DOI: 10.24996/ijs.2024.65.9.41

ISSN: 0067-2904

Overlapping Structure Detection in Protein-Protein Interaction Networks Using a Modified Version of Particle Swarm Optimization

Dhuha Abdulhadi Abduljabbar*, Inas Ali Abdulmunem

Department of Computer Science, College of Sciences, University of Baghdad, Baghdad, Iraq

Received: 4/9/2023 Accepted: 26/1/2024 Published: 30/9/2024

Abstract

In today's world, the science of bioinformatics is developing rapidly, especially with regard to the analysis and study of biological networks. Scientists have used various nature-inspired algorithms to find protein complexes in protein-protein interaction (PPI) networks. These networks help scientists guess the molecular function of unknown proteins and show how cells work regularly. It is very common in PPI networks for a protein to participate in multiple functions and belong to many complexes, and as a result, complexes may overlap in the PPI networks. However, developing an efficient and reliable method to address the problem of detecting overlapping protein complexes remains a challenge since it is considered a complex and hard optimization problem. One of the main difficulties in identifying overlapping protein complexes is the accuracy of the partitioning results. In order to accurately identify the overlapping structure of protein complexes, this paper has proposed an overlapping complex detection algorithm termed OCDPSO-Net, which is based on PSO-Net (a well-known modified version of the particle swarm optimization algorithm). The framework of the OCDPSO-Net method consists of three main steps, including an initialization strategy, a movement strategy for each particle, and enhancing search ability in order to expand the solution space. The proposed algorithm has employed the partition density concept for measuring the partitioning quality in PPI network complexes and tried to optimize the value of this quantity by applying the line graph concept of the original graph representing the protein interaction network. The OCDPSO-Net algorithm is applied to a Collins PPI network and the obtained results are compared with different state-of-the-art algorithms in terms of precision (P), recall (R), and F-measure ($F-measure$). Experimental results confirm that the proposed algorithm has good clustering performance and has outperformed most of the existing recent overlapping algorithms.

.

Keywords: Overlapping structure, Protein complex detection, Protein-protein interaction network (PPI), Particle swarm optimization algorithm.

الكشف عن الهياكل المتداخلة في شبكات التفاعل البروتينية باستعمال نسخة معدلة من خوارزمية تحسين سرب الجسيمات

ضحى عبدالهادي عبدالجبار* , ايناس علي عبدالمنعم

قسم علوم الحاسوب, كلية العلوم, جامعة بغداد, بغداد, العراق

*Email: dhuha.abd@sc.uobaghdad.edu.iq

الخلاصة

في عالم اليوم، يتطور علم المعلوماتية الحيوية بشكل سريع، خاصة فيما يتعلق بتحليل ودراسة الشبكات
الاحيائية. قام الباحثون بتطبيق خوارزميات مختلفة مستوحاة من الطبيعة من أجل تحديد مركبات البروتين في
شبكات تفاعل البروتين البروتين (PPI) ، والتي يمكن من خلالها التنبؤ بالوظيفة الجزيئية للبروتينات غير
المعرفة، بالإضافة إلى الكشف عن انتظام نشاط الخلية. من الشائع جدًا في شبكات PPI أن يشارك البروتين
في وظائف متعددة وينتمي إلى العديد من المركبات، ونتيجة لذلك، قد تتداخل المركبات في شبكات PPI. ومع
ذلك، فإن تطوير طريقة فعالة وموثوقة لمعالجة مشكلة اكتشاف البنية المتداخلة لمركبات البروتين لا يزال يمثل
تحديًا كونها تعتبر مشكلة تحسين معقدة و صعبة للغاية. إحدى الصعوبات الرئيسية في تحديد مركبات
البروتين المتداخلة هي دقة نتائج التجميع. من أجل تحديد البنية المتداخلة لمركبات البروتين بدقة، اقترح هذا
البحث خوارزمية كشف معقدة متداخلة تسمى OCDPSO-Net والتي تعتمد على خوارزمية PSO-Net (
نسخة معدلة معروفة من خوارزمية تحسين سرب الجسيمات). يتكون إطار الطريقة المقترحة OCDPSO-
Net من ثلاث خطوات رئيسية تشمل: استراتيجية التهيئة، واستراتيجية حركة البحث لكل جسيم، وتعزيز القدرة
على البحث من أجل توسيع مساحة الحل. استعملت الخوارزمية المقترحة مفهوم كثافة التقسيم لقياس جودة
تجميع المركبات في شبكاتPPI ، وحاولت تحسين قيمة هذه الكمية من خلال تطبيق مفهوم الرسم البياني
الخطي للرسم البياني الأصلي الذي يمثل شبكة تفاعل البروتين. تم تطبيق الخوارزمية المقترحةOCDPSO-
Netعلى شبكة Collins PPI ومقارنة النتائج التي تم الحصول عليها مع خوارزميات حديثة مختلفة من
حيث precision (P), recall (R) و F-measure . أكدت النتائج التجريبية أن الخوارزمية المقترحة
OCDPSO-Netتتمتع بأداء واعد وإمكانيات كبيرة في تحديد البنية المتداخلة للمركبات البروتينية في شبكات
PPI.

1. Introduction

Many PPI networks have been mined as a result of the rapid advances in data processing, experimental methods, and computing technology [1, 2]. In accordance with previous studies, it has been stated that PPI networks can be created as complex, scale-free networks that meet small-world properties and high degrees of clustering [3]. In general, PPI networks represent a group of proteins connected among themselves by interactions [4–6], and the proteins within PPI networks are connected in two different types of cluster organizational structures, namely protein complexes and functional modules [7]. For simplicity, the terms protein complexes and functional modules have been used in this study as a single concept. Due to the completion of the Human Genome Project, many studies on the discovery of protein functional modules have become one of the hottest issues in computer science and life sciences [8–15].

Since many structures in the PPI network complexes may overlap, which means that a protein may belong to more than one complex, the identification of overlapping structures of protein complexes has broad prospects in different interesting applications such as the prediction of therapeutic targets, disease-causing genes, and the biological function of proteins. Many researchers have developed various algorithms in recent years that employ computational methodology to find protein complexes in PPI networks that have overlapping organizational structures [16]. Network clustering is one of the most effective and widely used methods for exploring and studying the overlapping topological and functional characteristics of PPI networks, among countless other alternatives [17, 18].

Mathematically, a graph is considered an efficient way, and often used in practice, to represent a complex network, where the graph nodes represent the network objects and the graph edges correspond to the connections between them [19–22]. Based on that, overlapping graph data can be represented using the line graph concept of the original graph, which allows nodes to be present in multiple clusters [19]. Recent years have witnessed great interest in solving the problem of detecting the overlapping structures of functional modules or protein complexes in PPI networks [23–33], but developing efficient and reliable methods for addressing the aforementioned problem remains a difficult and significant scientific problem in computational biology [34]. One of the main difficulties in discovering overlapping protein complexes is the accuracy of the partitioning results.

