An Evolutionary Algorithm for Complex Detection in PPI Networks

Sree Vaishnavi Govindraj, Vignesh Sankara Narayanan and Vyshakh Gireesh
Otto-von-Guericke University, sree.govindaraj, vignesh.sankara, vyshakh.gireesh@st.ovgu.de

I. INTRODUCTION

The identification of protein complexes within protein-protein interaction (PPI) networks is crucial for understanding cellular functions and malfunctions, which can lead to the development of more effective treatments for various diseases. This task remains a significant challenge due to the complex nature of PPI networks, characterized by large datasets and potential spurious interactions. Traditional high-throughput experimental methods, such as two-hybrid screening, phage display, and mass spectrometry, have made significant strides in mapping these interactions, producing vast amounts of data. However, these methods are often expensive, time-consuming, and prone to errors, necessitating the development of robust computational approaches to complement and enhance experimental techniques.

Computational methods for detecting functional modules in PPI networks have evolved, with clustering techniques playing a pivotal role. Clustering aims to capture the essential characteristics, topology, and functions of networking systems. Despite the inherent difficulty of the graph clustering problem, which is NP-hard, recent advances have shown that evolutionary algorithms (EAs) are particularly effective for this purpose. EAs offer a promising alternative to traditional methods due to their ability to handle the complex and often undefined nature of clusters within PPI networks.

Most state-of-the-art methods for partitioning PPI networks into protein complexes rely on either very general graphical properties or highly specific biological semantics, or a combination of both. These methods, while effective to some extent, often overlook the intrinsic topological properties of protein complexes. Additionally, many existing EAs treat the complex detection problem as a single-objective optimization problem. This approach, although beneficial, does not fully exploit the multi-faceted nature of PPI networks and the inherent complexities of protein interactions.

In this research, we propose a novel approach that combines a ranking algorithm with a single-objective evolutionary algorithm to effectively detect protein complexes in PPI networks. Our method leverages the inherent topological properties of protein complexes, which are often underutilized in existing approaches. By integrating a ranking algorithm, we aim to improve the initial population of the evolutionary algorithm, thereby enhancing the quality and accuracy of the detected complexes.

To demonstrate the effectiveness of our proposed method, we conducted experiments on well-known PPI networks, including those from Saccharomyces cerevisiae (yeast). We also used reference sets of benchmark complexes and generated new random networks to evaluate the impact of perturbing protein interactions. Our results indicate that the proposed algorithm achieves a higher predictive level of matched protein complexes compared to several state-of-the-art methods. Furthermore, the integration of the ranking algorithm significantly improves the performance of the evolutionary algorithm, leading to more accurate and reliable detection of protein complexes.

This paper is organized as follows. The Introduction provides an overview of the significance and challenges associated with detecting protein complexes in PPI networks. The background section of our paper outlines the importance of protein complex identification in PPI networks, reviews advances in proteomics technologies, discusses various traditional and advanced computational methods for detecting protein complexes, and introduces our novel approach combining a ranking algorithm with a single-objective evolutionary algorithm to address existing challenges in the field. The Literature Review examines existing methods and state-of-the-art techniques for protein complex detection, highlighting their strengths and limitations. The Proposed Methodology showcases the concepts of evolutionary algorithms and ranking algorithm, as well as showcases the working mechanism of our proposed methodology. Finally, the References section compiles all the sources cited throughout the paper, providing a comprehensive bibliography for further reading and validation of the discussed concepts.

II. BACKGROUND

The identification of protein complexes in protein-protein interaction (PPI) networks is pivotal for advancing our understanding of cellular mechanisms and disease pathogenesis. Advances in proteomics technologies, such as two-hybrid screening, phage display, and mass

spectrometry, have generated extensive datasets of protein interactions, particularly in model organisms like Saccharomyces cerevisiae. These datasets, which include thousands of protein interactions, necessitate sophisticated bioinformatics tools for storage, management, visualization, and analysis to facilitate knowledge discovery and functional insights.

