Evolutionary Graph-Clustering vs Evolutionary Cluster-Detection Approaches for Community Identification in PPI Networks

Marwa Ben M'barek, Sana Ben Hmida, Amel Borgi, Marta Rukoz

Evolutionary Graph-Clustering vs Evolutionary Cluster-Detection Approaches for Community Identification in PPI Networks

Marwa Ben M'Barek[1,2], Sana Ben Hmida[2], Amel Borgi[1], and Marta Rukoz[2]

[1] LIPAH, Faculté des Sciences de Tunis, Université de Tunis El Manar,2092, Tunis,Tunisia
[2] LAMSADE CNRS UMR 7243, Paris Dauphine University, PSL Research University, Place du Maréchal de Lattre de Tassigny, Paris, France

Abstract. Community detection in protein-protein interaction networks (PPIs) is an active area of research, and many studies have applied Genetic Algorithms (GAs) to this problem. This paper summarizes the different GA based approaches for community detection in PPIs and provides a taxonomy of these methods. Detailed comparative studies are then provided comparing an evolutionary graph-clustering approach (EGCPI) based on the partitioning paradigm and an evolutionary cluster-detection approach based on an evolutive and incremental search for potential communities in the graphs (GA-PPI-Net). The communities obtained by the two algorithms on Collins PPI network are compared according to the average similarity and interaction between genes, and also according to the recovery rate of known communities in some biological pathways. Experiments tests verify the effectiveness of the GA-PPI-Net approach compared with EGCPI approach.

Keywords: Community detection · Biological networks · PPI networks · Genetic Algorithm · GA-PPI-Net · EGCPI

1 Introduction

Networks can represent many systems such as biology, computer science, linguistics, etc. A network has a set of nodes and a set of edges, it is represented by a graph. The nodes represent the basic components of the system, and the edges represent the links between nodes according to a defined relationship [20]. For example, a network of interactions between proteins is generally represented as an interaction graph, where nodes represent proteins and edges represent pairwise interactions. When a system is modeled by a network, it helps to understand the system easily and identify hidden information. Lots of studies have done around networks and how to use them.

Networks have some common features. The most known feature is the existence of parts, or sub-graphs, more densely connected than others. These parts, which are a set of nodes and edges, are called communities. So, the community can be defined as a group of nodes much more strongly connected to each other

than other nodes [24]. These groups, or communities, are usually thought to correspond to groups of nodes that play similar roles or have similar functions within the network.

Communities detection (CD) in networks has become a new field of research [2, 10, 20, 31]. The goal of CD is to identify these groups in a way that is both meaningful and useful for understanding the structure and function of the network. Applications of CD include social network analysis, biology, and computer science, among others. In biological networks, for example, CD has been used to identify groups of genes that are involved in similar biological processes, or groups of proteins that interact to form functional complexes. In social networks, CD has been used to identify groups of individuals with similar interests or social roles.

The purpose of CD in protein-protein interaction networks (PPIs) is to better understand the biological mechanisms and pathways that underlie cellular processes. By identifying groups of proteins that are more likely to interact with each other, CD can help to identify functional modules or pathways within the network, and can provide a starting point for further experimental investigation. CD can also be used for network-based drug discovery and personalized medicine. By identifying communities of proteins that are involved in specific diseases, it may be possible to identify new drug targets or develop personalized treatments based on an individual's unique network structure. This work is multidisciplinary, as it brings the field of biology and computer science in the broad sense.

The community detection can be handled either by heuristic based methods or optimization based methods. Heuristic based algorithms use essentially the edge-betweenness [20]: its value is higher for inter-communities than for intracommunities. Then, communities are identified by sequentially removing edges that do not increase the edge-betweenness for inter-communities. The Optimization based methods maximize an objective function often computed from structural information in the network, such as the modularity (cf paragraph 2). However, computing such function for all possible partitions of the network is a NP-hard problem. In this context, Genetic Algorithms (GAs) have been considered as suitable approaches to obtain an optimal result [8]. Most of the proposed algorithms use the greedy or the partitioning paradigms. EGCPI (Evolutionary Game-based Community detection algorithm for Protein-protein Interaction networks) [11] is a known method in this category. However, recently, a new approach has been proposed by the GA-PPI-Net method [4, 5]. It consists of evolving groups of nodes by optimizing their structural and qualitative scores in order to obtain some potential communities in the graph. This approach is known as an incremental approach, since the communities are constructed iteratively along the evolution.

