

Cytoscape and its plugins

Aija Niissalo

Department of Computer Science

University of Helsinki

Finland

May 25, 2007

Abstract

Cytoscape is a software environment for integrated models of biomolecular interaction networks. It is developed and continually developing as an open source software project and offers an graphical user interface (GUI) for network modeling. This Java-based framework can visualize and integrate protein-protein, protein-DNA and genetic interactions with high-throughput expression data and molecular state information. With the help of Cytoscape and its plugins a researcher can e.g. identifying functional subnetworks in large-scale datasets. The cell interactomics can be visualized. Additionally, the produced images with Cytoscape are high quality and can be used in scientific papers. This article gives an overview of the tool and its plugins. This writing is not a manual or detailed user tutorial, because these can be found in the Cytoscape website (www.cytoscape.org), where also Cytoscape itself is downloadable.

1 Introduction

If we want to develop predictive biology - e.g. in genomics, proteomics or metabolomics and further on in medical and physiological understanding - one important principle is that in most cases it is not individual genes but rather biological pathways and networks that are responsible of the wide diversity of an organism's response or its phenotype [Quackenbush, 2007]. We must develop tools to understand the structures of the networks and the rules that govern the interactions between elements in a biological system. Biological knowledge integration and analysis from biodatabases lead to better interpretation

of experimental data of genomic and proteomic assays. It's easy to understand the importance of finding 'common languages' in integration. We need data representation methods and standards [Brazma et al, 2006]. Visualization of numerically non-intuitive or vast data is one natural way, which humans can easily interpret.

Nowadays high-throughput technologies and biodatabases provide us with increasing amount of genomic and proteomic data. A wide variety of different analysis tools are available for extracting knowledge out of that large-scale data. An individual scientist may, though, need some programming or other additional skills to use those tools or software packages. Furthermore, some useful tools are commercial, and academic researchers can't afford them. Especially visualization and appropriate partitioning/integration of the large-scale data give insight to the modeling of biological systems. This is particularly true for interaction networks which are fundamental in the understanding of cellular processes. Cytoscape [Shannon et al, 2003] is an easy-to-use platform for visualization biomolecular interaction networks. These interactions can be integrated with gene expression profiles and other functional genomics data. Cytoscape supports the development of plugin tools that extends the core functionality. Cytoscape was originally designed by Dr. Trey Ideker (Department of Bioengineering, University of California at San Diego, USA) in 1999 and made public in July, 2002. The methods presented in Ideker et al [Ideker et al, 2002] paper 'Discovering regulatory and signaling circuits in molecular interaction networks' are, also, implemented in Cytoscape. Cytoscape is now jointly developed with several groups and researchers.

In the next chapter I describe the requirements and basic utilities of Cytoscape. Following these I provide the short description of the plugin capabilities. In the subsequent chapters I give a more detailed picture of some of these useful plugin-features. At the end I present some of my own comments and a summary. All along the way I'm not intended to give user instructions - I try to sketch a picture how computer visualizing and analyzing tools used in comparative network analysis help us to create more understanding of the molecular mechanism in the cell. The reader can, in addition to my sketch, use different knowledge colors of his and hers own for a more detailed painting. By modeling cellular networks we get hypotheses which can then further on be tested in wet-lab experiments.

2 Requirements and basics

Cytoscape is a Java application that runs on Linux, Windows or Mac OS X and it's developed under the GNU LGPL (Lesser General Public License). Cytoscape is a stand-alone-tool but offers an open plugin-architecture, allowing anyone to add functionality by writing one. Plugin licenses are individual, but mostly free to use at least for academic purposes. Cytoscape requires Java version 1.5.0. Java and Cytoscape are freely downloadable and both installations are easy. Both platforms offer detailed tutorials for getting started (java.sun.com/docs/books/tutorial, www.cytoscape.org/tut/tutorial.php). Not much is required using the tools. For developing new features Java and Cytoscape offers great APIs (Application Program Interface) with HTML-documentations.

Also additional documentation for plugin writers is available in Cytoscape website. The source code of plugins can be available, but the policy is that it is not required. A developer needs only to provide other user with jar-file of the plugin. The jar-file needs then to be put in the plugin-folder of Cytoscape-files and it can then be used.

