別紙4

Sakata Y., Yoshioka W., Tohyama C., and Ohsako S.	Internal genomic sequence of human CYP1A1 gene is involved in superinduction of dioxin-induced CYP1A1 transcription by cycloheximide.	Biochem Biophys Res Comm	355	687-692	2007
Fujibuchi W, Kim H, Okada Y, Taniguchi T, Sone	H. High-performance gene expression module analysis tool and its application to chemical toxicity data.	Methods Mol Biol.	577	55-65.	2009
2: Alam MS, Ohsako S, Matsuwaki T, Zhu XB, Tsunekawa N, Kanai Y, Sone H, Tohyama C, Kurohmaru M.	Induction of spermatogenic cell apoptosis in prepubertal rat testes irrespective of testicular steroidogenesis: a possible estrogenic effect of di(n-butyl) phthalate.	Reproduction.	Feb;139(2)	427-37.	2010
3: Ohsako S, Fukuzawa N, Ishimura R, Kawakami T, Wu Q, Nagano R, Zaha H, Sone H, Yonemoto J, Tohyama C.	Comparative Contribution of the Aryl Hydrocarbon Receptor Gene to Perinatal Stage Development and Dioxin-Induced Toxicity Between the Urogenital Complex and Testis in the Mouse.	Biol Reprod.			
Sone H, Okura M, Zaha H, Fujibuchi W, Taniguchi T, Akanuma H, Nagano R, Ohsako S, Yonemoto J.	Profiles of Chemical Effects on Cells (pCEC): a toxicogenomics database with a toxicoinformatics system for risk evaluation and toxicity prediction of environmental chemicals.	J Toxicol Sci.	35(1)	43	2010
森拓哉、吉永淳、河原純子、溝井美穂、安達修一	日本人小児の多環芳香族炭化水素類曝露評価	日本衛生学雑誌	64	817-823	2009
Y. Suzuki, M. Niwa, J. Yoshinaga, C. Watanabe, Y. Mizumoto, S. Serizawa, H. Shiraishi	Exposure assessment of phthalate esters in Japanese pregnant women by using urinary metabolite analysis	Environmental Health and Preventive Medicine	14	180-187	2009

別紙4

書籍

著者氏名	論文タイトル名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年	ページ
Kojima Y, Mizuno K, Sasaki S, Hayashi Y, Kohri K	Gene therapy for male infertility; Potential and limitation	Lejeune T and Delvaux	Human spermatozoa: Maturation, Capacitation and	Nova Science Publishers, Inc.,	New York	in press	
Kojima Y, Casale P.	Pediatric robotic pyeloplasty for ureteropelvic junction obstruction	Ost MC	Laparoscopic reconstructive surgery in children and adults.	Springer	New York	in press	
Kojima Y, Casale P.	Robotic Bladder Surgery in Children	Hemal AK and Menon M	Robotic Urologic Surgery	Springer	New York	in press	
Kojima Y, Sasaki S, Kubota Y, Hayashi S, Kohri K.	New Insights into Alpha1-Adrenoceptor Subtypes and Pharmacogenomics of Benign Prostatic Hyperplasia	Morrison EP	Horizons in Cancer Research	Nova Science Publishers	New York	in press	

研究成果の刊行物・別刷り

Chapter 5

High-Performance Gene Expression Module Analysis Tool and Its Application to Chemical Toxicity Data

Wataru Fujibuchi, Hyeryung Kim, Yoshifumi Okada, Takeaki Taniguchi, and Hideko Sone

Summary

Gene clustering is one of the main themes of data mining approaches in bioinformatics. Although it has the power to analyze gene function, interpretation of the results becomes increasingly difficult when the number of experiments (samples) exceeds hundreds or more. A new type of clustering called "biclustering," where genes and experiments are coclustered in a large-scale of gene expression data, has been extensively studied in the last decade. We have developed "SAMURAI," an original program that detects all the biclusters or "gene modules" whose genes have similar expression patterns to query profile using the ultrafast data mining algorithm called Linear-time Closed itemset Miner (LCM). Using chemical toxicity dataset from J&J rat liver experiments, we compiled an exhaustive dictionary of gene modules by searching datasets of gene modules with each chemical exposure experiment as query. Through the module analysis, we found that our program can detect up/down-regulated gene sets that significantly represent particular GO functions or KEGG pathways, thereby unraveling reactions and mechanisms common to different toxicochemical treatments of hepatocytes.

Key words: Gene expression module, Biclustering, Chemical toxicity, Data mining, Linear time common itemset miner, Common reaction and mechanism

1. Introduction

Microarrays or other high-throughput gene expression analysis systems provide extensive information on gene expression differences under various experimental conditions, such as cell type, developmental stage, and reaction to stimulus. Recent easy access to such experimental techniques has promoted the accumulation of gene expression data in public gene expression data

Hisashi Koga (ed.), *Reverse Chemical Genetics*, Methods in Molecular Biology, vol. 577
DOI 10.1007/978-1-60761-232-2_5, © Humana Press, a part of Springer Science+Business Media, LLC 2009

repositories, such as GEO and ArrayExpress *(1, 2)*. Among available tools to analyze such large-scale data, a promising method called "biclustering" *(3)* has emerged and has been widely studied *(4–9)* for its ability to mine datasets containing hundreds of experiments. Biclustering detects common gene expression patterns or "gene expression motifs" that are represented in any combination of experiments. A subset of genes that contain a common expression pattern in a subset of experiments is called a "gene module."

We have developed a high-performance biclustering method that has high calculation speed and high biological evaluation accuracy. Our biclustering system called "SAMURAI (System for Assembling Modules by Ultra Rapid Algorithm on Itemsets)" *(10)* can search thousands of microarray data for gene modules in several seconds in most cases. In addition, the detected gene modules show surprisingly high accuracy of matching to known gene function groups in the study of 2,988 disease microarray data provided by the Critical Assessment of Microarray Data Analysis meeting *(11)*. Here we applied this system to a chemical toxicity dataset obtained from J&J rat liver experiments and compiled an exhaustive dictionary of gene modules by searching the dataset with each chemical exposure experiment as query. From the resultant gene modules, we found that there are a total of 92,100 modules of which 10,805 (11.7%) represent known functions (GO or KEGG) at a high significance threshold ($p < 1e-5.5$ and $p < 1e-4$, respectively).

2. Materials and Methods

To obtain and analyze gene modules from large-scale gene expression data, we first normalize and convert them into a unified file format. After obtaining formatted data, we discretize them to certain degrees of expression to find common expression patterns in a limited search space. Then, to reduce search space effectively and to detect gene modules in real time, we perform "query-and-database" search. In this approach, given query gene expression data by a user, all of the discretized values not common to the query in the database are erased, thus reducing database size extensively. This process is critical to the calculation of the following module detection process, giving exhaustive (not partial) results.

After fully reducing search space by discretizing and erasing data unrelated to query, we perform rapid data mining where common gene expression patterns that are preserved in all combinations of (maximum) experiments are exhaustively retrieved.

