BMC Bioinformatics

BioMed Central

Software

Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases

Seth I Berger†, Jeremy M Posner† and Avi Ma'ayan*

Address: Department of Pharmacology and Systems Therapeutics, Mount Sinai School of Medicine, 1425 Madison Avenue, New York, 10029, New York, USA

Email: Seth I Berger - seth.berger@mssm.edu; Jeremy M Posner - jposner@panix.com; Avi Ma'ayan* - avi.maayan@mssm.edu

* Corresponding author †Equal contributors

Published: 4 October 2007

BMC Bioinformatics 2007, **8**:372 doi:10.1186/1471-2105-8-372

Received: 7 May 2007
Accepted: 4 October 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/372

Abstract

Background: In recent years, mammalian protein-protein interaction network databases have been developed. The interactions in these databases are either extracted manually from low-throughput experimental biomedical research literature, extracted automatically from literature using techniques such as natural language processing (NLP), generated experimentally using high-throughput methods such as yeast-2-hybrid screens, or interactions are predicted using an assortment of computational approaches. Genes or proteins identified as significantly changing in proteomic experiments, or identified as susceptibility disease genes in genomic studies, can be placed in the context of protein interaction networks in order to assign these genes and proteins to pathways and protein complexes.

Results: Genes2Networks is a software system that integrates the content of ten mammalian interaction network datasets. Filtering techniques to prune low-confidence interactions were implemented. Genes2Networks is delivered as a web-based service using AJAX. The system can be used to extract relevant subnetworks created from "seed" lists of human Entrez gene symbols. The output includes a dynamic linkable three color web-based network map, with a statistical analysis report that identifies significant intermediate nodes used to connect the seed list.

Conclusion: Genes2Networks is powerful web-based software that can help experimental biologists to interpret lists of genes and proteins such as those commonly produced through genomic and proteomic experiments, as well as lists of genes and proteins associated with disease processes. This system can be used to find relationships between genes and proteins from seed lists, and predict additional genes or proteins that may play key roles in common pathways or protein complexes.

Background

The rapid increase in experimentally identified binary interactions between proteins has brought us to a stage where we are now able to start viewing how these interactions and components come together to form large functional regulatory networks [1]. However, it is impossible for researchers to keep up with the ever expanding literature. The emergence of high-throughput experimental technologies, such as yeast-2-hybrid screens [2,3], cDNA microarrays [4,5] and mass-spectrometry [6], as well as databases that mine legacy experimental literature [7,8] allow for the construction of large networks. Networks, formally graphs, are simple abstract representations of biomolecular interactions where cellular components are represented as nodes, and interactions connect these nodes through links.

The construction of cellular network datasets has several valuable uses. Network representation allows for extraction of global topological statistical and structural properties such as connectivity distribution [9], clustering [10], and the identification of network motifs [11] or graphlets [12]. These measurements provide clues about the design principles of intracellular organization. Interaction network datasets can also be used to predict unidentified interactions [13,14], or used as a starting point for quantitative computational modeling [15]. Additionally, interaction networks can assist in interpreting experimental results when identified lists of proteins or genes from proteomic or genomics experiments or computational studies can be placed in their contextual local interaction networks [16].

Methods

Our aim in developing the Genes2Network software is to provide cell- and molecular-experimental biologists as well as computational biologists with a user-friendly tool for creating subnetworks from lists of mammalian genes or proteins by connecting these genes or proteins using known protein-protein interactions. To accomplish this task we developed a large-scale high-quality mammalian protein-protein interaction database. This database was created by consolidating databases containing mostly low-throughput literature-based protein interaction data extracted manually by expert biologists, but also data generated from high-throughput methods. To develop Genes2Networks, we consolidated ten currently available mammalian protein interaction network datasets into one large dataset. To prune out interactions of low confidence, a simple filter was implemented. Genes2Networks is delivered as a web interface application. This tool can be used to extract relevant subnetworks given lists of gene or protein names. The input to the system is a list of Entrez gene symbols. The system uses the merged datasets made of selected databases to find interactions between the nodes in the seed list. The merged datasets can be filtered based on user preferences concerning the maximum number of interactions a reference can provide, and the minimum number of references required for interactions to be included. The resultant filtered dataset serves as a reference network for exploring, by depth-first traversal, paths between the seed nodes. Nodes that fall on paths shorter than a user defined path length between seed nodes are included as intermediates in the outputted subnetwork. The system's output includes a statistical analysis report, and a three color network map, highlighting the seed nodes in one color, the significant intermediates in another color, and the non-significant intermediates in a third color. The statistical analysis provides a list of intermediate nodes used to connect the gene names, sorted by significance of specificity to interact with nodes from the seed list. This process is illustrated in Figure 1.

