MOLECULAR & CELLULAR TOXICOLOGY, Vol. 1, No. 1, 13-16, March 2005

Toxicoinformatics: The Master Key for Toxicogenomics

Wan Sun Lee[1] **& Yang Seok Kim**[1]

[1]Istech21, 847, ᴊanghang-Dong Ilsan-Gu, Goyang,
KungKi-Do, Korea
Correspondence and requests for materials should be addressed to Y.-S. Kim (ysk m@istech21.com)

Accepted March 5, 2005

Abstract

The current vision of toxicogenomics is the development of methods or platforms to predict toxicity of uncharacterized chemicals by using '-omics' information in pre-clinical stage. Because each chemical has different ADME (absorption, distribution, mechanism, excretion) and experimental animals have lots of variation, precise prediction of chemical's toxicity based on '-omics' information and toxicity data of known chemicals is very difficult problem. So, the importance of bioinformatics is more emphasized on toxicogenomics than other functional genomics studies because these problems can not be solved only with experiments. Thus, toxicoinformatics covers all information-based analytical methods from gene expression (bioinformatics) to chemical structures (cheminformatics) and it also deals with the integration of wide range of experimental data for further extensive analyses. In this review, the overall strategy to toxicoinformatics is discussed.

Keywords: omics, bioinformatics, genomics

In current '-omics' study, bioinformatics is an essential tool for data manipulation, analysis and interpretation and its close connection to experimental biology is a common trend of current biology[1,2]. Therefore, the expectation that bioinformatics plays an important role in toxicogenomics as in many other functional genomics is very natural idea. However, in the viewpoint of bioinformatics, toxicogenomics is more sophisticated approach than other '-omics' studies and has many unsolved problems in current bioinformatics technology. The main reason is due to the characteristics of toxicogenomics. The current vision of toxicogenomics is the development of methods or platforms to predict toxicity of uncharacterized chemicals by using '-omics' information in pre-clinical stage[3,4,5,6]. Because each chemical has different ADME (absorption, distribution, mechanism, excretion) and experimental animals have lots of variation, precise prediction of chemical' s toxicity based on '-omics' information and toxicity data of known chemicals is very difficult problem. So, the importance of bioinformatics is more emphasized on toxicogenomics than other functional genomics studies because these problems can not be solved only with experiments.

Toxicoinformatics covers all information-based analytical methods from gene expression (bioinformatics) to chemical structures (cheminformatics) and it also deals with the integration of wide range of experimental data for further extensive analyses. Among them, the informatics for toxicotranscriptomics is a main issue in current toxicoinformatics research because microarray technology gives us a lot of gene expression information which may contain key information useful for toxicity prediction.

The Experimental Design

The first issue of toxicotranscriptomics is how to properly design experiments to get crucial information from expression profiles. It involves decisions of time points, the number of experimental animals to get credible results, the dosages of chemical, and the platform of arrays. The most difficult part of experimental design is that, in most case, the final goal of research is not just understanding the toxicity mechanism of given chemicals but predicting toxicity of new chemicals based on expression profile of given chemicals. Because each chemical have different ADME properties and the expression profiles of each gene change by time and dose even though they have similar toxicity, it is really difficult to get standard experimental condition for toxicogenomics. Theoretically, experimental design with many different time points, different doses, and repeated experiments should be helpful for the toxicity detection of many different chemicals. Due to the high costs of toxicogenomics researches, every researcher faces the problem of efficient design for given purpose.

Thus, the investigation of publicly available data is essential before actual experiment. The pre-investigation of the genes related to given chemicals can be generally performed by Pubmed keyword search. Since Pubmed deals with huge amounts of published papers, however, searched results tend to contain many false-positive results and manual investigation of searched results is another time-consuming process. CTD (The comprehensive Toxicogenomics Database, http://ctd.mdibl.org/) is a good source to get integrated information for toxicogenomics research[7,8]. CTD gathered the relational information among genes, diseases and chemicals using text mining technologies and it is freely available through web interface.

The Basic Statistical Analysis

The basic statistical analysis processes of Microarray such as pre-processing, normalization, differently expressed gene finding, and clustering have been well established and there are many free and com-mercial software programs. Among them, ArrayTrack is one of the most comprehensive and suitable for toxicogenomics research than any other free software[9]. It is equipped with a very easy data import interface for both one-dye and two-dye experiments and visual output interface (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/). Furthermore, its analysis module is connected with the database module so that researchers can easily analyze and store their own data within a unified interface. Because it was originally designed for toxicogenomics research, another benefit of ArrayTrack is that it can store chemical information that is crucial for toxicology researches in the molecular level. In the statistical aspect, BRB ArrayTools is a powerful microarray analysis package (http://linus.nci.nih.gov/BRB-ArrayTools.html). It supports both supervised and unsupervised approaches such as permutation t-test or F-test, cluster analysis, principal component analysis, and so on. Users can perform data cleaning, normalization, class comparison, class prediction, survival test, and other statistical analyses on their data depending on their specific experiment design. Annotations can also be retrieved from NCI or Affymetrix if available. It works as an Excel add-in and is very easy to use.

