Nucleic Acids Research Advance Access published May 19, 2009

Nucleic Acids Research, 2009, **1–6**
doi:10.1093/nar/gkp402

SNOW, a web-based tool for the statistical analysis of protein–protein interaction networks

Pablo Minguez[1], Stefan Götz[1,2], David Montaner[1], Fatima Al-Shahrour[1] and Joaquin Dopazo[1,2,3,]*

[1]Department of Bioinformatics and Genomics, Centro de Investigación Príncipe Felipe (CIPF), [2]CIBER de Enfermedades Raras (CIBERER) and [3]Functional Genomics Node (INB) at CIPF, Valencia, Spain

Received January 21, 2009; Revised April 22, 2009; Accepted May 2, 2009

ABSTRACT

Understanding the structure and the dynamics of the complex intercellular network of interactions that contributes to the structure and function of a living cell is one of the main challenges of today's biology. SNOW inputs a collection of protein (or gene) identifiers and, by using the interactome as scaffold, draws the connections among them, calculates several relevant network parameters and, as a novelty among the rest of tools of its class, it estimates their statistical significance. The parameters calculated for each node are: connectivity, betweenness and clustering coefficient. It also calculates the number of components, number of bicomponents and articulation points. An interactive network viewer is also available to explore the resulting network. SNOW is available at http://snow.bioinfo.cipf.es.

INTRODUCTION

It is widely accepted that most of the biological functionality of the cell arises from complex interactions between their molecular components that define operational interacting entities (1). Understanding the structure and the dynamics of the complex intercellular network of interactions that contribute to the structure and function of a living cell is one of the main challenges of today's biology (2) and constitutes the objective of systems biology (3). Alterations in the network of protein interactions are of special relevance in many diseases. For example, it has recently been demonstrated that tumorigenesis takes place in a specific and organized way that encompasses the precise down-regulation of groups of topologically-associated proteins (4). For the last years, there has been an enormous interest in the exploration of the interactome of model organisms such as *Saccharomyces cerevisiae* (5–7), *Drosophila melanogaster* (8,9), *Caenorhabditis elegans* (10) or human (11,12), just to cite a few examples. Thus, for several model organisms, a reasonable and operative description of the interactome is available. Several tools exist for network representation and calculation of network parameters, such as the popular Cytoscape (13) and its plugins (14) and other programs (15), but most of them are stand alone applications [see a recent review in (16)]. Also, several web tools for network analysis and visualization have been reported (see Supplementary Table 1) with functionalities that rank from mere network viewers to more sophisticated programs that perform different network parameter calculations (17–22). Tools of this type are enormously useful for the interpretation of genomic experiments. For example, if a number of genes has been found to be activated in a microarray experiment, their analysis in the context of the interactome can give clues on their possible role as a protein complex, as a signalling pathway, etc. However, the conclusions extracted from the simple visualization of a network of the calculation of some of its parameters are subjective without the proper statistical support. This feature, the proper statistical analysis of networks, still remains to be addressed by a web tool. Here we introduce Studying Networks in the Omics World (SNOW), a unique tool specifically designed to offer visualization and analysis of protein–protein interacting (PPI) networks, including their statistical analysis.

DESCRIPTION OF THE TOOL

Functionality of SNOW

Essentially, SNOW takes a list of proteins (or genes) and maps them onto an interactome of reference. This interactome can be the human interactome (in two versions, see databases subsection) or any other user-defined interactome. Once the list is mapped, SNOW calculates several relevant network parameters for the proteins in the contexts of the interactome and the minimum connected network (MCN) defined by the proteins. The corresponding tests are performed to assess the significance of the parameters calculated (see below).

