GOVP1200511505

과제번호 : UCPDM0070213-2003034-6

자생식물 유전자 D/B 구축 및 유전정보처리기술 개발

Construction and Development of Plant Gene Index and Bioinformatics System

한국생명공학연구원

과 학 기 술 부

제 출 문

과학기술부 장관 귀하

본 보고서를 " 자생식물 유전자 D/B 구축 및 유전정보처리기술 개발"과제의 보고서로 제출합니다.

2002 . 8. 30.

주 관 연 구 기 관 명 : 한국생명공학연구원

주관연구책임자 : 최 도 일

연 구 원 : 이 상 협

〃 : 허 철 구

〃 : 정 영 희

〃 : 박 정 디

〃 : 정 은 숙

〃 : 성 은 수

〃 : 오 상 근

〃 : 이 소 영

〃 : 김 영 철

〃 : 변 상 진

〃 : 김 지 협

〃 : 김 수 용

보고서 초록

식물유래 유전정보의 대량발굴 및 대량 발굴된 유전정보 처리기술을 개발 연구를 수행하여 고추, 인삼, 고구마, 참깨 및 개똥쑥 유래의 EST 를 6 만여 건 발굴하여 데이터베이스를 구축하였으며 사업단의 홈페이지를 통하여 공개하였고 식물유전정보처리기술을 개발하여 EST 분석시스템, InterPro Scan 프로그램, Internal BLAST system, Plant Gene Index, Gene Expression profile DB, 유전자칩 분석 시스템 등을 구축하여 본 사업단의 제 3 분야 과제에서 대량으로 발굴되는 자생식물의 유전정보 및 이에 파생되는 결과물을 분석 집약하였음. 현재 개발된 모든 분석 시스템과 데이터베이스가 http://plant.pdrc.re.kr 을 통하여 공개되고 있으며 국내의 식물생명공학 분야의 연구를 위한 공공서비스 기능을 수행하고 있음.

<table>
<tr><td>과제관리번호</td><td></td><td>해당단계
연구기간</td><td></td><td>단계 구분</td><td>(해당단계) / (총단계)</td></tr>
<tr><td rowspan="2">연구사업명</td><td>중 사 업 명</td><td colspan="4">21C 프론티어연구개발사업</td></tr>
<tr><td>세부사업명</td><td colspan="4">자생식물이용기술개발사업</td></tr>
<tr><td rowspan="2">연구과제명</td><td>중 과 제 명</td><td colspan="4">중과제가 있을 경우에는 기재</td></tr>
<tr><td>세부(단위) 과제명</td><td colspan="4">자생식물 유전자 D/B 구축 및 유전정보처리기술 개발</td></tr>
<tr><td>연구책임자</td><td>최 도 일</td><td>해당단계 참여연구원수</td><td>총 : 30 명
내부 : 5 명
외부 : 25 명</td><td>해당단계 연구비</td><td>정부: 2,400,000 천원
기업: 0 천원
계: 2,400,000 천원</td></tr>
<tr><td>연구기관명 및 소속부서명</td><td colspan="2">한국생명공학연구원</td><td>참여기업명</td><td colspan="2"></td></tr>
<tr><td>국제공동연구</td><td colspan="5">상대국명 : 상대국연구기관명 :</td></tr>
<tr><td>위 탁 연 구</td><td colspan="5">연구기관명 : 연구책임자 :</td></tr>
<tr><td colspan="5">요약(연구결과를 중심으로 개조식 500 자이내)</td><td>보고서면수</td></tr>
<tr><td colspan="6">식물유래 유전정보의 대량발굴 및 대량 발굴된 유전정보 처리기술을 개발 연구를 수행하여 고추, 인삼, 고구마, 참깨 및 개똥쑥 유래의 EST 를 6 만 여 건 발굴하여 데이터베이스를 구축하였으며 사업단의 홈페이지를 통하여 공개하였고 식물유전정보처리기술을 개발하여 EST 분석시스템, InterPro Scan 프로그램, Internal BLAST system, Plant Gene Index, Gene Expression profile DB, 유전자칩 분석 시스템 등을 구축하여 본 사업단의 제 3 분야 과제에서 대량으로 발굴되는 자생식물의 유전정보 및 이에 파생되는 결과물을 분석 집약하였음. 현재 개발된 모든 분석 시스템과 데이터베이스가 http://plant.pdrc.re.kr 을 통하여 공개되고 있으며 국내의 식물생명공학 분야의 연구를 위한 공공서비스 기능을 수행하고 있음.</td></tr>
<tr><td rowspan="2">색 인 어
(각 5 개이상)</td><td>한글</td><td colspan="4">유전체, 유전정보, 자생식물, 생물정보처리기술</td></tr>
<tr><td>영어</td><td colspan="4">genomics, genome information, endemic plants, bioinformatics</td></tr>
</table>

요 약 문

Ⅰ. 제 목

자생식물 유전자 D/B 구축 및 유전정보처리기술 개발

Ⅱ. 연구개발의 목적 및 필요성

- 국내의 식물 게놈연구가 활성화되어 유전자 데이터가 양산되고 있음
- 국내 연구진에 의해 확보된 유전자가 벼, 배추를 비롯하여 각각 수천에서 수만 건의 데이터를 확보하고 있음
- 데이터의 대량생산에 대해 지원할 수 있는 분석 시스템이 부재하고 계속 선진국에 의존하고 있는 실정임
- 대규모 EST 발굴과 뒤따르는 유전자 Chip 연구는 양산된 데이터 처리를 위한 전문적인 전산프로그램의 제작과 computation 의 지원을 필요로 하고 있음
- 본 사업단의 과제를 통하여 연간 최소 수천에서 수만 건의 EST 가 발굴될 예정이며 이를 위한 분석시스템 확립이 절실히 필요함.

Ⅲ. 연구개발의 내용 및 범위

연차	연구목표	연구 내용 및 범위
1 차 년도 (2000)	- 식물 유용 유전자 DB 구축 및 EST 분석시스템 개발	-cDNA library 제작 및 연구 시스템 구축 -Plant EST DB 구축 및 자동분석시스템 -사용자 인터페이스 개발 -대규모 유전자발굴을 위한 연구기반구축 -EST clustering 및 multialignment 분석 - Internal BLAST 시스템구축
2 차 년도 (2001)	- Gene Expression profiling 시스템 개발	-Gene expression profile 을 위한 Algorithm -고추 EST DB(총 12,526 unigene) 구축 -Web 을 통한 EST DB 공개 -인삼 EST(6,384 개 unigene) DB 구축 -참깨 EST DB 구축 및 공개 -고구마 EST DB 구축 및 공개 -Gene expression profile 정보처리기술 개발
3 차 년도 (2002)	- EST 및 DNA Chip 정보 분석을 위한 통합 시스템개발	-타 과제 도출 EST 의 D/B 화 -Plant Gene Index 구축 -식물에 대한 Organism specific 한 EST 분석 및 서비스 -Plant Gene Index 의 웹을 통한 공개 -유전자 칩 데이터 베이스구축 -Gene Expression 및 관련정보 분석 DB -웹 인터페이스를 이용한 정보분석서비스

Ⅳ. 연구개발결과

-cDNA library 제작 및 연구 시스템 구축
-저항성 관련 식물 EST의 대량발굴
-Plant EST DB 구축 및 자동분석시스템
-사용자 인터페이스 개발
-대규모 유전자발굴을 위한 연구기반구축
-EST clustering 및 multialignment 분석시스템 구축
- Internal BLAST 시스템구축
-Gene expression profile을 위한 Algorithm
-고추 EST DB(총 12,526 unigene) 구축
-Web을 통한 EST DB 공개
-인삼 EST(6,384개 unigene) DB 구축
-참깨 EST DB 구축 및 공개
-고구마 EST DB 구축 및 공개
-Gene expression profile을 위한 정보처리기술 개발
-타 과제 도출 EST의 D/B화
-Plant Gene Index 구축
-Arabiopdis를 비롯한 8종의 식물에 대한 Organism specific한 EST분석 및 서비스
-Plant Gene Index의 웹을 통한 공개
-유전자 칩 데이터 베이스구축
-유전자 클론 분양
-Gene Expression 및 관련정보 분석 DB
-웹 인터페이스를 이용한 정보분석서비스

Ⅴ. 연구개발결과의 활용계획

· 발굴된 유전자의 활용계획

자생식물이용기술개발 사업 1 단계를 통하여 인삼, 고추, 고구마, 개똥쑥, 참깨 및 야생종 담배를 통틸어 발굴된 유전자는 약 6 만 여 개에 이르며 독립(independent) 유전자를 기준으로 해도 약 3 만여 개에 달함. 현재 모든 유전자는 일련의 분석과정을 거쳐 자생식물 사업단의 홈페이지 (http://plant.pdrc.re.kr)에 데이터 베이스로 구축되어 있으며 사용자 편의에 따라 Key Word 또는 다양한 Category 를 이용하여 검색할 수 있게 되어 있으며 각 연구자에게 요청하여 실물 유전자를 획득할 수 있어 향 후 식물의 기능유전체 연구에 유용하게 활용 될 것임. 특히 고추의 유전자의 경우는 5 천 개의 유전자가 실려진 유전자 칩을 작물기능유전체연구사업단과 공동으로 개발하였으며 유전자발현 프로파일을 대량생산하여 데이터베이스화 된 상태로 당장 유전자의 기능을 유추할 수 있는 데이터가 제공되고 있으므로 고추연구자가 언제든지 직접 활용 할 수 있는 체계를 갖추고 있음.

· 데이터 베이스 및 생물정보처리기술의 활용

본 과제를 통하여 식물유전정보처리의 근간이 되는 EST 분석, DB 화 및 유전자발현 프로파일분석 기술 등이 개발되었으며, 이는 향후 식물유전체기능연구에 광범위하고 지속

적으로 활용될 수 있을 것으로 판단됨. 특히 국내에서는 처음으로 유전자칩 데이터 프로세싱기술과 발굴된 데이터를 데이터베이스화하여 공개함으로써 앞으로 다양하게 발굴되는 유전자발현 프로파일의 데이터베이스를 구축하는 모델로 활용될 수 있을 것으로 전망함.

S U M M A R Y

I. Title: Construction and Development of Plant Gene Index and Bioinformatics System

II. Objectives and Need for research

The overall goal of this project is to produce several thousands of EST data from plants and develop analysis system. Many of gene data has been produced from plants and increased rapidly in Korea. To develop knowledge of plant genomics, need more information such as EST and expression profiles. Until now, for analysis of these data, we have to use advanced technology of other countries because of analysis systems for mass data are not yet developed in Korea. First of all, development of analysis system of EST and gene chip data need for plant research.

III. Research contents

1. Objectives of 1st year (2000): Development of plant gene DB and analysis system.

-Construct of cDNA library and development of research system.
-Construct of plant EST DB and development of analysis system.
-Development of user interface.
-Development of research infra for massive gene data production.
-Development of EST clustering, multi-alignment and internal BLAST system

2. Objectives of 2nd year (2001): Development of gene expression profiling system.

-Develop knowledge of algorithm for gene expression profiling.
-Construct of hot pepper EST DB and open information by Web.
-Construct of ginseng, sesame and sweet potato EST DB and open information by Web.
-Development of bioinformatics system for gene expression profiling.

3. Objectives of 3rd year (2002): Development of bioinformatics system for EST and the gene chip analysis.

-Construct of other plants EST DB
-Construct of plant gene index and open information by Web.
-Analysis and service of information for plant organism specific EST.
-Construct of the gene chip DB.
-Construct of DB for gene expression and related information analysis.
-Service of bioinformatics analysis by Web interface.

IV. Results

-Construct of cDNA library and development of research system.

-Massive production of ESTs, which are related to disease resistant.
-Development of plant EST DB and analysis system.
-Development of research infra for mass gene data production
-Development of EST clustering, multi-alignment and internal BLAST system
-Construct of hot pepper EST (total 12,526) DB and open information by Web.
-Construct of ginseng, sesame and sweet potato EST DB and open information by Web.
-Development of bioinformatics system for gene expression profiling.
-Construct of plant gene index and open information by Web.
-Service of plant organism specific EST analysis information in 8 species.
-Service of gene clone distribution.
-Construct of the gene chip DB.
-Construct of DB for gene expression and related information analysis.
-Service of bioinformatics analysis by Web interface.

V. Application of results

1. Application of attaining genes

Over sixty thousand EST clones were produced from several plants such as hot pepper, sesame, ginseng, *Artemisia annua* and wild-type tobacco and independent genes were about thirty thousand of them. All of the EST information is open on Internet home page of Plant Diversity Research Center and any user can access and search the information by key words or variety categories and also, received the EST clones by application. It might be potential benefit for functional genomics research. Especially in hot pepper. 5K gene chips, which were 5000 of hot pepper genes were arrayed in a slide. were developed by collaboration with Crop Functional Genomics Center. Many of expression profiles were produced using the gene chip and DB was constructed. Researcher can deduce function of unknown gene and application for other research using the DB.

2. Application of DB and bioinformatics technology

Technologies of EST analysis, DB construction and gene expression profile analysis were developed from this project. The DB and bioinformatics technology can be used for functional genomics research. And technology for gene chip data processing and DB can be used model system of other expression profile DB.

C O N T E N T S

목 차

제 1 장 연구개발과제의 개요

가. 연구개발의 중요성

- 국내의 식물 게놈연구가 활성화되어 유전자 데이터가 양산되고 있음
- 국내 연구진에 의해 확보된 유전자가 벼, 배추를 비롯하여 각각 수천에서 수만 건의 데이터를 확보하고 있음
- 데이터의 대량생산에 대해 지원할 수 있는 분석 시스템이 부재하고 계속 선진국에 의존하고 있는 실정임
- 대규모 EST 발굴과 뒤따르는 유전자 Chip 연구는 양산된 데이터 처리를 위한 전문적인 전산프로그램의 제작과 computation 의 지원을 필요로 하고 있음
- 본 사업단의 과제를 통하여 연간 최소 수천에서 수만 건의 EST 가 발굴될 예정이며 이를 위한 분석시스템 확립이 절실히 필요함.

나. 지금까지의 연구개발 실적

- 본 연구팀은 식물 EST 발굴 연구를 통하여 현재까지 3,000 개의 유전자 단편을 확보하고 있음
- 유전자 chip 기술을 확보하고 있으며 prototype data 를 확보하고 있음
- 대량유전자 분석 시스템을 개발하여 현재 수 백 개 이상의 염기서열을 일시에 분석할 수 있는 시스템을 구축하고 있음
- 연구소 홈페이지를 이용한 BLAST, Contig assembly 등 다양한 유전정보 분석서비스를 제공하고 있음
- 단백질 데이터베이스 분석을 위한 전체 단백질서열을 D/B 로 확보하고 있으며 internal BLAST 를 이용한 EST 대량분석이 가능함

다. 앞으로의 전망

- 국내연구과제로 벼, 배추, 고추의 EST 확보 및 게놈연구가 진행중이며 인삼, 오가피 등 약용식물의 게놈연구가 본 사업단 과제로 진행될 예정임
- 국내의 식물게놈 연구가 가속화되어 유전정보처리기술의 확보가 식물게놈연구 성공의 관건이 될 것임
- 벼, 애기장대 등 전세계적인 식물전체게놈 프로젝트의 완성과 더불어 막대한 유전정보가 획득 가능해질 것임
- 세계적으로 새로이 진행되는 다수의 식물게놈정보도 수년 내에 공개될것임
- 결국 유전정보처리기술의 확보여부가 홍수처럼 밀려드는 정보를 이용한 유전자기능규명 및 지적재산권획득에 결정적 관건이 될 것임.

제 2 장 국내외 기술개발 현황

가. 기술의 정의

생물정보학은 생명공학(Biotechnology)과 정보공학 (Information Technology)이 합쳐진 학문으로 컴퓨터를 기반으로 하는 정보공학 기술을 생물학에 접목하여 대량의 생물학 정보를 분석활용하는 기술을 의미한다.

나. 기술의 동향

0 생물정보학은 생명과학 및 유전체 관련 지식이 폭발적으로 늘어남에 따라 이러한 정보를 효과적으로 관리하기 위해 탄생한 학문이며 이미 생물산업 분야의 핵심응용 기술로 인정받고 있다. 초기에는 단순한 유전자 서열분석을 위한 컴퓨터 프로그램 개발이 주목적이었으나 분자생물학의 급속한 발전과 Genome projects 의 가속화를 통하여 생물정보량이 기하급수적으로 늘어남에 따라 생물정보의 생산으로부터 정보응용까지 생물학 전 분야에 걸쳐 활용되고 있다. 관련 software 시장도 급성장하여 PANGEA, MILLENIUM, NETGENICS 등 많은 생물학 software 개발회사들이 급성장하고 있고 새로운 벤처회사가 설립되고 있다. 초기의 생물정보학의 결과물들과 분석소프트웨어들은 정보공유의 정신에 의해 대부분 인터넷을 통해 무료로 생물학자들에게 제공되었지만 점차 정보분석의 중요성이 커짐에 따라 상업화하는 추세에 있다.

0 생물정보학에서 가장 기초적인 분야는 생물학 정보 데이터 베이스를 만들고 관리하는 컴퓨터 생물학으로 다양한 생물의 유전자 동정, 신규발견 단백질 및 RNA 구조와 기능 예측방법 개발, 단백질서열 분류, 단백질 3 차원 구조모델 개발, 그리고 진화단계 조사목적의 단백질 계통발생도 작성을 다루고 있어 생물정보학의 초석이 되고 있다. 이 분야에서 선두를 달리고 있는 기업은 Compugen(이스라엘)으로 유전자서열 기능분석을 위한 고성능 연산장치를 개발하고 있다. 생물정보학 분야에서 가장 성장 잠재력이 큰 분야로는 단백질의 발현, 단백질 동정, 단백질 구조 및 작용 등을 다루는 단백질학 분야로 전망되고 있다. 유전학, 단백질학, 조합화학 그리고 대량검색시스템 등에 대한 토대를 제공함에 따라 21 세기에 각광받는 분야로 자리를 잡을 것으로 예상되어 관련 하드웨어, 데이터베이스 그리고 소프트웨어가 거대시장을 형성할 것으로 예측된다.

0 미국은 1988 년 분자생물정보에 대한 공용 데이터베이스를 생산 관리하는 NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov)를 설립 운영 중에 있으며 NCBI 의 활동은 크게 기초연구와 데이터베이스/프로그램개발 그리고 교육 기능으로 나눌 수 있다. 데이터 베이스와 프로그램개발분야는 유전체 연구와 함께 매우 중요해진 NCBI 기능의 하나이다. 대표적인 데이터 베이스로는 1992 년에 만들어진 유전자들 및 단백질의 서열정보를 담고 있는 GenBank 이며 유럽의 EMBL, 일본의 DDBJ 와 데이터를 공유하고 있으며, 염기서열에 관한 특허에 관해서도 관여하고 있다. 이에 덧붙여 Pubmed (생물학 및 의학관련 논문의 제목 및 초록검색기능), Online Mendelian

Inheritance in Man (OMIM, 각종 유전자에 관련된 질병, 변이에 관한 목록), Molecular Modeling DataBase (MMDB, 단백질 3 차원 구조), Unique human Gene sequence collection (UniGene, 인간 유전자중 중복되지 않은 목록), Gene map of the human genome (인간 유전자의 염색체 위치 정보), Taxanomy browser (각종 생물의 진화관계 및 분지 관계정보), Cancer Genome Anantomy Project (CGAP, 암 발생관련유전자의 발현 및 변이에 관한 정보)를 다루고 있다.

0 TIGR (The Institute for Genome Research, http://www.tigr.org)는 1992 년에 설립된 비영리연구기관으로 NIH 의 NHGRI 와 함께 미국의 유전체 연구를 주도하는 기관이다 주 임무는 유전체의 구조, 기능 및 비교 유전체학 연구이다. 1995 년 독감을 일으키는 *Haemophilus influenze* 바이러스 유전체의 전 염기서열을 해독하므로 써 유전체 해독의 장을 열었으며, 이후 각종 미생물의 유전체 및 식물의 모델생물인 *Arabidpsis* 유전체의 해독에도 참여하였다. 따라서 이들의 생물정보학 기술은 NCBI 와 함께 미국에서 선도적인 역할을 하고 있으며 유전체를 해독하는데 필요한 각종 프로그램을 개발하여 각국의 연구기관에는 무료로 배포하고 있으며, 이들이 축적한 데이터도 비영리의 조건에 배포한다. 이 기관에서는 웹을 통해서도 보유하고 있는 데이터에 대한 다양한 검색 서비스를 제공한다.

0 TAIR (The Arabidopsis Information Resource, http://www.arabidopsis.org)는 식물의 유전체 만을 위한 그 중에서도 식물 유전체 연구에서 모델로 사용되는 *Arabidopsis thaliana* 만에 한정하여 운영하는 특이한 페이지이나, 식물의 유전체를 연구하는 연구자에게 있어서는 필수 불가결한 사이트이며, 컨소시엄을 통해 얻어진 *Arabidopsis* 데이터의 처리는 유럽의 MIPS 에서 제공하는 데이터와 함께 중요한 자원이다. 이곳은 앞에서 언급한 NCBI 나 TIGR 와는 달리 세 군데의 기관이 협력하여 운영되는 조직이다. 먼저 전산처리에 관해서는 지금은 그 명맥만이 유지되고 있는 NCGR (National Center for Genome Research)와 *Arabidopsis* 의 종자를 보유하고 있는 ABRC (Arabidopsis Biological Resource Center) 그리고 Stanford 대학의 식물연구기관인 Carnegie 연구소가 TAIR 를 구성한다. 따라서 TAIR 의 특이한 점은 데이터의 제공과 함께 연구자들에게 필요한 클론이나 종자에 대한 정보를 함께 제공하는 점에 있다.

0 1998 년 설립된 Celera genomics (http://www.celera.com/)는 불과 9 개월만에 인간 게놈의 염기서열을 분석하여 유명해진 기업으로 인간 게놈 외에 초파리, 쥐의 유전체 정보 및 관련 서비스를 제공하고 있다. Celera 는 Celera Discovery System 개발을 통해 통합정보시스템 구축하고 있고, 최근 단백질체 기술역량 강화 및 신약개발을 통해, 안정적이고 다양한 수익원 창출 추진 중이며, 유전체 뿐만 아니라 단백질체까지 정보 제공 영역을 넓히고 있으며 2002 년까지 100 개의 신약타깃 발굴을 목표로 의약 개발 분야에 직접 관여하기 시작하였다. Celera 의 기업가치는 2001 년 6 월 8 일 현재 2000 년 말에 비하여 35% 증가한 약 30 억 달러에 이르는 등 높이 평가되고 있다.

0 EBI (European Bioinformatics Institute, http://www.ebi.ac.uk)는 EMBL 의 일부분을 형성하는 비영리 학술 연구 기관이며, 독일 하이델 베르그에 설립되었던 EMBL 의 염기서열 데이터 라이브러리에 그 뿌리를 두고 있다. 서비스, 연구와 산업의 세 가지 프로그램으로 운영되고 있다. 서비스 프로그램에서는 데이터의 수집과 이용을 지원할 수 있도록 생물학적인 데이터베이스들 및 정보서비스를 제공할 수 있는 시스템을 구축하고 유지 보수하는데 초점을 맞추고 있다. 연구 프로그램에서는 전산분자생물학 (computational molecular biology)에서 첨단분야에 대한 순수 및 응용 연구를 수행한다 이와 같은 학술 활동은 진화, 게놈 비교, 유전자 예측, 단백질 모티브, 대사 경로 , 서열과 구조 비고, 분자생물학에서의 병렬연산, 단백질 3 차 구조 및 분자생물학적 서열분

석, 새로운 데이터 베이스, 새로운 방식의 데이터 베이스들의 연동 도구에 대한 것들을 포함한다. 산업 프로그램에서는 표준적인 기술들을 산업체에 전파하는데 역점을 두고 있다. EBI 에서는 다양한 데이터 베이스 및 서열 분석 도구들을 갖추고 있어서 생물학자들에게 매우 유용한 정보를 제공하고 있다. 현재 진행중인 주요 연구분야는 다음과 같다.

- CORBA
- Computational Genomics
- Structural Genomics
- Microarray Informatics
- Protein Sequence and Structural Analysis.

○ 엑스파시 (ExPASy; http://www.expasy.ch)는 SIB (Swiss Insititute of Bioinformiatics)에서 운영하는 생물정보 서비스 시스템이다. 여기서는 생물정보학 전반에 데이터베이스와 분석 소프트웨어들을 개발하고, 높은 수준의 생물정보 연구 프로그램을 유지하며, 학술 연구자들과 공동연구를 수행하고, 생물정보분야의 과학자들을 양성하기 위한 교육 및 세미나 과정의 마련 등의 목표를 가지고 활동하고 있다. SWISS-PROT 과 TrEMBL 같은 데이터 베이스 검색을 제공하며, 각종 소프트웨어들은 단백질분석을 위주로 하고 있고, Peptide mass spectrum, 단백질 3 차 구조 분석 및 모티브 분석 등도 포함하고 있다. ExPASy 에서 제공하는 주요 데이터 베이스로는 SWISS-PROT and TrEMBL (Protein knowledgebase), PROSITE (Protein families and domains 정보 제공), SWISS-2DPAGE (Two-dimensional polyacrylamide gel electrophoresis 정보 제공), SWISS-3DIMAGE (3D images of proteins and other biological macromolecules 정보 제공), SWISS-MODEL Repository (Automatically generated protein models), CD40Lbase (CD40 ligand defects), ENZYME (Enzyme nomenclature), SeqAnalRef (Sequence analysis bibliographic references) 등이 있다.

○ MIPS (Munich Information Center for Protein Sequence, http://www.mips.biochem.mpg.de/)는 독일의 GSF (National Research Center for Environment and Health)의 지원을 받는 바이오인포메틱스 연구소이다. PIR DB 유지, 유로판 (Eurofan: European Yeast Functional Analysis Program), ESSA (European Sequencing of Arabidopsis thaliana, 효모의 기능분석을 위한 독일 네트워크, 유럽 분자생물학 네트워크(EMBnet), 통합 데이터 베이스, 생물학적 데이터 베이스들의 연계, Neurospora crassa 에 대한 독일의 Sequencing 등의 프로젝트에 관여하고 있다. 여기서는 개발된 소프트웨어를 제공하기보다는 유전자 발현 데이터를 비롯한 다양한 데이터의 분석 결과를 제공해 주고 있다. 현재 진행중인 주요 연구 분야는 다음과 같다.

- Saccharomyces cerevisiae (Yeast) DB 구축
- Human Genome Analysis Project
- German Human cDNA Project
- PIR-International DB 구축
- Expression Analysis (MouseExpress)

○ 1986 년 일본유전학연구소 (National Institute of Genetics)에서부터 DNA 데이터 은행 기능을 수행하면서 미국 NCBI 와 유럽의 EMBL 과 국제 공동 협력을 통하여 DNA 관련 정보를 연구하기 위하여 DDBJ (DNA Data Bank of Japan, http://www.ddbj.nig.ac.jp/)설립 운영하고 있다. DDBJ 는 데이터베이스 관리 및 해외 협력을 담당하고 DNA 관련 분석

및 생물 정보 연구 기능은 CIB (Center for Information Biology)와 공동으로 운영 중이며 DNA 데이터 분석, 유전자 기능 연구, 생물학 데이터베이스 개발 등 일본 생물정보의 집대성 및 분석, 가공, 서비스를 수행하는 대표 기관이다. GenBank 에서 제공하는 데이터가 80% 정도이며 일본에 직접 모이는 DNA 정보는 20% 정도로 미국 의존도가 높다. 하지만 미국 GenBank 와 서비스하는 자료 형식이 같아서 국내에서도 일부 이용이 가능한 기능을 가지고 있다.

○ KEGG (Kyoto Encyclopedia of Genes and Genomes, http://www.ddbj.nig.ac.jp/)는 Kyoto 대학의 Kanehisa 박사팀에 의해 운영되는 pathway 관련 genome database 이다. KEGG 는 pathway 에 관련된 모든 정보들을 현재 급속도로 진행중인 genome project 의 결과물들과 잘 조합하여 제공하고 있으며, Java 기술등 최신 전산 기술들을 이용하여 사용자들에게 편리한 interface 를 제공하고 있다. 현재 KEGG 에서 진행중인 project 는 다음의 네 가지로 요약할 수 있다.