A variety of computational methods have been developed to detect protein complexes within these vast PPI networks. Traditional approaches include Markov Clustering (MCL), which uses random walks to identify clusters in protein interaction networks, and Molecular Complex Detection (MCODE), which identifies complexes as dense regions grown from highly-weighted vertices. Other notable methods include Clustering based on Maximal Cliques (CMC), Affinity Propagation (AP), ClusterONE, which uses overlapping neighborhood expansion, and the restricted neighborhood search clustering (RNSC) algorithm. Each of these methods has demonstrated varying degrees of success in detecting protein complexes, yet they often assume that protein complexes correspond to dense subgraphs in the interaction network. This assumption can limit their ability to detect complexes with few members or sparse interactions.

More recent methods have introduced advanced techniques to improve detection accuracy. For example, the Affinity Propagation algorithm has shown promise by identifying clusters based on the similarity of data points. ClusterONE and other methods like the repeated random walks (RRW) algorithm and CFinder, which is based on the clique percolation method, have improved performance in detecting overlapping complexes. Despite these advancements, many methods still face challenges in accurately identifying overlapping complexes—a significant limitation given that proteins often participate in multiple functional modules.

To address these challenges, we propose a novel approach that combines a ranking algorithm with a single-objective evolutionary algorithm (EA) to detect protein complexes in PPI networks more effectively. Evolutionary algorithms, inspired by the principles of natural evolution, are well-suited for optimization problems in computational biology. They operate on a population of potential solutions, applying selection, recombination, and mutation operators to evolve the population towards higher fitness, defined by an appropriate fitness function.

Understanding the topological properties of PPI networks and the principles of evolutionary algorithms forms the core of our approach. PPI networks are characterized by clusters of densely interconnected proteins, with sparse connections between clusters. This modular structure is often quantified using modularity-based methods, where modularity serves as a single-objective function reflecting the quality of the network's partitioning into complexes.

Our method leverages these topological insights and the adaptive search capabilities of evolutionary algorithms to achieve more accurate and efficient detection of protein complexes. By overcoming the limitations of previous methods, such as the inability to detect overlapping complexes, our approach holds the potential to provide deeper insights into the organization and dynamics of cellular networks.

In conclusion, the integration of a ranking algorithm with a single-objective evolutionary algorithm represents a significant advancement in the field of protein complex detection in PPI networks. This approach not only improves the initial solution set for the evolutionary algorithm but also enhances the detection accuracy of overlapping complexes, providing a more comprehensive understanding of cellular functions and interactions.

III. LITERATURE REVIEW

Proteins are a vital component of living beings, and they play a key role in managerial and executing most biological processes. Many clustering algorithms have been designed to detect meaningful groups of proteins from PPI networks. There are three main approaches to cluster graphs aiming protein complex detection. The first one is to find subgraphs with specified connections which are called network motifs and is described as a complex or a part of it. Clique is one of the subgraphs detected using this approach. Due to the time complexity of this approach, its application is limited nowadays. The second one is graph-growing in which a cluster grows and completes around a vertex as a seed using greedy search algorithms. The third one includes several variants. The algorithms try to minimize or maximize measures of a specified cluster such as connection density, edge cut, or a metric distance between nodes. Generally, the algorithms aim to optimize an objective function for the entire graph.

Bader and Hogue proposed the Molecular Complex Detection (MCODE) algorithm to detect densely connected regions as molecular complexes in large PPI networks [2]. MCODE consists of two main steps: network weighting and complex detection. In network weighting, all vertices are assigned weights based on their local network density. This is followed by outward traversal from a locally dense seed protein to isolate the dense regions. King et al. proposed the restricted neighbourhood search clustering (RNSC) algorithm that partitions the nodes of the network into clusters, based on a low-cost clustering function called homogeneity value [3]. It starts with initial random clusters and then randomly moves a protein from one cluster to another, satisfying a minimum deleterious of the cost function. Altaf-Ul-Amin et al. suggested a deterministic

algorithm to select initial clusters as the seed highest weighted node or highest degree node [4]. The cluster then gradually grows by adding neighbour nodes to the cluster one by one depending on neighbour priority. A cluster then continues to expand until cluster density and/or cluster property violates the initial constraint, at which point a new cluster starts to form from the remaining nodes of the original graph.

