Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1161~1167

Web-based DNA Microarray Data Analysis Tool

Ki Hyun Ryu[1)] · Hee Chang Park[2)]

Abstract

Since microarray data structures are various and complicative, the data are generally stored in databases for approaching to and controlling the data effectively. But we have some difficulties to analyze and control the data when the data are stored in the several database management systems. The existing analysis tools for DNA microarray data have many difficult problems by complicated instructions, and dependency on data types and operating system, and high cost, etc.

In this paper, we design and implement the web-based analysis tool for obtaining to useful information from DNA microarray data. When we use this tool, we can analyze effectively DNA microarray data without special knowledge and education for data types and analytical methods.

Keywords : Analysis tool, Bioinformatics, Database, Java, Microarray

1. 서론

생명정보학은 90년대 후반부에 들어오면서 HGP(human genome project)의 활성화와 유전자 마이크로어레이 등 제반 기술들의 발달로 인해 더욱 더 풍부한 데이터들을 활용할 수 있게 되었고, 이러한 데이터들로부터 유용한 정보를 얻고자 하고 있다. 여기서 유전자 마이크로어레이(DNA microarray)는 세포나 조직 내의 수천 개 유전자의 발현 양상(gene expression pattern)을 한번에 볼 수 있게 하는 도구이다 (Schema(2000)).

생명정보학의 연구자의 전공 분야는 생물학에서부터 미생물학, 수학, 컴퓨터 공학, 통계학 등에 이르기까지 매우 다양하다. 각기 다른 전공 분야를 가진 연구자들이 기존에 개발되어진 분석 도구를 가지고 마이크로어레이 데이터를 분석을 통해 유용한

1) Graduate Student, Department of Computer Engineering, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : bioman@changwon.ac.kr

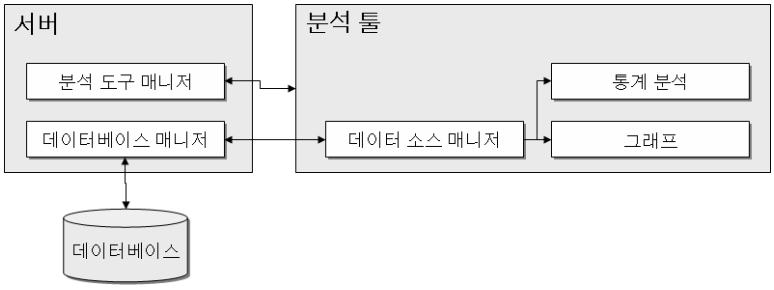
2) Corresponding Author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@changwon.ac.kr

정보를 얻고자 할 때, 컴퓨터나 통계학의 사전 지식이 없는 연구자들은 시스템 운용에 있어서 많은 어려움을 겪고 있다. Ruy와 Park(2006)이 기술한 바와 같이 기존에 개발되어진 분석 도구 몇 가지를 살펴보면, SAS는 방대한 양의 프로시저(procedure)와 이에 따른 다양한 옵션이 존재하는 반면에, 사용료가 비싸고, 초보자가 사용하기에는 쉽지 않으며, 웹 환경을 지원하지 않아 프로그램 실행 환경에 제약이 있다. SAS와 더불어 많이 쓰이는 SPSS는 원하는 통계 결과를 신속하고 용이하게 얻어낼 수 있으나, 윈도우라는 플랫폼(platform)으로 한정되어 있고 높은 사용료(김규곤(2001))를 지불해야 한다. 마지막으로 엑셀은 그래픽 환경에서의 수식 작업과 편리한 계산 기능을 제공하여 많은 사람들이 이용하고 있다. 그러나 워크시트(worksheet)당 65,538행과 256열로 제한되어 있어 수천 개의 컬럼(column)을 가진 유전자 마이크로어레이 데이터를 한 번에 다 읽어오지 못한다.

또한 마이크로어레이 데이터 형태가 파일 또는 데이터베이스로 매우 다양하지만, 한정적인 데이터 형태를 받아들이는 기존의 분석 도구로 인해 분석할 데이터 형태를 도구에 맞추어서 작성하거나, 변환 프로그램을 이용해야 하는 번거로움이 있다. 이에 본 논문에서는 Java Web Start 기술을 이용하여 다양한 데이터 형태를 받아들일 수 있는 인터페이스와 웹이 되는 환경이라면 언제 어디서나 사용할 수 있는 분석 도구를 개발하고자 한다. 논문의 구성은 다음과 같다. 먼저 2장에서는 제안하는 시스템의 특징과 전체적인 구조에 대해서 기술한 후, 3장에서 구현된 분석 도구를 고찰하고자 한다. 마지막으로 4장에서 결론을 맺고 향후 과제에 대하여 언급한다.