However, the gene modules retrieved in this step have rigid patterns with no relaxation (i.e., noise), which contrasts to real data containing noise and biological flexibility. Thus, in the next step, these "core modules" are compared with each other and merged into bigger modules containing noise. As the final output, we obtain each module consisting of a subset of genes and a subset of experiments. By scrutinizing both gene functions and experimental relationships in each module, we can formulate new and interesting hypotheses on gene functions and experimental groups.

2.1. Rat Chemical Exposure Dataset and Normalization

1. Download gene expression data of chemical effects on rat liver by J & J from the Web site: http://cebs.niehs.nih.gov/. To retrieve the data, select the subject "J&J Hepatotoxicant Library" in the "Display All Studies" page or limit data by the organization name "Johnson and Johnson" in the "Study Characteristics" page. There are 133 toxicochemical groups containing 964 microarray experiments. Go to "View Selected Microarray Data by Studies/Experiments" from the bottom of page, select the dataset, and go to "View Details about Selected Experiment(s)." Then, download microarray data files after entering "Click to Download" page. As the file size is 117.3 MB, downloading will take 5 min to hours depending on the user's network conditions.
2. Unzip downloaded files and select files to analyze. In this study, we select only experiments that have both chemical exposure done and control data collected on the same day. The number of such experiments is 298. Among 9,215 probes in the dataset, we delete low-abundance genes that show no expression in all of the 298 experiments and select only 7,614 probes. Every pair of gene expression values is transformed into log-fold-change abundance by subtracting control values from the chemical exposure after taking $\log_2$ and subtracting the median value in each array (*see* Note 1). Finally, the log-fold-change values are normalized to Z-score by $(x - \text{mean})/\text{SD}$.
3. To perform gene functional analysis in later process, check if each probe has a link to UniGene database. Only 5,832 probes have links to UniGene database. Then, to remove probe redundancies, take the average if multiple probes correspond to the same UniGene ID. As a result, we obtain 2,497 averaged probes that have links to UniGene database for 298 chemical exposure experiments.

2.2. Formatting Database by Discretization

1. Convert normalized dataset into rank-ordered discrete data within each experiment. To do this, select one experiment and sort genes by expression value. Then, put all the genes into 10,000 bins of the same size and assign every gene a rank value equal to the number of bins (1st–10,000th from low to high) that it belongs to.

2. Select one gene and make a distribution of rank values for all the 298 experiments. Then, set a discretization parameter and thresholds. In this study, we set the degree of discretization at 3 (±1, 0) and the thresholds at 3% or 5% from each side of top and bottom (i.e., 6% or 10% in total) in rank value distribution. Using these parameters, we assign discrete values to rank values.

2.3. Query Data and Database Compression

1. Given query gene expression data, discretize gene expression values using the same procedure as that employed in the above database formatting process. Use the same degree and thresholds to discretize query data as that used in the database discretization.
2. Compare discretized query to discretized database at the gene level. Then, delete all discrete values that differ from the query in the database. In addition, delete all zero values in the database. This procedure compresses the database to an extremely small size (*see* **Note 2**).

2.4. LCM for Data Mining of Core Gene Modules

1. LCM *(12, 13)* is an ultrafast algorithm for data mining that has been used to retrieve maximum common itemsets from a large list of itemsets called a transaction database (*see* Note 3). To apply this algorithm, assign item names to all the existing combinations of gene names and discrete values in the gene expression database. For example, we give item names "a – 1" and "a – 2" to the case that gene "a" has values of both "–1" and "–2" in the database. This procedure converts each gene expression experiment datum into an itemset list that symbolizes gene expression status consisting of gene names and their discrete values.
2. Write itemset list to a file in the format of one experiment in one line. Then, download the LCM program from the Web site: http://research.nii.ac.jp/~uno/codes.htm. Run the LCM program with the itemset list file with parameters of minimum size of gene modules to extract: *m* genes × *n* experiments. For example, the input command to run LCM on a Linux machine is: % lcm CqI –l *m* [input_file] *n* [output_file] (*see* **Note 3**).

2.5. Merging Redundant Modules

1. Raw gene modules (core modules) extracted by LCM are expected to be highly redundant. To reduce almost the same or quite similar gene modules in output data, merge them if they meet conditions specified by the following procedure. First, sort gene modules by size. Then, select the largest module and merge it with other modules one by one from large to small ones. Suppose we are merging modules A and B. If A and B share genes g1 and g2 and experiments e1 and e2 but A has another gene g3 and B has another experiment e3, the merged module will have genes g1, g2, and g3 and experiments e1, e2,

and e3 by adding missing gene (g3) and experiment (e3) to B and A, respectively. However, if any gene or experiment of the merged module contains inconsistent values, such as missing or different discrete values, the percentages of inconsistencies in each line and row of the merged module should be checked. If the inconsistency in every line and row is less than the threshold (in this study, 0.4 and 0.5 are adopted), execute the merge and replace the larger module with it. (Delete the smaller one.) Repeat this "check and merge" process for this larger module until it reaches the smallest one.

2. Repeat the above "check and merge" process from the next largest module to the smallest module. Once the smallest module is reached, sort the modules by size again and perform "check and merge" from the largest module to the smallest one for a new list of modules.
3. Repeat the above "sort, check, and merge" until no merge happens. Then, output the final (merged) modules to a file (*see* Note 4). The whole process from formatting data to merging modules is illustrated in Fig. 1. Here an example of two degrees (high and low expressions, or +1 and −1) of discretization is shown.

