Pamukkale Univ Muh Bilim Derg, 28(2), 292-298, 2022

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

# Evaluation of Sub-Network search programs in epilepsy-related GWAS dataset

## Epilepsi ile ilgili GWAS veri kümesinde alt ağ arama programlarının değerlendirmesi

*Beyhan ADANUR DEDETURK[1*], Burcu BAKİR GUNGOR[1]*

[1]Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey.
beyhan.adanur@agu.edu.tr, burcu.gungor@agu.edu.tr

Received/Geliş Tarihi: 08.03.2021 Revision/Düzeltme Tarihi: 02.06.2021 doi: 10.5505/pajes.2021.56424
Accepted/Kabul Tarihi: 28.06.2021 Research Article/Araştırma Makalesi

**Abstract**

*The active sub-network detection aims to find a group of interconnected genes of disease-related genes in a protein-protein interaction network. In recent years, several algorithms have been developed for this problem. In this study, the analysis of disease-specific sub-network identification programs is evaluated using epilepsy data set. Under the same conditions and with the same data set, 9 different programs are run and results of their Greedy algorithm, Genetic algorithm, Simulated Annealing Algorithm, MCC (Maximal Clique Centrality) algorithm, MCODE (Molecular Complex Detection) algorithm, and PEWCC (Protein Complex Detection using Weighted Clustering Coefficient) algorithm are shown. The top-scoring 5 modules of each program, are compared using fold enrichment analysis and normalized mutual information. Also, the identified subnetworks are functionally enriched using a hypergeometric test, and hence, disease-associated biological pathways are identified. In addition, running times and features of the programs are comparatively evaluated.*

**Keywords:** Protein-Protein interaction networks, Active sub-network search, Functional enrichment analysis, Fold enrichment, Normalized mutual information.

**Öz**

*Aktif alt ağ tespiti, bir protein-protein etkileşim ağında hastalıkla ilgili genlerin birbirine bağlı bir grup genini bulmayı amaçlamaktadır. Son yıllarda bu problem için çeşitli algoritmalar geliştirilmiştir. Bu çalışmada, hastalığa özgü alt ağ tanımlama programlarının analizleri epilepsi veri seti kullanılarak değerlendirilmiştir. Aynı koşullar altında ve aynı veri seti ile 9 farklı program çalıştırılmış ve bu programların Greedy algoritması, Genetik algoritma, Simüle Tavlama Algoritması, MCC (Maximal Clique Centrality) algoritması, MCODE (Molecular Complex Detection) algoritması ve PEWCC (Protein Complex) Ağırlıklı Kümeleme Katsayısı) algoritması sonuçları gösterilmiştir. Her programın en yüksek puan alan 5 modülü, kat zenginleştirme analizi ve normalleştirilmiş karşılıklı bilgi kullanılarak karşılaştırılmıştır. Aynı zamanda tanımlanan alt ağlar, hipergeometrik test kullanılarak fonksiyonel olarak zenginleştirilmiş ve hastalıkla ilişkili biyolojik yollar belirlenmeye çalışılmıştır. Ayrıca programların çalışma süreleri ve özellikleri karşılaştırmalı olarak değerlendirilmiştir.*

**Anahtar Kelimeler:** Protein-Protein etkileşim ağları, Aktif alt-ağ araması, Foksiyonel zenginleştirme analizi, Kat zenginleştirme, Normalleştirilmiş karşılıklı bilgi.

## 1 Introduction

Understanding life's secrets have always been a key problem on which several disciplines have collaborated. Bioinformatics and genomics are fields of study that look into the secrets of life using biological data. The link between diseases in an organism and the causative gene or mutations can be established via bioinformatics analysis. Scientists can perform disease predictions and evolutionary processes of disease by applying these analyses to -omics or GWAS data. As a result, by developing personalized medication and treatment, a disease can be prevented, and its effects can be reduced before it poses a serious hazard. Epilepsy is a serious and prevalent neurological disease that is linked to psychiatric comorbidities. The number of persons suffering from epilepsy has risen to 65 million all over the world. Despite the discovery of numerous anti-epilepsy treatments, roughly 30% of patients cannot be cured or show no response to treatments due to the development of pharmacoresistance during therapy [1]. So, novel and effective treatments based on the pathogenesis of epilepsy are urgently needed. In order to develop a new and effective treatment, first of all, it is necessary to determine whether there are genes responsible for this disease. If there is, the relationship between these genes and the disease should be examined. For this purpose, computational analyzes are performed. One of such analyzes is the search for active modules containing disease-specific proteins using the information from high throughput methods performed in the wet laboratory, e.g., microarray studies, RNA-seq, genome-wide association studies (GWAS).

