Dual-population Firefly Algorithm Based on Gender Differences for Detecting Protein Complexes

Qiwen Zhang and Xinxin Guo

***Abstract*—To address the problem of high false positive/negative rate and low accuracy in protein complex detection, we propose the Dual-population Firefly Algorithm Based on Gender Differences (DFAGD) based on the unique core-attachment structure of protein complexes and the biological properties of fireflies. This method divides the detection of protein complexes into two phases. Firstly, a global search of the male population is used to detect the core proteins, and then a local search of the female population is used to detect the attachment proteins, which improves the accuracy of detection. In the male population strategy, the population diversity is redefined, and when the diversity falls below the threshold, a spring model is introduced to bring the population into the repulsion phase so that it does not fall into a local optimum. The female population selects elite and excellent individuals from the detection results of the male population to perform guided neighborhood searches, which can effectively improve detection accuracy. Finally, the effectiveness of the protein complex detection method is tested by comparing it to eight classical detection methods using four datasets of Saccharomyces cerevisiae proteins.**

***Index Terms*—protein complex, firefly algorithm, dual-population, core-attachment structure, spring model**

I. Introduction

Protein complexes play a critical role in a multitude of intracellular biological processes [1]. Examining these complexes provides insights into cell function and organization, aiding in the identification of disease-related genes and establishing connections between drugs and diseases [2]. Nevertheless, traditional techniques for uncovering protein complexes, like co-immunoprecipitation and mass spectrometry [3], are costly in terms of reagents and demand significant time and labor.

With the rise of high-throughput technologies, a substantial volume of data on protein-protein interactions (PPI) has been generated [4]. In 2001, Legrain et al. [5] depicted PPI as an undirected graph by transforming extensive PPI data into a network structure, thus initiating protein complex detection through computational methods. While these computational approaches can identify some protein complexes, they may be prone to false positives/negatives and might not detect sparsely connected protein complexes within the network. In 2006, Gavin et al. [6] conducted a comprehensive examination of protein complex structure and revealed that a protein complex comprises a core-attachment model, where the core represents a densely interconnected subgraph, and the attachment consists of proteins with sparse connections to the core. Subsequently, various detection methods based on the core-attachment structure have been introduced. For example, methodologies like COACH [7], WPNCA [8], and NRAGE-WPN [9] extract the core of protein complexes from neighboring protein structures, then identify the attachment proteins, and ultimately combine certain protein complexes with significant overlap to derive final detection outcomes.

In recent years, optimization algorithms have garnered significant attention from researchers both domestically and internationally due to their straightforward structure and ease of implementation [10-11]. Protein complexes are commonly identified using assays centered around the core-attachment structure. To address complex protein-related challenges, researchers have integrated swarm intelligence optimization algorithms with the core-attachment framework. This integration taps into the adaptability and optimization capabilities offered by swarm intelligence algorithms. In 2019, Lei et al. [12] introduced the MFOC algorithm, which employs layered concepts to delineate the core of a protein complex and identifies auxiliary roles of protein nodes through the moth-flame optimization algorithm. Similarly, in the same year, the IFPA algorithm [13] was developed, determining protein complex cores based on core set density and mimicking the process of pollen landing on compatible flowers. It identifies the nearest core proteins for attachment proteins using the enhanced flower pollination algorithm (FPA). In 2022, Wang et al. [14] presented the MP-AHSA method to detect protein complexes with multiple properties (MP) by recognizing the core of protein complexes with MCL, formulating strategies for protein complex assembly to detect attachment proteins, and ultimately creating an adaptive and acoustic search algorithm to optimize the parameters of the MP algorithm.

Swarm intelligence optimization algorithms have made significant strides in detecting protein complexes.

Manuscript received November 18, 2023; revised April 4, 2024.

This work is supported by National Natural Science Foundation of China 62063021 (Research on HMS Scheduling Optimization and Control and Intelligence System in Manufacturing IoT Environment) and National Natural Science Foundation of China 62162040 (Research on Terrain Representation and Dissemination Adaptive Scheduling Strategy for Large-scale Social Network Influence Adaptability).

Qiwen Zhang is an associate professor of Lanzhou University of Technology, Gansu, China (e-mail:823869941@qq.com).

Xinxin Guo is a postgraduate student of Electronic Information of Lanzhou University of Technology, Gansu, China (corresponding author to provide phone: 18893461573; e-mail: 2177742671@qq.com).

Nonetheless, these methods have not fully leveraged complex structures, and inherent algorithmic limitations impact detection accuracy. To address these challenges, we introduce a pioneering methodology known as the Dual-Population Firefly Algorithm based on Gender Differences for Detecting Protein Complexes (DFAGD-DPC). This innovative approach aims to boost the precision of protein complex identification by surmounting current method constraints. The key advancements encompass three core components: (1) Seed proteins initialization: leveraging the concept of topological potential to weigh the PPI network and select seed proteins. (2) Male population strategy: reintroducing population diversity when faced with diminishing male population diversity that tends towards local optima; upon reaching a predetermined threshold, the spring model is introduced to transition the population into the repulsion phase, thereby determining the core of the protein complex. (3) Female population strategy: utilizing the optimal region established by the male population strategy as the female population's vicinity, employing elite male individuals to guide female counterparts towards optimal solutions. The female population strategy is instrumental in characterizing the attachment of protein complexes.

The remaining sections of this article are organized as follows: Section 2 provides an overview of pertinent foundational works; Section 3 improves the firefly algorithm for protein complex detection; Section 4 showcases and evaluates the experimental outcomes; and lastly, the article concludes by summarizing key findings and outlining future research directions.

II. Related Works

A. Protein-protein Interaction Network

The PPI network is depicted through undirected graphs $G=(V,E)$, where V and E are the set of nodes and the set of edges of the graphs G, respectively [15].

Definition 1 (Weighted degree ($d_{\omega}(v)$)): The weighted degree of a node v is the sum of the edge weights of the neighbor nodes directly connected to it. The formula is shown in (1):

$$d_{\omega}(v)=\sum_{u\in N(v)}\omega(u,v) \tag{1}$$

(Average weighted degree ($d_{avg\_\omega}(v)$)): The average weighted degree is defined as:

$$d_{avg\_\omega}(v)=\frac{\sum_{u\in N(v)}\omega(u,v)}{n} \tag{2}$$

Where $\omega(u,v)$ is the weight of the interaction edge between the node v and its neighbor u, and n is the total number of nodes.

Definition 2 (Core cluster density (*dens*)): The density of a core cluster is defined as formula (3):

$$dens(cc)=\frac{2\times|E|}{|V|\times(|V|-1)} \tag{3}$$

Where $|E|$ is the number of edges in the core cluster and $|V|$ is the number of nodes in the core cluster.

Definition 3 (Adhesion (*ds*)): A measure of whether a node v can be added to the core cluster as an attachment protein. The formula is defined as follows:

$$ds_{(v,CSS)}=\frac{E_{in}-E_{out}}{E_{in}+E_{out}}=2\frac{E_{in}}{E_{in}+E_{out}}-1 \tag{4}$$

Where E_{in} is the number of edges that node v interacts with nodes in the core cluster, and E_{out} is the number of edges that node v interacts with nodes outside the core cluster.

