MicroArray의 직관적 시각적 분석을 위한 웹 기반 분석 도구

이승원[a], 박준형[a], 김현진[a], 강병철[a], 박희경[b], 김인주[a], 김철민[a]

[a]부산대학교 의과대 부산 지놈 센터

[b]동서대학교 응용 생명공학부

[c]진인

Web-based microarray analysis using the virtual chip viewer and bioconductor.

Seung-Won Lee[a],Jun-Hyung Park[a], Hyun-Jin Kim[a], Byeong-Chul Kang[b], Hee-Kyung Park[c],In-Ju Kim[a], Cheol-Min Kim[a]

[a]Busan Genome Center, College of Medicine, Pusan National University, Busan, South Korea

[b]Devision of Applied Bioengineering, Dongseo University, Busan,South Korea

[c]Institute for Genomics Medicine, GeneIn. Co., Ltd., Busan, South Korea

▪ **Abstract**

DNA microarray 칩은 신약 개발, 유전적 질환 진단, Bio-molecular 상호작용 연구, 유전자의 기능 연구 등 폭넓게 사용되고 있다. 이 논문은 cDNA mi mcroarray 데이터를 분석하기 위한 웹형태의 시스템 개발에 대한 내용을 다룬다. 하나의 cDNA microarr ay에는 수 백에서 수 만개의 유전자가 심어져 있으며, 데이터를 분석할 때 대량의 데이터와 다양한 형태의 오류로 인해서 데이터간의 차이를 보정하는 분석 도구와 통계적 기법들이 사용되어야 한다.

본 논문에서는 가상 칩 뷰어를 이용하여 실제 microarray 데이터의 foreground intensity에서 백그라운드의 intensity를 제거하여 일반화된 칩 이미지를 생성한다. 이 가상 칩 뷰어는 여러 가지 필터 효과와 서로 다른 두 형광의 차이를 조정하는 global normalization 기법을 사용하여 발현 유전자 분석을 시각적으로 할 수 있고, 중복된 마이크로어레이 칩 데이터를 통하여 시간이 많이 걸리는 분석 전 칩의 유효성을 검토할 수 있다. 칩 데이터의 normalization을 위한 통계 방법으로 R 통계 도구와 linear 모델을 사용하여 microarray 칩의 유전자 발현 양상을 분석한다. 통계적 방법을 사용하지 않은 데이터를 추출, 이 데이터의 패턴 그래프 그리고 발현 레벨을 분류하여 마이크로어레이의 각 스팟의 유효성 검토의 정확성을 높였다. 이 시스템은 칩의 유효성 검토, 스팟의 유효성 검토, 유전자 선정에 대해 분석의 용이성과 정확성을 높일 수 있었다.

▪ *Keywords:*

마이크로어레이, cDNA, 유전자 발현, linear model , R, Bioconductor

▪ 1. 서론

DNA microarray는 병렬적으로 수 천개 유전자들의 발현 양상을 측정하는 장치이다. DNA microarray는 스텐포드 대학에서 schena와 공동연구자들에 의해서 1990년 초에 개발되었고, photolithography, ink je tting 그리고 contact printing(lemieux et al., 19 98; Schena et al., 1998)기술은 microarray 생산을 위한 중요한 기술로 현재 사용되고 있다.

DNA micoarray는 유전체 전체의 유전자들이나 다양한 유전자들의 세트들을 빠르고 대량의 스크린 하는 것이 가능하게 되었고, bio-molecular 상호작용에 대한 네트워크들을 조사할 수 있으며, 또한 DNA microarray는 다양한 시간과 조건에서 전체적인 유전자 발현 패턴을 조사할 수 있다.

DNA 마이크로어레이의 이러한 장점으로 인해, drug 개발, 진단, 비교유전체학, 기능유전체학등 많은 분야에서 사용되고 있다.

DNA 마이크로어레이에는 수백에서 수만 개의 유전정보 및 발현에 대한 데이터가 생겨난다. 이 데이터의 대량성과, 분석의 복잡성으로 인해 다양한 분석 도구들이 이용되어야 하며, 분석 절차 및 복잡한 도구의 사용이 쉽지 않다.

