ISSN: 0067-2904

An Evolutionary-Based Mutation with Functional Annotation to Identify Protein Complexes Within PPI Networks

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

Received: 17/10/2022 Accepted: 17/12/2022 Published: 30/10/2023

Abstract

The research deals with an evolutionary-based mutation with functional annotation to identify protein complexes within PPI networks. An important field of research in computational biology is the difficult and fundamental challenge of revealing complexes in protein interaction networks. The complex detection models that have been developed to tackle challenges are mostly dependent on topological properties and rarely use the biological properties of PPI networks. This research aims to push the evolutionary algorithm to its maximum by employing gene ontology (GO) to communicate across proteins based on biological information similarity for direct genes. The outcomes show that the suggested method can be utilized to improve the predictability of the complexes identified. The GO functional annotation of proteins as a heuristic guide is injected into the framework of single-objective evolutionary algorithms (EAs), while the complex detection community score (CS) model works as a fitness function in EAs. In the experiments, we analyzed the performance of our proposed algorithm when applied to the publicly accessible yeast protein networks. The results show a considerable improvement in the detection ability of complexes in the PPI network.

Keywords: Complex detection, evolutionary algorithm, protein complexes, gene ontology, functional annotation.

الطفرة القائمة على الخوارزمية التطورية مع التوصيف الوظيفي لتحديد معقدات البروتين في الشبكة التفاعل البروتينية

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

الخلاصة

يتناول البحث الطفرة القائمة على الخوارزمية التطورية مع التوصيف الوظيفي لتحديد معقدات البروتين في شبكة التفاعل البروتينية (PPI). أحد المجالات المهمة للبحث في علم الأحياء الحسابي هو التحدي الصعب المتمثل في الكشف عن المجمعات في شبكات تفاعل البروتين. تم تطوير نماذج الكشف المعقدة للتعامل مع التحديات ، لكنها تعتمد في الغالب على الخصائص الطوبولوجية ، ونادرًا ما تستعمل الخصائص البيولوجية لشبكات تفاعل البروتين. يهدف هذا البحث إلى دفع الخوارزمية التطورية إلى أقصى حد من خلال استعمال علم

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

الوجود الجيني (GO) للتواصل عبر البروتينات بناءً على تشابه المعلومات البيولوجية للجينات المباشرة. تظهر النتائج أنه يمكن استعمال الطريقة المقترحة لتحسين إمكانية التنبؤ بالمجمعات المحددة. يتم إدخال التعليق التوضيحي الوظيفي للبروتينات GO كدليل إرشادي في إطار خوارزميات تطورية (EAs) ذات الهدف الواحد ، بينما يعمل نموذج درجة مجتمع الاكتشاف المعقد (CS) كوظيفة لياقة في .EAs في التجارب ، نقوم بتحليل أداء الخوارزمية المقترحة عند تطبيقها على شبكات بروتين الخميرة المتاحة. تظهر النتائج تحسينًا كبيرًا في قدرة الكشف عن المجمعات في شبكة (PPI).

1. Introduction

One of the subjects that has received a lot of attention recently is biological network analysis, which has also provided a lot of valuable and detailed data that defines specific interactions among cells' many components. However, this data needs to be formatted correctly to be useful. In the field of collaboration between computer scientists and biologists, biological networks of protein-protein interactions (PPIs) are one of the current trends that have been accepted. These particular forms of biological networks are utilized for modeling and simplification [1], [2], [3], and [4]. Recently, several approaches have been used to entirely visualize biological networks. They collect pertinent data from these networks to give a thorough understanding of the intricate biological processes that take place inside the cell, helping us to comprehend its behavior and development on a biological level [5].

One of the main challenges facing scientific research on biological networks is complex detection in PPI networks, and as a result, there is competition to develop effective algorithms that accurately predict the structure of these networks. An overview of the key techniques has been provided, organized, and debated. The most notable ones that show remarkable performance against other competing methods are population-based random search algorithms [6]. Complex discovery in PPI networks has been the subject of extensive research [7] and [8]. Proteins can collaborate with one another to carry out the same biological function, or they can be involved in specific biological functions [9].

In general, a set of genes, whose number varies depending on the protein, make up each protein. In a protein network, functionally similar genes are tightly associated with one another, and in the case of a disruption, these interacting proteins may cause the same process or disease phenotype [10] and [11]. Gene ontology (GO) is a structured method for classifying genes based on their various properties, such as their functions; each gene is given a specific code in this system. According to their biological characteristics, the information regarding gene function is separated into three groups in the GO structure [12].