B. Firefly Algorithm

The Firefly Algorithm (FA) is a computational method that refines solutions by mimicking the light-emission behavior observed in fireflies in nature [16]. The mathematical model for the FA is established based on three key criteria. Initially, all fireflies within the algorithm are regarded as gender-neutral. Secondly, the attraction and luminosity between fireflies are directly correlated. Lastly, the brightness of fireflies is intricately linked to the objective function being optimized.

Definition 4 (Attraction): The attraction of firefly j to firefly i is defined as formula (5):

$$\beta_{ij}(r_{ij})=\beta_0 e^{-\gamma r_{ij}^2} \tag{5}$$

Where β_0 is the maximum attraction, i.e., the attraction of the firefly at the light source ($r=0$); γ is the light absorption coefficient; r_{ij} is the Cartesian distance between the firefly i and the firefly j:

$$r_{ij}=x_i-x_j=\sqrt{\sum_{k=1}^{d}\left(x_{i,k}-x_{j,k}\right)^2} \tag{6}$$

Definition 5 (Position update): As the firefly i is attracted to the firefly j, the firefly i moves towards it and updates its position, with the update formula as follows (7):

$$x_i(t+1)=x_i(t)+\beta_0 e^{-\gamma r_{ij}^2}\left(x_j(t)-x_i(t)\right)+\partial\varepsilon_i \tag{7}$$

Where t is the number of iterations of the algorithm, ∂ is the coefficient of the random term, and ε_i is the random number that comes from a uniform distribution.

III. Improved Firefly Algorithm for Detecting Protein Complexes

A. Constructing Weighted PPI Network

PPI data can originate from various experimental techniques, occasionally accompanied by false positive or negative noise within the dataset [17]. To mitigate the effects of such noise, weights are assigned to the PPI network. In this network, interactions between nodes are depicted as a topological potential field, where nodes exert influence on one another. Within networks demonstrating modular characteristics, nodes' influence is confined to specific regions. As the distance separating nodes expands, their influence diminishes accordingly.

Following the concept of topological potential, the PPI network incorporates this attribute in its weighting, with the weight formula outlined as follows:

$$\omega(u,v) = m_u \times e^{-(\frac{d_{uv}}{\sigma})^2} \tag{8}$$

Where m_u denotes the quality of the node u, d_{uv} describes the shortest path between node u and node v, and σ controls the influence range of each node. Typically, the shortest path between protein functional modules is less than or equal to 2, so to ensure that the effect between proteins is not greater than 2, the σ takes a value of 0.9428 [18].

When calculating the topological potential, the value of m_u is usually 1. To remove noise from the data and create a more realistic network, m_u is redefined using the Pearson correlation coefficient, GO annotations, and subcellular localization information:

$$m_{\mathrm{u}} = \frac{PCC(u,v) + CGO(u,v) + CSL(u,v)}{3} \tag{9}$$

Where $PCC(u,v)$ is the Pearson correlation coefficient, which measures the co-expression characteristics of protein pairs [19]. The greater the correlation between two adjacent proteins, the greater the *PCC* value corresponding to their interaction edges, u_i and v_i are the expression values of proteins u and a v at a time point i, respectively. μ_u and μ_v are the mean values of gene expression of protein u and v, respectively. The formula is shown in (10):

$$PCC(u,v) = \frac{\sum_{i=1}^{n}(u_i - \mu_u)(v_i - \mu_v)}{\sqrt{\sum_{i=1}^{n}(u_i - \mu_u)^2} \times \sqrt{\sum_{i=1}^{n}(v_i - \mu_v)^2}} \tag{10}$$

$CGO(u,v)$ is GO annotations that evaluate the similarity between two interacting proteins [20]. The larger the number of common GO annotations, the stronger the interaction between protein pairs. GO_u and GO_v are two sets of GO terms annotating protein u and protein v, respectively, and denote the set of common GO annotations between them. The formula is shown in (11):

$$CGO(u,v) = \begin{cases} \dfrac{|GO_u \cap GO_v|^2}{|GO_u| \times |GO_v|}, |GO_u| > 0 and |GO_v| > 0 \\ 0, otherwise \end{cases} \tag{11}$$

$CSL(u,v)$ the information about subcellular localization, and if two interacting proteins have the same subcellular localization, the interaction between them is more reliable [20]. SL_u and SL_v represents the subcellular localization set of protein u and protein v, respectively, and $SL_u \cap SL_v$ denotes the set of common subcellular localizations between them. The formula is shown in (12):

$$CSL(u,v) = \begin{cases} \dfrac{|SL_u \cap SL_v|^2}{|SL_u| \times |SL_v|}, |SL_u| > 0 and |SL_v| > 0 \\ 0, otherwise \end{cases} \tag{12}$$

B. DFAGD-DPC

In both humans and animals, notable differences exist between the sexes. This distinction extends to fireflies, where distinct male and female populations are present. Typically, male fireflies possess wings and tend to soar higher in the sky to survey their surroundings, while female fireflies, lacking wings, remain closer to the ground [21]. Consequently, two separate subpopulations emerge, each constituting half of the total population. The male population serves as the primary reservoir of core proteins, whereas the female population acts as the primary source of attachment proteins. Core proteins are readily identified within the male population through a comprehensive search, while attachment proteins are more commonly located within the female population through a refined search conducted locally. The behavioral characteristics of fireflies correspond to the features of DFAGD-DPC, as illustrated in Table 1:

TABLE 1. THE BEHAVIORAL CHARACTERISTICS OF FIREFLIES CORRESPOND TO THE FEATURES OF DFAGD-DPC.

Behavioral characteristics of fireflies	DFAGD-DPC
Fireflies	Protein nodes in the PPI network
The movement of the firefly	Detection of protein complexes
The movement of male firefly	Search for core proteins
The movement of female firefly	Search for attachment proteins

(1) Initialize Seed Proteins

To initialize the positions of fireflies, we leverage the node weight data from the PPI network. Nodes with a weighted degree surpassing the average weighted degree are designated as seed proteins, with each seed protein representing an individual firefly.

(2) Male Population Strategy for Extended Core Clusters

When employing the swarm intelligence optimization algorithm to detect protein complexes, firefly entities gravitate towards the optimal individual under its guidance, causing a gradual contraction of the exploration space. This phenomenon results in a dense clustering of individuals around the optimal entity, leading to a swift reduction in diversity. Consequently, the algorithm often converges towards a local optimum, neglecting crucial regions within the search space and failing to accurately pinpoint the protein complex core. To surmount this challenge, population diversity is redefined based on the spatial data of individuals, and a spring model is introduced to disperse firefly entities from densely clustered areas. This approach augments population diversity, facilitating the algorithm's escape from local optima.

a) Population diversity

Assessing the level of aggregation within a population can be accomplished by evaluating diversity. A diminished diversity value signifies heightened aggregation, whereas an elevated diversity value suggests wider dispersion among individuals. Through factoring in the spatial positioning of individuals, we have recalibrated population diversity and established specific quantitative benchmarks. Formula (13) delineates the concept of diversity.

$$diversity(t)=\frac{1}{N\cdot\sqrt{D}}\cdot\sum_{i=1}^{N}\sqrt{\sum_{j=1}^{D}\left(p_i^j(t)-\overline{p_j(t)}\right)^2},$$
$$p_i^j(t)=\left(x_i^j(t)-x_{min}^j(t)\right)\Big/\left(x_{max}^j(t)-x_{min}^j(t)\right), \quad (13)$$
$$\overline{p_j(t)}=\sum_{i=1}^{N}p_i^j(t)\Big/N$$

Where N is the population size, D is the dimensionality of the problem, $p_i^j(t)$ is the value after linear normalization in the j th dimension, $x_{\max}^j(t)$ and $x_{\min}^j(t)$ are the maximum and minimum values in the j th dimension, and $\overline{p_j(t)}$ is the average value of all fireflies after linear normalization in the j th dimension.

b) Spring model

A spring model is established between the optimal individual and other particles. And if we assume that there are N firefly individuals, then they are $N-1$ spring models. When the population's diversity falls below a certain threshold, the spring model is implemented to increase the distribution of individuals and promote diversity.