Adamcsek et al. proposed an independent platform for locating overlapping groups of interconnected nodes[5]. This strategy merges nodes into clusters using the Clique Percolation Method (CPM) of Palla et al. to form k (for k-clique) percolation clusters of the network[6]..

Pizzuti and Rombo proposed three local search co-clustering strategies. The basic concept behind these co-clustering methods is to search for dense sub-matrices in the adjacency matrix. The quality of sub-matrices differs from one strategy to another depending on the contribution of the proteins to improve the quality function[7-9].. Additionally, Pizzuti and Rombo stated the complex detection problem in PPI networks as a single-objective optimization problem and devised the methodology of evolutionary algorithms (EAs) to solve the formulated problem[10]. They formulated different topology-based quality functions, including connection density, edge cut, and a metric distance between nodes, as fitness models. Their investigations showed that EAs have more detection ability than traditional complex detection algorithms[10]. Although Pizzuti and Rombo showed that EA has advantages over other complex detection algorithms, they presented EA in its more general form. The main contribution of this paper is to introduce a heuristic operator to be injected into the general framework of the EA to improve its detection ability.

Attea and Abdullah improved EA-based complex detection by integrating a heuristic operator into the EA framework, leading to enhanced detection ability[1]. Bader and Hogue's automated method for finding molecular complexes in large PPI networks further refined the network weighting and complex detection steps to improve accuracy[2]. Hanna and Zaki developed a ranking algorithm combined with a refined merging procedure to detect protein complexes, focusing on enhancing the accuracy and reliability of the detected complexes[3]. However, these methods have certain limitations. For example, while Attea and Abdullah's approach improves the EA's performance, it may still suffer from premature convergence issues[1]. Bader and Hogue's method, although accurate, might not handle overlapping clusters effectively[2]. Hanna and Zaki's ranking algorithm enhances detection but may require substantial computational resources for large networks[3].

Nepusz et al. introduced the ClusterONE (Clustering with Overlapping Neighbourhood Expansion) algorithm to identify overlapping protein complexes from PPI networks [7]. ClusterONE employs a greedy growth process, starting with a seed and incrementally adding nodes to the cluster based on a cohesiveness score. This method is effective in detecting protein complexes of varying densities and sizes. However, the reliance on a greedy algorithm can lead to suboptimal clusters, as the local choices made during cluster expansion might not always contribute to the global optimum. Moreover, the performance of ClusterONE can be sensitive to the initial seed selection, potentially affecting the consistency of the detected complexes.

Liu et al. developed the Core-Attachment based method for protein complex identification (COACH) [11]. COACH identifies protein complexes by first detecting dense core structures within the network and then attaching peripheral proteins to these cores. This two-step process helps in forming more complete complexes. Nevertheless, the method's dependence on accurately identifying core structures means that any errors or ambiguities in the core detection phase can propagate through to the final complex detection. Additionally, the attachment phase might sometimes add peripheral proteins that do not significantly contribute to the biological relevance of the detected complexes.

Li et al. presented the ClusterMaker algorithm, which combines topological features with functional annotation to enhance the accuracy of protein complex prediction [12]. ClusterMaker initially uses spectral clustering to generate clusters based on topological properties and then refines these clusters by incorporating functional similarity data. While this hybrid approach can improve the biological relevance of the detected complexes, it also introduces complexity in balancing the influence of topological and functional data. The need for high-quality functional annotation data can also limit the algorithm's applicability in cases where such data is sparse or noisy. Furthermore, the spectral clustering component can be computationally intensive, especially for large PPI networks.

IV. PROPOSED METHODOLOGY

In this research, we propose a novel approach that combines a ranking algorithm with a single-objective evolutionary algorithm to effectively detect protein complexes in PPI networks. Our method leverages the inherent topological properties of protein complexes, which are often underutilized in existing approaches. By integrating a ranking algorithm as a heuristic operator, we aim to improve the initial population of the evolutionary algorithm, thereby enhancing the quality and accuracy of the detected complexes.