Accordingly, the main contributions of this paper are: (1) Enhancing the detection accuracy of the overlapping protein complexes by proposing an overlapping complex detection algorithm termed OCDPSO-Net using the modified version of particle swarm optimization (PSO-Net) that was proposed by Abdollahpouri et al. [35]. (2) Employing the concept of partition density D [36] to assess the division quality of the captured complexes in the PPI network, and optimizing the value of this quantity by applying the line graph concept $L(G)$ of the graph G [37, 19] representing the protein interaction network. (3) Applying the proposed algorithm to the Collins PPI network and comparing the obtained results with different state-of-the-art algorithms (namely, DMCL-EHO [33], K-means [32], MCODE [26], MCL [24], and ClusterOne [27]) in terms of R, P and $F - measure$. Experimental results confirm that the proposed algorithm, OCDPSO-Net, has promising performance and great potential for identifying complexes in PPI networks. The rest of the paper is processed as follows: Section 2 presents some recent related works. Section 3 describes the proposed OCDPSO-Net algorithm in detail. Section 4 presents the settings for the experiments, including the dataset used to evaluate the proposed algorithm, the evaluation metrics, and the experimental results obtained. The last section summarizes the paper with concluding remarks.

2. Related Works

Over the past years, a lot of algorithms have been proposed to address the problem of detecting overlapping or non-overlapping protein complexes in PPI networks [23]. For instance, Nepuse et al. [27] proposed the ClusterOne algorithm, which took advantage of the well-known greedy strategy. In general, the vertex or protein having the highest degree in the entire network was selected as the seed, and then, using a function with a greedy approach, some vertices were inserted or deleted from the seed. All the pair complexes in which the Overlap Score function's value exceeds a predetermined threshold are merged in order to discover overlapping clusters. Finally, complexes with a number of proteins lower than 3 or a density value below a predetermined threshold are eliminated. In the face of multiobjective optimization functions, Cao et al. [28] introduced the MOEPGA algorithm. The main aspect of the MOEPGA algorithm is to take advantage of various properties related to the network topology, such as size, name, distance, length, and density.

At first, the MOEPGA algorithm analyzed the structure of the PPI network and, based on the multiple topological characteristics of the network, formulated a multiobjective function. In each subgraph, three steps are taken, which are: generating an initial population, mutation, and selection. For clustering a PPI network by taking into account both biological properties and topological patterns. Sikandar et al. [29] introduced such an algorithm in which decision tree learners were used to model each complex subgraph. The authors employed a training set of known complexes to build decision trees, using the strategy of divide and conquer, in a best-first and depth-first manner.

Gu et al. [30] developed a Markov-based clustering algorithm that identified overlapping structures using link similarity information. They have demonstrated in their study that the proposed technique can discover more biologically significant modules in PPI networks.

Ramadan et al. [31] proposed a genetic algorithm for predicting protein complexes in a PPI network where its population was generated through spectral and random methods. In addition, genetic operations and different objective functions were used to improve the clustering quality. At first, the algorithm employed spectral clustering to determine the minimum cut that occurs between the subgraphs because, in general, the main aim of clustering is to capture subgraphs with the highest intraconnections. Accordingly, the authors used eigenvector, Laplacian, and diagonal matrices to derive the network's minimum cut [23]. Kalaivani et al. [32] adopted the k-means algorithm as an efficient algorithm for protein complex detection.

In 2019, Rani et al. [33] proposed the DMCL-EHO approach, which used Markov Clustering-based Elephant Optimization to apply clustering analysis to each subnetwork and find dynamic protein complexes. Shirmohammady et al. [23] recently proposed PPI-GA, a new clustering algorithm based on the genetic algorithm. To reduce the search space, they also came up with a new way to encode solutions. In addition, the authors adopted a new multiobjective quality function to optimize both cluster intraconnections in terms of maximization language and cluster interconnections in terms of minimization language.

3.The Proposed OCDPSO-Net to Identify Overlapping Protein Complexes

The PSO algorithm is an exciting and ever-expanding research topic that falls under the category of swarm intelligence algorithms, meaning "working with a set of particles." It has been the focus of attention for many scholars from various domains to tackle different challenging issues [38–42]. There are two crucial steps in the modified version of particle swarm optimization (PSO-Net) that Abdollahpouri et al. [35] proposed: (1) initialization and (2) moving. In PSO-Net, the particles of the algorithm are approaching their local and global best positions by taking part in crossover operations. After that, to spread within the search space, a mutation operator is applied to each particle. PSO-Net optimizes each particle based on the modularity measure for which it is selected as a fitness function.

Based on the PSO version of Abdollahpouri et al. [35], the framework of the proposed OCDPSO-Net algorithm consists of three main steps, including: initialization strategy (i.e., particle structure representation scheme (as shown in Section 3.1) and fitness computation (as shown in Section 3.2)), movement strategy for each particle (i.e., search strategy) (as shown in Section 3.3), and enhancing search ability in order to expand the solution space (as shown in Section 3.4). Figure 1 shows the suggested method's flowchart, and details of each step are presented in the subsections that follow.

