Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Web Server**

GS2PATH: A web-based integrated analysis tool for finding functional relationships using gene ontology and biochemical pathway data

Jin Ok Yang[1], Charny Park[1], Byungwook Lee[1], Sangsoo Kim[2], Jong Bhak[1,*], Hyun Goo Woo[3]

[1]Korean BioInformation Center, KRIBB, Daejeon 305-806, Korea; [2]Department of Bioinformatics, Soongsil University, Seoul, Korea; [3]Laboratory of Experimental Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA; Jong Bhak* - E-mail: jong@kribb.re.kr; * Corresponding author

received November 30, 2007; revised December 07, 2007; accepted December 30, 2007; published online December 30, 2007

Abstract:
GS2PATH is a Web-based pipeline tool to permit functional enrichment of a given gene set from prior knowledge databases, including gene ontology (GO) database and biological pathway databases. The tool also provides an estimation of gene set enrichment, in GO terms, from the databases of the KEGG and BioCarta pathways, which may allow users to compute and compare functional over-representations. This is especially useful in the perspective of biological pathways such as metabolic, signal transduction, genetic information processing, environmental information processing, cellular process, disease, and drug development. It provides relevant images of biochemical pathways with highlighting of the gene set by customized colors, which can directly assist in the visualization of functional alteration.

Availability: The GS2PATH system is freely available at http://array.kobic.re.kr:8080/arrayport/gs2path/.

Keywords: gene set enrichment test; gene ontology; pathways; integrated search

Background:
Much research effort has been devoted to the analysis functional relationships among sets of genes. Gene Ontology (GO) data, and information on genes belonging to known metabolic pathways **[1-3]**, have been most valuable. The GO database and associated tools have been used to find functional relationships among genes, because genes annotated by the same GO term are probably involved in the same biological process, have similar molecular functions, and are located in the same cellular compartments. **[4]** Pathway data can be exploited to infer functional relationships among genes because genes with similar functional relationships are likely to belong to the same pathway **[5]**.

An enrichment test is useful for investigating which specific GO term is associated with a given gene set. **[6, 7]** Gene set enrichment derives its power by focusing on gene sets, that is, groups of genes that share common biological functions, chromosomal locations, or regulation. **[8, 9]** Gene set enrichment evaluates microarray data at the level of gene sets defined on prior biological knowledge such as co-involvement in biochemical pathways or co-expression in previous experiments, and reveals many common biological pathways. **[8, 9]** As for pathways, the representative databases and search tools are KEGG **[2]** and BioCarta **[3]**. The KEGG database integrates current knowledge on molecular interaction networks in biological processes and the BioCarta database provides dynamic graphical models of molecular and cellular pathways.

Currently, there is no integrated tool for providing comprehensive results using both GO terms and pathways. Thus, users have to manually check which gene sets are associated with particular GO terms or defined pathways. This can be time-consuming and error-prone. We therefore developed an integrated search tool, termed GS2PAPTH, for analyzing functional relationships in gene sets and providing comprehensive results. GS2PATH provides 1) a hyper-geometric test for gene set enrichment with GO and pathways databases; 2) a dual search for upregulated- and down-regulated gene sets; 3) an analysis of gene sets using biological pathways based on GO terms; 4) various filtering options for GO terms including the number of descendant nodes, the statistical weight of GO terms, and statistical values mapping gene sets with particular GO terms to biological pathways; and, 5) user-specified coloring for genes on a pathway.

Input:
The GS2PATH consists of one internal database (a mapping database) in MySQL DBMS and four components: a Query Processor, a GO Accessor, a BioCarta Accessor, and a KEGG Accessor. The web interface and the web-based usage of all functions of GS2PATH were developed in JAVA and JSP. The main web page provides end-users with query input along with options to select specific organism and GO terms or pathways for searching paired gene sets. When users acquire a gene set, they can search gene identity lists

against GO categories and pathways in the main Web interface of GS2PATH (Figure 1). In the first step, users select the GO database or a pathway database. Next, users select the organism of interest such as human, mouse, rat, or yeast. Third, users call up single or multiple gene lists. Finally, users click the search button. The tabular results of the gene set enrichment are then displayed and linked to KEGG and BioCarta pathway information. This system calculates statistical values for GO terms and supports several GO term filtering options related to GO terms, allowing users to connect the mapping gene set in each GO term to the biological databases of KEGG and BioCarta.

