How to Cite:
Fathima, H., Kavitha, V., Nallusamy, C., Suryaprabha, D., Jeyanthi, P., Leelavathi, M., & Musthafa, A. S. (2022). Using PPI-GA for identifying protein complexes in protein-protein interaction networks. *International Journal of Health Sciences*, *6*(S3), 11790–11800. https://doi.org/10.53730/ijhs.v6nS3.8873

Using PPI-GA for identifying protein complexes in protein-protein interaction networks

Fathima H.
Assistant Professor, Department of Computer Applications, KSR College of Arts and Science, Tiruchengode,
Tamil Nadu -637 215
Email: fathi.fathimahussain@gmail.com

Dr. V. Kavitha
Assistant Professor, Department of Information Technology, Nehru Arts and Science College, Coimbatore
Email: sekarkavi09@gmail.com

C. Nallusamy
Professor, Department of Information Technology, K.S.Rangasamy College of Technology,
Tiruchengode
Email: nallu80@yahoo.com

Dr. D. Suryaprabha
Assistant Professor, Department of Computer Applications, Nehru Arts and Science College, Coimbatore
Email: nascdrdsuryaprabha@nehrucolleges.com

P. Jeyanthi
Assistant Professor, Department of Information Technology, Nehru Arts and Science College, Coimbatore
Email: nascpjeyanthics@nehrucolleges.com

M. Leelavathi
Assistant Professor, Department of Information Technology, Nehru Arts and Science College, Coimbatore
Email: leelavathi.mphil86@gmail.com

A. Syed Musthafa
Associate Professor, Department of Information Technology, M.Kumarasamy College of Engineering,
Karur
Email: syedmusthafait@gmail.com

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 18 Feb 2022, Manuscript revised: 27 April 2022, Accepted for publication: 9 June 2022

***Abstract*---**The major focus of an interdisciplinary field of study termed proteomics is a comprehensive analysis of proteins to evaluate their genetic variety, examine their distinctions, and respond to stresses. Proteomics' major goal is to use protein chemistry, bioinformatics, and biology to detect and measure proteins as well as analyze their post-translational modifications and interactions. Any disruption in the interacting network of proteins can lead to biological abnormalities and diseases like Alzheimer's and cancer. The majority of existing computational approaches for detecting protein complexes are based on certain topological properties of protein-protein networks (PPI). In this paper, it initially provides a novel encoding technique for representing the clarification and solution, and next we recommend the concept of PPI-GA, an innovative clustering based algorithm which is based on genetic approach algorithm which employs a innovative multi based objective excellence function to find out the protein complexes. A different two gold based standard and the real-world based datasets are been used to evaluate proposed algorithm. The obtained result shows that the suggested method can discover key protein complexes and that it offers more accurate results than existing protein complex identification techniques.

***Keywords*---**protein-protein networks, genetic approach, protein based complexes, clustering approach.

Introduction

Proteins are necessary nutrients for carrying out biological functions. They're also important for understanding molecular functioning and biological processes. Proteins in biological organisms are usually arranged into protein complexes. Protein-protein interaction occurs when two or more proteins come into physical contact due to electrostatic forces or chemical processes. Biological functions are frequently the result of such interactions. The huge molecular based complexes, that are prepared of very numerous structured oriented proteins which interact with each other, it performs different key based molecular actions within the cell, like DNA based replication and the even based protein metabolism.

Protein interactions are significant because they increase our understanding of illnesses and can lead to the development of new therapeutic strategies. Although there are numerous experimental and computational methods for detecting protein complexes, each has its own set of benefits and drawbacks. The protein based complexes could be detected by the experimental models such as the yeast based two-hybrid screening with the tandem based affinity purification through the mass spectrometry [1]. The nature and behaviour of these interactions can be investigated using a variety of biophysical approaches. These approaches have some drawbacks, which are detailed in [2].

PPI data derived from trials, for example, often has significant false-negative and false-positive rates, reducing the accuracy of the experiment results. In spite of the truth that different computational density-oriented approaches to find the

connected based sub graphs in the PPI networks been projected more than a last few decade, still few challenges in the identifying the protein complexes related to the network's major topology and properties, that reduces the accurateness of the results. In theory, graph theory is used to model the interactions between artifacts (here proteins). With the result, the recognized computational based methods are been employed in the conjunction with the experimental methods.

