Genomics & Informatics Vol. 4(3) 129-132, September 2006

GraPT: Genomic InteRpreter about Predictive Toxicology

Jung Hoon Woo[1,3], Yu Rang Park[1], Yong Jung[1], Jihun Kim[1] and Ju Han Kim[1, 2]*

[1]Seoul National University Biomedical Informatics (SNUBI), [2]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea, [3]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-742, Korea

Abstract

Toxicogenomics has recently emerged in the field of toxicology and the DNA microarray technique has become common strategy for *predictive toxicology* which studies molecular mechanism caused by exposure of chemical or environmental stress. Although microarray experiment offers extensive genomic information to the researchers, yet high dimensional characteristic of the data often makes it hard to extract meaningful result. Therefore we developed toxicant enrichment analysis similar to the common enrichment approach. We also developed web-based system graPT to enable considerable prediction of toxic endpoints of experimental chemical.

Keywords: toxicogenomics, predictive toxicology, high-dimensional data, toxicant enrichment analysis, web-based system, prediction, toxic endpoints

Introduction

As genomics technologies have been gradually integrated into conventional toxicology, the new era so-called *toxicogenomicsis* emerged in this field of study. Especially DNA microarray which explains thousands of transcripts' changes has become a well established method in biological research fields. Gene expression is a sensitive indicator of toxicant exposure, disease state, and cellular metabolism, and thus represents a unique way of characterizing how cells and organisms adapt to changes in external environment (Lettieri, 2006). Proponents of toxicogenomics aim to apply mRNA expression technology to study chemical effects in biological systems (Afshri *et al.*, 1999; Lovet, 2000).

Predictive toxicology, predicting toxic endpoints caused by unknown chemical exposure, has been the main issue of conventional study, and accordingly it still is a challenge of toxicogenomics (Laura *et al.*, 2004). Comparing gene expression patterns generated by microarray between model organisms stimulated with toxicant or environmental stress and control have been widely used strategy for prediction of toxicity of new and existing chemicals (Fielden *et al.*, 2001).

However, it is unlikely to get meaningful information directly fromhigh-dimensional gene expression data. The enrichment approach with cluster analysis (-apowerful technique for dimension reduction) which uses gene ontology (GO), and pathway information has emerged as a result. The relationship between Gene and toxicant can be sources of different kinds of enrichment approach.

The major purpose of toxicant enrichment analysis is to identify biological function of experimental toxicants and to get vital information such as prediction of toxic endpoints. We developed toxicant enrichment strategy and the graPT based on the gene-toxicant relationship in order to provide information on the association between toxic endpoints and unknown chemicals.

Methods

Public toxicogenomics data localization and integration

We localized CTD (Comparative Toxicogenomics Database-*http://ctd.mdibl.org/*) data and *CHE* (*http://database.healthandenvironment.org/*) Toxic data used for determining relationship between gene, toxicant, and disease data types. We localized *Entrez Gene* (*http://www.ncbi.nlm.nih.gov/entrez/*), *RefSeq* (*http://www.ncbi.nlm.nih.gov/RefSeq/*), and *MeSH* (*http://www.nlm.nih.gov/mesh/*) data for annotating the three data types respectively and constructed relational database by integrating these data sources (Fig.1).

Toxicant enrichment test

Frequencies of toxicant terms within the dataset are calculated and compared with reference frequencies. The probability of obtaining by chance a number of k of related genes for given toxicant term among a dataset size n, knowing that reference dataset contains m such related genes out of N genes, is then calculated. This

*Corresponding author: E-mail juhan@snu.ac.kr, Tel +82-2-740-8320, Fax +82-2-742-5947
Accepted 20 August 2006

