식물생명공학회지 제33권 제3호
J Plant Biotechnol
Vol. 33, No. 3, 209-222 (2006)

식물 생명공학과 생물정보학

김정은[1], 백효정[1], 김영철[1], 허철구[1*]
[1]한국생명공학연구원 식물유전체연구센터

Plant Biotechnology and Bioinformatics

JungEun Kim[1], Hyojung Paik[1], Young-Cheol Kim[1], and Cheol-Goo Hur[1*]
[1]Plant Genome Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-333, Korea

ABSTRACT The whole genome sequence was completed in arabidopsis and rice. Large amounts of EST data have been available from many other plants. Also, vast quantities of diverse biological data have been generated by various '-omics' technologies such as transcriptomics, proteomics, and metabolomics. Bioinformatics plays an essential role in extracting useful information from these tremendous amounts of biological data. In this review we introduced experimental methods to generate massive data, applications to plant science such as plant disease resistance and molecular breeding and bioinformatics tools and web sites available in plant biotechnology R&D. We concluded that new experimental methods and bioinfomation analysis techniques have made major contributions to the development of plant biotechnology and that bioinformatics has become a critical factor in plant biotechnology R&D.

서 론

생명현상의 비밀을 담고 있는 DNA의 구조가 왓슨과 크릭에 의하여 발견 (Watson and Crick 1953)된 이후로, 생명공학은 분자 생물학과 생화학을 기반으로 다양한 생명체로부터 생명현상을 이해하기 위한 연구가 급속도로 발전하고 있다. 이런 생명현상에 대한 궁금증을 해결하기 위해 과거에는 한 유전자가 암호화하고 있는 정보와 그 산물이 다양하고 복잡한 대사 과정에서 어떤 역할을 수행하는지를 규명하는 것이 주된 목적이었다. 그러나 생명 과학 기술의 발달로 인하여 다양한 생명체로부터 대량의 생물학적 데이터를 얻을 수 있게 됨으로써 생명체를 전체적 이해를 위한 연구가 시도되고있다. 생물정보학에서는 대량의 데이터를 효율적으로 가공, 저장, 분석하기 위한 컴퓨터공학, 통계, 수학적 모델을 제시하고 있으며 (Benton 1996), 그 결과, 종 간의 비교 분석, 생명체 내에서 상호관계를 고려한 생명현상을 분석을 함으로써 복합적으로 해명하고자 하였다. 이러한 연구 방향은 식물 생명공학 분야에서도 예외 없이 적용되고 있으며, 이젠 많은 양의 데이터를 생성, 분석하는 것이 식물 생명공학을 연구에 중요하게 사용되고 있다.

식물에서 애기장대와 벼의 전체 게놈 염기서열 분석이 완료 (The Arabidopsis Genome Initiative 2000, Goff et al. 2002)되었고, 토마토, 옥수수, *Medicago truncatula* (Cannon et al. 2005)와 같은 중요한 식물들의 게놈 염기서열 분석이 진행되고 있다. 또한, 다양한 식물 종으로부터 엄청난 양의 expressed sequence tag (EST)들이 공개되어 이용되고 있다 (http://www.tigr.org, http://www.ncbi.nlm.nih.gov). 특히 우리나라는 고추와 같은 중요 작물에 대한 많은 EST 데이터를 생산 및 확보하고 있으며, 서열 기반의 다양한 정보 분석 기술을 토대로 토마토 전체 게놈 염기서열 결정 및 분석을 위한 국제 컨소시엄에서 chromosome 2번을 맡아 분석함으

*Corresponding author Tel 042-879-8560 Fax 042-879-8569
E-mail: hurlee@kribb.re.kr

로써 식물 genomics 연구에 참여하고 있다 (Lee et al. 2004, http://genepool.kribb.re.kr/new/). EST와 게놈 염기서열들은 각 유전자의 발현을 하나씩 실험하던 고전적 연구 방법을 고집적 DNA chip을 이용하여 수 백개에서 수 만개에 이르는 유전자의 발현 양상을 동시에 분석할 수 있게 하였다 (Richmond and Somerville 2000, Seki et al. 2004). 이와 더불어 high-throughput 분석은 단백질과 metabolite 수준에서도 이루어져 다양한 proteome 및 metabolome 분석이 식물 연구에서도 수행되기 시작하였다 (Park 2004, Nobeli and Thornton 2006). 이러한 데이터는 GenBank와 SwissProt (http://www.expasy.org/sprot/)과 같은 공용 데이터베이스에 축적되어 있으며, 일부 분석 도구와 함께 world-wide-web 상에서 다운로드 받아 사용할 수 있다.

우리는 아래에서 식물 생명공학의 연구에 있어서 많은 양의 생물학적 정보를 생산하는 실험적 방법들과 이런 데이터들을 생물정보학적으로 분석하고 이용할 수 있는 몇몇 구체적인 사례들을 기술하고자 한다. 첫 번째로, 식물 연구에서 대량의 데이터 생성에 필요한 실험 방법들을 소개하고, 두 번째로 얻어진 생물학적 정보를 분석한 결과를 이용할 수 있는 예로 식물 병과 분자 육종의 예를 들어 설명할 것이다. 그리고 마지막으로 얻어진 방대한 양의 생물학적 정보를 생물정보학적으로 분석하는 방법과 식물 생명공학 연구에 도움이 될 유용한 데이터베이스들을 소개할 것이다.

식물 연구에서의 기능 유전체학

1990년 대에 시작한 Human Genome Project가 예상보다 빨리 완성되면서, 식물에서도 모델 식물인 애기장대와 중요 작물인 벼를 이용한 전체 게놈 염기서열 분석이 완료되었고 (The Arabidopsis Initiative 2000, Goff et al. 2002, Yu et al. 2002), 다른 토마토, 옥수수, *Medicago truncatula* (Cannon et al. 2005)등의 식물에서도 전체 게놈 염기 서열 분석이 진행중이다. 전체 게놈 서열 데이터는 물리적 지도 (physical map)을 작성함으로써 한 생명체의 게놈에 존재하는 유전자의 분포와 특징을 분석하고, 다른 종과 비교하여 종간의 특이성과 연관성를 밝히기위한 연구에 이용되고 있다. 또 다른 genome level의 염기서열 분석 중인 하나인 EST는 1991년부터 연구되기 시작하였으며 (Adams et al. 1991), 전체 게놈 염기서열 결정에 비해 시간적, 경제적 비용이 적게 드는 반면, gene finding, gene expression level 분석, 상황 또는 조직 특이적 유전자 분석 등 다양한 분석이이 가능하기 때문에 여러 식물체에서 데이터가 만들어지고 있다 (http://www.ncbi.nlm.nih.gov/dbEST/, http://www.tigr.org). 기능 유전체학 (functional genomics)에서는 이렇게 만들어진 서열 데이터를 이용하여 전체 게놈에 존재하는 유전자들의 기능을 연구하고자 하며, 이를 위하여 다양한 생물정보학적 분석과 생물학적인 실험이 동시에 요구된다. 기능 유전체학은 proteomics, metabolomics, DNA microarray를 이용한 expression profiling 등을 통해 전체 게놈 수준에서 high-throughput의 데이터를 분석하는 연구 분야들을 포함하고 있다. 그러므로, 아래에서는 다양한 식물로부터 생성된 방대한 양의 생물학적 정보들을 생물정보학적 기법이 식물 생명공학 연구에 어떻게 이용할 수 있으며, 분석 기법과 사례를 소개하고자 한다.

