APPLICATION NOTE

Genomics & Informatics Vol. 5(3) 133-136, September 2007

GSnet: An Integrated Tool for Gene Set Analysis and Visualization

Yoon Jeong Choi[1], Hyun Goo Woo[2] and Ungsik Yu[3]*

[1]Nano Materials Simulation and Fabrication Laboratory, Department of Materials Science and Engineering, KAIST, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea, [2]Laboratory of Experimental Carcinogenesis, National Cancer Institute, NIH, USA [3]Korean BioInformation Center, Korea Research Institute of Bioscience and Biotechnology, Daejeon 305-806, Korea

Abstract

The Gene Set network viewer (GSnet) visualizes the functional enrichment of a given gene set with a protein interaction network and is implemented as a plug-in for the Cytoscape platform. The functional enrichment of a given gene set is calculated using a hypergeometric test based on the Gene Ontology annotation. The protein interaction network is estimated using public data. Set operations allow a complex protein interaction network to be decomposed into a functionally-enriched module of interest. GSnet provides a new framework for gene set analysis by integrating a priori knowledge of a biological network with functional enrichment analysis.

Availability: GSnet is freely available at http://www.kobic.re.kr/gsnet(*contact:* yjchoi@kribb.re.kr)
Supplementary Information: http://www.kobic.re.kr/gsnet

Keywords: microarray analysis, gene set analysis, protein-protein interaction network, network visualization

High-throughput data from microarray experiments provide information about the expression of genes, with gene sets being defined based on the presence of coexpression in these experiments. Several methods have been proposed for the functional analysis of gene sets, including the chi-square test, the hypergeometric test, and the Gene Set Enrichment Analysis (GSEA) (Boyle *et al.*, 2004; Curtis *et al.*, 2005; Subramanian *et al.*, 2005). These methods find which given gene sets are significantly associated with a priori knowledge data, such as the Gene Ontology (GO) classification (Ashburner *et al.*, 2000), measure the statistical enrichment of genes in a gene set by calculating the *p*-value using a statistical method, and produce more robust and interpretable information from microarray analyses (Bammler *et al.*, 2005). However, the current state-of-the-art ontological analysis is subject to conceptual limitations (Khatri and Drăghici, 2005), and the key regulators of a gene set can not be elucidated from expression profile alone. The proteinprotein interaction network is a type of scale-free architecture that follows a power-law distribution (Jeong *et al.*, 2001). These scale-free networks are characterized by the existence of hubs with multiple interacting partners, with the genes in the hubs playing a functional key role in maintaining cell functions (Jeong *et al.*, 2001). It is important to understand the relationships and topology of gene expression in the complex hierarchical web of protein interactions (Jeong *et al.*, 2001; Nikolsky *et al.*, 2005; Rhodes and Chinnaiyan, 2005). Moreover, the presence of coexpression in multiple data sets is indicative of functional relatedness (Lee *et al.*, 2004), which implies that hub genes in functionally-clustered modules are functionally significant and worthy of investigation. However, to the best of our knowledge, there is no available tool that can integrate the functional analyses with analyses of proteinprotein interaction networks. The Gene Set network viewer (GSnet) has been designed to provide an analytical framework for integrating gene set enrichment analysis with the analysis of proteinprotein interaction networks. GSnet characterizes and visualizes functional enrichment with the properties of an interaction network. It analyzes the GO term enrichment in a given gene set, and then constructs a proteinprotein interaction network of the gene set. Furthermore, set operations can be used in GSnet to select interesting genes included in a particular enriched GO term on the network graph in user-defined colors. Our approach facilitates the selection of putative target genes that have functional importance in a given gene set.

GSnet has been developed as a plug-in for Cytoscape, which is an open-source bioinformatics software platform for visualizing molecular interaction networks (available at http://www.cytoscape.org) (Shannon *et al.*, 2003). GSnet is freely available at http://www.kobic.re.kr/gsnet. GSnet takes a list or a file of Entrez Gene IDs or symbols, and also supports the two-input-set mode for the analysis of upregulated and downregulated gene sets. GSnet currently supports two protein-protein interaction databases for four species: human, mouse, rat, and yeast. These databases

