**ISSN: 0067-2904**

# Evolutionary-based Gene Ontology for Complex Detection in Protein-Protein Interaction Networks

**Mustafa Abdulhussein Kadhim *, Rawaa Dawoud Al-Dabbagh**
*Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq*

Received: 10/11/2022 Accepted: 19/3/2023 Published: 29/2/2024

**Abstract**

Complex detection in protein-protein interaction (PPI) networks is one of the major issues facing scientific study in biological networks. In PPINs, proteins are distributed differently as groups (complexes). These groups can be identified as having a great internal density in the number of edges inside the groups while having the least possible number of edges between these groups. The most common methods for finding such complexes are evolutionary algorithms (EAs), which have been used widely in literature for this objective. Despite the reliability of these complicated detection models, they are mostly based on topological (graph) qualities, and the biological implications of the PPI networks have been rarely explored. In this research, EA with mutation-based gene ontology is developed, particularly in the mutation part where the functional annotation of the protein has been considered using gene ontology structure. The experimental results prove the reliability of the proposed method using standard validation measures. It also outperforms the state-of-the-art method in terms of the prediction ability and quality of the complexes found.

**Keywords:** Complex detection, Evolutionary algorithm, Protein complexes, Gene ontology, Functional Annotation.

# خوارزمية تطورية قائمة على علم الوجود الجيني لكشف المعقدات في شبكات التفاعل البروتينية

**مصطفى عبد الحسين كاظم*, رواء داود الدباغ**
قسم علوم الحاسوب, كلية العلوم, جامعة بغداد, بغداد, العراق

**الخلاصة**

يعد اكتشاف المعقدات في شبكات تفاعل البروتين – بروتين (PPINs) أحدى المشاكل الرئيسية التي تواجه الدراسة العلمية في الشبكات البيولوجية. في PPINs، يتم توزيع البروتينات بشكل مختلف كمجموعات (معقدات) حيث يمكن تحديد هذه المجموعات على أنها تتمتع بكثافة داخلية كبيرة في عدد الحواف داخل المجموعات مع وجود أقل عدد ممكن من الحواف بين هذه المجموعات. احدى الطرق الأكثر شيوعًا للعثور على مثل هذه المجمعات هي الخوارزميات التطورية (EAs)، والتي تم استعمالها على نطاق واسع في الدراسات السابقة لتحقيق هذا الهدف. وعلى الرغم من موثوقية هذه النماذج في كشف المعقدات، إلا أنها تعتمد في الغالب على الصفات الطوبولوجية، ونادرًا ما تم استكشاف الآثار البيولوجية لشبكات PPI. في هذا البحث، تم تطوير EA مع علم الوجود الجيني القائم على الطفرة، لا سيما في جزء الطفرة حيث تم الأخذ بنظر الاعتبار التعليق التوضيحي الوظيفي للبروتين باستعمال هيكلية علم الوجود الجيني. أثبتت النتائج التجريبية موثوقية الطريقة المقترحة باستعمال

*Email: Mustafa.a@sc.uobaghdad.edu.iq

مقاييس التحقق القياسية. كما أنها تفوقت أيضًا على أحدث الأساليب السابقة من حيث القدرة على التنبؤ وجودة المعقدات المستكشفة.

## 1. Introduction

Biological networks of protein-protein interactions (PPINs) are one of the current trends that have been embraced in the field of cooperation between computer scientists and biologists. These special types of biological networks are used to model the physical interactions of a cell's components [1–4]. Recently, several methodologies have been employed to obtain a complete visualization of biological networks. They extract relevant information from these networks to provide a deep sense of the complicated biological function inside the cell, thereby understanding its biological behavior and developments [5].

Proteins often interact with other proteins to perform the same biological function, or they can be connected to certain biological processes [6]. Generally, each protein is made up of a group of genes, whose number varies depending on the protein. Biologists have discovered that functionally similar genes are closely coupled to one another in a protein network, and these interacting proteins may carry out the same process or disease phenotype in the event of disruption [7] and [8]. Gene ontology (GO) is a structured way of organizing genes according to their different characteristics, such as their functions; it assigns a unique code to each gene. In the GO structure, the information about gene function is divided into three groups based on their biological traits, and each gene is derived from a lineage of genes known as ancestors, which are linked in a way that prevents cycles; hence, this structure is realized as a directed acyclic graph (DAG) of expressions and their interactions [9].

Complexity detection in PPI networks is one of the primary difficulties that has taken up a field in the scientific research of biological networks, wherefore there is a competition to develop powerful algorithms that produce accurate information about the structure of these networks. This is because many diseases are the consequence of changes in the interaction patterns of proteins, and the identification of such interactions contributes to many applications in disease diagnosis [4]. Considerable attempts have been carried out to detect complexes (clusters) in PPI networks [10] and [11]. An overview of the main methods has been presented, categorized, and discussed. The prominent ones that show significant performance against other competing methods are evolutionary algorithms (EAs) [12]. In this paper, a new approach to complex cluster detection in a PPI network based on EAs is proposed. This proposed method employs EA with a topology-based fitness function to extract complexes accompanied by heuristically based mutations, where functional annotation, which is obtained from GO to indicate connections between proteins based on their biological information similarity, is the key guide for evolving good solutions. An experimental performance evaluation has been provided based on some of the well-known validation measures to prove the reliability of our proposed method. Also, the results returned by the EA with a heuristically based mutation have been compared with those obtained by the main EA to further assess its performance.