*To whom correspondence should be addressed. Tel: +34 963289680; Fax: +34 963289701; Email: jdopazo@cipf.es

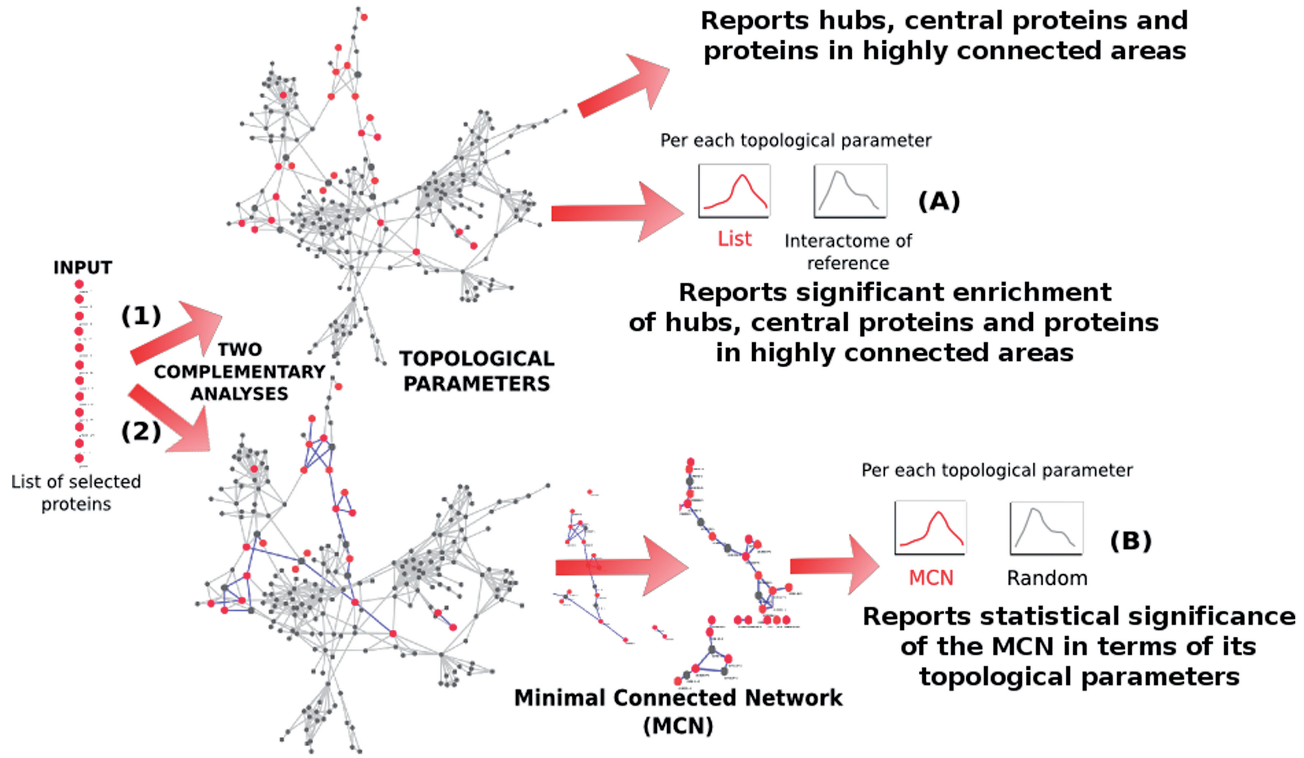

Reports hubs, central proteins and proteins in highly connected areas
Per each topological parameter
List
Interactome of reference
(A)
Reports significant enrichment of hubs, central proteins and proteins in highly connected areas
INPUT
(1)
TWO COMPLEMENTARY ANALYSES
(2)
TOPOLOGICAL PARAMETERS
List of selected proteins
Per each topological parameter
MCN
Random
(B)
Reports statistical significance of the MCN in terms of its topological parameters
Minimal Connected Network (MCN)

Figure 1. Schema of the analysis steps implemented in SNOW. Two complementary analyses are carried out: (1) the distribution of parameters (connection degree, betweenness centrality, clustering coefficient and number of components) of the list of proteins are tested against the rest of proteins of the interactome of reference within the context of connections of the complete interactome of reference and (2) The list of proteins is mapped onto the reference interactome and the MCN is found. The distributions of networks parameters in this MCN are tested against their random expectation (see 'Testing strategy for the network parameters' section in the Supplementary Data). See Supplementary Figure 2 for a more detailed schema.

