유전자 온톨로지를 이용한 마이크로어레이 데이터의 유전자 기능 분석 시스템의 개발

이종근[O]* 박성수* 홍동완** 윤지희*
* 한림대학교 컴퓨터공학과
** 송곡대학 미디어컨텐츠과
{jeikei[O], sspark}@hallym.ac.kr dwhong@songgok.ac.kr, jhyoon@hallym.ac.kr

Development of a Gene's Functional Classifying System for a Microarray Data using a Gene Ontology

Jong-Keun Lee[O]* S.S Park* D.W. Hong** J.H. Yoon*
* Dept. of Computer Engineering, Hallym University
** Dept. of Media Contents, Songgok College

요 약

마이크로어레이 실험은 수 천에서 수 만개의 유전자 발현 결과를 동시에 측정할 수 있어 질병의 발현 형질 분류 등에 유용하게 이용되고 있다. 그러나 마이크로어레이 실험은 동일한 플랫폼의 실험이라 할지라도 환경 등에 따라 그 실험 결과에 차이가 나는 등 오차를 항상 포함하고 있다. 또한 마이크로어레이 실험은 아직 고가의 실험으로 분류되어 다수의 샘플에 대한 반복 실험 결과를 얻기 어려운 상황이다. 따라서 이종의 플랫폼, 데이터 포맷, 정규화 기법 등이 서로 다른 데이터를 효율적으로 통합하여 유용한 정보를 추출하는 새로운 방식의 개발이 필요하다. 본 논문은 이와 같은 문제를 해결하기 위한 기초 단계 연구 결과이다. 마이크로어레이 실험 데이터로부터 통계적 방법을 이용하여 유의(informative) 유전자를 추출하고 유전자 온톨로지(Gene Ontology : GO)와의 연계를 통하여 유전자 정보의 기능적 분류 결과를 사용자에게 제공하는 유전자 기능 분석 시스템의 설계 및 구현 방안을 보인다. 본 시스템의 실험방법에서는 3-Fold Filtering 기법을 통하여 발현 차가 큰 유전자를 추출하고, t-검정 기법에 의하여 이들 유전자를 순위화 하였으며, 이 중 상위 100개의 유전자를 유의 유전자로 추출하였다. 다음, 이 들 유의 유전자의 t-검정 값을 GO의 유전자 기능을 나타내는 해당 텀(term)에 가중치로 부과하여 각 유전자들과 기능적으로 연관성이 높은 텀들을 추출한다. 또한 본 연구의 유효성을 검증하기 위하여 본 시스템에 의한 마이크로어레이 데이터 분석 결과를 전문가에 의한 유전자 기능 분석 결과와 비교한다.

1. 서 론

20세기 후반 미국, 일본, 유럽 등이 중심이 되어 생명현상을 관장하는 유전자 정보를 얻기 위해 인간 유전체 사업(Human Genome Project)이 시작되었고, 2003년 4월에는 인간이 가지고 있는 30억 쌍의 유전체 염기서열 해독이 완료되었다. 이와 같은 인간 유전체 염기 서열이 해독됨으로써 인간생명의 실체를 이해하는 기초가 마련되었으며 더불어 암과 같은 질병의 원인분석 및 진단방법과 치료제 개발을 위한 새로운 토대가 마련되었다고 할 수 있다.

그러나 인간의 복잡한 생명현상을 규명하기 위해서는 유전체의 서열정보만으로는 부족하며, 단백질 간의 상호작용 및 유전자 발현 여부 등 추가적인 생물학적 정보들이 필요하다. 따라서 분석이 완료된 후 유전체의 기능을 밝히려는 기능 유전체학이 중요한 연구주제가 되었다. 기능 유전체학의 주요 실험/분석 도구 중 하나로 DNA 마이크로어레이(MicroArray)를 들 수 있다. DNA 마이크로어레이는 염기서열을 알고 있는 올리고뉴클레오타이드(oligonucleotide), 게놈 DNA 분자를 소형 기판 위에 고밀도로 배열해 놓은 것으로 대량의 유전자 발현 상황을 총체적으로 탐색할 수 있다. 마이크로어레이 데이터는 동시에 수 만개 유전자의 발현 값을 포함하고 있기 때문에 질병의 발현 형질 분류에 매우 유용하게 쓰이며, 기존 의료인이나 생물학자들이 진행해온 실험방식보다 유전자의 기능 분석에 필요한 시간과 어려움을 획기적으로 줄일 수 있다[1].