A new method for complex discovery in PPI networks based on evolutionary algorithms (EAs) is put forth in this paper. This suggested approach uses EA with a topology-based fitness function to extract complexes along with heuristically based mutations, where functional annotation acquired from GO to indicate connections between proteins based on their biological information similarity is the key guide for evolving useful solutions. An experimental evaluation has been offered based on some well-known validation metrics. Additionally, comparisons between the main EA's results and those returned by the EA with heuristically based mutations have been made.

2. Previous work to determine the complexes in PPI networks

For the objective of extracting complexes from PPINs, various clustering techniques have been tested, analyzed, and developed in the literature (such as population-based stochastic search (PS), cost-based local search (CL), etc.). In the typical form of PPI networks, edges are

used to describe the interactions between proteins, i.e., connecting the appropriate nodes, whereas nodes are used to represent the proteins. According to this illustration, protein complexes resemble dense protein clusters in which proteins interact with each other more than the rest of the network. Finding dense regions in PPINs has been thought to be possible through the detection of protein clusters. In [6], certain early methods were discussed and classified into five distinct classes according to various topology-based fitness functions. Population-based stochastic search (PS) is the most promising class with an exceptional result, in which the genetic algorithm (GA) is the base method for most developed algorithms that have been applied to PPINs.

In another study, [13] proposed the genetic algorithm GA-PPI-Net and then compared it with three clustering methods for population discovery (MCL, RNSC, and Cluster). In computational tests, the GA-PPI-Net genetic algorithm achieved excellent results for discovering complexes from the PPI network. Recently, [14] has also proposed a new heuristic approach named "locally assisted heuristic," considering the topology, which is premised on actual protein interaction and the concept of protein pairing to characterize the search space. The main idea is to classify each pair of proteins with respect to their topological similarities into two classes: intra-delineation pairs and inter-delineation pairs. Two proteins with high sequence similarities are likely to form an intra-delineation structure; otherwise, they can form an inter-delineation pair. The influence of this approach on improving the ability of multiple existing optimization models in single and multi-objective EAs has thoroughly been studied to detect complexes in PPI networks.

3. Individual representation and phenotype

The individual is represented in the form of an array with a length equal to the number of proteins in the network. The protein number is an index of the array, and the content represents the protein number to which this protein is connected. Figure 1 shows that the first place of the array (individual) includes the number 3, indicating that the first protein is linked to the third, and that the second position contains the number 6, indicating that the second protein is linked to the sixth. We can save information about a protein's location in any existing cluster (phenotype). This information is represented as an array with a length equal to the number of proteins in the network. Figure 2 shows that the third place of the array includes the number 1, indicating that the third protein is inside the first cluster, and that the fourth place of the array includes the number 3, indicating that the fourth protein is inside the third cluster.