Developing a high-quality large-scale mammalian protein interaction network

We used only mammalian (mouse/rat/human) interactions recorded in the following datasets: BIND [17], HPRD [18], IntAct [19], DIP [20], MINT [21], Rual et al. [22], Stelzl et al. [23], Ma'ayan et al. [24], PDZBase [25], and PPID [19,26]. All interactions from these databases/datasets were determined experimentally and include a PubMed reference to the primary research article that describes the experiments used to identify the interactions. Some of the databases contain interactions that were manually extracted from the literature (e.g. HPRD); some datasets are the result of high throughput experimental data (e.g. Rual et al. and Stelzl et al.); whereas some databases contain both low and high-throughput interactions (e.g. BIND, IntAct, and DIP). Consolidation of interactions from the ten different network databases was accomplished by combining human/mouse/rat Entrez gene symbols using information from Swiss-Prot [27]. The consolidated network created from the ten datasets contains 44,877 interactions and 11,033 nodes. This network is stored in a structured text flat-file-space-delimited format. This file is loaded into the program using a hash data structure implemented in c language for fast loading and access. We do not include in this initial implementation datasets of interactions created via *in-silico ab-initio* interaction prediction methods or model organisms orthologs interactions such as those collected in OPHID [27], HPID [28], IntNetDB [29], and POINT [30]. The datasets we used describe mostly binary interactions, but in rare cases complexes containing more than two proteins are listed. These were excluded from the merged dataset. Nodes in the ten datasets are provided with accession codes linking them to entries describing genes and proteins in databases such as Swiss-Prot [31] and NCBI's Entrez Gene [32]. HPRD [18] and PPID [19,26] are not included in the public web interface application since these databases require a license for redistribution. Currently, HPRD and PPID data are only