The data treatment of the variance among individual experimental animals is one of the key issues in statistical analysis of toxicogenomics. Theoretically animals of same species are expected to show same gene expression patterns in reaction to single chemical but there exist variance among expression patterns in most cases,. So the statistical consideration is needed for extracting common change of expression patterns from each animal. In some experimental design, pooling mRNAs before hybridization is proposed in order to exclude individual variance, however, pooling is statistically dangerous and must be applied to data analysis with care because generally 3 to 5 animals are small number for achieving statistical significance.

Clustering process of expression profile is essential for gathering genes that may lie in a same biological process and clustering results give us the insight of the mechanism study of toxicology. There are a lot of clustering technologies such as hierarchical[10], K-means[11], and SOM (Self Organizing Map[12]). It should be noted that different clustering methods usually produce different results which may bring confusion to researchers. Unfortunately, there is no outstanding clustering method for all kinds of Microarray data because the performances of each clustering method vary according to data characteristics. Extracting consensus results from different clustering methods can be a solution for achieving more reliable results and several tools have been developed for this purpose.

Assigning biological meaning to each cluster is the next step after clustering. Recently, some informatics tools are developed for automatic assignment of biological meaning of clusters. GOODIES (http://istech21.com/en/download/download_a01_02.html) is free software that can interpret clustering results based on Gene Ontology[13]. The tree-shaped output is very helpful for easy identification of cluster structures and users can check the robustness of clusters by simply observing the divergence of cluster trees.

Mapping of clustered genes to the previously known pathway is an essential step for mechanism study from the clustering results. GenMAPP can apply user-defined color schema to its pathway maps, based on gene expression data[14]. A utility tool that helps users to prepare the data set ready for GenAMPP from Affymetrix chips are integrated into ICBR AnalyzeIt software. Currently, GenMAPP program only works on its own preloaded pathway maps. It doesn't take pathway maps in KEGG (http://www.genome.jp/kegg/) database[15,16] unless users would like to manually construct GenMAPP formatted files for them. ICBR bioinformatics group is seeking a way to integrate KEGG database into AnalyzeIt software so that KEGG pathway maps can be used efficiently in microarray data analysis.

Classification

Classification is the key technology for prediction of given chemical's toxicity based on expression profiles of known chemicals. The classification steps can be divided into three successive steps; gene selection, classification, and error estimation. Gene selection is the first step in which informative genes for classifications from the whole array spots are extracted. Classification is the main process to

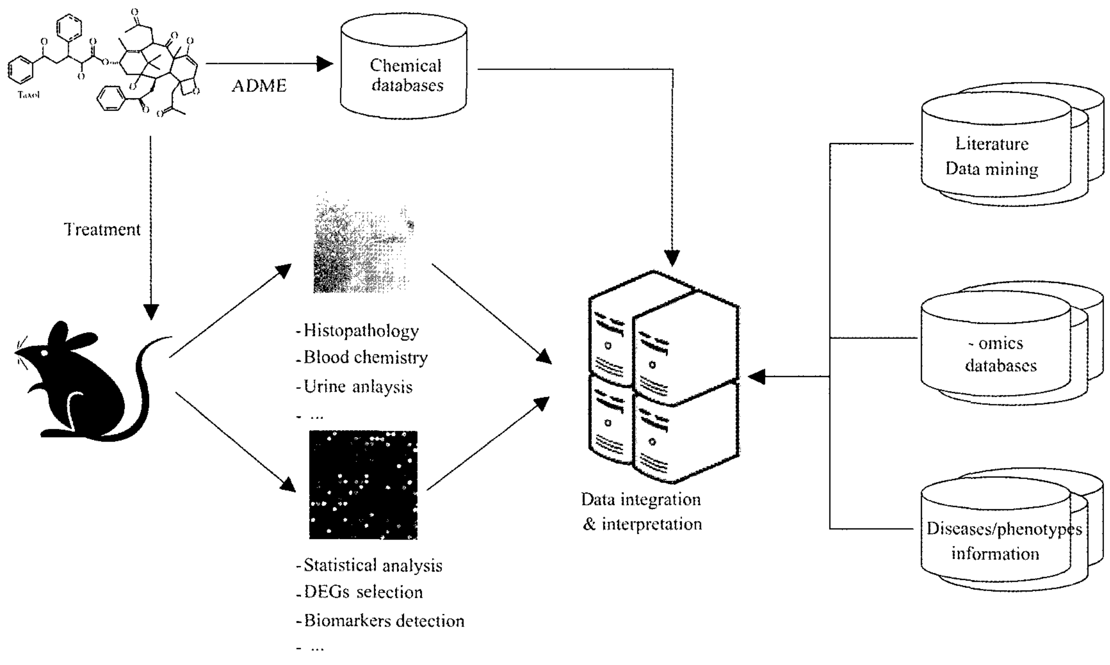

Fig. 1. The Role of Toxicoinformatics for the Toxicotranscriptomics Research

classify each group using selected genes and the accuracy of classification can be measured by error estimation. Many algorithms have already been developed for each step and applied to toxicogenomics and other classification analyses.