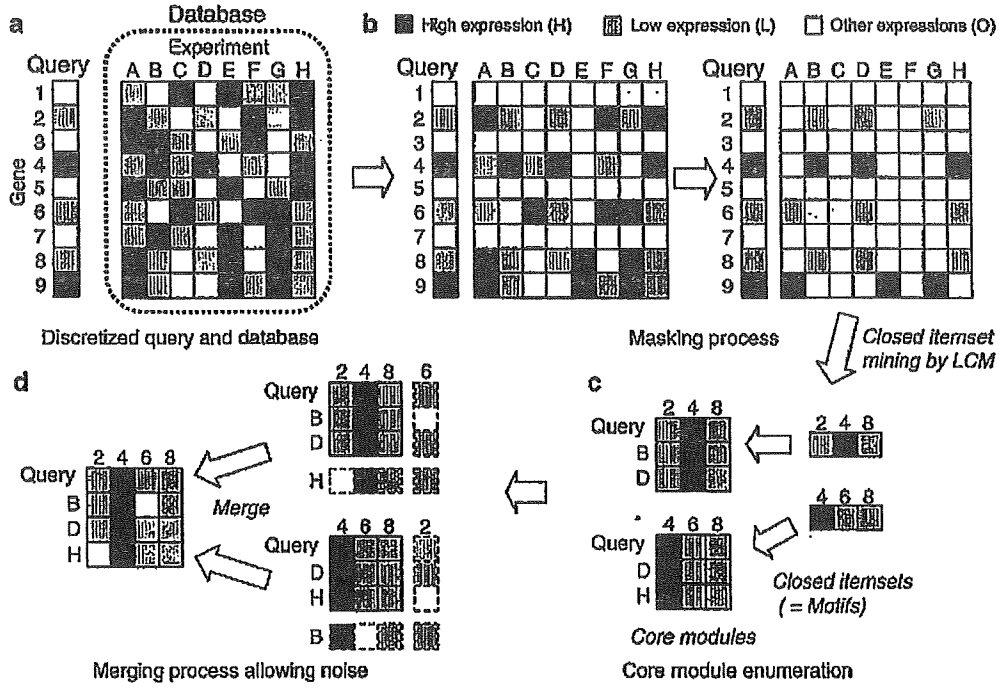

Fig. 1. Whole scheme of gene module search from gene expression database. The system consists of four steps (a) discretization of query and database, (b) database masking, (c) enumeration of core modules, and (d) module merging. In this example, the query and the database are discretized into only three degrees "High (+1)," "Low (−1)," and "Other expressions (0)".

2.6. Evaluation by GO and KEGG

Once the final set of modules is obtained, it is necessary to check the biological validities of those modules to verify if the selected parameters (discretization degree, noise threshold, module size, etc.) work properly. To approach this, compare each gene module with known biological functions or pathways to investigate if genes in a single module are statistically "enriched" for a particular category of functions. Here we describe the method of performing categorical enrichment analysis based on Gene Ontology (GO) functions and KEGG biological pathways.

1. Download necessary files from two FTP sites (a) gene2unigene and gene2go.gz files from ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/ for gene name conversion and GO term information, respectively; and (b) gene_map.tab for KEGG pathway map index information and rno_gene_map.tab files for rat specific pathway gene list from ftp://ftp.genome.ad.jp/pub/kegg/pathway/.
2. Take one gene expression module. Convert UniGene IDs into EntrezGene IDs via the cross-reference list in gene2unigene file (*see* **Note 5**).
3. Assign GO function terms to the converted (Entrez-) genes. Obtain four parameters (a) U, the number of EntrezGenes found in the input gene expression data (2,497 UniGenes); (b) M, the number of EntrezGenes assigned to each GO term; (c) k, the number of EntrezGenes in each gene module; and (d) m, the number of EntrezGenes in each GO term found in each module.
4. Assign KEGG pathway map names to the converted (Entrez-) genes. Obtain the four parameters for each pathway map in the same way as the GO terms.
5. Evaluate each GO term and KEGG pathway map names for each gene module with standard hypergeometric distribution statistics by giving m as positives out of k samples and M known positives among the total population of U. To critically assess p value threshold, shuffle gene names (UniGene IDs) in the input dataset and do the same statistical analysis for each module, and use them as a null-distribution model.
6. The numbers of extracted modules for 298 query chemicals under two different discretization thresholds (3% and 5%) are summarized in **Table 1**. Two different noise ratios in the module merging process are also tested in both data.

2.7. Analysis of Common Reactions Among Chemicals by Gene Modules

The main objective in gene module analysis of reverse chemical genomics is to find new functional relationships between different chemicals. Once a set of gene modules annotated by GO and KEGG is obtained, check the common function names for every combination of query chemicals with a significant p value threshold.

Table 1
Numbers of obtained modules under various parameter conditions

	Noise ratio	
Discretization	0.4	0.5
3%	2,859	2,088
5%	92,100	61,852

1. First, assign *p* values to each gene module by GO and KEGG categorical enrichment analysis as described in 2.6. Choose the most significant (the smallest) *p* value among various functional candidates for a single gene module. Then, plot each module by its (−log) *p* value, as shown in Fig. 2a.
2. Plot the same module in which gene IDs are shuffled by its *p* value, as shown in Fig. 2b. Compare the two plots. Set the *p* value threshold at the critical point where raw modules are still observed but gene-shuffled modules disappear. Here, we arbitrarily choose 1e − 5.5 and 1e − 4 as the thresholds for GO and KEGG, respectively, for the modules extracted by the parameters of 5% discretization and 0.4 of noise ratio.
3. Take every pair of query chemicals and check if they share common function or pathway names at the above threshold. Store the results.
4. Find biological hypotheses that can explain the obtained relationships among two or more chemicals. For example, in our data, some modules obtained from a query of Benzbromarone are found to affect the pathway of "Biosynthesis of unsaturated fatty acids." This pathway was also found with queries of Fenbufen, Clofibrate, and Dichloroacetate. Three genes are involved in the obtained modules: acyl-CoA thioesterase 3 (Rn.11326), acyl-CoA thioesterase 7 (Rn.6024), and acyl-Coenzyme A oxidase 1, palmitoyl (Rn.31796), all of which are important in unsaturated fatty acid metabolism. This result and information from literature suggest that these chemicals could affect the rat liver in a similar manner that is related to carcinogenesis via PPARα-derived oxidative reaction. Figure 3b is an example of the graphical output of these modules produced by the web-based GUI version of the SAMURAI system.

2.8. Usage of SAMURAI System

To enhance the above analysis in a coordinated way, we have developed a gene module extraction and evaluation system called SAMURAI. The free trial version of the program coded in java/C++/Perl is available from http://samurai.cbrc.jp/download/SAMURAI-Progressive/ (*see* Note 6). Here we describe briefly the usage of the system.