Toward discovering the protein based complexes, an undirected oriented graph is been used to imitate the set of every protein-protein based interactions in a specified organism, with every vertex and the edge representing the protein with the interaction among between the two proteins, correspondingly. The protein based detection in an interaction among the protein based network is very difficult owing to a multiplicity of the variables. A rate of the false-positive and false-negative which results in pulling out the interactive data be often very high. Each protein can also be a part of many protein complexes. The most of all computational based approaches to detect protein based complexes depend upon the complexes or the extracting clusters from the PPI based networks.

The objective of the PPI network based clustering is to recognize a collection of proteins so as to interact and that are all involved in the similar biological procedure or to perform a convinced biological based function. Due to the NP-hard nature of detecting protein complexes, evolutionary methods such as the genetic algorithm are used to develop near-optimal solutions [3]. To find protein complexes using PPI networks, this research provides a graph clustering technique with a unique multi objective quality function and a new compacted encoding method. In this article, we employ the genetic based algorithm to discover clusters for the protein based complexes model. When it is compared to the existing clustering methods, our results suggest that the proposed technique outperforms them. With the compared based approaches, this approach has the maximum computational accuracy. Clusters of various densities and complexities can be detected using the method. Our approach is scalable, which means it can cluster a big network without causing performance issues.

The below listed are the paper's contributions: [1] Introducing a novel multi objective based quality function to identify the protein complexes; [2] introducing a new encoding based approach to display the solutions in the evolutionary method to the genetic based algorithm, with the intention to reduce the search space. The existing evolutionary based algorithms provide a good solution to space of where the n denotes the number of the proteins. The proposed encoding in this paper lowers the state space to, which is substantially smaller than. This decrease in the search space makes it easier for the GA to solve the problem and converge faster and identify acceptable solutions in earlier generations.

Literature Review

Proteins are an essential component of all living things, and they are involved in the management and execution of nearly all biological processes. To find the relevant groups of the proteins from the PPI based networks, numerous clustering techniques been developed [4]. There are three major ways to cluster graphs with the goal of detecting protein complexes. The first is to locate sub graphs with

specific connections, known as network motifs, which can be characterized as a complex or a subset of it. Clique is been a technique of the sub graphs which is discovered by this scheme. This approach's application is currently limited due to its temporal complexity. The second is graph-growing, which involves employing greedy search methods to grow and complete a cluster around a vertex as a seed. The third one has a number of variations. The methods attempt to minimise or maximise cluster parameters like connection density, edge cut, or metric distance between nodes. In general, the algorithms seek to maximise a graph's objective function.

The MCL based technique were developed [5], that uses the expansion and the inflation based operators to distinguish the random based walk process based on the graph and to locate the protein families. This approach is proved to be one of the best ways for clustering a graph in [6]. In [7], the proposed the RNCS based algorithm, that uses the cost based function and the local search algorithm which is to distribute the vertices in the graph. [8] employed the Newman average edge technique to recover protein complexes in another study. The edges with the greatest average value are deleted iteratively until the necessary clusters are achieved with this approach.

The MCODE based algorithm was proposed [10], and that is the one of the primary and the most well-known algorithms to predict protein based complexes. A greedy technique is used in this algorithm to create a cluster around a seed protein. The weighting vertices, the forecasting based clusters, and the preprocessing to the filter or to add a vertex into clusters which are the major three fundamental steps in the algorithm. The SPICi algorithm was proposed [15] to cluster massive biological networks. The clustering process begins with the production of an initial seed pair, which is then supplemented with nodes. The CPA algorithm, developed, employs the degree and distance length between two nodes in a cluster. A cluster is created, then deleted from the network, and the process is repeated until no edges remain. The seed to form a cluster is the node with the highest weight.

The script then searches for and adds surrounding nodes to the seed. If the function based value for multiple nodes is similar, the node among the longer distance length is been added to the current based cluster, and if the distance lengths are similar, all nodes are been added to the cluster. The ClusterOne based algorithm, proposed by [15], takes advantage of a greedy approach. In most cases, the highest-degree seed vertex in the network is chosen. Following that, a function with a greedy approach is used to add or remove vertices from the seed. To discover overlapping in clusters, all pair complexes with an Overlap Score value greater than a specific threshold are added together. Finally, complexes with fewer than three proteins or a density below a specified level are eliminated.