이 논문의 목적은 대량의 DNA 마이크로어레이 데이터를 직관적이고, 시각적인 방법으로 분석할 수 있는 시스템의 개발에 관한 것이다. 마이크로

어레이 데이터를 가상 칩 이미지로 재현함으로써 관련 정보들을 쉽게 시각적으로 분석할 수 있고, 수행 옵션 값을 입력하여 신속하게 다양한 분석 정보를 얻을 수 있다. 데이터의 다양한 수집 방법을 통하여 분석 결과의 신뢰성을 높였다. 이 논문은 분석 방법을 쉽고, 정확하면서 직관적으로 분석하는 방법을 제시한다.

본 논문에서 연구 개발된 시스템은 웹 기반으로 구현되었으며, 아래 사이트에서 확인할 수 있다.

Availability : http://164.125.47.97/webma/index.php

▪ 2. 관련연구

▪ 2.1 SNOMAD

SNOMAD는 한 가지 또는 두 가지의 형광색을 가진 마이크로어레이 칩의 유전자 발현에 관한 분석도구이다. 백그라운드 intensity 제거, Global Normalization, Z-Score를 사용하여 분석한다. 이 도구는 R 통계 도구를 사용하였다. 다양한 그래프를 통하여 시각적으로 분석과정을 점검할 수 있다. 그래프의 사용은 전체적인 스팟들의 양상을 확인할 수는 있으나 직접적으로 그래프를 이용할 수 없으며, 보관 및 관리할 수는 없고, 또한 분석결과를 복수의 통계를 통해 얻기가 힘들다는 단점이 있다.

▪ 2.2 DNMAD

DNMAD는 R 통계 도구를 사용하여, normalization 방법으로 global loess normalization 방법과 print-tip loess normalization 방법을 사용한다. 칩의 normalization 방법에 중점을 두었고, 유전자 발현 양상에 대한 분석은 가능하지만 GEPAS라는 분석도구를 사용한다. 따라서 DNMAD는 마이크로어레이 칩 발현 양상에 대한 분석 중간 단계이다.

▪ 3. 구현

본 논문 크게 3가지로 분류(그림1)할 수 있다. 첫째는 로그인 부분으로 로그인, 사용자 관리, 사용자별 작업 공간을 확보하는 부분이고, 둘째는 프로젝트 관리 부분으로 프로젝트 생성 및 실험정보를 저장하는 기능을 담당하고, 셋째는 분석 부분으로 프로젝트 단위이며, 분석 작업 단위이다.

분석 부분은 가상 칩 뷰어, 데이터 수집, normalization 처리, noramlization 검토 및 관찰, 분석 결과 부분으로 구성되어 있다.

▪ 3.1 개발환경

Apache web server 1.3.22 버전을 사용하였으며, 데이터베이스 시스템은 Mysql 서버를 사용하였다. 또한 시스템 구현을 위한 언어는 PHP를 사용하였다. 이에 통계적 기법을 사용하기 위해 널리 사용하고 있는 R 시스템과 분석 패키지는 bioconductor의 limma를 사용하였다.

