mathematics

MDPI

Article

A Method for Detecting Overlapping Protein Complexes Based on an Adaptive Improved FCM Clustering Algorithm

Caixia Wang [1], Rongquan Wang [2,*] and Kaiying Jiang [2]

[1] School of International Economics, China Foreign Affairs University, 24 Zhanlan Road, Xicheng District, Beijing 100037, China; wangcaixia@cfau.edu.cn
[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, 30 Xueyuan Road, Haidian District, Beijing 100083, China
* Correspondence: rongquanwang@ustb.edu.cn

Abstract: A protein complex can be regarded as a functional module developed by interacting proteins. The protein complex has attracted significant attention in bioinformatics as a critical substance in life activities. Identifying protein complexes in protein–protein interaction (PPI) networks is vital in life sciences and biological activities. Therefore, significant efforts have been made recently in biological experimental methods and computing methods to detect protein complexes accurately. This study proposed a new method for PPI networks to facilitate the processing and development of the following algorithms. Then, a combination of the improved density peaks clustering algorithm (DPC) and the fuzzy C-means clustering algorithm (FCM) was proposed to overcome the shortcomings of the traditional FCM algorithm. In other words, the rationality of results obtained using the FCM algorithm is closely related to the selection of cluster centers. The objective function of the FCM algorithm was redesigned based on 'high cohesion' and 'low coupling'. An adaptive parameter-adjusting algorithm was designed to optimize the parameters of the proposed detection algorithm. This algorithm is denoted as the DFPO algorithm (DPC-FCM Parameter Optimization). Finally, the performance of the DFPO algorithm was evaluated using multiple metrics and compared with over ten state-of-the-art protein complex detection algorithms. Experimental results indicate that the proposed DFPO algorithm exhibits improved detection accuracy compared with other algorithms.

Keywords: protein–protein interaction network; protein complexes; fuzzy clustering algorithm; density peaks clustering algorithm; parameter optimization; swarm intelligence optimization algorithm

MSC: 05C85; 68W50

Academic Editor: Ioannis Tsoulos

Received: 5 December 2024
Revised: 5 January 2025
Accepted: 6 January 2025
Published: 9 January 2025

Citation: Wang, C.; Wang, R.; Jiang, K. A Method for Detecting Overlapping Protein Complexes Based on an Adaptive Improved FCM Clustering Algorithm. *Mathematics* **2025**, *13*, 196. https://doi.org/10.3390/math13020196

1. Introduction

As one of the most fundamental interdisciplinary topics, complex network analysis exhibits excellent theoretical significance and application prospects; through investigations of complex networks and artificial intelligence in theory and application, identifying community structures in complex networks has attracted significant attention. A complex network was once defined as comprising several or all of the following features: self-organization, self-similarity, small world, and scale-free [1]. Indeed, a complex network is obtained by abstracting and summarizing complex systems in the practical world. Analysis methods for complex networks can be employed to describe daily life and social phenomena, and they have been widely applied as they are easy yet effective.

Detection of community structure refers to the process where a complex network is divided into multiple subgraphs. In other words, nodes in a complex network would be present in various community structures owing to their features. Nodes in the same community structure are closely connected, while those in different community structures are sparsely connected [2–5], resulting in the idea of 'high cohesion, low coupling'. As an essential structural feature of complex networks, community structures are naturally present in complex networks of various types. For instance, community structure represents pages or attributes with identical or similar subject categories in the World Wide Web (WWW); people with identical interests in social networks (WeChat or Weibo) [6]; and a protein complex in protein–protein interaction (PPI) networks. Overall, the detection of community structures exhibits excellent significance to a thorough understanding of functions and features of a complex network, determination of its intrinsic topological structure, clarification of its potential, and prediction of its behaviors [7].

Various PPI datasets have been generated with the popularity of high-throughput rapid detection technologies. These datasets can be naturally expressed as a network. Specifically, proteins are represented by network nodes, and their interactions appear as edges in the network. Generally, closely connected subgraphs in the PPI network are assumed to be real protein complexes. Identifying such protein complexes in the PPI network is significant for understanding life activities and exploring life science. The PPI network is a 'small-world' network [8,9], and dense subgraphs or modules thereof typically correspond to protein complexes [10,11]. Mining protein complexes in the PPI network has excellent theoretical significance for investigating the complex activities of life, interpretation of life mysteries, and understanding the intrinsic organizations and processes of complex life networks at the system level. Additionally, detecting protein complexes is of great scientific and commercial value for recognizing pathogenic genes, phenotypic effects of gene mutations, and the evolution of biological networks [12].

During the completion of life activities, few proteins function alone; instead, most proteins [13] interact with each other to form protein complexes or functional modules, thus realizing their main functions [14]. A protein complex refers to a group of proteins that aggregate and interact with each other to achieve a specific physiological process; such an aggregate exists under particular space and time conditions. Hence, studies of protein complexes facilitate understanding of cell activities and function realization processes and would make significant contributions to life science and medicine in the future. For instance, RNA polymerases are responsible for the synthesis of RNA, which takes place during transcription; proteasomes are responsible for molecular degradation. Both polymerases and proteasomes are common protein complexes. Traditional methods for detecting protein complexes rely on biotechnologies such as TAP-MS (tandem affinity purification) [15]. These methods are limited because some protein complexes with weak binding may not be detected; yeast protein two-hybrid analysis [16] is readily exposed to low detection sensitivity of false positive and false negative protein complexes. Meanwhile, detecting protein complexes by biotechnologies is costly and time-consuming, and thus, it cannot meet the needs in the field of proteomics [17]. Currently, computing-based data mining methods are widely applied. Due to its low cost and short processing time, computer technology can rapidly and accurately establish PPI networks, and computing-based methods for detecting protein complexes have been developed [18].

Over the past decades, various computing-based methods have been proposed to automatically detect protein complexes in PPI networks to overcome the drawbacks and limitations of the experiment-based methods. As protein complexes are more likely to be present in local PPI network areas with high density, various density-subgraph-based protein complex detection algorithms have been proposed. However, these algorithms

are barely applicable to detecting overlapping protein complexes and are incapable of automatic and adaptive parameter settings according to the input PPI network. Therefore, a novel protein complex detection algorithm that combines the fuzzy C-means clustering (FCM) algorithm with swarm intelligence optimization (SIO) can solve the above urgent problems.

1.1. Related Works

The state-of-the-art protein complex detection algorithms are discussed from several perspectives: methods for detecting protein complexes based on unsupervised and supervised learning, methods for detecting protein complexes based on the SIO algorithm, and methods for detecting protein complexes based on the FCM algorithm.

1.1.1. Protein Complex Detection Methods Based on Unsupervised Learning

In recent years, various methods have been proposed for detecting protein complexes. Such methods are based on the hypothesis that the dense subgraphs in PPI networks are the protein complexes to be detected. Based on this hypothesis, various methods based on unsupervised learning have been proposed for detecting protein complexes.

Bader and Hogue proposed the protein complex detection algorithm (MCODE) [19]. This algorithm has three steps. First, the local neighborhood density of the node was calculated, and a weight was assigned to the node according to the obtained density. Then, nodes with large weights were employed as seeds. Finally, seeds were expanded for detection. Proposed by Liu et al. in 2009 [20], CMC achieves the detection of protein complexes by identifying the largest aggregate in the PPI network. The Markov clustering algorithm (MCL) [21] achieves detection of protein complexes by simulation of a random walk in the PPI network. In addition, Omranian and Nikoloski [22] proposed an algorithm called CUBCO+, which employs GO semantic similarity to retain biologically relevant interactions and uses link prediction approaches to predict protein complexes. Most of the protein–protein interaction networks heavily suffer from noise. Wang et al. [23] constructed a cross-species ortholog relation matrix and transferred GO terms from other species to evaluate the confidence of PPIs, then used the PPI filter strategy to clean the PPI network. Finally, they constructed a weighted clean PPI network, which is used for detecting protein complexes.

Li et al. [24] proposed IPCA, which achieves the detection of protein complexes by seed selection and local searching. SPICi is a clustering method for rapid clustering in biological networks with small storage space occupied; it mainly achieves the detection of clustering structures based on density function and support function [25]. ClusterONE [26] achieves the detection of dense subgraphs in PPI networks by using a greedy algorithm, and these subgraphs with higher cohesiveness scores are regarded as protein complexes. CPredictor2.0 [27] is an algorithm that effectively identifies protein complexes in the PPI network. Zhan et al. [28] proposed a partially shared signed network clustering model, which considered PPI signs and identified the common and unique protein complexes in different states. It is an effective way to jointly detect protein complexes from multiple state-specific signed PPI networks.

Specifically, proteins are first grouped based on their function annotation information, and density subgraphs were detected using the Markov algorithm and employed as protein complexes. As a nonparametric greedy approximation algorithm, PC2P [29] converts protein complex detection into network segmentation and can detect double-chain spanning subgraphs comprising sparse and dense subgraphs. Lyu et al. [30] firstly calculated the balanced weights to replace the original weights and divide the original PPIN into small

PPINs, then enumerated the connected subset of each small PPIN and remove similar PPINs This method, called BOPS, identifies potential protein complexes based on cohesion.

In COACH, the core of the protein complex is first detected, then attachment proteins are introduced and mined [31]. Peng et al. [32] proposed the WPNCA based on the core-attachment structure. First, density subgraphs are detected and used as the core of protein complexes. Then, attachment proteins are recognized based on the core of protein complexes. Owing to high proportions of false positive and false negative interactions in PPI networks (50%) [33], algorithms such as PEWCC [34] have been proposed to mitigate the influences of false positive and false negative interactions on detection of protein complexes. The topological structure of the PPI network was employed, and weights were assigned to interactions to improve the reliability of edges, thus enhancing the detection accuracy of protein complexes. ICJointLE [35] is a classical method for detecting protein complexes based on the co-expression and co-localization of proteins in the same protein complex. SE-DMTG [36] is a seed-extending algorithm that achieves the detection of protein complexes by a combinatorial function. The proposed MPC-C [37] is based on the 3-sigma principles. In addition, based on core attachment and second-order neighbors, Yang Yu and Dezhou Kong [38] combined the resource allocation with gene expression to detect protein complexes from the PPI network. Herein, a series of time-series subnetworks are established using gene expression data. These time-series subnetworks constitute a dynamic PPI network; static and dynamic protein complexes in original static and dynamic PPI networks were detected, respectively.

Nevertheless, the methods mentioned above are based on unsupervised learning and exhibit the following limitations: (1) the accuracy is reasonable only if interacting edges are highly reliable; (2) the detected protein complex has a relatively single topological structure; (3) known topological features of protein complexes can barely be effectively utilized and learned.

