GOVP1200506031

M101KG010001-03K070100520

작물유전체기능연구사업

Gene mining 시스템 개발 및 미생물 기능 분석에 관한 연구

Development of gene mining system

한국생명공학연구원

과 학 기 술 부

제 출 문

과학기술부 장관 귀하

본 보고서를 과학기술부 특정연구개발사업 21세기프론티어연구개발사업 중 과학기술부와 농촌진흥청이 지원하는 작물유전체기능연구사업단 "Gene mining 시스템 개발 및 미생물 기능 분석에 관한 연구"과제 (과제번호 CG1222)의 단계보고서로 제출합니다.

2004. 8. 30

주관연구기관명 : 한국생명공학연구원

주관연구책임자 : 허 철구

연 구 원 : 변 상진

〃 : 신 윤희

〃 : 이 규열

〃 : 정 태훈

〃 : 박 선용

〃 : 홍 성의

〃 : 이 효수

과제관리번호		해당단계 연구기간	2001.08.01 ~2004.06.30	단계 구분	(해당단계) / (총단계)
연구사업명	중 사 업 명	21C프론티어연구개발사업			
	세부사업명	작물유전체연구개발사업			
연구과제명	중 과 제 명	중과제가 있을 경우에는 기재 (단위과제일 경우에는 아래 기재)			
	세부(단위)과제명	Gene mining 시스템 개발 및 미생물 기능 분석에 관한 연구			
연구책임자	허 철구	해당단계 참여연구원수	총 : 6 명 내부 : 1 명 외부 : 5 명	해당단계 연구비	정부: 천원 기업: 천원 계: 천원
연구기관명 및 소속부서명	한국생명공학연구원 유전체연구센터		참여기업명		
국제공동연구	상대국명 :		상대국연구기관명 :		
위 탁 연 구	연구기관명 :		연구책임자 :		

요약(연구결과를 중심으로 개조식 500자이내) | 보고서 면수 |

작물관련 유전자 연구에 필수적인 Plant EST데이타와 DNA Chip을 이용한 gene expression profile, Proteomics연구에 가장 널리 활용되는 Peptide Mass fingerprinting 등을 통합하여 분석할 수 있는 시스템 구축을 1단계 목표로 하였다.

본 연구에서는 Arsbidopsis를 비롯한 총 9종 137만건, 11개의 Tissue별로 EST 분석을 BLAST, GO, 독일 MIPS 기준의 Protein function분류 , Coding region예측 등의 정보를 제공하는 방대한 양의 데이터를 분석하여 놓았다. 또 상용 package를 대체할 수 있는 DNA Chip 분석용 S/W를 자체 개발하였다. 이 S/W는 CG1221 최도일박사 연구 결과 및 KRIBB자체 데이터 분석을 통하여 우수성이 검증되어 실용적으로 사용하고 있다. 또 MALDI-TOF에서 나온 데이터를 분석할 수 있는 Peptide mass fingerprinting 분석용 S/W를 자체 개발하여 국내에서 직접 생산하여 연구하는 EST, MALDI-TOF데이타와 DIP protein protein interaction data를 통합 분석할 수 있는 시스템을 개발하여 기술료 계약과 프로그램 등록 실적을 나타내었다. 아울러 DNA Chip분석이 후 co-expression 된 데이터의 co-regulation 관련 Motif를 분석하였고, 기능해석에 필요한 BLASTX, Gene Ontology, MIPS Funtion Catalog 에서 제공하는 yeast 각protein 의 description 과 mipsfuncat number를 조합하여 yeast mips function catalog 완성 Arabidopsis thaliana protein sequence와 mipsfuncat number이용하여 각 plant의 mips catalog 형성시에 Arabidopsis의 protein sequence에 blastx로 homology check 하여 match 되는 Arabidopsis funcat number를 부여하여 mips function catalog 완성, 관련 promoter 정보 등을 제작하여 분석에 활용하도록 하였으며 gene analysis, gene expression profiling, proteomics의 데이터가 통합 분석이 가능하도록 시스템 개발하였다.

색 인 어 (각 5개 이상)	한 글	유전자 분석, 마이크로어레이, 프로테옴, 기능카타로그, 통합분석
	영 어	EST analysis, Microarray, Proteome, functional catalog, integrated genome

요 약 문

Ⅰ. 제 목

Gene mining 시스템 개발 및 미생물 기능 분석에 관한 연구

Ⅱ. 연구개발의 목적 및 필요성

가. Plant 관련 유전자를 기능 해석을 위하여 EST, DNA Chip 데이터, Proteomics데이타, Protein interaction, Metabolic Pathway 데이터, Data mining 기법이용 등을 통합 분석하여 실험실에 효율적인 정보 제공을 목표로 한다.

나. 필요성

Plant 유전자 기능 해석 시스템을 위하여 아래와 같은 필수 S/W개발이 필요하다.

- ❑ DNA Chip 분석용 S/W개발
- ❑ Peptide Mass fingerprinting S/W개발, Functional Catalog 제작
- ❑ EST, DNA chip 정보, Peptide Mass spectrum, gene function catalog의 통합 시스템 구현
- ❑ co-expression된 정보의 co-regulation region 예측 시스템 구축

Ⅲ. 연구개발의 내용 및 범위

구 분	연구개발 세부목표	추진 실적
2001	- 벼,콩등 작물유전체 EST 데이타베이스 구축, 분석 시스템 개발 및 서비스 - DNA Chip image분석에서 Clustering 기법까지의 데이터분석 work-flow개발	- Arabidopsis, rice, soybean등 에 대한 EST 분석 완료 - DNA Chip 분석에 필요한 Data Quality Assessment기능, Data Normalization기능, Diffrentially expression gene기능, Clustering 기능, Classification기능 등에 통합 버전 개발 (프로그램 등록 2건 완료)
2002	○애기장대,콩, 벼관련 Proteomics 연구를 위한 peptide mass spectrum 예측 시스템 개발	○ 구축 완료 (기술료 3000만원 계약 완료)
	○EST 데이터베이스와 애기장대, 콩,벼등 기존의 공개 식물 EST 를 Protein으로 변환하여 Peptide mass spectrum시스템과 연결한 DB구축	○ 구축 완료 (EST분석, DNA Chip분석, Proteomics분석 상호 연동성 있는 시스템 개발)
	○DNA Chip을 이용한 Plant gene expression profiling 시스템 개발	○ 구축 완료
2003	❑ 벼,콩, 애기 장대 gene function catalog를 개발하여 보다 빠른 EST 기능을 추정하 하는 S/W개발	❑ MIPS, Ontology 기능 분류 S/W 를 이용하여 Catalog 완성 (Yeast, Arabidopsis용)
	❑ EST데이타와 DNA Chip정보, gene function catalog와 통합 시스템 구현	❑ 구현 완료
	❑ DNA Chip해석에 따른 regulation region 예측 시스템 구축 및 co-expression 검증	❑ Chip 데이터의 Clustering에 대한 co-regulation정보 분석을 위하여 Gibbs sampling, MEME,등을 이용한 DNA motif 분석 완료 - TRANSFAC를 이용한 TF binding motif 분석 완료

Ⅳ. 연구개발결과

- Plant EST 데이터 구축
- Microarray용 데이터 분석용 S/W 국내 원천 기술 개발
- Peptide mass spectrum 분석 S/W 국내 원천 기술 개발
- Integrated genome analysis를 위한 데이터베이스 개발

Ⅴ. 연구개발결과의 활용계획

- 국내 식물 유전체 기능 연구 관련 기관에게 적극적인 서비스 활용
- 국내 자체 개발된 Mciroarraye 데이터 분석용 S/W보급으로
  수입 대체 효과

SUMMARY

Title : Development of gene mining system

Objectives

To support the researchers working with the crop related species, we aimed providing valuable information at a goal of the first step of the CFGC project by using and analyzing public plant EST sequences, microarray data, proteomic data, and protein interaction data. As a result, we developed the softwares for the microarray analysis and the peptide mass fingerprinting (PMF). We also constructed the systems for the predicted regulatory motifs of co-expressed genes from the results of microarray experiments, plant function catalogs, and EST analyses of crop-related species and nine representative plant species. In addition, we constructed the integrated systems on the basis of gene function catalog inked with microarray and PMF data reciprocally. All of our results are freely available at http://crop.kribb.re.kr as a global integrated web interface.

Background

ESTanalyses Functioncatalogs

The gene analysis based on ESTs involves clustering and assembly of EST sequences, chromosomal mapping of consensus sequences obtained from clustering and assembly, and finally gene annotation process. For the EST clustering process, we used StackPACK software from SANBI to make virtual mRNA candidate with high coverage and high quality. To analyze these sequence data further, these process are needed as follow: homology search of consensus sequences using BLASTX against NCBI NR database, protein function categorization according to MIPS (Munnish Information) and Gene Ontology(GO) catalog, chromosomal mapping using sim4 tool and the upsteam sequence analysis obtained from genomic mapping. We applied the public softwares such as Gibbs sampling, Multiple EM for Motif Elicitation (MEME), and TRANSFAC to our research for promoter analysis.

DNAchipanalysis

In the case of clustering analysis of microarray study, validation is one of important data processing step. The notion is generally accepted that co-expressed genes are under the same transcriptional control and are probably categorized into same or similar functional group biologically or biochemically. Thus, the efforts to find common regulatory motifs in their promoter region, or functional grouping the genes in the same cluster are common process for the interpretation and validation of microarray data.

Peptidemassfingerprinting

Public protein sequence database such as SWISS-PROT is used practically for the protein identification from the result of Matrix-Assisted Laser Desorption Ionization-Time Of Flight (MALDI-TOF)data, which is one of popular proteomic studies. However, for the less of protein information for the specific plant species in these databases it is needed to construct the private protein database containing sufficient protein information for interpreting massive PMF results about each specific plant species. Thus we tried to make the protein database by translating enormous coding region sequences obtained from EST analysis and the PMF software working on these databases.

Therefore, in this study, we tried to make the individual systems about EST based data analysis, regulatory motif information from chromosomal mapping of ESTs, microarray data

and PMF information from bench works at first and finally integrate these individual we constructed the web-based integrated systems in which the results from gene analysis, microarray analysis, and proteomic analysis are reciprocally connected and complemented with each other.

Conclusion

We developed the integrated systems based on the gene catalogs from EST analysis results which are reciprocally connected with microarray analysis data and PMF data. Gene function catalogs help making customized cDNA microarray and the output of microarray experiments are directly analyzed on these system. In addition, the data from MALDI-TOF can be easily explained in our systems in an interactive mode with EST based information, microarray based information and vice versa.

C O N T E N T S

목 차

제 1 장 연구개발과제의 개요

기능유전체 연구가 본격화되면서 EST를 이용한 유용 유전자 발굴에 대한 연구와 대량 유전자 발현 양상을 보기 위한 Microarray 실험에 따른 분석 작업이 원활하게 되어야하고 단백질 분석을 위한 Peptide mass spectrum 분석 작업 등이 상호 연동이 되어 유전자 기능 연구에 실질적인 기초 연구가 되는 것이다. 특히 EST관련 연구는 Human 500만건 이상, mouse/rat 400만건이상, 식물 관련 EST는 약 200만건이상 미국 NCBI GenBank에 공개되어 있다. 따라서 가장 중요한 서열 수준에서의 유전자 기능 연구가 필수적인데 과거 15년간 꾸준히 연구가 진행되고 있으며 연구 대상 또한 종별로 확대되고 있는 상황이다. 이에 대한 연구가 국내에서도 상당한 진척을 보이고 있는 시점에서 해외 공개된 EST를 체계적으로 연구하여 미지의 기능을 밝히는데 필요한 정보를 가공하고 서비스하는 일은 매우 중요한 과제라고 판단된다.

대량으로 신속하게 발달되는 EST databases는 이미 알려져 있거나 혹은 아직 알려지지 않은 유전자에 대한 풍부한 정보를 포함하고 있다. 공공 데이터베이스에 수백만의 EST 정보가 존재하고, 적어도 몇 개의 개별적인 데이터베이스가 있다. EST 서열은 서로 다른 생물종이나 세포, 기관에서 발현되는 유전자를 나타내는 DNA단편 서열을 읽어 꼬리표 (tag)를 염기 쌍에 대응함으로써 염색체 DNA로부터 정확한 위치를 찾아내는 데 사용되기도 하고, drug 개발에 이용되기도 한다. EST는 약 300-500bp의 DNA 서열 조각으로서 발현된 유전자의 한쪽 끝을 단 한차례 읽어 들인 것이다. EST 데이터베이스에는 많은 새로운 유전자들이 포함되어 있는 것이 확인되었다 (Bailleul *et al.*, 1997; Lin *et al.*, 1997 Wu *et al.*, 1997; Yamada *et al.*, 1997).

EST 데이터베이스가 대량으로 공급되고 있지만, 이러한 데이터들은 가공되지 않은 상태에서는 다음과 같은 두 가지의 비효율적인 이유를 가지고 있다. 첫째, EST는 한 방향으로 sequencing을 수행하였기 때문에 sequencing error율이 높다. 둘째, EST 데이터 자체는 한 유전자의 일부분이므로 서로간에 충분히 중복 될 수 있고, 이러한 중복성 (redundancy)때문에 가공되지 않은 EST 데이터베이스는 비효율적일 수 밖에 없다. 이러한 중복성을 관리하는 강력한 방법은 같은 유전자로부터 유래된 EST cluster들을 모아서 좀 더 긴 가상의 cDNA 서열을 만드는 것이다. 각각의 EST 데이터보다는 assembly 작업을 거친 consensus sequence의 몇 가지 장점이 있다. 첫째, 분석하기 위한 서열을 좀더 단순화 할 수 있다. 둘째, 모아진 서열은 각각의 EST보다 더 길고, 코딩 서열을 해석할 수 있는 더 많은 가능성을 내포하고있다. 셋째, 각각의 EST에 존재하던 error가 assembly과정을 거치는 동안 충분히 제거될 수 있다. 넷째, 가상의 cDNA서열은 실험실에서 cloning할 수 있는 full-length 일 가능성도 있다.

NCBI에서 개발한 UniGene의 EST clustering 방법은 서열 유사성에 근거한 일반적인 clustering과정만 거칠 뿐 cluster로부터 consensus sequence를 생성하지는 않는다 (Boguski and Schuler, 1995). 유전체 연구소 (The Institute for Genome Research, TIGR)는 매우 엄격한 clustering 방법을 적용하여 확실한 consensus를 만드는 전략을 사용한다 (Quackenbush et al., 2000; Quackenbush et al., 2001). 남아프리카 국립생물정보학연구소 (South African National Bioinformatics Institute, SANBI) 의STACK (Sequence Tag Alignment and Consensus Knowledgebase) 에서는 UniGene과 TIGR의 병합시스템을 이용하여 clustering을 시도하고

있다. STACK의 clustering 시스템은 loose clustering방법을 이용하여 보다 긴 consensus서열을 만들어낼 수 있고 alternative splicing form을 예측할 수 있는 반면, paralog가 포함될 수 있다는 단점이 있다. 이러한 시스템을 이용하여 masking, loose clustering, assembly, alignment, alignment analysis for variation 을 위한 일련의 과정을 지원하는 프로그램 패키지인 StackPACK을 내놓았다. StackPACK의 clustering은 3가지 과정으로 진행이 된다 (Ramesh *et al.*, 2002). 준비 과정으로 vector sequence를 제거한 다음 첫 번째 과정으로 d2_cluster 프로그램을 이용하여 유사성을 바탕으로 loose clustering을 시행한다. d2-cluster는 대부분의 서열 비교 알고리즘과는 다른 word-based 방법을 사용하여, 최소한 99%의 sensitivity와 selectivity를 가지며, 많은 EST를 빠르게 clustering할 수 있고, alternative form을 예측하는데 효율적이라고 보고된 바 있다 (Burke *et al.*, 1999). 두 번째 과정에서는 PHRAP 프로그램을 이용하여 각각의 cluster에 속한EST들을 assemble 하는 작업이 수행된다. 세 번째 과정은 CRAW 프로그램을 이용하여 alternative splicing sequence나 같은 유전자의 유사한 variant들을 동정해 내는 과정을 거친다 (Miller *et al.*, 1999).

Functional category는 genome이 완전하게 밝혀지지 않은 상황에서 종 전체의 genome structure를 분석하고자 할 때 유용한 database로 사용된다. Gene Ontology (GO)는 GO는 유전자의 기능을 크게molecular function, biological process, cellular component 의 범주로 나누고 각각의 범주에 계층적인 controlled vocabulary를 확립하였다. 이들 범주는 서로 배타적인 것이 아니기 때문에, 하나의 유전자나 유전자의 산물이 세 개의 category에 각각 annotation되어 종 내의 전체적인 유전자 구성을 분석하는 genome annotation을 연구하는데 효율적이다 (Ashburner *et al.*, 2000). MIPS (The Munich Information Center for Protein Sequences)는 이미 알려진 protein function을 폭 넓고 다양한 관점으로 고려하였다. 본 연구 과제에서는 Plant 11종 137만건을 분석하고 Microarray , peptide mass spectrum, gene index 분석 결과를 통합하기 위하여 애기장대풀, 콩, 벼 등을 보다 더 세밀한 분석을 하여 공개하였다.

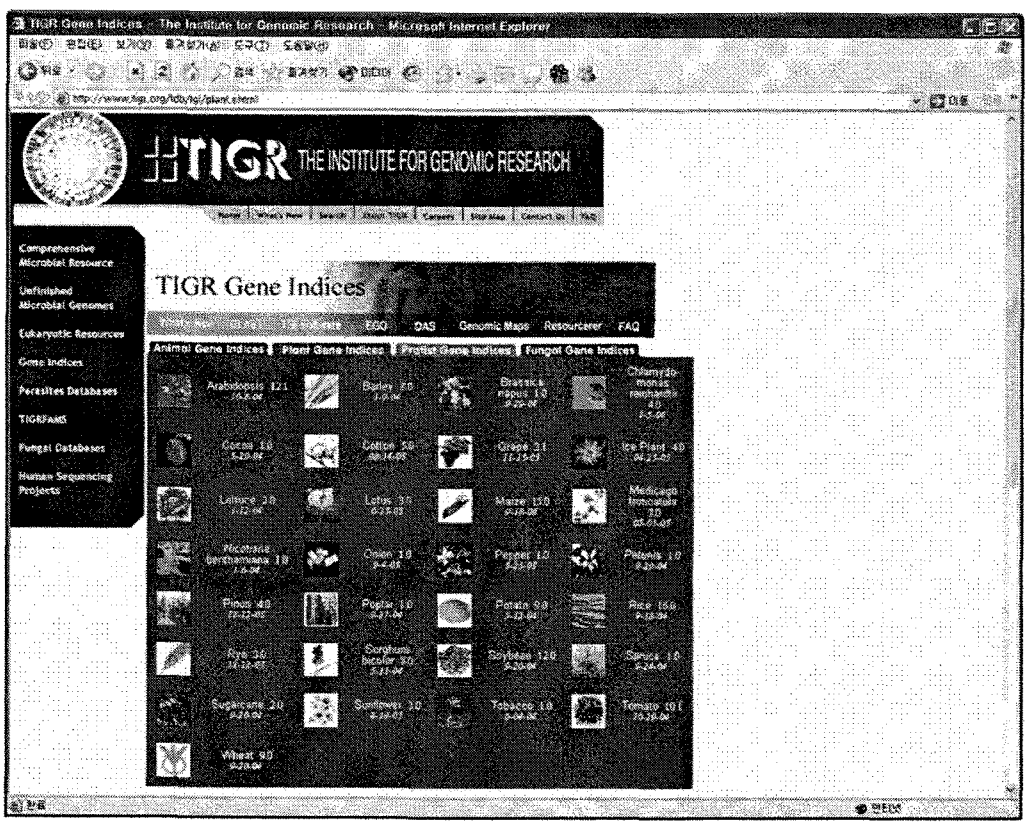
또 Microarary 같은 경우 주로 고추 관련한 데이터 분석을 처리하고 여기서 연구된 결과를 기타 작물 마이크로어레이 분석에 활용하고자 하였다. 그러나 microarary 데이터 분석을 위하여 상용 프로그램을 사용하여 분석하였으나 기능이 미비하고 사용자의 제한점이나 새로운 알고리즘 개발 시 부가적으로 활용할 방법이 없기 때문에 BioConduct개념을 이용한 자체 S/W를 개발하였다. 이미지 처리 단계(Preprocessing), 정규화단계(Normalization), 유용유전자 발현 정보 추출 단계(Diffrentially expression gene), 클러스링 단계(Clustering), 분류 단계 (Classification) 등에 관한 전 과정을 사용자가 쉽게 활용할 수 있도록 개발하였으며 새로운 알고리즘을 첨가 할 수 있게 공개용 R language로 개발하였다.

프로테옴 분야에 있어서 아직 국내에서 대량의 데이터는 생산되고 있지 않지만 Post-Genome연구 분야에서 가장 기본적인 요소 기술로 활용되고 있는 분야로 각광 받고 있다. 하지만 자체 개발된 프로그램이 없을 경우 국내 유전체 연구자와 단백질체 연구자 들에게 많은 어려움이 예상된다. 해외에서 제공하지 않는 데이터를 연구 할 경우 시급한 대책이 필요한 것이다.

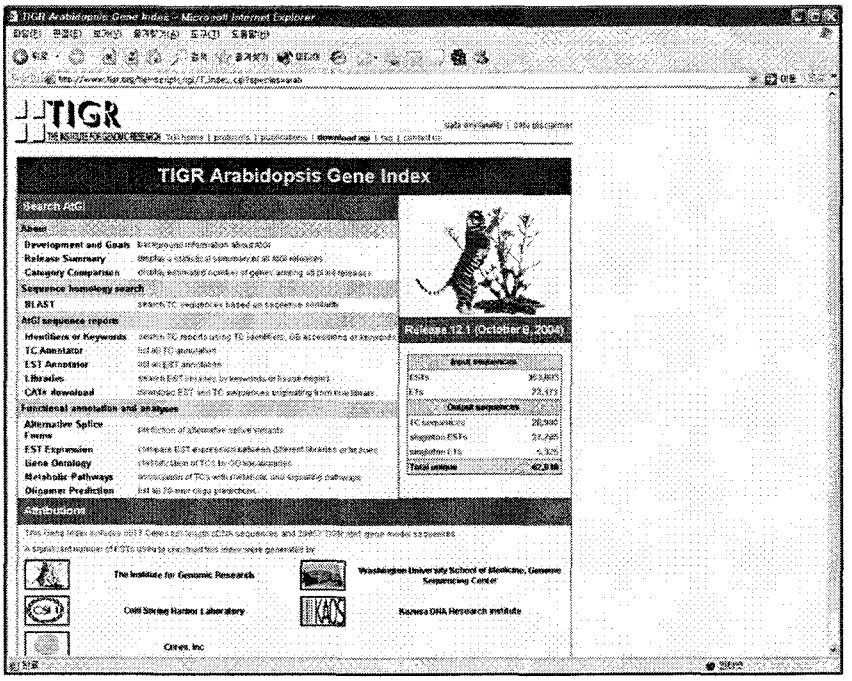
제 2 장 국내외 기술개발 현황

* 국내·외 관련분야에 대한 기술개발현황과 연구결과가 국내·외 기술개발현황에서 차지하는 위치 등을 기술

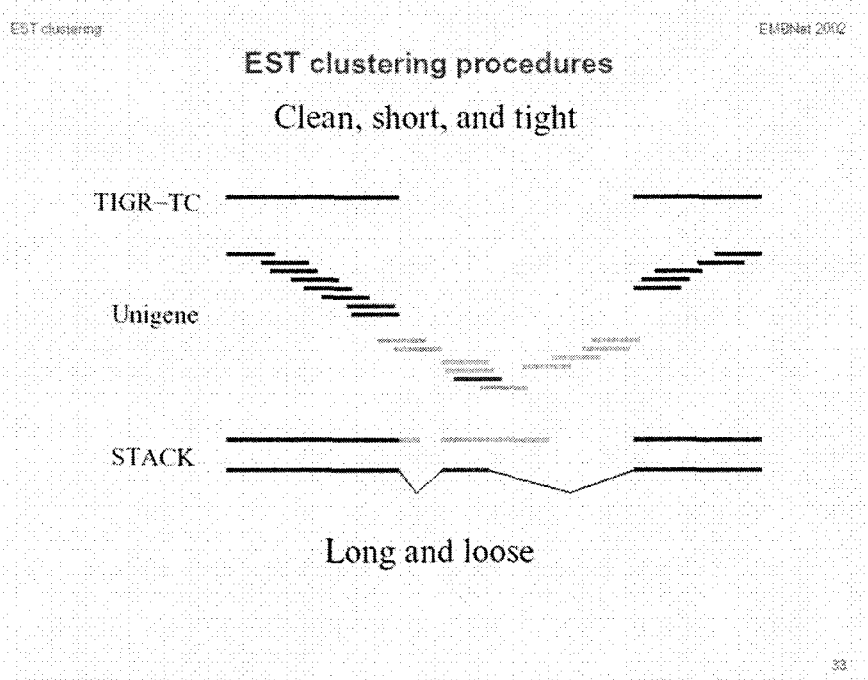
EST분석의 경우 국내에서도 인간유전체관련하여 프론티어 사업단, 농촌과학원, 국립보건원, 기업체 연구소 등에서 매우 활발하게 진행되고 있고 고유의 Unigene을 찾는 연구에 적극 활용되고 있다. 해외에서는 약 10여년 전부터 주요 연구 기관에서 꾸준히 연구 개발을 하고 있어 기술 수준면에서는 거의 대등한 면을 보이고 있으나 EST를 분석하는 S/W개발측면에서는 국내 기술의 전무한 실정에 있다. 가장 대표적인 연구 기관으로서는 미국의 NCBI와 TIGR를 들 수 있는데 몇 년전부터 남아프리카공화국의 SANBI (South African National Bioinformatics Institute)에서 StackPack이라는 S/W를 개발하면서 거의 전 세계적으로 표준화된 EST 분석 기법으로 자리잡고 있다.



