TEST DB:
The intelligent data management system for Toxicogenomics
독성유전체학 연구를 위한 지능적 데이터 관리 시스템

Wan-Seon Lee[1], Ki-Seon Jeon[1], Chan-Hwi Um[1], Seung-Young Hwang[2], Jin-Wook Jung[3], Seung-Jun Kim[2], Kyung-Sun Kang[4], Joon-Suk Park[4], Jae-Woong Hwang[4], Jong-Soo Kang[5], Gyoung-Jae Lee[5], Kum-Jin Chon[5] and Yang-Suk Kim[1*]

[1]Bioinformatics Unit, ISTECH Inc., #704, Hyundai Town Vill, 848-1 Janghang, Ilsan, Goyang, Gyeonggido, 411-380, Korea

[2]GENOCHECK Co.Ltd., #630 HBI, Hanyang Univ., Ansan, Kyunggido, 425-791, Korea

[3]Hanyang University, Ansan, Kyunggido, 425-791, Korea

[4]Lab of Stem Cell & tumor Biology, Department of Veterinary Public Health, College of Veterinary Medicine, Seoul National Uni., San 56-1, Shilim, Kwanak, Seoul 151-742, Korea

[5]Shin-Won Scientific Co.Ltd., #303, Dong-Wha B/D, 1580-11, Seocho3dong, Seocho, Seoul, 137-875, Korea

*To whom correspondence should be addressed. E-mail: yskim@istech21.com

Abstract

Toxicogenomics is now emerging as one of the most important genomics application because the toxicity test based on gene expression profiles is expected more precise and efficient than current histopathological approach in pre-clinical phase. One of the challenging points in Toxicogenomics is the construction of intelligent database management system which can deal with very heterogeneous and complex data from many different experimental and information sources.

Here we present a new Toxicogenomics database developed as a part of 'Toxicogenomics for Efficient Safety Test (TEST) project'. The TEST database is especially focused on the connectivity of heterogeneous data and intelligent query system which enables users to get inspiration from the complex data sets. The database deals with four kinds of information; compound information, histopathological information, gene expression information, and annotation information.

Currently, TEST database has Toxicogenomics information for 12 molecules with 4 efficacy classes; anti cancer, antibiotic, hypotension, and gastric ulcer. Users can easily access all kinds of detailed information about these compounds and simultaneously, users can also check the confidence of retrieved information by browsing the quality of experimental data and toxicity grade of gene generated from our toxicology annotation system. Intelligent query system is designed for multiple comparisons of

experimental data because the comparison of experimental data according to histopathological toxicity, compounds, efficacy, and individual variation is crucial to find common genetic characteristics.

Our presented system can be a good information source for the study of toxicology mechanism in the genome-wide level and also can be utilized for the design of toxicity test chip.

Introduction

'Toxicogenomics'는 독성에 관련된 신체 내의 현상을 유전체 수준에서 연구하는 분야로, 특히 신약 개발에서 전임상 및 임상 과정의 독성 및 안전성 연구에서 신약 개발 과정 중 가장 큰 걸림돌로 작용하고 있는 다수의 신약 후보물질에서 성공 가능성이 높은 신약 후보 물질을 보다 빠르고 정확하게 선별하여 작업의 효율성을 극대화 할 수 있는 획기적인 학문이다. 'Toxicogenomics' 연구는 보통 cDNA microarray을 이용한 약물과 시간 등에 의한 gene expression pattern을 분석하는 high-throughput 기술과 접목하여 진행되며 동시에 동물 실험 데이터, 유전자 정보 데이터 등 방대하고 다양한 데이터를 활용하는 분야이다. 이들 데이터 간의 관계를 잘 정리하고 활용하는 작업은 'Toxicogenomics' 연구에서 가장 중요한 부분 중의 하나라 할 수 있다.

본 논문에서는 2002년부터 진행중인 'Toxicogenomics for Efficient Safety Test(TEST)' project를 통해 산출된 데이터를 통합하여 관리하는 TEST database에 대해 소개하고자 한다. TEST project는 약물에 대한 해부학적 독성 분석 결과와 유전자 발현 변화간의 상관관계 분석을 통해 전임상 단계에서 '신약 후보 물질의 독성을 판정할 수 있는 진단칩의 개발 및 제품화'를 목적으로 한다. TEST 데이터베이스는 이 프로젝트를 통해 산출된 이종 데이터간의 연결 관계에 대해 초점을 두고 있으며 이들 정보를 최적으로 활용할 수 있는 데이터 통합 및 지능적 쿼리 시스템으로 구성되어 있다.

