AUT Journal of Mathematics and Computing

AUT J. Math. Comput., 3(1) (2022) 59-76
DOI: 10.22060/AJMC.2021.20101.1053

The assessment of essential genes in the stability of PPI networks using critical node detection problem

Javad Rezaei[a], Fatemeh Zare-Mirakabad[*a], Sayed-Amir Marashi[b], Seyed Ali MirHassani[a]

[a] *Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran*
[b] *Department of Biotechnology, College of Science, University of Tehran, Tehran, Iran*

ABSTRACT: Essential genes and proteins as their products encode the basic functions of a cell in a variety of conditions and are vital for the survival of a cell. Analyzing the characteristics of these proteins provides important biological information. An interesting analysis is to demonstrate the correlation between the topological importance of a protein in protein-protein interaction networks and its essentiality. Different centrality criteria such as degree, betweenness, closeness, and eigenvector centralities are used to investigate such a correlation. Despite the remarkable results obtained by these methods, it is shown that the centrality criteria in scale-free networks show a high level of correlations which indicate that they share similar topological information of the networks. In this paper, we use a different approach for analyzing this correlation and use a well-known problem in the field of graph theory, Critical Node Detection Problem and solve it on the protein-protein interaction networks to obtain a subset of proteins called critical nodes which have the most effect on the network stability. Our results show that essential proteins have a more prominent presence in the set of critical nodes than what is expected at random samples. Furthermore, the essential proteins represented in the set of critical nodes have a different distribution of topological properties compared to the essential proteins recovered by the centrality-based methods. All the source codes and data are available at "http://bioinformatics.aut.ac.ir/CNDP_PPI_networks/".

Review History:
Received:31 May 2021
Accepted:31 August 2021
Available Online:01 February 2022

Keywords:
Essential genes
Protein-protein interaction network
Centrality
Critical node
Network stability

1. Introduction

With the advent of high throughput technologies and whole-genome sequencing of different species, complete information on proteins is available at the genome level. Understanding the functionality of a living cell now requires studying this level of data on proteins, analyzing the way they cooperate, and modeling them as a system of interacting components which is the main purpose of systems biology [5, 37]. Systems biology, which can be presented in the form of molecular networks, frequently deals with topological and structural analysis using the graph-theoretical concepts [38] to generate new hypotheses about how biological systems are organized and also review, support, or reject the previous biological hypotheses [6].

An interesting challenge in analyzing protein-protein interaction (PPI) networks is to demonstrate the correlation between the topological importance of a protein (a gene product) and its essentiality [7, 8, 9, 13]. Essential genes are vital for the survival of a cell and encode the basic functions of a cell in a variety of conditions. In synthetic biology, essential genes provide insights into the minimal genome required to create a cell with the self-replication capability [1, 2]. Analyzing the characteristics of these genes provide important biological information in explaining how genotype affects phenotype [3], identification of genes related to human diseases [2], and discovering attractive

*Corresponding author.
E-mail addresses: ja.rezaei@aut.ac.ir, f.zare@aut.ac.ir, marashi@ut.ac.ir, a_mirhassani@aut.ac.ir

drug targets for new antibiotics [4].

In analyzing the hypothesis about the correlation between the importance of a protein in the PPI network and its essentiality, different topological properties of biological networks, specially centrality criteria such as degree, betweenness, closeness, and eigenvector centralities [10, 11, 12] are used to reflect the protein topological importance. In that way, the network nodes are ranked according to one of the centrality criteria (the top nodes are considered to be the most important nodes in the network), and a set of the highly ranked nodes is considered as a set containing essential proteins. The more essential proteins in such a set, the more confident the criterion confirms the above hypothesis. Jeong et al. [13] conducted the first research on the *Saccharomyces cerevisiae* PPI network, resulting in the *centrality-lethality rule*. This group found that the frequency of essential proteins in the set of highly-ranked nodes, according to degree centrality, was higher than what expected at random which is also confirmed by others [14, 15, 16, 17]. Estrada [18] used degree, closeness, betweenness, eigenvector, and information centralities, and in a similar study, Zaidi et al. [21] considered degree, closeness, betweenness, pagerank, and Katz centralities, and showed that all centralities are correlated with the essentiality of proteins, meaningfully. Kendall's tau and Spearman's rho rank correlation coefficient are applied by Elena et al. [19] on the data of six different PPI networks of *Saccharomyces cerevisiae*. In another study, Altaf-Ul-Amin et al. [20] used the ROC analysis method to demonstrate the relationship between the centrality and the essentiality of proteins.

Despite the remarkable results obtained by the methods above, none of the centrality criteria used in the literature are recognized as a standard criterion for determining the importance of a node in a network [28]. Furthermore, PPI networks tend to be scale-free which means that a small fraction of the nodes in the network have high degree values while the rest of the nodes having relatively lower degree values are connected to the high-degree nodes. It is shown that the centrality criteria in scale-free networks show a high level of correlations which indicate that they share similar topological information of the networks [61, 62, 63]. In this regard, another concept is used to show the importance of a node in the network, and that is the role of the node in the stability of the network. One of the basic studies on complex networks is to evaluate the stability of the networks against nodes failure or targeted attacks on its nodes. In order to quantify the stability of the network against node removal, network connectivity metrics such as the number of connected components, the size of the largest connected component, and the pairwise connectivity are considered [22, 23, 24]. In the following, we use the two words connectivity and stability many times instead of each other. In the literature, targeted attack on the basis of various centrality criteria and their effect on the network connectivity is studied [11, 12, 25, 26, 27, 28], and it is shown that these centrality criteria are unable to obtain the critical nodes whose removal mostly degrade the network [28].

In this respect, the critical node detection problem (CNDP) is defined. In CNDP, the aim is to obtain a subset of nodes whose removal results in the greatest reduction in the network connectivity [29, 30, 31]. CNDP has many applications [39] in different types of networks, including social networks for finding the most influential users [32], communication networks for preventing the spread of viruses, terrorist networks for disrupting the network [33], and the real-world networks for minimizing the spread of infections [34, 35]. Many exact and heuristic algorithms have been proposed to solve CNDP, which are reviewed in detail by the Lalou study [36]. Furthermore, a complete overview on CNDP is conducted by Rezaei et al. [39].

In this paper, instead of ranking proteins based on a centrality criterion, we employ CNDP on PPI networks for the first time and propose a different approach to investigate the correlation between essential proteins and their topological properties in a PPI network. Due to the NP-Hardness of CNDP [22, 23, 39] and the inability of the exact methods to solve this problem on large networks, we propose a genetic algorithm to solve CNDP and obtain a set of critical nodes. We then explore the frequency of essential proteins in the set of critical nodes found by the genetic algorithm. The results show that the essential proteins play a more effective role in the stability of the PPI network than what expected at random. In addition, we analyze the set of essential proteins that contributed to the optimal solution of CNDP and show that these essential proteins have different topological properties compared to the essential proteins found by ranking methods using the centrality criteria. For example, the degree distribution of the essential proteins in the optimal solution of CNDP is similar to the degree distribution of the whole essential proteins of the PPI network. In contrast, the essential proteins presented in the sets of highly-ranked proteins (regardless of the centrality criterion used) have significantly high degree values and therefore follow a different degree distribution. In other words, this observation distinguishes our approach from the other ranking methods in that we cover essential proteins that none of the ranking methods are able to recover.

In the next section, we define CNDP formally and present the details of a genetic algorithm to solve it. Section 3 prepares experimental results to show the presence of essential proteins among the critical nodes returned by the genetic algorithm.

2. Materials and Methods

In this section, we first explain some of the basic concepts and notations, then define CNDP and the most frequently used centralities formally, and finally, provide the details of a genetic algorithm for solving CNDP.

2.1. PPI network as undirected graph

Among many biological networks, PPI network is one of the most studied networks [45]. A PPI network is considered to be a biological system consisting of proteins as its interacting components and the functional/physical interaction between the proteins. It is represented as an undirected graph $G(V, E)$ where V and E are the set of nodes and the set of edges, respectively, where

$$V = \{1, 2, \ldots, N\},$$

$$E \subseteq \{\{u, v\} | u, v \in V, u \neq v \text{ where the corresponding proteins } u \text{ and } v \text{ interact with each other}\}.$$

To calculate the network connectivity, we use the concept of the connected component in graph G. A subgraph C of graph G with the set of nodes $V(C)$ is called a connected component of G if and only if for any arbitrary nodes $u, v \in V(C)$ there is at least one path in G and furthermore, there is no path between any arbitrary node $u \in V(C)$ and an arbitrary node $v \in V \setminus V(C)$. The set of the components of G is represented by $Comps(G)$. If L is the number of connected components of G, then the precise definition of $Comps(G)$ is provided as follows:

$$Comps(G) = \{C_1, C_2, \ldots, C_L\}.$$

We can now provide the formal definition of CNDP as follows.