The rest of this paper is organized as follows. Section 2 is devoted to providing an ample background for the related work concerning our research. Section 3 describes in detail the proposed heuristic-based GO operator and its integration with the main EA algorithm. Experimental evaluation, comparisons, and discussion are illustrated in Section 4. Section 5 concludes the paper.

## 2. PPINs complex detection: Related works

In the literature, different clustering algorithms have been tested, analyzed, and designed for the purpose of extracting complexes from PPINs. Some early approaches have been presented

in [12] and distinguished into five specific classes with respect to different topology-based fitness functions. The most promising class with an outstanding result is population-based stochastic search (PS), where the genetic algorithm (GA) is the base algorithm for most of the developed algorithms that have been applied to PPINs. For instance, chaotic GA (CGA) has been proposed in [13] to detect as many protein complexes (near maximal cliques) as possible through adopting chaotic variables to boost the range of the initial population as the primary modification. Then, to prevent the disturbance of good solutions in every generation, CGA added chaotic disturbance based on standard GA operations to put solutions with the highest fitness values directly in the next generation. Binary representation has been selected to describe the solution, where 0 is the score for the edge being survived in the next generation and 1 otherwise. Hence, the length of the chromosome is the number of the network edges. Another approach that has been indicated by Pizzuti in [12] is the improved immune genetic algorithm (IGA) proposed in [14]. IGA detects protein complexes using the concept of creating a population of variable-length antibodies inspired by the artificial immune system. Each antibody is represented by a permutation of integers (the vertices), and additional bits with value 0 to indicate separation between two vertices belonging to the same cluster or with value 1 to indicate separation between two clusters. Therefore, each antibody has a length of $2\lambda-1$, where $\lambda$ denotes the number of vertices in the odd positions and $\lambda - 1$ is the number of separators in the even positions. And because antibodies have variable lengths, the value -1 is used to indicate values that are not contributing to the calculation of the fitness function. Besides these two approaches, Pizzuti has also referred to her past work with Rombo (named GA-PPI) proposed in [15] and [16], wherein they conducted extensive experimental evaluations using different topological-based fitness functions in an attempt to deeply inspect and explore the capability of GA to detect complexes in PPI networks. For the individual representation, they adopted the concept of "graph adjacency," where each individual consists of $\mu$ genes corresponding to the nodes (proteins) of the graph modeling the PPI network. Two values $i$ and $j$ have been used to denote a link between two proteins $i$ and $j$, that is the $i^{th}(j)$ protein belongs to the same cluster $k$ of protein $i$, as illustrated in Figure 1.

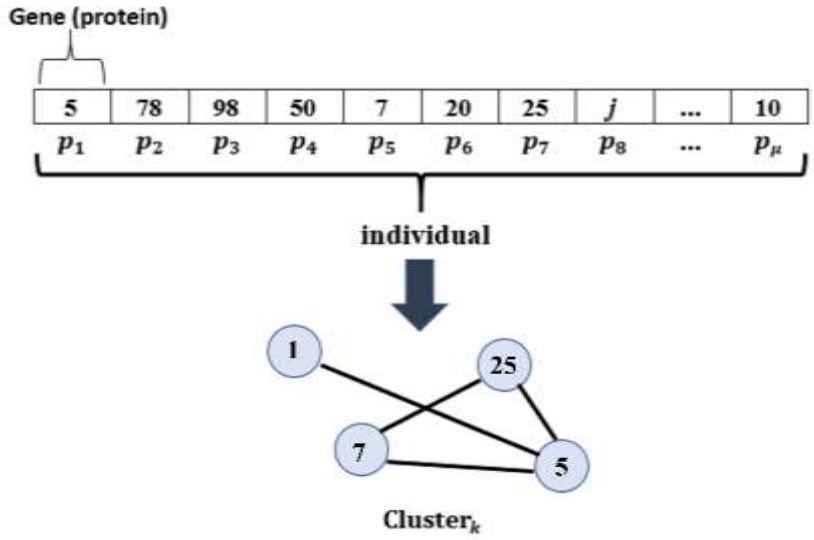



**Figure 1:** Individual representation

Recently, a new method to detect complexes in PPI networks has been introduced in [17] based on the well-known decomposition-based multi-objective evolutionary algorithm (MOEA/D) of [18]. In this method, MOEA/D has been coupled with two topologically conflicting objectives

concerning inter- and intra-community scores to rank the associated clusters of the PPI network accordingly.

According to what has already been reviewed, it is clear that no research has been conducted on the use of EA mutations that are based on the functional annotation of genes to enrich the PPI network clustering analysis and enhance the exploitation ability of the algorithm.