미국 TIGR의 대용량 Gene indices 프로젝트 서비스 화면을 예를 보면 현재까지 연구 등록된 모든 EST데이타를 분석하여 서비스를 하고 있다.

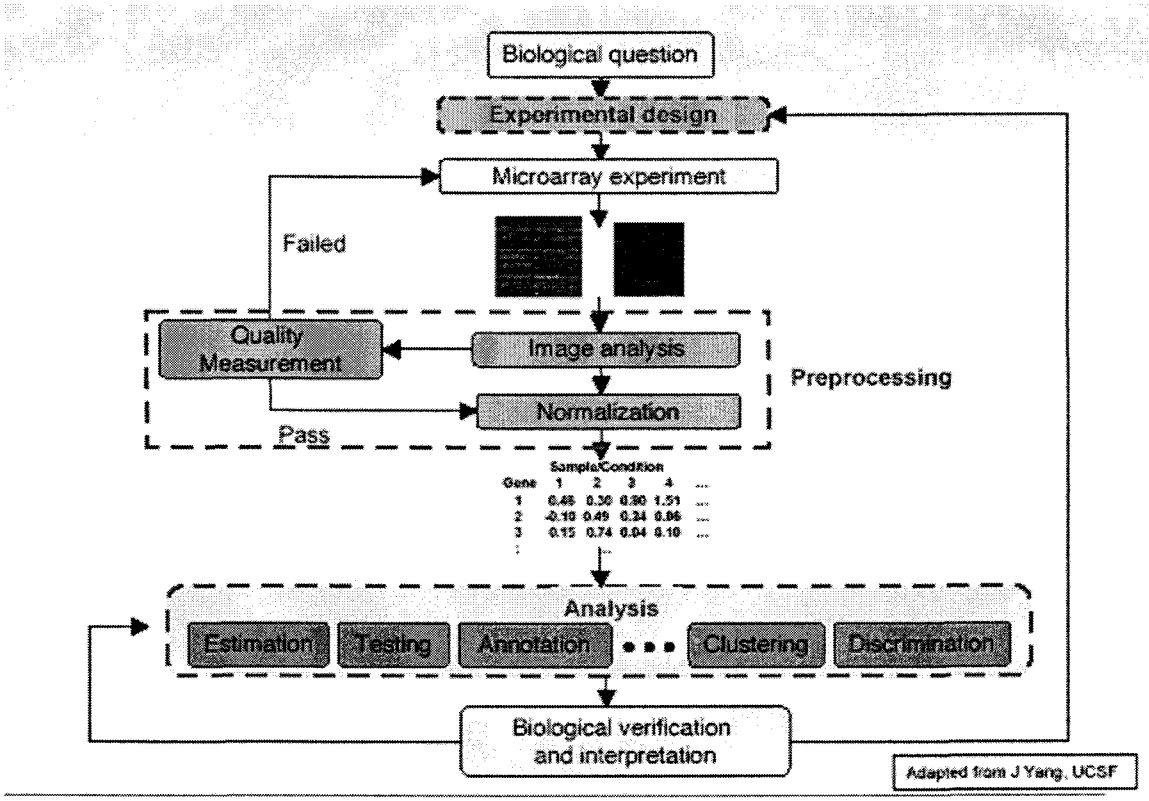


TIGR gene indices 서비스 내용 중에 Arabidopsis를 예를 들어 보인 것으로 alternative splicing , Gene Ontology, metabolic pathway 등 기능유전체연구에 필수적인 정보를 개발 서비스하고 있어 국내에서도 시급히 연구 개발에 투자를 하여 적극적인 대처가 필요한 실정이다. 본 연구에서는 이 TIGR 방식과 별도로 SANBI에서 개발된 StackPack S/W를 이용한 EST gene indices project 수행하고 있다.



EST 분석에 있어서 NCBI UniGene, TIGR의 방식, SANBI의 STACK 방식의 유전자 분석 결과를 비교 한 것으로 비교적 SATCK방식이 유용유전자 발굴에 효율적이라는

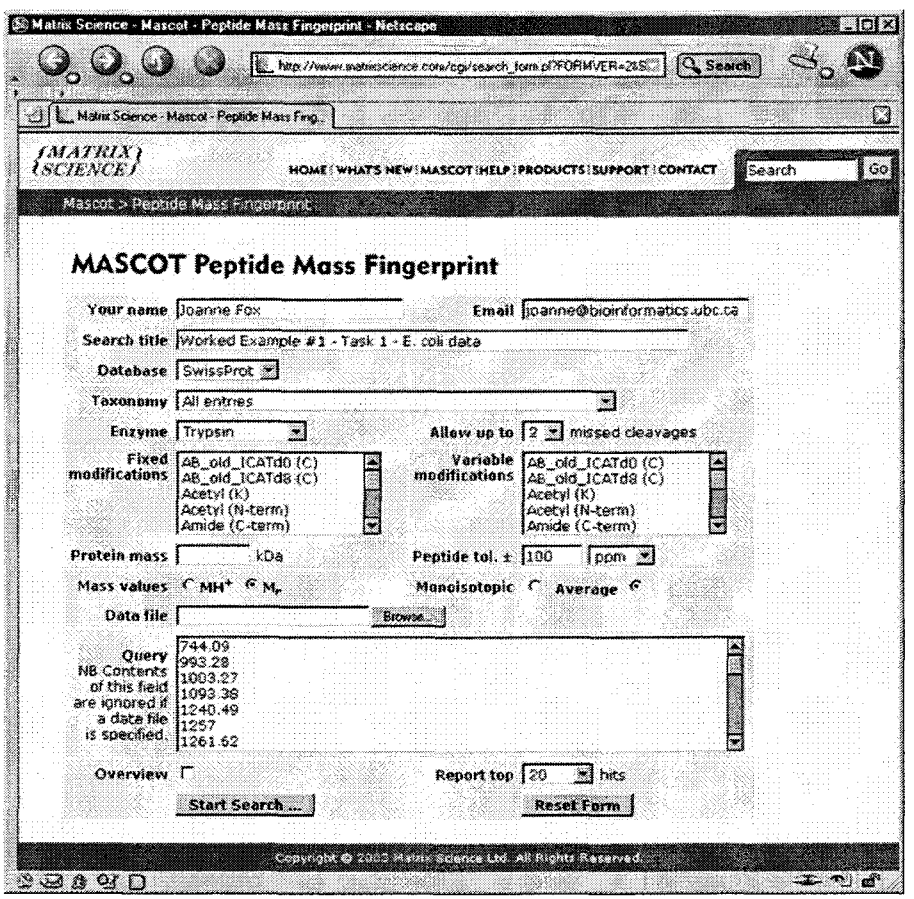
평가 때문에 본 연구에서도 SANBI의 STACK 방식을 택하고 있다.



Microarray 분석에 관한 프로그램의 사용은 국내에서도 많은 기관들이 필수적인 연구 방식으로 활용되고 있지만 국내에서 상용 S/W를 사용하는 것에 대한 문제점을 파악하고 원천 기술을 확보한 다음 새로운 기술에 따른 데이터 분석 기법을 고도화 하여 생명공학 연구자들에게 보다 정확한 gene expression 정보를 분석하는 것이 급선무라고 보여지면 이 분야에 대한 연구 개발은 국내 많은 대학 및 연구 기관에서 연구를 하고 있고 해외에서는 이미 기술 개발 자체가 고도화 되어 있다. 하지만 실직적으로 상용으로 보급된 분석용 S/W는 불편한 점이 많고 새로운 알고리즘에 대한 수정이 불가능하고 분석를 다양하게 보여주지 못하는 점 등을 고려할 때 편리하고 수정이 간편한 자체 S/W 개발이 필요하였다. 따라서 본 연구에서는 위 그림에서 보여주는 mciroarray 분석 전 과정을 연구하여 실질적으로 필요한 S/W를 개발하였다. 그 상세한 내용은 제 3 장에서 설명하게 될 것이다.

프로테옴 연구 분야에 있어서 MALDI-TOF를 이용한 데이터 분석 기법이 보편화 되면서 단백질 질량에 따른 Protein sequeince 인증을 위한 S/W가 필요한데 이미 세계적으로 우수성이 입증된 MOWSE, MASCOT, PROFOUND 등 약 10여개의 프로그램이 개발되어 있다. 그러나 이들 S/W들은 대부분 웹상에서 DB를 검색하여 서비스

하는 형태이며 분석 가능한 생물도 공개된 데이터베이스를 벗어나지 못하고 있어 국내에서 수행하는 고유의 유전체, 단백질체 연구 분야에 있어서는 무용지물이 되고 있다. 예를 들어 인삼 프로테옴 연구를 위하여서는 필수적으로 인삼 ES 프로젝트를 통하여 얻어진 EST를 gene index 작업을 하고 conceptual translation 작업을 거친 후 얻어진 단백질 서열을 데이터베이스화하여 직접 연구 할 수 밖에 없는 실정이다. 본 연구에서도 해외 공개된 S/W에 대응하기 위하여 새로운 알고리즘을 개발하고 적용하여 Peptide Mass Scpetrum 분석용 S/W를 개발하였다. 그리고 EST 분석, Microarray, Peptide mass spectrum 분석 등을 통합하고 유전자 분석에 따른 기능해석을 지원하기 위하여 gene ontology, MIPS function catalog 분석 지원은 물론 microarray 분석에 따른 co-regulation정보까지 분석하여 서비스하는 통합 서비스 시스템을 구축한 것이 매우 중요한 내용으로 판단된다.



PMF 분석에 관한 MASCOT 웹사이트의 예.

제 3 장 연구개발수행 내용 및 결과

가. 벼, 콩등 작물유전체 EST 데이터베이스 구축, 분석

1차년도를 통하여 작물 및 식물 유전체 분석을 자생식물사업단과 연동하여 개발하여 사용자들에게 유용유전자 정보를 제공하고 있다.

Plant 9종 , 11개 Tissue에 관한 EST 137만건 분석 완료

Status Report for Plant Organism

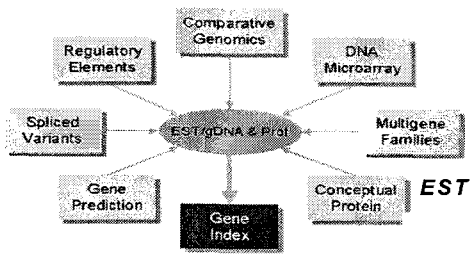
Species	Total Seqs	Cluster	Seqs in Cluster	Analysed Data		Full length clue
				Consensus	Singletons	
Arabidopsis thaliana	172,477	18,947	156,761	20,630	15,716	2,931
Glycine max	256,445	16,551	232,503	23,480	23,942	3,100
Hordeum vulgare	236,771	15,259	220,987	18,094	15,784	1,602
Lycopersicon esculentum	147,317	13,265	135,384	15,600	11,933	2,445
Medicago truncatula	172,364	14,485	152,474	16,737	19,890	2,314
Oryza sativa	97,583	10,656	80,087	12,140	17,496	520
Solanum tuberosum	73,057	10,123	64,106	12,219	8,951	1,644
Triticum aestivum	171,377	17,230	148,009	24,152	23,368	2,502
Zea mays	43,607	5,092	37,005	6,402	6,602	594
Total	1,370,998	121,608	1,227,316	149,454	143,682	17,652

Status Report for Plant Tissue

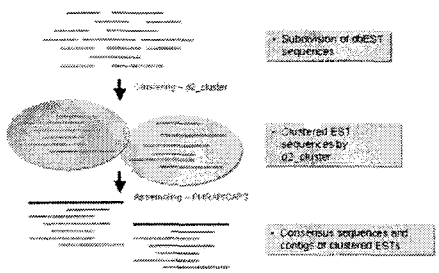
Tissue	Total Seqs	Cluster	Seqs in Cluster	Analysed Data		Full length clue
				Consensus	Singletons	
Aboveground	15,759	2,440	10,800	2,529	4,969	265
Callus	19,797	2,849	11,983	3,146	7,814	296
Flower	174,621	23,149	132,410	30,012	42,211	3,178
Fruit	36,753	5,126	25,853	5,629	10,900	627
Leaf	96,159	12,053	63,360	14,332	32,799	1,819
Nodule	9,268	1,314	6,617	1,451	2,651	298
Root	140,992	19,863	103,034	23,538	37,958	3,897
Seed	119,198	12,684	84,232	15,590	34,966	1,817
Seedling	97,348	11,382	74,362	14,695	22,986	2,284
Shoot	6,445	900	3,419	983	3,026	150
Stem	6,982	771	2,294	815	4,688	68
Total	723,382	92,551	518,304	112,720	204,968	14,899

GENE INDEX

1차년도 결과 DB화 및 2차년도
PMS 분석을 위한 개발 사항



EST clustering using STACKPACK



EST 분석을 위한 절차 및 활용도를 그림으로 나타내고 있다.
벼의 경우를 예를 들어서 분석 결과들을 도식화 하여 나타내고 있다.

Gene Index : Browsing

Introduction

Plant

- Statistics & Acknowledgement
- Gene Indexes Browsing/Table
  - By Tissue
    - Total / Species Unique
    - Dicotyledon/Monocotyledon
  - By Species
    - Total / Tissue Unique
- Search Gene Indexes
  - By Keyword
  - By User Sequences
- Function Categorization
  - PyFact
- Splice Variation
- CDS Candidate
- Categorized EST Subset
  - Subset Display
  - Download

Human

Mouse

KSPEPPER5kChip

Sweet Potato

Sesame

Ensembl - MIRROR

Swiss-Prot

Links

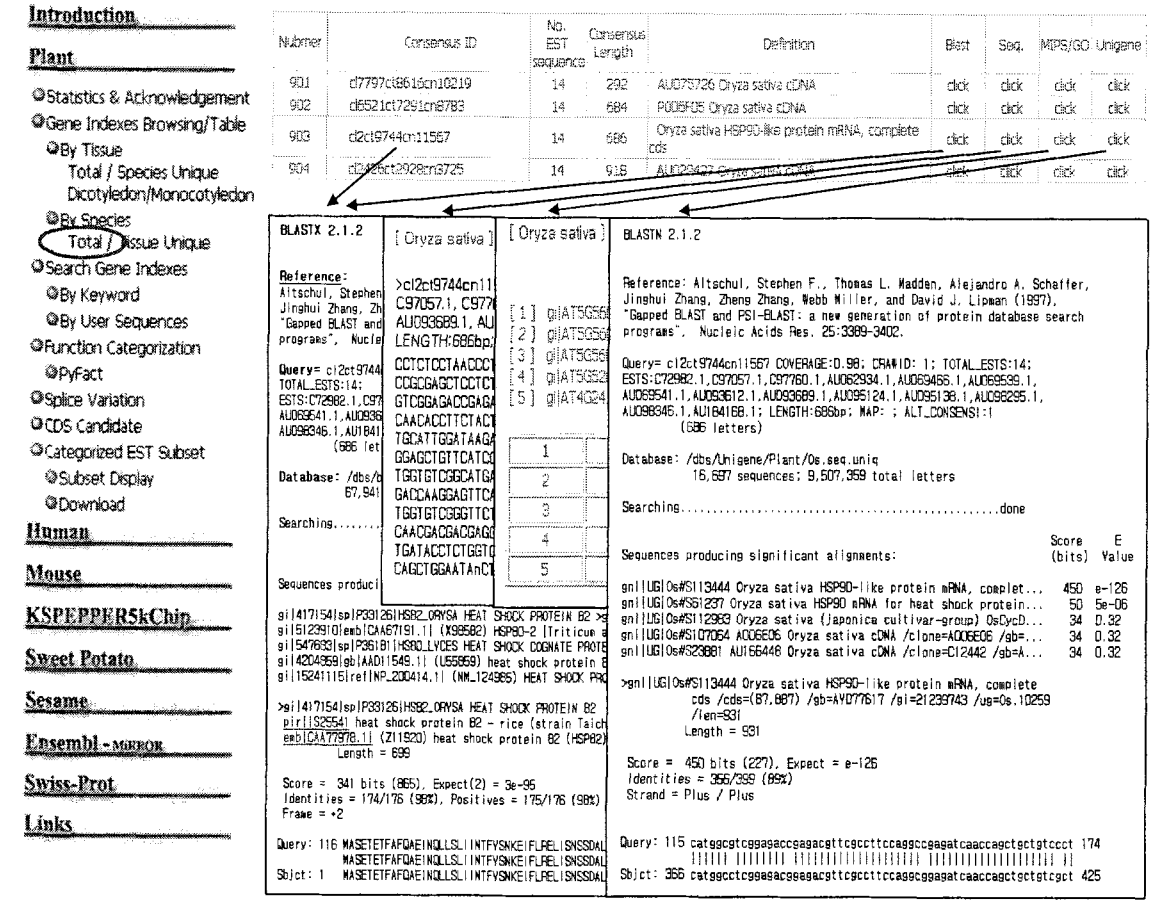
Gene Index Browsing/Table : Total by Species

[Oryza sativa]

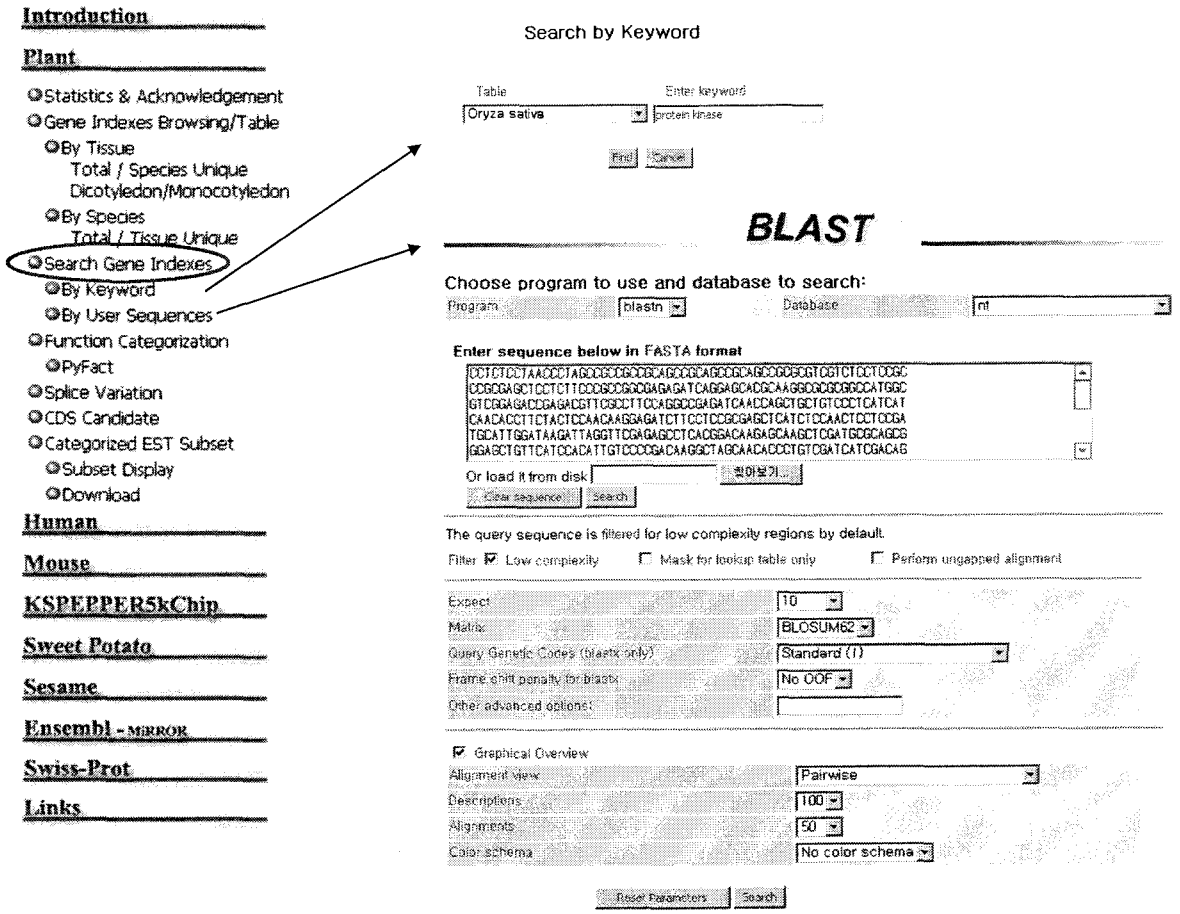
1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | NEXT >>

Number	Consensus ID	No. EST sequence	Consensus Length	Definition	Blast	Seq.	MIPS/GO	Unigene
1	d2ct9945cn12419	345	1914	K0010G03 Oryza sativa cDNA	click	click	click	click
2	d1386ct1913cn2440	241	1685	Oryza sativa aldolase mRNA, complete cds	click	click	click	click
3	d2ct9946cn12437	239	1441	Oryza sativa tRNA intron-encoded homing endonuclease mRNA, partial cds	click	click	click	click
4	d2ct9944cn12396	235	896	Oryza sativa DNA fragment with a miscellaneous signal and an open reading frame	click	click	click	click
5	d2ct9941cn12376	230	732	Oryza sativa mRNA for prolamin, complete cds, strain kinmaze RM1	click	click	click	click
6	d2ct9943cn12390	229	1750	Oryza sativa polyubiquitin (Rubq1) mRNA, complete cds	click	click	click	click
7	d2ct9939cn12266	222	860	H1033B05 Oryza sativa cDNA	click	click	click	click
8	d2ct9942cn12382	202	771	Oryza sativa (japonica cultivar-group) mRNA for allergenic protein, complete cds, clone RA5b	click	click	click	click
9	d1940ct2664cn3936	193	1248	O.sativa Waxy mRNA	click	click	click	click
10	d2ct9938cn12264	187	777	Oryza sativa metallothionein-like protein mRNA, complete cds	click	click	click	click
11	d2ct9936cn12227	186	2660	OF05G08 Oryza sativa cDNA, 5' end	click	click	click	click
12	d2ct9934cn12212	185	1280	Oryza sativa (japonica cultivar-group) mRNA for type I light-harvesting chlorophyll a	click	click	click	click
13	d2ct9935cn12216	183	1679	Rice mRNA for glutelin	click	click	click	click
14	d2ct9931cn12174	156	1232	AU068678 Oryza sativa cDNA	click	click	click	click
15	d1758ct2281cn2934	154	1193	AU103864 Oryza sativa cDNA	click	click	click	click
16	d2ct9930cn12163	152	1792	Oryza sativa OsrcaA2 mRNA for RubisCO activase small subunit precursor, complete cds	click	click	click	click
17	d2ct9927cn12132	141	2088	Rice mRNA for preproglutelin	click	click	click	click
18	d2ct9921cn12076	135	1071	Oryza sativa chlorophyll a-b binding protein mRNA, complete cds	click	click	click	click
19	d2ct9920cn12071	134	1534	Rice mRNA for aldolase C-1, complete cds	click	click	click	click
20	d154ct963cn1261	133	1005	Oryza sativa glycine-rich protein (OSGRP1) mRNA, complete cds	click	click	click	click

