MOLECULAR & CELLULAR TOXICOLOGY, Vol. 3, No. 3, 208-213, September 2007

SOP (Search of Omics Pathway): A Web-based Tool for Visualization of KEGG Pathway Diagrams of Omics Data

Jun-Sub Kim[1], Hye-Jung Yeom[1], Seung-Jun Kim[1], Ji-Hoon Kim[1], Hye-Won Park[1], Moon-Ju Oh[1] & Seung Yong Hwang[1]

[1]Department of Biochemistry, Hanyang University & GenoCheck Co. Ltd., Sangrok-gu, Ansan, Gyeonggi-do 426-791, Korea
Correspondence and requests for materials should be addressed to S.Y. Hwang (syhwang@hanyang.ac.kr)

Accepted 24 July 2007

Abstract

With the help of a development and popularization of microarray technology that enable to us to simultaneously investigate the expression pattern of thousands of genes, the toxicogenomics experimenters can interpret the genome-scale interaction between genes exposed in toxicant or toxicant-related environment. The ultimate and primary goal of toxicogenomics identifies functional context among the group of genes that are differentially or similarly co-expressed under the specific toxic substance. On the other side, public reference databases with transcriptom, proteom, and biological pathway information are needed for the analysis of these complex omics data. However, due to the heterogeneous and independent nature of these databases, it is hard to individually analyze a large omics annotations and their pathway information. Fortunately, several web sites of the public database provide information linked to other. Nevertheless it involves not only approriate information but also unnecessary information to users. Therefore, the systematically integrated database that is suitable to a demand of experimenters is needed. For these reasons, we propose SOP (Search of Omics Pathway) database system which is constructed as the integrated biological database converting heterogeneous feature of public databases into combined feature. In addition, SOP offers user-friendly web interfaces which enable users to submit gene queries for biological interpretation of gene lists derived from omics experiments. Outputs of SOP web interface are supported as the omics annotation table and the visualized pathway maps of KEGG PATHWAY database. We believe that SOP will appear as a helpful tool to perform biological interpretation of genes or proteins traced to omics experiments, lead to new discoveries from their pathway analysis, and design new hypothesis for a next toxicogenomics experiments.

Keywords: Database, KEGG, Omics, Pathway, Toxicogenomics

Omics is called a general term for a broad discipline of science for investigating the interactions of biological information in various omes that is a word for describing very large-scale data in biology[1].

Toxicogenomics is highlighted as a new scientific field in which researcher examines how the genomic elements respond to their environmental stressors or toxicants[2-4]. It has been described as the approach of response between genome and toxic substances, the measurement of gene expression patterns in the toxic environment, and the study of discovery of toxicant pathways according to the gene expression changes. In this regards, toxicogenomics includes genomics for genome, transcriptomics for mRNA complement, proteomics for proteome, and bioinformatics[5-7]. As a result of microarray technology has become most popular in last decade, toxicogenomics makes it possible to inspect the expression of large groups of gene at a time and provide experimenters the important key to enhance discovery of toxicant pathways and biomarkers of specific chemical and drug targets[8].

As a rule, toxicity in biological pathway can results from the expression change not only in a single or few genes but also in relationship of gene interactions, the toxicogenomics researchers focus on interpreting functions of known or unknown genes, expression pattern, and their metabolic pathway information by searching and comparing with omics databases such as GenBank[9] for nucleotides and proteins, Swiss-Prot[10] for gene products crated protein sequence database, KEGG (Kyoto Encyclopedia of Genes and Genomes)[11] for biological pathway which is a collection of metabolism, genetic information processing, environmental information processing, cellular process, human diseases and drug development and GO (Gene Ontology)[12] for gene ontology terms assigned to the three ontologies, molecular function, cellular component or biological process.