-분자생물학과 세포생물학의 연구를 통해 밝혀진 pathway 에 관련된 information 들을 모아 구축한 PATHWAY database 의 운영 (Metabolic pathway 와 regulatory pathway 에 관련된 정보들을 구분하여 제공)

- Pathway 에 관련된 유전자들 중 sequence 가 밝혀진 유전자들의 gene catalog 를 만들어 제공 (GENES database)
- Pathway 에 관련된 화합물들을 정리하여 Pathway 데이터베이스와 결합시킨 LIGAND 데이터베이스 제공
- Pathway comparision, pathway reconstruction, pathway design 등 관련된 bioinformatics technology 들의 개발

제 3 장 연구개발수행 내용 및 결과

가. 자생식물이용기술개발사업에서 생산된 EST 분석 정보

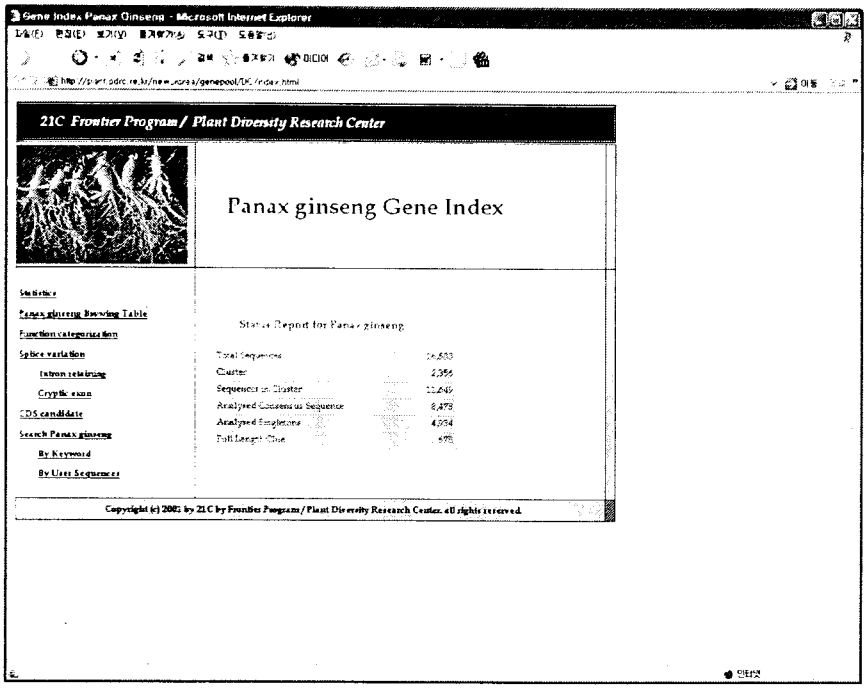
1. 관계형 데이타베이스 구축에서 Flat-file DB형태로 전환
   - 전체 스키마는 RDB가 아니므로 flat-file형태의 디렉토리별 자료 저장
   - word-기반의 자료만 RDB에 저장하는 형태임.

2. 웹 기반에서 검색 가능한 시스템으로 전환 운영

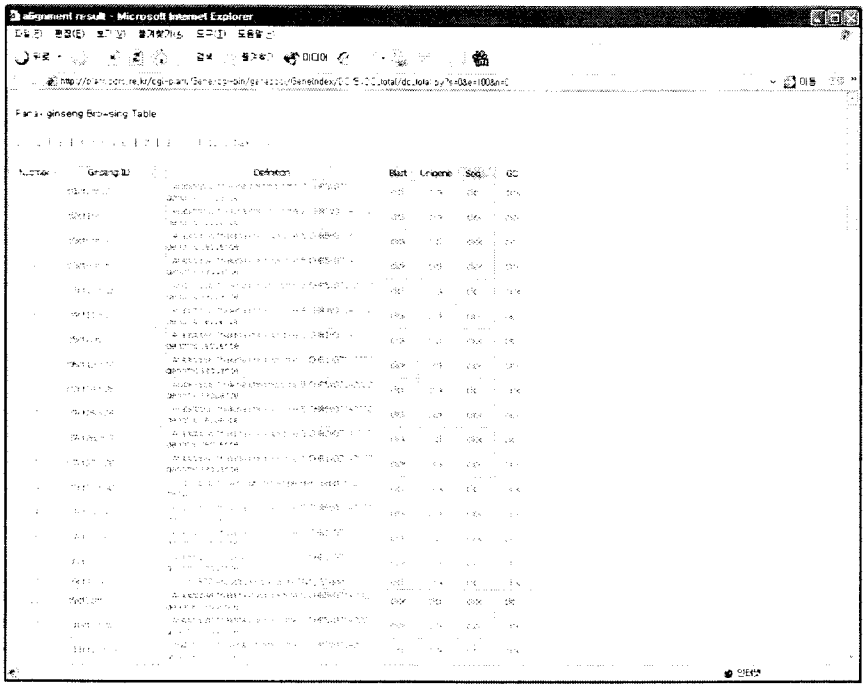
3. 제 3 대 과제 연구자들의 데이터 수집 및 분석 지원 후
   - NCBI 및 국가유전체정보센터에 등록 절차를 거치고 있슴.
   - 단, 연구자들의 특허 및 논문 등록 시 까지는 유보.
   - 항상 특허와 논문 시점이전에 공개시 외국 기관에서 즉시 분석하여 버리는 국가적인 경쟁력 약화 우려되는 문제점. (해외 기관들이 분석 능력이 탁월하여 현재까지 등록되는 정보는 국내에서 연구자들이 분석 후 공개가 바람직하다고 판단됨.)

4. 외국도 분석이 다 되는 시점까지는 NCBI에 등록 유보 추세이며Washington University와 같이 정부와 계약하여 데이터를 전문적으로 생산하는 기관은 72시간 내에 raw데이터를 직접 등록함. 이는 데이터를 생산하기 위한 별도의 연구비 계약이 필요한 것임.

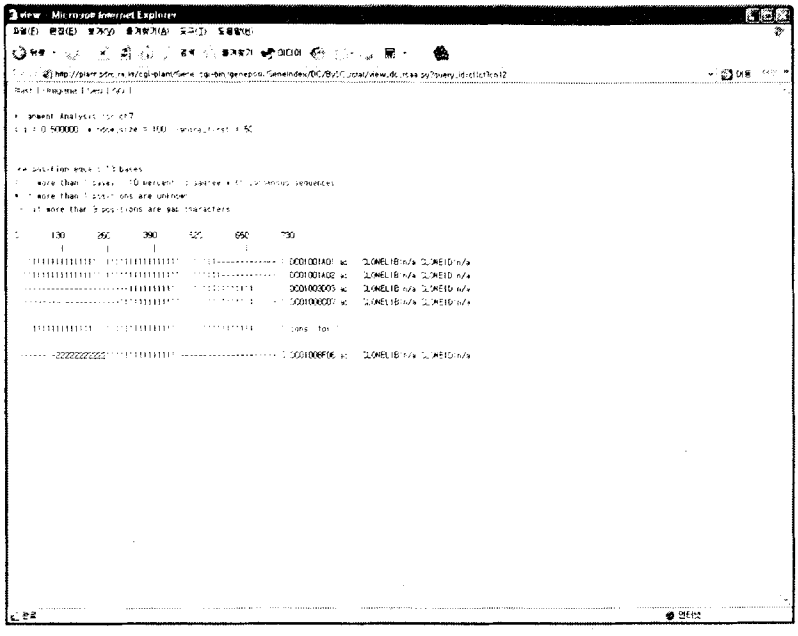
가) 인삼의 초기 화면

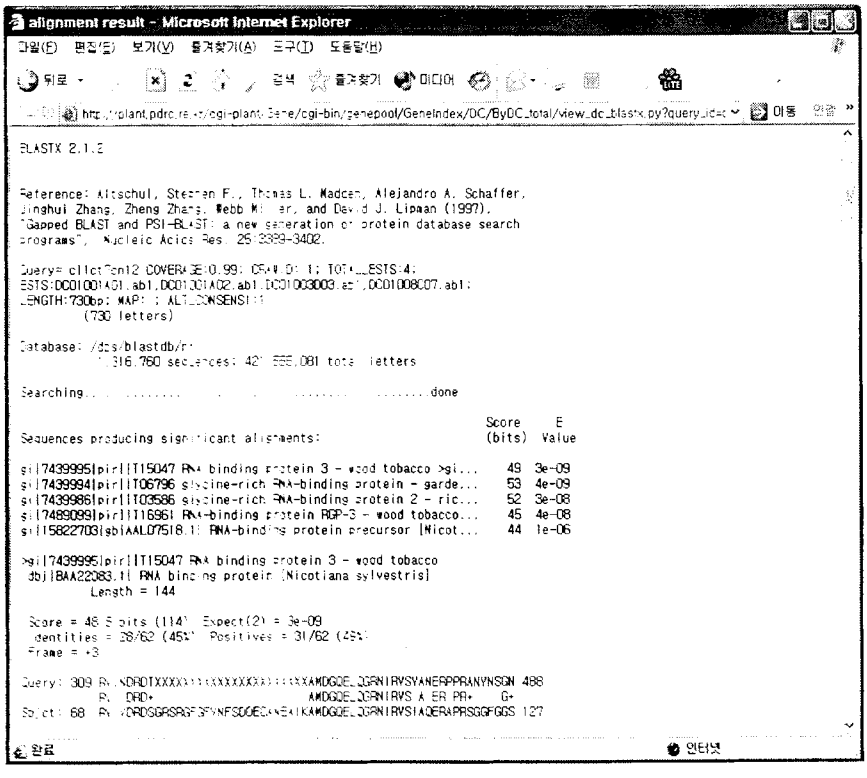


나) 전체 분석된 검색 화면

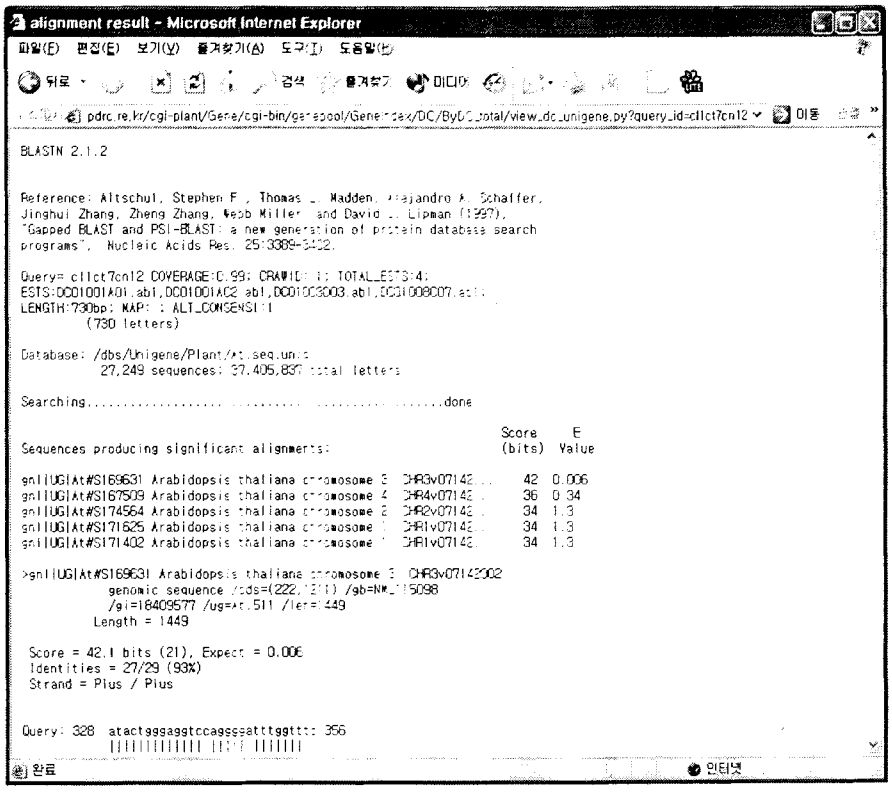


다) EST clustering 결과화면



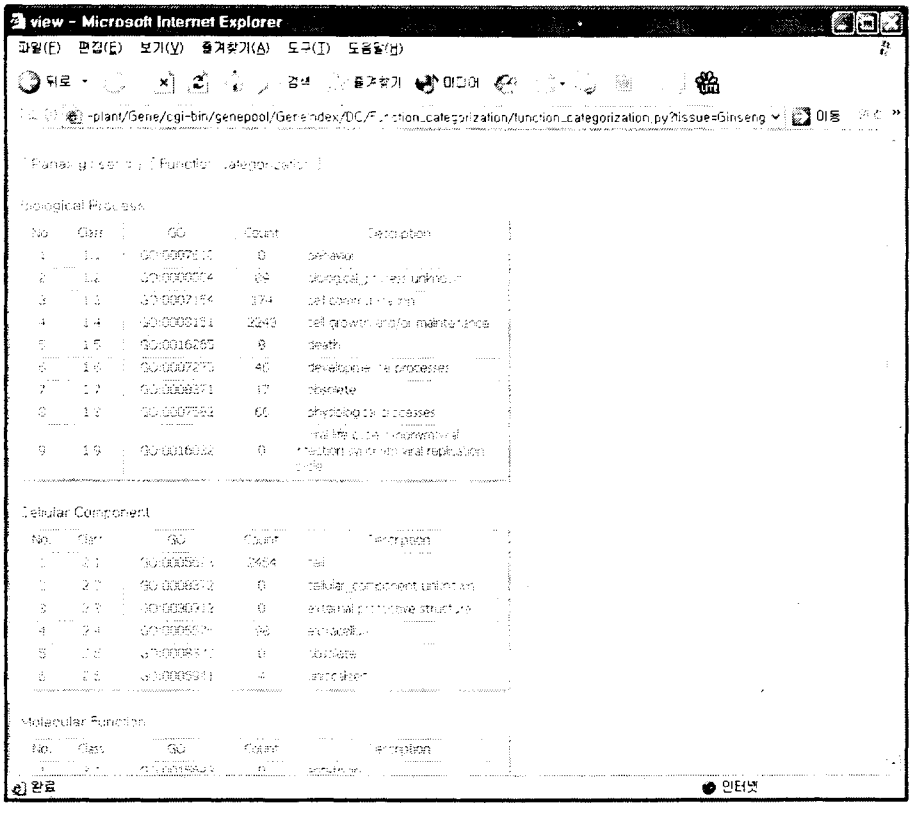


라) BLAST 분석 화면

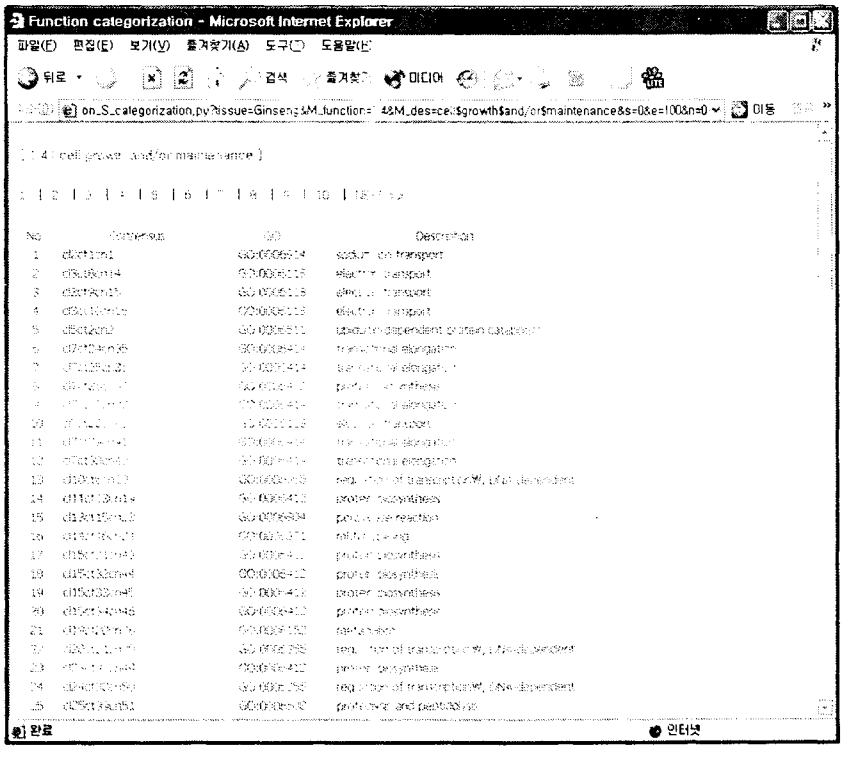

alignment result - Microsoft Internet Explorer
파일(F) 편집(E) 보기(V) 즐겨찾기(A) 도구(T) 도움말(H)
뒤로
검색
즐겨찾기
미디어
pdrc.re.kr/cgi-plant/Gene/cgi-bin/genespool/GeneIndex/DC/ByDC_total/view_dc_unigene.py?query_id=cllct7cn12
이동
BLASTN 2.1.2
Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs", Nucleic Acids Res. 25:3389-3402.
Query= cllct7cn12 COVERAGE:0.99; CRAWID: 1; TOTAL_ESTS:4;
ESTS:DC01001A01.ab1,DC01001AC2.ab1,DC01003003.ab1,DC01008C07.ab1;
LENGTH:730bp; MAP: ; ALT_CONSENSI:1
(730 letters)
Database: /dbs/Unigene/Plant/At.seq.unis
27,249 sequences; 37,405,837 total letters
Searching.....done
Sequences producing significant alignments:
Score (bits)
E Value
gnl|UG|At#S169631 Arabidopsis thaliana chromosome 3 CHR3v07142.. 42 0.006
gnl|UG|At#S167509 Arabidopsis thaliana chromosome 4 CHR4v07142.. 36 0.34
gnl|UG|At#S174564 Arabidopsis thaliana chromosome 2 CHR2v07142.. 34 1.3
gnl|UG|At#S171625 Arabidopsis thaliana chromosome 1 CHR1v07142.. 34 1.3
gnl|UG|At#S171402 Arabidopsis thaliana chromosome 1 CHR1v07142.. 34 1.3
>gnl|UG|At#S169631 Arabidopsis thaliana chromosome 3 CHR3v07142002
genomic sequence /cds=(222,1311) /gb=NM_115098
/gi=18409577 /ug=At.511 /len=1449
Length = 1449
Score = 42.1 bits (21), Expect = 0.006
Identities = 27/29 (93%)
Strand = Plus / Plus
Query: 328 atactgggaggtccagggaatttggttta 356
완료
인터넷

NCBI에서 제공한 UNIGENE 분석 화면

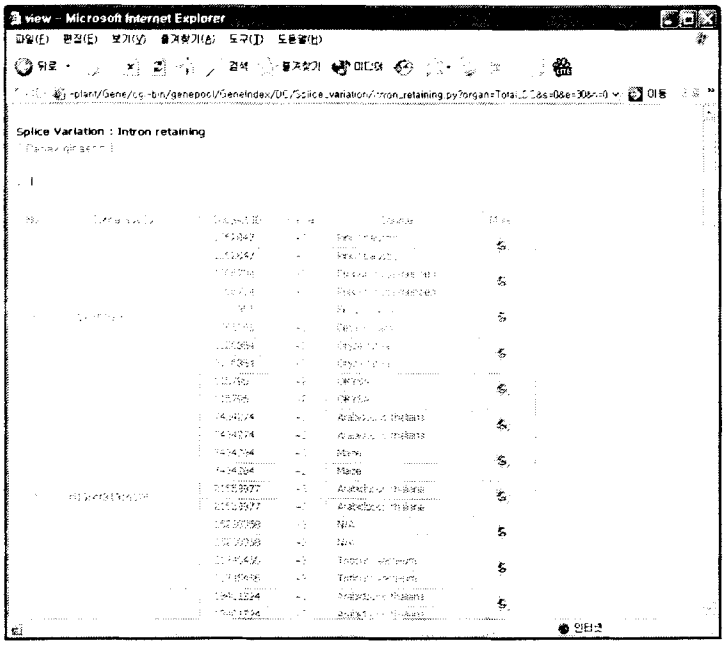
가) 기능 분석 (GO를 이용한 분석 내용)

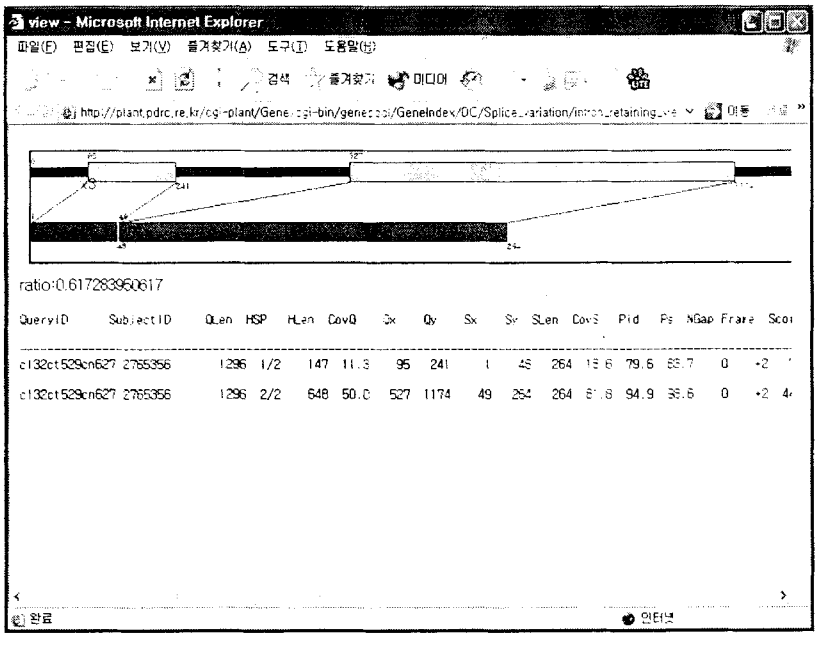


나) GO로 분석된 Cell growth 기능의 예제

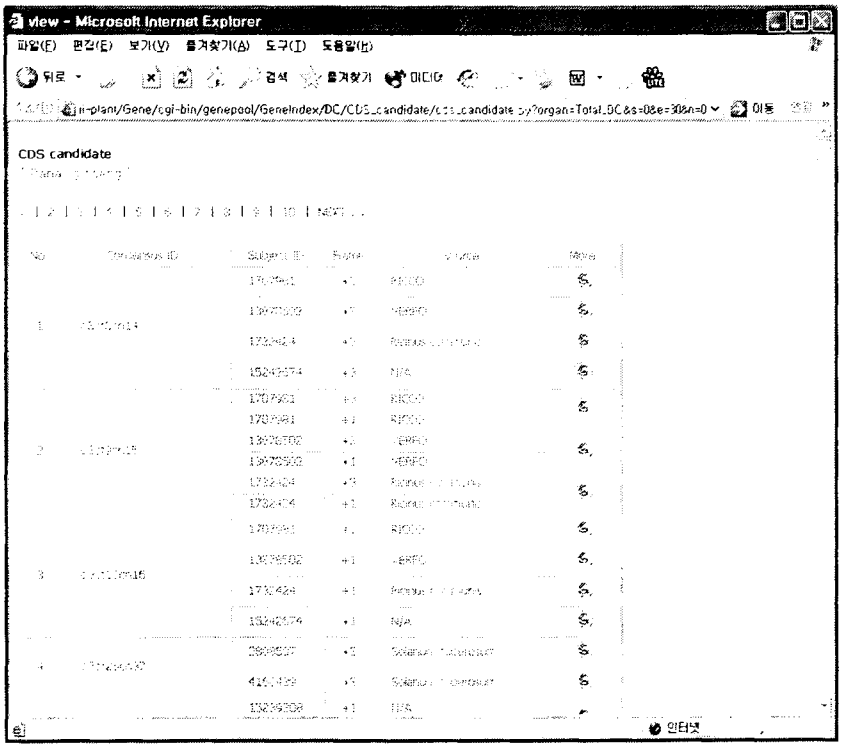


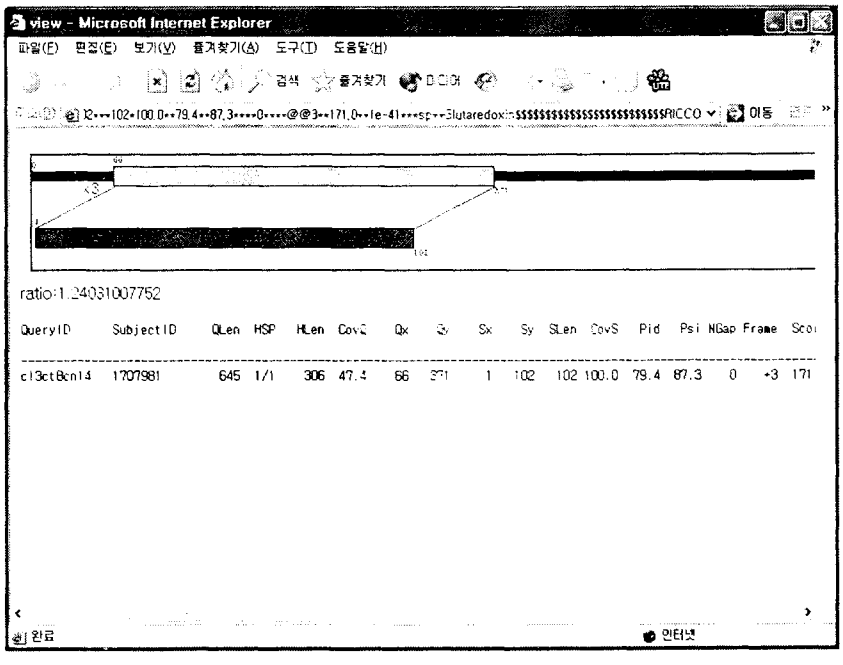
다) Splice variation 분석 결과



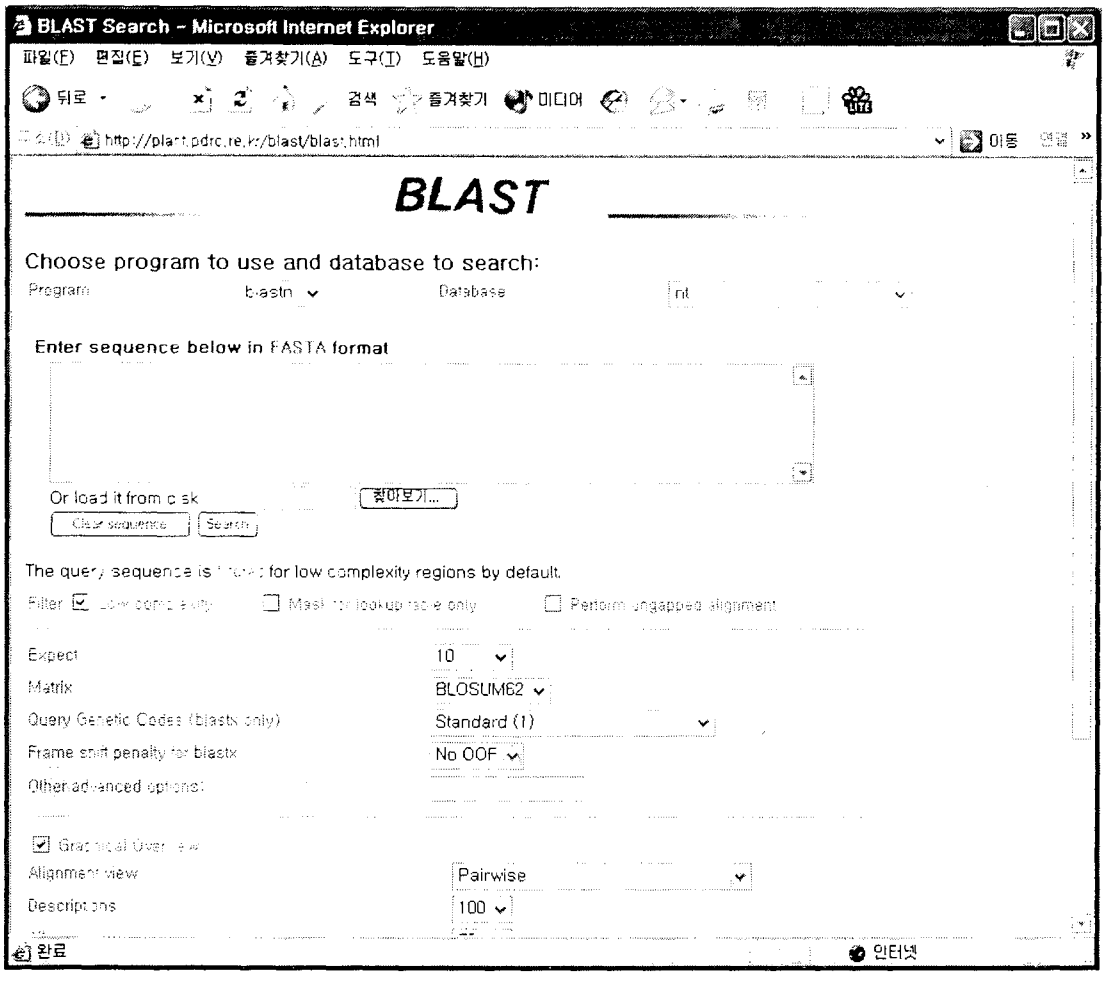


라) Coding region 분석 결과





마) BLAST를 이용한 서열 분석



나. Gene Index 작업 과정 (Plant)

1. Datasets 구성

Blast database의 est_other(ftp://ftp.ncbi.nih.gov/blast/db/)로부터 9개의 organism을 분류하고 난 후, Unigene cDNA library(5종)를 기준으로 하여 Library와 Tissue 별로 분류하였다. Unigene cDNA library의 tissue 개수를 줄이기 위해 same tissue라 생각되어지는 것들을 합하여 large tissue classification을 하였다. ESTs library Tissue classification과 tissue category는 Table1과 Table2에 보여주고 있다.

