Kor. J. Microbiol. Biotechnol.
Vol. 31, No. 2, 182-190 (2003)

마이크로어레이 실험 및 분석 데이터 처리를 위한 통합 관리 시스템의 설계와 구현

이미경[1]* · 최정현[1] · 조환규[1,2]
[1]부산대학교 전자계산학과, [2]부산대학교 컴퓨터 및 정보통신연구소

Design and Implementation of Integrated System for Microarray Data. Lee, Mi-Kyung[1]*, Jeong-Hyeon Choi, and Hwan-Gue Cho. *[1]Department of Computer Science, Pusan National University, Busan, Korea, [2]Research Institute of Computer, Information, and Communiation, Pusan National University, Busan, Korea* – As DNA microarrays are widely used recently, the amount of microarray data is exponentially increasing. Until now, however, no domestic system is available for the efficient management of such data. Because the number of experimental data in a specific laboratory is limited, it is necessary to avoid redundant experiments and to accumulate the results using a shared data management system for microarrays. In this paper, a system named WEMA (WEb management of MicroArrays) was designed and implemented to manage and process the microarray data. WEMA system was designed to include the basic feature of MIAME (Minimal Information About a Microarray Experiment), and general data units were also defined in the system in order to systematically manage the data. The WEMA system has three main features: efficient management of microarray data, integration of input/ouput data, and metafile processing. The system was tested with actual microarray data produced by a molecular biology laboratory, and we found that the biologists could systematically manage and easily analyze the microarray data. As a consequence, the researchers could reduce the cost of data exchange and communication.

Key words : Microarray database, MIAME

마이크로어레이 칩은 수천 개 이상의 유전자 발현 변이를 단 한번의 실험으로 확인할 수 있어서 최근에 각광받는 기술이다. 마이크로어레이 기술이 발전함에 따라 마이크로어레이 이미지 데이터와 이미지 분석 데이터들이 급격히 늘어나고 있다[7, 10]. 그러나 국내에서는 그 데이터들을 효율적으로 관리하기 위한 시스템이 개발되어 공개된 경우가 없다. 그리고 마이크로어레이 실험은 한 실험실에서 분석하고 연구할 수 있는 유전자의 수가 제한되어 있으므로 서로 다른 연구실에서 실험한 연구 결과들을 공유함으로써 실험의 중복을 막을 수 있고 그 연구 결과들을 축적할 수 있다. 또한 마이크로어레이 실험은 마이크로어레이 이미지 생성(마이크로어레이 칩 제작, 샘플 준비, hybridization, 스캐닝), 이미지 분석, 통계 분석, 생물학적 의미 추출이라는 4단계를 거치므로 각 분야의 연구자들이 원만하게 데이터를 교환하고, 긴밀히 의사 소통을 할 수 있어야 실험의 목적에 맞는 연구 결과를 쉽고 효율적으로 추출해 낼 수 있다.

국외에서는 이러한 문제들을 인식하고 마이크로어레이 데이터를 효율적으로 관리하기 위한 시스템이 개발되어 공개된 경우가 많고 그 대표적인 예로 Stanford Microarray Database, ARGUS, ArrayDB, GeneX, Partisan arrayLIMS 등이 있다. Stanford Microarray Database와 ARGUS는 마이크로어레이 데이터를 저장 분류하고 각 스팟을 확대하여 상세 정보와 함께 보여 주는 가시화 기능들을 제공한다[2, 4, 11]. 그러나 생물학자, 전산학자, 통계학자 등과 같은 연구자들간의 의사 소통 기능을 제대로 지원하지 못한다. ArrayDB는 NHGRI(National Human Genome Research Institute)에서 개발한 것으로 분석 결과를 자바 애플릿을 통해 웹으로 확인할 수 있다[1]. 자바 애플릿에는 두 개의 모듈이 제공되는데, Experiment Viewer는 하나의 실험 데이터를 히스토그램 형태로 보여주고, Multi-Experiment Viewer는 여러 번 실험을 수행하는 경우 마이크로어레이에 나타나는 발현 양상의 패턴 변화를 한 눈에 관찰 할 수 있다. ArrayDB는 가시화의 기능은 뛰어나지만 데이터의 계층적인 관리 기능이 약하다. GeneX는 NCGR(National Center for Genome Resources)와 캘리포니아 대학의 Computational Genomics Group에서 개발한 시스템으로 마이크로어레이 데이터 관리 시스템에서 필요한 거의 모든 기능을 제공하고 있다[3]. 그러나 GeneX는 대규모의 마이크로어레이 데이터 관리 시스템을 목적으로 설계된 것으로 일반적인 소규모의 생물학 연구실에서 설치하여 사용하기에

*Corresponding author
Tel: 82-42-864-2524, Fax: 82-42-866-9241
E-mail: mklee@smallsoft.co.kr

는 어려운 점이 있다. Partisan arrayLIMS는 마이크로어레이 실험과 분석의 전과정에서 생성되는 데이터를 효율적으로 관리할 수 있는 LIMS(Laboratory Information Management System)를 목표로 설계된 시스템이다[9]. 그것은 마이크로어레이 데이터를 projects, experiments 등으로 나누어서 계층적으로 관리하고 있으며, 데이터의 보안을 위해서 사용자별로 데이터의 소유권을 관리하는 등 실제로 마이크로어레이 데이터를 관리하기 위한 편의 기능을 제공하고 있다. BASE(BioArray Software Environment)는 다른 데이터 서버 시스템과 마찬가지로 마이크로어레이 관련 데이터를 저장 관리하기 위한 시스템으로 마이크로어레이에 사용된 유전자 정보와 실험 정보를 효율적으로 저장하도록 설계되었다[5]. Table 1은 마이크로어레이 데이터 서버 시스템을 기능별로 비교한 것이다.

본 논문은 국내에서 처음으로 소규모의 생물학 연구실에서 마이크로어레이 실험과 분석에 관련된 데이터를 효율적으로 관리하고 그것의 통합적 분석을 용이하게 할 수 있는 웹 기반의 시스템을 소개한다.

