DNA chip 데이터 분석을 위한 Web-Bioconductor System 설계

Design of Web-Bioconductor System for DNA chip data analysis

신동훈[1], 박준형[2,3], 강병철[2], 신창진[4], 김철민[2,3]
[1] 국립암센터 암정보연구과
[2] 부산대학교 대학원 생물정보협동과정
[3] 부산지놈센터
[4] (주)진인 유전체 의학 연구소
Dong-Hoon Shin[1], Jun-Hyung Park[1,2], Byeong-Chul Kang[1]
Chang-Jin Shin[3], and Cheol-Min Kim[1,2,3]
[1] Cancer Information Branch, National Cancer Center Research Institute
[2] Interdisciplinary Program of Bioinformatics, Graduate School, Pusan Nat'l University
[3] Busan Genome Center
[4] Institute for Genomics Medicine, GeneIn. Co., Ltd.,
E-mail : micro_dna@hotmail.com

요 약

Web-Bioconductor System은 유전자 분석에 대한 통계적 모듈과 그래픽 환경을 제공하는 R언어와 DNA chip 데이터의 분석을 수행하는 Bioconductor 패키지를 이용하여 웹으로 DNA chip 데이터를 분석할 수 있도록 설계한 시스템이다. 본 시스템은 DNA chip 데이터의 분석을 위해 사용자 계정 모듈, 데이터 입력 모듈, 전 처리 모듈, 유전자 차등 발현 분석 모듈, 결과 출력 모듈로 구성되어 있으며, 분석된 결과물은 HTML, 이미지, XLS 파일 형태로 제공된다. 웹을 이용하여 DNA chip 분석을 수행함으로써 인터넷이 가능한 곳이면 시간과 장소의 구분이 없이 DNA chip 데이터 분석이 가능하며, 인터넷으로 DNA chip 데이터 분석 자료를 공유할 수 있음으로 연구자들의 상호 의견 교환을 바탕으로 효율적인 분석이 가능할 것이다. 또한 기존의 R언어와 Bioconductor가 전산 지식이 부족한 사람들에게는 접근하기 어려운 점을 웹 인터페이스로 간단하게 구현함으로써 DNA chip 데이터 분석에 있어 용이성과 효율성을 증대하고 있다.

1. 서론

DNA chip은 기존의 분자 생물학적 지식에, 현대에 엄청난 발전을 한 기계 및 전자공학의 기술을 접목해서 개발되었으며, 기계의 자동화와 전자 제어 기술 등을 이용하여, 적게는 수백 개 부터 많게는 수십만 개의 DNA를 아주 작은 공간에 집어 넣을 수 있게 만든 것이다. 이렇게 분자생물학과 공학 기술의 결합으로 탄생한 DNA chip은 생체 조직 표본으로부터 수천 개의 유전자 발현 양상을 동시에 관찰할 수 있는 도구로 개발되었으며, post genome 시대에 하루 동안 수백 개 이상 밝혀지는 새로운 유전정보들이나 모든 유전 암호가 밝혀진 생물들을 기존의 방법들로 연구하기에는 너무나 많은 시간이 요구되는 상황을 극복하기 위하여 최근에 개발된 방법 중 하나이다.

오늘날, 생물학 관련 연구에 DNA chip을 이용한 유전자 프로파일링 기술이 도입됨으로써 수많은 생체분자의 작용에 대한 이해뿐만 아니라, 질병의 원인을 분자 레벨에서 찾아내는 데에 큰 도움이 되고 있다. 이 과정에서 얻어진 복잡하고 방대한 양의 생물학적 데이터를 전통적인 실험

및 분석방법으로 연구하기에는 어려움이 있다.

따라서 이러한 문제점을 극복하기 위해 통계 모듈을 도입한 많은 분석 프로그램들이 개발되었다. 각 분석 프로그램들은 거의 동일한 데이터 분석 알고리즘을 채택하고 있으나, 분석과정에서 차이점이 있으며, 출력 결과의 표현 방법에서 차이점을 보이고 있다.

이와 함께 무료로 사용가능한 기존의 공개 분석 프로그램들은 수정이 불가능한 완성된 버전을 제공하기 때문에 사용자의 요구를 있을 때 쉽게 반영하지 못하는 단점을 가지고 있다. 또한 일반 프로그램과 같이 프로그램이 설치되어 있는 컴퓨터에서만 분석이 가능하다.

Web-Bioconductor System은 DNA chip 데이터 분석 도구인 R언어를 사용하였으며, R언어는 DNA chip 데이터에 대한 정확한 값을 예측하기 위해 도입된 통계분석 전용 프로그램 중 하나로 공개용으로 사용 가능한 패키지이다. 또한 Bioconductor 패키지를 사용하였는데 이는 R언어 관련 프로젝트로 생성되었으며, DNA chip 데이터의 통합 분석을 목적으로 개발된 R 기반 라이브러리 패키지이다.