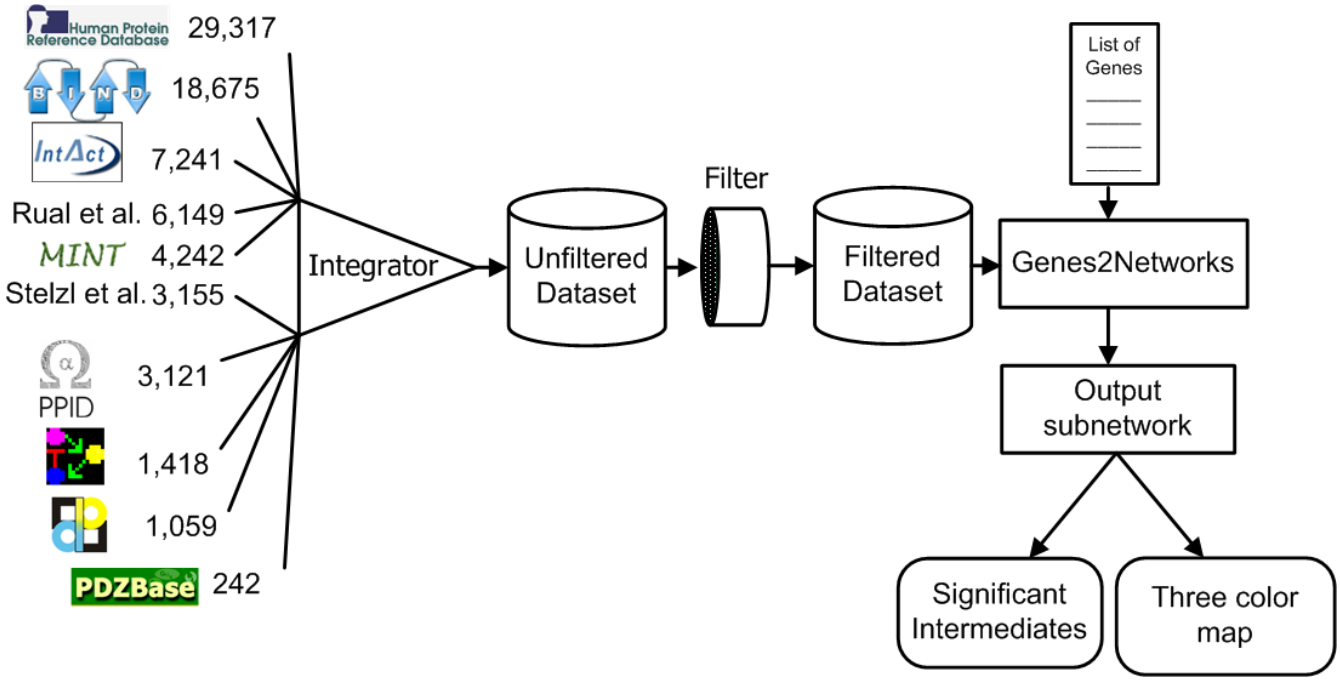

Figure 1
Ten mammalian PPI network datasets were consolidated into one dataset, and then filtered by excluding interactions originating from articles that contributed many interactions, or by excluding interactions with few references. The filtered merged dataset is then used to analyze lists of gene or protein names by outputting a subnetwork with nodes in three different colors: seed, significant, insignificant. The output also includes a statistical report that ranks intermediate nodes based on their specificity to interact with the seed list.

available to internal users at Mount Sinai School of Medicine.

Filtering

Many of the interactions and components listed in the ten databases that we used are the result of high-throughput experiments such as yeast-2-hybrid screens [2,3], and mass-spectrometry [6]. These interactions are considered low-quality since these techniques often report many false positives [33]. Thus, we applied a simple filtering approach allowing users to exclude interactions originating from articles that provide many interactions, and/or include only interactions reported by several different papers. The rationale for this filtering approach is the assumption that a research article that reports many interactions is likely reporting the results of a high-throughput technique which tends to produce many false positives. Alternatively, interactions that are reported in many different research articles, and appear in multiple databases, can be given more confidence because these interactions have been reported multiple times independently. Hence, users may select to include only interactions from low-throughput studies with multiple references to improve the reliability of the consolidated network. Users are presented with list-boxes and text-boxes that allow adjustment of the filtering thresholds. More sophisticated filtering techniques implementing machine learning technologies such as support vector machines (SVM) [34], and taking into account more knowledge about the interactions (i.e. experimental method used, impact factor of journals, etc.) are planned for future implementations.

Web interface

To enhance accessibility to the core Genes2Networks software, we developed a state-of-the-art web-based interface. This interface allows users to input lists of human Entrez Gene symbols in a textbox or through uploading a text file. As genes are added, the system validates the entries using NCBI's e-utils. The validation is achieved by searching the NCBI gene database, with the input entry, while restricting the organism to human. If an exact match is not found, the user is presented with a list of suggestions with links to choose the intended matching entry. By clicking on a highlighted gene symbol from the list of suggestions, the gene can be added to the seed list.

Using the merged consolidated network reference database, the program outputs subnetworks that describe all found interactions and nodes on paths connecting the list of inputted gene symbols. The web interface provides users

with full access to configure which databases to include in the consolidated reference network that is used to connect the genes. Additionally, users can upload other network databases for inclusion in the reference dataset. These additional networks can be consolidated with the provided networks. The output subnetwork is visualized using a dynamical web-enable AJAX viewer called AVIS [35]. The viewer allows browsing, zooming and panning, and linking to interaction resources. The user can configure the colors of the outputted nodes so that the seed-list genes, intermediate genes that are above a specified Z-score and the rest of the nodes are displayed in different colors. The user can also adjust the maximum number of steps/hops to use in order to find paths between the nodes in the seed list to connect the seed list genes. Steps/hops are the number links (not nodes) needed to connect the inputted seed list. Additionally, the program outputs a statistical report that ranks intermediates used to connect the genes based on their specificity to interact with the seed list. As the user adjusts the settings, changes in the resulting network are automatically redisplayed. A representative screenshot of the system is illustrated in Figure 2.

Significant intermediates

The output subnetworks produced by Genes2Networks contain nodes, mostly proteins, which were not originally provided by the user as input. We call these nodes "intermediate nodes". Some of these intermediate nodes may be present in the output subnetwork because these intermediates are highly connected nodes (hubs) in the consolidated reference network used to connect the seed-list. On the other hand, intermediate nodes may be specific to interact with components from the inputted seed list. If these intermediates are specific, it may be beneficial for the user to identify them as potential specific regulators and specific participants in pathways, protein complexes and modules involving the input seed list. For this, Genes2Networks outputs a Z-score value of the significance of intermediates in the outputted subnetwork. The Z-score is computed for each intermediate node using a binomial proportions test [36] as follows:

$$z = \frac{\left(\frac{a}{c} - \frac{b}{d}\right)}{\sqrt{\frac{\frac{b}{d}\cdot\left(1-\frac{b}{d}\right)}{d}}} \quad (1)$$

Where "a" equals the links from the intermediate node being examined to nodes from the input seed list, "b" equals the total links for the intermediate node in the consolidated background reference network, "c" is the total links in the outputted subnetwork, and "d" is the total links in the consolidated background reference network. The outputted ranked list of intermediates is displayed underneath the subnetwork map viewer.