In this study, we perform a systematic quantitative and qualitative evaluation of the capability of a clustering evolutionary method (EGCPI) and an incremental evolutionary approach (GA-PPI-Net) for inferring communities from PPI networks. EGCPI was proposed for detecting communities in PPI networks

using evolutionary game theory. The method aims to find the optimal partition of nodes into communities by modeling the interactions between nodes and using a genetic algorithm to get for the best solution [11]. The GA-PPI-Net is a Genetic Algorithm that allows to find one or several communities on a PPIs based on a search strategy. It uses the similarity score as well as the interaction score between proteins or genes and tries to find the best community by maximizing the concept of community measure [5]. To compare the two algorithms, we use the collins dataset[3]. This dataset concerns the genes of the yeast species Saccharomyces Cerevisiae (yeast).

Algorithms like EGCPI, identify protein community in PPI networks based only on network topologies. However, the GA-PPI-Net is not only based on graphical topology but also on semantic similarity between nodes. To our knowledge, GA-PPI-Net is the only communities' detection method that uses both semantic and topological measures. The main contribution of this paper is to evaluate and to compare two genetics based approaches in order to prove that GA-PPI-Net approach is suitable for the CD problem. The comparison is based on Three specific evaluation measures: i) the recovered percentage of each identified communities in an existing networks by using DAVID Tools [12]; ii) the semantic similarity measure and iii) the interaction score.

The rest of the paper is organized as follows. The next section presents an overview of the existing community detection algorithms. Section 3 provides a description of the two evolutionary approaches used in this study. In section 4, experimental results on real data set are presented and analyzed. Finally, section 5 reports the conclusion.

2 Background

The task for network community detection is to divide the whole network into small parts or groups, which are also called communities. In the literature, there is no uniform definition for community, but in academic domain, a community (also called a cluster or a module) is defined as a group of nodes that are connected densely inside the group but connected sparely with the rest of the network. Radicchi et al. [24] propose two definitions of community. These definitions are based on the degree of a node (or valency)[4]. In the first definition, a community is a subgraph in a strong sense: each node has more connections within the community than the rest of the graph. In the second definition, a community is a subgraph in a weak sense: the sum of all incident edges in a node is greater than the sum of the out edges.

2.1 Overview of Community Detection Methods

Many community detection methods in networks have been developed over the years. Each method has advantages and disadvantages (simplicity, efficiency, run

[3] https ://thebiogrid.org/

[4] The degree of a node is the number of edges incident to the node.

time etc.). The literature survey of community detection methods in graphs is divided into two categories: methods based on analytical or computational approaches and those based on evolutionary approaches [31, 2, 21]. The Analytical approaches use well defines rules that systematically explore the search space. Most of them are based on the hierarchical clustering techniques. These methods group nodes of a graph by minimizing the number of links between the different groups, so that the nodes of the same group or cluster are as similar as possible, while the nodes of different communities are as different as possible. As examples of clustering-based methods using graph modularity and density, we list Markov Clustering (MCL [30]), Restricted Neighborhood Search Clustering (RNSC [15]) and ClusterOne [19]. These methods have been applied to biological networks (PPI) to identify protein communities in PPI networks [31]. A recent comparative study of these methods with an evolutionary approach is published in [6] Other computational techniques have been proposed using different approaches, such as Random walk [23] or spectral clustering [32].

Evolutionary methods apply a global search algorithms that implicitly sample the search space and try to zoom on interesting regions based on the quality of the sampled solutions. Similarly, most of Evolutionary techniques are based on graph clustering strategy, except for GA-PPI-Net that uses an iterative an incremental strategy to identify clusters in the graph [5, 11]. As the main objective of this paper is to evaluate and to compare two GA based approaches, namely GA-PPI-Net vs EGCPI, the Evolutionary Approaches for community detection are presented in more details in next section.

A complete review of community detection methods in complex networks could be found in [2].
Most of clustering methods, either computational or evolutionary, are based on the modularity metric (Q). Modularity is a metric to measure the quality of partitioning a graph into communities. It is mainly used in social network analysis. It is introduced by M. E. J. Newman [20] and is defined in eq 1. It is described as the proportion of edges incident on a given class minus the value that this same proportion would have been if the edges were randomly arranged between the nodes of the graph.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(C_i, C_j) \tag{1}$$

Where:

- m is the total number of edges in the network
- A_{ij} is the ij^{th} element of the adjacency matrix of the network (it indicates whether the pair of nodes i and j are adjacent or not in the graph).
- k_i and k_j are the degrees of nodes i and j, respectively
- C_i and C_j are the communities to which nodes i and j belong
- $\delta(C_i, C_j)$ is the Kronecker delta function, which is 1 if $C_i = C_j$ and 0 otherwise

The modularity has been used as an optimization function for several community detection tasks in graphs [[20], [28],[17]]. The goal is to find, among all

possible partitions, the one with the best modularity. Another quality function has also been used in [22]. It consists in the determination of a global measure of the quality of a division within communities called *community score*.

