Biotechnol Res 2019; Vol 5(3):22-27
eISSN 2395-6763

ORIGINAL RESEARCH

A Multi-Level Layout Algorithm for Identifying Highly Interacted Human Protein Complexes and Various Protein Pathways

Dr. Sudeepa E.S*

Assistant Professor, Nehru Arts and Science College, Thirumalayampalayam, Tamil Nadu, India

***Corresponding Author email: *sudeepa.es@gmail.com*

• Received: 21 June 2019 • Revised: 19 July 2019 • Accepted: 24 July 2019 • Published: 1 August 2019 •

ABSTRACT

Discovering highly interacting proteins and its pathways is a challenge for computational biologists. Identifying interactions among proteins and its pathways has been found to be useful for drug development. In our study, we collected data of 40,788 protein - protein interactions from HPRD and IntAct. This pooled off data was loaded into cytoscape (version 2.63) to visualize the human interactome network through the grid layout method. By using Connected Component (CC) algorithm, the largest and the highly connected human network were found. From 36,945 binary protein – protein interactions, 89 highly connected modules were selected, which are of high score value. The total numbers of proteins in each of the 89 modules were found out. The proteins from all the 89 clusters were classified into 2 categories - normal and diseased pathways. From all these 89 clusters, 1350 proteins were obtained and were unique, of which 374 proteins are in the normal pathways and 976 proteins in diseased pathways. But 976 proteins were found in both normal and diseased pathways. Further computational studies can help to understand the changes that occur in the proteins to become in the diseased pathways.

KEYWORDS: *Human interactome, HPRD, IntAct, CCA, MCODE, Cytoscape*

INTRODUCTION

Protein interactions and its pathways provide a valuable resource for understanding cellular function, signaling pathways, modeling of protein complex structures and various biochemical processes. Protein networks characterize physical relationships between proteins that are in direct binding contact or in a complex and it is essential to understand how gene functions and regulations are integrated at the level of an organism. Changes in the observed modularity of the human protein interaction networks and pathways have been used to predict biological and clinical outcomes, such as brain and breast cancer (Taylor *et al*., 2009 and Dutkowski and Ideker, 2011). The technologies like two-hybrid systems (Legrain and Selig, 2000), protein complementation assays, protein arrays or tandem affinity purification allow the generation of large quantities of protein interaction data. However, for research works,

researchers develop their own systems for the storage, representation and analysis of protein interactions and its pathways data.
The goal of our study of proteomics is to elucidate the structure, interactions and functions of all proteins within human intreractome. One strategy to determine the role of protein is to identify the protein-protein interactions and its pathways. The increasing use of high-throughput and large-scale computational-based studies has generated huge amount of data stored in a number of different databases. Also a number of layout algorithms are available in popular network analysis platforms, such as Cytoscape, but it remains poorly understood how well their solutions reflect the biological processes that give rise to the network connectivity structure. A challenge for computational biology is to explore this huge data and to uncover biologically relevant interactions.

Conventionally, these layout algorithms are specifically designed for a particular network type, such as gene regulatory networks or signalling pathways (Hosoyama *et al*., 2003, Kojima *et al*., 2007), metabolic pathways or biochemical networks (Paley and Karp, 2006, Villéger *et al*., 2010, Rocha *et al*., 2010), or phylogenetic networks (Gambette and Huson, 2008). Algorithms have also been introduced for grid layouts (He *et al*., 2010), or detailed visualization of small networks (Dannenfelser *et al*., 2011). There is no universal layout for finding out which of the module arranges the best, so that the multiple layout algorithms have the role.

Cytoscape network helps in analysis and visualization for the biological community because of its ease of use, compatibility with and direct access to many network formats and databases, as well as straightforward extensibility through open-source plug-in development (Cline *et al*., 2007, Shannon, 2003). Highly connected networks, in particular those with dense or clustered connectivity structure, are more difficult to visualize, often resulting in 'hairball' network layouts (Suderman and Hallett, 2007, Pavlopoulos *et al*., 2008, Merico *et al*., 2009).