2.2. CNDP and its genetic algorithm

Suppose B is a positive number where $B \leq |V|$. In CNDP, we are looking for a subset of nodes denoted by S, $S \subseteq V$ and $|S| \leq B$, whose removal mostly degrade the network according to the network connectivity metrics. Many criteria have been used in the literature to measure the network connectivity, the most important of which are the number of connected components, the size of the largest connected component, and the pairwise connectivity [22, 23, 24]. These three connectivity metrics are correlated; as the number of connected components increases, both the size of the largest connected component and the pairwise connectivity decrease [58]. Therefore, in this article, we focus on CNDP where the size of the largest connected component is considered as the connectivity metric. Using the notations above, CNDP can be defined as follows:

Problem 2.1 (CNDP). *Given graph G and positive number B, find a set $S^* \subseteq V$ such that $f(G[\overline{S^*}])$ is minimum:*

$$S^* = \underset{S \subseteq V,\ |S| \leq B}{\operatorname{argmin}} f(G[\overline{S}])$$

where the set of the nodes that we want to remove from G is illustrated with S, the subgraph induces by the remaining nodes ($u \in V \setminus S$) is shown by $G[\overline{S}]$, and $f(G)$ denotes the size of the largest component of G:

$$f(G) = \max_{C \in Comps(G)} |V(C)|.$$

Remark 2.2. *Based on the definition of CNDP, the optimal solution of the problem, S^*, may have B or less than B nodes ($|S^*| \leq B$) but it is simple to show that we have always an optimal solution S^* with $|S^*| = B$ [39]. To standardize further comparisons in this paper, hereafter we assume $|S| = B$ in the definition of CNDP.*

As we stated in the introduction section, there are many exact algorithms for providing optimal solutions of CNDP, but due to the NP-Hardness of the problem [39], exact approaches fail in solving CNDP in a tractable time. To the knowledge of the authors, the most powerful exact algorithm is capable of solving CNDP on graphs with up to 300 nodes [39]. Therefore, for real-world networks consisting of thousands to millions of nodes, the heuristic algorithms are remedial. In this regard, we prepare a genetic algorithm for CNDP to solve it on the large PPI networks.

Genetic algorithm is a type of iterative optimization algorithms to find the optimal solution(s) of a computational problem, first used by Holland and his colleagues in the 1960s. This algorithm can be considered as a branch of evolutionary algorithms, in that they are inspired by biological processes of reproduction and natural selection. The implementation of a genetic algorithm usually begins with the production of a population of "chromosomes" (each chromosome, a set of properties, is the genetic representation of a possible solution of the problem). The "initial population" of chromosomes in genetic algorithms is usually randomly generated. Then, using the two operators of

crossover and mutation, the members of the next generation are "reproduced". This process continues iteratively until the stopping condition of the algorithm is met and it reaches the optimal solution of the problem. In the generation-to-generation phase, each chromosome is evaluated using a "fitness function" and the chromosomes with higher values of the fitness function that better represent the "optimal solutions" of the problem have a better chance of reproduction (are selected with a higher probability to generate new chromosomes for the next generation). In the following we explain our genetic algorithm in more details corresponding to CNDP.

2.3. The proposed genetic algorithm on CNDP

The elements of the proposed genetic algorithm for solving CNDP are described as follows.

- **Genetic representation for a feasible solution of CNDP**

  A feasible solution of CNDP is a subset of nodes S, $S \subseteq V$ and $|S| = B$. Considering each node as a property, any feasible solution of CNDP is represented as a chromosome of B properties which we denote by ch (see Fig. 1).

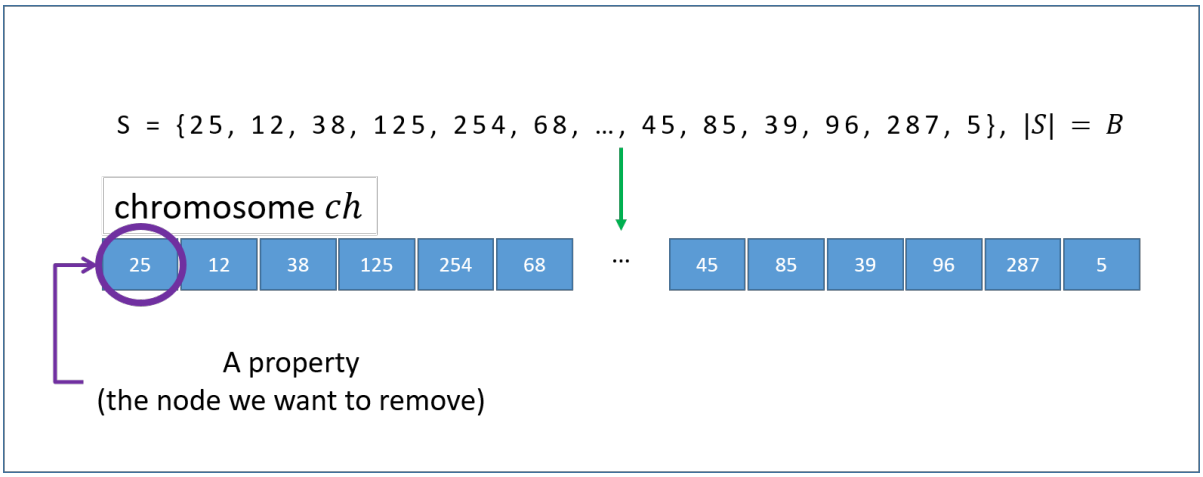


Figure 1: representing a feasible solution of CNDP as a chromosome

- **Fitness function**

  The fitness function for a chromosome ch is the size of the largest connected component of the network after removing the corresponding nodes of ch from the network. If we show the set of nodes in ch by $V(ch)$, then the fitness of ch is defined to be

$$fit(ch) = f(G[V \setminus V(ch)]).$$

- **Population**

  At each iteration t ($t = 1, 2, 3, \ldots$) of the genetic algorithm, we deal with a population of $P = 8000$ chromosomes where each chromosome represents a subset $S \subseteq V$, $|S| = B$. The population in iteration t is displayed using the notation pop_t (see Fig. 2) where

$$pop_t = \{ch_1, ch_2, \ldots, ch_P\}.$$

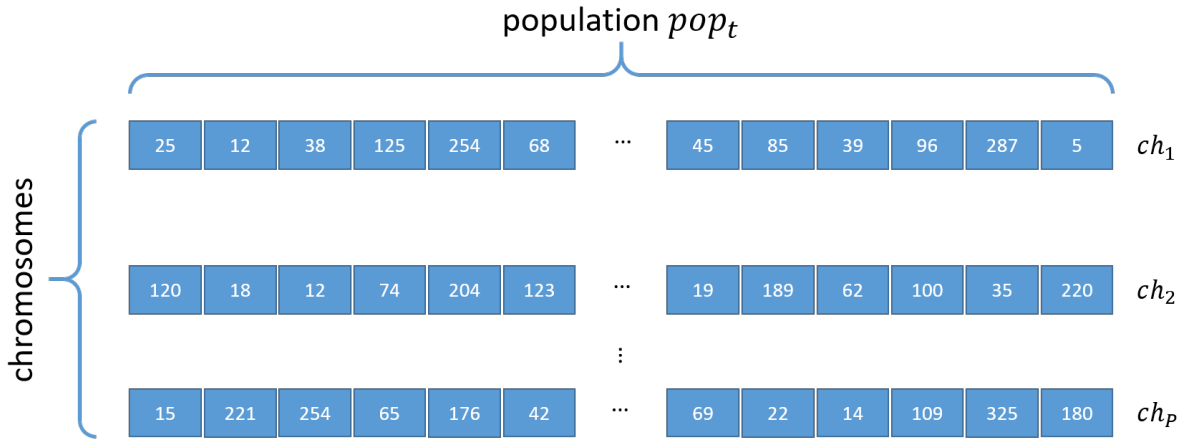


Figure 2: The population of chromosomes in the genetic algorithm

- **Initial population**

  At the starting point of the genetic algorithm (iteration $t = 1$), we create an initial population of P chromosomes where the nodes of each chromosome are selected from V, randomly with no duplicates.

- **Stopping condition**

  The algorithm terminates if its stopping condition is met. In this paper, the condition for termination of the algorithm is to achieve more than 10 consecutive iterations, with no improvement in the corresponding populations.