*Corresponding author: E-mail ungsik@kribb.re.kr
Tel +82-42-879-8520, Fax +82-42-879-8519
Accepted 28 August 2007

are generated from the Biomolecular Interaction Network Database (Alfarano *et al.*, 2005) and the Agilent literature search engine (Agilent Technologies, 2005 http://www.agilent.com/labs/research/mtl/projects/sysbio/sysinformatics/litsearch.html) using PubMed (Wheeler *et al.*, 2005) as a literature reference. For the GO functional enrichment analysis of the gene set, we adopted a hypergeometric test (Tavazoie et al., 1999), which is one of the most widely-used methods. The hypergeometric test calculates the cumulative probability (*p*-value) of a given specific number of genes from a single gene set in a whole gene list with the hypergeometric distribution. For the correction of the significance of multiple testing, the false discovery rate (FDR) of the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) is also calculated. GSnet provides options for filtering out some enriched functions, including the cut-off *p*-value, the number of genes within a GO term, and the levels of the GO-term hierarchy. The filtered GO terms are then categorized into molecular function, biological process, and cellular component. The categorized results are displayed in order of ascending *p*-value as a tabbed-table in a results window, with two such windows being shown in the two-input-set mode. The proteinprotein interaction network of given genes is constructed simultaneously using the protein interaction database and visualized in Cytoscape. To customize the visualization of sub-networks or modules, users can highlight them by selecting the GO function of interest with a user-defined color. A set with more than onefunction of interest can be defined by set operations, which allow users to decompose the network and refine functional modules of interest, as shown in Fig. 1. Set operations can be carried out among genes of GO terms in the same or different GO categories. The set of nodes is selected by checking the "Set A" or "Set B" checkbox beside a GO term, and the type of set operation is specified by marking the appropriate checkbox in the view-option. The selected nodes (i.e., proteins) can be further manipulated to create a child network or to merge them into another network using the layout functionality supported by Cytoscape. To identify the hub genes in a given network, GSnet also provides the number of interacting partners of each gene as a new attribute—the number of edges. Attribute information is shown on the node attribute browser of Cytoscape. We briefly summarize some specific aspects of GSnet in Table 1.

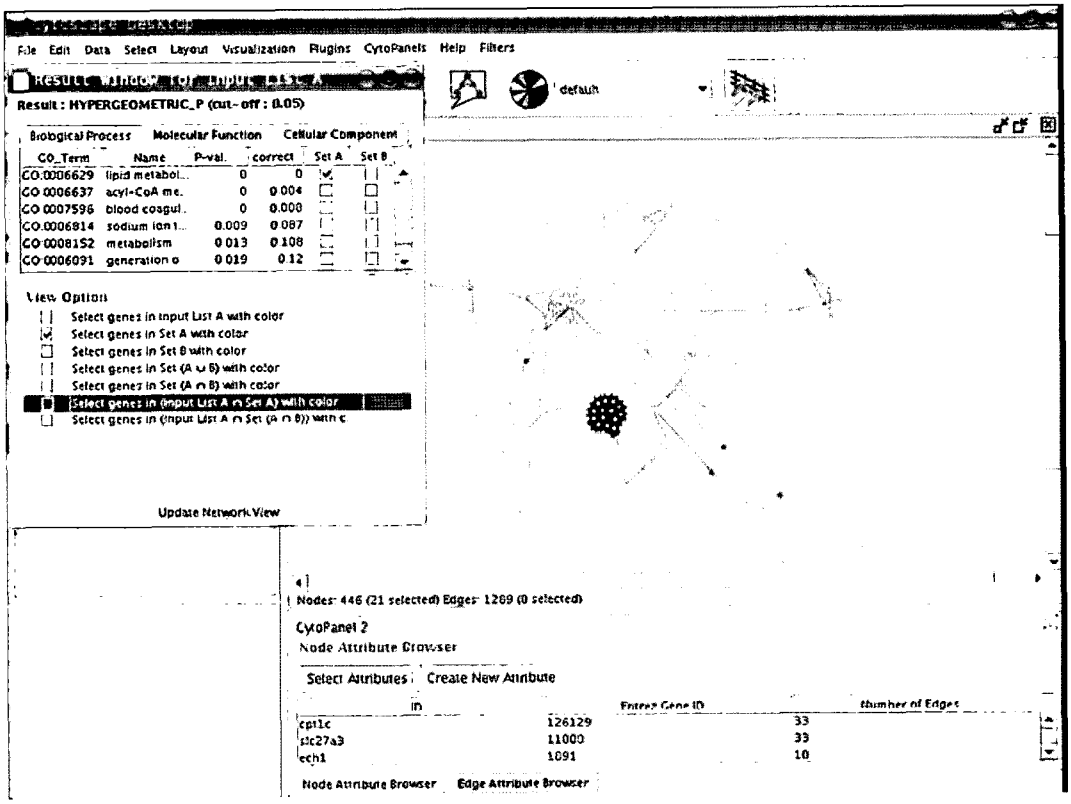

Fig. 1. An example GSnet results window and network view in Cytoscape. Nodes in the GO:0006629 term are shown in green and nodes both in GO:0006629 and the input list are shown in red.