Unigene cDNA library가 없는 4종 중 Mt(http://www.medicago.org/)와Gm (http://129.186.26.94/soybean%20EST/libraries.html)은 다른 cDNA library를 기준으로 분류하였다.

Table 1. Construction of EST library tissue classification

Organism	No. of total EST	No. of library	No. of library tissue	No. of large tissue classification
Arabidopsis thaliana	172,477	93	28	8
Glycine max	256,445	68		8
Hordeum vulgare	236,771	102	58	10
Lycopersicon esculentum	147,317			5
Medicago truncatula	172,364	31		9
Oryza sativa	97,583	86	23	9
Solanum tuberosum	73,057			7
Triticum aestivum	171,377	76	44	9
Zea mays	43,607	62	32	9
Total	1,370,998			

Table 2. Tissue category

Tissue	At	Gm	Hv	Le	Mt	Os	St	Ta	Zm
Aboveground	15,769								
Callus			11,511			7,210		1,076	
Nodule					9,268				
Flower	23,267	25,754	21,196	17,259	X	15,688	X	70,385	1,072
Fruit		1,379	8,032	23,462	1,915	5	1	1	1,958
Leaf	2,678	25,486	6,985	430	6,003	20,124	16,320	14,411	1,722
Root	21,463	22,073	11,284	5,227	49,780	1,774	10,191	19,195	5
Seed	11,254	24,178	18,230	6	2,672	16,874	1	45,982	1
Seedling	3,206	77,731						7,691	8,720
Shoot	326		41			131		5,947	X
Stem		2,958	1,188			2,832	X		4
tuber							X		
Other	93,737	25,121	22,458		2,143	30,195		6,689	29,429
Unclassified	777	51765	135,846	100,933	100583	2,750	44544		696
Total	172,477	256,445	236,771	147,317	172,364	97,583	73,057	171,377	43,607

x : Tissue category에는 있었는데 분류되지 않은 Tissue

2. StackPack tool을 이용한 consensus sequence 생성

위와 같이 분류한 tissue 와 organism을 SANBI(http://www.sanbi.ac.za)에서

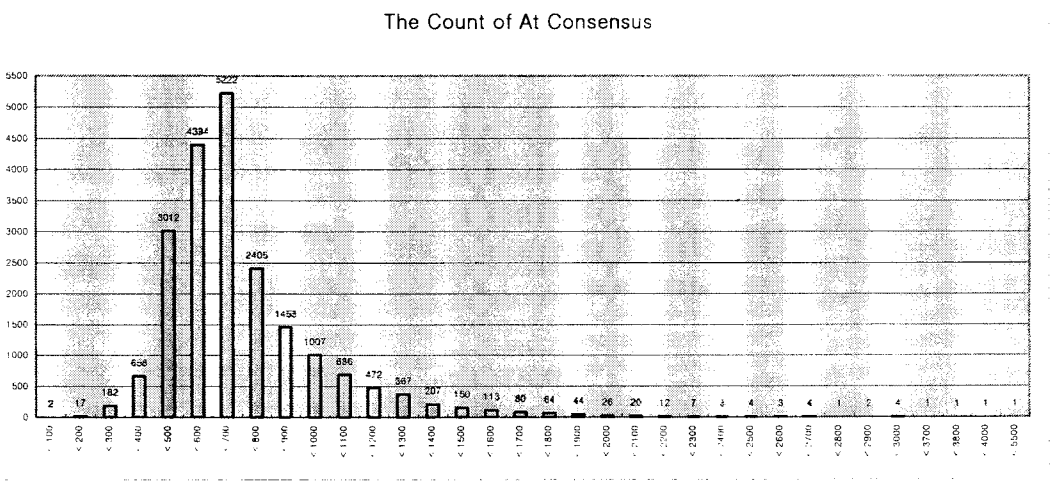
제공하는 StackPack(version 2.1.1) tool 사용하여 consensus sequence를 생성하였다.

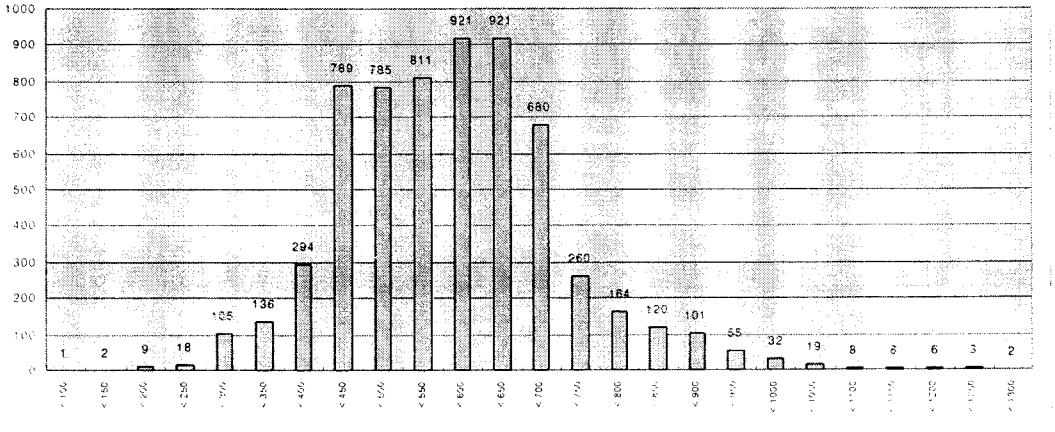
Masking(crossmatch) - cluster (d2 cluster) - assembly (phrap) - alignment analysis(craw)

3. Consensus sequence 고려

Gene을 cover할 수 있는 consensus sequence를 선별하기 위해 consensus sequence를 구성하는 est 개수와 consensus sequence length를 고려하기로 결정하였으며, 그들의 분포와 개수를 분석한 결과 length는 400bp이상으로 하고 EST 개수는 1개 이상 되는 consensus sequence를 분석대상으로 결정하였다.

Ex>





4. Consensus sequence의 분류

1번에서 분류된 tissue 와 organism을 shelve db로 만들어 놓은 후(accession number : organism name , accession : tissue name), 3번을 고려한 consensus sequence(file name : ooo.RNRF_P : P.list)의 ESTs가 same tissue cluster, same organism cluster, same mono/dicotyledon인 것만을 parsing하여 tissue unique, organism unique, mono/dicotyledon unique로 분류하였다.

>cl2ct3cn3 COVERAGE:0.99; CRAWID: 1; TOTAL_ESTS:3; ESTS:AA296711,BG941662,BG943218; LENGTH:393bp; MAP: ; ALT_CONSENSI:0
GAGTCTTTGGAGTAAGATGCCCTTGGAAGGGATGAGCAAACGGAAAGAGAAAGTGAAGTG
TCCAGGAGGGAGAGAATCCTGACGACGGCGTTCGCGGGAGTCCGCCGGAAGACTACAGGC
TTGGACAGGTCGCCAGTAGCTTATTTCGCGGCGAACACCATTCCAGAGGTGGCACCGGTC
GGCTGGCGTCCCTCTTCAGTTCTCTGGAGCCCCAGATTCAACCCGTGTACGTGCCTGTGC
CTAAACAAACCATCAAAAAACGAAACGGAATGAGGAGGAAGAAAGTACATCCCAGATTG
AAAGACCACTTTCGCAAGAACCTGCCAAAAAAGTGATAGCGAAGAAGAAACACACTAACG
CAGAAAAAAAGTTGGCAGACAGGGAAAGCGCTC

Table 4. Organism Unique

	Callus	Flower	Fruit	Leaf	Root	Seed	Seedling	Shoot	Stem	Total
Arabidopsis thaliana		3,774		50	3,265	1,446	399	21		8955
Glycine max		3,692	114	3,685	3,286	2,958	10,881		320	24936
Hordeum vulgare	1,871	2,212	1,318	845	1,452	1,966		2	162	9,828
Lycopersicon esculentum		3,025	3,634	16	581	1				7257
Medicago truncatula			256	890	7,607	346				9,099
Oryza sativa	927	2,388		3,197	326	1,383		9	333	8,563
Solanum tuberosum				2,974	1,598					4,572
Triticum aestivum	98	12,332		1,684	3,304	5,800	1,661	946		25,825
Zea mays		130	292	240			1,730			2,392
Total	2,896	27,553	5,614	13,581	21,419	13,900	14,671	978	815	
Monocotyledon	250	2,422	4	579	1,156	1,647	11	5		
Dicotyledon		9	5	164	94	33	5			

Total consensus 와 Unique consensus 비교

Status Report for Plant Tissue

Tissue	Total Seqs	Cluster	Seqs in Cluster	Analysed Data Consensus	Singletons	Full length clue
Aboveground	15,769	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	827
Leaf	96,159	12,053	68,360	14,332	32,799	1,819
Nodule	9,268	1,314	6,617	1,451	2,651	298
Root	140,992	19,883	103,034	23,538	37,958	3,897
Seed	119,198	12,684	84,230	15,590	34,964	1,817
Seedling	97,348	11,382	74,362	14,695	22,986	2,384
Shoot	6,445	900	3,419	983	3,026	150
Stem	6,982	771	2,294	815	4,688	68
Total	723,332	92,551	518,364	112,720	204,968	14,899

Table 3. Tissue Unique

Tissue	At	Gm	Hv	Le	Mt	Os	St	Ta	Zm	total
Callus			333			104		25		462
Aboveground	179									179
Nodule					34					34
Flower	505	245	541	408		250		5,679	41	7,669
Fruit		3	134	468	3				163	771
Leaf	29	421	125	3	2	98	578	556	115	1,927
Root	395	268	225	135	230	9	631	1,114	X	3,007
Seed	132	458	248		2	158		2,155	X	3,153
Seedling	42	1,461						331	582	2,416
shoot			X			1		244		246
Stem		44	1			13			X	58
Total	1,283	2,900	1,607	1,014	271	633	1,209	10,104	901	
	6.2%	12.4%	8.9%	6.5%	1.6%	5.2%	9.9%	41.8%	14.1%	

Total consensus 와 Unique consensus 비교

Status Report for Plant Organism

Species	Total Seqs	Cluster	Seqs in Cluster	Analysed Data: Consensus	Analysed Data: Singletons	Full length clue
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

5. protein database searches

4번에서 나온 consensus sequence를 HTC_Bio blast 프로그램(blastx)을 사용하여 NCBI의 nr database에서 paring한 Plant_NR 9종과 sequence similarity search를 하였으며 expectation cut off value는 조정하지 않고 alignment된 상위 5개만을 선택하여 보여주었다.

6. Unigene database searches

4번에서 나온 consensus sequence를 HTC_Bio blast(blastn) 프로그램을 사용하여 NCBI의 Unigene database (5종)와 sequence similarity search를 하였으며 expectation cut off value는 조정하지 않고 alignment된 상위 5개만을 선택하여

보여주었다.

7. Function categorization

4번에서 나온 consensus sequence를 HTC_Bio blast(blastx) 프로그램을 사용하여 At Function database와 sequence similarity search를 하였으며 expectation cut off value는 조정하지 않고 alignment된 상위 5개만을 선택하여 MuSeqBox (Blast output을 파싱하는 프로그램)프로그램을 실행하였다. MuSeqBox의 결과에서 e-value : 1e-10인 것들을 선택하여 function categorization의 재료로 사용하였다. 위의 재료를 NBI에서 재 분류한 MIPSCode(MIPS)와 OntologyCode(GO_TAIR)에 matching 시켜 function categorization을 만들었다.

* AraFunc.db
* MipsCode
* GO code

8. Signal pathway prediction of human EST consensus based on homology

EST consensus를 통해 novel gene를 찾고 mRNA의 발현 양상을 분석하는 작업이 가능하다. 그러나 궁극적으로 mRNA를 통한 protein의 기능과 그 protein들의 biological process를 밝혀내는 작업이 필요하다. NBI의 Gene Index는 EST consensus를 통해 protein의 기능과 biological process의 핵심적인 역할을 하는 signal pathway에 대한 예측이 가능하다.

Signal pathway 예측을 위해 Biocarta corporation (http://www.biocarta.com)에서 제공하는 255개의 biological pathway에 관여하는 gene들을 database로 구축하고 EST consensus들과 blastx 분석을 하였다. 그 결과 human bcell의 consensus의 하나인 c137ct688cn3897(Consensus ID)의 경우 muscle contraction에 관여하는 actin임을 알 수 있었으며, pathway prediction을 통해 actin이 death signal인 TNF/fas signal을 통해 caspase6, 7, 8등에 의해 cleavage된다는 것과 PI3K의 조절을 통해 organization되며, 세포의 migration에 관여함을 알 수 있다. 또한 actin이 bcell뿐만 아니라 brain, ovary, colon등과 같은 기관에서도 발현되는 정보를 얻을 수 있으며. 이는 c137ct688cn3897 consensus가 여러 기관에서 발현되는 것으로 유추할 수 있다. 마지막으로 pathway 예측은 CGAP의 정보와 link되어 actin이 brain cancer, colon cancer등에 관여함을 볼 수 있다.

9. Splice Variation

Plant_NR과 sequence similarity search한 blast결과를 MuSeqBox 프로그램을 사용하여 Potential alternative spliced transcript(retained intron, exon skipped)를 추출하였다. 이 데이터로부터 overlap되는 protein sequence(retained intron)와 consensus sequence (skipped exon)를 제거하였다.

10. CDS Candidate

Plant_NR과 sequence similarity search한 blsast결과 상위 5개를 MuSeqBox 프로그램을 사용하여 full length coding sequence를 추출하였다.

11. Keyword Search

Plant_NR과 sequence similarity search한 blast결과의 definition line을 분리하여 Oracle db에 넣고 사용자가 Keyword 검색을 통해 원하는 정보를 찾을 수 있게 하였다.

참고> db schema

12. Sequence Search

user sequence를 Public database 뿐 아니라 NBI의 Gene Index database를 선택하여 sequence similarity search를 할 수 있도록 제공하였다. Search된 sequence는 consensus sequence의 모든 정보와 link되어있다.

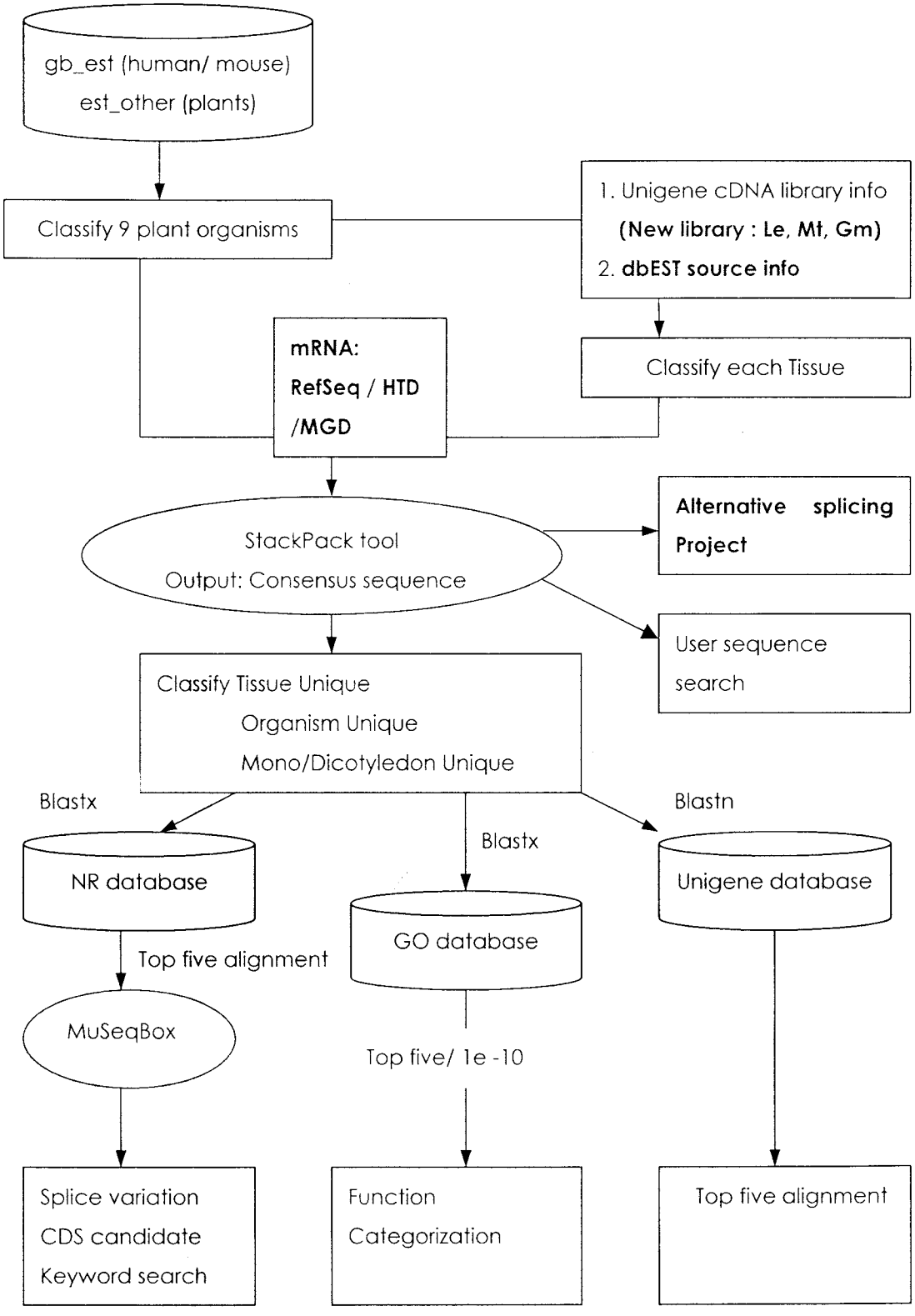
13. Sequence Download

각 organism의 tissue 별로 분류된 ESTs를 다운로드

Entry Number and Release of Database

Database		Release	Entry
est_other		6월	4,869,383
	Arabidopsis_thaliana		172,477
	Glycine_max		256,445
	Hordeum_vulgare		236,771
	Lycopersicon_esculentum		147,317
	Medicago_truncatula		172,364
	Oryza_sativa		97,583
	Solanum_tuberosum		73,057
	Triticum_aestivum		171,377
	Zea_mays		43,607
NR		9월	1,044,014
Plant_NR		9월	67,941
	NRArabidopsis.seq		43,361
	NROryza.seq		15,910
	NRZea.seq		2,438
	NRLycopersicon.seq		1,258
	NRGlycine.seq		1,155
	NRTriticum.seq		1,143
	NRHordeum.seq		1,117
	NRSolanum.seq		1,006
	NRMedicago.seq		553
AraFunc.db		4월	26,060
UniGene		9월	
	At.seq.uniq		27,586
	Hv.seq.uniq		7,681
	Os.seq.uniq		16,697
	Ta.seq.uniq		13,203
	Zm.seq.uniq		12,011
	Hs.seq.uniq		104,170
	Ms.seq.uniq		88,844

New version of Gene Index



나. PyFACT: A Tool for Function Assignment and Classification to a Sequence using Dictionary-Based Approach

1. Abstract

The length of the current lists of databases is dramatically increasing by the generation of huge amount of sequences from high-throughput experiment like single-pass EST sequencing. The ESTs are, of course, need to be annotated for using in further research and usually this is performed by sequence similarity search using BLAST. For its size and complexity, however, the BLAST output is not easy to handle and analyze. Therefore, there have been needs for a tool to summarize and categorize annotated sequences into functional categories automatically, and we developed PyFACT (Function Assignment & Classification Tool). PyFACT was developed using Python programming language and its XML parser, and it consists of three basic modules: first for dictionaries, second for function assignment, and the last for *Arabidopsis thaliana* gene-coded BLAST DB. PyFACT parse the XML output generated from BLAST searching on the BLAST DB and assign functions to each query sequence using our dictionary of gene-function category which is constructed based on *Arabidopsis* Functional Catalog in ' Munich Information center for Protein Sequences (MIPS)' . Added to these basic functions are two modules: one for displaying Gene Ontology (GO) codes related to each gene according to ' The Arabidopsis Information Resource (TAIR)' gene-GO mapping table, and the other for visualizing neighboring genes on a chromosome. PyFACT will be useful in large-scale sequences annotation.

Availability: The program is available over the web at http://www.ncgi.re.kr/PyFACT/

Contact: hurlee@kribb.re.kr

2. Introduction

Currently massive sequence data such as EST and cDNA are being generated by the advance of high-throughput sequencing techniques. They need to be annotated and this process is primarily done by BLAST search (Altschul *et. al.*, 1997). In the case of large-scale sequences analysis, researchers have difficulty in analyzing the results of BLAST search caused by the size and complexity of the outputs. To solve this problem we developed PyFACT, which can parse, filter analyze, and summarize BLAST results via a convenient and intuitive Graphic User Interface (GUI). It also provides functional catalogues of many sequences

using knowledge-based dictionaries. Furthermore, every output of PyFACT can be edited and saved, so users can use PyFACT as a workbench for sequence annotation.

가) Program Overview

PyFACT consists of three parts, *Arabidopsis thaliana* gene-coded BLAST database, dictionaries, and function assignment and classification module, and its overall architecture can be shown in Figure 1a. As shown in Figure 1a, PyFACT parses and filters the XML-format result of BLAST search against *Arabidopsis thaliana* gene-coded BLAST database, and analyze and summarize it with dictionaries. We describe each part more specifically in the following paragraphs.

나) Arabidopsis thaliana gene-coded BLAST database

The DB is the core part of PyFACT, and is searched by BLAST to find and assign the function category of a sequence. The *Arabidopsis thaliana* (*AT*) gene code in the title of a sequence, e.g. AT1G00001, is a staring point for function annotation, and using it, we can gather information on function category and GO annotation respectively from MAtDB (Schoof *et. al.*, 2002) and TAIR (Huala *et. al.*, 2001). The BLAST DB was constructed as follows:

1. Make the mapping table of accession and gene code by parsing GenBank files.
2. Extract *Arabidopsis thaliana* sequences from nr DB.
3. Substitute gene code for gi number in sequence title using the mapping table. (E.g >gi|AT1G35625|ref|NP_174799.1| (NM_103263) integral ...

다) Dictionaries

In PyFACT, a dictionary is a key-value data structure containing *AT* gene-related information extracted from large, complicated, and difficult-to-parse flat files of GenBank, MAtDB, and TAIR. The advantage of dictionaries is that they are simplified and customized for users, and easily expanded and replaced for further analysis. PyFACT has three types of dictionaries: first for gene-function category, second for gene-GO annotation, and the last for gene-position, and their data structures can be shown in Figure 1b.

라) Function Assignment and Classification

After the BLAST search, PyFACT selects *n* hit reference sequences over a user-defined threshold. Then, it extracts *AT* gene codes from reference sequence titles, and looks up the gene-function category dictionary using them and assigns matched

categories to the sequences. Statistics for all input sequences can be calculated and be displayed in a window. PyFACT summarizes the results of all sequences into 20 main categories and 110 subcategories according to the MAtDB functional catalogue, and draws a pie graph for one category (Figure 1c).

바) Other Features

In addition to MAtDB function category, PyFACT provides information on nr title, GO annotation on one line, and this gives a chance to compare all annotations at the same time. Second feature is that users can save, sort, edit, and filter BLAST results, and the third that users can go the NCBI website for further information related with GenBank accession number using a web browser. Finally, PyFACT shows all genes on a chromosome with annotations of GO and MAtDB.

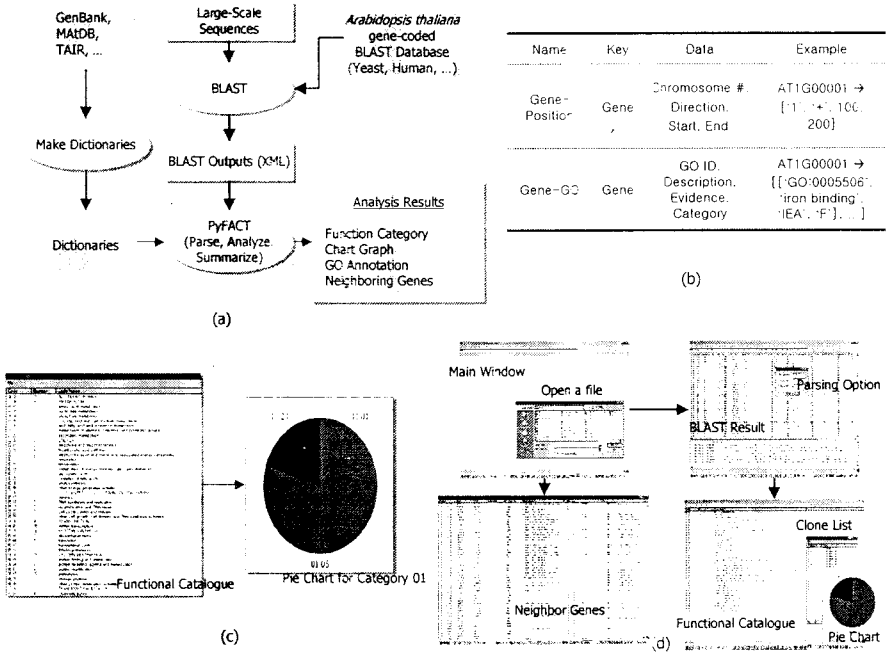


Fig. 1. (a) Overall architecture. (b) Dictionary type & structure, (c) Functional catalogue (d) Usage Example

바) Implementation

PyFACT was implemented using Python programming language, and wxPython was used for GUI. To parse the XML-format blast output, it uses XML Document Object Model (DOM) parser in the PyXML package, and uses the ReportLab package for chart graph drawing. PyFACT runs on Windows and it will be ported for Linux as

soon as possible.

3. Discussion & Future Works

PyFACT can be used in some areas such as clone selection for DNA microarray, function assignment on cluster of DNA microarray (co-operated with R-MAT, Kim *et. al.*, 2002), re-annotation & annotation transfer, and protein families analysis. This version of PyFACT assigns function category using knowledge-based dictionary, and this means it can't in the case of no information in the dictionary. Therefore, we will add algorithms for function category prediction using context later. In addition, we expand dictionaries for Arabidopsis thaliana to dictionaries for Human and Yeast for comparative genomics.

4. References

Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schä ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389-3402.

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 Res.*, 30, 91-93.

Huala, E., Dickerman, A., Garcia-Hernandez, M., Weems, D., Reiser, L., LaFond, F., Hanley, D., Kiphart, D., Zhuang, J., Huang, W., Mueller, L., Bhattacharyya, D., Bhaya, D., Sobral, B., Beavis, B., Somerville, C. and Rhee, S. Y. (2001) The Arabidopsis Information Resource (TAIR): A comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, 29, 102-105.

ReportLab http://www.reportlab.com/

Gene Ontology http://www.geneontology.org/

Kim, S., Kim J., Park, C. and Hur, C., R-MAT: Development of a GUI-based Microarray Data Analysis Tool Using R-Language. for the 10th International Conference on Intelligent Systems for Molecular Biology. (Poster)

다. An open source microarray data analysis system with GUI:Quintet

Abstract

We address Quintet, 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 equally important 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.

Correspondence: Cheol-Goo Hur. E-mail: hurlee@kribb.re.kr

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 equally important 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 module of 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 and users 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

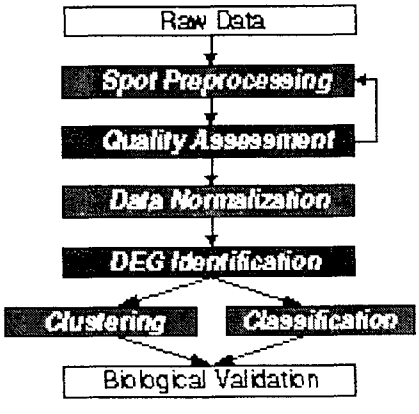


Figure 1. Simplified data analysis model.