Also from CSC's (the Finnish IT center for science) website (www.csc.fi), we can find useful material, slides, exercises and videos, when we familiarize ourself with Cytoscape.

2.1 Core features

The central organizing metaphor of Cytoscape is a network graph [Shannon et al, 2003]. Molecular species represented as nodes and molecular interactions represented as links/edges between those nodes. Core software components offer basic functionality for integrating arbitrary data on the graph and visual representation of that. Data are integrated with the graph model using attributes. Name value pairs map node or edge names to specific data values. Functionality includes methods for graph layout and setting visual properties according to node or edge attributes. One can also customize network data and display those by using different visual styles. Cytoscape supports several automated network layout algorithms, e.g. spring-embedder-layout. Selectively displaying subset of large interaction network can make graph more comprehensive. There are also selection and filtering tools, such as a differential expression filter by expression data values. In the figure 1 we can see

basic functionality alternatives in the menu bars and in the panel layout. These are adjustable and many attributes can be determined by the user. From the menu additional tools are available. Cytoscape basics are related to visual-

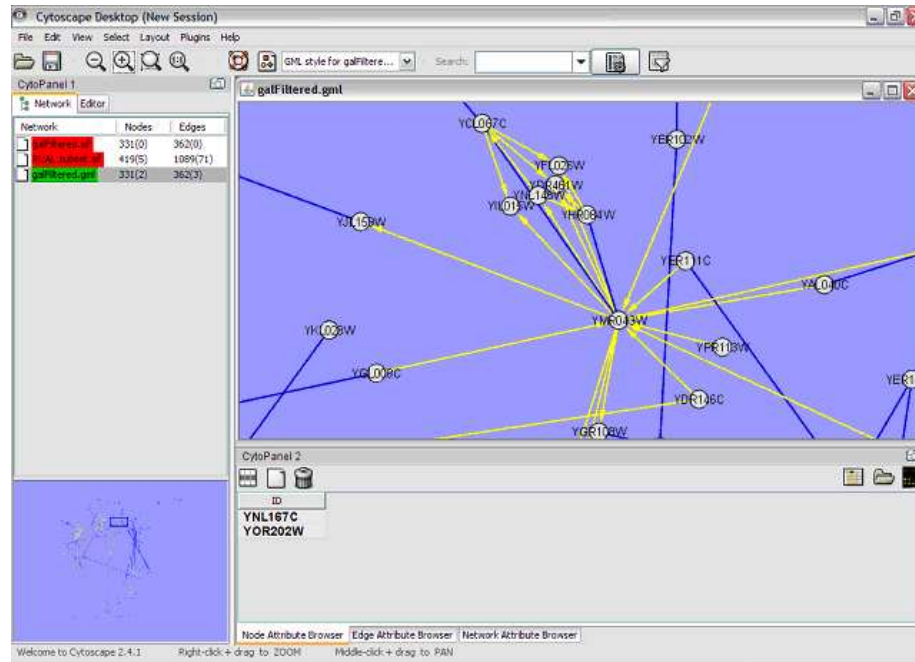


Figure 1: GUI of Cytoscape

ization but when used with its plugins Cytoscape can be used to analyzing purposes as well. Integrating network data with other data e.g. expression data [Schlitt and Brazma, 2006] and Gene Ontology annotations (e.g. using BiNGO, a Cytoscape-plugin [Maere, 2005]) is great advantage in analysis.

2.2 Biological network and expression data

Getting interactions from public databases is the self evident feature, which a tool like this ought to have. Protein protein interactions from BIND and TRANSFAC databases are suitable for Cytoscape. Functional annotations from Gene Ontology (GO) and KEGG databases can be used with Cytoscape. Biological models represented with Systems Biology Markup Language (SBML), BioPAX, PSI-MI, delimited text or Excel-files can be used. The other way round, some other tool export Cytoscape compatible formats, e.g BRM (The

Bioinformatics Resource Manager, [Shah, 2007]) and iArray (Integrative Array Analyser, [Pan et al, 2006]). These were examples - more accurate data format compatibility can be found in the Cytoscape site and few examples comes further in this article.

A quite comprehensive general (not just Cytoscape) data format and standards review recently published can be found in [Brazma et al, 2006], many of the acronyms mentioned here comes more understandable reading it.