현재 고가인 마이크로어레이 실험은 샘플의 수가 제한적이고, 또한 동일한 생물학적 주제에 대한 실험이라도 실험 환경, 플랫폼 등에 따라 분석결과가 다를 수 있다. 따라서 이종의 플랫폼, 데이터 포맷, 정규화 기법 등이 서로 다른 데이터를 효율적으로 통합하여 유용한 정보를 추출할 수 있는 새로운 방법론의 개발이 필요하다. 본 논문은 이와 같은 문제를 해결하기 위한 기초 단계 연구 결과이다. 본 연구에서는 통계적 분석 기법을 기반으로 하여 유전자 온톨로지(Gene Ontology : GO)를 활용한 새로운 마이크로어레이 유전자 기능 분석 시스템의 개발 방식을 제안한다.

제안된 시스템에서는 다음과 같은 3단계 과정에 의하여 마이크로어레이 데이터로부터 유의 유전자를 추출하고 이들 유의 유전자와 연관 있는 유전자 기능(function) 정보를 추출하여 사용자에게 제공한다. 우선 1 단계에서

본 연구는 과학기술부 과학재단 목적기초연구(R01-2006-000-11106-0) 지원으로 수행되었음.

는 마이크로어레이 실험에서 유전자 샘플 그룹 간 발현 값의 차이를 크게 나타내는 유전자가 의미 있는 유전자라는 가정에 의하여 유전자 그룹 간 평균 차이가 3배 이상인 것을 추출하는 3-fold filtering 기법을 적용한다. 다음 2 단계에서는 t-검정 방식[2]을 사용하여 이 들 유전자를 순위화한 후, 전체 순위 중 상위 특정 순위까지의 유전자를 유의 유전자로 선별한다. 다음 3 단계 에서는 GO를 활용하여 선별된 유의 유전자의 기능 분석을 수행한다. 이 단계에서 GO 검색을 위하여 Probe Set ID와 유전자 심볼의 매칭 과정이 필요하며, 이를 위하여 NBN(National Bioinformatics Network)에서 제공하는 MAD_SCHEMA의 일부를 수입(Import), 사용하였다[3].

본 논문의 구성은 다음과 같다. 2장은 관련 연구로 본 연구의 배경 지식에 대해 설명한다. 3장에서는 개발 시스템의 유의 유전자 추출법 및 기능 분석 방법 등 실험 방법에 대해 언급하며, 4장에서는 실험 결과를 분석하고, 5장에서는 본 시스템의 구조와 사용자 인터페이스에 설명하고, 마지막으로 6장에서는 결론과 향후 연구 계획을 보인다.

2. 관련연구

2.1 마이크로어레이 데이터

DNA 마이크로어레이는 염기서열을 알고 있는 DNA 분자를 소형 기판위에 고밀도로 배열해 놓은 것이다. 마이크로어레이는 기판에 붙이는 유전물질의 크기에 따라 cDNA 칩과 올리고뉴클레오타이드 칩으로 나눌 수 있다. cDNA 칩은 최소한 500bp(base pair) 이상의 유전자(full-length open leading frame 또는 EST)가 붙어져 있고, 올리고뉴클레오타이드 칩은 약 15~20개의 염기들로 이루어진 올리고뉴클레오타이드가 붙여져 있다. 마이크로어레이는 대량의 유전자 발현 상황을 총체적으로 탐색할 수 있다. 또한 유전자 발현 형태 확인, 유전자 기능 예측, 질병 관련 유전자 발굴 등 생명현상과 관련된 유전체 수준의 연구를 가능하게 한다[4].

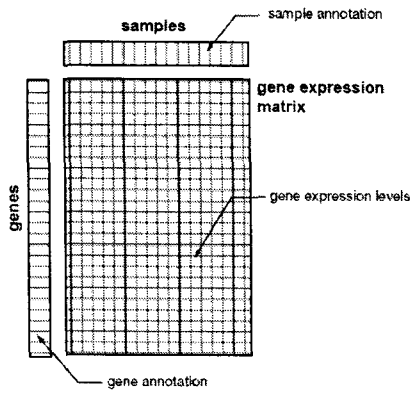

그림 1. 마이크로어레이 구조도

그림 1.과 같이 마이크로어레이는 유전자 발현 데이터를 해석하기 위해 유전자에 관한 정보와 생물학적 시료와 시료가 얻어진 실험 조건에 대한 주석 그리고 특정 실험 조건에서 유전자들의 발현 정도를 담고 있는 유전자 발현 매트릭스에 대한 정보를 포함하고 있다[5].