System and Methods

WEMA(WEb management of MicroArray) 시스템은 마이크로어레이 관련 데이터의 처리 및 관리를 효율적으로 하기 위해 웹 환경에서 운영되는 시스템이다. 이 시스템은 리눅스 환경에서 자바로 개발되었고 웹과 자바 환경을 지원하는 모든 시스템에서 사용 가능하며 소규모의 마이크로어레이 데이터 서버 시스템으로 사용할 수 있다.

Fig. 1은 WEMA 시스템의 주요 화면이다. 그림의 왼쪽은 WEMA의 기본 메뉴를 보여주는 것으로 Projects, Community가 있다. Projects에서는 현재까지 등록된 project, experiment, work의 목록이 있고, Community에서는 의사소통을 위한 게시판, 자료실 등이 있다. Table 2는 WEMA의 메뉴들을 정리한 것이다. 그림의 가운데는 선택된 project의 계층적 구조를 트리 형태로 보여주는 것으로, WEMA의 데이터 단위인 experiment와 work 들에 대한 리스트로 구성된다. 그림의 오른쪽은 선택된 experiment나 work에 대한 상세한 정보가 보여진다. Fig. 1에 나타난 정보에는 Work에 등

Table 1. Comparison among microarray data management systems.

기 능	상세 기능	SMD	Array DB	GeneX	array LIMS	WEMA
데이터 저장 및 관리	실험 관련 파일	O	O	O	O	O
	이미지 파일	O	O	O	O	O
	이미지 분석 파일	O	O	O	O	O
	통계 분석 파일	X	X	O	O	O
	생물 분석 파일	X	X	O	X	O
가시화	스팟의 확대 및 상세 정보	O	O	O	X	X
	클러스터링 결과	O	X	O	X	X
검색	유전자 정보 파일, 이미지 분석 결과 파일 검색	O	O	O	O	X
	검색 결과 파일로 적기	O	X	O	X	X
표준화	데이터 표준화	O	X	O	O	X
	Ontology	X	X	O	O	△
데이터 처리	정규화	O	X	O	O	X
	반복 실험 데이터 처리	X	X	O	X	X
	메타 파일 생성	X	X	X	X	O
	통계 프로그램과 연동	X	O	O	X	X
웹 연동	외부 저장소와 연결	O	O	O	O	X
	이미지 분석 툴과 연결	X	X	X	O	△
사용자 편의 기능	계층적 데이터 관리	O	△	O	O	O
	사용자 관리 및 보안	O	△	O	O	O
	일정 및 리포팅 기능	X	X	X	O	O
	데이터 백업 및 복구	X	X	X	X	O
	파일 이름 자동 부여	X	X	X	X	O

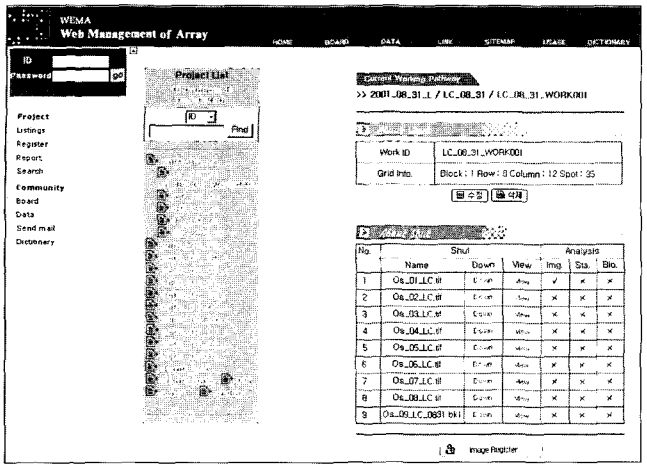

Fig. 1. WEMA main Interface. The left of the figure shows the main menu, the center shows the hierarchical structure of the registered project, the right shows the detailed Information about a selected project, experiment, or work.

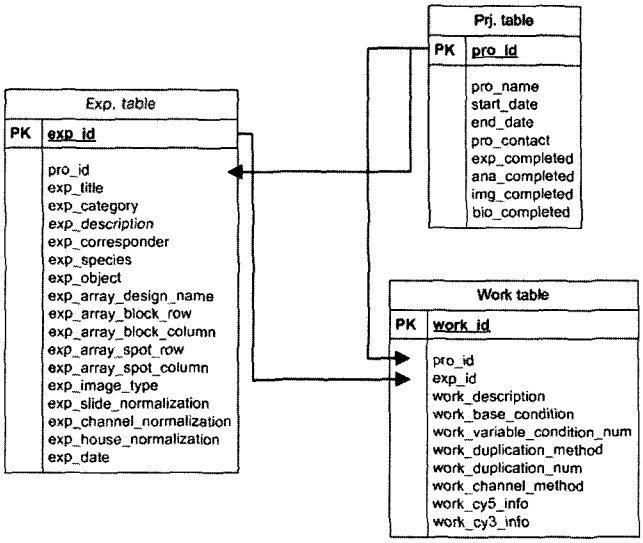

Fig. 2. Data Schema of WEMA. WEMA has 3 tables: Prj., Exp., and Work. Each table takes a identifier(id) as a primary key(PK). The arrow means the relation between two tables.

록된 마이크로어레이 이미지 및 각각의 분석 파일들을 볼 수 있으며 이미지를 다운 받아 보거나 웹을 통해서 이미지를 볼 수도 있다. 분석 파일이 업로드 된 경우와 업로드 되지 않은 경우에 해당 아이콘을 통해 쉽게 확인할 수 있으므로 작업의 진척도를 쉽게 확인할 수 있다.