2.2 Evolutionary Approaches for Community Detection

The Evolutionary Algorithm's (EA) ability to solve various problems has brought considerable popularity for them in solving optimization problems. These methods start from random individuals, and through keeping and combining the fittest and eliminating the weak solutions, narrow the search space to desired solutions [8]. This seemingly simple logic has shown to be able to find remarkable results for complicated problems. In recent years, several EA-based methods have been proposed to solve CD problems. A majority of evolutionary-based community detection methods are based on Genetic Algorithms (GA). GAs can be applied to community detection in networks by treating the problem as an optimization problem, where the objective is to find a partition of the network's nodes into communities that maximize a quality function. The quality function is typically chosen such as nodes within a community should be densely connected to each other, while nodes in different communities should be sparsely connected. A synthetic taxonomy of the different category of evolutionary methods is given in the following figure 1.

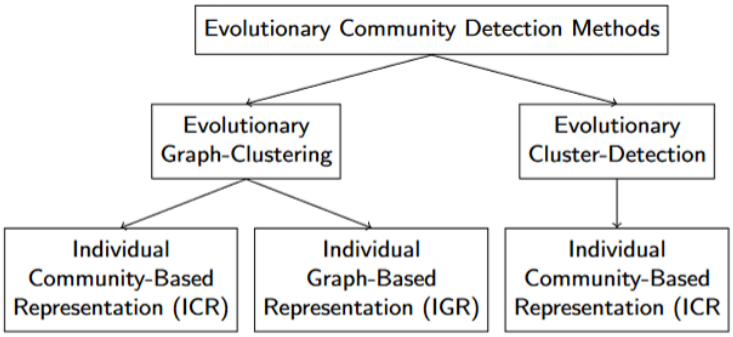


Fig. 1. Synthetic taxonomy of the different community detection methods.

The basic idea is to define a representation for each potential partition of the network and to evolve new candidate solutions by two mechanisms:

- the production of new solutions by the genetic operators (crossover and mutation)
- the selection of new solutions that replace earlier solutions in order to improve the overall quality of the population and preserve its diversity.

The fitness of each candidate solution is evaluated based on the quality function, and the fittest solutions are selected for further evolution. The modularity is the commonly used objective function in GA-based approaches. However, other quality functions have been used with GAs such as the conductance that measures the degree to which a community is internally connected and externally disconnected, or the edge betweenness that measures the importance of an edge in the network, based on the number of shortest paths that pass through the edge. The edge betweenness objective function can be used to identify edges that act as bottlenecks or bridges between communities, and can be used in combination with other objective functions such as modularity or conductance.

To design the individuals (solutions) of the population to be evolved, two main strategies were used by the GAs. In the first one, a graph partitioning solution is defined by the whole population. Thus, each individual represents a part of the graph (a cluster). Any modification of a cluster naturally modifies the other clusters. We denote this strategy in the taxonomy (fig 1) as the Individual-Community based representation (ICR). In the second strategy, each individual is a full clustering solution. So an individual represents the full clustered graph. We denote this strategy in the taxonomy (fig 1) as the Individual Graph-based Representation (IGR). This representation is expensive and difficult to scale to very large graphs. The simplest and least expensive representation that only evolves potential communities in the graph is proposed by GA-PPI-Net presented in the next section. Below some methods in the first two categories.

To the best of our knowledge, Bingol et al. made the first attempt to use EAs to detect the communities of a network [28] [29]. They have implemented an algorithm where individuals are graph partitions (IGR) designed with arrays of integer values. Each value indicates which community a particular node belongs to, and the modularity is the fitness function. Liu et al. introduced, in [17], an algorithm in which the optimal partition with the best modularity is obtained through successive bipartitions of the graph. A bipartition given by the GA is accepted only if it increases the total modularity of the graph. In [22], Pizzuti proposed a genetic algorithm named GA-Net to discover communities in social networks. GA-net is one of the famous community detection algorithms based on evolutionary methods, that doesn't use modularity as its fitness function. Two new measures are introduced, the community score and the notion of safe individual. The first one measures the quality of the partitioning of the network into communities and define the objective function for the algorithm.

The application of GA to PPI networks clustering started with MCODE-Evo. This method is an extension of the MCODE algorithm [3] for PPIs Clustering. Based on ICR, it applies a GA-based search to optimize the density and size of candidate communities, and then applies a refinement step using a local search strategy to iteratively remove or add nodes to the communities, while maintaining or increasing their fitness [16]. Later, Genetic Algorithm for Modularity Maximization (GAMMA) [7] is introduced. GAMMA is an ICR-based approach that uses modularity as the objective function for community detection in PPI networks. The algorithm starts with a random partition of the PPI network

into communities, and then iteratively applies genetic operators to generate new partitions with higher modularity.