- **Creating the next generation**

  If the stopping condition of the genetic algorithm is not met at the current iteration, t, the algorithm goes to the next iteration, $t = t + 1$, and generates a new population of P chromosomes. The new population is generated as follows: 90 percents of the new population is generated by applying the crossover operator on the chromosomes of the current population. Two chromosomes from the current population are selected to produce two offsprings. This operation is repeated until a set of offsprings of size P is created of which 90 percents of the members with the best fitness function values are transferred to the new population. To make sure that our algorithm does not fall into a local optimum, the remaining 10 percents of the new population are chromosomes that are generated randomly. Once the members of the new population are generated, the mutation operator is applied on the whole population. The crossover and mutation operators are explained in more detail:

  1. **Crossover**
     To reproduce two new offspring (new solutions) using the crossover operator, two chromosomes must be selected as parents. Selecting the parents is based on the law of natural selection where better chromosomes in the current generation have a higher chance of crossover and reproduction. In the current genetic algorithm, we use the roulette wheel selection method to select the parents using the following scoring approach. For the ith chromosome (ch_i) in the current generation (pop_t), we consider the chance of being selected equal to
     $$\frac{|V| - fit(ch_i)}{fit_{total}}$$
     where $|V|$ denotes the number of nodes in G and fit_{total} equals $\sum_{ch \in pop_t} |V| - fit(ch)$. Once two parents are selected, the crossover operation is performed to produce two new offsprings. The crossover on two chromosomes X and Y is as follows (Fig. 3): Three loci on chromosomes X and Y (identical positions on both chromosomes) are chosen randomly which divide each parent into 4 segments. Then, the corresponding segments of the two parents are exchanged with the probability of 0.5. Note that in this way, the offsprings may be the same as parents when no segment is exchanged.

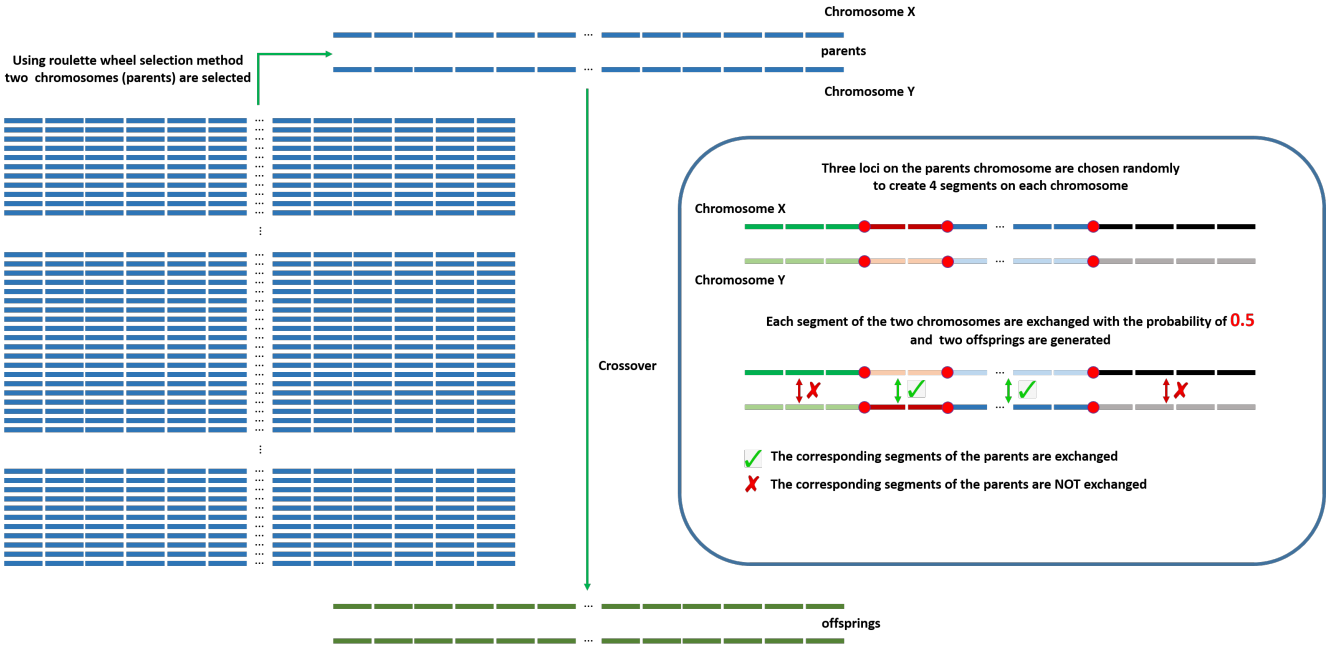


Figure 3: 90 percents of the population are generated from the reproduction by applying the crossover operator.

2. **Mutation**
   And, as a last operator of the genetic algorithm, the mutation operator is applied to each member of the new population. As we stated before, each chromosome is a representation of a subset of nodes S, $S \subseteq V$ and $|S| = B$. The mutation of node $u \in S$ is exchanging that node with a node $v \in V \setminus S$. For each chromosome, a random number M, between 1 and 100, is generated. Then, M nodes of the set S are selected randomly and mutated with the probability of 0.1.

Given the concepts above, the genetic algorithm for solving CNDP can be seen schematically in Figure 4.

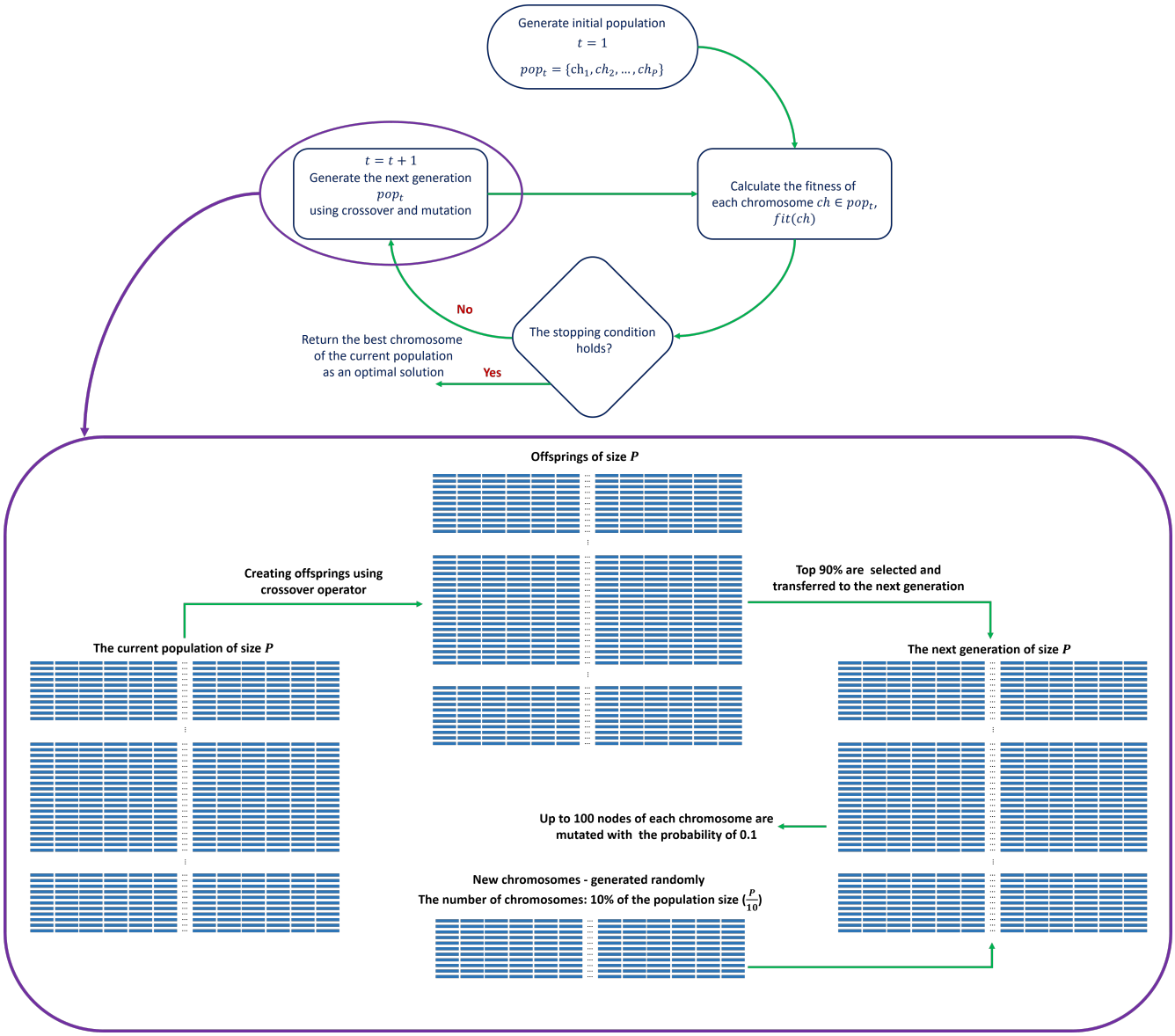


Figure 4: Schematic view of the genetic algorithm for solving CNDP.

By solving CNDP on a PPI network, we obtain B critical nodes as an optimal solution of CNDP whose removal mostly reduces the network connectivity. In order to investigate how the critical nodes affect the network stability, we need a benchmark. In this regard, we use centrality criteria to rank and sort the proteins of the PPI network based on each centrality and remove the B top highly-ranked nodes from the network to measure their effects on the network connectivity. Therefore, in the next subsection, the most frequently used centralities are described in detail.