Alternatively two lists of proteins can be compared by testing for significant deviations in their respective parameters' distributions. Figure 1 shows a schema of the analysis steps implemented in SNOW.

Input

SNOW inputs a collection of protein (or gene) identifiers in plain text. Most of the standard protein and gene identifiers are accepted given that the tool uses the database of identifiers of Babelomics (23).

Alternatively, a user-defined interactome can be provided to SNOW. It can be uploaded either as a list of protein–protein (or gene–gene) interactions tab-delimited in plain text or in the popular SIF format, used by Cytoscape (see http://www.cytoscape.org/cgi-bin/moin.cgi/Cytoscape_User_Manual/Network_Formats).

Calculation of network parameters and statistics

In particular, SNOW tests four topological parameters: (i) Node connections degree, which was computed as the number of edges (interaction events) for a node; (ii) Betweenness, which depends on the number of shortest pathways passing through a given node; (iii) Clustering coefficient, which measures the connectivity of the neighbourhood and (iv) Number of components of the network (see Additional methods for a detailed description on the calculation of these parameters). SNOW also finds the number of bicomponents and the articulation points and gives detailed descriptions of them. The parameters calculated account for different network properties. For example, signalling networks tend to have high connectivity and low clustering coefficient while metabolic networks have higher clustering coefficients.

The three first parameters (degree, betweenness and clustering coefficient) are calculated for all the nodes in the network. Thus, a distribution of values for any of these parameters can be derived for each particular network, which can give information on the network properties. Consequently, the comparison of two networks by contrasting how different they are in terms of their characteristic parameters is straightforward by means of a Kolmogorov–Smirnov test. The potential biological relevance of a particular network can therefore be obtained by comparing the network parameter distributions to the corresponding empirical distributions derived from ten thousand networks with the same number of nodes and a random protein composition (see 'Testing strategy for the network parameters' section in the Supplementary Data). By default the network is compared to the interactome of reference and to a network of random composition. However, the direct comparison of two networks is also possible.

In addition, the same analysis can be performed by adding one (or two or even three) intermediate nodes (proteins originally not included in the list that can actually link two or more proteins of the list), which is useful in proteomics analysis where often not all of the proteins can be identified in an experiment.

Output

The program outputs the average parameter values, their significance and boxplots representing the comparison of their actual distributions in the network studied to the complete interactome (top boxplots Figure 2) and to a random network of the same size (bottom boxplots, Figure 2). SNOW outputs the number of components, bicomponents and articulation points. The program also provides exhaustive information on these parameters as well as functional information on the proteins in the list. Moreover, information about the shortest paths and articulation points is also available. A low resolution viewer of the network is present in the main results web page (Figure 2). This viewer can be opened in a new window (see below). SNOW includes 12 examples in which different sub-networks along with their corresponding parameters can be visualized (see Tutorial link in the main menu of the SNOW program). A particularly interesting example can be found in the additional case study information. SNOW was used to analyze 168 genes induced by the over-expression of BRCA1, one of the most studied cancer-related genes. The application of conventional functional enrichment analysis (24) failed to detect any functionality significantly over-represented among the genes. Nevertheless, SNOW detected a significant concentration of highly connected ($P < 0.0001$) and central ($P = 0.0017$) proteins in the resulting MCN (see Additional Case Study for details).

Visualization

An interactive viewer for the connections defining the network studied can also be opened from the results page. All the components of the network are displayed in different layouts. The network orientation can be interactively changed. Gene names can be displayed and connecting proteins (intermediate nodes) can be included or excluded from the graph. Supplementary Figure 1 shows two possible layouts of the same network obtained from cell-cycle dependent genes regulated following exposure to serum in a variety of human fibroblast cell lines (data available from the third example included in the Tutorial and examples link of the main page of the SNOW program, see section above). The topmost view shows the menu with the options and the functional (gene ontology) and topological information on any protein which is displayed when placing the mouse over it.