마이크로어레이는 미지의 유전자를 포함한 수많은 유전자의 활성도를 동시에 측정하여, 특정조건 전후의 발현 상태를 비교하거나, 서로 다른 조직 혹은 이웃한 조직에서의 발현 프로파일을 비교하거나, 다른 종에서 관련 유전자의 발현을 비교하거나, 시계열 발현 프로파일을 분석하는 등의 다양한 방법에 의해서 유전자 정보를 대량 처리하여 새로운 지식을 찾아낼 수 있다.

2.2 유전자 온톨로지(GO)

각각의 바이오 정보 시스템에 저장된 데이터들은 서로 독립적인 것이 아니라 의미상 연관성이 있어 관련 자료들을 함께 보일 수 있는 서비스가 필요하다. 예를 들어, 생명 유지에 필요한 단백질이나 유전자 등은 하나의 종에만 국한된 것이 아니라 여러 종에 공통적으로 존재하므로 모든 종에서 서로 연관된 데이터에 대한 용어의 확립이 필요하다 연관 정보는 연관된 바이오 데이터의 연결에 대한 주석정보를 담고 있는 것으로, 여러 바이오 정보 시스템의 데이터 간 정보 공유와 교환에 사용된다. 이와 같이 바이오 정보 통합에 필요한 지식 정보를 유전자 온톨로지(GO)[6]라 한다. 즉, GO는 단백질 간의 상호 작용을 표현한 유전자 네트워크상에서 그 위치 등의 정보를 데이터베이스에 저장한 지식정보이다.

유전자 온톨로지는 이미 3만여 개의 생물학 용어와 용어 간의 관계 정보를 갖고 있으며 Biological Process, Molecular Function, Cellular Component의 세 가지 범주로 나누어져 있다. 각 범주는 텀(term)이라 불리는 노드(node)들의 집합으로 이루어져 있으며, 이들 노드들은 is_a 관련(혹은 part_of)으로 연결된 DAG(Directed Acyclic Graph) 형태를 갖는다. 텀은 생물학적 과정이나 기능, 세포내 구조 등을 표현하고 있으며, 각 텀과 연관된 단백질, 유전자 정보가 연관(association) 정보의 형태로 링크되어 있다. 초기엔 다양한 종의 유전체 주석의 표준화를 위해 만들어진 GO는 최근 발현 정보 분석의 대표적인 도구로 사용되고 있다.

2.3 기존 연구

마이크로어레이 분석 시스템은 초기에는 단순한 발현 데이터 분석을 목적으로 하였으나, 최근에는 유전자의 생물학적 기능 분석 등을 목적으로 하는 등, 그 개발 목적이 점차 확대되어 나가고 있다. 이들 시스템은 대부분 생물학적 상호 운용의 프로파일로 GO를 사용하고 있으며, 이와 같은 형태의 대표적인 시스템으로 "GO-Mapper[7]"와 "Onto-Express[8]" 등을 들 수 있다. 이들 시스템에서는 일반적으로 특정 유전자 혹은 샘플 데이터를 분석하여 GO를 이용한 다양한 기능 분석 결과를 제공하고 있다. 그러나 이들 기존의 마이크로어

레이 분석 시스템은 실험을 통해 얻어진 단일 샘플 별로 분석을 진행하는 것으로 유전자의 기능적 분석에 초기화 단계라 할 수 있다. 마이크로어레이 실험은 동일 플랫폼의 실험이라 할지라도 실험 환경 등의 영향에 따라 상이한 결과를 얻을 수 있어, 샘플 수가 많을 경우 그 신뢰성을 높일 수 있다고 할 수 있다. 따라서 샘플 그룹 전체를 분석을 대상으로 하는 등 확장된 기능을 갖는 분석 시스템의 개발이 필요하다.

3. 실험 방법

본 시스템 HMDA(Hallym Microarray Data Analysis)는 "3-Fold Filtering", "t-test & Ranking", "Gene Ontology Mapper", "Gene Knowledge-base"의 4개의 기능 모듈로 구성된다. 각 모듈의 데이터 분석/가공 과정은 다음과 같다.

