Superclusteroid 2.0: A Web Tool for Processing Big Biological Networks

Maria Tserirzoglou-Thoma, Konstantinos Theofilatos, Eleni Tsitsouli, Georgios Panges-Tserres, Christos Alexakos, Charalampos Moschopoulos, Georgios Alexopoulos, Konstantinos Giannoulis, Spiros Likothanassis, Seferina Mavroudi

Superclusteroid 2.0: A Web tool for processing big biological networks

Maria Tserirzoglou-Thoma[1], Konstantinos Theofilatos[2,*], Eleni Tsitsouli[1], Georgios Panges-Tserres[1], Christos Alexakos[2], Charalampos Moschopoulos[1], Georgios Alexopoulos[1], Konstantinos Giannoulis[1], Spiros Likothanassis[1,*] and Seferina Mavroudi[3]

[1]Department of Computer Engineering and Informatics, University of Patras, Greece
{thom, tsitsoue, panges, mosxopul, alexopo, giannul, likothan}@ceid.upatras.gr
[2]InSyBio Ltd, 109 Uxbridge Road, London, United Kingdom
{k.theofilatos, c.alexakos}@insybio.com
[3]Department of Social Work, Technological Institute of Western Greece, Greece
mavroudi@teiwest.gr

Abstract. Biological networks have been the most prevalent model to analyze the complexity of cellular mechanisms. The expansion of the existing knowledge on known intracellular players such as genes, RNA molecules and proteins as long as the continued study on their interactions has increased lately the ability to construct big biological networks of increased complexity. Many web tools have been introduced in the last decade but they are incomplete, as they do not provide all required features for a full research study neither they can handle the big and complex nature of these networks and the increased needs of researchers for fast and uninterrupted analysis. In the present paper, the new version of the Superclusteroid tool is presented which includes among others new visualization features, network comparison tools and new clustering algorithms. Moreover, a new strategy is proposed to deal with the necessity of handling effectively the increased work load of the tool as long as to improve the speed in the two most time consuming steps: network visualization and network clustering.

Keywords: Web tool, biological networks, protein-protein interaction networks, network clustering, network visualization

1 Introduction

Network visualization is a fundamental method that helps scientists in understanding biological networks and important properties in underlying biochemical processes. Molecules such as DNA, RNA, proteins, metabolites and interactions between them are related to highly important biological networks. Whenever such molecules are connected by physical interactions, they form molecular interaction networks that are generally classified by the nature of the compounds involved. Many biological networks have been characterized in detail: Protein-Protein Interaction (PPI), Gene co-expression

networks (Transcript-Transcript association networks), Gene regulatory networks (DNA-protein interaction networks), protein phosphorylation, metabolic interactions, and genetic interaction networks [1].

The PPIs represent the interaction between proteins: e.g. the formulation of protein complexes and the activation of one protein by another protein. These interactions are essential to almost every process in a cell, thus understanding of them is crucial. Such a network can be defined as an un-directed graph G = (V,E) where V is the set of proteins represented as nodes and E the set of interactions represented as edges. Graphs of a whole cell PPIs are complex and difficult to be generated. For this reason bioinformatics tools have been developed to simplify the difficult task of visualization, such as Cytoscape [2] which is an open-source software commonly used. Large scale identification of PPIs generates hundreds of thousands interactions, which were collected together in specialized biological databases, such as BIND and DIP, that are continuously updated in order to provide complete interactomes [3].

The Gene Co-expression networks represent the pairs of genes which show a similar expression pattern across samples, since the transcript levels of two co-expressed genes rise and fall together across samples. Gene co-expression networks are of biological interest since co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex. These networks can be defined as un-directed graphs where each node corresponds to a gene and a pair of nodes is connected with an edge when there is a significant co-expression relationship between them. Modules or the highly connected subgraphs in gene co-expression networks correspond to clusters of genes that have a similar function or involve in a common biological process which causes many interactions among themselves. Gene co-expression networks are usually constructed using datasets generated by high throughput gene expression profiling technologies such as Micro-array or RNA-Seq. [4].