Protein-protein interaction (PPI) networks present the interactions among proteins, based on their operation in the cell. In these PPI networks, the active module search aims to find disease-related subnetworks that contain most of the highly affected nodes (proteins) and their interaction partners with medium effect on the disease [2]. The active subnetwork search problem requires two main inputs, i) protein-protein interaction network, ii) node scores of proteins that indicate the statistical significance of a protein for the disease under investigation [3]-[5]. Most methods often use undirected graphs of protein-protein interaction networks and the node scores are used as the weight of the nodes. Via defining a score for a subnetwork, the search step of active subnetwork search tries to find the sub-network with the maximum score [6]. The active subnetwork search is an NP-hard problem and many

*Corresponding author/Yazışılan Yazar

methods have been developed especially focusing on the search step. In this study using epilepsy-related GWAS dataset; performance results of Greedy algorithm, Genetic algorithm, Simulated Annealing Algorithm, MCC (Maximal Clique Centrality) algorithm, MCODE (Molecular Complex Detection) algorithm, and PEWCC (Protein Complex Detection using Weighted Clustering Coefficient) algorithm are compared. The names of programs used are as follows; jActiveModules [7], PINBPA [8], MCODE [9], PEWCC [10], ActiveSubnetworkGA [11], ClusterViz [12], CytoHubba [13], CytoMOBAS [14] and PathFindR [15]. Results are evaluated using functional enrichment analysis using BINGO [16], fold enrichment analysis [17], and normalized mutual information [18],[19]. In our previous work, GO enrichment analysis results of programs are shown in detail [20]. The operation requirements and the parameters of each active module search program are different. We also reviewed SubNet [21], MSIGNET [22], CytoGTA [23], BMRF-Net [24], dmGWAS [25], COSINE [26], Prize-Collecting Steiner Forest [27] and MAGENTA [28] programs. Different input parameters are required in each program according to the data set used, hence these programs are not included in this study but for different data sets, these programs are among the most used. The above-mentioned analysis is summarized in Figure 1.

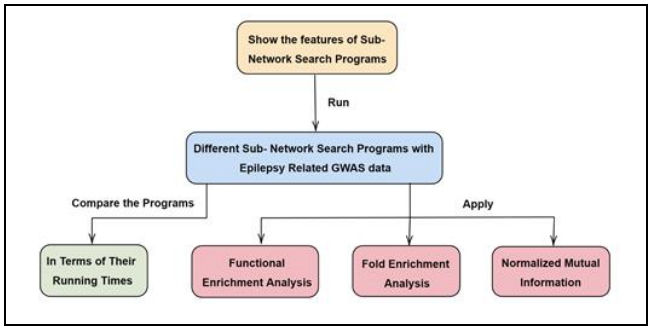


Figure 1. Shows the summary processing of our work.

## 2 Methods

### 2.1 Datasets

A human protein-protein interaction network and epilepsy-related GWAS dataset [29] are used to run all sub-network search programs in this study. The PPI network dataset [30],[31] contains 10175 different genes, while the GWAS dataset includes 4494 epilepsy-related genes and their importance scores. 224 genes related to idiopathic generalized epilepsy (IGE) are obtained from DISEASE [32] database and used as reference genes in fold enrichment analysis.

### 2.2 Sub-Network search programs

In this section, the active subnetwork search programs that are compared in this study are briefly introduced with their basic features. All programs are Cytoscape [33] plugins except ActiveSubnetworkGA and PathFindR.

#### 2.2.1 jActiveModules

This method is the pioneer of subnetwork search algorithms and combines statistical measurements with a search algorithm to find high-scoring modules. jActiveModules firstly calculates a z-score and uses this score in the search step based on simulated annealing and genetic algorithms. The purpose of the simulated annealing algorithm is to search for the highest-ranked subnetwork, and the greedy search expands a subnetwork by adding one of its adjacent genes to maximize a feature based on mutual information. In our study, we ran this program with the simulated annealing algorithm with its default parameters.