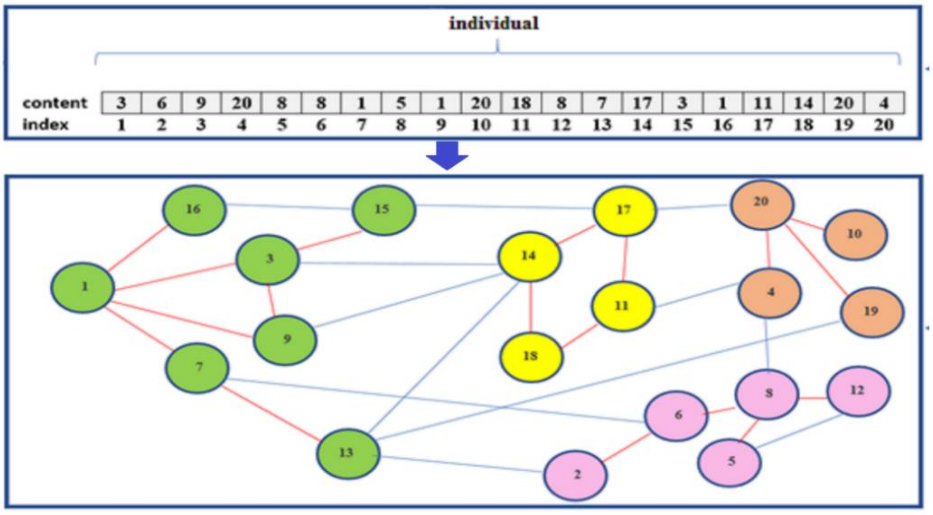


Figure 1: individual representation

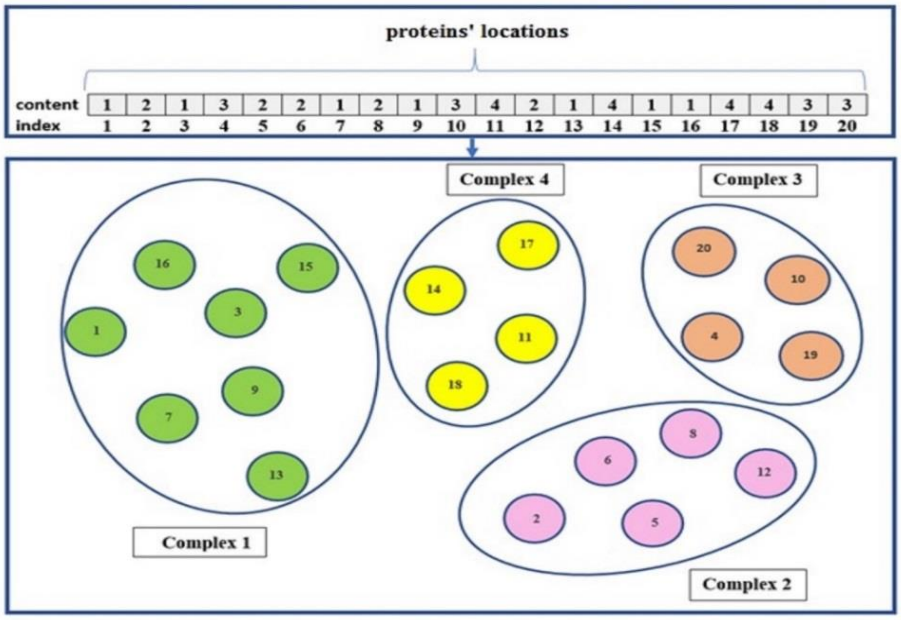


Figure 2: Proteins' locations (phenotype)

4. EA with functional annotation based on mutations

EAs are strategies for dealing with NP-hard issues, which are those that cannot be resolved quickly or require a long time to resolve. EAs have historically been utilized to resolve challenging engineering optimization issues since their primary technique is analogous to natural selection in that it constantly keeps the strong organs while eliminating the unfit ones from subsequent generations. EAs basically partition the search space of an optimization problem $\mathcal{F}(X)$ into a collection of solutions, which is indicated Ω and denoted as the search space size, while |Ω|∈ ℕ denotes the number of candidate solutions. The evolution job is often applied to a population of individuals, or $\mathbb{P}^{pz}$, with a size of pz, where $pz \in \mathbb{N}$ and $\mathbb{P}^{pz} = (P_1, P_2, \ldots, P_{pz})$, that is randomly generated. Each individual P is the genotype representation of its matching phenotype X Figure 2; the individual representation is shown in Figure 1, and these phenotypes are assessed using a fitness (objective) function that yields values used to explore various regions of the search space.

In this study, the community score (CS), an optimization problem, is utilized as a metric to assess each cluster's edge density in relation to its size. In other words, CS offers a method of partitioning that considers the edge density of each cluster in relation to its size. CS is described in Equation (1) as a maximizing problem, $max\ \mathcal{F}(X)$ to assess a candidate solution P, where P is a collection of c_1, c_2, …, c_l clusters, and l is the number of clusters in P.

$$max\ \ CS(X) = \sum_{i=1}^{l} \left(\frac{2*v_i}{c_i}\right)^2 \quad (1)$$

Where C_i is the cluster's cardinality and v_i is the number of internal edges for the cluster c_i. The purpose of the CS model in EA is to measure the effectiveness of solutions in order to get results that gauge how robust these solutions are at resolving challenging detection problems. Based on these outcomes, EA behavior is geared towards improving the likelihood that good individuals will be present in upcoming generations. As a result, population $\mathbb{P}_g$ is subjected to a set of procedures collectively referred to as population transformation to produce a new population $\mathbb{P}_{g+1}$, where g is the generation index. The existing population is first filtered using tournament selection (S), which transfers the good solutions into a mating pool while

maintaining the same population size. Then, uniform crossover (C) with P_c probability is used to maintain the diversity of the solutions. Lastly, to boost gene variation and to make sure that the population won't settle into a local optimum, the mutation (M) operator is applied with P_m probability. All nominee solutions in $₱_g$ are passed through these consecutive operators. Lastly, populations continue to alter throughout each generation up to the maximum number of generations, and $₱^*$ should have the near-optimal solution $P^* = [p_1^*, p_2^*, \dots, p_N^*]$ [15].