▪ 3.2 가상 칩 뷰어

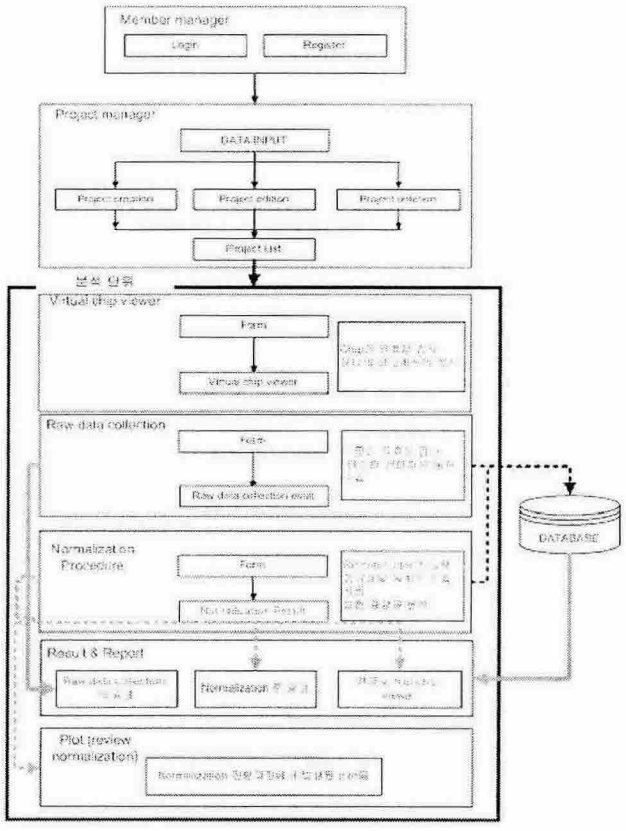

그림1. 전체 분석 흐름도

마이크로어레이 칩 정보를 가상 chip 이미지로 재현 하여 보여준다. 가상 chip 이미지는 log ratio 값, intensity값, Dye-swap 표현 그리고 chip 안에서 normalization 값을 통하여 filtering하거나 spot 이미지를 표현된다. 이 프로세서는 chip 간의 패턴을 비교하는데 아주 용이하다. 오류가 있는 chip이 있을 경우 패턴을 통하여 직관적으로 판단할 수 있다. 이 분석은 분석 이전의 chip을 유효성유무 확인을 통하여 분석할 chip 목록에서 제거할 수 있다는 장점이 있다. spot image 패턴에 의해서 발현 양상에 대한 분석이 가능하고 단일 chip의 분석도 가능하다. 그림2의 Cy5의 intensity는 RGB의 red로, Cy3는 green으로 변환하여 나타냈다.

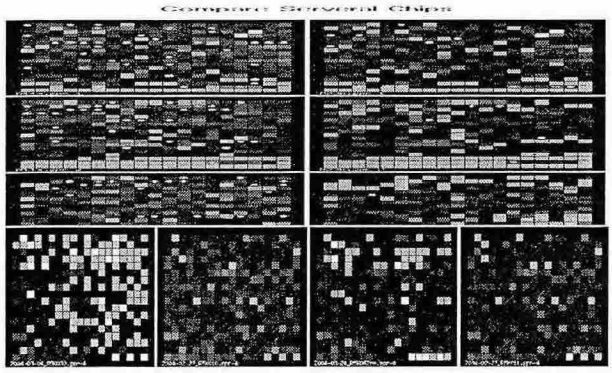

그림 2. 가상 칩 이미지

회색의 사각형 모양은 intensity로 필터링된 것을 나타내고 사각형 영역 안의 원은 log-ratio 항목의 값을 나타낸다. 그림 2의 상단 이미지는 2가지 칩의 첫 번째 Block을 나타내고 정확성과 통계적 분석을 위해서 반복 실험한 칩이다. 2가지 이미지에서 칩의 발현 양상이 차이가 있음을 알 수 있고, 아래 4개의 칩은 두 번의 duplication 실험과 2번의 dye-swap 실험한 것이다. 가장 왼쪽의 이미지는 회색이 많이 분포하고 있으며, 이는 전체적 강도가 약하고, 재실험을 검토해야 됨을 나타내고 있다.

▪ 3.3 칩 데이터 수집

Chip들의 intensity, log ratio에 대한 데이터, ratio의 레벨에 의한 데이터 분류, log-ratio에 대한 평균값, 그리고 spot 각각의 intensity 값들의 합을 계산한다. 이 기능은 통계방법으로 보여줄 수 없는 정보들을 수집하여 spot의 유효성을 raw 데이터를 가지고 판단할 수 있다. 또한 그래프를 통하여 intensity와 log-ratio 변화의 패턴을 시각적으로 확인할 수 있다.