However, owing to the heterogeneous and independent natures of public databases[13,14], it is not only trivial and tedious to compare the genomic-scale omics

data with the databases, but is hard to substantially analyze the biological significance of the omics data at a time. For example, in analyzing one gene expression profile which is preprocessed and normalized from a microarray results, it may be time consuming task: the collection of the functional information obtained from the different databases and the connection gene IDs to protein IDs and gene IDs/protein IDs to the KEGG pathway categories.

In recent years, numerous studies have attempted to construct the local databases that have been systemically integrated for user-based convenience[15-19]. Several stand-alone applications based on the Linux or Windows operating systems such as GenMAPP have been developed as pathway-based visualization tool for analyzing microarray data[20]. However, there have been a few researches regarding to support user-friendly web-applications for the visualization of KEGG pathway maps in the genomic-level omics data.

In this overall perspective, we have developed SOP (Search of Omics Pathway) database system to integrate public omics databases, to retrieve a set of query genes that arise from omics experiment such as microarray, and to approach pathway-oriented data analysis rather than traditional gene-centric data analysis. As a consequence, SOP provides: 1) a cross-linked biological annotation of all the genes identifier, proteins identifier and KEGG pathway categories, 2) a visualized pathway image of complex omics data based on KEGG reference pathways to overview gene expression profile.

SOP is a web-based search tool that is made up of three display panels (Figure 1): Query Panel, Viewer Panel and Annotation Panel. It is accessible from the web site at http://www.koreagene.co.kr/omics/sop/.

Query Panel

Query Panel contains three options (Organism, Type, Query) for a user query on the left side of window. Currently, Organism option is available for three organisms (*Homo sapiens, Mus musculus, Rattus norvegicus*). Type option is the drop down box with which