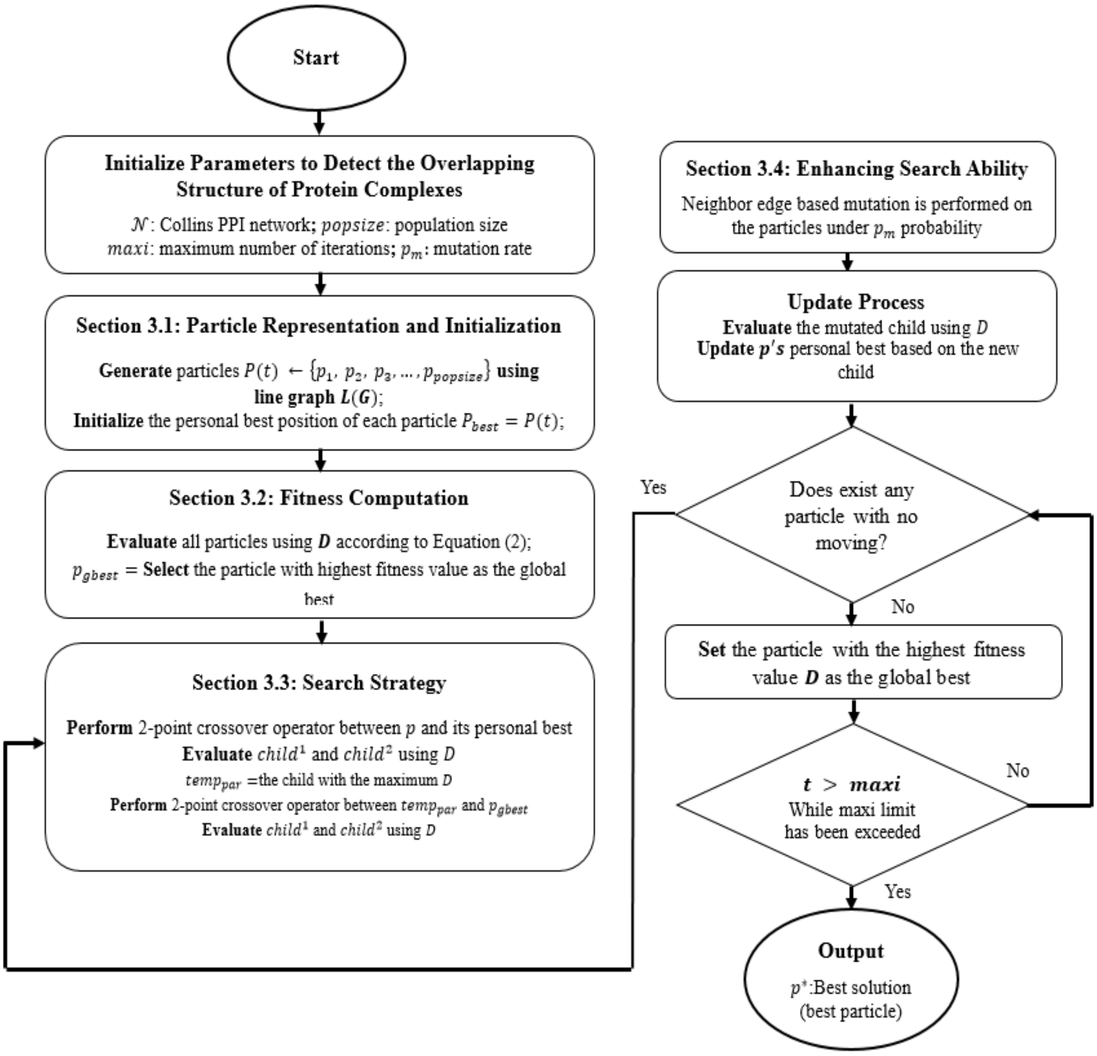

Start
Initialize Parameters to Detect the Overlapping Structure of Protein Complexes
$\mathcal{N}$: Collins PPI network; $popsize$: population size
$maxi$: maximum number of iterations; p_m: mutation rate
Section 3.1: Particle Representation and Initialization
Generate particles $P(t) \leftarrow \{p_1, p_2, p_3, \ldots, p_{popsize}\}$ using line graph $L(G)$;
Initialize the personal best position of each particle $P_{best} = P(t)$;
Section 3.2: Fitness Computation
Evaluate all particles using D according to Equation (2);
p_{gbest} = Select the particle with highest fitness value as the global best
Section 3.3: Search Strategy
Perform 2-point crossover operator between p and its personal best
Evaluate $child^1$ and $child^2$ using D
$temp_{par}$ =the child with the maximum D
Perform 2-point crossover operator between $temp_{par}$ and p_{gbest}
Evaluate $child^1$ and $child^2$ using D
Section 3.4: Enhancing Search Ability
Neighbor edge based mutation is performed on the particles under p_m probability
Update Process
Evaluate the mutated child using D
Update $p's$ personal best based on the new child
Yes
Does exist any particle with no moving?
No
Set the particle with the highest fitness value D as the global best
No
$t > maxi$
While maxi limit has been exceeded
Yes
Output
p^*:Best solution (best particle)

Figure 1: Flowchart of the proposed OCDPSO-Net algorithm. The algorithm is based on the PSO version of [35]

3.1Particle Representation and Initialization

Given a graph $G = (V, E)$ representing a PPI network $\mathcal{N}$et, the proposed PSO algorithm has been applied to the line graph $L(G)$ of G. The $L(G)$ of an undirected graph G is another graph $L(G)$ where: 1) each $L(G)$'s node is an edge of G, 2) two nodes of $L(G)$ are adjacent if and only if their corresponding edges in G have a common end node. Thus, a $L(G)$ represents the adjacency between G's edges. The clustering approach based on the line graph yields an overlapping graph partitioning of the original graph G, thus allowing nodes to be present in multiple clusters [37]. OCDPSO-Net encodes on edges of the network, where a particle p in the population consists of m edges $\{p_0, p_1, \cdots, p_i, \cdots, p_{m-1}\}$, in which $i \in \{0, \cdots, m - 1\}$ is the edges' identifier (i.e., links), m is the edges' number, and each p_i at random selects one from the edges adjacent to edge i.

According to graph theory, if two edges share one node in an undirected graph, then they are adjacent. A genotype is transferred to an edge complex partition during the decoding phase. Since p_i has the value j, edge i and edge j share a common node. Therefore, they ought to be classified under the same component. During the decoding process, it is necessary to determine all linked components. All edges that belong to the same connected component are mapped into a single complex. This decoding step can be done in linear time [36]. In the process of initializing a population, a set of particles is randomly generated. For each particle p, at random, each p_i is assigned one of its adjacent links.

3.2Fitness Computation

In OCDPSO-Net, the objective function's value represents the particles' fitness, which indicates the particles' merit. The objective function directs the random search of OCDPSO-Net. In overlapping protein complexes, one protein is allowed to belong to more than one complex, which makes the conventional complex (or community) definitions unreasonable [36, 43]. As a result, a new definition of community has been adopted and employed in many nature-inspired algorithms to detect overlapping communities. Ahn *et al.* [43] proposed the partition density D for overlapping communities, which evaluates the edge density within communities. For a network with M links, suppose $C = \{c_1, \dots.., c_k\}$ as a partition of the network's links into k subsets, $m_{c_i} = |c_i|$ is the number of links in subset c_i, $n_{c_i} = \left|\cup_{e_{ab} \in c_i} \{a, b\}\right|$ refers to the nodes' number that incident to links in subset c_i. Note that $\sum_{i=1}^{k} m_{c_i} = M$ and $\sum_{i=1}^{k} n_{c_i} \geq N$ (assuming no unconnected nodes). Accordingly, D_{c_i} refers to the link density of subset c_i, Equation (1) shows its definition [36, 43]:

$$D_{c_i} = \frac{m_{c_i} - (n_{c_i} - 1)}{n_{c_i}(n_{c_i} - 1)/2 - (n_{c_i} - 1)} \qquad (1)$$