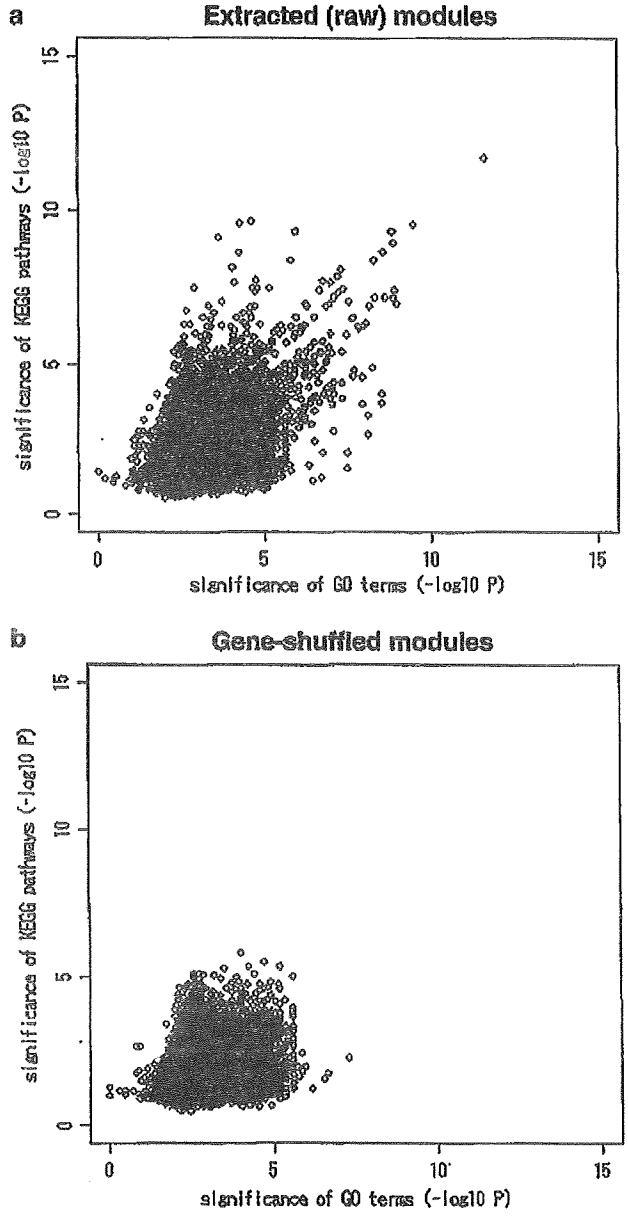

Fig. 2. Analysis of significant gene modules by GO and KEGG function groups. Each gene module (with 5% discretization threshold and 0.4 noise ratio) is evaluated with GO terms and KEGG pathway maps based on hypergeometric distribution statistics and plotted by the most significant p values. A total of 92,100 obtained gene modules are plotted in (a), and the same gene modules in which their gene ids are shuffled are plotted in (b). p values are dramatically increased due to randomness in (b). There is a significant correlation between p values of GO and KEGG in only (a). We arbitrarily choose $1e-5.5$ and $1e-4$ as the threshold for GO and KEGG, respectively.

1. Download SAMURAI program from the above site. Select SAMURAI-P program. Uncompress and untar the frozen file on your Linux machine. Go to the expanded directory and type "make compile" to compile all the programs in the system.

2. Transform your gene expression dataset into the "CellMontage" format *(14)* where a gene expression profile consists of a single description line, starting with ">," followed by one or more data lines. The data consist of elements separated by a space or a new line. Each element consists of a UniGene identifier and an expression value separated by a colon. See Fig. 3a for example data.

a

An example of partial profile in CellMontage format, where only three genes and their expression values are shown.

```
>Clofibrate_600mg_HybGrpOA2
Rn.98209:0.26226233    Rn.53257:0.550381825
Rn.94195:-0.285033381
```

A gene expression profile consists of a single description line, starting with ">", followed by one or more data lines. The data consist of elements, separated by a space or a new line. Each element consists of a UniGene identifier and an expression value separated by a colon.

b

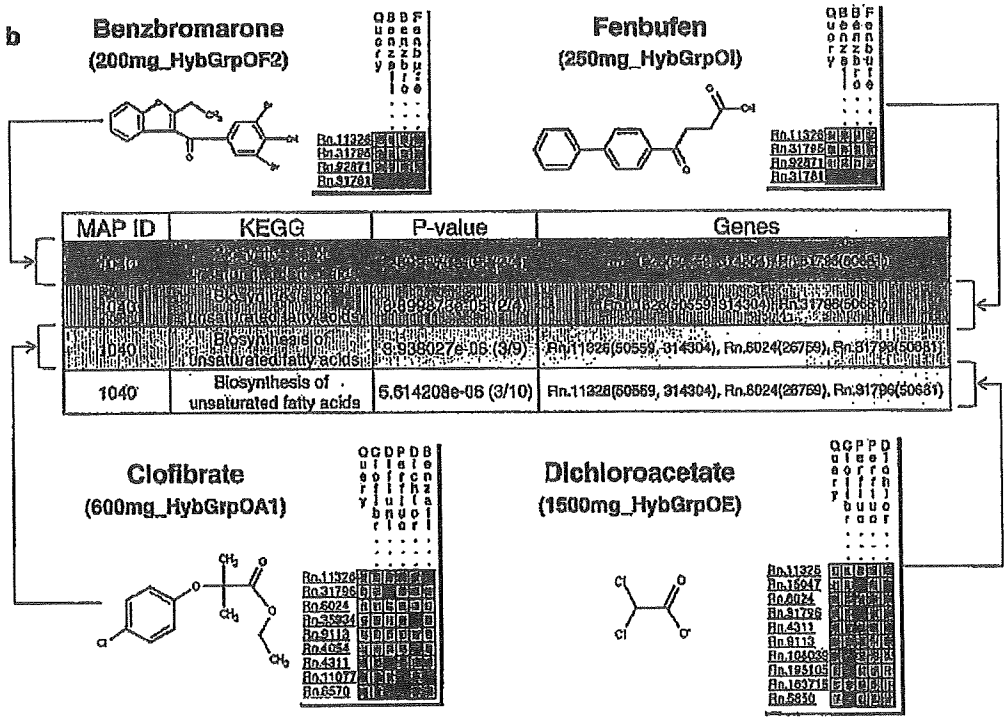

Fig. 3. Examples of input and output of SAMURAI-GUI system. An example of analysis with queries tnts (a) the CellMontage format as input and (b) graphical output of modules that share common biological function. Four example modules are searched by Benzbromarone, Fenbufen, Clofibrate, and Dichloroacetate. KEGG pathway functions for each module with the first and the second most significant *p* values are only shown.

3. Format data as described in 2.2 "Formatting Database by Discretization." Use the "formatdb" command. To discretize your data with the thresholds, as exemplified in 2.2, type "%formatdb dataset_file 0.03 (or 0.05)." This command takes a while and creates both the discretized data "dataset_file.db" and its gene index file "dataset_file.idx."
4. Run the LCM algorithm to extract core modules and merge them by typing: "% java –Xmx2000m xSamurai –M –i dataset_file.idx –d dataset_file.db –q query_file –n noise_threshold –s minimum_module –r result_dir."

   The query_file must also be written in the CellMontage format. The "noise_threshold" must be in the range of [0,1]. The "minimum_module" parameter sets the minimum size of gene modules to output. The "result_dir" parameter indicates the directory to write the results of final gene modules.
5. To perform module evaluations with GO and KEGG, execute the command: "%EA = GO_KEGG_test GO_KEGG_test/assignKEGG.pl dataset_file.idx P result_dir/*."

   The "GO_KEGG_test" is the directory for GO/KEGG evaluation package (*see* Note 7).
6. The GUI-based web version for multi-CPU calculation and module visualization in color is also available from a commercial site. To view free test results, visit: http://samurai.cbrc.jp/ and try the "Module search from a large-scale database" Web page (*again, see* Note 6).

3. Notes

1. Actually, the purpose of subtracting median values and Z-transformation in each data is only to improve visualization; they do not change the results as the following discretization process is based on rank values.
2. With an average query that represents only 10% of its discrete values are active (up- or down-regulated) genes, it is estimated that the database will be reduced to ($0.05 \times 0.05 \times 2 =$) 0.5% of its original size.
3. LCM program usually takes several seconds to finish, but sometimes takes several minutes. A faster program is currently available from the developers' Web site.
4. Many versions of merging processes are available. For example, implementing merge after finishing all the "checks" of module pairs is an option. Take the best approach depending on the situation (computational resource, purpose, etc.).