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 much trouble since every scanning software has a mechanism to export slide data into text format files 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

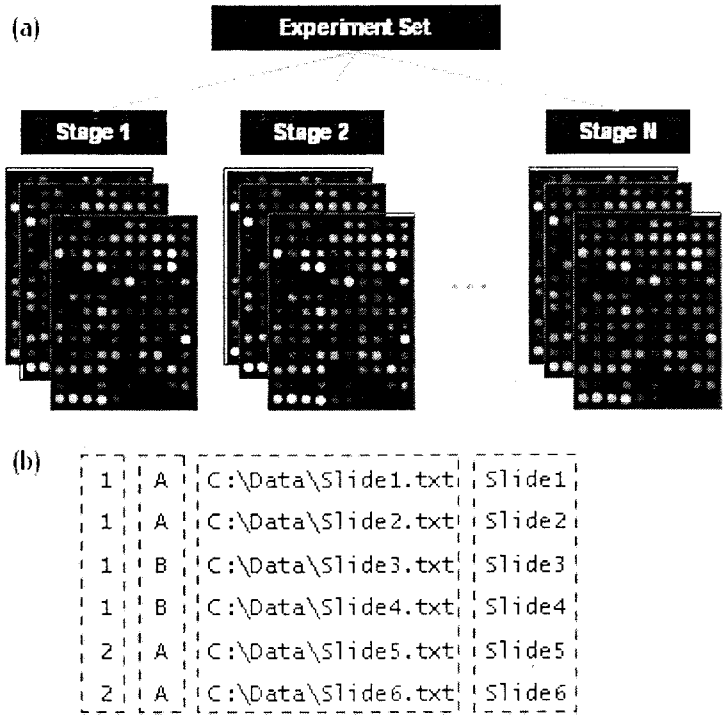


Figure 2. (a) Schematic experiment data model assumed in Quintet anc (b) part of a sample configuration file composed of 6 slides.

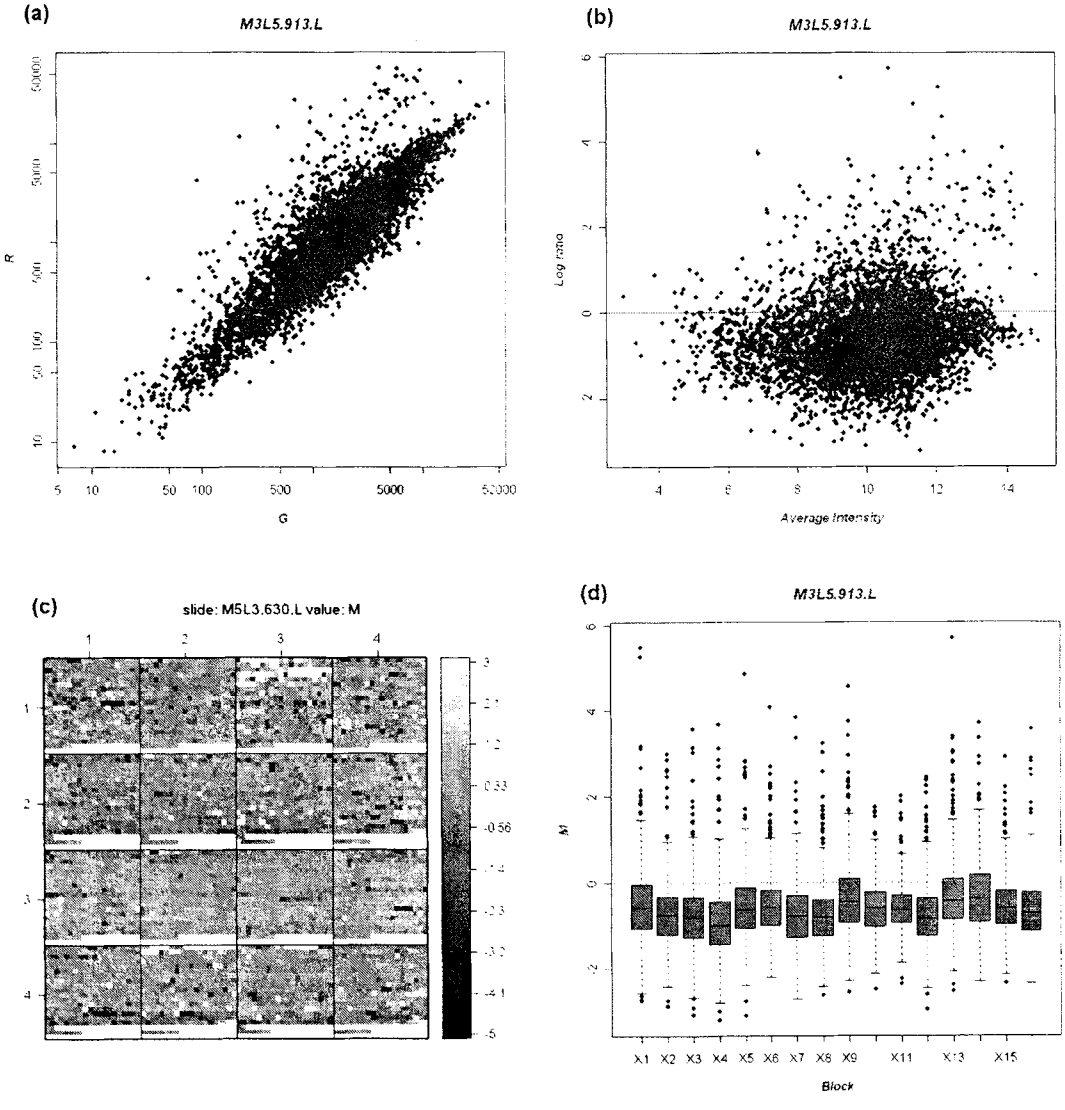
Schematic data organization assumed in Quintet is appeared in Figure 2(a). In general, a microarray experiment is composed of multiple stages. For example, each time point can be considered as a stage in the case of time-series experiments [7] and each individual condition can be considered as a stage 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 and collecting gene expression data on a comprehensive scale since data sharing can be drastically restricted without quality assurance [9, 10, 11].



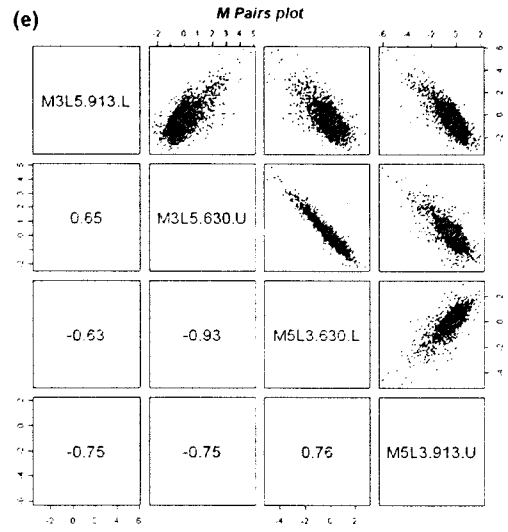


Figure 3. Sample plots used in QA module. (a) scatter plot of background-corrected green and red intensities, (b) RI (log ratio vs average intensity) plot, (c) 2D image plot of spot log ratio values, (d) block-by-block box plot of log ratio values of a slide, (e) pairs plot of log ratios with correlation coefficients among a group of slides.

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 $M=log_2R/G$ and average intensity by $A=(log_2R\times G)/2$. 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 if block-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 differential expression 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.

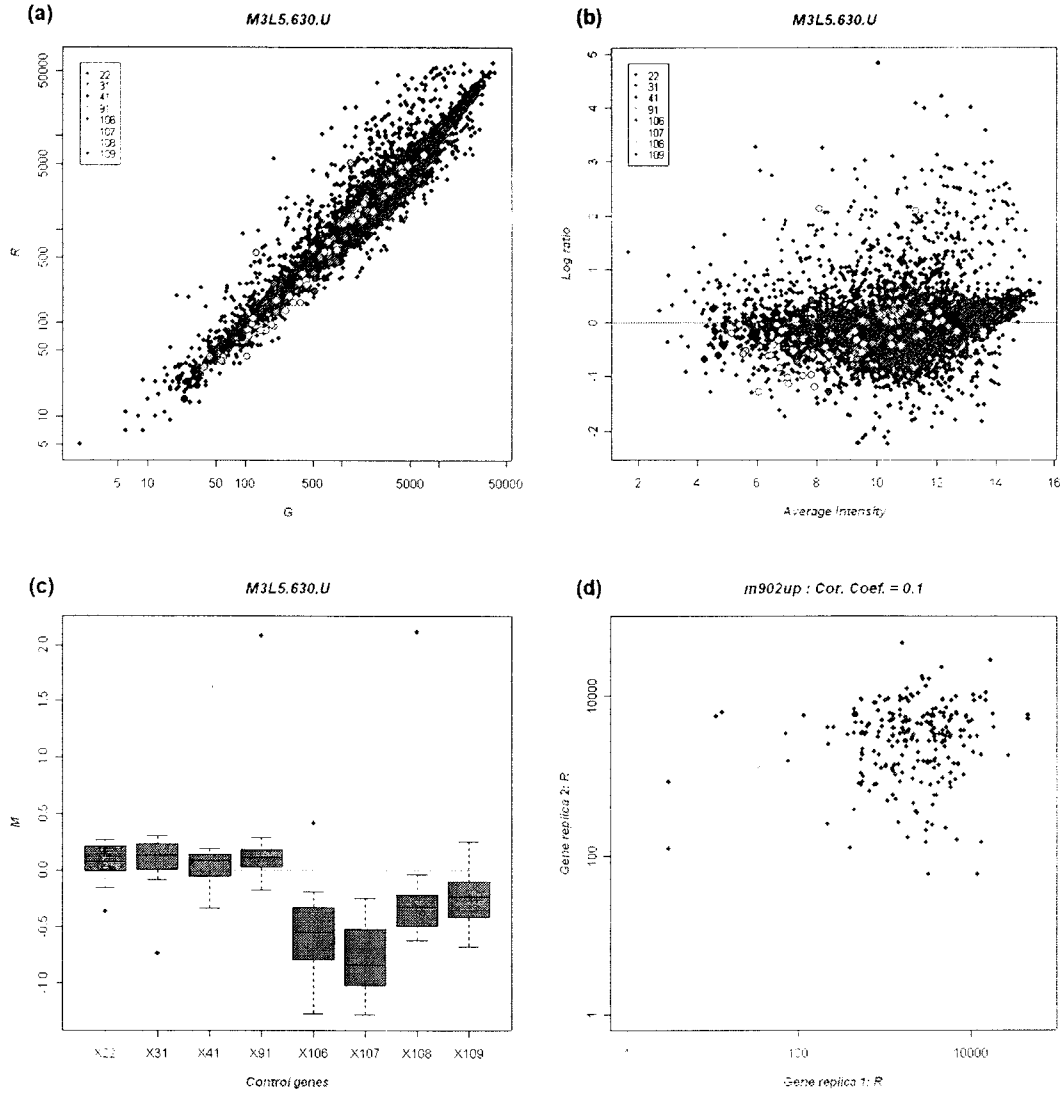


Figure 4. Sample plots for QA module using replicated genes. (a) red vs green intensity scatter plot for control genes (b) RI plot for control genes (c) box plot of log ratio values for control genes (d) self-against-self red intensity scatter plot of duplicated genes

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 comparison of 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 log_2R and log_2G. However, if only log ratio M values are normalized, large variations in average intensities will mask the true difference between log_2R and log_2G, 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 log_2R and log_2G values are restored from normalized A and M values.

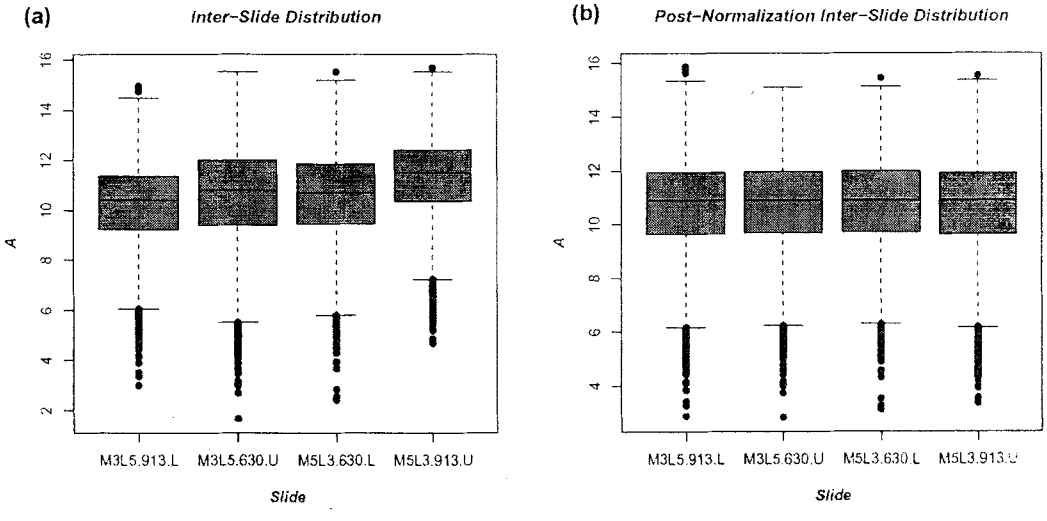
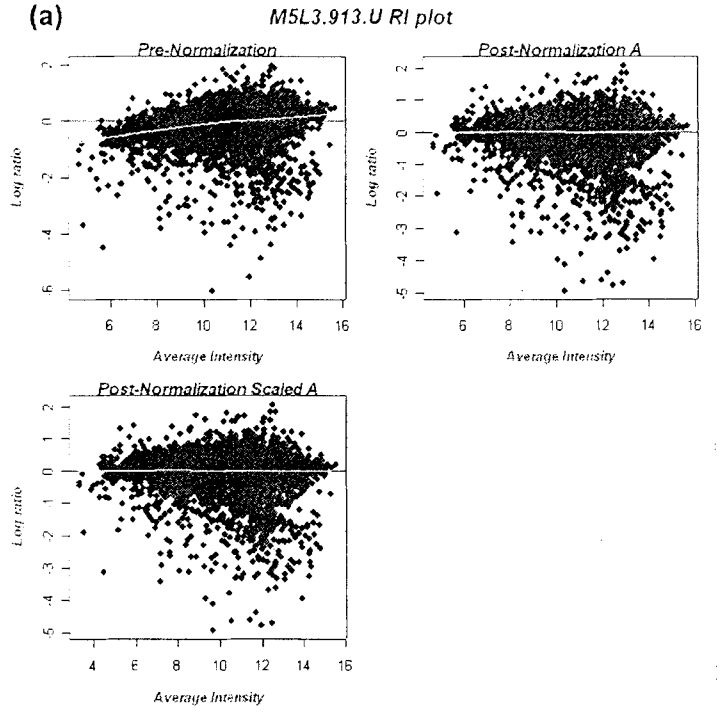


Figure 5. Box plot of average intensity A values across a sample experiment set before (a) and after (b) global A normalization.

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 will 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 to quality 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.



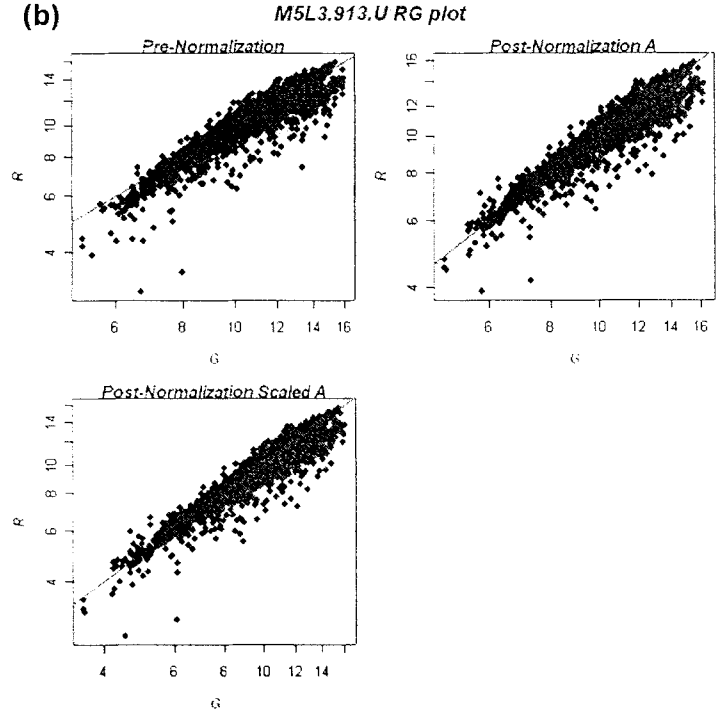


Figure 6. Normalization summary plots of a sample slide. (a) shows RI plots before any normalization, after log ratio normalization in the absence of global *A* normalization, and after log ratio and global *A* normalization while (b) shows corresponding red-vs-green scatter plots.

In Figure 6, results of normalization procedures are summarized. Figure 6(a) shows AM plots before normalization, after normalization without global *A* normalization and 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-slide algorithms 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 local z-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 residuals from 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 the AM 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 algorithms are 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 should be 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 available DEG 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.

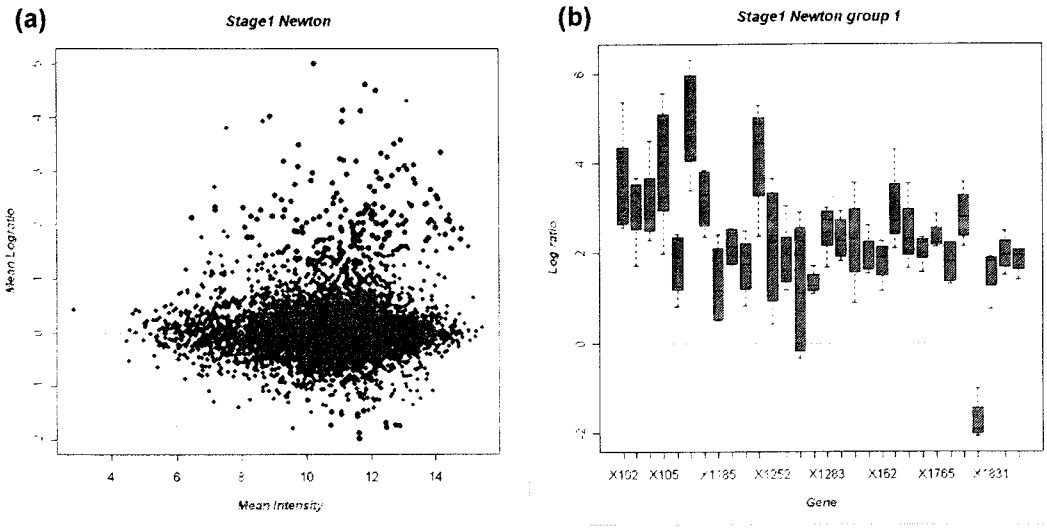


Figure 7. Supplementary plots for assessing DEG characteristics. (a) RI plot and (b) box plot of DEGs. In (a), DEGs are depicted in red.

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 the plots 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 [47]. 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*-regulatory elements 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 element to 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 different clusters. 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.

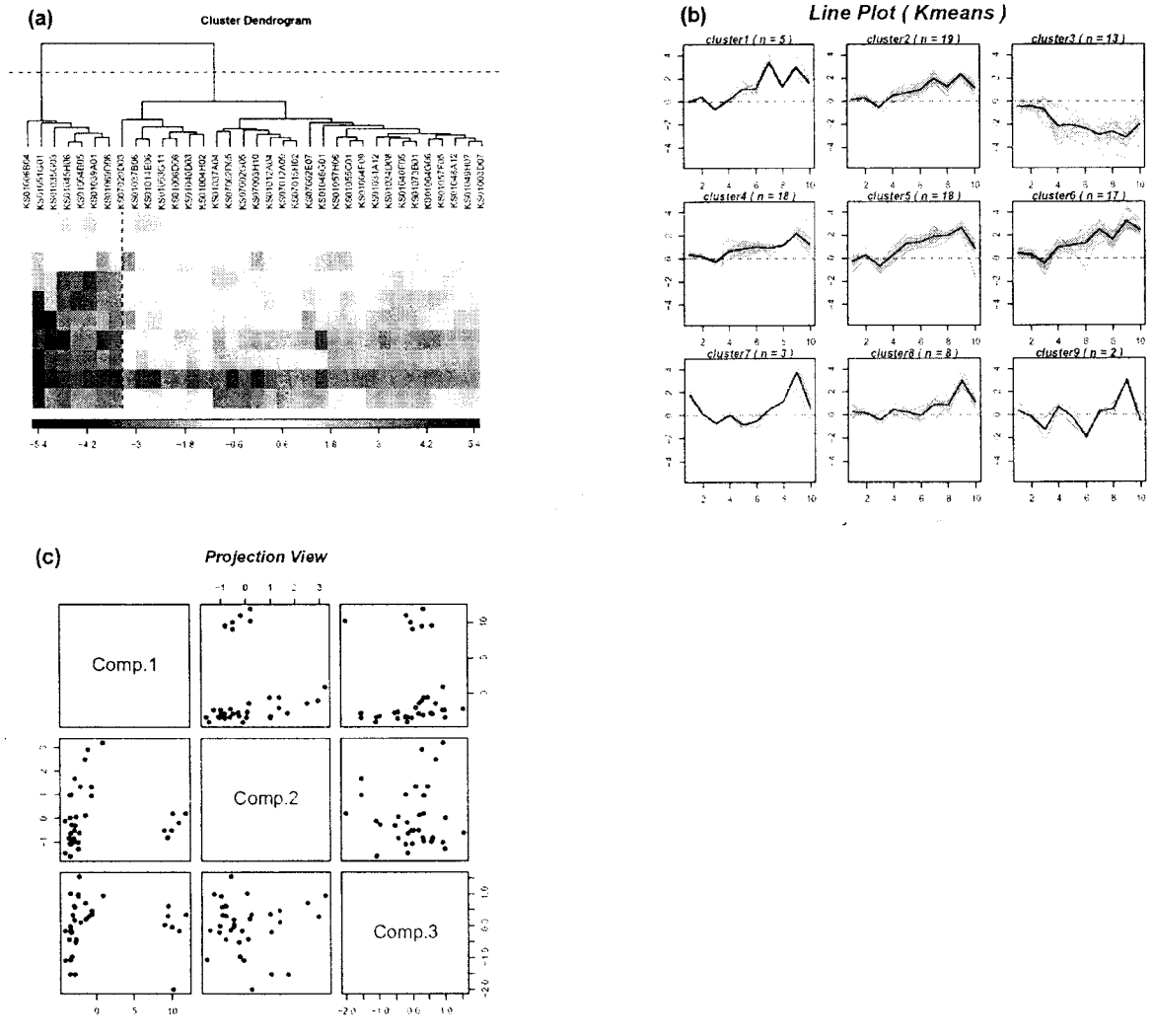


Figure 8. Typical cluster result presentation plots. (a) Dendrogram with 2D image plot

(b) Line plot showing gene expression patterns in each cluster (c) projection map using 3 principal components.

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 biological point 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 comprise the 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 functions and 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 the learned 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 the error 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 indispensable categories 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.

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

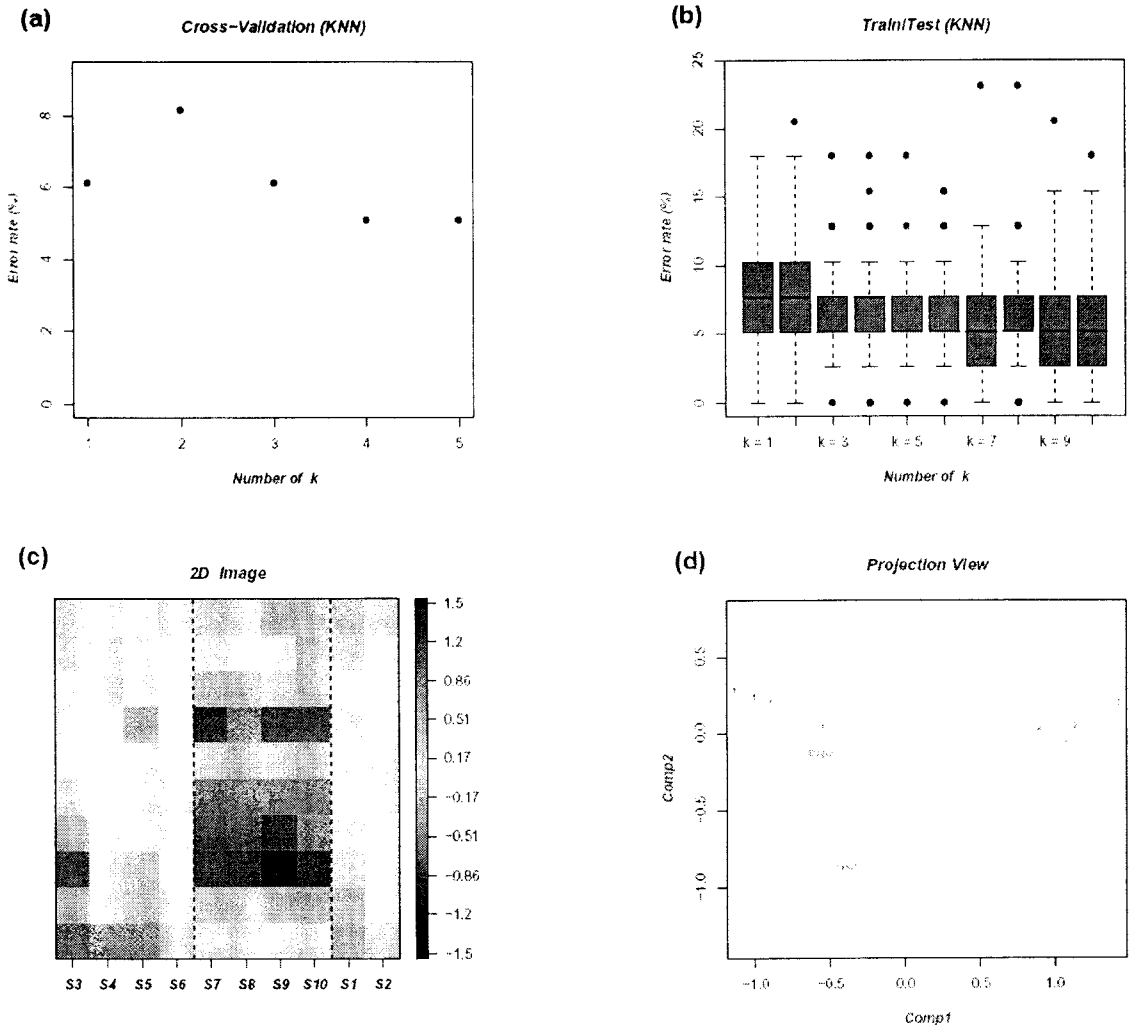

(a)
Cross-Validation (KNN)
Error rate (%)
Number of k
(b)
Train/Test (KNN)
Error rate (%)
k = 1
k = 3
k = 5
k = 7
k = 9
Number of k
(c)
2D Image
S3 S4 S5 S6 S7 S8 S9 S10 S1 S2
(d)
Projection View
Comp2
Comp1
(e)
Cross-Validation (Bar Plot)
Error rate (%)
KNN SVM NNET CART DLDA DQDA FLDA
Train/Test-Validation (Box Plot)
Error rate (%)
DLDA DQDA FLDA KNN SVM NNET CART

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 compoenets.

라. Computational analysis of neighboring genes on *Arabidopsis thaliana* chromosomes 4 and 5: Their genomic association as functional subunits

1. ABSTRACT

The genes related to specific events or pathways in bacteria are frequently localized proximate to the genome of their neighbors, as with the structures known as operon, but eukaryotic genes seem to be independent of their neighbors, and are dispersed randomly throughout genomes. Although cases are rare, the findings from structures similar to prokaryotic operons in the nematode genome, and the clustering of housekeeping genes on human genome, lead us to assess the genomic association of genes as functional subunits. We evaluated the genomic association of neighboring genes on chromosomes 4 and 5 of *Arabidopsis thaliana* with, and without, consideration of the scaffold/matrix-attached regions (S/MAR) loci, respectively. The observed number of functionally identical bigrams and trigrams were significantly higher than expected, and these results were verified statistically by calculating *p*-values for weighted random distributions. The observed frequency of functionally identical bigrams and trigrams were much higher in chromosome 4 than in chromosome 5, but the frequencies with, and without, consideration of the S/MAR in each chromosome were similar. In this study, a genomic association among functionally related neighboring genes in *Arabidopsis thaliana* was suggested.