Other GA-based approaches have been proposed for CD in PPIs this last decade, but they still using modularity as objective function, such as HGAC (Hierarchical Genetic Algorithm for Community Detection) [13] or PPI-GA (Clustering Genetic Algorithm to Identify Protein Complexes within PPI) [27]. However, despite the extensive use of modularity, it suffers from scalability problems, which indicates that modularity can't detect communities smaller than a specific scale [9]. Therefore, efforts to propose another measure continue, such with GA-PPI-Net presented in the following section.

3 GA-PPI-Net vs EGCPI

The main objective of this paper is to compare two Evolutionary approaches to identify communities in PPI networks belonging to the two different evolutionary categories in the DC taxonomy described in section 2. The first one, ECGPI [11], is an Evolutionary Graph-Clustering EA using Individual-Graph based representation (IGR). The second one is GA-PPI-Net [5] that is an Evolutionary Cluster-Detection GA using Individual-Community based representation (ICR). Both methods are detailed below.

3.1 EGCPI

EGCPI (Evolutionary Game-based Community detection algorithm for Protein-protein Interaction networks) is a recently proposed method for detecting communities in PPI networks using evolutionary game theory. The method aims to find the optimal partition of nodes into communities by modeling the interactions between nodes as a non-cooperative game, and using a genetic algorithm to search for the best solution [11].

EGCPI performs the task of community protein identification in several steps.

- An Attributed PPI network Graph (APPIG) is constructed based on the PPI network and the attribute information that can be obtained from the Gene Ontology database.
- A weighted Attributed PPI Graph (wAPPIG) is then constructed. Given an APPIG, EGCPI determines a weight to each edge in the graph based on the degree of topological similarity. It measures the weight of each pair of interacting proteins according to how much they are connected.
- An evolutionary algorithm is applied to identify dense graph clusters by maximizing the overall degree of topological similarity in each cluster.

EGCPI evolves an optimal graph clustering arrangement in several steps. To begin, a number of individuals is first initialized. Each individual in the population is encoded as a vector of integers, where each integer represents the community to which the corresponding node belongs. A population of random

solutions is generated. Then, it starts with the definition of a fitness function, which quantifies the quality of a partition of nodes into communities. The evolutionary game is then played by iteratively applying the following steps:

1. Reproduction: The fittest individuals in the population are selected to reproduce, and their offspring are generated through crossover and mutation.
2. Fitness evaluation: The fitness of each offspring is evaluated using the fitness function.
3. Game interaction: Each offspring competes with a randomly chosen individual from the previous generation, and the fitter individual is selected to survive to the next generation.

The algorithm terminates when a stopping criterion is met, such as a maximum number of generations or a threshold fitness value. The EGCPI method uses two parameters to control the algorithm's behavior: population size and mutation rate.

EGCPI applies an evolutionary graph clustering. The graph partitioning evolves by optimizing a topological measure called Independence of Cluster (IoC). Thus, in GA semantic similarity is not taken into account during evolution.

$$IoC_{c_i} = \frac{\sum\limits_{v_j, v_k \in C\,i} W_{jk}}{\sum\limits_{v_j \in C\,i} W_{jk}} \tag{2}$$

$$IoC_{wAPPIG} = \frac{\sum\limits_{i=1}^{S} n_{vi}}{n_v} IoC_{ci} \tag{3}$$

With:

- c_i: the i^{th} cluster,
- w_{jk}: the weight assigned to the interaction e_{jk} between the gene j and the gene k
- n_{vi} : the total number of nodes in the cluster i.
- n_v : the total number of nodes in the graph wAPPIG.
- S : the number of clusters.
- IoC_{ci}: represents the total weight of the intra-interactions compared to all the interactions linking the genes of the c_i cluster
- IoC_{wAPPIG}: measures the independence between clusters c_i with respect to the other clusters of the population. The latter allows to decrease the interdependence between two clusters.

After obtaining a set of clusters from the wAPPIG graph, EGCPI performs an additional step to identify gene communities. These communities are detected by calculating the degree of homogeneity between each pair of genes in the c_i cluster.