데이터 스키마

마이크로어레이 실험은 마이크로어레이 칩 제작, 샘플 준비, hybridization, 스캐닝, 이미지 분석의 5단계로 이루어지는데 각 단계별로 실험에서 채택하는 실험 방법이나 protocol들은 다양하다. 따라서 실험의 각 단계에서 생성되는 정보가 이후 연구 결과의 재해석과 재사용성을 위해 정확히 기술되어서 저장되어야 한다. 이를 위해 MGED(Microarray Gene Expression Data Society)에서는 MIAME(Minimal Information About a Microarray Experiment)로 불리는 모임을 결성하여 마이크로어레이 데이터의 관리 및 분석을 용이하게 하기 위해 마이크로어레이 실험에 수록되어야 할 최소한의 정보들을 정의하였다[8]. WEMA 시스템은 마이크로어레이 데이터를 정확하게 기술하고 저장하기 위해서 MIAME의 정보를 바탕으로 데이터 스키마를 설계하였다. 다음은 MIAME에 포함되어 있는 정보를 나타내고 Fig. 2는 WEMA에서 정의한 데이터 스키마이다.

1. Experimental Design : 실험에 관련된 일반적인 정보들을 나타내는 것으로 실험자, 실험 종류 등을 포함하며 Project, Experiment 테이블에 정의되어 있다.
2. Array Design : 마이크로어레이 제작 시에 사용된 어레이 디자인 및 protocol 정보를 나타내는 것으로 Experiment 테이블에 정의되어 있다.
3. Samples : 마이크로어레이에 심어진 유전자 정보를 나타내는 것으로 유전자 정보 파일을 통해 저장된다.
4. Hybridization : 마이크로어레이 실험을 할 때 hybridization에 사용된 방법에 관한 정보를 나타내는 것으로

Table 2. Menu of WEMA system.

메뉴	하위 메뉴	설 명
Project	Listings	현재 등록된 project, experiment, work 목록을 볼 수 있는 화면.
	Register	Project를 등록할 수 있는 화면.
	Report	현재 등록된 project, experiment, work, shot 등에 관한 정보를 문서 형태로 출력해 주는 화면
	Search	등록된 데이터를 검색하는 화면
Community	Board	연구자들이 의견을 나눌 수 있는 게시판
	Data	연구자들이 자료를 공유할 수 있는 자료실
	Send Mail	연구자들 사이에 메일을 교환할 수 있는 화면
	Dictionary	마이크로어레이 실험과 분석에 관련된 혼란스러운 용어들을 정의해 두는 화면

Project menu consists of listings, register, report, and search. Community menu consists of board, data, send mail, and dictionary.

실험 방법을 기술한 파일에 기록된다.

5. Measurements : 마이크로어레이 이미지 생성시 스캐닝 방법 및 이미지 분석 방법을 나타내는 것으로 실험 방법을 기술한 파일에 기록된다.
6. Controls : 정규화 방법에 관한 정보를 나타내는 것으로 Experiment 테이블에 정의되어 있다.

데이터 단위

WEMA 시스템에서는 마이크로어레이 데이터를 계층적으로 관리하고 마이크로어레이 제작의 실험 방법과 protocol에 따른 여러 변수들을 고려하기 위해서 다양한 실험 방법을 통합해서 나타낼 수 있는 공통적인 데이터 단위를 정의하였다. Fig. 3은 마이크로어레이 데이터 관리를 위한 단위를 표현한 것이다. 데이터의 단위는 Shot, Line, Work, Experiment, Project은 5개로 나누어진다.

1. Shot : 마이크로어레이 데이터의 가장 작은 단위로서 물리적으로 R,G 채널을 가진 마이크로어레이 이미지를 나타낸다. 마이크로어레이 이미지에서 R,G를 따로 분리하여 실험한 경우에는 하나의 shot이 물리적으로 2개의 이미지가 된다.
2. Line : 하나의 실험 조건인 기본 조건(base condition)에 변화 조건(variable condition)를 가한 shot들의 셋이다. 예를 들면 기본 조건은 heat shock이고 변화 조건이 시간이면 line의 각 shot은 heat shock을 시간대별로 가해서 얻어진 이미지이다. 첫번째 shot은 1시간 후, 두 번째 shot은 2시간 후, 여섯번째 shot은 6시간 후를 의미한다. 이렇게 실험해서 얻어진 6개의 이미지를 모두 합쳐서 line이라고 한다.
3. Work : 기본 조건이 같고 마이크로어레이 칩 디자인이 같은 shot들의 셋이다. 예를 들면 heat shock를 실험 조건으로 하고 같은 어레이 디자인 구조 즉, 4×4의 블록과 각 블록 내에 12×12의 스팟을 가지는 어레이에 같은 유전자들을 심어놓고 실험한 이미지들을 모두 합쳐서 work라고 한다.
4. Experiment(Exp.) : 서로 같은 어레이 디자인을 가진 마이크로어레이 칩에 다양한 기본 조건(cold stress, treatment)으로 실험한 work들의 셋이다. 서로 같은 마이크로어레이 칩 디자인을 가진 shot에 다양한 실험 조건 즉, heat shock, cold stress 등을 주어서 실험을 한 work들을 모두 합쳐서 Experiment라고 한다.
5. Project(Prj.) : 생물학자들이 실험을 통해 최종적인 결과를 얻기 위해 하는 일련의 모든 실험 과정으로 Experiment가 하나 혹은 그 이상 모인 것으로 마이크로어레이 디자인이 다른 즉, 다른 유전자가 심어져 있는 마이크로어레이 칩으로 실험한 Experiment들의 셋이다.

위의 단위를 통해서 마이크로어레이의 특정한 데이터 셋을 표현하는 방법은 다음과 같다.

Chip (Prj 이름, Base Condition 이름, Variable Condition 이름, 채널 이름, Replication 번호)

예를 들면, Chip(Ga, bio, HeatShock, Var5, R, Rep3)는 Ga 프로젝트의 bio라는 Experiment에서 heat shock을 5시간 가한 실험에서 세 번째 반복 실험한 R 채널의 shot, Chip(Ga, bio, Temp, *, R, *)는 Ga 프로젝트의 bio라는 Experiment에서 temperature를 조건으로 하는 work 중에서 R 채널의 shot를 나타낸다.