The CFinder algorithm was proposed [15]. This method is based on the clique notion. A cluster can be thought of as a small entity that is fully connected. The minimum number of common proteins is determined using a parameter. First, the method extracts all of the network's subgraphs. The number of common nodes between two cliques is then extracted using a matrix. Finally, cliques that share nodes are considered nearby, and the two cliques are merged. Pizzuti

introduced a modified version of the restricted neighbourhood search clustering (RNSC) technique with new operators. Instead of assessing the cost function at each node, the cost function is rebuilt and applied to a group of nodes. The results showed that, while there were fewer clusters predicted than using the RNSC approach, the clusters were of good quality. The MOEPGA algorithm, which is based on a multiobjective function, was presented by Cao. The algorithm's fundamental idea is to take into account the network's name, size, density, and distance length, among other topological aspects. The programme first evaluates an interaction protein network before formulating a multiobjective function based on the network's topological qualities. In each subgraph, the initial population, mutation, and selection are all done in three steps.

The MFOC is a moth-flame optimization-based protein complex prediction algorithm proposed by Lei. First, the MFOC creates a weighted dynamic PPI network by combining topological and biological data. The moths are then allowed to fly spirally around the flames to create the complexes, using a layer-by-layer technique to discover the cores of protein complexes as the flames. For clustering a PPI network, Sikandar described an approach that considers both topological patterns and biological properties. Each complicated subgraph is modelled by decision tree learners in their method. They employed the divide and conquer technique to build decision trees using a training set of known complexity in a depth-first and best-first way. Gu proposed a Markov-based clustering technique that identifies overlapping structures based on link similarity. In this research, they demonstrated that their suggested strategy may uncover more biologically significant modules in PPI networks. [3] It suggested an algorithm in which a proposed genetic algorithm clusters the network and its population is generated using random and spectral methods. Genetic procedures and goal functions are used to improve clustering quality. Because the goal of clustering is to locate subgraphs with a lot of intraconnections, the technique starts by using spectral clustering to determine the shortest path between two subgraphs. The network's minimal cut is calculated using eigenvector, Laplacian, and diagonal matrices.

Proposed Algorithm

We offer a new PPI network clustering approach based on the evolutionary algorithm in this section. The genetic algorithm (GA) is a computational model based on biological evolution that outperforms traditional optimization approaches in many ways. GA can adapt to complex optimization issues, allowing it to address a wide range of problems. In most circumstances, each GA begins with a randomly selected initial population that evolves to achieve a specific objective, such as optimising a function, by applying genetic operators (crossover and mutation) to the available chromosomes and resulting in the creation of a new population. The process is repeated until the algorithm satisfies the termination condition, at which point the new population is reset on the prior one. We use an evolutionary algorithm (PPI-GA) to find the near-optimal solution to the challenge of discovering protein complexes in interactive PPI networks, based on evolutionary algorithms' adequate performance in finding a near-optimal solution to NP-complete problems. The operators utilised in the PPI-GA are discussed in the following sections:

Encoding: As a basic chromosome, GA encodes all of the possible solutions to a given problem. The genetic algorithm's performance is highly dependent on the encoding method used. A genetic algorithm can converge to unsatisfactory results due to a poor encoding technique. We introduce a new encoding method for representing solutions in PPI networks clustering. Each node (i.e., protein) in the PPI networks is given a unique numerical identifier to accomplish this. These IDs control where in the encoded string that node is utilised. A partition of the PPI network is encoded as a string, with each string being a permutation of N numbers. An encoding, P, is defined in formal terms as follows: The number of proteins in the PPI networks is N, and (1 iN) has a value in the range [1, N]. The number of proteins indicated by is equal to the length of each chromosome. Algorithm 1 demonstrates how to decode a chromosome into a clustering algorithm.

```
Let P is a chromosome which holds a permutation of PPI
network.
node 1 to N
For each element i in the chromosome
{
If (P[i] ≥ i) then
}
    Create a new cluster and assign node i to it.
  Else
    {
Assign node i to the same cluster as node P[i]
}
```