#### 2.2.2 PINBPA (Protein interaction network-based pathway analysis)

This method is the first Cytoscape app intended to analyze GWAS information on the network. It provides an easy interface for a complicated set of analyzes and can be used in a broad variety of situations, enabling genomic researchers to conduct post-GWAS analyzes in a simplified, thorough, and reproducible manner. The program works with a greedy algorithm and it has six basic steps for the analysis, i.e., building a manhattan chart, sequencing the coordinates of all genes, generating the subnetworks of first-order networks based on the threshold value, checking the statistical importance test of subnetworks, applying the restart algorithm and identifying modules that enriched with important genes using z-scores.

#### 2.2.3 MCODE (Molecular complex detection)

The MCODE detects tightly linked areas in big networks of protein-protein interaction that can represent molecular complexes. The technique is based on vertex weighting from a locally dense seed protein by local neighborhood density and outward traversal to isolate the thick areas according to specified parameters. The algorithm has the benefit over other graph clustering techniques of having a guided mode that enables cluster fine-tuning of interest without considering the remainder of the network and enables cluster interconnectivity to be examined, which applies to protein networks. It consists of three stages: peak weight, complex estimation, and optionally post-treatment.

#### 2.2.4 PEWCC (Protein complex detection using weighted clustering coefficent)

The method is a novel mining algorithm for identifying groups such as protein complexes. First, the algorithm evaluates the accuracy of interaction information and then predicts protein complexes based on the weighted clustering coefficient idea. This method can be used for all kinds of diseases. PEWCC first assesses the reliability of protein interaction data using the PE measure which is a new measure to protein pairs interaction reliability then it detects protein complexes using a weighted aggregation coefficient.

#### 2.2.5 ActiveSubnetworkGA

The method is a novel genetic algorithm technique in that crossover branch swapping, individual addition mutation, pruning, and two-stage architecture are introduced. It is similar to jActiveModules, but the search step using a genetic algorithm is more advanced. The goal of the method is to get the best solution which refers to a subnetwork with the maximum score. The algorithm is executed up to the threshold value specified by the user. The best solutions are identified as the first population and the genetic algorithm is run once again for best results. Good solutions are combined with branch swapping crossover.

#### 2.2.6 ClusterViz

This method is used to find extremely interconnected areas, protein complexes, or functional modules in a network. To do more associated studies, ClusterViz fascinates the comparison of the outcomes of distinct algorithms. It has three clustering

algorithms, i.e., FAG-EC (fast agglomerate algorithm based on the edge clustering coefficients), EAGLE and MCODE. In our study, ClusterViz runs with FAG-EC algorithm. Since the edge clustering coefficient is local variables, FAC-EC has a low time complexity and can handle large PPI networks.

#### 2.2.7 CytoHubba

The method provides eleven different topological analysis methods. It is possible to divide eleven techniques into two main classifications as local and global methods. A local rank method only looks at the relationship between the node and its direct neighbors to calculate the score of a node within a network vice versa the global method looks at the relationship between the node and all networks. Among the eleven different analysis methods, the MCC (Molecular Complex Detection) has the best performance in predictive accuracy so, in this study, we preferred to use the MCC method.

#### 2.2.8 CytoMoBAS

The main concept of the suggested technique is to evaluate the association of each interaction with the disease in the network and to take into consideration the association of background disease as an approximation for statistical significance. The CytoMoBAS proposes a scoring system that integrates parameter-free disease connection and network connectivity. It includes an approximation of the statistical importance of this integrated score.

#### 2.2.9 PathFindR

The pathfindR is a tool that uses active subnetworks to analyze pathway enrichment. It defines gene sets forming active subnetworks in a network of protein-protein interactions using the user's list of genes. It conducts pathway enrichment analysis on the gene sets which are recognized. It also maps user information on the enhanced pathways using the R package pathview and makes path diagrams together with the mapped genes. Since many of the enhanced pathways are generally biologically linked, pathfindR also provides features for clustering these pathways and identifying representative pathways in clusters. This program has three search algorithms, i.e., greedy, simulated annealing, and genetic algorithm and its scoring scheme is based on the z-score. In this study, pathFindR is run with each algorithm**.**

### 2.3 Functional enrichment analysis

After the identification of the highest-scoring active sub-networks, interpretation of results from a biological perspective is very important. Functional Enrichment Analysis is a common technique used to show whether this sub-network is biologically meaningful [34],[35]. In this technique, using hypergeometric test and Bonferroni correction, the set of identified genes in the subnetwork are compared with the set of genes that are known to be part of a biological pathway or a Gene Ontology (GO) term. Hence, the biological relevance of the identified sub-network in terms of disease development is assessed. In this study, a widely used functional enrichment program, BINGO is used. GO terms including BP (Biological Process), MF (molecular function), and CC (Cellular Component) are preferred for functional enrichment due to our data set.