1.1.2. Protein Complex Detection Methods Based on Supervised Learning

Although methods based on unsupervised learning are not affected by the insufficient and unskilled nature of protein complex training models, they cannot effectively learn the structures of standard protein complexes; instead, they can only detect protein complexes with simple topological structures. Recently, increasing known protein complexes have been reported, and training the model of protein complexes has become increasingly normative. As a result, several methods for detecting protein complexes based on supervised learning have been proposed. Yu et al. [39] reported the detection of protein complexes by using a trained regression model and some identified initial clusters. Lei et al. [40] reported a semi-learning algorithm that achieves the detection of protein complexes based on a trained neural network model. ClusterEPs [41] can effectively distinguish real protein complexes from random subgraphs and determine whether a specific subgraph is a protein complex. Dong et al. reported ClusterSS [42], which combines a neural network with a local modularity graph and achieves the detection of protein complexes based on the combination and searching strategies. Liu et al. [43] reported a supervised learning algorithm based on a network embedding method and random forest model. Sikandar et al. [44] proposed a protein complex detection algorithm based on decision trees and biological information. Wang et al. [45] proposed a supervised learning method based on network representation learning and the gene ontology knowledge of known protein complexes, which can predict new protein complexes. Palukuri et al. [46] developed and evaluated a reinforcement learning pipeline algorithm that is trained to calculate the value of different subgraphs encountered while walking on the network to reconstruct known complexes and then scales the reinforcement learning pipeline to search for novel protein complexes. Based on the

topological characteristics of the subgraph, Sahoo et al. [47] proposed a decision-tree-based method that is effective and efficient in identifying protein complexes from large-scale PPI networks not only on the human Database of Interacting Proteins but also on Biological General Repository for Interaction Datasets. At the same time, Chen et al. [48] presented an adaptive convolution graph network to predict protein functional modules that effectively integrate protein gene ontology attributes and network topology.

Nevertheless, methods for the detection of protein complexes based on supervised learning have several limitations: (1) low detection accuracy; (2) poor extraction of effective topological features describing protein complexes; and (3) the dataset of protein complexes available for training is small, and the trained model is overfitted.

1.1.3. Protein Complex Detection Methods Based on the SIO Algorithm

Researchers working on SIO algorithms would observe and imitate the behaviors of social animals to achieve optimization. Due to excellent optimization performance and robustness, the SIO algorithms have been widely applied to detect protein complexes. Currently, the SIO algorithms are applied in identifying protein complexes from two perspectives.

One is detection of protein complexes by using the SIO algorithms. In 2015, Ramadan et al. [49] proposed a method for detection of protein complex by combining genetic algorithm and drosophila optimization clustering algorithm [50]; in 2017, Zhang et al. proposed a novel method based on firefly algorithm [51]; in 2017, Zhao et al. [52] proposed the improved cuckoo search clustering (ICSC) algorithm; in 2019, Lei et al. proposed a method for prediction of protein complexes based on the moth-flame optimization (MFO) algorithm [53]; and in 2022, Feng et al. [54] developed a new MP-DE algorithm, which generated protein complex cores using Markov clustering and searched for attached proteins using a differential evolution algorithm.

The other is optimizing parameters proposed by the detection methods using the SIO algorithms. In 2015, Lei et al. [55] proposed the ISHC clustering method and in 2016, Lei et al. [56] proposed the F-MCL clustering algorithm, which are based on Markov clustering and firefly algorithms. Nevertheless, such algorithms exhibited several limitations. For instance, these algorithms could have displayed better local optimization capability despite strong global optimization capability.

1.1.4. Protein Complex Detection Methods Based on the FCM Algorithm

Due to good accuracy and high efficiency, FCM-based clustering algorithms have attracted significant attention recently. Lei et al. proposed a clustering model [57], which combines the optimization mechanism of an artificial bee colony (ABC) with a fuzzy membership matrix for the detection of protein complexes. Mao et al. [58] designed a clustering algorithm based on a fuzzy ant colony algorithm for mining protein complexes and proposed a novel objective function. Hu et al. [59] proposed a method that realizes the detection of protein complexes in the yeast PPI network by using a fuzzy clustering algorithm. Zhang et al. [60] investigated the detection of community structures in the complex network using FCM. In this study, overlapping protein complexes were detected using the fuzzy clustering algorithm owing to its excellent performance in graph clustering tasks [61]. However, a fuzzy clustering algorithm is limited by several issues. For instance, the performance of the FCM algorithm is closely related to the selection of the initial clustering center and the number of clusters; the design of the objective function is not aligned with the detection of protein complexes; and most parameter setting strategies are based on manual parameter adjustment, which is not flexible enough.

1.2. Motivation and Innovation

According to public protein complex databases, overlapping proteins are widely present for different protein complexes. In other words, a considerable amount of overlapping protein complexes is present in PPI networks. As a result, the detection of overlapping protein complexes is exceptionally challenging. Therefore, it is urgent to develop an overlapping protein complex detection algorithm that further considers overlapping factors to enhance the detection accuracy of overlapping protein complexes.

The FCM algorithm can detect overlapping protein complexes due to its fuzzy membership matrix. Meanwhile, the FCM algorithm exhibits higher accuracy and applicability than other clustering algorithms. Nevertheless, the clustering center and cluster number were generated by random initialization in traditional FCM algorithms. As a result, the selection of the initial clustering center and cluster number directly affects the clustering algorithm's performance. The clustering center and number could be determined using the traditional density peaks clustering (DPC) algorithm in the FCM algorithm. However, the traditional DPC algorithm involves tedious distance calculation and is not directly applicable to the PPI networks. Hence, this paper proposed an improved DPC algorithm to determine the clustering center and number. This method calculated the local densities of all sample points and compared them with those of their neighbor proteins. Then, proteins with a local density higher than their neighbor protein were selected as clustering centers. Additionally, proteins with local density over the threshold and no other clustering centers present in their vicinities were employed as initial clustering centers for the FCM algorithm; the number of initial clustering centers was used as the initial cluster number of the FCM algorithm. The objective function of the traditional FCM algorithm only considers that the points within one specific category should be closely related to each other and ignores that the distances between different classes should be significant.

In this paper, an objective function considering both high cohesion and low coupling was designed and used to guide clustering by the FCM algorithm. In the proposed improved FCM detection algorithm, the input PPI network was preprocessed, the clustering center and cluster number of the FCM algorithm were initialized by the improved DPC algorithm based on the drawbacks of the current FCM algorithm, and an objective function suitable for the detection of protein complex was proposed, thus achieving optimization of membership matrix by the objective function and FCM algorithm. Additionally, a protein division strategy was proposed to detect overlapping protein complexes.

Most traditional methods for protein complex detection require manual parameter adjustment, which has two issues. First, manual parameter adjustment needs to be more flexible and adaptively set parameters for different PPI networks according to the input PPI network. Inspired by the ABC algorithm, we proposed an adaptive parameter-adjusting algorithm for automatic parameter optimization of the proposed detection algorithm. Specifically, global parameter optimization and local parameter optimization were proposed. In the way of roulette, global optimization was executed; then, the probability of global optimization was reduced, and the probability of local optimization was raised. In this way, the adaptive parameter-adjusting algorithm can rapidly identify the approximate ranges of parameters at the beginning, and local optima of parameters can be determined in local ranges.

To validate the effectiveness of the proposed DFPO algorithm, evaluation metrics such as F-measure, accuracy (ACC), maximum matching ratio (MMR), Frac, coverage rate (CR), and their sum (Total score) were employed to analyze the algorithm's performance and compare methods in the detection of protein complexes. The experimental results revealed that the proposed DFPO algorithm was superior to other excellent algorithms regarding the evaluation metrics involved.

The main contributions are summarized as follows:

- This paper proposes a new method for protein complex detection, the DFPO algorithm (DPC-FCM Parameter Optimization).
- This paper proposes a new method that combines the improved density peaks clustering algorithm (DPC) and the fuzzy C-means clustering algorithm (FCM) to overcome the shortcomings of the traditional FCM algorithm.
- This paper designed a new objective function for the FCM algorithm based on 'high cohesion' and 'low coupling'.
- An adaptive parameter-adjusting algorithm was designed to optimize the parameters of the proposed DFPO algorithm.

The remaining chapters of this paper are as follows. Section 2 briefly introduces the algorithms involved in this study and describes the proposed algorithm. Section 3 presents the experiment and results, including datasets, evaluation metrics, and results analysis. Section 4 gives conclusions and a future outlook.

2. Materials and Methods

2.1. Problem Definition

The PPI network can be represented by a classical graph model $G = (V, E)$ corresponding to a binary tuple. Herein, V corresponds to the protein aggregate in PPI network G, and E corresponds to the interactions in PPI network G. The interconnection relationship in G can be expressed by e_{ij}. If $e_{ij} \in E$, then v_i and v_j have a connection. In this study, the adjacent matrix $D = [d_{ij}](1 \leq i, j \leq n)$ was used to describe G. d_{ij} can be determined by Equation (1):

$$d_{ij} = \begin{cases} 1, & e_{ij} \in E \\ 0, & others \end{cases} \tag{1}$$

Identification of protein complexes in the PPI network can be regarded as a soft division of V proteins in the PPI network into K cluster sets, namely $V = \cup_{k=1}^{K} C_k$. Each cluster C_k is regarded as one detected protein complex. Figure 1 illustrates the mining of protein complexes in the PPI network.

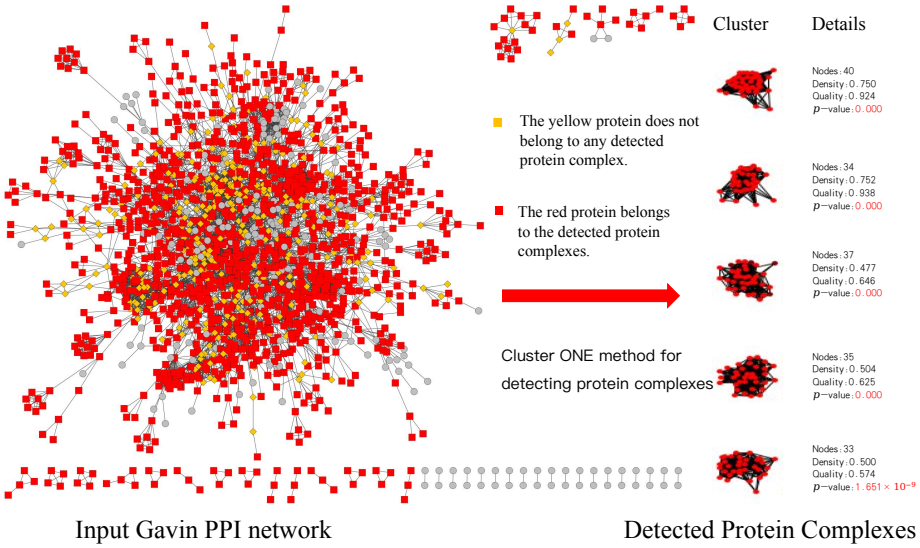

Figure 1. Schematic of the identification of protein complexes in the PPI network based on Cytoscape.

