# Genetic Algorithm for Community Detection in Biological Networks

Marwa Ben M'barek, Amel Borgi, Walid Bedhiafi, Sana Ben Hmida

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 195–204

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2018, 3-5 September 2018, Belgrade, Serbia

# Genetic Algorithm for Community Detection in Biological Networks

Marwa BEN M'BAREK[a], Amel BORGI[a,b], Walid BEDHIAFI[c,d], Sana BEN HMIDA[e]

[a]*Université de Tunis El Manar, Faculté des Sciences de Tunis, LIPAH, 2092, Tunis, Tunisie*
[b]*Université de Tunis El Manar, Institut Supérieur d'Informatique, 1002, Tunis, Tunisie*
[c]*Université de Tunis El Manar, Faculté des Sciences de Tunis, Laboratoire de Génétique Immunologie et Pathologies Humaines, 2092,Tunis, Tunisie*
[d]*Department of Immunology-Immunopathology-Immunotherapy, Sorbonne Universités, UPMC Univ Paris 06, INSERM, UMRS959, Paris, France*
[e]*Université Paris Dauphine, PSL Research University, CNRS, UMR 7243, LAMSADE, Paris, France*

**Abstract**

We are interested in the detection of communities in biological networks. We focus more precisely on gene interaction networks. They represent protein-protein or gene-gene interactions. A community in such networks corresponds to a set of proteins or genes that collaborate at the same cellular function. Our goal is to identify such network or community from gene annotation sources such as Gene Ontology (GO). In this paper, we propose a Genetic Algorithm (GA) based approach to discover communities in a gene interaction network. Special solution coding and mutation operator are introduced. Otherwise, we propose a specific fitness function based on similarity measure and interaction value between genes. Experiments on real data extracted from the well-known Kyoto Encyclopedia of Genes and Genomes (KEGG) database show the ability of the proposed method to successfully detect existing or even new communities.



*Keywords:* community detection; biological networks; Gene Ontology; Genetic Algorithm; Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

## 1. Introduction

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of real networks system's representation is the existence of areas more densely connected than others. These areas are usually called communities [1].

In this paper, we are interested in the detection of communities in biological networks. More precisely we focus on gene interaction networks. These communities give us an idea about the perception of the network's structure. They


10.1016/j.procs.2018.07.233

correspond to a set of proteins or genes having the same specific function within the cell. This work is multidisciplinary as it brings the field of biology and computer science in the broad sense.

The analyzes performed in [2] showed that genes in the same community of the KEGG [3] database are semantically similar and are interacting. From this hypothesis, our first work consisted in characterizing the similarity between genes that are annotated by terms of GO. We tested different similarity measures to determine the most suitable measure for our problem. Our goal is to identify gene communities having biological sense (involved in the same biological process) from gene annotation sources. To achieve this task, we combine three levels of information:

1. Semantic level: information contained in biological ontologies such as Gene Ontology GO [4].
2. Functional level: information contained in public databases describing the interactions of genes such as Search Tool for Recurring Instances of Neighboring Gene (STRING) database [5].
3. Networks level: information contained in pathway databases that present community of genes such as KEGG database [3].

More precisely, this work is articulated in three main phases:

- "Data acquisition": it consists in extracting data from GO and STRING databases and saving them into a new database specific to this project.
- "Semantic similarity between genes": it consists in determining the semantic similarity between a community of genes based on GO. Several measures have been tested in order to identify the most appropriate measure of similarity between genes that will be adopted.
- "GA for detecting gene communities": it consists in proposing a new method for detecting communities of genes based on a genetic algorithm and using the similarity measure found in the previous step.

Informally, communities are groups of nodes that are connected densely inside the group but connected sparely with the rest of the network. Radicchi et al. [6] propose two definitions of community. These definitions are based on the degree of a node (or valency) that is the number of edges incident to the node. 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. In recent years, the problem of community detection has been receiving a lot of attention and many different approaches based on Genetic Algorithms GA have been proposed [7], [8], and [9]. These methods have been proposed to detect community structures in social networks. In [7] and [8] the authors present a GA that uses a fitness function based on the network modularity proposed by Newman and Girvan [10]. A different approach is described in [9] where a concept of community score is introduced as a fitness function able to identify densely connected groups of nodes. The approach searches for an optimal partitioning of the network by maximizing the community score. The concept of community score has proved to be very efficient.