### 2.4 Fold enrichment

Another method used to interpret active sub-network search program results is Fold Enrichment. This test is important to find the overlap rate between the identified subnetworks and the reference dataset [17]. In this study, the reference set is selected as the idiopathic generalized epilepsy genes, known in the literature. Fold enrichment is calculated as in (Eq 1);

A= Total number of genes overlapped in the current module with reduced reference data set that consists of reducing repetitive genes to a single gene,
B= Total number of nodes in the reduced protein-protein interaction dataset that consists of reducing repetitive genes to a single gene,
C= Total number of genes overlapped in the reduced protein-protein interaction dataset and reduced reference dataset,
D= Total number of nodes in the modüle.

$$FoldEnrichment = \frac{A\,.\,B}{C\,.\,D} \tag{1}$$

Our protein-protein interaction network includes 10175 nodes (proteins). The reference data set contains 224 proteins associated with IGE and 151 of these proteins exist in the PPI network. So, in our study, (Eq 1) is replaced with (Eq 2).

$$FoldEnrichment = \frac{A\,.\,10175}{151\,.\,B} \tag{2}$$

### 2.5 Normalized mutual information (NMI)

As a performance measure for assessing the predicted modules of the sub-network search algorithms, we are using the normalized-mutual information (NMI). The NMI is important because it gives information about the consistency of the results between different methods vice versa it does not allow to get an idea of the absolute quality of the identified subnetworks [18],[19].

Let us show two of the subnetwork identification methods with U and V. Let's assume that these methods predict |R| and |C| subnetworks and that the common protein numbers within these identified subnetworks are shown in Table 1.

Table 1. A contingency table which defines the overlap between two methods, U and V.

| U/V | $V_1 \ldots V_2 \ldots V_C$ | SUMs |
|---|---|---|
| $U_1$ | $N_{11} \ldots n_{12} \ldots n_{1c}$ | $a_1$ |
| $U_2$ | $N_{21} \ldots n_{22} \ldots n_{2c}$ | $a_2$ |
| . | . | . |
| . | . | . |
| . | . | . |
| $U_R$ | $n_{R1} \ldots n_{R2} \ldots n_{RC}$ | $a_R$ |
| SUMs | $b_1 \ldots b_2 b_c$ | N |

i.e., the second subnetwork of U method $U_2$ and the first subnetwork of V method $V_1$ share $n_{21}$ common proteins.

The normalized mutual information is calculated as shown in (Eg 6) using (Eg 3), (Eg 4), and (Eg 5) [36].

$$H(U) = -\sum_{i=1}^{R} \frac{a_i}{N}\left(\log\frac{a_i}{N}\right) \tag{3}$$

$$H(V) = \sum_{i=1}^{C} \frac{b_i}{N}\left(\log\frac{b_i}{N}\right) \tag{4}$$

$$I(U,V) = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{N}\left(\log\frac{n_{ij}/N}{a_i b_j/N^2}\right) \quad (5)$$

$$NMI = \frac{I(U,V)}{H(U)+H(V)} \quad (6)$$

Via normalizing the calculated mutual information, the NMI value is converted to a range of [0, 1]. While the value zero in NMI indicates that the subnetworks identified by the related methods are independent of each other, the value one in NMI indicates that the subnetworks identified by two different methods are the same with each other.

## 3 Results

In this study, nine different active subnetwork search programs are comparatively evaluated using an epilepsy-associated GWAS data set to help understand the underlying mechanism of this disease. The evaluated programs are jActiveModules, ActiveSubnetworkGA, ClusterViz, PINBPA, MCODE, PEWCC, CytoHubba, CytoMOBAS, and PathFindR. As shown in Table 2, most of these programs are implemented in Java and most of them have GUI support using Cytoscape and R-studio. Also in Table 2, the algorithms shown in dark are used in this study. The running time results of programs are shown in Table 3, and they are obtained using an HP-TPN-C129 computer. Any applications were blocked from running in the background while a program was run. All programs run with their default parameters except CytoHubba. In CytoHubba only Top 50 node(s) are ranked by MCC. As a result the programs running stage, we compared the obtained modules according to certain criteria and made inferences about which active subnet program could be more efficient for epilepsy disease. In terms of running time, pathFindR has the best performance (the average running time of its three different algorithms), as shown in Table 3. For this reason, data sets with big data input and parameters suitable for the program, pathFindR should be preferred. After the running stage is completed, functional enrichment analysis is applied on the modules which have the maximum score of each program (one module is selected for each program) and reference data set. The aim of this analysis is to measure the importance of the identified sub-networks, and to compare the set of identified genes with the set of genes that are known to be part of a biological pathway. In another sense, using GO terms, the biological relevance of the identified sub-network in terms of disease development can be assessed. If diversity and accuracy of this obtained biological relevance increase, also treatment options developed for the epilepsy disease increases. For these reasons, the identification of different types of GO terms is critical for understanding disease biological mechanisms and inferring gene function. Figure 2 shows the total number of GO terms of the modules and the reference data set. According to Figure 2. ActiveSubnetworkGA has the highest value with 1116 different GO terms so it can identify the different types of the GO Terms for the top-scoring subnetworks.