According to Hooke's law, spring force is defined as:

$$F=k_s\cdot\Delta x \quad (14)$$

Where k_s is the stiffness coefficient of the spring. Considering that the spring force increases with the increase of the stiffness coefficient, the interaction between two nodes in the PPI network increases with the increase of the number of public neighbors. If a spring model is established between node i (current particle) and node j (optimal particle) in the network, then the product $k_i\cdot k_j$ of the degrees of node i and node j can be considered as the stiffness coefficient of the spring. $\triangle x$ is the compression of the spring. In the t iteration, the individual firefly moves towards the optimal individual, the distance traveled is the compression of the spring, then the spring compression is: $\Delta x=\| x(t)-x(t-1)\|$. Here the F is treated as a scalar, considering only the size and not the direction, and the F is defined as:

$$F=k_i\cdot k_j\cdot\Delta x \quad (15)$$

When the level of diversity decreases to a certain point, individual fireflies tend to move away from optimal individuals. This means that the male population's iterative process can be divided into two phases: attraction and repulsion. The rebound coefficient is determined based on the concepts of diversity and spring force, as described in formula (16):

$$\zeta=e^{diversity(t)/F} \quad (16)$$

c) Male population strategy based on the spring model

When the population diversity is above the threshold d_{low}, the positional movement of individuals follows the basic firefly algorithm, this phase is the attraction phase. The position update is defined by the formula (17):

$$x_i(t+1)=x_i(t)+\beta_0 e^{-\gamma r_{ij}^2}\left(x_j(t)-x_i(t)\right)+\alpha\xi_i,\ diversity>d_{low} \quad (17)$$

When the population diversity is below the threshold d_{low}, the individuals near the optimal individual bounce in the opposite direction of compression, and the distance of bouncing is affected by the rebound coefficient, this phase is the repulsion phase. The position of the rebounded individual is according to the formula (18):

$$x_i(t+1)=x_i(t)+\zeta\left(x_i(t-1)-x_i(t)\right)+\alpha\xi_i,\ diversity<d_{low} \quad (18)$$

d) Male population strategy for extended core clusters

For each male firefly, the position of the neighboring node with the highest PCC value in the corresponding core cluster is chosen as the optimal position. If the PCC value at the optimal position is greater than 0, the firefly moves to the optimal position, this node is clustered to the core cluster, otherwise, the iteration of this firefly is terminated. Repeat the above process, and update the firefly position, until the density of core cluster is less than the given threshold, which means the end of this iteration phase. The threshold controls the number of iterations and the size of the core cluster, which is fixed at 0.7 [22].

The pseudo code for determining core proteins in male populations is shown in Algorithm 1.

Algorithm 1. The pseudocode for determining core proteins in male populations.

Input: weighted PPI network
Output: detected core clusters

a) for each firefly c do
b) if $dens(cc)>threshold$
c) x_{neib} = neighbor nodes of the core cluster corresponding to firefly c
d) for each node x_i in x_{neib}
e) calculated PCC according to formula (10)
f) end for
g) x_{best} = the node with the largest PCC
h) if $PCC(x_{best},c)>0$
i) add x_{best} to the core cluster
j) update the position of firefly c according to formula (17) or (18)
k) else
l) end the search for male firefly c
m) end if
n) end if
o) end for

(3) Female Population Strategy to Identify Attachment Proteins

By conducting a thorough exploration within the male population, we can identify an approximate optimal region. However, optimizing the precision of this solution is imperative. Upon locating a near-optimal solution, investigating the neighborhood of the individual reaching the local optimum could potentially reveal the global optimal solution. Given that female individuals lack wings, their search capabilities are confined to their immediate vicinity. To enhance the algorithm's accuracy, the optimal region identified through the male population search strategy serves as the neighborhood for the female population. Elite and outstanding individuals from the male population provide

guidance to the female counterparts as they navigate towards the optimal solution.

a) Position update

The search equation for the female population strategy is explicitly depicted in formula (19) as part of the algorithm design.

$$y_i(t+1) = y_i(t) + \lambda_1\beta_0 e^{-\gamma r_{ij}^2}\left(x_{elite}(t) - y_i(t)\right) + \lambda_2\beta_0 e^{-\gamma r_{ij}^2}\left(x_{excellent}(t) - y_i(t)\right) \quad (19)$$

Where $y_i(t)$ is the position vector of the last iteration of the female i. $x_{elite}(t)$ is an elite individual, $x_{excellent}(t)$ is an excellent individual. This means that female fireflies can learn useful information from high-quality solutions to update their location. λ_1 and λ_2 are random numbers with a uniform distribution between [0, 1], satisfying $\lambda_1 + \lambda_2 = 1$.

b) Female population strategy to identify attachment proteins

Algorithm 2. The pseudocode for determining the attachment proteins for female populations.

Input: weighted PPI Network
Output: detected protein complex

p) for each firefly c do
q) $s_{\text{elite}} \rightarrow$ sort nodes of core cluster corresponding to firefly c according to the ds
r) $s_{\text{excellent}} \rightarrow$ sort the neighbor nodes of the core cluster corresponding to female firefly c according to the ds
s) x_{elite} = the largest ds in s_{elite}
t) $x_{excellent}$ = the largest ds in $s_{\text{excellent}}$
u) $x_{\text{center}} = x_{excellent}$
v) if $ds(x_{center}, c) > 0$
w) add x_{center} as an attachment protein to the core cluster
x) update the position of firefly c according to formula (19)
y) else
z) end the search for firefly c
aa) end if
bb) end for

The nodes in the core cluster corresponding to each female firefly and the neighboring nodes of the corresponding core cluster are ranked by the size of the ds. The nodes with the largest ds are selected as the elite individual x_{elite} and the excellent individual $x_{excellent}$, respectively. The excellent individual $x_{excellent}$ is used as the position of the central firefly, and if the ds of the central firefly is greater than 0, the firefly position is updated and added to the core cluster as an attachment protein, iterated to update the central firefly position and other firefly positions, and if the ds at the central firefly position is less than 0, the iteration of the firefly ends, and the corresponding core cluster of the firefly is a protein complex.

The final protein complex is formed when the female firefly goes through the above steps and deletes the protein complex with less than three nodes. The pseudocode for determining the attachment proteins for female populations is shown in Algorithm 2.

IV. Simulation Experiment and Result Analysis

The simulation experiment was conducted in the following operating environment: Windows 10 operating system, Intel (R) Core (TM) i5-11300H @ 3.10GHz processor, with 16GB of physical memory. The algorithms were executed using PyCharm Professional 2022 and implemented in Python 2.7.

