International Journal of
Applied Mathematics,
Electronics and Computers
ISSN: 2147-8228 http://ijamec.atscience.org
Advanced Technology and Science
Original Research Paper

Network motif detection in PPI networks and effect of R parameter on system performance

Yilmaz Atay*[1], Halife Kodaz[2]

Accepted 10th August 2015 **DOI: 10.18100/ijamec.05406**

Abstract: Bioinformatics is an area on which lots of researches have been done and it is used widely in drug design, cancer treatment, disease detection, functional analysis, phylogenetic research and food and cell mutation changes. The field benefits from the partnership of disciplines like computer science, biology, genetics, mathematics and statistics. The incredible rise in areas of computer science such as artificial intelligence, graph theory, software and hardware technology and kind of has provided also a direct contribution to Bioinformatics. The subject of this study is to examine the *FANMOD* tool which discovers and analyses network motifs and to carry out performance analysis which is depending on the *R* (number of random networks) parameter. Four different protein-protein interaction (*PPI*) networks were used in the experiments. The *R* parameter used in the experiments were taken as 100, 1000, 10000 and 100000 and the obtained results are presented both graphically and quantitatively. Furthermore the average experiment time of each network and characteristic features of the obtained motifs are given at the end of the paper.

Keywords: *Bioinformatics, Complex networks, FANMOD, Graph isomorphism, Network motif detection, Protein-protein interaction networks, System performance analysis*.

1. Introduction

Complex networks are defined as networks in which relationships between the nodes of the network cannot be achieved with simple tools, many nodes having complex relationship with each other and having lots of nodes and edges. Biological, economical, physical, social and technological networks are some of complex networks. Examples of complex networks are given in Table 1.

Table 1. Complex networks

Networks	Nodes	Edges
Biological networks	Molecular Structures	Functional relationships
Internet	Computers	Network connections
World Wide Web	Web pages	Links
Ecosystem	Life forms	Relationships
Transportation systems	Locations	Ways (Connections)
Social networks	People	Social relationships

Many critical functions performed by organisms are managed by a complex network of interactions between the various biochemical molecules [1]. These biological networks are the basis of this study. Biological networks are concerned with molecules such as DNA, RNA, protein and genes and also their relation to one another. There are many important biological networks. For example, regulatory networks, metabolic networks, signal transduction networks, networks of protein structure, protein-protein interaction (*PPI*) networks and etc. [2]. We used *PPI* networks in our experiments. Understanding the operational functions which is a result of the interactions in biological networks depends on understanding the structures called network motifs [3]. Therefore to detect network motif, it is often needed to solve the subgraph isomorphism problem which is defined as NP-complete [4, 5, 12].

In this study *FANMOD* [6] which effectively solves the above problem was applied on the *PPI* networks which are given in test data section. The most important issue focused in this study is that, to what extent *FANMOD* tool uses *CPU* and *RAM* resources with different *R* parameter values. Network motif detection, *FANMOD*, test data, experimental results and conclusions are respectively the next sections of the study.

2. Network Motif Detection

Subgraphs found in complex networks, are statistically significant and the number of its available in destination network is greater than the average number of its available in random networks are called network motifs [3].

Sample candidate motif was given in Fig. 1. In the figure, the structure can be considered as a candidate motif with 0-2-7 members.

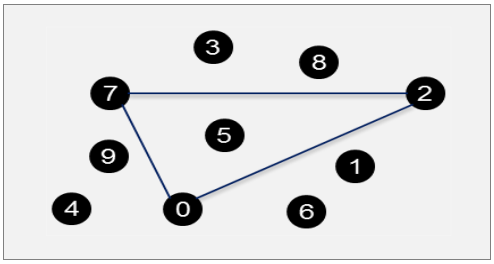

Figure 1. Sample candidate motif

[1,2] *Computer Engineering Department, Engineering Faculty, Selçuk University, Campus, 42031, Konya/Turkey*
** Corresponding Author: Email: hkodaz@selcuk.edu.tr*
Note: This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'15) held in Antalya (Turkey), August 04-07, 2015

2.1 Frequency Concepts

Network motifs are identified by three different frequency concepts in the literature [2]. Different methods are designed to find motifs with different concepts. Sample candidate motif and network are given in Fig. 2. In the network these substructures (subgraphs) has been named as FC-1, FC-2, FC-3, FC-4 and FC-5 respectively. Concepts and properties of F1, F2, F3 are given in Table 2. In this study, we have tried to find suitable motifs according to F1 concept. These motifs are composed of substructures of size-3.