The Gene Regulatory networks (DNA-Protein interaction networks) represent the DNA segments in a cell which interact with each other indirectly, through their RNA and protein expression products and with other substances in the cell to govern the gene expression levels of mRNA and proteins. The modeling techniques for such networks involves the use of Coupled Ordinary Differential Equations, Boolean networks, Bayesian networks, Graphical Gaussian models, Stochastic Gene network et al. The Gene Regulatory networks can be defined as graphs in which the un-directed edge connects two genes, representing a biochemical process such as a reaction, transformation, interaction, activation or inhibition [5].

The problem of analyzing biological networks is a big data problem for two reasons. First, it is related to intensive time-consuming analysis of big datasets as for example a PPI network can implicate more than 20000 proteins and more than 200000 interactions. Second, this analysis is not an one off procedure. New versions of biological networks are being available daily as new interactions are being studied and thus the tool's workload is extremely big. Additionally, if we consider the vast number of different organisms as well as the enormous number of biological conditions we can easily understand the "big data" nature of the problem.

In the present paper, we introduce a new version of the Superclusteroid tool. Superclusteroid is a web based tool which enables the analysis of various types of biological networks including the aforementioned categories with the main constraint being that only undirected networks can be used. This new version is differentiated from the previous one by including a new more advanced algorithm for biological network clustering, providing new features for the analysis and using a new mechanism to handle high workload of very large networks.

2 Existing tools for the analysis of biological networks

Large scale biological studies produce huge amounts of data that reveal various layers of molecular interaction networks. As we saw graphs have been used to represent, study and integrate such biological networks, which leads in large-scale analyses. Specialized tools are required to extract and compare information obtained from multiple data sources, and apply various statistical parameters treatments to describe and understand networks properties. Following are the basic web-based or standalone tools for analyzing biological networks, which are based on graphs and their visualization.

NETAL [6], is a new graph-based method for global alignment of protein-protein interaction networks. It uses a greedy method, based on the alignment scoring matrix, which is derived from both biological and topological information of input networks to find the best global network alignment.

NeAT (Network Analysis tool) [7], is another tool which provides a user-friendly web access to a collection of modular tools for the analysis of networks (graphs) and clusters (e.g. microarray clusters, functional classes, etc.). This tool is designed to cope with large datasets and provide a flexible toolbox for analyzing biological networks stored in various databases (protein interactions, regulation and metabolism) or obtained from high-throughput experiments (two-hybrid, mass-spectrometry and microarrays). The web interface interconnects the programs in predefined analysis flows, enabling to address a series of questions about networks of interest.

GraphWeb [8], is a public web server for biological network analysis and module discovery. It provides methods to integrate heterogeneous and multispecies data for constructing directed and undirected, weighted and un-weighted networks, to discover network modules using a variety of algorithms and topological filters and interpret modules using functional knowledge of the Gene Ontology and pathways, as well as regulatory features such as binding motifs and microRNA targets.

Giba [9], is a clustering tool that offers the ability to detect important protein modules such as protein complexes. GIBA implements a two-steps strategy, where in the first one the whole protein – protein interaction graph is divided into clusters and in the second step these clusters are filtered and only the ones considered important are kept.

jClust [10] is an application which provides access to a set of widely used clustering and clique finding algorithms. The toolbox allows a range of filtering procedures to be applied and is combined with an advanced implementation of the Medusa interactive visualization module. These implemented algorithms are k-Means, Affinity Propaga-

tion, Bron–Kerbosch, MULIC, Restricted neighborhood search cluster algorithm, Markov clustering and Spectral clustering, while the supported filtering procedures are haircut, outside–inside, best neighbors and density control operations. The tool provides a powerful tool for data analysis and information extraction.

Cluto [11] is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters. Cluto is well-suited for clustering data sets arising in many diverse application areas including information retrieval, customer purchasing transactions, web, GIS, science, and biology. Cluto's distribution consists of both stand-alone programs and a library via which an application program can access directly the various clustering and analysis algorithms implemented in Cluto.

VisANT [12], is an application for integrating biomolecular interaction data into a cohesive, graphical interface. This software features a multi-tiered architecture for data flexibility, separating back-end modules for data retrieval from a front-end visualization and analysis package. This system is integrated with standard databases for organized annotation, including GenBank, KEGG and SwissProt. It provides a general tool for mining and visualizing such data in the context of sequence, pathway, structure, and associated annotations. Interaction and predicted association data can be combined, overlaid, manipulated and analyzed using a variety of built-in functions.