3.2 GA-PPI-NET

GA-PPI-Net is an evolutionary algorithm, that aim at identifying communities of proteins [5]. It allows finding communities having different sizes. It uses the similarity measures as well as the interaction measure between proteins and tries to find the best protein community by maximizing the concept of community measure. Thus, it allows combining two level of information:

1. Semantic level: information contained in biological ontologies such as Gene Ontology (GO) [1] and information obtained by the use of a similarity measure such as GO-based similarity of gene sets (GS2) [25]. It assesses the semantic similarity between proteins or genes.
2. Functional level: information contained in public databases describing the interactions of proteins or genes, such as Search Tool for Recurring Instances of Neighboring Gene (STRING) database [18].

GA-PPI-Net uses an Individual-Community based representation (ICR) and it is based on a specific structure for representing a community. Thus, a solution S is a community, and its quality is assessed thanks to the community measure. This one is used as the fitness function. The community measure, denoted F of a solution, S is computed using equations 4.

$$F(S) = W_1\ AVGSim(S) + W_2\ AVGInteraction(S) \tag{4}$$

where W_1 and W_2 are weights $\in [0, 1]$

$$AVGSim(S) = \sum_{i,j\ \in\ [1,n],\ i \neq j} SIM_{GS2}(G_i, G_j)/n \tag{5}$$

Where: i) G_i and G_j are two different genes in the community S; ii) n: the size of the community S; iii) $SIM_{GS2}(G_i, G_j)$: the similarity value between two genes (G_i, G_j) in S, it is calculated using the semantic similarity measure GS2 [25];

$$AVGInteraction(S) = \sum_{i,j\ \in\ [1,n],\ i \neq j} InteractionValue(G_i, G_j)/n \tag{6}$$

Where: $InteractionValue(G_i, G_j)$ is the value of an interaction between two genes (G_i, G_j) in S extracted from STRING Database [18].

This concept provides a solution of communities that are semantically similar and interacting. Moreover, it is based on a new genetic operation that is a specific mutation operator. The algorithm outputs the final community by selectively exploring the search space [5].

4 Comparative study

4.1 Experimental protocol

In this section, we propose a comparative study of the algorithm *GA-PPI-Net* with the EGCPI method using the Collins dataset. The latter was downloaded

from the BioGRID database platform [5]. The Collins dataset concerns the genes of the yeast species Saccharomyces cerevisiae (yeast). It is composed of 1620 genes and 9064 interactions.

The results obtained by EGCPI on Collins dataset are available online at [6]. In order to apply GA-PPI-NET, a preprocessing step on the Collins dataset is needed in order to determine the semantic similarity using the $GS2$ method and the interaction values from the String database.

The experimental protocol used is as follows:

- We execute *GA-PPI-Net* 20 times using the Collins dataset, and we retain the best detected community each time.
- To perform a fair comparison, the value of common parameters were considered the same for all methods. GA-PPI-Net and EGCPI is run using the same parameters described in [11] namely population size = 100, $PC = 0.6$ and number of generations = 30.
- We retrieve from the data available online all the clusters obtained by EGCPI. Then, we filter the identified communities in order to obtain communities with a size greater than five genes.
- Randomly select 20 communities from online available data of EGCPI having the same size as the communities detected by GA-PPI-Net.

4.2 Performance measures

We propose to use different metrics to report the comparison results. These metrics inspect the performance of the EGCPI and the GA-PPI-Net with respect to structural and functional quality as well as biological relevance of the identified communities. First, we check if the identified communities exist in real biological pathway databases such as KEGG Reactome, etc. This checking is achieved by the DAVID tools (Database for Annotation Visualization and Integrated Discovery), which compares this community with others in different databases and gives the percentage of proteins that belong to the existing communities in those databases. DAVID bioinformatics resources consist of an integrated biological knowledge-base and analytic tools that aim at systematically extracting biological meaning from large gene/protein lists. It is the most popular functional annotation programs used by biologists [26]. It takes as input a list of proteins and exploits the functional annotations available on these genes in a public database in order to find common functions that are sufficiently specific to these genes.

Moreover, to evaluate these two approaches, we assessed the functional and structural quality of the obtained communities by calculating the semantic similarity and interaction score for each pair of proteins within predicted communities. We then compared the distribution of these values across all communities being studied. The semantic similarity between two proteins was measured by

[5] https://thebiogrid.org/
[6] https://github.com/he-tiantian/EGCPI.

their respective Gene Ontology (GO) annotation terms using the GS2 method[5]. GS2 quantifies the similarity of the Gene Ontology annotations among a set of proteins by averaging the contribution of each all gene's Gene Ontology terms and their ancestor terms with respect to the Gene Ontology vocabulary graph [25]. The semantic similarity of a community was summarized by the average of the semantic similarity (*AVGSim*) of the protein pairs in the community, as defined in (eq 5). To determine the structural quality of an identified community, we used the string database. For each community, we calculated the average interaction score (*AVGInteraction*) of the protein pairs within that community, as defined in (eq 6).