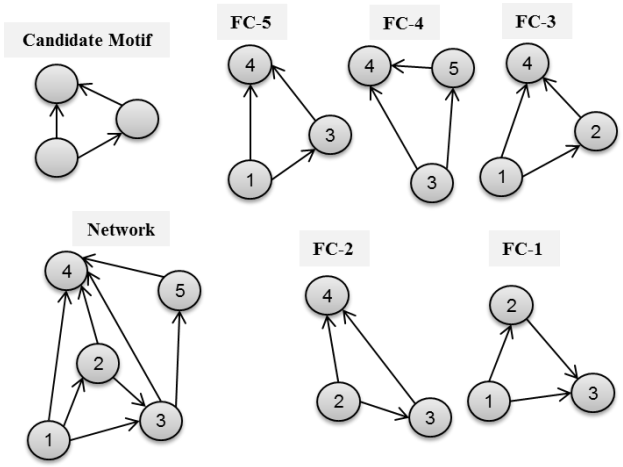

Figure 2. Candidate motifs in the destination network

Table 2. Motif frequencies under different frequency concepts

Concepts	Node Overlap	Edge Overlap	Frequency	Candidate Motifs
F_1	✓	✓	5	FC-1, FC-2, FC-3, FC-4, FC-5
F_2	✓	×	2	Either {FC-1, FC-4} or {FC-3, FC-4}
F_3	×	×	1	One of {FC-1, FC-2, FC-3, FC-4, FC-5}

2.2 General Algorithm about Network Motif Detection

Network motifs when compared to a random graph with a similar degree distribution are higher in number at a given graph and significant substructures [13]. The methods and tools developed for the detection of these structures generally include the following steps:

- Destination network simulation
- Definition of frequency concepts
- Generating of random networks (*R*)
- Candidate motifs identification and frequency census
- Performing statistical significance tests
- Motif identification and so on.

In general program flow in Fig. 3 is the basis of algorithms and tools which are used in the determination of motifs in destination network [8].

Require: Graph G and integers k and R
Ensure: Motifs of size k in graph G

1: SUBGRAPHCENSUS(k, G)
2: **for** $i := 1$ to R **do**
3: $R_i :=$ GENERATESIMILARRANDOMNETWORK(G)
4: SUBGRAPHCENSUS(k, R_i)
5: CALCULATESIGNIFICANCEMOTIFS()

Figure 3. General program flow for network motif detection [8]

2.3 *FANMOD*

FANMOD tool which was recommended by Wernicke and Rasche in 2006 is very fast and efficient network motif detection tool [6]. This tool can be applied to both directed and undirected networks and can be detected up to size-8 motifs. This algorithm offers a new approach which is related to the unbiased sampling method and only works for induced subgraphs in the networks, neglecting non-induced subgraphs. This tool can also be applied to color graph structures. *FANMOD* uses RAND-ESU [7] approach which was proposed in 2005 for motif census and sampling (to enumerate and sample subgraphs).
RAND-ESU is based on the ESU search tree. Thus the motif search method of a network which consists of 5 nodes (Fig. 4) should be examined according to the search tree given in Fig. 5 [8].