5. The conversion of genes from UniGene into EntrezGenes generates often more than single (one-to-one) correspondences.
6. The full license and web-enhanced server versions are available from HPC Solutions Inc. (http://www.hpc-sol.co.jp/).
7. Before running GO/KEGG evaluation script, do not forget to download necessary data files, including gene2unigene, gene2go.gz, and KEGG pathway maps. To do this, add your species code to "getKEGGmaps.pl" script in the "GO_KEGG_test" directory and then type "make compile" at the top directory.

Acknowledgments

We would like to thank Dr. Takeaki Uno at National Institute of Informatics for advice and kindly providing the LCM program for free use.

References

1. Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. (2005) NCBI GEO: mining millions of expression profiles – database and tools. *Nucleic Acids Res.* 33 (Database issue), D562–D566.
2. Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G.G., Oezcimen, A., Rocca-Serra, P., and Sansone, S.A. (2003) ArrayExpress – a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.* 31, 68–71.
3. Cheng, Y., and Church, G. (2000) Biclustering of expression data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 93–103.
4. Tanay, A., Sharan, R., and Shamir, R. (2002) Discovering statistically significant biclusters in gene expression data. *Bioinformatics* 18, S136–S144.
5. Ben-Dor, A., Chor, B., Karp, R., and Yakhini, Z. (2002) Discovering local structure in gene expression data: the order-preserving submatrix problem. *Proceedings of the 6th Annual International Conference on Computational Biology, ACM Press, New York, NY, USA*, 49–57.
6. Murali, T.M., and Kasif, S. (2003) Extracting conserved gene expression motifs from gene expression data. *Pac. Symp. Biocomput.* 8, 77–88.
7. Ihmels, J., Bergmann, S., and Brkai, N. (2004) Defining transcription modules using large-scale gene expression data. *Bioinformatics* 20, 1993–2003.
8. Wu, C.J., and Kasif, S. (2005) GEMS: a web server for biclustering analysis of expression data. *Nucleic Acids Res.* 33, W596–W599.
9. Prelic, A. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics* 22, 1122–1129.
10. Okada, Y., and Fujibuchi, W. (2007) Mining a Large-scale Microarray Database for Similar Gene Expression Modules to Find Distant Relationships between Down Syndrome and Huntington's Disease. *Proceedings of Critical Assessment of Microarray Data Analysis 07, Valencia, Spain.*
11. http://camda.bioinfo.cipf.es/camda07/
12. Uno, T., Asai, T., Uchida, Y., and Arimura, H, (2004) An efficient algorithm for enumerating closed patterns in transaction databases, *Lecture Notes in Artificial Intelligence* 3245, 16–31.
13. Uno, T., Kiyomi, M., and Arimura, H, (2002) LCM ver.2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets, *IEEE ICDM'04 Workshop FIMI'04* 126.
14. Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H. and Horton, P. (2007) CellMontage: Similar Expression Profile Search Server, *Bioinformatics* 23, 3103–3104.

The Journal of Toxicological Sciences (J. Toxicol. Sci.)
Vol.35, No.1, 115-123, 2010

Toxicogenomics/proteomics Report

Profiles of Chemical Effects on Cells (pCEC): a toxicogenomics database with a toxicoinformatics system for risk evaluation and toxicity prediction of environmental chemicals

Hideko Sone[1], Masahiro Okura[1], Hiroko Zaha[1], Wataru Fujibuchi[2], Takeaki Taniguchi[3], Hiromi Akanuma[1], Reiko Nagano[1], Seiichiro Ohsako[4] and Junzo Yonemoto[1]

[1]Research Center for Environmental Risk, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba, Ibaraki 305-8506, Japan
[2]Advanced Industrial Science and Technology (AIST), Computational Biology Research Center, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan
[3]Mitsubishi Research Institute, Inc., Practice Areas and Industry Sectors, 3-6 Otemachi 2-chome Chiyoda-ku, Tokyo 100-8141, Japan
[4]Graduate School and Faculty of Medicine, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

(Received September 3, 2009; Accepted October 21, 2009)

ABSTRACT — Profiles of Chemical Effects on Cells (pCEC) is a toxicogenomics database with a system of classifying chemicals that have effects on human health. This database stores and handles gene expression profiling information and categories of toxicity data. Chemicals are classified according to the specific tissues and cells they affect, the gene expression changes they induce, their toxicity and biological functions in this database system. The pCEC system also analyzes relationships between chemicals and the genes they affect in specific tissues and cells. The reason why we developed pCEC is to support decision-making within the context of environmental regulation. Especially, exposure to environmental chemicals during fetal and newborn development may result in a predisposition to various disorders such as cancer, learning disabilities and allergies later in life. The identification and prediction of hazardous chemicals using limited information are important issues in human health risk management. Therefore, various toxicity information including lethal dose 50 (LD50), toxicity pathways and pathological data were loaded into pCEC. pCEC is also a facility for query, analysis and prediction of unknown toxicochemical reaction pathways and biomarkers which are based on toxicoinformatical data mining approaches. This database is available online at http://project.nies.go.jp/eCA/cgi-bin/index.cgi. The current version of the database has information on the hepatotoxicity, reproductive toxicity and embryotoxicity of chemicals.

Key words: Toxicogenomics, Toxicoinformatics, Risk assessment, Toxicity prediction, Chemical profiling

INTRODUCTION

Risk assessment for human health is now standing in front of a new door of toxicity testing in the 21st century. Many scientists at universities and regulatory agents have proposed the need for a paradigm-shift from the old-fashioned-style toxicity test using many animals and high doses of chemicals (Andersen and Krewski, 2009; Bushnell *et al.*, 2007; Fentem *et al.*, 2004). Exposure to environmental chemicals during fetal and newborn development may result in a predisposition to various disorders such as cancer, learning disabilities and allergies later in life. The identification and prediction of these chemicals using limited information are important issues in human