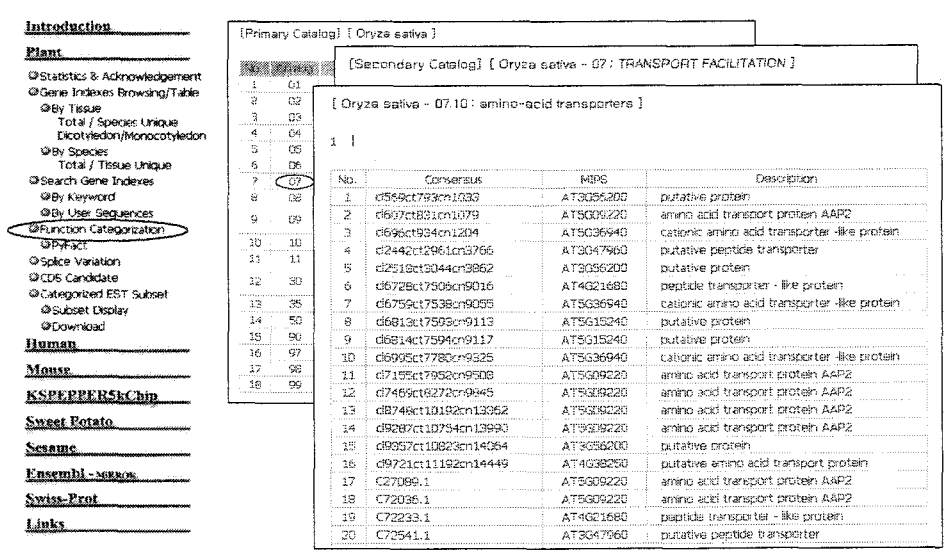
Gene Index : Browsing



Gene Index : Search



Gene Index : Function Categorization



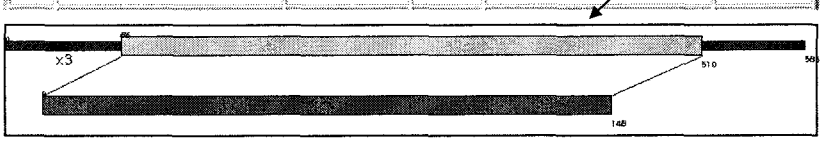
Gene Index : CDS Candidate

- **Introduction**
- **Plant**
  - Statistics & Acknowledgement
  - Gene Indexes Browsing/Table
    - By Tissue
      Total / Species Unique
      Dicotyledon/Monocotyledon
    - By Species
      Total / Tissue Unique
  - Search Gene Indexes
    - By Keyword
    - By User Sequences
  - Function Categorization
    - PyFact
  - Splice Variation
  - CDS Candidate
  - Categorized EST Subset
    - Subset Display
    - Download
- **Human**
- **Mouse**
- **KSPEPPER5kChip**
- **Sweet Potato**
- **Sesame**
- **Ensembl - MIRROR**
- **Swiss-Prot**
- **Links**

CDS candidate
[Oryza sativa]

1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | NEXT >>

No.	Consensus ID	Subject ID	Frame	Source	More
1	d2ct9454cn11152	729205	+1	ORYSA	
		729205	+3	ORYSA	
2	d2ct9495cn11206	130958	+1	ORYSA	
		82502	+1	Rice	
		1842176	+1	Oryza sativa (japonica c	
3	d2ct9532cn11252	485518	+2	Rice	
		18404062	+2	Arabidopsis thaliana	
		20330759	+2	Oryza sativa (japonica c	
4	d2ct9560cn11287	485518	+1	Rice	
		485518	+3	Rice	
		18404062	+1	Arabidopsis thaliana	
		18404062	+3	Arabidopsis thaliana	
		20330759	+1	Oryza sativa (japonica c	
		20330759	+3	Oryza sativa (japonica c	
5	d2ct9561cn11289	21740572	+1	Oryza sativa	
		21740569	+1	Oryza sativa	
		21740571	+1	Oryza sativa	



Gene Index : Alternative Splicing

Splice Variation

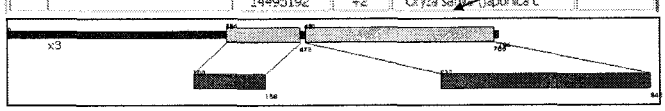
Species	Intron Retaining	Cryptic exon
Arabidopsis thaliana		
Hordeum vulgare		
Oryza sativa		

Splice Variation : Cryptic exon

[Oryza sativa]

1 | 2 | 3 |

No.	Consensus ID	Protein ID	Frame	Source	More
1	d2ct9368cn11065	100454	+1	Potato	
		100454	+1	Potato	
2	d2ct9596cn11344	15235241	+3	Arabidopsis thaliana	
		15235241	+1	Arabidopsis thaliana	
		19347816	+3	Arabidopsis thaliana	
		19347816	+1	Arabidopsis thaliana	
		15227815	+3	N/A	
		15227815	+1	N/A	
3	d92ct132cn160	14495192	+2	Oryza sativa (japonica c	
		14495192	+2	Oryza sativa (japonica c	



나. Peptide mass spectrum 분석을 위한 프로그램 개발 결과

프로테옴 연구에 있어서 가장 많이 활용되는 MALDI-TOF 분석에 나온 단백질 질량에 대한 서열을 판별하는 분석 기법으로 이에 국내 원천 기술이 필요하여 연구하게 된 것으로 앞으로 단백질 상호 작용 및 기능 해석 시스템과 결합하여 기능이 확대되어 나갈 것이다.

MALDI-TOF란?

최근에 개발된 질량분석기법중 하나로서 절차가 간단하고 단백질 같은 고분자의 질량을 빠르고 정밀하게 측정할 수 있도록 고안되어 있다. 고분자를 이온화하는 MALDI와 이온화된 고분자의 질량을 측정하는 장치인 TOF, 두 부분으로 나눌 수 있다.

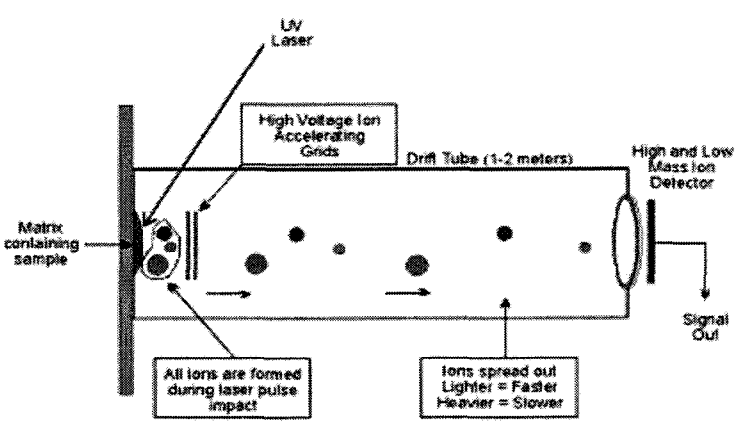
MALDI : matrix assisted laser distortion/ionization

올해 노벨 화학상을 수상한 일본의 다나카 고이치 등에 의해 개발된 방법이다. 고분자 즉 단백질을 적절한 화학물질(matrix)과 섞고 진공을 만들면 결정화된다. 여기에 레이저를 조사하여 단백질이 고체상태에서 떨어져 나와 이온화되게 하는 방법을 말한다.

TOF: Time Of Flight

이온화된 물질의 질량을 측정하는 장치이다. 이온화된 물질을 일정한 전위차로 가속시켜 검출기에 도달할 때까지 걸리는 비행시간을 측정하는 장치이다.

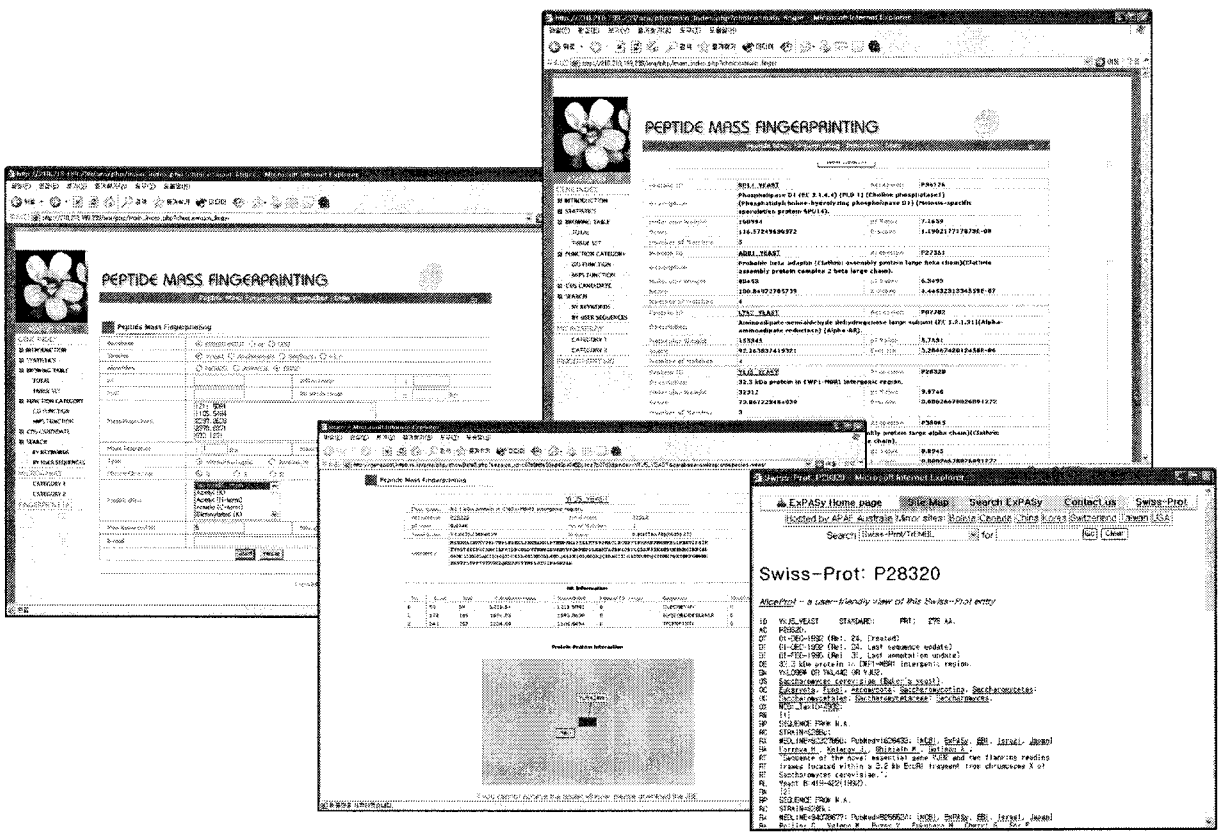
도달시간은 질량의 제곱근에 비례한다.



본 연구에서 개발된 내용을 아래에 기술한다.

Improved algorithms for the identification of yeast proteins and significant transcription factor and motif analysis

Peptide mass fingerprinting S/W개발 및 DB 구축 완료



With the rapid development of MS technology, the demand for more sophisticated MS interpretation algorithm hasgrown as well. We developed advanced molecular weight search (AMWISE) that makes up for the weakness of molecular weight search (MOWSE) and fingerprinting using binomial distribution (fBIND).

AMWISE and fBIND improved the performance up to maximum 49% and 2% as compared to the established algorithms, respectively. Moreover, we also suggest the statistical approach to define the significance of transcription factor and motif in the identified protein based on the Gene Ontology (GO).

As the tendency in biology researches has been moved from analysis on few genes/proteins to macroanalysis on more extensive genes/proteins, MS has been recognized as key biotechnology. In particular, MS can be considered asthe most important tool in performing proteomics to understand biological phenomenon of living organisms at the large-scale protein level. Accordingly, a number of recent researches presented several software and algorithms that can effectively analyze the results of MSand relevant researches have been actively continued [1-3]. While PepSea and PeptIdent/

MultiIdent identifiedproteins on the basis of the number of matches between peptide molecular weight in database and peptide from experiments, they didn't support accurate analysis [4-6]. MOWSE used scoring method that reflected characteristics of database beyond simple matches as calculating frequencies of peptide generated from proteins at 10kD intervals [7]. The researches based on probability are MASCOT that identifiedmore reliable analysis as calculating the probability that random hit could occur, ProFound that calculated probability values of proteins using Bayesian approach and research by Wool *et al* that made scoring method using binomial distribution [8-10].

While MOWSE scoring algorithm that is widely used in general is simple and relatively accurate, it has the limit to depend on relatively large masses among peptide fragments in experiments. In accordance with the research by Pappin *et al.*, most proteins are distributed around 50kD and peptides created as cutting these proteins in trypsin are mainly small fragments [7]. As reported above, yeast proteins also showed the same distribution (data not shown). While most peptides from experiments are small fragments, fragments don't have substantial effects on MOWSE score, but a small number of fragments with relatively large molecular weight fragment significantly influence on MOWSE score. As shown in Figure 1(a), for the performance of MOWSE scoring, its accuracy is substantially decreasedwhen molecular weights of peptides are small. This phenomenon is also appeared according to intact molecular weight of protein as shown in Figure 1(b). The reason is that proteins with larger molecular weights create more peptide fragments and on thisoccasion, a number of fragments with lower molecular weights are generated. This MOWSE scoring algorithm is also applied to other peptide mass fingerprinting (PMF) programs such as MASCOT and MS-Fit [8, 11]. Although scoring algorithm in the research by Wool *et al.* was based on relatively simple probability, its performance was very remarkable as shown in Figure 1 [10]. It demonstrated uniform accuracies regardless of peptide fragment ranges and intact molecular weights of proteins as shownin Figure 1. However, its accuracy tended to be slightly declined because of repetitive sequences. Accordingly, this study developed AMWISE and fBIND to identify proteins more accurately as making up for disadvantages ofexisting algorithms.

Peptide pool to analyze PMF was generated using SWISS-PROT 4.2 release and nonredundant (nr) of NCBI. Then, only yeast proteins were collected in SWISS-PROT 4.2 and NCBI nr and then, theoretical peptide pool created by trypsin activity was built up. This peptide data set consisted of only peptides in 500-4000Da that was experimentally significant among molecular weights of peptides. Moreover, in consideration of errors caused by biological experiments, this study assumed the missed cleavage level of trypsin to level 2 and peptides generated on this level were included in peptide pool. Next, peptide pool generated by the process above and data related to yeast protein were used to construct database using MySQL 4.2.13 DBMS. Yeast proteins in the database were 5,406 for SWISS-PROT 4.2 and 9,007 for NCBI nr. Peptide pool generated by trypsin consisted

of 718,106 and 1,040,173 records for SWISS-PROT 4.2 and NCBI nr, respectively.

Scoring method used in AMWISE is the same as MOWSE in the aspect that both of them are based on frequencies of fragments generated as dividing intact protein molecular weight and peptide molecular weight by the unit of 10kD and 100Da, respectively. However, scoring method in AMWISE calculated weight scores of each peptide molecular weight using the expressionbelow and then, the weight scores were included in scores.

On the assumption that theoretical peptide mass of ith hit of protein K is m_i and m_i is included in jth peptide range and kth protein range,

F_{jk} = frequency of fragment range j and intact protein range k, 5=j=40, 0=k=N,

N = maximum of intact protein molecular weight/10000

$$V_{jk} = F_{jk} / \max\{F_{5k}, F_{6k}, ..., F_{40k}\}$$

$$W_j = \log_{10} \sum_{k=0}^{N} F_{jk}$$

Score of protein K = mean of intact protein molecular weight X $\prod_{i=0}^{n} (V_{jk} \times W_j)$ of m_i / molecular weight of protein K, n = number of hit

Like the research by Wool *et al.*, scoring method in fBIND calculated probabilities with respect to binomial distribution as described below and applied overlap penalty to reduce impacts from repetitive sequences and proportional relation between molecular weights of proteins and the number of peptides [10].

$$P(N,r) = \binom{N}{r} p^r (1-p)^{N-r} \times e^k$$

N = total number of peptides in a protein

r = number of random match

p = number of match / number of peptide in database

k = frequency of overlapping match in a protein

Next, the results by AMWISE and fBIND were compared to those by MOWSE and researches by Wool *et al.* in order to evaluateperformances of AMWISE and fBIND that were modified above. The test sets for analysis and comparison of performance randomly selected 100 proteins from SWISS-PROT 4.2 release and then 10 peptides from each theoretical peptide pool. Next, random error values among e = {-0.999, -0.998, , 0, , 0.998, 0.999}in the calculated mass of

each selected peptide were added up for the test sets. These test sets were divided into the set by molecular weight of a peptide (<1500Da~<4000Da) and that by molecular weight (0kD~200kD) of a protein and each set was independently analyzed. It was assumed that missed cleavage was 0, mass tolerance was 1Da and there was no modification. Consequently, as shown in Figure 2, AMWISE demonstrated far higher performances than MOWSE. As shown in Figure 2(a, b), MOWSE showed low accuracies of 40.4% (SWISS-PROT 4.2 release) and 12% (nr) in PMF with peptides of less than 1500Da. The accuracies of MOWSE were gradually improved in accordance with increases of molecular weights of peptides. For peptides below 4000Da, the performances of SWISS-PROT 4.2 release and nr were improved up to 92% and 79%, respectively. This phenomenon in accordance with peptide ranges also influenced on performances according to intact protein molecular weights. As shown in Figure 2(c, d), as the sizes of proteins became larger, performances were gradually decreased. Then, in the range between 190kD and 200kD, the accuracies in SWISS-PROT 4.2 release and nr reached 57% and 31%, respectively. Meanwhile, as shown in Figure 2(a, b), AMWISE solving underestimate on small peptides substantially made up for disadvantages of MOWSE. Then, it demonstrated accuracies of 88% and 62% for SWISS-PROT 4.2 release and nr evenin peptide regions below 1500Da and 97% and 91% for SWISS-PROT 4.2 release and nr in peptide regions below 4000Da, respectively. These accuracies were up to 47% and 49% higher than those in MOWSE for SWISS-PROT 4.2 release and nr, respectively. However, the accuracies slightly depended on peptide molecular weights even in AMWISE. This suggested that the underestimate issue was not completely solved. As shown in Figure 2, the performances of fBIND didn't show significant differences as compared to researches by Wool et al. It is because scoring method of Woolshowed higher performances of 99.1% and 97.9% in average in accordance with peptide ranges and protein ranges, respectively, when SWISS-PROT 4.2 release was applied, as shown in Figure 1. fBIND that considered influences by increases of random hit rates in accordance with increases of peptide fragments and repetitive sequences demonstrated higher performances as compared to researches by Wool *et al.*As illustrated in Figure 2(a, b), the performances in accordance with peptide ranges were 99.6% and 97.1% for SWISS-PROT 4.2 release and nr, respectively, with increases of about 1% and 2% for SWISS-PROT 4.2 release and nr, respectively, as compared to study of Wool *et al.* The performances in accordance with sizes of proteins were also improved. The performances were 99.6% and 90.9% in average for SWISS-PROT 4.2 release and nr, respectively, with increases of up to 10% and 7% for SWISS-PROT 4.2 release and nr, respectively, as compared to researches by Wool *et al.*

This study tried to identify proteins more accurately as improving existing algorithms and analyzed significant transcription factor/motif analysis to provide useful information on identified proteins. MATCH and PATCH of TRANSFAC, representative transcription factor analysis tools, are very useful to find out transcription factors binding on specific sequences using position-specific matrixes and patterns [12, 13]. Moreover, InterPro, the representative motif database, is important

database collecting motifs in each protein [14]. However, all of them don't provide information on how much important it is the contribution of transcription factors or motifs identified by those tools and database on functions. In accordance with analysis of transcription factors existing on upstream of yeast ORF by MATCH and PATCHin reality, it was observed that a number of transcription factors were abundantly appeared regardless of specific functions including that HIF-1 was appeared up to 1614 times on total 5,406 upstream. Therefore, this study tried to analyze significant transcription factors and motifs contributing on specific functions using cumulative hypergeometricprobability distribution as described below.

First of all, yeast ORF and proteins were divided into relatively detailed function categories of about 105 with reference to data annotated with respect to the process among GO terminology in Saccharomyces Genome Database (SGD) [15]. Then, mapping of ORF region on chromosome was conducted using sim4 to acquire transcription factors binding on upstream of ORF included in each function category [16]. As reported by Zhu *et al.*, -1000 regions froma translation start site was considered as upstream region and transcription factors in relevant upstream region were analyzed by MATCH and PATCH [17]. Motifs of each protein were acquired from InterPro database. Transcription factors and frequencies of motifs in each function category segmented into 105 categories were calculated and then, transcription factors andmotifs characteristically appeared in each function category were analyzed through cumulative hypergeometric probability distribution as shown below.

$$P\{x=i\}=\frac{\binom{m}{i}\binom{N-m}{r-i}}{\binom{N}{r}}$$

N = Total number of ORF/proteins

r = Number of ORF/proteins in a specific category

m = Number of specific transcription factors/motifs identified in total ORF/proteins

i = Number of specific transcription factors/motifs identified in ORF/proteins in a specific category

$$P\{X \geq i\}=\sum_{j=i}^{r} P\{X=j\}$$

When $P\{X \geq i\} \leq 0.001$ satisfied, relevant motif is are considered as the motifs specifically generated in specific function categories. Table 1 described the examples of transcription factors and motifs specifically generated in each function acquired by the process above.

Table 1 describedmotifs and relevant characteristics existing in each protein. O13527 has integrase motif as the protein included in DNA recombination among 105 segmented categories. As a result

of analysis using cumulative hypergeometric probability distribution, integrase catalytic domain is a specific motif that is appeared especially a lot in DNA recombination category and it is considered that it contributes on the functions related to DNA recombination of O13527. This result is the same as GO mapping results of InterPro. P00330 falls under GO:0006113 fermentation according to GO mapping of SGD and includes in energy pathway, the upper category. Zinc-containing alcohol dehydrogenasesuperfamily, a motif found in P00330, is especially appeared a lot in energy pathway category, alcohol metabolism and aldehyde metabolism category. It contributes on functions related to energy pathway of P00330 and has the possibility to take part in other functions such as alcohol metabolism and aldehyde pathway. InterPro doesn't provide information on this motif. In accordance with GO mapping of SGD, P00410 was mapped to GO:0009060 aerobic respiration and included in energy pathway, the upper category. Copper center cu(A), an identified motif, was a specific motif neither to energy pathway category nor to other categories. However, cupredoxin, another motif, was appeared especially in the same category as P00410 and also the specific motif in ion transport and response to abiotic stimulus category. As explained above, this study provided information about contribution of motifs identified in each protein through cumulative hypergeometric probability distribution on functions and suggested to broaden function annotation more extensively. In accordance with annotation of yeast protein with respect to process among GO mapping data of InterPro, about 53% of total yeast proteins could be annotated [18]. Meanwhile, when significant factor/motif analysis of this study was combined with InterPro, about 83% of total proteins were covered.