In this paper, we propose a new community detection algorithm in biological networks based on GA, it tries to find the best community structure by maximizing the concept of community score. This score is not related to the density of groups as in [9] but it is based on semantic similarity and interaction between genes. The algorithm outputs the final community structure by selectively exploring the search space. Experiments on real networks show the ability of the genetic approach to correctly detect communities.

The paper is organized as follows. The next section provides the necessary background of the biological field and the used data to formalize the problem. In section 3, the notion of semantic similarity between genes is presented. Different approaches of semantic similarity measurements that allow the comparison of genes or genes products are tested in order to choose the best one. Section 4 depicts our main proposed algorithm for community detection. In section 5, experimental results on real data sets are presented and analyzed. Finally, Section 6 draws the conclusion.

## 2. Biological data preprocessing

In this section, we describe the used data acquiesced from different sources. Before that, let us present basic notions of biology. A biological pathway is a series of actions among molecules in a cell that leads to a certain product or a change in the cell. There are many types of biological pathways [11] such as metabolic pathways.

A biological network is a multiple biological pathways interacting with each other [12]. Our first objective is to get information about genes or genes product. As we said we combine three levels of information: semantic level, functional level and network level in order to get more information of gene interaction network' structure.

*2.1 Semantic Level: Gene Information*

*2.1.1 GO vocabulary structure*

GO is a major bioinformatics initiative to unify the representation of genes and genes products attributes across all species [4]. It provides a functional vocabulary for genes descriptions in terms of:

- Biological Process (BP): operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units (cells, tissues, organs …). In this work, we focus on BP collection.
- Cellular Component (CC): the parts of a cell or its extracellular environment.
- Molecular Function (MF): the elemental activities of a gene product at the molecular level, such as catalysis.

Each GO term within the ontology has a term name (which may be a word or string of words), a unique alphanumeric identifier (which start by ***GO :***), a definition with cited sources, and a namespace indicating the domain to which it belongs. Terms may also have synonyms (which are classed as being exactly equivalent to the term name, broader, narrower, or related), references to equivalent concepts in other databases, and comments on term meaning or usage. GO is structured as a collection of three directed acyclic graphs (DAGs), each representing a different ontology: (BP), (MF) and (CC). We use GO to extract data related to the BP aspects. We precisely focus on the relationship "is-a" and "part of" in order to identify the inheritance relationship between GO terms. To achieve this task, we implement two programs to extract the following data:

- The unique identifiers (ID) and the names of all the terms relating to the BP aspects. We obtain 28430 records.
- The inheritance relationship between the GO terms.

*2.1.2 Gene Ontology Annotation GOA*

The GOA is a project created by the European Bioinformatics Institute (EBI) that aims to provide assignments of terms from the GO resource to gene product [13]. It includes Swiss-Prot, TrEMBL and PIR-PSD that are the world's most highly annotated protein sequence databases, having archived and annotated more than a million proteins through a combination of manual and automatic techniques using the standardized vocabulary of the GO [14].

We use the GOA database to get a set of GO annotation for each gene of BP. For example the MEIKIN gene is identified by ID: 728637 and annotated by the following sets: "GO:0007060", "GO:0010789", "GO:0016321", "GO:0045143", "GO:0051754. These sets of terms represent the annotation of genes by GO.

*2.2 Functional level: Interaction between genes*

To study the interaction between genes, we use the database STRING. This database identifies genes and gene products interactions. It is one of the most complete resource that allows the exploration and visualization of the protein-protein or the gene-gene associations known and predicted according to different criteria in a bibliographic reference [5]. From this database, we extract couples of genes that are interacting, the mode of interaction between these couple of genes and the interaction value which defines the number of citation of this interaction in the literature.

*2.3 Network Level: Biological pathways databases*

Among the various biological pathways databases, we mention those that will be adopted.

Reactome: is an open source, open access, manually curated, peer-reviewed pathway database of human pathways and processes. The basic unit used to describe the data is the reaction [15].