Table 1. Summary of GSnet tool

Function		Short description
Input		Gene set list (one or two)
Option		
	Species (default: human)	Human, Mouse, Rat, and Yeast
	p-value	Statistical model for assessment of *p*-value
	Network DB (default: Agilent)	Interaction database (BIND, Agilent)
	P-value cut-off (default: 0.05)	Only GO terms with p-values less than cut-off are shown in the results window
	Matched genes cut-off	Only GO terms with the number of matched genes more than cut-off are shown.
	Level min~max cut-off	Only GO terms at the level between the specified minimum level and maximum level are shown.
Statistical method (*p*-value)		Hypergeometric test
Output		GO table window and protein-protein network window. These are interlinked.
View option		Selecting specific genes by performing set operation. Selected genes are highlighted in user-specified color.

GSnet is an integrated viewer for analyzing biological networks and the functional enrichment of gene sets. The integration of data derived from heterogeneous experimental sources may help to reduce the influence of noise inherent in experimental data. In addition, putative targets for further studies can be estimated by targeting the hub genes. GSnet can incorporate additional statistical models and databases as needed. We consider GSnet to be a valuable tool that provides a new analytical framework for expanding the interpretation of gene sets and biological networks.

Acknowledgments

The authors thank all members of the Cytoscape development team. This project was supported by the Korean Ministry of Science & Technology (MoST) under grant number M1052900001005N290001000.

References

Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M.J., Dumontier, M.R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic. A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montojo, J., Moore, S., Muskat, B., Ng, I., Paraiso, J.P., Parker, B., Pintilie, G., Pirone, R., Salama, J.J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B.F., and Hogue, C.W. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 33, D418-D424

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25, 25-29.

Bammler, T., Beyer, R.P., Bhattacharya, S., Boorman, G.A., Boyles, A., Bradford, B.U., Bumgarner, R.E., Bushel, P.R., Chaturvedi, K., Choi, D., Cunningham, M.L., Deng, S., Dressman, H.K., Fannin, R.D., Farin, F.M., Freedman, J.H., Fry, R.C., Harper, A., Humble, M.C., Hurban, P., Kavanagh, T.J., Kaufmann, W.K., Kerr, K.F., Jing, L., Lapidus, J.A., Lasarev, M.R., Li. J,, Li, Y.J., Lobenhofer, E.K., Lu, X., Malek, R.L., Milton, S., Nagalla, S.R., O'malley, J.P., Palmer, V.S., Pattee, P., Paules, R.S., Perou, C.M., Phillips, K., Qin, L.X., Qiu, Y., Quigley, S.D., Rodland, M., Rusyn, I., Samson, L.D., Schwartz, D.A., Shi, Y., Shin, J.L., Sieber, S.O., Slifer, S., Speer, M.C., Spencer, P.S., Sproles, D.I., Swenberg, J.A., Suk, W.A., Sullivan, R.C., Tian, R., Tennant, R.W., Todd, S.A., Tucker, C.J., Van Houten,B., Weis, B.K., Xuan, S., Zarbl, H. Members of the Toxicogenomics Research Consortium. (2005). Standardizing global gene expression analysis between laboratories and across platforms. *Nat. Methods* 2, 351-356.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stati. Soc.* B. 57, 280-300.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinderopen source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes *Bioinformatics* 20, 3710-3715.

Curtis, R.K., Orešič, M., and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data. *Trends Biotechnol.* 23, 429-435.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. (2001). Lethality and centrality in protein networks. *Nature* 411, 41-42.

Khatri, P. andDrăghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 21, 3587-3595.

Lee, H.K., Hsu, A. K., Sajdak, J., Qin, J., and Pavlidis, P. (2004). Co-expression analysis of human genes across many microarray data sets. *Genome Res.* 14, 1085-94.

Nikolsky, Y., Nikolskaya, T., and Bugrim, A. (2005). Biological networks and analysis of experimental data in drug discovery. *Drug Discov. Today* 10, 653-662.

Rhodes, D.R. and Chinnaiyan, A.M. (2005).Integrative analysis of the cancer transcriptome. *Nat. Genet.* 37,S31-S37.

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 Research* 13, 2498-2504.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci.* 102, 15545-15550.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., andChurch., G.M. (1999). Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281-285.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L., and Yaschenko, E.. (2005). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 33, D39-D45.