Abbreviations : bigram, cluster of two genes; trigram, cluster of three genes

2. INTRODUCTION

The more the genomes of various organisms are revealed, from complete sequencing, the more insights we gain into the sequences themselves. These include: the organization of genomes, the structure of genes and regulatory elements, and the conservation of gene order in evolution (1). The genome rearrangement in living organism is a progressive form of evolution, where genomes are constantly rearranged and shuffled (2). In bacterial genomes, the strength of genomic associations correlates with the strength of the functional associations between the genes. Several reports have suggested that genomic associations reflect functional association between their proteins (1, 3-8). In addition, Snel *et al.* (2002) obtained a protein interaction network by combining the pairwise interactions between proteins, predicted from the conserved co-occurrence of their genes in operons (9). The genomes of higher-order eukaryotes, like animals, plants and fungi, seem to be relatively disorganized, with the average gene generally assumed to be independent of its neighbors, with only a few exceptions, such as repeats of similar sequences caused by gene duplications, and a limited number of ancient gene clusters containing functionally related genes (2). However, it has been revealed that neighboring genes are occasionally assembled into regulatory units, called operons, in the nematode (10). The estimated proportion of genes, expressed as a part of operon, in *Caenorhabditis elegans* was 13-15% (10). In addition, correlation between transcriptome and protein-protein interactions was mapped for *Saccharomyces cerevisiae*, with genes from the same functional cluster showing a higher protein interaction density (11) In this respect, it is plausible that genes with similar transcription profiles may have a tendency to cluster in eukaryotic genomes (12,13), and it is suggested that functionally related proteins, encoded by neighboring genes either physically interact or are involved in a certain biological event. Although eukaryotic genes are not exactly the same as bacterial operons, it would be advantageous for the sets of genes involved in a certain biological process, to be localized as neighbors on the genome, with some conservation of gene order (13), where their expression might be regulated as a functional module.

Eukaryotic chromosomes at the interphase do not exist as condensed structures, but their relaxed chromatin is attached on the scaffold/matrix of the nucleus, and the looped structure can be dealt with as a functional domain of the chromosome or genome (14). Efforts to reveal the relationship

between gene regulatory mechanisms and the nuclear architecture have proved increased evidence (15), and the scaffold/matrix-attached region (S/MAR) has been suggested as one of the abundant regulatory DNA elements of the eukaryotic genome (16). S/MAR form the anchor points of loop domains, with domain sizes ranging from a few kilo bases, to more than one hundred (17), harbor one or more genes. However, there is no information on either the average gene number, or the functional relatedness between neighboring genes in a loop.

S/MARt DB deposits several hundred S/MAR containing sequences, extracted from original publications (14), and several bioinformatics methods form *in silico* S/MAR prediction have been developed, such as SMARTest (16) and MAR-Finder (18). These tools use several motifs in their library, including origin of replication, TG-rich sequences, curved DNA, linked DNA, topoisomerase II sites, and AT-rich sequences. These motifs, however, do not always appear on every known S/MAR containing sequences. Previously reported S/MAR consensus patterns were recently compared for their enrichment, and their MAR/SAR recognition signature (MRS) (19,20) verified as the most enriched motifs in the S/MAR containing sequences (21).

In the present study, we collected neighboring gene sets, with and without considering the S/MAR from chromosomes 4 and 5 of *Arabidopsis thaliana*, then analyzed the relation between their genomic and functional associations. The effects of S/MAR on the association of genes, with identical function sub-categories, were not confirmed, but it was suggested that genes in the same functional sub-category were assembled together, and with statistical significance.

3. MATERIALS AND METHODS

가) Data sources of Arabidopsis thaliana genome sequences and function annotation

Among five chromosomes of *Arabidopsis thaliana* we selected the chromosome 4 and 5 as the subject for analysis, as they are richer in annotation than the other three. The complete sequences of the two chromosomes were retrieved from the GenBank (Accession No. NC_003075.1 and NC_003076.2), and the function annotation information was retrieved from the Munich Information of Protein Sequences website (MIPS; http://mips.gsf.de/cgi-bin/proj/thal/search_funcat). From the GenBank flatfile, the features describing the sequence position and Arabidopsis Genome Initiative (AGI) code were extracted, and the AGI code linked to the annotation information and function category code from MIPS, such as enzyme category (EC) code. We used the function category codes, which subdivided 109 sub-categories from nineteen larger primary categories, including one additional customized category ' not found in the MAtDB,' which was assigned the code ' 00' . This information was saved in the form of a dictionary using an in-house program.

나) Collecting Bigrams and Trigrams

Pairs of two consecutive genes were collected from chromosomes 4 and 5 of *Arabidopsis thaliana*, and we refer to them as bigram in this study. Using similar strategy, three consecutive genes were collected, which we called trigrams. All the bigrams and trigrams were extracted from each frame according to their starting point. For example, bigrams from frame 1 were composed of the first-second, third-fourth, and so on, but those from frame 2 were composed of the second-third, fourth-fifth, and so on. We mapped function category code for the genes in the bigram and trigram using AGI code-function category code linking dictionary, then sorted them according to the number of function codes, and collected the ones not containing any' 98: Classification not yet clear-cut' , ' 99: Unclassified proteins' , or ' 00: Not found in MAtDB' . The bigrams and trigrams having an identical function sub-category were counted.

다) Prediction of scaffold/matrix attached region (S/MAR) loci

For the extraction of S/MAR motif weight matrices, we used S/MARt DB Professional 2.1 (Biobase GmbH, Germany; Release date: Jan. 21, 2002). We collected 55 entries corresponding to dicotyledonous plants, including 13 from *Arabidopsis thaliana*, from a total of 377 entries, and made them the subject for extracting the pattern of the 16-bp AWWRTAANNWWGNNNC and 8-bp

AATAAYAA sequences, which were reported as the MAR/SAR recognition signature, (MRS)-1 and MRS-2, respectively (20). The MRS-1 and MRS-2 matching sequences were extracted using the MATCH™ program of TRANSFAC Professional 6.2 (Biobase GmbH, Germany), with a default motif core similarity of 75%. Their weight matrices were generated, using the MATCH™ Profiler program, from 79 MRS-1 and 19 MRS-2 matching sequences. The MRS were used as they give the advantage of increasing the chance of uncovering the S/MAR data that would otherwise be unavailable, and the enrichment of the MRS is higher than any other S/MAR motifs in experimentally confirmed sequences (21). We selected the flanking sequences between non-overlapping genes as the targets for S/MAR prediction, but did not consider the S/MAR inside the coding sequences. These flanking sequences were extracted using positional information from the feature part of the GenBank flatfile. From these flanking sequences, both the MRS-1 and MRS-2 residing within 200 bp were colleted, without considering their orientation, using the MATCH™ program with a cutoff value of FN50 (MRS-1: 93%, MRS-2: 98.75%)

라) Collecting Bigrams and Trigrams considering S/MAR

For each predicted S/MAR, containing flanking sequences, the bigrams and trigrams at positions just before and after S/MAR were collected. In addition, bigrams that predicted where the S/MAR resides were also collected. We mapped the function category codes on these bigrams and trigrams, and then counted those that were functionally identical.

마) Statistical significance of bigrams and trigrams

To assess the statistical significance, *p*-values were calculated for the functionally identical bigrams. According to the Ge *et al.* (2001) (11), *p*-values for protein-protein interactions in *Saccaromyces cerevisiae* were evaluated assuming each pair has the same probability - i.e. a uniform random distribution. We applied this concept with some modification, due to the bias in some of the distribution of function categories in the *Arabidopsis thaliana* genome. We calculated of *p*-values, excluding unknown function categories, but considered sets of genes with known function categories, which we called *K*.

The algorithm for the *p*-values of the weighted random distribution is as follows:

1. Estimate probabilities for each function sub-category of *K*.
   A. Count all the bigrams (or trigrams) for which both (or all) gene functions are known, *N*, and count the respective frequencies for

each sub-category.

B. Divide each frequency by the total number of available bigrams (or trigrams). The calculated results are the estimated probabilities

2. Sum all the probabilities obtained to give the total probability of all the functionally identical bigrams (or trigrams), p.
3. Use a normal approximation to a binomial distribution to calculate p-values:

   Let I be the binomial random variable, with parameters p and i_0 being the true number of identical bigrams (or trigrams) in the data. The corresponding p-value is then given by the formula:

$$p = P(I > i_0) = \sum_{i=i_0+1}^{N} {}_N C_i \, p^i (1-p)^{N-i}$$

   With respect to p, the expected number of functionally identical bigrams (or trigrams) is pN, and I is approximately normally distributed, $N(pN, p(1-p)N)$. Hence,

$$p \approx P\left(Z > \frac{i_0 - pN}{\sqrt{p(1-p)N}} \right)$$

   where Z is a standard normal variable.

To show the distribution of the functional combination among neighboring genes, bigrams were assigned on the matrix according to their nineteen large function category codes. The frequencies of bigrams were normalized by scaling them down to 1000 pairs, and figured out as a twelve-color scale.

4. RESULTS

가) Bigrams without considering S/MAR

In chromosomes 4 and 5 of *Arabidopsis thalinana*, there are 3744 (22) and 5874 (23) non-overlapping genes, respectively. We collected bigrams from two frames according to the starting position. According to the MAtDB, however, some genes had more than one function category code assigned. These might come either from the ambiguity of the gene function making annotators difficult to define exactly, or from the multifunctional feature of gene products. Therefore, there is some increase in the number of bigrams and trigrams caused by combination between redundant functions, but these additional function assignments should not be ignored. We collected and accepted all the additional combined bigrams. The number of bigrams was 2178 for each frame in chromosome 4, and 3208 and 3213 for frame1 and 2 of chromosome 5 (Table 1). The average proportion of bigrams, excluding the 00, 98, and 99 function categories were 17.21 and 10.87% for chromosome 4 and 5, respectively (Table 1). From these, functionally identical bigrams, that is, bigrams composed of an identical function sub-category, were counted, and the proportions were 4.59 and 4.68% for frames 1 and 2 of chromosome 4, and 2.46 and 2.33% for frames 1 and 2 of chromosome 5, respectively (Table 1).

We evaluated the *p*-values of functionally identical bigrams to assess the statistical significance for a weighted random distribution. The *p*-values for frames 1 and 2 for chromosome 4 were 1.8279×10^{-95} and 1.219293×10^{-93}, and for chromosome 5 were 1.7615×10^{-47} and 1.3772×10^{-41}, respectively (Table 1). These *p*-values suggested that the probability of a genomic association of functionally identical bigrams, due to chance, was extremely low. The observed number of functionally identical bigrams was significantly higher than expected, even when the weighted random distribution was considered (Table 1). The observed number was higher in chromosome 4, although the chromosome 5 also had a higher than expected number.

We mapped all the bigrams on the diagonal matrices according to their nineteen large function categories in order to display their global genomic association (Figure 2). The diagonal pairs showed higher frequencies of bigrams, and the pairs on categories '01-metabolism' and '04-transcription' showed relatively higher frequencies than the other pairs. The metabolism and transcription categories are the first and second largest groups in both chromosomes 4 (01: 9.6%, 04: 5.6%) (22) 5 (01: 21.1%, 04: 18.6%) (23). Thus, it is plausible that those associated pairs would appear

more frequently. However, the diagonal pairs, i.e., composed of the same function category, appeared more frequently regardless of their function category and the proportion of the function category in each chromosome. This coincided with higher probability of co-localizations of genes, composed of identical function sub-categories, as functionally identical bigrams. The matrices of chromosome 4 showed clearer, denser pairs on the diagonal than those of chromosome 5, and the pairs related to the metabolism-01 and transcription-04 function categories showed a higher frequency than the other pairs. This occurred because there were more fully annotated genes in chromosome 4 than in chromosome 5, and there was bias in the proportion of function categories of annotated genes.

나) Prediction flanking sequences that containing S/MAR locus and collection of bigrams

To assess the effect of S/MAR on the co-localization of genes with an identical function sub-category, we predicted the S/MAR loci on chromosomes 4 and 5, and surveyed the bigrams on both sides of S/MAR. We collected the flanking sequences of the discrete non-overlapping genes, then assessed their S/MAR retention. The MATCH™ Profiler program generated five criteria for MRS-1 and MRS-2, and the cutoff value FN50 was selected following tests on the previously reported sequences. The sequences used for these tests were the plastocyanin (z83321), *ATB2* (z82043) and *ATH1* (z83320) genes of *Arabidopsis thaliana*, and they experimentally confirmed for their S/MAR retention (20). Using the FN50 criteria, the MATCH™ program correctly predicted all the experimentally confirmed S/MAR loci in the test sequences. The counts of flanking sequences containing S/MAR loci were 1119 and 1678 for chromosomes 4 and 5, respectively (Table 2). From this result, the densities of the S/MAR loci were calculated as one S/MAR locus per 15.5 kb for both chromosomes. This means two or three genes reside, on average, between two S/MAR loci, as the gene density is one per 4.6 kb and 4.4 kb for chromosomes 4 and 5, respectively (22, 23).

As a pivot, the S/MAR containing sequences give bigrams in both directions, so we collected the bigrams separately, before and after of the S/MAR containing flanking sequences. Additionally, we collected bigrams where the S/MAR resided in the middle of two genes. The proportions of functionally identical bigrams for chromosome 4 (Table 2) were higher (4.87~5.67%) than for chromosome 5 (2.06~2.41%). In chromosome 4, these proportions were similar, but slightly higher than in the case the S/MAR was

not considered, especially in the class of ' Across S/MAR' , suggesting S/MAR has some role in associating genes belonging to the same function category on the genome. In chromosome 5, however, the proportions were similar, but slightly lower than the cases that not consider the S/MAR, especially in the case of ' After S/MAR' . The *p*-values for functionally identical bigrams before, across and after the S/MAR on chromosome 4 (Table 2) were 7.3690×10^{-64} 3.8055×10^{-83} and 1.1369×10^{-60}, respectively, and for chromosome 5 were 1.0183×10^{-27}, 6.8648×10^{-27} and 2.2902×10^{-21}, respectively. Although these *p*-values were much higher than those cases where the non-S/MAR were considered, it was difficult to determine if the S/MAR affects the genomic association of the bigrams, because *p*-values were all extremely low. Nevertheless, these *p*-values suggested there was little probability of the appearance due to chance in either case, and the observed number of functionally identical bigrams was significantly higher than expected considering the weighted random distribution. The matrices of these classes (Figure 3) showed similar patterns to the cases where the S/MAR were not considered, and the diagonal pairs on chromosome 4 were denser than those on chromosome 5, as when the cases of the S/MAR was not considered.

다) Trigrams without considering S/MAR

As previously mentioned, the average interval of S/MAR loci in chromosomes 4 and 5 was 15.5 kb, and an average of two or three genes could reside in this interval. Thus, we extended the neighboring gene numbers to three, and assessed the association of three consecutive genes in their function. We divided cases into two classes, those where S/MAR were not considered and those where they were, for the analyses of trigrams.

For the cases where the S/MAR was not considered, we collected trigrams from three frames according to the start point. There were around 1500 trigrams for each frame in chromosome 4 and around 2190 trigrams in chromosome 5 (Table 3). The proportions of trigrams without the 00/98/99 categories, on average were 8.73 and 5.71% for chromosomes 4 and 5, respectively. This was about the half level of the bigrams because there were more chances of the 00, 98 and 99 function categories being neglected. The frequencies of functionally identical trigrams, on average were 1.55 and 0.59% for chromosomes 4 and 5. The *p*-values for the weighted random distributions were 1.2700×10^{-224}, 8.4649×10^{-264} and 9.5589×10^{-252} for frames 1, 2 and 3 of chromosome 4, and 9.9222×10^{-76}, 8.9400×10^{-105} and 1.2125×10^{-58} for chromosome 5, respectively. These *p*-values for both chromosomes suggested

the probability of a genomic association of functionally identical trigram, due to chance, is extremely low, and the observed number of functionally identical trigrams was significantly higher, statistically, than expected assuming the same weighted random distribution as with the bigrams.

라) Trigrams considering S/MAR

Using the same predicted S/MAR loci information as for the analyses of the bigrams, we collected trigrams before and after S/MAR from each chromosome. The frequencies of the trigrams with identical function sub-category were 1.79 and 1.48% for the trigrams before and after the S/MAR position in chromosome 4, and 0.66 and 0.38% for chromosome 5 (Table 4), respectively, which were similar to those cases when S/MAR were not considered. The *p*-values were 6.4265×10^{-254} and 2.4439×10^{-193} before and after S/MAR on chromosome 4, and 2.0943×10^{-88} and 3.0978×10^{-29} on chromosome 5. This indicated the probability of a genomic association of functionally identical trigrams due to chance to be extremely low, and the observed number of functionally identical trigrams was significantly higher, statistically, than expected assuming the same weighted random distribution as for the bigrams. With these results, however, it was not possible to suggest any correlation between the genomic association of genes belonging to an identical function sub-category and S/MAR locus, because of the little difference in the frequency of functionally identical trigrams and *p*-values between cases when the S/MAR considered or not.

5. DISCUSSION

The features of genes in eukaryotic genome are being revealed through the sequencing efforts, successive analyses by functional genomics and from *in silico* analysis. Operon-like structures of neighboring genes have been found in *Caenorhabditis elegans* (10), which suggests that similar organization could appear in the genome of other eukaryotic species. If those functionally related genes are assembled in a boundary on the genome, the regulation of their concerted expression at a higher level can be accomplished more easily, and the clustering of housekeeping genes of the human genome (13) can support this postulation. In addition, if there is any correlation between functions of the neighboring gene products, it could be used to predict both physical interactions between proteins, and protein function as the conservation of gene order in bacterial genomes is routinely used for the prediction of physical interactions of proteins, and the prediction of unknown function of neighboring gene (1). However, investigation on this theme, have not been widely addressed on eukaryotic genomes.

In the present study, we described the association of genes belonging to identical function sub-categories on chromosomes 4 and 5 of *Arabidopsis thaliana*. We initiated this study by focusing on two consecutive gene sets, because the gene sets composed of two consecutive genes are the smallest of neighbored gene pairs, which we defined as ' bigrams' in this study. The collections of bigrams were divided into two cases according to the consideration of the S/MAR. The reason the S/MAR was considered for the collection of bigrams was as a result of the looped structure of interphase chromatin, caused by attachment of S/MAR on the nuclear matrix, can be dealt with a functional subunit, and therefore the S/MAR are thought to be the tools that subdivide eukaryotic genomes into structural and functional domains (14). Thus, before collecting the bigrams we predicted the S/MAR loci of the whole sequences on chromosomes 4 and 5 of *Arabidopsis thaliana*. The S/MARt DB, the database for S/MAR, contains information fully extracted from original publications. However, we could not find all the possible S/MAR loci of *Arabidopsis thaliana* from this database, due to the number of entries for plants only being 55, including 13 entries for *Arabidopsis thaliana*. Therefore, we predicted S/MAR loci from an *in silico* method. There are a couple of prediction tools publicly available such as MAR-Finder and SMARTest. They use several S/MAR related motifs in their predictions,

but these features do not always appeared on every known S/MAR containing locus, and they are not adjusted to our subject, thus we had to devise another method. We used two MRS that had been reported as S/MAR motifs in *Arabidopsis thaliana* (19). These MRSs have been applied in other species (20), and furthermore, were defined as the most enriched motifs in S/MAR containing sequences (14). We extracted MRS matching sequences from 55 entries of dicotyledonous plants from the S/MARt DB, and made weight-matrices These weight matrices were tested on several experimentally confirmed S/MAR containing sequences, and FN_50 profiles from the MATCH™ Profiler correctly predicted all of the S/MAR loci on them. The MATCH program predicted S/MAR loci on chromosomes 4 and 5, and the average interval between S/MAR loci was 15.5 kb for both chromosomes. With this length of sequences, an average of three genes can reside because the gene densities for chromosomes 4 and 5 are 4.6 kb and 4.4 kb per gene, respectively. Thus, we extended the range of analyses to three consecutive gene sets, which we defined as ' trigrams' .

In the collection where the S/MAR was not considered, we tried on different two frames for bigrams. In the first frame, we chose the first and second genes for the first bigram, and the third and fourth for the second, and so on. In the second frame, we chose the second and third genes for the first bigram and so on for subsequent bigrams. Similarly, we chose three frames for the trigrams. In this way, we collected independent bigrams and trigrams and could calculate the p-values for the binomial random variable I, as described in materials and methods. The collections of bigrams and trigrams where the S/MAR was considered were also statistically independently extracted as they were separated by predefined S/MAR loci. Many of the collected bigrams and trigrams were excluded in this study, because we did not considered bigrams and trigrams with unknown or unclassified function categories. We calculated the proportions annotated genes, and the known function categories were assigned in the MAtDB, which were only about 32% (1190/3744) for chromosome 4 and 18% (1055/5874) for chromosome 5. The data for bigrams (Table 1 and 2) and trigrams (Table 3 and 4) showed the frequencies composing the genes belonged to identical function sub-categories and were similar regardless of whether the S/MAR was considered or not, which was contrary to our expectations. The differences were only the frequencies of identical bigrams or trigrams of chromosome 5 were much smaller than chromosome 4. Although it was not easily possible to conclude, it might that there were more unknown or unclassified genes on chromosome 5. The p-values

were evaluated to provide the statistical significance of the observed frequencies of functionally identical bigrams and trigrams. We first calculated *p*-values for random uniform distributions, as with the report by Ge *et al.* (11), where they evaluated *p*-values for the protein interacting pairs (and triplets) assuming each pair has the same probability in *Saccharomyces cerevisiae.* However, this assumption was not suitable for our study, because the *p*-values were extremely small when this assumption was made, and could not be used for calculations using our method. The other reason was that the proportion of annotated genes in the *Arabidopsis thaliana* genome are relatively small compared to those with the *Saccharomyces cerevisiae* genome whose gene functions are better understood, and furthermore their distribution is somewhat biased to a couple of function categories. Therefore, we provided another set of *p*-values for the weighted random distribution, and this assumption introduced a more realistic situation. In fact, excluding genes with unknown or unclassified functions, the function category for metabolism-01 was the largest in both chromosomes 4 and 5. We considered a set of genes with a known function sub-category, which we called *K*. Although the proportions of bigrams or trigrams consisting of an identical function sub-category were similar, of the total bigrams or trigrams available, when either the S/MAR was considered or not, the *p*-values were much different. If the S/MAR had some effect on the association of genes with respect to their function, the *p*-values when the S/MAR was considered should be much lower than when it is not, but the *p*-values when the S/MAR was considered were relatively higher. However, this did not mean the S/MAR affected negatively on the genomic association of genes with identical functions. This could be caused by differences in the number of bigrams or trigrams collected. If we were to try more bigrams (or trigrams), the situation becomes even further removed farther from the original assumption of the probability distribution - both for uniform and weighted random distribution. The real frequencies of the functionally identical bigrams and trigrams were much higher than expected, but the *p*-values suggested that these data were statistically significant. This suggested that regardless of the existence of S/MAR, there were significant associations of genes related in a certain cellular events on the genome. The clustering of housekeeping genes in human genomes has been reported, with suggestion that it might be advantageous to assemble housekeeping genes on some ' common ground' that remains in an open conformation across all cells

(13). The analysis of co-expressed genes suggested the possibilities of grouping genes as a functional module (24), and the accumulation of such data will resolve the relation between the genomic association of genes and their functional significance. Additionally, more analyses on the link between the higher-order chromatin structure, and the gene clustering on the genome, should be addressed to prove this relationship.

Despite the inadequacy of the annotation information, this study has shown the significant association of neighboring genes with identical function sub-categories on chromosomes 4 and 5 of the *Arabidopsis thaliana* genome. Using all the information on genome annotation from large-scale functional genomics the application of this strategy will reveal detailed and unbiased results, which complement experimental knowledge.

6. ACKNOWLEDGEMENT

This research was supported by a grant from Plant Diversity Research Center of 21st Century Frontier Research Program funded by Ministry of Science and Technology of Korean government.

7. REFERENCES

1. Dandekar,T., Snel,B., Huynen,M. and Bork,P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem. Sci.*, 23, 324-328.
2. Von Mering,C. and Bork,P. Genome organization: Teamed up for transcription. *Nature*, 417, 797 - 798.
3. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events *Nature*, 402, 86-90.
4. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W. , Yeates,T.O. and Eisenberg,D. (1999) Detecting Protein Function and Protein-Protein Interactions from Genome Sequences *Science*, 285, 751-753.
5. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA*, 96, 4285-4288.
6. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. USA*, 96, 2896-2901.
7. Huynen,M., Snel,B., LatheIII,W. and Bork,P. (2000) Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences. *Genome Res.*, 10, 1204-1210.
8. Yanai,I., Derti,A. and DeLisi,C. (2001) Genes linked by fusion vents are generally of the same functional category: A systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. USA*, 98, 7940-7945.
9. Snel,B., Bork,P. and Huynen,M.A. (2002) The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. USA*, 99, 5890-5895.
10. Blumental,T., Evans,D., Link,C.D., Guffanti,A., Lawson,D., Theirry-

Mieg,J., Chiu,W.L, Duke,K., Kiraly,M. and Kim,S. (2002) A global analysis of *Caenorhabditis elegans* operons. *Nature*, **417**, 851 - 854.

11. Ge,H., Liu,Z., Church,G.M. and Vidal,M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nature Genet.*, **29**, 482-486.
12. Cohen,B.A., Mitra,R.D., Hughes,J.D. and Church,G.M. (2000) A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. *Nature Genet.*, **26**, 183-186.
13. Lercher,M.J., Urrutia,A.O. and Hurst,L.D. (2002) Clustering of housekeeping genes provides a unified model of gene order in the human genome. *Nature Genet.*, **31**, 180-183.
14. Liebich,I., Bode,J., Frisch,M. and Wingender,E. (2002) S/MARt DB: a database on scaffold/matrix attached regions. *Nucleic Acids Res.*, **20**, 372-274.
15. Stein,G.S. (1998) Interrelationships of nuclear architecture with gene expression: Functional encounters on a long and winding road. *J. Cell. Biochem.*, **70**, 157-158.
16. Frisch,M., Frech,K., Klingenhoff,A., Cartharius,K., Liebich,I. and Werner,T. (2001) *In silico* prediction of scaffold/matrix attachment regions in large genomic sequences. *Genome Res.*, **12**, 349-354.
17. Bode,J., Kohwi,Y., Dickinson,L., Joh,T., Klehr,D., Mielke,C. and Kohwi-Shigematsu,T. (1992) Biological significance of unwinding capability of nuclear matrix-associating DNAs. *Science*, **255**, 195- 197.
18. Singh,G.B., Kramer,J.A. and Krawetz,S.A. (1997) Mathematical model to predict regions of chromatin attachment to the nuclear matrix. *Nucleic Acids Res.*, **25**, 1419-1425.
19. Van Drunen,C.M., Oosterling,R.W., Keultjes,G.M., Weisbeek,P.J, Van Driel,R. and Smeekens,S.C.M. (1997) Analysis of the chromatin domain organization around the platocyanin gene reveals an MAR-specific sequence element in *Arabidopsis thaliana*. *Nucleic Acids Res.*, **25**, 3904-3911.
20. Van Drunen,C.M., Sewalt,R.G.A.B., Oosterling,R.W., Weisbeek,P.J, Keultjes, G.M., Smeekens,S.C.M. and Van Driel,R. (1999) A bipartite sequence element associated with matrix/scaffold attachment regions. *Nucleic Acids Res.*, **27**, 2924-2930.
21. Liebich,I., Bode,J., Reuter,I. and Wingender,E. (2002) Evaluation of sequence motifs found in scaffold/matrix-attached regions (S/MARs).

Nucleic Acids Res., **30**, 3433-3442.

22. The European Union Arabidopsis Genome Sequencing Consortium & The Coldspring Harbor, Washington University in St Louis and PE Biosystems Arabidopsis Sequencing Consortium. (1999) Sequence and analysis of chromosome 4 of the plant Arabidopsis thaliana. *Nature*, 402, 769-777.
23. The Kazusa DNA Research Institute, The Cold Spring Harbor and Washington University in St Louis Sequencing Consortium and The European Union Arabidopsis Genome Sequencing Consortium. (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. *Nature*, 408, 823-826.
24. Thompson,H.G..R., Harris,J.W., Wold,B.J., Quake,S.R. and Brody,J.P. (2002) Identification and confirmation of a module of coexpressed genes. *Genome Res.*, 12, 1517-1522.

Figure Legend

Figure 1. Analysis flow of neighboring genes.

Figure 2. The distributions of the functional combinations of neighboring genes when the S/MAR was not considered. Bigrams were mapped according to their nineteen large function categories. Panel A and C, indicated by 'Total', are the matrices from both bigram frames. Panel B and D are the matrices for the first bigram frame, and C and F are for the second bigram frame. The color gradient in the upper-right corner of each panel shows the bigram density per one thousand gene pairs. Numbers on the vertical and horizontal axis indicate the large functional categories.

Figure 3. The distributions of the functional combinations between neighboring genes when the S/MAR was considered. Bigrams were mapped according to their nineteen large function categories. Panel A and C are the matrices for bigrams located before S/MAR loci. Panel B and D are the matrices for bigrams with the S/MAR in the middle of them, and C and F are for the bigram located after the S/MAR. Numbers on the vertical and horizontal axis indicate the large functional categories.