An excellent method for the detection of the protein complex is supposed to achieve a balance between the following aspects: (1) proteins in the same protein complex are

closely connected, while proteins in different protein complexes are sparsely connected; (2) proteins in the same protein complex have similar attributes, while proteins in different protein complexes have different attributes; and (3) different protein complexes may have overlapping proteins, indicating that overlapping proteins participate in the generation of multiple protein complexes.

2.2. Relevant Algorithms

2.2.1. DPC Algorithm

DPC is an algorithm for clustering based on fast search and density peak [62]. The core features of the DPC algorithm are as follows: (1) the density of nodes near the clustering center is lower than that of the clustering center and (2) the distances between two clustering centers are far [63]. The specific steps of the algorithm are as follows.

(1) Calculate similarity and establish a matrix

Assuming a dataset of $X = \{x_1, x_2, \cdots, x_n\}$ with a size of n, where d_{ij} is the Euclidean distance between x_i and x_j and denotes the distance measurement between them, then it is denoted in Equation (2):

$$d_{ij} = ||x_i - x_j||^2 \tag{2}$$

Then, the similarity of all data was calculated, and a matrix with distance as the content was established, and is defined in Equation (3):

$$D = [d_1, d_2, \ldots, d_n]^T \in R_{n*n} \tag{3}$$

where D is the established matrix, which is a symmetric one.

(2) Calculate the local density and relative distance of sample points

The local density can be calculated by Equation (4):

$$\begin{cases} \rho = \sum_j X(d_{ij} - d_c) \\ X(x) = \begin{cases} 1, x < 0 \\ 0, x \geq 0 \end{cases} \end{cases} \tag{4}$$

where d_c is the cut-off distance, the only input parameter that shall be set manually.

The relative distance is between the current sample point and the closest point with a larger local density. It can be calculated by Equation (5):

$$\sigma_i = \begin{cases} \min\limits_j(d_{ij}), if \ \ \exists j : \rho_j > \rho_i \\ \max\limits_j(d_{ij}), otherwise \end{cases} \tag{5}$$

A decision diagram was developed with σ and ρ as X and Y axes. Points with high σ and ρ in the decision diagram were selected as clustering centers of the DPC algorithm [64]. In contrast, other points were categorized into the classification containing high-density sample points closest to the respective point.

2.2.2. FCM Clustering Algorithm

In the FCM algorithm, the objective function was designed, the membership of all sample points for each clustering center was determined by constantly updating the membership matrix and the clustering center location, and the center point of the maximum membership was selected as its category to achieve clustering [65]. The steps are as follows.

$X = \{x_1, x_2, \cdots, x_n\}$ is present and $x_i = \{x_{i1}, \ldots, x_{in}\}$ is present for each x_i. The constraint was set as $\sum_{j=1}^{C} u_{ij} = 1$, where i is a set of natural numbers. The traditional objective function of FCM is as follows in Equation (6):

$$J(u, c, k) = \sum_{i=1}^{N} \sum_{j=1}^{K} u_{ij}^{m} d(x_i, c_j) \tag{6}$$

where K is the number of clustering category; u_{ij} is the membership matrix; m is the weighted index and $m > 1$; $d(x_i, c_j)$ is the Euclidean distance from x_i to c_j. Herein, x_i is the current sample point, and c_j is a clustering center.

u_{ij} and c_j can be described by Equations (7) and (8), respectively:

$$u_{ij} = [\sum_{k=1}^{c} (\frac{d(x_i, c_j)}{d(x_i, c_k)})^{\frac{1}{m-1}}]^{-1}, i = 1, 2, \ldots, N; j = 1, 2, \ldots, K \tag{7}$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} x_i}{\sum_{i=1}^{N} u_{ij}^{m}}, j = 1, 2, \ldots, K \tag{8}$$

In traditional FCM algorithms, the difference in membership matrices after two iterations serves as the termination condition. Before reaching the termination conditions, iterations continue until better clustering results are obtained.

2.3. DFPO Algorithm

Methods for the detection of overlapping protein complexes were investigated. The DFPO algorithm comprises preprocessing of the PPI network, determination of the initial clustering center and number of the FCM algorithm by using the improved DPC algorithm, protein clustering by using the FCM algorithm, parameter optimization of the detection algorithm by using adaptive parameter-adjusting algorithms, and evaluation of the effectiveness of the proposed algorithm based on various evaluation metrics. Figure 2 illustrates the flowchart of the present study.

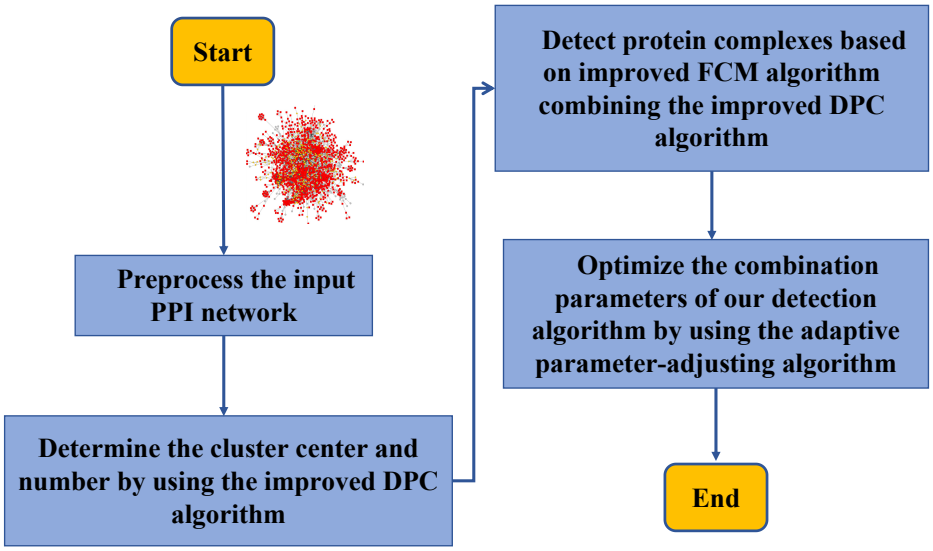

Figure 2. The flowchart of the presented DFPO algorithm.

In this study, protein complex detection is described as the optimization of an objective function. In the FCM algorithm, the selection of the initial clustering center significantly affects clustering performance. Hence, the improved DPC algorithm was proposed to

determine the initial clustering center and number in the FCM algorithm. Then, protein complexes were mined by using the FCM algorithm. Meanwhile, the FCM algorithm involves considerable parameters, and manual parameter adjustment is tedious and costly. Therefore, an adaptive parameter-adjusting algorithm was proposed in this study to optimize parameters in the FCM algorithm. Substeps of the proposed DFPO algorithm will be described below. In this study, the detection of the protein complex is described as the optimization of an objective function. In the FCM algorithm, the selection of the initial clustering center significantly affects clustering performance. Hence, the improved DPC algorithm was proposed to determine the initial clustering center and number in the FCM algorithm. Then, protein complexes were mined by using the FCM algorithm. Meanwhile, the details of the DPC-FCM algorithm are shown in Algorithm 1.

In addition, the FCM algorithm involves considerable parameters and manual parameter adjustment is tedious and costly. Therefore, in this study, an adaptive parameter adjustment algorithm was proposed to optimize the parameters in the FCM algorithm. Substeps of the proposed DFPO algorithm will be described below.

Algorithm 1 DPC-FCM algorithm

Input: G: the PPI network;
Output: LiPCs: the set of predicted protein complexes;
1: **initialize** Identified protein complexes, $IiPCs = \varnothing$; $param_K = 5$; $param_{density} = 0.04$; $param_{loss} = 0.045$; $param_{divide} = 1.9$;
2: **Step 1:** Preprocessing of PPI network, and construct a graph $G = (V, E)$;
3: **Step 2:** Generating the clustering center and number based on improved DPC algorithm;
4: Calculating the weight of interacting edges using Equation (13);
5: Calculation of local density of sample points using Equation (15);
6: Calculation of scores of protein sample points using Equation (16) and $param_{divide}$;
7: Selection of clustering center using Equation (17) and $param_{density}$;
8: **Step 3:** Improved FCM algorithm combining the improved DPC algorithm to detect protein complexes;
9: Determining initial clustering center and number of the FCM algorithm using Equation (18);
10: Calculating the value of objective function based on Equation (19);
11: Determination of category of each protein;
12: **Step 4:** The proposed adaptive parameter-adjusting algorithm is used to optimize combination parameters including $param_K$, $param_{density}$, $param_{loss}$, $param_{divide}$;
13: Calculating the value of objective function based on Equation (19) and $param_{loss}$;
14: Determination of category of each protein and $param_K$;
15: Obtain the initial identified protein complexes, $IiPCs$;
16: **return** Initial identified protein complexes, $IiPCs$.

2.3.1. Preprocessing of PPI Network

A data sample in PPI networks is a set of protein interactions instead of a discrete point. Hence, a PPI network is developed to serve as the input of the proposed detection algorithm. Since discrete data are the input required by the subsequent detection algorithm, while the data in the input PPI network dataset are graph data, it is necessary to preprocess the input PPI network to facilitate the subsequent processing of the detection algorithm.

After modeling and visualizing, the PPI network can be regarded as a Graph G, which comprises protein aggregate V and edge set E, and each maximal connected subgraph in the graph is highly likely to be the protein complex to be detected. First, an adjacent matrix with the size of $(n \times (n+1))$ was established to mark the presence of an edge between the two proteins (1 if an edge is present and 0 if no edge is present). The content of column $n+1$ is the density peak score of each protein, which can be calculated by the method described in Section 2.3.2. Upon establishment of the adjacent matrix, if the value at the ith ($i \leq n$) row and the jth ($j \leq n$) column was 0, no edge is present between the ith protein

and the *j*th protein; if the value was 1, an edge is present between the *i*th protein and the *j*th protein. The vector generated by the *i*th row is the vector of the *i*th protein and is used to calculate and process subsequent detection algorithms.

2.3.2. Improved DPC Algorithm

The improved DPC algorithm was developed based on the traditional DPC algorithm to determine the clustering center and cluster number in the improved FCM detection algorithm. The data entered in this paper are graph data, which differ from the discrete data used by DPC; thus, data representation conversion is required. The FCM and automatic parameter optimization algorithms also require continuous iterations, resulting in massive time consumption. Hence, the proposed algorithm shall accurately identify the clustering center and cluster number to reduce time consumption and facilitate subsequent algorithm optimization.

In DPC, the local density of each point and its distance from the closest point with a more significant local density were calculated, and the point with a more significant density and longer distance was employed as the clustering center. Similarly, the local density of each point was calculated in this study, and this local density was regarded as the initial score for this point. However, graph data were used in the proposed algorithm, and the determination of local density is different from that in the original DPC algorithm. First, traverse the neighbor protein of this sample point to verify the presence of a protein with a higher local density. If so, this neighbor protein is more suitable as a clustering center. In other words, the effect of this sample point as a clustering center is slightly worse. Hence, the initial score of this point can be reduced, and the reduced result can be used to calculate the final protein score. Based on this score, parameters are set, the distance is determined, and an appropriate clustering center is selected.