Biocarta: concerns several species. It makes it possible to visualize, construct or identify the networks mapping the known genomic and proteomic relationships. It offers a synthesis of these paths and represents them by graphs [16].

Wikipathway: is a database of multi-species metabolic pathways. This database includes, on the one hand, metabolic pathways available in other databases such as Reactome or KEGG, and on the other hand patterns created by users through a graphical editing tool [17].

KEGG (Kyoto Encyclopedia of Genes and Genomes): is a knowledge base for systematic analysis of gene functions, linking genomic information with higher order functional information. The genomic information is stored in the GENES database, which is a collection of gene catalogs for all the completely sequenced genomes and some partial genomes with up-to-date annotation of gene functions [3].

The biological pathway database used to test the proposed approach is KEGG as it was the one proposed by our biology expert. In this database, we focused on the biological pathway. The other biological pathway databases are used to validate the experimental results as explained in section 5.

### *2.4 Summary of extracted data*

In this section, we present a summary of the extracted and used data:

- A gene is described by an ID, a name and a set of terms that annotate it. We have 17404 genes. For example, the description of the MEKIN gene is: ID: 728637 || NAME: MEIKIN || Terms that annotate it: ["GO:0007060", "GO:0010789", "GO:0016321", "GO:0045143", "GO:0051754"].
- The data related to the interaction between two genes. For example the interaction between the HSPA1A gene and the GRPEL1 gene is: NameGene1: "HSPA1A" || NameGene2: "GRPEL1" || Interaction: "reaction" || InteractionScore: 900
- The biological pathway is described by a source and a set of genes (pathway). These data are extracted from the KEGG database.

Fig. 1. summarizes the sources of these extracted data. Our goal is to get information about a gene. First, we get a set of GO terms that identify such gene from GO and GOA. Then, we get the interaction between a couples of genes from STRING database. We will proceed to the presentation of different approaches of semantic similarity, in the next section.

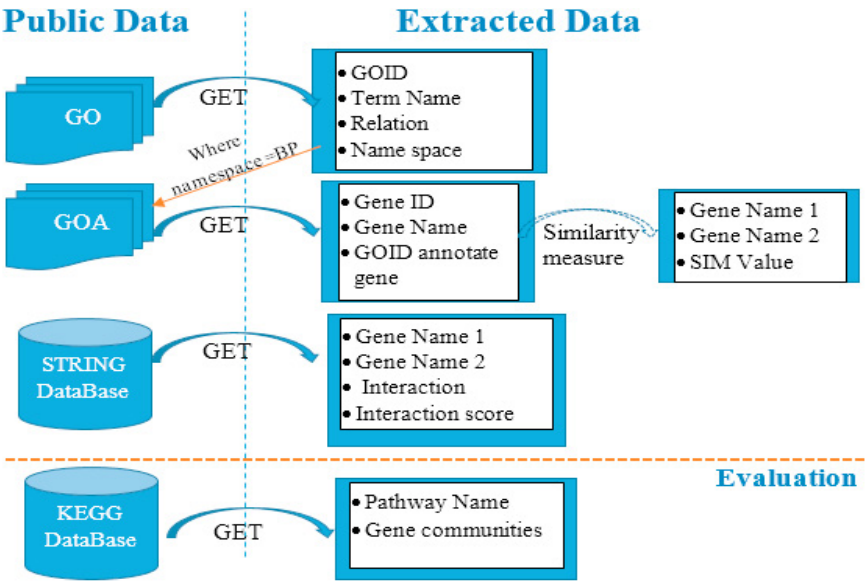


Fig. 1. Summary of extracted data.

## 3. Semantic Similarity based on GO

In [2] the authors showed that genes of the same community are semantically similar and interact with each other. We, therefore, assumed that genes belonging to the same community are similar and tried to find the best similarity measure between genes.

In this section, we define the notion of semantic similarity between genes. Then, we present different approaches of semantic similarity measurements that allow the comparison of genes or genes products. After that, we present the tests realized in order to choose the most appropriate similarity measure to retain for the rest of the work.