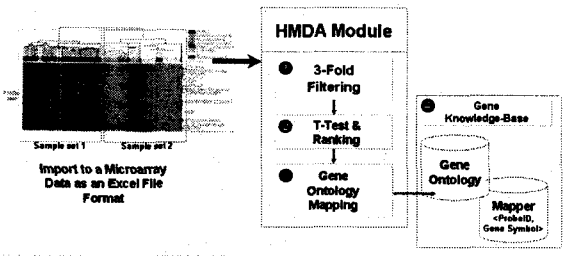

그림 2. 데이터 분석 과정

① 3-Fold Filtering : 마이크로소프트 엑셀 형태로 저장된 마이크로어레이 데이터를 수입(Import)한 후, 3-Fold Filtering 모듈은 유의성을 가진 유전자 샘플은 유전자 그룹 간 발현차가 큰 것으로 가정하여 (식 1)에 의하여 그룹 간 평균비가 3배 이상인 유전자 그룹을 정제한다.

$$\log_{10}\frac{average_{control\ group}}{average_{test\ group}} \geq \log_{10}3 \quad \cdots\cdots\cdots\cdots \text{(식1)}$$

② T-Test & Ranking : t-검정은 두 집단 간 평균이 통계적으로 유의한 차이를 보이고 있는지의 여부를 검증할 때 이용된다. ①에서 필터링된 유전자 데이터에 대하여 각 그룹의 데이터 분포를 고려하기 위하여 (식 2)에 의하여 평균값의 차(difference between the means)를 두 그룹의 평균값의 차의 표준 편차(standard deviation of the difference between the means)로 나눈 t-검정값을 구한다. 다음, 이 값에 의하여 각 유전자의 순위(rank)가 결정되며, 휴리스틱 기법에 의하게 상위로부터 임의의 개수를 지정하여 유의 유전자를 결정한다. 본 시스템은 상위 100개의 데이터를 유의 유전자로 결정하였다.

$$t-value = \frac{\left|average_{control\ group} - average_{test\ group}\right|}{\sqrt{\frac{var_{control\ group}}{n_{control\ group}} + \frac{var_{test\ group}}{n_{test\ group}}}} \quad \text{..(식2)}$$

③ Gene Ontology Mapper : ②에서 결정된 유전자들을 GO와 연계시켜 그 기능 정보를 검색한다. 마이크로어레이 데이터의 "Probe Set ID"에 해당하는 유전자 심볼을 검색어로 GO를 검색하여 해당 유전자가 연관 정보로 등록되어 있는 IS_A 계층 구조 상의 텀을 검색한다. 다음, 검색된 해당 텀과 그의 모든 조상 텀에 연관된 유전자의 t-test 값을 가중치로 부가한다. 이와 같은 과정에 의하여 모든 유의 유전자에 의한 가중치 부여가 완료된 후, GO 상의 가중치가 부여된 모든 텀에 대하여 다음 (식 3)에 의하여 유의도(significant value: SV)를 결정한다. 여기에서 n과 m은 각각 전체 유의 유전자의 개수와 해당 텀에 연관된 유의 유전자의 개수를 나타낸다.

$$sv = \frac{\sum_{j=1}^{m} t-value(gene_j)}{count(matching\ gene)} \Big/ \frac{\sum_{i=1}^{n} t-value(gene_i)}{count(selected\ all\ genes)} \quad \cdots\cdots\cdots\cdots \text{(식3)}$$

그림 3의 예를 이용하여 이 과정을 간단히 설명하면 다음과 같다. 그림에 보이는 바와 같이 3개의 유의 유전자 Gene_b, Gene_c, Gene_d가 추출되어 각각 1.2, 3.2, 2.7의 t-test 값을 가지고 있다고 가정한다. GO의 해당 텀에 이 들의 t-test 값을 가중치로 부과하면, 그림과 같이 각 텀에 가중치가 할당되게 된다. 다음 각 텀에 대하여 SV를 구하면 표 1.과 같은 결과를 얻게 되며, 이들 중 특정 임계치보다 큰 SV를 갖는 텀을 유의 기능을 갖는 텀으로 결정한다. 표 1.에 의하여 "GO:Term_C", "GO:Term_D" 등이 유의 기능을 갖는 텀으로 결정될 수 있다.