2. 웹 기반 데이터 분석 도구

이 절에서는 웹을 기반으로 하는 유전자 마이크로어레이 데이터 분석 도구의 전체적인 구조와 특징에 대하여 기술하고자 한다. 웹을 기반으로 운용되는 유전자 마이크로어레이 분석 도구의 구조는 <그림 1>과 같다.



<그림 1> 유전자 마이크로어레이 분석 도구의 구조와 동작

1) 서버 : 사용자가 분석 도구 요청 시에 다운로드를 제공하는 분석 도구 매니저와 데이터베이스 연결 및 해제를 관리하는 데이터베이스 매니저가 있다.

2) 분석 툴 : 데이터 소스 매니저(data source manager)는 사용자가 유전자 마이크로어레이 데이터가 들어 있는 파일(file) 또는 원격 데이터베이스(database)에서 분석

할 데이터를 추출하고 통합할 수 있도록 데이터 선택 및 편집 인터페이스를 제공하는 단계이다. 통계 분석(statistic analysis)은 선택 및 편집된 데이터의 통계적 분석 결과를 텍스트(text) 형태로 확인 할 수 있는 단계이다. 그래프는 통계 분석에서 분석된 결과를 사용자가 좀 더 알아보기 쉽도록 차트 또는 막대 형태의 분석 결과를 보여주는 단계이다.

본 연구에서 제안하는 유전자 마이크로어레이 데이터 분석 도구는 파일과 데이터베이스를 읽고 원하는 정보를 선택해서 볼 수 있도록 하여, 데이터를 분석하여 정보를 얻는 시간을 단축시킬 수 있다. 또한 파일과 데이터베이스 정보를 통합할 수 있으며, 플랫폼 독립적이다. 분석 도구를 자바(java) 언어(Flanagan(2000))로 개발하여 어느 환경에서도 동일한 결과물을 보장받을 수 있고, 운영의 폭을 넓힐 수 있게 하였다. 본 시스템에서는 데이터베이스에 접근하기 위한 수단으로 JDBC(java database connectivity, Reese(2000))를 이용하여 동일한 인터페이스로 데이터베이스에 접근하기 때문에 특정한 DBMS(database management system)에 종속적이지 않아 추가적인 구축이 불필요하다.

3. 웹 기반 분석 도구의 세부 기능

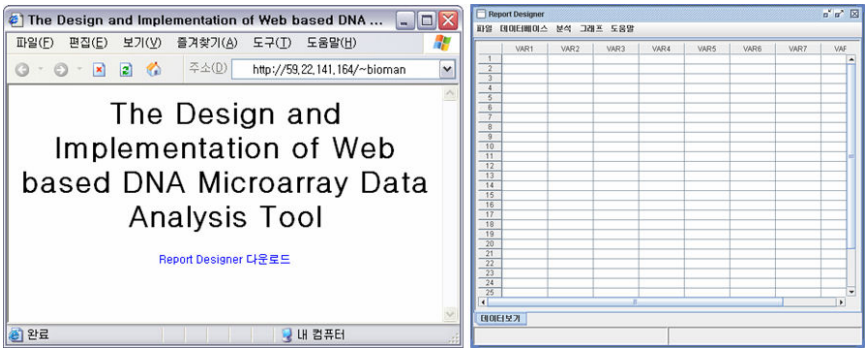
본 시스템의 구현을 위해 Java 2 SDK Standard Edition 1.5가 프로그래밍에 사용되었고 데이터베이스로는 MySQL4.1이 사용되었으며, Apache 1.3이 웹 서버로 사용되었다.

3.1 서버

Java Web Start로 구현된 분석 도구 매니저와 데이터베이스 연결을 관리하는 데이터베이스 매니저에 대해 기술하고자 한다.

1) 분석 도구 매니저

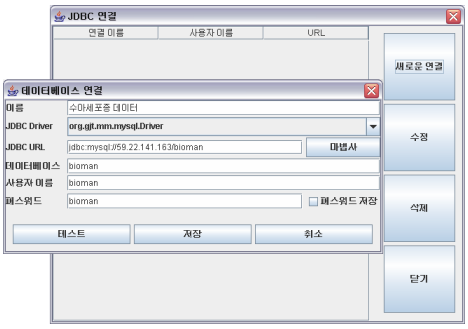
<그림 2>와 같이 분석 도구를 제공하는 서버에 접속해서 JNLP(java network launching protocol) 링크 파일을 클릭하면 분석 도구를 다운로드 하여 사용할 수 있다.