Most of the aforementioned tools cover only a subset of the analysis related to biological networks, while most of them are not PPI network specific and some of them include obsolete algorithmic solutions. Superclusteroid 1.0 [13] is a web tool dedicated to data processing of protein-protein interaction networks which was initially introduced to cover all these caveats. The tool is implemented in the GNU/Linux environment and is written in Perl. It supports various input file formats and provides the following services. First, clustering, by choosing one of the available clustering algorithms. Second, PPI network visualization, such as the original network or other DOT files and also the algorithms results can be automatically visualized or can be downloaded for later use. Third, protein cluster function prediction in which the user, by choosing a specific protein, may continue with the analysis by implementing the Majority Vote Prediction Algorithm (MVPA) [14] or the Hypergeometric Distribution Prediction Algorithm (HDPA) [15].

3 Superclusteroid 2.0: A web tool for processing big protein-protein interaction networks

3.1 Superclusteroid 1.0

The Superclusteroid 1.0 [13] uploads and manipulates input of PPI data, which can have one the following formats: tab-delimited text files, adjacency matrices in text files, DOT files-using the DOT network description languages and SIF files, a popular tab-delimited text file mostly used in Cytoscape.

It offers a selection of widely used clustering algorithms to process the input data. These algorithms are the MCL -Markov Cluster [15], the Restricted Neighbourhood

Search Clustering Algorithm –RNSC [16], the Highly Connected Subgraphs Algorithm -HCS [17] and the SideS, a variation of HCS which uses a statistical model to express the statistical significance of a cluster [18]. The resulting files are tab-delimited data with two columns, one for the name of the cluster and one for the protein belonging to that cluster.

The results from the clustering algorithm can be automatically visualized or can be downloaded for later use. Also the original network or other DOT files can be viewed by choosing the "visualize" tab. In either case, a java applet named "ZGRViewer" [19] is used to support the "fdp" and "twopi" GraphViz/DOT tools for spring model and radial layouts respectively.

The analysis of the network data can be continued, since the user can choose a specific protein and implement the Majority Vote Prediction Algorithm (MVPA) [14] or the Hypergeometric Distribution Prediction Algorithm (HDPA) [15]. Both methods are applied only on PPI data with Uniprot IDs and for the S. cerevisiae organism.

A help page is available on the tool which contains explicit instructions describing its services and a comprehensive list of the web services available, along with their description and the access URL for each of them. Additionally, the web tool provides demo data to help the user to understand its functionality.

Despite the success of the first version of Superclusteroid tool it lacked some extra functionalities in order to enable its users to realize the whole analysis of PPI networks without the need of referring to external tools. Moreover, new algorithms have emerged for some of the tasks of analyzing PPI networks, such as new clustering algorithms for overlapping clusters, since the initial launching of Superclusteroid 1.0 and some of them were needed to be incorporated in it. Finally, the successful adoption of Superclusteroid tool from the scientific community (more than 200 unique users/visitors per month) raised some performance issues and more sophisticated solutions were required to manage work load and improve the efficiency of the tool. For all these reasons, a new version of Superclusteroid was essential and the new features incorporated in it are described in the next section.

3.2 New Features incorporated in Superclusteroid 2.0

EEMC

The Evolutionary Enhanced Markov Clustering (EEMC) [20] is a hybrid combination of an adaptive evolutionary algorithm and a state-of-the-art clustering algorithm. It is based on the MCL algorithm [15] which is one of the most commonly used methods in clustering PPI graphs in order to predict protein complexes. Although it has some strong limitations, with the most important one being its restriction to assign each protein to only one protein complex. To overcome the above MCL's problem, Moschopoulos et al. (2008) [21], proposed the Enhanced Markov Clustering (EMC) method which is an improvement of the MCL algorithm. In specific, it deploys the MCL algorithm to make an initial clustering and then it improves it by applying 4 different filtering methods: density filter, haircut, best neighbor and cutting edge operators. The last method requires tuning of their parameters and it is not able to function on weighted graphs.

The EEMC is a fully unsupervised method, which is a combination of an adaptive Genetic Algorithm and an extension of the EMC method in which, the filtering methods of the EMC algorithm were adjusted to enable handling of weighted PPI graphs. The Genetic Algorithm was used to optimize on parallel the inflation rate and the parameters of the filtering methods of EMC algorithm.