Consider the following example of encoding: Because the initial value in P in the above permutation encoding is greater than 1, a new cluster called cluster 1 is generated and node 1 is assigned to it. Node 2 is assigned to a new cluster since the second value in P is 3, which is more than 2. (cluster 2). Node 3 is in the same cluster as node 2 and belongs to it. Because the fourth and fifth values in P are 1 and 4, nodes 4 and 5 are assigned to the same clusters as nodes 1 and 4. Algorithm 2 shows the encoding method described in the above.

```
Input: a PPI network with N proteins and K complexes (clusters)
Output: a Chromosome P with N genes
Begin
Create an empty chromosome C with N genes.
{For i = 1 to k}
A is an array containing nodes in cluster #i
P is the number of nodes in A
Sort A, in ascending order
{For j = 1 to p-1
```

Assign A(j) to C(A(j + 1))
Next j
Assign A(p) to C(A(1))
Next i
}

By looking into the suggested encoding, it becomes clear that for a particular PPI network with n nodes, the search space formed by the encoding for the genetic algorithm may comprise up to chromosomes, which is significantly less than the chromosomes produced by the existing value-based encoding. Because the search space is smaller, this makes it easier for the GA to solve the problem and converge faster and identify acceptable answers during earlier generations. Algorithm 3 depicts the pseudocode of our algorithm. In each iteration, the evolutionary operators (selection, crossover, and mutation) are applied to the population, respectively, and the population is steered toward the optimal solution.

Output: a clustered PPI

The steps of the algorithm are as follows:

- Generate the initial population
- Evaluate the fitness of the chromosomes using Equation
- Generate the intermediate population using the standard roulette whee selection strategy for selecting potentially useful chromosomes fo recombination
- Apply crossover operator on the intermediate
- Apply mutation operator on the intermediate generation
- Apply elitism operator
- Replace the old population with a new population
- If the number of iterations is less than (8)If the number of iterations

Initial population and operators

To start the population, chromosomes with the population's number are generated at random. Each chromosome is an equal-probability permutation of N integers. The tournament selection method is used to choose chromosomes. Initially, a certain and limited number of chromosomes (tournament size) is chosen at random. The selected chromosomes are then compared, and the best-fitting chromosome is picked. We require a crossover operator to retain the permutation form in each chromosome during the evolutionary process of the genetic algorithm because the proposed encoding method is a permutation. A two-point crossover operator is employed in the PPI-GA. Take a look at the following two parents: P1: P2: To create two offspring O1 and O2, the algorithm produces two places k1 and k2 at random.

Results

Collins, an interaction protein network, is used to test the PPI-GA clustering algorithm. The BioGrid dataset, which has 1879 vertices and 140849 edges and is one of the most important and large-scale datasets in protein-protein networks, was used to create this network. The Collins network contains 1004 proteins and 8319 protein interactions. The network has an average degree of 16.57 (where the degree of a node in the network equals the number of edges connecting to a vertex) and a density of 0.016. (which is defined as the number of network interactions per total number of possible edges of network). The Cytoscape software was used to simulate this network, which is modelled as an interactive graph. Cytoscape is a bioinformatics software platform for visualising molecular interaction networks that is free and open-source. We need a gold standard to evaluate clustering algorithms so that we can compare the proposed algorithm to existing techniques. We also employ the reference clusters CYC2008 and MIPS, which are two of the most prominent standards for determining the complexity of interacting protein networks. The coherence of the collected clusters is further assessed using the Gene Ontology (GO) cellular component ontology.

Evaluation

Precision (also known as positive predictive value), Recall (also known as sensitivity), and F-measure assessment measures are used to assess the quality of predicted clusters. These criteria are based on the number of nodes in a pair (here proteins). "These metrics strive to estimate whether the prediction of this pair as being in the same cluster was right with regard to the underlying true categories in the data for each pair of proteins that share at least one cluster in the overlapping clustering findings" (i.e., reference cluster). Precision is defined as the percentage of pairings accurately placed in the same cluster, Recall is the percentage of real pairs found, and F-measure is the harmonic mean of Precision and Recall"

Stability

The performance of an algorithm is said to be correct if the algorithm is stable. The term "algorithm stability" refers to the fact that the algorithm's findings are not dispersed, meaning that if the algorithm is run numerous times for the same sample and under the same conditions, the quality of the results is consistent across all runs. We collect the results of 20 different runs to demonstrate the algorithm's stability. The Collins data sample in the CYC2008 and MIPS tests yielded the following findings..