As large-scale protein researches have been analyzed, researches on instrument related to MS, experimental techniques and PMF algorithm have been actively studied. In particular, we need the algorithm to identify accurate proteins without being sensitive to experimental errors in order to identify effective proteins. In accordance with the development of MS, a wide range of research results from algorithms of simple matches to algorithms based on complicated probabilities has been reported. Each algorithm demonstrates unique characteristics in accordance with scoring methods. For example, MOWSE scoring method based on frequency that consists of mainly peptide fragments show lower performances, but that with peptide fragments of high molecular weights demonstrates significantly higher performances. Moreover, for PMF based on binomial distribution presented in the researches by Wool *et al*, the performance tends to be slightly decreased when measured peptide is randomly matched to proteins with repetitive sequences. Consequently, this study developed the system showing better performances as making up for scoring methods based on MOWSE and binomial distribution among PMF algorithms that have been studied until now. Furthermore, this study tried to provide information for interpreting peptide fingerprinting results for researchers as analyzing significant factors/motifs related to regulations and activities of identified proteins as well as for accurate identification of proteins. In the future, this study will

provide information on AMWSIE and fBIND with respect to protein-protein interaction data and protein sub-cellular localization information. AMWISE and fBIND is available at http://plant.pdrc.re.kr:8888/peptMass/yeastPMF/ and supplementary information is available at http://plant.pdrc.re.kr:8888/peptMass/yeastPMF/Instruction/instruction.html

Figure 1. Performance Comparison of MOWSE and Wool *et al.*'s study.

(a) Performance comparison of MOWSE (squares) and Wool *et al.*'s study (circles) according to peptide ranges using SWISS-PROT 4.2 release; (b) Performance comparison of MOWSE (squares) and Wool *et al.*'s study (circles) according to protein molecular weight using SWISS-PROT 4.2 release.

Figure 2. Performance Comparison of Algorithms.

(a) Performance of MOWSE (squares), AMWISE (diamonds), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance with peptide ranges using SWISS-PROT 4.2 release; (b) Performance of MOWSE (squares), AMWISE (diamonds), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance with peptide ranges using nr; (c) Performance of MOWSE (squares), AMWISE (diamonds), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance with protein molecular weight using SWISS-PROT 4.2 release; (d) Performance of MOWSE (squares), AMWISE (diamonds), Wool *et al.*'s study (circles) and fBIND (triangles) in accordance protein molecular weight using nr.

Figure 1

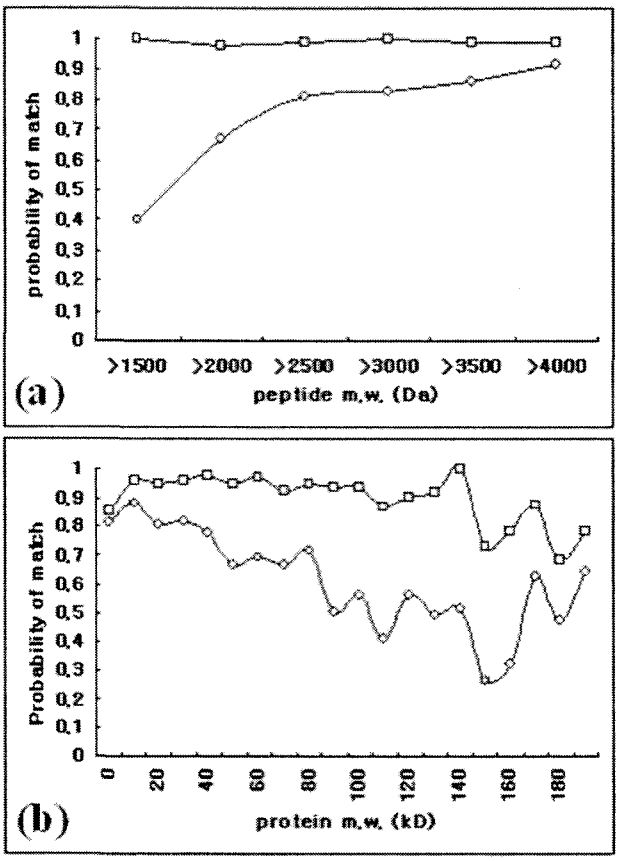

probability of match
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
>1500
>2000
>2500
>3000
>3500
>4000
peptide m.w. (Da)
(a)
Probability of match
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
20
40
60
80
100
120
140
160
180
protein m.w. (kD)
(b)

Figure 2

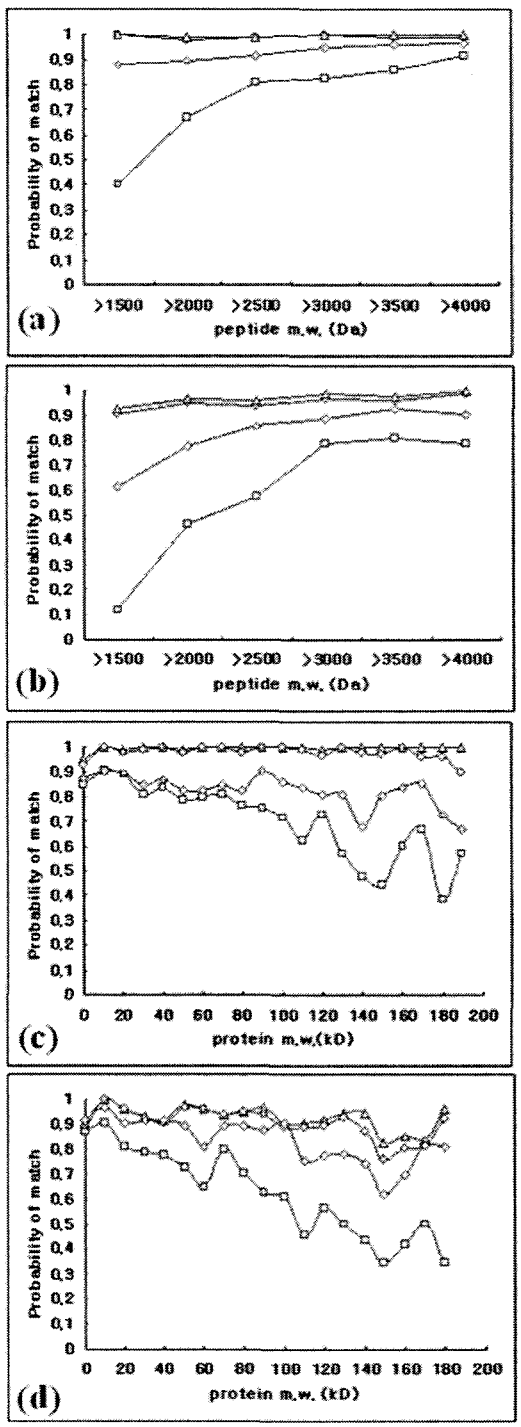
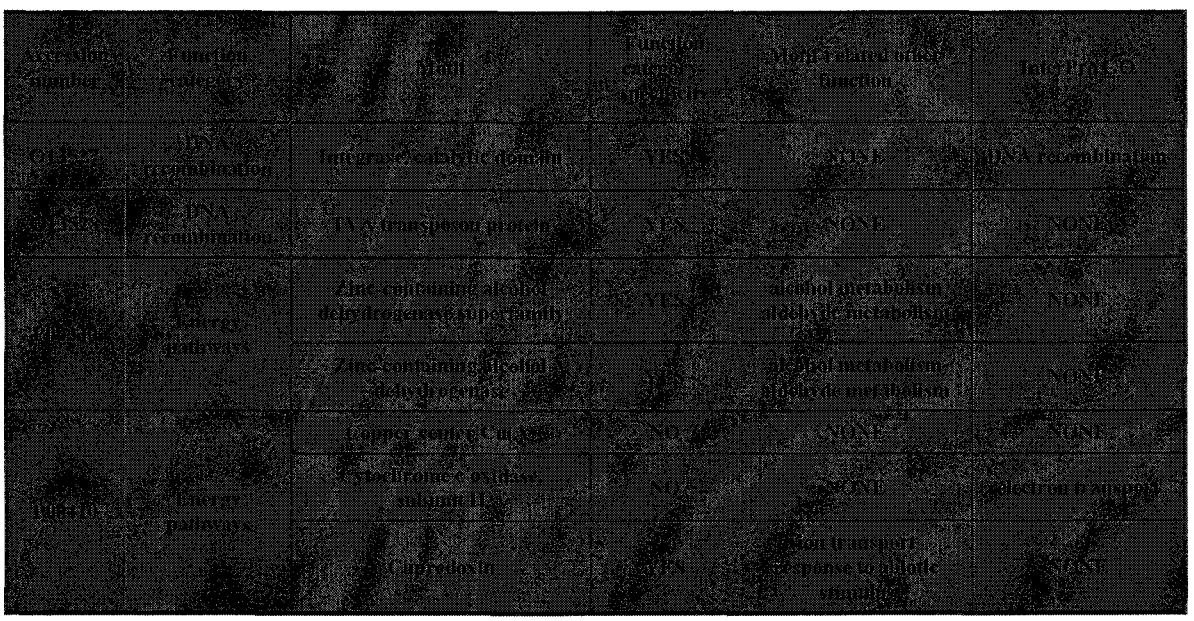

Probability of match
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
>1500 >2000 >2500 >3000 >3500 >4000
peptide m.w. (Da)
(a)
Probability of match
>1500 >2000 >2500 >3000 >3500 >4000
peptide m.w. (Da)
(b)
Probability of match
0 20 40 60 80 100 120 140 160 180 200
protein m.w.(kD)
(c)
Probability of match
0 20 40 60 80 100 120 140 160 180 200
protein m.w. (kD)
(d)

Table 1. Description about function specificity of motifs of protein that is within a particular function category and comparison to InterPro GO mapping

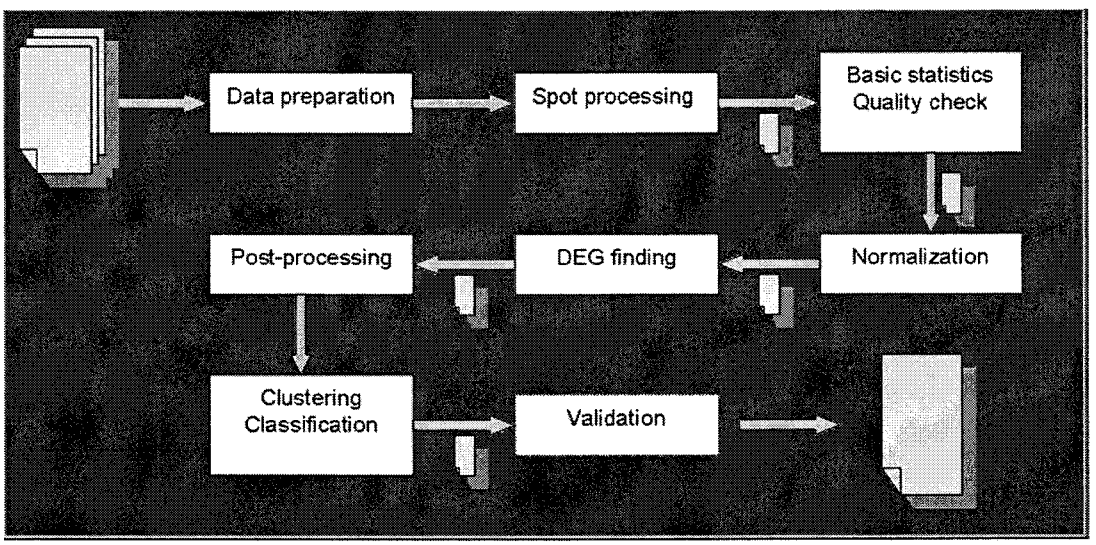


This work was supported by a grant (CG1222) from Crop Functional Genomics Center, by a grant (PF0300501-00) from the Plant Diversity Research Center of the 21st Century Frontier Research Program funded by MOST of the Korea government. We thank Sun Yong Park at the Korean Research Institute of Bioscience and Biotechnology for the advice.

다. Microarray 데이터 분석을 위한 S/W개발

기존의 상용 Microarray 프로그램의 경우 약 2 000만원 정도의 가격으로 판매되면서 기능이 상당히 제약되어 있고 새로운 알고리즘을 첨가할 경우 불가능한 점이 불편하여 본 과제를 통하여 R 기반 통계 프로그램을 이용하여 Microarray 핵심 프로그램을 개발하게 되었다.

본 연구에서 개발한 Pipeline형 DNA microarray 분석 툴 기본 모델
데이터 분석 모델



기본적인 분석 툴 개념

- 최대한 공개 소프트웨어(R 등)를 활용, 개발과 관리에 드는 자원을 줄임
- 기본적으로 유닉스 기반의 분석 파이프라인을 고려
- 분석의 방법과 내용을 가능한 표준화 (보고서 양식의 표준화)
- 실제 데이터분석의 파이프라인에 필요한 컴포넌트 함수와 wrapper 함수를 개발하는 데 주력
- 대량의 자료를 처리하도록 고려

각 단계별 Task 및 수행방법

l **Data preparation**

n 본격적인 대량의 데이터분석에 앞서 분석에 필요한 데이터를 표준화하여 동일한 포맷으로 맞춤

n **이미지 처리**

u 이미지 만 전달된 데이터에 대해 프로세싱을 수행 수치데이터를 획득

u 수행방법: 개별 파일에 대해 ***Imagene*** 소프트웨어를 이용하여 수행 (작업 부하가 큼)

n **데이터 파일 포맷 정리**

- u 모든 파일의 포맷을일정하게 맞춤 (예: txt 타입, xls 타입 등)
- u 수행방법: 수작업 (작업 부하가 큼) 혹은 대량의 처리를위한 새로운 처리함수 필요

n **데이터 확인**

- u 여러 슬라이드 데이터를 처리하므로 각 파일의 내용이 일치하는지를 확인. 특히 각 파일의 유전자 이름과 각 열의 이름이 모두 동일한지를 확인
- u 수행방법: 새로운 처리함수 필요

n **중복유전자 확인**

- u 각 슬라이드에 중복으로 심어진 유전자 확인
- u 향후 데이터 품질분석 등에 유용
- u 수행방법: 새로운 처리함수 필요

n **각 유전자 블록표시 통일**

- u 각 블록의 id가 순차적인 번호(1, 2,)로 지정된 경우, "메타 행 메타 열" 형태의 정보를 추가
- u 각 블록의 id가 "메타 행 메타 열"의 형태로 지정된 경우, 각 블록에 순차적인 번호를 부여
- u 수행방법: 새로운 처리함수 필요

l **Spot processing**

n 각 슬라이드별로 오류인 스팟에 대해 표지를 붙여 향후 분석에서 활용할 수 있도록 함

n 오류 스팟의 예

- u 배경강도에 비해 스팟강도가 작은 스팟
- u 각 색깔별로 스팟강도가 스캐너의 최대값에 가까우면서 두 색깔 사이의 차이가적은 스팟
- u 각 색깔별로 신호와 잡음 사이의비(signal-to-noise ratio)가 작은 스팟
- u 각 색깔별로 스팟강도의 분포와 배경강도의 분포 사이에 통계적인 차이가 적은 스팟 (예: t-검증의 p-값이 큰 스팟)
- u 각 색깔별로 스팟강도가 네거티브 컨트롤의 스팟강도에 비해 작은 스팟
- u 특정 색깔의 형광물질에 특이적으로 반응을 보이는 스팟(dye-swap 실험을 수행한 경우에 해당)

n 수행방법: 새로운 처리함수 필요

l **Basic statistics & quality check**

n 각 슬라이드 데이터의 품질을 확인하고자 필요한 기초통계조사를 수행

n 데이터의 품질을 조사하여 품질이 떨어지는 슬라이드를 제거, 분석결과의 신뢰도를 높이기 위한 중요한 과정

n 조사내용의 예

- u 각 슬라이드별
  - l 유전자별 기본적인 데이터의 분포 (예: R, G의 스팟과 배경에서의 강도 분포, A(= log2RG), M(= log2R/G)의 분포에 대한 히스토그램 혹은 밀도함수그림
  - l 블록별 기본적인 데이터의 분포에 대한 히스토그램 혹은 밀도함수그림
  - l A-M 그림
  - l 중복유전자(예: 컨트롤유전자) 발현의 균일성
  - l 정상 스팟의 수
- u 동일한 실험조건에서의 중복슬라이드 사이에서
  - l 동일한 실험조건의 임의의 두 슬라이드 사이에서 각 유전자별로 R, G, A, M의 분포 (S-plus 혹은 R에서 pairs plot) 및 상관계수
  - l 동일한 실험조건의 임의의 두 슬라이드 사이에서 각 블록별로 R, G, A, M의 분포 및 상관계수
- u 서로 다른 실험조건에서의 슬라이드세트 사이에서
  - l 임의의 두 슬라이드 사이에서 각 유전자별로 R, G, A, M의 분포 (S-plus 혹은 R에서 pairs plot) 및 상관계수
  - l 임의의 두 슬라이드 사이에서 각 블록별로 R, G, A, M의 분포 및 상관계수

l **Normalization**
- n 각 슬라이드 데이터를 상호 비교가 가능하도록 규격화함
- n 수행방법: *Bioconductor*의 *marrayNorm* 패키지를 활용

l **DEG(Differentially Expressed Gene) Finding**
- n 각 슬라이드별 혹은두 실험조건별로 발현량에 차이가 나는 유전자를 선정
- n 전체 분석의 과정에서 가장 중요한 단계
- n 수행방법: *Bioconductor*의 *edd, genefilter, multtest, ROC* 등의 패키지와 기타 패키지를 활용

l **Post-processing**
- n 유의한 발현차를 보이는 유전자의 선정이 끝난 후 데이터를 처리
- n 대표적으로 발현행렬구성과 없는 데이터의 처리 등을 수행
- n **발현행렬구성 (Expression matrix construction)**
  - u 각 슬라이드별로 관리된 자료를 clustering / classification에서 활용할 수 있도록 발현행렬(expression matrix)형태로 변환
  - u 수행방법: 새로운 처리함수 필요
- n **없는 데이터 처리 (Missing value imputation)**
  - u 발현행렬에서 값이 없는 셀을 포함하는 유전자를 처리
  - u 처리의 예: 값이 없는 셀의 값을 적당히 추론하여 메우거나 혹은 값이 없는 셀

이 있는 유전자를 향후 처리에서 배제

- u 수행방법: 새로운 처리함수 필요

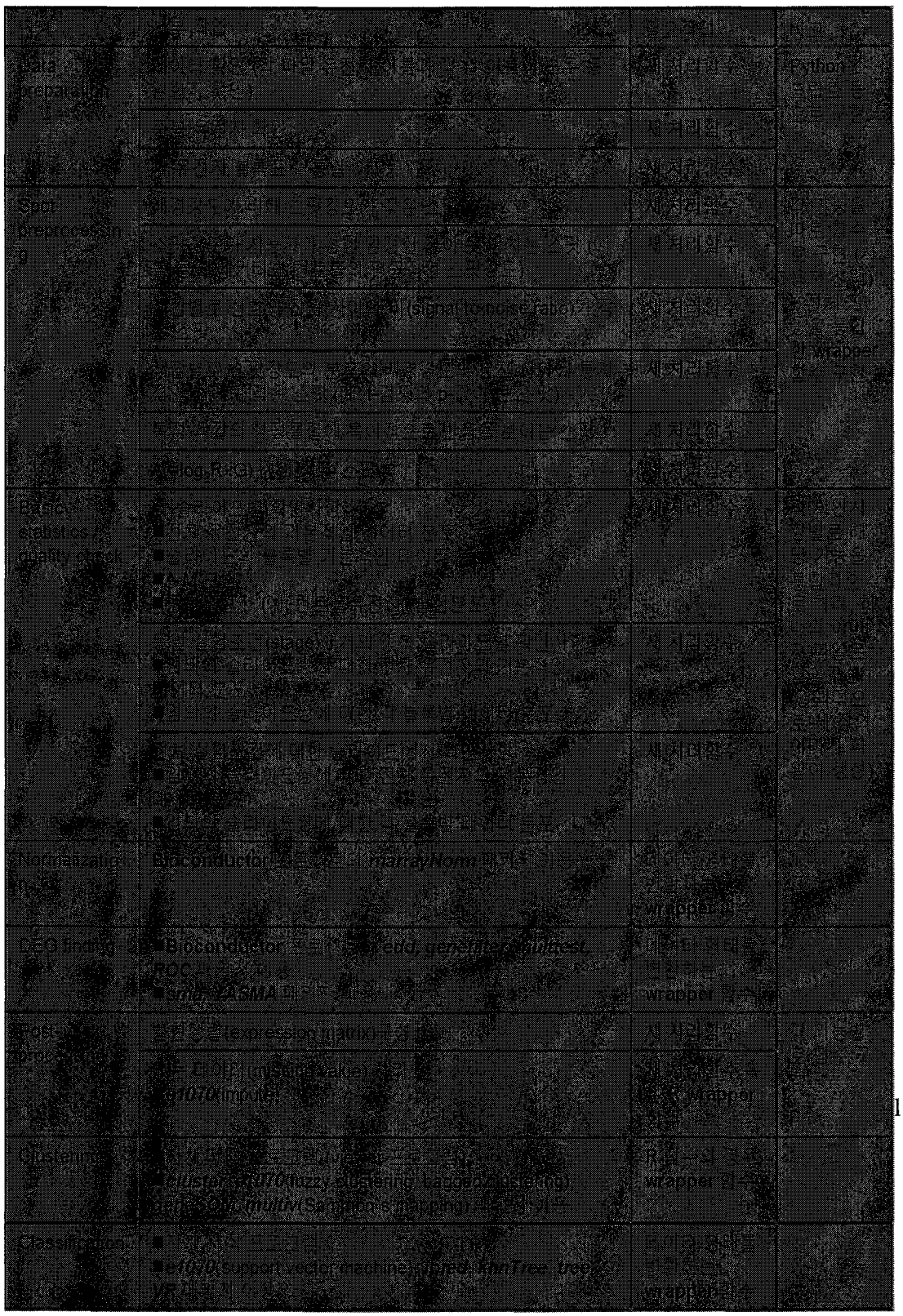
l **Clustering / Classification**

- n 시계열 혹은 다양한실험조건에 의한 발현행렬을 이용, 클러스터링 혹은 분류 등의 작업을 수행
- n 수행방법: R의 각종 패키지 활용

각 단계별 필요작업과 일정

l Raw data 인수 시 필수 데이터

- n 각 슬라이드별 이미지 데이터 (jpg 타입 권장)
- n 각 슬라이드별 수치 데이터 (txt 혹은xls 타입 권장)
- n 기타 MIAME 기본 사항 데이터 (txt 혹은xls 타입 권장) 기본 포맷은 NCGI에서 제공



기타 단계별 필요작업과 일정

An open source microarray data analysis system with GUI: Quintet

Abstract

We addressQuintet, an R-based unified cDNA microarray data analysis system with GUI. Five principal categories of microarray data analysis have been coherently integrated in Quintet: data processing steps such as faulty spot filtering and normalization, data quality assessment (QA), identification of differentially expressed genes (DEGs), clustering of gene expression profiles and classification of samples. Though many microarray data analysis systems normally consider DEG identification and clustering/classification the most important problems, we emphasize that data processing and QA are equallyimportant and should be incorporated into the regular-base data analysis practices because microarray data are very noisy. In each analysis category, customized plots and statistical summaries are also given for users convenience. Using these plots and summaries, analysis results can be easily examined for their biological plausibility and compared with other results. Since Quintet is written in R, it is highly extendable so that users can insert new algorithms and experiment them with minimal efforts. Also, the GUI makes it easy to learn and use and since R-language and its GUI engine, Tcl/Tk, are available in all operating systems, Quintet is OS-independent too.

Introduction

DNA microarray is the *de facto* standard technology for high-throughput functional genomics in the post-genomic era [1]. Since the microarray experiment is highly evolved and requires multiple handling steps each of which is a potential source of fluctuation which undermines the reliability of the data itself [2], much effort has been exerted to understand the sources of variability and minimize them to produce high-quality reproducible data [3].

In order for this technology to be fruitful, however, reliable analysis of the data is as important as the production of high-quality data itself. Due to the high-throughput character of the microarray data, this requires maturity in numerous statistical techniques, not to mention the data processing chores. Also required is the dexterity in scrutinizing various pertinent biological information so that one can successfully reconstruct the 'big picture"of biological processes fragmentally reflected in the data. Considering these, a system that can provide analytic capability as well as informatic capability is crucial for an effective, versatile analysis of microarray data. In addition, since many new approaches appear almost daily by researchers, an ideal system should be extandable enough so that new techniques can be experimented with minimal efforts.