5. EA mutation-based GO

In this study, we modified the population genes (proteins) using the GO functional annotation of proteins as a heuristic guide. So, we gave it the name "mutation-based GO." In order to improve the individual's quality, as indicated in Algorithm 1, our suggested complex detection technique combines the entire EA procedure with the protein functional annotation interfered with in the mutation operation. Traditionally, a direct set of genes is used to functionally represent each protein. As shown in Figure 3, these genes are divided into three groups by the gene ontology: molecular function (MF), cellular component (CC), and biological process (BP). Then, using this protein's characteristics, a symmetric matrix $\mathbb{M}$ of dimension $N \times N$ has been created, with the entries of the matrix being values derived from the functional biological data of GO, which provides the intensity (ratio) of proteins' interactions.

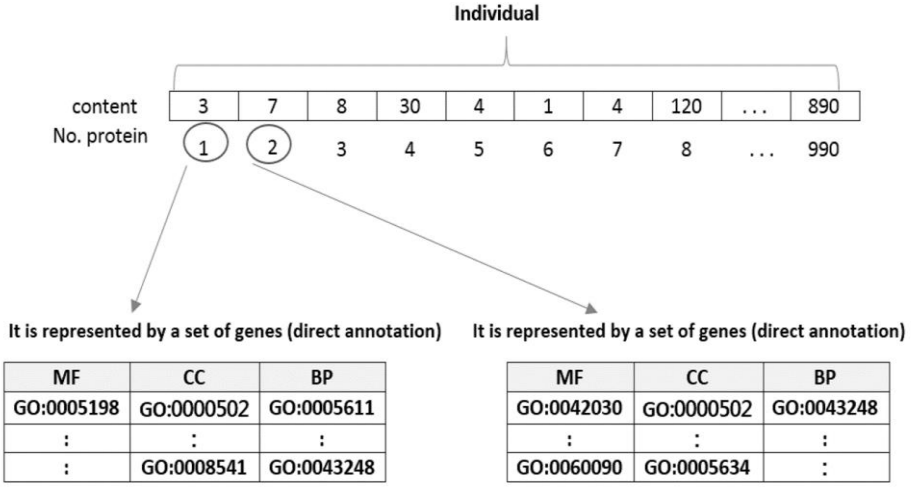


Figure 3: Proteins' direct genes GO data

The proteins' index is used to label the rows and columns of the matrix $\mathbb{M}$, which contains values between 0 and 1. Both entries (i, j) and (j, i) are given the same value if the proteins i and j interact topologically with one another. The power of the similarity scale among proteins with regard to the three gene ontology classes is therefore the foundation of our novel methodology. We used the well-known Jaccard similarity (JS), as shown in Eq. (2)., to determine the degree of similarity between proteins i and j.

$$JS(i_i, j) = \frac{|GD_i \cap GD_j|}{|GD_i \cup GD_j|} \quad (2)$$

Where GD_i and GD_j represent the set of genes (direct) of proteins i and j, respectively. As a result, when any of the proteins involved in an individual $P_k (k \in \{1, \dots, ps\}$ undergo a mutation, algorithm 1 determines whether to move the protein $p_{i,k}$ $(i \in \{1, \dots, N\})$ from its current complex to one of the other complexes or keep it in the current complex. Eq. (3) is the sum of the values in $\mathbb{M}$ conforming to the intersection of the protein $p_{i,k}$ with the proteins that are topologically linked to it in the candidate complex.

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

Where $inco$ are the proteins' indexes, which are situated in the candidate complex and have a direct connection to the protein $p_{i,k}$. Calculate the sum of the values in $\mathbb{M}$ conforming to the intersection of the protein $p_{i,k}$ with the proteins that are topologically linked to it in the candidate complex using Eq. (4).

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

Where $outco$ are the proteins' indexes, which are directly related to proteins and are situated outside of the candidate complex.

Algorithm 1 :$mutation$

Input: P, $\boldsymbol{p_m}$, $\mathbb{M}$
Output: **new individual** P^*

```
for i = 1 to length individual P
    if ( rand ≤ p_m)
        Compute Sm_in & Sm_out for current complex that has pi; // used Eq. (3) & Eq. (4)
        If Sm_in < Sm_out
            compute diffr_old = Sm_in - Sm_out ;
            new complex = current complex;
            For each other complexes c in P
                Compute diffr_new = Sm_in - Sm_out;
                If diffr_new > diffr_old
                    diffr_old = diffr_new;
                    new complex = complex c;
                end
            end
        end
        new_complex(p_i) = new complex;
        for each protein directly connected with pi and placed in new complex
            Protein Number that connected with p_i = Protein Number that has Max M(pi,protein number connected with pi );
        end
    end
end
return P*;
```

6. Results and Discussion

In order to enhance the quality of solutions and the overall efficiency of EA for resolving the complex detection problem in PPI networks, a mutation-based heuristic guide that considers the biological information in GO has been proposed in this study. As a result, we focused on the CS model using the yeast network PPI_YD dataset [16] and [17]. This PPI network now has 990 proteins and 4687 interactions. Our findings were compared with those of the 81 complexes that serve as the gold standard. Additionally, a comparison has been made between the effectiveness of the standard EA and the effectiveness of our suggested approach.

