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

Pablo Minguez¹, Stefan Götz¹,², David Montaner¹, Fatima Al-Shahrour¹ and Joaquin Dopazo¹,²,³,*

¹Department of Bioinformatics and Genomics, Centro de Investigación Principce Felipe (CIPF), ²CIBER de Enfermedades Raras (CIBERER) and ³Functional 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: jpdopazo@cipf.es

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
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

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.
six top categories of experimental methods described and a filtered, more confident scaffold interactome. The scaffold interactome, which includes all the available PPIs, data to generate two scaffold interactomes: a non-filtered DIP (28) and MINT (29). We used this collection of PPI Human PPI datasets were downloaded from five main Databases (gene ontology) and topological information on any pro-

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 inter-acitively 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 pro-vide the scientific community with web tools, such as the Babelomics suite (38), for the functional profiling of genomic experiments. SNOW broadens the conventional
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.rticc.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 Raras is an initiative of the ISCIII. The RTICC is an initiative of the ISCIII. The RTICC is an initiative of the ISCIII. The National Institute of Raras is an initiative of the ISCIII. The RTICC is an initiative of the ISCIII.

SUPPLEMENTARY DATA

at http://snow.bioinfo.cipf.es.

REFERENCES