A limitation of the conventionally used algorithms is relatively lengthy running times in the largest network graphs. For instance, generating the layout took almost 7 minutes for the largest yeast genetic interaction network (4319 nodes with 74,984 edges) and 8 minutes for the human HPRD PPI network (5699 nodes with 19,779 edges) (Johannes, 2012). While cytoscape modules reduced computation times, further speed-ups could be achieved by linking specific C functions to the Java implementation (Su *et al*., 2010), or by using hardware-based graphics acceleration (Brown *et al*., 2009).

There are a large number of proteins involved in various pathways in normal as well as diseased pathways. All proteins are continually ‘turning over’; they are being hydrolyzed to their constituent amino acids and replaced by new synthesis. Overall, the rates of protein synthesis and degradation must be balanced precisely because even a small decrease in synthesis or a small acceleration of degradation, if sustained, can result in various changes. Abnormal proteins in the body, which are often precursors to disease, are first tagged to be degraded by the protein ubiquitin before they can be recognized. The changes in the protein structure and its constituents can occur, which leads into diseased pathways (Nandi, 2006).

We suggest that a multilevel layout algorithm can be utilized to provide both visually attractive and biologically relevant networks. In the present work, a systematic comparative evaluation on various large-scale networks, originating both from physical and genetic interaction mappings, various protein pathways could provide users with a practical guidance on how to choose a preferable layout algorithm for different network types and their characteristic properties. To facilitate drawing of networks with several thousands of nodes, we improved the computational complexity of the multilevel approach through the use of efficient Mcode module. Here we present multi layout algorithm method to find out the most largely connected protein interactions and its pathways.

Source of data

(a) HPRD

Human Protein Reference Database (HPRD) contains manually acquired scientific information pertaining to the biology of most human proteins. HPRD has grown to over 36,500 unique protein-protein interactions annotated for 25,000 proteins including 6,360 isoforms by the end of 2006 (Mathivanan *et al*., 2006). Totally 41,327 protein – protein interactions are there at present. More than 50% of molecules annotated in HPRD have at least one protein – protein interactions and 10% have more than 10 protein – protein interaction. Experiments for protein – protein interactions are broadly grouped into three categories namely in vitro, in vivo and yeast two hybrids (Y2H). HPRD data is available for download in tab delimited and XML file formats shown in Figure.1.

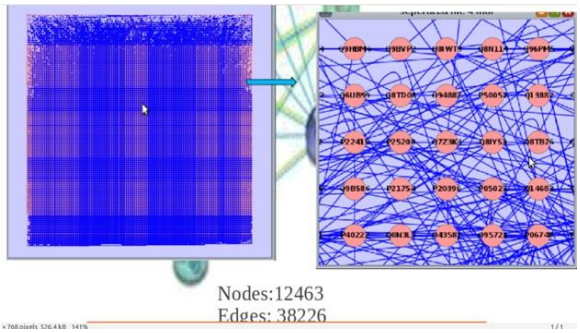

Figure 1: Human interactome in cytoscape software (version 2.63) with grid layout.

(b) IntAct

IntAct provides an open source database and toolkit for the storage, presentation and analysis of protein interactions. The web interface provides both textual and graphical representations of protein interactions, and allows exploring interaction networks in the context of the Gene Ontology (*GO*) annotations of the interacting proteins. IntAct contains approx 22,000 binary and complex protein – protein interactions. IntAct data is available in XML file formats.

The IntAct data model has three main components: - Experiment, Interaction and Interactor. An Experiment may have only a single interaction. Hundreds of interactions occur in the large-scale experiments. An Interactor is a biological entity participating in an Interaction, usually a protein, but potentially also a DNA sequence, or a small molecule. An Interaction contains one or more Interactors participating in the Interaction. The representation of interactions is not limited to binary interactions; data on multi-protein interactions, e.g. the results of tandem affinity purification experiments (Gravin *et al*., 2002), can be represented as one interaction, without artificially splitting them up into several binary interactions.