식물에서 Proteome에 대한 연구

Proteomics는 대상 생물의 특정 세포나 조직에 존재하는 모든 단백질들을 high-throughput 기술을 이용하여 체계적으로 분석하는 연구분야이다 (Patterson and Aebersold 2003). 다양한 생물체에서 전체 게놈 염기서열 분석이 완료됨과 동시에 전체 프로테옴을 분석하고자 하는 연구는 기능 유전체학에서 매우 중요한 분야로 대두되었고, 다양한 단백질의 특성을 규명하고자 하는 실험 방법들의 개발되어 대량의 단백질 기능 분석이 가능하게 하게 되었다. Proteomics는 많은 양의 DNA 염기서열 분석 결과로부터 예측된 단백질의 아미노산 서열들을 이용하여 구축된 단백질 데이터베이스에서 시작되었으나, 점차 단백질의 구조 예측이나 발현 분석에 대한 데이터베이스도 소개되었으며, 최근에 post-genomics 주요 분야로써 연구 기술이 급속도로 발전되고 있다 (Edwards and Batley 2004). 식물에서의 proteomics 연구는 동물이나 효모 같은 다른 생물들에 비해 아직 연구가 많이 되어 있지 않다. 그러나 최근 몇몇 모델 식물을 기반으로 식물 단백체를 이용한 구조와 기능 분석을 위한 연구 노력들이 시도되고 있다. 식물 proteomics 연구는 주로 2차원 전기영동법 (2-dimensional electrophoresis, 2-DE)을 이용하여 원형질 막, 핵, 미토콘드리아, 또는 엽록체와 같은 식물 세포의 일부 구획으로부터 부분적으로 이루어지고 있다 (Park 2004). 최근에는 mass spectrometry (MS)와 matrix-associated laser desorption ionization (MALDI) time of flight (TOF) 분석법 (Karas and Hillenkamp 1988) 등을 이용한 식물 단백질 분석 논문들이 발표되고 있다. Peltier (2000) 연구팀은 2-DE와 mass spectrometry (MS)를 이용하여 완두 엽록체의 thylakoid 단백질을 분석한 결과 18개의 새로운 엽록체 단백질을 동정하였으며, 애기장대로부터 서로 다른 subunit들로 이루어진 거대 단백질의 각 subunit을 동정하였다 (Peltier et al. 2001). 엽록체와 더불어 애기장대의 핵, 미토콘드리아 프로테옴 분석도 2-DE와 MALDI-TOF MS를 통하여 새로운 단백질들이 동정되고 있다 (Kruft et al 2001, Bae et al 2003). 이외에도 식물의 다양한 세포 구획으로

부터 수 많은 단백질 정보들이 생성되고 있으며, 실험적 분석 외에도 ChloroP, PSORT, 또는 SignalP와 같은 생물정보학적 예측 프로그램들을 이용하여 식물 세포구획 내에서 작용 될 것으로 예상되는 단백질군의 분석도 가능하게 되었다. 앞으로 식물 생명공학 연구에서 식물 전체 단백질의 분석을 위하여 multidimensional promtein identification technology (Wolters et al. 2001), 또는 liquid chromatography (LC)-MS/MS와 같은 2차원 전기영동법의 문제점을 극복한 더욱 진보된 단백질 분석 기술을 이용하여 많은 단백질 정보들이 식물로부터 생성될 것이다. 이와 같이 방대한 단백체 데이터는 게놈 및 전사체 정보와 더불어 식물 생명공학 연구에 중요한 연구 도구로써 제공될 것이다. 또한 많은 정보들을 연구자들이 편리하게 사용할 수 있도록 생물정보학적 분석과 데이터베이스화 기술 또한 진보하고 있다.

Metabolomics

Metabolomics는 한 생명체에서 유래한 대사물질의 전체를 의미하는 metabolome을 연구하는 학문으로, genome, proteome, transcriptome과 더불어 기능 유전체학에서 중요한 분야 작용 하고 있다 (Nobeli and Thornton 2006). Metabolome의 연구 목적은 몇 가지로 설명될 수 있다. 첫번째는 세포로부터 생성되는 대사물질은 세포 상태와 직접적으로 연관 되어 있어 생명체의 표현형에 직접적인 영향을 줄 수 있으며, 때로는 대사물질이 유전자의 발현이나 단백질의 영향과는 무관하게 물질 대사에 영향을 줄 수도 있기 때문에 metabolite 자체를 연구하는 것은 매우 중요하다 (Goodacre et al. 2004). 두 번째로, metabolome에 대한 연구는 중요한 대사산물의 생체 내 합성 경로를 추적 및 연구함으로써 조절 가능한 새로운 대사경로를 찾을 수 있고, 네트워크 조절을 통해 유용한 대사산물이나 천연 물질을 다량으로 얻을 수 있다. 세 번째로, 전사체 또는 단백체 발현 분석과 함께 특정 상황에서 생명체의 상태 변화를 모니터링하고, 이를 다양하게 이용할 수 있다.

Metabolome을 연구하는 방법은 목적에 따라 metabolite target analysis, metabolite profiling, 그리고 metabolic fingerprinting 등이 있다 (Shulaev 2006). Nuclear magnetic resonance는 spectra는 넓은 범위의 유기 화합물에 대한 동정과 정량을 동시에 할 수 있어서 분석의 어려움에도 불구하고 metabolite fingerprinting에 광범위하게 사용되고 있다 (Viant et al. 2003). MS는 direct-injection MS나 다른 크로마토토그래피 또는 전기영동법과 병행하여 생물학적인 시료를 분석할 수 있고, 매우 높은 sensitivity와 넓은 범위의 metabolite들에 적용될 수 있다는 장점때문에 많은 metabolomics 연구, 특히 metabolic fingerprinting과 metabolite profiling에 사용되고 있다 (Glinski et al. 2006). MS는 또한 gas chromatography (GC-MS)와 liquid chromatography (LC-MS)와 함께 분석 방법으로 사용되고 있다 (Nobeli and Thornton 2006). capillary electrophoresis-MS (CE-MS)는 짧은 시간에 1-20 nl의 매우 적은 시료로 높은 해상력의 분석 결과를 보여 줄 수 있어서 metabolome 연구에 중요한 분석 기술로 이용되고 있다 (Terabe et al. 2001). 이러한 데이터들은 역시 대량으로 생산되고 있으며, 수학적, 또는 통계적인 방법들과 함께 다음과 같은 생물정보학적 도구들을 이용하여 분석된다. 예를 들어, 주로 GC-MS 결과로 얻어지는 가공되지 않은 데이터들은 AMDIS (automated mass spectral deconvolution and identification system, http://chemdata.nist.gov/mass-spc/amdis/) 분석 프로그램을 이용하여 processing 된다 (Holket et al. 1999). ESI-LC-MS 데이터 분석에는 component detection algorithm (CODA) 또는 windowed mass selection method (WMSM)가 사용된다 (Windig et al. 1996; Fleming et al. 1999). 이렇게 만들어진 분석 데이터들은 world-wide-web상에서 연구자들이 이용할 수 있도록 데이터베이스로 구축 되어 있다. 이런 데이터베이스들은 주로 metabolic pathway 데이터베이스나 pathway viewer들로 이루어져 있는데, 대표적인 데이터베이스로는 KEGG (http://www.genome.ad.jp/kegg/), KaPPA-View (http://kpv.kazusa.or.jp/kappa-view/), DOME (http://medicago.vbi.vt.edu) 등이 있다.

Metabolomics는 다른 '-omics'들 보다 비교적 최근에 발전하기 시작한 학문으로, 특히 식물 분야에서 metabolomics는 앞으로 더욱 발굴해야 할 데이터들이 많다. 이를 위해 다양한 분석 기술들이 개발되어야 하고, 이와 동등하게 생성된 데이터를 분석하기 위한 새로운 생물정보학적 분석 기법들이 개발되어야 할 것으로 생각된다.

식물병 연구를 위한 생물정보학

식물 연구는 발생학, 분화학, 생리학, 병리학 등, 다양한 분야에서 수행되고 있다. 이런 다양한 식물 연구의 궁극적인 목표는 안정적이면서 풍요로운 식량 자원의 확보를 통하여 인간 삶을 한 층 더 윤택하게 만들고자 함이다. 따라서 식물 병 저항성에 관한 연구는 식물의 다른 연구 분야 못지 않게 매우 중요하다. 식물은 다양한 병원체의 침입에 대항하여 국부적인 또는 전신적인 저항성 반응을 일으킨다. 식물의 저항성 반응에서 대표적인 현상 중 하나는 과민성 반응 (hypersensitive response, HR)으로, 식물의 저항성 유전자 산물과 이에 상응하는 병원체의 비 병원성 유전자 산물의 상호 작용으로 인하여 나타나는 식물의 빠른 세포 사멸반응이다 (Dangl et al 1996). 식물은 저항성 유전자 산물을 통하여 병원체를 인식한 후, 복잡하고 다양한 신호 전달 과정을 통해 pathogenesis-

related (PR) 유전자와 같은 생체 방어와 관련된 유전자들의 발현을 증가시키고, phytoalexin과 같은 항미생물 활성을 갖는 물질들을 생성한다. 또한 활성 산소군, 살리실산과 같은 세포 내 신호 전달 물질들이 작용하여 식물 전체에 생체 방어 체계를 활성화시켜 2차적인 병원체의 침입에 전신 저항성을 가지는 전신 획득 저항성을 유도한다 (Yang et al. 1997).