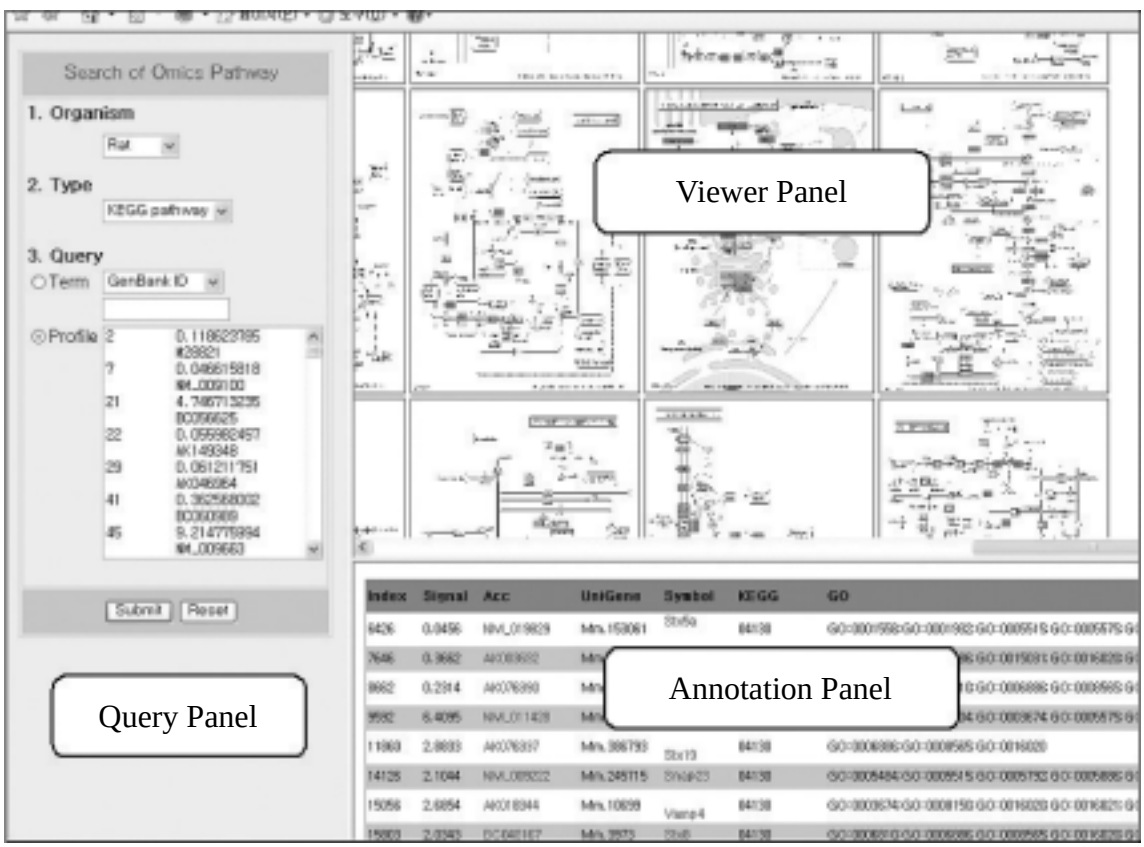

Figure 1. Main frame of the SOP web application. SOP main web page supports tree panels (Query Panel, Viewer Panel, Annotation Panel) on web browser. In Query Panel user can submit query to SOP server by selecting Organism option (human, mouse, rat) and inputting user query in Query options (a single identifier or gene express profile). If user query is successfully submitted to SOP web server, the results of the query appear to Viewer Panel with a matrix pathway image and Annotation Panel displayed biological annotation table.