Figure-4 findings (a), (b), (c), (d), (e), and (f) illustrate how the suggested technique performs when compared to the EA for the PPI_YD network when the overlapping score (OS)

threshold is changed from 0.1 to 0.8. This figure (Figure-4) makes it obvious that the suggested EA's detection reliability is greater than that of the traditional EA in terms of recall, precision, F-measure, RecN, PrecN, and Fn-measure. Because of our strategy of selecting the proper complex at the time of the mutation, the protein is always positioned in a complex with other proteins that are connected and have functions that are more similar to its own. While canonical EA selects the complex at random, if the protein is linked to it, it is transferred to a random complex. In the Figures (a, b, and c), the results of the recall, precision, and F-measure are in close proximity to the canonical at the threshold of 0.1 and then begin to increase greatly with the increase of the threshold, which means that the ratio of complexes matching between the golden standard complexes and our results is greater than the canonical. In the Figures (d, e, and f), the results of the RecN, PreN, and Fn-measure in our algorithm are better than the Canonical, which means that the ratio of proteins matched between the golden standard complexes and our results is greater than the Canonical.

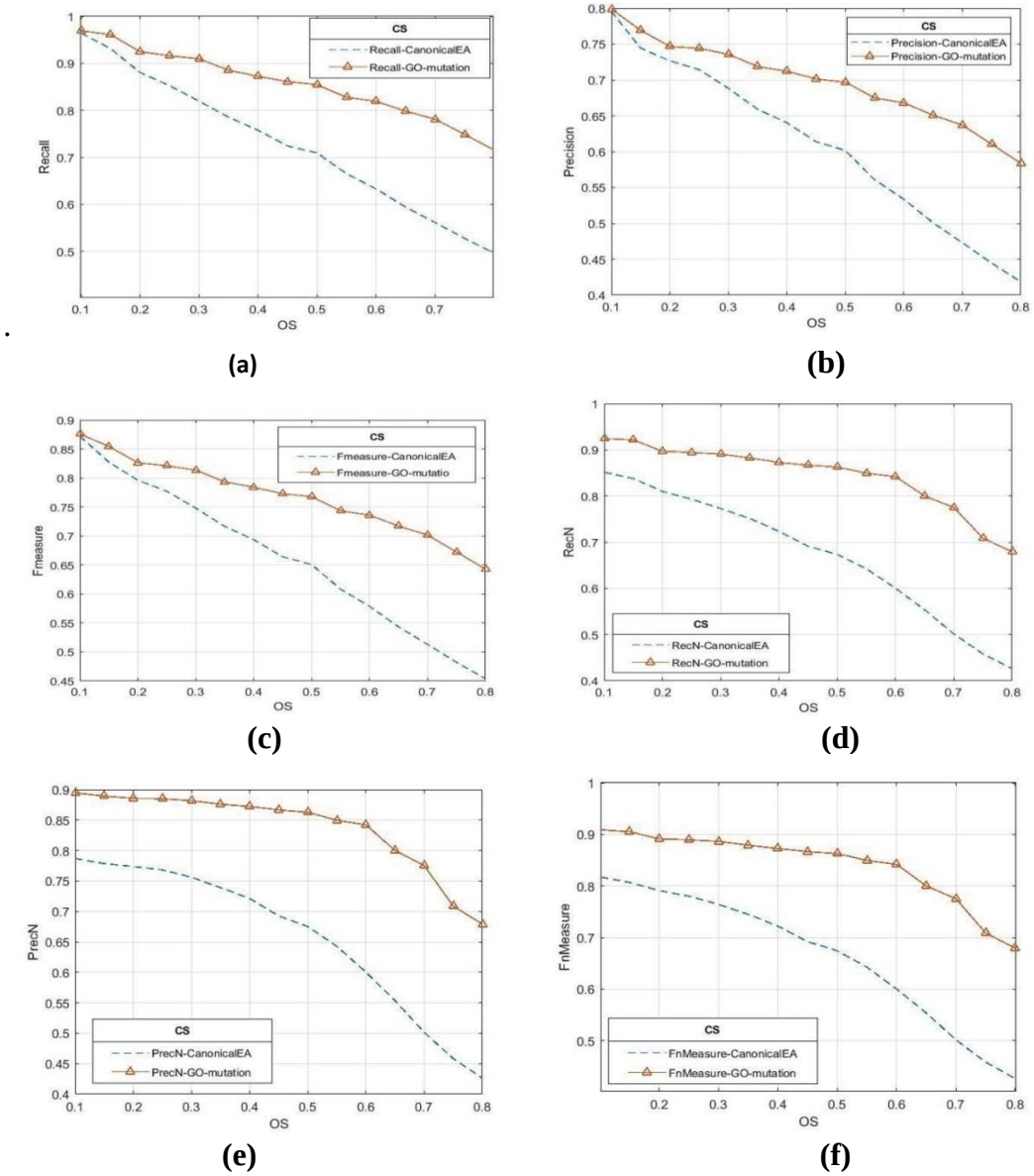
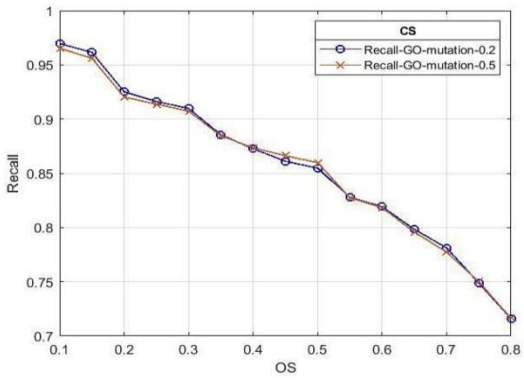