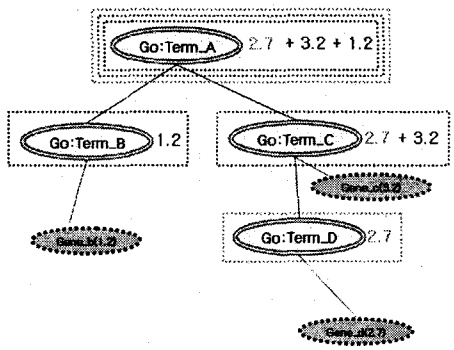

그림 3. 연관 텀 추출을 위한 가중치 부여

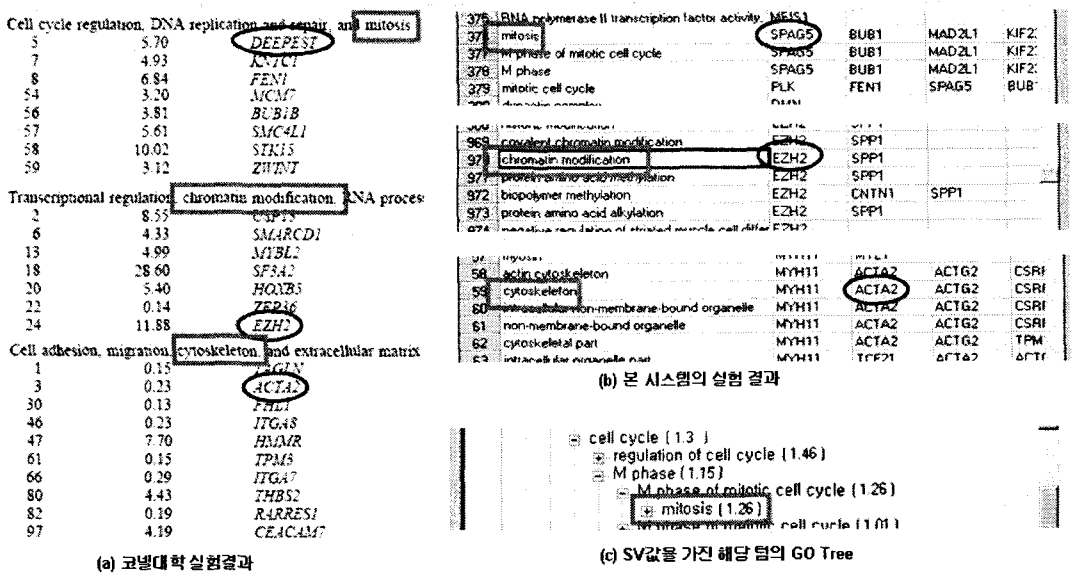

그림 5. 타당성 검증 결과

표 1. SV(Significant Value) 추출

Term	Related Gene's T-Test Value	Weighted value's average	SV
GO:Term_A	none	2.36	1
GO:Term_B	1.2	1.2	0.508
GO:Term_C	3.2	2.95	1.25
GO:Term_D	2.7	2.7	1.144

④ Gene Knowledge-base : 마이크로어레이 데이터의 "Probe Set ID"와 GO의 유전자 심볼 간 매칭을 위하여 TIGR(The Institute for Genomic Research)의 NBN에서 제공하는 MAD 스키마를 수입하여 매칭 정보 지식베이스(Matching Information Knowledge-Base)를 구축하였다. 그림 4.는 MAD 스키마의 개념 단계(Conceptual Level)를 보인 것으로 본 시스템에서는 "Clone", "Gene", "Probe", "Probe_source" 테이블을 수입, 사용하였다.