2.4. Centralities

In this article, we use four very common criteria in analyzing different kinds of networks which are degree, betweenness, closeness, and eigenvector centralities [40, 41, 42, 43, 44]:

- Degree Centrality of node u, $DC(u)$, indicates the number of direct connections that u makes with other nodes of G.
- Betweenness Centrality of node u, $BC(u)$, represents the fraction of all the shortest paths of G which pass through u.
- Closeness Centrality of node u, $CC(u)$, measures the mean distance from u to other nodes of G where the standard measure of the distance between u and v is the length of the shortest path between the two nodes.

- Eigenvector Centrality of node u, $EC(u)$, is a more sophisticated view of centrality. In this criterion, the importance of node u is a function of the importance of its direct neighbors. If $N(u)$ is the set of adjacent nodes to u, then the $EC(u)$ is calculated from the following equation.

$$EC(u) = \frac{1}{\lambda} \sum_{t \in N(u)} EC(t) \tag{1}$$

Remark 2.3. *Beside the value of λ in the definition of EC, there is a fundamental question about the calculation of EC and it is how to calculate the importance of the first node of the graph. At first glance, calculating EC seems impossible. This problem can be solved with a simple trick. We show the adjacency matrix of G with A so that its elements a_{uv} indicate the presence/absence of an edge between the two nodes u and v. In other words, $a_{uv} = 1$ iff there is an edge between the nodes u and v. Now, the Eq. (1) can be rewritten as follows.*

$$EC(u) = \frac{1}{\lambda} \sum_{t \in V} a_{ut} EC(t) = \frac{1}{\lambda} \sum_{t \in V} a_{tu} EC(t)$$

$$\Downarrow$$

$$\begin{pmatrix} EC(1) \\ \vdots \\ EC(u) \\ \vdots \\ EC(n) \end{pmatrix} = \frac{1}{\lambda} \begin{pmatrix} a_{11}EC(1) + a_{12}EC(2) + \cdots + a_{1n}EC(n) \\ \vdots \\ a_{u1}EC(1) + a_{u2}EC(2) + \cdots + a_{un}EC(n) \\ \vdots \\ a_{n1}EC(1) + a_{n2}EC(2) + \cdots + a_{nn}EC(n) \end{pmatrix}$$

If we show $\begin{pmatrix} EC(1) \\ \vdots \\ EC(n) \end{pmatrix}$ *by X, then*

$$X = \frac{1}{\lambda} AX \Rightarrow \lambda X = AX.$$

In the above statement, if λ is one of the eigenvalues of matrix A, then X will be the corresponding eigenvector. If all the values of X are positive, it fulfills the conditions of eigenvector centrality definition.

In the next section, we apply the genetic algorithm to solve CNDP on PPI networks and then discuss the presence of essential proteins in the optimal solutions.

3. Result

For the experimental results, we implement the genetic algorithm as explained in section 2.3 using the Boost Graph Library [59] in Microsoft Visual Studio 2017. The algorithm is then applied to solve CNDP on the PPI networks of two species E. coli and S. cerevisiae extracted from the Database of Interacting Proteins (DIP) [56]. In addition, the performance of the algorithm in finding critical nodes and the presence of essential proteins in the set of critical nodes are explored. In the following, we first describe the data sets of protein-protein interactions of the two species and then investigate the role of essential proteins in the stability of the PPI networks. To measure the stability of the networks, we use the size of the largest connected component as the connectivity metric.

3.1. Data

We focus on the PPI networks of two species E. coli and S. cerevisiae which are extracted from DIP [56], a biological database that lists experimentally-identified interactions between proteins and combines the information from different sources to create a consistent set of protein-protein interactions. To label the proteins as essential/non-essential, the essential genes data of these species are collected from DEG database [49]. Further description of the data is as follows.

3.1.1. PPI networks

In the raw data of protein-protein interactions in DIP for E. coli, 12246 interactions between 2924 proteins have been reported. In Fig. 5, we illustrate a schematic view of the PPI network which is plotted using Cytoscape software [57]. The network in question is a fragmented network having a main component with several small islands. Since we are going to solve CNDP on this network, the small islands are of no interest and we remove them from the network such that only the main component of the network is remained. Furthermore, there exist some interactions which are actually self-loops and are removed from the data set. Therefore, the final network of E. coli consists of 11501 interactions between 2524 proteins. The same progress is made for clearing the PPI network of S. cerevisiae, leading to a final network of 22523 interactions between 5059 proteins.

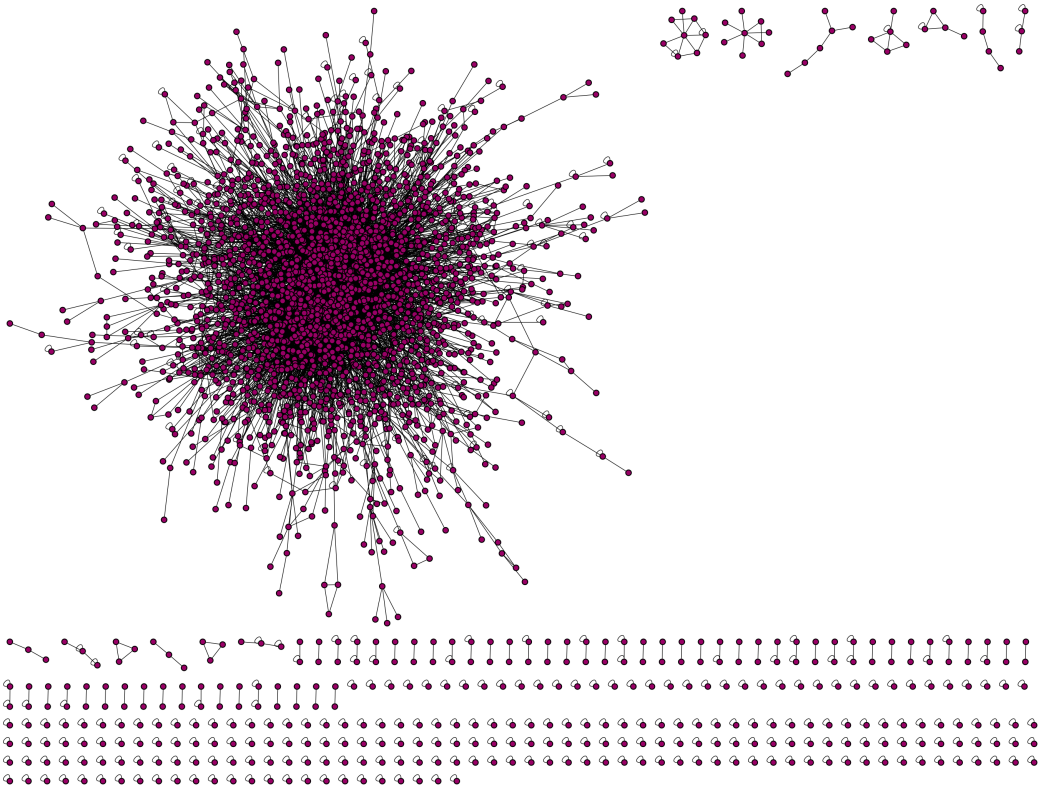


Figure 5: Schematic view of E. coli PPI network

3.1.2. Essential genes

The essential genes information are extracted from DEG database [49]. The data set of E. coli genes essentiality (being essential or non-essential) includes 11888 experiment reports on 4323 genes. In this data set, there is only one report for a fraction of genes, more than one experiment has been reported for some of the genes and also there is no information reported for the remaining genes of E. coli. For genes with more than one report, if the number of reports about their essentiality (non-essentiality) is higher than the number of reports about their non-essentiality (essentiality), those genes are considered as essentials (non-essentials). If the number of reports on essentiality and non-essentiality of a gene is equal, we consider the essentiality status of that gene as unknown. In that way, we have 306 essential genes, 3987 non-essential genes and 30 genes with an unclear status in the data set extracted from DEG for E. coli. DEG has provided the genes with STRING ids while in DIP the main used ids are UniProtKB ids. Therefore, to label proteins in the PPI networks as essential or non-essential and also as unknown, we use Uniprot ID Mapping [60] to map DIP ids to STRING ids. The status of proteins for which there is no corresponding gene with reports about its essentiality in DEG, are also considered to be unknown. In this respect, of 2524 proteins in E. coli PPI network, 371 and 2044 proteins are essential and non-essential, respectively and there are 109 proteins with the unknown label. Table 1 gives an overview of the information about E. coli and S. cerevisiae PPI networks and the essentiality status of their proteins. It should be noted that the PPI network of each of these two species is the main component of the raw PPI network where small fragmented islands are removed.

Table 1: E.coli and S. cerevisiae PPI networks and their proteins essentiality information

	interactions	proteins	essential proteins	non-essential proteins	unknowns
E. coli	11501	2524	371 (15%)	2044	109
S. cerevisiae	22523	5059	967 (19%)	3500	592

3.2. Experiments and results

The experiments in this paper are performed in two parts. In the first part, we apply the genetic algorithm to identify a set of critical nodes and investigate the effect of critical node removal on the network stability compared with removing nodes based on different centralities. In the second part, the participation of essential proteins in the set of critical nodes, which indicates the effectiveness of essential proteins in the network stability, is explained. By considering the set of critical nodes as central elements of the network, these results confirm the centrality-lethality rule.