The partition density, D, is the average of D_{c_i}, over all communities, weighted by the fraction of links presenting in each community. Equation (2) shows its definition [36, 43]:

$$D = \frac{2}{M} \sum_{i=1}^{k} m_{c_i} \frac{m_{c_i} - (n_{c_i} - 1)}{(n_{c_i} - 2)(n_{c_i} - 1)} \qquad (2)$$

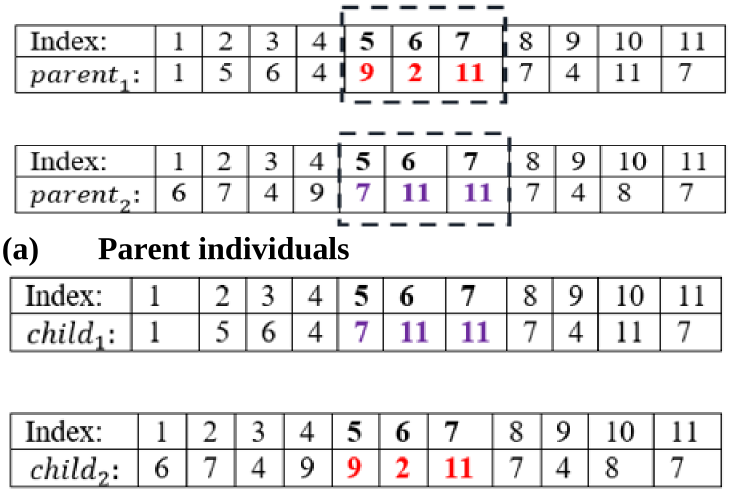
Equation (2) does not possess a resolution limit since each term is local in c_i [43].

3.3 Search Strategy

The search strategy of the PSO is based on moving each particle toward its local best and also moving it toward the swarm's global best. Genetic operators like crossover and mutation based on line graph representation are used to move each particle toward the best positions. The main steps of the movement strategy in the OCDPSO-Net are presented as follows [35, 44, 45].

3.3.1 Moving toward personal best

Firstly, each particle executes a two-point crossover with its personal best. The outcome of the crossover process is the production of two new particles. After that, the fitness values of the obtained outcomes are compared, and the solution that achieves a higher value of fitness, or higher partition density (D), is selected as a temporary position for the present addressed particle. Figure 2 illustrates an example of how the two-point crossover operator can be applied. Figure 2 (a) shows two arbitrary solutions, $parent_1$ and $parent_2$ considered as parents, and two random points (i.e., two random edges' identifiers) $i = 5$ and $j = 7$ are determined. Figure 2(b) illustrates the crossover operator. To generate $child_1$ (the first child), the values of the edges from the particle's beginning to i (the first crossover point) are copied from the parent $parent_1$, the part from i to j (the second crossover point) is copied from the parent $parent_2$, and the rest is copied from $parent_1$. This procedure is carried out in reverse order to create $child_2$ (the second child) [35, 44].



Index:	1	2	3	4	5	6	7	8	9	10	11
$parent_1$:	1	5	6	4	9	2	11	7	4	11	7

Index:	1	2	3	4	5	6	7	8	9	10	11
$parent_2$:	6	7	4	9	7	11	11	7	4	8	7

(a) Parent individuals

Index:	1	2	3	4	5	6	7	8	9	10	11
$child_1$:	1	5	6	4	7	11	11	7	4	11	7

Index:	1	2	3	4	5	6	7	8	9	10	11
$child_2$:	6	7	4	9	9	2	11	7	4	8	7

(b) Crossover operator

Figure 2: Illustration of a 2-point crossover (a) illustrates the representation of two arbitrary particles. (b) Shows the crossover operator

3.3.2 Moving toward global best

In this step, a two-point crossover is carried out between the global best and each particle in the population so as to move the particles toward the identified global best point. After this process, two new solutions are generated, and the temporary state of the present particle is chosen from the resulting solutions based on the highest fitness value (i.e., the highest partition density, D value).

3.4 Enhancing Search Ability

Finally, neighbor edge-based mutation is applied to the particles under the predetermined probability of the mutation operator [46] in order to move the solutions around the entire search space. To ensure that there are only possible and feasible solutions in the solution space, a gene or edge identifier i is randomly chosen for the selected mutated particle, where the possible values of gene i are restricted to its adjacent edges. The outcome of the mutation operator is compared to the native personal best value of the mutated particle, and the personal best value is replaced by the mutated particle value if the mutated particle outperforms it. Otherwise, the personal best value stays the same. After all the particles have been moved and their personal bests have been updated, fitness values are then computed again using the partition density (D) measure. The particle with the highest fitness value is chosen as the global best of the entire swarm. This procedure is carried out over a predetermined number of iterations [35, 44].

4. Experiments and Evaluation

4.1 Dataset

Since there are numerous high-confidence datasets of the PPI network from the yeast organism [47], they are selected and studied in this paper. In the research experiment, the Collins PPI network derived from the BioGrid dataset has been used [31, 48] to test the performance of the proposed overlapping clustering algorithm. There are 1,004 proteins (or nodes) and 8,319 interactions (or edges) in this network. It has an average degree equal to 16.57, where a node's degree in a network is determined by the number of links it has with other nodes; on the other hand, this network has a density of 0.016, which is the ratio of total connections to connections that may potentially exist in the network. The network has been illustrated as an interactive graph in Figure 3 using Gephi software. Gephi is a software package written in Java for network analysis and visualization. High-quality datasets are needed as gold standards or reference complexes to validate the quality of the clustering approaches [31]. Collections of yeast protein complexes taken from literature and listed in the MIPS (known yeast genome database) have been used [49], in addition to a hand-curated reference complex set termed CYC2008 [50]. Table 1 presents a summary of the characteristics of the dataset used.