One example of integrated network-genetic data usage in Cytoscape is following: Users may select nodes involved in a threshold number of interactions, nodes that share a particular GO annotation, or nodes whose gene expression levels change significantly in one or more conditions according to p-values loaded with the gene expression data.

3 Plugins

The before mentioned plugins add variety of visual, data format and algorithmic capabilities. Plugins are separate works which use Cytoscape as a Java code library. Ideker et al [Ideker et al, 2002] used Cytoscape with plugin, now named jActiveModules, for discovering regulatory and signaling circuit in molecular interaction networks in 2002. So, the plugin-idea was there from the beginning. This particular plugin identify significant active subnetworks.

Today there are 38 plugins available in the Cytoscape website (www.cytoscape.org/plugins2.php). These are divided to five categories: analysis, network and attribute I/O, network inference, functional enrichment and communication/scripting plugins. And they are used for analyzing existing networks, importing networks and attributes in different file formats, inferring new networks, functional enrichment of networks and communicating with or scripting Cytoscape, respectively.

For instance, interfaces for other programming languages are provided by a plugin, called CytoTalk, which runs a simple internal XML-RPC server from within Cytoscape that allows the current network and its various attributes to be manipulated from an external process that is XML-RPC capable. So interaction with Cytoscape from Perl, Python, R, shell scripts, C or C++ programs or external Java processes is possible. Many plugins for different database interaction are developed for both public and custom databases.

There is a plugin, tYNA, that allows one to upload networks to and download

networks from the tYNA database. TYNA is a Web system for managing, comparing and mining multiple networks, both directed and undirected. It can also identify defective cliques, calculate global statistics and identify hubs (nodes with many connections) and bottlenecks [Yip et al, 2006]. Combining the network sharing capability and other unique features provided by tYNA and the advanced visualization and analysis facilities of Cytoscape can be powerful.

A plugin, called cPath, query, retrieve and visualize interactions from the MSKCC Cancer Pathway database. The Agilent Literature Search plugin builds networks by extracting interactions from scientific literature.

Celebral is a plugin that can automatically generate a view of the network in the style of traditional pathway diagrams from an interaction network and sub cellular localization annotation. It gives an intuitive interface for the exploration of a biological pathway of a system [Barsky et al, 2007]. Localization information can be catered e.g. from HPRD database. GenePro [Vlasblom et al, 2006] plugin provides several integrative and interactive visualization and analysis tools for PPI networks.

And one example more, named Motif Finder, runs a Gibbs sampling motif detector on sequences for nodes in a Cytoscape network. In the next sub chapters we can find a few more examples of the usefulness of the plugins.

3.1 Network creation

BioNetBuilder [Avila-Campillo et al, 2007] offers an interface to create biological networks integrated from several databases. Networks can be created for about 1500 organisms, e.g. common model organisms and human. Databases that can be used are for instance: DIP, BIND, Prolinks, KEGG, HPRD, The BioGrid and GO. Standardizing in interaction databases aren't complete and naming conventions either. So, tools are needed, at least those that aren't expensive. This plugin allows for creation of networks composed of metabolic relationships, protein and protein-DNA interactions and associations from comparative genomics. After creation the network can be saved, viewed, annotated or analyzed by other features of Cytoscape. For instance, the CyGaggle plugin allows access to many non-Cytoscape analysis tools.

BioNetBuilder consists of a client and a secure Java servlet. XML-RPC (Apache Software Foundation, 2006) is used for communication between the client and servlet. The servlet consists of several database handlers, which make queries to read-only interaction MySQL databases. With a synonym-

resolution system handler the plugin can integrate data from databases that identify their genes with different ID types. BioNetBuilder doesn't require a rigid database schema, file-format or data-model that new data sources have to conform to. New database interfaces can be added to the server with source data from several possible formats being used with little formatting costs. As a part of the tool the tool-implementers maintain a server which responds to requests made by users. They also provide database initialization and updating tools for users' own mirror BioNetBuilder servlet and databases. The plugin is robust and scalable solution for building and visualizing networks for species for which public data can be found.

3.2 Domain interaction networks

The known functions of the the interacting protein domains provides important information on cellular function of protein interactions and complexes. It is useful to decompose protein-protein interactions into their constituent domains before being able to functionally characterize them further and to model the spatial structure of protein complexes.

