respectively. Here, we use the person correlation coefficient (Wang et al., 2013) to calculate the co-expression value $PCC(v, u)$, as defined in **Eq. 3**:

$$PCC(v,u) = \frac{\sum_{i=1}^{n}(v_i - \bar{v})*(u_i - \bar{u})}{\sqrt{\sum_{i=1}^{n}(v_i - \bar{v})^2}*\sqrt{\sum_{i=1}^{n}(u_i - \bar{u})^2}}, \quad (3)$$

where $\bar{v}$ and $\bar{u}$ are the average gene expressions of proteins v and u in n time points, respectively. $PCC(v, u)$ indicates the co-expression of the vector representation between two interacting proteins. As the value of $PCC(v, u)$ ranges from −1 to 1, we set $PCC(v,u) = \frac{PCC(v,u)+1.0}{2}$ to set $PCC(v, u)$ in [0,1]. The higher the value of $PCC(v, u)$, the larger the probability of co-expression of proteins v and u and formation of a protein complex.

2.2.2.1.3 Protein Functional Similarity. From the perspective of protein function, we used GO-slim data to reflect the functional similarity of proteins. Moreover, we generated an attribute matrix $O \in R^{N\times M}$, where N denotes the number of proteins in the PPI network and M denotes the number of GO slim attributes. Based on matrix O, we constructed a protein attribute affinity matrix, $S \in R^{N\times N}$. Each entry $FS(v, u)$ reflects the GO slim attribute similarity between proteins v and u, as defined in **Eq. 4**:

$$FS(v,u) = \frac{\sum_{k=1}^{M} o_{vk} * o_{uk}}{\sqrt{\sum_{k=1}^{M} o_{vk}^2} * \sqrt{\sum_{k=1}^{M} o_{uk}^2}}. \quad (4)$$

2.2.2.1.4 Protein Subcellular Location Similarity. Generally, if two interacting proteins have the same subcellular location, the interaction between the proteins is more reliable. Proteins in the protein complex should be localized in the same inner cellular compartment. Here, we defined the subcellular location similarity $SL(v, u)$, and is defined in **Eq. 5**:

$$SL(v,u) = \frac{|SL(v) \cap SL(u)|^2}{|SL(v)|*|SL(u)|} \quad (5)$$

where $|SL(v)|$ and $|SL(u)|$ denote the number of subcellular localizations of proteins v and u, respectively. $|SL(v) \cap SL(u)|$ represents the number of common subcellular localization attributes between proteins v and u.

The edges whose weight is 0 are seen as noise and are removed from the PPI network, and the rest of edges whose weight $W(v, u)$ are expressed by **Eq. 6**:

$$W(v,u) = \frac{HCN(v,u) + PCC(v,u) + FS(v,u) + SL(v,u)}{4}. \quad (6)$$

Finally, the weighted PPI networks were constructed, and the reliability of the PPI networks was enhanced.

2.2.2.2 Memetic Algorithm

A memetic algorithm (Li et al., 2014) is typically a hybrid-local heuristic search method used for optimization. Generally, memetic algorithms consist of a genetic algorithm and local optimization strategy. Here, the genetic algorithm is the global search method, which can explore a reliable estimate of the global optimum, but it does not obtain an optimal individual in the explored search space quickly. Therefore, the local optimal search strategy is typically used to accelerate searching and find the best individual in the local search space. In this paper, we present an improved memetic algorithm that can create new individuals that are located in new promising regions in the global search space

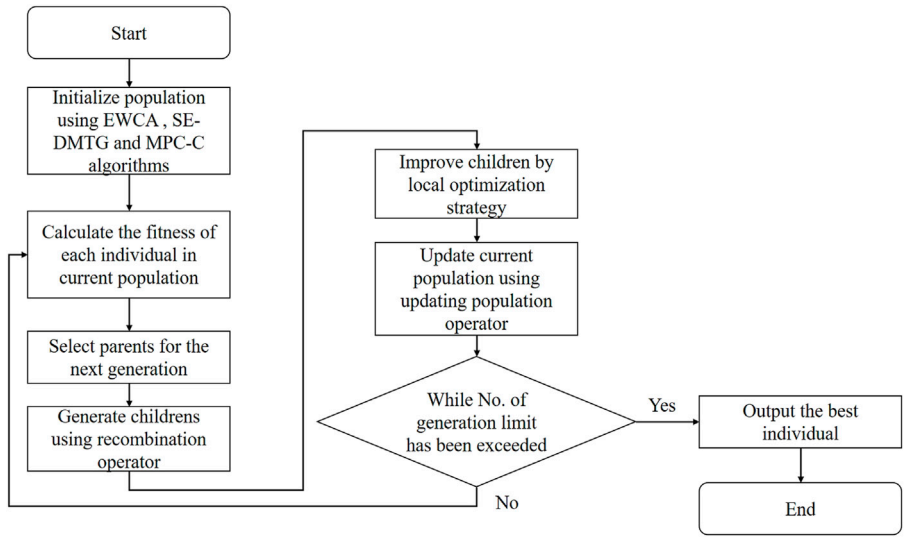

FIGURE 1 | The flowchart of IMA framework.

and search around the newly generated individuals to optimize individuals of better quality (Gach and Hao, 2012).

2.2.2.3 The IMA Algorithm for Complex Detection

2.2.2.3.1 The Framework of IMA Algorithm. The memetic algorithm is a valuable framework for dealing with combinatorial optimization problems. It provides a good balance between searching for diversification and optimization by employing a genetic algorithm and the local optimization strategy (Žalik and Žalik, 2018). We designed a fitness function to evaluate the quality of the clusters. In addition, we propose a recombination operator and local optimization strategy for detecting protein complexes in the PPI network. Our goal is to maximize the individual's $FS_{fitness}$ (see **Eq. 13**) using an improved memetic algorithm. Algorithm 1 is the main framework of IMA. The primary input of the IMA algorithm was the PPI network. A flowchart of Algorithm 1 is presented in **Figure 1**.

IMA begins with initial populations (line 1, see section generating the initial population) and then repeats an iterative procedure many times (generations) (lines 3–18). Two parent individuals in the current population are selected using a tournament selection strategy for each generation (line 4, see section selection operator). Two new children individuals were generated based on two selected parents using the recombination operator (line 5, see section recombination operator). Then, we select the children with the highest fitness function (line 6, see section fitness function: **Eq. 13**), and if its fitness function is larger than the 80% fitness of the individual with the maximum fitness, it is further improved using the local optimization strategy we proposed (line 7–11, see section local optimization strategy). Finally, a roulette wheel selection strategy is used to update the population (line 12, see section updating population operator). The individual with the highest fitness function (line 13, **Eq. 13**) mined during the search process is always recorded (lines 14–16). The entire IMA stops when *iter* consecutive generations (line 3, Max_{Iter}). In the following subsections, we provide more details on the parts of the IMA algorithm.

2.2.2.3.2 Generating Initial Population. As we know, most of existing protein complex detection methods are based on unsupervised learning, and they can only identify protein complexes with a single topology. However, real protein complexes have a variety of topologies. That is why we proposed the IMA. In fact, IMA algorithm is a typical swarm intelligence optimization algorithm, which needs to build an initial population *P*. In order to build an initial population, we should follow two basic requirements: 1) The individuals in the initial population should be high quality. It means that the protein complex detection method should produce the high accuracy initial population; 2) Individuals in the initial population need to be diverse. Obtaining initial population method should identify the protein complexes with different topological structures. Based on the above two points, we choose EWCA (Wang et al., 2019a), SE-DMTG (Wang et al., 2019b) and MPC-C (Wang et al., 2020) as the methods to generate the initial population.

In fact, according to their references, we can see that these three algorithms not only have high protein complex detection accuracy, but also can identify the protein complexes with different topological structures. And their advantageous than other existing methods are shown in their literatures (Wang et al., 2019a; b, 2020). For example, EWCA can identify protein complexes with core-attachment structure, SE-DMTG can identify protein complexes with high density and modularity and MPC-C can predict static protein complexes and dynamic protein complexes with community structure. Moreover, the identification accuracy of these three algorithms is also excellent in the existing algorithms. Therefore, they can ensure the generation of protein complexes with high quality and

different topological structures. That is why we chose these algorithms.

The first approach is the EWCA, which generates *o* individuals by using different values of *ss* (structural similarity) threshold and ranges from 0.3 to 0.62 with 0.04 increment. The second approach is the SE-DMTG algorithm, which generates *o* individuals by randomly constructing seed queues, and then uses its seed-extended algorithm to generate individuals. The last method is the MPC-C algorithm, which is used to construct *o* individuals by setting different values of the filter $Score(C)$ cutoff (from 0.1 to 0.8 with a step size of 0.1). Finally, we obtain $Pop = 3 \times o$ individuals and combine them to create the initial population, where the parameter *o* is 8.

Algorithm 1. The pseudo-code of general framework of IMA.

Input: A static weighted PPI network, i.e., $G = (V, E, W)$,$Max_{Iter} = 60$,$Pop = 24$, EWCA, SE-DMTG, and MPC-C algorithms.
Output: A history best individual, $Ahbi$.
Initialize Population $P = \{P_1, P_2, ..., P_{Pop}\}$ by using EWCA, SE-DMTG and MPC-C algorithms; /* Generating initial population */
Initialize $iter = 0$, the history best individual $Ahbi = \varnothing$;
while $iter \leqslant Max_{Iter}$ **do**
  Select two parents from P using tournament selection strategy, ie. $parent_1, parent_2$;/* selection operator */
  $Childrens$ = Recombination operator($parent_1, parent_2$);/* Recombination operator generate a new offspring */
  Search the $Childrens_{max}$ with the highest fitness function from $Childrens$;
  if $FS_{fitness}(Childrens_{max}) \geq FS_{fitness}(P_{best}) * 0.8$ **then**
    $Children_{opt}$ =Local optimization strategy($Childrens_{max}$); /* Optimal individual offspring */
  else
    $Children_{opt} = Childrens_{max}$
  end if
  P = Updating population strategy(P, $Children_{opt}$);/* Regeneration population */
  Find the individual P_{best} with having maximum fitness in current population P;
  if $FS_{fitness}(P_{best}) \geq FS_{fitness}(Ahbi)$ **then**
    $Ahbi = P_{best}$;
  end if
  $iter = iter + 1$;
end while
return Output the history best individual, $Ahbi$.