Databases

Human PPI datasets were downloaded from five main public databases: HPRD (25), IntAct (26), BIND (27), DIP (28) and MINT (29). We used this collection of PPI data to generate two scaffold interactomes: a non-filtered scaffold interactome, which includes all the available PPIs; and a filtered, more confident scaffold interactome. The six top categories of experimental methods described in the Molecular Interaction (MI) Ontology (30) plus the categories *in vivo* and *in vitro* from HPRD can be used as confidence indicators. Thus, only PPIs verified by at least two of these categories were considered in the filtered scaffold interactome. Protein identifiers can be directly mapped to Ensembl transcript identifiers that can easily be linked to many other gene or protein identifiers. Since in many genomic experiments only data on genes (but not on particular transcripts) is available, we build up another two high and low confidence interactomes by mapping transcripts onto genes. Conceptually, each gene can potentially interact with all the interaction partners of all of its transcripts. Transcript- and gene-based interactomes are similar but not identical.

Implementation details

The SNOW web interface was implemented using Perl and JavaScript languages. The queries to the server are implemented as web services written in Perl and Pyhton. Boost Graph libraries (http://www.boost.org/libs/graph/doc/index.html) were used to perform graph parameters calculation. SNOW accesses these libraries through BGL-Python bindings (http://www.osl.iu.edu/~dgregor/bgl-python/). SNOW uses R programming language to perform the statistics and mysql to store interactome network parameters. The visualization applet was implemented in Java using TouchGraph libraries (http://touchgraph.sourceforge.net).