DomainNetworkBuilder [Albrecht et al, 2005] plugin decomposes protein networks into domain-domain interactions and generates a new network of interacting domains. Basically, it transforms each protein node into a chain of consecutive domain nodes and constructs a putative network of interacting domain nodes.

It queries in-house database, which contains protein information from the UniProt database, domain information from the Pfam database, and domain interaction information from the InterDom database. Like the interaction type 'pp' used by Cytoscape for protein-protein interaction edge, the plugin introduces three new edge types for domain nodes. There is 'dl' for a domain linker between domain nodes of the same protein, 'pl' for a protein linker between a protein and domain node the same protein, and 'dd' for a domain-domain interaction between different proteins. The coloring schema as well as the different shapes of protein and domain nodes and interaction edges can be changed using the visualization tools of Cytoscape. The domain network can be saved in file formats supported by Cytoscape.

So, the plugin provides tools for investigating and visualizing protein interactions on detailed molecular level of domain and binding sites. This helps in the validation and functional analysis of observed and predicted protein inter-

actions. This is important when new experiments are designed and 3D models are constructed.

3.3 Expression data

Expression Correlation Network plugin clusters expression data. More of this can be found from Cytoscape site. It enables the users to correlate genes or conditions in an expression matrix file which is loaded into Cytoscape. The correlations are visualized as a network. A condition correlation network is an good alternate way of representing expression condition clustering results. Cluster visualization can sometimes be easier in that way than the normal heatmap view.

In addition to previous, using other Cytoscape features by coloring nodes according expression data and defining node shape by expression significance data Cytoscape user can put different expression results in more comprehensive context.

3.4 Clustering

Network clustering is available through the MCODE plugin. It finds clusters of highly interconnected regions in networks. The properties of MCODE are listed in MCODE [www-page](http://www.baderlab.org/Software/MCODE) ([baderlab.org/Software/MCODE](http://www.baderlab.org/Software/MCODE)): fast network clustering, fine-tuning of results with numerous node-scoring and cluster-finding parameters, interactive cluster boundary and content exploration, multiple result set management and cluster sub-network creation and plain text export. Clusters mean different things in different types of networks. Clusters in a protein-protein interaction network have been shown to represent protein complexes and parts of pathways. Clusters in a protein similarity network represent protein families. Visualization of those may give a new insight and hypothesis to the user of Cytoscape. A good user manual can be found in the MCODE [www-pages](http://www.baderlab.org/Software/MCODE).

The algorithm used in MCODE operates in three stages, vertex weighting, complex prediction and optionally post-processing to filter or add proteins in the resulting complexes by certain connectivity criteria. A network of interacting molecules can be intuitively modeled as a graph, where vertices are molecules and edges are molecular interactions. To find locally dense regions of a graph, MCODE uses a vertex-weighting scheme based on the clustering coefficient, which measures 'cliquishness' of the neighborhood of a vertex. A clique is defined

as a maximally connected graph. The first stage of MCODE, vertex weighting, weights all vertices based on their local network density. The second stage, molecular complex prediction, takes as input the vertex weighted graph, seeds a complex with the highest weighted vertex and recursively moves outward from the seed vertex, including vertices in the complex whose weight is above a given threshold. The third stage is post-processing. Complexes are filtered if they do not contain at least a 2-core (graph of minimum degree 2). A detailed description of the algorithm (these few sentences are from there) can be found in [Bader and Hogue, 2003].

3.5 GO, Gene Ontology

Gene Ontology is a three-fold conceptual hierarchical vocabulary system of genes applying commonly to different species. It is a acyclic relationship graph, where the connections mean is-a and part-of relations. It is continuously commonly, by a project of biological community, curated so that the IDs of the conceptions don't get changed. The curation is based on evidences. It has three separate classifications within itself. Genes have three-fold classification according to which biological process they belong, what molecular function they have and in what cellular component they belong. A good place to learn more is the community web site in www.geneontology.org.

BiNGO (the Biological Networks Gene Ontology tool) [Maere, 2005] determines which Gene Ontology (GO) categories are statistically over-represented in a set of genes. It provides annotation for a wide range of organisms. BiNGO supports the use of GOSlim ontologies, as well as custom ontologies and annotations. The graphs can be viewed, laid out, modified and saved in various manners. The default annotations are parsed from the GO information available from NCBI. Additionally the before mentioned other classification systems can be used. BiNGO uses, for instance, Bonferroni correction to control false positive rate in the statistical analysis. Other evidence should be used along the suggestions that BiNGO gives in the interpretation of the results.