Output:

GS2PATH provides integrated search facilities over the GO database, the KEGG and the BioCarta databases. We offer an alternative method for data enrichment which incorporates these prior knowledge databases. GS2PATH conducts a hyper-geometric test for gene set enrichment, and retrieves a set of GO terms relevant for the comparison of a dual gene set (e.g. up-regulated and down-regulated, or normal and abnormal). In addition, GS2PATH also displays statistical values mapping a gene set, in each GO term, to the KEGG and BioCarta databases. Users may also specify various filtering options for GO terms, including the number of descendant nodes, and may narrow the choice of GO terms to refine search results. Additionally, users can obtain dynamic graphical images of interacting genes from both BioCarta and KEGG. To assist in interpretation, users can choose colors to highlight genes of interest in input gene sets. If a user clicks a link to pathway results, GS2PATH retrieves images of corresponding pathways, highlighting genes in the input gene sets using user-specified coloring. It shows the results of a KEGG pathway interrogation and a BioCarta download with genes colored, respectively.

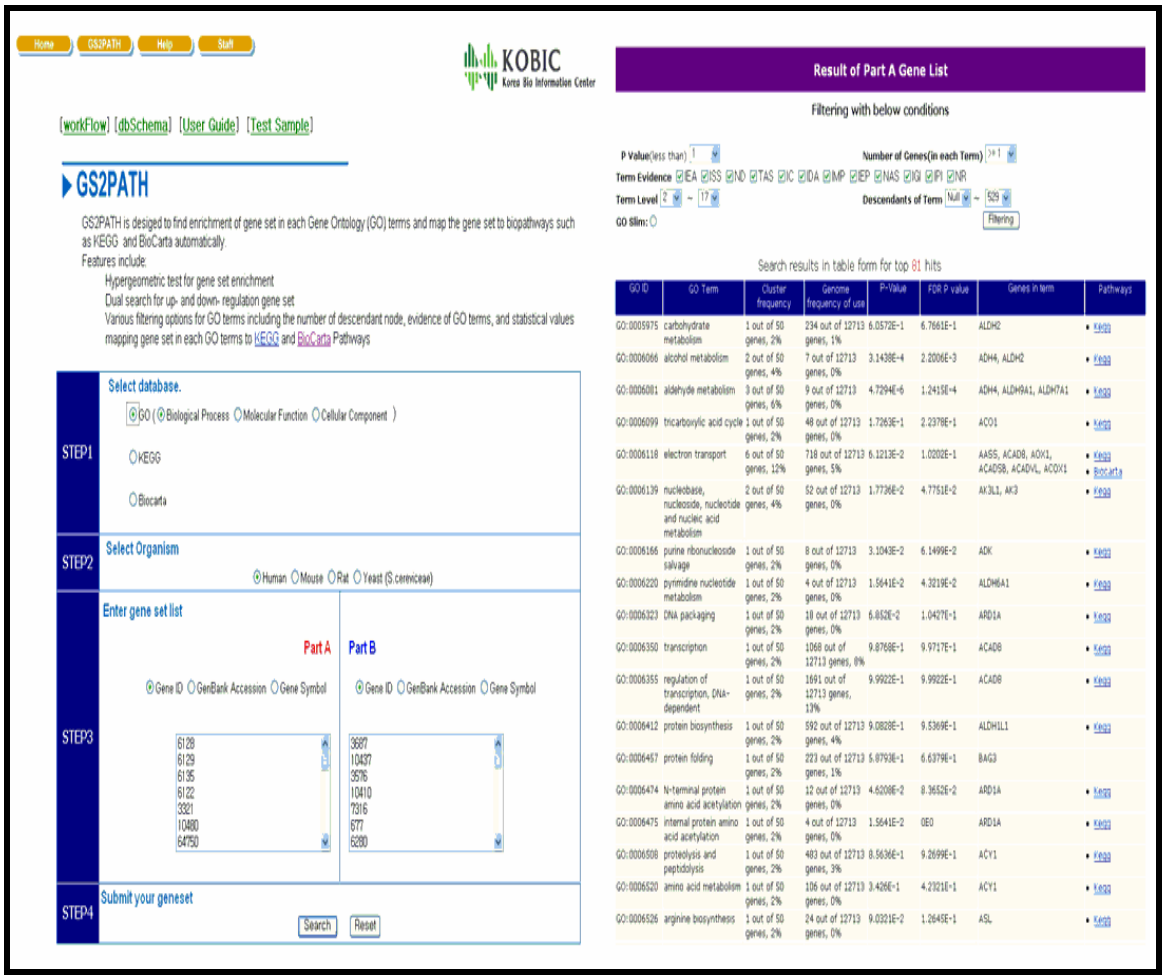

Figure 1: (Left) Web interface of GS2PATH. Users can select the databases for GO or KEGG including maps for biological processes. (Right) GS2PATH returns the results for query list. The results contain GO term with GO ID, Term, correlated p-value, cluster frequency, genes annotated to term, and pathways associated with genes in terms. There are various filtering options for GO terms including the number of descendant nodes, evidence of GO terms, statistical values mapping genes in each GO term to biological pathways. Users can connect GO results to KEGG and BioCarta pathway databases, which are commonly used biological pathways databases.

Caveat and future development:

GS2PATH is a Web-based integrated tool that performs gene set enrichment using GO terms and pathways associated with the GO terms. GS2PATH is better at

capturing integrated information, and defined biological functions, than the two-pronged approaches (separate GO and pathway analyses) currently in use. GS2PATH therefore helps in the understanding of biological differences among gene sets, and permits disease association studies by linking of GO terms to biological pathways associated with sub-gene sets. This system is freely available on the Web and will be regularly upgraded with new GO databases for new organisms and pathway database information.

Acknowledgment:
We would like to thank our colleagues at the Korean Bio Information Center (KOBIC) for helpful comments and discussions. This work was supported by the Bio-infrastructure Program of the Korea Ministry of Science and Technology (MOST), the KRIBB Research Initiative Program of Korea, 07-38 Program by Korea Agency Digital Opportunity and Promotion, and M10508040002-07N0804-0021 grant by MOST.

References:

[01] M. Ashburner, *et al*., *Nature genetics*, 25: 25 (2000) [PMID: 10802651]
[02] M. Kanehisa & S. Goto, *Nucleic acids research*, 28: 27 (2000) [PMID: 10592173]
[03] http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
[04] O. G. Troyanskaya, *et al.*, *Proc Natl Acad Sci USA.*, 100: 8348 (2003) [PMID: 12826619]
[05] J. Tomfohr, *et al.*, *BMC Bioinformatics*, 6: 225 (2005) [PMID: 16156896]
[06] A. Subramanian, *et al.*, *Proc Natl Acad Sci USA.*, 102: 15545 (2005) [PMID:16199517]
[07] A. D. Windt, *et al.*, *DNA Cell Biol*., 26: 765 (2007) [PMID: 17867930]
[08] S. Y. Kim & D. J. Volsky, *BMC Bioinformatics*, 6: 144 (2005) [PMID: 15941488]
[09] Y. Benjamini, *et al.*, *Behavioural Brain Research*, 125: 279 (2001) [PMID: 11682119]

Edited by T. Nandi, H. Tan, T. W. Tan & S. Ranganathan
Citation: Yang *et al.*, Bioinformation 2(5): 194-196 (2007)