이러한 식물 병 저항성 기작에 관한 연구에서도 생물정보학적 도구를 이용하여 생체 방어체계에 관여하는 후보 유전자를 얻고자 하는 노력들이 시도되고 있다. 예를들면, 병 저항성 반응에서 병원체를 인식하는 것으로 알려진 resistance gene (R-gene) 에 대해 nucleotide-binding site와 leucine rich repeat (NBS-LRR)와 같은 domain 정보를 이용하여 게놈 분석이 끝난 애기장대와 벼의 유전체에서 분석한 결과 각각 145개와 535개의 R-gene이 screening 되었다 (Meyers et al. 2003, Zhuo et al. 2004). 이러한 게놈 수준에서의 분석은 다른 저항성 관련 유전자들의 분석에 대한 동기를 부여하게 되고, 나아가서 유전자 그룹간의 상호 관계를 생물정보학적 분석 기법들을 이용하여 예측할 수 있다. 생물 정보학적 기법을 이용한 비교 분석은 식물에서 잘 알려져 있지 않은 과민성 반응을 동물에서 유사한 반응기작인 apoptosis와 비교 분석함으로써 아직 기능이 알려져 있지 않거나 명확하지 않은 식물 유전자에 대한 기능을 예측가능 하도록 하였다. 비록 식물과 동물의 면역체계가 다르기는 하지만, 일부 실험적 증거들은 식물의 세포 사멸 과정이 동물과 유사할 것이라는 추측을 가능하게 한다 (Kawai-Yamada et al. 2001, Ausubel 2005). 도메인 기반으로 동물의 apoptosis와 관련된 단백질들을 분석하여 만든 데이터 베이스 (http://www.apoptosis-db.org)에서 보면 동물의 세포 사멸에 관여하는 것으로 알려진 단백질들이 가지고 있는 특징적인 도메인이나 유사한 구조를 가지고 있는 식물의 단백질들이 제시되어 있다. 이런 몇 가지 근거들은 식물의 세포 사멸 및 병 저항성과 관련된 새로운 정보들을 찾을 수 있는 가능성을 제시해주고 있다.

식물의 병 저항성 반응을 전체적으로 이해하기 위하여 EST data 분석과 DNA chip 분석은 매우 유용하게 사용되고 있다. 이러한 데이터는 특정 병원체의 침입 상황에서 저항성 반응이 일어나는 식물 조직으로부터 EST를 library를 제작하거나, DNA chip을 제작함으로써 병원체 침입 상황에 특이적으로 반응하는 유전자 군을 발굴할 수 있다. EST의 경우 전체 EST full에서 consensus를 만든 후 통계적 분석을 통해 상황 특이적으로 발현 되는 유전자들을 예측 할 수 있다. 이렇게 예측 된 일부 유전자들은 실험적으로 증명된 바 있다 (unpublished data). DNA chip의 경우 preprocessing과 normalization 후 병원체 침입 상황과 비 침입 상황에서의 유전자 발현 비를 구하고, 침입 상황에서 특이적으로 많이 또는 적게 발현 되는 유전자 군을 찾을 수 있다 (Maleck et al. 2000, Lee et al. 2004, Marathe et al. 2004). 또한 이러한 유전자 군의 발현 양상이 유사한 것들을 clustering에 의해 분류할 수 있으며 (Maleck et al. 2000, Lee et al. 2004, Marathe et al. 2004), 이렇게 분류된 유전자군들은 promoter 분석 등에 의해 유전자 level에서 이러한 차이를 보이는 증거를 찾을 수 있을 것으로 기대된다.

생물정보학적인 기술의 개발과 분석은 식물 병 저항성뿐만 아니라 다양한 식물 현상에 대한 데이터를 수집하고, 알려져 있는 데이터베이스의 정보들을 이용하여 검색, 비교 가능하게 함으로써 식물 생명공학 연구에 필요한 유용한 정보들을 축적해 나가는데 도움을 줄 수 있을 것이다.

식물의 분자 육종을 위한 생물정보학

고등 식물, 특히 작물의 경우에 있어 육종 이란, 형질전환과 같은 유전공학기술에 의한 목적 형질 (target trait) 의 획득이 아닌, 교배와 유전법칙을 이용해 형질을 획득하는 과정이다. 형질을 획득하여 신품종을 육종하였다는 것은 표현형을 주관하는 유전서열을 확보하고, 대립 유전자 (allele) 다형성 (polymorphism) 에 의해 목적한 표현형 획득이 직접적으로 증명되었음을 의미하며 최종적으로 새로운 품종의 종자를 생산한다.

육종 연구는 목적 형질을 포함하는 개체와 그렇지 않은 개체와의 대립 유전자 (allele) 의 비교에 의해 이루어지며, 대립 유전자좌의 변이에 의한 표현형의 변화 및 상관관계를 분석함으로써 증명된다. 이를 위해 자연 집단의 자가 교배 (selfing) 보다는 인위적인 교배 (cross) 를 유도하여 지도 작성 집단 (mapping population) 을 구성한다. Fig. 1에서와 같이 구성된 집단은 molecular marker를 이용하여 목적 표현형과 형질에 관련된 유전거리 (genetic distance)가 가까운 marker를 동정하여 후보 유전자 (candidate gene)를 선발 하게 된다 (Silivo and Roberto 2005). 선발된 서열은 형질의 표현형 효과에 대한 직접적 기능 증명이 될 수 없으므로 knock-out transformation, northern hybridization, cDNA microarray, RNAi 와 같은 실험을 통해 이를 증명하게 된다 (Fig.1. validation of candidate gene).

Molecular marker를 통해 유전자좌 변이를 비교 분석하여 목적 형질과 유전거리 (genetic distance)가 가까운 molecular marker를 동정하는 일련의 과정을 MAS (Marker Assisted Selection)라고 하며 positional cloning, map-base cloning 이라고도 한다. 전통 육종의 경우 표현형에 의한 개체선발을 통해 최종의 형질을 획득하였으므로 실험자의 주관과 표현형 판단의 모호함과 같은 문제가 있었으나 MAS 의 경우 유전형과 표현형의 상관관계를 유전적 확률로 접근하므로 보다 효과적인 방법이라 할 수 있다.

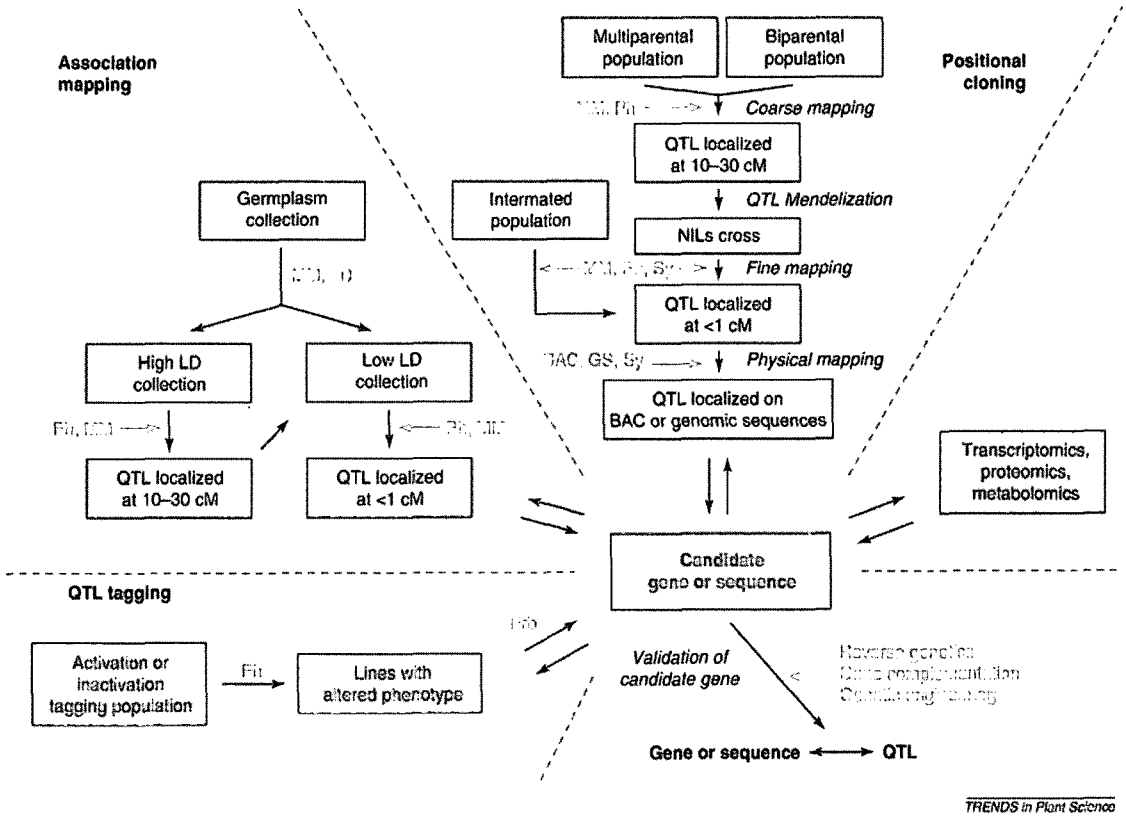

Figure 1. Flow-chart of molecular breeding. Ph; phenotyping, MM; molecular marker, GS; genomic sequence, Sy; Synteny, LD; linkage disequilibrium (Silivo et al. 2005).