To demonstrate the effectiveness of our proposed method, we plan to conduct experiments on well-known PPI networks, including those from Saccharomyces cerevisiae (yeast). We will also use reference sets of benchmark

complexes and generate new random networks to evaluate the impact of perturbing protein interactions. We anticipate that the proposed algorithm will achieve a higher predictive level of matched protein complexes compared to several state-of-the-art methods. Furthermore, we expect that the integration of the ranking algorithm will significantly improve the performance of the evolutionary algorithm, leading to more accurate and reliable detection of protein complexes.

Our research aims to present a significant advancement in the field of protein complex detection by addressing the limitations of existing methods. By integrating a ranking algorithm with an evolutionary algorithm, we expect to not only improve the initial population but also enhance the overall detection accuracy and reliability. We anticipate that this novel approach will demonstrate superior performance on benchmark datasets, making it a valuable tool for bioinformatics researchers studying PPI networks.

1. EVOLUTIONARY ALGORITHMS

Evolutionary algorithms (EAs) are heuristic optimization methods inspired by the principles of natural evolution, such as competition, selection, reproduction, and mutation. These algorithms aim to improve the fitness of a population of potential solutions by exploring the search space associated with an optimization problem. In the context of protein complex detection in protein-protein interaction (PPI) networks, a single objective EA can be particularly effective.

EAs operate by partitioning the search space into a finite set of points, with the population representing a small, arbitrary subset of these points. The population is denoted as $I_\mu = (I_1, I_2, \ldots, I_\mu)$, where each individual I_i represents a potential solution. A fitness function, $F(I)$, evaluates different regions in the search space based on the processed population. The primary goal of the EA is to enhance this fitness function through iterative transformations.

The EA employs three main operators: selection, recombination (crossover), and mutation, denoted as $\text{Trans}: I_\mu \to I_\mu$. The process begins with the selection operator s, which chooses individuals from better regions of the search space. The recombination operator, r, combines two selected individuals to produce new offspring, while the mutation operator, m, introduces slight perturbations to explore new regions of the search space.

Traditional mutation operators alter an allele of a mutated gene I_i to any value j provided that proteins I_i and j interact in the PPI network (i.e.,$A(I_i, j) = 1$). However, to enhance the performance and reliability of the EA, domain knowledge must be integrated into the design process. In this paper, we propose a heuristic perturbation operator, $h: I \to I$ specifically tailored for the complex detection problem in PPI networks.

The proposed heuristic perturbation, termed the protein complex attraction and repulsion operator, is designed to satisfy the topological properties at both the complex and protein levels. This operator aims to release sparse interactions within the complex and accurately delineate its boundaries, as many protein complexes in nature are small, comprising only two or three proteins.

This single objective EA follows a general iterative loop until a stopping condition is reached, denoted as $\iota: I_\mu \to$ {true,false} Throughout this iterative process, the population undergoes successive transformations, continually evolving towards better solutions. By employing a single objective EA with the proposed heuristic perturbation operator, this methodology aims to overcome limitations found in traditional clustering algorithms. Specifically, the mutation operator facilitates exploration of the search space by perturbing solutions toward nearby regions, thereby avoiding local optima and enhancing the robustness of the complex detection process.

Algorithm 1 The General Framework of EA

Input: $\mu, \Theta_s, \Theta_r, \Theta_m, p_c, p_m, \iota$
Output: optimal I^*;
$t \leftarrow 0$;
Initialize population $\mathbb{I}^\mu(t) \leftarrow (I_1, I_2, \ldots, I_\mu)$;
for $i \leftarrow i \in (1, 2, \ldots, \mu)$ **do**
  Evaluate $F(I_i(t))$;
end for
while $\iota(\mathbb{I}^\mu(t)) \neq true$) **do**
  $t \leftarrow t + 1$;
  for $i \leftarrow i \in (1, 2, \ldots, \mu)$ **do**
    $I_{i,1}(t) \leftarrow \Theta_s(\mathbb{I}^\mu(t-1))$
    $I_{i,2}(t) \leftarrow \Theta_s(\mathbb{I}^\mu(t-1))$
    $I_i'(t) \leftarrow \Theta_r(I_{i,1}(t), I_{i,2}(t), p_c)$;
    $I_i'(t) \leftarrow \Theta_m(I_i(t), p_m)$;
    Evaluate $F(I_i'(t))$;
  end for
end while
Return $I^*(t)$