## 3. EA with mutation-based functional annotation

EAs are methods for solving problems of the NP-hard type: problems that cannot be solved in a certain amount of time or that take a lengthy period to solve. EAs' main procedure is comparable to natural selection in that it always retains the strong organs while removing the unfit ones from future generations; as a result, EAs have always been used to solve complex engineering optimization problems.

Basically, EAs divide the search space of an optimization problem $\mathcal{F}(X)$ into a set of solutions that is denoted by Ω and referred to as the "search space size," where $|\Omega| \in \mathbb{N}$ denotes the number of candidate solutions. Customarily, the evolution task is performed on a randomly generated subgroup of Ω known as population of individuals, designated by $\mathbb{P}^{ps}$, with a size of $ps$, where $ps \in \mathbb{N}$, and $\mathbb{P}^{ps} = (P_1, P_2, \dots, P_{ps})$. Each individual $P$ is the genotype representation, the individual representation is depicted in Figure 1, along with its corresponding phenotype $X$, in which these phenotypes are evaluated using a fitness (objective) function that produces values used to explore different parts of the search space. In this work, the optimization problem $Min\ \mathcal{F}(X)$, namely internal density (ID), is used as a metric for measuring a cluster's internal edge density. That is, ID provides a strategy for partitioning based on the density of the internal cluster edges. Equation (1) describes ID as a minimization problem to evaluate a candidate solution $P$, where $P$ is a set $Co = \{c_1, c_2, \dots, c_l\}$ of clusters and $l$ is the number of clusters in $P$.

$$Min\ ID(Co) = \sum_{i=1}^{l} 1 - \frac{v_i}{C_i(C_i - 1)/2} \quad (1)$$

Where $v_i$ is the number of cluster $c_i$'s internal edges, and $C_i$ is the cardinality of cluster $c_i$. The role of the ID model in EA is to evaluate the quality of solutions and produce outputs that assess the robustness of these solutions to solve the complex detection problem. Depending on these outputs, EA behavior is directed towards increasing the chance for good individuals to appear in future generations. Consequently, an implementation of a series of operations known as population transformation is performed on population $\mathbb{P}_g$ to generate a new population $\mathbb{P}_{g+1}$, where $g$ is the generation index. Firstly, tournament selection ($S$) is applied to filter out the good solutions in the current population with respect to the same population size and transfer them into a mating pool. A uniform crossover ($C$) with $P_c$ probability is then applied to maintain solution diversity. Finally, the mutation ($M$) operator is performed with $P_m$ probability to increase the gene's variations and likely to ensure that the population is not going to fall into a local optimum. All candidate solutions in $\mathbb{P}_g$ are subjected to these successive operators. Eventually, population transformation is continued throughout each generation until a maximum number of generations $G$ is reached, and $\mathbb{P}^*$ should contain the near-optimal solution, $P^* = [p_1^*, p_2^*, \dots, p_N^*]$ [19].

### 3.1 EA mutation-based GO

In this work, a protein's functional annotation in GO is used as a heuristic guide for the mutation operator to alter the population genes (proteins); therefore, we named it mutation-based GO. As such, our proposed complex detection technique combines the complete procedure of EA with the protein functional annotation intervened in the mutation operation as a process for improving the individual quality, as stated in Algorithm 1.

Conventionally, every protein is functionally represented by a direct set of genes. According to the Gene Ontology, these genes are categorized into three classes: molecular function (MF), cellular component (CC), and biological process (BP). These three classes of genes inherit their traits from their ancestors, and so forth; as such, these genes are structured into a directed acyclic graph (DAG), as depicted in Figure 2. Then, this protein's characteristic has been exploited to construct a symmetric matrix $M$ of size $N \times N$, in which its entries are values generated from the functional biological data of GO that give the intensity (ratio) of proteins' interactions. The rows and columns of this matrix are labeled with the proteins' index and hold values between 0 and 1. If protein $i$ has a mutual topological interaction with protein $j$, then both entries $(i, j)$ and $(j, i)$ are assigned to the same value. Therefore, our new method is based on the strength of the similarity scale between proteins with respect to the three gene ontology classes, each with its own direct lineage as DAG.