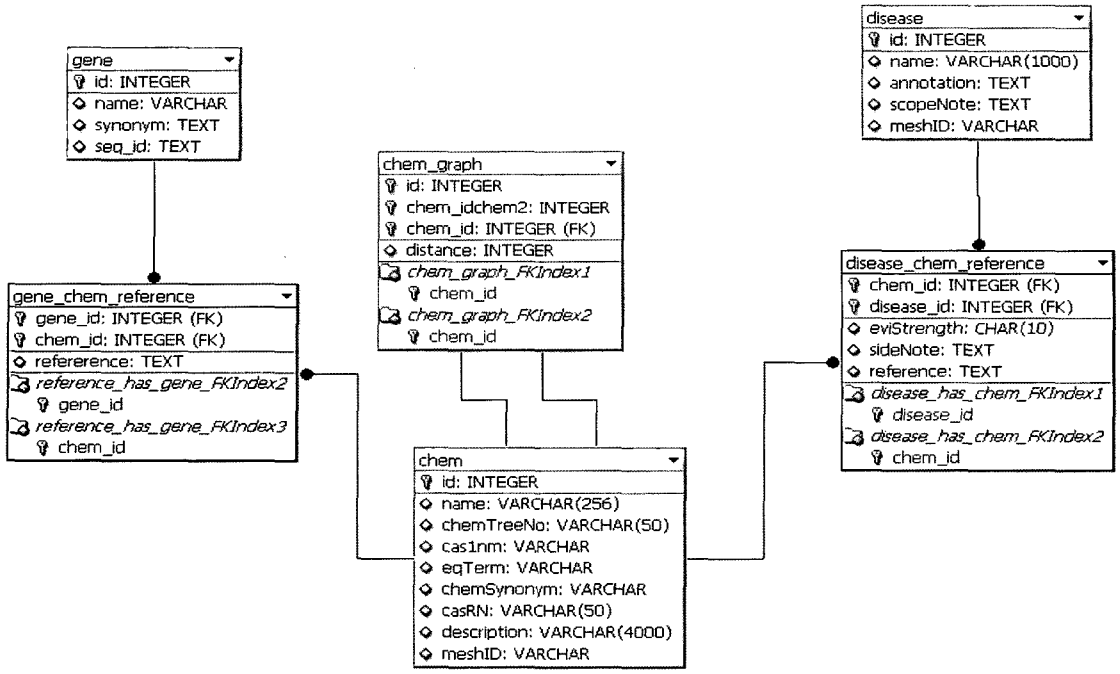

Fig. 1. Entity-Relationship diagram of database in the *gra*PT. Gene, chemicals (toxicant), and disease are the three data types of the *gra*PT. Both gene and disease have respective relationship withchemical. Information regarding the hierarchy of chemicals is also included in the system.

probability follows the hypergeometric distribution described in Eq. (1):

$$\Pr\{x=k\}=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{m}{k}} \quad (1)$$

where the random variable X represents the number of genes within a given gene subset, related with a given toxicant term. Because this approach simultaneously tests the statistical significances of the associations of a set of genes to multiple toxicants, multiple hypothesis testing problems should be considered. We applied FDR to offer a much reliable statistical testing (Benjamini *et al.*, 1995). The percentage of such toxicants selected by chance is the FDR, and adjusted P-value threshold was decided by determining the FDR (Storey *et al.*, 2003)

Input and output

Input

A list of differentially expressed genes (DEGs) is produced as a common result of DNA microarray experiments. The input of operation in *gra*PTis the DEGs list and the cut-off threshold about p-value. The ID for genes must contain at least one of GenBank accession number, SwissProt ID or TrEMBL ID.

Output

The *gra*PT produces a list of best matching toxicants for input DEGs list with statistical significance scores of none random association (Fig. 2). Relevant toxicants are listed in ascending order of p-values (and adjusted p-value for multiple testing problems). Users are provided with additional information on the listed toxicants. Inputted genes information is hyperlinked to an automated annotation page provided by NCBI gene centric database, *Entrez Gene*.

Application

Our first application was the data set by Kharasch *et al.* (2006). These data were collected for the purpose of identifying genes related in nephrotoxicity from haloalkene fluoromethyl-2,2-difluoro-1-(trifluoromethyl)-vinyl ether (FDVE). They performed t-test to select differentially expressed genes according to the significance assigned at p-value<0.05. We applied toxicant enrichment analysis to the 12 DEGs to find out significantly enriched toxicants

