cDNA 마이크로어레이 데이터의 분석과 관리 시스템: cMAMS

김상배 O 김효미 이은정 김영진 박정선 박윤주 정호열 고인송
역학정보실, 유전체연구부, 국립보건연구원
ksb003@korea.com O, ippiy@hanmail.net, sys305@hotmail.com, inthistime@lycos.co.kr,
oct1001@yahoo.co.kr, pj518@freechal.com, hyjung@ngri.re.kr, insong@nih.go.kr

cDNA Microarray data Analysis and Management System: cMAMS

S.B Kim O H.M Kim E.j. Lee Y.J Kim J.S Park Y.J Park H.Y Jung I.S. Koh
Division of Epidemiology and Bioinformatics,
National Genome Research Institute, Korea National Institute of Health

요 약

마이크로어레이 기술은 근래에 개발된 신기술로써 동시에 수천-수만 개의 유전자 발현을 측정할 수 있어 다양한 생물학적 연구에 이용되고 있다. 여러 단계의 실험 과정과 이를 통해 얻은 다량의 데이터를 처리하기 위해서는 이를 효율적으로 관리, 저장, 분석할 수 있는 통합 정보 관리 시스템을 필요로 한다. 현재 외국에서는 몇몇 관리시스템이 개발되어 있고, 국내에서도 WEMA 등이 있지만 아직 데이터 관리부분에 기능이 치우쳐 있다. 따라서 우리는 복잡한 자료구조를 가지는 마이크로어레이의 실험 정보와 각 단계별 처리 정보 등을 사용자의 관점에서 효과적이고 체계적으로 관리할 수 있고, 데이터 정규화 및 다양한 통계적 분석 기능을 갖춰 불필요한 시간과 비용을 줄임으로써 마이크로어레이 연구에 도움을 주고자 통합 분석관리 시스템 cMAMS (cDNA Microarray Analysis and Management System)를 개발하였다. 웹 기반으로 구현된 cMAMS는 데이터를 저장, 관리하는 부분과 데이터를 분석하는 부분, 그리고 모든 관련 정보가 저장되는 데이터베이스 부분으로 구성되어 있다. 데이터관리부분에서는 WEMA의 계층적 데이터구조를 도입해 관리의 효율성을 높이고 시스템의 이용자를 시스템운영자, 프로젝트관리자, 일반사용자로 구분하여 데이터 접근을 제한함으로써 보안성을 높였다. 통계처리 언어 R로 구현된 데이터분석 부분은 7 단계의 다양한 분석(전처리, 정규화, 가시화, 군집분석, 판별분석, 특이적 발현 유전자 선별, 마이크로어레이 간의 상관분석)이 가능하도록 구현하였고, 분석결과는 데이터베이스에 저장되어 추후에 검토 및 연구자간의 공유가 가능하도록 하였다. 데이터베이스는 실험정보가 저장된 데이터베이스, 분석결과가 저장된 데이터베이스, 그리고 유전자 정보 탐색을 위한 데이터베이스로 분류해 데이터를 효율적으로 관리할 수 있게 하였다. 본 시스템은 LiNUX를 운영체계로 하고 데이터베이스는 MYSQL로 하여 JSP, Perl, 통계처리 언어인 R로 구현되었다.

1. 서 론

1.1 마이크로어레이 기술 및 시스템 필요성

DNA 마이크로어레이는 생물학 분야에서 유전자의 발현 청사진을 제공함으로써 유전자들의 기능 연구뿐 아니라 독성연구, 신약개발, 질병진단 등의 의학분야에 응용되고 있다[1]. 마이크로어레이 실험은 마이크로어레이 칩 제작 단계, 시료 준비 단계, Hybridization 단계, 스캐닝 단계, 각 유전자 발현 분석 단계로 구분되며, 최종적 결과를 얻기까지 많은 과정을 거친다[2]. 따라서 한번의 실험에서 나오는 데이터는 실험 결과 이외에 종합적인 실험정보가 포함된 방대한 양으로 복잡한 데이터 구조를 가지므로 이러한 데이터를 효율적으로 관리하기 위한 실험실 정보 통합 관리 시스템 (LIMS: Labarotory Information Management System)이 요구된다.