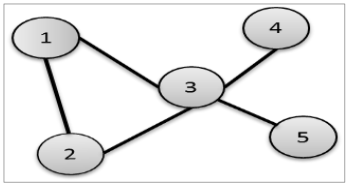

Figure 4. Sample network (5 nodes)

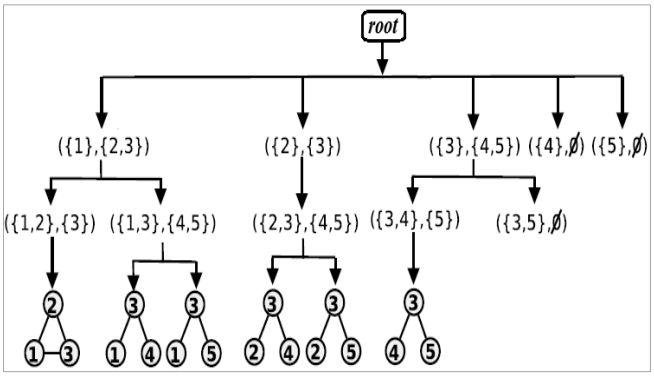

Figure 5. ESU search tree of size-3 candidate motifs in the network [8]

The pseudo code of RAND-ESU is given in Fig. 6 [8, 9].

Input : A graph G = (V, E), integer k>0

'RAND-ESU function'

{

For each vertex v ∈ V(G) do

$V_{Ext} := \{u \in N(v) : u > v\}$

with probability P_1 EXTENDSUBGRAPH$(\{v\}, V_{Ext}, v)$

}

procedure EXTENDSUBGRAPH(V_{Subg}, V_{Ext}, v)

{

If $|V_{Subgraph}| = k$ **then**

INCREMENTCOUNT(canonicalLabeling(V_{Subg}))

Else

While $V_{Extension} \neq \emptyset$ **do**

remove random chosen $w \in V_{Ext}$

$V'_{ext} := V_{ext} \cup \{u \in N_{excl}(w, V_{subg}) : u > v\}$

$V'_{Subg} := V_{Subg} \cup \{w\}$

with prob. $P_{|V'_{Subg}|}$ EXTENDSUBGRAPH$(V'_{Subg}\}, V'_{Ext}, v)$

}

Figure 6. RAND-ESU Algorithm [8], [9]

3. Test Data

In experiments 4 different *PPI* networks were used. These networks are protein interactions relative to living creatures like Caenorhabditis Elegans, E. Coli, Helicobacter Pylori and Saccharomyces Cerevisiae (yeast) respectively [10]. General properties of networks like source of data, number of nodes and edges have been given in Table 3. The data which are used in this study were taken from *txt* format and converted to integer numbers between 1 to *n* representatively for specific application (*n* represents the maximum number of node) [10].

Table 3. Protein-protein interactions (data sets features)

Name	Num. of Nodes	Num. of Edges	Source and Date
Caenorhabditis Elegans	375*	437	*DIP*, 2000-2001
E. Coli	270	716	*The Emili Lab*, 2005
Helicobacter Pylori	732	1465	*DIP*, 2000-2001
Saccharomyces Cerevisiae (yeast)	4142	8099	*DIP*, 2000-2001

* In the program this network was considered to have 375 nodes.

4. Experimental Results

In this study, by using *FANMOD* tool motif of size-3 have been found in different four *PPI* networks in very short period of time. The most important objective in this study is to test random network number (*R*) parameters (to be 100, 1000, 10000 and 100000) in all networks which plays a very important role while testing the subgraphs' statistical significance that is taken as candidate motifs. Value of *R* parameter is generally taken as 100 [11] and 1000 [8] by default. In this study effect of CPU and RAM resources of *FANMOD* tool in different *R* values were investigated and the results were presented graphically (in the figures).

Experiments have made on a computer which has a 4 GB of memory (RAM) and a processor (CPU) speed of 3.10 GHz. Total runtimes (in seconds) for each network according to different *R* values are given in Table 4.