MAS에 이용되는 마커의 종류는 RFLP (Restriction Fragment Length Polymorphism), AFLP (Amplify Fragment Length Polymorphism), SSR (Simple Sequence Repeat), RAPD (Restriction Amplify Polymorphic DNA) 등이 molecular marker로 개발되어 있으며, 이러한 마커를 통해 옥수수 (Silivo et al. 2002), 고추 (Charless et al. 2005), 벼 (Wan et al. 2006) 등의 많은 작물에서의 형질 동정이 지속적으로 이루어져 왔다. 일반적인 molecular marker의 경우 서열상의 패턴변화를 이용한 random DNA marker 였으나 functional marker의 경우 기능이 알려진 유전자 서열을 이용하여 개발되었기 때문에 MAS 최종단계인 후보 유전자의 기능동정에 유리하다 (Jeppe et al. 2003). 따라서 다양한 생물학적 정보를 접목시킨 마커의 개발이 MAS 과정에 활용될 것으로 예상되며, in silico 분석에 의한 대량의 molecular marker 후보를 데이터 베이스화한 PlantMarkers (Stephen et al. 2005)가 구현되어 있다. 또한, microarray의 발현정보를 이용한 후보 유전자의 선발 기법 (Wayne et al. 2002)의 개발, 시스템 생물학 (systemic biology)에 의한 형질 분석 기법이 지속적으로 연구되고 있으며 그 결과, 육종 연구는 과거에비해 빠른 속도로 전개될 것으로 예상된다.

반면 경제적 효용이 큰 수량의 증대와 같은 양적 형질 (Quantitative Trait Loci QTL) 에 대한 동정은 다양한 환경 효과와 형질에 대한 정의, 주요 형질 (major QTL)을 발견하기 위한 종합적인 분석이 요구된다 (Cuartero et al. 2006). 따라서, 각각의 유전자를 marker로 활용한 functional marker 또는 array 정보와 같은 단편적 정보를 이용한 형질동정의 방법으로는 양적 형질의 분석에서는 큰 발전을 이루기 어려울 것으로 판단된다. 그 이유는 첫째, 쌍자엽 (dicotyledon) 식물의 모델식물인 애기장대 (*Arabidopsis thaliana*)의 경우 전체 유전체 중 기능이 밝혀 진 유전자의 수는 MAS에 필요한 molecular marker 요구량보다 극히 적고, 증가량이 크지 않다. 둘째, 다양한 '-omics' 정보를 MAS 에 이용하기 위한 연구는 아직 미흡 하며, 모든 정보를 연구자 자신이 분석 할 경우 많은 시간과 노력이 요구된다.

Fig. 2. 는 단백체학 (proteomics), 전사체학 (trascriptomics), 유전체학 (genomics) 등의 다양한 정보를 MAS에 활용하기 위한 데이터베이스 또는 기초 시스템의 개념도이다 (Rajee et al. 2005). 이러한 시스템 생물학적 분석을 통해 MAS 기반의 다양한 -omics 정보의 통합이 가능해 진다면 선발된 trait의 기능 증명에 중요한 역할을 할 것으로 생각 된다. 따라서, 생물정보학 (bioinformatics, systemics biology) 분야는 육종의 기능 분석을 위한 융합 system에 필연적으로 요구될 것이며, 다양한 기능 분석을 위한 데이터를 조직적, 복합적으로 모델링 함으로써 육종 실험 설계와 MAS에 의해 선발된 trait 동정을

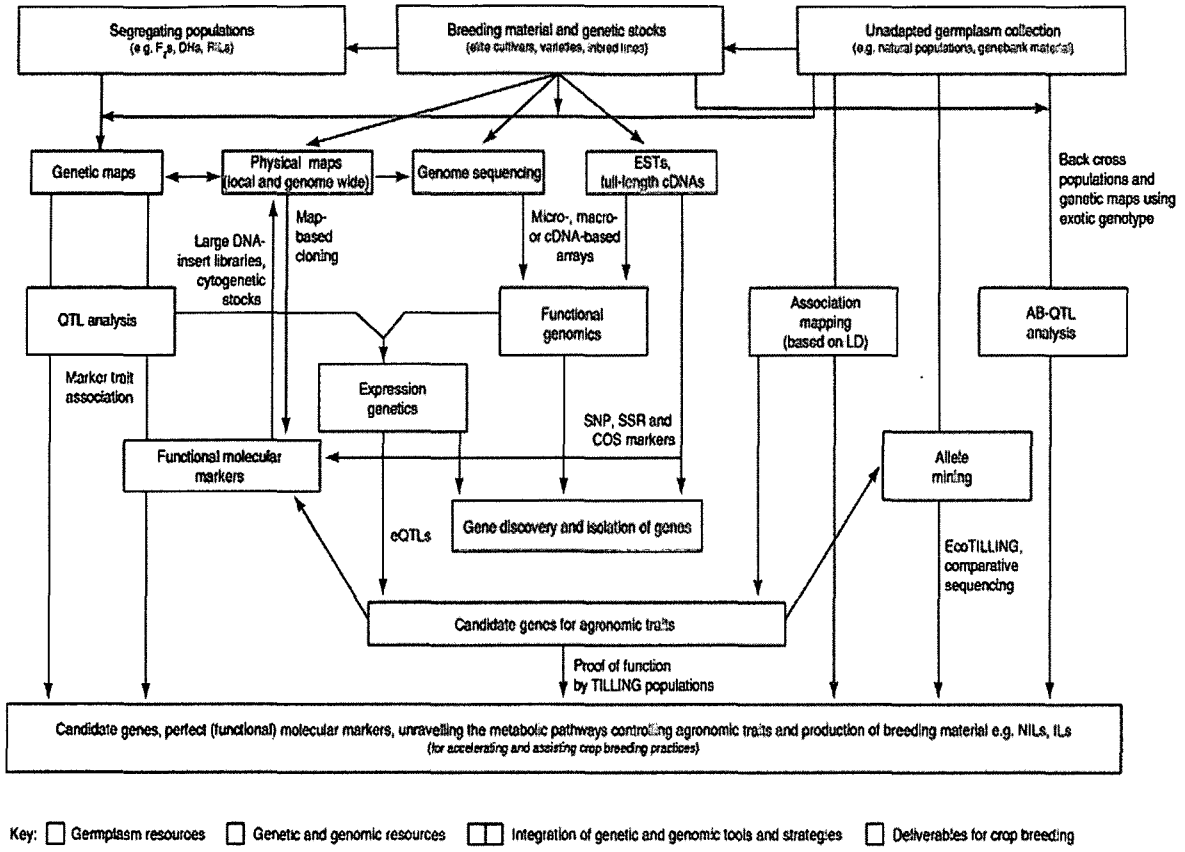

Figure 2. Integrated systemic approach for effective MAS (Marker Assisted Selection) (Rajee et al. 2005).

보다 효율적으로 연구 할 수 있도록 할 것으로 예상된다.

식물 생명공학 연구를 위한 생물정보

최근 대량의 생물 데이터를 만들어내는 생명 공학 기술이 발전하면서, 생명체를 복합적으로 이해하고자 하는 노력이 많이 이루어지고 있다. 식물 연구 분야에서도 지난 2000년 Arabidopsis Genome Initiative에서 애기장대 (The_Arabidopsis _Genome_Initiative 2000)와 Rice (Goff et al. 2002, Yu et al. 2002)의 전체 게놈 염기서열 분석을 성공적을 마치면서 토마토, 옥수수, *Medicago truncatula*, 사탕수수 (Bedell et al. 2005)도 whole genome sequencing 및 annotation이 진행 중이다. 또한, transcriptome을 보기 위한 full-length cDNA와 EST, 특정 조직이나 환경에서의 유전체 발현양상을 보기 위한 microarray나 SAGE, 단백체 연구를 위한 2D-gel 전기영동 등의 시스템 수준에서 식물체를 이해하기 위한 연구를 도울 수 있는 다양한 방법이 개발되어 있다. 이러한 연구 경향은 생물정보학이라는 새로운 분야를 탄생시켰으며, 생물 정보에서는 대량의 데이터를 자동으로 처리, 가공하기 위한 통계나 컴퓨터 사이언스, 엔지니어링의 개념을 도입하여 데이터 가공 및 분석, 저장부터 데이터를 이용한 이론적 방법론 개발, 수학적 모델링 설립, 컴퓨터 시뮬레이션을 이용한 생명체 이해를 위한 연구가 활발히 진행되고 있다 (Rhee et al. 2006). 이 장에서는 식물체의 이해를 위한 genomics, transcriptomics, proteomics 연구가 어떻게 진행되고 활용방안에 대해 설명하고, 이와 관련된 생물정보 도구, 데이터베이스를 함께 소개하고자 한다.