2. RANKING ALGORITHM

In our method, the ranking algorithm enhances the initial population of the EA by prioritizing proteins based on their centrality in the network. This prioritization ensures that the EA starts with a more informed set of potential solutions, increasing the likelihood of detecting biologically relevant complexes. By focusing on central proteins, which are often key players in multiple interactions, the algorithm effectively narrows down the search space to the most promising candidates. Additionally, our approach specifically addresses the detection of overlapping complexes, which are crucial for accurately representing the multifaceted roles of proteins in cellular processes.

Overlapping complexes are particularly significant because many proteins are involved in multiple functional modules, and capturing this overlap is essential for a realistic and comprehensive understanding of protein interactions. This feature of our method not only enhances the biological relevance of the detected complexes but also provides insights into the dynamic and interconnected nature of cellular networks. Through this dual focus on centrality and overlap, our approach aims to achieve a more accurate and holistic mapping of protein-protein interactions.

The steps involved in our ranking algorithm are as follows:

1.**Pruning**: Removing unreliable protein interactions from the dataset using a technique that assigns weights based on the network topology. Interactions with weights below a specified threshold are considered unreliable.

2.**Filtering**: Identifying and removing bridge, fjord, and shore proteins that could add noise to the network.

3.**Protein Prioritization:** Ranking proteins using a ranking algorithm analogous to the PageRank algorithm. This step identifies essential proteins likely to play key roles in cellular functions, which serve as starting points for complex formation.

4.**Cluster Identification**: Forming detected complexes based on the ranked proteins, using a model that allows proteins to belong to more than one complex.

5.**Data Preparation:** Filtering the set of predicted complexes to remove possible duplicates generated due to the overlap assumption.

Pseudocode of the Merge-by-Cohesiveness algorithm

```
Merge_by_Cohesiveness (C1, C2, merging_threshold)
ep1 = (set of essential proteins in C1)
ep2 = (set of essential proteins in C2)
if size(ep1) > size(ep2) then
    larger_set = ep1
else larger_set = ep2
end if
ep = ep1 ∪ ep2
if size(ep)>size(larger_set)*merging_threshold then
    C = C1 ∪ C2
    for prot in C do
        N_prot = (set of neighbors of prot)
        for n_prot in N_prot do
            C' = C ∪ {n_prot}
            if Cohesive(C') ≥ Cohesive(C) then
                C = C ∪ {n_prot}
            end if
        end for
    end for
end if
```

6.**Merging by Cohesiveness**: Merging detected complexes whose overlap exceeds a merging threshold, and iteratively extending the resultant complex using a defined merging procedure.

7.**Refinement**: Filtering the refined set of predicted complexes to remove possible replicated copies resulting from the merging step.

3. INTEGRATION OF PROTEIN RANKING INTO EVOLUTIONARY ALGORITHM FOR COMPLEX ALLOCATION

In our new approach, we will leverage protein ranking to allocate proteins to each complex. This method involves providing the ranks obtained from the protein ranking algorithm to the evolutionary algorithm. By doing so, we ensure that the natural selection process remains intact, but with a significant modification: instead of using traditional mutation, we incorporate the protein ranking algorithm.

This integration allows us to:

- Utilize the strengths of evolutionary algorithms, such as natural selection, to optimize complex formation.
- Ensure that protein allocation is guided by their rankings, which can improve the overall efficiency and accuracy of the process.
- Maintain the dynamic and adaptive nature of evolutionary algorithms while benefiting from the insights provided by protein ranking.
- By combining these two powerful methodologies, we aim to enhance the effectiveness of protein complex formation, leading to better results and more efficient computational processes.

4. PARAMETERS USED IN THE APPROACH

A (Adjacency Matrix): Represents the interaction between proteins, where each element indicates the presence(1) or absence(0) of an interaction.