Table 4. Total runtimes according to different *R* values (*s*)

Name/*R*	100	1000	10000	100000
Caenorhabditis Elegans	0.191	1.886	18.66	186.924
E. Coli	0.617	6.043	60.063	603.127
Helicobacter Pylori	1.203	11.86	118.164	1199.249
Saccharomyces Cerevisiae (yeast)	11.835	116.974	1172.041	11708.69

When Table 4 is carefully examined, as number of edges in the network increases, runtime of program also increases rapidly. However, the increment is not parallel. For example, the number of edges of Caenorhabditis Elegans living creature is 437 and the number of edges of E. Coli living creature is 716. According to (*716:437*) ratio the running time of the program is expected to increase to about 2 times more. But as seen in Table 4 the increment is between 3 and 6 times. This ratio has increased more with the increase of the number of edges in the network (This ratio is about 11 times when Saccharomyces Cerevisiae (yeast) living creature compared with Helicobacter Pylori).

In the experiments, each candidate motif is represented by the *ID* according to the number between 1 to *m* (*m* represents the maximum number of candidate motifs). Obtained results from experiments were kept on a *txt* file. Name of this file is *NetworkName_RNumber_Results.txt*.

For example, for Caenorhabditis Elegans living creature and *R*=100 the obtained results were kept in the file by name of *elegans_R100_Results.txt*. As a result of experiments the motifs (According to the *ID* number) and their numbers for every networks is given in Table 5.

Table 5. The obtained motif IDs and numbers

Name - Number of Motif	Motifs (*ID*)
Caenorhabditis Elegans - **10**	6, 14, 36, 12, 78, 164, 38, 46, 174, 102
E. Coli - **4**	36, 12, 6, 38
Helicobacter Pylori - **4**	6, 36, 12, 38
Saccharomyces Cerevisiae (yeast) - **13**	6, 12, 36, 14, 38, 164, 140, 46, 78, 174, 102, 238, 166

Another program was coded in Microsoft .NET environment to determine the usage rates and amount of system resources. Utilization of system resources (CPU and RAM) by *FANMOD* tool according to different *R* values for each interaction network is shown in following figures. Each figure was created according to 4 different *R* values. Please refer to; for Caenorhabditis Elegans Fig. 7 and Fig. 8, for E. Coli Fig. 9 and Fig. 10, for Helicobacter Pylori Fig. 11 and Fig. 12, for Saccharomyces Cerevisiae (yeast) Fig. 13 and Fig. 14. When the results of experiments are examined, there is no sufficient change in CPU

usage as the network size increases but the usage of RAM increases rapidly.