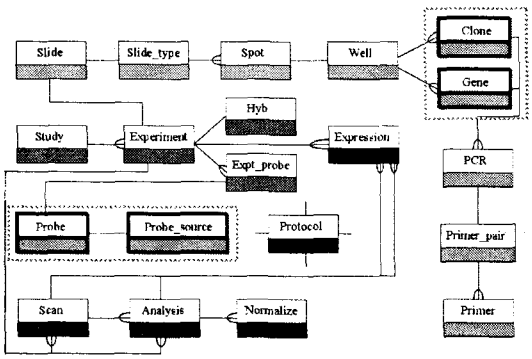

그림 4. MAD(MicroArray Data in NBN) 스키마

4. 실험 결과 및 분석

제안된 마이크로어레이 데이터 분석 방식의 유효성을 검증하기 위하여 참고논문 [9]의 실험 결과와 비교 검증을 수행하였다. 코넬대학에서 수행된 참고 논문 [9]의 실험에서는 초기, 말기의 전립선 암(prostate cancer) 환자에 대한 마이크로어레이 실험을 수행하여 그 두 그룹의 유전자 발현 결과를 분석하고 아울러 이들 유전자의 기능적 분석 결과를 보이고 있다. 실험에서는 "Affymetrix" 사의 "U95A" 단일 플랫폼의 마이크로어레이를 사용하였으며, 3개의 "noncancerous prostate", 23개의 "primary prostate cancers", 9개의 "metastatic prostate cancers"의 총 35개의 샘플 결과를 보이고 있다. 여기에서 "primary"는 원래의 암세포를 의미하는 것이고, "metastatic"은 암세포가 전이되거나 다른 조직에 새로운 암세포가 형성되었을 때를 의미한다[10]. 다시 23개의 "primary prostate cancers"는 재발 가능한(recurrence) 샘플 9개, 재발되지 않는(non-recurrence) 샘플 14개로 나누어진다. 본 실험에서는 "metastatic prostate cancers" 샘플 9개와 "primary prostate cancer" 샘플 중 재발되지 않는 샘플 14개의 마이크로어레이 데이터를 사용하였다.

다음의 그림 5.에 참고 논문 [9]의 실험 결과와 본 시스템의 분석 결과를 비교하여 보인다. 그림 5-(a)는 참고문헌 [9]에 실린 실험 결과의 일부로서 추출된 유의 유전자를 기능별로 분류하여 표시하고 있다. 이 논문에서 유전자의 기능 분류는 전문가에 의한 수작업에 의하여 이루어진 것이다. 그림 5-(b)는 본 시스템의 GO를 사용한 분석 결과의 일부로서 우리의 결과는 그림 5-(a)의 기능 분석 결과를 대부분 포함하고 있음을 알 수 있다. 예