1.2 국내외 연구현황

현재 마이크로어레이 실험 관리용으로 외국에서 개발된 관리 시스템으로는 BASE, SMD, ARGUS, ArrayDB 등 여러 개가 있으나 국내에서는 WEMA, ISMIG 외에는 특별한 것이 없는 실정이다[3~8]. 이런 시스템들은 마이크로어레이 관련 데이터의 데이터베이스화를 통해서 데이터를 관리, 분류, 탐색만을 할 수 있고, ISMIG의 경우 유전자 상호간의 네트워크를 분석할 수 있다는 장점이 있지만, 일반 실험자들이 실제로 절실히 필요로 하는 데이터의 정규화 및 보다 다양한 통계적 분석 기능의 연동이 결여되어 있다 [7, 8]. 일반적으로 마이크로어레이 데이터를 분석하기 위해서는 각 분석단계별로 다른 프로그램을 이용해야 하는 번거로움이 있다. 본 연구에서는 WEMA의 데이터관리 체계를 도입하여 데이터 관리의 편리성을 증대시키고, 데이터 접근의 차별화를 통해 자료의 보안성을 강화하였으며, 여러 단계 의 분석을 하나의 시스템으로 가능하게 하는 분석 기능을 추가 함으로써 일반 실험자의 입장에서 사용하기 쉽고 분석도 용이한 효율적인 분석관리시스템 cMAMS (cDNA Microarray data analysis and Management System)을 구현하였다.

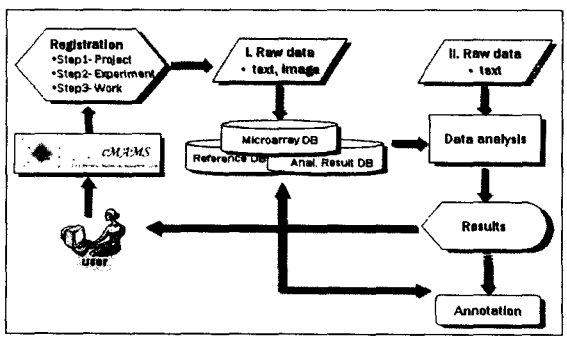


그림 1 cMAMS의 시스템 개요도

2. 본 론

2.1 시스템 개요 및 개발 언어

우리가 개발한 cMAMS는 마이크로어레이 실험 정보 및 결과를 저장, 관리, 분석하기 위해 웹 기반으로 구현된 통합 시스템으로써 데이터를 저장, 관리하는 부분(Data management module), 데이터를 분석하는 부분(Data analysis module), 그리고 모든 관련 정보가 저장되는 데이터베이스 부분(Database)으로 구성되어 있다. 본 시스템은 LINUX를 운영체계로 하고 데이터베이스는 MYSQL로 하여 JSP(Java Server Page), Perl, 통계처리 언어인 R로 구현 되었다.

사용자는 웹을 통해 본 시스템에 접속하여 사용자 등록 및 사용자 인증을 받은 후, 실험정보를 등록하고 마이크로어레이 이미지 파일 및 분석 파일을 업로드하여 데이터베이스에 저장할 수 있게 하였다. 또한 데이터 분석 시 저장된 데이터 목록을 선택하여 이를 다양한 통계적 방법으로 분석할 수 있도록 하였다 (그림 1).

2.2 시스템 체계 및 기능

cMAMS의 데이터 구조는 실험 방법, 재료, 조건에 따라 다양한 실험 결과를 통합하여 표현할 수 있는 Project, Experiment, Work 의 계층적 데이터 구조와 연구자 상호간의 의견제시, 요구 또는 지시사항 등의 전달 수단으로 게시판과 쪽지 기능 등을 WEMA 시스템에서 도입하였다 [3].

데이터의 보안성을 고려하여, 시스템을 이용하는 대상을 시스템운영자(System admin),프로젝트관리자(Project admin), 일반사용자(User)로 구분하였다. 시스템운영자는 전체 시스템 관리, 프로젝트관리자와 일반회원의 인증 및 관리, 공지사항 처리 등을 담당하고, 프로젝트관리자는 프로젝트와 프로젝트 하위단계인 Experiment의 등록과 삭제 등을 할 수 있다. 일반 이용자는 하나의 프로젝트에 속하며 실험(Experiment)만을 등록, 삭제하고 데이터를 업로드 할 수 있도록 권한을 설정하였다 (그림 2).