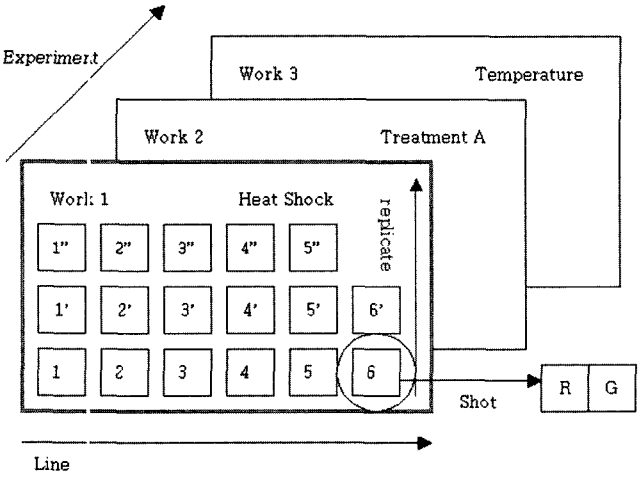

Fig. 3. The data unit of WEMA: Project, Experiment, Work, Line, and Shot. The whole picture shows 1 project composed of 1 experiment because 6 shots have a same microarray design. Experiment consists of 3 works which have a 3 base condition(heat shock, treatment A, temperature). Each work has 3 replicate microarrays and 3 lines. Each line has 6 shots.

WEMA의 주요 기능

WEMA 시스템은 마이크로어레이 관련 데이터를 저장 및 분류하고 있으며 그 외에 마이크로어레이 분석에 필요한 데이터 처리 및 검색 등을 지원한다. WEMA 시스템에서 제공하는 기능은 다음과 같다.

데이터 저장 기능

이미지 데이터와 분석 데이터를 저장, 분류 및 관리한다. 마이크로어레이 데이터는 다음의 네 가지로 나뉘어진다.

1) 마이크로어레이 실험에 관련된 데이터
   a. 실험 방법을 기술한 데이터
   b. 실험에 사용된 clone의 id 및 accession number를 기술한 데이터
2) 마이크로어레이 이미지 데이터 : 실험을 통해 만들어진 마이크로어레이 이미지

3) 마이크로어레이 이미지 분석 데이터 : 이미지 데이터로부터 각 스팟의 강도, 분산, R/G ratio를 분석한 결과 파일
4) 마이크로어레이 클러스터링 결과 데이터 : 이미지 분석 데이터로서 각 조건에서 유사한 발현 패턴을 보이는 유전자들을 묶어서 그에 관한 정보들을 기록한 데이터

WEMA 시스템은 위의 데이터들을 계층적인 구조에 따라 효율적으로 관리할 뿐만 아니라, 다른 분석 프로그램을 통해 만들어진 분석 데이터들은 키워드로 분류하여 저장한다. 예를 들어 마이크로어레이 이미지 분석 시에 GenePix와 ImaGene 프로그램을 사용하는 경우 각 2개의 이미지 분석 파일이 생성된다. WEMA에서는 여러 개의 분석 파일을 키워드와 함께 저장하도록 설계하였다.

파일 이름 자동 생성(Auto File Naming)

생물학자가 마이크로어레이 실험을 하고 그 이미지에 적절한 이름을 부여하지 않은 경우에 특정한 이미지 데이터를 나중에 찾으려면 많은 시간을 낭비하게 된다. 대부분의 생물학자들이 데이터를 저장할 때 실험의 조건을 고려하지 않고 날짜, 알파벳 및 숫자등을 데이터 이름으로 정의해 둔다. 실제로 실험 결과 데이터에 적절한 이름을 부여하는 것은 생물학자들로 하여금 실험 외의 노력과 시간을 낭비하게 하는데, 이와 같은 문제를 극복하기 위해서 WEMA 시스템은 사용자가 데이터를 업로드 할 때, 데이터의 특징을 가장 잘 나타낼 수 있는 이름을 데이터에 자동으로 부여해 준다. 데이터에 이름을 부여하는 방법은 앞에서 설명했던(Prj. 이름, Exp 이름, Base Condition 이름, Variable Condition 이름, 채널 종류, 반복 횟수) 방식이다. 예를 들면 Ga_bio_HeatShock_Var5_R_Rep3는 Ga 프로젝트의 bio라는 Experiment에서 heat shock을 5시간 가한 실험에서 세 번째 반복 실험한 R 채널의 shot이고, Ga_bio_Temp_Var0_G_Rep5는 Ga 프로젝트의 bio라는 Experiment에서 온도를 0으로 한 work 중에서 다섯번째 반복 실험한 G 채널의 shot을 나타낸다.

데이터 입출력 통합 기능

마이크로어레이에서 생물학자들이 최종적으로 의미 있는 결과를 도출하기 위해서는 마이크로어레이 실험, 이미지 분석, 통계 분석, 생물학적 의미 도출이라는 4단계를 거쳐야 한다. 위와 같은 작업은 서로 다른 연구자들간의 의사 소통과 긴밀한 협력을 바탕으로 이루어져야 하는데, 서로 다른 작업 환경과 학문적 배경을 가진 연구자들 사이에서는 혼란이 야기될 수 밖에 없다. 이와 같은 문제를 극복하기 위해서는 서로 약속된 하나의 의사 소통 공간이 필요하며 서로 통일된 방법을 통해서 데이터의 교환과 분석이 이루어져야 한다.