A. Parameter Setting

Parameter settings play a crucial role in determining the performance of an algorithm. In the DFAGD algorithm, the diversity threshold is a key parameter that needs to be established. The selection of the diversity threshold directly impacts the algorithm's search scope and efficiency in problem-solving, necessitating rational adjustments based on specific characteristics and requirements of the problem at hand. By setting the diversity threshold d_{low}, one can strike a balance between the algorithm's exploration and exploitation capabilities, thereby optimizing performance and enhancing solution quality. During the algorithm's iterative process, when the population diversity declines below the predefined threshold, the algorithm introduces a spring model to boost diversity within the population. This operation is akin to the "mutation" operation in evolutionary algorithms and aids in steering the algorithm away from local optima to explore broader solution spaces. Therefore, the diversity threshold is typically set to a small value to ensure timely adjustments to population diversity. To determine the optimal diversity threshold, this study employed the Friedman test method, considering a range of potential threshold options, including {0.06, 0.07, 0.08, 0.09, 0.10, 0.11, 0.12}. Four representative benchmark functions were chosen for testing, evaluating convergence performance under different thresholds to derive corresponding mean ranking. A lower mean ranking indicates superior overall optimization performance of the algorithm. Experimental comparisons revealed that when the diversity threshold d_{low} was set to 0.08, the mean ranking across the tested functions were minimized. Thus, to maximize the performance of the DFAGD algorithm, the diversity threshold d_{low} was set at 0.08. This critical configuration ensures the algorithm maintains population diversity while efficiently conducting global searches and effectively avoiding the pitfalls of local optima.

TABLE 2. THE RESULTS OF THE FRIDMAN TEST AT DIFFERENT THRESHOLDS WERE ANALYZED

d_{low}	BEST FITNESS				MEAN RANKING
	f5	f11	f15	F18	
0.06	7.31E+02	5.95E-01	8.00E-04	3.00E+00	2.38
0.07	1.76E+03	4.35E-01	8.00E-04	3.02E+00	2.38
0.08	1.08E+03	3.15E-01	1.40E-03	3.00E+00	**2.25**
0.09	1.97E+03	5.61E-01	3.60E-03	3.00E+00	3.50
0.10	3.74E+03	8.21E-01	1.80E-03	4.81E+00	6.00
0.11	4.20E+03	6.90E-01	1.20E-02	1.11E+01	6.50
0.12	1.91E+03	6.58E-01	1.37E-02	3.18E+00	5.00