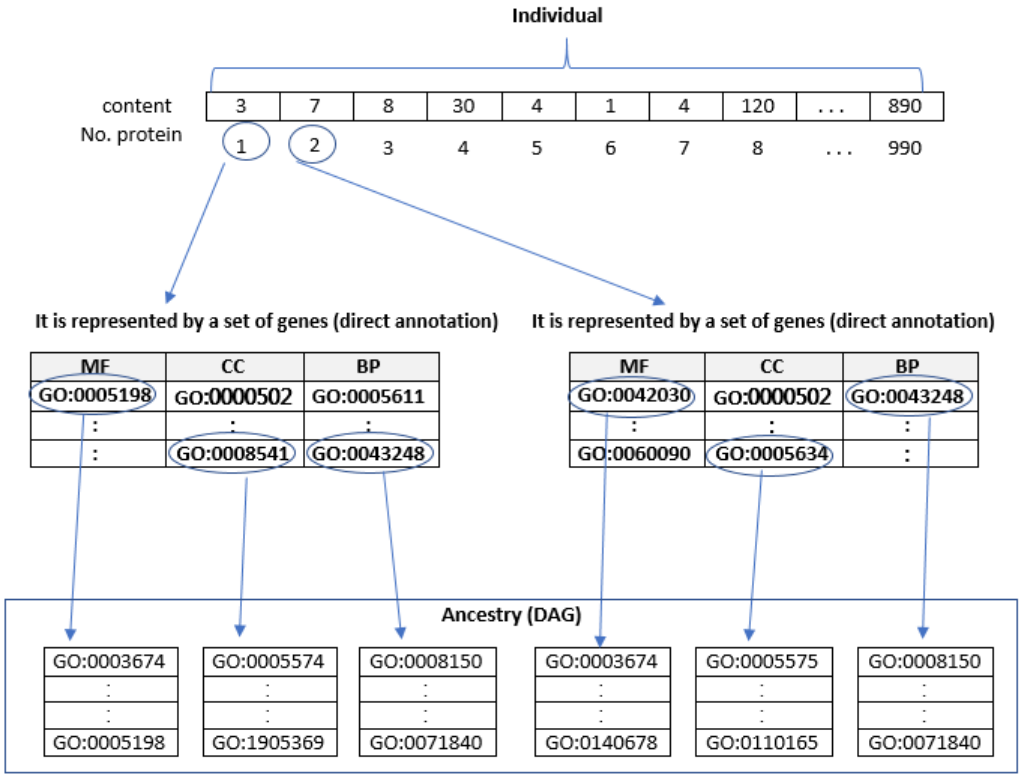



**Figure 2:** Proteins' direct and indirect genes as derived from GO data

To calculate the similarity intensity of interaction between proteins $i$ and $j$, we adopted the well-known Jaccard similarity (JS), as described in Eq. (2).

$$JS(i_i, j) = \frac{|G_i \cap G_j|}{|G_i \cup G_j|} \tag{2}$$

Where $G_i$ and $G_j$ denote the group of genes (direct and indirect) for proteins $i$ and $j$, respectively.

As such, when mutation occurs, with regard to $P_m$, on any of the proteins involved in individual, $P_k(k \in \{1, \dots, ps\}$, this protein $p_{i,k}$ $(i \in \{1, \dots, N\})$ is relocated from its current complex $c_{cur}$ to one of the remaining complexes based on a set of steps that must be performed over the $l$ complexes of $P_k$, as follows:

- For $c_{cur}$, calculate using Eq. (3) the sum $S_{in}$ of the values in $M$ corresponding to the intersection of $p_{i,k}$ with the proteins in $P_c = \{p_{j,k} | (p_{j,k}, p_{i,k})\}$ where $j \in \{1, \dots, N\}$ that are topologically connected to it and both $p_{j,k}$ and $p_{i,k}$ are in $c_{cur}$.

$$S_{in} = \sum_{in=\{inco\}} M(p_{i,k}, p_{in,k}) \quad (3)$$

Where $inco$ are the proteins' indexes that are directly connected with $p_{i,k}$ and located in $c_{cur}$. Likewise, calculate using Eq. (4) the sum $S_{out}$ of the values in $M$ corresponding to the intersection of protein $p_{i,k}$ with the proteins in $P_c$ when $p_{j,k}$ and $p_{i,k}$ are in different complexes.

$$S_{out} = \sum_{o=\{outco\}} M(p_{i,k}, p_{o,k}) \quad (4)$$

Where $outco$ are the proteins' indexes that are directly connected to $p_{i,k}$ and located outside $c_{cur}$. If $S_{in} > S_{out}$ then $p_{i,k}$ is kept in the current complex $c_{cur}$; otherwise, mutation occurs based on the following calculations:

- Equation 5 is used to calculate the difference $diff_{c_{cur}}$ of $S_{in}$ and $S_{out}$ for complex $c_{cur}$.

$$diff_{c_{cur}} = S_{in} - S_{out} \quad (5)$$

Likewise, for the remaining complexes $l - 1$, calculate using Eq. (5) the $diff_{c_m}$ where $c_m \in Co$ and $m \neq cur$.

- For the given set of $l$ distinct elements $diff = \{diff_{c_i}\}_{i=1}^{l}$, the maximum value of this set is denoted as $max_i\ diff_i$ and is equal to the last element of a sorted (i.e., ordered) version of $diff$ (i.e. the output is reordering $\langle diff'_{c_1}, diff'_{c_2}, \cdots, diff'_{c_l} \rangle$ of the given set such that $diff'_{c_1} \leq diff'_{c_2} \leq \cdots \leq diff'_{c_l}$). Hence, the value of $diff'_{c_l}$ represents the likelihood of relocating protein $p_{i,k}$ to complex $l$.
- Then, and before transferring $p_{i,k}$ to a new complex, $p_{i,k}$ must be reconnected with a new protein in $P_c$ with respect to the following conditions:

**1.** The new protein is already located in the new complex.

**2.** In case of finding more than one protein that satisfies the above condition, the protein with the highest value in $M$ at its intersection with $p_{i,k}$ is selected.