WEMA 시스템에서는 생물학자, 전산학자, 통계학자들이 서로 혼란을 야기하지 않도록 WEMA 시스템을 통해 데이터를 교환하도록 하였다. 마이크로어레이를 실험하는 연구자는 이미지를 생성하고 그것을 앞에서 정의한 데이터 단위

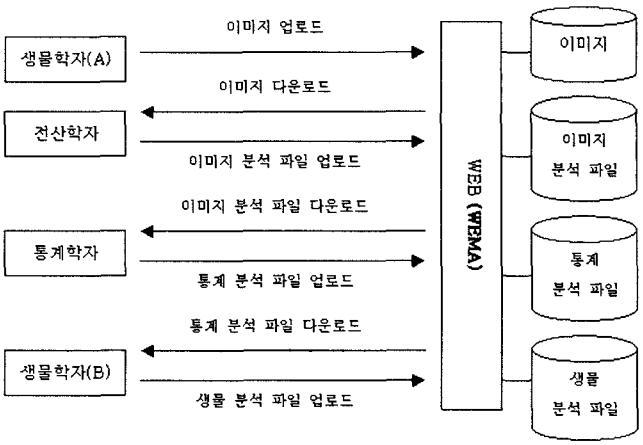

Fig. 4. Integration of input/output data and collaboration for microarray analysis through WEMA. All researchers exchange the data without confusion and easily communicate through WEMA.

에 맞게 데이터를 업로드 한다. 그리고 이미지 분석을 담당하는 연구자는 마이크로어레이 실험자가 올린 이미지를 다운로드 받아서 각각의 이미지 분석 프로그램을 이용하여 결과 파일을 생성하여 업로드 한다. 그 후에 통계 담당자는 이미지 분석 결과 파일을 다운 로드 받아 그것을 처리하고 결과 파일을 업로드 하게 된다. 이러한 방법은 각 분야의 연구자들이 작업의 진척도와 일정들을 일일이 연락해서 확인해야 하는 불편함과 시간적 낭비들을 해결해 줄 수 있다. 그 외에 community의 게시판을 이용하여 연구자들 사이에 의견을 나눌 수 있으며 자료실을 이용해서 필요한 정보 및 자료를 공유할 수 있다(Fig. 4).

메타 파일 생성(MeatFile Processing)

마이크로어레이 실험과 분석을 통해서 여러 개의 데이터가 만들어지면 생물학자들은 그 데이터들의 조합을 통해서 최종적으로 생물학적인 의미를 도출하게 된다. 예를 들어 특정한 유전자의 시간대별 발현 양상을 관찰하고자 하는 경우 각 시간대별 파일을 모두 확인해야 하는 번거로움이 있다. 그리고 실험의 정확성을 위해서 반복 실험을 한 경우 반복 실험 마다 만들어지는 분석 파일들에서 Cy3와 Cy5의 ratio를 살펴보고자 하는 경우에도 모든 분석 파일을 찾아보아야 한다.

WEMA 시스템은 위와 같은 문제를 극복하고자 사용자가 원하는 데이터의 내용을 조합하는 메타 데이터를 생성하고 사용자가 쉽게 데이터를 분석할 수 있도록 한다. 사용자가 발현 양상을 살펴보고자 하는 유전자, 알고자 하는 발현 정보(R/G 강도, R/G Ratio)의 요소 및 이미지 분석 파일을 선택하면 WEMA 시스템은 사용자가 선택한 유전자의 발현 정보를 선택한 파일에서 골라서 하나의 새로운 메타 파일로 만들어 준다. 생성된 메타 파일은 텍스트 파일 형태로 엑셀이

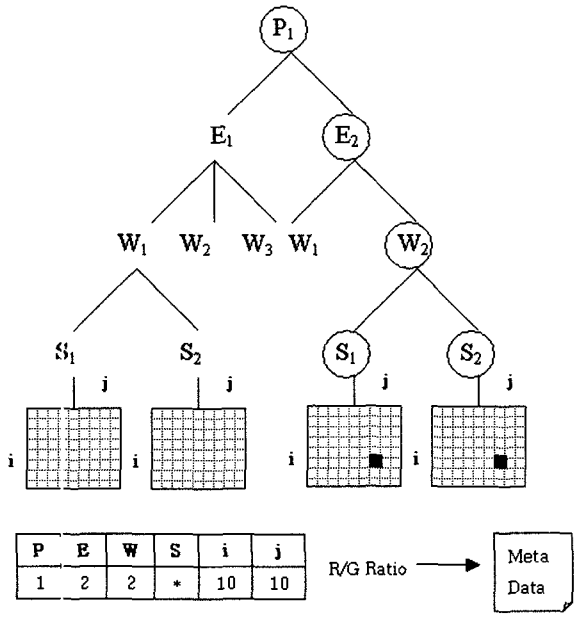

Fig. 5. MetaFile Processing: P - Project, E - Experiment, W - Work, S - Shot. If we want to know the expression level of the gene located in (10,10) for all shots in circled W2, metafile processing makes the metafile combining 2 circled shots' image analysis file.

나 텍스트 편집기 등을 통해서 쉽게 확인할 수 있고 분석 프로그램의 입력으로 사용될 수 있다. Fig. 5는 메타 파일 생성을 예를 들어 표현한 것이다.

사용자 편의 기능

WEMA 시스템은 생물학자들이 실험 외적으로 드는 시간과 노력을 절약하기 위해서 여러 가지 사용자 편의 기능들을 제공한다. 첫째, 리포팅 기능은 사용자가 현재 작업의 진행 상황 즉, 업로드 된 이미지 및 분석 파일의 수 등을 알기 쉽게 문서 형태로 출력해 준다. 리포팅 기능을 이용하면 생물학자는 모든 연구자들에게 연락을 취해서 확인하지 않고도 손쉽게 작업의 진척도를 알 수 있으며 그것을 문서화 할 수 있다(Fig. 6).