2.2.2.3.3 Fitness Function. We first define the fitness function of the IMA method; this fitness function should reflect the topological properties of protein complexes in PPI networks. Generally, a high-quality protein complex is a group of proteins that are densely interconnected but only sparsely connected with the rest of the PPI network (Li et al., 2008; Nepusz et al., 2012; Wang et al., 2019b, 2020). Meanwhile, in identifying various topological properties of protein complexes, the combination of multiple single objective functions can compensate for the shortcomings of a single objective function, which leads to an improved quality of the identified protein complexes. Therefore, we propose a multi-objective function (**Eq. 12**) by integrating five objective functions (**Eqs 7–11** to describe the topological properties of protein complexes: *cohesiveness*(*C*), *density*(*C*), *AIEW*(*C*), *ABEW*(*C*), and *AWM*(*C*).

$C = (V_C, E_C, W_C)$, where V_C is the set of proteins in cluster *C*, E_C is the set of interactions in cluster *C*, and W_C is the set of weights between the pair of proteins. According to previous studies (Nepusz et al., 2012; Wang et al., 2019b), the cohesiveness score is defined as **Eq. 7**:

$$cohesiveness(C) = \frac{W_{in}(C)}{W_{in}(C) + W_{out}(C)}, \tag{7}$$

where $W_{in}(C)$ is the sum of the weights of all edges among *C*, and $W_{out}(C)$ is the sum of the weights of the edges connecting nodes in *C* to other nodes in the rest of the PPI network.

(Li et al., 2008; Liu et al., 2009; Wang et al., 2019b) hypothesized that the higher the density of the cluster is, the more likely the cluster is a protein complex in the PPI network. The weighted density of cluster *C* is defined by **Eq. 8**:

$$density(C) = \frac{2 * W_{in}(C)}{|V_C| * (|V_C| - 1)} \tag{8}$$

where $W_{in}(C)$ is the sum of the weights of the edges between them, and V_C is the number of nodes in *C*.

In this paper, we propose three measures to estimate the likelihood of a local cluster *C* being a protein complex. First, it is the average inner edge weight (AIEW), and it can estimate the reliability of the internal edges of the cluster *C*. This is defined in **Eq. 9**:

$$AIEW(C) = \frac{W_{in}(C)}{|E_C|}, \tag{9}$$

where $W_{in}(C)$ is the sum of the weights of the edges among them, and $|E_C|$ is the number of edges in cluster *C*. $AIEW(C)$ is the average weight of the inner edges in cluster *C*.

Second, it is the average border edge weight (ABEW), and it can measure the reliability of the border edges of cluster *C*. This is defined in **Eq. 10**:

$$ABEW(C) = \frac{W_{out}(C)}{|BE_C|}, \tag{10}$$

where $|BE_C| = \{(u, v)|u \in C, v \notin C\}$ is the number of border edges that connect the cluster *C* with the rest of the PPI network, and $W_{out}(C)$ is the sum of the weights of the edges connecting nodes in *C* to the neighbor of the cluster *C*. $ABEW(C)$ is the average weight of the border edges in cluster *C*.

Third, it is the average weighted modularity (AWM), which indicates that the cluster *C* is highly average weight connected among them and has a low average weight interaction with the rest of the network. This is defined in **Eq. 11**:

$$AWM(C) = \frac{AIEW(C)}{AIEW(C) + ABEW(C)}, \tag{11}$$

Based on these objective functions, we propose a fitness function that combines these single objective functions to assess the possibility of a cluster *C* being a protein complex. This fitness function is denoted by **Eq. 12**:

$$FF(C) = cohesiveness(C) + density(C) + AIEW(C) - ABEW(C) + AWM(C), \tag{12}$$

here, the *density*(*C*) and *AIEW*(*C*) seek a dense intra-connection topological structure, whereas *ABEW*(*C*) identifies the sparse topological structure inter-connecting with the rest of the PPI network. *cohesiveness*(*C*) and *AWM*(*C*) are used to identify the topological structures with densely interconnected nodes that are sparsely connected to the rest of the PPI network.

As a result, $FF(C_i)$ is used to identify protein complexes with various topological structures in the individual *FS*. Finally, the fitness function ($FS_{fitness}$) of the individual is the sum of $FF(C_i)$, which is defined in **Eq. 13**:

$$FS_{fitness} = \sum_{i=1}^{k} FF(C_i), \tag{13}$$

where C_i represents the ith cluster in the individual *FS*. $FF(C_i)$ is a multi-objective function that is designed to capture the community structure of protein complexes (a group of nodes has better internal connectivity than external connectivity). k is the total number of protein complexes found in an individual *FS*.

In this study, the goal of our IMA method is to find the individual with the maximum value of $FS_{fitness}$. Generally, the higher the $FS_{fitness}$ of an individual *FS*, and the better the quality of the individual. Therefore, the protein complex detection problem can be regarded as an optimization problem by maximizing the value of $FS_{fitness}$ in a PPI network.

2.2.2.3.4 Selection Operator. The selection operator is an essential operation used in the memetic algorithm. The main idea of the selection strategy is that the better an individual, the higher is its chance of being a parent. In the IMA algorithm, the recombination operator and updating population operator need individual selection strategies.

In the parent selection operator, a selection strategy is required to select two parent individuals from the current population to generate individual children. The fundamental principle of this operation is that individuals with a higher fitness function are more likely to be selected as parent individuals. Here, we use a binary tournament selection strategy to select the parent individuals. Meanwhile, the recombination operator also uses a binary tournament selection strategy to select protein complexes from composite parent individuals to create individual children.

Lastly, the population updating strategy needs to update the population according to the fitness of individuals in current population and generate children individuals. In this updating process, we use the roulette wheel selection strategy to update the current population to balance the relatively good individuals and avoid precocity.

2.2.2.3.5 Recombination Operator. The recombination operator is the critical diversification mechanism of memetic algorithm. An effective recombination operator should generate not only diversified solutions but also transfer significant components from parents to children (Hao, 2012; Li et al., 2014). The recombination operator is responsible for combining the genetic material of several individuals (usually two individuals) to create a new children (Spears and De Jong, 1995). New children inherit many high-quality protein complexes from their parents. Additionally, the recombination operator plays a vital role in the effectiveness of the memetic algorithm in the global search space. Traditional recombination operators, having uniform crossover and two-point crossover, are challenging to convey the excellent protein complexes of two parents to a new children simultaneously. This method is less suitable for protein complex detection. Therefore, we present a recombination operator based on the fitness function, whose children can inherit the better protein complexes of their parents. This operator plays an essential role in preventing the algorithm from being trapped in an optimal local solution and exploring the global search space. The main idea of this operator is to take the protein complexes from two parents as the genetic material and try to retain the high-quality protein complexes in parents for the new children. The recombination operator is described in Algorithm 2.

Algorithm 2. Recombination operator.

Input: A weighted PPI network, i.e., $G = (V, E, W)$, Two parents $Parent1$, $Parent2$.
Output: The two new offsprings, $Childrens$.
1: **Initialize** $Children = \varnothing$;
2: Combining two parents, $parent1$, $parent2$ into a composite parent, $compositeparent$;
3: $len_{low} = min(len(parent1), len(parent2))$;
4: $len_{up} = max(len(parent1), len(parent2))$;
5: $Children_{len}$ = Randomly generating a number between len_{low} and len_{up}, and it is used to determine the number of protein complexes in the new children;
6: We calculate the FF (Equation (12)) of each protein complex in $compositeparent$, and sort them based on their score, $compositeparent_{score}$;
7: According to $compositeparent_{score}$, $compositeparent$, we use binary tournament selection strategy to create two new offsprings, $Childrens$;
8: **return** The two new offsprings, $Childrens$.

Let *parent*1 and *parent*2 represent the parents, and let *len*(*parent*1) and *len*(*parent*2) be the number of protein complexes in each parent. Parents *parent*1 and *parent*2 are merged into a composite parent. Note that if there are redundant protein complexes, and we only leave one. As a result, we obtain a composite parent individual *compositeparent* in line 2. Based on the length of *parent*1 (*len*(*parent*1)) and the length of *parent*2 (*len*(*parent*2)), we determine the length of the new children ($children_{len}$) by randomly generating a number between len_{low} and len_{up} in lines 3–5. The multi-objective function (**Eq. 12**) of each protein complex in the *compositeparent* individual is calculated and sorted in line 6. Next, according to their multi-objective function (**Eq. 12**) and *compositeparent*, we use the binary tournament selection strategy to create two new offspring in line 7.

2.2.2.3.6 Local Optimization strategy. To improve the quality of the generated offspring, we presented a new local optimization strategy to obtain better offspring. This strategy is different from the general sense of local search strategies, such as the hill-climbing strategy and simulated annealing strategy. The purpose of the local optimization strategy is to obtain offspring of relatively high quality. Here, each protein complex in children is optimized using a multi-objective function (**Eq. 12**) and a local optimization strategy. The local optimization strategy of the IMA is applied to new children as shown in Algorithm 3.

In this process, for each protein complex $Childrens_i$ in the $Childrens_{\max}$, we optimize it using the following steps in lines 3–22. First, we find inner nodes (*Innernodes*) that belong to the $Childrens_i$ and connect at least one protein in the rest of the PPI networks in line 9. Then, we find the *inner_max* by improving the multi-objective function (see **Eq. 12**) maximum in *Innernodes* in line 10, and then we remove the *inner_max* from *Clusterdel* in line 11. Second, we find boundary nodes (*Boundarynodes*), which is the set of proteins that connect at least one inner protein of the current $Childrens_i$ in line 15. Then, we detect the $boundary_{\max}$ by increasing the multi-objective function (see **Eq. 12**) maximum in line 16, and we insert the $boundary_{\max}$ into *Clusteradd* in line 17. We repeat the above two steps until the protein complex $Childrens_i$ does no change, and if it is not changed (it is

considered a locally optimal cluster) or *iteration* > 20 in lines 19–22, it is an identified protein complex $Opcs_i$ in line 24. Next, we use the local optimization strategy to optimize the rest of the protein complexes in $Childrens_{\max}$ in line 2–25. If the fitness function of the optimized individual *Opcs* is not larger than that of the $Childrens_{\max}$, for the individual $Childrens_{\max}$, its local optimization strategy is deemed invalid in lines 26–29. Finally, we output this optimal child, *Opcs*, in line 30.

Algorithm 3. The local optimization strategy.