Child: The offspring generated in each generation of the evolutionary algorithm.

IndicesInteractionProtein: Indices of proteins involved in interactions.

MaxNumberInteractionProtein: The maximum number of interactions a protein can have.

NumInteractionProtein: The current number of interactions for a given protein.

Pm (Probability of Mutation): Set to 0.2, this parameter defines the likelihood of mutation occurring during the evolutionary process.

d, Damping factor: Set to 0.85, used in the ranking algorithm to account for the probability of continuing the random walk.

ε: A small threshold value (0.001) used for convergence criteria in iterative processes.

In our new approach, we will leverage protein ranking to allocate proteins to each complex. This method involves providing the ranks obtained from the protein ranking algorithm to the evolutionary algorithm. By doing so, we ensure that the natural selection process remains intact, but with a significant modification: instead of using traditional mutation, we incorporate the protein ranking algorithm.

5. LIMITATIONS OF OUR APPROACH

1. **Scalability Issues with Large Datasets:**

The approach may not perform efficiently with large datasets due to the increased computational complexity. As the size of the dataset grows, the time and resources required to rank proteins and run the evolutionary algorithm will also increase, potentially leading to scalability issues.

2. **Sensitivity to Noise in the Dataset:**

The presence of noise in the dataset can adversely affect the accuracy of protein rankings. Erroneous or noisy data may lead to incorrect rankings, which can subsequently impact the performance of the evolutionary algorithm and result in suboptimal complex formations.

3. **Challenges with Excessive Overlapping Complexes**:

If there are a significant number of overlapping complexes, the approach may encounter difficulties in effectively allocating proteins. Overlapping complexes can complicate the ranking and complex migration process, potentially leading to conflicts or ambiguities in protein allocation.

V. REFERENCES

1. B. A. Attea and Q. Z. Abdullah, "Improving the performance of evolutionary-based complex detection models in protein–protein interaction networks," Iraqi Journal of Science, vol. 57, no. 4A, pp. 2513-2528, 2016.

2. G. D. Bader and C. W. V. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinformatics, vol. 4, no. 2, pp. 1-27, 2003.

3. King, A.D., Pržulj, N., & Jurisica, I. (2004). Protein complex prediction via cost-based clustering. Bioinformatics, 20(17), 3013-3020.

4. A.Altaf-Ul-Amin, K. Nishikata, T. Koma, T. Miyasato, S. Kanaya, and T. Aoki, "Detecting protein complexes in protein interaction networks using a deterministic algorithm," Journal of Bioinformatics and Computational Biology, vol. 7, no. 2, pp. 199-217, 2009.

5. A.Adamcsek, G. Palla, I. J. Farkas, I. Derényi, and T. Vicsek, "CFinder: locating cliques and overlapping modules in biological networks," Bioinformatics, vol. 22, no. 8, pp. 1021-1023, 2006.

6. K. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," Nature, vol. 435, no. 7043, pp. 814-818, 2005.

7. C. Pizzuti and M. Rombo, "Co-clustering algorithms for detecting protein complexes in PPI networks," Journal of Computational Biology, vol. 17, no. 4, pp. 527-539, 2010.

8. C. Pizzuti and M. Rombo, "A co-clustering approach for mining large biological networks," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 2, pp. 523-535, 2011.

9. C. Pizzuti and M. Rombo, "A novel co-clustering algorithm for discovering protein complexes," BMC Bioinformatics, vol. 13, no. S17, pp. 1-16, 2012.

10. M. A. Altaf-Ul-Amin, K. K. Mizuguchi, S. Kanaya, and T. Aoki, "An algorithm for finding clusters of proteins in the protein interaction networks," Journal of Theoretical Biology, vol. 248, no. 3, pp. 454-464, 2007.

11. Liu, G., Wong, L., & Chua, H.N. (2009). Complex discovery from PPI networks with core-attachment structures. *BMC Bioinformatics, 10*, 169.

12. Li, M., Chen, J., Wang, J., Hu, B., & Chen, G. (2008). Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. *BMC Bioinformatics, 9*, 398.