Discussion and Conclusion

Several commercial and academic initiatives have been attempting to address the need for integration, consolidation, visualization, querying and organization of information about binary mammalian protein-protein interactions and signaling pathways from sparse sources. For example, Cytoscape [37] is Java-based desktop software for protein and gene network visualization. Cytoscape's several plug-ins allow for analysis and integration of experimental data as well as incorporation with Gene Ontology [38]. One Cytoscape plug-in, called cPath [39], is a data warehouse that joins together databases stored in PSI-MI XML format [19]. Other similar software platforms include: PIANA [40], Pathway Studio [7], ProViz [41], PATIKA [42], and Ingenuity. Some are commercial products and some were developed by academic laboratories and are freely available. Genes2Networks provides several advantages over existing systems; the consolidated network made from the ten databases, after filtering, is a high quality yet comprehensive dataset; the user interface is an intuitive web-based Web 2.0 enabled application; the systems is free for academic users; the system provides predictions about intermediate components and their involvement with the proteins and genes from seed lists by ranking intermediates according to their specificity to interact with the seed list. Genes2Networks is suitable for analysis of diverse proteomic and genomic experimental results. The web interface and visualization provide easy access and a user friendly environment eliminating the need for training.

Availability and requirements

Project name: Genes2Networks

Project home page: http://actin.pharm.mssm.edu/genes2networks

Operating system: Platform independent

Programming language: C, JavaScript, PHP, Perl

Other requirements: The HPRD and PPID dataset are only available to Mount Sinai School of Medicine users due to licensing restrictions.

License: GNU GPL

Any restrictions to use by non-academics: License needed. Users should contact technology@mssm.edu

Competing interests

The author(s) declares that there are no competing interests.