Input: A weighted PPI network, i.e., $G = (V, E, W)$, $Childrens_{max}$.
Output: A set of optimized protein complexes, $Opcs$.
1: **Initialize** $Opcs = \varnothing$,i = 1;
2: **for** each protein complex $Childrens_i \in Childrens_{max}$ **do**
3: **Initialize** $mark$ = 1,$iteration$ = 0;
4: **Initialize** current cluster, $CC = Childrens_i$;
5: **while** $mark$ **do**
6: **Step 1**: Removing inner nodes process:
7: **Initialize** $Clusterdel = CC$;
8: **if** $len(Clusterdel) \geq 3$ **then**
9: Find its inner nodes based on $Clusterdel$, i.e., $Innernodes$;
10: Search the inner node with improving multi-objective function (Equation (12)) maximum in $Innernodes$,$inner\_max$;
11: Remove the $inner\_max$ from $Clusterdel$;
12: **end if**
13: **Step 2**: Adding boundary nodes process:
14: **Initialize** $Clusteradd = Clusterdel$;
15: Find its boundary nodes based on $Clusteradd$, i.e., $Boundarynodes$;
16: Search the boundary node with increasing multi-objective function (Equation (12)) maximum in $Boundarynodes$,$boundary\_max$;
17: Insert the $boundary\_max$ into $Clusteradd$;
18: $iteration = iteration + 1$
19: **if** $CC == Clusteradd$ or $iteration > 20$ **then**
20: $CC = Clusteradd$;
21: $mark$ = 0;
22: **end if**
23: **end while**
24: $Opcs_i = CC, i = i + 1$;
25: **end for**
26: Calculate the fitness function of the $Opcs$, $fitness_{Opcs}$ and the $Childrens_{max}$, $fitness_{Childrens_{max}}$;
27: **if** $fitness_{Childrens_{max}} > fitness_{Opcs}$ **then**
28: $Opcs = Childrens_{max}$;
29: **end if**
30: **return** Output this optimal offspring, $Opcs$.

2.2.2.3.7 Updating Population Operator. Population diversity is also a vital issue for memetic algorithms to effectively avoid prematurity. When a new child individual is produced with the recombination operator and local optimization strategy, the fitness function (**Eq. 13**) of the new children and the current population are calculated, respectively. All of these are sorted by their fitness functions. Moreover, to avoid premature convergence, we employ a roulette wheel selection strategy to update the population. Here, the size of the new population in each iteration is the same as that of the original population. The roulette wheel selection strategy can balance the diversity of the population and guarantee that individuals with higher fitness always have a greater probability of being retained in the population.

3 EXPERIMENTS AND RESULTS

In the experiment, our operational environment was a windows 10 operating system with an Intel(R) Core(TM) i7-9700 CPU with a physical memory of 16 GB, and a speed of the processor was 3.60 GHz. The IMA was run on PyCharm Community Edition 2017.2.2. The IMA was implemented using the Python 3.

3.1 Evaluation Metrics

There are several statistical matching-based metrics that estimate the quality of the detected protein complexes based on different protein complex detection methods. Meanwhile, biological relevance-based metrics, which are supplementary to statistical matching-based metrics, are used to evaluate the biological significance of identified protein complexes.

If a detected protein complex *ipc* and a known protein complex *kpc* contain common proteins each other, their overlapping score ($OS(ipc, kpc)$) is calculated using **Eq. 14**:

$$OS(ipc, kpc) = \frac{|V_{ipc} \cap V_{kpc}|^2}{|V_{ipc}| \times |V_{kpc}|}, \tag{14}$$

where *ipc* and *kpc* are the protein set of *ipc* and the protein set of *kpc*, respectively. If $OS(ipc, kpc) \geq \lambda$, *ipc* is matched with *kpc*, where λ is a threshold.

The F-measure is the harmonic mean of precision and recall, and it can be calculated using **Eq. 15**:

$$F - measure = \frac{2 \times precision \times recall}{precision + recall}, \tag{15}$$

For more details, please see (Lei et al., 2019a). An identified protein complex is considered to match a standard protein complex where the overlap score $OS(ipc, kpc)$ is larger than 0.2 (Lei et al., 2019a).

The coverage rate (CR) was used to measure the number of proteins in the standard protein complexes that could be covered by the identified protein complexes (Peng et al., 2014). This is defined in **Eq. 16**:

$$CR = \frac{\sum_{s=1}^{|S|} max\{T_{st}\}}{\sum_{s=1}^{|S|} N_s}, \tag{16}$$

For more details on these parameters, please refer to reference (Peng et al., 2014).

Generally, a higher *Sn* indicates that the identified protein complexes cover the proteins in the standard protein complexes better. In contrast, a higher *PPV* indicates that the identified protein complexes are more likely to be actual protein complexes. Accuracy (ACC) is the geometric average of *PPV* and *Sn*, which is denoted by **Eq. 17**:

$$ACC = \sqrt{Sn \times PPV}. \tag{17}$$

The maximum matching ratio (MMR) (Nepusz et al., 2012) can measure the overlap matching between standard protein complexes and detected protein complexes based on maximal one-to-one mapping. It can deal with the case that a known protein complex split into different parts in the identified protein complexes, because only one part is matched with the known protein complex.

Jaccard (Wang et al., 2019b) was used to quantify the overlap between the detected protein complexes and known protein complexes. In fact, *Jaccard* is defined as the harmonic mean of the *JaccardC* of the identified protein complexes and the *JaccardG* of standard protein complexes, and it is used to evaluate the clustering results. *Jaccard* is calculated using **Eq. 18**:

$$Jaccard = \frac{2 \times (JaccardC \times JaccardG)}{JaccardC + JaccardG}. \tag{18}$$

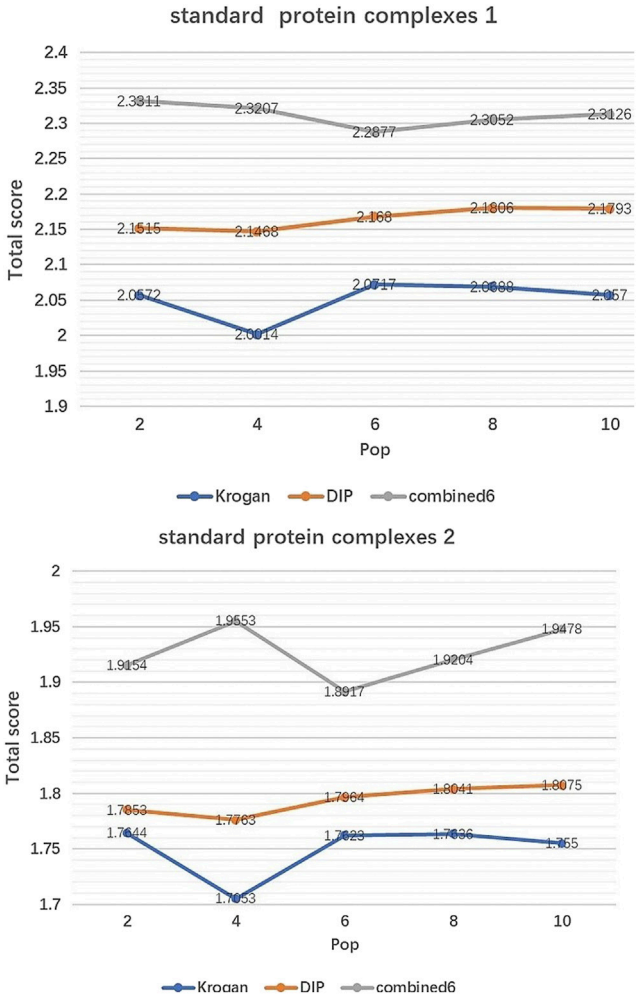

FIGURE 2 | The effect of Pop. It shows how the variation of parameter Pop affect the Total score of IMA method in standard protein complexes 1 and standard protein complexes 2.

As a result, the performance of the detection method is evaluated by the total score, which is calculated using **Eq. 19** (Wang et al., 2020):

$$Total\ score = F - measure + CR + ACC + MMR + Jaccard. \tag{19}$$

In this paper, the *p*-value is used to estimate the biological relevance of the identified protein complexes, and it is denoted by **Eq. 20**:

$$p - value = 1 - \sum_{k=0}^{K-1} \frac{\binom{F}{k}\binom{N-F}{C-k}}{\binom{N}{C}}, \tag{20}$$

For a more detailed explanation of these parameters, please refer to references (Lei et al., 2019a; Wang et al., 2020). If the *p*-value of the protein complex is less than 0.01, it means that the protein complex has biological significance.

The co-localization score (*CL*) (Krumsiek et al., 2008) is denoted as the maximal fraction of proteins in a protein complex that is found at the same location (Friedel et al., 2009). For all the detected protein complexes by different methods, the average co-localization score was computed using **Eq. 21**:

$$CL = \frac{\sum_{j=1}^{m} max_{i=1}^{n} l_{i,j}}{\sum_{j=1}^{m} N_j}, \tag{21}$$

where $l_{i,j}$ is the number of proteins in the detected protein complex j allocated to the localization group i, N_j is the number of proteins in the detected protein complex j, and m and n are the number of detected protein complexes and localization groups, respectively. The final localization score was calculated as the geometric mean of the co-localization scores based on the Huh datasets (Huh et al., 2003).

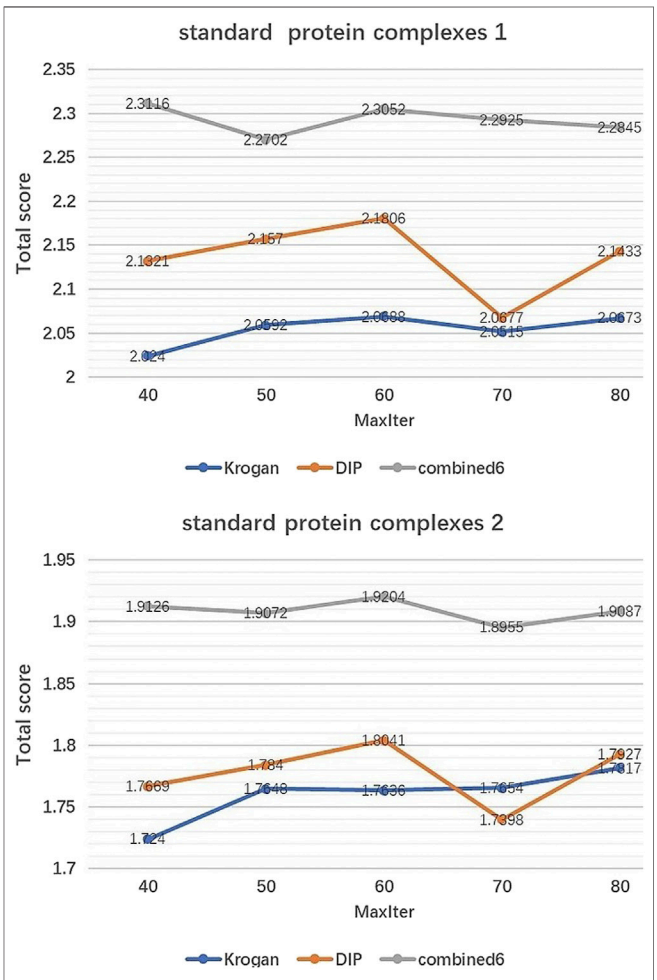

FIGURE 3 | The effect of MaxIter. It shows how the variation of parameter MaxIter affect the Total score of IMA method in standard protein complexes 1 and standard protein complexes 2.