**Algorithm 1** EA with mutation-based functional annotation

**Input:** Population size $ps$, individual length $N$, crossover probability $P_c$, mutation probability $P_m$;
**Output:** best individual $P^*$;
**Initialization:** for $i = 1, \dots, ps$ do sample $P^{(i)} \in \{0 \dots 990\}^N$ uniformly at random (u.a.r.) and proteins interactions constraint that must be satisfied; Set $\mathbb{P}_0 = \left\{P_0^{(1)}, P_0^{(2)} \dots P_0^{(ps)}\right\}$;
**Evaluate** $f(\mathbb{P}_0)$ ; **/* using Eq. (1) */**
**Optimization:** for $g = 1, 2 \dots G$ do
  $\mathbb{P}_g \leftarrow S(\mathbb{P}_{g-1})$;
  $\mathbb{P}'_g \leftarrow C\{P_c\}(\mathbb{P}_g)$;
  $\mathbb{P}''_g \leftarrow Mfnot\{P_m\}(\mathbb{P}'_g)$; /***a call for Algorithm 2*/**
  **Evaluate** $f(\mathbb{P}''_g)$;
  $g \leftarrow g + 1$;
**end**
**return** $P^*$

As such, the mutation process occurs and transfers a protein from one complex to another if it finds a complex with proteins that are more similar in functionality, as depicted in Algorithm 2.

**Algorithm 2** Mutation based functional annotation ($Mfnot$) procedure

**Input:** Individual $P$, mutation probability $p_m$;
**Output:** An updated individual $P'$

```
for i = 1,2 ... N do
    if ( rand_i ≤ p_m)
        Compute S_in(c_cur) and S_out(c_cur) of protein i        // use Eqs. 3 and 4
        If S_in(c_cur) < S_out(c_cur)
            Compute diff_old = S_in(c_cur) - S_out(c_cur)        // use Eq. 5
            new_complex = c_cur
            For each c (m ≠ cur) of P
                Compute diff_new = S_in(c_m) - S_out(c_m)
                If diff_new > diff_old                           //finding the max diff value
                    diff_old = diff_new
                    new_complex = c_m
                end
            end
        end
        Add p_i to new_complex
        For each protein j directly connected with p_i and located in new_complex
            Mc = Max_connect (p_j, p_i);
        end
        Connect p_i to the protein with the highest Mc
    end
end
return P'
```

## 4. Results and Discussion

In this work, a mutation-based heuristic has been proposed to consider the biological information in GO to improve the quality of solutions and the overall performance of EA for solving the complex detection problem in PPI networks. As such, we have examined the ID

model on the dataset of the yeast network PPI_YD [20] and [21]. This PPI network has been filtered out, and now it contains 990 proteins with 4687 interactions. Our results have been compared to those of the 81 golden standard complexes. Also, a comparison has been established between the performance of our proposed approach and the performance of the standard EA.

The results depicted in Figure 3 show the performance of the proposed method against the EA for the PPI_YD network when the overlapping score (OS) threshold is varied from 0.1 to 0.8. According to this figure, the detection reliability of the proposed EA is higher than that of the canonical EA in terms of Recall, Precision, Measure, RecN, PrecN, and Fn-measure. The reason is that when the mutation occurs, our proposed method chooses the appropriate complex, where the protein is placed in the complex that has connected proteins that are more similar in functionality to it. While canonical EA chooses the complex randomly, it transfers the protein to a random complex if it has only proteins connected to it. In Figure 3, the result of the recall of our proposed EA is almost equivalent to the canonical EA at the threshold of 0.1, and then it begins to increase slightly with the increase in threshold. This means that the ratio of complexes matching the golden standard complexes in our results is greater than that of the canonical.

In Figure 3, the result of the precision of our proposed algorithm is also equivalent to the Canonical EA from the threshold of 0.1 to 0.35 and then begins to increase slightly with the increase of the threshold, which means that the ratio of complexes matching our work in the golden standard complexes is higher than the Canonical. Figure 3 depicts the comparison results of F-measure between our proposed algorithm and canonical EA, which show that canonical EA is almost equivalent to our method at the top, with a slight difference at the bottom of the plot when the threshold increases. Figure 3 depicts the results of the RecN between the two algorithms and shows the obvious difference in the performances, with our proposed method being the best. Also, for Figure 3, the comparison results of the PrecN and $F_n$-measure between our proposed algorithm and that of the Canonical EA show the improvement in the performance of the EA-based functional annotation. This means that the similarity ratio of the proteins distributed into the candidate complexes of our work significantly matches the golden standard complexes better than the canonical EA. This indicates the amount of correction and quality that was added to the traditional EA at the protein level.

On the other hand, another experiment has been conducted when the mutation probability $P_m$ is increased gradually from 0.2 to 0.5. The result of this experiment is depicted in Figure 4, which shows that the performance of the proposed method has improved because the probability of transferring the mutated protein to a more appropriate complex has increased.