T-Test

This test analyses the average of two independent samples with a normal distribution using statistics. In other words, the averages obtained from random samples are examined in this test. This implies we choose samples from two separate communities at random and compare their means. This method is based on a normal distribution and is best suited for small samples when the comparator data has a normal distribution. The hypothesis is rejected and the

mean of the data has no significant difference if the variable value Sig. (2-tailed) is less than the planned error level 0.05; otherwise, there is insufficient reason to reject the hypothesis. Given the higher Sig. (2-tailed) from 0.05 in all cases, the data has a normal distribution and the -test can be utilised. The outcomes of the numerous runs are separated into two categories (first half and the second half). Using the IBM statistical software SPSS 21, we run the -test on the samples. As shown in the tables, the value Sig. (2-tailed) is greater than 0.05 in all situations, therefore there is no reason to reject the premise that the data mean is not significantly different

Clustering Quality

We use the PPI-GA algorithm with Gene Ontology components to assess the biological importance of clusters in the Collins network. To locate the most important GO terms, the GO term finder is used. The p value is used to determine the likelihood of a given GO word enriching a projected complex by chance. The SGD GO Term software modules are used to calculate this function. The hypergeometric distribution defines the p value as follows: where is the number of proteins in the PPI network, is the size of a list of proteins, marks to the reference term of interest (the number of proteins in a GO term), is the number of proteins annotated with the same GO term, is the number of proteins in an extracted cluster, and is the number of proteins shared.As a result, the closer the value is to zero, the more biologically important the anticipated complexes are. A cutoff value is used to distinguish significant from inconsequential groups when determining biological relevance. Only matches with a value of less than 0.05 are statistically significant.

Conclusion

Within PPI networks, detecting protein complexes is an important analysis tool that allows us to understand cellular architecture and biological processes. We proposed a new method for detecting and predicting protein complexes in PPI networks in this research. To optimize cluster intraconnections and decrease cluster interconnections, we devised a multiobjective function. The objective function performed better than the other techniques, according to the findings. Our clustering approach is more accurate than the algorithms used by MCODE, MCL, ClusterOne, and Ramadan et al., resulting in better output and a higher F-score. We believe that the proposed method could be applied to other biological networks, such as metabolic networks, in the future. For overlapping proteins, the proposed technique can be improved (so that a protein may belong to several clusters). Instead of genetic algorithms, we will try to apply some of the methods provided in to the PPI networks clustering problem, as well as other current methods.

References

1. J. Zahiri, A. Emamjomeh, S. Bagheri et al., “Protein complex prediction: a survey,” *Genomics*, vol. 112, no. 1, pp. 174–183, 2020.
2. Syed Musthafa A, "Block chain Based Secure Data Transmission among Internet of Vehicles," 2022 2nd International Conference on Innovative