본 논문에서는 DNA chip을 분석하는 여러 가지 분석 프로그램들을 벤치마킹한 뒤, R언어와 Bioconductor를 이용하여 일반 사용자들이 쉽게 분석할 수 있고, 사용자들의 요구를 쉽게 반영할 수 있으며, 웹을 이용하여 시간과 장소에 구애받지 않으면서 데이터를 분석할 수 있는 시스템을 설계, 개발하였다.

2. Web-Bioconductor System의 구성

Web-Bioconductor System은 R언어와 Bioconductor 패키지를 이용하여 DNA chip 데이터의 입출력, Bioconductor 라이브러리 호출 및 분석 수행, 사용자 계정 관리를 웹으로 통합하여 실행하는 시스템이다.

2.1 사용자 계정 관리

웹으로 DNA chip을 분석하는 시스템이므로 여러 사용자들이 인터넷이 가능한 장소에서 시간의 구애 없이 사용가능하다는 장점이 있지만, 외부의 무분별한 접근은 시스템의 마비와 함께 분석과정에서 장애를 유발시킬 수 있다. 또한 외부인에게 공개하고 싶지 않은 데이터를 노출할 수 있는 문제점을 안고 있다. 따라서 이러한 문제를 해결하기 위해 사용자에게 계정을 부여하고, 로긴 후 분석을 가능하게 하는 사용자 계정 관리 시스템을 구현하였다.

2.2 웹/분석 서버의 구성

데이터의 보안을 위하여 접근권한을 갖는 사용자 계정을 생성, 관리하며 분석을 위한 데이터 파일과 일괄처리 분석을 위한 리스트 파일을 입력 받는 웹서버와 입력받은 파일의 분석을 수행하는 분석 서버의 구성은 그림 1과 같다. 또한 표 1은 Web-Bioconductor System의 사양을 나타내고 있다.

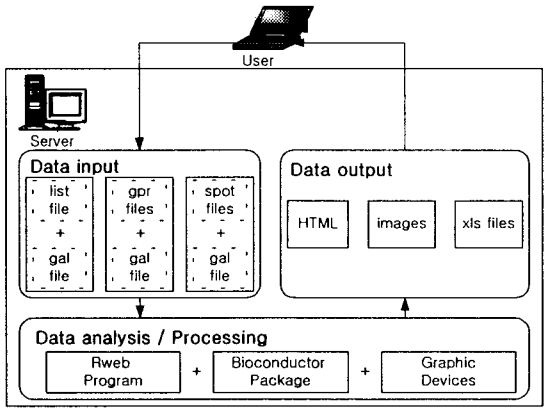

그림 1 Web-Bioconductor System 구조

표 1 Web-Bioconductor System 사양

구성요소	프로그램 및 버전
O.S	RedHat Linux 9.0
R Language	R 1.7.1
Bioconductor	Ver 1.3
Rweb	Rweb1.03
Perl	ver 5.6
Graphic Device	bitmap
Graphic Interface	Tcl/Tk(ver8.0)
Database	MySQL
Web-Server	Apache 2.0.48
HardWare	Intel Pentium III 800MHz
Memory	256MByte

3. Web-Bioconductor System의 설계

본 Web-Bioconductor System은 사용자 계정 모듈, 데이터 입력 모듈, 전 처리 모듈, 유전자 차등 발현 분석 모듈, 결과 출력 모듈로 설계되었으며, 각 모듈은 순차적으로 구성되어 있어 사용자가 쉽게 데이터의 입력부터 분석 그리고 결과물의 출력까지 수행할 수 있다.

다섯 가지 모듈의 기본적인 역할은 다음과 같다.

- **사용자 계정 모듈** : Web-Bioconductor의 웹 서버를 통해 접속된 사용자들의 계정을 등록, 관리한다.
- **데이터 입력 모듈** : 웹 서버를 통해 DNA chip 데이터 파일과 데이터 리스트 파일을 업로드하고, 기본적인 데이터 필터링을 수행한다.
- **전 처리 모듈** : 입력된 데이터 파일들을 정규화하며, 그 결과를 파일로 생성한다.
- **유전자 차등 발현 분석 모듈** : 정규화 과정에서 생성된 파일을 이용하여 PCA, 클러스터링, 필터링 등 DNA chip을 직접적인 분석이 수행된다.
- **결과 출력 모듈** : 최종 분석 결과를 웹으로 출력하고 결과 파일을 이미지, HTML, XLS형태로 생성하고 저장한다.