Correspondence: Hideko Sone (E-mail: hsone@nies.go.jp)

health risk management (Krewski *et al.*, 2009; Woodruff *et al.*, 2008). The National Research Council (NRC) of the National Academies released a report entitled, "Toxicity testing in the 21st Century" in 2007 (NRC, 2007). The central components of their vision are toxicity pathway and targeted testing, and the main components of their vision are collection and computational modeling of physical and chemical properties, environmental concentrations and stability, routes of human exposure, the potential for bioaccumulation, metabolites and molecular interactions (Kavlock *et al.*, 2008). The United States Environmental Protection Agency (U.S. EPA) has already developed and publicly opened access to Aggregated Computational Toxicology Resource (ACToR), which is a database holding essentially all publicly available information on the identity, structure, physical-chemical properties, *in vitro* assay results, and *in vivo* toxicology data of chemicals (Judson *et al.*, 2008, 2009). In Europe, the European Union established a new regulation concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH) (Ahlers *et al.*, 2008), in which all relevant industrial chemicals must now be assessed by industries themselves and the industries are responsible for risk assessment. REACH sets certain minimum data requirements in order to achieve a high level of protection for human health and the environment. All available data from the different steps have to be integrated to come to an overall conclusion on the toxicity of the chemical (Mattingly, 2009; Maxwell *et al.*, 2008). In this background, SuperToxic (Schmidt *et al.*, 2009) and ToxRTool (Schneider *et al.*, 2009) from Germany and MMSINC (Masciocchi *et al.*, 2009) from Italy were released. SuperToxic is a database that compiles a wide range of compounds with large amounts of bioassay data, and ToxRTool is a tool that assesses the reliability of toxicological data. MMSINC is a chemical structure database, where data on chemicals are appropriately stored and annotated. Thus, databases that collected chemical properties and toxicological data are being used to forecast the unknown toxicity of unstudied chemicals. In order to predict the unknown toxicity of chemicals, categorization according to toxicity is an important step after completing data collection, as is making various templates of toxic type or toxicity pathways (Wullenweber *et al.*, 2008). The reliance on toxicity pathway perturbations for human health risk assessment will require sufficient understanding of such pathways to explain phenotypic outcomes in animals. The successful prediction of chemical toxicity will require the development of three areas: a comprehensive suite of toxicity pathway assays, analytical tools for the data, and regulatory and political infrastructure enabling their use in health risk assessment (Mattingly, 2009, Maxwell *et al.*, 2008). To solve these problems, genomic analysis such as microarray data analysis is one strong approach (Wang, 2008; Waters and Jackson, 2008).

Toxicogenomics has been widely used for elucidating the molecular and cellular actions of chemicals and other environmental stressors on biological systems (Guyton *et al.*, 2009; Waters and Fostel, 2004; Waters *et al.*, 2003). Classification of known or new toxicants based on signatures of gene expression will be a basis for predicting toxicity before any potential functional damage (Aardema and MacGregor, 2002; Benigni *et al.*, 2007; Lambert *et al.*, 2009). Therefore, we developed a database accompanied by software for classification of known or new toxic chemicals based on signatures of gene expression with other toxicology data. The system was named the Profiles of Chemical Effects on Cells (pCEC) system. pCEC shows classifications of chemicals that act on particular cells from the viewpoint of gene expression signatures. pCEC also stores and handles gene expression profiling information and categorizations of bioassay data to elucidate toxicity. The compiled data in pCEC have been organized into a variety of chemical groups that are classified according to the type of molecular pathway or type of toxicity (Fig. 1). In comparison with other databases (Supplementary Table 1), intuitive visualization of gene expression information using clustering techniques such as a self-organizing map and minimum-spanning tree in pCEC is very unique and allows us to make toxicity predictions and find biomarkers more easily. Thus, this is a unique database containing information on the health effects of chemicals combined with gene alteration profiles in specific tissues and cell types. This is directly useful for risk evaluation and assessment of the effect of chemicals on human health. All of the data and search functions of pCEC are accessible through a user-friendly web interface that we describe later in this paper.

MATERIALS AND METHODS

pCEC was established on the basis of data from publicly available resources at the websites, Gene Expression Omnibus (GEO) database (Barrett *et al.*, 2005, 2007) in National Center for Biotechnology Information (NCBI), The Distributed Structure-Searchable Toxicity (DSSTox) (Williams-Devane *et al.*, 2009b; Williams-DeVane *et al.*, 2009a), the Chemical Effects in Biological Systems (CEBS) database in the National Institute of Environmental Health Sciences (NIEHS) (Fostel, 2008; Waters *et al.*, 2008), NCI60 project in the National Cancer Institute (Shoemaker, 2006) and the Toxicology data network

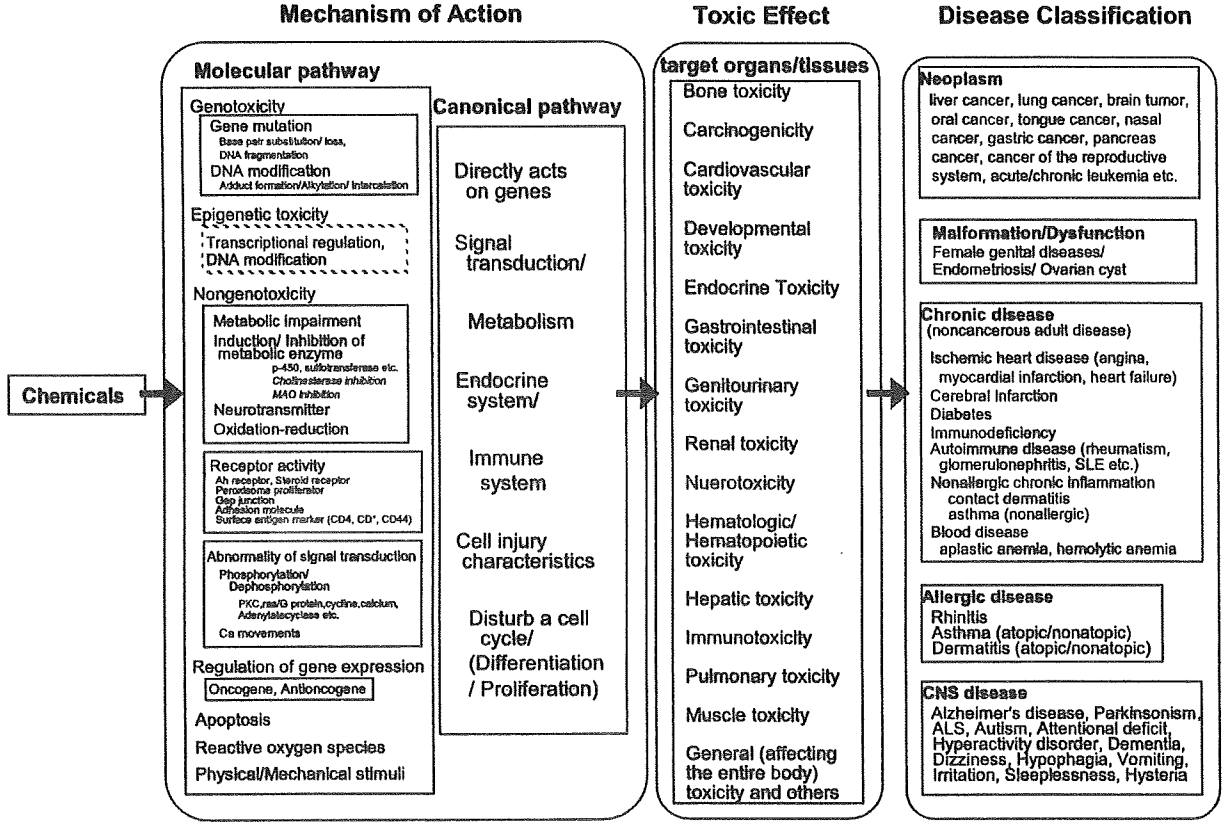

Fig. 1. Categorization of chemicals by their toxicity and molecular mechanism of action. The flow of pictures represents the molecular pathway, toxicity types and disease classification of a chemical when exposure to the chemical affects cells, tissues and organs. This flow influences the structure of pCEC. Neurotoxicity is subclassified as CNS (central nervous system including brain and spinal cord), PNS (peripheral nervous system including motor nerves and sensory nerves) and ANS (autonomic nervous system including sympathetic nerve, parasympathetic nerve).