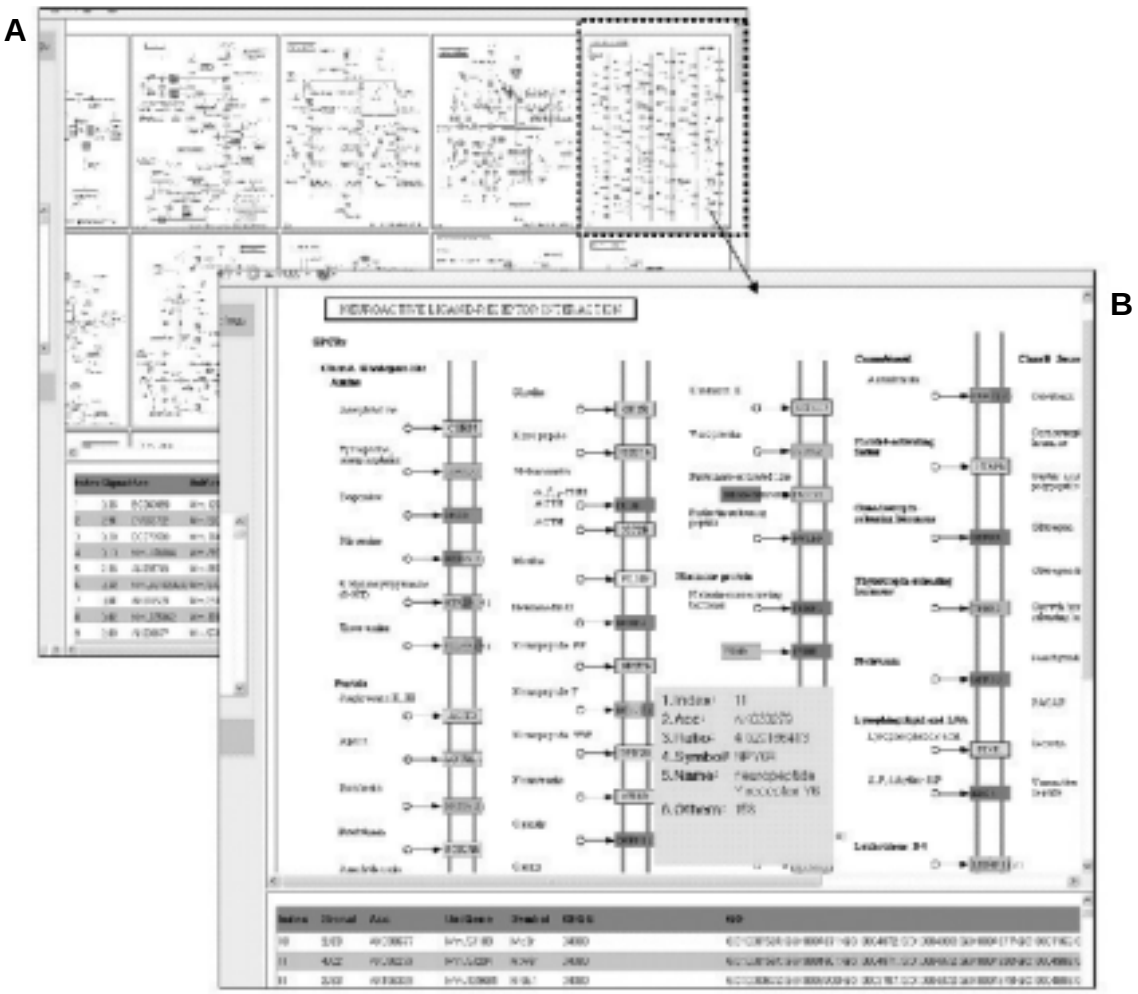

Figure 2. KEGG pathway outputs of user interface of the SOP. When user selects the KEGG type item of Query options, SOP exhibits a matrix that consists of several thumbnail images of KEGG pathway maps. User can observe the entire overview of pathways in the Viewer Panel at a glance (A). By clicking one thumbnail image of the matrix of thumbnails, a full-sized scalable KEGG pathway image can be obtained and user can simultaneously navigate the table of gene annotation in the Annotation Panel (B).

user is able to select a KEGG pathway item. Query option includes alternative radio buttons (Term, Profile). While the Term radio button of this option requires a choose of gene identifier such GenBank accession number, UniGene ID or official gene symbol, the Profile radio button of this option demands a gene expression profile resulted from microarray experiment. This text area box must be inputted as the tab-delimited text table that is made of three columns: integer for unique identifier, decimal for gene expression signal ratio and character for gene identifier in left to right order.

Viewer Panel

Viewer Panel presents the results of user query transmitted from Query Panel (Figure 2). The results belong to given user query are illustrated as a matrix of the thumbnail pathway images on this Panel. User can see an enlarged pathway image by clicking one thumbnail image on the matrix. By means of the graphical tool (Image::Magick module) of Perl, SOP draws these new pathway maps based on coordinates of genes or gene products extracted from the species-specific KEGG reference pathway diagrams. In general, the original KEGG reference pathways are constructed as a set of graphical symbols (circle, rectangle, line, etc.). A rectangle (box) is an enzyme (gene or gene product) with the EC (Enzyme Commission) number inside and a circle is a metabolic compound. The biological relationship between them is illustrat-

ed as line. For the visualization of a color-coded pathway maps reflecting the gene expression pattern, SOP is portrayed by changing the color of the rectangle according to the gene expression signal ratio in the profile queried by user. The boxes on the edited map are colored green for down regulation (≤ 0.50), yellow for normal regulation (> 0.50 and < 2.00), and red for up regulation (≥ 2.00). In processing pathway map, if one box includes multiple genes, it is vertically split for indicating them.

Annotation Panel

Annotation Panel shows a biological annotations table traced to NCBI, Swiss-Prot and KEGG GENE database. This table involves gene index, gene expression signal ratio, gene ID, UniGene ID, KEGG map numbers, GO IDs, and so on. The data in the table is linked to the web site reference database that offers more detailed information.