3.2.1. applying the genetic algorithm on PPI networks

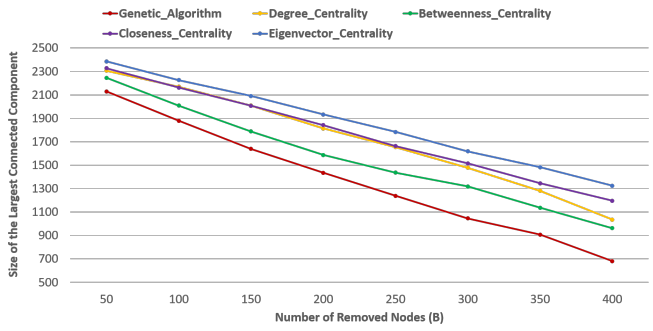
In this part, we first apply the genetic algorithm on the PPI networks of E. coli and S. cerevisiae for different values of B to find a set of B critical nodes in each network. To ensure that the results are reliable and robust, we repeat the genetic algorithm on the PPI network of E. coli three times.

Remark 3.1. *Since E. coli and S. cerevisiae PPI networks contain 371 and 967 essential proteins, respectively, this part of the experiment is conducted for* $B = 50, 100, \ldots 400$ *in E. coli and for* $B = 50, 100, \ldots, 950$ *in S. cerevisiae PPI networks.*

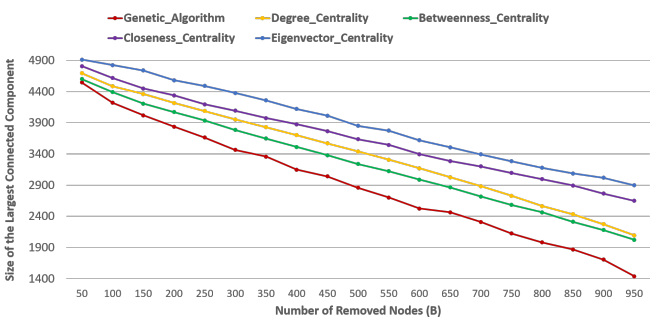
To evaluate the performance of the genetic algorithm, the sets of B top highly-ranked nodes obtained based on degree, betweenness, closeness, and eigenvector centralities are also considered as follows. Consider the degree centrality as the example. We calculate the degree centrality for each node of G (using the Python package Networkx [64]) and sort the nodes in ascending order of the calculated values, and then the B top highly-ranked nodes are chosen.

Remark 3.2. *For the simplicity of the explanation, hereafter, we use the notation* $S_{(B,GA)}$ *to show the set of* B *critical nodes returned by the genetic algorithm. Furthermore,* $S_{(B,DC)}, S_{(B,BC)}, S_{(B,CC)}$, *and* $S_{(B,EC)}$ *are used to show the* B *top highly-ranked nodes obtained based on degree, betweenness, closeness, and eigenvector centralities, respectively.*

We plot $f(G[\overline{S_{(B,GA)}}])$, $f(G[\overline{S_{(B,DC)}}])$, $f(G[\overline{S_{(B,BC)}}])$, $f(G[\overline{S_{(B,CC)}}])$, and $f(G[\overline{S_{(B,EC)}}])$ on both E. coli and S. cerevisiae PPI networks and for different values of B. The level of destruction in the network structure (the reduction in the size of the largest connected component) caused by the removal of critical nodes, as depicted in Figs. 6a and 6b, well determines the genetic algorithm efficiency in finding important nodes.



(a) Node removal on the PPI network of E. coli.



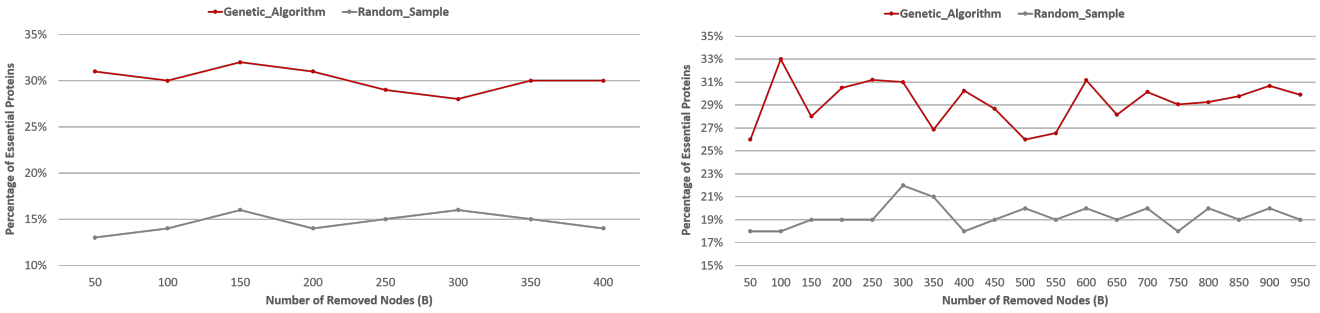
(b) Node removal on the PPI network of S. cerevisiae.

Figure 6: The effect of node removal on the largest connected component of PPI networks: critical node removal vs. node removal based on degree, betweenness, closeness, and eigenvector centralities. Figs. 6a and 6b depict the results on E. coli and S. cerevisiae PPI networks, respectively.

In the next part of the experiment, we analyze how essential proteins are represented in the set of critical nodes found by the genetic algorithm, and also provide evidence to show that the topological properties of the essential proteins represented in the set of critical nodes are different from the topological properties of the essential proteins recovered by the centrality-based methods.

3.2.2. Indicating the effectiveness of essential proteins in network stability

At the first step, we create five random samples of size B, for different values of B, and determine the percentage of essential proteins observed in the random samples. Then, this result is compared to the portion of essential proteins in the set of critical nodes, of size B, obtained in the previous part. Figs. 7a and 7b confirm that the percentage of essential proteins in the set of critical nodes is greater than what is expected to be observed by chance, indicating that essential proteins have a meaningful effect on the PPI network stability. If we call the set of critical nodes as the set of central nodes in the network, this result confirms the centrality-lethality rule from another perspective. Meghanathan et al. and Oldham et al. have conducted two separate studies and showed that the degree, betweenness, closeness and eigenvector centralities are to some extent correlated [62, 63]. We just have to show that the essential proteins covered by our approach are different from the essential proteins which are represented in the sets $S_{(B,DC)}$, $S_{(B,BC)}$, $S_{(B,CC)}$, and $S_{(B,EC)}$.



(a) Results on the PPI network of E. coli.

(b) Results on the PPI network of S. cerevisiae.

Figure 7: For different values of B, the percentage of essential proteins in the critical nodes found by the genetic algorithm is displayed compared with the presence of essential proteins in random samples of size B. Figs. 7a and 7b depict the results on E. coli and S. cerevisiae PPI networks, respectively.

At the second step, we investigate the topological properties of essential proteins contributing to the set of critical nodes of size B compared with the topological properties of the essential proteins represented in the set of B highly-ranked proteins for each centrality criterion.
In this regard, we consider the mentioned four centrality criteria as reflections of node topological properties. For example, the degree centrality value of a node is one of its topological properties. We consider all proteins of S. cerevisiae PPI network and calculate their degree centrality values. Then, the probability density functions (PDF) for the degree centrality values of the essential proteins in $S_{(B,GA)}$ ($B = 50, 100, 150,$ and 200) and the essential proteins in $S_{(B,DC)}$ are estimated using the class "gaussian-kde" in Python Statistical package "scipy.stats" [1] as follows. The degree centrality values of essential proteins in $S_{(B,GA)}$ are stored in a vector X and fed into "gaussian-kde" as univariate data points to estimate the PDF. The same procedure is done for the essential proteins represented in $S_{(B,DC)}$. To make the comparison more meaningful, we also estimate the PDF for the degree centrality values of all essential proteins of the PPI network through the same procedure. Fig. 8 depicts the three estimated PDFs and can be considered as a confirmation for the difference between the topological properties of essential proteins in $S_{(B,GA)}$ and the topological properties of essential proteins in $S_{(B,DC)}$. To provide more details, we conduct four sub-experiments in which degree, betweenness, closeness, and eigenvector centralities are used to reflect topological properties of essential proteins included in the set of critical nodes and the sets of highly-ranked proteins.