(1) Calculation of weights of interacting edges

The weight of interacting edges in the PPI network was calculated to calculate protein density based on this weight. Indeed, this weight considers four attributed information of proteins, including subcellular localization information, gene co-expression data information, functional annotation information, and shared neighbor information of two interacting proteins. In this way, interacting edges were weighted and used as the score of the reliability of interacting edges.

In Equation (9), $weight_{SL}$ represents each protein's subcellular localization attribute weight. Two proteins that have various attributes of the same location are readily exposed to interaction and are defined in Equation (9):

$$weight_{SL} = \frac{n_{ij}}{min(n_i, n_j)} \tag{9}$$

where n_{ij} defines the number of subcellular localization attributes shared by the two proteins, n_i defines the number of subcellular localization attributes of protein i, and n_j defines the number of subcellular localization attributes of protein j.

As the weight of gene expression information of two interacting proteins, $weight_{GE}$ can describe the co-expression similarity of two proteins, which was calculated based on the person correlation coefficient and is defined in Equation (10):

$$weight_{GE} = \frac{\sum(x_i - \overline{x})(y_i - \overline{y}))}{\sqrt{(\sum x_i - \overline{x})^2}\sqrt{(\sum y_i - \overline{y})^2}} \tag{10}$$

$weight_{CN}$ describes the topological structure similarity of two interacting proteins, meaning that the interaction similarity of the two proteins is depicted by the number of neighbor nodes shared by the two proteins, and is defined in Equation (11):

$$weight_{CN} = \frac{n_{ij}^2}{(n_i) \times (n_j)} \tag{11}$$

where n_{ij} denotes the number of neighbor nodes shared by the two proteins, n_i denotes the number of neighbor nodes the ith protein, and n_j denotes the number of neighbor nodes of the jth protein.

$weight_{GO}$ contains a series of functional annotation attributes of proteins. The $weight_{GO}$ values of the two proteins were vectorized, and the cosine similarity of the two protein vectors was calculated to determine the interaction similarity of two interacting proteins, which is defined in Equation (12):

$$weight_{GO} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i^2} \times \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{12}$$

The average weight was obtained by combining the four calculation methods mentioned above, and the method was employed for the calculation of weights of interacting edges in the proposed algorithm and is defined in Equation (13):

$$weight = \frac{weight_{SL} + weight_{GE} + weight_{CN} + weight_{GO}}{4} \tag{13}$$

(2) Calculation of the local density of sample points

In a specific maximal connected subgraph, one protein sample point was selected, as well as the sum of weights of all edges of neighbor proteins adjacent to this protein sample point (see Equation (14)). Then, the sum was divided by the number of all possible edges to obtain the local density of the protein sample point. In other words, the total weight of the edges of the protein sample point and its neighbor proteins was assigned to all possible edges, which is consistent with the definition of density in Equation (15):

$$Sw = \sum_{i=1}^{N} weight_i \tag{14}$$

$$\rho = \frac{2 \times Sw}{n_s \times (n_s - 1)} \tag{15}$$

where $weight_i$ denotes the weight of each edge in the current neighbor subgraph, Sw denotes the sum of weights of all edges in the neighbor subgraph consisting of the protein sample point and its direct neighbor proteins, and n_s denotes the number of proteins in the current neighbor subgraph.

(3) Calculation of scores of protein sample points

In the traditional DPC algorithm, a protein sample point with a density more significant than that of the protein sample point and the minimum distance shall be identified after obtaining the local density of the protein sample point, and the distance between the two protein points shall be calculated. The clustering center shall be determined from local density and distance perspectives. However, the calculation of the distance between two protein nodes in the proposed algorithm is complicated as the data used are graph data. Additionally, successive identification of qualified protein points and distance calculation in the case of a dataset containing various protein sample points could

be time-consuming. As the FCM and adaptive parameter-adjusting algorithm would be considerably time-consuming, it is essential to propose an appropriate method to reduce time costs.

As discussed, the selection of a clustering center is closely related to the local density. Indeed, a protein with high local density is suitable as the clustering center. However, various proteins in one dense subgraph may have large local densities, and 'low coupling' would be violated if all of them were used as clustering centers. Hence, a protein with higher local density shall be identified in this dense subgraph as the clustering center. In the proposed algorithm, the distance of each protein pair is not calculated. Hence, all proteins that are mutually adjacent nodes may be selected as the clustering centers, which is irrational if a threshold is used as the criteria for selecting clustering centers (proteins with a density greater than the threshold are selected as clustering centers). It is proposed in this study that for the current protein sample point (x_i), the presence of a neighbor protein (x_j) with a higher local density suggests that x_i is more unsuitable as a clustering center in this dense subgraph compared with x_j. It has been confirmed that proteins with high local density are suitable as clustering centers. Therefore, it can be deduced that a protein unsuitable for a clustering center is supposed to have a low local density. As a result, the probability of a protein being selected as a clustering center can be reduced by reducing its local density. In this study, the local density of sample point (ρ_i) obtained in Step 1 was employed as the initial score of xi and successively compared with the initial scores of its neighbor proteins. If a protein with a higher initial score is found, the score of the current protein is reduced, and the score obtained after traversing all neighbor nodes is the final score of the current protein (*Score*), as defined in Equation (16):

$$Score = \frac{Score_0}{param_{divide}} \tag{16}$$

where *Score*0 is the initial score and $param_{divide}$ is a parameter reflecting the scaling factor of the density score. $param_{divide}$ was employed as a parameter of the adaptive parameter-adjusting algorithm and optimized.

(4) Selection of the clustering center

$param_{density}$ was set to be a parameter for selecting the clustering center. Specifically, protein sample points with scores more significant than $param_{density}$ were regarded as clustering centers, and it was determined whether these proteins were suitable as clustering centers. For protein x_i, which has been selected as a clustering center, x_j, which is the next protein to be selected as a clustering center, cannot be selected as a clustering center if it has an over-small distance from x_i (<1.0 in this case), in order to ensure that the distance between two clustering centers is not over-small. The low coupling clustering algorithm would eventually select several clustering centers. These protein points would be used as initial clustering centers in the FCM algorithm, and the number of these protein points would be used as an input parameter. The distance of two protein centers can be calculated by Equation (17):

$$d_{ij} = ||x_i - x_j||^2 \tag{17}$$

2.3.3. Improved FCM Algorithm Combining the Improved DPC Algorithm

Despite its good performance in tackling specific clustering issues, traditional FCM algorithms exhibit some defects. For instance, the selection of initial points significantly affects the results. Specifically, good clustering performance can be expected in rational selection of random initial points; if random initial points are not appropriately selected (i.e., several initial points at edges), the clustering performance would not be as expected. Additionally, the clustering center number selection significantly affects the algorithm's

performance. This study combined the proposed improved DPC algorithm with the FCM algorithm. Specifically, the clustering center was determined by the improved DPC algorithm, and the number of FCM algorithms for reclustering was used as the initial clustering center.

Additionally, traditional FCM algorithms rely on determining the objective function based on the product of membership and Euclidean distance of the two protein points. For this reason, traditional FCM algorithms only consider the high cohesion state within each community structure; that is, the points in the same community structure are closely connected, while the fact that points in different community structures are sparsely connected is not considered. This study proposes a new novel objective function for clustering so that both high cohesion and low coupling are considered.

(1) Determination of the initial clustering center and number of the FCM algorithm

As mentioned in Section 2.3.2, clustering center and number were determined using the improved DPC algorithm and used as parameter input of the FCM algorithm. Herein, the clustering center set and the number of clustering centers were set to be the initial clustering center and initial cluster number, respectively. This method is superior to the FCM algorithm with a randomly initialized clustering center and number in terms of clustering performance.

(2) Design of the objective function

Traditional FCM algorithm considers 'high cohesion' but not 'low coupling' of different categories. In this study, the design of the objective function would integrate 'high cohesion, low coupling' with the detection of the protein complex to optimize an objective function. For 'high cohesion', the product of membership and Euclidean distance of the two points (see Equation (6)) was used for calculation. For 'low coupling', the following method is proposed: for x_j, the closest clustering center protein (x_k) was identified and the offset direction of x_j can be obtained by subtracting the vector of x_k from the vector of x_j. In this way, the distances between clustering centers are large enough to meet the requirements of 'low coupling', and it is defined in Equation (18):

$$A = \sum_{j=1}^{K} \min_{j \neq k} \|c_j - c_k\|^2 \tag{18}$$

where K represents the number of proteins as clustering center, c_j represents the vector of current clustering center protein, c_k represents the vector of the closest clustering center protein, and $\|c_j - c_k\|^2$ represents the distance from c_j to c_k.

As a function designed to realize high cohesion in clustering, J should be as small as possible; as a function designed to realize high cohesion in clustering, A should be as large as possible, and it can realize low coupling in different clusters. Additionally, an overlarge distance between two clustering centers may lead to cases where marginal points of edges serve as clustering centers, resulting in poor clustering performances. In this study, the weights of J were set as one and A is set as a value less than 1 in the design of objective function to achieve improved clustering performance, and is defined in Equation (19):

$$J_{loss} = J - param_{loss} A \tag{19}$$

where $param_{loss}$ is a parameter less than one, and it denotes the weight of A in the objective function. $param_{loss}$ as one of the parameters was adjusted in the adaptive parameter-adjusting algorithm.

(3) Determination of the category of each protein

In traditional clustering of discrete points, the membership matrix of each protein and all clustering centers was determined, while the clustering center where the maximum membership is located was selected; the membership matrix was classified into the community where the clustering center is located. This is not the case for PPI networks. Indeed, each protein in the PPI network may participate in the generation of multiple protein complexes. In other words, each protein can belong to multiple protein complexes, meaning that this protein could be overlapping. First, the weights of each protein belonging to different clustering centers were calculated according to the membership matrix. Then, all clustering centers were sorted in descending order according to this weight. After that, clustering centers were placed at the top of the list according to the weight, and this protein was added to the protein complexes represented by these clustering centers. Herein, this number was set to be $param_K$, which reflects the number of protein complexes to which each protein belongs. Clustering of the input PPI network was conducted using the improved FCM algorithm according to the objective function to obtain protein complexes finally.

2.3.4. Adaptive Parameter Adjusting Algorithm

The details of the Adaptive Parameter Adjusting Algorithm are shown in Algorithm 2.

(1) Fitness function

Inspired by the ABC algorithm, we proposed an adaptive parameter-adjusting algorithm PO (Parameter Optimization). First, an appropriate fitness function was designed and selected. Each protein complex's weighted modularity score (modularity) was calculated [26], and the sum of weighted modularity scores of all protein complexes in the detected protein complexes was determined. Then, the sum of the weighted modularity score was divided by the number of detected protein complexes to obtain the fitness of parameter adjustment in the DFPO algorithm.