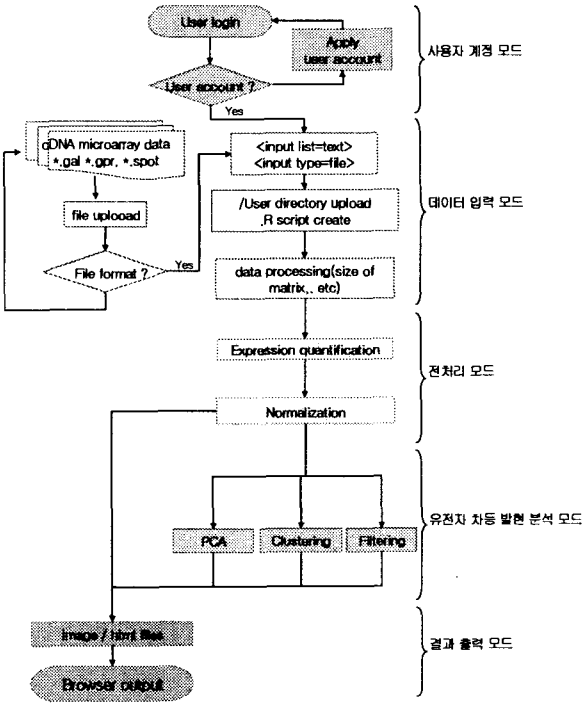

그림 2 Web-Bioconductor 모듈의 구성 및 연결

3.1 사용자 계정 모듈

사용자 계정 모듈은 MySQL 데이터베이스를 이용하여 사용자 개별 ID 및 패스워드를 부여하고 관리하고 있으며, 사용자의 데이터에 대한 분석 및 결과에 대한 보안을 고려하여 구성하였다.

3.2 데이터 입력 모듈

개별적인 입력 데이터의 형식은 DNA chip 이미지 분석에 가장 널리 사용되고 있는 GenePix 프로그램의 DNA chip 데이터 파일 gpr과 Spot 프로그램의 DNA chip 데이터 파일 spot을 받아들인다. Gpr과 spot은 DNA chip을 스캐닝하여 얻어진 이미지를 분석하여 수치화한 데이터 파일이며, 이 데이터 파일을 이용하여 유전자의 발현 패턴을 분석한다. 또한 gpr과 spot은 DNA chip 상의 각 probe 위치 정보와 probe 명을 기록한 gal 파일을 공통적으로 취하고 있다. 따라서 입력파일로 gal 파일과 함께 gpr파일 또는 spot파일을 업로드하여 다음 단계로 넘어간다.

3.3 전 처리 모듈

DNA chip 데이터의 정확한 분석 결과를 얻기 위해서는 분석 이전에 유전자 발현에 영향을 미치는 여러 가지 요인들을 보정하고 제거해야 한다. 이를 데이터 정규화(Normalization)라고 하며, 여러 가지의 정규화 기법들이 사용된다.

본 시스템에서는 스케일 정규화(scale normalization)와 위치 정규화(location normalizatin)을 이용하였다.

3.3.1 스케일 정규화

스케일 정규화는 단일 슬라이드 데이터에서 일반적으로 Cy5의 log-ratio와 Cy3의 log-ratio의 2차원 plot으로, $log_2R = log_2G$ 의 직선상에서 벗어난 유전자의 양을 관찰하려고 할 때 사용하며, $M = log_2(R/G)$, $A = log_2(R \cdot G)^{1/2}$에 나타난 결과에서 0 값을 기준선으로 유전자 데이터의 관찰에 적용되는 기법이다.

3.3.2 위치 정규화

DNA chip은 여러개의 블록으로 구성되어 있는데, 이때 각각의 블록은 서로 다른 프린트-팁(print-tip) 또는 핀(pin)을 사용하여 제작되어진다. 이때 프린트-팁 사이에는 팁 홈이 넓어지거나 많은 시간이 지난 후 프린팅의 변형 같은 시스테믹한 변형이 존재한다. 위치 정규화는 이를 해결하기 위한 방법으로 프린트-팁 그룹 내에서 loess 위치 정규화의 스케일 과정을 거치게 된다.

3.4. 유전자 차등 발현 분석 모듈

DNA chip 데이터를 분석하는 단계로서, 여러 통계적 분석 기법을 사용하여 분석 처리 조건에 따른 유전자 유사도 및 발현 패턴 변화를 분석할 때 사용하는 방법이다. 유전자 데이터를 분석함에 있어서 일반적인 통계 접근 방식은 주 요소 분석(PCA), 클러스터링(clustering), 유전자 필터링(gene filtering), 분류(classification) 등이 있으며, 본 시스템에서는 주 요소 분석 및 클러스터링과 유전자 필터링에 포함되어 있는 세부 기법

을 중심으로 시스템을 구현하였다.

그림 3은 정규화된 데이터에 대한 클러스터링의 분석 흐름도를 보이고 있으며, 계층적 방법과 k-means, SOM과 같은 비계층적인 방법으로 클러스터링이 가능하다.