This work is supported by Ministry of Health & Welfare.

Toxicogenomics for Efficient Safety Test(TEST) Database

TEST 데이터베이스는 다양한 Toxicogenomics 연구 결과를 확인하고 정리하기 및 정보를 사용하기 편리하도록 간결한 형식으로 구성하는 작업에 필수적인 데이터베이스이다. 본 데이터베이스는 다른 형태의 네 가지 종류의 데이터베이스 - compound, animal experiment, microarray, annotation information- 를 통합 구성하고 있다.

Data source

TEST 데이터베이스는 샘플 준비부터 histopathology 데이터 및 유전자 발현 패턴 분석 결과에 이르기까지 2002년부터 진행중인 TEST 프로젝트로부터 산출된 모든 데이터를 저장 관리한다.

등급	내용	Clone 수
A	Paper set에서 추출한 clone set	248
B	Function annotation 수행으로 추출한 clone set	1871
C	Known clone set	1005
D	Unknown clone set	1620

표1. 4.8k clone 등급

Compound database

위궤양, 고혈압, 항생제, 항암제 등 총 4가지 분류의 약물을 각 3종씩 선택하여 각 약물에 대한 CAS(Chemical Abstracts Service) number, MTD(Maximum tolerated dose), NOL(Normal Operating Loss), 구조 등에 대한 기본적인 정보와 약물에 대한 실험 정보를 저장한다.

Animal experiment database

각각의 약물에 대해 Spring-Dawley Rat 암수 다섯 마리씩 1주에서 2주간 투여 후, 임상 증상 및 병리학적 분석 데이터를 저장하고 관리하는 데이터베이스이다.

Microarray database

각 장기로부터 뽑은 total RNA에 대한 발현 정보를 저장한 데이터베이스로 초기 raw data에서부터 데이터 quality 및 유의 발현 유전자(Different Expressed Gene) 목록까지, cDNA microarray 실험에서 분석까지 전반의 데이터를 관리한다.

Annotation database

cDNA microarray에 심겨진 clone set은 16k ResGen clone set으로부터 'Gene Ontology(GO)'를 통한 '기능적 분류 알고리즘(functional classification algorithm)', 독성 관련 키워드, 그리고 관련 문헌 등을 이용하여 용하여 4.8k의 clone을 선별하였다. 이렇게 선별된 clone은 선별 기준에 따라 4가지 등급(표1)으로 나누어 관리되고, public database를 이용한 조직 별 발현 정보 이용 및 데이터베이스 연동이 가능하다

Integrated management system

TEST 프로젝트를 통해 다뤄진 데이터들은 각각 독립적인 네 가지 형태의 데이터로 구분되는데, 데이터 간의 상관관계 분석을 위해서는 독립적인 데이터를 통합적으로 관리하여 보여주는 것은 필수적이다(그림1).

TEST 데이터베이스는 이들 데이터간의 연결점을 찾아 네 데이터베이스들 중 어느 한 데이터베이스의 데이터를 통해서도 다른 세 데이터베이스의 관련 정보를 함께 추출, 정리 및 분석 할 수 있는 시스템이다(그림 2). 예를 들어, annotation 데이터베이스를 통해 특정 clone에 대한 정보를 볼 경우, 선택한 clone에 대한 약물 별 동물실험 및 microarray 데이터와 결과를 함께 제공 받을 수 있다.