Table 1: Summary of the characteristics of the dataset used

Network Category	Biological
Network Name	Collins PPI network
Network Pattern	Overlapping complex structure
Number of Proteins	1,004
Number of Interactions	8,319
Average Degree	16.57
Density	0.016
Reference Complex Sets	MIPS and CYC2008

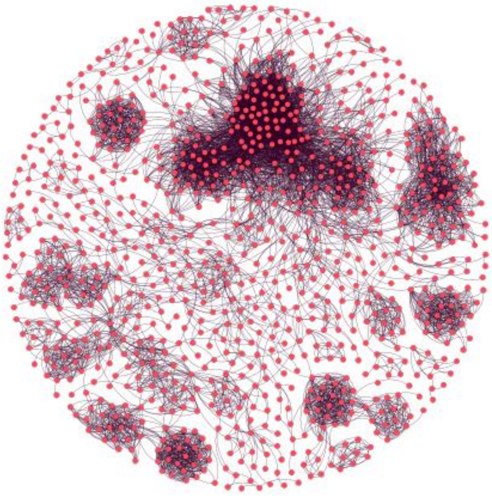


Figure 3: Collins interactive network in Gephi software

4.2 Evaluation Measures

The P (i.e., positive predictive value), R (i.e., sensitivity), and $F-measure$ assessment metrics are used for assessing the predicted clusters' quality. All of these criteria will be calculated over pairs of nodes (in this case, proteins). According to the obtained overlapped clustering results, for each protein's pair that participates in at least one complex, these evaluation measures attempt to determine if the prediction of this protein's pair is correct, that is, if it is mapped to the same cluster of underlying complexes in the database (i.e., the true reference complexes) [23]. Given a predicted cluster C and a reference protein complex G, let true positive (TP) refer to the set of proteins common to the complex C and complex G, false positive (FP) refers to the set of proteins present only in the complex C, and true negative (TN) refers to the set of proteins that are members of the complex G but are not members of complex C [31]. Hence, $P, R,$ and $F-measure$ can be formally expressed as [23, 31]:

$$P = \frac{TP}{TP+FP} \qquad (3)$$

$$R = \frac{TP}{TP+TN} \qquad (4)$$

$$F-measure = \frac{2\times(P\times R)}{P+R} \qquad (5)$$

4.3 The Results

In this section, the analysis of the results of experiments on the Collins network is presented in detail. The proposed algorithm (OCDPSO-Net) is simulated using the MATLAB platform. The number of iterations in OCDPSO-Net was set to 50, and the size of the population was customized to 100. In addition, we have discussed the results obtained by the OCDPSO-Net algorithm on the Collins PPI network when mutation rate $p_m = \{0.0,0.3,0.5,0.7,1\}$. Since the OCDPSO-Net algorithm is a random optimization method, all test results achieved by the proposed method are computed over twenty individual runs.

Figures 4 and 5 show clustering results achieved by OCDPSO-Net on MIPS and CYC2008 complexes reference, respectively, in terms of $R, P,$ and $F - measure$ when $p_m = \{0.0,0.3,0.5,0.7,1\}$. These figures show that, given the different measures, the results obtained by the proposed algorithm in different individual runs are not significantly different. When $p_m = 0.7$, the proposed algorithm reported the best average values in terms of $F - measure$ which are equal to 0.4431 and 0.6603 when using MIPS and CYC2008 as reference complexes, respectively. Therefore, we set $p_m = 0.7$ in this study.

According to [23], the algorithm's stability points out that the algorithm's performance is correct. When an algorithm is stable, the results it produces are not scattered, which means that even if it is executed many times on the same sample and under the same settings, the results' quality is not very different in all the tests and will be within about the same range. To look more closely and to demonstrate the proposed algorithm's stability, the results of twenty individual runs in terms of $F - measure$ have been collected when $p_m = 0.7$ and presented in Figures 6 and 7 for the Collins data sample when CYC2008 and MIPS have been used as reference clusters, respectively. As shown in Figures 6 and 7, the obtained $F - measure$ values are within about one range and not much different, so the solutions obtained from the proposed algorithm have good stability.

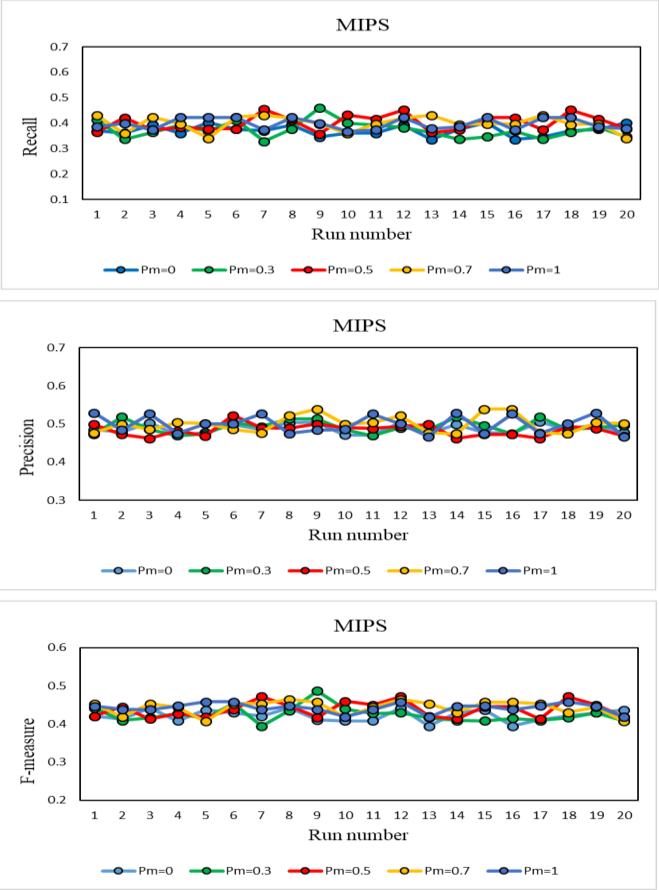


Figure 4: Collins data sample clustering results in twenty individual runs on MIPS complexes referenced in terms of $R, P,$ and $F - measure$ when $p_m = \{0,0.3,0.5,0.7,1\}$

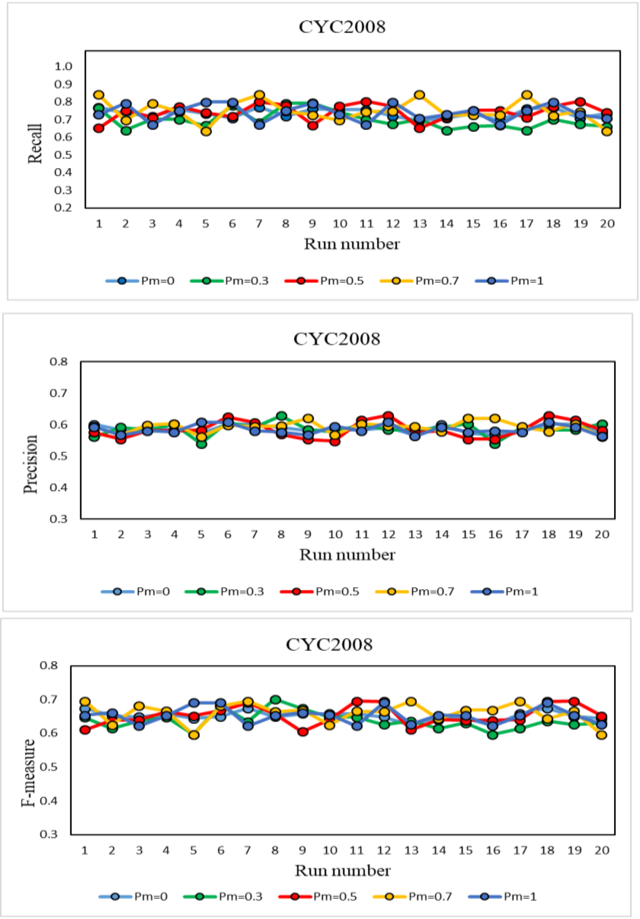


Figure 5: Collins data sample clustering results in twenty individual runs on CYC2008 complexes referenced in terms of $R, P,$ and $F - measure$ when $p_m \; = \{0,0.3,0.5,0.7,1\}$