Discussion

The various omics data in genomic level were individually stored in the heterogeneous public databases which individually provide only their genomic information such as transcriptom, proteom, biological pathway, and gene ontology. Although they offer the results of search with one query (gene identifier, proten identifier, or gene symbol) on their web sites, the results are too limited in the point of the functional prediction of the omics data which have a large volume of genomic identifiers and expression values to analyze biological validation that may be divided into functional information of one gene, relationship of genomic elements, and reaction in a particular pathway. Specially, in toxicogenomics researchers it is perhaps important to approach the expression patterns of large groups of genes within pathways in the toxic environment. Therefore for the purpose of the biological pathway analysis of high-throughput screening data such as microarray or protein chip experimental results, it is essential to integrate these public omics databases. In recent years, there have been more researches regarding the construction of homology database and the production of computational applications for analyzing genomic experimental data. Of computational software, a web-based application is slower than a stand-alone application in aspect of running time. But a web tool is simpler to run than a stand-alone tool to. Also in user-friendly as well as portable point of view, if only users are accessible via web, they are able to realize their purpose wherever it may be.

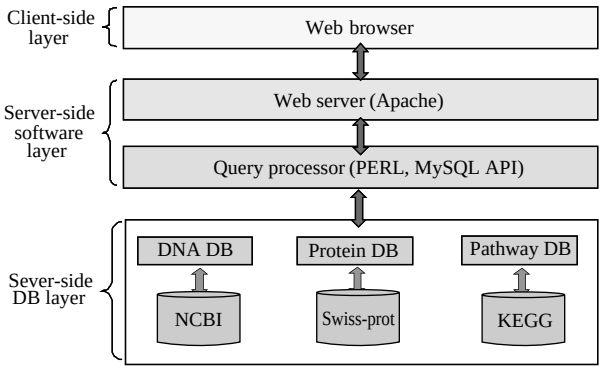

Figure 3. Schematic overview of SOP database system. Server-side DB layer was manipulated by Perl and MySQL. Server-side software layer is written in Perl for compiling user query and executing command. Client-side layer gets user query through web browser.

In this insight, we have built up an integrated database system consisting of a biological annotation database and a pathway database, that has supplied users with web-based tool to interpret of omics data based on KEGG GENE/PATHWAY database. SOP is a flexible and user-friendly web tool.

Although the usable species of SOP are restricted with human, mouse and rat, we are planning on adding additional species in future. In addition to KEGG GENE/PATHWAY database by means of a source data set of SOP database system, we will add KEGG LIGAND database which consists of several databases which are REACTION for substrate-product relations, COMPOUND for metabolites and chemical compounds, and ENZYME for enzyme molecules, DRUG for chemical structures of all approved drugs[24].

Methods

SOP database system is built as a server-client architecture system on Linux (Fadora core 5.0). Script for the back-end and web-based query interface is coded with Perl (Practical Extraction and Reporting Language: http://www.perl.org) programming language combining a MySQL database management system.

This database system is based on traditional three layers architecture of client-server database system that is composed of sever-side DB layer, sever-side software layer, and client-side layer (Figure 3).

Sever-side DB Layer

Sever-side DB layer is played a dominant role in connection between biological annotation database

and KEGG pathway database. This layer is the front-end of a relational database (MySQL) containing biological annotation data originated from the publicly available resources that are NCBI's UniGene[21], Locuslink[22] for nucleotides annotation, Swiss-Prot for proteins annotation, KEGG GENES database for genes in high-quality genomes and KEGG PATHWAY database for pathway groups[23]. On the one hand, as these public biological databases are continuously updated in the public domain, we designed to continuously download the original data sources from their FTP sites and autoupdate our database system to keep the information in the most recent and accurate status.

For creating biological annotation table and pathway table on this layer, Perl script of SOP parses the source files for mapping gene to protein and gene/protein to KEGG pathway categories and to extracts the values of gene coordinate from the original KEGG pathway maps. Finally Perl script converts parsed files into new files, and then automatically stores in SOP database.

Sever-side Software Layer

Sever-side software layer is grouped into the Query Processor (Perl MySQL API) and the Web Server. The former is entirely coded with Perl with a several Perl modules that are comprised DBD::mysql, Image ::Magick, and CGI module. This module is able to freely use at CPAN (Comprehensive Perl Archive Network; Comprehensive Perl Archive Network) search site. DBD::mysql module serves as an interface between the Perl and the MySQL programming API that comes with the MySQL relational database management system. Image::Magick module is an graphical interface to read, manipulate, or write an image. CGI module provides shortcut functions that produce HTML and passes user query to Perl script. The functions of this layer are to compile a user query to executable command that computer is able to decipher, extract results data from the annotation tables of sever-side DB layer and deliver the results data to Web Server.