<그림 2> Java Web Start를 이용한 분석 도구 실행

2) 데이터베이스 매니저

<그림 3>과 같이 사용자의 데이터베이스 서버에 접속하고, 데이터베이스 서버의 종류가 다수일 경우 연결 목록을 만들어 관리할 수 있다.



<그림 3> 데이터베이스 연결 및 관리

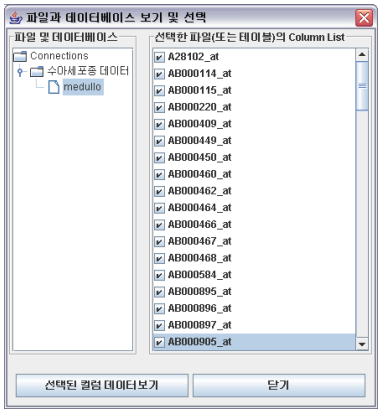
3.2 분석 도구

데이터 소스 매니저, 통계 분석 그리고 그래프의 기능들을 기술하고자 한다.

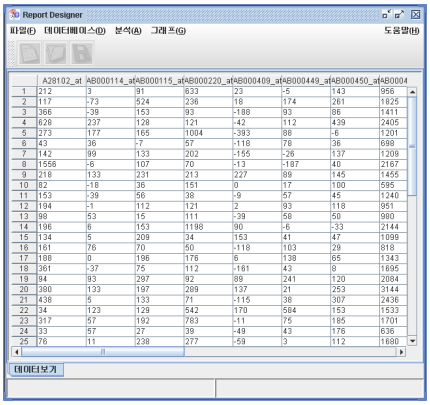
1) 데이터 소스 매니저

데이터 소스 매니저는 데이터 추출과 데이터 통합 기능을 제공한다.

(1) 데이터 추출 : <그림 4>와 같이 데이터베이스와 테이블 목록(또는 파일 목록)이 왼쪽 패널에 나타나며, 왼쪽 패널에서 선택된 테이블(또는 파일)의 컬럼 목록들이 오른쪽 패널에 나타난다. 원하는 컬럼 목록을 체크한 후에 “선택된 컬럼 데이터 보기” 버튼을 누르면 해당 컬럼의 정보들이 <그림 5>와 같이 데이터 테이블에 나타난다.

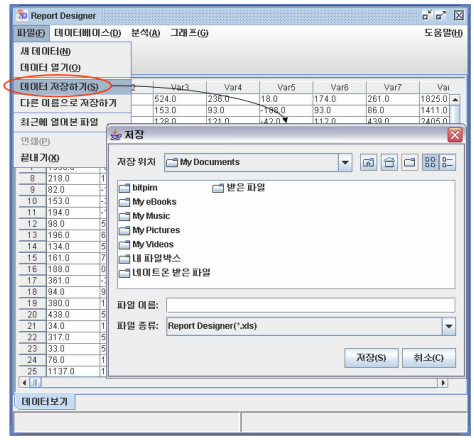


<그림 4> 데이터 추출 대화 상자



<그림 5> 데이터 테이블

(2) 데이터 통합 : (1)에서 추출한 정보가 표현된 데이터 테이블의 내용을 <그림 6>과 같이 파일로 저장할 수 있다.

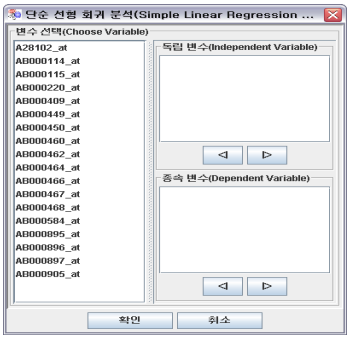


<그림 6> 데이터 통합을 위한 대화 상자

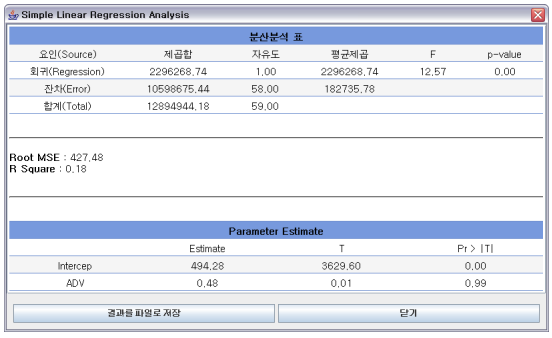
2) 통계 분석

기술통계, 도수분포표, 분산분석, 회귀분석 그리고 상관분석을 제공한다. 이들 기법들은 데이터를 분석하고 결과를 얻는 과정이 동일하므로 여기서는 기술 통계량의 예를 들어 설명한다.