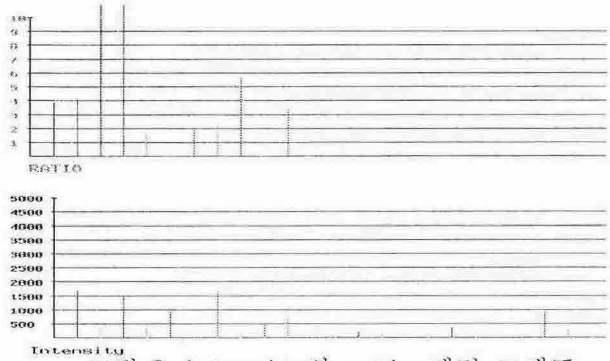

그림 3. intensity와 ratio 패턴 그래프

그림 3의 위 패널은 RATIO그래프로서 Cy5(Red Color)가 Cy3(Green Color)보다 intensity가 높다는 것을 나타낸다. 길이는 비율을 나타낸다. 그림3의 아래 패널은 intensity 강도를 나타낸다.

▪ 3.4 Normalization 절차

Within normalization method로서 print-tip normalization method와 between normalization method로서 scale normalization method을 사용하여 처리한다. 이 부분에서는 R를 통하여 bioconductor의 limma 패키지를 사용하여 normalization를 한다. limma패키지는 linear model을 사용한 cDNA 칩 분석 패키지이며, 널리 사용되고 있는 신뢰성이 높은 패키지이다. 이 패키지를 사용함으로써 신뢰성을 높이고, 이 절차를 통하여 normalization 진행 과정을 plot image로 생성시켜 확인할 수 있다. 이때 생성된 plot image는 normalization observation 메뉴를 통하여 확인할 수 있다. 발현 양상을 확인할 수 있는 통계결과는 M, A, t-statistic, B-statistic, P-value을 포함한다. M은 log-ratio을 나타내고, A는 Cy5와 Cy3 intensity의 log 값을 취한 것에 평균값이다.

▪ 3.5 Normalization 검토

Normalization 절차상에서 R과 bioconductor 실행 과정에서 생성된 plot들을 볼 수 있다.

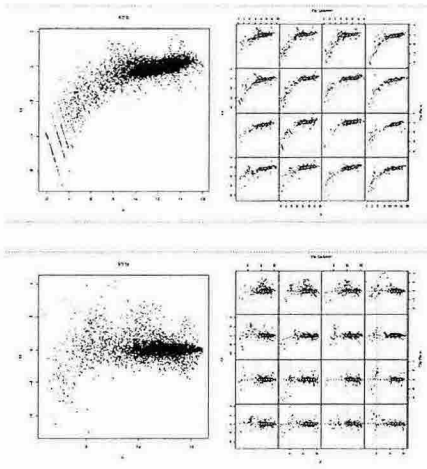

그림 4. print-tip normalization

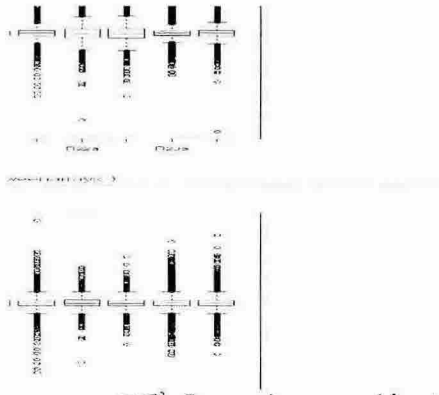

그림 5. scale normalization

정상적이고, 적합하게 normalization이 진행 되었는

지에 대해서 검토할 수 있다. 그림 4는 위는 칩에서 각 블럭에 대한 print-tip normalization 전, 아래는 normalization후의 plot들이다. 그림 5는 칩 간의 scale normalization 전후의 plot이다.