Intelligent query system

TEST 데이터베이스는 지능적 쿼리 연산이 가능한 시스템으로 다양한 데이터의 다중 비교가 가능하므로 데이터를 보다 효율적으

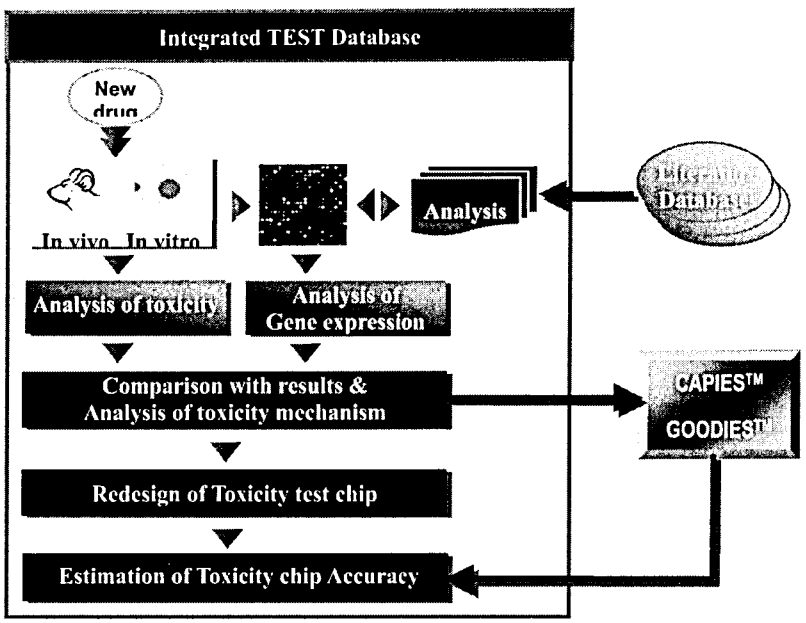

그림1. 통합 데이터베이스의 구성 내용

*CAPIES™: Gene Ontology Oriented Development with Innovative Endeavors & System developed by ISTECH,Inc