TABLE 3 | Parameters of each method used in the study.

ID	Algorithms	Parameter
1	MCODE	(default setting)
2	MCL	inflation = 2 (author suggestions)
3	IPCA	S = 3, p = 2,T_{in} = 0.6 (author suggestions)
4	COACH	w = 0.225 (default setting)
5	ClusterONE	Density = auto, Overlap threshold = 0.8 (author suggestions)
6	PEWCC	Overlap = 0.8,-r = 0.1,Re-join = 0.3 (author suggestions)
7	ProRank+	AdjstCD threshold = 0.45 (author suggestions)
8	WPNCA	lambda = 0.3,size = 3 (author suggestions)
9	WEC	Balance factor (λ) = 0.8, Edge weight (T_w) = 0.7, Enrichment (T_e) = 0.8, Filtering (T_f) = 0.9 (author suggestions)
10	EWCA	Structural similarity (ss = 0.4) (author suggestions)
11	SE-DMTG	Size = 3 (author suggestions)
12	MPC-C	Overlap threshold = 0.8 (author suggestions)
13	IMA	Generations (Max_{Iter}) = 60, Population size (Pop) = 24 (default setting)

3.2 The Effects of Parameters

IMA includes two parameters which need to be tuned: Generations(MaxIter) and Population size (Pop). Pop controls the number of initial population by EWCA, SE-DMTG, MPC-C. MaxIter determines the number of iterations of population optimization. **Figures 2**, **3** show how performance is influenced by these parameters in three PPI networks and two standard protein complexes.

Effect of Pop. **Figure 2** shows how Total score changes with the value of Pop. We can see that Total score is not very sensitive to Pop, especially when Pop falls in [6, 10]. So in our experiments, Pop is set to 8 by default.

Effect of MaxIter. **Figure 3** shows the changing trend of Total score when MaxIter increases from 40 to 80. We can see that for the three PINs, the fluctuations of Total score are not significant. In our experiments, we set MaxIter = 60 by default.

According to **Figures 2**, **3**, we can see that the Total score is not very sensitive to the changing trend of MaxIter and Pop, so we only set MaxIter = 60 and Pop = 8 as the default value. To avoid evaluation bias and overestimation of the performance, we do not tune the parameter to a particular dataset and set them as the default value in the all experiments. For more details on parameters setting, please see https://github.com/RongquanWang/IMA/Additional file 1.rar.

3.2.1 Comparison With Competitive Algorithms

To demonstrate the performance of the IMA, we compared it with 12 state-of-the-art methods. These include MCODE (Bader and Hogue, 2003), MCL (Van Dongen, 2000), IPCA (Li et al., 2008), COACH (Wu et al., 2009), ClusterONE (Nepusz et al., 2012), PEWCC (Zaki et al., 2013), ProRank+ (Hanna and Zaki, 2014), WPNCA (Peng et al., 2014), WEC (Keretsu and Sarmah, 2016), EWCA (Wang et al., 2019a), SE-DMTG (Wang et al., 2019b), and MPC-C (Wang et al., 2020). The total score is used for a more comprehensive evaluation of the different methods. We obtained the software implementations for all the compared methods, and their parameters are shown in **Table 3**. Although better results could probably be obtained by fine-tuning these parameters, we only use default or suggestion thresholds to maintain the fairness of different algorithms.

The performance of the methods was compared on four PPI networks based on two standard protein complexes. The experimental results of our IMA and other methods on these PPI networks are listed in **Tables 4**, **5**, and the highest value of each metric of each PPI network is in bold.

First, we compared them with the standard protein complexes 1. As shown in **Table 4**, IMA outperformed the other algorithms on the Krogan dataset. IMA obtained a F-measure of 0.6272, CR of 0.3917, MMR of 0.3266, Jaccard of 0.4373, and total score of 2.0688, which were obviously superior to other detection algorithms. MCL predicted 370 protein complexes and achieved the highest ACC of 0.3192. Second, we compared the 13 approaches using the DIP dataset. IMA detected 1338 protein complexes, and achieved the highest F-measure, MMR, Jaccard, and Total scores, respectively. ClusterONE found 904 protein complexes and achieved a CR of 0.5062 and ACC of 0.3270, the best performance in terms of CR and ACC. However, it only achieved a F-measure of 0.5118, MMR of 0.1467, Jaccard of 0.3297, and total score of 1.8214, which were lower than those obtained using the IMA method. Second, we compared our IMA and other methods using the combined6 dataset. **Table 4** shows than the results obtained by using combined6 dataset are similar to those obtained using the DIP dataset. IMA detected 1054 protein complexes and achieved the highest F-measure, MMR, Jaccard, and total score with values of 0.7256, 0.3364, 0.4869, and 2.3052, respectively. IPCA found 2160 protein complexes and achieved a better CR of 0.5106. ClusterONE predicted 648 protein complexes, achieving a ACC of 0.3306, which was the highest. Finally, we also used the WI-PHI dataset to evaluate the performance of all methods, and IMA identified 2561 protein complexes, and IMA scores of F-measure, CR, MMR, Jaccard, and total score were higher than those determined by the other methods, and they were 0.7503, 0.6223, 0.3965, 0.4828, and 2.5579, respectively. As for the ACC, which is among the top three, only lower than SE-DMTG and ClusterONE. From the above analysis, we found that the IMA algorithm achieved the best performance in the most evaluation metrics, with the exception of CR and ACC in some cases. Therefore, these results demonstrate that the IMA outperforms the base and could be an excellent approach to detect protein complexes in PPI networks. More evaluation metrics are made available in the https://github.com/RongquanWang/IMA/Additional file 2.

TABLE 4 | Performance of different algorithms with respect to standard protein complexes 1.

Algorithms	Num	F-measure	CR	ACC	MMR	Jaccard	Total score
Krogan							
MCODE	39	0.3414	0.2140	0.1994	0.0351	0.2359	1.0258
MCL	370	0.4004	0.3895	**0.3192**	0.1123	0.2902	1.5117
IPCA	582	0.5573	0.3389	0.2653	0.2368	0.3713	1.7696
COACH	345	0.5254	0.3473	0.2667	0.1824	0.3556	1.6775
ClusterONE	240	0.4694	0.3085	0.2829	0.1262	0.3324	1.5194
PEWCC	383	0.5289	0.3231	0.2554	0.2194	0.3786	1.7053
ProRank+	357	0.5448	0.3660	0.2718	0.2018	0.3544	1.7388
WPNCA	369	0.5446	0.3897	0.2758	0.1663	0.3415	1.7179
WEC	516	0.5440	0.3442	0.2637	0.2573	0.4005	1.8089
EWCA	676	0.5883	0.3782	0.2769	0.3019	0.4073	1.9525
SE-DMTG	372	0.5878	0.3504	0.2820	0.2284	0.4058	1.8781
MPC-C	458	0.6010	0.3760	0.2814	0.2344	0.3882	1.8810
IMA-unweighted	767	0.5917	0.3827	0.2843	0.3221	0.4174	1.9982
IMA	773	**0.6272**	**0.3917**	0.2859	**0.3266**	**0.4373**	**2.0688**
DIP							
MCODE	26	0.1300	0.2193	0.1337	0.0103	0.1292	0.6224
MCL	628	0.3106	0.3578	0.2684	0.0752	0.2155	1.2275
IPCA	1242	0.5741	0.3519	0.2404	0.2096	0.3004	1.6764
COACH	329	0.5850	0.3697	0.2462	0.1254	0.3305	1.6568
ClusterONE	904	0.5118	**0.5062**	**0.3270**	0.1467	0.3297	1.8214
PEWCC	648	0.6004	0.3783	0.2438	0.1938	0.3514	1.7677
ProRank+	167	0.3123	0.2115	0.1870	0.0452	0.2007	0.9567
WPNCA	623	0.5888	0.4307	0.2594	0.1807	0.3360	1.7955
WEC	253	0.4185	0.3104	0.2309	0.0953	0.3078	1.3628
EWCA	964	0.6428	0.4374	0.2691	0.2534	0.3723	1.9750
SE-DMTG	869	0.6309	0.3822	0.2674	0.2264	0.3573	1.8482
MPC-C	1477	0.6632	0.4413	0.2729	0.2716	0.3537	2.0027
IMA-unweighted	1569	0.6861	0.4488	0.2731	0.2957	0.3959	2.0995
IMA	1338	**0.7196**	0.4528	0.2820	**0.3028**	**0.4234**	**2.1806**
combined6							
MCODE	63	0.2483	0.3441	0.1762	0.0313	0.1832	0.9833
MCL	508	0.3606	0.4628	0.3098	0.0909	0.2871	1.5114
IPCA	2160	0.7218	**0.5106**	0.2783	0.3093	0.4396	2.2597
COACH	682	0.5623	0.4839	0.2653	0.2035	0.3703	1.8855
ClusterONE	648	0.4165	0.5098	**0.3306**	0.1235	0.3173	1.6979
PEWCC	737	0.6586	0.4713	0.2594	0.2661	0.4370	2.0924
ProRank+	472	0.5837	0.3898	0.2394	0.2162	0.4363	1.8657
WPNCA	898	0.5912	0.5725	0.2720	0.1872	0.3306	1.9537
WEC	544	0.5614	0.4367	0.2504	0.2001	0.4120	1.8609
EWCA	935	0.6860	0.5058	0.2771	0.3097	0.4534	2.2321
SE-DMTG	490	0.6854	0.4347	0.2767	0.2326	0.4517	2.2321
MPC-C	1008	0.7001	0.4871	0.2769	0.2820	0.4438	2.1899
IMA-unweighted	1183	0.7097	0.4746	0.2771	0.3341	0.4755	2.271
IMA	1054	**0.7256**	0.4829	0.2734	**0.3364**	**0.4869**	**2.3052**
WI-PHI							
MCODE	124	0.1095	0.4282	0.1720	0.0142	0.1095	0.8333
MCL	772	0.2597	0.4323	0.2960	0.0647	0.2246	1.2773
IPCA	2181	0.5361	0.5789	0.2819	0.2604	0.3585	2.0156
COACH	1353	0.4689	0.6095	0.2752	0.1803	0.3144	1.8483
ClusterONE	1313	0.1813	0.4908	0.3103	0.0689	0.2065	1.2577
PEWCC	1813	0.5440	0.5943	0.2757	0.2535	0.3516	2.0192
ProRank+	255	0.1814	0.2038	0.1801	0.0259	0.1719	0.7630
WPNCA	1813	0.5385	0.6198	0.2834	0.2257	0.3432	2.0106
WEC	729	0.3700	0.4830	0.2353	0.0969	0.2987	1.4839
EWCA	964	0.6428	0.4374	0.2691	0.2534	0.3723	1.9750
SE-DMTG	774	0.4945	0.5198	**0.3107**	0.1816	0.3827	1.8894
MPC-C	2560	0.6068	0.5054	0.2793	0.2013	0.3668	1.9597
IMA-unweighted	3316	0.6769	0.5924	0.2983	0.3841	0.4423	2.3941
IMA	2561	**0.7503**	**0.6223**	0.3060	**0.3965**	**0.4828**	**2.5579**

The bold values are the highest value of each metric of each PPI network.