파일이나 데이터베이스에서 읽어 들인 정보가 표현된 데이터 테이블에서 회귀분석 결과를 구하고자 할 때, <그림 7>과 같이 독립 변수와 종속 변수를 선택해서 그 결과를 <그림 8>과 같이 화면에서 확인 할 수 있으며, 파일로 저장할 수도 있다.



<그림 7> 변수 선택 대화상자

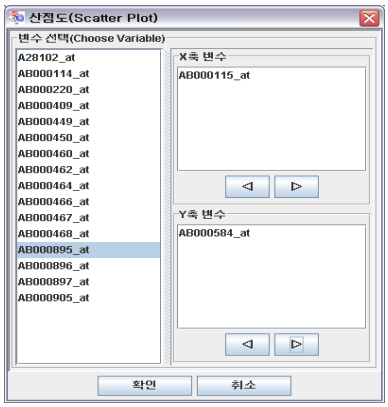


<그림 8> 회귀 분석 결과

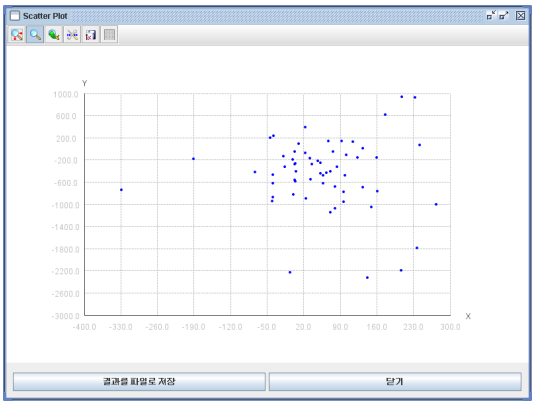
3) 그래프

히스토그램, 파이 차트 그리고 산점도 그래프를 제공한다. 이들 결과를 얻는 과정이 동일하므로 여기서는 산점도의 예를 들어 설명한다.

파일이나 데이터베이스에서 읽어 들인 정보가 표현된 데이터 테이블에서 산점도를 보고자 할 때, <그림 9>와 같이 X축과 Y축에 놓을 변수를 선택해서 그 결과를 <그림 10>과 같이 화면에서 확인할 수 있으며, 파일로 저장할 수도 있다.



<그림 9> 변수 선택 대화상자



<그림 10> 산점도 결과

4. 결론 및 향후 연구

본 논문에서는 웹을 기반으로 한 유전자 마이크로어레이 데이터를 분석하는 도구를 설계하고 구현하였다. 분석 도구의 구현을 위해 자바 언어를 사용하여 소스 프로그램의 변경 없이 플랫폼 독립적으로 운용이 가능하도록 하였다. 또한 데이터 분석에 필요한 최소한의 통계 분석 도구를 제공하고, 직관적이고 사용하기 쉬운 인터페이스를 제공함으로써 누구나 시스템 운용에 불편함이 없도록 하였다. 데이터 소스를 파일로 제한하지 않고, 데이터베이스의 데이터를 가져와서 분석할 수 있으며, 필요한 경우 직

접 데이터를 수정하고 파일로 저장함으로써 데이터 형태에 구애받지 않도록 하였다.

향후 연구과제로는 유전자 마이크로어레이 데이터 분석을 위해 데이터마이닝에서 주로 쓰이는 방법 중에서 연관성 규칙과 군집 및 분류분석이 가능하도록 개선되어야 할 것이다.

참고문헌

1. 김규곤 (2001), 비모수통계를 위한 한글형 통계분석 시스템의 연구 개발, *The Journal of Korean Data Analysis Society*, Vol. 3, No. 4, pp.457-476.
2. 양회석, 표선영 (2000), *퍼펙트 JSP*, 한빛미디어.
3. Flanagan, D. (2000), *Java Examples in a Nutshell*, 2nd Edition, O'Reilly & Associates.
4. Reese, G. (2000), *Database Programming with JDBC and JAVA*, 2nd Edition, O'Reilly & Associates.
5. Ruy, K. H. and Park, H. C. (2006), Network-based Microarray Data Analysis Tool, *Journal of the Korean Data & Information Science Society*, Vol. 17, No. 1, pp.53-62.
6. Schema, M. (2000), (ed.), *Microarray Biochip Technology*, Eaton Publising, MA.

[2006년 10월 접수, 2006년 11월 채택]