둘째, 데이터의 백업 및 복구 기능이다. 마이크로어레이 데이터는 실험에 드는 비용이 적지 않고 실험의 방법이 까다로우므로 데이터를 분실하는 경우 금전적, 시간적 손실이 크다. 이에 대비하여 데이터를 백업해 두고 이후에 백업된 데이터를 복원할 수 있는 기능을 제공한다. 생물학자들이 수작업으로 데이터를 디스크나 CD 등에 복사하지 않고도 웹을 통해 한 번의 클릭으로 데이터를 손쉽게 복구할 수 있으므로 상당히 편리한 기능이라 할 수 있다. 그리고 기존의 WEMA 시스템에 있던 마이크로어레이 관련 데이터를 시스템 고장 등의 이유로 다른 WEMA 시스템으로 옮기고자 할 때, 데이터베이스의 지식을 가지고 있지 않은 생물학자의 경우에는 어려움을 겪을 수가 있다. WEMA 시스템에서는 이

Report on Project Progress

by on 2002-12-27

Project ID	2001_07_01_L	Project 이름	GFP expression
Project 시작일	2001-07-01	Project 책임자	bcchung
설명	2001_07_01_L		

Experiment 1

Experiment ID	CO_07_01	Experiment 이름	Control
Experiment 날짜	2001-07-01	Experiment 책임자	bcchung
Treatment	CO	Shut 수	6
Experiment 대상	plant	세부대상	Os
설명	Control		
Work	Shut		
CO_07_01_WORK001	Os_06_CO_0701.tif, Os_05_CO_0701.tif, Os_04_CO_0701.tif, Os_03_CO_0701.tif, Os_02_CO_0701.tif, Os_01_CO_0701.tif,		

Fig. 6. Reporting on information about a registered project, experiment, work, and shot.

같은 문제에 대비하여 손쉽게 데이터를 옮길 수 있는 transfer 기능을 제공한다. Transfer 기능을 이용하면 웹을 통해 쉽게 데이터를 옮길 수 있게 된다. 그리고 WEMA 시스템이 아닌 다른 서버에 저장되어 있던 데이터를 Import 기능을 통해 WEMA로 쉽게 데이터를 옮길 수 있다. 저장되어 있던 데이터의 구조와 WEMA 시스템의 데이터 구조를 파일로 정의해 두면 WEMA 시스템은 그 파일을 읽어서 WEMA 구조에 맞게 데이터를 옮겨주게 된다.

WEMA에 추가할 기능

Biological Q & A

생물학자들이 마이크로어레이 분석을 통해서 의미있는 결론을 얻기 위해서는 이미지 분석 파일을 여러 차례 분석하고 의미있는 결과들을 축적해 두어야 한다. 생물학자들이 최종적으로 얻고자 하는 정보는 "유전자 A가 target1과 target2 중에서 어떤 것과 더 반응을 많이 하였는가?", "어떤 유전자들이 target1에서 더 많이 발현되었는가?", "유전자 A의 시간대별 발현 양상은 어떠한가?" 등과 같은 것인데, 이러한 질문들을 일반화하여서 마이크로어레이 데이터의 분석 모델을 설계하고 마이크로어레이 데이터 분석을 용이하게 한다. 그리고 이러한 질문과 답들은 파일로 적어 두어서 이 후에 다른 연구자가 중복적으로 분석하는 일이 없도록 한다.

검색

마이크로어레이 데이터는 엄청나게 많은 이미지와 분석 데이터로 이루어져 있다. 그 데이터에서 의미 있는 결과를 도출하기 위해서는 데이터 필터링과 검색이 필수적이다. 대부분의 마이크로어레이 데이터 처리 시스템들이 다양한 기준으로 마이크로어레이 데이터를 검색할 수 있도록 한다. 검색의 조건은 앞에서 정의한 데이터 단위(Project, Experiment, Work) 뿐만 아니라, 키워드(사용자가 project 생성시

등록한 키워드), 유전자의 이름, 이미지 분석 결과 파일에서 특정 강도 이상의 스팟들을 검색하여 보여줌으로써 사용자가 편리하게 생물학적인 의미를 도출할 수 있도록 도와준다.

정규화 및 반복 실험 데이터의 처리

마이크로어레이 실험은 실험 환경이나 마이크로어레이 이미지를 스캐닝 할 때의 차이 등으로 인해 똑같은 조건과 유전자로 실험을 하더라도 각 스팟들에 대한 발현 정도의 값이 일정하지는 않다. 이러한 문제점을 최소화하기 위해서 발현 정도 값의 정규화 (normalization)이 필요하게 된다. 일반적인 정규화 방법은 전체 데이터의 평균값이 1이 되도록 전체 데이터의 값을 일정한 비율로 수정하는 방법인데, 마이크로어레이 데이터 처리 시스템은 마이크로어레이 이미지 분석 데이터를 정규화하여 각 스팟의 강도를 보정한다. 그리고 마이크로어레이 실험의 경우 데이터의 정확성을 위해서 반복 실험을 하는 경우가 많은데, WEMA 시스템은 반복 실험한 여러 개 데이터의 평균 값으로 하나의 데이터를 만들어 준다.

Application and Results

WEMA 시스템을 통해 실제로 한 분자 생물학 연구실에서 연구하고 있는 생물 정보학 프로젝트의 마이크로어레이 이미지 및 분석 데이터를 관리하였다. 이 연구실에서 수행한 실험은 특정한 식물의 유전자에 특정한 온도에 의한 스트레스를 8번의 시간대별로 가하고 유전자들의 발현 변화 양상을 조사하는 것이다. 실험에 사용된 마이크로어레이는 4×4의 블록을 가지며 각 블록에 10×10의 유전자를 심어놓아서 총 1600개의 유전자를 가진다. 이 실험에서는 데이터의 정확성을 위해서 2번의 반복 실험과 dye swap을 수행하였다. 이러한 방식으로 생성된 데이터는 TIFF 타입의 이미지 64장과 실험에 사용된 유전자의 이름과 실험 정보들을 기술한 파일들이다. 그리고 실험에서 생성된 64개의 이미지를 이미지 분석 프로그램을 통해 각 스팟의 강도를 수치화하여 이미지에 해당하는 각각의 분석 결과 파일을 생성하였다. 이 실험과 분석에 WEMA 시스템을 사용하여 연구를 수행하였다.

먼저 생물학자들은 실험의 목표를 세우고 그 목표에 맞는 실험 방법을 선택하고 마이크로어레이에 심을 유전자들을 설계하였다. 생물학자들은 이후에 그 실험을 재사용할 수 있도록 실험 방법을 정확히 기술한 데이터와 마이크로어레이에 심은 각 유전자 정보 파일들을 생성하였다. 그리고 실험을 통해 64개의 마이크로어레이 이미지가 생성되었다. 생물학자들은 마이크로어레이 관련 데이터를 WEMA를 통해 관리하기 위해서 Fig. 7과 같이 Project를 생성하였다. Project에는 실험의 제목, 목적 등의 일반적인 정보들을 등록하였다.