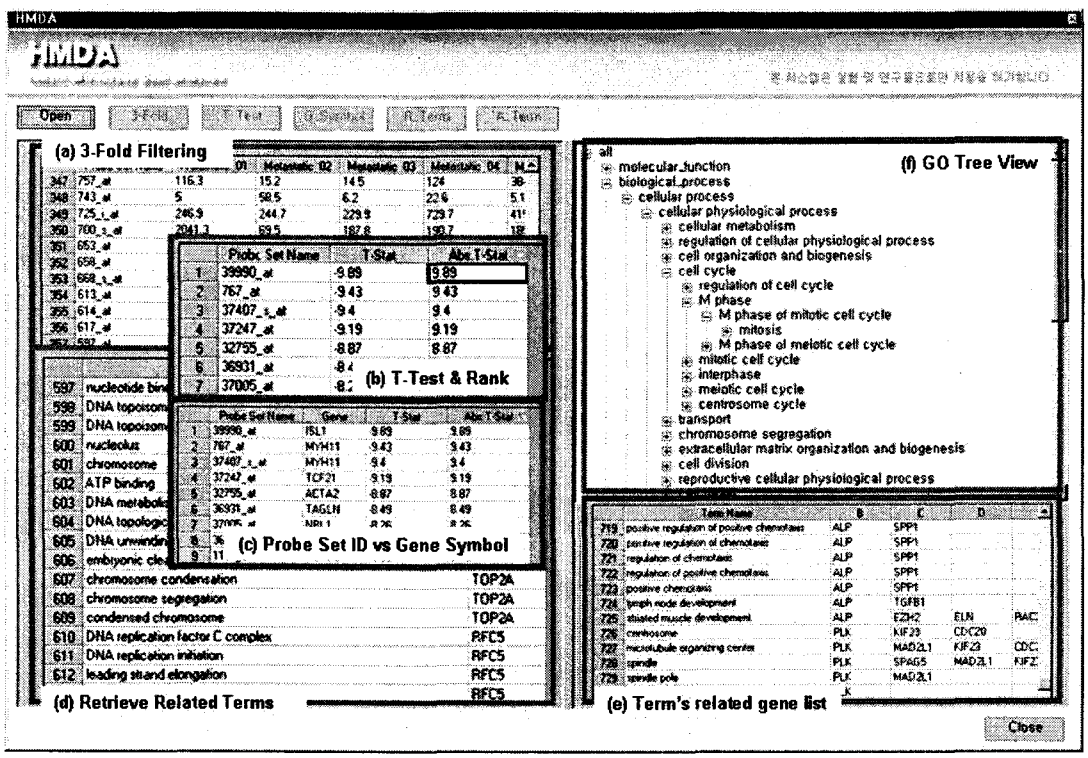

그림 7. 사용자 인터페이스

를 들어 그림 5-(a)의 "Cell Cycle regulation, DNA replication and repair, mitosis" 와 유전자 온톨로지에서 검색한 그림 5-(b)의 "regulation of cell cycle, DNA replication, DNA repair, mitosis"가 동일한 기능을 하는 텀임을 확인하였으며, 그림 5-(a)의 'DEEPEST', 'BUB1B', 'EZH2', 'ACTA2' 등의 유전자는 그림 5-(b)의 'SPAG5', 'BUB1', 'EZH2', 'ACTA2' 등의 유전자와 각각 동일한 유전자임을 알 수 있다. 또한 그림 5-(b)의 각 행의 처음 열은 추출된 텀의 순위를 보이고 있어 이 들 텀을 체계적 분석에 활용될 수 있다. 다음의 그림 5- (c)는 유전자 온톨로지에서 검색한 텀들을 GO 트리 구조로 표현한 것으로 이 결과로부터 추출된 텀들의 상호 위치를 파악할 수 있으며, 이 때 각 텀의 SV를 함께 출력하여 이들의 주요도를 비교, 분석하는 것이 가능하다.

5. 시스템 구조 및 사용자 인터페이스

5.1 시스템 구조

본 시스템의 구조도는 그림 6.과 같으며 크게 HMDA 엔진과 저장소로 구분한다. HMDA 엔진은 통계적 방법으로 마이크로어레이 데이터 처리를 하는 "Data Analysis module", 유의 유전자의 연관된 텀을 검색하기 위한 "Term Extract module"로 구성되며, 저장소에는 마이크로어레이 데이터와 유전자 온톨로지, 매칭 정보 지식베이스가 저장, 관리된다. 각 모듈의 세부 기능은 다음과 같다.

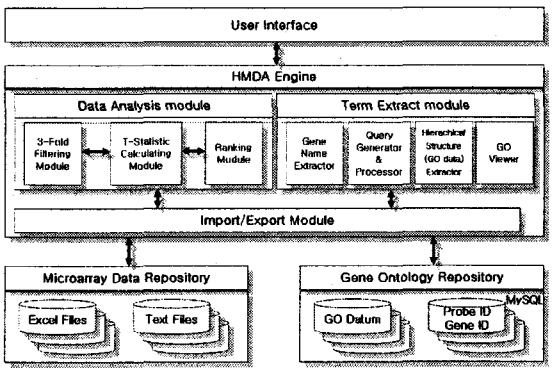

그림 6. 시스템 구조도

■ 3-Fold Filtering Module : 수입한 유전자 데이터에서 발현차가 큰 데이터를 추출하는 모듈이다. 두 그룹 간 평균 차이가 3배 이상이 되는 유전자를 추출한다.

■ T-Statistic Calculating & Ranking Module: t-검정

계산 모듈은 두 집단 간 평균이 통계적으로 유의한 차이를 보이고 있는지를 검증하기 위하여 t-검정 값을 추출한다. 계산된 t-검정 값들은 Ranking Module에 의해 큰 값부터 정렬된다.

■ Gene Name Extractor : 마이크로어레이 데이터의 Probe set ID에 관련된 유전자 심볼을 매칭 정보 지식베이스에서 추출한다.

■ Query Generator & Processor : 유전자 온톨로지에서 관련된 텀을 검색하기 위하여 Gene Name Extractor에서 추출된 유전자 심볼을 검색어로 한 질의어 작성 및 처리를 한다.

■ Hierarchical Structure Extractor : 방향성을 갖는 비순환 그래프 형태의 텀 구조를 사용자에게 비쥬얼하게 제공하기 위하여 추출 텀에 대한 계층 구조 정보를 추출한다.

■ GO Viewer : Hierarchical Structure Extractor에서 추출한 텀의 계층 구조를 사용자 인터페이스에서 트리 형태로 표현한다.