Genomics

유전체 연구는 whole genome sequence나 EST를 기반으로 한 유전자 구조연구 대한 연구, 단백질 family 구조에 의한 기능 연구가 가장 활발이 진행되고 있으며, 이러한 자원은 뒤에서 설명할 transcriptome, proteome 등의 연구를 위한 원천 자원이기 때문에 매우 중요하다.

Genome Sequencing

대량의 DNA 서열을 분석 처리하기 위해서는 DNA를 random하게 잘라 clone을 만든 후 대량으로 sequencing을 한

Table 1. *Ab initio* gene prediction programs

Name	Method	Web site	Reference
AUGUSTUS	Generalized Hidden Markov Model (GHMM)	http://augustus.gobics.de.	(Stanke et al. 2006)
GENSCAN	semi Markov Model	http://genes.mit.edu/GENSCAN.html	(Burge and Karlin 1997)
GRAIL	Neural network	http://compbio.ornl.gov/Grail-1.3/	(Xu et al. 1994)
GenLang	Definite clause grammer	http://www.cbil.upenn.edu/genlang/genlang_home.html	(Dong and Searls 1994)
GenView	Linear combination		(Ronneberg et al. 2001)
GenFinder	Dynamic programming	http://www.bioscience.org/urllists/genefind.htm	
GeneID	Perceptron, rules	http://www1.imim.es/geneid.html	(Guigo et al. 1992)
GeneMark	5th-Markov	http://opal.biology.gatech.edu/GeneMark/genemark24.cgi	(Besemer and Borodovsky 2005)
GeneParser	Neural networks	http://beagle.colorado.edu/eesnyder/GeneParser.html	(Snyder and Stormo 1995)
Genie	GHMM	http://www.fruitfly.org/seq_tools/genie.html	(Kulp et al. 1996)
Glimmer	Interpolated Markov models(IMMMs)	salzberg@cs.jhu.edu	(Xu et al. 1994)
MORGAN	Decision tree	http://www.cs.jhu.edu/labs/compbio/morgan.html	(Salzberg et al. 1998)
MZEF	Quadratic discriminant analysis	http://argon.cshl.org/genefinder/	(Zhang 1997)
NetPlantGene	Combined Neural Networks	http://www.cbs.dtu.dk/services/NetPGene/	(Hebsgaard et al. 1996)
VEIL	Hidern markov models	http://www.cs.jhu.edu/labs/compbio/veil.html	(Henderson et al. 1997)

다. Sequence 결과는 chromatogram으로 'abi' 파일로 컴퓨터에 저장되는데, 이 정보를 서열정보로 바꾸기 위해서는 Phred 프로그램 (Ewing and Green 1998) 에 의해 DNA 서열정보를 얻을 수 있다. 이러한 서열 정보는 overlapping segments에 의해 assemble되는데, 관련된 프로그램은 genomic sequence assemble에는 phrap이 주로 이용이 되고 있으며, EST sequence의 경우 CAP3 (Huang and Madan 1999), STACK -pack (Burke et al. 1999) 등이 연구 목적에 따라 다양하게 사용되고 있다. 이외에도 DNA assembly를 위해 Arachne (Jaffe et al. 2003), GAP4 (http://staden.sourceforge.net/overview.html), AMOS (http://www.tigr.org/software/AMOS/) 등이 개발, 이용되고 있다.

Gene Finding and Annotation

다양한 종의 whole-genome project에서 exon, intron, UTR 등을 포함한 유전자의 구조를 밝히고, boundary를 정하는 것은 매우 중요한 작업이다. 전형적인 방법은 기존의 실험이나 annotation 과정에서 알려진 다른 종의 유전자 서열이나 cDNA, EST 등을 이용해 서열 유사성을 이용한 방법이 사용되고 있으나, 알려진 유전자의 종류가 제한적이고, low similarity에 의한 false positive의 우려가 높다. 그렇기 때문에 whole genome sequence로부터 common protein coding transcripts의 recognize feature를 찾는 software를 사용하는 'ab initio gene discovery' 방법이 많이 제시 되었다 (Mathe et al. 2002). 이러한 소프트웨어는 codon-bias, transcriptional and translational initiation motif, 3' polyadenylation sites, exon-intron boundary 영역에 있는 splicing 보존적 서열을 학습하거나 패턴을 인식시킴으로써 유전자 구조를 찾기 위한 다양한 알고리즘이 개발되었다 (Rogic et al. 2001, Mathe et al. 2002)(Table 1). 하지만, 각 종에 따라 유전자의 구조적 특성에 차이가 나기 때문에 일괄적인 프로그램을 이용할 경우 정확성이 떨어지는 단점이 있다. 'ab initio gene discovery' 방법 이외에도 full-length cDNA, EST, 잠재적 단백질 (potential protein)의 homologous와 같은 transcript evidence를 이용한 'structural annotation' 방법이 있다 (Rhee et al. 2006). 특히, 이러한 방법에 의해 모델링 된 유전자는 앞에서 언급한 프로그램을 training 시키기 위한 sample set으로도 이용되고 있으며, 새로운 트레이닝에 의한 gene prediction 방법은 ab initio algorithms을 이용한 gene finding의 정확성을 높여줄 수 있다.

Protein Classification

단백질은 DNA 서열에 비해 진화적으로 더 잘 보존되어 있는 아미노산 서열로 이루어져 있으며, 특히 같은 family 또는 superfamily 내에서는 도메인이나 3차구조를 공유하고 있다 (Liu and Rost 2003). 이러한 특징은 초기 단백질 분석에서 homology를 기반으로 한 단백질 분류를 가능하게 하였다. Homology를 기반으로 한 단백질 분류로 널리 알려진 데이터베이스는 Orthologous Groups of proteins (COG)으로 gapped-blast를 이용해 실험적으로 알려진 단백질의 기능과 orthologous한 protein의 relationship을 밝힘으로써 분류하고 있다 (Tatusov et al. 2000). 현재 COG에서는 단백질의 17개의 기능적 분류 체계를 가지고 있는데 그 중 가장 큰 카테고리는 'uncharacterized'이다 (Tatusov et al. 2000). 이러한 문제가 발생되는 이유는 homology를 기반으로 한 단백질 분류는 house-keeping function을 가진 단백질처럼 진화적으로 매우 보존적인 서열을 가진 단백질에서 유리한 algorithm이기 때문이다 (Tatusov et al. 2000). 진화적으로 거리가 먼 단백질의 경우 전체 단백질 sequence의 similarity가 낮기 때문에 homology 기반의 분류는 어렵기 때문에 단백질의 기능이나 active sites와 직접적으로 연관된 'motif'를 이용한 profile을 만든 후 profile을 기반으로 단백질 분류를 하는 것이 정확성이 더 높은 것으로 나타났다. 따라서 profile을 잘 만들기 위한 알고리즘 및 tool이 많이 개발, 이용되고 있는 추세이다 (Table 2). 하지만 많은 단백질의 경우 multiple domain으로 이루어져 있기 때문에 하나의 motif로 단백질을 characterization 할 수 없는 부분적 난제가 있다. 위와 같은 여러 가지 문제점 때문에 단백질을 가장 잘 분류 할 수 있는 알고리즘을 정의 하기는 쉬운 일이 아니다. 따라서 단백질 분류와 관련된 데이터베이스도 개발자의 관점에 따라 short-sequence motifs, structural domain-like regions, interaction, active site 등 특성을 살린 데이터베이스가 다양하게 존재한다 (Table 3). 또한 최근에는 Protein Kinase Resource (PKR) (Smith et al. 1997), protease 분류를 위한 MEROPS (Rawlings et al. 2006), transcription factor 관련 TRANSFAC (Matys et al. 2003), esterases와 lipases를 모

Table 2. method and software for protein classification

1. pairwise comparison • FASTA (http://fasta.biocch.verginia.edu) • BLAST (http://www.ncbi.nlm.hih.gov/blast)
2. sequence • PSI-BLAST (http://www.ncbi.nlm.nih.gov/BLAST) • HMMER (http://hmmer.wustl.edu) • SAM(http://www.cse.ucsc.edu) META-MEME(http://metameme.sdsc.edu)
3. profile-profile alignment

Table 3. Availablity of databases and methods

Classification scheme	URL
Short sequence motifs	
PROSITE	http://www.expasy.ch/prosite
Block	http://block.fhcrc.org/block
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS
Structural domain like regions	
Pfam-A	http://pfam.wustl.edu
TIGRFAM	http://www.tigr.org/TIGRFAMs
SBASE	http://smart.embl-heidelberg.de
DOMO	http://infobiogen.fr/services/domo
ProDom	http://prodes.toulouse.inra.fr/prodom/doc/prodom.htm
GeneRAGE	http://ebi.ac.uk/research/cgg/services/rage
TribeMCL	http://www.ebi.ac.uk/research/cgg/tribe
CHOP	http://cubic.bioc.columbia.edu/db/chop
Integration	
InterPro	http://www.ebi.ac.uk/interpro
MetaFam	http://metafam.ahc.umn.edu
3D-structure	
PDB	http://www.rcsb.org/pdb
SCOP	http://scop.berkeley.edu/
CATH	http://www.cathdb.info
FSSP	http://www.fssp.org
DSSP	http://swift.cmbi.ru.nl/gv/dssp/
Active sites	
EzCatDB	http://mbs.cbrc.jp/EzCatDB

은 MELDB (Kang et al. 2006) 같이 하나의 large-family의 단백질을 기준으로 해당 단백질의 특성을 잘 대표할 수 있는 알고리즘을 복합적으로 이용하여 분류하는 데이터베이스도 늘어나는 추세이다.