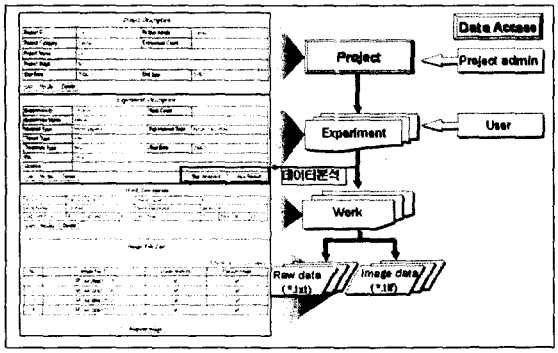


그림 2 cMAMS의 데이터 구조 및 접근 권한

cMAMS에서 마이크로어레이 데이터의 통계적 분석은 편의에 따라 두 가지 경로 중 하나를 선택 할 수 있다 (그림 1). 첫째, 데이터 등록 후 데이터베이스에 데이터를 저장한 다음 필요에 따라 데이터를 선택하여 분석하는 방법이다. 계층적으로 분류, 저장된 마이크로어레이 데이터는 Experiment 단계에서 통계적 분석이 가능 하도록 하였다. Experiment의 하위 단계에 속한 모든 데이터 중 분석을 원하는 것만을 선택하여 통계 분석이 가능하도록 하였다. 둘째, 데이터가 많지 않고 등록절차 없이 간편한 분석을 원할 경우, 사용자가 직접 이미지 분석 파일(Raw 파일)을 업로드 하여 분석하는 방법이다. 사용자 편의성을 고려하여 데이터를 따로 가공하지 않고 Raw 파일을 그대로 처리하게 하여 쉽게 분석할 수 있도록 하였다. 현재 인식 가능한 이미지 분석 소프트웨어의 파일형식은 Imagene, Genepix, Spot 등이다. 모든 분석 결과는 그래픽 이미지와 엑셀파일 형식으로 다운로드 할 수 있다.

cMAMS에서 제공되는 분석방법은 통계언어인 R의 팩키지로 구현되었고, 크게 7 부분으로 구성되었다 [9](그림 3). 구성요소는 전처리(Preprocessing), 정규화(Normalization), 가시화(Visualization), 군집분석(Clustering), 판별분석(Classification), 특이적 발현 유전자 선별(Gene selection), 마이크로어레이 간의 상관분석(Correlation between slides) 등이다 [10, 11](표1, 2).

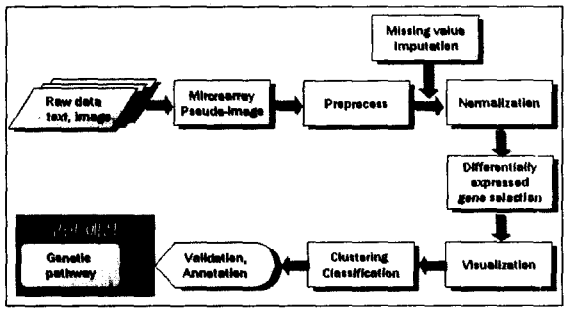


그림 3 cMAMS의 데이터 분석 흐름도

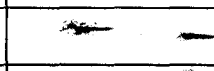
표 1. cMAMS에 구현된 데이터 분석 기능

구분	분석기능
Preprocessing	▪ Flag ▪ Imputation
Normalization	▪ Within-slice normalization ➢ Median, lowess, pin-wise-lowess, scaled pin-wise lowess ▪ Between-slide normalization
Visualization	▪ Pseudo-images ▪ Spot intensity & ratio histogram ▪ PCA (Principal Component Analysis) ▪ MDS (Multidimensional Scaling)
Clustering	▪ Hierarchical ▪ K-means ▪ SOM (Self-Organizing Maps)
Classification	▪ KNN (K-Nearest Neighbor) ▪ SVM (Support Vector Machine)
Gene selection	▪ T-statistics ▪ F-statistics
Correlation	▪ Correlation analysis between slides