The EEMC starts with the creation of the initial population ie. the chromosomes with binary representation. Then the adjusted filters from the EMC algorithm are applied on the chromosomes. These filtering methods are density filter, haircut, best neighbor and cutting edge operators. The next step is the evaluations of the chromosomes, with an evolutionary framework using unsupervised fitness value which will produce scaled fitness, to assign high values for high performance clustering. The selection operator is then taking place on the algorithm. The roulette wheel selection assigns probabilities of selection in each chromosome proportional to its performance. The variation operators used are the two-point crossover and the binary mutation. A dynamic control of the mutation parameter is used, to estimate the variation which is applied in the mutation probability for each iteration of the EEMC. Finally the termination criteria of the algorithm are a combination of the maximum number of generations to be reached and a convergence criterion.

New functions of Version 2.0

Superclusteroid 2.0 includes a set of new features which enable users to conduct all required biological network analysis without the need to use other tools.

To begin with, a new way of cluster visualization is proposed, using the Cytoscape program [22], a program for analyzing and visualizing network data. Specifically, the users can now not only to see the nodes of a cluster but also their interactions and their weights in the cases of weighted networks. This can enable a more detailed analysis of the cluster and its connectivity.

Furthermore, with the new version of Superclusteroid, the users have the ability to compare and evaluate clusters, with a collection of human core protein complexes from the CORUM database [23] or with a dataset which the user chooses to upload. The uploaded datasets should be in a tab delimited format, with every row representing a cluster and including its nodes separated with tabs. The calculated metrics include the standard metrics of sensitivity, positive predictive value, arithmetic accuracy, geometric accuracy, separation [24].

Another newly introduced feature is the linkage of clustering results with the he Gene Set Enrichment Analysis (GSEA) [25] tool. With this feature the users are able to evaluate and characterize the importance of gene sets, i.e., gene groups with a common biological function, chromosomal location or setting.

Moreover, the pipeline has been completed with a network comparison option which allow the comparison of similar biological networks which have either be constructed under different biological conditions or refer to different organisms. This analysis, additionally to the standard network metrics calculation, such as clustering coefficient, also offers a graph with the degree distributions of the networks under comparison and

allow the biomarker discovery by locating the nodes and edges which have been differentiated between the under comparison networks.

Some other features of secondary significance which have been included in the current version of Superclusteroid tools for the meta-analysis of protein clustering results such as a haircut filter to discard nodes with low connectivity, a neighborhood analysis tool to locate nodes being connected with a specific node of interest and cluster function prediction tool which is based on the HDPA algorithm but is only applicable on *Homo Sapiens* and *Saccharomyces cerevisiae*.

The most intensive tasks for the analysis of biological networks are network clustering and network visualization. In order to satisfy the increased demands of Superclusteroid's users for these two types of analysis, Superclusteroid 2.0 introduce a queuing mechanism presented in Figure 1. In specific, the Superclusteroid web server, using iteratively the min cut algorithm splits the initial network to sub-networks and provides them to the virtual infrastructure dedicated to clustering and visualization analysis. When the analysis is conducted then it is provided to Superclusteroid's web server which undertakes to assemble the results and present them to the user. File sharing between web server and virtual infrastructure is done via a Network File System (NFS) memory unit. RabbitMQ server [26] is utilized in order to handle efficiently the work load of the Virtual Infrastructure while all required scripts are written in Python programming language version 2.7.