Client-side Layer

In SOP web-based tool, a client interface is a web browser which enables user to access to SOP database system on internet. SOP has been tested with Perl and Javascript and designed to work with IE 6.0 web browser or higher.

Acknowledgements

This subject is supported by Korea Ministry of Environment as "The Eco-technopia 21 project".

References

1. -omics. Wikipedia, the free encyclopedia Available at http://en.wikipedia.org/wiki/-omics. Accessed July 10 (2007).
2. Aardema, M. J. & MacGregor, J. T. Toxicology and genetic toxicology in the new era of "toxicogenomics": impact of "-omics" technologies. *Mutat Res* **499**:13-25 (2002).
3. Burchiel, S. W. *et al*. Analysis of genetic and epigenetic mechanisms of toxicity: potential roles of toxicogenomics and proteomics in toxicology. *Toxicol Sci* **59**:193-195 (2001).
4. Ulrich, R. & Friend, S. H. Toxicogenomics and drug discovery: will new technologies help us produce better drugs? *Nat Rev Drug Discov* **1**:84-88 (2002).
5. Marchant, G. E. Toxicogenomics and toxic torts. *Trends Biotechnol* **20**:329-323 (2002).
6. Orphanides, G. Application of gene expression profile to toxicology. *Toxciol Lett* **140**:145-148 (2003).
7. Tong, W. *et al*. Development of public toxicogenomics software for microarray data management and analysis. *Mutat Res* **549**:241-253 (2004).
8. Lee, K. M., Kim, J. H. & Kang, D. Design issues in toxicogenomics using DNA microarray experiment. *Toxicol Appl Pharmacol* **207**:200-208 (2005).
9. Benson, D. A. *et al*. GenBank. *Nucleic Acids Res* **35**: 21-25 (2007).
10. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**:365-370 (2003).
11. Kanehisa, M. The KEGG database. *Novartis Found Symp* **247**:91-103, 119-128, 244-252 (2002).
12. Harris, M. A. *et al*. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* **32**: 258-261 (2004).
13. Waters, M. *et al*. System's toxicology and the Chemical Effects in Biological Systems (CEBS) knowledge base. *EHP Toxicogenomics* **111**:15-28 (2003).
14. Lee, W. S. *et al*. The intelligent data management system for toxicogenomics. *J Vet Med Sci* **66**:1335-1338 (2004).
15. Arakawa, K. *et al*. KEGG-based pathway visualization tool for complex omics data. *In Silico Biol* **5**:419 -423 (2005).
16. Arakawa, K. *et al*. G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics* **19**:305-306 (2003).
17. Klukas, C. & Schreiber, F. Dynamic exploration and editing of KEGG pathway diagrams. *Bioinformatics* **23**:344-350 (2007).
18. Kono, N., Arakawa, K. & Tomita, M. MEGU: pathway mapping web-service based on KEGG and SVG. *In Silico Biol* **6**:621-625 (2006).

19. Altermann, E. & Klaenhammer, T. R. PathwayVoyager: pathway mapping using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database. *BMC Genomics* **6**:60 (2005).
20. Salomonis, N. *et al*. GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* **24**:217 (2007).
21. Pontius, J. U., Wagner, L. & Schuler, G. D. UniGene: a unified view of the transcriptome. In: The NCBI Handbook. Bethesda (MD): *National Center for Biotechnology Information* (2003).
22. Pruitt, K. D., Katz, K. S., Sicotte, H. & Maglott, D. R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* **16**:44-47 (2000).
23. Kanehisa, M. *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34**:354-357 (2006).
24. Goto, S. *et al*. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* **30**:402-404 (2002).