전처리 과정에서는 플래그(Flag) 처리, 결측값 대치(Missing value imputation) 등이 가능하다. 정규화에서는 하나의 마이크로어레이 내에서 수행하는 정규화(within-slide)와 여러 개의 마이크로어레이 간에 하는 정규화(between-slide)가 가능하다. 하나의 마이크로어레이 내에서 이뤄지는 정규화는 4 가지 (median, lowess, pin-wise lowess, scaled pin-wise lowess) 중 데이터에 따라 하나를 선택할 수 있다. 가시화는 데이터 분석 후 제공되는 그래픽 기능으로 이미지 분석 후 얻은 스팟 시그널의 Intensity를 정규화 전, 후로 하여 마이크로어레이 이미지, 히스토그램, M-A plot, Box plot 등이 표현된다. 또한 다차원 데이터의 경우 차원을 축소하여 그래픽으로 보여주는 주성분분석(PCA: Principal Component Analysis)과 다차원척도법(MDS: Multidimensional Scaling)을 지원한다. 군집분석에서는 계층적 군집분석(hierarchical clustering), K-평균군집분석(K-means clustering), 자기조직화지도(Self-Organizing Maps) 등이 제공된다. 판별분석에는 K-근접이웃분석(KNN: K-Nearest Neighbor), SVM분석(Support Vector Machine) 등을 지원한다. 두 종류의 실험 조건에서 특이적으로 발현된 유전자를 선별하기

위해 t-검정법과 F-검정법을 이용할 수 있으며 반복실험이나 다른 조건의 마이 크로어레이 간의 상관분석도 가능하다. 분석한 모든 결과는 데이터베이스에 저장되고 언제든지 참고할 수 있도록 하여 연구자들간의 중복된 분석을 줄임으로써 연구효율을 높일 수 있도록 하였다.

표 2. cMAMS를 이용한 마이크로어레이 데이터 분석방법

분석방법	분석결과의 이미지
cDNA microarray image	
Histogram	
Box plot	
Normalization(MA plot)	
Correlation	
Gene selection	
Visualization	
Clustering	

유전자 탐색기능을 추가하여 분석결과 얻은 유전자들에 대한 다양한 정보(유전자이름, 설명, Map 위치, GenBank ID, RefSeq NM, RefSeq NP, Locus ID, Cluster ID, STS ID, SwissProt ID, NCBI GI, OMIM ID)를 확인할 수 있고, 이런 정보가 외부데이터베이스에 링크되어 있기 때문에 보다 자세한 정보 탐색이 가능하도록 하였다. 또한 이들 유전자의 이미지 분석 결과로 얻은 스팟의 수치 값도 함께 검색할 수 있도록 구현하였다.

본 시스템의 데이터베이스는 3가지로 분류되어 있다. 마이크로어레이 실험에 관한 정보, 이미지 및 분석 파일을 저장한 데이터베이스, 통계적 분석 결과를 저장한 데이터베이스, 그리고 유전자 정보 탐색을 위해 NCBI나 SwissProt 등의 다양한 정보를 저장한 참조용 데이터베이스로 구성되어 있다.

3. 결 론

본 시스템은 복잡한 자료구조를 가지는 마이크로어레이 실험에 대한 정보와 각 단계별 처리 정보 등을 사용자의 관점에서 효과적이고 체계적으로 관리, 분석하는 기능을 제공하고 불필요한 시간과 비용을 줄임으로써 마이크로어레이 연구에 도움을 주고자 개발하였다. 또한 하나의 프로젝트에 여러 기관들이 참여한 경우에 각 기관의 실험 정보와 결과를 cMAMS에 등록, 저장 한 후 언제든지 접속하여 실험 데이터의 검토, 분석하고, 그 결과를 데이터베이스화하여 공유할 수 있도록 구현하였다.