[1] https://docs.scipy.org/doc/scipy/reference/stats.html

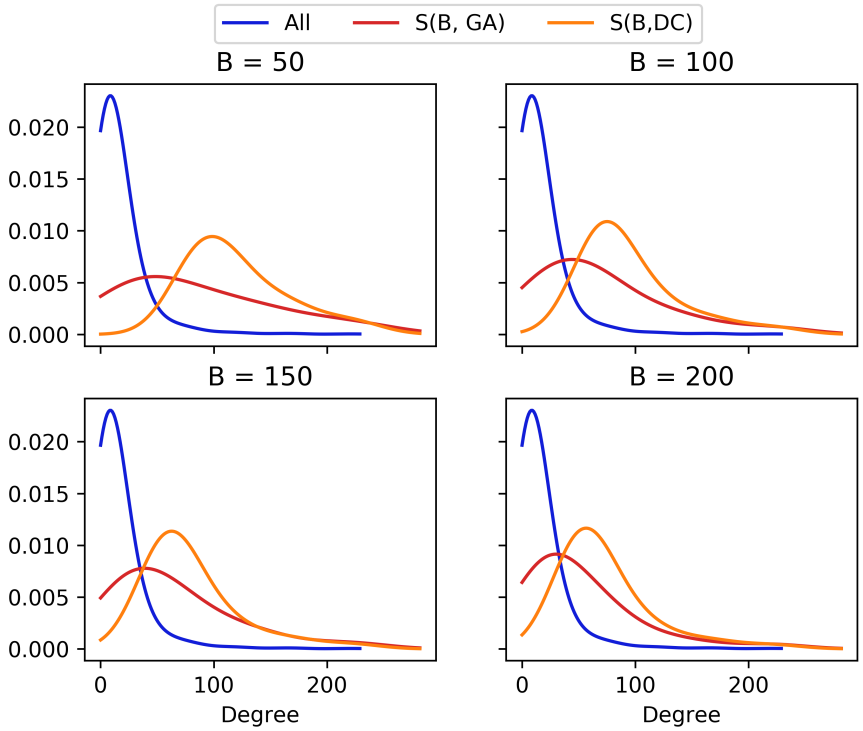
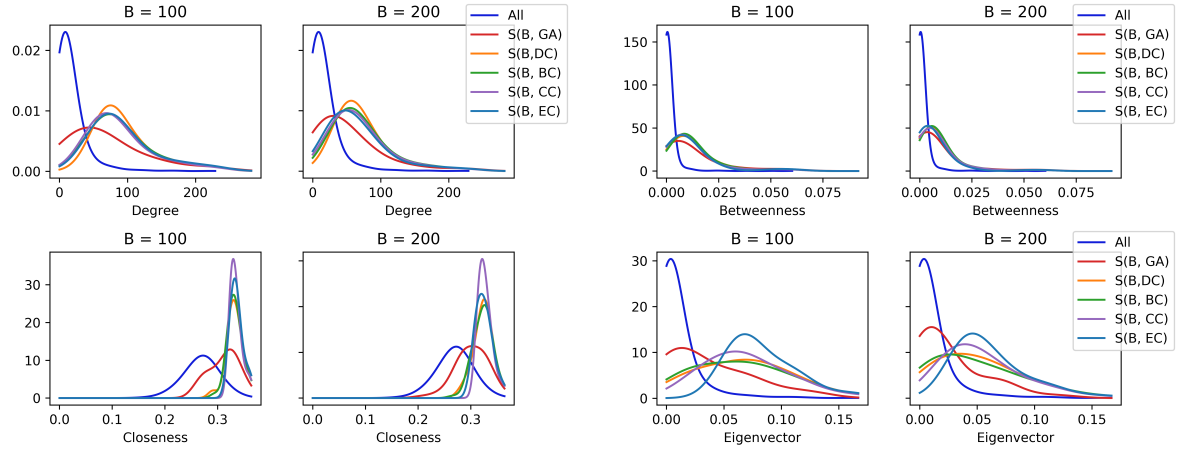


Figure 8: The probability density function (PDF) for degree centrality value of **all essential proteins** in the PPI network of S. cerevisiae (**All**), along with the PDFs for degree centrality value of **essential proteins** represented in $S_{(B,GA)}$ and $S_{(B,DC)}$.

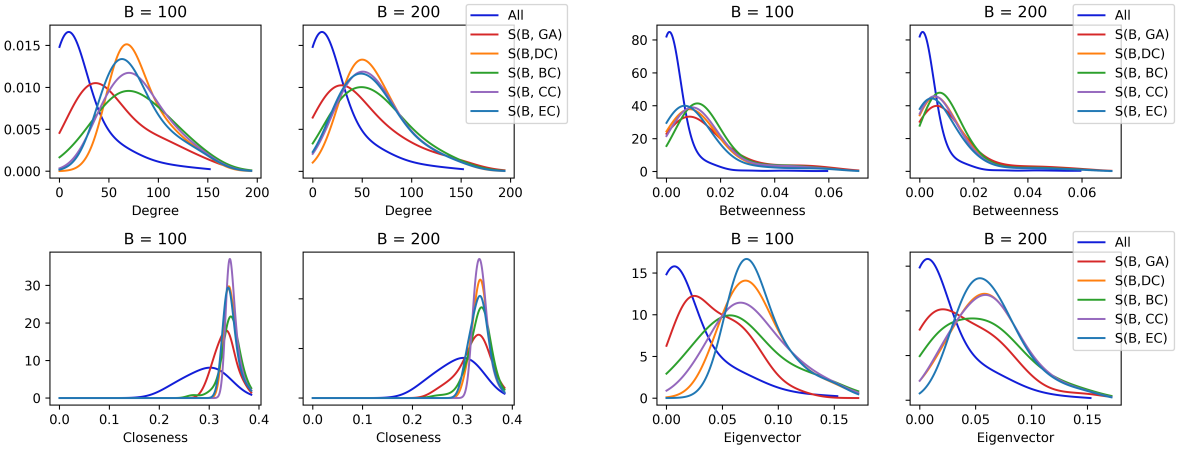
In the four sub-experiments, the degree, betweenness, closeness, and eigenvector centrality values of all proteins of the S. cerevisiae and E. coli PPI networks are calculated. Then, for each centrality measure and PPI network, the PDF of all essential proteins of the network along with the PDFs of essential proteins of $S_{(B,GA)}$, $S_{(B,DC)}$, $S_{(B,BC)}$, $S_{(B,CC)}$, and $S_{(B,EC)}$ are estimated ($B = 100$ and 200). Figs. 9a and 9b show the mentioned PDFs in S. cerevisiae and E. coli PPI networks, respectively.

The results show that the essential proteins covered by the genetic algorithm have a variety of degree, closeness, and eigenvector values, and are therefore distinct from the essential proteins found by the ranking-based methods in which almost essential proteins presented in the set of B top highly-ranked nodes share a similar distribution of degree, closeness and eigenvector values. Furthermore, the distribution of each centrality measure in the set of essential proteins of $S_{(B,GA)}$ tends to be similar to the distribution of all essential proteins of the corresponding species which confirms the fact that by applying CNDP, we are covering some essential proteins which other centrality-based methods are unable to recover. The only centrality measure in which the essential proteins of $S_{(B,GA)}, S_{(B,DC)}, S_{(B,BC)}, S_{(B,CC)}$, and $S_{(B,EC)}$ behave similarly is the betweenness measure. However, in the corresponding PDFs, there is a slight difference between the essential proteins of $S_{(B,GA)}$ and the essential proteins of the other sets of B top highly-ranked proteins.

In addition to the PDFs displayed in Fig. 9, we use box plot to show the distribution of degree, betweenness, closeness and eigenvector values for the set of essential proteins of $S_{(B,GA)}, S_{(B,DC)}, S_{(B,BC)}, S_{(B,CC)}$, and $S_{(B,EC)}$ in S. cerevisiae and E. coli which are displayed in Figs. 10a and 10b, respectively. Each plot corresponds to one centrality measure as a reflection of a topological property. For the set of essential proteins in the $S_{(B,DC)}, S_{(B,BC)}, S_{(B,CC)}$, and $S_{(B,EC)}$, we consider their median values according to the corresponding centrality measure of that plot and draw a blue horizontal line which shows the average value of the medians. This line intends to show the similarity between the distributions of topological properties of essential proteins obtained by the centrality-based methods. Similar to what is seen in PDFs, these figures also indicate a different distribution of the topological properties of essential proteins in $S_{(B,GA)}$ (for degree, closeness and eigenvector measures).