TABLE 5 | Performance of different algorithms with respect to standard protein complexes 2.

Algorithms	Num	F-measure	CR	ACC	MMR	Jaccard	Total score
Krogan							
MCODE	39	0.2317	0.1863	0.1861	0.0238	0.1982	0.8260
MCL	370	0.3214	0.3534	**0.3088**	0.0792	0.2559	1.3187
IPCA	582	0.4606	0.3097	0.2405	0.1545	0.3271	1.4924
COACH	345	0.4369	0.3166	0.2441	0.1228	0.3136	1.4340
ClusterONE	240	0.3913	0.2729	0.2756	0.0887	0.2826	1.3110
PEWCC	383	0.4228	0.2913	0.2343	0.1418	0.3247	1.4150
ProRank+	357	0.4435	0.3282	0.2621	0.1306	0.3067	1.4711
WPNCA	369	0.4361	**0.3572**	0.2614	0.1083	0.2960	1.4590
WEC	516	0.4344	0.3022	0.2465	0.1559	0.3351	1.4741
EWCA	676	0.5112	0.3483	0.2663	0.1986	0.3607	1.6851
SE-DMTG	372	0.5060	0.3092	0.2684	0.1445	0.3471	1.5852
MPC-C	458	0.5252	0.3354	0.2706	0.1583	0.3338	1.6233
IMA-unweighted	767	0.5091	0.3488	0.2709	0.2078	0.3665	1.7031
IMA	773	**0.5452**	0.3526	0.2733	**0.2135**	**0.3790**	**1.7636**
DIP							
MCODE	26	0.1061	0.1982	0.1205	0.0071	0.1114	0.5433
MCL	628	0.2409	0.3025	0.2504	0.0482	0.1921	1.0341
IPCA	1242	0.4516	0.3196	0.2304	0.1298	0.2674	1.3989
COACH	329	0.4703	0.3184	0.2307	0.0800	0.2829	1.3823
ClusterONE	904	0.4232	**0.4358**	0.2937	0.0972	0.2874	1.5373
PEWCC	648	0.4812	0.3336	0.2329	0.1125	0.2986	1.4588
ProRank+	167	0.2506	0.1895	0.1802	0.0323	0.1784	0.8310
WPNCA	623	0.4603	0.3709	0.2472	0.1065	0.2866	1.4715
WEC	253	0.2921	0.2588	0.2422	0.0526	0.2497	1.0954
EWCA	964	0.5334	0.3812	0.2536	0.1522	0.3226	1.6429
SE-DMTG	869	0.5305	0.3403	0.2562	0.1382	0.3108	1.5697
MPC-C	1477	0.5692	0.3799	0.2538	0.1706	0.3050	1.6785
IMA-unweighted	1569	0.5763	0.3873	0.2583	0.1855	0.336	1.7435
IMA	1338	**0.6064**	0.3828	**0.2710**	**0.1894**	**0.3545**	**1.8041**
combined6							
MCODE	63	0.1771	0.2943	0.1642	0.0213	0.1543	0.8113
MCL	508	0.2902	0.4078	0.2966	0.0629	0.2605	1.3181
IPCA	2160	0.5641	0.4521	0.2824	0.1854	0.3725	1.8567
COACH	682	0.4454	0.4171	0.2626	0.1229	0.3184	1.5665
ClusterONE	648	0.3454	0.4385	**0.3145**	0.0885	0.2881	1.4752
PEWCC	737	0.5223	0.4064	0.2588	0.1531	0.3739	1.7145
ProRank+	472	0.4697	0.3305	0.2322	0.1237	0.3631	1.5194
WPNCA	898	0.4968	**0.5117**	0.2822	0.1182	0.3140	1.7231
WEC	544	0.4324	0.3842	0.2577	0.1172	0.3562	1.5478
EWCA	935	0.5657	0.4523	0.2810	0.1846	0.3971	1.8807
SE-DMTG	490	0.5568	0.3747	0.2782	0.1414	0.3793	1.7306
MPC-C	1008	0.5964	0.4077	0.2677	0.1752	0.3756	1.8225
IMA-unweighted	1183	0.602	0.3999	0.2695	0.1993	0.3988	1.8694
IMA	1054	**0.6127**	0.4138	0.2828	**0.2046**	**0.4066**	**1.9204**
WI-PHI							
MCODE	124	0.0766	0.3701	0.1606	0.0086	0.0938	0.7096
MCL	772	0.2116	0.3563	0.2776	0.0445	0.2042	1.0941
IPCA	2181	0.4655	0.4970	0.2830	0.1634	0.3250	1.7340
COACH	1353	0.3577	0.5228	0.2559	0.1115	0.2821	1.5300
ClusterONE	1313	0.1571	0.4179	0.2920	0.0481	0.1956	1.1108
PEWCC	1813	0.4464	0.5215	0.2615	0.1495	0.3189	1.6979
ProRank+	255	0.1397	0.1732	0.1714	0.0190	0.1502	0.6536
WPNCA	1813	0.4285	**0.5445**	0.2776	0.1346	0.3093	1.6945
WEC	729	0.2914	0.4329	0.2523	0.0590	0.2767	1.3122
EWCA	2347	0.4346	0.5382	0.2840	0.1679	0.3295	1.7542
SE-DMTG	774	0.4252	0.4397	0.2894	0.1183	0.3307	1.6033
MPC-C	2560	0.6068	0.5054	0.2793	0.2013	0.3668	1.9597
IMA-unweighted	3316	0.5822	0.5041	0.2902	0.2427	0.3838	2.0029
IMA	2561	**0.6721**	0.5342	**0.3023**	**0.2547**	**0.4232**	**2.1866**

The bold values are the highest value of each metric of each PPI network.

TABLE 6 | Functional enrichment of the protein complexes identified using different algorithms.

Algorithms	Num	< E-20	< E-15	< E-10	< E-5	Significant
Krogan						
MCODE	39	6 (15.38%)	8 (20.51%)	14 (35.89%)	24 (61.53%)	33 (84.61%)
MCL	370	43 (11.62%)	72 (19.46%)	125 (33.78%)	246 (66.48%)	275 (74.32%)
IPCA	582	108 (18.56%)	162 (27.84%)	244 (41.93%)	445 (76.47%)	485 (83.34%)
COACH	345	66 (19.13%)	107 (31.01%)	165 (47.82%)	272 (78.83%)	295 (85.5%)
ClusterONE	240	62 (25.83%)	92 (38.33%)	130 (54.16%)	199 (82.91%)	212 (88.33%)
PEWCC	383	144 (37.6%)	196 (51.18%)	275 (71.81%)	357 (93.22%)	374 (97.66%)
ProRank+	357	88 (24.65%)	119 (33.33%)	184 (51.54%)	283 (79.27%)	311 (87.11%)
WPNCA	369	62 (16.8%)	100 (27.1%)	167 (45.26%)	290 (78.59%)	311 (84.28%)
WEC	516	133 (25.78%)	186 (36.05%)	262 (50.78%)	421 (81.59%)	447 (86.63%)
EWCA	676	149 (22.04%)	216 (31.95%)	323 (47.78%)	529 (78.25%)	564 (83.43%)
SE-DMTG	372	80 (21.51%)	110 (29.57%)	161 (43.28%)	282 (75.81%)	301 (80.92%)
MPC-C	458	130 (28.38%)	199 (43.45%)	293 (63.97%)	442 (96.5%)	**449(98.03%)**
IMA-unweighted	767	219 (28.55%)	301 (39.24%)	457 (59.58%)	681 (88.78%)	712 (92.82%)
IMA	773	226 (29.24%)	319 (41.27%)	501 (64.81%)	719 (93.01%)	735 (95.08%)
DIP						
MCODE	26	8 (30.77%)	12 (46.15%)	14 (53.84%)	19 (73.07%)	19 (73.07%)
MCL	628	118 (18.79%)	184 (29.3%)	279 (44.43%)	443 (70.54%)	485 (77.23%)
IPCA	1242	147 (11.84%)	315 (25.37%)	556 (44.77%)	972 (78.26%)	1039 (83.65%)
COACH	329	75 (22.8%)	122 (37.09%)	177 (53.81%)	290 (88.16%)	305 (92.72%)
ClusterONE	904	137 (15.15%)	201 (22.23%)	337 (37.27%)	690 (76.32%)	772 (85.39%)
PEWCC	648	153 (23.61%)	247 (38.12%)	376 (58.03%)	572 (88.28%)	597 (92.14%)
ProRank+	167	23 (13.77%)	38 (22.75%)	63 (37.72%)	129 (77.24%)	138 (82.63%)
WPNCA	623	156 (25.04%)	242 (38.84%)	370 (59.39%)	562 (90.21%)	590 (94.7%)
WEC	253	97 (38.34%)	121 (47.83%)	149 (58.9%)	195 (77.08%)	209 (82.61%)
EWCA	964	172 (17.84%)	284 (29.46%)	477 (49.48%)	823 (85.37%)	866 (89.83%)
SE-DMTG	869	142 (16.34%)	213 (24.51%)	358 (41.2%)	708 (81.48%)	770 (88.61%)
MPC-C	1477	323 (21.87%)	538 (36.43%)	906 (61.35%)	1398 (94.66%)	1445 (97.84%)
IMA-unweighted	1569	327 (20.84%)	495 (31.55%)	810 (51.63%)	1430 (91.15%)	1492 (95.1%)
IMA	1338	382 (28.55%)	577 (43.12%)	897 (67.04%)	1305 (97.53%)	**1324 (98.95%)**
combined6						
MCODE	63	26 (41.27%)	31 (49.21%)	42 (66.67%)	57 (90.48%)	60 (95.24%)
MCL	508	129 (25.39%)	162 (31.89%)	209 (41.14%)	323 (63.58%)	349 (68.7%)
IPCA	2160	579 (26.81%)	784 (36.3%)	1145 (53.01%)	1923 (89.03%)	2027 (93.84%)
COACH	682	156 (22.87%)	196 (28.74%)	290 (42.52%)	520 (76.24%)	575 (84.3%)
ClusterONE	648	148 (22.84%)	208 (32.1%)	258 (39.82%)	420 (64.82%)	461 (71.15%)
PEWCC	737	285 (38.67%)	375 (50.88%)	505 (68.52%)	688 (93.35%)	707 (95.93%)
ProRank+	472	255 (54.03%)	324 (68.65%)	395 (83.69%)	443 (93.86%)	452 (95.77%)
WPNCA	898	375 (41.76%)	493 (54.9%)	609 (67.82%)	797 (88.76%)	829 (92.32%)
WEC	544	235 (43.2%)	273 (50.19%)	310 (56.99%)	400 (73.53%)	423 (77.76%)
EWCA	935	274 (29.3%)	337 (36.04%)	437 (46.74%)	721 (77.11%)	770 (82.35%)
SE-DMTG	490	147 (30.0%)	199 (40.61%)	248 (50.61%)	431 (87.96%)	455 (92.86%)
MPC-C	1008	311 (30.85%)	437 (43.35%)	651 (64.58%)	969 (96.13%)	993 (98.51%)
IMA-unweighted	1183	370 (31.28%)	547 (46.24%)	798 (67.46%)	1117 (94.43%)	1152 (97.39%)
IMA	1054	387 (36.72%)	557 (52.85%)	771 (73.15%)	1032 (97.91%)	**1042 (98.86%)**
WI-PHI						
MCODE	124	24 (19.35%)	29 (23.38%)	40 (32.25%)	58 (46.77%)	64 (51.61%)
MCL	772	25 (3.24%)	35 (4.54%)	74 (9.59%)	234 (30.32%)	287 (37.19%)
IPCA	2181	411 (18.84%)	550 (25.21%)	807 (36.99%)	1259 (57.71%)	1345 (61.65%)
COACH	1353	303 (22.39%)	422 (31.19%)	591 (43.68%)	921 (68.07%)	989 (73.1%)
ClusterONE	1313	198 (15.08%)	256 (19.5%)	342 (26.05%)	555 (42.27%)	635 (48.36%)
PEWCC	1813	435 (23.99%)	627 (34.58%)	906 (49.97%)	1297 (71.54%)	1363 (75.18%)
ProRank+	255	53 (20.78%)	60 (23.53%)	83 (32.55%)	145 (56.86%)	156 (61.17%)
WPNCA	1813	429 (23.66%)	594 (32.76%)	834 (46.0%)	1253 (69.11%)	1336 (73.69%)
WEC	729	215 (29.49%)	264 (36.21%)	337 (46.22%)	478 (65.56%)	501 (68.72%)
EWCA	2347	474 (20.2%)	675 (28.76%)	950 (40.48%)	1428 (60.85%)	1532 (65.28%)
SE-DMTG	774	87 (11.24%)	146 (18.86%)	255 (32.94%)	496 (64.08%)	540 (69.76%)
MPC-C	2560	452 (17.66%)	766 (29.93%)	1382 (53.99%)	2163 (84.5%)	2239 (87.47%)
IMA-unweighted	3316	715 (21.56%)	1062 (32.02%)	1656 (49.93%)	2646 (79.79%)	2815 (84.89%)
IMA	2561	847 (33.07%)	1243 (48.53%)	1732 (67.62%)	2326 (90.81%)	**2379 (92.88%)**