Transcriptomics

transcriotomics는 세포 또는 조직의 특이적 상황에서 유전자의 발현 양상에 대해서 연구하는 분야로, 주로 micrarray에 의해 얻어진 데이터를 clustering (Eisen et al. 1998, Wang et al. 2002) 과정을 통해 유사한 pattern으로 발현되는 유전자 군으로 만든 후, 이러한 유전자 group의 cis-element sequence analysis (Zhang et al. 2005, Haberer et al. 2006), gene ontology (Tatusov et al. 2003), pathway information과 결합하여 의미 있는 분석 결과를 얻고 있다 (Rhee et al. 2006). 예를 들면, Micrarray를 이용한 스트레스 상황에서의 유전자 발현 양상 실험을 한 데이터를 발현 양이 유사한 유전자 groups으로 clustering한 후 promoter영역을 조사한 결과 WRKY, bZip, ERF와 같은 stress 상황에서 유전자발현을 조절하는 것으로 알려진 transcription factor의 cis-element가 관측되었다 (Chen et al. 2002). 이러한 결과는 co-expression된 유전자 set이 co-regulation되었다는 결론을 유추할 수 있게 한다. 뿐만 아니라, clustering 결과를 KEGG pathway map (Ogata et al. 1999)과 연동하면 signaling cascade나 metabolism pathway상의 가까이 위치한 유전자 set의 expression level의 차이를 해석함으로써 development stage (Breyne et al. 2002)나 metabolism에 대한 연구도 가능하다. Microarray 분석 결과는 작물의 pathogen infection 상황이나 stress 상황에서 중요한 역할을 할 것으로 추정되는 novel gene을 찾는데 도움을 줄 수 있을 것으로 생각된다. 최근에는 alternative splicing을 고려한 chip도 생산되고 있어 microarray를 이용한 transcriptome 연구의 활용도가 더욱 높아질 전망이다 (Jaffe et al. 2003). Microarray는 image processing, normalization, clustering (Eisen et al. 1998, Wang et al. 2002) 의 과정을 통해 데이터가 얻어지는데, microarray 실험이 많이 공급되면서 이러한 데이터 분석을 위한 CaARRAY (http://caarray.nci.nih.gov/), BASE, Bioconductor (http://www.bioconductor.org)와 같은 공용프로그램과 함께 gene Traffic, GeneSpring (http://www.agilent.com/chem/genespring), Affymetrix's GeneChip Operating Software (GCOS)와 같은 상업적 프로그램도 많이 개발되어 있다. 식물체 연구의 microarray 실험은 주로 식물의 development stage에 대한 연구 (Breyne et al. 2002), stress 또는 pathogen infected 상황에서의 유전자 발현 양상에 대한 연구가 주를 이루고 있다. 이러한 연구 결과는 table 4와 같은 데이트베이스에서 쉽게 다운로드 받아 연구에 이용 할 수 있다. 이 이외에도 transcriptome 연구는 EST를 이용한 alternative splicing 분류 (Gupta et al. 2004), digital gene expression profile (Audic and Claverie 1997) 조직 특이적 또는 상황 특이적 유전자 발굴 (Jeong et al. 2006), SAGE를 이용한 정량적 분석 (Nielsen et al. 2006)도 많이 수행되고 있다.

Proteomics

프로테오믹스 연구는 주어진 환경에서 발현 된 단백질들의 large scale로 동정하고, 기능을 연구하는 분야로, mRNA의 발현으로 예측할 수 없는 단백질 발현, methylatioin, pho-

Table 4. Plant microarray database

TAIR	http://www.arabidopsis.org/tools/bulk/microarray/analysis/index.jsp
EBI Microarray	http://www.ebi.ac.uk/microarray/index.htm
CATMA	http://www.catma.org/
Genevestigator	https://www.genevestigator.ethz.ch
AtGenExpress	http://web.uni-frankfurt.de/fb15/botanik/mcb/AFGN/atgenex.htm
TIGR AT Array	http://atarrays.tigr.org/
Expression Profiling of Plant Disease Resistance Pathways -	http://www.fastlane.nsf.gov/servlet/showaward?award=0114783
Pathoarrays	http://aztec.stanford.edu/cold/cgi-bin/data.cgi
ColdArrayDB	http://www.osmid.org/
Osmostic Stress Microarray	http://stress-genomics.org/stress.fls/expression/expression.html
Stress Genomics Consortium, USA	http://www-biology.ucsd.edu/labs/schroeder/guardcellchips.html
ABA-dependent Guard Cell and Mesophyll Cell Expression Arrays	http://arexdb.org "target="_blank
AREX (Philip Benfey, USA)	http://www.bpp.msu.edu/Seed/SeedArray.htm
Developing Seeds Array Data	http://www.arabidopsis.org/info/expression/ada_gcn.jsp
Gene Expression in ada 2b-1 and gcn 5-1 mutants	http://bioinformatics.med.yale.edu/pathmapa.htm
PathMAPA (Yale Univ., USA)	

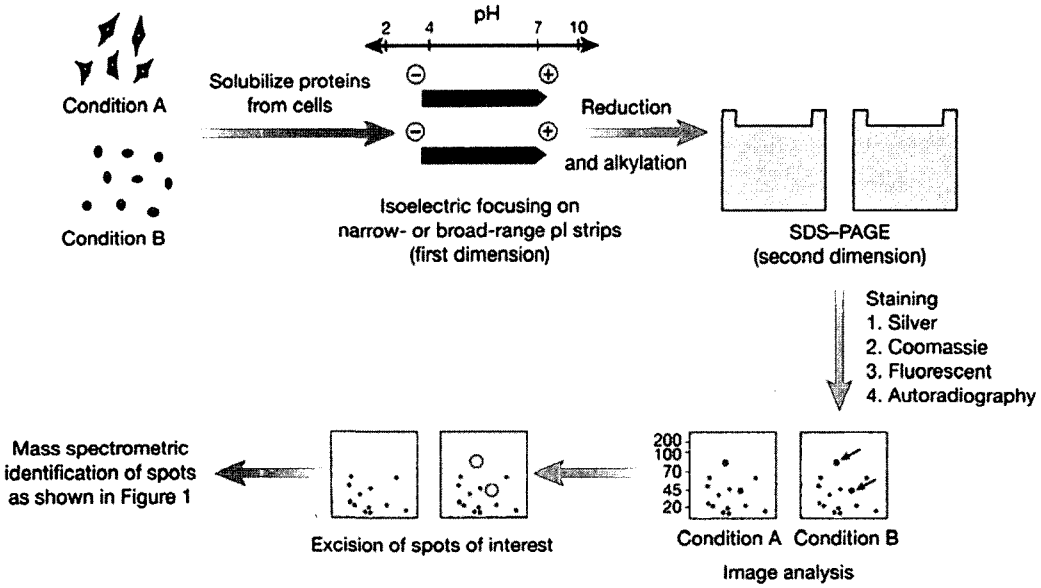

Figure 3. Proteomics study (image from brin.uams.edu/images/Proteomics%202.png)