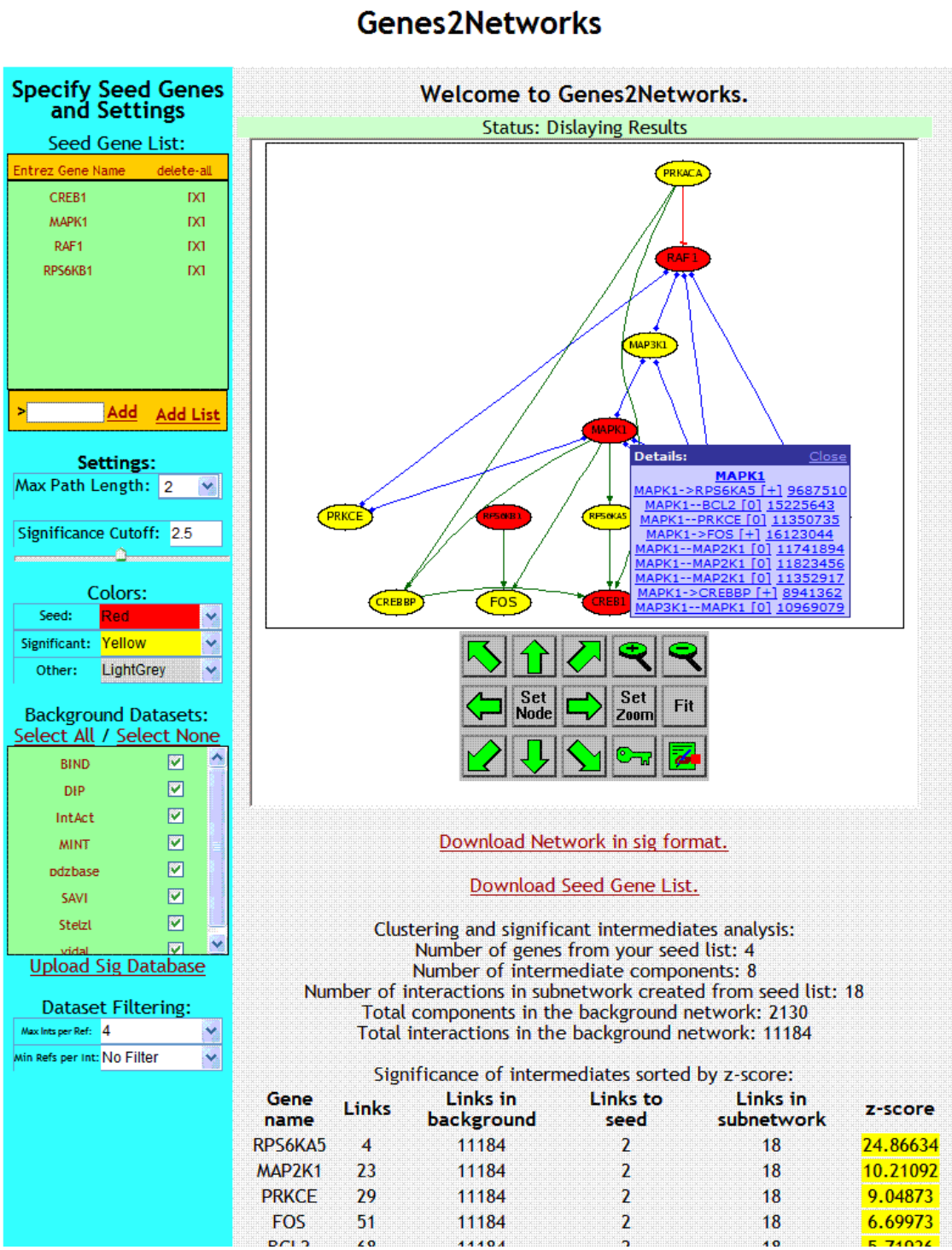

Figure 2
Genes2Networks web interface. The interface allows users to input a list of human Entrez Gene symbols, entered in a textbox or through a text file (top left). As genes are added, using the merged consolidated reference network made of different protein-protein interaction network databases, the program outputs a network map that visualize known interactions that "connect" the list of gene symbols from the seed list, and a statistical report that ranks intermediates based on their specificity to interact with the seed list.

Authors' contributions

AM designed and supervised the study, wrote the manuscript, implemented the significant intermediates statistical algorithm, and wrote the code for the initial Genes2Network prototype. JMP re-implemented and upgraded the code that merges and filters the datasets including implementing the hash function for fast loading of the datasets. JMP also rewrote the code to construct subnetworks from lists of gene names. SB designed and implemented the web interface including the AVIS visualization tool as well as upgraded the Genes2Network software to support listing of databases for interaction. SB also provided useful comments to the written manuscript.

Acknowledgements

This research was supported by NIH Grant No. GM-054508 and an advanced center grant from NYSTAR to Ravi Iyengar. We thank the anonymous reviewers for their useful comments.