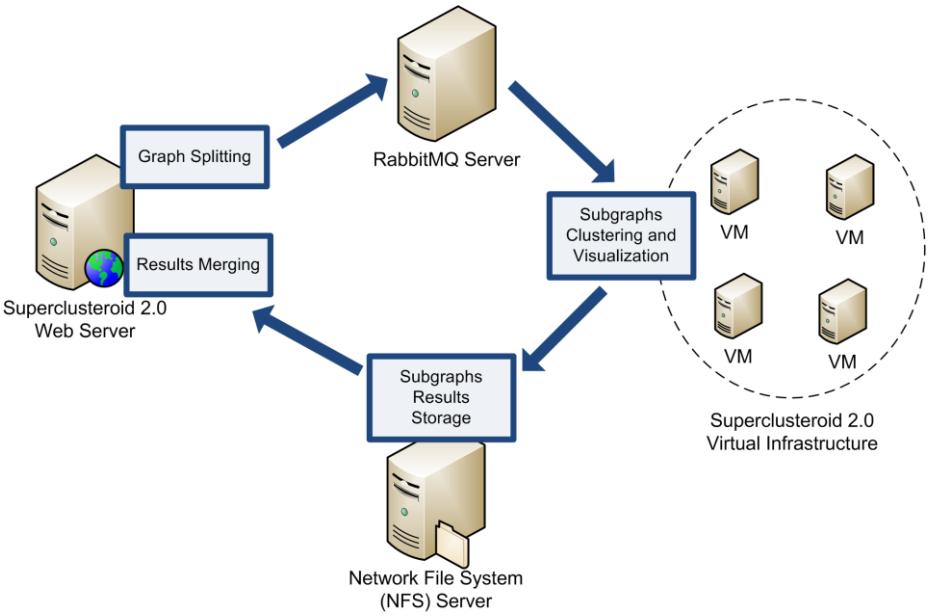

Fig. 1. Proposed Superclusteroid 2.0 architecture for handling effectively increased work load on time intensive biological network analysis tasks

4 Results

Superclusteroid 2.0 has been tested using as input the PPI graph of the Yeast organism as described in [27]. The dataset contains 1430 proteins and 26531 interactions between them. Experimental results [20] have proved the superiority of the EEMC algorithms in all examined clustering metrics. Figures 2-4 depict some screenshots of the analysis.

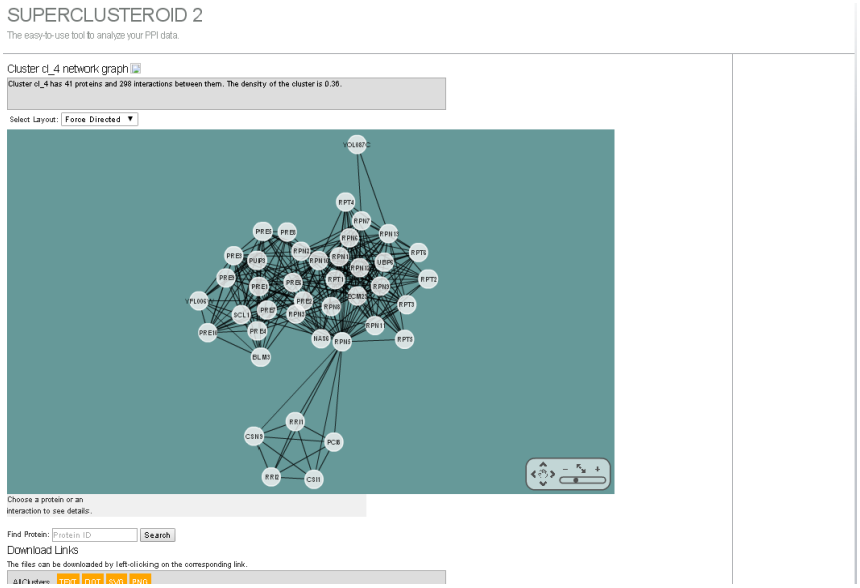

Fig. 2. Superclusteroid's new protein function prediction view.

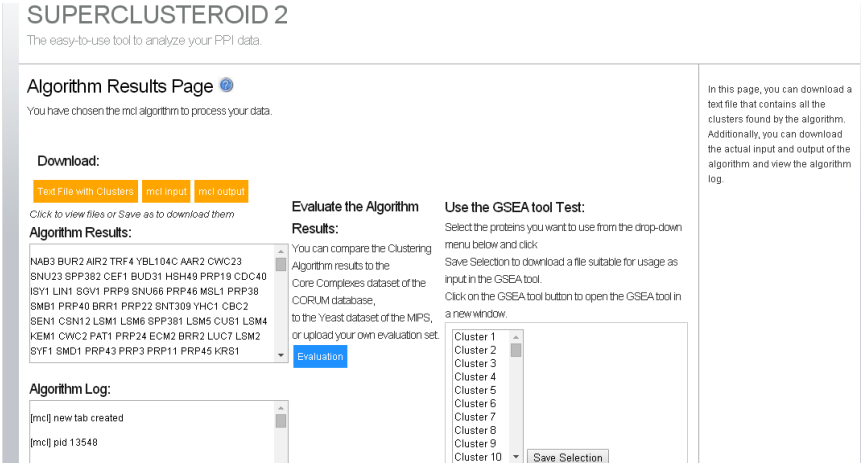

Fig. 3. Newly designed view for showing clustering results and allowing algorithms evaluation and GSEA analysis.