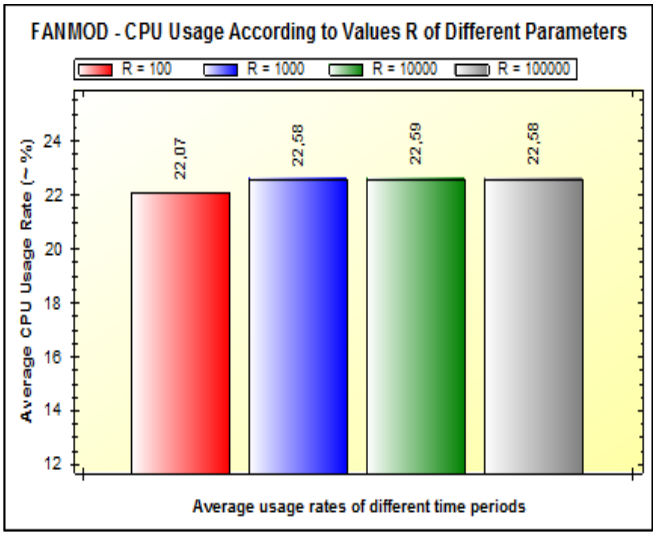

Figure 7. *Caenorhabditis Elegans PPI* network - CPU usage rates

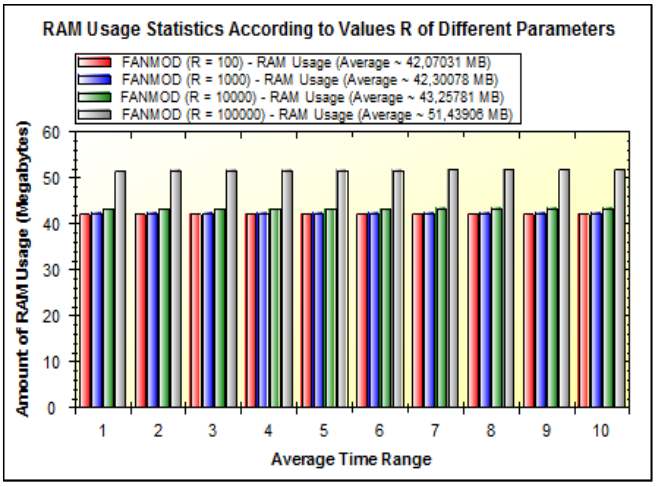

Figure 8. *Caenorhabditis Elegans PPI* network - RAM usage

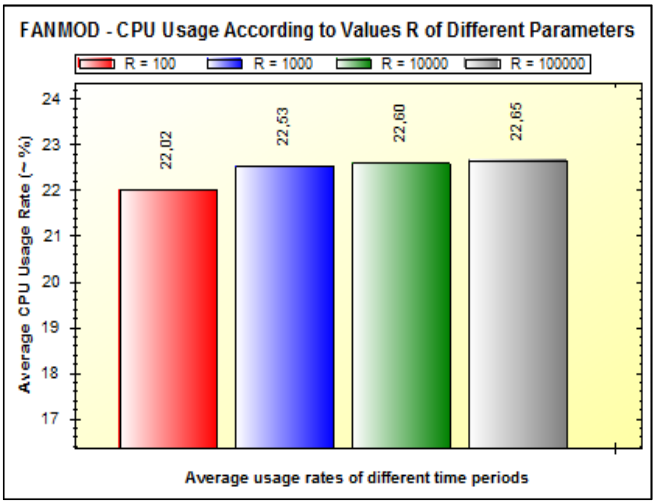

Figure 9. *E. Coli PPI* network - CPU usage rates

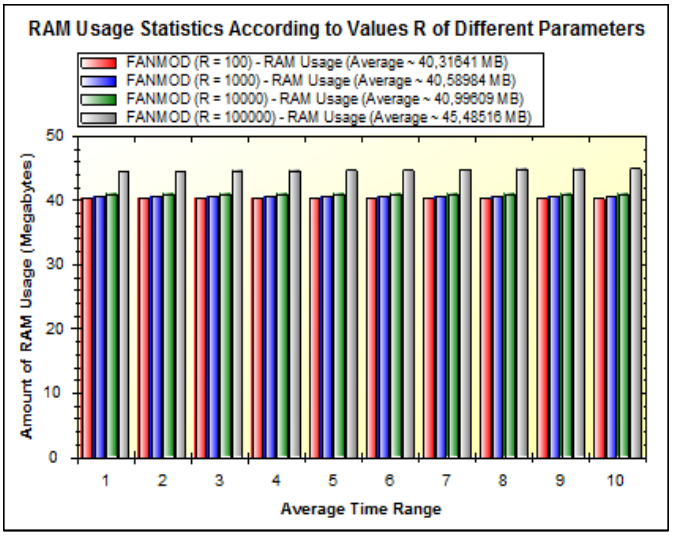

Figure 10. *E. Coli PPI* network - RAM usage

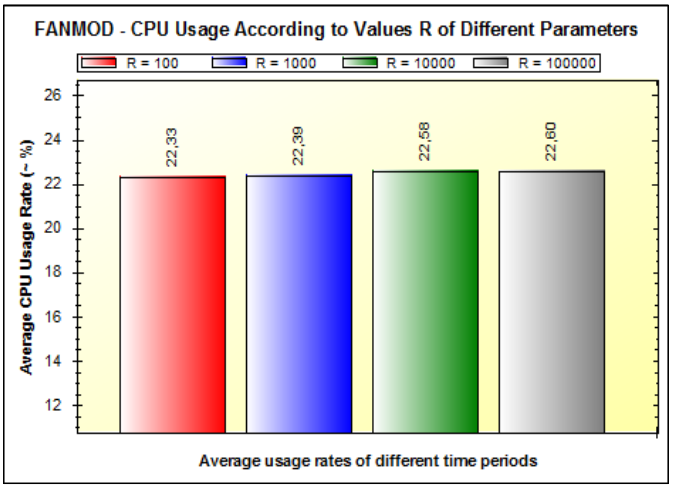

Figure 11. *Helicobacter Pylori PPI* network - CPU usage rates

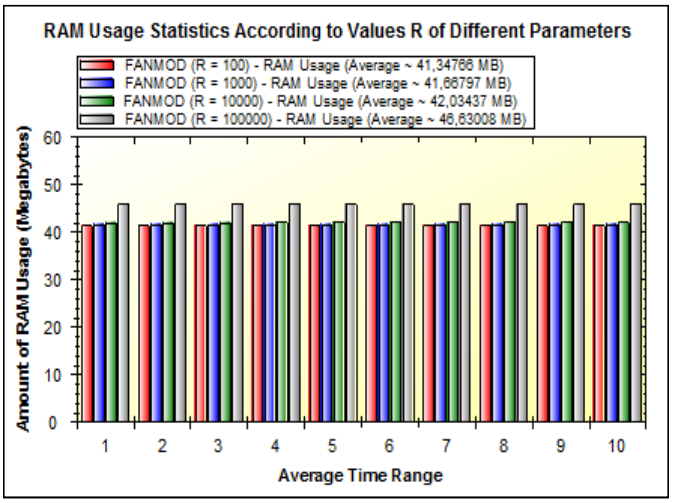

Figure 12. *Helicobacter Pylori PPI* network - RAM usage

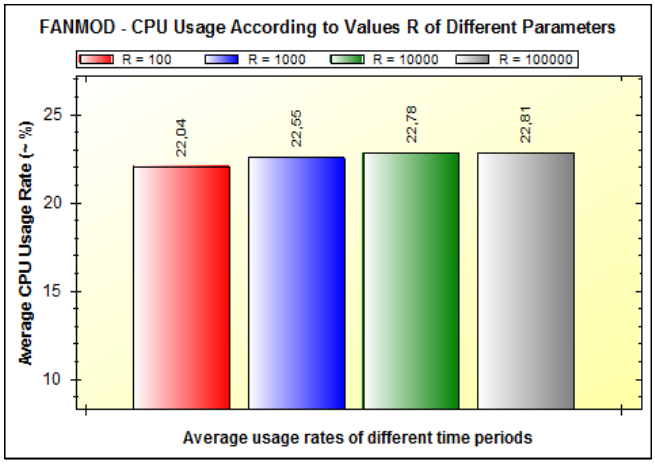