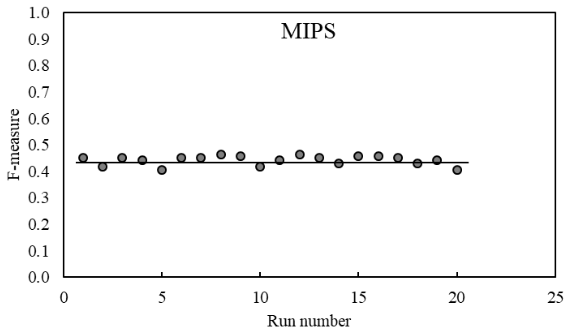


Figure 6: The $F - measure$ stability diagram for the MIPS data sample

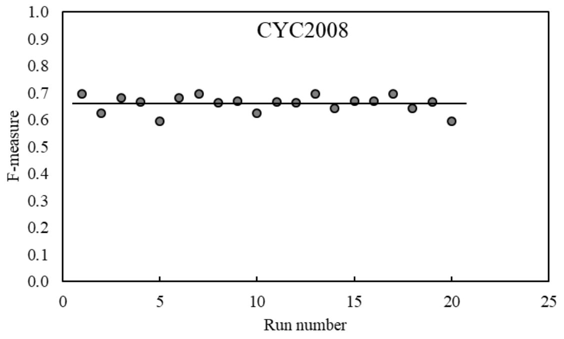


Figure 7: The $F - measure$ stability diagram for the CYC2008 data sample

Lastly, the proposed algorithm (OCDPSO-Net) has been compared with some recent state-of-the-art algorithms. Comparison algorithms fall into the category of overlapping community detection algorithms, including DMCL-EHO [33], K-means [32], MCODE [26], MCL [24], and ClusterOne [27]. The comparison results in terms of $P, R,$ and $F - measure$ are presented in Table 2. It should be mentioned that the results of all the counterpart state-of-the-art algorithms in Table 2 are taken from the reference [23]. In general, compared to the other state-of-the-art algorithms, our proposed algorithm reports the highest results in terms of P , R and $F - measure$ in clustering MIPS references and in terms of R and $F - measure$ in clustering CYC2008 references, where the first highest result in terms of P has been reported by the K-means clustering algorithm. This is due to the fact that the particles' representation of OCDPSO-Net guarantees convergence, and employing the PSO algorithm as a search engine guarantees fast convergence. In addition, selecting partition density D as an objective function ensures the quality of the overlapping complex detection and helps in finding protein complex structures that are closer to the gold standard ones.

Looking closely, Table 2 demonstrates that the OCDPSO-Net clustering algorithm has the same accuracy values as the MCODE algorithm, which are 0.48 and 0.59 in clustering the MIPS and CYC2008 references, respectively. For the R measure, OCDPSO-Net records a value of 0.43 and 0.84 compared to MCODE results of 0.27 and 0.66. Regarding the $F-measure$, it is obviously clear from Table 2 that the proposed algorithm has yielded a higher value than the MCODE algorithm. On the other hand, the suggested algorithm exhibits higher values in terms of $P, R,$ and $F-measure$ when compared to MCL, which employs the random walk process for clustering, as well as higher $P, R,$ and $F-measure$ values when compared to ClusterOne, which is based on network density.

As for the comparison of the OCDPSO-Net result with DMCL-EHO and K-means clustering, it is shown that OCDPSO-Net recorded the best solutions in all cases except in clustering the CYC2008 in terms of P. In summary, OCDPSO-Net has presented good clustering performance and outperformed the existing recent overlapping algorithms.

Table 2: Comparison of OCDPSO-Net algorithm results with existing recent overlapping algorithms

Method	**CYC2008**			**MIPS**		
	R	P	$F-m$	R	P	$F-m$
DMCL-EHO	0.74	0.61	0.66	0.42	0.41	0.41
K-means clustering	0.55	**0.67**	0.6	0.28	0.4	0.32
MCODE	0.66	0.59	0.63	0.27	**0.48**	0.35
MCL	0.65	0.45	0.54	0.27	0.34	0.3
ClusterOne	0.55	0.43	0.49	0.2	0.34	0.25
OCDPSO-Net (our)	**0.84**	0.59	**0.69**	**0.43**	**0.48**	**0.45**

5 Conclusion

The main contribution of this research is to propose a robust clustering algorithm, termed OCDPSO-Net, to identify the overlapping structures of protein complexes in PPI networks. The proposed algorithm makes use of the new version of the PSO algorithm to improve the clustering accuracy of protein complexes, uses the concept of partition density to measure the clustering quality in complexes of a PPI network, and attempts to maximize this quantity by applying the PSO algorithm to the line graph $L(G)$ of the graph G representing the protein interaction network. The OCDPSO-Net algorithm has been applied to the Collins PPI network and compared with a number of competent state-of-the-art algorithms in terms of $R, P,$ and $F-measure$. Experimental results have shown that OCDPSO-Net has promising performance and great potential for identifying complexes in the Collins PPI network. Future research will focus on enhancing the algorithm's performance by developing a local problem-specific operator and integrating it with multiobjective models. Also, it is interesting to investigate the ability of the proposed algorithm in other types of complex biological networks, such as metabolic networks.

References

[1] Y. Chen, W. Wang, J. Liu, J. Feng, and X. Gong, "Protein interface complementarity and gene duplication improve link prediction of protein–protein interaction network," *Front Genet.*, vol. 11-2020, pp. 291, 2020. Doi: 10.3389/fgene.2020.00291.

[2] M. N. Mohamed, & M. F. Altaee, “Evaluation of Caspase 8 Role as a Gene and Protein in Chronic Myeloid Leukemia Incidence,” *Iraqi Journal of Science,* vol. 64, no. 8, pp. 3914–3924, 2023. https://doi.org/10.24996/ijs.2023.64.8.18.

[3] J. Ji, A. Zhang, C. Liu, X. Quan, and Z. Liu, “Survey: functional module detection from protein-protein interaction networks,” *IEEE Trans. Knowl. Data Eng*., vol. 26, no. 2, pp. 261–277, 2012. Doi: 10.1109/tkde.2012.225.