In this article, we present an R-based unified cDNA microarray data analysis system, Quintet, the first result of our on-going project to build up a customized microarray data analysis suite. As the name suggests, th five indispensable categories of data analysis have been coherently integrated in Quintet: data processings including filtering and normalization, customized set of data quality assessments (QAs), identification of differentially expressed genes (DEGs), clustering of gene expression profiles, and classification of samples using a small set of gene expression patterns.

Though many microarray data analysis systems claim DEG identification and clustering/classification the most important problems, we emphasize that data processing and QA are equallyimportant and should be incorporated into the regular-base data analysis practices because the microarray data are quite noisy [2, 4]. Under this rationale, some set of data processing and QA procedures are implemented in Quintet and constitute the core functionality moduleof Quintet. Quintet is written in R which is virtually the standard platform for microarray data analysis now. Since many new algorithms are also written in R, they can be inserted into Quintet without much trouble andusers can extend its functionality for their own needs. The GUI makes it easy to learn and use Quintet. Furthermore, Quintet is OS-independent since R-language and its GUI engine adapted for Quintet, Tcl/Tk, are available in all operating systems.

Overview of Quintet: Data Analysis Model

Figure 1 is about here.

A simplified data analysis model we have projected in Quintet is depicted in Figure 1. In this Figure, procedures that are carried out in Quintet are depicted in colored boxes. We have not implemented any image analysis functionality in Quintet and the data analysis starts from a set of text slide data files. According to our experience, the absence of image analysis module does not cause muchtrouble since every scanning software has a mechanism to export slide data into text formatfiles and detailed examination of data variables, not the visual inspection of microarray images, provide thorough understanding of microarray data. Quintet retains all the variables that scanning softwares provide for each gene since previously unused variables may turn out to be important for particular purposes, especially in QA steps. For example, a popular microarray image analysis software from Axon Instruments [5], GenePix, provides 43 variables for each gene, which enables a detailed understanding of spot intensities and their characteristics.

For each slide data, we first mark genes that are doubted to be erroneous from various criteria. Then, we check the quality of each slide data using various plots and statistical summaries. The error spot flagging and QA procedures should be iterated until no further quality improvements are evidenced. We consider the inter-operation of data processing and quality improvement check is very important to avoid data "over-processing" since any data processing can introduce unwanted artifacts which cannot be amended in downstream analysis steps. We apply normalization procedures to remaining genes according to the algorithm developed by Yang *et al.* [6]. This is an effort to remedy systematic artifacts that may have been introduced by signal extraction procedures. Normalized data are the basis for downstream analysis steps like DEG identification and clustering/classification.

Downstream analysis steps are quite straightforward.First, DEGs are identified. Since DEGs constitute basic elements for subsequent analysis steps, reliable identification of DEGs is of utmost importance. Furthermore, since different algorithms produce different DEG sets, multiple algorithms are supplied in Quintet and users can select their own DEG set among them based on the statistical characteristics revealed by auxiliary plots provided in Quintet. Using the identified DEGs, a gene expression matrix is constructed and clustering/classification is carried out. As such, the DEG identification procedure is a dimension-reduction step in this sense. In clustering/classification, we also supplied multiple algorithms and users can experiment different algorithms to survey possible variations in clustering/classification results.

Data organization

Figure 2 is about here.

Schematic data organization assumed in Quintet is appeared in Figure 2(a). In general, a microarray experiment is composed of multiple stages. For example, each time point can be considered as a stage in the case of time-series experiments [7] and each individual condition can be considered as a stage in the case of experiments composed of multiple conditions [8]. Furthermore, each stage is usually composed of multiple slides. Some slides in the stage can be replicates and the others can be dye-swaps.

Since the number of stages and data compositions in each stage can be arbitrary, we needed a simple but flexible method to import all necessary slides at one stroke. At the same time, the stage-slide relationship should be stored also. For this purpose, simple configuration file approach is used in Quintet (Figure 2(b)). This configuration file is a simple

text file with each line composed of 4 columns. In this file, stage information appears in the first column, experiment types (replicate/dye-swap) in the second, full paths of slide data file in the third and slide aliases to be used internally in Quintet in the last column. Replicates are designated by 'A' while dye-swaps are designated by 'B' in the second column. Using this information, all relevant slide data are imported into Quintet in a batch mode. The data organization is stored for later use also.

QA Module

QA of microarray data has not been considered as an important problem of micorarray data analysis by itself, which explains the lack of established standard procedure for QA. However, QA can be a decisive factor in establishing the reliability of analysis results performed using highly evolved algorithms because 'nothing can compensate for poor-quality data regardless of the sophistication of the analysis'[4]. Furthermore, QA itself is very important in constructing large centralized databases andcollecting gene expression data on a comprehensive scale since data sharing can be drastically restricted without quality assurance [9, 10, 11].

Figure 3 is about here.

In Quintet, QA module is one of the five core functional module. Although there is no established procedure for QA and methods implemented in Quintet are rather exploratory, QA methods implemented in Quintet were very successful in understanding the data quality according to our experience. QA module relies heavily on various statistical plots and summaries of particular variables like spot intensities, background intensities and log ratios. Some major plots used in QA module are depicted in Figure 3. In Figure 3(a), we show a scatter plot between background-corrected green intensity and background-corrected red intensity of a slide in log-log scale. Though we show a scatter plot between green and red intensities, any combination of variables can be used, which can be very useful in exploratory investigation of data characteristics. In Figure 3(b), we show an AM (average intensity vs log ratio) plot. Log ratio of a spot is given by $M=log2R/G$ and average intensity by $A=(log2RG)/2$. What is well-known is that this plot is the 45 degree clock-wise rotation of Figure 3(a) and it is much easier to apprehend the data characteristics because the diagonal line in Figure 3(a) is now a horizontal line at $y=0$. Since only a small number of genes is assumed to be differentially regulated in any microarray experiment, we can check if the main axis of data distribution is distorted through this plot and determine if data normalization of this slide is necessary. In Figure 3(c), we show a 2D image plot of spot log ratio values for a slide. Through this plot, we can check whether the data distribution shows any spatial bias due to

improper treatments in data handling steps. In Figure 3(d), we show a block-by-block box plot of log ratio values for a slide. This plot shows block-by-block variability level of log ratio values within a slide and we can determine ifblock-wise centering and block-wise scaling of log ratios should be carried out to the slide. In Figure 3(e), we show a pairs plot of log ratio values and correlation coefficients among a group of slides. Through this plot, we can check if there is a clear distinction between the data distributions between replicates and those between independent slides. This result is very useful since it is directly related to the reproducibility and specificity of microarray data under analysis. In (b) to (e), any numerical variables other than log ratio can also be used instead.

In assessing the data quality of a slide, replicated genes can provide the clearest information since, though spotted at various positions within the same slide, they should show very similar behavior in every aspect. Position-dependent dissimilarity and variability between variables of the same replicated genes can be used as a quality measure. Because of this, we implemented two special menus to examine the characteristics of replicated genes. We classify replicated genes into two types: controls and simple duplicates. Genes that are repeatedly spotted for *special purposes* are called controls and genes that are replicated without such consideration are called simple duplicates. Examples of control genes are the positive or negative controls defined by Schena [12] and other controls used for defining reference differentialexpression levels [13]. Therefore, by comparing the observed differential expression levels of control genes with their expected differential expression levels and by measuring the variability of observed differential expression levels over control genes representing the same reference level, we can estimate the quality of accuracy in differential expression levels recorded for a slide. Contrary to this, simple duplicates cannot be used to estimate the data quality by measuring discrepancy between the observed and expected differential expression levels, However, they can be used in assessing data quality by examining the correlation of variables between values obtained from different positions.

Figure 4 is about here.

In Figure 4, we show sample plots for controls and simple duplicates generated in Quintet. In Figure 4(a), we show a background-corrected green intensity vs background-corrected red intensity scatter plot for control genes along with the scatter plot for all genes. Different control genes are depicted in different colors so that they can be differentiated easily. Unfortunately, the control genes are not used to indicate specific reference differential expression levels in this case and they align along the diagonal line. In Figure 4(b), we show an AM plot for control genes along with the AM plot for all genes. Though the control genes

seem to align along the diagonal line in Figure 4(a), they show discernable deviation from the horizontal red line in Figure 4(b). The log ratio distributions of all control genes are shown in Figure 4(c) using a box plot. From this Figure, we can notice that some of control genes, especially those whose average intensity is small, show deviations from 0. Similarly, if control genes are used as specific reference differential levels, the discrepancy between observed differential levels and expected differential levels and the amount of fluctuations of observed differential levels of a control gene from its mean value can be used as a definite evidence of data quality. In Figure 4(d), we show a sample self-vs-self red intensity scatter plot of duplicated genes. Contrary to what is expected, this self-self scatter plot does not show clear correlation (correlation coefficient = 0.1), which means that the data quality seem to be doubtful. There are other plots to help users understand the distribution of differences between values of the same duplicated gene in Quintet.

Data Processing Module

Quintet's data processing module is composed of two parts: spot preprocessing and data normalization. In spot preprocessing part, Quintet filters out faulty spots that can undermine the reliability of analysis results. What actually takes place is that Quintet marks suspicious spots based on several criteria separately and keeps all the mark results so that users can select a suitable combination of error flags under their own discretion. In normalization part, Quintet carries out the local regression (LOWESS) fit normalization procedures developed by Yang *et al.* [6] to remaining spots.

Following list of flagging criteria are supplied in Quintet:

- **BG error:** spots whose local background intensities are larger than spot intensities in any of the dyes are marked as errors. Since each spot is composed of many small pixels, Quintet normally uses the median of pixel intensities in spot region as the representative spot intensity and the median of pixel intensities in local background region as the representative local background region.
- **Maximum intensity error:**spots whose spot intensities reach the scanner detection limit in any of the dyes are marked as errors because, for these spots, we are not able to say whether such spot intensities are exactly the scanner maximum limit or beyond it.
- **Control spots**
- **Original error:**spots that experimenters marked already are automatically flagged as errors. Normally spots that are marred by dusts, finger prints and scratches are marked.
- **SNR (signal-to-noise ratio) error:**spots whose ratio between background-corrected

intensity and local background intensity is smaller than a user-specified threshold are marked as errors. Although all spots marked as SNR error spots may not be considered as errors, credibility of analysis results can be gained by excluding less informative spots.

- **Outlier error [4]:**spots whose log ratio differences (sums) between two replicated (dye-swapped) slides of a stage deviate from the expected value, 0, are marked as errors. These error spots clearly represent the results of inconsistency during experimental procedures and should be removed from further analysis.
- **Median-vs-mean ratio error [14]:**spots whose ratio between the mean signal intensity and median signal intensity in any of the dyes is smaller than a user-specified threshold are marked as errors. This criterion is inspired by the result in Tran *et al.* [14], which claims that 'a simple ratio between the mean and median signal intensities may be the best way to eliminate inaccurate microarray signals'.
- **FG-vs-BG error [15]:**spots whose statistical test between the spot region pixel intensities and the local background region pixel intensities do not show clear distinction are marked as errors. In Quintet, t-test is used.

After removing error spots, remaining data should be normalized to minimize systematic variations in the measured gene expression levels. What we hope is that biological differences can be more easily distinguished, as well as the comparisonof expression levels across slides can be accomplished more easly as a result of normalization. The procedure that is implemented in Quintet is basically the one developed by Yang *et al.* [6]. This normalization procedure is composed of three parts:pin-block centering, pin-block scaling and file scaling. Pin-block centering is used to correct the distortions in main axis of the AM plot by applying the LOWESS fit to the AM values of genes located within each print tip group of a slide. Then, pin-block scaling is carried out to reduce variations of log ratio variances across pin-blocks of a slide by multiplying pin-block-specific scaling factors. Finally, file scaling is carried out to reduce variations of log ratio variances across files by multiplying file-specific scale factors. In pin-block scaling and file scaling, scaling factors are calculated using median absolute deviations (MAD) [6].

To these basic normalization procedures, we supplemented another step in Quintet: global normalization of average intensity *A*. This is motivated by the fact that the DEG identification in cDNA microarrrays should be based on comparison between values of *log2R* and *log2G*. However, if only log ratio *M* values are normalized, large variations in average intensities willmask the true difference between *log2R* and *log2G*, which would

result in high levels of false positives and false negatives. To avoid this problem, we should normalize average intensities as well as log ratios. The procedure proceeds like this: first, the *A* value scales of all slides are normalized through file-specific scale factors calculated using MAD. Then, resulting *A* values are adjusted so that the median *A* of all genes in each slide becomes the mean of median *A* values of all slides. Results before and after the global *A*normalization is shown in Figure 5 (a) and (b). In DEG identification, *log2R* and *log2G* values are restored from normalized *A* and *M* values.

Figure 5 is here.

The global normalization of average intensity is performed to all slides under analysis in Quintet. However, other three basic normalization procedures are selectively applied to individual slides based on a user-specified configuration. Based on QA results, users should classify slides into three groups: pin-block centering group, pin-block scaling group and removal group. As the name suggests, pin-block centering wil be performed to the first group, pin-block scaling will be performed to the second group and slides in the third group will be removed from further analysis due toquality problems. This classification is possible since QA results give clear view of the type of normalization procedures that should be applied to individual slides. File scaling is performed afterwards if necessary. Data transformations after each normalization step can be examined using a set of statistical plots in Quintet. These diagnostic plots include scatter plot, AM plot, inter-slide box plot, histogram and pairs plot.

Figure 6 is here.

In Figure 6, results of normalization procedures are summarized. Figure 6(a) shows AM plots before normalization, after normalization without global *A* normalizationand after full normalization. Figure 6(b) shows corresponding RG scatter plots, respectively. Yellow lines in Figure 6(a) depict LOWESS fit lines between average intensity and log ratio. These plots clearly show that systematic trends in AM data are largely corrected. Furthermore, plots in Figure 6(b) show the global A normalization works quite well, comparing plots before and after normalization. Especially, the green and red intensities show similar distribution range without much change in data distribution characteristics.

The rationale of Quintet's data processing module is that though the data processing procedures are expected to remove non-biological artifacts and to remedy data distortions that occurred during data preparation and signal acquisition steps, it is also highly probable

that they introduce unwanted new artifacts into the data. Therefore users should be very cautious to avoid "over-processing"the data and every data processing result should be checked for its legitimacy using suitable examination procedures.

DEG Identification Module

DEGs are genes whose expression levels show clear difference between reference and experiment samples. Since observed differential expression is normally interpreted as a result of biological response to the experimental condition under study, the DEG identification is one of the most crucial tasks of microarray data analysis. Because of this, many DEG identification algorithms have been developed, and we are trying to supply as many available algorithms as possible in Quintet.

One perplexing factor while implementing DEG identification algorithms in Quintet is the number of replicates in each comparison unit since some algorithms can be applied only to single slides (single-slide algorithms) while others intrinsically need multiple slides (multiple-slide algorithms). Therefore multiply-replicated comparison units cause another complication for single-slide algorithms. Furthermore, since we cannot measure the level of confidence without replicates [16], single-slide algorithms are of limited value compared with multiple-slide algorithms. Nevertheless, we have included some single-slide algorithms in Quintet since, in general, they are easy to implement and their results are easy to interpret and gain in reliability can be achieved by imposing more stringent cutoff values. In addition, the absence of statistical significance should not prevent single-slidealgorithms from being utilized because comparison units in most published microarray data so far are single slides whose results have been quite successful in elucidating various previously unknown global genetic pictures.

Currently, the following algorithms are implemented in Quintet:

- **Fold change type:** genes whose differential expression levels are beyond a cutoff value are declared as differentially expressed in fold change type algorithms. Multiple algorithms can be categorized into this type.
  - **Generic algorithm:**in generic fold change algorithm, genes whose log ratios are beyond a user-specified cutoff value are selected as DEGs. This is the oldest algorithm for selecting DEGs and still widely used by many researchers though criticisms have been filed by many researchers [17].
  - **Z-test type:**z-scores in statistics measure the number of standard deviations a data point is away from the mean. In z-test type algorithms, genes whose z-transformed log ratios are beyond a cutoff are selected as DEGs [4]. Two

types of z-tests are implemented in Quintet: global z-test and local z-test. In global z-test algorithm, standard deviation is calculated using log ratios of whole-slide. Therefore the global z-test is essentially the same to the generic fold change algorithm. The difference between the two algorithms lies on the cutoff. In the case of generic fold change algorithm, the cutoff is given in the unit of fold change while the cutoff is given in the unit of standard deviations in the case of global z-test. To the contrary, the localz-test reflects the intensity-dependent change of variability by applying z-test to groups of genes clustered according to their intensity levels [18].

- **Sapir-Churchill algorithm:**Sapir and Churchill [19] applied EM algorithm to residualsfrom orthogonal regression and separated them into common and differentially expressed components. Though internal details are quite different from the generic fold change approach, resulting cutoff is given by two horizontal lines symmetric to *y=0* in theAM plot, similar to the generic fold change approach.

- **Newton's algorithm:**Newton et al. [20] considered the problem of inferring fold changes in gene expression from cDNA microarray data within a Bayesian hierarchical model and significant expression changes are identified by deriving the posterior odds of change. Though original algorithm considers only single slides, recent generalization in an R package, YASMA, dissolves this restriction [21]. This generalized version is implemented in Quintet.
- **T-test [22]:**t-test is a representative parametric hypothesis test method assessing whether two groups of data are statistically different from each other or not. In Quintet, the comparison groups can be two-color fluorescence signals as well as log ratios in two different experimental samples.
- **Wilcoxon rank sum test (RST) [23]:** Wilcoxon RST is a non-parametric analogue of t-test which permits robust hypothesis testing between two groups. According to Troyanskaya *et al.* [23], this test algorithm appears to be very conservative and can be advantageous when subsequent biological validation procedures are concerned.
- **Significance analysis of microarrays (SAM) [24]:**SAM identifies DEGs using a statistic similar to t-score and statistical significance estimation using permutations of repeated measurements. Although the algorithm is similar to the t-test, it also gives the level of false discoveries called false discovery rate (FDR [25]) by identifying nonsense genes through the permutations, which makes it popular in recent years.

Since these algorithmsare based on statistical arguments, false positives and false negatives cannot be avoided. Furthermore, according to our experience, different algorithms

produce different DEG results, which makes the identification of optimal DEG set very difficult. Therefore one shouldbe very careful in interpreting the result of any single algorithm and we strongly recommend users to try to use as many different algorithms as possible and compare them very cautiously to get a robust result. For this reason, we are trying to implement as many availableDEG identification algorithms as possible in Quintet. In addition, we are developing a method to integrate the results of individual algorithms, hoping to get more robust set of DEGs thinking that only robust DEGs not false positives and false negatives of each algorithm will be will be selected by many different algorithms.

Figure 7 is here.

In Quintet, we also supplement auxiliary plots to help users get intuitive understanding of statistical characteristics of the DEG set under consideration. In Figure 7, we show some of the plots. In Figure 7(a), we show the average AM plot where x-axis (y-axis) represents the average of A (M) values. In this plot, DEGs are represented in red while all other genes are represented in green. Therefore one can gain a rough understanding of statistical characteristics of DEGs from this plot. If more detailed information of DEGs is desired, one can turn to the box plot shown in Figure 7(b). In this plot, the distribution of log ratios for each DEG is represented in a box. Combining theplots shown in Figure 7(a) and (b), one can understand DEG characteristics more thoroughly, which will be of help in determining optimal DEG sets.

Clustering Module

Clustering is one of the most widely used methods in gene expression analysis [4, 7]. The rationale is that when genes are grouped into clusters according to their levels of similarity in expression profiles, the co-expression of genes within each cluster can be interpreted as a result of co-regulation, which provides greater insight into their biological relationship. For instance, if two or more genes have similar expression patterns in different experimental conditions or at different time points, these genes may be co-regulated and even be functionally related. Furthermore, as transcription is regulated mainly by the binding of transcription factors (TFs) to the promoter region, clustering of gene expression profiles can be very useful in identifying *cis*-regulatoryelements in the promoters, providing more insight to gene function and regulation networks [26]. Recent breakthrough in this line of approach has made it possible to infer condition-specific regulatory modules in a simple eukaryote by combining clustering results and *cis*-regulatory element patterns in promoter regions under a probabilistic graphical model

[27].

When experimental samples are clustered using gene expression profiles, it is an unsupervised learning (also known as class discovery in pattern recognition) problem where *a priori*unknown number of classes among samples should be identified using gene expression profiles. This problem is of an utmost practical importance since it is directly related to disease classification using gene expression profiles [28, 29]. Most current disease classifications are primarily based on phenotypic characteristics. As such, current disease classes cannot explain markedly different clinical courses and treatment responses observed among patients with the same disease. Since gene expression profiles represent a molecular portrait of biological mechanism, disease sample clustering based on gene expression profile can be advantageous. In particular, clustering disease samples can elucidate previously uncharacterized disease subtypes, which can be beneficial in diagnosing disease types or disease progress stages.

Since the seminal work of Eisen *et al.*[7], clustering has been extensively used in microarray data analysis and culminated a lot of successful results. The spectrum of clustering algorithms that has been used in microarray data analyses is very wide. This entails novel algorithms such as self-organizing maps (SOM) [30], clustering affinity search technique (CAST) [31], minimum spanning tree (MST) [32] as well as conventional algorithms such as hierarchical clustering[7] and k-means clustering [33], to mention a few. Because of this, many non-commercial and commercial systems regard clustering module as their core functionality [34, 35] and Quintet provides following clustering algorithms currently:

- **K-means clustering:** starting from *K*randomly chosen organizing centers (centroids), this algorithm iterates between two steps. First it tries to partition elements so that the summation of distances from each elementto its nearest centroid becomes minimum. Then *K*centroids are recalculated using present cluster partition. In each cluster, centroid is given by the mean of all elements. This is a greedy algorithm and every expression profile is assigned to one of the clusters.
- **Partitioning around medoids (PAM) clustering [36]:**PAM clustering is very similar to the K-means clustering. In the case of PAM clustering, a medoid is given by the median of all elements contained in a cluster. Therefore, the clusters are quite robust and exceptional elements within a cluster do not contribute much in calculating the medoid.
- **Hierarchical clustering:**contrary to K-means and PAM clustering, this is an agglomerative clustering algorithm, by iteratively merging two most similar clusters at

each step until all elements form one large cluster. Initially each element is assigned to its own cluster. Since there are many different ways to merge two most similar clusters, this should be specified in advance. In Quintet, only the most common merging methods are implemented: single linkage, complete linkage, and average linkage [37].

- **Self-organizing map (SOM) clustering [30, 38]:**SOM is an unsupervised neural network algorithm trying to find prototype vectors that represent the input data set and continuous mapping from input to a lattice at the same time. The lattice structure is self-organized according to the weight vectors assigned to each lattice point, starting from random positions. As a result of self-organization, similar vectors come close to each other in the lattice while dissimilar ones move away from each other.
- **Clustering affinity search technique (CAST) [31]:** CAST is a kind of adaptive agglomerative clustering algorithm. Among unassigned elements, elements whose average similarity(affinity) from the current cluster core does not damage the cluster coherence will be added to it. However elements whose affinity from the cluster core is below a tolerance level will be removed from the cluster among elements assigned to the current cluster. The addition and removal steps will be iterated until a stable cluster results are obtained.

Despite its popularity, there remain lots of statistical issues on the clustering of gene expression data [39] and it is very difficult to choose which cluster results to use in subsequent analyses without supplementary information since different clustering algorithms produce differentclusters. Currently, we are working on this problem in two directions. First, we are trying to develop an algorithm to compare results from different clusters and produce a robust one. Second, we are trying to implement a module supplying cluster validation measures and determine optimal cluster result based on them.