The association analysis of expression data with biochemical and/or histopathological data can make classification step more precise and specific for prediction of new chemical's toxicity. Through the association study between expression data and biochemical and/or histopathological data, one can perform sub-grouping of expression profile according to specific toxicity. Suppose that several chemicals have many different efficacies and toxicities but that they commonly bring necrosis *in vivo*. We can extract the expression profiles about necrosis by comparing expression profiles and histopathological data. This type of analysis is not focused on specific chemicals but on molecular mechanism, it is more widely applicable for the prediction of new chemical's toxicity.

Database and Data Standardization

]For current toxicogenomics studies, it is essential to integrate different types of '-omics' data such as microarray and protein 2-D for comprehensive understanding. Due to their heterogeneity and dispersion, however, gathering and curating '-omics' data are one of the most challenging issues in toxicogenomics. Toxgate is an integrated information delivery system for '-omics' data for Toxicogenomics research[17]. It provides curated and integrated information for same target or similar class of toxicant derived from different types of experiments. This can be used as a model system for the construction of toxicogenomics database.

The data exchange is another challenging issue. The MIAME/TOX document is based on the MIAME 1.1document that was produced by the MGED (microarray gene expression database) Society. The goal of MIAME (minimum information about microarray experiment) is to outline the minimum information required to interpret unambiguously and potentially reproduce and verify an array based gene expression monitoring experiment[18]. Although details for each experiment may be various, MIAME aims to define the core that is common to most experiments. MIAME /TOX is an expanded set of MIAME and designed for proper use of MIAME to toxicogenomics research. Thus, this standardized protocol may be helpful for exchanging toxicogenomics data.

Toxicoinformatics is the master key for toxicogenomics research because it is involved in every experimental step of research. Many algorithms and tools have already been developed for the toxicogenomics but key issues such as precise prediction of toxicity and data integration are not fully resolved yet. The most urgent problem is the construction of integrated database based on standardized experimental and analytical protocols among different array platforms, and this will leads the application of more sophisticated information technologies to toxicogenomics field.

References

1. Roos, D.S. Computational biology. Bioinformatics--trying to swim in a sea of data. *Science* **291**, 1260-1 (2001).
2. Kanehisa, M. & Bork, P. Bioinformatics in the post-sequence era. *Nat Genet.* **33** Suppl, 305-10 (2003).
3. Guerreiro, N., Staedtler, F., Grenet, O., Kehren, J. & Chibout, S.D. Toxicogenomics in drug development. *Toxicol Pathol.* **31**, 471-9 (2003).
4. Suter, L., Babiss, L.E. & Wheeldon, E.B. Toxicogenomics in predictive toxicology in drug development. *Chem Biol.* **11**, 161-71 (2004).
5. Yang, Y., Blomme, E.A. & Waring, J.F. Toxicogenomics in drug discovery: from preclinical studies to clinical trials. *Chem Biol Interact.* **150**, 71-85 (2004).
6. Lakkis, M.M., DeCristofaro, M.F., Ahr, H.J. & Mansfield, T.A. Application of toxicogenomics to drug development. *Expert Rev Mol Diagn.* **2**, 337-45 (2002).
7. Mattingly, C.J., Colby, G.T., Rosenstein, M.C., Forrest, J.N., Jr. & Boyer, J.L. Promoting comparative molecular studies in environmental health research: an overview of the comparative toxicogenomics database (CTD). *Pharmacogenomics.* **4**, 5-8 (2004).
8. Mattingly, C.J., Colby, G.T., Forrest, J.N. & Boyer, J. L. The Comparative Toxicogenomics Database (CTD). *Environ Health Perspect.* **111**, 793-5 (2003).
9. Tong, W. *et al.* ArrayTrack--supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research. *Environ Health Perspect.* **111**, 1819-26 (2003).
10. Eisen, M. B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci. USA* **95**, 14863-8 (1998).
11. Lu, Y., Lu, S., Fotouhi, F., Deng, Y. & Brown, S.J. Incremental genetic K-means algorithm and its application in gene expression data analysis. *BMC Bioinformatics.* **5**, 172 (2004).
12. P. Tamayo *et al.* Interpreting patterns of gene expression with SOMs-methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA,* **96**, 2907-2912 (1999).
13. Lee, S.G., Hur, J.U. & Kim, Y.S. A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics.* **20**, 381-8 (2004).
14. Dahlquist, K.D., Salomonis, N., Vranizan, K., Lawlor, S.C. & Conklin, B.R. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat Genet.* **31**, 19-20 (2002).
15. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**, D277-80 (2004).
16. Kanehisa, M. The KEGG database. *Novartis Found Symp.* 247, 91-101 discussion 101-3, 119-28, 244-52 (2002).
17. http://www.istech.info/ToxGate/
18. Knudsen, T. B. & Daston, G. P. MIAME guidelines. *Reprod Toxicol.* **19**, 263 (2005).