Table 1. Statistics on bigrams of chromosomes 4 and 5 without considering S/MAR

	Chromosome 4			Chromosome 5		
	Frame 1	Frame 2	Total	Frame 1	Frame 2	Tota
Expected No. of	16.7720	17.5968		18.3054	18.3054	
Observed No. of	100 (4.59%)	102 (4.68%)	202 (4.64%	79 (2.46%)	75 (2.33%)	154 (2.4
No. of bigrams	366 (16.80%)	384 (17.63%)	750 (17.21	349 (10.88%)	349 (10.86%)	698 (10
No of total	2178	2178	4356	3208	3213	6421
P-value	P(Z>20.804 8) ≈	P(Z>20.598 2) ≈		P(Z>14.573 3) ≈	P(Z>13.612 9) ≈	

Table 2. Statistics on bigrams of chromosomes 4 and 5 considering S/MAR

	Chromosome 4			**Chromosome 5**		
	Before	**Across**	**After S/MAR**	**Before**	**Across**	**After**
No. of predicted S/MAR loci		1119			1678	
Expected No. of functionally identical bigrams	9.3483	8.7984	9.9899	9.6509	8.3922	8.9691
Observed No. of functionally identical bigrams	60 (4.87%)	65 (5.67%)	61 (4.87%)	43 (2.41%)	39 (2.28%)	37 (2.06%)
No. of bigrams w/o 00/98/99 categories	204 (16.55%)	192 (16.80%)	218 (17.41%)	184 (10.32%)	160 (9.37%)	171 (9.53%)
No of total Bigrams	1233	1143	1252	1783	1707	1794
P-value	P(Z>16.9595) ≈ 7.3690×10^{-64}	P(Z>19.3969) ≈ 3.8055×10^{-83}	P(Z>16.5220) ≈ 1.1369×10^{-60}	P(Z>11.0280) ≈ 1.0183×10^{-27}	P(Z>10.8541) ≈ 6.8648×10^{-27}	P(Z>9.6153) ≈ 2.2902×10^{-21}

Table 3. Statistics on trigrams of chromosome 4 and 5, without considering S/MAR.

	Frame 1	Frame 2	Frame 3	Total
Chromosome 4				
Expected No. of functionally	0.4540	0.4611	0.4824	
Observed No. of functionally	22 (1.46%)	24 (1.59%)	24 (1.60%)	70 (1.55%)
No. of trigrams w/o 00/98/99 categories	128 (8.49%)	130 (8.64%)	136 (9.05%)	394 (8.73%)
No. of total trigrams	1507	1505	1503	4515
P-value	P(Z>32.0332) ≈	P(Z>34.7262) ≈	P(Z>33.9207) ≈	
Chromosome 5				
Expected No. of functionally	0.4611	0.4469	0.4256	
Observed No. of functionally	13 (0.59%)	15 (0.68%)	11 (0.50%)	39 (0.59%)
No. of trigrams w/o 00/98/99 categories	130 (5.91%)	126 (5.75%)	120 (5.47%)	376 (5.71%)
No. of total trigrams	2199	2190	2192	6581
P-value	P(Z>18.4977) ≈	P(Z>21.8087) ≈	P(Z>16.2376) ≈	

Table 4. Statistics on trigrams considering S/MAR of chromosomes 4 and 5

	Chromosome 4		Chromosome 5	
	Before	After S/MAR	Before	After S/MAR
Expected No. of functionally identical trigrams	0.4398	0.3937	0.3405	0.3476
Observed No. of functionally identical trigrams	23 (1.79%)	19 (1.48%)	12 (0.66%)	7 (0.38%)
No. of trigrams w/o 00/98/99 cat.	124 (9.66%)	111 (8.67%)	96 (5.26%)	98 (5.36%)
No. of total trigrams	1283	1281	1824	1828
P-value	$P(Z>34.0767) \approx 6.4265\times10^{-254}$	$P(Z>29.7065) \approx 2.4439\times10^{-193}$	$P(Z>20.0165) \approx 2.0943\times10^{-88}$	$P(Z>11.3029) \approx 3.0978\times10^{-29}$

Figure 1

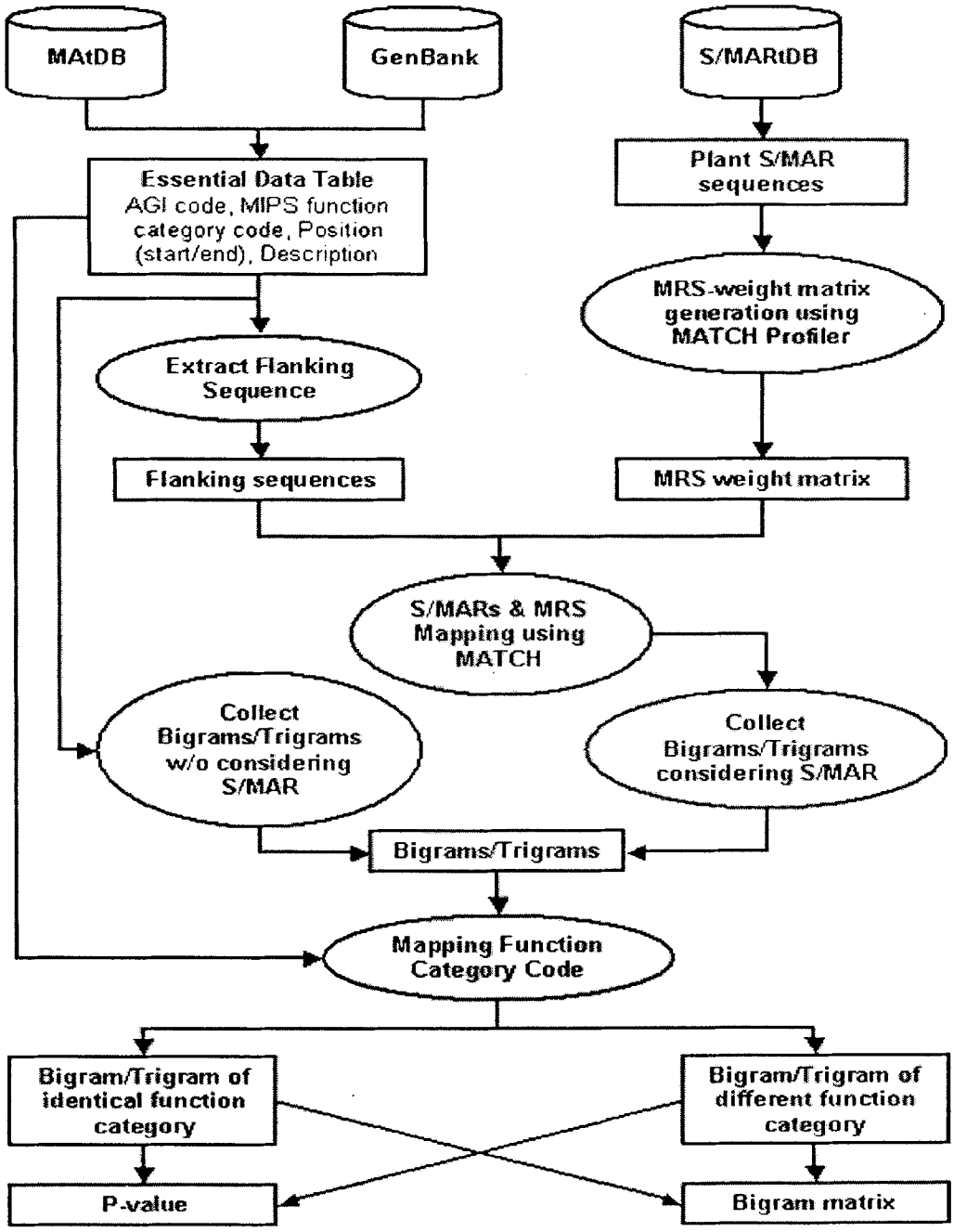

MAtDB
GenBank
S/MARtDB
Essential Data Table
AGI code, MIPS function category code, Position (start/end), Description
Plant S/MAR sequences
MRS-weight matrix generation using MATCH Profiler
Extract Flanking Sequence
Flanking sequences
MRS weight matrix
S/MARs & MRS Mapping using MATCH
Collect Bigrams/Trigrams w/o considering S/MAR
Collect Bigrams/Trigrams considering S/MAR
Bigrams/Trigrams
Mapping Function Category Code
Bigram/Trigram of identical function category
Bigram/Trigram of different function category
P-value
Bigram matrix

Figure 2

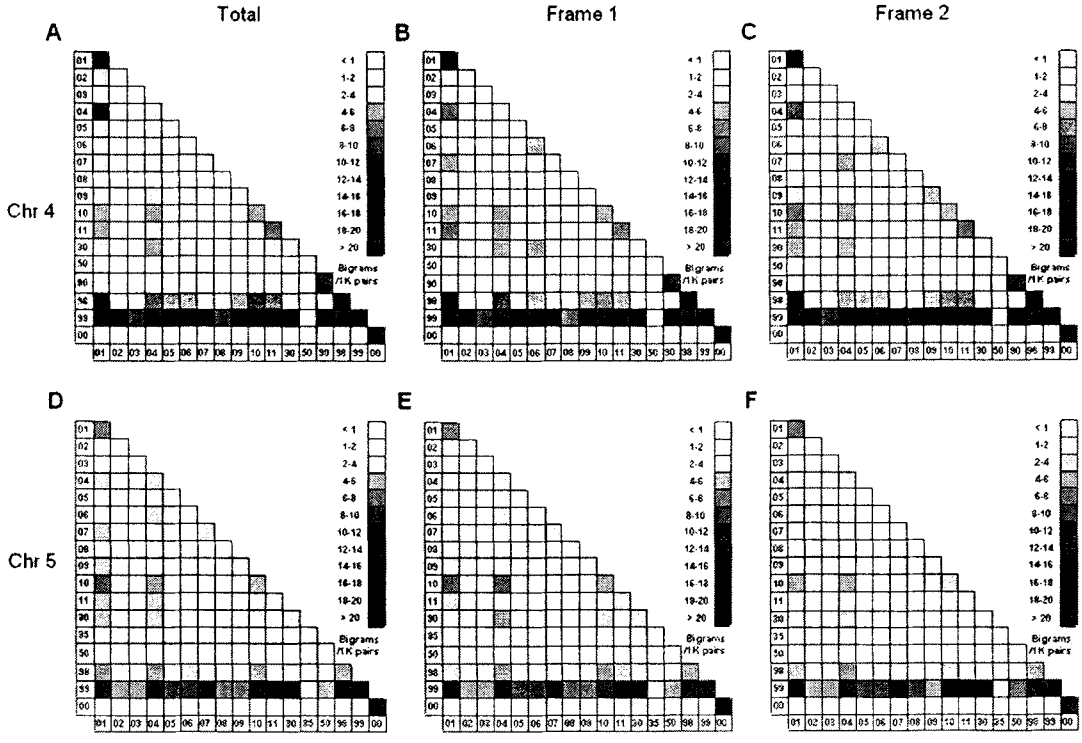
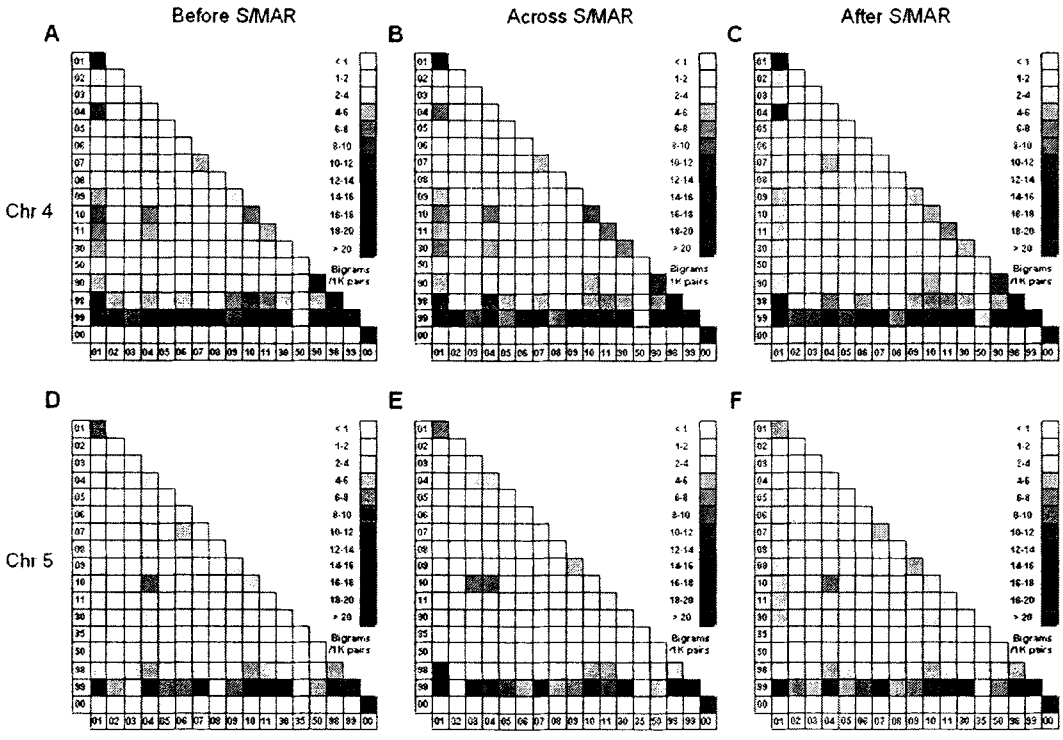


Figure 3


Before S/MAR
Across S/MAR
After S/MAR
A
B
C
D
E
F
Chr 4
Chr 5
Bigrams /1K pairs

마. EST and microarray analysis of pathogen responsive genes in hot pepper non-host resistance against soybean pustule pathogen.

1. ABSTRACT

Large-scale single-pass sequencing of cDNA has proven to be a useful tool for discovery of new genes and understanding of biological mechanisms. As a first step to understand the complexity of plant defense mechanism, expressed sequence tags (EST) were generated from hot pepper leaf cDNA library constructed from combined leaves collected at different time points after inoculation with soybean pustule pathogen (*Xanthomonas campestris* pv. *glycines*). To increase gene diversity, ESTs were also generated from cDNA libraries constructed from anthers and flower buds using dye termination method. Among total of 10,061 generated ESTs, 8,525 had sufficient quality for analysis, and clustering analysis revealed that 55% of total ESTs (4685) were unique. BLASTX analysis revealed that 74% of ESTs had significant sequence similarities with known proteins present in NCBI nr database. In addition, 1,265 tentative full-length cDNAs were also identified from EST analysis. Functional classification of ESTs derived from pathogen-infected pepper leaves revealed that about 25% of the ESTs represented disease- or defense-related genes. Furthermore, 323 (7%) ESTs were identified as tentative hot pepper specific sequences. Here, we describe detailed sequence analysis data, which are the first work in hot pepper plant species. Although we focused on the genes related to plant defense response, our data are the useful depository for further comparative studies or other purposes.

[The sequence data in this paper have been submitted to the dbEST database under . The sequences and detail information are also available at http://plant.pdrc.re.kr/Gene]

2. INTRODUCTION

As sessile organisms, plants should consistently respond to local biotic and abiotic stresses by changing molecular and cellular responses. Through these processes, plants have evolved various defense mechanisms. The plant defense responses are not the simple expression of defense-related genes, but orchestrated reactions of transcriptional activation of multiple genes, accumulation of secondary metabolites, hypersensitive response, and systemic acquired resistance (Lam et al. 1989; Dixon 1986; Dangl and Jones 2001; Ryal et al. 1996). This complexity of the plant defense responses hardly yields effective strategies leading to generation of plants with improved disease tolerance (Somssich and Hahlbrock 1998). Therefore, identification and analysis of the complete set of genes involved in the defense processes is an essential step toward understanding of the whole scheme of plant defense mechanisms and generation of disease resistant plants.

During past several years, lots of efforts have been put into sequencing of genomes including plant species (Venter et al. 2001; TAGI 2000). As a result of these endeavors, the first complete genome sequences of flowering plant (*Arabidopsis thaliana*) have been unveiled (TAGI 2000). Although entire genomic sequences of Arabidopsis are available, sequencing and analysis of the cDNA sequences (EST) is still invaluable, especially for other plant species. Moreover, recent developments in DNA sequencing and sequence analysis tools also emphasized usefulness and effectiveness of single-pass cDNA sequencing. Therefore, expressed sequence tag (EST) projects were widely launched to analyze expressed genes of different stages and tissues (Cooke et al., 1996; Sasaki et al. 1994; Van de Loo et al. 1995; Allona et al. 1998; Mekhedov et al. 2000; Kwak et al. 1997; Zhang et al. 2001). As a result of these efforts, more than 1,000,000 ESTs were identified from more than 30 different plant species (http://www.ncbi.nlm.nih.gov/dbEST/).

To gain insight on plant defense mechanisms, we performed random EST sequencing to isolate genes at the onset of hypersensitive response (HR) expressed from hot pepper plant. The collected ESTs were compared for sequence homology to Arabidopsis database (http://mips.gsf.de/desc/thal) to classify their function. In addition, various sequence analysis tools and methods were applied to this study. The comprehensive analysis might provide a useful tool to study not only molecular mechanisms of plant defense, but also those of other stress reactions or hormone responses, since the signal transduction pathway of plant defense is partially overlapping with those of other abiotic or hormonal signaling pathway (Genoud and Metraux 1999; Thomma et al., 2001). The overall EST analysis information was opened to public through our web site (http://plant.pdrc.re.kr/Gene) for free access.

3. RESULTS AND DISCUSSION

Generation, Quality Assessment and Clustering of ESTs

To isolate series of genes involved in defense response of hot pepper plant, we constructed three different libraries. One library (KS01) was generated from hot pepper leaf RNA prepared after inoculation of soybean pustule pathogen (Xcg; *Xanthomonas campestry* pv *glycine*) and two others are generated from flower buds (KS07) and anthers (KS08), respectively. We attempted to analyze these three different libraries to obtain more information about differential expression of defense-related genes in hot pepper plant.

Before performing detailed analysis of ESTs, sequences possibly originated from non-nuclear organelle were detected by searching EST sequences against mitocondrial, chloroplast, and ribosomal RNA sequences using BLASTN algorithm (cut-off value was e-10). As shown in Table 1, all 3 libraries have less than 4.3% possible contamination of organelle or ribosomal RNA, and ribosomal RNA sequences occupied half of all possible contaminated sequences. Specifically, KS01, KS07 and KS08 libraries have 2.0%, 6.6% and 6.8% of possible contaminated sequences, respectively. Analysis of ESTs using in-house developed program, 13% of our ESTs (1265 EST) contain entire open reading frame of each gene and identified as tentative full-length cDNAs (Table 2). The full-length cDNA candidates were divided into two groups. First group was characterized by presence of Met (translation initiation codon) in the first position of the encoded protein sequence, and this group composed 45% of tentative

Table I. Possible contaminated sequences in hot pepper ESTs generated in this study.

Contaminated Sources	KS 01 library	KS 07 library	KS 08 library	Total[d]
rRNA[a]	50 (0.9)[e]	82 (4.0)	31 (2.7)	182 (2.1)
Mt DNA[b]	42 (0.8)	25 (1.2)	24 (2.1)	108 (1.3)
Cp DNA[c]	14 (0.3)	29 (1.4)	23 (2.0)	79 (0.9)
Sum	106 (2.0)	136 (6.6)	78 (6.8)	369 (4.3)

[a]rRNA represents the ribosomal RNA.
[b]Mt DNA represents Mitochondrial DNA.
[c]Cp DNA represents plastid DNA.
[d] Total represents the mixed all three, KS01, KS07 and KS08 libraries and information of each library was described in " material and method" .
[e]Parentheses indicates the percentage of analysis result.

full-length clones. Second group was the ESTs which are absence of Met (translation

initiation codon) in the first position of the encoded protein sequence, but 5' end of query sequence (EST) have more than three times of nucleotide compared to the corresponding subject sequence. Even though this approach is not quite accurate (Ablett et al. 2000) but it is a quick and convenient way of determining possible full-length cDNA directly from the BLASTX output without lengthy analysis.

Table 2. Quality of hot pepper EST.

	Library			
	KS01	KS07	KS08	Total
Total EST	5368	2017	1140	8525
Cluster (%)[a]	3056 (57%)	1189 (59%)	901 (79%)	4685 (55%)
Singleton (%)[a]	2163 (40%)	853 (42%)	817 (72%)	3287 (39%)
Tentative Full Length cDNA (%)[b]	1009 (17%)	173 (8%)	83 (7%)	1265 (15%)
G+C Content	41%	43%	41%	42%
Average Length	524 bp	548 bp	402 bp	506 bp

[a]cluster and singleton represents the unique sequences with or without redundancy
[b]The percentage was out of 8,525 ESTs and detailed information about tentative fu

Average G+C contents of hot pepper ESTs were 42% and were similar to those of soybean and arabidopsis (Qutob et al. 2000; TAGI 2000. Table2). G+C content could be one good marker for separating plant cDNA from mixed sources such as cDNAs from fungal pathogen-infected plant tissues. As an example, ESTs from *Phytophthora* infected soybean leaves, plant cDNAs were easily distinguished from *Phytophthora* cDNA because of the high G+C content of *Phytophthora sojae* (60%), which is 18% higher than that of soybean. Single pass sequencing from 5'-end cDNA resulted in high quality sequences of average 516 bp (Table 2). After removing low-quality sequences (PHRED cut-off value 0.05), 8,525 cDNA sequences were selected and analyzed for their redundancy. From this analysis, 4,685 clusters were obtained which include 3287 singletone sequences (Table 2). Although most of clusters (81%) were composed of less than 5 redundant sequences, 19% of clusters were composed of more than 5 redundant sequences (Figure 1).

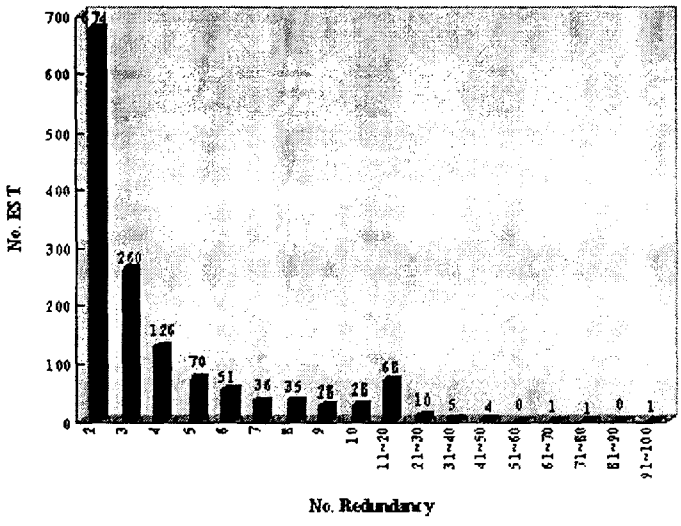


Figure 1. Distribution of EST redundancy. X-axis indicates the number of same sequences corresponding to each EST (redundancy). Y-axis indicates the number of EST clusters. About half (674) of 1,398 unique ESTs have only two same sequences in our EST database. The highest redundancy was occurred only one time and the redundancy was more than 90 times.

Since we have sequenced randomly selected ESTs, the redundancy might indicate the levels of mRNA expression. Our clustering analysis revealed that 55% of the ESTs were redundant but this result could be under- or over-estimated. The former case could be explained by imprecise clustering algorithm because CAP3 program (http://genome.cs.mtu.edu/cap/cap3.html) accidentally could pick up highly homologous gene family members as one cluster. The latter case is also possible because all of our EST data were obtained by 5'-end single pass sequencing and most of our unique ESTs comes from 5' -end region of corresponding transcripts, but some of them might come from different regions of the same transcripts by partial length cDNA synthesis during library preparation and subsequently clustered as different groups. In order to get more accurate unique gene information, additional sequencing and detailed analysis will be needed. Sequencing of 3'-untranslated region and/or detailed sequence alignment within one cluster may give clues to identify independently transcribed gene(s) from the same cluster.

4. Analysis of Simple Sequences Repeat

Simple sequence repeats (SSRs or microsatellites) are tandemly repeated nucleotide sequences of 2-4 bp length that vary in number of repeats and are flanked by conserved

DNA sequences (Tautz 1989). Due to the ubiquity of SSRs and their usefulness as genetic markers, extensive analyses of genome wide distribution of SSRs have been demonstrated for a variety of plant species (Akkaya et al. 1992; Liu et al. 1996; Milbourne et al. 1998; Morgante and Olivieri 1993; Senoir et al. 1996; Zhao and Kochert 1992) and considered as a tool for marker assisted selection and germplasm assessment (Cardle et al. 2000). In order to find SSR markers, we

Table 3A. Analysis of repeated sequences from pepper ESTs generated in this study.
(A) A distribution of repeated sequences.

Repeated type (total number)	Nucleotide composition	No. of appearance
low complexity (201)	AT rich	198
	GC rich	3
Di-repeat (74)	(CA)n	3
	(GA)n	63
	(TA)n	8
Tri-repeat (102)	(CAA)n	11
	(CAG)n	14
	(CAT)n	15
	(CGA)n	1
	(CGG)n	6
	(GAA)n	20
	(GGA)n	8
	(TAA)n	9
	(TAG)n	1
	(TGG)n	17
Tetra-repeat (12)	(CAAT)n	1
	(CATA)n	1
	(GAAA)n	5
	(TAAA)n	4
	(TTAA)n	1
Penta-repeat (14)	(CAAAA)n	1
	(CAATA)n	1
	(CAATT)n	1
	(CACAA)n	1
	(CACCC)n	1
	(GAAAA)n	4

	(GAGAA)n	2
	(GGAAA)n	1
	(TAAAA)n	1
	(TTAAA)n	1
Hexa-repeat (22)	(TAAAAA)n	3
	(CACCAT)n	2
	(CAGAGA)n	1
	(CCCCAA)n	2
	(CCCCAG)n	1
	(CCCGAA)n	1
	(GGAGAA)n	7
	(GGGAGA)n	2
	(TGGGGG)n	3
Total (425)		425

" Repeatmasker" was used to analyze repeat sequences at default parameters. Each repeated sequences were composed of more than 20 nucleotides.

analyzed pepper EST database using RepeatMasker software (http://ftp.genome.washington.edu/RM/RepeatMasker.html) for searching all possible low complexity and repeated sequences. We found 425 repeated sequences. Among them, 48% were AT-, GC- rich low complexity repeated sequences and 52% were simple sequence repeats including di-, tri-, tetra-, penta-, and hexa-nucleotide repeat (Table 3A). The most abundant simple sequence repeats were tri-nucleotide and di-nucleotide composed of 79% of all simple sequence repeats. In tri- and di-nucleotide repeats, the sequence of most abundant repeats was (GA)n, (GAA)n, (TGG)n, (CAT)n, and (CAG)n (Table 3A). When we compared our information of repeated sequences to Cardle' s work (Cardle et al. 2000), the interesting results were found. The proportion of di-, and tri-nucleotide repeated sequences was 87%, which is similar to other species, but the percentage of di-nucleotide repeated sequences in pepper were 37% at the highest, compared to 26% in average in other species (Table 3B). The average distance of repeated sequences were 11 kb, which is most similar to that of tomato, one of the solanaceous plant.

Table 3B. Comparative analysis of Simple Sequence Repeat (SSR)

	Source					
	Pepper	Rice[a]	Maize[a]	Soybean[a]	Tomato[a]	Cotton[a]
Di-nucleotide repeat	74	657	140	147	84	53
Tri-nucleotide repeat	102	3,747	478	311	289	157
Tetra-nucleotide repeat	12	498	126	30	24	21
Penta-nucleotide repeat	14	230	46	9	2	8
Total SSRs	202	5,132	790	497	399	239
Total EST sequences	4,685	45,033	14,950	9,611	9,100	8,083
Average length of sequence (bp)	506	380	430	380	490	590
Total length (kb)	2,371	17,304	6,411	3,675	4,444	4,788
Average distance (kb)	11.7	3.4	8.1	7.4	11.1	20.0

[a]All the information was shown in previous paper of cardle et al. (Genetics 156: 847-854)

5. Comparison of Expression of ESTs in Different Libraries

The EST occurrence in a specific library represents the expression levels of its corresponding gene in a specific situation, and is called an electronic Northern blot (Ewing et al. 1999). Since we used three different libraries, the expression profiling of ESTs can be divided into seven categories (Table 4). 88% of ESTs were expressed only in each specific library, about 2% of ESTs were commonly expressed in all three libraries and the remaining 10% of ESTs were expressed in two libraries among three (Table 4). When we searched library-specific ESTs, 56%, 17%, and 15% of total ESTs were only expressed in KS01, KS07, and KS08 library, respectively (Table 4). Since our interests are on the genes related to cell death, defense and

Table 4. Expression profile of ESTs in three different libraries.