5.2 사용자 인터페이스

본 시스템은 Visual Studio .net 2003 버전을 사용하여 개발하였으며, 마이크로어레이 데이터 수입에는 Microsoft JET OLEDB 4.0을 사용하였고, 유전자 온톨로지 및 유전자 매칭 지식 베이스는 MySQL 5.0을 이용하여 구축하였다. 본 시스템의 사용자 인터페이스는 유전자 데이터와 연관된 텀의 계층구조를 사용자에게 비쥬얼하게 제공하여 이해도와 분석의 정확성을 높일 수 있다.

그림 7.은 본 시스템의 사용자 인터페이스 화면을 캡쳐한 것으로 각 버튼의 구성은 다음과 같다. "open" 버튼을 클릭하여 마이크로어레이 데이터를 수입한다. "3-Fold" 버튼을 클릭하면 3배 이상의 평균값 차를 갖는 유전자 샘플을 필터링하는데, (a)는 평균 값의 차가 3배 이상의 유전자 샘플을 추출한 결과를 보이고 있다. "T-Test" 버튼은 t-검정 값을 추출하고, 순위를 결정하여 (b)의 결과를 획득한다. "G_Symbol" 버튼은 (c)와 같이 Probe Set ID의 매칭 유전자 심볼을 추출한다. "R_Term" 버튼은 (d)와 같이 유전자 데이터에 연관된 해당 텀을 추출한다. "A_Term"은 (e)와 같이 해당 텀에 대한 유전자 리스트를 보이며, (f)와 같은 트리 형태로 연관 텀의 계층 구조를 보인다. 사용자 인터페이스 내 각 창(Window)의 구성은 각 버튼의 이벤트 발생 후 처리된 결과를 보여주는 "Excel Display" 부분, 유전자 온톨로지의 계층 구조를 보이는 "GO View" 부분, 시스템의 중간 진행 상태를 보이는 "Status of Processing" 창으로 이루어져 있다.

6. 결론 및 향후 계획

제안된 마이크로어레이 분석 시스템은 마이크로어레이 데이터를 입력으로 받아 3-Fold Filtering, t-검정 등의 통계적 방법을 기반으로 유의 유전자를 추출하고 이들 유의 유전자와 매칭되는 텀을 GO에서 검색하여 유전자 기능 분석을 수행한다. 또한 이 들 분석 결과를 단계적인 다양한 방식으로 출력하여 유연성 있는 분석, 검색 환경을 제공한다. 본 시스템은 현재 단일 플랫폼의 실험 결과만을 분석 대상으로 하고 있다. 금후 이종의 플랫폼, 데이터 포맷, 정규화 기법 등이 서로 다른 마이크로어레이 데이터를 통합하여 이들의 기능적 분석을 수행할 수 있도록 시스템 기능을 확장할 예정이다. 또한 시스템 분석 기능의 타당성을 보이기 위한 방식으로 전문가에 의한 분석과 함께 "Hyper-Geometric Distribution", "Binomial Distribution", "Chi-Square Distribution" 등의 통계적 모델에 기반한 타당성 검증을 수행할 예정이다.

참고문헌

[1] 박태성, 이승연, 김기웅, 이성곤, 최호식, 윤단규, "마이크로어레이 자료의 통계적 분석," 자유아카데미, pp. 2~43, 2005.

[2] http://www.socialresearchmethods.net/kb/stat_t.htm

[3] http://www.nbn.ac.za/Education/14-microarray-2004/

[4] 김동훈, 권호정, "DNA chip의 생명과학연구 및 신약개발에의 활용," Life Science & Biotechnology, pp. 12~19, 2002.

[5] Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., etc., "Minimum information about a microarray experiment (MIAME) - toward standards for microarray data," nature genetics, vol.29, pp. 365~371, 2001.

[6] http://www.geneontology.org

[7] Smid M. and Lambert C.J. Dorssers, "GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms," Bioinformaics, Vol.20, No.16, pp. 2618~2625, 2004.

[8] Khatri P., Draghici S., Ostermeier C. and Stephen A. Krawetz, "Profiling Gene Expression Using Onto-Express," Genomics, Vol.79, No.2, pp. 266~270, 2002.

[9] LaTulippe E., Satgopan J., Smith A., Scher H., Scardino P., Reuter V., and William L. Gerald, "Comprehensive Gene Expression Analysis of Prostate Cancer Reveals Distinct Transcriptional Programs Associated with Metastatic Disease," Cancer Research, vol.62, pp. 4499~4506, 2002.

[10] http://www.cancer.gov/cancertopics/factsheet/Sites-Types/metastatic