Table 5. Internet sites used for proteomic analysis

Internet site	Useful utilities	Site address
Protein prospector	Seqrch databases with PMF and MS/MS data. Provie theoretical data for a known peptide sequence	http://prospector:ucsf.edu
PROWL	Search databases with PMF and MS/MS data. Provide theretical data and useful protocols	http://prowl.rockefeller.edu/PROWLprowl.html
ExPASy	PMF database searches, theoretical tools and links to manty other sites	http://us.expasy.org
PeptideSearch	Search PMF data or a sequence tag	http://www.mann.embl-heidelberg.de/GroupPages
Mascot	Search databases with PMF and Ms/MS data and sequence tags	http://www.matrixscience.com
PepMAPPER	Search databases with PMF data	http://wolf.bms.umist.ac.uk/mapper
MOWSE	Search databases with PMF data	http://srs.hgmp.mrc.ac.uk/rost/preditprotein/
PredictProtein server	Gives theoretical fragments and data on a protein amino acid sequence	http://www.ebi.ac.uk/rost/predictprotein/
BLASAT	Search sequence tag	http://www.ncbi.nlm.nih.gov/BLAST/

sphorylation과 같은 단백질 변형, alternative splicing에 의한 isoforms에 의한 분석 등이 이루어지며 세포내의 동적인 생명현상의 이해를 돕는 핵심 연구분야 이다 (Lisacek et al. 2006, Palagi et al. 2006). 주로 서로 다른 condition에서의 세포로부터 단백질을 분리하여 2차원 전기영동에 의한 spot의 양적 또는 질적 차이를 비교하고, interest spot으로부터 단백질을 분리하여 질량분석법 (Mass spectrometry analysis)에 의해 단백질을 동정하는 방법을 많이 이용하고 있다 (Newton et al. 2004, Rhee et al. 2006) (Fig. 3). Proteome 분석을 위한 유용한 site는 table 5와 같고(Newton et al. 2004), 식물의 2D-PAGE 데이터베이스는 table 6과 같다. 이외에도 yeast two hybrid를 이용한 interactome 연구, protein array (Nagayama 1997, Walter et al. 2000)에 의한 프로테오믹스 연구도 많이 이용되고 있다.

Table 6. 2D-PAGE databases

Database	Web site
ANU-2DPAGE	http://semele.anu.edu.au/2d/2d.html
Rice Proteome Database	http://gene64.dna.affrc.go.jp/RPD/database_en.html
Arabidopsis Seed Proteome	http://seed.proteome.free.fr/
GABI primary database	http://gabi.rzpd.de/projects/Arabidopsis_Proteomics/
NASC Proteomics database for Arabidopsis data	http://proteomics.arabidopsis.info/
Plant Proteomics database (PROTICdb)	http://cms.moulon.inra.fr/proticdb/Protic/home/
2-D PAGE of Medicago truncatula and other plants	http://www.pierroton.inra.fr/genetics/2D/
Plant plasma Membrane Database (PPMdb)	http://www.noble.org/2dpage/
Mt Proteomics	http://www.mtproteomics.com/

결론과 전망

식물로부터 생성되는 방대한 생물학적 데이터들은 식물 생명공학 연구의 중요한 재료로 이용되기 시작했다. 애기장대와 벼에서 전체 게놈 염기서열이 분석되고, 뒤이어 다양한 다른 식물로부터 이런 분석이 진행되고 있다. 초기 생물정보학은 많은 양의 염기서열 정보를 분석에서 출발하였으나 앞서 언급한 것과 같이 다양한 '-omics' 기술에 의하여 전사체, 단백체, 또는 metabolite의 정보들이 생성되고 있다. 이러한 정보들은 여러 가지 조건하에서 식물의 전체적인 유전자 발현이나 대사 과정의 차이를 비교하는데 이용 될 수 있고, 원하는 조건에서 특이적으로 발현되는 유전자, 단백질, 또는 대사 산물들을 얻을 수 있게 되었다. 이런 결과물들은 각각의 기능과 역할을 규명할 수 있지만, 전체적인 생명현상을 이해한다는 측면에서 여러가지 -omics 기술에서 나온 데이터들이 서로 유기적 관계를 이루면서 연구되어야만 한다. 그러기 위해서는 방대한 데이터를 모아서 정렬을 하는 일과 같은 단순한 생물정학적 분석 기술에서 여러 가지 -omics 기술로부터 나온 정보들을 유기적으로 상호 연결시킬 수 있는 새로운 생물정보학적 분석 기술들이 개발되고 적용되어야 할 것이다. 또한, 다양한 표현형을 high-throughput으로 스크린닝하고 분석할 수 있는 phenomics가 앞선 -omics들과 상호 협력적으로 발전해야 할 것이다. 식물에서 이런 phenomics가 가능하게 된 좋은 예는 virus-induced gene silencing 방법 (Liu et al. 2002)을 이용하는 것이다. 이 방법은 서로 다른 많은 유전자들의 돌연변이 표현형을 짧은 시간 동안에 빠르게 확인할 수 있는 스크리닝 방법이라는 점에서 생물정보학적 분석 결과들에 대한 high-throughput으로 유전자의 기능 연구를 가능하게 해주고 있다.

이제 다른 연구 분야와 마찬가지로 식물 생명공학 연구 분야에서 생물정보학은 필수 불가결한 동반자가 되어가고 있다. 그러므로 실험 방법과 생물정보학을 위한 컴퓨터 사이언스의 새로운 기술 개발에 생물학자와 생물정보학자 간의 긴밀하고 유기적인 협력관계를 더욱 발전시켜 나가야 할 것이라 생각한다.

적 요

애기 장대와 벼의 전체 게놈 염기서열 분석이 완료되었고, 다량의 EST 데이터가 많은 식물에서 이용 가능하게 되었다. 또한, 방대한 양의 다양한 생물학적 데이터들이 transcriptomics, proteomics, metabolomics와 같은 여러 '-omics' 기술에 의하여 만들어져 왔다. 생물정보학은 이런 방대한 양의 생물학적 데이터로부터 유용한 정보를 얻는데 필수적이고도 매우 중요한 역할을 수행한다. 이 총설에서, 우리는 대량의 데이터를 생성하는 실험적 방법들과, 식물 병 저항성과 분자 육종과 같은 식물 연구분야로의 응용, 그리고 식물 생명공학의 연구 개발에 유용한 생물정보학적 기술과 인터넷 정보 사이트들을 소개하였다. 우리는 새로운 실험 방법들과 생물정보학적 분석 기술들이 식물 생명공학 발전에 중요하게 기여할 것으로 기대하고 있으며, 생물정보학은 식물 생명공학의 연구 개발에 있어서 결정적인 요소가 될 것이라 생각한다.

인용문헌

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC (1991) Complementary-DNA sequencing expressed sequence tags and human genome project. Science 252: 1651-1656

Audic S, Claverie JM (1997) The significance of digital gene *expression profiles. Genome Res* 7: 986-995

Ausubel FM (2005) Are innate immune signaling pathways in plants and animals conserved? Nature Immunol 6: 973-979

Bae MS, Cho EJ, Choi E-Y, Park OK (2003) *Analysis* of the

Arabidopsis nuclear proteome and its response to cold stress. Plant J. 36: 652-663

Bedell JA, Budiman MA, Nunberg A, Citek RW, Robbins D, Jones J, Flick E, Rholfing T, Fries J, Bradford K, McMenamy J, Smith M, Holeman H, Roe BA, Wiley G, Korf IF, Rabinowicz PD, Lakey N, McCombie WR, Jeddeloh JA,Martienssen RA. (2005) Sorghum genome sequencing by methylation filtration. PLoS Biol 3: e13

Benton D (1996) Bioinformatics-principles and potential of a new multidisciplinary tool. Trends Biotechnol 14: 261-272

Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33: W451-454

Breyne P, Dreesen R, Vandepoele K, De Veylder L, Van Breusegem F, Callewaert L, Rombauts S, Raes J, Cannoot B, Engler G, Inze D, Zabeau M (2002) Transcriptome analysis during cell division in plants. Proc Natl Acad Sci 99: 14825-14830

Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94

Burke J, Davison D, Hide W (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. Genome Res 9: 1135-1142

Cannon SB, Crow JA, Heuer ML, Wang X, Cannon EK, Dwan C, Lamblin AF, Vasdewani J, Mudge J, Cook A, Gish J, Cheung F, Kenton S, Kunau TM, Brown D, May GD, Kim D, Cook DR, Roe BA, Town CD, Young ND, Retzel EF (2005) Databases and information integration for the Medicago truncatula genome and transcriptome. Plant Physiol 138: 38-46

Charless S Jr, Kang BC, Liu K, Mazourek M, Moore SL, Yoo EY, Kim BD, Paran I, Jhan MM (2005) The Pun1 gene for pungency in pepper encodes a putative acyltransferase. Plant J 42: 675-688

Chen W, Provart NJ, Glazebrook J, Katagiri F, Chang HS, Eulgem T, Mauch F, Luan S, Zou G, Whitham SA, Budworth PR, Tao Y, Xie Z, Chen X, Lam S, Kreps JA, Harper JF, Si-Ammour A, Mauch-Mani B, Heinlein M, Kobayashi K, Hohn T, Dangl JL, Wang X, Zhu T (2002) Expression profile matrix of Arabidopsis transcription factor genes suggests their putative functions in response to environmental stresses. Plant Cell 14: 559-574