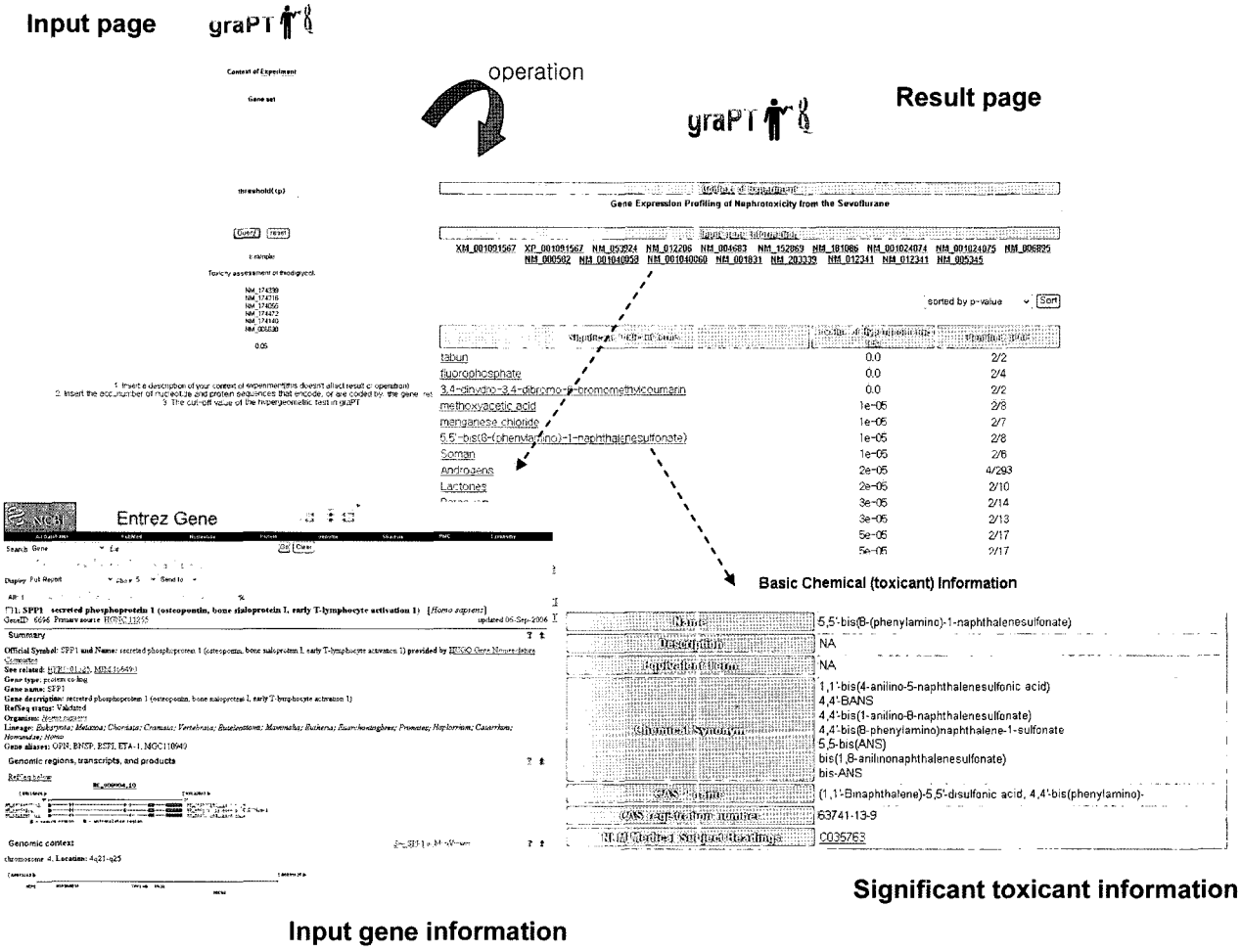

Fig. 2. graPT user interface. Users determine the differentially expressed genes and threshold of scoring test in the input page. Output of operation is a list of toxicants sorted by p-value. graPT provides summary information for listed significant toxicants through its internal annotating system. Users are offered with external link for the information of input genes through the NCBI gene centric database, Entrez Gene.

-through genomic information. Significance was assigned at a minimum 2 identified nodes among mapped genes per each toxicant and p-value<0.001. Second application was the data set generated by Thukral *et al.* (2005). They selected 9 DEGs with biomarkers of nephrotoxicity and like the preceding we tested the DEGs with graPT. 8 and 10 toxicants each were significantly enriched with two experimental data sets (Table 1).

