Genomics & Informatics Vol. 4(1) 40-44, March 2006

XPERNATO-TOX: an Integrated Toxicogenomics Knowledgebase

Jung Hoon Woo[1], Hyeouneui Kim[2], Gu Kong[3] and Ju Han Kim[1,4]*

[1]Seoul National University Biomedical Informatics (SNUBI) and [4]Human Genome Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea
[2]Graduate Program in Health Informatics, University of Minnesota, Minneapolis, MN 55455, USA
[3]Department of Pathology, College of Medicine and Molecular Biomarker Research Center, Hanyang University, Seoul 133-791, Korea

Summary

Toxicogenomics combines transcriptome, proteome and metabolome profiling with conventional toxicology to investigate the interaction between biological molecules and toxicant or environmental stress in disease caution. Toxicogenomics faces the problems of comparison and integration across different sources of data. Cause of unusual characteristics of toxicogenomic data, researcher should be assisted by data analysis and annotation for getting meaningful information. There are already existing repositories which claim to stand for toxicogenomics database. However, those just contain limited abilities for toxicogenomic research. For supporting toxicologist who comes up against toxicogenomic data flood, now we propose novel toxicogenomics knowledgebase system, XPERANTO-TOX. XPERANTO-TOX is an integrated system for toxicogenomic data management and analysis. It is composed of three distinct but closely connected parts. Firstly, Data Storage System is for reposit many kinds of '-omics' data and conventional toxicology data. Secondly, Data Analysis System consists of analytical modules for integrated toxicogenomics data. At last, Data Annotation System is for giving extensive insight of data to researcher.

Keywords: toxicogenomics, data comparison, integration, analysis, annotation, knowledgebase

Introduction

Toxicology is study to understand the relationship between toxicants, and human disease susceptibility. A critical part of this study is the characterization of the adverse effects at the level of the organism, the tissue, the cell, and the molecular makeup of the cell. Thus, studies in toxicology measure effects on body weight and food consumption of an organism, on individual organ weights, on microscopic histopathology of tissues, and on cell viability, necrosis, and apoptosis (Waters *et al.*, 2004).

Recently, cause of the appearance of new '-omics' technology, toxicologist could get the extensive information of molecular level (Hamadeh *et al.*, 2003). That is, 'new -omics technology', such as transcriptomics, proteomics and metabolomics have been developed, therefore toxicologist can measure thousands of movement of transcriptome or proteome (Aardema *et al.*, 2002). Obviously, variations at cellular and organism level are caused by critical changes of mRNAs or proteins. So the application of new technologies to conventional toxicology gives insight into 'cause and effect chain' of organism.

As mentioned of changes of toxicology, the new term Toxicogenomics has been evolved. Toxicogenomics could be demonstrated of integration of conventional

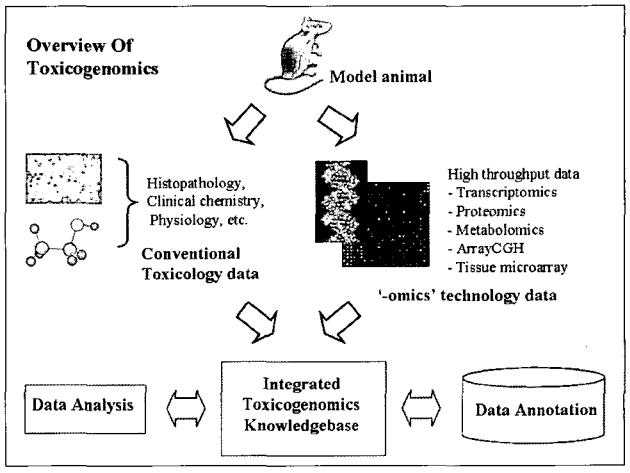

Fig. 1. The overview of toxicogenomics
From model animal, data flow of clockwise rotation indicates data gathering with '-new omics' technology. Opposite flow indicates data gathering in conventional toxicology. Data Analysis represented on bottom left of the figure reduce numerous but unfocused genomic data into significant one. Data Annotation represented on bottom right. Analysts easily get insight from toxicogenomic data which are just an array of numbers and tiny indices. Arrows mean the data flow.

*Corresponding author: E-mail juhan@snu.ac.kr,
Tel +82-2-740-8320, Fax +82-2-742-5947
Accepted 3 March 2006

Table 1. Existing toxicogenomics repositories.