다음으로 생성한 project 아래에 Experiment를 생성하였는

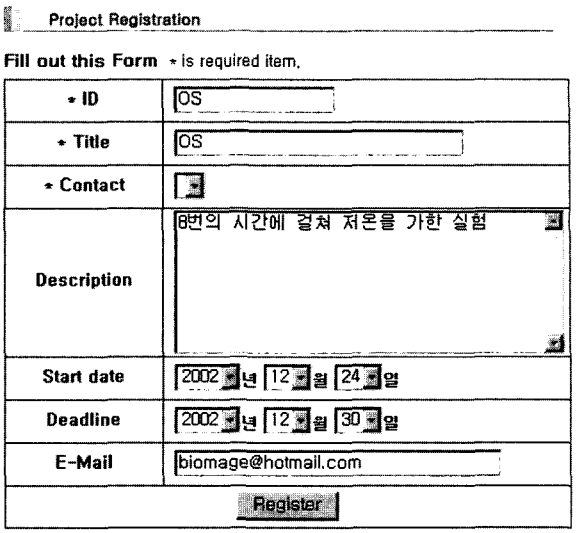

Fig. 7. Interface for registration of Project. Users input the general information about a project: id, title, contact, date etc.

데, 실험의 상세한 정보와 마이크로어레이 디자인 정보 등을 등록하였다. 이 연구실은 실험의 정확성을 위해 R,G 채널을 바꿔서 2번의 실험을 수행하였으므로 마이크로어레이 디자인이 다른 두 개의 Experiment를 생성하였다(Fig. 8).

각각의 Experiment 아래에는 8번의 시간대별로 저온을 가

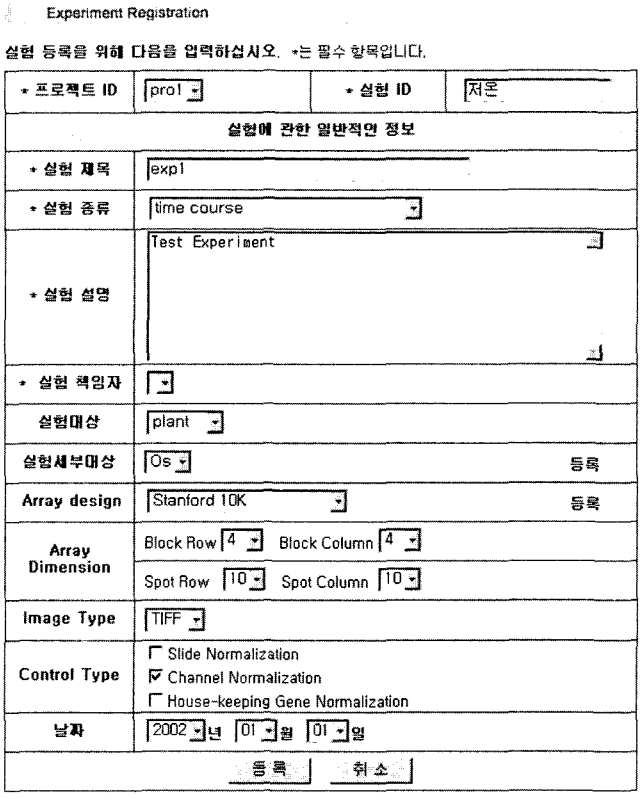

Fig. 8. Interface for registration of Experiment. Users input the detailed information about a experiment: id, title, kinds, microarray design, image type etc.

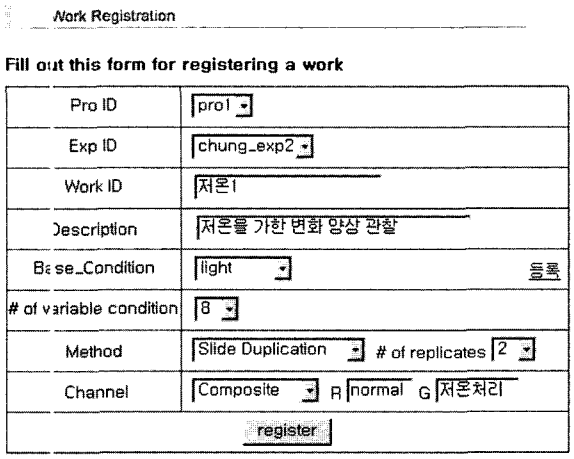

Fig. 9. Interface for registration of work. Users input the detailed information about a work: id, description, base condition, variable condition, method of duplication etc.

해서 8개의 shot이 있는 work를 생성하였다(Fig. 9). 다음으로 생물학자는 마이크로어레이 실험을 통해 각각의 shot에 맞는 이미지를 업로드 하였다. 업로드 되는 이미지 데이터는 데이터의 특성에 맞게 자동으로 이름이 부여되므로 수작업으로 데이터를 관리하는 데서 오는 혼란을 방지할 수 있었다.

전산학자들은 생물학자들이 올려놓은 이미지를 웹을 통해 다운로드 받아서 이미지 분석 프로그램으로 분석하고 그 결과를 WEMA에 업로드 하였다. 위와 같은 중앙 서버에 의한 데이터 관리는 생물학자, 전산학자, 통계학자들이 데이터 교환을 위해 불필요하게 연락을 취하거나 의사 소통을 하는 시간을 절약할 수 있게 하였다.

WEMA 시스템을 채택하기 전에 생물학자들은 수작업으로 이미지 및 분석 데이터들을 관리하였다. 수작업으로 데이터를 관리하는 경우에는 실험 방법에 따라 이미지를 구분하기도 어려웠을 뿐더러 실험 외적으로 드는 많은 시간과 노력을 낭비하게 하였다. 그러나 WEMA 시스템을 도입 한 후 생물학자는 데이터 분류 및 관리에 드는 상당한 시간을 절약할 있었다. 또한 WEMA 시스템을 사용하여 전산학자와 통계학자 등과 공동 연구를 통일된 방법으로 원활히 수행할 수 있어 좀더 빠르고 통합적인 연구 결과를 얻을 수 있다.