RESULTS AND DISCUSSION

Protein protein interactions are involved in the supramolecular assemblies (collagens, elastic fibers, actin filaments), in the building of molecular machines (molecular motors, ribosomes, proteasome) and biological processes such as immunity (antigen–antibody interaction), metabolism (enzyme–substrate interaction), signaling (interaction of messenger molecules, hormones, neurotransmitters with their cognate receptors) and gene expression (DNA–protein interactions (Hanahan and Weinberg, 2000). This work was carried out by collecting all true binary interactions between proteins from HPRD and IntAct.Keshava Prasad *et al*., had reported similar work and has been published on 2009. 39,240 binary protein interactions were collected from HPRD and 7,214 binary protein interactions from IntAct. Binary protein interactions from IntAct also have been published by Bowen *et al*., 2009. There are many databases, which have large amount of interaction data between proteins and more than 16 million citations in the Medline database (Zhou and He, 2008). Lehner and Fraser (2004) described a network of over 70,000 predicted physical interactions between proteins and around 6,200 human proteins generated using the data from lower eukaryotic protein–interaction maps. Among the data bases available, HPRD and IntAct are manually currated, validated and a large amount of true proteins interaction are present. The annotations of both the data bases (HPRD and IntAct) were different. Kerrien *et al*. on 2007 gave an uniport ID for the proteins.Personalized inbuilt perl programmes and Linux operating system helped to identify the uniport ID for all the proteins that have been taken from interaction network databases. From 46,454 interaction databases, the redundant interactions were removed and finally 40,788 protein - protein interactions were obtained and (Liu *et al*., 2010). There are several protein-protein interaction information extraction systems and tools for biomedical literature mining are available on the web (Zhou and He, 2008).

The final sets of data were loaded into cytoscape software (version 2.6). Cytoscape helped us to improve drawing of large-scale networks of all human protein interactions. Moreover, the biological evaluation developed here enabled one to assess the layout algorithms from any existing data or in future graph drawing algorithm to optimize their performance for a given network type or structure. Grid analysis helped us to take important decision where there isn't a clear and obvious result. The human interactome obtained had 12463 nodes and 38226 edges are present (Shannon, 2003).

The small and disconnected clusters which are found along with the interactions were eliminated using CCA module. Bonetta published a similar report on 2010. So the way to find the highly interconnected network became very easy, so that the way to find the highly interconnected network very easily. Figure.2 shows the highly interacting networks using CCA module.

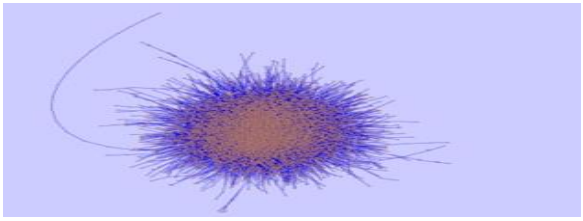

Figure 2: The highly interacting network.

The densely connected regions in large protein-protein interaction networks that may represent molecular complexes using MCODE. This algorithm has the advantage over other graph clustering methods that allows isolating clusters of interest. Figure.3 shows 189 highly interacting protein cluster using MCODE. Among these 89 clusters are taken, which are having the score more than one.

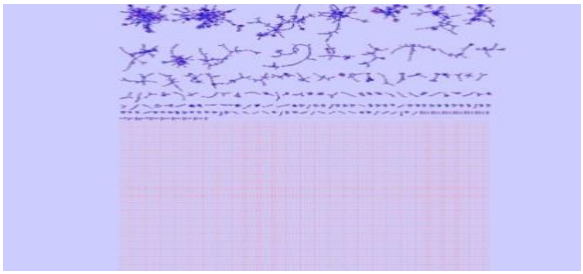

Figure 3: The 189 highly interacting protein clusters.