Clustering results are presented in several different forms in Quintet. First, individual clusters are reported in separate text format external files with corresponding differential expression levels so that users can scrutinize individual genes contained within each cluster in detail and use clustering results in other programs. Second, clusters are depicted in several plots. Typical plots normally adopted in cluster representation are shown in Figure 8. The dendrogram attached 2D image plot shown in Figure 8(a) is the most well-known representation format of hierarchical clustering. For non-agglomerative clustering algorithms like K-means clustering, only the 2D image plot is shown in Quintet. However, according to our experience, line plots shown in Figure 8(b) are more

informative since detailed comparison of expression profiles is possible. The red line in each line plot designates the mean expression profile of corresponding cluster. The projection plot shown in Figure 8(c) is another convenient plot that can be used to check if clusters are well-separated in low-dimensional plots using a few principal components. As such, this plot can be used as a kind of cluster validation measure.

Figure 8 is here.

One often-neglected chore in clustering is the construction of gene expression matrix (GEM) based on the results from upstream analysis such as DEG identification. Though this seems simple, one needs to decide two things. First, one needs to select the genes that should be included in the GEM. If all the genes are used, genes that remain in their basal expression level increase noise in clustering, which results in unreliable clusters. However, if only DEGs are used, since DEG sets from different stages are different, one needs to decide which DEGs to be included in GEM. In Quintet, if only DEGs are used, genes that show differential expression in any of the stages are included in GEM construction by default. Second, one has to decide whether missing values should be filled up using some surrogate values or not while constructing GEM [40]. Although missing value imputation may not be justifiable from biologicalpoint of view, one has to discard substantial amount of genes without it just because only a few stage values are missing. We are implementing the weighted k-nearest neighbor imputation algorithm (KNNimpute)now but recommend that this procedure should be used very cautiously since unwanted artifacts can be introduced and affect the result. This algorithm is selected because many researchers report that KNN works better than other competing algorithms [35, 40]

Classification Module

Classification is a process assigning objects to known classes based on the measurements made on it. In microarray data analysis, objects that should be classified are experimental samples and measurements are gene expression profiles. For example, classifying tissue samples according to their gene expression profiles has produced promising results for cancer diagnostics [28, 29]. Since accurate diagnosis can affect the treatment course and the probability of survival, the tissue sample classification based on the gene expression profile has a tremendous practical importance. Also, expression profile based disease classification can be used as a generic framework for disease diagnosis, contrary to simple morphology based disease classification. Furthermore, since classification based on gene expression profiles will provide a genomic view of

molecular mechanism involved in the phenotypic disease progress, distinctive expression profiles can be used as a molecular portrait of a disease and genes that show clear difference between two molecular phenotypes can be used as disease markers.

Typical sample classification is carried out in multiple steps. First, genes that show distinctively different expression levels between samples in one class and those in the other are selected since the majority of genes are assumed to exhibit basal expression levels across samples. This gene selection process is exactly what we do in the DEG identification step and the DEG identification module is used for this job in Quintet. These selected genes are used in the gene expression matrix assumbly. Then, classification function (classifier) is constructed using samples whose class relationship is known. The samples whose class relationship is known comprisethe learning set (LS) while samples whose class relationship is unknown comprise the test set (TS). Choosing a classifier, the error between predicted and true classes in LS is exhaustively refined to obtain the most optimal parameters. Finally, classes of samples in TS are predicted using the classifier.

There are many good candidate classifiers already [41] and the following list of algorithms is implemented in Quintet:

- **Fisher linear discriminant analysis (FLDA):** FLDA is a method to find a linear transform of measured multiple variables such that linear transformations of gene expressions drawn from two classes are separated as widely as possible. Because of its simplicity, FLDA is the most popular approach in classification.
- **Maximum likelihood discriminant analysis (MLDA):**in MLDA, an object is assigned to a class to which the class membership conditional probability (likelihood) of that object is maximum. Generally, the conditional probability density functions are given by multivariate normal functionsand the MLDA can be subdivided into three special cases: class probability functions with the same covariance matrix, class probability functions with diagonal covariance matrix, and class probability functions with the same diagonal covariance matrix. The first case is the FLDA given above, and latter two cases are referred as diagonal quadratic discriminant analysis (DQDA) and diagonal linear discriminant analysis (DLDA), respectively.
- **K-nearest neighbor (KNN) method:**in KNN, an object is assigned to a class which the majority of *K*nearest objects are belonged to. The distance between two objects can be Euclidean distance or one minus Pearson correlation. The number *K*is selected by optimizing error rates in learning phase.
- **Classification and regression tree (CART) [42]:** CART is a binary tree classifier

which builds classification and regression trees depending on the type of measurement variables. If the measurement variable is categorical then classification trees are created and if the measurement variable is continuous then regression trees are created. At each non-terminal node, binary segregation of measurement variables takes place and each terminal node contains the label of a class to which an object is assigned. Though CART is rather algorithmically involved, it is widely used in classifying objects since the results are quite intuitive.

- **Artificial neural network (ANN) [43]:**ANN is a collection of interconnected model neurons that emulate some of the observed properties of biological neurons which work as basic information processing units in mammalian brain. The network structure is designed to imitate the learning process in biological systems where the synaptic connection weights between neurons are altered to represent knowledge contained in thelearned examples. In the learning phase, synaptic weights are modified to reduce the errors between predicted and true class memberships. In this way, the connection weights are used as the knowledge base necessary to classify un-learned samples.
- **Support vector machine (SVM):** SVM is one of the most recent developments in pattern recognition field as a general purpose tool for feature classification [44]. It tries to separate a given set of two-class training data with a hyper-plane and, if such linear separation is impossible, it tries to build a hyper-plane classifier in a high-dimensional 'feature space' to which each measurement data is projected using a non-linear mapping function (kernel). SVMs have been shown to perform well in many areas of biological analysis [45, 46] and have also been quite successful in the analysis of microarray data [47, 48].

Though Quintet can handle only two-class classification problems now, efforts to incorporate multiple-class classification problems are underway. In the course of refining classifiers in learning phase, one needs to minimize errors between known class assignments and predicted classes. Since only the class assignments of learning set is usually known, one randomly splits the learning set into two classes, pseudo learning set and pseudo test set, constructs classifiers using the pseudo learning set only and estimates the error rate using the pseudo test set. In Quintet, the random separation of learning set can be selected between two different ways: cross-validation and test-train set type. In the cross-validation type learning, the learning set is divided into K subsets (K-fold cross validation) of equal size from the start. The classifier is then learned at K times, each time using one subset in turn as a test set. To the contrary, in the test-train set type learning the learning set is divided at every turn into two different subsets (pseudo

learning set and pseudo test set) and the classifier is learned using the pseudo learning set while the error rate is estimated on the pseudo test set. This whole process is repeated a number of times to calculate the error rate distribution.

Besides the core classification functional module, Quintet supplies auxiliary plots to be used in assessing the performance of specific classifiers. In Figure 9, we show some of the plots. In Figure 9(a) and (b), we show the error rate profiles calculated through cross-validation and train-test set type learning as the parameter *K* is varied in KNN classification, respectively. Conferring to theerror rate profiles shown in Figure 9(a) and (b), one can select the optimum value of *K*. Figures 9(c) and (d) depict the 2D image and projection view of gene expression matrix used in classification at the optimum value *K*, respectively. It is clear that the test set samples should be members of the class 1. The error rate profiles across different classification algorithms at their respective optimum parameters are shown in Figure 9(e). This can be used as a measure of classifier performance.

Conclusion

Quintet is an unified cDNA microarray data analysis system capable of carrying out five indispensablecategories of microarray data analysis seamlessly: data processing steps such as faulty spot filtering and normalization, data quality assessment, identification of differentially expressed genes, clustering of gene expression profiles and classification of samples. Though many existing tools of microarray data analysis emphasize their capacity to carry out three core categories of data analysis (DEG identification, clustering and classification), Quintet is geared to perform data preprocessing and QA also. In particular, QA is crucial for enhancing the reliability of analysis results and sharing gene expression data using centralized data bases since nothing can compensate for poor-quality data no matter how sophisticated the analysis is. We insist that data processings and QA should be incorporated into the regular-base data analysis practices. To help users intuitively understand data characteristics, we provide lots of plots and statistical summaries. In addition, since Quintet is written in R, it is highly flexible so that users can experiment new algorithms in Quintet with minimal efforts. Also, the GUI will make it easy to learn and use Quintet and since R-language and its GUI engine, Tcl/Tk, are available in all operating systems, Quintet is OS-independent.

Acknowledgements

We acknowledge the following R-language software packages: BioConductor project, cluster, e1071, GeneSom, lattice, MASS, mva, nnet, rpart, sma, YASMA. This work was

supported by a grant (PF003301-00) from Plant Diversity Research Center of 21st Century Frontier Research Program funded by Ministry of Science and Technology of Korean Government.

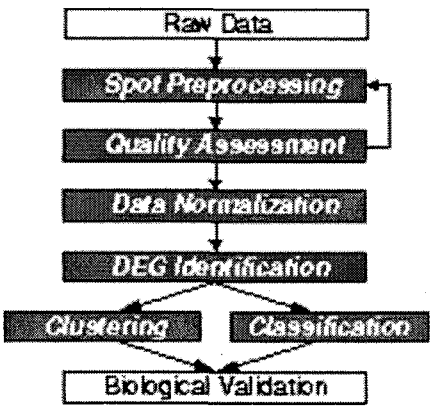
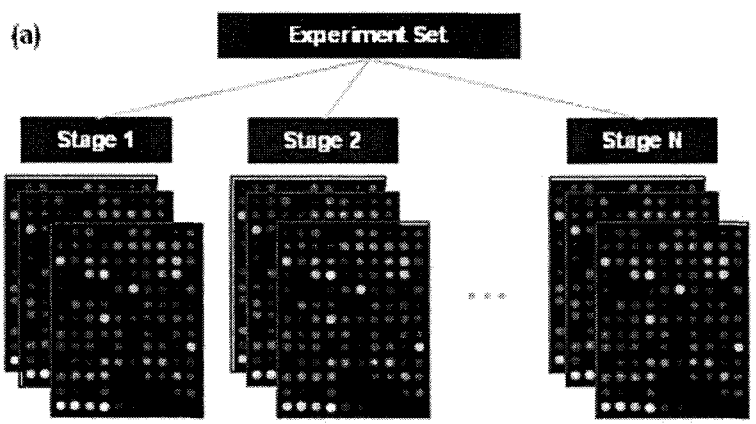


Figure 1. Simplified data analysis model.



(b)

1	A	C:\Data\Slide1.txt	Slide1
1	A	C:\Data\Slide2.txt	Slide2
1	B	C:\Data\Slide3.txt	Slide3
1	B	C:\Data\Slide4.txt	Slide4
2	A	C:\Data\Slide5.txt	Slide5
2	A	C:\Data\Slide6.txt	Slide6

Figure 2. (a) Schematic experiment data model assumed in Quintet and (b) part of a sample configuration file composed of 6 slides.

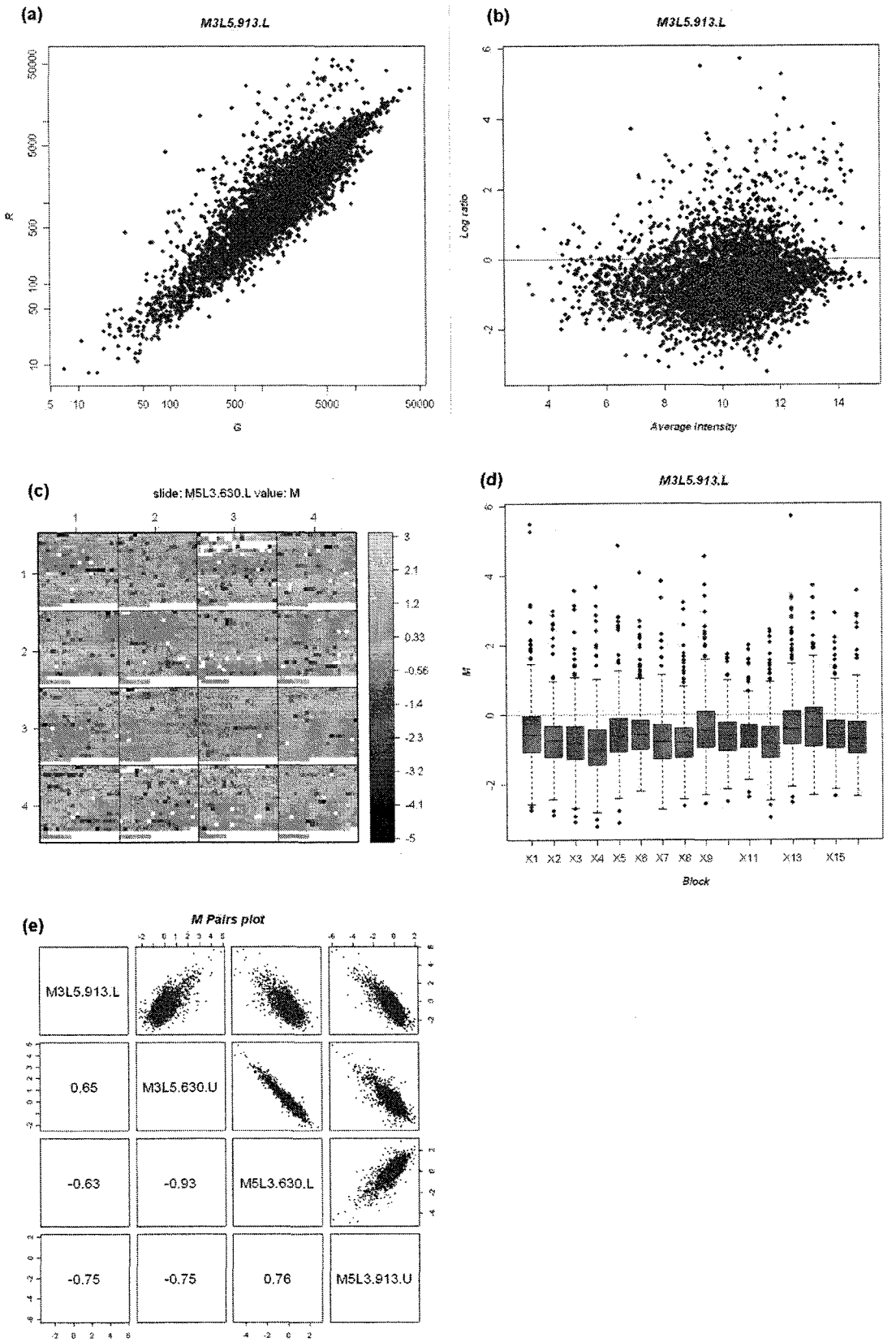


Figure 3. Sample plots used in QA module. (a) scatter plot of background-corrected green and red

intensities, (b) RI (log ratio vs average intensity) plot, (c) 2D image plot of spot log ratio values, (d) block-by-block box plot of log ratio values of a slide, (e) pairs plot of log ratios with correlation coefficients among a group of slides.

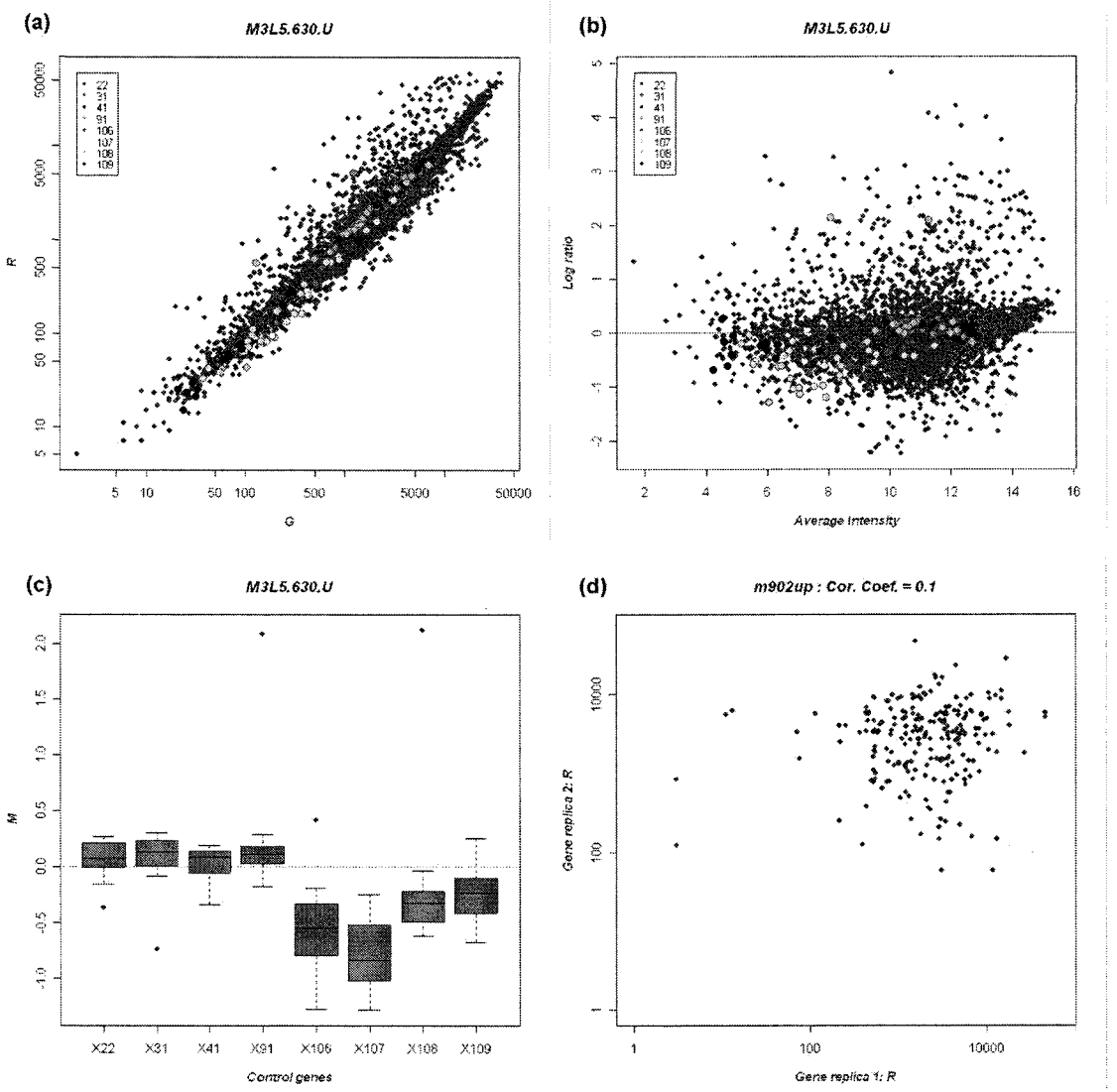


Figure 4. Sample plots for QA module using replicated genes. (a) red vs green intensity scatter plot for control genes (b) RI plot for control genes (c) box plot of log ratio values for control genes (d) self-against-self red intensity scatter plot of duplicated genes

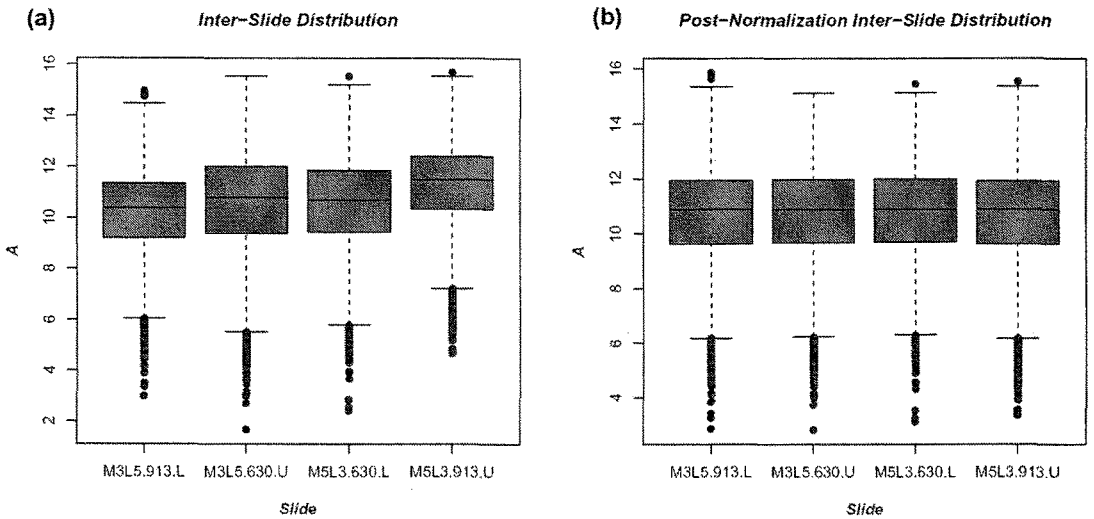
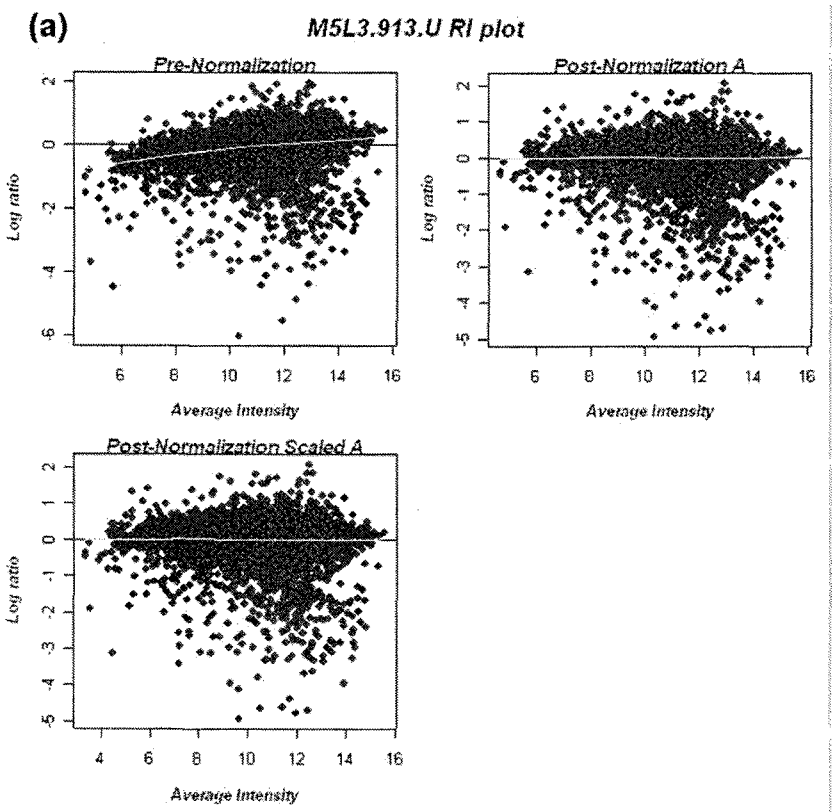


Figure 5. Box plot of average intensity *A*values across a sample experiment set before (a) and after (b) global A normalization.



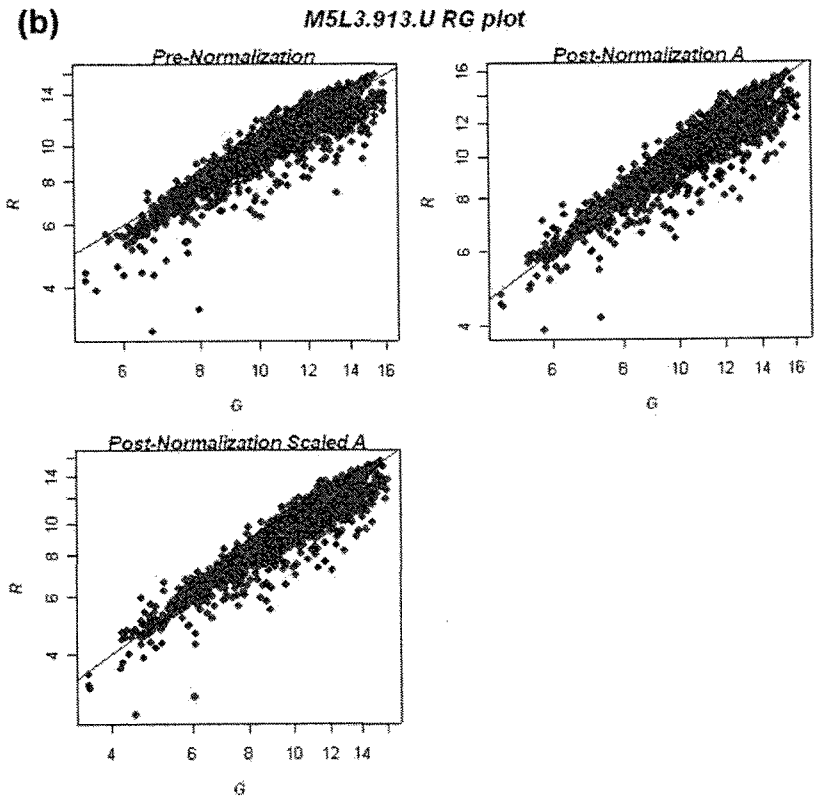


Figure 6. Normalization summary plots of a sample slide. (a) shows RI plots before any normalization, after log ratio normalization in the absence of global *A*normalization, and after log ratio and global *A* normalization while (b) shows corresponding red-vs-green scatter plots.