4.3 Results and discussions

In this section, we comprehensively evaluate and compare the performance of the GA-PPI-Net [5] with the clustering evolutionary method EGCPI on widely used Collins dataset of real PPI network data. Since these two approaches depend on EA parameters, we have used the same parameters proposed in the study of Tiantian [11].

As said before, to inspect the performance of these two approaches with respect to functional and structural quality of the resulting communities, we first, determined the semantic similarity and the interaction score in each predicted community (i.e. clusters). The different computed scores are summarized in table 1.

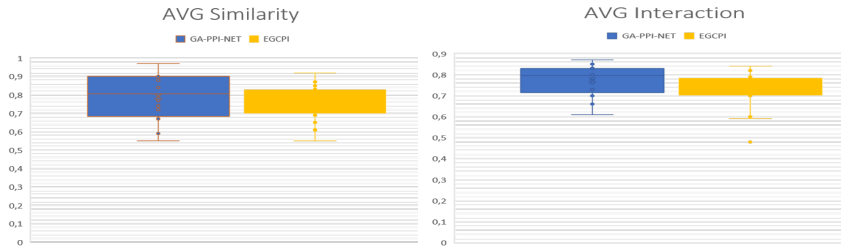
The results are summarized with a box plot in figure 2.

The comparison of the results of the two approaches can be made only by comparing the general quality of the obtained communities by the proposed measure of performance based on the similarity and the interaction scores. It should be noted that the similarity is not used by EGCPI but is calculated later on the obtained clusters to identify the communities. The similarity score' values varied between 0.55 and 0.97 for GA-PPI-NET and between 0.55 and 0.92 for EGCPI. Nevertheless, the GA-PPI-Net exhibited the largest average semantic similarity with respect to the EGCPI approach, for example where the size of the identified community is equal to 128, 57 and 30. As for the interaction score, which is used by both methods, it varies between 0.66 and 0.87 for GA-PPINET and between 0.48 and 0.82 for EGCPI. This result reflects that GA-PPI-NET performs slightly better according to this score than the other method. This superiority is confirmed by figure 2 illustrating the median value for both cases. Thus, the quality of the communities given by the two methods has a close competition, with a slight superiority of GA-PPI-NET.

Otherwise, the performance of both GA-PPI-Net and EGCPI approaches with respect to biological relevance of the obtained communities are evaluated by checking whether they exist in multi-species biological pathways such as KEGG [14], Reactome and Ec Number. This assessment is performed using the DAVID tool [12]. Each new identified community is presented to the DAVID tools, which compares this community with others in different biological databases and gives the percentage of proteins that belong to the existing communities in those

Table 1. Semantic similarity and interaction values of the obtained communities with GA-PPI-Net and EGCPI

Method	community size	AVG Similarity	AVG Interaction
GA-PPI-Net	9	0.88	0.71
	30	0.9	0.86
	7	0.67	0.61
	14	0.73	0.7
	8	0.59	0.8
	19	0.84	0.81
	128	**0.97**	**0.85**
	57	**0.91**	**0.83**
	8	0,72	0,66
	10	0,74	0,73
	38	0,67	0,7
	12	0,55	0,76
	7	0,77	0,84
	10	0,91	0,8
	25	**0,91**	**0,87**
	22	0,89	0,8
	12	0,67	0,78
	18	0,84	0,76
	30	0,9	0,79
	16	0,77	0,83
EGCPI: Step 1	9	0,78	0,48
	31	0,76	0,61
	7	0,7	0,6
	14	0,55	0,79
	8	0,69	0,72
	19	0,74	0,71
	123	0,92	0,59
	53	0,86	0,73
	8	0,71	0,74
	10	0,83	0,7
	39	0,85	0,72
	11	0,61	0,76
	7	0,73	0,82
	10	0,81	0,79
	25	0,87	0,84
	22	0,81	0,76
	12	0,65	0,72
	19	0,82	0,72
	29	0,82	0,8
	15	0,79	0,74
EGCPI: Step 2	10	0,92	0,81
	17	0,73	0,92



(a) AVG Similarity (b) AVG Interaction

Fig. 2. Synthesis of AVG Similarity and AVG Interaction measures recorded with EGCPI and GA-PPI-Net.

biological databases. The table 2 describes the minimum and maximum percent recovery rate of each approach in different biological databases.

Table 2. Min and Max recovery rate of the obtained communities (Collins dataset).