First, the sum of weights of all edges in each protein complex s ($weight_{in}$) was calculated. For proteins in the current protein complex s, if a neighbor node that does not belong to the current protein complex s is present, the sum of weights of all edges connecting neighbor proteins and the current protein complex s ($weight_{out}$) was calculated. The weighted module score of the current protein complex s can be obtained by dividing the former by the sum, as shown in Equation (20):

$$modularity(s) = \frac{weight_{in}}{weight_{out}} \tag{20}$$

Finally, weighted module scores of all detected protein complexes were determined and divided by the number of protein complexes, and the obtained average modularity score of protein complex was used as the fitness of the DFPO algorithm (M) is defined in Equation (21):

$$M = \frac{\sum_{k=1}^{n} modularity(s_k)}{N} \tag{21}$$

where $weight_{in}$ defines the sum of weights of all edges in one detected protein complex, $weight_{out}$ defines the sum of weights of all edges of proteins in one detected protein complex and neighbor proteins in this detected protein complex, and N defines the number of detected protein complexes.

Algorithm 2 Adaptive parameter-adjusting algorithm

Input: The weighted PPI network, $G(V, E, W)$; Initial identified protein complexes, $IiPCs$; $param_K$ = 5; $param_{density}$ = 0.04; $param_{loss}$ = 0.045; $param_{divide}$ = 1.9; $param_K^{stride}$ = 1; $param_{density}^{stride}$ = 0.01; $param_{divide}^{stride}$ = 0.05; $param_{loss}^{stride}$ = 0.005; $param_K^{min}$ = 3; $param_K^{max}$ = 7; $param_{density}^{min}$ = 0.01; $param_{density}^{max}$ = 0.3; $param_{divide}^{min}$ = 0.8; $param_{divide}^{max}$ = 1.5; $param_{loss}^{min}$ = 0.02; $param_{loss}^{max}$ = 0.06; $epoch_{max}$ = 100; R_{value} = 200;
Output: Identified protein complexes, $IPCs$;

initialize $i = 0$; $J_{loss}^{best} = 0.0$
while $epoch < epoch_{max}$ **do**
  if $np.random.uniform(0, 500) > R_{value}$ **then**
    $which_{param}$ = np.random.uniform(0, 8);
    if $which_{param} < 2$ **then**
      $param_K = param_K + np.random.uniform(-param_K^{stride}, param_K^{stride})$
    else if $which_{param} < 4$ **then**
      $param_{density} = param_{density} + np.random.uniform(-param_{density}^{stride}, param_{density}^{stride})$
    else if $which_{param} < 6$ **then**
      $param_{divide} = param_{divide} + np.random.uniform(-param_{divide}^{stride}, param_{divide}^{stride})$
    else
      $param_{loss} = param_{loss} + np.random.uniform(-param_{loss}^{stride}, param_{loss}^{stride})$
    end if
  else
    $param_K = np.random.uniform(param_K^{min}, param_K^{max})$
    $param_{density} = np.random.uniform(param_{density}^{min}, param_{density}^{max})$
    $param_{divide} = np.random.uniform(param_{divide}^{min}, param_{divide}^{max})$
    $param_{loss} = np.random.uniform(param_{loss}^{min}, param_{loss}^{max})$
  end if
  if $R_{value} > 50$ **then**
    $R_{value} = R_{value} - 1$
  end if
  current identified protein complexes ($CIPCs$), J_{loss}^{i} = DPC-FCM Algorithm 1 ($param_K, param_{density}, param_{divide}, param_{loss}$)
  if $J_{loss}^{i} > J_{loss}^{best}$ **then**
    $param_K^{best} = param_K$
    $param_{density}^{best} = param_{density}$
    $param_{divide}^{best} = param_{divide}$
    $param_{loss}^{best} = param_{loss}$
  else
    $param_K = param_K^{best}$
    $param_{density} = param_{density}^{best}$
    $param_{divide} = param_{divide}^{best}$
    $param_{loss} = param_{loss}^{best}$
  end if
  i = i + 1;
end while
Obtain the finally identified protein complexes ($FIPCs$), J_{loss}^{best} = DPC-FCM Algorithm 1 ($param_K^{best}, param_{density}^{best}, param_{divide}^{best}, param_{loss}^{best}$);
return Finally Identified protein complexes, $FIPCs$.

(2) Adaptive parameter-adjusting algorithm

As an SIO algorithm, the ABC algorithm categorizes bees during honey collection as foragers, observers, or scouts. Its objective is to identify the nectar source with the most materials. After determining the fitness of each nectar source, foragers would continue to search for new sources based on the greedy strategy. Then, observers would select

one nectar source according to the probability. A new nectar source would be acquired by random disturbance at this nectar source, accompanied by a calculation of its fitness. If observers fail to find a better source after multiple trials, the forager who locates this source will become a scout to find a new one.

The proposed DFPO algorithm involves identifying new nectar sources by observer disturbance, as in the ABC algorithm, and global optimization to avoid local optimal solutions. As shown in Figure 3, there are six steps: (1) the iteration times of the adaptive parameter-adjusting algorithm were determined and used as termination conditions of the proposed algorithm. The iteration times shall be sufficient to make the results convincing. Meanwhile, the FCM algorithm would also iterate multiple times to search for an optimal solution as the improved DPC algorithm identifies clustering centers for the FCM algorithm. Therefore, the iteration times of the FCM algorithm can be a manageable size, as an optimal solution can be obtained after several iterations, and further iterations, which are time-consuming, can barely identify better results. Overall, the iteration times were set to be 100 in this study to balance the time and performance. (2) A constant C_1 was set, and a random number (R_1) was generated within the designated range. If $R_1 > C_1$, local parameter optimization was executed; otherwise, global parameter optimization was executed. Meanwhile, C_1 decreased as the iteration times increased so that it is guaranteed that the probability of identifying an optimal solution in the global scope decreases during the entire process of parameter optimization. In other words, we hope that the adaptive parameter-adjusting algorithm can first identify an appropriate parameter in the global scope to determine an approximate range before adjusting each parameter to identify a more suitable parameter combination. (3) Global parameter optimization refers to the random disturbance of all parameters in the DFPO algorithm. In local parameter optimization, a random number (R_2) of 0∼8 is generated: if $0 \leq R_2 < 2$, the first parameter $param_K$ is disturbed; if $2 \leq R_2 < 4$, the second parameter $param_{density}$ is disturbed; if $4 \leq R_2 < 6$, the third parameter $param_{divide}$ is disturbed; if $6 \leq R_2 < 8$, the fourth parameter $param_{loss}$ is disturbed. After parameter disturbance in each iteration, the DFPO algorithm's fitness was calculated using this parameter combination. (4) The fitness of the DFPO algorithm was recalculated for parameter combination after each disturbance. Then, the old fitness was substituted by the new fitness, and the new parameter combination substituted the old parameter combination if the new fitness was larger than the old fitness; otherwise, no steps were taken. (5) Iterations are terminated if the iteration times meet the termination conditions; otherwise, iterations are continued. (6) Clustering of the input PPI network was executed with the optimized parameter combination as the input parameter in the DFPO algorithm to obtain the detected protein complex set.

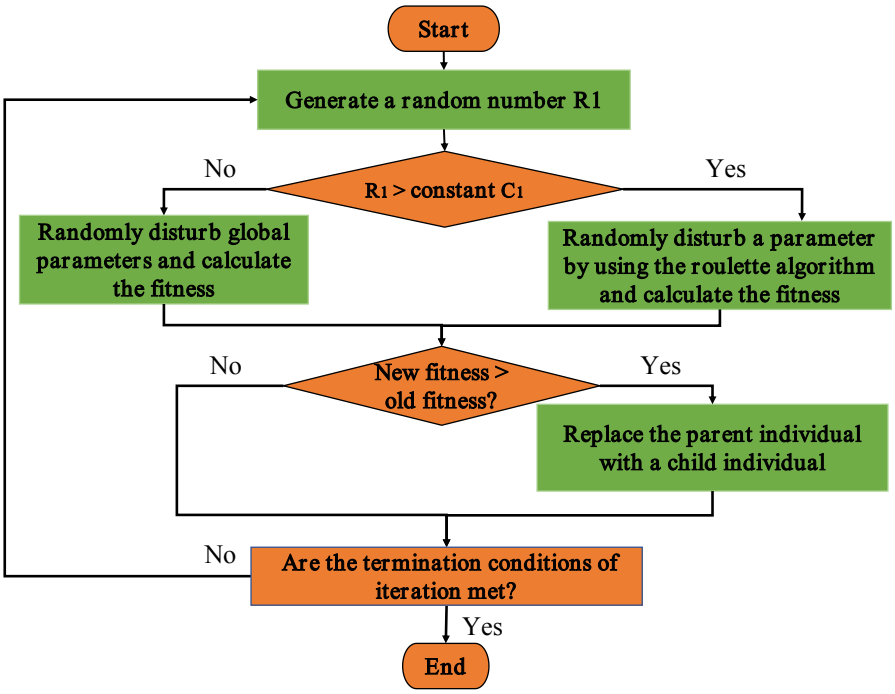

Figure 3. The flowchart of the adaptive parameter-adjusting algorithm for DFPO.

3. Results

3.1. Experimental Datasets

Three PPI networks (Collins [66], Gavin [67], and Krogan [12]) were employed for verification of the proposed algorithm. In these datasets, self-interaction and repeated interaction of proteins are removed. Table 1 shows detailed information on these PPI networks.

Table 1. The information of PPI networks used in this paper.

Datasets	The Number of Proteins	The Number of Interactions	Density
Krogan	2674	7075	0.00198
Gavin	1855	7669	0.00446
Collins	1622	9074	0.00690

This study selected two standard protein complexes with high coverages to assess protein complexes in the PPI network detected by the proposed detection algorithm. The standard protein complexes 1 comprises known protein complexes, such as TAP06 [67], MIPS [68], SGD [69], and ALOY [70]; the standard protein complexes 2 comprises Wodak-database [71] and PINdb and GO complexes [72] datasets, as shown in Table 2. In most cases, the standard protein complexes can serve as a data source as a standard dataset can provide reliable evidence for physical interactions.

Table 2. The information of standard protein complexes used in this paper.

Datasets	The Number of Protein Complexes	The Number of Proteins	Average Size
standard protein complexes 1	812	2773	8.92
standard protein complexes 2	1045	2778	8.97

The two standard protein complex datasets can be employed to effectively verify the performance of the protein complex detection algorithm and assess the matching rate of detected protein complexes and the protein complexes in the standard protein complexes.