[4] Abdulateef, D. A. Alzubaydi, & B. A. Attea “An Evolutionary Algorithm with Gene Ontology-Aware Crossover Operator for Protein Complex Detection,” *Iraqi Journal of Science*, vol. 64, no. 4, pp: 1975-1987, 2023. DOI: 10.24996/ijs.2023.64.4.341975-1987.

[5] M. A. Kadhim, & R. D. Al-Dabbagh, “An Evolutionary-Based Mutation with Functional Annotation to Identify Protein Complexes Within PPI Networks,” *Iraqi Journal of Science*, vol. 64, no. 10, pp: 5416-5427, 2023. DOI: 10.24996/ijs.2023.64.10.436316-6327.

[6] A. S. Alhendi, “ A Review: Protein Identification by LC-MS: Principles, Instrumentation, and Applications: Protein Identification by LC-MS,” *Iraqi Journal of Science,* vol. 61, no. 10, pp. 2448–2466, 2020. DOI: 10.24996/ijs.2020.61.10.2.

[7] S. Ray, M. De, and A. Mukhopadhyay, “A multiobjective go based approach to protein complex detection,” *Procedia Technology,* vol. 4, pp. 555-560, 2012. Doi: 10.1016/j.protcy.2012.05.088.

[8] K. Ying, and S. Lin, “Maximizing cohesion and separation for detecting protein functional modules in protein-protein interaction networks,” *PLoS One*, vol. 15, no. 10, pp. e0240628, 2020. Doi: 10.1371/journal.pone.0240628.

[9] A. Elmsallati, C., and J. Kalita, “Global alignment of protein-protein interaction networks: A survey,” *IEEE/ACM Transactions on Computational Biology and Bioinfor-matics*, vol. 13, no. 4, pp. 689-705, 2016. https://doi.org/10.1109/TCBB.2015.2474391.

[10] R. Wang, H., Ma & C. Wang, “An improved memetic algorithm for detecting protein complexes in protein interaction networks,” *Frontiers in genetics,* vol. 12, p. 794354, 2021.

[11] R. Wang, C. Wang, & H. Ma, “Detecting protein complexes with multiple properties by an adaptive harmony search algorithm,” *BMC bioinformatics*, vol. 23, no. 1, pp. 1-32, 2022.

[12] R. Wang, H. Ma, & C. Wang, “ An ensemble learning framework for detecting protein complexes from PPI networks,” *Frontiers in Genetics*, vol. 13, p. 839949, 2022.

[13] M. S. Islam, M. R. Islam, & A. S. Ali, “Protein complex prediction in large protein–protein interaction network,” *Informatics in Medicine Unlocked,* vol. 30, p. 100947, 2022. https://doi.org/10.1016/j.imu.2022.100947.

[14] Y. Karakuş, & V. Altunta, “Protein complex detection from protein-protein interaction networks with machine learning methods,” 2023. Doi: 10.5505/pajes.2023.56887

[15] T. R. Sahoo, S. Patra, & S. Vipsita , “Decision tree classifier based on topological characteristics of subgraph for the mining of protein complexes from large scale PPI networks,” *Computational Biology and Chemistry,* vol. 106, p. 107935, 2023. https://doi.org/10.1016/j.compbiolchem.2023.107935

[16] H. Wu, L. Gao, J. Dong, & X. Yang, “Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks,” *PloS one,* vol. 9, no.3, p. e91856, 2014

[17] S. S. Bhowmick, and B. S. Seah, “Clustering and summarizing protein-protein interaction networks: a survey,” *IEEE Trans. Knowl. Data Eng*., vol. 28, no. 3, pp. 638–658, 2015. Doi: 10.1109/tkde.2015.2492559.

[18] Y. Wang, Q. Chen, L. Yang, S. Yang, K. He and X. Xie, “Overlapping Structures Detection in Protein-Protein Interaction Networks Using Community Detection Algorithm Based on Neighbor Clustering Coefficient,” *Frontiers in genetics,* vol. 12, p. 689515, 2021. https://doi.org/10.3389/fgene.2021.689515.

[19] J. Xie, S. Kelley, & B. K. Szymanski, “Overlapping community detection in networks: The state-of-the-art and comparative study,” *Acm computing surveys (csur),*vol. 45, no. 4, pp. 1-35, 2013.

[20] D. A. Abduljabbar, S. Z. M. Hashim, and R. Sallehuddin, “An Evolutionary Algorithm for Community Detection Using an Improved Mutation Operator,” In *2019 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pp. 406-410, 2019. IEEE.

[21] D.A. Abduljabbar, S. Z. M. Hashim, and R. Sallehuddin, “An enhanced evolutionary algorithm with local heuristic approach for detecting community in complex networks,” *Journal of Theoretical and Applied Information Technology,* vol. 97, no. 20, pp. 2452-2466, 2019.

[22] W. Zheng, J. Sun, Q. Zhang, & Z. Xu, "Continuous Encoding for Overlapping Community Detection in Attributed Network," *IEEE Transactions on Cybernetics,* vol. 53, no. 9, pp. 5469-5482, Sept. 2023, Doi: 10.1109/TCYB.2022.3155646.

[23] N. Shirmohammady, H. Izadkhah, and A. Isazadeh, "PPI-GA: A Novel Clustering Algorithm to Identify Protein Complexes within Protein-Protein Interaction Networks Using Genetic Algorithm," *Complexity,* vol. 2021, p. 2132516, 2021. https://doi.org/10.1155/2021/2132516.

[24] S. V. Dongen, "A cluster algorithm for graphs," *Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science, Amsterdam, The Netherlands,* 2000.

[25] S. Brohee and V. Helen, "Evaluation of clustering algorithms for protein-protein interaction networks," *BMC Bioinformatics*., vol. 7, p. 488, 2006. https://doi.org/10.1186/1471-2105-7-488.

[26] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," *BMC Bioinformatics*., vol. 4, no. 1, pp. 1-27, 2003. https://doi.org/10.1186/1471-2105-4-2.

[27] T. Nepuse, H. Yu, and A. Paccanaro, "Detection overlapping protein complexes in the protein-protein interaction network," *Nature Methods,* vol. 9, no. 5, p. 471, 2012. https://doi.org/10.1038/nmeth.1938.

[28] B. Cao, J. Luo, C. Liang, S. Wang, and D. Song, "MOEPGA: A novel method to detect protein complexes in yeast protein–protein interaction networks based on MultiObjective Evolutionary Programming Genetic Algorithm," *Computational Biology and Chemistry,* vol. 58, pp. 173–181, 2015. https://doi.org/10.1016/j.compbiolchem.2015.06.006.