Table 1. Result of toxicant enrichment analysis using DEGs sets generated by previous studies

	DEGs[a]	Significantly enr iched toxicants[b]
Kharasch *et al.*	GAPDH, CFTR2, KIM , RGN , TNF, HNMT , SPP1 , HSP70 1B , CLU, SCF21m1, CRFG , Hspa1a	tcdA protein, Clostridium difficile, Carbamates Hyaluronic Acid, Kainic Acid, sodium arsenite, geldanamycin Antihypertensive Agents, Hypoglycemic Agents
Thukral *et al.*	GST-pi 2, Slc21a1, Slc22a2, Slc21a7 Osteopontin, Kim1, Timp1, Regucalcin, C8	Procainamide, Sulfobromophthalein, estrone sulfate, Quinidine, Cadmium Chloride, Leukotriene C4, Testosterone Acetaminophen, Taurocholic Acid, Choline

[a] Indicates that differentially expressed genes are selected by author's own criterion.
[b] Indicates selected toxicants by two criteria; hypergeometric p value <0.001 and number of identified nodes >= 2

Discussion

The common toxicogenomics microarray analysis generates set of genes stimulated by the experimental chemical. The *gra*PTitself does not perform any statistical test for selecting differentially expressed genes so that the users should select them by their own statistical criterion. There is no unique standard of scoring and selecting genes-set, several kinds of trials are required to produce best result in *gra*PT.

Through performing statistical test based on hypergeometric distribution, this system permits the automatic ranking of all toxicant terms, as well as the evaluation of the significance of their occurrence within the dataset. We applied the data sets from previous nephrotoxicity related studies, and observed that significantly enriched toxicants were partly related to nephrotoxicity. Our result suggests that if high ranked toxicants show relationship with experimental chemical based on the genomic information, researcher can predict toxic endpoints of unknown chemicals more effectively.

Acknowledgements

This study was supported by a grant from Korea Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (0412-MI01-0416-0002). J. W. and Y. J's research activity is partly supported by a grant from the TGRC project, Korea Food & Drug Administration, Republic of Korea.

Reference

Afshari. C.A., Nuwaysir, E.F., and Barrett, J.C. (1999). Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. *Cancer Res.* 59, 4750-4760.

Benjamini,Y. and Hochberg,Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Statist.Soc.* 57, 289-300.

Fielden, M.R. and Zacharewski, T.R. (2001).Challenges and limitation of genen expression profiling in mechanistic and predictive toxicology. *Toxicological Sciences* 60, 6-10.

Hamadeh, H.K., Amin, R.P., Paules, R.S., and Afshari, C.A. (2002). An overview of toxicogenomics. *Curr. Issues Mol. Biol.* 4, 45-56.

Kharasch, E.D., Schroeder, J.L., Bammler, T., Beyer, R., and Srinouanprachanh, S. (2006). Gene Expression Profiling of Nephrotoxicity from the Sevoflurane Degradation Product Fluoromethyl-2,2-difluoro-1-(trifluoro-methyl) vinyl Ether ("Compound A") in Rats. *Toxicol. Sci.* 90, 419-431.

Laura, S., Lee, E.B., and Eric, B.W. (2004). Toxicogenomics in Predictive Toxicology in Drug Development. *Chemistry and Biology* 11, 161-171.

Lettieri, T. (2006).Recent Applications of DNA Microarray Technology to Toxicology and Ecotoxicololgy. *Environmental Health Perspective* 114, 4-9.

Lovett, R.A. (2000). Toxicogenomics. Toxicogists brace for genomics revolution. *Science* 289, 536-537.

Mattingly, C.J., Colby, G.T., Forrest, J.N., and Boyer, J.L. (2003). The Comparative Toxicogenomics Database (CTD). *Environ Health Perspect* 111, 793-795.

Storey, J.D. and Tibshirani, R. (2003). Statistical significance for genome wide studies. *Proc. Natl . Acad. Sci. USA* 100, 9440-9445.

Thukral, S.K., Nordone, P.J., Hu, R., Sullivan, L., Galambos, E., Fitzpatrick, V.D., Healy, L., Bass, M.B., Cosenza, M.E., and Afshari, C.A. (2005). Prediction of Nephrotoxicant Action and Identification of Candidate Toxicity-Related Biomarkers. *Toxicol. Pathol.* 33, 343-355.