3.2. Evaluation Metrics

(1) F-measure

As a harmonic mean of accuracy and recall, the *F-measure* is typically comprehensive for assessment methods. Meanwhile, the *F-measure* can assess the overall performance of detection methods. The *F-measure* is defined in Equation (22):

$$F\text{-}measure = \frac{2 \times precision \times recall}{precision + recall}. \tag{22}$$

(2) Accuracy

The sensitivity can be effectively enhanced if all proteins are in the same protein complex. Assigning these proteins to their respective protein complex can maximize the positive predictive values. Therefore, it is essential to balance their impacts on detection algorithms based on detection accuracy; that is, the geometric mean of sensitivity and positive predictive value (ACC). T_{ij} is the number of proteins. These proteins are included in the standard protein complex S_i and the detected protein complex D_j. Then, Sn and PPV are calculated by $Sn = \frac{\sum_{i=1}^{|S|} \max_{j=1}^{|I|} \{T_{ij}\}}{\sum_{i=1}^{|S|} N_i}$ and $PPV = \frac{\sum_{j=1}^{|D|} \max_{i=1}^{|S|} \{T_{ij}\}}{\sum_{j=1}^{|D|} \sum_{i=1}^{|S|} T_{ij}}$, respectively. ACC is defined by Equation (23):

$$ACC = \sqrt{Sn \times PPV}. \tag{23}$$

(3) MMR

Nepusz et al. [26] proposed the MMR as an evaluation metric. Herein, a bipartite graph reflected the matching degrees of all standard protein complexes with protein complexes detected by the protein complex detection algorithm. Indeed, the bipartite graph comprises standard protein complexes and detected protein complexes. According to this detection method, a subset of the maximum weight-matching edges in this bipartite graph was selected, and the selected edges indicate the maximum matching degree between the standard protein complexes and the detected protein complexes. As a result, the standard protein complex does not match any other detected protein complex, and vice versa. The MMR of the standard and detected protein complexes is essentially the ratio of the sum of weights of all selected edges and the number of standard protein complexes. Overall, the proposed method can accurately and effectively assess the matching of detected protein complexes with standard protein complexes.

(4) Coverage rate

Coverage rate (CR) [73,74] reflects the number of proteins in the standard protein complexes covered by proteins in the detected protein complexes. A high CR indicates many proteins in the standard protein complex covered by the protein complex detection algorithm, suggesting good detection performance. With a standard protein complex set (S) and a protein complex detection set (P), an optimized matching matrix (T) was established so that each element in T denotes the number of proteins shared by the standard protein complex and the jth detected protein complex is maximized. CR is defined by Equation (24):

$$CR = \frac{\sum_{s=1}^{|S|} max\{T_{st}\}}{\sum_{s=1}^{|S|} N_s}, \tag{24}$$

(5) Frac score

The Frac score [26] is used to compare the percentage of standard protein complexes matched with detected protein complexes. Herein, S and D are the standard and predicted protein complexes, respectively. The threshold (ω) was set to be 0.25 so that at least 50% proteins in the standard protein complex can be the same as the matched detected protein complex. The Frac score can be obtained by Equation (25):

$$Frac\ score = \frac{N_s}{|S|}, N_s = |s|s \in S, \exists d \in D, OS(d,s) \leq w \tag{25}$$

(6) Total score

To assess the performance of protein complex detection algorithms from multiple aspects, we consider multiple evaluation metrics, including the *F-measure*, *ACC*, *MMR*, *CR*, and *Frac*, and add them to obtain the Total score, which was employed to verify algorithm performance. Intuitively, a high Total score denotes strong recognition capability and good performance of this algorithm. The total score can be calculated by Equation (26):

$$Total\ score = F\text{-}measure + ACC + MMR + Frac + CR. \tag{26}$$

3.3. Experimental Results

3.3.1. Selection of Compared Algorithms and Parameter Setting

To compare the proposed algorithm with other compared algorithms, 12 protein complex detection algorithms (including MCODE [19], CMC [20], SPICi [25], CPredictor2.0 [27], PC2P [29], COACH [31], WPNCA [32], PEWCC [34], ICJointLE [35], SE-DMTG [36], MPC-C [37] and ClusterEPs [41] were selected, and their datasets were aligned with the proposed algorithm's dataset. Meanwhile, these algorithms were applied in these datasets to detect protein complexes in the PPI network. All parameters were set as suggested so that all algorithms could perform rationally. Table 3 lists the parameters set.

Table 3. Parameters of each method used in the study.

ID	Algorithm	Parameters
1	MCODE	minimum cluster size = 3 (default setting)
2	COACH	w = 0.225
3	CMC	*min_deg_ratio* = 1, *min_size* = 3, *overlap_thres* = 0.5, $merge_t hres$ = 0.25
4	SPICi	Graph mode = 0, minimum support threshold = 0.5, minimum cluster size = 3, minimum density threshold = 0.5
5	PEWCC	Overlap = 0.8, −r = 0.1, Rejoin = 0.3
6	WPNCA	lambda = 0.3, minimum cluster size = 3
7	CPredictor2.0	func_lvl = 6, Overlap threshold = 0.8, size = 3 (default setting)
8	ClusterEPs	NEPs of Complexes (minimum st = 0.4,maximum st = 0.05); NEPs of noncomplexes (maximum st = 0.05, minimum st = 0.4); maximum overlap = 0.9,Maximum size = 100
9	SE-DMTG	minimum cluster size = 3
10	ICJointLE	−L = 1, −r = 999, −d = 0.3, −c = 0.7, −f = 0.75, −p = 0.3; −m = 0.08, −u = 0.01, −e = 0.9, size = 3 (author suggestions)
11	MPC-C	Overlap threshold = 0.8, minimum cluster size = 3
12	PC2P	Minimum cluster size = 3

3.3.2. Comparison of Experimental Results Based on Statistical Metrics

Aimed at the evaluation mentioned above metrics, the performances of the DFPO algorithm and multiple protein complex detection algorithms on two standard protein complex datasets were determined and compared. As shown in Table 4, with standard protein complexes, one of which is a real protein complex, *MMR* and *CR* of the proposed DFPO algorithm in the Collins dataset were excellent; the DFPO algorithm was superior to all other algorithms in *Total score*. Meanwhile, the *Total score* of the DFPO algorithm on the Gavin dataset was excellent, and the DFPO algorithm was superior to all other algorithms in terms of *F-measure*, *MMR*, and *CR*; the sum of the metric score (*Total score*) of the DFPO

algorithm was ranked first among all algorithms involved. Additionally, the *Total score* of the DFPO algorithm in the Krogan dataset was superior to those of all other algorithms; the *CR* of the DFPO algorithm in the Krogan dataset was significantly superior to that of other algorithms. Overall, the proposed algorithm was superior to most compared algorithms in most of the evaluation metrics (especially *CR*) and the *Total score*.

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

Methods	Num	F-Measure	ACC	MMR	CR	Frac	Total Score
Collins							
ClusterEPs	587	0.5386	0.2378	0.2696	0.2384	0.5778	1.8622
CMC	177	0.6326	0.3535	0.2141	0.4585	0.7022	2.3610
COACH	251	0.6452	0.3629	0.2466	0.4674	0.7178	2.4399
CPredictor2.0	237	0.6355	0.3482	0.2710	0.4652	0.6644	2.3843
ICJointLE	214	0.6238	0.2963	0.2547	0.3420	0.6667	2.1834
MCODE	111	0.5887	0.3280	0.1575	0.4072	0.5533	2.0347
MPC-C	274	0.6206	0.3351	0.2657	0.4590	0.6178	2.2981
PC2P	159	0.6466	**0.3728**	0.2011	0.4808	0.7000	2.4014
PEWCC	426	0.6230	0.3446	0.2930	0.4530	**0.7489**	2.4626
SE-DMTG	167	**0.6468**	0.3339	0.2343	0.4123	0.6578	2.2851
SPICi	121	0.5954	0.3451	0.1651	0.4244	0.6267	2.1567
WPNCA	269	0.6199	0.3678	0.2051	0.5222	0.6644	2.3794
DFPO	361	0.6356	0.3125	**0.3052**	**0.5403**	0.7133	**2.5067**
Gavin							
ClusterEPs	271	0.6014	0.2841	0.2166	0.3656	0.6077	2.0754
CMC	294	0.5844	0.3487	0.2229	0.4501	0.7398	2.3458
COACH	361	0.6578	0.3266	0.2772	0.4428	0.7581	2.4625
CPredictor2.0	254	0.6268	0.3128	0.2285	0.3750	0.6037	2.1467
ICJointLE	243	0.6329	0.2989	0.2619	0.3557	0.6280	2.1774
MCODE	122	0.4864	0.3010	0.1205	0.3765	0.4411	1.7254
MPC-C	398	0.6369	0.3146	0.3068	0.4160	0.6098	2.2840
PC2P	219	0.5769	**0.3551**	0.1825	0.4439	0.6443	2.2026
PEWCC	664	0.6576	0.3146	0.3538	0.4316	**0.7744**	2.5321
SE-DMTG	214	0.6394	0.3187	0.2398	0.3769	0.6606	2.2354
SPICi	189	0.5777	0.3401	0.1693	0.4157	0.6341	2.1370
WPNCA	484	0.6428	0.3114	0.2557	0.4949	0.6504	2.3552
DFPO	549	**0.6590**	0.3141	**0.3563**	**0.5254**	0.7297	**2.5925**
Krogan							
ClusterEPs	410	0.5836	0.2621	0.2209	0.3352	0.5728	1.9747
CMC	264	0.4819	0.2978	0.1584	0.3656	0.5955	1.8991
COACH	345	0.5254	0.2667	0.2151	0.3473	0.5917	1.9462
CPredictor2.0	221	0.5878	0.2793	0.2119	0.3044	0.5747	1.9581
ICJointLE	216	0.5389	0.2284	0.1936	0.2206	0.5142	1.6957
MCODE	39	0.3414	0.1994	0.0403	0.2140	0.2476	1.0427
MPC-C	456	**0.5982**	0.2816	**0.2848**	0.3760	0.5955	2.1362
PC2P	249	0.4356	**0.2970**	0.1337	0.3458	0.5217	1.7338
PEWCC	389	0.5244	0.2534	0.1466	0.3208	0.4216	1.6561
SE-DMTG	372	0.5878	0.2821	0.2777	0.3504	**0.6730**	2.1710
SPICi	224	0.4444	0.2883	0.1167	0.3315	0.5180	1.6989
WPNCA	369	0.5446	0.2758	0.1912	0.3897	0.5520	1.9533
DFPO	372	0.5346	0.2686	0.2368	**0.6738**	0.5501	**2.2688**

As shown in Table 5, the performance of different algorithms in standard protein complexes 2 was lower than those of all algorithms in standard protein complexes 1. This can be attributed to the fact that standard protein complex 2 contains many protein complexes, resulting in low matching accuracy of different detection algorithms. The Total score of the proposed DFPO algorithm was higher than those of other algorithms. On the Gavin dataset, the proposed DFPO algorithm had relatively high scores in *MMR* and *CR* (scores of *CR* and *MMR* were ranked first), and its *Total score* was more significant than those of the compared algorithms. On the Krogan dataset, the *CR* score of the DFPO algorithm was significantly larger than those of the compared algorithms, while other

scores of the DFPO algorithm were reasonable; the *Total score* of the DFPO algorithm was ranked first.