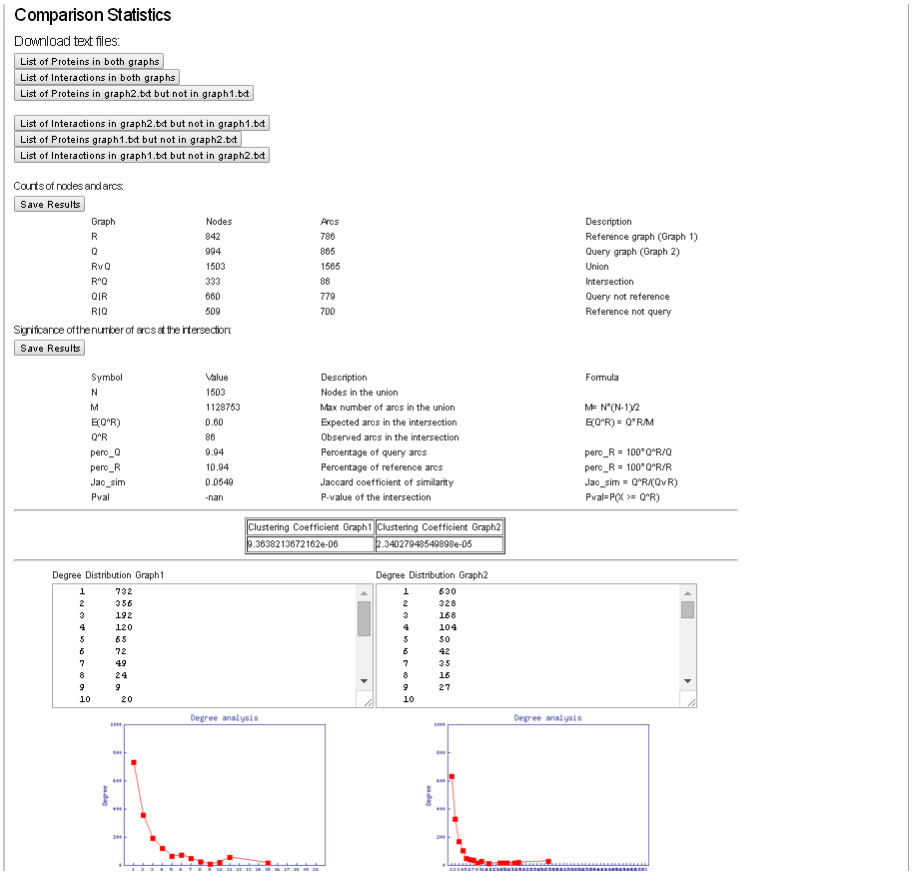

Fig. 4. Superclusteroid ‘s Biological Network comparison module. This figure shows the comparison of two yeast PPI networks. This analysis allows for a network comparison using general network metrics such as clustering coefficient as well as a comparison using the node degree distribution figures and a more detailed analysis on the union, intersection and differences between the two examined networks.

5 Discussion

In this paper we have presented a new version of the Superclusteroid web tool. This new version extends significantly the previous one in order to provide to the users all the necessary tools to perform a full analysis and meta-analysis on undirected weighted biological networks. Moreover, a new more elaborated network clustering solution was added to allow the discovery of overlapping clusters. This clustering solution was proved to overcome the constraints of previous clustering solutions achieving higher clustering metrics.

Finally, in order to handle the increased traffic of the web tool, we have implemented and incorporated in it an architecture which faces effectively the work load on the analysis of the time-consuming steps of graph visualization and clustering. As a future research work, our research team has already started to work on a more elaborated scheduling solution to manage more effectively the visualization and clustering tasks.

6 Acknowledgments

InSyBio participates in the NBG seeds program by the National Bank of Greece.