Another plugin's, Golorize, layout algorithm that determines the placement of the nodes based on both their connection and class structure can be used with BiNGO. Golorize uses GO annotations as source of external class information to direct the layout process and to emphasize the biological function of the nodes [Garcia et al, 2007].

3.6 Network topology

In cell biology, it is useful to see whether the connectivity of genes of one functional type is similar to some characteristic shape, like a feed-forward loop [Ferro et al, 2007]. NetMatch plugin searches biological networks for subcomponents matching a given query, queries may also be approximate. NetMatch supports subgraph matching queries against a target network, which are previously loaded into the Cytoscape workspace. Approximate queries are special subgraphs that may contain nodes and edges labeled with a special wildcard symbol. The plugin handles target and query graphs with multi-edges, loops and a list of attributes for each edge and node.

The tool can be set to search labeled or unlabeled, directed or undirected networks. Users may query in NetMatch by loading from an existing file, importing from the Cytoscape workspace or drawing using the NetMatch query drawing tool.

4 Conclusion

There are future plans and numerous suggestions for plugin implementations in the Cytoscape site. They arise from the users' needs and that is seen to be a good point when developing useful tool in bioinformatics. John Quackenbush's and some of his colleagues' opinions of standardizing [Quackenbush et al, 2006, Quackenbush, 2006] seems to apply to this issue as well.

Although I haven't seriously used Cytoscape, I have found it useful with some other tools that user can make his or hers own notes along the data and results. Maybe some simple note editor which can also make notes on an image (produced by Cytoscape) would be useful. A simple-small-size Java-implementation of such as a plugin isn't too hard, but serious implementation needs usage of Cytoscape, so the real need of it comes clear. Maybe some kind of custom annotations (general version, not just related to some plugins - as now seems to be) would be useful, too. I expect, we'll see fine quality animated network function visualizations as our knowledge of interactomics grows.

I'd like to see beyond the visual images which Cytoscape so nicely gives. A review of Sharan and Ideker [Sharan and Ideker, 2006] concerning network comparison gives a nice broader vision related to some issues written in this article. I suggest it for further readings. They categorize network comparison to alignment, integration and querying. The meaning and goals of these three

comparisons are ('shortened version'): *alignment* meaning the global identification of similar/dissimilar regions to find functional protein modules and study as well as predict network interaction and evolution, *integration* meaning the process of combining networks (at least two networks of different types for the same species) and finding interrelations and interaction predictions, *querying* meaning comparing subnetwork module versus a network and identifying duplicated or conserved modules. The writers compare the more mature sequence analysis and the 'youngster' biological networks analysis, the first one having more easier structure to interpret.

The network comparison and evolution considered from the theoretical point of view have quite significant history - much older than this new interactomics history. About the network evolution, Barabasi and Oltavai have written a review [Barabasi and Oltavai, 2004], where they consider gene duplication as main reason for the evolution of the so called scale free networks. Another fundamental issue in 'gene nature' is the sequence mutation occurrence. This naturally changes the interface between interacting proteins. At least these two things should be in mind, when interpreting the visualized networks.

5 Summary

Researchers may report the developing project if they use Cytoscape. A quite large user article set can be found in the Cytoscape website. Among others the research articles concern identification of protein complexes, phenotyping and interactome mapping, phenotype analysis using network motifs derived from changes in regulatory network dynamics, systematic interpretation of genetic interactions using protein networks and conserved patterns of protein interaction in multiple species (www.cytoscape.org/pubs.php). All this kind of usage suggest that Cytoscape is a powerful graph layout-analysis tools. Just to remind ourself from the beginning:

It can organize multiple networks. In those large networks navigation is easy. The networks can be filtered and graph node and edge attributes mapped to visual attributes. Visual styles can be defined for later use.

A graph can have node and edge attributes e.g. expression data, interaction type and GO function mapping and those can be visualized with various node/edge size, shape, color or font. User can take continuous gene expression data and visualize it as continuous node colors.

In the website a huge list of features that have been discussed for future inclusion into Cytoscape can be found. This list is prioritized each year. The idea of developing according to users needs seems powerful.