### *3.1 Notions of semantic similarity*

A gene can be annotated by several GO terms. To calculate the similarity between two genes, we need to use an approach allowing to compare sets of terms that annotate these genes thus we can quantify the similarity between these sets. In literature, there are three main approaches for measuring semantic similarity between the objects of an ontology [18] and [19]. The first are node-based approaches: the main data sources are the nodes and their properties. One concept commonly used in these approaches is information content, which measures how specific and informative a term is. The most prevalent node-based approaches are Resnik's [19], Lin's [20], Rel [21] and Jiang &Conrath's [22] methods. They were originally developed for the WorldNet, and then applied to GO [23]. The second approaches are edge-based approaches: they are based mainly on counting the number of edges in the graph path between two terms. The most common technique selects either the shortest path or the average of all paths when more than one path exists [24]. Among this family of approaches, there are the method of Rada [25] and the one of Wu & Palmer [24]. The third family of approaches are hybrid ones: Wang and al. [26] developed a hybrid measure in which each edge is given a weight according to the type of relationship. For a given term $c_1$ and its ancestor $c_a$, the authors define the semantic contribution of $c_a$ to $c_1$, as the product of all edge weights in the "best'' path from $c_a$ to $c_1$, where the ''best'' path is the one that maximizes the product. Semantic similarity between two terms is then calculated by summing the semantic contributions of all common ancestors to each of the terms and dividing by the total semantic contribution of each term's ancestors to that term. Ruths and al. [27] propose GS2 (GO-based similarity of gene sets), a novel GO-based measure of gene set similarity. The measure quantifies the similarity of the GO annotations among a set of genes by averaging the contribution of each gene's GO terms and their ancestor terms with respect to the GO vocabulary graph.

### *3.2 Implementing and testing similarity measurements*

Using a similarity measurement method, we determine the semantic similarity between two genes by comparing the set of annotation terms that define them.

To interpret the obtained results, a threshold must be defined. The used threshold is 0.5 (the midpoint of the interval in which the measure takes its values): if the value of similarity between two genes is greater than or equal to 0.5 then these two genes are similar, else they are not. Each similarity measurement method will take as input a set of genes and will give as output a symmetric similarity matrix. In order to choose a measure of similarity, we performed tests on eleven gene networks extracted from the database KEGG. We know that these groups of genes form communities and should have similar genes. We compute the semantic similarity using the Resnik, Jiang, Rel, Lin, Wang and GS2 methods. To achieve this task we used the open source package GOSemSim [28] that implements the Resnik, Jiang, Rel, Lin and Wang similarity measure. And we developed a tool that implements the GS2 method. From each similarity matrix, we determine the number of similar couples of genes. An example representing the result of these tests are described in Table1 (an example of two networks among the eleven tested networks). In this table, average rate refers to the average percentage of similar gene pairs obtained for each method. For example, with GS2 similarity measure, we found for the first network (R1) 1675 similar pairs of genes from 1891 pairs which corresponds to a rate of 87.63%. From these tests, we find that the similarity measure GS2 detects the largest number of similar genes on the eleven tested KEGG networks. So we decided to use the measure GS2 to characterize the similarity between genes in the rest of our work.

Table 1: Percentage of obtained similar gene pairs by different measures.

| Networks | Resnik | Jiang | Lin | Rel | Wang | GS2 |
|---|---|---|---|---|---|---|
| R1 | 61/1891 | 78/1891 | 68/1891 | 68/1891 | 855/1891 | 1657 /1891 |
| R2 | 13/55 | 33/55 | 37/55 | 31/55 | 31/55 | 49/55 |
| Average rate | 9.09% | 16.68% | 17.95% | 18.17% | 44.05% | 80.45% |

## 4. Proposed approach for detecting gene communities based on GA

In recent years, GAs have proved to be competitive alternative methods to traditional optimization and search techniques and they have been applied to many problems in diverse research and application areas such neural nets evolution, planning and scheduling, machine learning and pattern recognition [29].