Database	'-omics' data (transcriptomics)	Conventional data (histopathology)	Analysis System	Other knowledge For annotation
ArrayTrack	Included	Not included	Included (low level)	Gene library, toxicant library
CEBS	Included	Not included	Included (low level)	Not included
TOX · MAIMExpress	Included	Not included	Not included	Not included
CTD	Included	Not included	Not included	Gene-Chemical Cross reference
dbZach	Included	Not included	Not included	Included
EDGE	Included	Not included	Not included	Not included

toxicology and '-omics' technologies. To study of combining genomic changes and biological endpoints, collection of heterogeneous data has become challenge of toxicogenomics (Mattes *et al.*, 2004).

Fig. 1 shows the overview of current toxicogenomics. From model animal, data flow of clockwise rotation indicates data gathering with '-new omics' technology. There are proteomics, transcriptomics, metabolomics, arrayCGH, and tissue microarray which generate high-throughput expression data (Waters *et al.*, 2003). Opposite flow indicates data gathering in conventional toxicology. There could be histopathology, clinical chemistry, physiology and so on (Laura *et al.*, 2004). Those two processes yield heterogeneous data that cannot be combined easily, so well-designed knowledgebase which is able to store such types of data has been needed. Data Analysis represented on bottom left of the Fig. 1 reduce enormous but unfocused genomic data to significant one by several statistical and computational modules. Then analyzed data become small enough to handle. But in general, it is still too huge to be interpreted. Data Annotation is also important for that reason. Analysts easily get meaningful information from toxicogenomic data composed of just an array of numbers and tiny indices.

Currently, there are several repositories which claim to stand for 'Toxicogenomics database (or knowledgebase)' (Mattes *et al.*, 2004). However, they are still not enough to stand for toxicogenomics knowledgebase (Table 1). Most listed databases are not able to include conventional toxicology. Even though there are databases have analysis system, they perform only lower level analysis just as 'simple t-test'. For getting enough insight from high-throughput data, it should be able to perform higher level analysis. In addition, extra data for annotate primary one still insufficient in most cases. Therefore the actual circumstances, toxicologist cannot gather enough resources from existing repositories.

In this paper, we propose a novel system 'XPERANTO-TOX' to complement limitation of existing toxicogenomics repositories.

Methods

Design 'Data Storage System' using pre-existing data models

MIAME-TOX (http://www.mged.org) and TMA-OM (Lee *et al.*, 2005) were used as data standard for storage system. MIAME-TOX is an extended version of MIAME (Minimum Information About an Microarray Experiment). That is, MIAME-TOX is a guideline defining the minimum information required to interpret unambiguously and potentially reproduce and verify array-based toxicogenomic experiments. TMA-OM (Tissue Microarray - Object Model) is a data model for capturing tissue microarray experimental data and representing clinical and histopathology information of tissues. Otherwise there are no gold standard for conventional toxicology data, such as histopathology image and description. So we referenced sample description part of two mentioned data model.

Construct 'Data Analysis System' with statistical modules

There are several steps for getting meaningful information from high-dimensional data. Data analysis system in XPERANTO-TOX has been developed based on these steps. Statistical language R (current version R 2.2) was used for materializing most modules in analysis system. VSN (variance stabilizing transformation) (Huber *et al.*, 2002), KNN (K Nearest Neighbor), cyclic lowess, global lowess and several baseline normalization algorithms (Bolstard *et al.*, 2003) were used for data preprocessing step. Significant Analysis for Microarrays (Virginia *et al.*, 2001), cyberT (Baldi *et al.*, 2001), DEDS (Differential Expression via Distance Summary of Multiple Statistics Description) (Yee *et al.*, 2005), ANOVA (Analysis Of Variance), bayesANOVA (Baldi *et al.*, 2001), GSEA (Gene Set Enrichment Study) (Aravind *et al.*, 2005) and other algorithms were used for significant data analysis step. Hierarchical, K-means, SOM (Self Organizing Map), PCA (Principle Component Analysis), ICA (Independent Component Analysis) (Samir *et al.*,

2004) algorithms were used for clustering data elements. At last, SVM (Support Vector Machine) (Brown *et al.*, 2000) and PAM (Prediction Analysis of Microarray) (Robert *et al.*, 2002) were used for classification.

Localize public databases for organizing 'Data Annotation System'

We localized GenBank (http://www.ncbi.nlm.nih.gov/Genbank/index.html), UniGene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene), Entrez Gene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene), SWISSPROT (http://www.expasy.org/sprot/) databases for composing 'Gene' part in annotation system. CTD (Comparative Toxicogenomic Database) (Carolyn *et al.*, 2003), ArrayTrack (Wieda *et al.*, 2003) were localized for 'Toxicant' part. 'Disease' part was composed of OMIM (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM), MeSH term data. For connecting each part, we referenced CTD for gene-toxicant, PathMeSH for gene-disease and CHE (http://database.healthandenvironment.org/) for toxicant-disease connection.