(TOXNET) database (Tomasulo, 2005). By exploring comprehensively public datasets among these websites, we obtained three categories datasets. In other words, we could not find a battery of datasets about the other target organs such as kidneys and hearts. Three categories were datasets for the liver toxicity, the neuronal toxicity and the reproductive toxicity. To analyze expression values from large-scale gene expression data, we first normalize and convert them into a unified file format. Gene expression data normalized by Z score (*) transformation can be used directly in the calculation of significant changes in gene expression between different samples and experimental conditions. After obtaining formatted data, gene expression patterns for each chemical were used to create a heat map on a matrix using a self-organizing maps (SOM). The expression patterns extracted by the SOM were used to represent graphically Z-score values of the downregulated and upregulated genes. The heat-maps as an expression pattern of chemicals are then classified using a minimum spanning tree (MST) algorithm. For example, in "2007_rat_liver (102 chemicals, 2,488 genes)", data on the effects of chemicals on gene expression in the rat liver were downloaded from the website http://cebs.niehs.nih.gov/. There were a total of 133 chemical groups containing 964 microarray experiments. In the present study, 298 arrays among 9,215 probes in the dataset for both chemical exposure and control data were collected. After deleting low-abundance genes that showed no expression in all 298 experiments, a total of 7,614 probes were selected. Every pair of gene expression values was transformed into log-fold-change abundance by subtracting control values from the chemical exposure after taking log 2 and

subtracting the median value in each array. The log-fold-change values are normalized to Z-score by (x - mean) / SD. The gene expression data in the other projects were also normalized and converted into a unified file format.

The Z-score for gene expression values in cells can be approximated by the logarithm normal distribution with Zipf's law.

$Z = (x- \mu)/ \delta$

Where x is the gene expression value, μ is the mean expression value for all genes, and δ is the standard deviation of all genes.

Primary resources

The original microarray data sets used by this database were as follows, 2007_rat_liver: CEBS Accession 004-00002-0010-000-7; 2008_mouse_neuro: GSE367, GSE587, GSE1076, GSE1077, GSE1588, GSE1800, GSE3253, GSE3412, GSE5763; 2008_mouse_repro: GSE280, GSE438, GSE499, GSE3348, GSE4650; 2008_mouse_embryostem: GSE18503.

Supported web browsers

The following web browsers are fully supported, i.e., all of the features of pCEC including JavaScript and CSS styles should work properly: Internet Explorer 7, Firefox 3, Safari 3.

RESULTS AND DISCUSSION

pCEC (http://project.nies.go.jp/eCA/cgi-bin/index.cgi) has been developed as a database with a system of classifying chemicals that affect cells and induce gene expression changes, according to their toxicity and biological functions (Figs. 1 and 2). This database stores and handles gene expression profiling information and categorizes toxicity data. The system can separate chemicals into a variety of groups by the type of influence. Gene expression profiling was achieved by the SOM technique (Luo *et al.*, 2004; Törönen *et al.*, 1999), and the toxicity data are shown using MST algorithms (Xu *et al.*, 2001, 2002). The projects categorize chemicals according to the types of toxicity. The component in each option of pCEC reflects the framework of categorization of chemicals as shown in Fig. 1. All projects are tagged at the holder header by year, animal, and cell type, such as "2007 rat liver". The present latest version includes the following projects: 2007_rat_liver (102 chemicals, 2,488 genes); 2008_mouse_neuro (7 chemicals, 974 genes); 2008_mouse_repro (4 chemicals, 661 genes); 2008_mouse_embryostem (12 chemicals, 17,042 genes). The primary data can be linked to "the primary resource" in the top page of each project holder. pCEC has 4 tool boxes which are described below (Figs. 3, 4 and 5).

1) Chemicals. This section contains physiological and toxicity information on chemicals. The system categorizes data on chemicals that induce toxicity in animals, according to their mechanism of action, toxicity and structure (Fig. 3A).

2) Chemical Expression Neighbor. The user can compare the gene expression signature, which indicates gene expression changes affected by chemicals between different samples and experimental conditions (Fig. 3B). The gene expression signature of a chemical presents graphically genes whose expression is upregulated or downregulated with a SOM. The entire SOM signature for chemicals in each project is classified by the Minimum spading tree technique.

3) Correspondence Analysis (CA). CA is one of the multivariate analyses, a statistical visualization method of picturing the associations between the levels of a two-way contingency table. In this analysis, some measure of correspondence between the rows and columns of each datum value are mapped. The CA viewer of pCEC can display the results by plotting them in two-dimensional space (Fig. 3C).

4) Chemical Selector. The Chemical Selector allows the user to search for information on chemicals including physical and toxicity information according to information in the toxicity pathway. The Chemical Selector in pCEC provides three levels of categorization according to toxicity pathways, target organ toxicity and disease as shown in Fig. 1. For example, level 1 and level 2 represent classification of toxicity mechanisms and pathways according to the concept of categorization of toxicity at the molecular, cellular and tissue levels. In level 3, chemicals are listed according to the categories that the user had selected in levels 1 and 2 (Fig. 4).

5) Specific search options. If users selected one chemical in the site of "Chemical", there are many options to find the similar expression profiles and particular gene expression patterns (Fig. 5). As an example, Fig. 5 shows screenshots of toxicogenomic information for the chemical, antimycin A3 (Fig. 5A). Fig. 5B displayed networks for gene expression signatures of each chemical by using the shortest path problem algorism, which is the problem of finding a path between two nodes such that the sum of the weights of its constituent edges is minimized. Up and down regulated genes and distribution of expression values were listed in Fig. 5C. pCEC also directly link to SAMURAI, which is another program to find a module, and which is a minimal unit of common responsive genes affected by exposure to chemicals (Okada *et al.*, 2007).

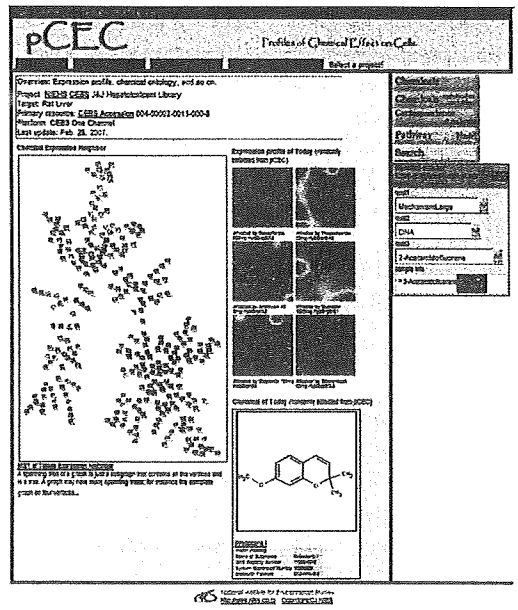

Fig. 2. Top page of Browser's window. Heat maps of the Gene expression patterns of chemicals are gathered by similarity of the map patterns.