The numbers of proteins present in each of the 89 clusters are found and all the proteins are identified (Bader *et al*., 2006). All the proteins obtained are unique. Each protein obtained is given to KEGG database, to identify its pathway, which provides valid information for find out the abnormal or disease causing proteins (Cerami *et al*., 2011). The proteins obtained from all the 89 clusters are classified into 2 categories – normal and diseased pathways. Wang *et al*.., 2010 also classified the proteins. From all these 89 clusters, 1350 proteins were obtained and were unique, in which 374 proteins are in

the normal pathways and 976 in diseased pathways. 976 proteins were present in both the normal and diseased pathways. Aldridge in 2006 also found out that certain proteins were found both in normal and diseased pathways.

Protein structures can reveal how disease processes affect large sets of molecular (Pawson and Warner, 2007) and cellular interactions (Kirouac *et al*., 2009). Since we have the set of proteins which are present in the Normal and Diseased pathway, further computational studies can be done to investigate what are the factors or the changes in the proteins occurred. Either structural or enzymatic changes in these proteins occured because of that they are found in normal and diseased pathways. Francesconi *et al*. on 2008 and Li *et al*. on 2008 studied how and what all changes and ccurred in the proteins, which caused various diseases.

CONCLUSION

By making use of the multilevel algorithm layout after visualizing biological networks, we can generate convenient visualizations for studying many pathways or network biology applications. We propose a network-based analysis for the comparisons of various proteins and its pathways by integrating protein-protein interactions and gene expression profiles. The availability and integration of high-throughput gene expression data and the genome-wide protein-protein interaction may help to understand the changes that occur in the proteins.

These interaction protein networks and pathways are complex systems, where new properties arise, which helps to find out various diseases. The systems biology which is ‘‘the study of an organism, viewed as an integrated and interacting network of genes, proteins and biochemical reactions which give rise to life’’ (http://www.systemsbiology.org/).

This interdisciplinary approach, involving techniques from the mathematical, computational, physical and engineering sciences is required to understand complex protein networks. An application of these networks is to provide information and to formulate hypotheses on human diseases and therapies (drug discovery and targeting) (Ideker and Sharan, 2008).

METHODS

1. Uniport ID for Integrated data from various databases

Both the HPRD and IntAct data are clubbed together so that a large amount of protein- protein interactions can be studied. But there may be the occurrence of repeated data, which cannot be identified since their annotations are different. Both HPRD and IntAct databases have different style of presenting data. It is a difficult task for most investigators to compare the voluminous data from these databases in order to conclude strengths and weaknesses of each database. Mathivanan and colleagues, 2006 helped biologists to choose among these databases based on their needs.

Personalized perl programs used in the background of Linux operating system allow manual searching for identifying the uniport ID, for all the individual proteins that have been taken from both the HPRD and IntAct protein – protein interaction data bases.

2. Use of Cytoscape

Cytoscape is a source for visualizing human protein interactome, using powerful visual styles. Expression data can be mapped to node color, label, border thickness, or border color, etc. according to user-configurable colors and visualization schemes. It helps to view large networks (100,000+ nodes and edges), very easily. It was originally created at the Institute of Systems Biology in Seattle in 2002. Now, it is developed by an international consortium of open source developers. Cytoscape Version 2.0 was initially released in 2004.

The Grid Layout manager is ideal for laying out objects in rows and columns, where each cell in the layout has the same size. Components are added to the layout from left to right, top to bottom. Grid Analysis (also known as Decision Matrix Analysis, Pugh Matrix Analysis) is a useful technique for making a decision. Grid analysis is the technique worked by getting, to list our options as rows on a table, and the factors we need consider as the columns. Then score for each option/factor combination, weight this score, and these scores are added up to give an overall score for the option. While this sounds complex, in reality the technique is quite easy to use. It is particularly beneficial where we have confusion, to take a decision for eg. a number of good alternatives to choose from, and many different factors to take into account. Being able to use Grid Analysis helps us to take decisions confidently and rationally, at a time when other people might be struggling to make a decision.