**GOODIES™: ClAss Prediction enhancement with Innovative Endeavors & System developed by ISTECH,Inc

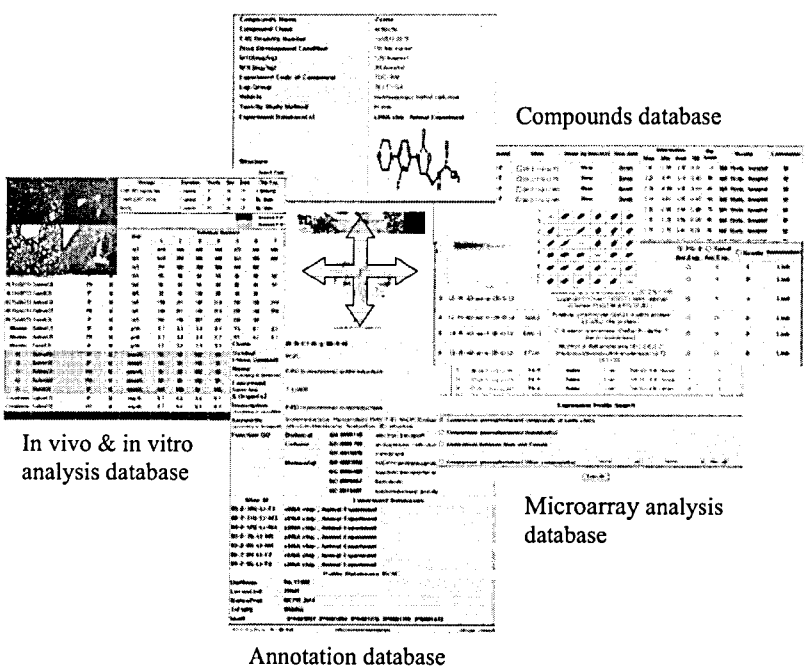

그림2. TEST 데이터베이스 데이터들간의 관계

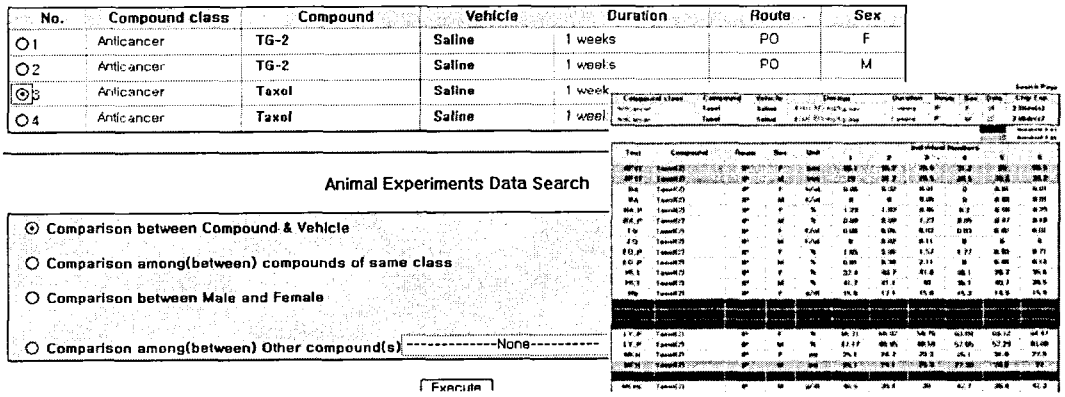

그림3. 동물 실험 데이터 검색 창과 결과 창

로 검색하고 관리토록 해주는 지능적 데이터 관리 시스템이다. 지능적 쿼리 검색은 데이터 타입이 보다 다양한 동물 실험 분석과 cDNA microarray 실험 분석 데이터베이스에 중점적으로 활용되고 있다.

동물 실험 데이터베이스의 쿼리 시스템은 우선 약물, 성별, 종 등의 기본 검색 후 최종 선택한 약물에 대해 약물(compound)과 용매(vehicle)간의 비교, 같은 분류(compound class)에 속하는 약물간의 비교 및 성별 비교 검색이 가능하고, 약물 분류에 상관없이 사용자가 최대 3개까지 자유롭게 약물에 대한 동물 실험 결과를 비교하는 것도 가능하다. 검색 결과는 food consumption, body weight, absolute organ weight, related organ weight, heamatology, blood chemistry, urinalysis, clinical observation, histopathology 등 기본적인 결과 수치 데이터에서부터 통계적 분석을 통한 유의적 결과 및 전문가의 소견까지 동물 실험 전반에 걸친 데이터를 총괄하여 보여준다(그림3).

cDNA Microarray 실험 데이터베이스의 쿼리 시스템은 microarray 이미지의 통계 처리를 통한 DEG 검색을 위한 기본 검색(basic search, 그림4A)과 같은 분류에 속하는 약물 사이의 microarray 실험 결과 비교, 개체 별 및 성별, 그리고 약물 분류에 상관없이 사용자가 선택한 3가지 약물에 대한 결과를 비교 검색하는 고급 검색(advanced search, 그림4B)으로 목적에 따라 나누어 진행할 수 있다. 기본 검색에서는 초기 이미지 수치 데이터 저장과 normalization 과정 전 후의 데이터 비교 및 quality 검사 자료를 수치와 그래프 데이터로 쉽게 확인할 수 있다. 그리고 사용자가 약물을 자유 선택하는 고급 검색에서는 같은 약물을 처리한 다른 개체들 사이에 중복되어 나오는 DEG만을 비교하는 'min' 검색과 같은 약물을 처리한 모든 개체에서 한번 이상 유의하게 나타난 DEG를 모두 선택하여 비교하는 'max'검색이 가능하다.

다양한 조건 검색이 가능한 지능적 쿼리 시스템은 이질적인 대량의 데이터를 동시에 비교 가능하게 함으로서 데이터간 관계 분석 작업에 소요되는 시간 단축은 물론 유용한 정보를 재생산하는 역할도 가능하다.