References

1. Ma'ayan A, Blitzer RD, Iyengar R: **TOWARD PREDICTIVE MODELS OF MAMMALIAN CELLS.** *Annual Review of Biophysics and Biomolecular Structure* 2005, **34(1):**319-349.
2. Fields S, Song O-k: **A novel genetic system to detect proteinÂ-protein interactions.** 1989, **340(6230):**245-246.
3. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y: **A comprehensive two-hybrid analysis to explore the yeast protein interactome.** *PNAS* 2001, **98(8):**4569-4574.
4. Brown PO, Botstein D: **Exploring the new world of the genome with DNA microarrays.** *Nat Genet* 1999, **21:**33-37.
5. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM: **Expression profiling using cDNA microarrays.** *Nat Genet* 1999, **21:**10-14.
6. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR: **Direct analysis of protein complexes using mass spectrometry.** 1999, **17(7):**676-682.
7. Nikitin A, Egorov S, Daraselia N, Mazo I: **Pathway studio – the analysis and navigation of molecular networks.** *Bioinformatics* 2003, **19(16):**2155-2157.
8. Marcotte EM, Xenarios I, Eisenberg D: **Mining literature for protein-protein interactions.** *Bioinformatics* 2001, **17(4):**359-363.
9. Barabsi A-L, Albert R: **Emergence of Scaling in Random Networks.** *Science* 1999, **286(5439):**509-512.
10. Watts DJ, Strogatz SH: **Collective dynamics of 'small-world' networks.** *Nature* 1998, **393(6684):**440-442.
11. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U: **Network Motifs: Simple Building Blocks of Complex Networks.** *Science* 2002, **298(5594):**824-827.
12. Przulj N, Corneil DG, Jurisica I: **Efficient estimation of graphlet frequency distributions in protein-protein interaction networks.** *Bioinformatics* 2006, **22(8):**974-980.
13. Albert I, Albert R: **Conserved network motifs allow protein-protein interaction prediction.** *Bioinformatics* 2004, **20(18):**3346-3352.
14. Yu H, Paccanaro A, Trifonov V, Gerstein M: **Predicting interactions in protein networks by completing defective cliques.** *Bioinformatics* 2006, **22(7):**823-829.
15. Eungdamrong NJ, Iyengar R: **Computational approaches for modeling regulatory cellular networks.** *Trends in Cell Biology* 2004, **14(12):**661-669.
16. Ulitsky I, Shamir R: **Identification of functional modules using network topology and high-throughput data.** *BMC Systems Biology* 2007, **1(1):**8.
17. Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database.** *Nucl Acids Res* 2003, **31(1):**248-250.
18. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, *et al.*: **Human protein reference database – 2006 update.** *Nucl Acids Res* 2006, **34(suppl_1):**D411-414.
19. Hermjakob H, Montecchi-Palazzi L, Bader G, Wojcik J, Salwinski L, Ceol A, Moore S, Orchard S, Sarkans U, von Mering C, *et al.*: **The HUPO PSI's Molecular Interaction format[mdash]a community standard for the representation of protein interaction data.** 2004, **22(2):**177-183.
20. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D: **DIP: the Database of Interacting Proteins.** *Nucl Acids Res* 2000, **28(1):**289-291.
21. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Letters Protein Domains* 2002, **513(1):**135-140.
22. Rual J-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al.*: **Towards a proteome-scale map of the human protein-protein interaction network.** 2005, **437(7062):**1173-1178.
23. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S: **A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome.** *Cell* 2005, **122(6):**957-968.
24. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, *et al.*: **Formation of Regulatory Patterns During Signal Propagation in a Mammalian Cellular Network.** *Science* 2005, **309(5737):**1078-1083.
25. Beuming T, Skrabanek L, Niv MY, Mukherjee P, Weinstein H: **PDZBase: a protein-protein interaction database for PDZ-domains.** *Bioinformatics* 2005, **21(6):**827-828.
26. Grant SG: **Systems biology in neuroscience: bridging genes to cognition.** *Current Opinion in Neurobiology* 2003, **13(5):**577-582.
27. Brown KR, Jurisica I: **Online Predicted Human Interaction Database.** *Bioinformatics* 2005, **21(9):**2076-2082.
28. Han K, Park B, Kim H, Hong J, Park J: **HPID: The Human Protein Interaction Database.** *Bioinformatics* 2004, **20(15):**2466-2470.
29. Xia K, Dong D, Han J-D: **IntNetDB v1.0: an integrated protein-protein interaction network database generated by a probabilistic model.** *BMC Bioinformatics* 2006, **7(1):**508.
30. Huang T-W, Tien A-C, Huang W-S, Lee Y-CG, Peng C-L, Tseng H-H, Kao C-Y, Huang C-YF: **POINT: a database for the prediction of protein-protein interactions based on the orthologous interactome.** *Bioinformatics* 2004, **20(17):**3273-3276.
31. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucl Acids Res* 2003, **31(1):**365-370.
32. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucl Acids Res* 2006, **34(suppl_1):**D16-20.
33. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P: **Comparative assessment of large-scale data sets of protein-protein interactions.** 2002, **417(6887):**399-403.
34. Boser BE, Guyon IM, Vapnik VN: **A training algorithm for optimal margin classifiers.** *Pittsburgh* 1992.
35. Berger SI, Iyengar R, Ma'ayan A: **AVIS: AJAX Viewer of Interactive Signaling Networks.** *Bioinformatics* 2007. btm444.
36. Rosner B: **Fundamentals of biostatistics.** Pacific Grove, CA: Duxbury; 2000.
37. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Res* 2003, **13(11):**2498-2504.
38. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks.** *Bioinformatics* 2005, **21(16):**3448-3449.
39. Cerami EG, Bader GD, Gross B, Sander C: **cPath: open source software for collecting, storing, and querying biological pathways.** *BMC Bioinformatics* 2006, **7:**497.
40. Aragues R, Jaeggi D, Oliva B: **PIANA: protein interactions and network analysis.** *Bioinformatics* 2006, **22(8):**1015-1017.
41. Iragne F, Nikolski M, Mathieu B, Auber D, Sherman D: **ProViz: protein interaction visualization and exploration.** *Bioinformatics* 2005, **21(2):**272-274.
42. Dogrusoz U, Erson EZ, Giral E, Demir E, Babur O, Cetintas A, Colak R: **PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization.** *Bioinformatics* 2006, **22(3):**374-375.