(a)CCA- Module

CCA Module is one of the important clustering algorithms for partitioning network based on similarity or distance values. This is a very simple "cluster" that finds all of the disconnected components of the network and treats each disconnected component as a cluster. It supports the array sources (all of the numeric edge attributes) and the Cytoscape Advanced Settings options (cluster attribute, metanodes with results, enable debugging). Standard XML file formats such as GraphML or XGMML can represent substructures. However, they are not easy to edit by hand for many biologists. We need to implement a simple table/text style file format which is editable on spread sheet programs (Micheal, 1992).

(b)MCODE

The MCODE algorithm finds highly interconnected regions in a network. The algorithm uses a three-stage process:-vertex weighing (weighs all of the nodes), molecular complex prediction and filters to improve the modules quality (Bader and Hogue, 2003). The modules score cut off is the important parameter for deciding the module shape, size etc.

The higher values of the score cut off enable us to identify the pathways and highly connected interacting modules. The modules having the score value greater than 1, is taken for the experiment, since they have higher biological relevance.

3. Identification and classification of proteins into various pathways

All the proteins in each of the 89 clusters were identified. Each protein from each cluster is given to KEGG database to identify its pathway. The pathways are classified into 32categories - Normal pathway and diseased pathway.

ACKNOWLEDGEMENT

None

REFERENCES

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., & Sorger, P. K. (2006). Physicochemical modelling of cell signalling pathways. Nature Cell Biology, 8,1195-1203.

Bader, G. D., Cary, M. P., & Sander, C. (2006). Pathguide a pathway resource list. Nucleic Acids Research, 34(1), D504-506.

Bader, G. D., & Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics, 4. 2.

Bonetta, L. (2010). Protein-protein interactions: Interactome under construction. Nature, 68, 851-854.

Bowen, N. J., Walker, L. D., & Matyunina, L. V. (2009). Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. BMC Med Genomics, 2, 71.

Brown, K. R., Otasek, D., Ali, M., Mc Guffin, M. J., Xie, W., Devani, B., Toch, I. L., & Jurisica, I. (2009). NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. Bioinformatics. 25, 3327-3329.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., & Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. Nucleic Acids Res, 39, D685-690.

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A, R., Vailaya, A., Wang, P, L., Adler, A., Conklin, B, R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G., J., Ideker, T. & Bader, G. D. (2007). Integration of biological networks and gene expression data using Cytoscape. Nat Protoc, 2, 2366-2382.

Dannenfelser, R., Lachmann, A., Szenk, M., & Ma'ayan, A. (2011). FNV: Light-weight Flash-based network and pathway viewer. Bioinformatics, 27, 1181-1182.

Dutkowski, J., & Ideker, T. (2011). Protein networks as logic functions in development and cancer. PLoS Comput Biol, 7, e1002180. doi: 10.1371/journal.pcbi.1002180.

Francesconi M, Remondini D, Neretti N, et al. (2008). Reconstructing networks of pathways via significance analysis of their intersections. BMC Bioinformatics, 9 Suppl 4, S9.

Gambette, P., & Huson, D. H. (2008). Improved layout of phylogenetic networks. IEEE/ACM Trans Comput Biol Bioinform, 5, 472-479.

Gavin, A. C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A. M., & Cruciat, C. M. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 415, 141-147.

Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100, 57-70.

He, S., Mei, J., Shi, G., Wang, Z., & Li, W. (2010). LucidDraw: efficiently visualizing complex biochemical networks within MATLAB. BMC Bioinformatics, 11, 31.

Hosoyama, N., Nasimul, N., & Iba, H. (2003). Layout search of a gene regulatory network for 3-D visualization. Genome Inform, 14, 103-113.

Ideker, T., & Sharan, R. (2008). Protein networks in disease. Genome Res, 18: 644-652.

Johannes, T., Heidi, V., Pekka, S., Olli, S. N. & Tero, A. (2012). A multilevel layout algorithm for visualizing physical and genetic interaction networks, with emphasis on their modular organization. BMC.Bioinformatics, 5, 2.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., Hermjakob, H. (2007). IntAct open source resource for molecular interaction data. Nucleic Acids Res, 35, D561-265.

Keshava Prasad, T. S., Goel, R., & Kandasamy, K. (2009). Human Protein Reference Database--2009 update. Nucleic Acids Res, 37, D767-72.