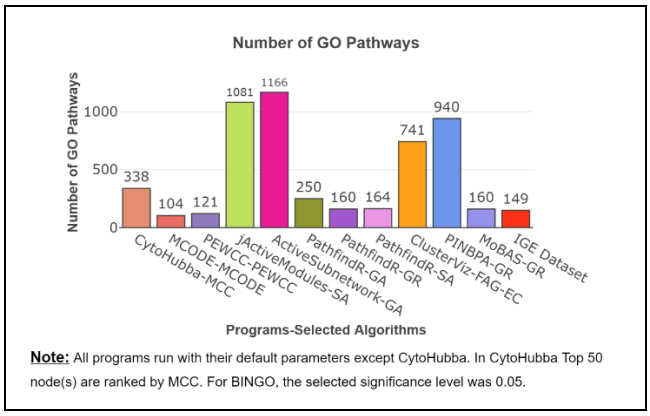


Figure 2. Number of the GO Terms identified for the top-scoring subnetworks of different algorithms.

As the third and fourth stage, fold enrichment and normalized mutual information analysis are performed for each program. In this respect, five modules that have the highest score of each program are selected to set a common input limit and to obtain meaningful results. Using fold enrichment analysis, we wanted to find overlapping rates between the selected subnetworks which are outputs of programs, and the reference dataset (224 genes) which is called idiopathic generalized epilepsy. For every five subnetworks of programs, fold enrichment results are computed and as a result, the average value is shown in Figure 3. The CytoMoBAS has the highest fold enrichment result. This result means that the subnetworks defined by CytoMoBAS overlap the reference dataset more than the subnetworks of other programs.

Table 2. Features of the sub-network identification programs.

| Programs | Programming Languages | Interface | Algorithms |
|---|---|---|---|
| PINBPA | Java | Cytoscape | Greedy Algorithm (GR) |
| CytoMOBAS | Java | Cytoscape | Greedy Algorithm (GR) |
| PathFindR | Java or R | Command-Line or R-Studio | Genetic (GA), Simulated Annealing (SA) and Greedy Algorithm (GR) |
| PEWCC | Java | Cytoscape | PEWCC |
| jActiveModules | Java | Cytoscape | Simulated Annealing (SA) and Genetic Algorithm (GA) |
| MCODE | Java | Command-Line or R-Studio | MCODE |
| CytoHubba | Java | Cytoscape | 11 different topological analysis methods.<br>MCC |
| ActiveSubnetworkGA | Java | Command Line | Genetic Algorithm (GA) |
| ClusterViz | Java | Cytoscape | FAG-EC, EAGLE, and MCODE |

Note: All programs run with their default parameters except CytoHubba. In CytoHubba Top 50 node(s) are ranked by MCC. To to run all sub-network search programs, a human PPI network and epilepsy-related GWAS dataset are used.

Table 3. Running time analyzes of the compared subnetwork identification programs based on hrs:mins:sec.

| Programs-Algorithms | Running Times |
|---|---|
| PINBPA-GR | 05:16:03 |
| CytoMOBAS-GR | 08:40:19 |
| PathFindR-GA/GR/SA | 00:00:27 |
| PEWCC-PEWCC | 00:13:22 |
| JActiveModules-SA | 00:08:04 |
| MCODE-MCODE | 00:10:44 |
| CytoHubba-MCC | 00:07:42 |
| ActiveSubnetwork-GA | 00:32:50 |
| ClusterViz-FAG-EC | 00:52:18 |

Note: All programs run with their default parameters except CytoHubba. In CytoHubba Top 50 node(s) are ranked by MCC.