Library	Number of ESTs
KS01 only	2,640
KS07 only	815
KS08 only	708
KS01 and KS07	305
KS07 and KS08	61
KS01 and KS08	72
KS01, KS07 and KS08	84
Total	4685

Disease resistance, we sequenced more clones from the KS01 library made from pathogen-infected leaf tissues, and focused on the analysis of the 2640 ESTs that were appeared only in KS01 library. We selected the ESTs, which appeared at least 6 times in KS01 library. Total 136 EST clusters were selected and categorized based on their putative functions (Supplementary Table 1). Seventy-two (53%) clusters were categorized into functional groups. Among them, defense-, metabolism-, and protein synthesis-related ESTs were the majority. Three functional categories including defense (21%), metabolism (8%), and protein synthesis (7%) cover 36% of KS01 derived abundant ESTs (Supplementary Table 1). This result accords with the fact that plants protect themselves *via* altering the metabolisms or gene expressions (Dixon 2001; Maleck and Dietrich, 1999; Bowles 1990). In KS01 library, disease and defense related genes were 21% of total 136 clustered ESTs. The most distinctive changes of gene expression in plant tissues showing disease resistance are PR (pathogenesis-related) proteins (Ward et al. 1991). In our ESTs from KS01 library, five different classes of PR proteins including PR-1, PR-10, chitinase, SAR8.2, and glucanase, were abundantly expressed. Furthermore, several disease resistance-related proteins, such as thionin, ubiquitin, catalase, glutathione-S-transferase, cytochrome P450, and 14-3-3 protein were also abundantly appeared in KS01 library (Oh et al., 1999; Becker et al., 2000; Wu et al., 1999; Levine et al., 1994; Whitbred and Schuler 2000; Roberts and Bowles 1999). During pathogen attacks, plant rapidly produces ethylene (Dong 1998). Ethylene biosynthetic pathway is well characterized and it is known that ACC oxidase (ACO) is the ethylene-forming enzyme (Kende 1993). We could find 8 different ACOs from KS01 library. Among them 5 ACOs were abundantly expressed (appeared more than 6 times) in KS01 library. This result may imply differential regulation of ACO family member during plant-pathogen interaction and disease resistance response of hot pepper plants. In contrast, we could not find any ACC synthase in our pathogen-infected library, which may imply negative regulation of ACC synthase expression by ethylene produced during pathogen infection (Yoon et al. 1997; Peck and Kende 1998).

Sixty-four (47%) clusters were not categorized to a specific functional group (Table 5). Among them, twenty EST clusters could not find any homologous sequence from established databases. These ESTs may imply the specific role(s) in defense responses of hot pepper plant. Although the transformation efficiency of hot pepper plant using *Agrobacterium* mediated transformation is relatively low, ectopic expression of specific genes in hot pepper plants could be sole method for functional characterization (Kim et al. 2001; Zhu et al. 1996). Others including unknown or hypothetical proteins might provide clues to uncover the unidentified part of plant defense mechanism. In order to identify their functional role(s) for these ESTs, over- and/or

under-expression analysis will be pursuit using appropriate hetelogous plants based on their sequence similarity to Arabidopsis, tomato and tobacco plants. Although difficult transformation of pepper plant forced us to use hetelogous organism for gene function identification, the tentative function(s) identified in hetelogus transgenic plants broaden our knowledge about plant-pathogen interactions.

6. Functional Categorization and Comparative Analysis of Hot Pepper ESTs

To assess similarity and differences of pepper ESTs compared with other eukaryotic genes or ESTs, we functionally categorized the hot pepper ESTs using MIPS Arabidopsis gene functional categories (http://mips.gsf.de/desc/thal). Since 2035 ESTs from our total ESTs did not match to the arabidopsis protein by using N2Tool (threshold 100), these sequences were not included for the functional cataloging. Following the MIPS classification, 2650 ESTs were assigned to their function. About half of sequences were assigned more than two functions.

Table 5.

EST ID[a]	Gene similarity[b]	Blastx (E value)[c]	Accession number
KS01002A02	Hypothetical protein	3.0E-52	AC005397
KS01002A07	Hypothetical protein	9.0E-58	AB024034
KS01002A08	Hypothetical protein	9.0E-17	AC000375
KS01002C04	Hypothetical protein	2.0E-35	AC061957
KS01002C07	Succinyl-CoA synthetase, alpha subunit	5.0E-46	AB007648
KS01002C09	N/A	N/A	N/A
KS01002D05	N/A	N/A	N/A
KS01002D07	Putative protein	1.0E-05	AB025604
KS01002D10	Unknown protein	3.0E-25	AF370507
KS01002E01	Putative protein	3.0E-14	AB052603
KS01002E09	N/A	N/A	N/A
KS01002E10	N/A	N/A	N/A
KS01002F04	N/A	N/A	N/A
KS01002G11	N/A	N/A	N/A
KS01002H02	Calcium-binding protein	2.0E-70	Z71395
KS01003A09	N/A	N/A	N/A
KS01003B06	N/A	N/A	N/A
KS01003B12	Hypothetical protein	1.0E-31	AC003028
KS01003C05	12-oxophytodienoate reductase (OPR1)	4.0E-39	AJ242551

KS01003D05	N/A	N/A	N/A
KS01003D07	Carbonic anhydrase	1.0E-57	AB012863
KS01003E07	Hypothetical protein	1.0E-71	AC013453
KS01003F11	N/A	N/A	N/A
KS01003F12	N/A	N/A	N/A
KS01003H06	Citrare synthase	1.0E-88	X84226
KS01003H08	Hypothetical protein	7.0E-41	AC008007
KS01004A10	100 kDa coactivator - like protein	2.0E-66	AB055904
KS01004D09	26S proteasom	2.0E-44	AB015476
KS01004G04	Putative protein kinase	1.0E-86	U28007
KS01004G05	Fimbriate-associated protein3	7.0E-29	Y14858
KS01004H09	N/A	N/A	N/A
KS01005B02	Hypothetical protein	3.0E-33	AB056451
KS01005B04	N/A	N/A	N/A
KS01005C02	N/A	N/A	N/A
KS01005C10	Ferredoxin(Fe-S) fd1 precursor	6.0E-42	Z46944
KS01005F11	Hypothetical protein	9.0E-26	X13934
KS01006F09	N/A	N/A	N/A
KS01006G07	N/A	N/A	N/A
KS01007E02	Proteasome subnite alpha type 6	2.0E-69	Y16644
KS01007F09	Putative protein	1.0E-72	AB017066
KS01007H01	N/A	N/A	N/A
KS01007H04	Unknown protein	4.0E-10	AC007060
KS01007H10	Chaperonin 60 beta chain precursor	1.0E-72	U46136
KS01008B08	ATP synthase epsilon chain	2.0E-18	AC025294
KS01008E09	N/A	N/A	N/A
KS01009C03	N/A	N/A	N/A
KS01012G02	Glucose-6-phosphate dehydrogenase	2.0E-63	X74421
KS01015A12	P-rich protein	5.0E-10	AB041516
KS01016H03	Terpene synthase-related protein	4.0E-69	AJ005588
KS01017H09	N/A	N/A	N/A
KS01018A09	N/A	N/A	N/A
KS01018F06	N/A	N/A	N/A
KS01018G07	N/A	N/A	N/A
KS01018G08	N/A	N/A	N/A
KS01033H03	N/A	N/A	N/A
KS01038E08	Keratin associate protein	1.0E-06	D89902

KS01042A10	Chitinase	1.0E-97	Z15139
KS01050C11	Delta fatty acid desaturase	1.0E-79	X92847
KS01052E09	N/A	N/A	N/A
KS01059G05	N/A	N/A	N/A
KS01063C10	N/A	N/A	N/A
KS01067B09	Stress related protein, putative	2.0E-52	AC009606
KS01067F12	N/A	N/A	N/A
KS01004E02	ABC family transporter - like protein	1.0E-39	AB010069
KS01005F03	Copine-like protein	4.0E-52	AC005397
KS01023C04	Major intrinsic protein2	1.0E-57	Y18312
KS01005D05	SNF2 transcrirtion factor	4.0E-25	AC005397
KS01011A02	N/A	N/A	N/A
KS01012E07	Blue copper binding protein	3.0E-24	U65511
KS01002H07	Dehydrin	5.0E-22	U69633
KS01003B11	CDPK	5.0E-04	AF219972
KS01034F09	dnaJ protein homolog atj3	7.0E-56	AF124139
KS01035G08	N/A	N/A	N/A
KS01005D12	N/A	N/A	N/A
KS01032D02	N/A	N/A	N/A
KS01004F12	Translation initiation factor	5.0E-45	AL109787
KS01038C04	60s ribosomal protein L34	6.0E-47	L27089
KS01045F08	Ribosomal protein S15	1.0E-71	AF051217
KS01045H06	Putative malate oxidoreductase	2.0E-98	AF001270
KS01046B10	60s ribosomal protein	4.0E-20	X95458
KS01048F07	40s ribosomal protein	6.0E-30	X76714
KS01054D12	N/A	N/A	N/A
KS01058E10	60S ribosomal protein L13, BBC1 protein	1.0E-87	X75162
KS01005C11	Putative histone deacetylase	4.0E-07	AF255711
KS01002G03	Adenosine kinase	3.0E-71	AF180895
KS01002H04	Starch phosphorylase H	8.0E-22	M69038
KS01003A06	Putative flavonol glucosyltransferase	1.0E-66	AJ310148
KS01003G02	Flavonoid 3',5'-hydroxylase -like protein	1.0E-44	AL080318
KS01004H05	Transketolase	1.0E-85	Y15781
KS01005D10	Diaminopimelate epimerase	5.0E-34	AL132966
KS01005G12	Omega-6 fatty acid desaturase	6.0E-45	AF192486
KS01007G07	Fructokinase	1.0E-64	Z12823

KS01008A10	Beta-D-glucan exohydrolase	4.0E-68	AL133292
KS01011D05	Allyl alcohol dehydrogenase	6.0E-22	AB036753
KS01012E12	N/A	N/A	N/A
KS01003B05	Plastidic aldolase	1.0E-11	AB027001
KS01004B05	Putative malate oxidoreductase	2.0E-78	AF001270
KS01007A06	Rubisco	3.0E-16	U08611
KS01007F12	N/A	N/A	N/A
KS01034A12	Plastidic aldolase	4.0E-78	AB027001
KS01002C05	ACC oxidase	2.0E-14	L21976
KS01002F09	ACC oxidase	7.0E-59	AB013101
KS01003F08	ACC oxidase	2.0E-63	X83229
KS01011F07	ACC oxidase	4.0E-37	X58885
KS01032F01	Putative cytochrome P450	1.0E-41	AB023038
KS01033A11	14-3-3 protein GF14	7.0E-95	Y11687
KS01034F04	Unknown protein	1.0E-08	AC006284
KS01035G03	N/A	N/A	N/A
KS01039G08	ACC oxidase	6.0E-62	AB013101
KS01047G08	Luminal binding protein	1.0E-52	X60058
KS01002D11	Germin-like protein	2.0E-47	U79114
KS01065F07	Ubiquitin-like protein SMT3	2.0E-41	AB010071
KS01002C11	Putative glutathione transferase	4.0E-51	X56265
KS01002D08	Beta-1,3-glucanase class I precursor	1.0E-39	M20619
KS01003B03	Thionin like protein	2.0E-15	AF112443
KS01003F01	Disease resistance gene BS2	7.0E-12	AF202179
KS01004B09	Harpin-induced protein-like	3.0E-78	AF212183
KS01007A09	Class IV chitinase (CHIV)	4.0E-15	X88803
KS01007G05	Sar8.2	5.0E-32	AF112868
KS01012E05	N/A	N/A	N/A
KS01013G04	Chitinase class II	9.0E-96	AF091235
KS01018A06	Catalase	2.0E-58	AF035255
KS01038D06	Beta-1,3-glucanase	7.0E-90	M63634
KS01038H06	Thioninn like protein	1.0E-36	AF112443
KS01040D08	PR-10	7.0E-50	AF244121
KS01043H09	Putative gamma-thionin precursor	2.0E-46	AF128239
KS01044E01	Chitinase Iib	2.0E-26	AB003194
KS01053A09	PR-1 precursor	3.0E-48	AF348141

KS01002E08	Cell wall protein	3.0E-08	AF242730
KS01002H12	N/A	N/A	N/A
KS01011B11	Tubulin alpha-3 chain	8.0E-71	AJ132399
KS01035B09	Cell wall protein	2.0E-25	AF242730
KS01041F04	Fruit reipening protein	9.0E-19	AF093141

Eventually, 47% of 2650 ESTs were assigned to their function using sequence similarity, while the rest (53%) were homologous to unclassified proteins (Figure 2). Among 47% of ESTs with

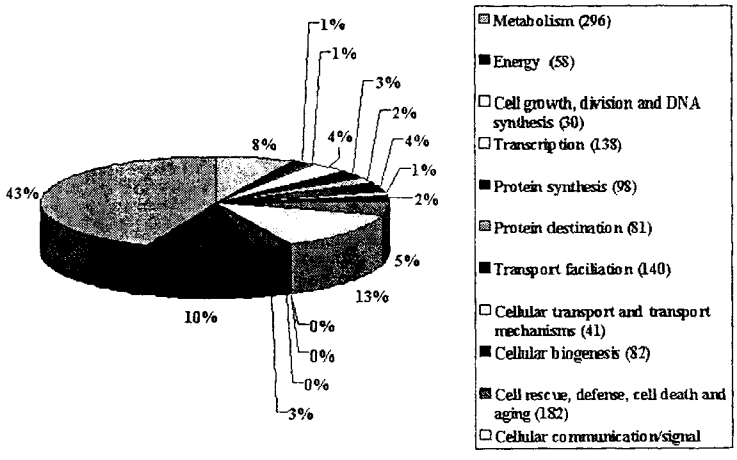


Figure 2. Functional catalog of hot pepper EST. Functional cataloging of hot pepper EST was done by sequence homology to the Arabidopsis proteins in MIPS database. Among 4685 unique ESTs, only 2650 were identified to homologous to Arabidopsis proteins and function(s) were assigned. Dual functions were assigned to some ESTs.

assigned function, about 30% of them could clearly assigned by the sequence similarity, but others required careful assignment by manual inspection. Signaling component (13%), metabolism (8%), plant defense (5%), transcription (4%), and transport facilitation (4%) are five main functional categories of hot pepper ESTs (Figure 2).

Major functional classes of hot pepper ESTs are similar to those of arabidopsis, although the percentages were different (TAGI 2000). Interestingly, we found metabolism covers the highest portion of Arabidopsis genes, but signaling component takes the highest portion of hot pepper in our study. In addition, growth-related genes are abundant in Arabidopsis genome, but rarely found in hot pepper ESTs. This difference

is likely caused by the direct comparison of Arabidopsis genome information and expressed gene information of hot pepper. When we looked over functional distribution of other plant ESTs, the relatively high portion of transport facilitation were found only in hot pepper, *Phytophthora sojae* infected soybean, and NaCl treated *Suaeda salsa* ESTs (Zhang et al. 2001; Qutob et al. 2000; Ujino-Ihara et al. 2000; Ablett et al. 2000; Covitz et al. 1998). In addition, early defense signaling regulates plasma membrane located ion channels to stimulate ion fluxes across the plasma membrane (Zimmermann et al. 1997; Jabs et al. 1997; Lee et al. 2001, Blatt et al. 1999; El-Maarouf et al. 2001). Although the majority of ESTs have unidentified functions, signaling and transport facilitation might act as a key process of stress (pathogen) response, flowering bud formation, and/or anther formation.

In order to find similarity/difference of hot pepper to other genomes or ESTs, hot pepper ESTs were compared to the sequences of 12 organisms from NCBI nr DB which include 8 different species of plants, human, *C. eligans*, *D. melanogaster*, *S. cerevisiae* (Figure 3). Although more than 58% of ESTs were highly homologous to plant genomes, about 3% of ESTs were homologous to non-plant sources. Since the size and diversity of information in dbEST is much larger than nr DB, searching sequence similarities in dbEST is likely to give more comprehensive results than searching nr DB. Total 8,525 hot pepper ESTs were clustered with ESTs of tomato, *Medicago truncatula*, potato, Arabidopsis, wheat, Maize, and rice. As we

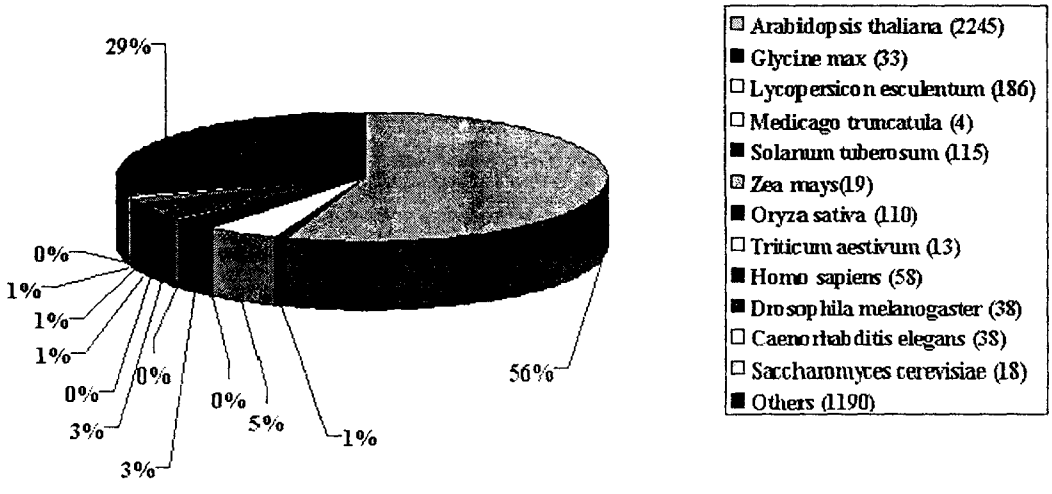


Figure 3. Inter-genome analysis. BLASTX searches were applied to find the similarity/difference between hot pepper and 12 organisms including 8 plants, 3 animals, and yeast. Each number (%) was generated according to BLASTX match. The cut-off value for this analysis was BLASTX score 35

expected, tomato has the highest number of hot pepper homologues (74%) and rice has the least homologues (56%; Figure 4). Then, we tried to identify tentative hot pepper-specific sequences. As a first step, we clustered hot pepper ESTs with other plant ESTs (Arabidopsis, Medicago, Maize, potato, rice, tomato and tomato) by using N2tool (cut-off threshold 100). After clustering un-clustered sequences were subjected to blastx analysis against NCBI nr DB, and the unmatched ESTs were considered as tentative hot pepper specific sequences. This analysis indicates that 7% (323 unique or clusters) of total clusters were possibly hot pepper-specific sequence candidates in current condition. This number could be exaggerated because of the characteristics of random EST sequencing. In other case, hot pepper genome may have more divergent genes compared to other plant species, which can be demonstrated if similar analysis is applied to other plant ESTs.

We also carried out comparative analysis of ESTs obtained from plants infected with various pathogens. In TIGR plant gene index (http://www.tigr.org/tdb/tgi.shtml), we could find three different ESTs information separately generated from *Phytophthora*-infected potato leaves

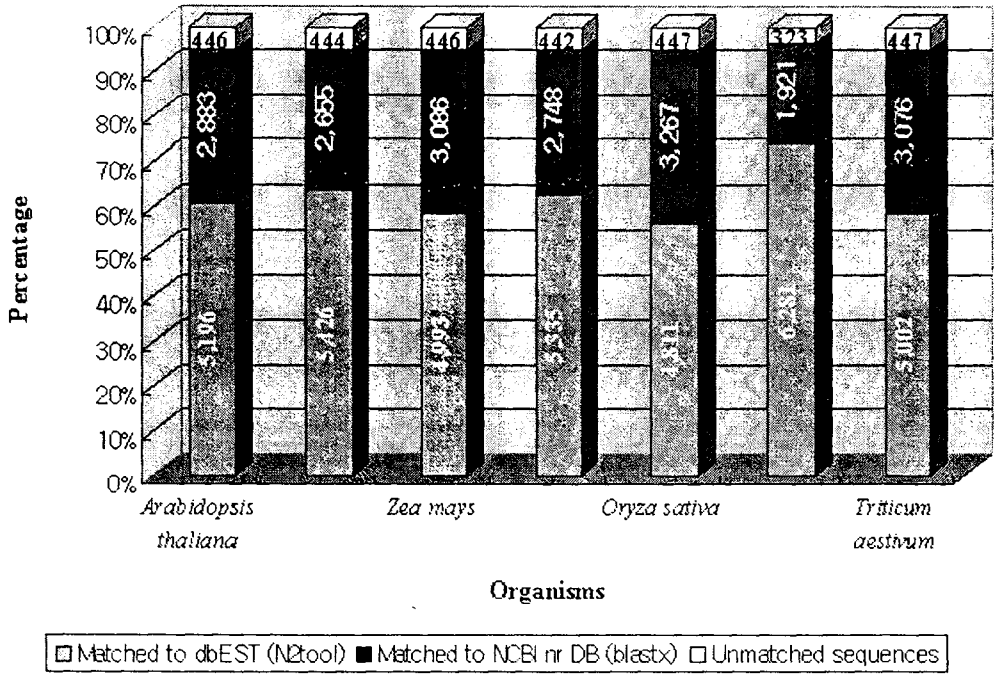


Figure 4. Candidate of hot pepper specific EST. Total 8,625 hot pepper ESTs were clustered separately with other 7 plant ESTs using N2Tool (threshold 100). The unclustered sequences were searched using BLASTX algorithm against NCBI nr database.

The final non-matched sequences (323 ESTs) were considered as hot pepper specific sequences and *Pseudomonas*-infected susceptible and resistant tomato leaves. When we compared the percentage of singleton and unique genes, Xcg-infected pepper leaves have about 10% more singleton and unique genes (Data not shown). In addition to ESTs generated from pathogen-infected plant tissues, we compared pepper ESTs with ESTs generated from nine different species of plants including tomato, rice, soybean, potato, medicago, maize, ice plant, and barley, which consist of 585,123 ESTs from 149 different libraries One distinctive point is that the proportion of singleton and unique sequences are higher in pepper than other plant organisms (Data not shown). Although most plant genome have relatively lower portion (below 50%) of single copy or unique DNA compared to animal genome, a study about pepper genome organization estimated that 65% of the pepper genome was composed of single-copy sequences (Walbot and Goldberg 1979; An et al. 1996). Based on An's previous work and our analysis data might indicate that higher portion of single and unique ESTs of pepper may imply the unique genome structure of hot pepper.

In conclusion, this work was the first genome level sequence analysis of hot pepper plant and this study indicates that pepper could have higher number of genes compared to other solanaceous plants and higher percentage of unique sequences, which do not match to the sequences of other plant species. In electronic Northern analysis, 136 EST clusters could be involved in plant defense mechanism, and their function will be pursuit. In addition to the sequence information and/or analysis, application of micro-array technologies will certainly accelerate our understanding of plant-pathogen interactions, and is being tried.

7. METHODS

가) Template Preparation and DNA Sequencing

The KS01 cDNA library was constructed from poly (A) RNA prepared from hot pepper (*Capsicum annuum* cv Bukang) leaves inoculated with soybean pustule pathogen, *Xanthomonas campestry* pv *glycine* (Suh et al 2001). Flowering buds (KS07) and anthers (KS08) libriaries were generated from *Capsicum annumm* cv HangKeun and purchased from Eugenetech (www.eugentech.com). Each cDNA library was transformed into *E. coli* SOLR strain (Clonetech, Palo Alto, California, USA) and 11,000 colonies were randomly picked using QPix (Genetix, New Milton, Hampshire, UK). Plasmid DNAs were prepared by using akaline lysis method and column purification. Automated fluorescence cycle sequencing reactions were done by using the ABI Bigdye cycle sequencing kit (PE Applied Biosystems) with T7 (5' -GTAATACGACTCACTATAGGG-3' ; for KS01 and KS08 libraries) and A1 primer (5' -CGCGTTTGGAATCACTACAGG-3' ; for KS07 library). These samples were ran and analyzed on ABI Prism 3700 DNA Analyzer.

나) Processing of the Sequences

All the following processes were done on SGI Origin 3200 Unix machine (SGI Korea, South Korea). The ABI formatted chromatogram sequences were fed into PHREP (Ewing et al., 1998). Subsequently, sequences that satisfy higher than 97% of quality and stretch more than 100 bp at the same time were collected for further analysis. Before the clustering of ESTs, these sequences were screened for the masking of vector sequences using RepeatMasker software (http://ftp.genome.washongton.edu/RM/RepeatMasker.html) at default parameters to screen out the interspersed repeats, vector and low-complexity DNA sequences that probably interfere with the clustering for their similar pattern. Our own developed program calculated the GC contents of ESTs. Possible contaminated sequences of non-pepper and non-cDNA were eliminated by searching BLASTN matches for the ESTs with strong homology (below E value e-10) to mitocondrial, chloroplast, and ribosomal RNA genes using Mendel DB (www.mendel.ac.uk.genomedb.html). Since the full-length cDNA candidates were divided into two types, we used two approaches. First, we found the appearance of Met (translation start codon) in the first position of subject sequence (homolog) of BLAST output. Second, we found the ESTs without appearing of translation start codon in subject sequence of BLAST output, but 5' end of query sequence (EST) have more than three times of nucleotide compared to the corresponding subject sequence. Then, we clustered ESTs into separated clusters with more than 100 bp of core sequences using CAP3 (http://genome.cs.mtu.edu/cap/cap3.html) and ICAtools (Parsons et al., 1995). The representative sequence of one cluster is parent sequence and the others are child sequences. In case of no child, they were classified as singleton.

다) Functional Classification of Hot Pepper EST and Comparative Analysis

In order to automatic process for functional classification of pepper ESTs, total 4685 unique ESTs were compared to arabidopsis sequences by using N2tool at threshold 100. Following the MIPS MatDB (http://mips.gsf.de/proj/thal/index.html) classification, 2650 ESTs were assigned to their functions. The same procedures were applied to functional cataloging of the abundantly expressed ESTs of KS01 library (total 136 ESTs).

For inter-genome analysis, both parent and singleton ESTs were searched against GenBank nr entries by BLASTX with e-value < e-20 and other parameters were default. Subsequently, the results were classified into 12 different groups according to the highest match with known genes of different organisms (Fig. 3). In contrast to BLASTX search to GenBank nr DB, we performed BLASTN search against 7 different species of plants ESTs in DB EST entries which include tomato, medicago, potato, arabidopsis, wheat, maize and rice, respectively. ESTs from each species of plant were mixed with pepper ESTs

generated in this study and clustered by using N2tool at threshold 100.

라) Acknowledgement

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. We thank to Drs. Chang-Hoon Kim, Jung-Keun Lee, and other members of genome informatics team in KRIBB for the computational analysis. We also thank to Dr. Xiaoqiu Huang at Dept. of Computer Science, Michigan Technological University for providing CAP3 program and Dr. David Gorden at Dept. of Molecular Biotechnology, University of Washington for providing Phred/Phrap/Consed package.