Results

XPERANTO-TOX is developed as a large system composed of closely connected three sub systems (Fig. 2). First is data storage system. It could be able to store histopathology data (image, description) and array based high throughput data. Second is higher level data analysis system. It directly connected with data storage system, so that analyst easily get enough information explained high dimensional '-omics' data. Last one is data annotation system which is connected with both storage and analysis system. Therefore if any researcher wants to get biological interpretations for stored or analyzed data, annotation system could be quite useful.

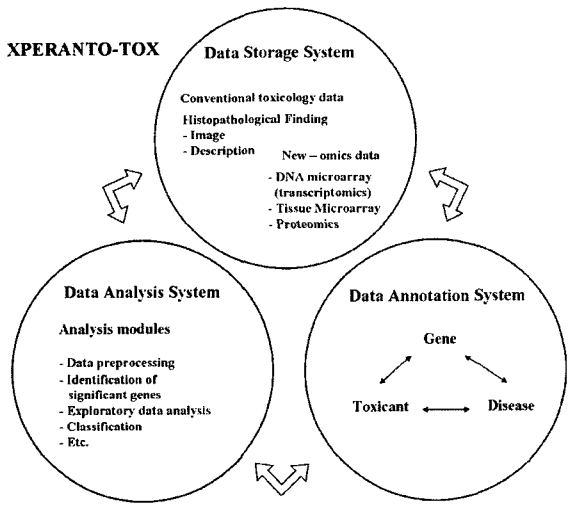

Fig. 2. The composition of whole system, XPERNATO-TOX XPERANTO-TOX is composed of three subsystems. They are 'Data Storage System', 'Data Analysis System' and 'Data Annotation System'. Many kinds of heterogeneous generated from toxicogenomic data stream could be stored 'Data Storage System'. Stored data could be effectively analyzed by 'Data Analysis System'. Data within XPRANTO-TOX could be explained or annotated by 'Data Annotation System'. Arrows between systems mean 'reference'

Data Storage System

Researcher could reposit heterogeneous data generated from various types of experiments. The system is designed with MIAME-TOX for storing data which is produced by array-based experiments such as DNA microarray, arrayCGH and so on. Currently, toxicogenomic research has included 'tissue microarray' experiment for efficient confirmation for '-omics' data. For instance, the differentially expressed gene sets derived by microarray data analysis could be easily certified by tissue microarray.

Conventional toxicology data would be included in database. For instance histopathology image and description could be stored in XPERANTO-TOX with own data model. Finally, XPERANTO-TOX could store heterogeneous array based high-throughput data, and conventional toxicology data. In addition, this repository would serve as a resource for discovery expression patterns of distinct molecules and comparison between the patterns. Simple and advanced query forms will be available to retrieve information about molecular profiling, including nucleotide and protein sequences as well as copy numbers of DNA fragments and metabolites.

Data Analysis System

Researchers have applied tens of classical and modified statistical modules for analyzing high-dimensional toxicog enomics data. No one knows about exact population of biological variables. Therefore several algorithms still have been developed in this area. We determined considerably generalized process for analyzing DNA microarray data, for instance.

Basically, goal of DNA microarray experiment (-explain thousands of transcripts' changes with specific condition) is finding differentially expressed gene sets. First step for this goal is data preprocessing through filtering, transformation, imputation and normalization. Input data should be filtered by determined standard (-generally commercial arrays have recommended filtering standard) or by flexible decision. Filtered data could be transformed by log2 or vsn module. Missing values generated by

previous steps are estimated with KNN imputation module. For normalization, cyclic lowess, global lowess, quantile and baseline normalization modules are available. After that, scoring statistics would be applied. Because selecting differentially expressed genes through distinct conditions is most significant parts among the whole process, numerous algorithms have been emerged. We materialized SAM, cyberT, simple t-test, DEDS, ANOVA, bayesANOVA, and so on as statistical scoring modules. In most cases, selected genes are still too many to interpret individually. For that reason, clustering has become essential in this field. Hierarchical, K-means, SOM, PCA, and ICA algorithms are available for clustering in this system. In addition, SVM and PAM also could be implemented for classification. Gene has become more important as a classifier for prognosis in medical fields, therefore analysis system in XPERANTO-TOX would be very useful for classification analysis with DNA microarray data.

Actually, there is no gold standard for analysis. Each algorithm has been shown different performance in different experiment. So we designed flexible system for extension. If new algorithm would be emerged, XPERANTO-TOX could easily attach new analysis modules.