In our approach of gene community detection, the population consists of individuals who are the solutions of the problem that is a set of genes that form a community. To evaluate a solution, we propose a fitness function based on a community score. The later uses the similarity value and the interaction score of every pair of genes making up the solution. Moreover, we modify the steps of GA to satisfy the needs of our algorithm. Thus, we propose a new mutation operator and insert some additional steps during the population initialization. The algorithm works as follows:

1. Start with an initial population of a set of genes which may be generated randomly,
2. Select parents from current population for mating,
3. Apply the crossover operator on the parents to generate new offsprings,
4. Apply mutation operator on the new off-springs generated,
5. Evaluate these off-springs, and replace the worst existing individuals in the population by these offsprings,
6. Repeat the process from the second step while the stop condition is not satisfied (stop condition: predefined number of generations or computation time reached).

We now explain the principle of each step of the proposed approach based on GA to detect gene communities.

### *4.1* Individual *Representation*

One of the most important decisions to make while implementing a GA is deciding how to represent our individuals. A solution of our problem is a community of genes. We represent it by a vector T that stores the genes names of a community, the average value of similarity (see equation1) and the average interaction score (see equation 2) of each two genes. This vector has (|V| + 2) elements where | V | is the number of genes in a community, it corresponds to an individual in GA terms. In this work, we only focus on communities having the same size (V). Fig. 2. illustrates the representation of an individual adopted in our algorithm.

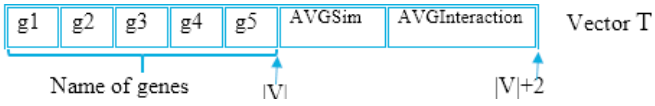


Fig. 2. Example of individual representation designing a community with five genes.

$$AVGSim = \frac{\sum_{i=1}^{n-1} Sim_{GS2}(G_i, G_{i+1})}{n} \quad (1)$$

$$AVGInteraction = \frac{\sum_{i=1}^{n-1} InteractionValue(G_i G_{i+1})}{n} \quad (2)$$

With: n = |V|: the number of genes in a community denoted by {g1, g2, g3, .., gn}.

$Sim_{GS2}$: the similarity value between two genes, it is calculated using the semantic similarity measure GS2.

$InteractionValue$: the score of an interaction between two genes extracted from STRING Database.

### 4.2 Population initialization

The population is defined as a two-dimensional array of individuals. It is a set of "individuals" that represent some potential solutions of the problem. In order to initialize this population, we must first recover randomly a group of genes having the same size from the created table "Pathway" extracted from KEGG database. Then, we compute the similarity value and get the interaction score of each two genes of this group from the created "interaction" table. Next, we calculate the average similarity value and the average interaction score of each group forming this population. Fig. 3. describes an example of the population with 4 individuals designing each a community of 5 genes.

| | | | | | | |
|---|---|---|---|---|---|---|
| IL2 | FOS | MALT1 | IL4 | MAPK1 | 0.764 | 0.581 |
| UGT1A7 | IL2RB | CALM1 | RPS6KA6 | TGFB1 | 0.851 | 0.541 |
| IL2RB | TBX21 | IL5 | ALDH1B1 | ALDH2 | 0.763 | 0.785 |
| CBL | SLK | AKT3 | UGDH | MIOX | 0.605 | 0.682 |

Fig. 3. Example of initial population with four individuals.

### 4.3 Fitness Function

The fitness function takes a candidate solution to the problem as input and produces as output how "fit" or how "good" the solution is with respect to the considered problem. The computation of the fitness value is done repeatedly in a GA and therefore it should be sufficiently fast. We propose to define a fitness function based on the computation of the average similarity value and the average interaction score of each two genes existing in the community. Indeed, we started from the hypothesis that genes in the same community are semantically similar and interact which other. It is defined as follows:

$$F = w_1\, AVGSim + w_2\, AVGInteraction \quad (3)$$

With: $AVGSim$ and $AVGInteraction$ defined in (1) and (2).

$w_1 and\ w_2 : weights \in [0,1]$

We performed several tests with different values of $w_1$and $w_2$ (($w_1$,$w_2$) = (0, 1), (1, 0) and (0.5, 0.5)). Then, we chose the values which give the best results in terms of the number of known networks recovered from KEGG. So, the value taken for the fitness function are $w_1$ = 0.5 and $w_2$ = 0.5.

### 4.4 Selection

Selection is the process of selecting parents which mate and recombine to create offsprings for the next generation. It is very crucial to the convergence rate of the GA as good parents drive individuals to better and fitter solutions. We used the tournament selection method because it's the most popular selection method in GA due to its efficiency and simple implementation [30].