Figure 13. *Saccharomyces Cerevisiae (yeast) PPI* network - CPU usage rates

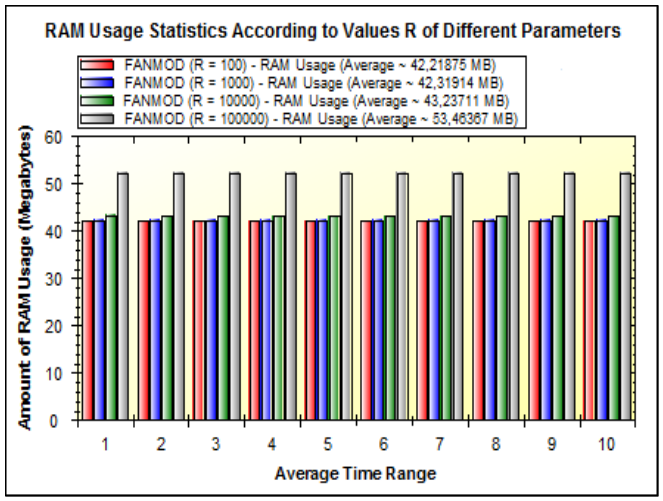

Figure 14. *Saccharomyces Cerevisiae (yeast) PPI* network - *RAM usage*

The ranges of CPU usage rate for all networks are usually between 22% and 23% when the results are analysed. It is obvious from the results above that the network size is not significantly affected by the CPU usage rate. RAM usage amount varies greatly up on the network sizes. While E. Coli and Helicobacter Pylori interaction networks use less than 50 *MB* of memory (for *R*=100000), Caenorhabditis Elegans and Saccharomyces Cerevisiae (yeast) interaction networks use more than 50 *MB* of memory capacity.