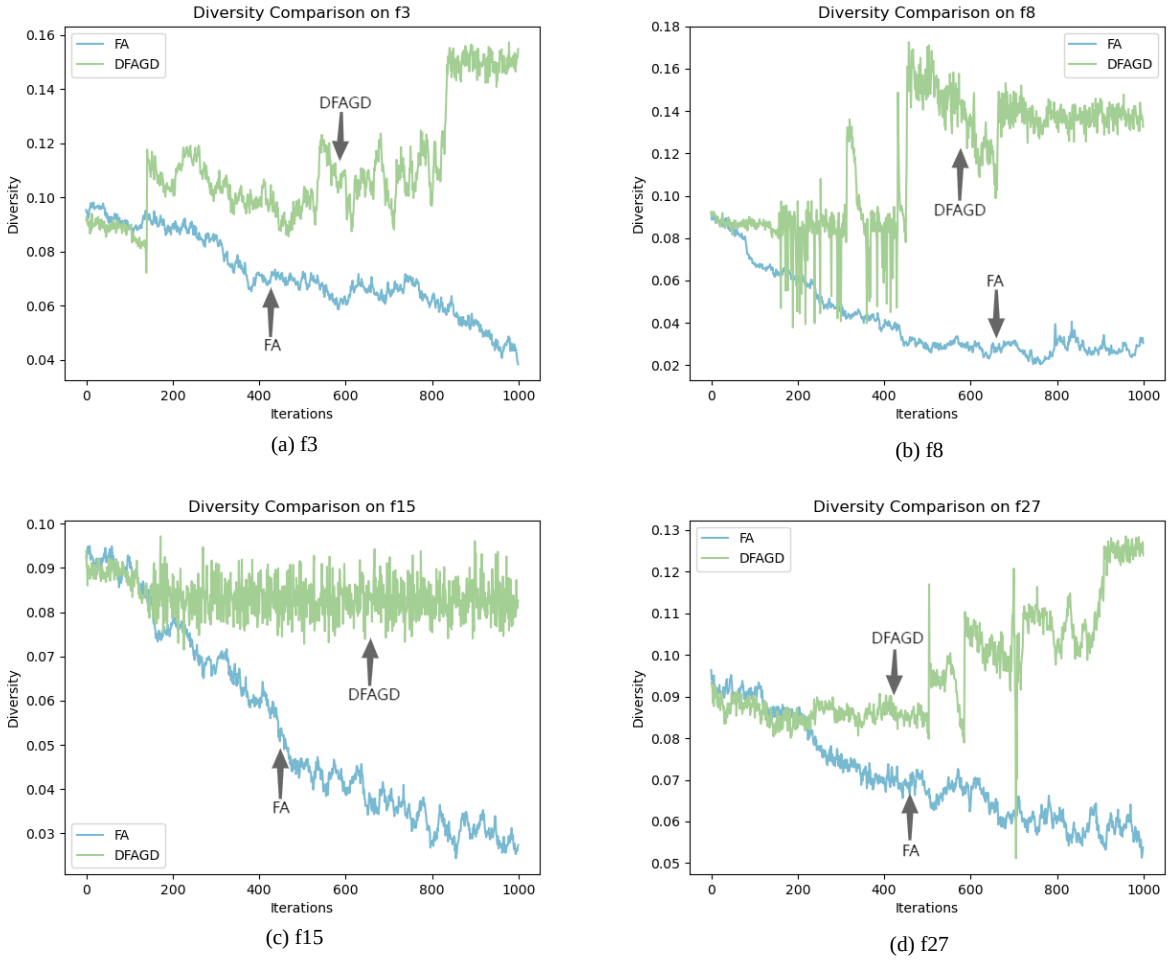


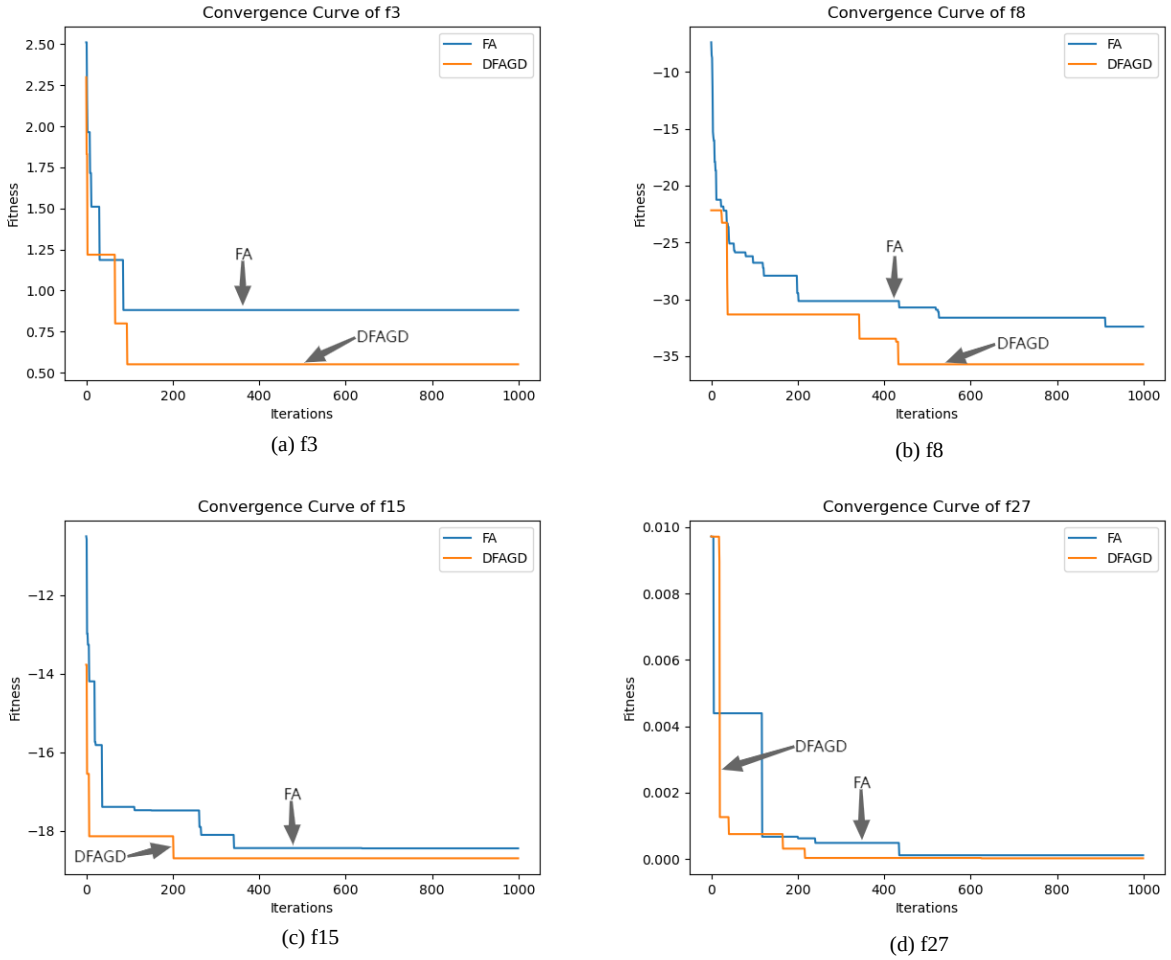
Fig. 1 the comparison of population diversity between FA and DFAGD across four distinct test functions

B. Population Diversity Analysis

To validate the effectiveness of the improved DFAGD algorithm, standard FA is compared with the enhanced DFAGD on four different test functions. When evaluating population diversity, the $diversity(t)$ is employed as a metric that effectively reflects the distribution status of individuals within the population. A smaller value of $diversity(t)$ indicates a higher tendency for individuals to cluster within a specific region, implying lower population diversity. Conversely, a larger $diversity(t)$ value suggests that individuals are more widely spread across the search space, indicating higher population diversity. By contrasting the variations in $diversity(t)$ values between FA and DFAGD on the same test functions, one can visually assess the algorithms' search capabilities and optimization effects. Fig. 1 illustrates the comparison of population diversity between FA and DFAGD across four distinct test functions. The experimental results reveal that, compared to FA, DFAGD significantly enhances population diversity during the iterative process. This enhancement not only allows DFAGD to thoroughly explore the search space and discover more potential optimal solutions but also effectively mitigates the risk of the population getting trapped in local optima.

C. Algorithm Convergence Analysis

To further validate the significant advantage of the improved DFAGD algorithm in terms of convergence speed, an in-depth experimental study was conducted using four widely representative test functions. As shown in Fig. 2, the graph provides a detailed comparison of the convergence characteristics between FA and DFAGD across the four different test functions. In the graph, the convergence curve of FA is depicted by a blue line, while the convergence curve of DFAGD is represented by an orange line. By comparing the convergence curves of both algorithms, it is evident that the enhanced DFAGD algorithm exhibits outstanding convergence performance across all four test functions compared to the standard FA. Specifically, whether in the early or later stages of iteration, DFAGD consistently approaches the optimal solution at a faster rate. This indicates that under the same number of iterations, the DFAGD algorithm can efficiently find an approximate optimal solution to the problem, significantly enhancing the overall execution efficiency of the algorithm. These experimental results not only further confirm the effectiveness and superiority of the DFAGD algorithm improvements, but also provide strong support for its application in the detection of protein complexes.



(a) f3 (b) f8 (c) f15 (d) f27

Fig. 2 the comparison of the convergence characteristics between FA and DFAGD across the four different test functions

D. Experimental Data and Evaluation Metrics

In this paper, we tested several datasets for Saccharomyces cerevisiae, including DIP [23], Gavin [24], Krogan [25], and MIPS [26]. We also used the standard dataset, CYC2008 [27], which consists of 408 protein complexes.

To assess how well the protein complexes have been detected, three commonly used statistical evaluation metrics are employed, namely precision, recall, and F-measure [28]. These metrics are measured on a scale of 0.0 to 1.0, with higher values indicating better detection methods and better performance of the complexes. The formulas used to calculate these evaluation metrics are given below:

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$F - measure = \frac{2(Precision \times Recall)}{Precision + Recall} \tag{22}$$

Where TP is the number of protein complexes detected by the algorithm that matches the standard protein complex, and FP is the number of protein complexes detected by the algorithm that do not match the standard protein complex, FN is the number of undetected protein complexes in the standard protein complex. F-measure is the harmonic mean of precision and recall.

E. Performance Comparison

To comprehensively evaluate the performance and effectiveness of the DFAGD-DPC method, a rigorous analysis was conducted using four diverse datasets. These datasets were chosen to represent a wide range of biological conditions and protein interactions, ensuring the reliability and generality of the results. To benchmark DFAGD-DPC against existing methods, ten established protein complex detection algorithms were selected for comparison. These included MCL [29], MCODE [30], ClusterONE [31], CSO [32], CORE [33], COACH [7], EWCA [34], MP-AHSA [14], NLPGE-WPN [35] and LCDA [36].

To better compare the results of protein complex detection, Table 3 compares the performance of DFAGD-DPC with ten classical methods on four datasets. The table highlights the top three experimental results in bold font, with the ranking indicated in the superscript, providing a clear and concise visualization of the method's superiority. This allows readers to quickly identify the most effective methods and gain a deeper understanding of the relative performance of each algorithm. The results show that DFAGD-DPC is a better protein complex detection method in terms of precision, recall, and F-measure in the MIPS dataset as compared to other methods. For the Gavin dataset, precision and recall are better than other methods, but the F-measure is even better.