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

Methods	Num	F-Measure	ACC	MMR	CR	Frac	Total Score
Collins							
ClusterEPs	587	0.4193	0.2219	0.1972	0.2154	0.4478	1.5016
CMC	177	0.4807	0.3382	0.1635	0.3901	0.5198	1.8923
COACH	251	0.4971	0.3182	0.1817	0.3961	0.5324	1.9255
CPredictor2.0	237	**0.5233**	0.3322	0.2097	0.4080	0.5360	2.0093
ICJointLE	214	0.4850	0.2771	0.1776	0.2832	0.4964	1.7193
MCODE	111	0.4344	0.3159	0.1251	0.3430	0.3975	1.6158
MPC-C	274	0.4971	0.3183	0.1904	0.3809	0.4730	1.8599
PC2P	159	0.5209	**0.3828**	0.1742	0.4318	0.5378	2.0475
PEWCC	426	0.4813	0.3178	0.2065	0.3921	**0.5522**	1.9499
SE-DMTG	167	0.5136	0.3256	0.1791	0.3460	0.4946	1.8589
SPICi	121	0.4396	0.3382	0.1259	0.3616	0.4496	1.7150
WPNCA	269	0.4602	0.3384	0.1596	0.4656	0.4658	1.8897
DFPO	361	0.4940	0.3154	**0.2159**	**0.4943**	0.5324	**2.0518**
Gavin							
ClusterEPs	271	0.4331	0.2715	0.1670	0.2906	0.4696	1.6318
CMC	294	0.3803	0.3301	0.1459	0.3575	0.4936	1.7073
COACH	361	0.4190	0.3257	0.1743	0.3505	0.5046	1.7733
CPredictor2.0	254	0.4802	0.2898	0.1721	0.3076	0.4843	1.7340
ICJointLE	243	0.4861	0.2834	0.1912	0.2920	0.5046	1.7573
MCODE	122	0.3143	0.2920	0.0863	0.3002	0.3057	1.2986
MPC-C	398	**0.4904**	0.3189	0.2128	0.3486	0.4936	1.8642
PC2P	219	0.4025	**0.3413**	0.1295	0.3610	0.4512	1.6855
PEWCC	664	0.4185	0.3137	0.2152	0.3483	**0.5304**	1.8260
SE-DMTG	214	0.4512	0.3188	0.1644	0.2997	0.4512	1.6853
SPICi	189	0.3819	0.3237	0.1158	0.3246	0.4199	1.5658
WPNCA	484	0.4217	0.3305	0.1670	0.4116	0.4567	1.7876
DFPO	549	0.4457	0.3062	**0.2225**	**0.4865**	0.4991	**1.9743**
Krogan							
ClusterEPs	410	0.4658	0.2390	0.1444	0.3021	0.4325	1.5839
CMC	264	0.3999	0.2732	0.1101	0.3192	0.4284	1.5308
COACH	345	0.4369	0.2441	0.1464	0.3166	0.4325	1.5765
CPredictor2.0	221	0.4918	0.2491	0.1396	0.2793	0.4353	1.5951
ICJointLE	216	0.4516	0.2147	0.1230	0.2083	0.3839	1.3815
MCODE	39	0.2317	0.1861	0.0271	0.1863	0.1405	0.7717
MPC-C	456	0.5178	0.2684	**0.1911**	0.3343	0.4520	1.7636
PC2P	249	0.3636	**0.2884**	0.0951	0.3141	0.3978	1.4589
PEWCC	389	0.4380	0.2358	0.0941	0.2950	0.2949	1.3560
SE-DMTG	372	**0.5060**	0.2685	0.1757	0.3093	**0.5007**	1.7602
SPICi	224	0.3484	0.2765	0.0818	0.2956	0.3491	1.3514
WPNCA	369	0.4361	0.2614	0.1250	0.3572	0.3936	1.5733
DFPO	372	0.4430	0.2619	0.1511	**0.6374**	0.3922	**1.8935**

In summary, the proposed algorithm exhibited high scores of *CR* and *Total score* metrics, which means it performed excellently in detecting protein complexes in the PPI network.

4. Conclusions

The protein complex is significant to exploring life activities and bioscience, and improving the rate and performance of protein complex detection is urgent. Various algorithms have been proposed for detecting protein complexes in the PPI network due to advances in science. Despite wide applications, these algorithms exhibit some limitations. This study proposed improvement from three perspectives: (1) current methods for de-

tecting protein complexes can barely detect overlapping protein complexes; (2) the FCM algorithm is sensitive to the selection of the initial clustering center and initial cluster number, and hence, the improved FCM algorithm was combined with the improved DPC algorithm, and a novel objective function was proposed; and (3) the improved FCM algorithm parameters were optimized using the adaptive parameter-adjusting algorithm. Finally, a dozen excellent methods for detecting protein complexes were employed to verify the effectiveness of the proposed detection algorithm. Specifically, these algorithms were applied to multiple PPI networks and the standard protein complexes, and their performances were assessed based on several evaluation metrics. The results demonstrated that the proposed algorithm was superior to other algorithms in terms of detection accuracy. Owing to advances in attention graph neural networks and machine learning, novel methods for detecting protein complexes will be proposed to enhance protein complex detection accuracy further.

Author Contributions: Conceptualization, K.J.; Methodology, C.W., R.W. and K.J.; Software, R.W. and K.J.; Validation, R.W.; Investigation, K.J.; Resources, K.J.; Data curation, R.W. and K.J.; Writing—original draft, C.W. and R.W.; Writing—review and editing, C.W.; Visualization, C.W. and R.W.; Supervision, C.W.; Project administration, C.W.; Funding acquisition, C.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by "the Fundamental Research Funds for the Central Universities" of China Foreign Affairs University (No. 3162022ZYQA01).

Data Availability Statement: The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Conflicts 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.

Abbreviations

The following abbreviations are used in this paper:

PPI	protein–protein interaction
DPC	density peaks clustering algorithm
FCM	fuzzy C-means clustering algorithm
ABC	artificial bee colony
DFPO	DPC-FCM Parameter Optimization
TAP-MS	tandem affinity purification
PPIN	protein–protein interaction network
WWW	World Wide Web
SIO	swarm intelligence optimization
MCODE	molecular complex detection
CMC	clustering based on maximal cliques
MCL	Markov clustering
CUBCO+	minimum CUt to detect Biclique spanned subgraphs as protein Complexes+
SPICi	'spicy', Speed and Performance In Clustering
ClusterONE	clustering with overlapping neighborhood expansion
PC2P	Protein Complexes from Coherent Partition
COACH	core-attachment based method
WPNCA	Weighted PageRank-Nibble algorithm and core attachment structure
PEWCC	PE-measure weighted clustering coefficient
SE-DMTG	Seed-Extended algorithm based on Density and Modularity with Topological structure and GO annotations
MPC-C	Mining Protein Complexes using a new Clustering model
ClusterEPs	EP-based clustering score and propose a search algorithm

ClusterSS	clustering with supervised and structural information
ICSC	improved cuckoo search clustering algorithm
MFO	moth-flame optimization
MP-DE	Markov clustering differential evolution(DE) algorithm
ISHC	light synchronization-based hierarchical clustering
F-MCL	MCL and its variants with Firefly algorithm
iOPTICS-GSO	improved Ordering Points to Identify the Clustering Structure (OPTICS) algorithm with Glowworm
GSO	swarm optimization algorithm
ACC	accuracy
MMR	maximum matching ratio
CR	coverage rate
Frac	Fraction
PO	Parameter Optimization
MIPS	Munich Information Center for Protein Sequence
SGD	Saccharomyces Genome Database
TAP06	Tandem Affinity Purification06

References

1. Wang, X.; Li, X.; Chen, G. *Complex Network Theory and Its Applications*; Tsinghua University Press: Bejing, China, 2006.
2. Girvan, M.; Newman, M.E.J. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 7821–7826. [CrossRef]
3. Guimera, R.; Nunes Amaral, L.A. Functional cartography of complex metabolic networks. *Nature* **2005**, *433*, 895–900. [CrossRef] [PubMed]
4. 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]
5. Radicchi, F.; Castellano, C.; Cecconi, F.; Loreto, V.; Parisi, D. Defining and identifying communities in networks. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 2658–2663. [CrossRef] [PubMed]
6. Zhao, J. Research on Adaptive Community Discovery Algorithms in Social Networks. Master's Thesis, Tianjin University, Tianjin, China, 2018.
7. Yu, Y. Research on Complex Recognition Algorithms in Protein Interaction Networks. Ph.D. Thesis, Harbin Institute of Technology, Harbin, China, 2014.
8. 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]
9. 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]
10. Barabasi, A.L.; Oltvai, Z.N. Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* **2004**, *5*, 101–113. [CrossRef] [PubMed]
11. 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]
12. Krogan, N.J.; Cagney, G.; Yu, H.; Zhong, G.; Guo, X.; Alexandr, L.; Li, J.; Pu, S.; Datta, N.; Tikuisis, A.P.; et al. Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature* **2006**, *440*, 637–643. [CrossRef] [PubMed]
13. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]
14. Eisenberg, D.; Marcotte, E.M.; Xenarios, I.; Yeates, T.O. Protein function in the post-genomic era. *Nature* **2000**, *405*, 823–826. [CrossRef] [PubMed]
15. Rigaut, G.; Shevchenko, A.; Rutz, B.; Wilm, M.; Mann, M.; Séraphin, B. A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **1999**, *17*, 1030–1032. [CrossRef]
16. Ito, T.; Chiba, T.; Ozawa, R.; Yoshida, M.; Hattori, M.; Sakaki, Y. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 4569–4574. [CrossRef] [PubMed]
17. Gavin, A.-C.; Bösche, M.; Krause, R.; Grandi, P.; Marzioch, M.; Bauer, A.; Schultz, J.; Rick, J.M.; Michon, A.-M.; Cruciat, C.-M.; et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **2002**, *415*, 141–147. [CrossRef] [PubMed]