References

- [Avila-Campillo et al, 2007] Avila-Campillo, I., Drew, K., Lin, J., Reiss, D. and Bonneau, R., Bionetbuilder: automatic integration of biological networks. *Bioinformatics*, 23, pages 392–393.
- [Albrecht et al, 2005] Albrecht, M., Huthmacher, C., Tosatto, S. and Lengauer, T., Decomposing protein networks into domain-domain interactions. *Bioinformatics*, 21, pages 3ii220–ii221.
- [Barsky et al, 2007] Barsky, A., Gardy, J., Hancock, R. and Munzner, T., Cerebral: a cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics*, 23, pages 1040–1042.
- [Bader and Hogue, 2003] Bader, G. and Hogue, C., An automated method for finding molecular complexes in large protein interaction networks. *Bioinformatics*, 4, pages doi:10.1186/1471-2105-4-2.
- [Brazma et al, 2006] Brazma, A., Krestyaninova, M. and Sarkans, U., Standards for systems biology. *Nature*, 7, pages 593–605.
- [Barabasi and Oltavai, 2004] Barabasi, A. and Oltavai, Z., Network biology: understanding the cell’s functional organization. *Nature Reviews*, 5, pages 102–113.
- [Ferro et al, 2007] Ferro, A., Giugno, R., Pulvirenti, A., Skripin, D., Bader, G. and Shasha, D., Netmatch:a cytoscape plugin for searching biological networks. *Bioinformatics*, 23, pages 910–912.
- [Garcia et al, 2007] Garcia, O., Saveanu, C., Cline, M., Fromont-Racine, M., Jacquier, A., Schwikowski, B. and Aittokallio, T., Golorize: a cytoscape plug-in for network visualization with gene ontology-based layout and coloring. *Bioinformatics*, 23, pages 394–396.
- [Ideker et al, 2002] Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A., Discovering regulatory and signaling circuit in molecular interaction networks. *Bioinformatics*, 14, pages 233–240.
- [Maere, 2005] Maere, S., Heymans, K. and Kuiper, M., Bingo: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21, pages 3448–3449.

- [Pan et al, 2006] Pan, F., Kamath, K., Zhang, K., Pulapura, S., Achar, A., Nunez-Iglesias, J., Huang, Y., Yan, X., Han, J., Hu, H., Xu, M., Hu, J. and Zhou, X., Integrative array analyzer: a software package for analysis of cross-platform and cross-species microarray data. *Bioinformatics*, 22, pages 1665–1667.
- [Quackenbush et al, 2006] Quackenbush, J., C., S., C., B., Brazma, A., R., G., W., H., Irizarry, R., Salit, M., Sherlock, G., Spellman, P. and N., W., Top-down standards will not serve systems biology. *Nature*, 440, page 24.
- [Quackenbush, 2006] Quackenbush, J., Standardizing the standards. *Molecular Systems Biology*, page doi: 10.1038.
- [Quackenbush, 2007] Quackenbush, J., Extracting biology from high-dimensional biological data. *The Journal of Experimental Biology*, 210, pages 1507–1517.
- [Schlitt and Brazma, 2006] Schlitt, T. and Brazma, A., Modelling in molecular biology: describing transcription regulatory networks at different scales. *Phil. Trans. R. Soc. B*, 361, pages 483–494.
- [Sharan and Ideker, 2006] Sharan, R. and Ideker, T., Modelling cellular machinery through biological network comparison. *Nature Biotechnology*, 24.
- [Shannon et al, 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N., J., W., Ramage, D., Amin, N., Schwikowski, B. and Ideker, T., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13,11(2003), pages 2498–2504.
- [Shah, 2007] Shah, A., Singhal, M., Klicker, K., Stephan, E., Wiley, H. and Waters, K., Enabling high-throughput data management for systems biology: The bioinformatics resource manager. *Bioinformatics*, 23, pages 906–909.
- [Vlasblom et al, 2006] Vlasblom, J., Wu, S., Pu, S., Superina, M., Liu, G., Orsi, C. and Wodak, S., Genepro: a cytoscape plug-in for advanced visualization and analysis of interaction networks. *Bioinformatics*, 22, pages 2178–2179.

- [Yip et al, 2006] Yip, K., Yu, H., Kim, P., Schultz, M. and Gerstein, M., The tyna platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics*, 22, pages 2968–2980.