8. REFERENCES

Ablett, E., G. Seaton, K. Scott, D. Shelton, M.W. Graham, P. Baverstock, L.S. Lee, and R. Henry. 2000 Analysis of grape ESTs: global gene expression patterns in leaf and berry. *Plant Science.* 159: 87-95

Akkaya, M.S., A.A. Bhagwat, and P.P. Cregan. 1992. Length polymorphisms of simple sequence repeat DNA in soybean. *Genetics.* 132: 1131-1139

Allona, I., M. Quinn, E. Shoop, K. Swope, S.S. Cry, J. Carlis, J. Riedl, E. Retzel, M.M. Campbell, R. Sederoff, and R.W. Whetten, 1998. Analysis of xylem formation in pine by cDNA sequencing. *Proc Natl Acad Sci USA.* 95: 9693-9698

An, C.S., C.S. Kim, and S.L. Go. 1996. Analysis of red pepper (*Capsicum annuum*) genome. *J. Plant Biol.* 39: 57-61

Becker, J., R. Kempf, W. Jeblick, and H. Kauss. 2000. Induction of competence for elicitation of defense responses in cucumber hypocotyls requires proteasome activity. *Plant J.* 21: 311-316

Blatt, M.R., A. Grabov, J. Brearley, K. Hammond-Kosack, and J.D. Jones. 1999. K+ channels of Cf-9 transgenic tobacco guard cells as targets for *Cladosporium fulvum* Avr9 elicitor-dependent signal transduction. *Plant J.* 19: 453-462

Bowles, D.J. 1990. Defense-related proteins in higher plants. *Annu Rev Biochem.* 59: 873-907

Cardle, L., L. Ramsay, D. Milbourne, M. Macaulay, D. Marshall, and R. Waugh. 2000. Computational and experimental characterization of physically clustered simple sequence repeats in plants. *Genetics.* 156: 847-854

Cooke, R., M. Raynal, M. Laudie, F. Grellet, M. Delseny, P.C. Morris, D. Guerrier, J. Giraudat, F. Quigley, G. Clabault, Y.F. Li, R. Mache, M. Krivitzky, et al. 1996. Further progress towards a catalogue of all *Arabidopsis* genes: analysis of a set of 5000 non-redundant ESTs. *Plant J.* 9: 101-124

Covitz, P.A., L.S. Smith, and S.R. Long. 1998. Expressed sequence tags from a root-hair-enriched *Medicago truncatula* cDNa library. *Plant Physiol.* 117: 1325-1332

Dangl, J.L. and J.D. Jones. 2001. Plant pathogens and integrated defence responses to infection. *Nature.* 411: 826-833

Dixon, R.A. 1986. Phytoalexin response; Elicitation, signaling, and control of host gene expression. *Biol. Rev. Camb. Philos. Soc.* 61: 239-292

Dixon, R.A. 2001. Natural products and plant disease resistance. *Nature.* 411: 843-847

Dong, X. 1998. SA, JA, ethylene, and disease resistance in plants. *Curr Opin Plan Biol.* 1: 316-323

El-Maarouf, H., M.A. Barny, J.P. Rona, and F. Bouteau. 2001. Harpin, a hypersensitive response elicitor from *Erwinia amylovora*, regulates ion channel activities in *Arabidopsis thaliana* suspension cells. *FEBS Lett.* 497: 82-84

Ewing, B., L. Hiller, M.C. Wendle, and P. Green. 1998. Base-calling of automated sequencer traces using *Phred.* I Accuracyy assesement. *Genome Res.* 8: 175-185
Ewing, B. and P. Green. 1998. Basecalling of automated sequencer traces using *phred.* II. Error probabilities. *Genome Res.* 8: 186-194

Ewing, R.M., A.B. Kahla, O. Poirot, F. Lopez, S. Audic, and J.M. Claverie. 1999. Large-scaling statistical analysis of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9: 950-959

Genoud, T. and J.P. Metraux. 1999. Crosstalk in plant cell signaling: structure and function of the genetic network. *Trends Plant Sci.* 3: 141-146

Huang, X. and A. Madan. 1999. CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877

Jabs, T., M. Tschope, C. Colling, K. Hahlbrock, and D, Scheel. 1997. Elicitor-stimulated ion fluxes and O2- from the oxidative burst are essential

components in triggering defense gene activation and phytoalexin synthesis in parsley *Proc Natl Acad Sci U S A* **94**: 4800-4805

Kende, H. 1993. Ethylene biosynthesis. *Ann Rev. Plant Pkysiolo. Plant Mol. Biol.* **44**: 283-307

Kim, S., S.R. Kim, C.S. An, Y.N. Hong, and K.W. Lee. 2001. Constitutive expression of rice MADS box gene using seed explants in hot pepper (*Capsicum annuum* L.). *Mol. Cells.* 12: 221-226

Kwak, J.M., S.A. Kim, S.W. Hong, and H.G. Nam. 1997. Evaluation of 515 expressed sequence tags obtained from guard cells of *Brassica campestris*. *Planta.* 202: 9-17

Lam, E, P.N. Benfley, P.M. Gilmartin, R.X. Fang, and N.H. Chua. 1989. Site specific mutations alter *in vivo* factor binding and change promoter expression pattern in transgenic plants. *Proc Natl Acad Sci.* 86: 7890-7894

Lee, J., D.F. Klessig, and T. Nurnberger. 2001. A harpin binding site in tobacco plasma membranes mediates activation of the pathogenesis-related gene hin1 independent of extracellular calcium but dependent on mitogen-activated protein kinase activity. *Plant Cell.* 13: 1079-93

Levine, A., R.Tenhaken, R. Dixon, and C. Lamb. 1994 H_2O_2 from the oxidative burst orchestrates the plant hypersensitive disease resistance response. *Cell.* 79: 583- 593

Liu, Z. W., R.M. Biyashev, and M.A. Saghai Maroof. 1996. Development of simple sequence repeat markers and their integration into a barley linkage map. *Theor. Appl. Gent.* 93: 869-876

Maleck, K. and R.A. Dietrich. 1999. Defense on multiple fronts: how do plants cope with diverse enemies? *Trend Plant Sci.* 4: 215-219

Mekhedov, S., O.M. de Ilarduya, and J. Ohlrogge. 2000. *Plant Physiol.* 122: 384-401

Milbourne, D., R.C. Meyer, A.J. Collins, L.D. Ramsay, C. Gebhardt et al. 1998. Isolation, characterization and mapping of simple sequence repeat loci in potato. *Mol. Gen. Genet.* 259: 233-245

Morgante, M. and A.M. Olivieri. 1993. PCR-amplified SSRs as markers in plant genetics. *Plant J.* 3: 175-182

Oh, B.J., M.K. Ko, I. Kostenyuk, B. Shin, and K.S. Kim. 1999. Coexpression of a defensin gene and a thionin-like via different signal transduction pathways in pepper and *Colletotrichum gloeosporioides* interactions. *Plant Mol. Biol.* 41: 313-319

Parsons, J.D., S. Brenner, and M.J. Bishop. 1992. Clustering cDNA sequences. *Comput. Applic Biosci.* 11: 603-613

Peck, S.C. and H. Kende 1998. A gene encoding 1-aminocyclopropane-1-carboxylate (ACC) synthase produces two transcripts: elucidation of a conserved response. *Plant J.* 14: 573-581

Qutob, D., P.T. Hraber, B.W.S. Sobral, and M. Gijzen. 2000. Comparative Analysis of expressed sequences in *Phytophthora sojae. Plant Physiol.* 123: 243-253

Roberts, M.R. and D.J. Bowles.1999. Fusicoccin, 14-3-3 proteins, and defense responses in tomato plants. *Plant Physiol.* 119: 1243-1250

Ryal, J.A., U.H. Neuenschwander, M.G. Willits, A. Molina, H.Y. Steiner, and M.D. Hunt. 1996. Systemic acquired reisistance. *Plant cell.*8: 1809-1819

Sasaki, T., J. Song, Y. KogaBan, et al. 1994. Toward cataloguing all rice genes: large-scale sequencing of randomly chosen rice cDNA from a callus cDNA library. *Plant J.* 6: 615-624

Senior, M. L., E.C.L. Chin, M. Lee, J.S.C. Smith, and C.W. Stuber. 1996. Simple sequence repeat markers developed from maize sequences found in the GENEBANK database: map construction. *Crop Sci.* 36: 1676-1683

Somssich, I.E. and K. Hahlbrock. 1998. Pathogen defense in plants- a paradigm of biological complexity. *Trends Plant Sci.* 3: 86-90

Suh, M.C., S.Y. Yi, S. Lee, W.S. Sim, H.S. Pai, and D. Choi. 2001 Pathogen-induced expression of plant ATP:citrate lyase. *FEBS Letters.* 488: 211-212

Tautz, D. 1989. Hypervariability of simple sequences as a general source of polymorphic DNA markers. *Nucleic Acids Res.* 17: 6463-6471

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 408: 796-815

Thomma, B.P., I.A. Penninckx, W.F. Broekaert, and B.P. Cammue. 2001. The complexity of disease signaling in Arabidopsis. *Curr Opin Immunol.* 13: 63-68

Ujino-Ihara, T., K. Yoshimura, Y. Ugawa, H. Yoshimaru, K. Nagasaka, and Y. Tsumura. 2000. *Plant Mol. Biol.* 43: 451-457

Van de Loo, F.J., S. Turner, and C. Somerville. 1995. Expressed sequence tags from developing castor seeds. *Plant Physiol.* 108: 1141-1150.

Venter, J.C., M.D. Adams, E.W. et al. 2001. The sequence of the human genome. *Science.* 291: 1304-135151.

Walbot, V. and R.B. Goldberg. 1979 Plant genome organization and its relationship to classical plant genetics. *In* Nucleic Acids in Plant. T. C. Hall and T. W. Davis (eds.) CRC Press Inc., Boca Rata, pp. 3-40

Ward, E. R., S.J. Uknes, S.C. Williams, S.D. Dincher, D.L. Wiederhold, D.C. Alexander, P. Ahl-Goy, J.P. Metraux, and J.H. Ryal. 1991. Coordinate gene activity in response to agent that induce systemic acquired resistance. *Plant Cell.* 3: 1085-1094

Whitbred, J.M. and M.A. Schuler. 2000. Molecular characterization of CYP73A9 and CYP82A1 P450 genes involved in plant defense in pea. *Plant Physiol.* 124: 47-58

Wu, G., B.J. Shortt, E.B. Lawrence, E.B. Levine, K.C. Fitzsimmons, and D.M. Shah. 1995. Disease resistance conferred by expression of a gene encoding H202- generating glucose oxidase in transgenic potato plants. *Plant Cell.* 7: 1357-1368

Yoon, I.S., H. Mori, J.H. Kim, B.G. Kang, and H. Imaseki. 1997. VR-ACS6 is an auxin-inducible 1-aminocyclopropane-1-carboxylate synthase gene in mungbean (Vigna radiata). *Plant Cell Physiol.* 38: 217-224

Zhang, L., X.L. Ma, Q. Zhang, C.L. Ma, P.P. Wang, Y.F. Sun, Y.X. Zhao, and H. Zhang. 2001. Expressed sequence tags from a NaCl-treated *Suaeda salsa* cDNA library. *Gene.* 267: 193-200

Zhao, X.P. and G. Kochert. 1992. Characterization and genetic mapping of a short, highly repeated, interspersed DNA-sequence from rice (*Oryza sativa* L.). *Mol. Gen. Genet.* 231: 353-359

Zimmermann, S., T. Nurnberger, J.M. Frachisse, W. Wirtz, J. Guern, R. Hedrich, and D. Scheel.1997. Receptor-mediated activation of a plant Ca(2+)-permeable ion channel involved in pathogen defense. *Proc Natl Acad Sci U S A.* 94: 2751-2755

Zhu, Y.X., Ou-Yang, W.J. Zhang, Y.F, and Chen, Z.L. 1996. Transgenic sweet pepper plants from *Agrobacterium* mediated transformation. *Plant Cell Rep.* 16: 71-75

제 4 장 목표달성도 및 관련분야에의 기여도

가. 연구목표 달성도

연차	연구목표	연구 내용 및 범위	달성도 (%)
1 차 년도 (2000)	- 식물 유용 유전자 DB 구축 및 EST 분석시스템 개발	-cDNA library 제작 및 연구 시스템 구축 -저항성 관련 식물 EST 의 대량발굴 -Plant EST DB 구축 및 자동분석시스템 -사용자 인터페이스 개발 -대규모 유전자발굴을 위한 연구기반구축 -잎 조직의 cDNA library 에서 6,600 EST 염기서열 발굴 -꽃 조직에서 2,000 EST 염기서열결정 -화분(anther) 조직에서 2,000 EST 염기서열 결정 -EST clustering 및 multialignment 분석 시스템 구축 - Internal BLAST 시스템구축	100
2 차 년도 (2001)	- Gene Expression profiling 시스템 개발	-Gene expression profile 을 위한 Algorithm -고추 flower bud EST 발굴(4,235 개) -고추 어린열매 EST(4,055 개) 발굴 -고추 뿌리 EST(4,320 개) 발굴 -고추 태좌 EST(4,279 개) 발굴 -고추 EST DB(총 12,526 unigene) 구축 -Web 을 통한 EST DB 공개 -인삼 EST(6,384 개 unigene) DB 구축 -참깨 EST DB 구축 및 공개 -고구마 EST DB 구축 및 공개 -Gene expression profile 을 위한 정보처리기술 개발	100
3 차 년도 (2002)	- EST 및 DNA Chip 정보 분석을 위한 통합 시스템 개발	-타 과제 도출 EST 의 D/B 화 -Plant Gene Index 구축 -Arabiopdis 를 비롯한 8 종의 식물에 대한 Organism specific 한 EST 분석 및 서비스 -Plant Gene Index 의 웹을 통한 공개 -유전자 칩 데이터 베이스구축 -유전자 클론 분양 -Gene Expression 및 관련정보 분석 DB -웹 인터페이스를 이용한 정보분석서비스	100

나. 연구개발 성과물 목록

논문(14 건) :

1. Lee. S., and Choi, D. (2001). Toward functional genomics of plant-pathogen interactions: Isolation and analysis of defense-related genes of hot pepper expressed during resistance against pathogen. Plant Pathology Journal 18:63-67.
2. Cho, H.-S., Yun, K.-M., Lee, S.-S., Kim, Y.-A., Hwang, I., Choi, D., and Pai H.-S. (2001). A novel dual-specificity protein kinase targeted to the chloroplast in tobacco. FEBS Letters 497:124-130.
3. Kim, Y.-C., Yi, S.-Y., Mang, H.-K., Seo, Y.-S., Kim, W.-T, and Choi, D. (2002). Pathogen-induced expression of cyclo-oxygenase homologue in hot pepper (*Capsicum annuum* cv. Pukang). Journal of Experimental Botany 53:383-385.
4. Lee, S.-J., Lee, M.-Y., Yi, S.-Y., Oh, S.-K., Choi, S.-H., Hur, N.-H., Choi, D., Min, B.-H., Yang, S.-K., and Harn, C.-H. (2002). PPI1: a novel pathogen-induced basic region-leucine zipper(bZIP) transcription factor from pepper. Molecular Plant Microbe-Interaction 15:540-548.
5. Yi, S.-Y., Yu, S.-H., and Choi, D. (2003). Involvement of hydrogen peroxide in repression of catalase in resistant tobacco following tobacco mosaic virus infection. Molecules and Cells 15(3): in press
6. Chung, E.-S., Kim, S.-Y., Yi, S.-Y., and Choi, D. (2003). *Cadhn*(*Capsicum annuum* dehydrin), an osmotic-stress gene in hot pepper plants. Molecules and Cells 15(3): in press
7. You. M.K., Hur, C.-G., Ahn, Y.S., Suh, M.C., Jeong, B.C., Shin, J.S., Bae, J.M. (2003) Identification of genes possibly related to storage root induction in sweetpotato. *FEBS letter* 536:101-105.
8. Suh*, M.C., Bae, J.M., Hur, C.G., Kang, C.W., Ohlrogge, J.B. (2003) Comparative analysis of Expressed Sequence Tags between *Sesamum indicum* and *Arabidopsis thaliana* developing seeds. *Plant Molecular Biology* (in press).
9. 임 소형, 신 민수, 박 경희, 고 성호, 허 철구* (2000) EST 를 이용한 Tissue expression 시스템 구축 한국정보과학회지, 8 월호 특집.
10. 허 철구 (2000) 분산 환경에서 Peptide Mass Mapping 에 의한 단백질 검증 시스템 설계 및 구. 한국멀티미디어학회, 제 3 권 2 호,571-574
11. 양 영렬, 허 철구* (2001) DNA Chip 데이터분석을 위한 유전자 발현 통합 프로그램 개발. 한국생물공학회지, Vol. 16, No 4.381-388.
12 양 영렬, 허 철구* (2001) DNA Chip 통합분석 프로그램을 이용한 효모의 세포주기 유전자 발현 데이터의 분석",(2001), 한국생물공학회지, Vol.16, No 6, 538-546.

13. 권경훈, 김승일, 김경욱, 김은아, 조건, 김진영, 김영환, 양덕춘, **허철구**, 유종신, 박영목 (2002) 인삼 모상근 프로테옴 데이터 분석:인삼 EST database 와의 통합 분석에 의한 단백질 동정. *한국식물공학회지*, Vol.29, No.3 p.161-170
14. Sung-Ho Goh, Tae-Hyung Kim, Jee-Hyub Kim, Dougu Nam, Doil Choi, Cheol-Goo Hur*(2003) Computational analysis of neighboring genes on *Arabidopsis thaliana* chromosomes 4 and 5: Their genomic association as unctional subunits. Genomics and Informatics 1:(in press).

특허

- 출원(1 건)

명칭: 담배에서 분리된 병저항성 반응시 특이적으로 발현되는 NgCDM1 단백질, 그를 코딩하는 유전자 및 프로모터
발 명 자: 최도일, 김영철, 정영희, 이상협
출원일자: 2002. 11. 4.
출원번호: 2002-67957
출원국가: 대한민국

생물정보 프로그램개발 (4 건)

1. PyFACT :A Tool for Function Assignment and Classification to a Sequence using a Dictionary-Based Approach.
2. MiDAS(가칭) :
   1)GUI 를 갖춘 R 기반의 cDNA 마이크로 어레이 분석 시스템.
   2) 데이터의 품질 평가, 전 처리 과정, 발현 유전자의 선정, clustering 및 classification 등 마이크로 어레이 데이터 분석의 전 과정을 Pile-line 형식으로 지원.
3. PMS : Peptide Mass Spectrum Analysis
4. EST 분석 시스템

제 5 장 연구개발결과의 활용계획

가. 발굴된 유전자의 활용계획

자생식물이용기술개발 사업 1 단계를 통하여 인삼, 고추, 고구마, 개똥쑥, 참깨 및 야생종 담배를 통털어 발굴된 유전자는 약 6 만 여 개에 이르며 독립(independent) 유전자를 기준으로 해도 약 3 만여 개에 달함. 현재 모든 유전자는 일련의 분석과정을 거쳐 자생식물 사업단의 홈페이지 (http://plant.pdrc.re.kr)에 데이터 베이스로 구축되어 있으며 사용자 편의에 따라 Key Word 또는 다양한 Category 를 이용하여 검색할 수 있게 되어 있으며 각 연구자에게 요청하여 실물 유전자를 획득할 수 있어 향 후 식물의 기능유전체 연구에 유용하게 활용 될 것임. 특히 고추의 유전자의 경우는 5 천 개의 유전자가 심겨진 유전자 칩을 작물기능유전체연구사업단과 공동으로 개발하였으며 유전자발현 프로파일을 대량생산하여 데이터베이스화 된 상태로 당장 유전자의 기능을 유추할 수 있는 데이터가 제공되고 있으므로 고추연구자가 언제든지 직접 활용 할 수 있는 체계를 갖추고 있음.

나. 데이터 베이스 및 생물정보처리기술의 활용

본 과제를 통하여 식물유전정보처리의 근간이 되는 EST 분석, DB 화 및 유전자발현 프로파일분석 기술 등이 개발되었으며, 이는 향후 식물유전체기능연구에 광범위하고 지속적으로 활용될 수 있을 것으로 판단됨. 특히 국내에서는 처음으로 유전자칩 데이터 프로세싱기술과 발굴된 데이터를 데이터베이스화하여 공개함으로써 앞으로 다양하게 발굴되는 유전자발현 프로파일의 데이터베이스를 구축하는 모델로 활용될 수 있을 것으로 전망함.

제 6 장 연구개발과정에서 수집한 해외과학기술정보

특기사항 없음

제 7 장 참고문헌

특정연구개발사업 연구결과 활용계획서

<table>
<tr><td rowspan="2">사업명</td><td>중사업명</td><td colspan="5">21C 프론티어연구개발사업</td></tr>
<tr><td>세부사업명</td><td colspan="5">자생식물이용기술개발사업</td></tr>
<tr><td colspan="2">과제명</td><td colspan="5">자생식물 유전자 D/B 구축 및 유전정보처리기술 개발</td></tr>
<tr><td colspan="2">연구기관</td><td colspan="2">한국생명공학연구원</td><td>연구책임자</td><td colspan="2">최 도 일</td></tr>
<tr><td colspan="2">총연구기간</td><td colspan="5">2002 년 7 월 1 일 ~ 2003 년 6 월 30 일 (36 개월)</td></tr>
<tr><td colspan="2" rowspan="2">총 연구비
(단위 : 천원)</td><td>정부출연금</td><td colspan="2">민간부담금</td><td colspan="2">합계</td></tr>
<tr><td>2,400,000</td><td colspan="2">0</td><td colspan="2">2,400,000</td></tr>
<tr><td colspan="2">기술분야</td><td colspan="5">생명공학</td></tr>
<tr><td colspan="2">참여기업</td><td colspan="5">해당사항 없음</td></tr>
<tr><td colspan="2">공동연구기관</td><td colspan="5">해당사항 없음</td></tr>
<tr><td colspan="2">위탁연구기관</td><td colspan="5">해당사항 없음</td></tr>
<tr><td colspan="2" rowspan="2">연구결과활용
(해당항목에(∨)
표시)</td><td>1. 기업화 ( )</td><td>2. 기술이전( )</td><td>3. 후속연구추진
(∨)</td><td colspan="2">4.타사업에
활용( )</td></tr>
<tr><td>5. 선행 및 기초
연구(∨)</td><td>6. 기타목적활용
(교육,연구)(∨)</td><td>7. 활용중단
(미활용)( )</td><td colspan="2">8. 기타( )</td></tr>
</table>

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

첨부 : 1. 연구결과 활용계획서 1 부.
2. 기술요약서 1 부

2003 년 9 월 4 일

연구책임자 : 최 도 일 (인)
연구기관장 : 양 규 환 (직인)

과학기술부장관 귀하

[첨부 1]

연구결과 활용계획서

1. 연구목표 및 내용

가. 연구목표

본 과제는 자생식물유래 유전정보의 대량발굴 및 대량 발굴된 유전정보를 처리할 수 있는 기술을 개발하여 본 사업단의 제 3 분야 과제 및 국가적인 식물게놈 연구를 위한 서비스 체계를 구축하는 것이 최종목표로 이를 위하여 두 가지 줄기의 연구를 수행할 것을 계획하였음.

나. 연구내용

정보처리프로그램 개발에 필요한 대규모 데이터 발굴로 식물의 비기주 저항성에 관여하는 유전자를 목표로 EST 의 대량발굴을 수행 할 예정이었으며 (1 단계 3 년에 걸쳐 약 20,000 개의 EST 발굴), 발굴된 유전자를 DNA chip 에 microarray 하여 유전자칩을 이용한 유전자의 대량발현 연구를 수행 하고자 함. 이렇게 얻어진 데이터는 공동연구과제인 유전정보처리기술 개발을 위한 Low Data 로 사용되어 자생식물 유전자종합 분석 시스템을 개발하는 재료로 사용될 것임. 공동연구를 통하여 발굴된 데이터 처리를 위하여 대규모 자생식물 유전정보 데이터베이스 구축, EST clustering system, 대량의 BLAST 분석 시스템, 생물정보 분석프로그램의 병렬연결을 이용한 종합 생물정보검색체계 (motif search, gene function prediction) 및 칩을 이용한 유전자 대량발현 시스템의 분석프로그램 등을 개발하여 국가 식물게놈연구의 기반이 되는 기술을 개발하여 공공서비스를 수행할 예정이었음. 자생식물 유전정보처리 기술은 검색체계가 개발되는 대로 자생식물 이용기술 개발사업단의 홈페이지를 이용하여 우선 사업단 과제를 수행하는 연구자 그리고 점차로 국내의 모든 연구자를 대상으로 공공 생물정보 분석 서비스를 수행하고자 함.

2. 연구수행결과 현황(연구종료시점까지)

가. 특허(실용신안) 등 자료목록

발명명칭	특허공고번호 출원(등록)번호	공고일자 출원(등록)일자	발명자 (출원인)	출원국	비고
담배에서 분리된 병저항성 반응시 특이적으로 발현되는 NgCDM1 단백질, 그를 코딩하는 유전자 및 프로모터	2002-67957	2002.11.04	최도일, 김영철, 정영희, 이상협	한국	
스트레스 저항성 전사인자 유전자, 단백질 및 이에 의해 형질전환된 스트레스 저항성 식물체	2003-20269	2003.03.31	최도일, 오상근, 박정미, 정영희, 이상협	한국	
스트레스 저항성 전사인자 유전자, 단백질 및 이에 의해 형질전환된 스트레스 저항성 식물체	2003-28792	2003.05.07	최도일, 이소영, 박정미, 정영희, 이상협	한국	

나. 프로그램 등록목록

프로그램 명칭	등록번호	등록일자	개발자	비고

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	SCI 게재 여부
		년 월 일				
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

해당사항 없음

4. 기술이전 및 연구결과 활용계획

가. 당해연도 활용계획 및 방법

-자생식물이용기술개발 사업 1 단계를 통하여 발굴된 유전자는 약 6 만 여 개에 이르며 독립(independent) 유전자를 기준으로 해도 약 3 만여 개에 달함. 현재 모든 유전자는 일련의 분석과정을 거쳐 자생식물 사업단의 홈페이지 (http://plant.pdrc.re.kr)에 데이터 베이스로 구축되어 있으며 사용자 편의에 따라 Key Word 또는 다양한 Category 를 이용하여 검색할 수 있게 되어 있으며 각 연구자에게 요청하여 실물 유전자를 획득할 수 있게 하였다.

-5 천 개의 고추 유전자가 심겨진 유전자 칩을 작물기능유전체연구사업단과 공동으로 개발하였으며 이를 이용하여 유전자발현 프로파일을 대량생산할 수 있게되었다.

-본 과제를 통하여 식물유전정보처리의 근간이 되는 EST 분석, 데이터베이스화 및 유전자발현 프로파일분석 기술 등이 개발되었으며 이를 이용한 분석이 활발하게 이루어지고 있다.

나. 차년도 이후 활용계획

-유전자 칩을 이용하여 생산된 유전자발현 프로파일 데이터베이스는 고추 연구자가 언제든지 직접 활용 할 수 있는 체계를 갖추고 있고 또한, 당장 유전자의 기능을 유추할 수 있는 데이터가 제공되고 있으므로 향 후 식물의 기능 유전체 연구에 유용하게 활용 될 것임. 특히 국내에서는 처음으로 유전자 칩 데이터 프로세싱기술과 발굴된 데이터를 데이터베이스화하여 공개 함으로써 앞으로 다양하게 발굴되는 유전자발현 프로파일의 데이터베이스를 구축하는 모델로 활용될 수 있을 것으로 전망함.

5. 기대효과

대규모 유전자 발굴을 위한 연구 시스템, Gene Expression profile 및 분석 기술을 개발함으로서 식물 유전체 연구의 기반을 마련하였으며, 이런 연구에 의해 만들어진 데이터를 Web 상으로 공개하고 여러 가지 분석 서비스를 함으로서 국내 연구자들에게 다양한 정보를 제공할 수 있게 되었고 이로 인한 식물생명공학연구를 활성화를 기대할 수 있겠다. 특히 fuctional genomis, metabolomis, proteomics 등의 연구 발전에 크게 기여 할 수 있는 기반을 마련하였다. 또 자체적인 데이터베이스 확립과 프로그램 개발, 또 web. 상에서 이 정보들을 공개함으로서 국내뿐 아니라 국제적으로도 연구발전에 기여할 수 있게 되었고 이로 인해 국제무대에서의 한국식물생명공학의 위상은 한층 더 높아질 수 있음을 예상할 수 있겠다.

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

[첨부 2]

기술 요약서

1. 기술의 명칭

※기술이란? 과제 수행결과 확보된 신기술, 산업재산권, 기술적 노하우 등 개발된 성과중 수요자에게 공급할 수 있는 형태의 기술을 의미함

2. 기술을 도출한 과제현황

과제관리번호				
과제명	자생식물 유전자 D/B 구축 및 유전정보처리기술 개발			
사업명	21C 프론티어연구개발사업			
세부사업명	자생식물이용기술개발사업			
연구기관	한국생명공학연구원		기관유형	정부출현연
참여기관(기업)				
총연구기간	3년			
총연구비	정부(2,400,000)천원 민간(0)천원 합계(2,400,000)천원			
연구책임자 1	성명	최 도 일	주민번호	
	근무기관 부서	식물유전체연구실	E-mail	doil@kribb.re.kr
	직위/직급	책임연구원	전화번호	042-860-4340
연구책임자 2	성명		주민번호	
	근무기관 부서		E-mail	
	직위/직급		전화번호	
실무연락책임자	성명		소속/부서	
	직위/직급		E-mail	
	전화번호		FAX	
	주소	(-)		

3. 기술의 주요내용

[기술의 개요]

[기술적 특징]

(1)

(2)

(3)

[용도· 이용분야]

(1)

(2)

(3)

■ 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	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				

주 의

1. 이 보고서는 과학기술부에서 시행한 특정연구개발사업의 연구보고서입니다.
2. 이 보고서 내용을 발표할 때에는 반드시 과학기술부에서 시행한 특정연구개발사업의 연구결과임을 밝혀야 합니다.
3. 국가과학기술 기밀유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안됩니다.