TABLE 3. COMPARES THE PERFORMANCE OF DFAGD-DPC WITH EIGHT CLASSICAL METHODS ON FOUR DATASETS.

datasets	algorithm	precision	recall	F-measure
DIP	MCL	0.185	0.517	0.272
	MCODE	0.618	0.232	0.337
	CORE	0.277	**0.738**[1]	0.403
	COACH	0.517	0.542	0.529
	ClusterONE	0.339	0.609	0.435
	CSO	**0.626**[3]	0.440	0.517
	EWCA	0.499	**0.701**[2]	**0.583**[3]
	MP-AHSA	**0.799**[1]	0.473	**0.594**[2]
	NLPGE-WPN	0.488	**0.669**[3]	0.564
	LCDA	0.608	0.552	0.578
	DFAGD-DPC	**0.670**[2]	**0.548**	**0.603**[1]
Gavin	MCL	0.642	0.441	0.523
	MCODE	**0.727**[1]	0.142	0.238
	CORE	0.574	0.434	0.494
	COACH	0.525	0.331	0.406
	ClusterONE	0.641	**0.480**[3]	0.549
	CSO	**0.645**[3]	0.302	0.411
	EWCA	0.607	0.402	0.484
	MP-AHSA	0.556	0.440	0.491
	NLPGE-WPN	0.632	**0.5042**[1]	**0.560**[1]
	LCDA	0.631	**0.4973**[2]	**0.556**[3]
	DFAGD-DPC	**0.700**[2]	0.465	**0.559**[2]
Krogan	MCL	0.456	**0.566**[1]	**0.505**[3]
	MCODE	**0.724**[2]	0.157	0.258
	CORE	0.412	**0.542**[2]	0.468
	COACH	0.617	0.343	0.441
	ClusterONE	0.463	0.523	0.491
	CSO	**0.726**[1]	0.331	0.445
	EWCA	**0.705**[3]	0.368	0.483
	MP-AHSA	0.555	0.461	**0.504**[3]
	NLPGE-WPN	0.578	**0.531**[3]	**0.553**[1]
	LCDA	0.589	0.520	**0.552**[2]
	DFAGD-DPC	**0.645**	**0.483**	**0.552**[2]
MIPS	MCL	0.202	0.545	0.295
	MCODE	0.447	0.115	0.183
	CORE	0.249	**0.624**[3]	0.356
	COACH	0.301	0.289	0.295
	ClusterONE	0.280	0.448	0.344
	CSO	**0.495**[2]	0.289	0.365
	EWCA	0.412	0.338	0.412
	MP-AHSA	0.401	**0.642**[2]	**0.493**[2]
	NLPGE-WPN	**0.485**[3]	0.599	**0.536**[2]
	LCDA	0.468	0.518	0.492
	DFAGD-DPC	**0.500**[1]	**0.644**[1]	**0.563**[1]

Although the values of precision and recall in the DIP and Krogan datasets are slightly lower than other methods, the F-measure is higher. Ultimately, the DFAGD-DPC method proves to be more effective in detecting protein complexes than other methods.

F. Comparison with Known Protein Complexes

To clearly demonstrate the algorithm's performance and the accuracy of our results. we conducted a thorough analysis to compare the accuracy of the DFAGD-DPC detection results with four other methods on the Krogan dataset. Specifically, we focused on the 265th protein complex from the CYC2008 standard protein complex. This complex consists of 12 protein nodes, including YNL232W, YOL021C, YHR081W, YGR158C, YHR069C, YOL142W, YDL111C, YCR035C, YDR280W, YGR095C, YOR001W, and YGR195W.

As shown in Fig. 3, a visual analysis of the results of the assays for known protein complexes and DFAGD-DPC and four other methods is shown. The blue nodes indicate the correctly detected proteins, the green nodes represent undetected proteins, and the pink nodes signify incorrectly detected proteins. As shown in the figure, the CORE method is less efficient, detecting only 2 standard proteins correctly. However, the MCODE and ClusterONE methods have shown improved detection efficiency by identifying 6 and 9 standard proteins correctly, respectively. The EWCA method has better detection results than the previous methods, detecting 11 standard proteins correctly but also incorrectly detecting 2 other proteins. The DFAGD-DPC method, on the other hand, detects all 12 standard proteins but also detects 2 proteins (YOR076C and YNR024W) incorrectly. This is mainly due to their stronger interactions and connections with more protein nodes in the complex. Ultimately, the

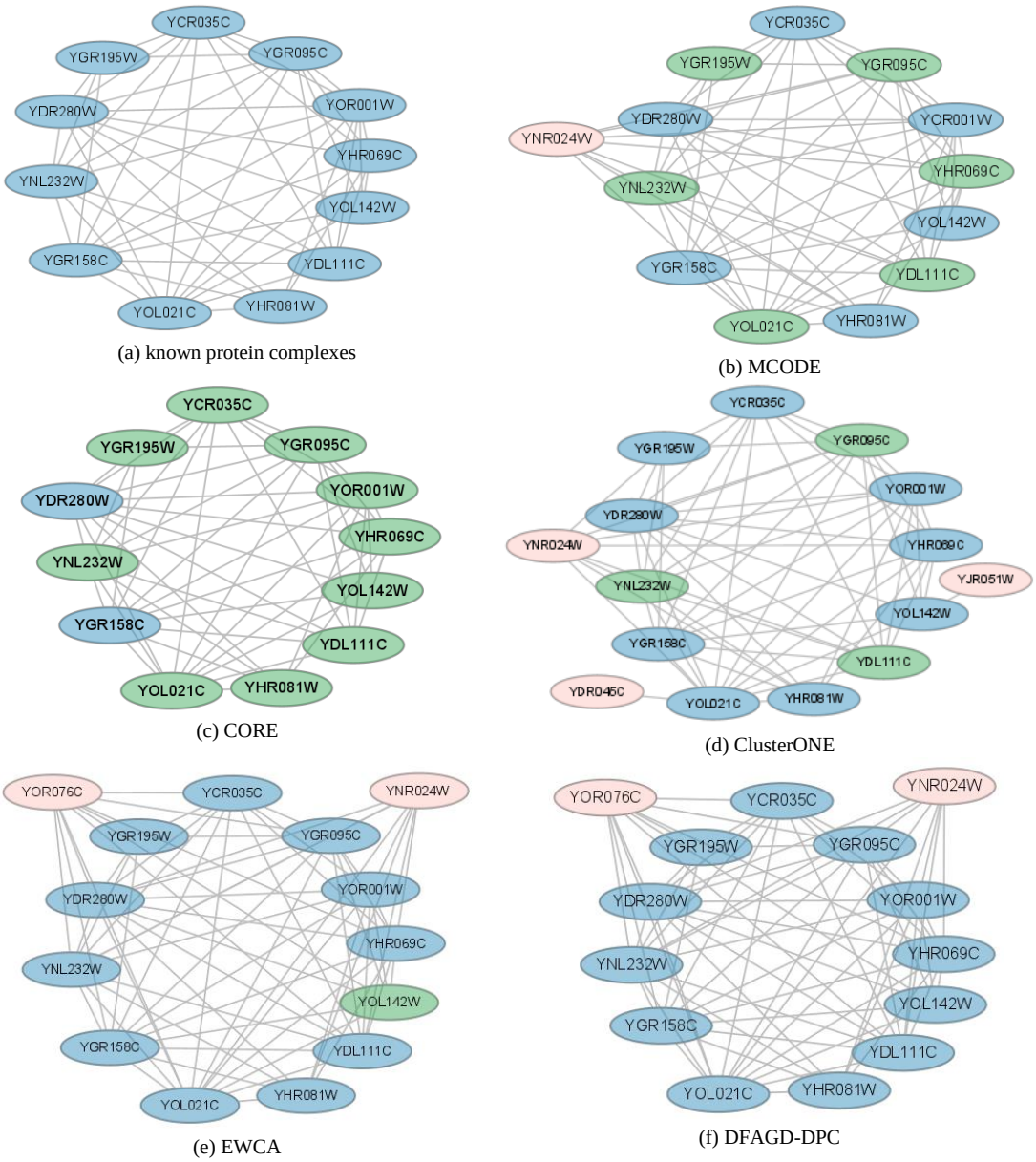
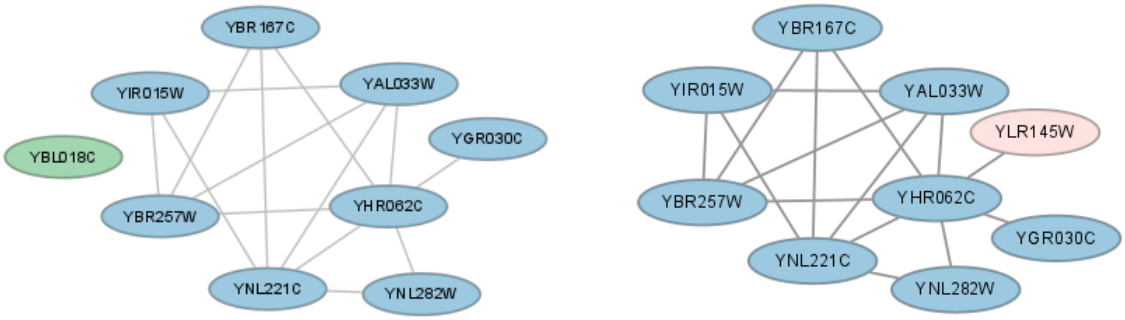