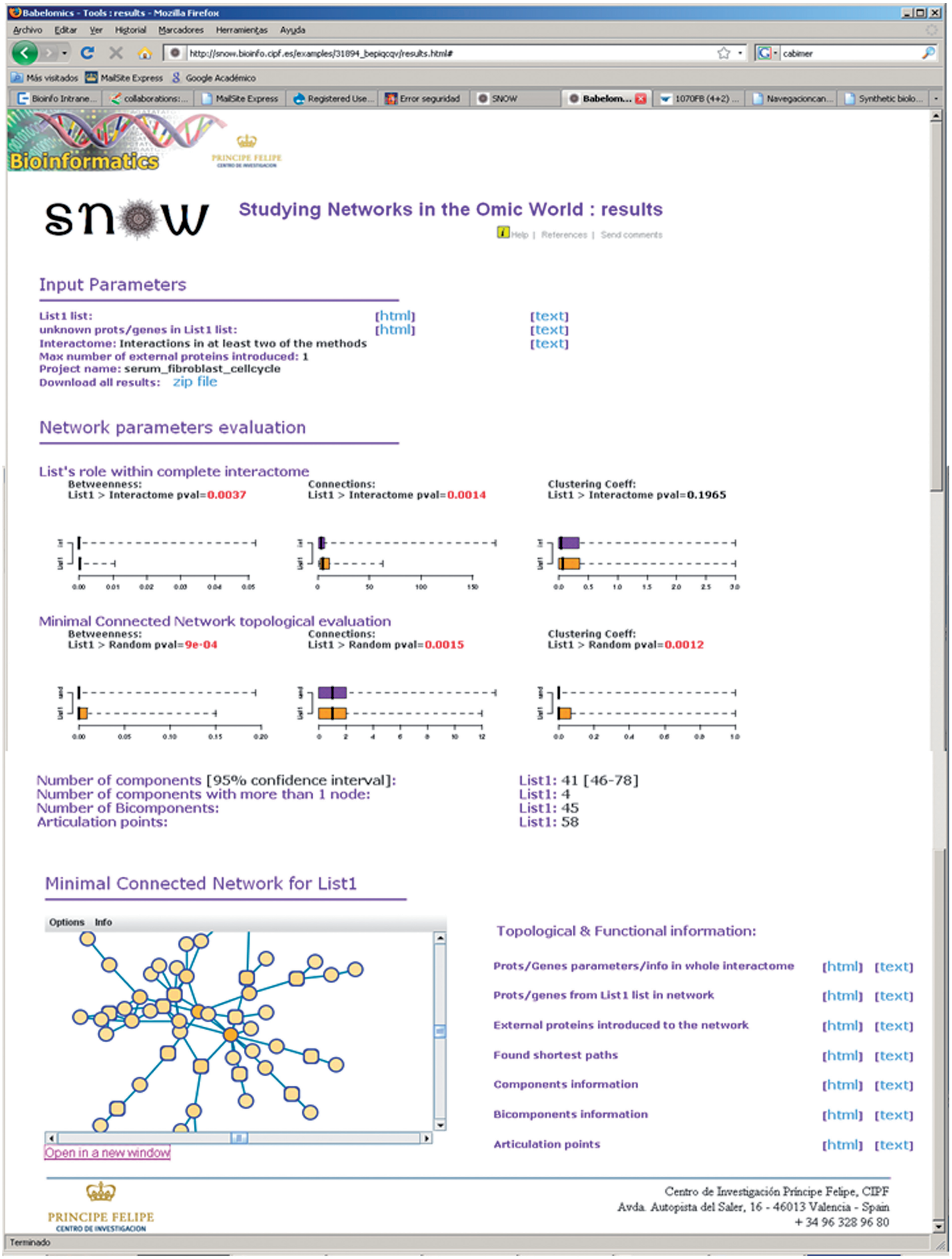
Other tools

There are several popular stand-alone tools for the visualization and analysis of networks, such as the popular Cytoscape (13,14) and other analogous tools (15,16). Also, in the last years, several web-based applications for this purpose have been reported in the literature (17–19,31). Some tools provide annotations (32), offer functional enrichment analysis (32,33) or are more focused to pathways (19,34,35) or provides some facilities for network comparisons (20). Several databases offer their own network viewers as is the case of STRING (36) or IntAct (26). Other tools allow for the calculation of some network parameters (20–22,37), although in some cases at the expense of losing interactivity in the visualization of the network (22). Supplementary Table 1 contains a list of web tools for network management with some of their most significant features.

DISCUSSION

SNOW is unique among its genre in the sense that it not only allows visualizing data in an interactive dynamic format while calculating relevant network parameters but it also finds their statistical significance. It can also be used to compare two networks in terms of their corresponding distributions of parameters. To our knowledge there are no other web-based tools with such features.

SNOW constitutes a step forward in our efforts to provide the scientific community with web tools, such as the Babelomics suite (38), for the functional profiling of genomic experiments. SNOW broadens the conventional



Babelomics - Tools : results - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://snow.bioinfo.cipf.es/examples/31894_bepiqcqv/results.html#
cabimer
Más visitados MailSite Express Google Académico
Bioinfo Intrane... collaborations:... MailSite Express Registered Use... Error seguridad SNOW Babelom... 1070FB (4+2) ... Navegacioncan... Synthetic biolo...
Bioinformatics
PRINCIPE FELIPE CENTRO DE INVESTIGACION
snow

Studying Networks in the Omic World : results

Help | References | Send comments

Input Parameters

List1 list: [html] [text]
unknown prots/genes in List1 list: [html] [text]
Interactome: Interactions in at least two of the methods [text]
Max number of external proteins introduced: 1
Project name: serum_fibroblast_cellcycle
Download all results: zip file

Network parameters evaluation

List's role within complete interactome

Betweenness:
List1 > Interactome pval=0.0037

Connections:
List1 > Interactome pval=0.0014

Clustering Coeff:
List1 > Interactome pval=0.1965

Minimal Connected Network topological evaluation

Betweenness:
List1 > Random pval=9e-04

Connections:
List1 > Random pval=0.0015

Clustering Coeff:
List1 > Random pval=0.0012

Number of components [95% confidence interval]: List1: 41 [46-78]
Number of components with more than 1 node: List1: 4
Number of Bicomponents: List1: 45
Articulation points: List1: 58

Minimal Connected Network for List1

Topological & Functional information:

Prots/Genes parameters/info in whole interactome [html] [text]
Prots/genes from List1 list in network [html] [text]
External proteins introduced to the network [html] [text]
Found shortest paths [html] [text]
Components information [html] [text]
Bicomponents information [html] [text]
Articulation points [html] [text]

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Centro de Investigación Príncipe Felipe, CIPF
Avda. Autopista del Saler, 16 - 46013 Valencia - Spain
+ 34 96 328 96 80

Figure 2. Output of SNOW—the web page contains a summary of the input data and the network parameters estimated along with the statistical significance for the connectivity, betweenness and clustering coefficient. Top boxplots represent the comparison of the network to the complete interactome and bottom boxplots account for the comparison of the network to a random network of the same size (see text). Component, bicomponents and articulation points are also provided. In addition, detailed information of genes in the list, genes included in the analysis, shortest pathways is also accessible from the web page. Finally an interactive network viewer can be launched from the page.

functional profiling based on gene modules (e.g. gene ontology or KEGG pathway gene sets) to gene sets based on PPIs. SNOW has been operative since last summer and it has been used by projects associated to the Spanish National Bioinformatics Institute (http://www.inab.org), the Spanish Network of Cancer (RTICC; http://www.rticcc.org) and the Network of Centres for Research in Rare Diseases (CIBERER, http://www.ciberer.es). SNOW is running in a high-end cluster with 10 dedicated Intel XEON Quad-Core CPUs at 2.0 GHz (summing up a total of 40 cores) with a large amount of RAM (total 60 GB). The program is available at http://snow.bioinfo.cipf.es.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Spanish Ministry of Science and Innovation (BIO2008-04212 to JD). The CIBER de Enfermedades Raras is an initiative of the ISCIII. The RTICC is an initiative of the ISCIII. The National Institute of Bioinformatics (www.inab.org) is a platform of Genoma España. Funding for open access charge: Spanish Ministry of Science and Innovation grant BIO2008-04212.

Conflict of interest statement. None declared.