### 4.5 Genetic operators

Crossover and mutation in GA are applied to generate new solutions for the next generation. Their goal is to both exploit the best solutions and to explore the search space. For this work, we used the common One Point Crossover: a random crossover point is selected and the tails within the two parents are swapped to get two new offsprings. This operator is usually applied with a high probability (*pc*). However, for mutation, we propose a new operator that can better meet the objectives of our problem.

Mutation may be defined as a small random tweak in the individual, to get a new solution. It is used to maintain and introduce diversity in the population and is usually applied with a low probability (*pm)*. If the probability is very high, the GA gets reduced to a random search. We now present a new mutation operator that is specific to our biological problem, it should allow a better exploration for the search space than the random mutation. Its goal is to maximize the chance of creating a better solution than the original one. To mutate a solution S, it alters only one gene at a time and uses a score function, denoted GS, applied to each gene in S. This score will help us to detect the gene

having the best score in a community. It is equal to the sum of the average similarity and the average interaction score of a gene in a community. It is defined as follows:

$$GS(G) = \frac{\sum_{i=1}^{n} Sim_{GS2}(G\ ,G_i)}{m} + \frac{\sum_{i=1}^{n} IneractionValue(G\ ,G_i)}{p} \quad (4)$$

With: $Sim_{GS2}(G\ ,G_i)$ : The similarity of a gene G compared to the other genes in the community.

$InteractionValue(G\ ,G_i)$: The interaction score of a gene G compared to the others in the community.

n: size of an individual (community)

m: number of similarity values different from 0.

p: number of interaction score values different from 0.

The mutation operator is applied according to the following steps:

1. Select in S the gene having the highest score GS that will be called "bestGene".
2. Randomly search a gene G' from the "interaction" table with which the "bestGene" interacts and G' ∉ S.
3. Get the gene having the lowest score GS in S, it will be called "worstGene".
4. Replace "worstGene" with the selected gene in the second step.

## 5. Experimental Results

In this section, we study the effectiveness of our approach on a real data set. We first carried out tests to tune the GA parameters. Different parameters values were tested: generation number set at 100, 300 and 500, size of the population set at 10, 20, 30, 70 and 100, crossover rate set at 0.5, 0.6, …1 and mutation rate set at 0.1, 0.2, …, 0.5. Based on these tests, we chose the combination of the parameters values giving the best results (highest values of fitness function), namely: population size 20, generation number 100, crossover rate 0.8 and mutation rate 0.2.

In order to check the ability of our approach to successfully detect the community structure of a network, we use genes that are present in communities from the reference pathway database KEGG. The evaluation consists to verify how the proposed method is likely to find gene communities existing in KEGG database. In fact, it is possible to detect communities of genes existing in KEGG database or a new community having high interaction and high similarity between its genes and that do not appear in KEGG. We performed tests to determine communities having 5, 10 and 16 genes. For each fixed size, we executed our approach 20 times. And, we retained each time the best community. So, we have twenty best communities of each size (5, 10 and 16). The results are shown in Table 2.

Table 2. Communities' detection: experimental results.

| Network Size | Networks belonging to KEGG | | Obtained new networks | |
|---|---|---|---|---|
| | Number of networks | Average fitness | Number of networks | Average fitness |
| 5 | 8/20 | 0,761 | 12/20 | 0,880 |
| 10 | 9/20 | 0,688 | 11/20 | 0,723 |
| 16 | 0/20 | - | 20/20 | 0,702 |