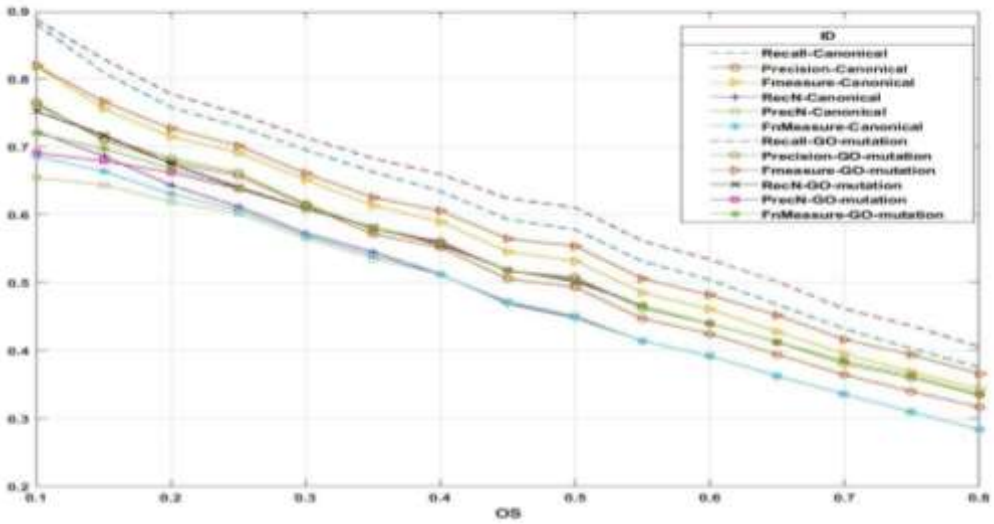



**Figure 3:** Comparison results between the performance of Canonical EA and EA with GO-Based of internal density model in terms of recall, precision, F-measure, RecN, PrecN and $F_n$-measure

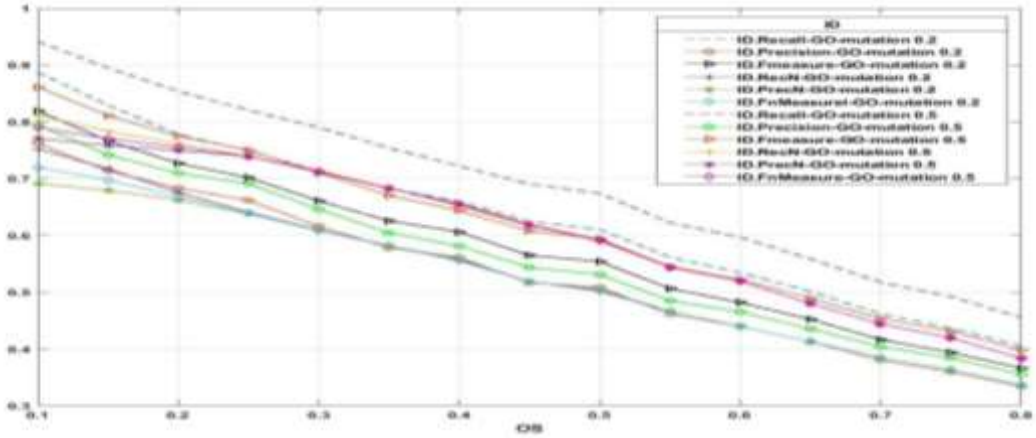


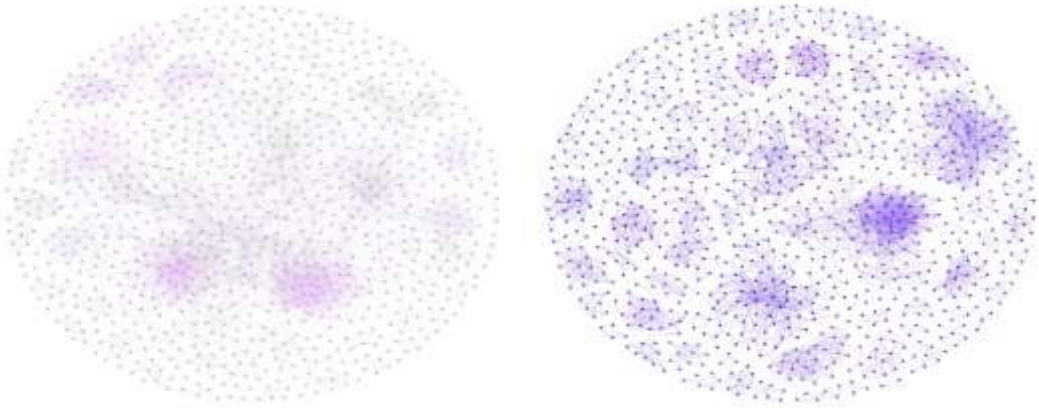

**Figure 4:** Comparison performance between EA with GO-Based with $P_m$ = 0.2 and EA with GO-Based with $P_m$ = 0.5 in internal density model in terms of recall, precision, F-measure, RecN, PrecN and $F_n$-measure)

Table 1 compares the results of our proposed EA-based GO mutation to those of Abduljabbar et al. [22] when $P_m = 0.5$ and the overlapping score (OS) is equal to 0.2. The results confirm that our proposed algorithm outperforms the results of [22] in terms of the three validation measures, Recall, Precision and F-measure as they have been provided in the literature. This means that our algorithm has a higher ability to determine the complexes for PPI_YD networks.