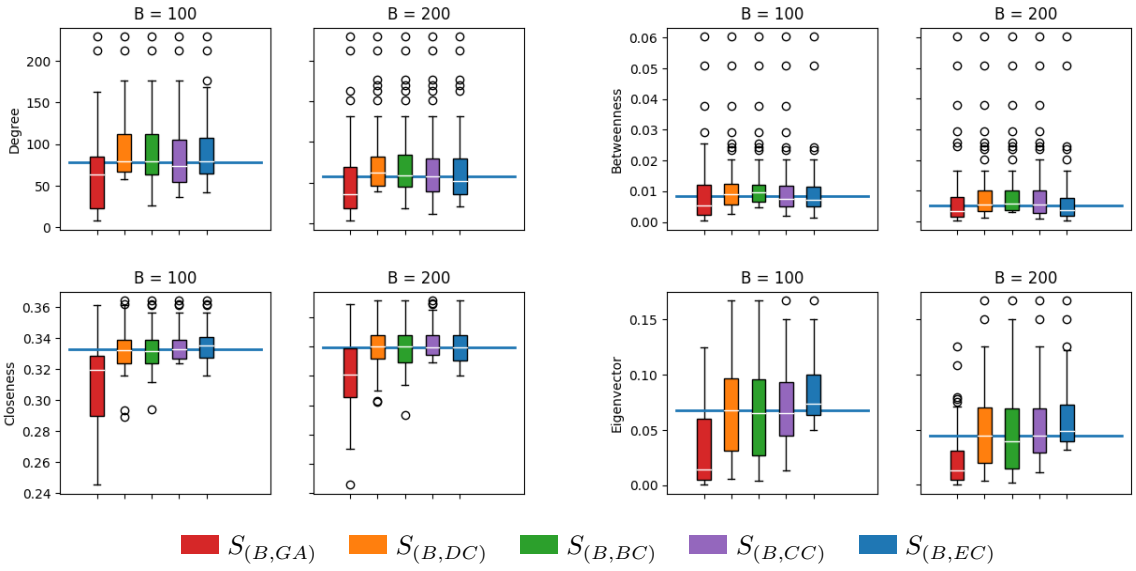


(a) Results on the PPI network of S. cerevisiae.

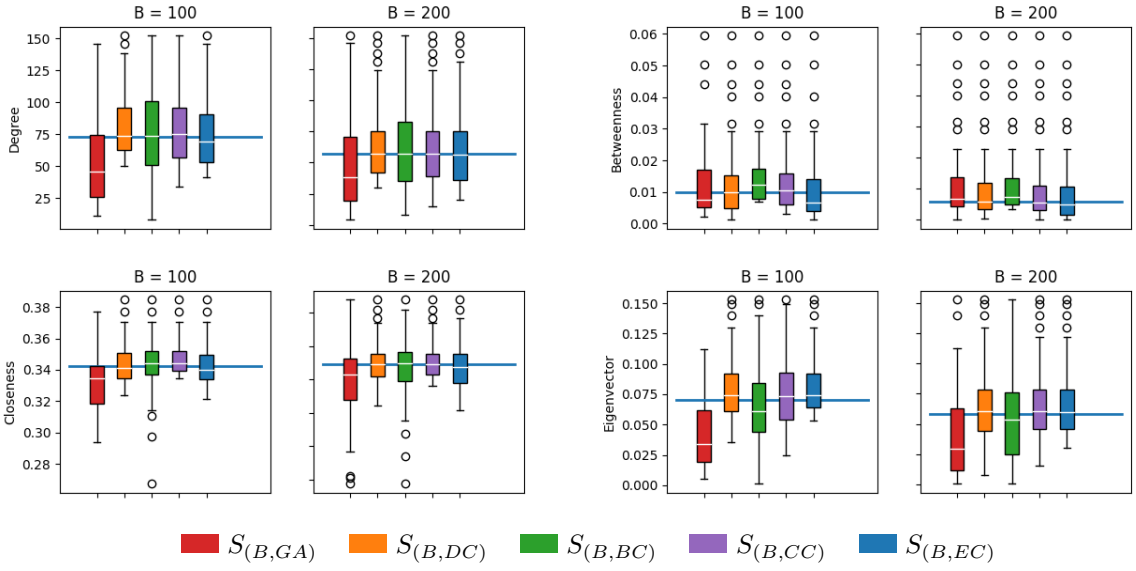


(b) Results on the PPI network of E. coli.

Figure 9: Part (a) depicts the probability density functions (PDFs) for degree, betweenness, closeness and eigenvector centrality value of **all essential proteins** in the PPI network of S. cerevisiae and **essential proteins** represented in $S_{(B,GA)}$, $S_{(B,DC)}$, $S_{(B,BC)}$, $S_{(B,CC)}$, and $S_{(B,EC)}$ ($B = 100$ and 200). Part (b) of the figure displays the same charts for the PPI network of E. coli.



(a) Results on the PPI network of S. cerevisiae.



(b) Results on the PPI network of E. coli.

Figure 10: Part (a) depicts the box plots for degree, betweenness, closeness and eigenvector centrality value of **essential proteins** represented in $S_{(B,GA)}$, $S_{(B,DC)}$, $S_{(B,BC)}$, $S_{(B,CC)}$, and $S_{(B,EC)}$ for the PPI network of S. cerevisiae ($B = 100$ and 200). Part (b) of the figure displays the same charts for the PPI network of E. coli. The blue horizontal line inside each plot, is the average of the medians in four groups $S_{(B,DC)}$, $S_{(B,BC)}$, $S_{(B,CC)}$, and $S_{(B,EC)}$.

4. Conclusion

An interesting challenge in analyzing protein-protein interaction (PPI) networks is to demonstrate the correlation between the topological importance of a protein and its essentiality. Previous studies use the centrality criteria to show such a correlation. Although they provide good insight on this issue but due to the scale-free property of PPI networks, all the criteria have high correlations and therefore, recover similar subsets of essential proteins. In this paper, we use a well-known problem in the field of graph theory, critical node detection problem (CNDP), and solve it on the PPI networks of two species E. coli and S. cerevisiae to cover a set of different essential proteins. In CNDP the aim is to find the set of critical nodes whose removal from the network most profoundly reduces its stability (connectivity). For the network connectivity metric, we consider the size of the largest connected component of the network. We implement a genetic algorithm to solve CNDP and find the critical nodes of the PPI networks. The results well show the efficiency of the genetic algorithm in solving CNDP. Finally, we measure the presence of

essential proteins in the resulting set of critical nodes. The results show that the presence of essential proteins in these sets is more than their presence in random samples, which indicates a significant role of essential proteins on the stability of the PPI network. Furthermore, the essential proteins represented in the set of critical nodes, have different topological properties compared with the essential proteins recovered by the centrality-based methods. For future research, the enrichment analysis of the critical nodes based on Gene Ontology will indicate how Gene Ontology terms such as antibiotic resistance are enriched in the set of critical nodes. Furthermore, PPI networks can be weighted using biological features to make the networks components more biologically meaningful.

References

[1] X. Tang, J. Wang, J. Zhong, Y. Pan, Predicting essential proteins based on weighted degree centrality, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11(2) (2013), 407-418.

[2] J. Wang, W. Peng, F. X. Wu, Computational approaches to predicting essential proteins: a survey, PROTEOMICS–Clinical Applications, 7(1-2) (2013), 181-192.

[3] J. Cheng, W. Wu, Y. Zhang, X. Li, X. Jiang, G. Wei, S. Tao, A new computational strategy for predicting essential genes, BMC genomics, 14(1) (2013), 1-13.

[4] L. Yang, J. Wang, H. Wang, Y. Lv, Y. Zuo, W. Jiang, Characterization of essential genes by topological properties in the perturbation sensitivity network, Biochemical and biophysical research communications, 448(4) (2014), 473-479.

[5] J. L. Snoep, H. V. Westerhoff, From isolation to integration, a systems biology approach for building the Silicon Cell, In Systems biology (pp. 13-30), Springer, Berlin, Heidelberg, 2005.

[6] E. Alm, A. P. Arkin, Biological networks, Current opinion in structural biology, 13(2) (2003), 193-202.

[7] S. Rasti, C. Vogiatzis, A survey of computational methods in protein–protein interaction networks, Annals of Operations Research, 276(1) (2019), 35-87.

[8] M. Ashtiani, A. Salehzadeh-Yazdi, Z. Razaghi-Moghadam, H. Hennig, O. Wolkenhauer, M. Mirzaie, M. Jafari, A systematic survey of centrality measures for protein-protein interaction networks, BMC systems biology, 12(1) (2018), 1-17.

[9] X. Zhao, Z. P. Liu, Analysis of topological parameters of complex disease genes reveals the importance of location in a biomolecular network, Genes, 10(2) (2019), 143.

[10] N. Matas, Comparing Network Centrality Measures as Tools for Identifying Key Concepts in Complex Networks: A Case of Wikipedia, Journal of Digital Information Management, 15(4) (2017).

[11] Y. Y. Liu, J. J. Slotine, A. L. Barabási, Control centrality and hierarchical structure in complex networks, (2012).

[12] S. Iyer, T. Killingback, B. Sundaram, Z. Wang, Attack robustness and centrality of complex networks, PloS one, 8(4) (2013) , e59613.

[13] H. Jeong, S. P. Mason, A. L. Barabási, Z. N. Oltvai, Lethality and centrality in protein networks, Nature, 411(6833) (2001), 41-42.

[14] N. N. Batada, L. D. Hurst, M. Tyers, Evolutionary and physiological importance of hub proteins, PLoS computational biology, 2(7) (2006), e88.

[15] M. W. Hahn, A. D. Kern, Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks, Molecular biology and evolution, 22(4) (2005), 803-806.

[16] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu, M. Gerstein, Genomic analysis of essentiality within protein networks, Trends in Genetics, 20(6) (2004), 227-231.

[17] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics, PLoS computational biology, 3(4) (2007), e59.