REFERENCES

1. Hartwell,L.H., Hopfield,J.J., Leibler,S. and Murray,A.W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.
2. Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.
3. Kitano,H. (2002) Computational systems biology. *Nature*, **420**, 206–210.
4. Hernandez,P., Huerta-Cepas,J., Montaner,D., Al-Shahrour,F., Valls,J., Gomez,L., Capella,G., Dopazo,J. and Pujana,M.A. (2007) Evidence for systems-level molecular mechanisms of tumorigenesis. *BMC Genomics*, **8**, 185.
5. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, **403**, 623–627.
6. Ito,T., Chiba,T., Ozawa,R., Yoshida,M., Hattori,M. and Sakaki,Y. (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
7. Reguly,T., Breitkreutz,A., Boucher,L., Breitkreutz,B.J., Hon,G.C., Myers,C.L., Parsons,A., Friesen,H., Oughtred,R., Tong,A. *et al.* (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.*, **5**, 11.
8. Formstecher,E., Aresta,S., Collura,V., Hamburger,A., Meil,A., Trehin,A., Reverdy,C., Betin,V., Maire,S., Brun,C. *et al.* (2005) Protein interaction mapping: a Drosophila case study. *Genome Res.*, **15**, 376–384.
9. Giot,L., Bader,J.S., Brouwer,C., Chaudhuri,A., Kuang,B., Li,Y., Hao,Y.L., Ooi,C.E., Godwin,B., Vitols,E. *et al.* (2003) A protein interaction map of Drosophila melanogaster. *Science*, **302**, 1727–1736.
10. Li,S., Armstrong,C.M., Bertin,N., Ge,H., Milstein,S., Boxem,M., Vidalain,P.O., Han,J.D., Chesneau,A., Hao,T. *et al.* (2004) A map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
11. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. *et al.* (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
12. Rual,J.F., Venkatesan,K., Hao,T., Hirozane-Kishikawa,T., Dricot,A., Li,N., Berriz,G.F., Gibbons,F.D., Dreze,M., Ayivi-Guedehoussou,N. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.
13. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
14. Assenov,Y., Ramirez,F., Schelhorn,S.E., Lengauer,T. and Albrecht,M. (2008) Computing topological parameters of biological networks. *Bioinformatics*, **24**, 282–284.
15. Junker,B.H., Koschutzki,D. and Schreiber,F. (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics*, **7**, 219.
16. Suderman,M. and Hallett,M. (2007) Tools for visually exploring biological networks. *Bioinformatics*, **23**, 2651–2659.
17. Hu,Z., Mellor,J., Wu,J. and DeLisi,C. (2004) VisANT: an online visualization and analysis tool for biological interaction data. *BMC Bioinformatics*, **5**, 17.
18. Breitkreutz,B.J., Stark,C. and Tyers,M. (2002) Osprey: a network visualization system. *Genome Biol.*, **3**, PREPRINT0012.
19. Dogrusoz,U., Erson,E.Z., Giral,E., Demir,E., Babur,O., Cetintas,A. and Colak,R. (2006) PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics*, **22**, 374–375.
20. Brohee,S., Faust,K., Lima-Mendez,G., Sand,O., Janky,R., Vanderstocken,G., Deville,Y. and van Helden,J. (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.*, **36**, W444–W451.
21. Wu,J., Vallenius,T., Ovaska,K., Westermarck,J., Makela,T.P. and Hautaniemi,S. (2009) Integrated network analysis platform for protein–protein interactions. *Nat. Methods*, **6**, 75–77.
22. Yip,K.Y., Yu,H., Kim,P.M., Schultz,M. and Gerstein,M. (2006) The tYNA platform for comparative interactomics: a web tool for managing, comparing and mining multiple networks. *Bioinformatics*, **22**, 2968–2970.
23. Al-Shahrour,F., Minguez,P., Tarraga,J., Montaner,D., Alloza,E., Vaquerizas,J.M., Conde,L., Blaschke,C., Vera,J. and Dopazo,J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res.*, **34**, W472–W476.
24. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
25. Peri,S., Navarro,J.D., Amanchy,R., Kristiansen,T.Z., Jonnalagadda,C.K., Surendranath,V., Niranjan,V., Muthusamy,B., Gandhi,T.K., Gronborg,M. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.*, **13**, 2363–2371.
26. Kerrien,S., Alam-Faruque,Y., Aranda,B., Bancarz,I., Bridge,A., Derow,C., Dimmer,E., Feuermann,M., Friedrichsen,A., Huntley,R. *et al.* (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res.*, **35**, D561–D565.
27. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
28. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
29. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
30. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al.* (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.

31. Iragne,F., Nikolski,M., Mathieu,B., Auber,D. and Sherman,D. (2005) ProViz: protein interaction visualization and exploration. *Bioinformatics*, **21**, 272–274.
32. Reimand,J., Tooming,L., Peterson,H., Adler,P. and Vilo,J. (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, **36**, W452–W459.
33. Hu,Z., Mellor,J., Wu,J., Yamada,T., Holloway,D. and Delisi,C. (2005) VisANT: data-integrating visual framework for biological networks and modules. *Nucleic Acids Res.*, **33**, W352–W357.
34. Hu,Z., Ng,D.M., Yamada,T., Chen,C., Kawashima,S., Mellor,J., Linghu,B., Kanehisa,M., Stuart,J.M. and DeLisi,C. (2007) VisANT 3.0: new modules for pathway visualization, editing, prediction and construction. *Nucleic Acids Res.*, **35**, W625–W632.
35. Hu,Z., Snitkin,E.S. and DeLisi,C. (2008) VisANT: an integrative framework for networks in systems biology. *Brief Bioinform.*, **9**, 317–325.
36. von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.*, **31**, 258–261.
37. Prieto,C. and De Las Rivas,J. (2006) APID: Agile protein interaction data analyzer. *Nucleic Acids Res.*, **34**, W298–W302.
38. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.*, **33**, W460–W464.