The bold values are the highest value of each metric of each PPI network.

TABLE 7 | The co-localization scores of protein complexes detected by different methods in four PPI networks.

Algorithms	Num	Krogan	Num	DIP	Num	combined6	Num	WI-PHI
Co-localization score								
MCODE	39	**0.7442**	26	0.6156	63	0.6481	124	0.5586
MCL	370	0.5820	628	0.5598	508	0.5182	772	0.5510
IPCA	582	0.6656	1242	0.6113	2160	0.5883	2181	0.5585
COACH	345	0.6587	329	0.6527	682	0.5833	1353	0.4960
ClusterONE	240	0.6716	904	0.5778	648	0.5277	1313	0.5207
PEWCC	383	0.7107	648	0.6271	737	0.6493	1813	0.5110
ProRank+	357	0.6779	167	**0.6933**	472	0.7377	255	0.5570
WPNCA	369	0.6245	623	0.6063	898	0.5270	1813	0.4997
WEC	516	0.7182	253	0.6520	544	0.6803	729	0.5039
EWCA	676	0.6960	964	0.6430	935	0.6707	964	0.5303
SE-DMTG	372	0.7164	869	0.6725	490	0.7247	774	0.6419
MPC-C	458	0.7315	1477	0.6503	1008	0.7178	2560	0.6111
IMA-unweighted	767	0.7052	1569	0.6556	1183	0.7090	3316	0.6199
IMA	773	0.7351	1338	0.6728	1054	**0.7440**	2561	**0.6506**

The bold values are the highest value of each metric of each PPI network.

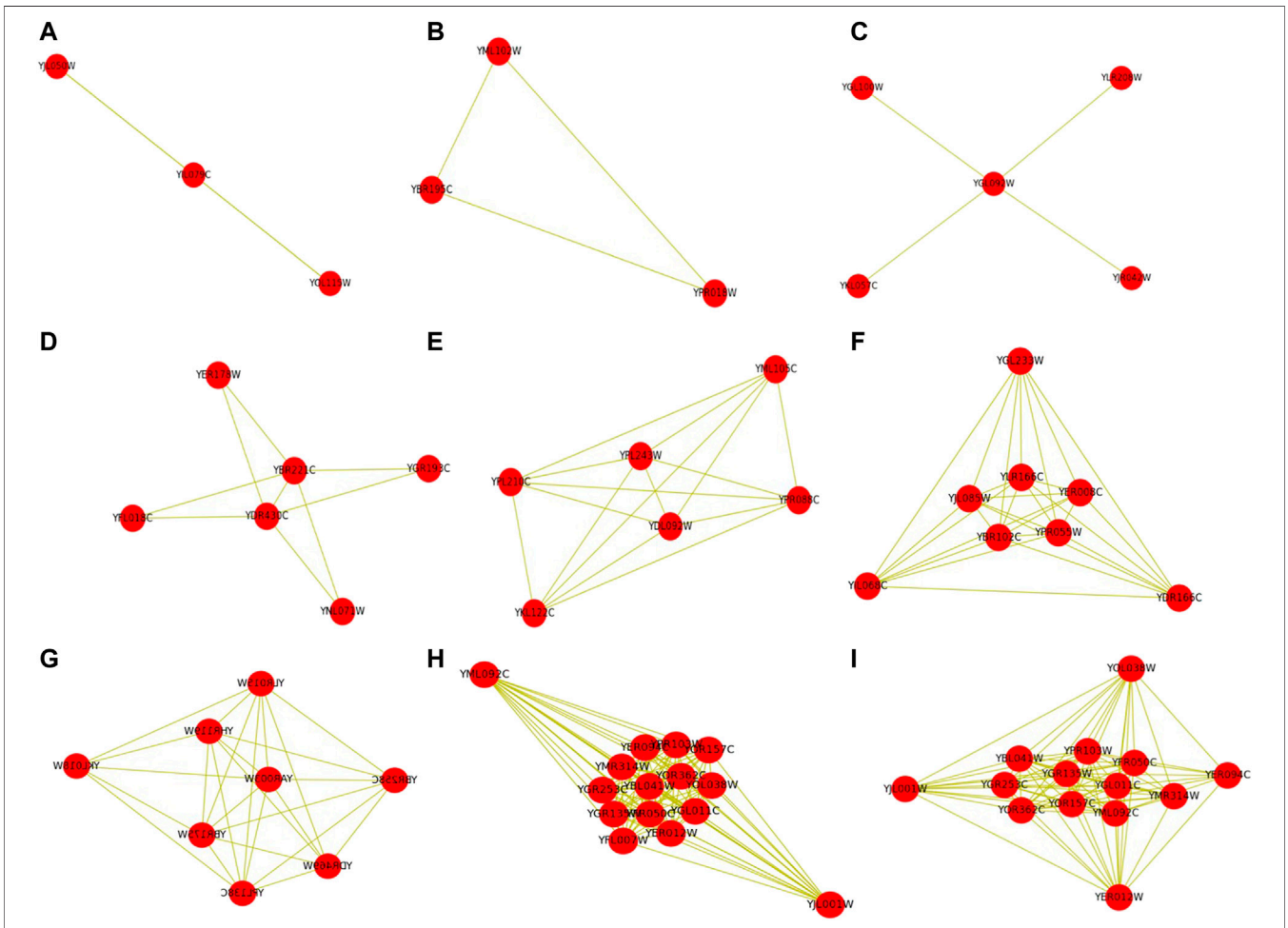

FIGURE 4 | Some examples of detected protein complexes with different topological structures by IMA. Note that these protein complexes contain the number of proteins between 3 and 14. Proteins that have binary interactions are connected by dashed lines.

The comparison results obtained using standard protein complexes 2 in **Table 5** are basically consistent with those obtained using standard protein complexes 1 in **Table 4**. This means that the performance of the proposed IMA is relatively stable. The IMA algorithm performs significantly well on four PPI networks, and it is competitive with the other algorithms in term of