**Table 1:** Performance Comparison between the proposed EA with GO-based mutation and the state-of-the art results with regard to ID model, PPI_YD network in terms of recall, precision, and F measure at OS= 0.2

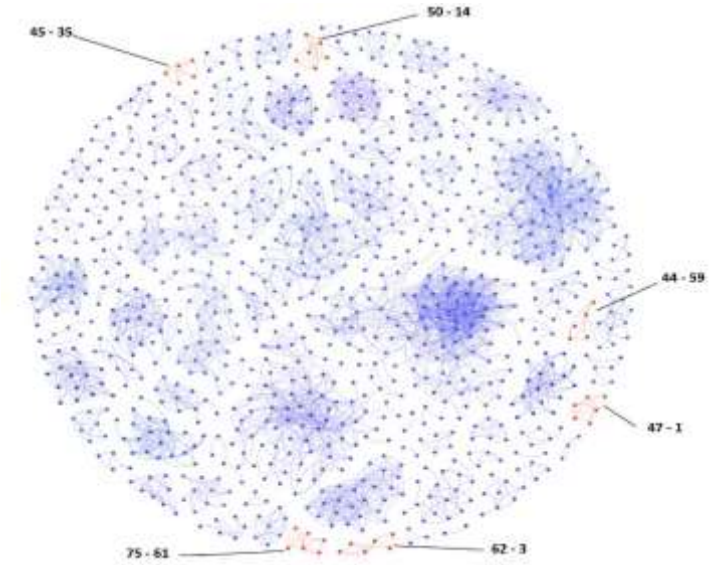
| Term | PGO = 0.5 [22] | Our proposed solution (Pm = 0.5) |
|---|---|---|
| **Recall** | 0.8256 | **0.8538** |
| **Precision** | 0.694 | **0.7096** |
| **F-measure** | 0.7533 | **0.7748** |

* **PGO**: represents the probability of the heuristic biological operator of [22].Figure 5 shows a sample of the original PPI network (in Figure 5 (a)) as a pictorial full network and after (in Figure 5 (b)) applying our proposed technique to this network and the detected complexes.

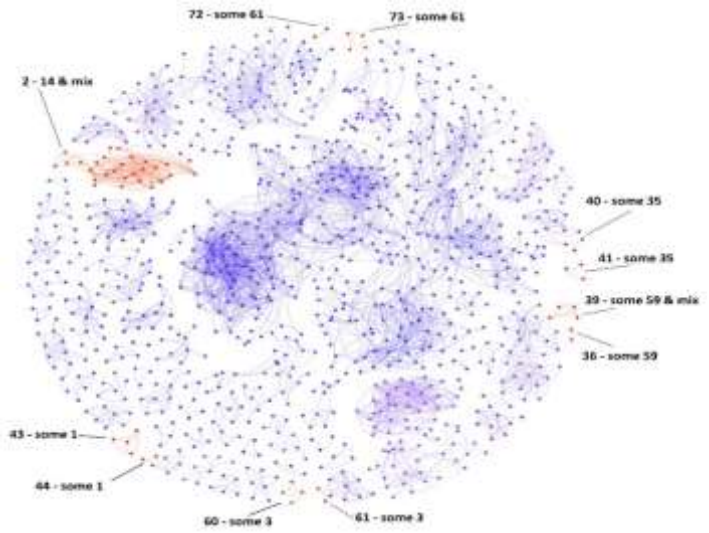


**(a)** **(b)**

**Figure 5:** (a) Original network with 4687 interactions, (b) PPI D1 as complexes with 3317 intra-complex interactions discovered using the GO mutation algorithm for model ID.




**(a)**




**(b)**

**Figure 6: (a)** PPI D1 With some complexes that are exactly identified using the GO-Based algorithm for model ID. (b) PPI D1 With some overlapping complexes using the canonical algorithm for model ID.

When comparing the canonical EA with our proposed method in terms of a solution for detecting the complexes using the ID model, we have noted that the GO-Based algorithm was able to determine exactly the complexes 1, 3, 14, 35, 59, and 61, as shown in Figure 6 (a). With the canonical algorithm, complex 1 was divided into two groups. The first is in complex 43, and the second is in complex 44. Likewise, complex 3 was divided into two groups. The first is in complex 60, and the second is in complex 61. Complex 14 was nested with a large group of proteins within a single complex. Complex 35 was divided into two groups. The first is in complex 40, and the second is in complex 41. Also, complex 59 was divided into two groups. The first is in complex 36, and the second is in complex 39 with the presence of another protein. Finally, complex 61 was also divided into two groups. The first is in complex 72, and the second is in complex 73. All these significant differences are shown in Figure 6 (b). In Figure 6, a comment consisting of two numbers separated by the minus sign is used. The first on the right represents the original complex number, while the second on the left represents the complex number resulting from the application of one of the two algorithms (GO-Based or Canonical).