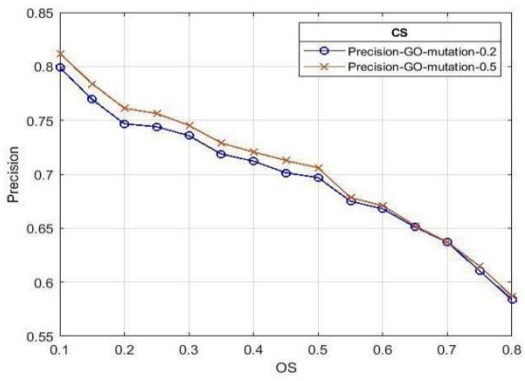
Figure 4: Comparison between the performance of Canonical and GO-mutation in CS model in terms of (a) recall, (b) precision, (c) F-measure, (d) RecN, (e) PrecN and (f) Fn-measure)

In Figure-5 (a, d, e, and f), the results of Recall, RecN, PreN, and Fn-measure in our algorithm are very close when Pm increases from 0.2 to 0.5. In Figures 2(b) and (c), the results

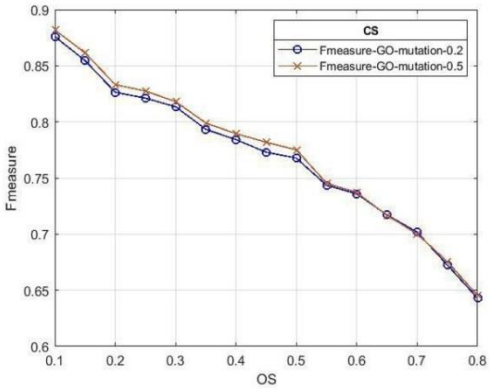
of precision and F-measure in our algorithm are better when Pm increases from 0.2 to 0.5 than the threshold of 0.1 to 0.55 and then begin to very nearly match with the increase of the threshold.



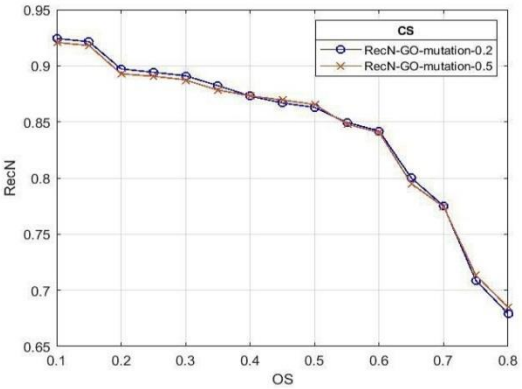
(a)



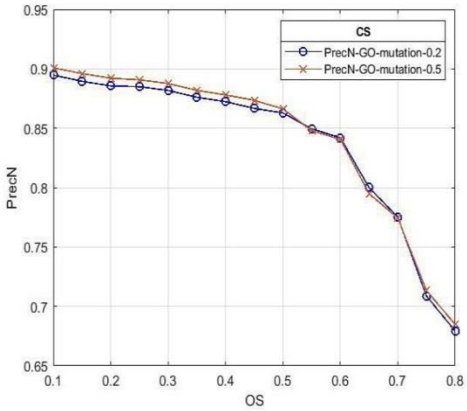
(b)