cMAMS의 장점은 첫째로 계층적인 데이터 체계를 만들고 접근 권한을 사용자에 따라 부여하여 정보의 효율적 관리성과 보안성을 높였고, 둘째로 프로젝트를 구성하는 Experiment 단위로 데이터를 선택하여 정규화를 포함한 다양한 통계분석이 가능하고 분석 후 결과가 목록으로 남아 있어 추후에 참고 자료로 이용할 수 있다는 것이다. 셋째로 유전자 탐색기능으로 통계분석 후 얻은 유전자에 대해 다양한 정보를 검색할 수 있고 외부데이터베이스와 연결해 보다 자세한 검색이 가능하다는 것이다. 마지막으로 쪽지, 게시판 기능과 같은 사용자 편의기능으로 연구자 상호간의 의견제시, 요구 또는 지시사항 등을 전달하는 수단으로 이용할 수 있고 기록으로 남아 있기 때문에 실험이나 결과 분석 후에 참고자료로 활용할 수 있는 장점이 있다.

앞으로의 시스템 개선계획은 첫째, 국제적으로 통용되는 마이크로어레이 표준데이터 모델인 MAGE-OM (Microarray Gene Expression Object Model)을 XML 형식으로 구현한 표준 데이터 교환 포맷인 MAGE-ML(Microarray Gene Expression Markup Language)을 지원하는 기능을 추가할 예정이다[12]. 둘째, 마이크로어레이 데이터와 유전자들의 SNP정보를 활용하여 특정 질병이나 조건에서 전체유전자의 발현양상을 분석하여 진단에 응용하는 기능을 추가할 예정이다. 그럴 경우 기존의 발현패턴만을 이용해 질병을 예측, 판단하는 것보다 더 높은 신뢰도를 보일 것으로 예상된다. 셋째, 마이크로어레이 데이터에 감추어진 중요 생물학적 정보를 얻기 위해 보다 정확하고 세밀한 유전자 주석 정보, 통계학적 분석 및 검증 방법을 추가할 계획이다. 넷째, 마이크로어레이 결과를 현재 알려진 대사 경로에 적용하여 생물학적 해석을 돕는 GenMAPP 같은 시스템을 응용함으로써 유전자 상호간의 네트워크 분석이 가능한 시스템으로 발전시켜 나갈 예정이다[13]. 현재 연구소 내에서만 cMAMS를 사용할 수 있도록 하였으나 기능을 추가, 보완하여 곧 일반에 공개할 계획이다.

4. 참고문헌

[1] Campbell CJ, Ghazal P. Molecular signatures for diagnosis of infection: application of microarray technology, J Appl Microbiol, 96(1):18-23, 2004.

[2] Forster T, Roy D, Ghazal P.Experiments using microarray technology: limitations and standard operating procedures, Endocrinol. 178(2):195-204, 2003.

[3] BioArray Software Environment (BASE), http://base.thep.lu.se/ index.phtml.

[4] The Stanford Microarray Database (SMD), http://genome-www5.stanford.edu/.

[5] Argus - A new database system for web-based analysis of multiple microarray datasets, http://vessels.bwh.harvard.edu/software/argus/default.htm.

[6] Data management and analysis for gene expression arrays (ArrayDB), http://genome.nhgri.nih.gov/arraydb/.

[7] 이미경, 최정현, 조환규, 마이크로어레이 실험 및 분석 데이터 처리를 위한 통합 관리 시스템의 설계와 구현, 한국미생물생명공학회, Vol 31, No 2, 182-190, 2003.

[8] Ji-Hung Kim, Kyung-Shin Lee, Hwan-Gue Cho, ISMIG : an Integrated System for Microarray data and Genetic network, [Poster]The Korean Society for Bioinformatics, p.315, 2003

[9] The R Project for Statistical Computing, http://cran.r-project.org

[10] Spellman PT, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks WL, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow BJ, Robinson A, Bassett D, Stoeckert CJ Jr, Brazma A. Design and implementation of microarray gene expression markup language (MAGE-ML), Genome Biol, 23:3(9), 2002.

[11] Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis, Trends Genet, 19(11):649-59, 2003.

[12] Krajewski P, Bocianowski J. Statistical methods for microarray assays, J Appl Genet, 43(3):269-78, 2002.

[13] Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, Genome Biol, 4(1):R7, 2003.