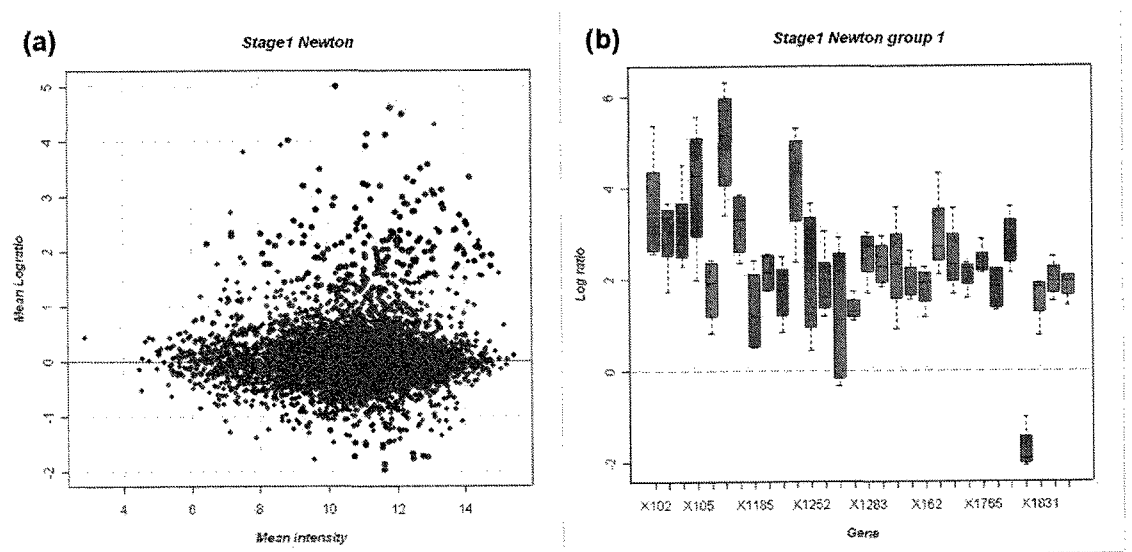


Figure 7. Supplementary plots for assessing DEG characteristics. (a) RI plot and (b) box plot of DEGs. In (a), DEGs are depicted in red.

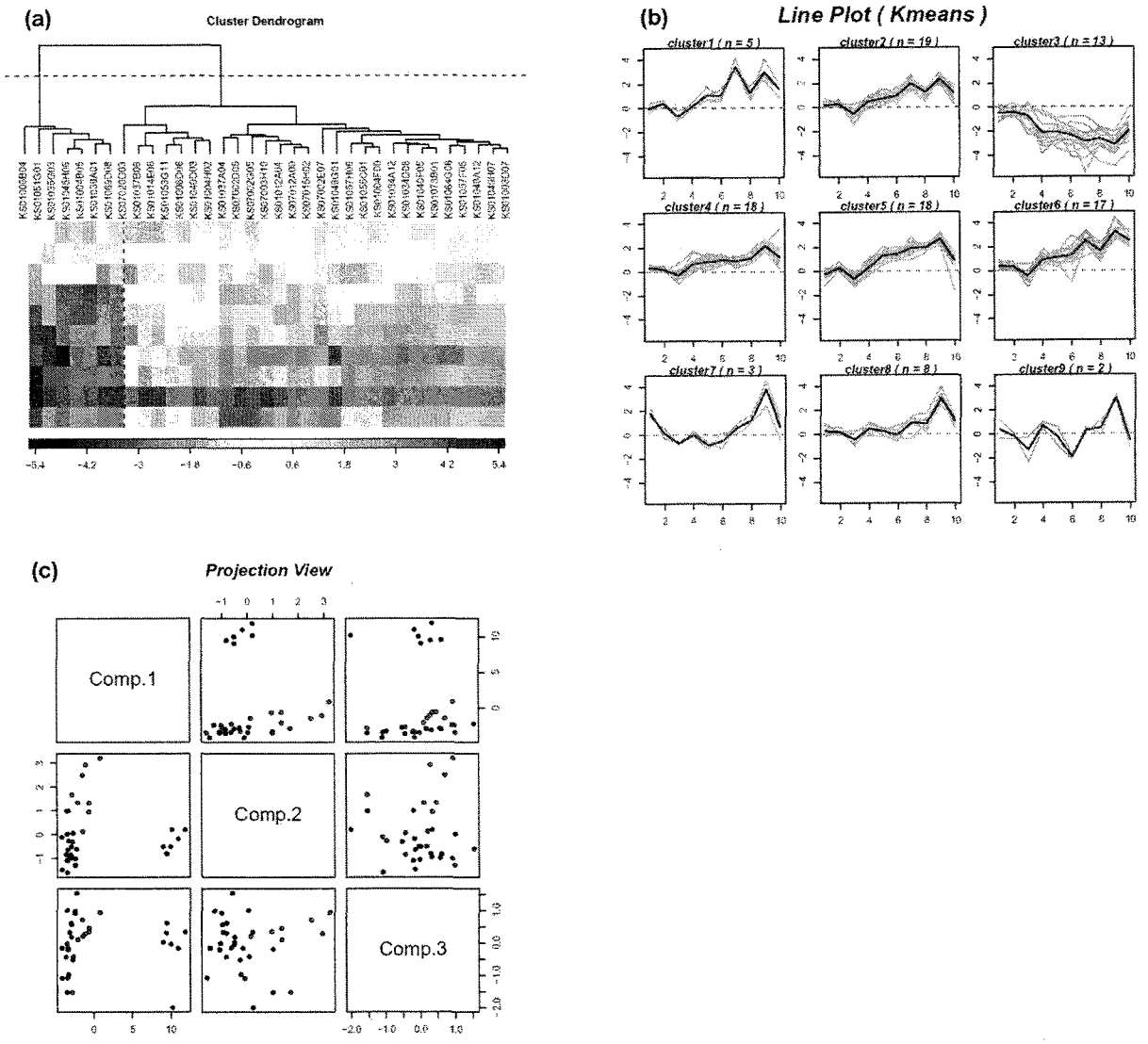


Figure 8. Typical cluster result presentation plots. (a) Dendrogram with 2D image plot (b) Line plot showing gene expression patterns in each cluster (c) projection map using 3 principal components.

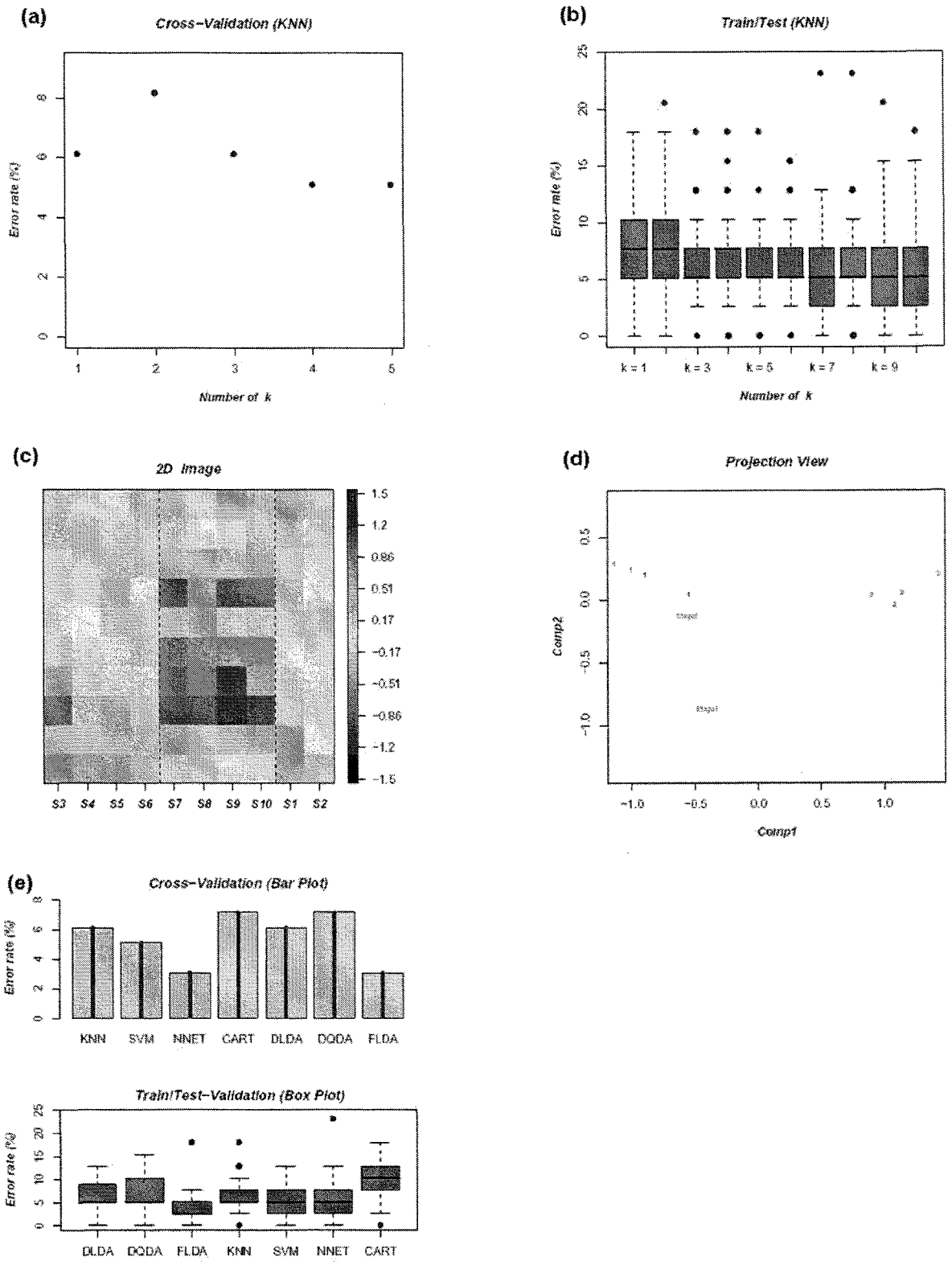


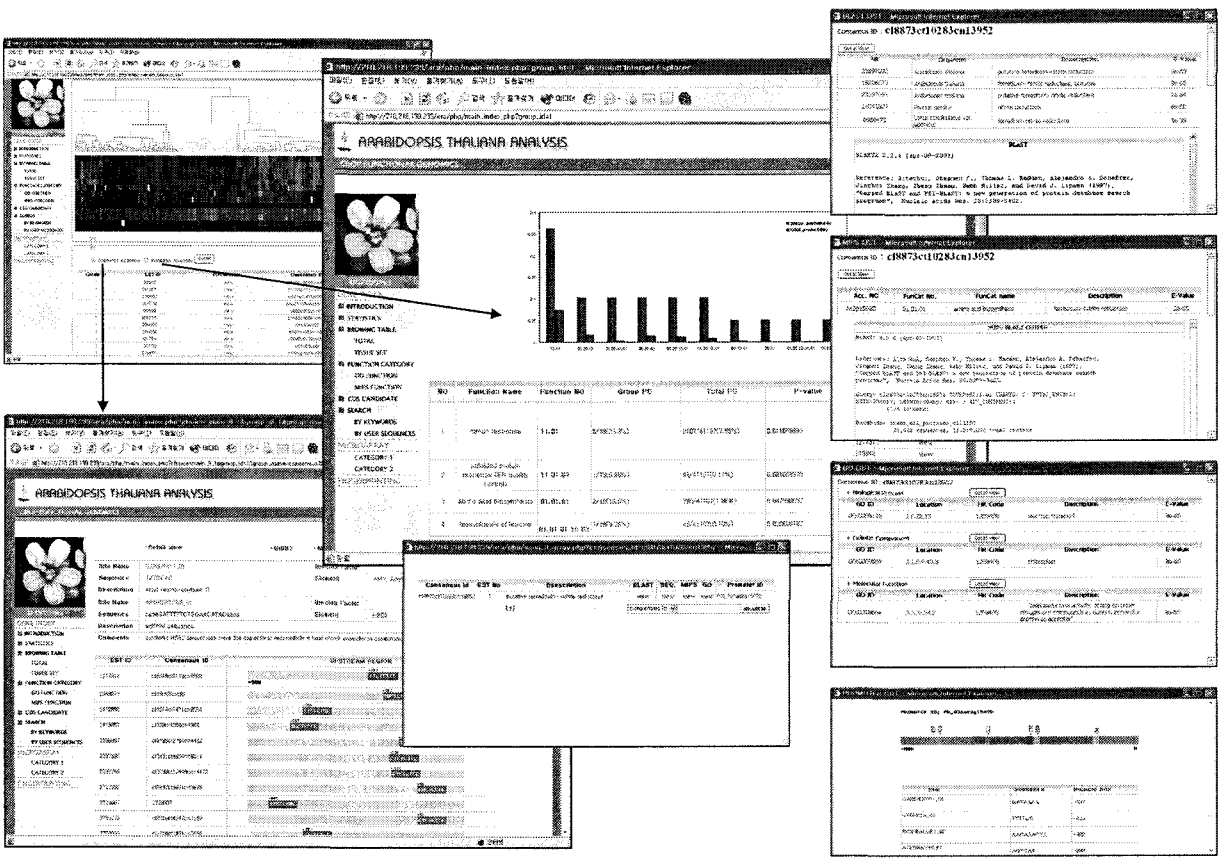
Figure 9. Auxiliary plots for classification module. (a) Cross-validation error rate profiles for each *K* in KNN (b) Train set / test set error rate profile for each *K*in KNN (c) 2D image plot of slides for two-class learning sets (left two groups: class 1 and class 2, respectively) and two test slides (right slides). In

this case, the two samples are classified as members of class 1. (c) Projection view of slides in learning sets and test sets using first two principal components.

라. Integrated genome analysis

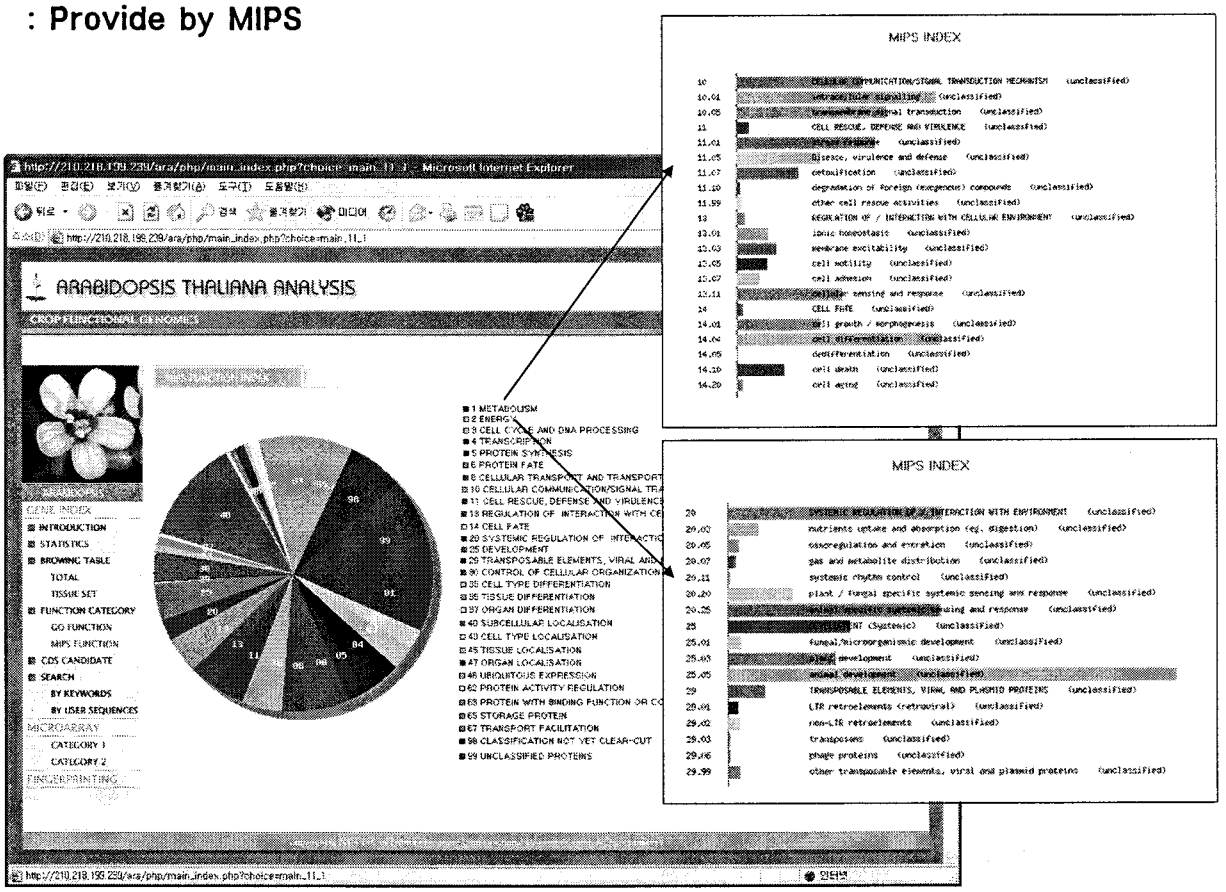
작물 관련 EST gene index 분석 결과와 gene expression 결과, Peptide mass spectrum 분석 결과 등 통합된 분석 결과를 보여주는 중요한 내용이다. 또 이를 바탕으로 One-Stop-Shopping 시스템을 만들어 나가는 과정이며 기능 유전체 연구에 기반이 된다.
벼, 콩, 애기장대풀 등을 EST, Microarray, peptide mass spectrum 분석 결과를 나타낸 것.
http://genepool.kribb.re.kr 홈페이지를 통하여 검색이 가능하다.

Microarray 분석에 따른 co-expression 정보와 co-regulation 정보 분석



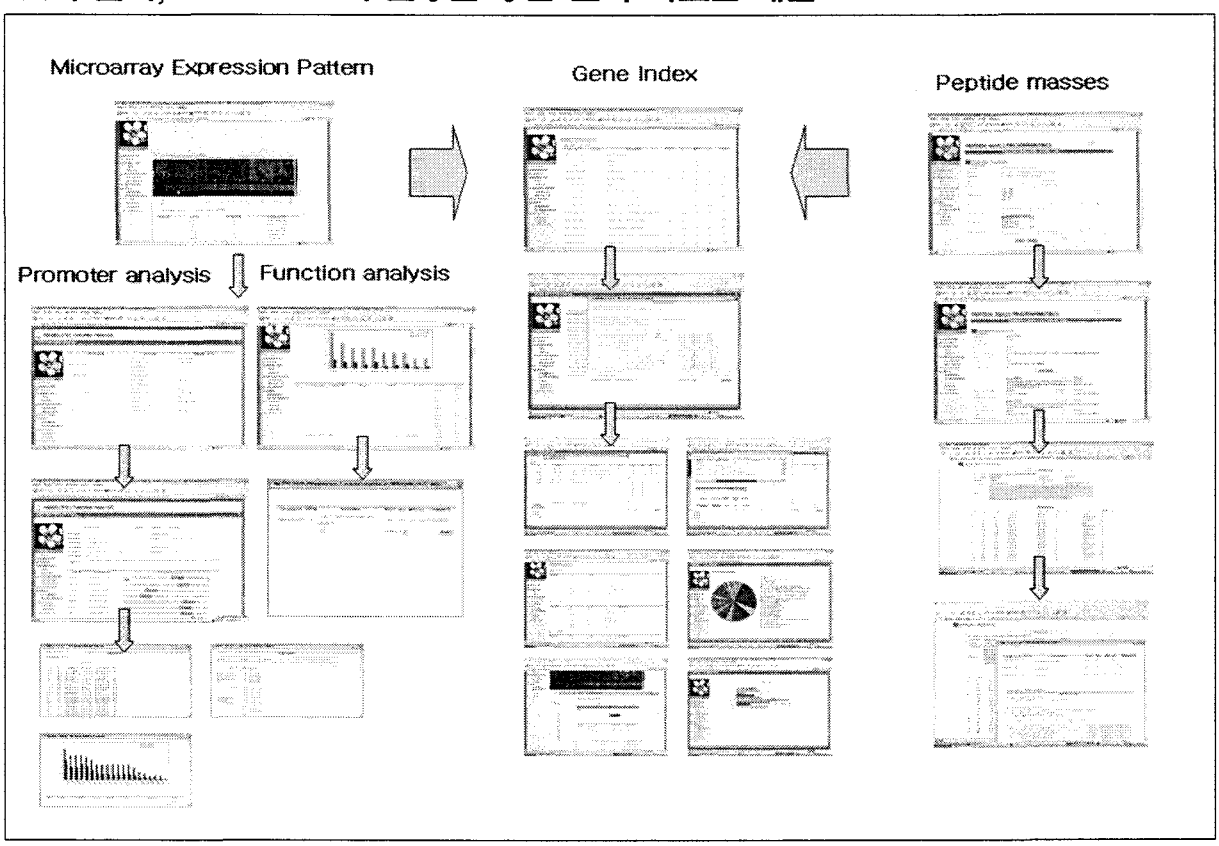
Arabidopsis, Yeast function catalog 제작 완료 및 Update 기능

: Provide by MIPS



EST 분석 결과 발생하는

Chromosome상의 regulation region 정보, EST 분석데이터,Microarray 분석 결과, MALDI-TOF와 연동한 통합 분석 시스템 개발



제 4 장 목표달성도 및 관련분야에의 기여도

구 분	연구개발 세부목표	추진 실적	달성도 (%)
2001	- 벼,콩등 작물유전체 EST 데이타베이스 구축, 분석 시스템 개발 및 서비스 - DNA Chip image분석에서 Clustering 기법까지의 데이터분석 work-flow개발	- Arabidopsis, rice, soybean등 에 대한 EST 분석 완료 - DNA Chip 분석에 필요한 Data Quality Assessment기능, Data Normalization기능, Diffrentially expression gene기능, Clustering 기능, Classification기능 등에 통합 버전 개발 (프로그램 등록 2건 완료)	150% 120%
2002	○애기장대,콩, 벼관련 Proteomics 연구를 위한 peptide mass spectrum 예측 시스템 개발	○ 구축 완료 (기술료 3000만원 계약 완료)	110%
	○EST 데이터베이스와 애기장대, 콩,벼등 기존의 공개 식물 EST를 Protein으로 변환하여 Peptide mass spectrum시스템과 연결한 DB구축	○ 구축 완료 (EST분석, DNA Chip분석, Proteomics분석 상호 연동성 있는 시스템 개발)	120%
	○DNA Chip을 이용한 Plant gene expression profiling 시스템 개발	○ 구축 완료	100%
2003	❑ 벼,콩, 애기 장대 gene function catalog를 개발하여 보다 빠른 EST 기능을 추정하는 S/W개발	❑ MIPS, Ontology 기능 분류 S/W를 이용하여 Catalog 완성 (Yeast, Arabidopsis용)	100%
	❑ EST데이타와 DNA Chip정보, gene function catalog와 통합 시스템 구현	❑ 구현 완료	100%
	❑ DNA Chip해석에 따른 regulation region 예측 시스템 구축 및 co-expression 검증	❑ Chip 데이터의 Clustering에 대한 co-regulation정보 분석을 위하여 Gibbs sampling, MEME,등을 이용한 DNA motif 분석 완료 - TRANSFAC를 이용한 TF binding motif 분석 완료	100%

제 5 장 연구개발결과의 활용계획

* 추가연구의 필요성, 타연구에의 응용, 기업화 추진방안을 기술

본 연구 과제의 결과는 작물유전체 2단계 사업중 가지과 유전체 국제콘소시움 과제 (책임자 : 최 도일박사)에서 적극적으로 활용할 방침이며, Integrated genome analysis 데이터베이스는 웹을 통하여 서비스를 할 예정이며, Microarray 분석용 S/W는 현재 약 250여장의 데이터를 분석하는 데 필수적인 도구로 활용하고 있고 논문 발표후 국내 대학들에게 적극적으로 배포 다음 기술 개발에 활용할 수 있게 조치를 취할 것이다. 또 Peptide mass spectrum분석용 S/W는 이미 기업체 기술이전을 마친 상태이며 추후 관련 기관들에게 시스템 판매와 더불어 상품화 가능하게 기술을 향상시킨 후 활용이 될 것이다.