Fig. 3 A visual analysis of the results of the assays for known protein complexes and DFAGD-DPC and four other methods

DFAGD-DPC method achieves the best performance in detecting protein complexes.

In order to determine which proteins in the protein complexes detected by the DFAGD-DPC method are correct or incorrect, the detected results are compared with known protein complexes from a standard library. Table 4 provides a comparison of some complexes with reference classes in the Krogan dataset. Upon examination of Table 4, it is evident that the DFAGD-DPC method has successfully identified protein complexes with serial numbers “1”, “2”, “4”, and “5”. However, in the test results for serial numbers “3” and “6”, YDR416W and YLR145W were incorrectly detected, and in the complex with serial number “6”, YBL018C was not detected at all. Based on these experimental findings, we can conclude that the DFAGD-DPC method is able to correctly identify most protein complexes, despite the presence of some individual complexes that may contain incorrect or missing proteins. However, the proteins within the standard protein complexes are detected correctly.

Fig. 4 visualizes the protein complex with serial number “6” in Table 4. Fig. 4(a) shows the interaction network of the standard protein complexes in the Krogan dataset. The green node represents undetected proteins, primarily because they do not interact with the other protein nodes in the complex. Fig. 4(b) displays the protein complex detected by the DFAGD-DPC method. The blue nodes indicate correctly detected proteins, while the pink node represents an incorrectly detected protein. Unfortunately, YLR145W is mistakenly detected due to its weak interaction with the complex, only having an association with the node YHR062C, leading to its incorrect detection.



(a) the interaction network of a standard protein complex in the Krogan dataset

(b) the protein complex detected by the DFAGD-DPC method

Fig. 4 Visualizes the protein complex with serial number "6" in Table 4

TABLE 4. COMPARISON OF SOME COMPLEXES WITH REFERENCE CLASSES IN THE KROGAN DATASET.

Num	Standard protein complex	Correct protein complexes	Incorrect protein complexes
1	YGR206W YLR119W YCL008C YPL065W	YCL008C YGR206W YLR119W YPL065W	
2	YKL052C YDR016C YKR083C YBR233W-A YDR320C-A YGR113W YGL061C YKL138C-A YDR201W YKR037C	YDR016C YDR320C-A YKR037C YDR201W YGL061C YGR113W YKL052C YKR083C YKL138C-A YBR233W-A	
3	YBR126C YDR074W YMR261C YML100W	YDR074W YBR126C YDR416W YML100W YMR261C	YDR416W
4	YGL223C YGR120C YER157W YPR105C YNL051W YNL041C YGL005C YML071C	YGL223C YER157W YPR105C YML071C YNL051W YNL041C YGR120C YGL005C	
5	YIL062C YLR370C YKL013C YNR035C YBR234C YDL029W YJR065C	YDL029W YJR065C YLR370C YIL062C YKL013C YBR234C YNR035C	
6	YNL221C YNL282W YBR257W YAL033W YGR030C YBR167C YBL018C YIR015W YHR062C	YHR062C YGR030C YLR145W YNL282W YBR257W YBR167C YAL033W YIR015W YNL221C	YLR145W

V. CONCLUSIONS

In this study, we present a novel method, called DFAGD-DPC, for detecting protein complexes. Our approach leverages the unique biological characteristics of fireflies and their inherent capacity to self-organize protein complexes. Initially, the method identifies seed nodes by assigning weights to the PPI network. Subsequently, the firefly population is segregated based on gender, with male fireflies being used to identify core proteins. This is achieved by redefining population diversity to assess individual aggregation levels and introducing a spring model that aids the algorithm in overcoming local optima. Simultaneously, female fireflies are employed to detect attachment proteins. Our methodology utilizes the optimal region highlighted by the male population strategy to guide the female population in the search for superiority. Experimental findings demonstrate the effectiveness of the DFAGD-DPC method in detecting protein complexes; however, the method detects more overlapping protein complexes, and this issue will be a crucial area of focus for future research.

REFERENCES

[1] J. Zahiri, A. Emamjomeh, S. Bagheri, A. Ivazeh, G. Mahdevar, "Protein complex prediction: A survey," Genomics, 2019, vol. 112, no. 1, pp. 174-183.

[2] L. Xue, X. Q. Tang, "A new framework for discovering protein complex and disease association via mining multiple databases," Interdisciplinary Sciences: Computational Life Sciences, 2021, vol. 13, no. 4, pp. 683-692.

[3] Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Millar, P. Taylor, K. Bennett and K. Boutilier, "Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry,"Nature, 2002, vol. 415, no. 6868, pp.180-183.

[4] L. Hu, S. Yang, X. Luo, H. Yuan, K. Sedraoui and M. Zhou, "A distributed framework for large-scale protein-protein interaction data analysis and prediction using mapreduce," IEEE/CAA Journal of Automatica Sinica, 2021, vol. 9, no.1, pp.160-172.

[5] P. Legrain, J. Wojcik, J. M. Gauthier, "Protein–protein interaction maps: a lead towards cellular functions," Trends in Genetics, 2001, vol. 17, no.6, pp.346-352.

[6] A. C. Gavin, P. Aloy, P. Grandi, R. Krause, M. Boesche, M. Marzioch, C. Rau, L. J. Jensen, S. Bastuck and B. Dümpelfeld, "Proteome survey reveals modularity of the yeast cell machinery," Nature, 2006, vol. 440, no. 7084, pp. 631-636.

[7] M. Wu, X. l. Li, C-k. Kwoh, S-k. Ng, "A core-attachment based method to detect protein complexes in PPI networks," BMC Bioinformatics, 2009, vol. 10, no.1, pp. 1-16.