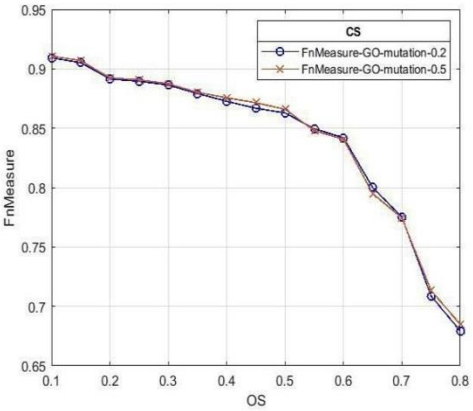
(c)



(d)



(e)



(f)

Figure 5: Comparison between the performance of GO- mutation with Pm = 0.2 and GO-Based with Pm = 0.5 in CS model in terms of (a) recall, (b) precision, (c) F-measure, (d) RecN, (e) PrecN and (f) Fn-measure)

In Table 1, when Pm = 0.5 and the overlapping score (OS) is equal to 0.2, the research results are compared with the findings of Abduljabbar et al. [18] using our proposed algorithm. The findings show that our suggested algorithm works better than that of [18] in terms of recall,

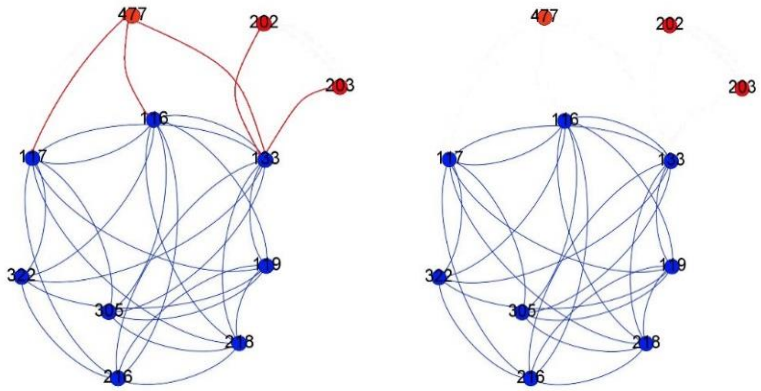
precision, and F-measure, which means that our system is better at figuring out the complexities of PPI-YD networks.

Table 1: The performance comparison for the community score model on PPI_YD (in terms of recall, precision, and F measure) at OS = 0.2

Term	PGO = 0.5 [18]	Our proposed solution (Pm = 0.5)
Recall	0.9026	**0.92.5**
Precision	0.746	**0.7612**
F-measure	0.8168	**0.8332**

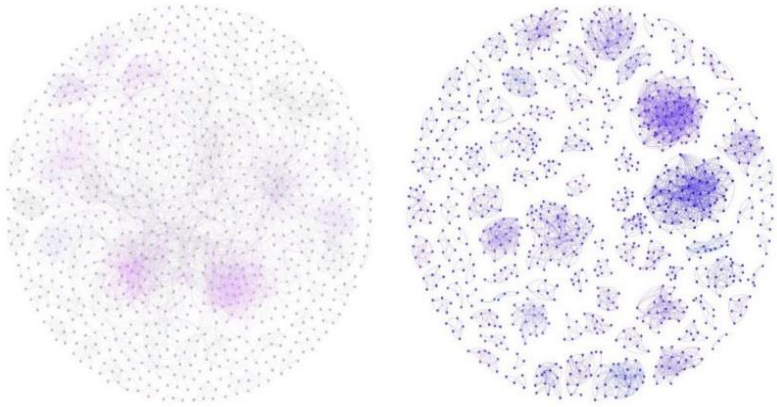
***PGO**: the probability of the heuristic biological operator.

To explain the main idea of the intra- and inter-delineation pairs, in the following example shown in Figure 6, the complex has eight proteins (blue circles) and three other proteins (red circles) that are located outside of the complex. (a) A correct golden complex structure with 25 intraconnections (blue links) among 8 proteins and 5 interconnections (red links), with 3 external proteins that belong to other complexes. (b) A correct complex structure with only intra-connections Protein names are shown as the protein's number in the complex structure. In this research, Figure 7 (a) represents the original network, which contained 4687 interactions (intra-delineation and inter-delineation pairs), whereas Figure 7(b) represents the network After using the GO mutation algorithm for model CS on PPI D1, it was found to have 3725 intra-complex interactions.



(a) **(b)**

Figure 6: Intra-delineation and inter-delineation protein pairs and their role in defining the intra and inter structure of a protein complex. (a) A correct golden complex structure. (b) A correct complex structure.