Data Annotation System

Such as histopathology data from conventional toxicology could be directly interpreted. Since the data is just a single image or description, researchers can recognize what it is and what it means. On the other hands, toxicogenomic data possess the characteristics of high dimension so, without annotation for stored or analyzed data, analyst would be in trouble with data flood. For example, general DNA microarray data matrix consists of myriads row and several tens of columns and the data matrix is just filled with numbers. Without an assistant, toxicologist fall into confusion with those kinds of data, so there should be suitable annotation system to make the data interpretably.

There are three data types in annotation system (Fig. 2). They are gene, toxicant and disease. Those data types are cross-referenced. Researcher can query about affected genes by specific toxicant or disease. Query for opposite direction also be possible. When researchers get differentially expressed genes by specific toxicant, they could be confirmed with above mentioned queries. Data in annotation system is kind of knowledge generated through bunch of experiments in wet laboratory. So information is definitely limited. Researchers could get inspiration about other effects of certain toxicant and function of unknown genes.

Discussion

Now toxicogenomics have the power and potentiality to revolutionize conventional toxicology. Several toxicologists begin to apply toxicogenomic approach to understand the relationship between environmental stress and human disease susceptibility; to identify useful biomarkers of disease and exposure to toxic substances; and to elucidate the molecular mechanisms of toxicity. In the vortex of this paradigm shifts, 'Toxicogenomic knowledgebase' is getting more important. To store, manage, analyze, and annotate toxicogenomic data, we recommend the novel system, XPERANTO-TOX. It has two significant differences as against the existing databases. First, XPERANTO-TOX could store not only different '-omics' data, but also conventional toxicology data. Second, 'Data Analysis and Annotation System' is included in XPERANTO-TOX system, so analyst could easily get an insight into toxicogenomic data. The whole interface is not enough to serve, but we are now performing pilot test. Right after that, toxicologist could fully use XPERANTO-TOX.

Acknowledgements

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

References

Michael, D.W. and Jennifer, M.F. (2004). TOXICOGE NOMICS AND SYSTEMS TOXICOLOGY: AIMS AND PROSPECTS, *Nature Genet.* 5, 936-948.

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.

Aardema, M.J. and MacGregor, J.T. (2002). Toxicology and genetic toxicology in the new era of 'toxicogenomics': impact of '-omics' technologies. *Mutat. Res.* 499, 13-25.

Mattes, W.B., Pettit, S.D., Sansone, S.A., Bushel, P.R., and Waters, M.D. (2004). Database development in toxicog enomics: issues and efforts. *Environ. Health Perspect.* 112, 495-505.

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

Michael, D.W., Kenneth, O., and Raymond, W.T. (2003). Toxicogenomic approach for assessing toxicant-related disease. *Mutation Research* 544, 415-424.

Lee, H.W., Park, Y.R., Sim, J.H., Park, R.W., Kim, W.H.,

and Kim, J.H. The Tissue Microarray Object Model: a data model for storage, analysis and exchange of tissue microarray experimental data, *Archives of Pathology and Laboratory Medicine*, accepted

Bolstad, B.M., Irizarry, R.A., Astrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185-193.

Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 1, 1-9.

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* 98, 5116-5121.

Baldi, P. and Long, A.D. (2001). A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized t-Test and Statistical Inferences of Gene Changes. *Bioinformatics* 17, 509-519.

Yang, Y.H., Xiao, Y., and Segal, M.R. (2005). Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics* 21, 1084-1093.

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. USA* 102, 15545-15550.

Saidi, S.A., Holland, C.M., Kreil, D.P., MacKay, D.J., Charnock-Jones, D.S., Print, C.G., and Smith, S.K. (2004). Independent component analysis of microarray data in the study of endometrial cancer. *Oncogene* 23, 6677-6683

Brown, M.P., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M. and Jr Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* 97, 262-267.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA* 99, 6567-6572.

Benson, D.A., Boguski, M., Lipman, D.J., and Ostell, J. (1994). GenBank. *Nucleic Acids Res.* 22, 3441-3444.

Mattingly, C.J., Colby, G.T., Rosenstein, M.C., Forrest, J.N.Jr, Boyer, J.L. (2003). The Comparative Toxicogeno mics Database (CTD). *Environmental Health Perspectives*. 111, 6.

Tong, W., Cao, X., Harris, S., Sun, H., Fang, H., Fuscoe, J., Harris, A., Hong, H., Xie, Q., Perkins, R., Shi, L., and Casciano, D. (2003). ArrayTrack-supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. *Environ Health Perspect.* 111, 1819-1826.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2001). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30, 52-55.