제 6 장 연구개발과정에서 수집한 해외과학기술정보

해당사항 없슴.

제 7 장 참고문헌

* 보고서 작성시 인용된 모든 참고 문헌을 열거한다

Gene index reference

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25, 1 : 25-9.

Bailleul, B., Akerblom, I., and Strosberg, A. D. (1997) The leptin receptor promoter controls expression of a distinct second protein. *Nucleic Acids Research*, 25, 14 : 2752-2758.

Boguski, M. S., Schuler,G. D. (1995) ESTablishing a human transcript map. *Nat Genet*, 10, 4 : 369-71.

Burke, J., Davison, D., Hide, W. (1999) d2_cluster: a validated method for clustering EST and full-length cDNAsequences. *Genome Res*, 9, 11 : 1135-42.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R; Gene Ontology Consortium. (2004) The Gene Ontology (GO) database and

informatics resource. *Nucleic Acids Res*, 1, 32 Database issue : D258-61.

Kantety, R. V., La Rota, M., Matthews, D. E., Sorrells, M. E. (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. *Plant Mol Biol*, 48, 5-6 : 501-10.

Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., Hide, W. A. (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res*, 9, 11 : 1143-55.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., Upton, J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res*, 28, 1 :141-5.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., White, J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res*, 29, 1 : 159-64.

Schoof, H., Zaccaria, P., Gundlach, H., Lemcke, K., Rudd, S., Kolesov, G., Arnold, R., Mewes, H. W., and Mayer, K. F. (2002) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Research*, 30, 1 : 91-3.

Yee, D. P., and Conklin, D. (1998) Automated clustering and assembly of large EST collections. *Proc Int Conf Intell Syst Mol Biol*, 6 : 203-11.

Peptide mass spectrum References

[1] Fenyo, D., *Curr. Opin. Biotechnol.* 2000, *11*,391-395.

[2] Rogers, M., Graham, J., Tonge, R.P., *Proteomics* 2003, *3*, 879-886.

[3] Cutler, P., Heald, G., White, I.R., Ruan, J., *Proteomics* 2003, *3*, 392-401.

[4] Mann, M., Hojrup, P., Roepstorff, P., *Biol. Mass Spectrom.* 1993, *22*, 338-4

[5] Wilkins, M.R., Gasteiger, E., Bairoch, A., Sanchez, J.C. *et al.*, *Methods Mol. Biol.* 1999, *112*, 531-552.

[6] Wilkins, M.R., Gasteiger, E., Wheeler, C.H., Lindskog, I. *et al.*, *Electrophoresis* 1998, *19*, 3199-3206.

[7] Pappin, D.D.J., Hjrup, P., Bleasby, A.J., *Curr. Biol.* 1993, *3*, 327-332.

[8] Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S., *Electrophoresis* 1999, *20*, 3551-3567.

[9] Zhang, W., Chait, B.T., *Anal. Chem.* 2000, *72*, 2482-2489.

[10] Wool, A., Smilansky, Z., *Proteomics* 2002, *2*, 1365-1373.

[11] Clauser, K.R., Baker, P., Burlingame, A.L., *Anal. Chem.* 1999, *71*, 2871-2882.

[12] Matys, V., Fricke, E., Geffers, R., Gossling, E. *et al.*, *Nucleic Acids Res.* 2003, *31*, 374-378.

[13] Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E. *et al.*, *Nucleic Acids Res.* 2003, *31*, 3576-3579.

[14] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., *et al.*, *Nucleic Acids Res.* 2003, *31*, 315-318.

[15] Dwight, S.S., Harris, M.A., Dolinsk,i K., Ball, C.A., *et al.*, *Nucleic Acids Res.* 2002, *30*, 69-72.

[16] Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., *et al.*, *Genome Res.* 1998, *8*, 967-974.

[17] Zhu, J., Zhang, M.Q., *Bioinformatics* 1999, *15*, 607-611.

[18] Camon, E., Magrane, M., Barrell, D., Binns, D., *et al.*, *Genome Res.* 2003, *13*, 662-672.

Microarray References

1. Brownstein MJ and Khodursky AB: *Functional Genomics: Methods and Protocols.* Humana Press; 2003
2. Draghici S, Kuklin A, Hoff B, Shams S: **Experimental design, analysis of variance and slide quality assessment in gene expression arrays.** *Curr Opin Drug Discov Devel* 2001, **4:**332-337
3. Wildsmith SE, Archer GE, Winkley AJ, Lane PW, Bugelski PJ: **Maximization of signal derived from cDNA microarrays.** *BioTechniques* 2000, **30:**202-208
4. Quackenbush J: **Microarray data normalization and transformation.** *Nat Genet* 2002, **Supple 32:**496-501
5. **Axon Instruments** [http://www.axon.com]
6. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15
7. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868
8. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257
9. Becker KG: **The sharing of cDNA microarray data.** *Nat Rev Neurosci* 2001, **2:**438-440
10. **GEO** [http://www.ncbi.nlm.nih.gov/geo/]
11. **ArrayExpress** [http://www.ebi.ac.uk/arrayexpress/]
12. Schena M: *Microarray Analysis.*Wiley; 2003
13. Evertsz E, Starink P, Gupta R, Watson D: **Technology and application of gene expression microarrays.** In *Microarray Biochip Technology.* ed. by Schena M. Eaton Publishing; 2000
14. Tran PH, Peiffer DA, Shin Y, Meek LM, Brody JP, Cho KWY: **Microarray optimizations: increasing spot accuracy and automated identification of true**

microarray signals. *Nucleic Acids Res* 2002, **30:**e54

15. Delenstarr G, Cattell H, Connell S, Dorsel A, Kincaid RH, Nguyen K, Sampas N, Schidel S, Shannon KW, Tu A, Wolber PK: **Estimation of the confidence limits of oligonucleotide microarray-based measurements of differential expression.** in *Microarrays: Optical Technologies and Informatics.* ed. by Bittner M, et al. *Proceedings of SPIE* 2001, **4266:**120-131
16. Lee MLT, Kuo FC, Whitmore GA, Sklar J: **Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations.** *Proc Natl Acad Sci USA* 2000, **97:**9834-9839
17. Hess KR, Zhang Wei, Baggerly KA, Stivers DN, Coombes KR: **Microarrays: handling the deluge of data and extracting reliable information.** *Trends Biotech* 2001, **19:**463-468
18. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, Sharov V, Saeed AI, White J, Li J, Lee NH, Yeatman TJ, Quackenbush J: **Within the fold: assessing differential expression measures and reproducibility in microarray assays.** *Genome Biol* 2002, **3:**research0062.1-0062.12
19. Sapir M and Churchill GA: **Estimating the posterior probability of differential gene expression from microarray data.** *Poster* 2000 [http://www.jax.org/research/churchill/pubs/index.html]
20. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37-52
21. **YASMA** [http://people.cryst.bbk.ac.uk/wernisch/yasma.html]
22. Dudoit S, Yang YH, Speed TP, Callow MJ: **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.** *Stat Sinica* 2002, **12:**111-139
23. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB: **Nonparametric methods for identifying differentially expressed genes in microarray data.** *Bioinformatics* 2002, **18**:1454-1461
24. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98**:5116-5121
25. Benjamini Y and Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Stat Soc B* 1995, **57:**289-300
26. Zhang MQ: **Large-scale gene expression data analysis: a new challenge to computational biologists.** *Genome Res* 1999, **9:**681-688
27. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific**

regulators from gene expression data. *Nat Genet* 2003, **34:**166-176

28. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537
29. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403:**503-511
30. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912
31. Ben-Dor A, Shamir R, Yakhini Z: **Clustering gene expression patterns.** *J Comput Biol* 1999, **6:**281-297
32. Xu Y, Olman V, Xu D: **Clustering gene expression data using a graph-theoretic approach: an pplication of minimum spanning trees.** *Bioinformatics* 2002, **18:**536-545
33. Tavazoie S, Huges JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285
34. Yeung KY, Medvedovic M, Bumgarner RE: **Clustering gene-expression data with repeated measurements.** *Genome Biol* 2003, **4:**R34
35. **GeneSight** [http://www.biodiscovery.com]
36. Kaufman L and Rousseeuw PJ: *Finding Groups in Data: an Introduction to Cluster Analysis.* John Wiley & Sons; 1990
37. Johnson RA and Wichern DW: *Applied Multivariate Statistical Analysis.* Prentice Hall; 1992
38. Kohonen T: *Self-Organizing Maps. (Series in Information Sciences, Vol 30)* Springer; 1997
39. Goldstein DR, Ghosh D, Conlon E: **Statistical issues in the clustering of gene expression data.** *Stat Sinica* 2002, **12:**219-240
40. Troyanskaya O, Canter M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525
41. Dudoit S, Fridlyand J, Speed TP: **Comparison of discrimination methods for the classification of tumors using gene expression data.** *JASA* 2002, **97:**77-87
42. Breiman L, Friedman JH, Olshen R, Stone CJ: *Classification and regression trees.* Wadsworth International Group; 1984

43. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS: **Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.** *Nat Med* 2001, **7:**673-679
44. Vapnik V: *The Nature of Statistical Learning Theory.* Springer-Verlag; 1995
45. Jaakkola T, Diekhans M, Haussler D: **Using the Fisher kernel method to detect remote protein homologies.** In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology* Menlo Park CA:AAAI Press; 1999
46. Zien A, Ratsch G, Mika S, Scholkopf B, Lemmen C, Smola A, Lengauer T, Muller K: **Engineering support vector machine kernels that recognize translation initiation sites.** *Bioinformatics* 2000, **16:**799-807
47. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci USA* 2000, **97:**262-267
48. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D: **Support vector machine classification and validation of cancer tissue samples using microarray expression data.** *Bioinformatics* 2000, **16:**906-914

특정연구개발사업 연구결과 활용계획서

<table>
<tr><td rowspan="2">사업명</td><td>중사업명</td><td colspan="4">21세기 프론티어 연구개발 사업</td></tr>
<tr><td>세부사업명</td><td colspan="4">작물유전체기능연구사업</td></tr>
<tr><td colspan="2">과제명</td><td colspan="4">Gene mining 시스템 개발 및 미생물 기능 분석에 관한 연구</td></tr>
<tr><td colspan="2">연구기관</td><td colspan="2">한국생명공학연구원</td><td>연구책임자</td><td>허 철구</td></tr>
<tr><td colspan="2">총연구기간</td><td colspan="4">2001년 8 월.1 일. ~ 2004년. 6 월. 30 일. (35개월)</td></tr>
<tr><td colspan="2" rowspan="2">총 연구비
(단위 : 천원)</td><td>정부출연금</td><td colspan="2">민간부담금</td><td>합계</td></tr>
<tr><td>300,000</td><td colspan="2"></td><td>300,000</td></tr>
<tr><td colspan="2">기술분야</td><td colspan="4">생명공학</td></tr>
<tr><td colspan="2">참여기업</td><td colspan="4"></td></tr>
<tr><td colspan="2">공동연구기관</td><td colspan="4"></td></tr>
<tr><td colspan="2">위탁연구기관</td><td colspan="4"></td></tr>
<tr><td colspan="2" rowspan="2">연구결과활용
(해당항목에(∨)
표시)</td><td>1. 기업화 ( )</td><td>2. 기술이전(O)</td><td>3. 후속연구추진( )</td><td colspan="2">4.타사업에 활용( )</td></tr>
<tr><td>5. 선행 및 기초연구(O)</td><td>6. 기타목적활용(교육,연구)( )</td><td>7. 활용중단(미활용)( )</td><td colspan="2">8. 기타( )</td></tr>
</table>

특정연구개발사업 처리규정 제 31조(연구개발결과의 보고) 제 2항에 의거 연구결과 활용계획서를 제출합니다.

첨부 : 1. 연구결과 활용계획서 1부.
2. 기술요약서 1부

2004 년 8 월 30 일

연구책임자 : 허 철구 (인)
연구기관장 : 양 규환 (직인)

과학기술부장관 귀하

[첨부1]

연구결과 활용계획서

1. 연구목표 및 내용

가. 과제의 최종 목표

Plant 관련 유전자를 기능 해석을 위하여 EST, DNA Chip 데이터, Proteomics데이타, Protein interaction, Metabolic Pathway 데이터, Data mining 기법이용 등을 통합 분석하여 실험실에 효율적인 정보 제공을 목표로 한다.

나. 과제의 1단계 목표

Plant 유전자 기능 해석 시스템 개발

- ❑ DNA Chip 분석용 S/W개발
- ❑ Peptide Mass fingerprinting S/W개발, Functional Catalog 제작
- ❑ EST, DNA chip 정보, Peptide Mass spectrum, gene function catalog의 통합 시스템 구현
- ❑ co-expression된 정보의 co-regulation region 예측 시스템 구축

2. 연구수행결과 현황(연구종료시점까지)

가. 특허(실용신안) 등 자료목록

나. 프로그램 등록목록

프로그램 명칭	등록번호	등록일자	개발자	비고
AMWISE 및 fBIND 기법을 이용한 Peptide Mass Fingerprinting Database 관리 프로그램	2004-01-12 -835	2004.1.12	홍 성의, 허철구, 최 도일	
AMWISE 및 fBIND 기법을 이용한 Peptide Mass Fingerprinting 프로그램	2004-01-12 -836	2004.1.12	홍 성의, 허 철구, 최 도일	
cDNA Microarray 데이터 Classification Tool	2004-01-22 -839	2004.1.22	박 선용, 허 철구, 최 도일	
cDNA Microarray 데이터 Clustering Tool	2004-01-22 -840	2004.1.22	박 선용, 허 철구, 최 도일	

다. 노하우 내역

라. 발생품 및 시작품 내역

마. 논문게재 및 발표 실적

○논문게재 실적(필요시 별지사용)

학술지 명칭	제목	게재연월일	호	발 행 기 관	국 명	SCI게재 여부
Functional & Integrative genomics	EST and microarray analysis of pathogen-responsive genes in hot pepper(*Capsinum annuum*) non-host resistance against soybean pustule pathgen(*Xanthomonas axonopodis* pv. *glycines*)	2004.10	4	Springer-Verlag Heldelberg	독일	
Molecules and Cells	Analysis of the Root Nodule-Enhanced Transcriptome from Soybean	2004.10	18	한국분자.세포생물학회	한국	SCI
계: 건수						

○학술회의 발표 실적(필요시 별지사용)

학술회의 명칭	제목	게재연월일	호	발 행 기 관	국 명
		년 월 일			
계: 건수					

3. 연구성과

- AMWISE 및 fBIND 기법에 의한 peptide mass spectrum 분석용 S/W 및 DB관리 기술, (주)위더스텍,
-계약 조건 : 선급실시료 3000만원 2004.3.31일까지 입금
실시자에게 "통상실시권" 을 부여 작물유전체기능연구사업단 만 무료 제공
경상실시료 : 총매출액의 3%
제 3자에게 양도시 : 대가의 25% 제공

4. 기술이전 및 연구결과 활용계획

가. 당해연도 활용계획(6하원칙에 따라 구체적으로 작성)

나. 활용방법

다. 차년도이후 활용계획(6하원칙에 따라 구체적으로 작성)

5. 기대효과

6. 문제점 및 건의사항(연구성과의 제고를 위한 제도·규정 및 연구관리 등의 개선점을 기재)

[첨부2]

기술 요약서

■ 기술의 명칭

■ 기술을 도출한 과제현황

과제관리번호	M101KG010001-03K070100520			
과제명	Gene mining 시스템 개발 및 미생물 기능 분석에 관한 연구			
사업명	21세기 프론티어연구개발사업			
세부사업명	작물유전체기능연구사업			
연구기관	한국생명공학연구원		기관유형	정부출연연구소
참여기관(기업)				
총연구기간	35개월			
총연구비	정부(300,000)천원 민간( )천원 합계( )천원			
연구책임자 1	성명	허 철구	주민번호	
	근무기관 부서	한국생명공학연구원 유전체연구센터	E-mail	hurlee@kribb.re.kr
	직위/직급	선임기술원	전화번호	042-879-8560
연구책임자 2	성명		주민번호	
	근무기관 부서		E-mail	
	직위/직급		전화번호	
실무연락책임자	성명		소속/부서	
	직위/직급		E-mail	
	전화번호		FAX	
	주소	(-)		

■ 기술의 주요내용

[기술의 개요]

Plant EST 데이터를 미국 NCBI GenBank에서 dbEST를 수집하고 식물별로 분류한 다음 종별, 조직별 유용유전자를 찾아내는 기술과 기능을 예측하는 것은 작물유전체연구에 가장 기본이 되는 것이며 해외에 비하여 연구비 규모가 작은 국내 상황에서는 적극적으로 활용할 가치가 매우 높은 데이터 가공 기술임. Microarray 분석용 S/W개발 기술은 이미 알려진 알고리즘들을 통합 버전으로 개발하고 상용제품에서 제공하지 못하는 내용을 첨가하여 국내 생산된 microarray 데이터 분석 지원에 효과적으로 활용하기 위해 개발되었으며, MALDI-TOF에서 나온 데이터를 이용하여 단백질 서열을 판별하는 기술은 국내 원천기술로 개발할 가치가 있는것이다. 또 위의 세가지 핵심기술을 통합하여 서열 데이터, microarray데이타, Peptide mass spectrum데이터를 생산하더라도 생물정보 기법으로 분석하기 쉽게 통합 개발하는 기술이 필요하다.

<기술적 특징>

(1) 대량의 Plant EST를 수집, 가공, 분석하여 유용유전자 발굴에 필요한 시스템 구축 및 데이터베이스 구축 기술 개발

(2) Microarray 데이터 분석의 전 과정을 공개용 R language로 개발하여 새로운 알로리즘 개발에 효과적으로 대처할 수 있고 상용제품의 수입대체효과가 매우 크다.

(3) MALDI-TOF를 이용한 프로테옴 분석 관련 연구자들에게 국내 생산된 데이터를 쉽게 분석할 수 있는 데이터베이스 구축이 원천기술 개발로 가능하게 되었다.

[용도 · 이용분야]

(1) Plant EST 가공 기법을 Human, Mouse 등 다른 생물의 EST분석 기법에 동일한 기술로 분석 할 수 있다. 이를 확장하면 alternative splicing 분석기술까지 확장이 가능하다.

(2) cDNA Chip을 이용한 실험자들에게 모두 적용이 가능하여 사용 범위가 매우 광범위하게 적용가능하다.

(3) 인삼을 비롯한 국내 유전체 연구과 프로테옴 연구자들에게 Peptide Mass Spectrum 분석 기술은 매우 활용성이 높다.

■ 기술의 분류

[기술코드] □□□ (3 Digit) (KISTEP 홈페이지 기술요약서용 기술분류표 참조)

[기술분야] (1개만 선택(∨로 표시)하여 주십시오)

□ 정보산업 □ 기계설비 □ 소재 □ 정밀화학 · 공정 □ 생명과학
□ 원자력 □ 자원 □ 에너지 □ 항공 · 우주 □ 해양
□ 교통 □ 보건 · 의료 □ 환경 □ 기초 · 원천 □ 기타

[기술의 활용유형] (1개만 선택(∨로 표시)하여 주십시오)

□ 신제품개발 □ 신공정개발 □ 기존제품개선 □ 기존공정개선
□ 기 타 ( )

[기술의 용도] (복수 선택(∨로 표시)가능합니다)

□ 기계설비 □ 부품소자 □ 원료재료 □ 소프트웨어
□ 가공처리기술 □ 자동화기술 □ 불량률 감소 등 현장애로기술
□ 제품설계기술 □ 공정설계기술 □ 기 타 ( )

■ 산업재산권 보유현황(기술과 관련한)

권리유형	명 칭	국가명	출원단계	일자	등록번호

* '권리유형'란에는 특허, 실용신안, 의장, 컴퓨터프로그램, 노하우 등을 선택하여 기재
* '출원단계'란에는 출원, 공개, 등록 등을 선택하여 기재

■ 기술이전 조건

이전형태	☐ 유상 ☐무상	최저기술료	30,000 천원
이전방식	☐ 소유권이전 ☐ 전용실시권 ☐ 통상실시권 ☐ 협의결정 ☐ 기타( )		
이전 소요기간	년 개월	실용화예상시기	년도
기술이전시 선행요건			

* 기술이전시 선행요건 : 기술이전을 위한 사전준비사항(필수 설비 및 장비, 전문가 확보 등)을 기술
* 실용화예상시기 : 기술을 활용한 대표적인 제품이 최초로 생산이 시작되는 시기를 기재

■ 기술의 개발단계 및 수준

[기술의 완성도] (1개만 선택(∨로 표시)하여 주십시오)

	① **기초, 탐색연구단계** : 특정용도를 위해 필요한 신 지식을 얻거나 기술적 가능성을 탐색하는 단계
	② **응용연구단계** : 기술적 가능성의 실증, 잠재적 실용화 가능성의 입증 등 실험실적 확인 단계
	③ **개발연구단계** : Prototype의 제작, Pilot Plant Test 등을 행하는 단계
	④ **기업화 준비단계** : 기업화에 필요한 양산화 기술 및 주변 기술까지도 확보하는 단계
	⑤ **상품화 완료단계**

[기술의 수명주기] (1개만 선택(∨로 표시)하여 주십시오)

	① **기술개념 정립기** : 기술의 잠재적 가능성만 있는 단계
	② **기술실험기** : 기술개발에 성공했으나 아직 실용성, 경제성 등이 확실치 않은 단계
	③ **기술적용 시작기**: 최초의 기술개발국에서만 활용되고 있는 단계
	④ **기술적용 성장기**: 기술개발국 및 일부 선진국에서 활용되고 있는단계
	⑤ **기술적용 성숙기**: 선진국사이에서 활발한 기술이전이 일어나며, 기술의 표준화가 되어가는 단계
	⑥ **기술적용 쇠퇴기**: 선진국에서 개도국으로 기술이전이 활발하게 일어나고, 선진국에서는 기술의 가치가 저하되나, 개도국에서는 아직 시장의 가치가 높은 기술

[기술발전 과정상의 기술수준] (1개만 선택(∨로 표시)하여 주십시오)

	① **외국기술의 모방단계** : 이미 외국에서 개발된 기술의 복제, reverse Eng.
	② **외국기술의 소화 · 흡수단계** : 국내시장구조나 특성에 적합하게 적응시킴
	③ **외국기술의 개선 · 개량단계** : 성능이나 기능을 개선시킴
	④ **신기술의 혁신 · 발명단계** : 국내 최초로 개발

■ 본 기술과 관련하여 추가로 확보되었거나 개발중인 기술

[기술개요]

기술명	
개발단계	□ 연구개발 계획 □ 연구개발 중 □ 연구개발 완료
기술개요	

[기술을 도출한 과제현황]

과제관리번호				
과제명				
사업명				
세부사업명				
연구기관			기관유형	
참여기관(기업)				
총연구기간				
총연구비	합계 : ( )백만원 - 정부 : ( )백만원 민간 : ( )백만원			
연구책임자	소속		성명	
	전화번호		E-mail	
연구개발 주요내용				