From table 2, we find that networks having size 5 or 10, correspond approximately to half of the communities existing in the KEGG database and correspond to real networks. In the case of networks having size 16, no community from the KEGG database was found. When a new network is found the question that arises is how it will be evaluated. Our biology expert, proposed to evaluate this network by checking if it exists in other biological pathway databases than KEGG. The biological databases used to evaluate our results are Biocarta, Reactome BBID and EC Number. Each new network $R_{new}$ founded is presented to the DAVID tools (Database for Annotation Visualization and Integrated Discovery), which compares this network with others in different databases and gives the percentage of $R_{new}{}'s$ genes that belong to the same community. DAVID bioinformatics resources consists 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 [31]. It takes as input a list of genes and exploits the functional annotations available on these genes in a public database such as BIB, KEGG Pathways, BioCarta Pathways etc., in order to find common functions that are sufficiently specific to these genes. For example, one of the new networks obtained with size 10 has the following percentages which describe the matching

of its genes in the communities of other databases: Biocarta (50%), BBID (50%), EC Number (75%) and Reactome Pathway (100%).These percentages are established by the division of the number of genes in the community concerned, which belong to other networks of biological pathways by the total number of genes forming these networks. We tested all the new networks obtained by our algorithm using DAVID tools. Tables 3, 4 and 5 below represent the minimum and the maximum percentage of genes that belong to a community in other biological pathways.

Table 3. Evaluation of new communities having size 5.

| Databases | Min Percentage | Max Percentage |
|---|---|---|
| Biocarta | 20% | 75% |
| REACTOME | 40% | 100% |
| BBID | 25% | 75% |
| EC Number | 25% | 60% |

Table 4. Evaluation of new communities having size 10.

| Databases | Min Percentage | Max Percentage |
|---|---|---|
| Biocarta | 20% | 100% |
| REACTOME | 70% | 90% |
| BBID | - | 10% |
| EC Number | 88.9% | 100% |

Table 5. Evaluation of new communities having size 16.

| Databases | Min Percentage | Max Percentage |
|---|---|---|
| Biocarta | 25% | 93,8% |
| REACTOME | 75% | 100% |
| BBID | 6.2% | 50% |
| EC Number | 88.9% | 100% |

The results presented in these tables show that the new networks obtained by our algorithm, those that don't exist in the KEGG database, correspond to some "parts" of real networks existing in other biological pathway databases, and in some cases to a complete network (percentage 100%). These results are considered very satisfactory by the biology expert. They constitute an initial validation of our algorithm, and show the relevance of the used fitness function. These tests should be supplemented on a larger scale with other datasets and different community sizes.

## 6. CONCLUSION

In this article, we have presented our work consisting of three parts. The first one focused on the data extraction from different sources. The second part dealt with the study of the semantic similarity between groups of genes that are annotated by the terms of GO. We were interested in determining the similarity between genes because it is considered as a characteristic of a gene community. There are several approaches that allow to compare sets of terms of an ontology in order to quantify the similarity between these sets. After testing different similarity methods, we adopted the GS2 method for our work. The last part presented the proposed approach based on GA to detect communities of interacting genes or proteins having same size. This approach introduced the concept of community score and searched for an optimal partitioning of the network by maximizing these scores. Our contribution in this paper is threefold. First, we applied GA to community detection in biological networks. Second, we proposed a community detection algorithm suitable for large networks. Third, we defined a specific mutation operator adapted to the considered biological problem. Dense communities presented in the network structure are obtained at the end of the algorithm by selectively exploring the search space, with the need to know in advance the community size. The experimental results showed the ability of the genetic approach to correctly detect communities having the same size. Future research will aim at applying multi-objective optimization to improve the quality of the results. Another perspective, which seems to be a logical continuation of this work, is the adaptation of the proposed approach to detect communities of genes having different sizes.

## References


[1] S. Fortunato, *Community detection in graphs*, Phys. Rep., vol. 486, no. 3–5, pp. 75–174, Feb. 2010.

[2] X. Guo, R. Liu, C. D. Shriver, H. Hu, and M. N. Liebman, *Assessing semantic similarity measures for the characterization of human regulatory pathways*, Bioinforma. Oxf. Engl., vol. 22, no. 8, pp. 967–973, Apr. 2006.

[3] M. Kanehisa and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes*, Nucleic Acids Res., vol. 28, no. 1, pp. 27–30, Jan. 2000.

[4] M. Ashburner et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*, Nat. Genet., vol. 25, no. 1, pp. 25–29, May 2000.

[5] C. von Mering, M. Huynen, D. Jaeggi, S. Schmidt, P. Bork, and B. Snel, *STRING: a database of predicted functional associations between proteins*, Nucleic Acids Res., vol. 31, no. 1, pp. 258–261, Jan. 2003.