[29] A. Sikandar, W. Anwar, U. I. Bajwa et al., "Decision tree based approaches for detecting protein complex in protein protein interaction network (PPI) via link and sequence analysis," *IEEE Access,* vol. 6, pp. 22108–22120, 2018. Doi: 10.1109/ACCESS.2018.2807811

[30] L. Gu, Y. Han, C. Wang, W. Chen, J. Jiao, and X. Yuan, "Module overlapping structure detection in PPI using an improved link similarity-based Markov clustering algorithm," *Neural Computing and Applications,* vol. 31, pp. 1481–1490, 2019. https://doi.org/10.1007/s00521-018-3508-z

[31] E. Ramadan, A. Naef, and M. Ahmed, "Protein complexes predictions within protein interaction networks using genetic algorithms," *BMC bioinformatics,* vol. 17, no. 7, pp. 481-489, 2016. https://doi.org/10.1186/s12859-016-1096-4.

[32] S. Kalaivani, D. Ramyachitra, and P. Manikandan, "K-means clustering: an efficient algorithm for protein complex detection," in *Progress in Computing, Analytics and Networking,* Springer, Singapore, vol. 710, pp. 449–459, 2018. https://doi.org/10.1007/978-981-10-7871-2_43

[33] R. R. Rani, D. Ramyachitra, and A. Brindhadevi, "Detection of dynamic protein complexes through Markov clustering based on elephant herd optimization approach," *Scientific Reports,* vol. 9, no. 1, p. 11106, 2019. https://doi.org/10.1038/s41598-019-47468-y.

[34] Y. Mao, and Y. Liu, "Functional module mining in uncertain PPI network based on fuzzy spectral clustering," *J. Comput*., vol. 31, no. 4, pp. 91–106, 2020. Doi: 10.3966/ 199115992020083104008.

[35] A. Abdollahpouri, S. Rahimi, S. M. Majd, and C. Salavati, "A modified particle swarm optimization algorithm for community detection in complex networks," In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction,* Springer, Cham, vol. 1101, pp. 11-27, 2018, Doi: 10.1007/978-3-319-99740-7_2.

[36] C. Shi, Y. Cai, D. Fu, Y. Dong, and B. Wu, "A link clustering based overlapping community detection algorithm," *Data & Knowledge Engineering,* vol. 87, pp. 394-404, 2013. https://doi.org/10.1016/j.datak.2013.05.004.

[37] C. Pizzuti, "Overlapped community detection in complex networks," In *Proceedings of the 11th Annual conference on Genetic and evolutionary computation,* pp. 859-866, July, 2009. https://doi.org/10.1145/1569901.1570019.

[38] A. J. Mohammed, K. I. Ghathwan, & Y. Yusof , "Optimal robot path planning using enhanced particle swarm optimization algorithm," *Iraqi Journal of Science,* vol. 61, no. 1, pp. 178-184, 2020. DOI: 10.24996/ijs.2020.61.1.20.

[39] H. A. Chachan, "Using Non-dominated Sorting Particle Swarm Optimization Algorithm II for Bi-objective Flow Shop Scheduling Problems," *Iraqi Journal of Science,* vol. 62, no. 1, pp. 275 - 288, 2021. DOI:10.24996/ijs.2021.62.1.26.

[40] R. N. Jawad, " Proposed Hybrid Technique in Cryptanalysis of Cryptosystem Based on PSO and SA*,"* *Iraqi Journal of Science*, vol. 63, no. 10, pp. 4547-4558, 2022. DOI: 10.24996/ijs.2022.63.10.37.

[41] S. A. Alsaidy, & N. A. Abdullah, " Power-Efficient Virtual Machine Placement in Cloud Datacenters using Heuristic Assisted Enhanced Discrete Particle Swarm Optimization*,"* *Iraqi Journal of Science*, vol. 63, no. 10, pp. 4499-4517, 2022. DOI: 10.24996/ijs.2022.63.10.34 4499-4517.

[42] D. A. Abduljabbar, "Parallel Particle Swarm Optimization Algorithm for Identifying Complex Communities in Biological Networks," *Iraqi Journal of Science*, vol. 65, no. 1, pp 1-17, 2024. DOI: 10.24996/ijs.2024.65.1.40.

[43] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, pp. 761-764, 2010. https://doi.org/10.1038/nature09182.

[44] S. Rahimi, A. Abdollahpouri and P. Moradi, "A multi-objective particle swarm optimization algorithm for community detection in complex networks," *Swarm and Evolutionary Computation*, vol. 39, pp. 297-309, 2018. https://doi.org/10.1016/j.swevo.2017.10.009.

[45] D. A. Abduljabbar, "Community Detection in Modular Complex Networks Using an Improved Particle Swarm Optimization Algorithm", *Iraqi Journal of Science*, vol. 64, no. 8, pp. 4228-4243, 2023. https://doi.org/10.24996/ijs.2023.64.8.41.

[46] D. A. Abduljabbar, S. Z. M. Hashim and R. Sallehuddin, "An enhanced evolutionary algorithm for detecting complexes in protein interaction networks with heuristic biological operator," In *Recent Advances on Soft Computing and Data Mining: Proceedings of the Fourth International Conference on Soft Computing and Data Mining (SCDM 2020)*, Melaka, Malaysia, vol. 978, pp. 334-345, January 22– 23, 2020. Springer International Publishing. https://doi.org/10.1007/978-3-030-36056-6_32.

[47] N. Zhang, and E. Bilsland, "Contributions of Saccharomyces cerevisiae to understanding mammalian gene function and therapy," In *Yeast Systems Biology*, vol. 759, pp 501-523, 2011. Humana Press. https://doi.org/10.1007/978-1-61779-173-4_28.

[48] T. Reguly, A. Breitkreutz, L. Boucher, B. J. Breitkreutz, G. C. Hon, C. L. Myers, ... and M. Tyers, "Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae," *Journal of biology*, vol. 5, no. 4, pp. 1-28, 2006. https://doi.org/10.1186/jbiol36.

[49] W. Mewes, D. Frishman, U. Güldener, G. Mannhaupt, K. Mayer, M. Mokrejs, and B. Weil, "MIPS: a database for genomes and protein sequences," *Nucleic acids research*, vol. 30, no. 1, pp. 31-34, 2002. https://doi.org/10.1093/nar/30.1.31.

[50] S. Pu, J. Wong, B. Turner, E. Cho, and S. J. Wodak, "Up-to-date catalogues of yeast protein complexes," *Nucleic acids research*, vol. 37, no. 3, pp. 825-831, 2009. https://doi.org/10.1093/nar/gkn1005.