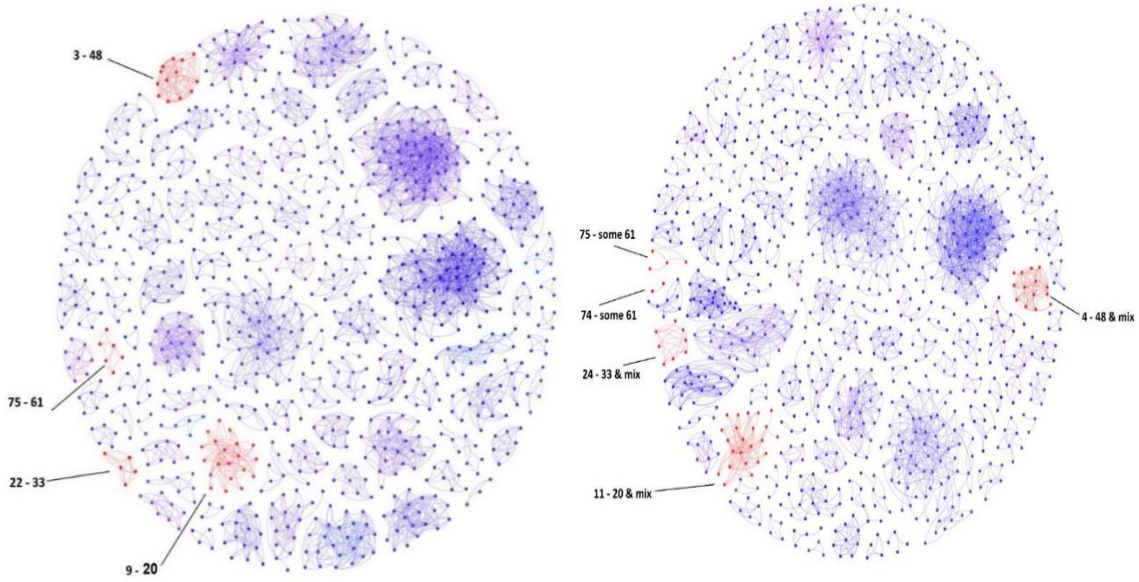


(a) **(b)**

Figure 7: (a) the original network and contained 4687 interactions, (b) After using the GO mutation algorithm for model CS on PPI D1, it was found to have 3725 intra-complex interactions.

For example, when comparing the canonical technique to our algorithm for detecting the complexes and for the model (CS), we note that in the GO-mutation algorithm, it was able to determine exactly Figure 8(a), the complexes No. 20, No. 33, No. 48, and No. 61. In the canonical algorithm shown in Figure 8(b), Complex No. 20 is mixed with proteins in cluster 11, Complex No. 33 is mixed with proteins in cluster 22, Complex No. 48 is mixed with proteins in cluster 3, and Complex No. 61 has been divided into two groups, the first in cluster 74 and the second in cluster 75.

Note: In Figure 8 (a) and (b), we'll use a comment with two integers separated by a negative sign. The number on the right is the original complex, while the cluster number for the GO mutation or canonical is shown on the left.



(a) **(b)**

Figure 8: (a) PPI D1 and some complexes that are precisely detected using the GO-Based algorithm for model CS, **(b)** PPI D1 and some overlapping or spread complexes using the canonical algorithm for model CS.

9. Conclusions

Complex discovery in protein interaction networks is a crucial research area in computational biology because it enables us to better understand the typical and abnormal molecular activities that take place in the complexes. We use topological information combined with biological function to solve the complex detection problem. Based on similarities in bio-functional informatics, the main contribution of this study is the introduction of biological data about proteins into the evolutionary algorithm that used the CS model. As a result, the functional annotation of proteins drawn from the gene ontology was used in this study to investigate how to discover protein complexes by adding them into the mutation operation. EA with GO-based mutations is the suggested technique for complex detection. Our algorithm has proven efficient in optimizing better solutions than the traditional EA for the PPI_YD network in the metrics of recall, precision, F-measure, RecN, PrecN, and Fn-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] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] M.B. M'barek, S.B. Hmida, A. Borgi and M. Rukoz, "GA-PPI-Net Approach vs Analytical Approaches for Community Detection in PPI Networks," *Procedia Computer Science*, vol. 192, pp. 903-912, 2021.

[14] A.H. Abdulateef, A.A. Bara'a, A.N. Rashid, and M. Al-Ani, "A new evolutionary algorithm with locally assisted heuristic for complex detection in protein interaction networks," *Applied Soft Computing*, vol. 73, pp. 1004-1025, 2018.

[15] 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.

[16] 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.

[17] 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.

[18] 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.