[6] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, *Defining and identifying communities in networks*, Proc. Natl. Acad. Sci. U. S. A., vol. 101, no. 9, pp. 2658–2663, Feb. 2004.

[7] M. Tasgin, A. Herdagdelen, and H. Bingol, *Community Detection in Complex Networks Using Genetic Algorithms*, In: Proc. of the European Conference on Complex Systems, Apr. 2006.

[8] X. Liu, D. Li, S. Wang, and Z. Tao, *Effective Algorithm for Detecting Community Structure in Complex Networks Based on GA and Clustering*, in Computational Science – ICCS, pp. 657–664, May.2007.

[9] C. Pizzuti, *GA-Net: A Genetic Algorithm for Community Detection in Social Networks*, in Parallel Problem Solving from Nature – PPSN X, pp. 1081–1090, Sep. 2008.

[10] M. Girvan and M. E. J. Newman, *Community structure in social and biological networks*, Proc. Natl. Acad. Sci. U. S. A., vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[11] Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health, *RNA | definition of RNA by Medical dictionary*, 2003.

[12] National Human Genome Research Institute (NHGRI), *Biological Pathways Fact Sheet, Aug. 2015*.

[13] E. Camon et al., *The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro*, Genome Res., vol. 13, no. 4, pp. 662–672, Apr. 2003.

[14] M. A. Harris, *The Gene Ontology (GO) database and informatics resource*, Nucleic Acids Res, vol. 32, pp. D258–D261, 2004.

[15] D. Croft et al., *Reactome: a database of reactions, pathways and biological processes*, Nucleic Acids Res., vol. 39, no. Database issue, pp. D691-697, Jan. 2011.

[16] D. Nishimura, *BioCarta*, Biotech Softw. Internet Rep., vol. 2, no. 3, pp. 117–120, Jun. 2001.

[17] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo, *WikiPathways: pathway editing for the people*, PLoS Biol., vol. 6, no. 7, p. e184, Jul. 2008.

[18] Y. Xu, M. Guo, W. Shi, X. Liu, and C. Wang, *A novel insight into Gene Ontology semantic similarity*, Genomics, vol. 101, no. 6, pp.368–375, Jun. 2013.

[19] P. Resnik, *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural*, J. Artif. Int. Res, Vol.11, no.1, pp. 95-130, Jul.1999.

[20] D. Lin, *An Information-Theoretic Definition of Similarity*, In Proceedings of the 15th International Conference on Machine Learning, pp. 296–304, 1998.

[21] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, *A new measure for functional similarity of gene products based on Gene Ontology*, BMC Bioinformatics, vol. 7, p. 302, Jun. 2006.

[22] Jiang, Jay J. and David W. *Conrath, Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy,* ROCLING, Oct.1997.

[23] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto, *Semantic Similarity in Biomedical Ontologies*, PLoS Comput. Biol., vol. 5, no. 7, Jul. 2009.

[24] Z. Wu and M. Palmer, *Verbs Semantics and Lexical Selection*, in Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 133–138, 1994.

[25] R. Rada, H. Mili, E. Bicknell, M. Blettner, *Development and application of a metric on semantic nets*, IEEE Trans. Syst. Man Cybern.19 pp.17–30, 1989.

[26] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, *A new method to measure the semantic similarity of GO terms*, Bioinformatics, vol. 23, no. 10, pp. 1274–1281, May 2007.

[27] T. Ruths, D. Ruths, and L. Nakhleh, *GS2: an efficiently computable measure of GO-based similarity of gene sets*, Bioinforma. Oxf. Engl., vol. 25, no. 9, pp. 1178–1184, May 2009.

[28] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu and S. Wang, *GOSemSim: an R package for measuring semantic similarity among GO terms and gene products*, Bioinformatics, Volume 26, Issue 7, pp. 976-978, Apr. 2010.

[29] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, 1st ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989.

[30] D. E. Goldberg and K. Deb, *A comparative analysis of selection schemes used in genetic algorithms*, in Foundations of Genetic Algorithms, pp. 69–93, 1991.

[31] B. T. Sherman et al., *DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis*, BMC Bioinformatics, vol. 8, p. 426, Nov. 2007.