Kirouac, D. C., Madlambayan, G. J., Yu, M., Sykes, EA., Ito, C., & Zandstra. (2009). PW: Cell-cell interaction networks regulate blood stem and progenitor cell fate. Mol Syst Biol, 5, 293.

Kojima, K., Nagasaki, M., Jeong, E., Kato, M., & Miyano, S. (2007). An efficient grid layout for biological networks utilizing various biological attributes. BMC Bioinforma.

Legrain, P., & Selig, L. (2000). Genome-wide protein interaction maps using two-hybrid systems. FEBS Lett, 25, 32-36.

Lehner, B., & Fraser, A. G. (2004). A first-draft human protein-interaction map. Genome Biol, 5, R63.

Li, Y., Agarwal, P., & Rajagopalan, D. (2008). A global pathway crosstalk network. Bioinformatics, 24, 1442-1447.

Liu, Z. P., Wang, Y., Zhang, X. S., & Chen, L. (2010). Identifying dysfunctional crosstalk of pathways in various regions of Alzheimer's disease brains. BMC Syst Biol, 4(2), S11.

Mathivanan, S., Periaswamy, B., & Gandhi, T. K. (2006). An evaluation of human protein-protein interaction data in the public domain. BMC Bioinformatics, 7(l5), S19.

Merico, D., Gfeller, D. & Bader, G., D. (2009) How to visually interpret biological data using networks. Nat Biotechnol.; 27, 921-924. doi: 10.1038/nbt.1567.

Michael, B., Dillencourt., Hannan., Samet., Markku., & Tamminen. (1992). A general approach to connected-component labeling for arbitrary image representations. J. ACM.

Nandi, D., Tahiliani, P., Kumar, A., & Chandu, D. (2006). The ubiquitin-proteasome system. J Biosci, 146(1), 137-140.

Paley, S. M. & Karp. P. D. (2006). The Pathways Tools cellular overview diagram and Omics Viewer. Nucleic Acids Res, 34, 3771-3778.

Pavlopoulos, G. A., Wegener, A. L., & Schneider, R. (2008). A survey of visualization tools for biological network analysis. BioData Mining, 1, 12.

Pawson, T., Warner, N. (2007). Oncogenic re-wiring of cellular signaling pathways. Oncogene, 26, 1268-1275.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., & Seraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. Nat. Biotechnol, 17, 1030–1032.

Rocha, I., Maia, P., Evangelista, P., Vilaca, P., Soares, S., Pinto, J. P., Nielsen, J., Patil, K. R., Ferreira, E. C., & Rocha, M. (2010). OptFlux: an open-source software platform for in silico metabolic engineering. BMC Syst Biol, 4, 45.

Samet, H., & Tamminen, M. (1988). Efficient Component Labeling of Images of Arbitrary Dimension Represented by Linear Bintrees. IEEE Transactions on Pattern Analysis and Machine Intelligence (TIEEE Trans). Pattern Anal. Mach. Intell, 10, 579.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res, 13, 2498-2504.

Su, G., Kuchinsky, A., Morris, J. H., States, D. J., & Meng, F. (2010). GLay: community structure analysis of biological networks. Bioinformatics, 26, 3135-3137.

Suderman, M., & Hallett, M. (2007). Tools for visually exploring biological network. Bioinformatics, 23, 2651-2659.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S., Pawson, T., Morris, Q., & Wrana, J. L. (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol, 27, 199-204. doi: 10.1038/nbt.1522.

Villéger, A. C, Pettifer, S. R., & Kell, D. B. (2010). Arcadia: a visualization tool for metabolic pathways. Bioinformatics, 26, 1470-1471.

Wang, Z., Li, Y., & Kong, D. (2010). Cross-talk between miRNA and Notch signaling pathways in tumor development and progression. Cancer Lett, 292, 141-148.

Zhou, D., & He. Y. (2008). Extracting interactions between proteins from the literature. J Biomed Inform, 41, 393-407.