Discussion

본 논문에서는 마이크로어레이 실험 및 분석 데이터들을 효율적으로 관리하기 위한 WEMA 시스템을 구현하였다. WEMA 시스템을 사용함으로써 다음과 같은 장점을 가졌다.

1. 컴퓨터 사용과 데이터 관리에 익숙하지 않은 생물학자들이 상당한 시간과 노력을 절약할 수 있었다.

2. 생물학자, 전산학자, 통계학자들이 WEMA 시스템을 통해서 데이터 교환 및 의사 소통을 원활히 할 수 있어서 연구자들 사이의 혼란을 방지할 수 있었다.

3. 데이터의 공유 및 축적으로 중복 연구를 막을 수 있으며, 이후 연구에 재사용될 수 있다.

4. 데이터 프로세싱을 통해서 특정 유전자의 발현 양상을 한번에 쉽게 파악할 수 있었다.

5. 데이터의 백업 및 리포팅 기능으로 실험 외에 드는 부가적인 시간과 노력을 절약할 수 있었다.

앞으로 계속 연구해야 할 과제는 마이크로어레이 데이터 표준안의 생성이다. GenBank 데이터베이스와 같이 공개된 데이터베이스에 보고되는 DNA나 단백질 서열 데이터의 경우에는 각 데이터베이스에 표준화된 형식으로 입력, 저장되고 있으므로 각 연구 그룹들이 데이터를 상호 공유할 수 있는 반면에 마이크로어레이에 관련된 실험을 하는 연구자들은 모두 다양한 형태의 마이크로어레이 기술과 분석 도구들을 사용하기 때문에 다른 연구자들이 그 데이터를 서로 공유하고 이용하는 데 한계가 있다. 다시 말해서 한 연구실에서 얻은 데이터를 다른 연구실에서 실험한 데이터와 비교할 수 있는 표준이 정해져 있지 않다. 이러한 문제를 극복하기 위해서 국외에는 마이크로어레이 실험에 대한 정보를 교환하기 위한 XML 표준안인 MAGE-ML 등이 만들어져 있다[6]. WEMA 시스템에도 마이크로어레이 데이터를 가장 잘 표현하고 공유할 수 있는 데이터 표준화 작업을 이루어야 한다.

요 약

마이크로어레이 기술이 널리 이용됨에 따라 마이크로어레이 이미지 데이터와 이미지 분석 데이터들이 급격히 늘어나고 있다. 그러나 국내에서는 그 데이터들을 효율적으로 관리하기 위한 시스템이 개발되어 공개된 경우가 없다. 그리고 마이크로어레이 실험은 한 실험실에서 분석하고 연구할 수 있는 유전자의 수가 제한되어 있으므로 서로 다른 연구실에서 실험한 연구 결과들을 공유함으로써 실험의 중복을 막을 수 있고 그 연구 결과들을 축적할 수 있다. 본 논문에서는 마이크로어레이 이미지 데이터를 처리 및 관리하기 위한 통합 시스템, WEMA(Web management of MicroArray)를 개발하였다. WEMA는 마이크로어레이 데이터 표준 규정의 제안인 MIAME(Minimal Information About a Microarray Experiment)에서 정의한 데이터 요소를 바탕으로 데이터 스키마를 설계하였으며 마이크로어레이 실험 설계에 따라 체계적으로 데이터를 관리하기 위해서 공동적인 데이터 단위를 정의하였다. WEMA의 주요 기능은 마이크로어레이 이미지 및 분석 데이터의 효율적인 관리, 데이터 입출력의 통합 기능, 메타 파일 생성 등이다. 본 WEMA 시스템을 이용해서 실제로 한 식물 분자 생물학 연구실에서 만들어내는 마이크로어레이 이미지 데이터를 처리, 관리한 결과 생물학자들이 마이크로어레이 데이터를 체계적으로 관리, 분석할 수 있었으며 연구자들간의 데이터 교환 및 의사 소

통이 원활히 이루어졌다.

감사의 말

본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어진 것임(02-PJ1-PG3-51207-0001).

REFERENCES

1. ArrayDB Software. http://genome.nhgri.nih.gov/arraydb/.
2. Gavin, S., H. B. Tima, K. Andrew, *et al.* 2001. The Stanford microarray database. *Nucleic Acids Research* **29(1)**: 152-155.
3. Harry, M., S. Jason, Z. Jiaye, *et al.* 2001. GeneX : An open source gene expression database and integrated tool set, *IBM Systems Journal* **40:** 552-569.
4. Jason, C., M. W. Griffin, A. G. Michael, *et al.* 2001. Argus–A new database system for web-based analysis of multiple microarray data sets. *Genome Research* **11**: 1603-1610.
5. Lao, H. S., T. Carl, V. C. Johan, *et al.* 2002. Bioarray software envir onment(base) : a platform for comprehensive management and analysis of microarray data. *Genome Biology* **3**: software0003.1-0003.6.
6. MAGE-ML (MicroArray and GeneExpression - Markup Language).
http ://www.mged.org/workgroups/mage/mage-ml.html.
7. Mark, S. 2000. *Microarray Biochip Technology.* Eaton publishing.
8. MIAME (Minimum Information About A Microarray Experiment).
http://www.mged.org/Workgroups/MIAME/miame.html.
9. Partisan arrayLIMS. http://www.clondiag.com.
10. Pierre, B. and G. W. Hatfield. 2002. DNA Microarrays and Gene Expression from experiments to data analysis and modeling. CAMBRIDGE University Press.
11. Stanford Microarray Database.
http://genome-www5.stanford.edu/microarray/smd.

(Received Dec. 28, 2002/Accepted May 25, 2003)