## 5. Conclusions

In this study, the detection of protein complexes has been investigated through incorporating the functional annotation among proteins extracted from the gene ontology in the evolution process itself and specifically in the mutation operation. The experimental results have proved that the mutation-based GO assisted the EA algorithm in applying the concept of transferring the protein to a complex with the highest functional similarity, which has a positive effect on the algorithm's capability. The development of the algorithm, particularly the mutation part with the addition of gene ontology, has enabled the application of the idea of selecting the best complex to locate the mutated protein at the time of mutation, resulting in higher-quality complexes than those produced by traditional methods when the mutation probability was increased from 0.2 to 0.5 in the ID model. This may be because we are exploring a larger space of solutions and exiting the local optimization area, which in turn improves their performance for this model. This new idea has proven efficient in optimizing better solutions than the classical EA and state-of-the-art algorithms for the PPI_YD network in terms of Precision, Recall and F-measure.

## References

**[1]** N. Atias and R. Sharan, "Comparative analysis of protein networks: hard problems, practical solutions," *Communications of the ACM*, Vol. 55, No. 5, pp. 88–97, 2012.

**[2]** R. De Virgilio and S. E. Rombo, "Approximate matching over biological RDF graphs," in *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012.

**[3]** B. A. Attea, A, D. Abbood, A. A. Hasan, C. Pizzuti, M. Al-Ani, S. Özdemir, and R. D. Al-Dabbagh," A review of heuristics and metaheuristics for community detection in complex networks: Current usage, emerging development and future directions," *Swarm and Evolutionary Computation*, Vol. 63, 100885, 2021.

**[4]** R. Sharan, I. Ulitsky and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, Vol. 3, No. 1, pp. 88, 2007.

**[5]** C. von Mering, R. Krause, B. Snel, M. Cornell, S. G. Oliver, S. Fields and P. Bork, "Comparative assessment of large-scale data sets of protein–protein interactions," *Nature*, Vol. 417, No. 6887, pp. 399–403, 2002.

**[6]** S. Tornw and H.W. Mewes, "Functional modules by relating protein interaction networks and gene expression," *Nucleic Acids Research*, Vol. 31, No. 21, pp. 6283–6289, 2003.

**[7]** A. L. Barabasi and Z.N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, Vol. 5, No. 2, pp. 101–113, 2004.

**[8]** M. Oti, B. Snel, M. A. Huynen and H.G. Brunner, "Predicting disease genes using protein–protein interactions," *J Med Genet*, Vol. 43, No. 8, pp. 691-698, 2006.

**[9]** M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, and G. Sherlock, "Gene ontology: tool for the unification of biology," *Nature genetics*, Vol. 25, No. 1, pp. 25-29, 2000.

**[10]** Q. Z. Abdullah and B. A. Attea, "A Heuristic Strategy for Improving the Performance of Evolutionary Based Complex Detection in Protein-Protein Interaction Networks," *Iraqi Journal of Science*, Vol. 57, No. 4A, pp. 2513-2528, 2016.

**[11]** A. H. Abdulateef, B. A. Attea, A. N. Rashid, "Heuristic Modularity for Complex Identification in Protein-Protein Interaction Networks," *Iraqi Journal of Science*, Vol. 60, No. 8, pp. 1846-1859, 2019.

**[12]** C. Pizzuti, and S.E. Rombo, "Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods," *Bioinformatics*, Vol. 30, No. 10, pp. 1343-1352, 2014.

**[13]** H. Liu and J. Liu, "Clustering protein interaction data through chaotic genetic algorithm," in *Asia-Pacific Conference on Simulated Evolution and Learning* , Springer, Berlin, Heidelberg, 2006.

**[14]** H. Ravaee, A. Masoudi-Nejad, S. Omidi and A. Moeini, "Improved immune genetic algorithm for clustering protein-protein interaction network," in *IEEE International Conference on BioInformatics and BioEngineering*, Philadelphia, PA, USA, 2010.

**[15]** C. Pizzuti and S. Rombo, "Experimental evaluation of topological-based fitness functions to detect complexes in PPI networks," in *Proceedings of the 14th annual conference on genetic and evolutionary computation*, 2012.

**[16]** C. Pizzuti and S.E. Rombo, "Restricted neighborhood search clustering revisited: An evolutionary computation perspective," in *APR International Conference on Pattern Recognition in Bioinformatics*, Springer, Berlin, Heidelberg, 2013.

**[17]** B. A. Attea and Q. Z. Abdullah, "Improving the performance of evolutionary-based complex detection models in protein–protein interaction networks," *Soft Computing*, Vol. 22, No. 11, pp. 3721-3744, 2018.

**[18]** Zhang, Q. and Li, H., "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on evolutionary computation,*, vol. 11, no. 6, pp. 712-731, 2007.

**[19]** C.A.C. Coello, G.B. Lamont and D.A. Van Veldhuizen, *Evolutionary algorithms for solving multi-objective problems, New York: Springer, 2007*, pp. 79-104.

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

**[21]** N. Zaki, J. Berengueres and D. Efimov, "Detection of protein complexes using a protein ranking algorithm," *Proteins: Structure, Function, and Bioinformatics*, Vol. 80, No. 10, pp. 2459-2468, 2012.

**[22]** 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 *International Conference on Soft Computing and Data Mining*, Cham, 2020.