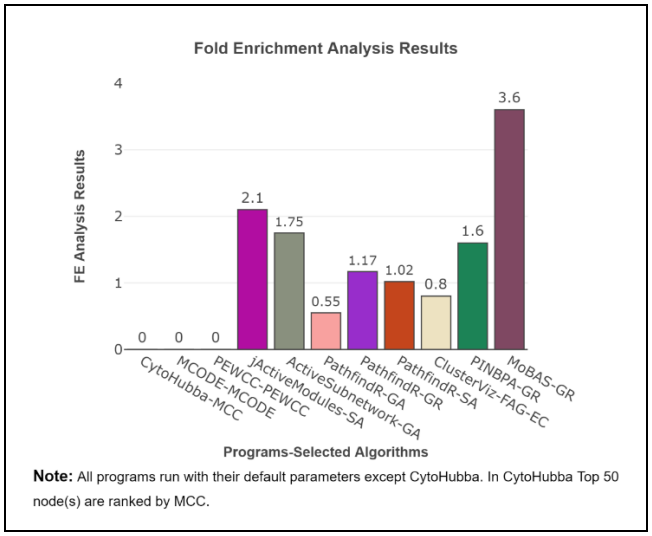


Figure 3. Fold Enrichment Analysis results of the subnetwork identification programs.

Lastly, We also wanted to assess the consistency of the subnetworks obtained via different search programs. For this goal, we used the selected subnetworks which are used in the fold enrichment analysis and applied the normalized mutual information on modules. In Figure 4, the NMI results are presented with a heatmap. NMI value is obtained in the range [0, 1].

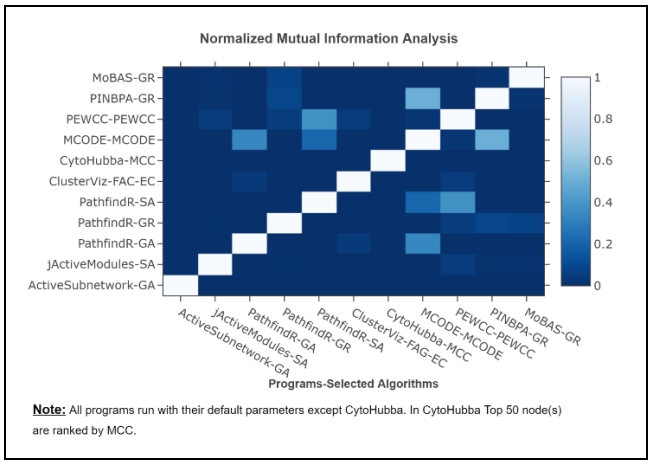


Figure 4. Normalized Mutual Information Analysis of the subnetworks identified via different programs.

They are are color-coded from dark blue (NMI=0) to light blue (NMI=1). While the value zero of NMI indicates that the subnetworks identified by the related methods are independent of each other, the value one of NMI indicates that the subnetworks identified by the related methods are the same as each other. As shown in Figure 4, we obtain the consistency of results between the methods according to the decreasing values as follows; PINBPA-MCODE, PEWCC-PathFindRSA, MCODE-PathFindRGA, MCODE-PathFindRSA. As a result, compared to others, the consistency of the results between the PINBPA-MCODE and PathFinRSA-PEWCC is the closest to 1 with a 0.5. Except for these, there is no consistency of results between other methods. It means that the results of PINBPA-MCODE and PathFinRSA-PEWCC methods are consistent. It may make sense to choose these binary method combinations in analyzes that require comparison.

## 4 Conclusions

Active sub-networks in the protein-protein interaction networks provide an insight into the basic principles of disease formation and development mechanisms. With the detection of related genes behind diseases, different treatments and the development of appropriate methods can be developed. For such reasons, active module finding programs get very popular. To define active subnetworks properly, the effective comparison of existing methods and the use of such analysis results while developing new active subnetwork search methods are necessary. For this purpose, in this study, we wanted to evaluate existing subnetwork search programs from different perspectives and get an idea about them. In summary, nine different active subnetwork search programs and their different types of algorithms are run. Firstly, features and the running time of programs are compared. Then the subnetworks are identified for all programs and we performed three evaluation analyses on the results, i.e., functional enrichment analysis, fold enrichment analysis, and normalized mutual information analysis. We used the functional enrichment analysis to show whether the founded sub-network is biologically meaningful.

We also aimed to find out how many proteins in the subnetworks overlap with known epilepsy genes. In this regard, fold enrichment analysis is applied. Finally, normalized mutual information analysis is performed to examines information about the consistency of the results between different methods.