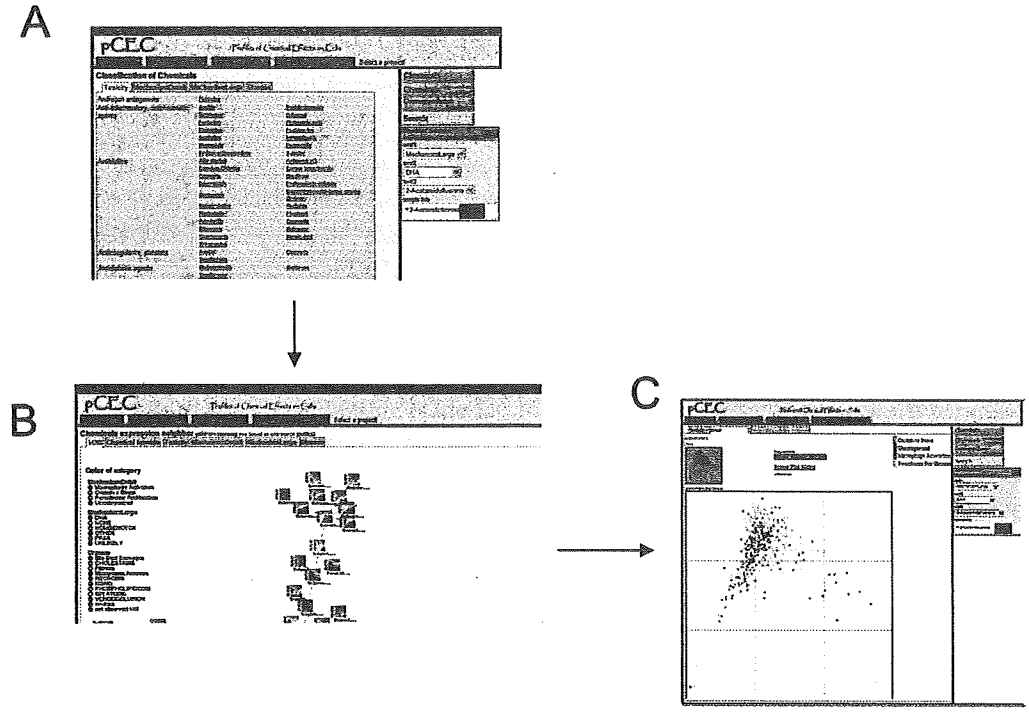

Fig. 3. "Chemicals" (A), "Chemicals Expression neighbor" (B), and "CA" (C). After clicking each button at the top-right corner, the respective page appears.

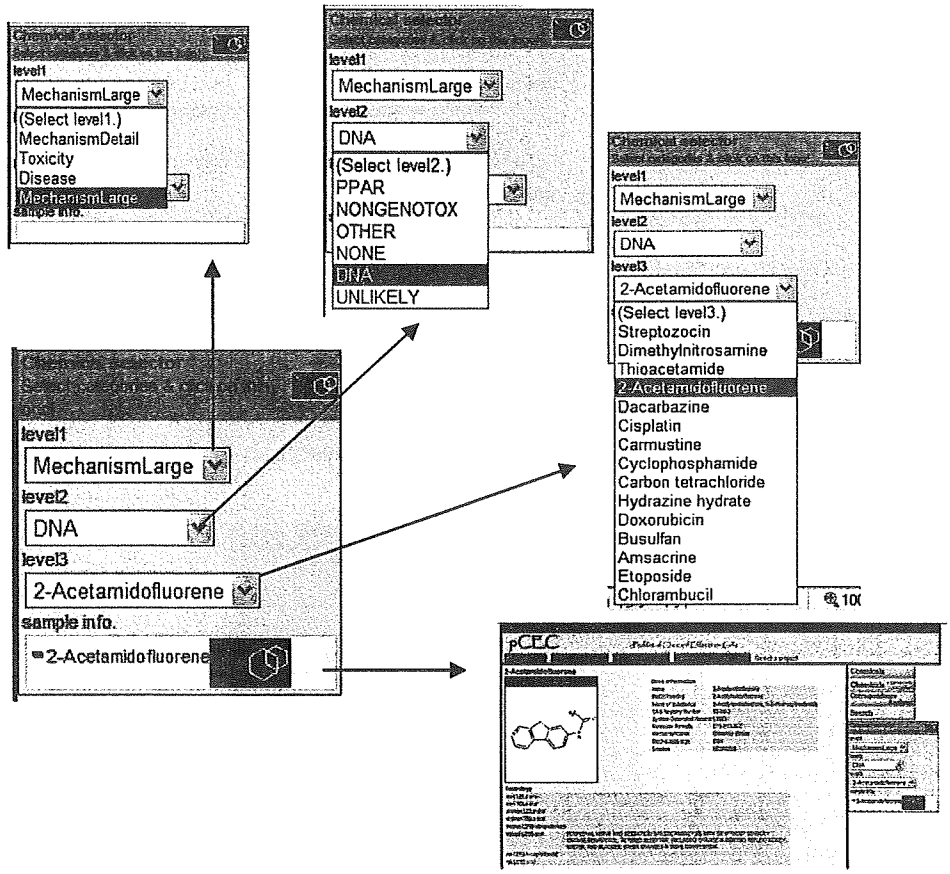

Fig. 4. Chemical selector this query option is categorized by the chemical at three levels. Level 1 indicates categories of chemical effects, which are the process during apparent toxic effects and disease outcomes from the first target on the cellular event as shown in Fig. 1. Level 2 indicates more detail categories than those of level 1 for mechanism of action and toxicity pathways. Level 3 indicates chemical categories based on the mother chemical structure.

The users can find specific molecular marker of chemical exposure from large-scale gene expression data using the SAMURAI program (Fig. 5D).

In a future plan, the pCEC database would be created to support decision-making within the context of environmental regulation, especially human health. Our research goals are to: (1) classify the different types of toxicity and the mechanism of action of environmental chemicals using analysis of gene expression induced by exposure to chemicals, as well as toxicological data; (2) develop analysis systems that categorize multiple profiling based on multidimensional information on the effects of chemicals on rodents and human health, including analysis of the relationship between toxicity data and related diseases; (3) create a high-quality categorized index for diverse chemical-based databases and lists of chemicals that are important for environmental regulation; (4) construct Health Effects Alert System (HEALS) that includes other systems to evaluate chemical effects using a variety of databases and algorithms. Further development of structure-activity and structure-toxicity databases as well as toxicity molecular endpoint computerized libraries is required.