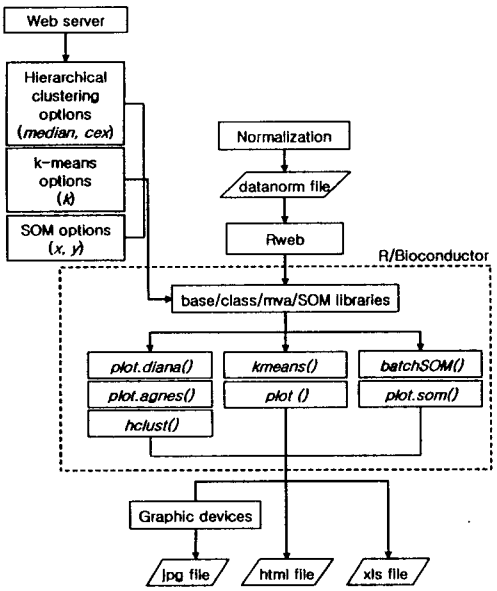

그림 3 클러스터링 분석 흐름도

3.5. 결과 출력 모듈

DNA chip 데이터 분석의 최종 결과를 나타내는 단계이며, 전 처리 과정의 정규화와 유전자 차등 발현 과정의 클러스터링, 필터링 단계에서 분석된 데이터 결과에 따른 그래프 및 테이블을 생성하고, 웹으로 구현한다. 최종 분석 결과는 이미지 파일, HTML, XLS 파일 형태로 생성된다.

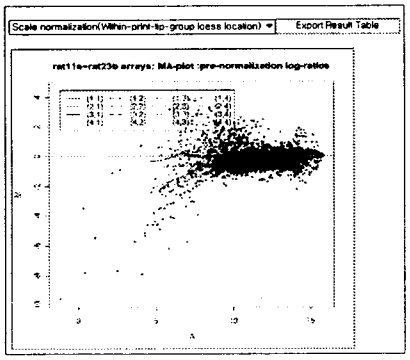

그림 4 Web-Bioconductor System의 결과 화면

4. 결과 및 고찰

본 논문에서 제시한 Web-Bioconductor System은 사용자 계정 모듈, 데이터 입력 모듈, 전 처리 모듈, 유전자 차등 발현 모듈, 결과 출력 모듈로 구성되어 있으며, 각각의 모듈을 호출함으로써 DNA chip 데이터 분석이 가능하다.

데이터 용량이 15MByte인 12개의 gpr 파일을 이용하여 사용자 계정으로 로그인 한 후 각 모듈들을 호출하여 분석을 수행하였다. 수행에 소요된 시간은 약 12분이었으며, 동일 샘플을 이용하여 서로 다른 곳에서 파일을 업로드 한 후 분석을 수행한 결과 동일한 결과를 얻을 수 있었다.

이 분석 시스템의 구축으로 얻어지는 효과는 다음과 같다.

첫째, 전산관련 지식이 부족한 생물 관련 연구자들이 분석하고자 하는 DNA chip 데이터 입력을 쉽게하여 유전자의 발현 패턴을 분석할 수 있다.

둘째, 고가의 DNA chip 데이터 분석 패키지가 설치되어 있지 않은 연구실에서 웹을 통해 시간과 장소의 구애 없이 분석이 용이하다.

셋째, 웹을 통해 데이터의 입출력이 가능하므로 원격으로 떨어져 있는 연구자들의 데이터 상호 공유가 가능하며, 또 다른 연구의 계기를 마련할 수 있다.

넷째, 분석한 데이터를 이용하여 연구자 고유의 데이터베이스를 구축할 수 있으며, SMD(Standford microarray database), GO, PubMed와 같은 데이터 분석의 의미를 부여할 수 있는 유용 데이터베이스와의 연계가 가능하다.

하지만 입력되는 데이터의 포맷이 GenePix와 Spot이라는 두 가지 상용화 포맷에 한정되어 있는 단점이 있다. 따라서 다양한 입력 파일 포맷에 대한 분석을 가능하게 하고, 유전자 발현 분석 기법에 이용되는 알고리즘을 사용자들이 쉽게 적용할 수 있도록 분석과정의 용이성과 분석 결과를 보다 쉽게 해석할 수 있도록 결과물을 표현하는 방법에 초점을 맞추는 연구가 요구된다.

5. 참고문헌

[1] Dudoit, S., Yang, Y. H., and Bolstad, B., 2002, Using R for the analysis of DNA microarray data. R News 2 (1), 24-32.

[2] Dudoit, S., Yang, Y.H., Calow, M.J., and Speed, T.P., 2002. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 12, 111-140.

[3] Jarno Tuimala and M.Minna Laine, 2003, A guidebook for DNA Microarray Data Analysis, Part II Analysis, Scientific Computing Ltd. 108 112