According to the study results we've achieved with the epilepsy data set, we can list our inferences as follows. In cases such as data sets with big data input and parameters suitable for the program, pathFindR may be preferred. ActiveSubnetworkGA has the highest number of GO Terms so it can identify the different types of GO Terms for the top-scoring subnetworks. The CytoMoBAS has the highest fold enrichment result so the subnetworks defined by CytoMoBAS overlap the reference dataset more than the subnetworks of other programs. Finally, compared to other programs, the results of PINBPA-MCODE

and PathFinRSA-PEWCC methods are consistent. So it may make sense to choose these binary method combinations in analyzes that require comparison. To help users choose a program based on their needs, the results of other sub-network search methods can be compared in future studies. Furthermore, similar analyzes can be made that consist of different lots of disease data sets using the identified proteins can be used as biomarkers, and some proteins in these sub-networks can be targeted for drug development studies.

## 5 Author contribution statements

In this work, the Burcu BAKIR GUNGOR contributed by her idea, designing article, spell check and evaluation of results; the Beyhan ADANUR DEDETURK contributed by literature review, running experiments, selection of parameters, writing the article.

## 6 Ethics committee approval and conflict of interest statement

The article does not require permission from the ethics committee and there is no conflict of interest with any person/institution.

## 7 References

[1] Zhang L, Li Y, Ye X, Bian L. "Bioinformatics analysis of microarray profiling identifies that the miR-203-3p target Ppp2ca aggravates seizure activity in mice". *Journal of Molecular Neuroscience*, 66(1), 146-154, 2018.

[2] Nguyen H, Shrestha S, Tran D, Sha A, Draghici SmNguyen, et al. "A comprehensive survey of tools and software for active subnetwork identification". *Frontiers in Genetics*, 10(155), 1-15, 2019.

[3] Ozisik O, Bakir-Gungor B, Diri B, Sezerman OU. "A genetic algorithm approach to active subnetwork search applied to GWAS data". *In: 2013 8th International Symposium on Health Informatics and Bioinformatics*, Ankara, Turkey, 25-27 September 2013.

[4] Bakir-Gungor B, Baykan B, I_seri SU, Tuncer FN, Sezerman OU. "Identifying SNP targeted pathways in partial epilepsies with genome-wide association study data". *Epilepsy Research*, 105(1-2), 92-102, 2013.

[5] Mitra K, Carvunis AR, Ramesh SK, Ideker T. "Integrative approaches for finding modular structure in biological networks". *Nature Reviews Genetics*, 14(10), 719-732, 2013.

[6] Nikolayeva I, Pla OG, Schwikowski B. "Network module identification-A widespread theoretical bias and best practices". *Methods*, 132, 19-25, 2008.

[7] Ideker T, Ozier O, Schwikowski B, Siegel AF. "Discovering regulatory and signaling circuits in molecular interaction Networks". *Bioinformatics,* 18(1), 233-240, 2002.

[8] Wang L, Matsushita T, Madireddy L, Mousavi P, Baranzini SE. "PINBPA: Cytoscape app for network analysis of GWAS data". *Bioinformatics*, 31(2), 262-264, 2014.

[9] Su G, Morris JH, Demchak B, Bader GD. "Biological network exploration with Cytoscape 3". *Current Protocols in Bioinformatics*, 47(1), 8-13, 2014.

[10] Zaki N, Emov D, Berengueres J. "Protein complex detection using interaction reliability assessment and weighted clustering coefficient". *BMC Bioinformatics*, 14(163), 1-9, 2013.

[11] Ozisik O, Bakir-Gungor B, Diri B, Sezerman UO. "Active subnetwork GA: A two stage genetic algorithm approach31 to active subnetwork search". *Current Bioinformatics*, 12(4), 320-328, 2017.

[12] Wang J, Zhong J, Chen G, Li M, Wu FX, Pan Y. "clusterviz: a cytoscape APP for cluster analysis of the biological network". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4), 815-822, 2014.

[13] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. "cytoHubba: identifying hub objects and sub-networks from complex interactome". *BMC Systems Biology*, 8(4), 4-11, 2014.

[14] Ayati M, Erten S, Chance MR, Koyutu RK M. "MOBAS: identification of disease-associated protein subnetworks using modularity-based scoring". *EURASIP Journal on Bioinformatics and Systems Biology*, 2015(1), 1-14, 2015.