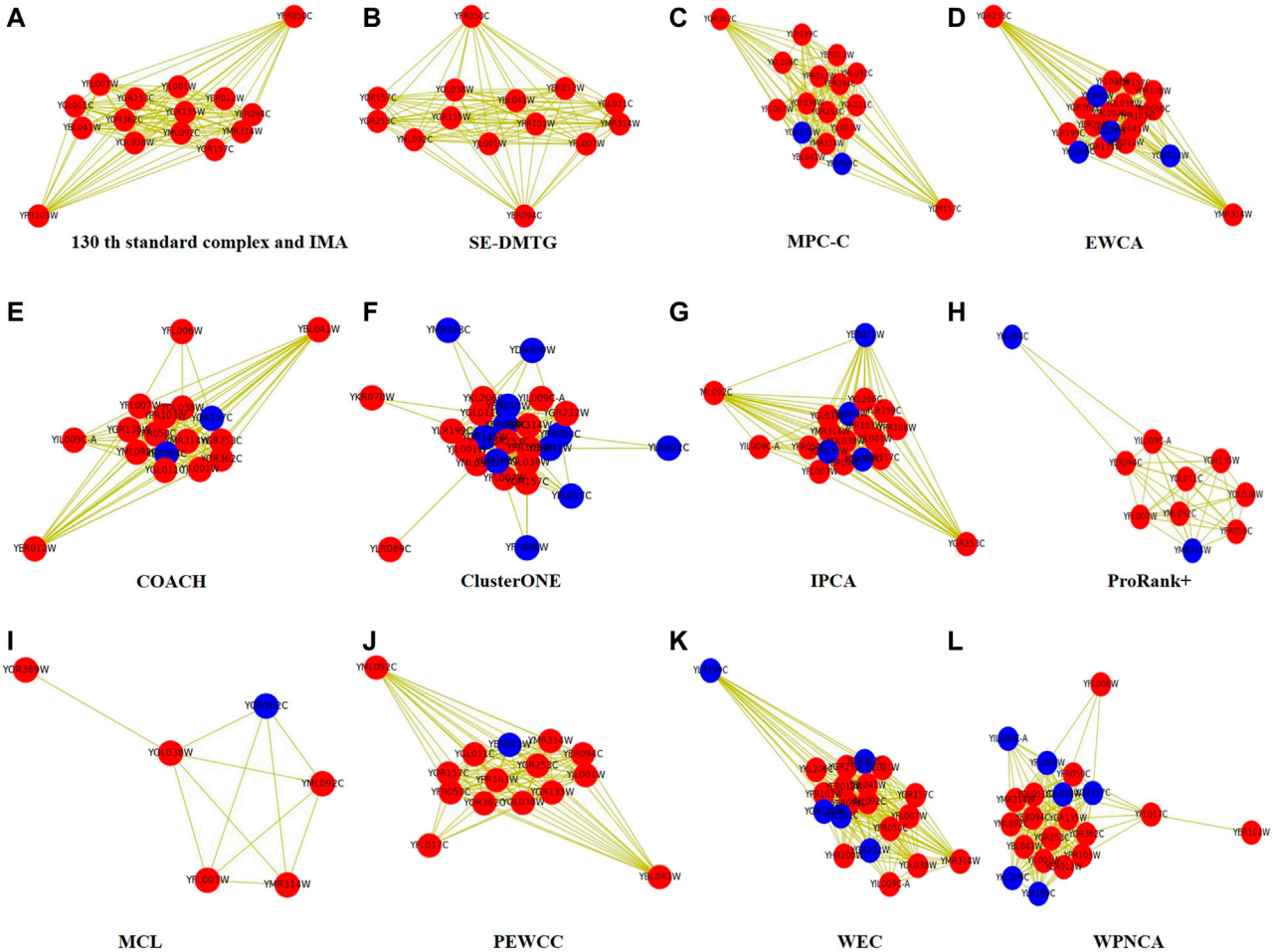

FIGURE 5 | The 130th protein complex in standard protein complexes 2 detected by different methods based on WI-PHI PPI network. Note that the red nodes represent the correctly identified proteins, and the blue nodes represent the wrongly identified proteins by method.

computational evaluation metrics. Additionally, in order to further verify the performance of our IMA algorithm, we also use CYC2008 protein complex dataset and MIPS protein complex dataset to evaluate these identification algorithms. The evaluation results are shown in https://github.com/RongquanWang/IMA/Additional file 4. From the experimental results, we can see that the performance of IMA algorithm on CYC2008 protein complex dataset and MIPS protein complex dataset is basically consistent with the performance on two datasets (standard protein complexes 1 and standard protein complexes 2). This experimental results show that IMA algorithm has strong adaptability and stability to different standard protein complexes.

3.3 Comparison With Functional Enrichment Analysis

We needed to conduct a multi-angle analysis for this statistic, because the *p*-value of the identified protein complexes is closely related to the size of the identified protein complexes (Wang et al., 2019b). For this purpose, the number of detected protein complexes (Num), the number of significantly identified protein complexes, and the percentage of significantly identified protein complexes with different *p*-values from 1E-2 to 1E-20 were used to analyze their functional enrichment. We used a *p*-value test to analyze the protein complexes discovered by the IMA, MCODE, MCL, IPCA, COACH, ClusterONE, PEWCC, ProRank+, WPNCA, WEC, EWCA, SE-DMTG, and MPC-C. The results of the *p*-values of these methods are shown in **Table 6**.

As shown in **Table 6**, the number of protein complexes that could be significantly detected by IMA was higher than that determined by the other methods in the four PPI networks. This means that IMA can detect more protein complexes with biological significance compared to other methods. Although some detected protein complexes do not match standard protein complexes currently, they are likely to be real protein complexes. As for the percentage of significantly detected protein complexes at different thresholds of the *p*-value from E-2 to E-20 in **Table 6**, we can conclude that IMA could detect a relatively higher proportion of protein complexes with biologically significance in most PPI

networks. The above analysis demonstrates that the IMA method could be a promising method for discovering new protein complexes with biological significance.

3.4 Comparison With Subcellular Location Score

According to the definition of colocalization score, it is based on the average colocalization of all detected protein complexes. It should be noted that the lower the number of detected protein complexes, the higher the colocalization score. Here, we used the ProCope tool (Krumsiek et al., 2008) to calculate the colocalization score.

Table 7 shows the average co-localization scores of protein complexes detected using various methods on localization dataset, (Huh et al., 2003). In Krogan, the best co-localization score of 0.7442 is obtained by the MCODE method, but MCODE only detected 39 predicted protein complexes, which was beneficial for achieving high the co-localization score, and IMA obtained a score of 0.7351, lower than MCODE. In DIP, ProRank + detected 357 protein complexes and obtained a co-localization score of 0.6933, which was better than that of all the other methods. In combined6, IMA method detected 1054 detected protein complexes and achieved the highest the co-localization score of 0.7440. In WI-PHI, IMA achieved the highest co-localization score, and the number of protein complexes was 2561. Based on the co-localization score of the detected protein complexes by IMA, it indicates that the proteins of protein complexes predicted by IMA have better localization consistency; these proteins in the same protein complex tend to carry out a similar function.

4 CASE STUDY AND DISCUSSION

IMA algorithm can detect protein complexes with multiple topological structures. **Figure 4** shows some examples of the detected protein complexes with different topological structures by using the IMA algorithm. Note that the standard protein complexes 1 and 2 are also detected by the IMA algorithm. These protein complexes contain the number of proteins between 3 and 14. These protein complexes with different topological structures include linear, triangular, star-like, rectangular, k-clique, dense subgraph, and core-attachment structure, and hybrid structure. Proteins that have binary interactions are connected by dashed lines. More examples can be found them at https://github.com/RongquanWang/IMA/Examples.

Figure 5 visualizes an example of the 130^{th} protein complex in standard protein complexes 1 in the WI-PHI dataset so as to display the detection result more clearly. **Figure 5A** shows that our IMA successfully detected all proteins correctly. **Figures 5B–L** illustrate the protein complexes identified by IMA, SE-DMTG, MPC-C, EWCA, COACH, ClusterONE, IPCA, ProRank+, MCL, PEWCC, WEC, and WPNCA, respectively. The red nodes represent the correctly identified proteins, and the blue nodes represent the wrongly identified proteins.

From **Figure 5**, we can see that SE-DMTG correctly identifies 14 proteins, but misidentifies a protein. Moreover, the other methods have inaccurately proteins. Our IMA can correctly identify almost all proteins, which suggests that the IMA algorithm is superior to the other comparative methods.

5 CONCLUSION

In this paper, we present a novel IMA method for identifying protein complexes in PPI network. The key idea of IMA is enabled us to design an improved memetic algorithm to optimize a fitness function for identifying protein complexes in PPI networks based on existing contending methods and a weighted PPI network. Here, an improved memetic algorithm is the cooperation of a genetic algorithm with a local optimization strategy. A genetic algorithm is used to improve the diversity of the population, and the local optimization strategy helps to locate better solutions more quickly. Furthermore, we designed a fitness function to overcome the limitations of a single objective function in estimating an individual's fitness. The experimental results show that IMA significantly outperforms the existing outstanding algorithms in various metrics. We will use graph neural networks (Zhang et al., 2021) and other evolutionary algorithms to improve the accuracy of protein complexes identified in the future.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/RongquanWang/IMA.

AUTHOR CONTRIBUTIONS

RW was responsible for the development phase of the main algorithm and drafted the article. HM and CW also revised the drafted article and approved the content for publication. All authors were responsible for the design of the algorithm.

FUNDING

This work was supported by the Fundamental Research Funds for the National Natural Science Foundation of China (No. U20B2062), the Fundamental Research Funds for the Central Universities (No. FRF-TP-20-064A1Z), the Civil Aviation Flight Academy of China (No. FZ2021ZZ05), the National Natural Science Foundation of China (No. 62172036). The funders provided the financial support to the research, but had no role in the design of the study, analysis, interpretations of data and in writing the manuscript.

REFERENCES

Abduljabbar, D. A., Hashim, S. Z. M., and Sallehuddin, R. (2020). "An Enhanced Evolutionary Algorithm for Detecting Complexes in Protein Interaction Networks with Heuristic Biological Operator," in International Conference on Soft Computing and Data Mining (Berlin, Germany: Springer), 334–345. doi:10.1007/978-3-030-36056-6_32

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., et al. (2004). Structure-based Assembly of Protein Complexes in Yeast. *Science* 303, 2026–2029. doi:10.1126/science.1092645

Bader, G. D., and Hogue, C. W. (2003). An Automated Method for Finding Molecular Complexes in Large Protein Interaction Networks. *BMC bioinformatics* 4, 2–27. doi:10.1186/1471-2105-4-2

Blatti, C., III, Emad, A., Berry, M. J., Gatzke, L., Epstein, M., Lanier, D., et al. (2020). Knowledge-guided Analysis of "omics" Data Using the KnowEnG Cloud Platform. *Plos Biol.* 18, e3000583. doi:10.1371/journal.pbio.3000583

Friedel, C. C., Krumsiek, J., and Zimmer, R. (2009). Bootstrapping the Interactome: Unsupervised Identification of Protein Complexes in Yeast. *J. Comput. Biol.* 16, 971–987. doi:10.1089/cmb.2009.0023

Gach, O., and Hao, J.-K. (2012). "A Memetic Algorithm for Community Detection in Complex Networks," in International conference on parallel problem solving from nature, Berlin, Heidelberg: Springer, 327–336. doi:10.1007/978-3-642-32964-7_33

Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., et al. (2006). Proteome Survey Reveals Modularity of the Yeast Cell Machinery. *Nature* 440, 631–636. doi:10.1038/nature04532

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., et al. (2002). Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes. *Nature* 415, 141–147. doi:10.1038/415141a

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., et al. (2019). CORUM: the Comprehensive Resource of Mammalian Protein Complexes-2019. *Nucleic Acids Res.* 47, D559–D563. doi:10.1093/nar/gky973

Hanna, E. M., and Zaki, N. (2014). Detecting Protein Complexes in Protein Interaction Networks Using a Ranking Algorithm with a Refined Merging Procedure. *BMC bioinformatics* 15, 204–211. doi:10.1186/1471-2105-15-204