[18] E. Estrada, Virtual identification of essential proteins within the protein interaction network of yeast, Proteomics, 6(1) (2006), 35-40.

[19] E. Zotenko, J. Mestre, D. P. O'Leary, T. M. Przytycka, Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality, PLoS computational biology, 4(8) (2008), e1000140.

[20] M. Altaf-Ul-Amin, S. H. Wijaya, D. F. Chandra, S. Kanaya, Centrality Values of Yeast Proteins in a PPI Network Are Related to Their Essentiality and Functions, Journal of Computer Aided Chemistry, 18 (2017), 94-109.

[21] F. S. Zaidi, U. Fatima, B. A. Usmani, A. R. Jafri, Comprehending Nodes Essentiality through Centrality Measures in Biological Networks, IJCSNS, 19(9) (2019), 65.

[22] A. Arulselvan, C. W. Commander, L. Elefteriadou, P. M. Pardalos, Detecting critical nodes in sparse graphs, Computers and Operations Research, 36(7) (2009), 2193-2200.

[23] S. Shen, J. C. Smith, R. Goli, Exact interdiction models and algorithms for disconnecting networks via node deletions, Discrete Optimization, 9(3) (2012), 172-188.

[24] A. Veremyev, V. Boginski, E. L. Pasiliao, Exact identification of critical nodes in sparse networks via new compact formulations, Optimization Letters, 8(4) (2014), 1245-1259.

[25] A. E. Motter, Y. C. Lai, Cascade-based attacks on complex networks, Physical Review E, 66(6) (2002), 065102.

[26] T. Nie, Z. Guo, K. Zhao, Z. M. Lu, New attack strategies for complex networks, Physica A: Statistical Mechanics and its Applications, 424 (2015), 248-253.

[27] B. Duan, J. Liu, M. Zhou, L. Ma, A comparative analysis of network robustness against different link attacks. Physica A: Statistical Mechanics and its Applications, 448 (2016), 144-153.

[28] S. W. SSun, Y. L. Ma, R. Q. Li, L. Wang, C. Y. Xia, Tabu search enhances network robustness under targeted attacks, Physica A: Statistical Mechanics and its Applications, 446 (2016), 82-91.

[29] S. P. Borgatti, Identifying sets of key players in a social network, Computational and Mathematical Organization Theory, 12(1) (2006), 21-34.

[30] M. Oosten, J. H. Rutten, F. C. Spieksma, Disconnecting graphs by removing vertices: a polyhedral approach, Statistica Neerlandica, 61(1) (2007), 35-60.

[31] T. N. Dinh, M. T. Thai, H. T. Nguyen, Bound and exact methods for assessing link vulnerability in complex networks, Journal of Combinatorial Optimization, 28(1) (2014), 3-24.

[32] R. Aringhieri, A. Grosso, P. Hosteins, R. Scatamacchia, A general evolutionary framework for different classes of critical node problems, Engineering Applications of Artificial Intelligence, 55 (2016), 128-145.

[33] Z. Zhu, Discovering the influential users oriented to viral marketing based on online social networks, Physica A: Statistical Mechanics and its Applications, 392(16) (2013), 3459-3469.

[34] M. Di Summa, A. Grosso, M. Locatelli, Complexity of the critical node problem over trees, Computers and Operations Research, 38(12) (2011), 1766-1774.

[35] M. Ventresca, D. Aleman, A randomized algorithm with local search for containment of pandemic disease spread, Computers and operations research, 48 (2014), 11-19.

[36] M. Lalou, M. A. Tahraoui, H. Kheddouci, The critical node detection problem in networks: A survey, Computer Science Review, 28 (2018), 92-117.

[37] R. Albert, Network inference, analysis, and modeling in systems biology, The Plant Cell, 19(11) (2007), 3327-3338.

[38] C. Wierling, T. Kessler, L. A. Ogilvie, B. M. Lange, M. L. Yaspo, H. Lehrach, Network and systems biology: essential steps in virtualising drug discovery and development, Drug Discovery Today: Technologies, 15 (2015), 33-40.

[39] J. Rezaei, F. Zare-Mirakabad, S. A. MirHassani, S. A. Marashi, EIA-CNDP: An exact iterative algorithm for critical node detection problem, Computers and Operations Research, 127 (2021), 105138.

[40] C. Y. Chen, J. J. Huang, A Novel Centrality for Finding Key Persons in a Social Network by the Bi-Directional Influence Map, Symmetry, 12(10) (2020), 1747.

[41] M. Ambriz-Rivas, N. Pastor, G. del Rio, Relating Protein Structure and Function Through a Bijection and Its Implications on Protein Structure Prediction, Protein Interactions, 349 (2012).

[42] E. Zajmi, The Ottoman Period in Albanian Historiography (1915-2015) (2018).

[43] F. Grando, Methods for the approximation of network centrality measures, (2018).

[44] H. Buchner, Displaying centrality of a network using orbital layout, Master's Thesis, University of Passau/National ICT Sydney, (2006).

[45] D. Vella, S. Marini, F. Vitali, D. Di Silvestre, G. Mauri, R. Bellazzi, MTGO: PPI network analysis via topological and functional module identification, Scientific reports, 8(1) (2018), 1-13.

[46] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, ... , C. V. Mering, STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, Nucleic acids research, 47(D1) (2019), D607-D613.

[47] W. H. Chen, G. Lu, X. Chen, X. M. Zhao, P. Bork, OGEE v2: an update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines, Nucleic acids research, gkw1013, (2016).

[48] S. Gurumayum, P. Jiang, X. Hao, T. L. Campos, N. D. Young, P. K. Korhonen, ..., W. H. Chen, OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines, Nucleic Acids Research, 49(D1) (2021), D998-D1003.

[49] H. Luo, Y. Lin, F. Gao, C. T. Zhang, R. Zhang, DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements, Nucleic acids research, 42(D1) (2014), D574-D580.

[50] T. Murali, S. Pacifico, J. Yu, S. Guest, G. G. Roberts, R. L. Finley, DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila, Nucleic acids research, 39(suppl_1) (2011), D736-D743.

[51] J. Yu, S. Pacifico, G. Liu, R. L. Finley, DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions, BMC genomics, 9(1) (2008), 1-9.

[52] S. Orchard, M. Ammari, B. Aranda, L. Breuza, L. Briganti, F. Broackes-Carter, ..., H. Hermjakob, The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases, Nucleic acids research, 42(D1) (2014), D358-D363.

[53] Y. López, K. Nakai, A. Patil, HitPredict version 4: comprehensive reliability scoring of physical protein–protein interactions from more than 100 species, Database, (2015).

[54] R. Oughtred, C. Stark, B. J. Breitkreutz, J. Rust, L. Boucher, C. Chang, ..., M. Tyers, The BioGRID interaction database: 2019 update, Nucleic acids research, 47(D1) (2019), D529-D541.

[55] O. Aromolaran, T. Beder, M. Oswald, J. Oyelade, E. Adebiyi, R. Koenig, Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features, Computational and structural biotechnology journal, 18 (2020), 612-621.

[56] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, D. Eisenberg, The database of interacting proteins: 2004 update, Nucleic acids research, 32(suppl_1) (2004), D449-D451.

[57] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, ..., T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome research, 13(11) (2003), 2498-2504.

[58] L. Faramondi, G. Oliva, R. Setola, F. Pascucci, A. E. Amideo, M. P. Scaparra, Performance analysis of single and multi-objective approaches for the critical node detection problem, In International Conference on Optimization and Decision Science (pp. 315-324), (2017) Springer, Cham.

[59] J. G. Siek, L. Q. Lee, A. Lumsdaine, The boost graph library: User guide and reference manual, portable documents, Pearson Education, (2001).

[60] S. Pundir, M. J. Martin, C. O'Donovan, UniProt Consortium, UniProt tools, Current protocols in bioinformatics, 53(1) (2016), 1-29.

[61] S. Gündüç, R. Eryiğit, Time dependent correlations between the probability of a node being infected and its centrality measures, Physica A: Statistical Mechanics and its Applications, 563 (2021), 125483.

[62] N. Meghanathan, Correlation coefficient analysis of centrality metrics for complex network graphs, In Computer Science On-line Conference (pp. 11-20) (2015). Springer, Cham.

[63] S. Oldham, B. Fulcher, L. Parkes, A. Arnatkeviciute, C. Suo, A. Fornito, Consistency and differences between centrality measures across distinct classes of networks, PloS one, 14(7) (2019), e0220061.

[64] D. A. Schult, P. Swart, Exploring network structure, dynamics, and function using NetworkX, In Proceedings of the 7th Python in science conferences (SciPy 2008) (Vol. 2008, pp. 11-16).

Please cite this article using:

Javad Rezaei, Fatemeh Zare-Mirakabad, Sayed-Amir Marashi, Seyed Ali MirHassani, The assessment of essential genes in the stability of PPI networks using critical node detection problem, AUT J. Math. Comput., 3(1) (2022) 59-76
DOI: 10.22060/AJMC.2021.20101.1053