[15] Ulgen E, Ozisik O, Sezerman OU. "pathfindR: An R Package for pathway enrichment analysis utilizing active subnetworks". *BioRxiv*, 2018. https://doi.org/10.1101/272450.

[16] Maere S, Heymans K, Kuiper M. "BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks". *Bioinformatics*, 21(16), 3448-3449, 2006.

[17] He H, Lin D, Zhang J, Wang YP, Deng HW. "Comparison of statistical methods for subnetwork detection in the integration of gene expression and protein interaction network". *BMC Bioinformatics*, 18(1), 149, 1-6, 2017.

[18] Tripathi S, Moutari S, Dehmer M, Emmert-Streib F. "Comparison of module detection algorithms in protein networks and investigation of the biological meaning of predicted modules". *BMC Bioinformatics*, 17(1), 1-18, 2016.

[19] Taya F, de Souza J, Thakor NV, Bezerianos A. "Comparison method for community detection on brain networks from neuroimaging data". *Applied Network Science,* 1(1), 1-20, 2016.

[20] Adanur B, Gungor BB. "Comparison of disease-specific sub-network ıdentification programs". *In 2018 3rd International Conference on Computer Science and Engineering,* Sarajevo, Bosnia, 20-23 September 2018.

[21] Zhang Q, Zhang ZD. "SubNet: a Java application for subnetwork extraction". *Bioinformatics*, 29(19), 2509-2511, 2013.

[22] Chen X, Xuan J. "MSIGNET: a Metropolis sampling-based method for global optimal significant network identification". *BioRxiv*, 2018. https://doi.org/10.1101/260844.

[23] Farahmand S, Foroughmand-Araabi MH, Goliaei S, Razaghi-Moghadam Z. "CytoGTA: a cytoscape plugin for identifying discriminative subnetwork markers using a game-theoretic approach". *PloS one*, 12(10), 1-12, 2017.

[24] Shi X, Barnes RO, Chen L, Shajahan-Haq AN, Hilakivi-Clarke L et al. "BMRF-Net: a software tool for identification of protein interaction subnetworks by a bagging Markov random field-based method". *Bioinformatics*, 31(14), 2412-2414, 2015.

[25] Wang Q, Yu H, Zhao Z, Jia P. "EW_dmGWAS: edge-weighted dense module search for genome-wide association studies and gene expression proles". *Bioinformatics*, 31(15), 2591-2594, 2015.

[26] Ma H, Schadt EE, Kaplan LM, Zhao H. "COSINE: COndition-SpecIc sub-NEtwork identification using a global optimization method". *Bioinformatics*, 27(9), 1290-1298, 2011.

[27] Akhmedov M, Kedaigle A, Chong RE, Montemanni R, Bertoni F, Fraenkel E, Kwee I. "PCSF: An R-package for network-based interpretation of high-throughput data". *PLoS Computational Biology*, 13(7), 1-7, 2017.

[28] Segre AV, Groop L, Mootha VK, Daly MJ, Altshuler D et al. "Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits". *PLoS Genetics*, 6(8), 1-19, 2010.

[29] Kasperaviciute D, Catarino CB, Heinzen EL, Depondt C, Cavalleri GL et al. "Common genetic variation and susceptibility to partial epilepsies: a genome-wide association study". *Brain*, 133(7), 2136-2147, 2010.

[30] Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. "A human protein-protein interaction network: a resource for annotating the proteome". *Cell,* 122(6), 957-968, 2005.

[31] Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A et al. "Towards a proteome-scale map of the human protein-protein interaction network". *Nature*, 437(7062), 1173-1178, 2005.

[32] Novo Nordisk Foundation Center for Protein Research. "DISEASES (Disease-Gene Associations Mined From Literature)". https://diseases.jensenlab.org (08.03.2021).

[33] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks". *Genome Research* 13(11), 2498-2504, 2003.

[34] Manda S, Michael D, Jadhao S, Nagaraj, SH. "Functional enrichment analysis". *Encyclopedia of Bioinformatics and Computational Biology*, 2019. https://doi.org/10.1016/B978-0-12-809633-8.20097-6.

[35] Pietro H. Guzzi. "Functional Enrichment Analysis Methods". *Encyclopedia of Bioinformatics and Computational Biology,* 2019. https://doi.org/10.1016/B978-0-12-809633-8.20404-4.

[36] Vinh NX, Epps J, Bailey J. "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance". *Journal of Machine Learning Research*, 11, 2837-2854, 2010.