Methods	biologicals databases	% Min	% Max
GA-PPI-Net	Ec Number	2.4%	25.0%
	KEGG	12.5%	**100%**
	Reactome	5.3%	**100%**
EGCPI (Step 1: clustering)	Ec Number	2.5%	28.6%
	KEGG	5.1%	**100%**
	Reactome	6.9%	**100%**
EGCPI (Step 2 : Community detection)	Ec Number	-	-
	KEGG	-	100%
	Reactome	30%	100%

The results presented in Table 2 show that the identified communities correspond to some "parts" of real communities existing in other biological pathway databases, and in some cases to a complete network (percentage 100%). Therefore, the two methods were able to efficiently rebuilt communities existing in real biological pathway databases. GA-PPI-Net approach, as well the EGCPI, achieve the highest percentage 100% in two pathway databases: KEGG and Reactome. The worst percentage value is of 2.4% which corresponds to some "parts" of the real communities. Nevertheless, GA-PPI-Net exhibited the largest Percentage Min in the KEGG database. These tests should be supplemented on a larger scale with other datasets and different communities.

To conclude, the results show the capability of the GA-PPI-Net approach to effectively deal with community detection in PPI networks. Further extensions will be proposed to detect networks with larger size and identify new networks not yet known in the public biological databases.

5 Conclusion

With the continuous growth of the complexity and size of the available networks to be explored, the use of evolutionary algorithms as a powerful meta-heuristic for graph clustering is expanding, and several solutions have been proposed for community detection. This paper classifies the different community detection approaches based on GA in a general taxonomy, with a synthetic overview of the evolutionary methods. It presents in details two evolutionary methods based on different solution encoding and using different fitness function, EGCPI and GA-PPI-Net. A comparative study is then performed on the Collins PPI network. The biological relevance of the obtained communities by the two methods is similar. The qualities of the detected communities by the two methods, in terms of the average of the similarities and the average of the interactions between the genes, are very close, with a slight superiority of the GA-PPI-Net method. The advantage of the latter lies essentially in the simplicity of the representation of the solutions (communities) that makes it able to scale up for very large networks.

Future works aim to extend this comparative study to some analytical approaches for DC, such as Markov Clustering [30], Restricted Neighborhood Search Clustering [15] and ClusterOne [19].

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. **25**(1), 25–29 (May 2000). https://doi.org/10.1038/75556
2. Attea, B.A., Abbood, A.D., Hasan, A.A., Pizzuti, C., Al-Ani, M., Ã–zdemir, S., Al-Dabbagh, R.D.: A review of heuristics and metaheuristics for community detection in complex networks: Current usage, emerging development and future directions. Swarm and Evolutionary Computation **63**, 100885 (2021). https://doi.org/https://doi.org/10.1016/j.swevo.2021.100885, https://www.sciencedirect.com/science/article/pii/S2210650221000468
3. Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC bioinformatics **4**(1), 1–27 (2003)
4. Ben M'barek, M., Borgi, A., Bedhiafi, W., Ben Hmida, S.: Genetic Algorithm for Community Detection in Biological Networks. Proc Computer Science **126**, 195–204 (2018). https://doi.org/https://doi.org/10.1016/j.procs.2018.07.233, knowledge-Based and Intelligent Information Engineering Systems: Proc of the 22nd Inter Conf, KES-2018, Belgrade, Serbia
5. Ben M'barek, M., Borgi, A., Ben Hmida, S., Rukoz, M.: Genetic algorithm to detect different sizes' communities from protein-protein interaction networks. In: Proc of the 14th Inter Conf on Software Technologies - Volume 1: ICSOFT,. pp. 359–370. SciTePress (2019)