5. Conclusions

This study focuses on subject of network motif detection which is often used for biological networks analysis. For this purpose, *FANMOD* motif detection tool which gives acceptable results and is used quite often in literature was used in the experiments. In experiment for each of 4 different networks different four *R* parameters were used and obtained results were given in the experimental results section. When results of this study are examined it's observed that as the network size increases RAM capacity is an important source for network motif detection but as CPU usage rate does not change sufficiently, it does not play a valuable role on motif detection. However, beside CPU usage rate, multi-core systems and parallel working environments save a great amount of time while detecting network motifs and its examples are available in literatures.

In further studies, different search methods can be improved adhering to the existing network features algorithms like representing network in the memory, concept selection and random network production. Capacity utilization of multi-dimensional network of system memory (RAM) can be tested and compared to other algorithms.

Acknowledgements

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK, 2211-C Domestic Doctoral Scholarship Program Intended for Priority Areas, Project No. *1649B031402383*) and Selcuk University OYP Coordination Unit (Project No. *2013-OYP-057*).

References

[1] Zhu, X., Gerstein, M. and Snyder, M., Getting connected: analysis and principles of biological networks, Genes & Dev, 21: 1010-1024, 2007.

[2] Wong, E., Baur, B., Quader, S. and Huang, C-H., Biological network motif detection: principles and practice, Briefings in bioinformatics, Vol 13, No 2, 202-215, 2011.

[3] Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D. and Alon U., Network motifs: Simple building blocks of complex networks, Science, 298:824–827, 2002.

[4] Grochow, J. A. and Kellis, M., "Network motif discovery using subgraph enumeration and symmetry-breaking." Research in Computational Molecular Biology. Springer Berlin Heidelberg, 2007.

[5] Hasan, M. M., Kavurucu, Y. and Kahveci, T., A scalable method for discovering significant subnetworks, BMC systems biology, 7(Suppl 4), S3, 2013.

[6] Wernicke, S. and Rasche, F., *FANMOD*: a tool for fast network motif detection. Bioinformatics, 22(9): 1152-1153, 2006.

[7] Wernicke, S., A faster algorithm for detecting network motifs, In Proceedings of the 5th Workshop on Algorithms in Bioinformatics (WABI'05), Lecture Notes in Bioinformatics, Vol. 3692, pp. 165–177, 2005.

[8] Ribeiro, P., Efficient and Scalable Algorithms for Network Motifs Discovery, Diss. PhD thesis, University of Porto, 2011.

[9] Zuba, M., A Comparative Study of Network Motif Detection Tools, UConn Bio-Grid, REU Summer, 2009.

[10] The COSIN Network data and Analysis, Available: http://pil.phys.uniroma1.it/~gcalda/cosinsite/extra/data/proteins/

[11] Kashtan, N., Itzkovitz, S., Milo, R. and Alon, U., Network Motif Detection Tool: mfinder Tool Guide. Weizmann Institute of Science, Depts of Mol Cell Bio and Comp Sci & Applied Math, Rehovot, Israel (2002-2005).

[12] Kavurucu, Y., Network Motifs and Indexing Techniques in Biological Networks, Journal of Naval Science and Engineering, 8.2 (2012): 87-102, 2012.

[13] Itzhack, R., Mogilevski, Y. and Louzoun, Y., An optimal algorithm for counting network motifs, Physica A: Statistical Mechanics and its Applications, 381, 482-490, 2007.