18. Pan, Y.; Liu, D.; Deng, L. Accurate prediction of functional effects for variants by combining gradient tree boosting with optimal neighborhood properties. *PLoS ONE* **2017**, *12*, e0179314. [CrossRef]
19. Bader, G.D.; Hogue, C.W.V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinform.* **2003**, *4*, 1–27. [CrossRef] [PubMed]
20. Liu, G.; Wong, L.; Chua, H.N. Complex discovery from weighted PPI networks. *Bioinformatics* **2009**, *25*, 1891–1897. [CrossRef]
21. Van Dongen, S.M. Graph Clustering by Flow Simulation. Ph.D. Thesis, University of Utrecht, Utrecht, The Netherlands, 2000.
22. Omranian, S.; Nikoloski, Z. CUBCO+: Prediction of protein complexes based on min-cut network partitioning into biclique spanned subgraphs. *Appl. Netw. Sci.* **2022**, *7*, 71. [CrossRef]
23. Wang, W.; Meng, X.; Xiang, J.; Shuai, Y.; Bedru, H.D.; Li, M. CACO: A core-attachment method with cross-species functional ortholog information to detect human protein complexes. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 4569–4578. [CrossRef] [PubMed]
24. 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.* **2010**, *9*, 1105–1111. [CrossRef]
25. Jiang, P.; Singh, M. SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics* **2018**, *26*, 7821–7826. [CrossRef]
26. Nepusz, T.; Yu, H.; Paccanaro, A. Detecting overlapping protein complexes in protein–protein interaction networks. *Nat. Methods* **2012**, *9*, 471–472. [CrossRef] [PubMed]
27. Xu, B.; Wang, Y.; Wang, Z.; Zhou, J.; Zhou, S.; Guan, J. An effective approach to detecting both small and large complexes from protein–protein interaction networks. *BMC Bioinform.* **2017**, *18*, 19–28. [CrossRef] [PubMed]
28. Zhan, Y.; Liu, J.; Wu, M.; Tan, C.S.H.; Li, X.; Ou-Yang, L. A partially shared joint clustering framework for detecting protein complexes from multiple state-specific signed interaction networks. *Comput. Biol. Med.* **2023**, *159*, 106936. [CrossRef]
29. Omranian, S.; Angeleska, A.; Nikoloski, Z. PC2P: Parameter-free network-based prediction of protein complexes. *Bioinformatics* **2021**, *37*, 73–81. [CrossRef]
30. Lyu, J.; Yao, Z.; Liang, B.; Liu, Y.; Zhang, Y. Small protein complex prediction algorithm based on protein–protein interaction network segmentation. *BMC Bioinform.* **2022**, *23*, 405. [CrossRef]
31. 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]
32. Peng, W.; Wang, J.; Zhao, B.; Wang, L. Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *12*, 179–192. [CrossRef]
33. Von Mering, C.; Krause, R.; Snel, B.; Cornell, M.; Oliver, S.G.; Fields, S.; Bork, P. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **2002**, *417*, 399–403. [CrossRef]
34. Zaki, N.; Efimov, D.; Berengueres, J. Protein complex detection using interaction reliability assessment and weighted clustering coefficient. *BMC Bioinform.* **2013**, *14*, 163. [CrossRef]
35. Zhang, J.; Zhong, C.; Huang, Y.; Lin, H.X.; Wang, M. A method for identifying protein complexes with the features of joint co-localization and joint co-expression in static PPI networks. *Comput. Biol. Med.* **2019**, *111*, 103333. [CrossRef] [PubMed]
36. Wang, R.; Wang, C.; Sun, L.; Liu, G. A seed-extended algorithm for detecting protein complexes based on density and modularity with topological structure and GO annotations. *BMC Genom.* **2019**, *20*, 637. [CrossRef]
37. Wang, R.; Wang, C.; Liu, G. A novel graph clustering method with a greedy heuristic search algorithm for mining protein complexes from dynamic and static PPI networks. *Inf. Sci.* **2020**, *522*, 275–298. [CrossRef]
38. Yu, Y.; Kong, D. Protein complexes detection based on node local properties and gene expression in PPI weighted networks. *BMC Bioinform.* **2022**, *23*, 24. [CrossRef]
39. Yu, F.Y.; Yang, Z.H.; Tang, N.; Lin, H.F.; Wang, J.; Yang, Z.W. Predicting protein complex in protein interaction network-a supervised learning based method. *BMC Syst. Biol.* **2014**, *8*, S4. [CrossRef] [PubMed]
40. Shi, L.; Lei, X.; Zhang, A. Protein complex detection with semi-supervised learning in protein interaction networks. *Proteome Sci.* **2011**, *9*, S5. [CrossRef]
41. Liu, Q.; Song, J.; Li, J. Using contrast patterns between true complexes and random subgraphs in PPI networks to predict unknown protein complexes. *Sci. Rep.* **2016**, *6*, 21223. [CrossRef]
42. Dong, Y.; Sun, Y.; Qin, C. Predicting protein complexes using a supervised learning method combined with local structural information. *PloS ONE* **2018**, *13*, e0194124. [CrossRef] [PubMed]
43. Liu, X.; Yang, Z.; Sang, S.; Zhou, Z.; Wang, L.; Zhang, Y.; Lin, H.; Wang, J.; Xu, B. Identifying protein complexes based on node embeddings obtained from protein–protein interaction networks. *BMC Bioinform.* **2018**, *19*, 332. [CrossRef] [PubMed]
44. Sikandar, A.; Anwar, W.; Bajwa, U.I.; Wang, X.; Sikandar, M.; Yao, L.; Jiang, Z.L.; Zhang, C. Decision tree based approaches for detecting protein complex in protein protein interaction network (PPI) via link and sequence analysis. *IEEE Access* **2018**, *6*, 22108–22120. [CrossRef]
45. Wang, X.; Zhang, Y.; Zhou, P.; Liu, X. A supervised protein complex prediction method with network representation learning and gene ontology knowledge. *BMC Bioinform.* **2022**, *23*, 300. [CrossRef] [PubMed]

46. Palukuri, M.V.; Patil, R.S.; Marcotte, E.M. Molecular complex detection in protein interaction networks through reinforcement learning. *BMC Bioinform.* **2023**, *24*, 306. [CrossRef]
47. Sahoo, T.R.; Patra, S.; Vipsita, S. Decision tree classifier based on topological characteristics of subgraph for the mining of protein complexes from large scale PPI networks. *Comput. Biol. Chem.* **2023**, *106*, 107935. [CrossRef]
48. Chen, H.; Cai, Y.; Ji, C.; Selvaraj, G.; Wei, D.; Wu, H. AdaPPI: Identification of novel protein functional modules via adaptive graph convolution networks in a protein–protein interaction network. *Brief. Bioinform.* **2023**, *24*, bbac523. [CrossRef]
49. Ramadan, E.; Naef, A.; Ahmed, M. Protein complexes predictions within protein interaction networks using genetic algorithms. *BMC Bioinform.* **2016**, *17*, 481–489. [CrossRef]
50. Lei, X.; Ding, Y.; Fujita, H.; Zhang, A. Identification of dynamic protein complexes based on fruit fly optimization algorithm. *Knowl.-Based Syst.* **2016**, *105*, 270–277. [CrossRef]
51. Zhang, Y.; Lei, X.; Tan, Y. Firefly clustering method for mining protein complexes. In Proceedings of the Advances in Swarm Intelligence: 8th International Conference, ICSI 2017, Fukuoka, Japan, 27 July–1 August 2017; Proceedings, Part I 8; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; pp. 601–610.
52. Zhao, J.; Lei, X.; Wu, F.-X. Predicting protein complexes in weighted dynamic PPI networks based on ICSC. *Complexity* **2017**, *2017*, 4120506 [CrossRef]
53. Lei, X.; Fang, M.; Fujita, H. Moth–flame optimization-based algorithm with synthetic dynamic PPI networks for discovering protein complexes. *Knowl.-Based Syst.* **2019**, *172*, 76–85. [CrossRef]
54. Feng, Z.; Tuo, S.; Chen, T. A New Model Based on Differential Evolutionary Algorithm and Markov Clustering for Identifying Protein Complexes. In *Proceedings of the 2023 42nd Chinese Control Conference (CCC)*; IEEE: New York, NY, USA, 2023; pp. 6789–6794.
55. Lei, X.; Ying, C.; Wu, F.-X.; Xu, J. Clustering PPI data by combining FA and SHC method. *BMC Genom.* **2015**, *16*, S3. [CrossRef] [PubMed]
56. Lei, X.; Wang, F.; Wu, F.-X.; Zhang, A.; Pedrycz, W. Protein complex identification through Markov clustering with firefly algorithm on dynamic protein–protein interaction networks. *Inf. Sci.* **2020**, *329*, 303–316. [CrossRef]
57. Lei, X.; Wu, F.-X.; Tian, J.; Zhao, J. ABC and IFC: Modules detection method for PPI network. *BioMed Res. Int.* **2014**, *2014*, 968173. [CrossRef] [PubMed]
58. Mao, Y.; Deng, Q.; Liu, Y. Mining Weighted Protein Complexes Based on Fuzzy Ant Colony Clustering Algorithm. In Proceedings of the Smart City and Informatization: 7th International Conference, iSCI 2019, Guangzhou, China, 12–15 November 2019; Proceedings 7; Springer: Singapore, 2019; pp. 557–569.
59. Hu, L.; Yuan, X.; Xiong, S. Identifying overlapping protein complexes in yeast protein interaction network via fuzzy clustering. In Proceedings of the 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Naples, Italy, 9–12 July 2017; pp. 1–6.
60. Zhang, S.; Wang, R.-S.; Zhang, X.-S. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Phys. A Stat. Mech. Its Appl.* **2007**, *374*, 483–490. [CrossRef]
61. Hu, L.; Chan, K.C.C. Fuzzy clustering in a complex network based on content relevance and link structures. *IEEE Trans. Fuzzy Syst.* **2015**, *24*, 456–470. [CrossRef]
62. Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496. [CrossRef] [PubMed]
63. Du, M.; Ding, S.; Xu, X.; Xue, Y. Density peaks clustering using geodesic distances. *Int. J. Mach. Learn. Cybern.* **2018**, *9*, 1335–1349. [CrossRef]
64. Xu, X.; Ding, S.; Ding, L. Research progress in density peak clustering algorithms. *J. Softw.* **2020**, *33*, 1800–1816.
65. Kesemen, O.; Tezel, Ö.; Özkul, E. Fuzzy c-means clustering algorithm for directional data (FCM4DD). *Expert Syst. Appl.* **2016**, *58*, 76–82. [CrossRef]
66. Collins, S.R.; Kemmeren, P.; Zhao, X.-C.; Greenblatt, J.F.; Spencer, F.; Holstege, F.C.P.; Weissman, J.S.; Krogan, N.J. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteom.* **2007**, *6*, 439–450. [CrossRef] [PubMed]
67. 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]
68. Mewes, H.-W.; Amid, C.; Arnold, R.; Frishman, D.; Güldener, U.; Mannhaupt, G.; Münsterkötter, M.; Pagel, P.; Strack, N.; Stümpflen, V.; et al. MIPS: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **2004**, *32*, D41–D44. [CrossRef]
69. Hong, E.L.; Balakrishnan, R.; Dong, Q.; Christie, K.R.; Park, J.; Binkley, G.; Costanzo, M.C.; Dwight, S.S.; Engel, S.R.; Fisk, D.G.; et al. Gene Ontology annotations at SGD: New data sources and annotation methods. *Nucleic Acids Res.* **2007**, *36*, D577–D581. [CrossRef] [PubMed]
70. 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]

71. 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]
72. 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]
73. 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]
74. Brohee, S.; Van Helden, J. Evaluation of clustering algorithms for protein–protein interaction networks. *BMC Bioinform.* **2006**, *7*, 488. [CrossRef]