Hao, J.-K. (2012). "Memetic Algorithms in Discrete Optimization," in Handbook of memetic algorithms, 73–94. doi:10.1007/978-3-642-23247-3_6

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., et al. (2007). Gene Ontology Annotations at Sgd: New Data Sources and Annotation Methods. *Nucleic Acids Res.* 36, D577–D581. doi:10.1093/nar/gkm909

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., et al. (2003). Global Analysis of Protein Localization in Budding Yeast. *Nature* 425, 686–691. doi:10.1038/nature02026

Keretsu, S., and Sarmah, R. (2016). Weighted Edge Based Clustering to Identify Protein Complexes in Protein-Protein Interaction Networks Incorporating Gene Expression Profile. *Comput. Biol. Chem.* 65, 69–79. doi:10.1016/j.compbiolchem.2016.10.001

Kiemer, L., Costa, S., Ueffing, M., and Cesareni, G. (2007). Wi-phi: a Weighted Yeast Interactome Enriched for Direct Physical Interactions. *Proteomics* 7, 932–943. doi:10.1002/pmic.200600448

King, A. D., Przulj, N., and Jurisica, I. (2004). Protein Complex Prediction via Cost-Based Clustering. *Bioinformatics* 20, 3013–3020. doi:10.1093/bioinformatics/bth351

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., et al. (2006). Global Landscape of Protein Complexes in the Yeast saccharomyces Cerevisiae. *Nature* 440, 637–643. doi:10.1038/nature04670

Krumsiek, J., Friedel, C. C., and Zimmer, R. (2008). ProCope--protein Complex Prediction and Evaluation. *Bioinformatics* 24, 2115–2116. doi:10.1093/bioinformatics/btn376

Lei, X., Fang, M., Guo, L., and Wu, F. X. (2019b). Protein Complex Detection Based on Flower Pollination Mechanism in Multi-Relation Reconstructed Dynamic Protein Networks. *BMC bioinformatics* 20, 131–174. doi:10.1186/s12859-019-2649-0

Lei, X., Ding, Y., Fujita, H., and Zhang, A. (2016a). Identification of Dynamic Protein Complexes Based on Fruit Fly Optimization Algorithm. *Knowledge-Based Syst.* 105, 270–277. doi:10.1016/j.knosys.2016.05.019

Lei, X., Fang, M., and Fujita, H. (2019a). Moth-flame Optimization-Based Algorithm with Synthetic Dynamic PPI Networks for Discovering Protein Complexes. *Knowledge-Based Syst.* 172, 76–85. doi:10.1016/j.knosys.2019.02.011

Lei, X., Wang, F., Wu, F.-X., Zhang, A., and Pedrycz, W. (2016b). Protein Complex Identification through Markov Clustering with Firefly Algorithm on Dynamic Protein-Protein Interaction Networks. *Inf. Sci.* 329, 303–316. doi:10.1016/j.ins.2015.09.028

Lei, X., Zhang, Y., Cheng, S., Wu, F.-X., and Pedrycz, W. (2018). Topology Potential Based Seed-Growth Method to Identify Protein Complexes on Dynamic Ppi Data. *Inf. Sci.* 425, 140–153. doi:10.1016/j.ins.2017.10.013

Li, M., Chen, J. E., Wang, J. X., Hu, B., and Chen, G. (2008). Modifying the Dpclus Algorithm for Identifying Protein Complexes Based on New Topological Structures. *BMC bioinformatics* 9, 398. doi:10.1186/1471-2105-9-398

Li, Y., Jiao, L., Li, P., and Wu, B. (2014). A Hybrid Memetic Algorithm for Global Optimization. *Neurocomputing* 134, 132–139. doi:10.1016/j.neucom.2012.12.068

Liu, G., Liu, B., Li, A., Wang, X., Yu, J., and Zhou, X. (2021). Identifying Protein Complexes with clear Module Structure Using Pairwise Constraints in Protein Interaction Networks. *Front. Genet.* 12. doi:10.3389/fgene.2021.664786

Liu, G., Wong, L., and Chua, H. N. (2009). Complex Discovery from Weighted Ppi Networks. *Bioinformatics* 25, 1891–1897. doi:10.1093/bioinformatics/btp311

Ma, C. Y., Chen, Y. P., Berger, B., and Liao, C. S. (2017). Identification of Protein Complexes by Integrating Multiple Alignment of Protein Interaction Networks. *Bioinformatics* 33, 1681–1688. doi:10.1093/bioinformatics/btx043

Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Güldener, U., Mannhaupt, G., et al. (2004). Mips: Analysis and Annotation of Proteins from Whole Genomes. *Nucleic Acids Res.* 32, D41–D44. doi:10.1093/nar/gkh092

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting Overlapping Protein Complexes in Protein-Protein Interaction Networks. *Nat. Methods* 9, 471–472. doi:10.1038/nmeth.1938

Peng, W., Wang, J., Zhao, B., and Wang, L. (2014). Identification of Protein Complexes Using Weighted Pagerank-Nibble Algorithm and Core-Attachment Structure. *Ieee/acm Trans. Comput. Biol. Bioinform* 12, 179–192. doi:10.1109/TCBB.2014.2343954

Pu, S., Wong, J., Turner, B., Cho, E., and Wodak, S. J. (2009). Up-to-date Catalogues of Yeast Protein Complexes. *Nucleic Acids Res.* 37, 825–831. doi:10.1093/nar/gkn1005

Ramadan, E., Naef, A., and Ahmed, M. (2016). Protein Complexes Predictions within Protein Interaction Networks Using Genetic Algorithms. *BMC bioinformatics* 17 (Suppl. 7), 269–489. doi:10.1186/s12859-016-1096-4

SabziNezhad, A., and Jalili, S. (2020). Dpct: a Dynamic Method for Detecting Protein Complexes from Tap-Aware Weighted Ppi Network. *Front. Genet.* 11, 567. doi:10.3389/fgene.2020.00567

Samanta, M. P., and Liang, S. (2003). Predicting Protein Functions from Redundancies in Large-Scale Protein Interaction Networks. *Proc. Natl. Acad. Sci.* 100, 12579–12583. doi:10.1073/pnas.2132527100

Spears, W. M., and De Jong, K. D. (1995). *On the Virtues of Parameterized Uniform Crossover*. Washington, DC: Naval Research Lab.

Spirin, V., and Mirny, L. A. (2003). Protein Complexes and Functional Modules in Molecular Networks. *Proc. Natl. Acad. Sci.* 100, 12123–12128. doi:10.1073/pnas.2032324100

Srihari, S. M. (2012). *Integrating Biological Insights with Topological Characteristics for Improved Complex Prediction from Protein Interaction Networks*. Singapore: Citeseer.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., et al. (2019). Random Walk with Restart on Multiplex and Heterogeneous Biological Networks. *Bioinformatics* 35, 497–505. doi:10.1093/bioinformatics/bty637

Van Dongen, S. M. (2000). *Graph Clustering by Flow Simulation*.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., et al. (2002). Comparative Assessment of Large-Scale Data Sets of Protein-Protein Interactions. *Nature* 417, 399–403. doi:10.1038/nature750

Wang, J., Peng, X., Li, M., and Pan, Y. (2013). Construction and Application of Dynamic Protein Interaction Network Based on Time Course Gene Expression Data. *Proteomics* 13, 301–312. doi:10.1002/pmic.201200277

Wang, J., Peng, X., Peng, W., and Wu, F.-X. (2014). Dynamic Protein Interaction Network Construction and Applications. *Proteomics* 14, 338–352. doi:10.1002/pmic.201300257

Wang, R., Liu, G., and Wang, C. (2019a). Identifying Protein Complexes Based on an Edge Weight Algorithm and Core-Attachment Structure. *BMC bioinformatics* 20, 471. doi:10.1186/s12859-019-3007-y

Wang, R., Wang, C., Sun, L., and Liu, G. (2019b). A Seed-Extended Algorithm for Detecting Protein Complexes Based on Density and Modularity with Topological Structure and Go Annotations. *BMC genomics* 20, 637. doi:10.1186/s12864-019-5956-y

Wang, R., Wang, C., and Liu, G. (2020). A Novel Graph Clustering Method with a Greedy Heuristic Search Algorithm for Mining Protein Complexes from Dynamic and Static Ppi Networks. *Inf. Sci.* 522, 275–298. doi:10.1016/j.ins.2020.02.063

Wu, M., Li, X., Kwoh, C. K., and Ng, S. K. (2009). A Core-Attachment Based Method to Detect Protein Complexes in Ppi Networks. *BMC bioinformatics* 10, 169. doi:10.1186/1471-2105-10-169

Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002). Dip, the Database of Interacting Proteins: a Research Tool for Studying Cellular Networks of Protein Interactions. *Nucleic Acids Res.* 30, 303–305. doi:10.1093/nar/30.1.303

Zaki, N., Efimov, D., and Berengueres, J. (2013). Protein Complex Detection Using Interaction Reliability Assessment and Weighted Clustering Coefficient. *BMC bioinformatics* 14, 163–169. doi:10.1186/1471-2105-14-163

Žalik, K. R., and Žalik, B. (2018). Memetic Algorithm Using Node Entropy and Partition Entropy for Community Detection in Networks. *Inf. Sci.* 445, 38–49.

Zhang, X. M., Liang, L., Liu, L., and Tang, M. J. (2021). Graph Neural Networks and Their Current Applications in Bioinformatics. *Front. Genet.* 12, 690049. doi:10.3389/fgene.2021.690049

Zhang, Y., Lei, X., and Tan, Y. (2017). "Firefly Clustering Method for Mining Protein Complexes," in International Conference on Swarm Intelligence (Berlin, Germany: Springer), 601–610. doi:10.1007/978-3-319-61824-1_65

Zhang, Y., Lin, H., Yang, Z., Wang, J., Li, Y., and Xu, B. (2013a). Protein Complex Prediction in Large Ontology Attributed Protein-Protein Interaction Networks. *Ieee/acm Trans. Comput. Biol. Bioinf.* 10, 729–741. doi:10.1109/tcbb.2013.86

Zhang, Y., Lin, H., Yang, Z., Wang, J., and Xu, B. (2013b). Integrating Multiple Biomedical Resources for Protein Complex Prediction. *IEEE Int. Conf. Bioinformatics Biomed.* 456, 459. doi:10.1109/bibm.2013.6732535

Zhao, J., Lei, X., and Wu, F.-X. (2017). Predicting Protein Complexes in Weighted Dynamic Ppi Networks Based on Icsc. *Complexity* 2017, 4120506. doi:10.1155/2017/4120506

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.