7 References

1. Zhu, X, Gerstein, M. and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. Genes & Dev. 2007. 21: 1010-1024
2. Kohl, M., Wiese, S., Warscheid, B. (2011). Cytoscape: Software for Visualization and Analysis of Biological Networks. Data Mining in Proteomics. Methods in Molecular Biology 696. pp. 291–303.
3. Mason, O., & Verwoerd, M. (2007). Graph theory and networks in biology.Systems Biology, IET, 1(2), 89-119.
4. Stuart, J. M., Segal, E., Koller, D., & Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. science, 302(5643), 249-255.
5. Roy, S., Bhattacharyya, D. K., & Kalita, J. K. (2014). Reconstruction of gene co-expression network from microarray data using local expression patterns.BMC bioinformatics, 15(7), 1.
6. Neyshabur, B., Khadem, A., Hashemifar, S., & Arab, S. S. (2013). NETAL: a new graph-based method for global alignment of protein–protein interaction networks. Bioinformatics, 29(13), 1654-1662.
7. Brohée, S., Faust, K., Lima-Mendez, G., Sand, O., Vanderstocken, G., Deville, Y., & van Helden, J. (2008). NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. Nucleic acids research,36(suppl 2), W444-W451.
8. Reimand, J., Tooming, L., Peterson, H., Adler, P., & Vilo, J. (2008). GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. Nucleic acids research, 36(suppl 2), W452-W459.
9. Moschopoulos, C. N., Pavlopoulos, G. A., Schneider, R., Likothanassis, S. D., & Kossida, S. (2009). GIBA: a clustering tool for detecting protein complexes. BMC bioinformatics, 10(6), 1..
10. Pavlopoulos, G. A., Moschopoulos, C. N., Hooper, S. D., Schneider, R., & Kossida, S. (2009). jClust: a clustering and visualization toolbox.Bioinformatics, 25(15), 1994-1996.
11. Zhao, Y., & Karypis, G. (2005). Data clustering in life sciences. Molecular biotechnology, 31(1), 55-80.
12. Hu, Z., Mellor, J., Wu, J., & DeLisi, C. (2004). VisANT: an online visualization and analysis tool for biological interaction data. BMC bioinformatics, 5(1), 1.
13. Ropodi, A., Sakkos, N., Moschopoulos, C., Magklaras, G., & Kossida, S. (2011). Superclusteroid: a Web tool dedicated to data processing of protein-protein interaction networks. EMBnet. journal, 17(2), pp-10.

14. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., ... & Li, G. (2003). Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic acids research, 31(9), 2443-2450.
15. Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic acids research,30(7), 1575-1584.
16. King, A. D. (2004). Graph clustering with restricted neighbourhood search(Doctoral dissertation, University of Toronto).
17. Hartuv, E., Schmitt, A., Lange, J., Meier-Ewert, S., Lehrach, H., & Shamir, R. (1999, April). An algorithm for clustering cDNAs for gene expression analysis. In Proceedings of the third annual international conference on Computational molecular biology (pp. 188-197). ACM.
18. Koyutürk, M., Szpankowski, W., & Grama, A. (2007). Assessing significance of connectivity and conservation in protein interaction networks. Journal of Computational Biology, 14(6), 747-764.
19. Pietriga, E. (2005). Zgrviewer-a 2.5 D graph visualizer for the DOT language.
20. Theofilatos, K., Pavlopoulou, N., Papasavvas, C., Likothanassis, S., Dimitrakopoulos, C., Georgopoulos, E., ... & Mavroudi, S. (2015). Predicting protein complexes from weighted protein–protein interaction graphs with a novel unsupervised methodology: Evolutionary enhanced Markov clustering.Artificial intelligence in medicine, 63(3), 181-189.
21. Moschopoulos, C. N., Pavlopoulos, G. A., Likothanassis, S. D., & Kossida, S. (2008, October). An enhanced Markov clustering method for detecting protein complexes. In BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on (pp. 1-6). IEEE.
22. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 13(11), 2498-2504.
23. Ruepp A., Weagele B, et al. (2009) “CORUM: the comprehensive resource of mammalian protein complexes—2009”, Nucleic Acids Res. 2010 Jan;38(Database issue):D497-501
24. Brohee, S., & Van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC bioinformatics, 7(1), 1.
25. Eric S. Lander et al, “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles”, PNAS 2005 102 (43) 15545-15550; published ahead of print September 30, 2005.
26. Russell, J., & Cohn, R. (2012). Rabbitmq. Book on Demand.
27. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., ... & Edelmann, A. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature, 440(7084), 631-636.