▪ 3.6 분석 결과 및 레포트

전체 분석한 결과를 확인할 수 있다. 데이터 수집과 normalization 과정에서 생성된 결과들을 통합하여 확인할 수 있다. 통합된 결과는 normalization 처리결과의 신뢰성을 확인할 수 있다.
normalization 후 결과를 확인할 수 있는 가상 칩 뷰어가 있다. 가상 칩 뷰어의 분석 전 칩의 유효성을 평가하는 뷰어와는 달리 이 뷰어는 통계 결과에 대한 다양한 옵션을 통하여 후보 유전자를 추출하는데 용이하도록 제작되어 있다. AND 와 OR 연산자를 가지고 있어, 다른 통계적 기법을 사용한 결과와 필터를 동시에 할 수 있다. 이 결과 파일을 다운로드 할 수 있고, 데이터베이스에 보관이 된다. log ratio값들에 의해서는 원으로, 그밖에 통계 값에 의해 흰색의 사각으로 표시한다.

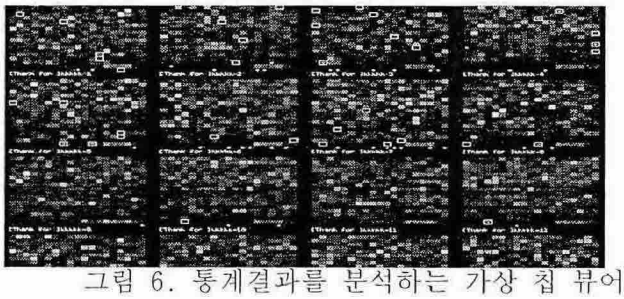

그림 6. 통계결과를 분석하는 가상 칩 뷰어

4. 결론

가상 칩 뷰어를 사용하여 여러 장의 칩을 비교분석한 결과 실험과정이나 칩 이미지 처리과정에서 잘못된 칩을 추출하여 제거할 수 있었다. 분석 전의 가상 칩 뷰어는 오류 칩을 제거할 수 있어 시간적으로 효율적이고, 분석의 정확성을 높일 수 있다. 스팟의 유효성을 위해 발현 레벨 검사와 그림5의 패턴 검사가 스팟 유효성 검사에 효율적 이였고, 후보 유전자 추출 방법으로 다양한 통계적 기법, log-ratio 분석, 필터기능을 동시에 사용함으로서 분석의 효율성과 정확성을 높였다. 이 도구는 인터넷에서 누구나 사용이 가능하다.

▪ 참고문헌

[1] Colantuoni,C., Henry,G., Zeger,S. and Pevsner,J.(2002)SNOMAD(standardization and Normalization of MicroArray Data):web-accessible gene expression data analysis. Bioinformatics, 18, 1540-1541.

[2] Firth,D. (2004) CGIwithR:gacilities for the use of R to write CGI scripts.

[3] Juan M. Vaquerizas, Joaquin Dopazo and Ramon Diaz-Uriarte. (2004) DNMAD: web-based diagnosis and normalization for microarray data. Bioinformatics, 20, 3656-3658.

[4] Herrero,J., A1-Shahrour,F., Diaz-Uriarte,R., Mateos, A., Vaquerizas,J. M., Santotyo,J. and Dopazo,J. (2003a) GEPAS: a web-based resource for microarray gene expression data analysis. Nucleic Acids Res., 31, 3461-3467.

[5] R Development Core Team (2004) R: a language and environment for statistical computing. Vienna, Austria.

[6] Smyth,G.K., Thorne,N.P. and Wettenhall,J. (2004) Limma: linear models for microarray data version 1.6.6, User's Guide.

[7] Smyth,G.k., Yang,Y.H. and Speed,T. (2003) Statistical issues in cDNA microarray data analysis. Methods Mol. Biol., 224, 111-136.

[8] Gordon K. Smyth and Terry Speed. (2003) Normalization of cDNA Microarray Data. Methods 31, 265-273.

[9] GenePix Pro microarray and array analysis software. Axon Instruments Inc. http://www.axon.com.

[10] Lonnstedt, I. and Speed, T.P. (2002). Replicated microarray data. Statistica Sinica 12, 31-46.