Practices in Technology and Management (ICIPTM), 2022, pp. 765-769, doi: 10.1109/ICIPTM54933.2022.9753890.
3. X. Lei, M. Fang, and H. Fujita, “Moth-flame optimization-based algorithm with synthetic dynamic PPI networks for discovering protein complexes,” *Knowledge-Based Systems*, vol. 172, pp. 76–85, 2019.
4. R. Senthil Kumar, Syed Musthafa A, "Recursive CNN Model to Detect Anomaly Detection in X-Ray Security Image," 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), 2022, pp. 742-747, doi: 10.1109/ICIPTM54933.2022.9754033.
5. L. Gu, Y. Han, C. Wang, W. Chen, J. Jiao, and X. Yuan, “Module overlapping structure detection in PPI using an improved link similarity-based Markov clustering algorithm,” *Neural Computing and Applications*, vol. 31, no. 5, pp. 1481–1490, 2019.
6. A.Syed Musthafa, "Oryza Sativa Leaf Disease Detection using Transfer Learning," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 1-7, doi: 10.1109/ICSCDS53736.2022.9760972.
7. S. Kalaivani, D. Ramyachitra, and P. Manikandan, “K-means clustering: an efficient algorithm for protein complex detection,” in *Progress in Computing, Analytics and Networking*, pp. 449–459, Springer, Singapore, 2018.
8. H. W. Mewes, “The bioinformatics of the yeast genome-A historical perspective,” *Yeast*, vol. 36, no. 4, pp. 161–165, 2019.
9. Murugan G, Syed Musthafa A, Abdul Jaleel D, Sathiya Kumar C, “Tourist Spot Proposal System Using Text Mining”, International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, No. 2, ISSN 2278-3091, Page 1358- 1364, April 2020
10. F. A. Paniagua, “The null hypothesis is always rejected with statistical tricks: why do you need it?*” Interamerican Journal of Psychology*, vol. 53, no. 1, pp. 17–27, 2019.
11. N. Teymourian, H. Izadkhah, and A. Isazadeh, “A fast clustering algorithm for modularization of large-scale software systems,” *IEEE Transactions on Software Engineering*, p. 1, 2020.
12. Sengan, Sudhakar, Priya, V. Syed Musthafa, A., Ravi, Logesh, Palani, Saravanan, “A Fuzzy based High-Resolution Multi-View Deep CNN for Breast Cancer Diagnosis through SVM Classifier on Visual Analysis”, Journal of Intelligent & Fuzzy Systems, IOS Press, 10.3233/JIFS-189174, Page 1-14, September 2020
13. B. Pourasghar, H. Izadkhah, A. Isazadeh, and S. Lotfi, “A graph-based clustering algorithm for software systems modularization,” *Information and Software Technology*, vol. 133, Article ID 106469, 2021.
14. Keerthana Nandakumar, Viji Vinod, Syed Musthafa Akbar Batcha, Dilip Kumar Sharma, Mohanraj Elangovan,, “Securing data in transit using data-in-transit defender architecture for cloud communication”, Soft Computing (2021), ISSN: 14327643, 14337479, doi.org/10.1007/s00500-021-05928-6, june 2021
15. S. Pourbahrami and S. Azimpour, “A new method for detection of clustering based on four zones Apollonius circle,” *Iran Journal of Computer Science*, vol. 3, no. 1, pp. 59–64, 2020.
16. M. Bodaghi and K. Samieefar, “Meta-heuristic bus transportation algorithm,” *Iran Journal of Computer Science*, vol. 2, no. 1, pp. 23–32, 2019.

17. Murugan G, Syed Musthafa.A, Mohanraj.B, Priyanga.S, Krishnan.C, “High Security Distributed MANETs using Channel De-noiser and Multi-Mobile-Rate Synthesizer”, International Journal of Advanced Trends in Computer Science and Engineering, Volume 9, No. 2, ISSN 2278-3091, Page 1346-1351, April 2020
18. R. Nithya, K. Amudha, A. Syed Musthafa, Dilip Kumar Sharma, Edwin Hernan Ramirez-Asis, Priya Velayutham, V, “An Optimized Fuzzy based Ant Colony Algorithm for 5G-MANET”, Computers, Materials & Continua, Volume 70, issue 01, DOI:10.32604/cmc.2022.019221, September 2021
19. Rinartha, K., Suryasa, W., & Kartika, L. G. S. (2018). Comparative Analysis of String Similarity on Dynamic Query Suggestions. In *2018 Electrical Power, Electronics, Communications, Controls and Informatics Seminar (EECCIS)* (pp. 399-404). IEEE.
20. Suryasa, I. W., Rodríguez-Gámez, M., & Koldoris, T. (2021). Get vaccinated when it is your turn and follow the local guidelines. *International Journal of Health Sciences*, *5*(3), x-xv. https://doi.org/10.53730/ijhs.v5n3.2938
21. Eviasty, N., Daud, N. A., Hidayanty, H., Hadju, V., Maddeppungeng, M., & Bahar, B. (2021). The effect of personal coaching on increasing knowledge, attitudes, and actions about lactation nutrition, uterial involution, and lochea in public mothers. International Journal of Health & Medical Sciences, 4(1), 155-163. https://doi.org/10.31295/ijhms.v4n1.1669