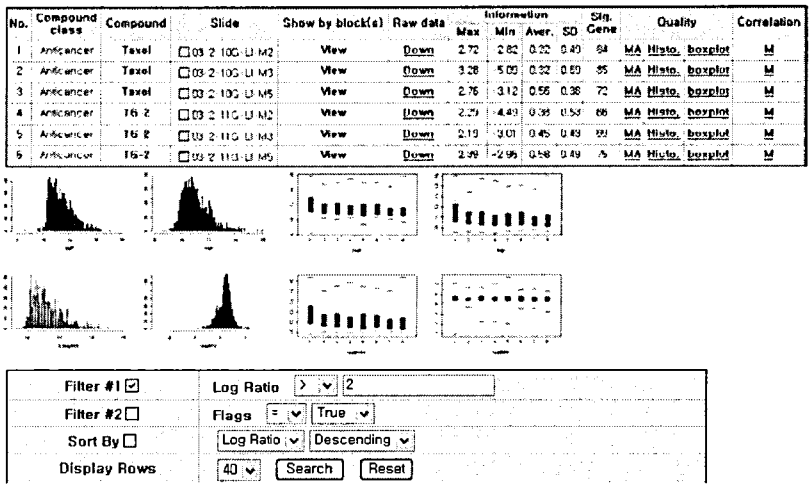

그림4A. cDNA microarray데이터 기본 검색

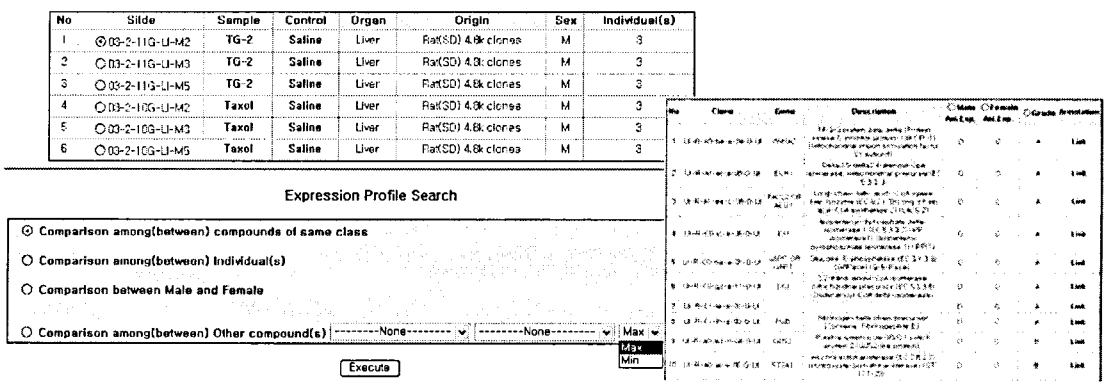

그림4B. cDNA microarray 데이터 고급 검색

Discussion

'Toxicogenomics'의 데이터를 다루기가 어려운 것은 데이터의 방대함 보다는 데이터 타입이 다양하고 이질적인 점이다. 흩어져서 관리되는 독특한 형식의 데이터들을 분석 시 활용하는 일은 추가적인 시간과 노력을 필요로 한다. 따라서 이들 데이터들의 관계를 잘 정리하는 작업은 데이터를 분석하는 일 만큼 중요한 역할을 한다.

TEST 데이터베이스는 compound와 clone 정보, cDNA microarray과 동물실험 초기 데이터부터 분석 결과 데이터까지 모든 정보를 총괄하여 관리하는 데이터베이스이다. 같은 조건 혹은 다른 조건간의 데이터 비교가 용이하며 매년 진행되는 연구에 따른 데이터 량 증가와 상관없이 추가 분석 및 비교가 가능하다.

TEST 데이터베이스는 진행중인 과제를 통한 데이터 뿐만 아니라 외부의 다른 독성

데이터와 다른 방법을 통한 데이터, 예를 들면 'Chemical structure'를 활용하거나 대사 물질을 소변이나 체액으로부터 독성에 의한 genotype, phenotype을 확인하는 방법에 의한 결과 등 cDNA microarray 이외의 다른 방법들을 통해 얻은 데이터와의 비교 및 교환을 가능토록 그 기능을 확대하여 독성유전체학 연구를 위한 통합 관리 데이터베이스로서의 역할을 하고자 한다.

References

[1] Robert R. Young, Genetic toxicology: Web resources, *Toxicology*, *173*, 2002, 103-121

[2] Kevin T. Morgan, Toxicogenomics, Drug Discovery, and the Pathologist, *Toxicologic pathology*, *30*, 2002, 15

[3] Robert C. Sills, Quality Review Procedures Necessary for Rodent Pathology Databses and Toxicogenomics Studies: The National Toxicology Program Experience, *Toxicological Pathology*, *30*, 2002 88

[4] Kenneth Olden, Genomics: implications for toxicology, *Mutation Research*, *473*, 2001, 3-10

[5] Lewis L. Smith, Key challenges for toxicologist in the 21st century, *TRENDS in Pharmacological Sciences*, *22*, 2001

[6] Amy Francis, Toxico-Logic Cutting-edge technologies in molecular toxicology promise to speed drug, *The Scientist*, *14(1)18*, 2000

[6] Jan Kimber, The principles and Practice of Toxicogenomics: Applications and Opportunities, *Toxicological Science*, *54*, 2000, 277-283

[8] Joseph F. Contrera, Use of Toxicological Information in Drug Design, *Journal of Molecular Graphics and Modeling*, 18, 2000, 605