[8] P. Wei, J. Wang, B. Zhao, L. Wang, "Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure," IEEE/ACM Transactions on

Computational Biology and Bioinformatics, 2014, vol. 12, no. 1, pp. 179-192.
[9] Y. Yu, and D. Kong, "Protein complexes detection based on node local properties and gene expression in PPI weighted networks," BMC bioinformatics, 2022, vol. 23, no. 1, pp. 1-15.
[10] Q. Dan, H. Y. Li, and H. F. Chen. "Multi-objective Differential Evolution Algorithm Based on Affinity Propagation Clustering," IAENG International Journal of Applied Mathematics, 2023, vol. 53, no. 4, pp. 1408-1417.
[11] P. D. Kusuma, R. A. Nugrahaeni, and A. Dinimaharawati. "Cone Search: A Simple Metaheuristic Optimization Algorithm," IAENG International Journal of Applied Mathematics, 2022, vol. 52, no. 4, pp. 838-845.
[12] X. J. Lei, M. Fang, H. Fujita, "Moth–flame optimization-based algorithm with synthetic dynamic PPI networks for discovering protein complexes," Knowledge-Based Systems, 2019, vol. 172, pp. 76-85.
[13] X. J. Lei, M. Fang, L. Guo, Wu, F. X, "Protein complex detection based on flower pollination mechanism in multi-relation reconstructed dynamic protein networks," BMC Bioinformatics, 2019, vol. 20, pp. 63-74.
[14] R, Wang, C. Wang, H Ma. "Detecting protein complexes with multiple properties by an adaptive harmony search algorithm," BMC Bioinformatics, 2022, vol. 23, no. 1, pp. 414.
[15] S. Dilmaghani, M. R. Brust, C., Kieffer, E. Danoy, D. Grégoire, B. Pascal, "From communities to protein complexes: A local community detection algorithm on PPI networks," PLOS ONE, 2022, vol. 17, no. 1, pp. e0260484.
[16] X. S. Yang, "Nature-inspired metaheuristic algorithms," Luniver Press, 2010.
[17] Z. Wu, Q. Liao, B. Liu, "idenPC-MIIP: identify protein complexes from weighted PPI networks using mutual important interacting partner relation," Briefings in Bioinformatics, 2021, vol. 22, no. 2, pp. 1972-1983.
[18] M. Li, Y. Lu, J. X. Wang, F. X. Wu, Y. Pan, "A topology potential-based method for identifying essential proteins from PPI networks," IEEE/ACM Transactions on computational biology and bioinformatics, 2015, vol. 12, no.2, pp.372-383.
[19] R. Wang, H. Ma, C. Wang, "An improved memetic algorithm for detecting protein complexes in protein interaction networks," Frontiers in Genetics, 2021, vol. 12, pp. 794354.
[20] X. Lei, X. Yang, F. X. Wu, "Artificial fish swarm optimization based method to identify essential proteins," IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2018, vol. 17, no. 2, pp. 495-505.
[21] C. F. Wang, W. X. Song, "A novel firefly algorithm based on gender difference and its convergence," Applied Soft Computing, 2019, vol. 80, pp. 107-124.
[22] M. Altaf-Ul-Amin, Y. Shinbo, K. Mihara, K. Kurokawa, S. Kanaya, "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," BMC Bioinformatics, 2006, vol. 7, pp. 1-13.
[23] L. Salwinski, "PathBLAST: a tool for alignment of protein interaction networks," Nucleic Acids Research, 2004, vol. 32, no. suppl_2, pp. W83-W88.
[24] K. Schleinkofer, T. Dandekar, "Proteome survey reveals modularity of the yeast cell machinery - Global landscape of protein complexes in the yeast," Chemtracts, 2006, vol. 19, no. 12, pp. 469-473.
[25] N. J. Krogan, G. Cagney, H. Yu, G. Q. Zhong, X. H. Guo, "Global landscape of protein complexes in the yeast Saccharomyces cerevisiae," Nature, 2006, vol. 440, no. 7084, pp. 637-643.
[26] U. Güldener, M. Münsterkötter, M. Oesterheld, H. Mewes, V. Stümpflen and H. Yu, "MPact: the MIPS protein interaction resource on yeast," Nucleic acids research, 2006, vol. 34, no. suppl_1, pp. D435-D441.
[27] S. Omranian, A. Angeleska, Z. Nikoloski, "PC2P: parameter-free network-based prediction of protein complexes," Bioinformatics, 2021, vol. 37, no. 1, pp. 73-81.
[28] M. Pellegrini, D. Haynor, J. M. Johnson, "Protein interaction networks," Expert review of proteomics, 2004, vol. 1, no. 2, pp. 239-249.
[29] A. J. Enright, S. Van Dongen, C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," Nucleic Acids Research, 2002, vol. 30, no. 7, pp. 1575-1584.
[30] G. D. Bader, C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinformatics, 2003, vol. 4, no. 1, pp. 1-27.
[31] J. Wang, L. Min, J. Chen, P. Yi, "A Fast Hierarchical Clustering Algorithm for Functional Modules Discovery in Protein Interaction Networks," IEEE/ACM Trans Comput Biol Bioinform, 2011, vol. 8, no. 3, pp. 607.
[32] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. P. Li, "Protein Complex Prediction in Large Ontology Attributed Protein-Protein Interaction Networks," IEEE/ACM Transactions on Computational Biology & Bioinformatics, 2013, vol. 10, no. 3, pp. 729-741.
[33] H. C. M. Leung, Q. Xiang, S. M. Yiu, F. Y. L. Chin, "Predicting Protein Complexes from PPI data: a core-attachment approach," Journal of Computational Biology, 2009, vol. 16, no. 2, pp. 133-44.
[34] R. Q. Wang, G. Liu, C. Wang, "Identifying protein complexes based on an edge weight algorithm and core-attachment structure," BMC Bioinformatics, 2019, vol. 20, no. 1, pp. 471.
[35] Y. Yang, K. Dezhou. "Protein Complexes Detection Based on Node Local Properties and Gene Expression On PPI Weighted Networks," Research Squares, 2021.
[36] S. Dilmaghani, M. R. Brust, C. H. C. Ribeiro. "From communities to protein complexes: A local community detection algorithm on PPI networks," Plos one, 2022, vol. 17, no. 1, pp. e0260484.

Qiwen Zhang (1975.2-), male, master degree, master tutor. He graduated from Gansu University of Technology, majoring in computer and applied technology, with a bachelor's degree in engineering in 1999. He received the M.S. degree in computer and applied technology from Lanzhou University of Technology in 2005. Main research directions: intelligent information processing, knowledge discovery, computational intelligence.
He presided over or participated in the completion of a batch of projects such as the National Science and Technology Support Program, the National Natural Science Foundation of China, and the Natural Science Foundation of Gansu Province, and more than 20 horizontal projects, and won the Gansu Province University Science and Technology Progress Award (first prize). Published more than 40 academic papers.

Xinxin Guo (1998.2-), female, from Huining, Gansu, is a postgraduate student of computer technology of Lanzhou University of Technology. Her research interests include protein networks and swarm intelligence optimization algorithms.
At present, she has published 2 SCI papers, 1 Chinese core paper, and won provincial awards in 2 competitions.