6. Ben M'barek, M.B., Hmida, S.B., Borgi, A., Rukoz, M.: GA-PPI-Net approach vs analytical approaches for community detection in PPI networks. Procedia Computer Science pp. 903–912 (2021). https://doi.org/https://doi.org/10.1016/j.procs.2021.08.093
7. Bilal, S., Abdelouahab, M.: Evolutionary algorithm and modularity for detecting communities in networks. Physica A: Statistical Mechanics and its Applications **473**, 89–96 (2017). https://doi.org/https://doi.org/10.1016/j.physa.2017.01.018, https://www.sciencedirect.com/science/article/pii/S0378437117300249
8. Cai, Q., Ma, L., Gong, M., Tian, D.: A survey on network community detection based on evolutionary computation. Int. J. Bio-Inspired Comput. **8**(2), 84–98 (May 2016). https://doi.org/10.1504/IJBIC.2016.076329
9. Fortunato, S., Hric, D.: Community detection in networks: A user guide. Physics Reports **659**, 1–44 (Nov 2016). https://doi.org/10.1016/j.physrep.2016.09.002, arXiv: 1608.00163
10. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. Proc. Natl. Acad. Sci. U.S.A. **99**(12), 7821–7826 (Jun 2002). https://doi.org/10.1073/pnas.122653799
11. He, T., Chan, K.C.C.: Evolutionary graph clustering for protein complex identification. IEEE/ACM Trans on Comp Biology and Bioinfo **15**(3), 892–904 (2018). https://doi.org/10.1109/TCBB.2016.2642107
12. Jiao, X., Sherman, B.T., Huang, D.W., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A.: DAVID-WS: a stateful web service to facilitate gene/protein list analysis. Bioinfo **28**(13), 1805–1806 (Jul 2012). https://doi.org/10.1093/bioinformatics/bts251
13. Jin, D., Wang, T., Cao, L., Zhang, Y.: Hgac: A hierarchical genetic algorithm for overlapping community detection in social networks. Information Sciences **258**, 26–42 (2014)
14. Kanehisa, M., Goto, S.: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28**(1), 27–30 (Jan 2000)
15. King, A.D., Pržulj, N., Jurisica, I.: Protein complex prediction via cost-based clustering. Bioinformatics **20**(17), 3013–3020 (2004)
16. Leiserson, M.D., Vandin, F., Wu, H.T., Dobson, J.R., Eldridge, J.V., Thomas, J.L., Papoutsaki, A., Kim, Y.S., Niu, B., McLellan, M., et al.: Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. Nature genetics **47**(2), 106–114 (2015)
17. Liu, X., Li, D., Wang, S., Tao, Z.: Effective algorithm for detecting community structure in complex networks based on ga and clustering. In: Inter Conf on Computational Science. pp. 657–664. Springer (2007)
18. Mering, C.v., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B.: STRING: a database of predicted functional associations between proteins. Nucl. Acids Res. **31**(1), 258–261 (Jan 2003). https://doi.org/10.1093/nar/gkg034
19. Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods **9**(5), 471 (2012)
20. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E **69**(2) (Feb 2004). https://doi.org/10.1103/PhysRevE.69.026113, arXiv: cond-mat/0308217
21. Pizzuti, C.: Evolutionary Computation for Community Detection in Networks: A Review. IEEE Transactions on Evolutionary Computation **22**(3), 464–483 (Jun 2018). https://doi.org/10.1109/TEVC.2017.2737600

22. Pizzuti, C.: Ga-net: A genetic algorithm for community detection in social networks. In: Inter conf on parallel problem solving from nature. pp. 1081–1090. Springer (2008). https://doi.org/10.1007/978-3-540-87700-4$_1$07
23. Pons, P., Latapy, M.: Computing communities in large networks using random walks. J. Graph Algorithms Appl. **10**(2), 191–218 (2006)
24. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., Parisi, D.: Defining and identifying communities in networks. PNAS **101**(9), 2658–2663 (Feb 2004). https://doi.org/10.1073/pnas.0400054101
25. Ruths, T., Ruths, D., Nakhleh, L.: GS2: an efficiently computable measure of GO-based similarity of gene sets. Bioinfo **25**(9), 1178–1184 (May 2009). https://doi.org/10.1093/bioinformatics/btp128
26. Sherman, B.T., Huang, D.W., Tan, Q., Guo, Y., Bour, S., Liu, D., Stephens, R., Baseler, M.W., Lane, H.C., Lempicki, R.A.: DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinfo **8**, 426 (Nov 2007). https://doi.org/10.1186/1471-2105-8-426
27. Shirmohammady, N., Izadkhah, H., Isazadeh, A.: PPI-GA: A novel clustering algorithm to identify protein complexes within protein-protein interaction networks using genetic algorithm. Complex. **2021**, 2132516:1–2132516:14 (2021). https://doi.org/10.1155/2021/2132516, https://doi.org/10.1155/2021/2132516
28. Tasgin, M., Bingol, H.: Community Detection in Complex Networks using Genetic Algorithm. arXiv:cond-mat/0604419 (Apr 2006), arXiv: cond-mat/0604419
29. Tasgin, M., Herdagdelen, A., Bingol, H.: Community Detection in Complex Networks Using Genetic Algorithms. arXiv:0711.0491 [physics] (Nov 2007), arXiv: 0711.0491
30. Van Dongen, S.M.: Graph clustering by flow simulation. Ph.D. thesis, Utrecht University Repository (2000)
31. Wu, Z., Liao, Q., Liu, B.: A comprehensive review and evaluation of computational methods for identifying protein complexes from protein–protein interaction networks. Briefings in bioinfo **21**(5), 1531–1548 (2020)
32. Zhang, Y., Levina, E., Zhu, J.: Detecting overlapping communities in networks using spectral methods. SIAM Journal on Mathematics of Data Science **2**(2), 265–283 (2020)