Cuartero J, Bolarin MC, Asins MJ, Moreno V (2006) Increasing salt tolerance in the tomato. J Exp Bot 57: 1045-1058

Dangl JL, Dietrich RA, Richberg MH (1996) Death don't have no mercy: cell death programs in plant-microbe interactions. Plant Cell 8: 1793-1807

Dong S, Earls DB (1994) Gene structure prediction by linguistic methods. Genomics 23: 540-551

Edwards D, Batley J (2004) Plant bioinformatics: from genome to phenome. Trends Biotech 22: 232-237

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Nat Acad Sci 95: 14863-14868

Ewing B, Reen P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186-194

Fleming CM, Kowalski BR, Apffel A, Hancock WS (1999) Windowed mass selection method: a new data processing algorithm for liquid chromatography?mass spectrometry data. J Chromatogr A 849: 71-85

Glinski M, Weckwerth W (2006) The role of mass spectrometry in plant systems biology. Mass Spectrom Rev 25: 173-214

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchison D, Martin C, Katagiri F, Lange BM, Moughamer T, Xia Y, Budworth P, Zhong J, Miguel T, Paszkowski U, Zhang S, Colbert M, Sun WL, Chen L, Cooper B, Park S, Wood TC, Mao L, Quail P, Wing R, Dean R, Yu Y, Zharkikh A, Shen R, Sahasrabudhe S, Thomas A, Cannings R, Gutin A, Pruss D, Reid J, Tavtigian S, Mitchell J, Eldredge G, Scholl T, Miller RM, Bhatnagar S, Adey N, Rubano T, Tusneem N, Robinson R, Feldhaus J, Macalma T, Oliphant A, Briggs S (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 296: 92-100

Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. Trends Biotechnol 22: 245-252

Guigo R, Knudsen S, Drake N, Smith T (1992) Prediction of gene structure. J Mol Biol 226: 141-157

Gupta S, Zink D, Korn B, Vingron M, Haas SA (2004) Genome wide identification and classification of alternative splicing based on EST data. Bioinformatics 20: 2579-2585

Haberer G, Mader MT, Kosarev P, Spannagl M, Yang L, Mayer KFX (2006). Large-scale cis-element Detection by Analysis of Correlated Expression and Sequence Conservation between Arabidopsis and *Brassica oleracea* (pp. pp.106.085639)

Halket JM, Przyborowska A, Stein SE, Mallard WG, Down S, Chalmers RA (1999) Deconvolution gas chromatography/ mass spectrometry of urinary organic acids?potential for pattern recognition and automated identification of metabolic disorders. Rapid Commun Mass Spectrom 13: 279-84

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P,Brunak S (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res 24: 3439-3452

Henderson J, Sahasrabudhe S, Fasman KH (1997) Finding genes in human DNA with a hidden Markov model. J Comput Biol 4: 127-141

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

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP, Zody MC,Lander ES (2003) Whole-genome

sequence assembly for mammalian genomes: Arachne 2. Genome Res 13: 91-96

Jeong SC, Yang K, Park JY, Han KS, Yu S, Hwang TY, Hur CG, Kim SH, Park PB, Kim HM, Park YI, Liu JR (2006) Structure, expression, and mapping of two nodule-specific genes identified by mining public soybean EST databases. Gene (in press)

Jeppe RA, Thomas L (2003) Functional markers in plants. Trends Plant Sci 8: 544-560

Kang HY, Kim JF, Kim MH, Park SH, Oh TK, Hur CG (2006) MELDB: a database for microbial esterases and lipases. FEBS Lett 580: 2736-2740

Karas M, Hillenkamp F (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem 60: 2299-2301

Kawai-Yamada M, Jin L, Yoshinaga K, Hirata A, Uchimiya H (2001) Mammalian Bax-induced plant cell death can be down-regulated by overexpression of Arabidopsis Bax Inhibitor-1 (AtBI-1). Proc Nat Acad Sci 98: 12295-12300

Kruft V, Eubel H, Jansch L, Werhahn W, Braun HP (2001) Proteomic approach to identify novel mitochondrial proteins in *Arabidopsis*. Plant Physiol 127: 1694-1710

Kulp D, Haussler D, Reese MG, Eeckman FH (1996) A generalized hidden Markov model for the recognition of human genes in DNA. Proc Int Conf Intell Syst Mol Biol 4: 134-142

Lee S, Kim SY, Chung E, Joung YH, Pai HS, Hur CG, Choi D (2004) EST and microarray analyses of pathogen-responsive genes in hot pepper(*Capsicum annuum* L.) non-host resistance against soybean pustule pathogen (*Xanthomonas axonopodis* pv. *glycines*). Funct Integr Genomics 4: 196-205

Lisacek F, Cohen-Boulakia S, Appel RD (2006) Proteome informatics II: Bioinformatics for comparative proteomics. Proteomics 6: 5445-5466

Liu J, Rost B (2003) Domains, motifs and clusters in the protein universe. Curr Opin Chem Biol 7: 5-11

Liu Y, Schiff M, Marathe R, Dinesh-Kumar SP (2002) Tobacco *Rar1*, *EDS1*, and *NPR1/NIM1* like genes are required for N-mediated resistance to tobacco mosaic virus. Plant J 30: 415-429

Marathe R, Guan Z, Anandalakshmi R, Zhao H, Dinesh-Kumar SP (2004) Study of *Arabidopsis thaliana* resistome in response to cucumber mosaic virus infection using whole genome microarray. Plant Mol Biol 55: 501-520

Maleck K, Levine A, Eulgem T, Morgan A, Schmid J, Lawton KA, Dangle JL, Dietrich RA (2000) The transcriptome of Arabidopsis thaliana during systemic acquired resistance. Nature Genetics 26: 403-410

Mathe C, Sagot MF, Schiex T, Rouze P (2002) Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res 30: 4103-4117

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res 31: 374-378

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-Wide Analysis of NBS-LRR Encoding Genes in Arabidopsis. Plant Cell 15: 809-834

Nagayama K (1997) Protein arrays: concepts and subjects. Adv Biophys 34: 3-23

Newton RP, Brenton AG, Smith CJ, Dudley E (2004) Plant proteome analysis by mass spectrometry: principles, problems, pitfalls and recent developments. Phytochemistry 65: 1449-1485

Nielsen KL, Hogh AL, Emmersen J (2006). DeepSAGE-digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples (pp. gkl714)

Norbeli I, Thornton JM (2006) A bioinformatician's view of the metabolome. BioEssays 28: 534-545

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29-34

Palagi PM, Hernandez P, Walther D, Appel RD. (2006) Proteome informatics I: Bioinformatics tools for processing experimental data. Proteomics 6: 5435-5444

Park OK (2004) Proteomic studies in plants. J Biochem Mol Biol. 37: 133-138

Patterson SD, Aebersold RH (2003) Proteomics: the first decade and beyond. Nat. Genet. Suppl. 33: 311-323

Peltier JB, Friso G, Kalume DE, Roepstorff P, Nilsson F, Adamska I, van Wijk KJ (2000) Protomics of the chloroplast: systematic identification and targeting analysis of luminal and peripheral thylakoid proteins. Plant Cell 12: 319-342

Peltier JB, Ytterberg J, Liberles DA, Roepstorff P, van Wijk KJ (2001) Identification of a 350 kDa ClpP protease complex with 10 different Clp isoforms in chloroplasts of *Arabidopsis thaliana*. J Biol Chem 276: 16318-16327

Rajee KV, Andrease G, Mark ES (2005) Genomics-assisted breeding for crop improvements. Tends Plant Sci 10: 621-630

Rhee SY, Dickerson J, Xu D. (2006) Bioinformatics and Its Applications in Plant Biology. Annu Rev Plant Biol 57: 335-360

Richmond T, Somerville S (2000) Chasing the dream: plant EST microarrays. Curr Opin Plant Biol 3: 108-116

Rogic S, Mackworth AK, Ouellette FB. (2001) Evaluation of gene-finding programs on mammalian sequences. Genome Res 11: 817-832

Ronneberg TA, Freeland SJ, Landweber LF. (2001) Genview and Gencode : a pair of programs to test theories of genetic code evolution. Bioinformatics 17: 280-281

Salzberg S, Delcher AL, Fasman KH, Henderson J. (1998) A decision tree system for finding genes in DNA. J Comput Biol 5: 667-680

Seki M, Satou M, Sakurai T, Akiyama K, Iida K, Ishida J, Nakajima M, Enju A, Narusaka M, Fujita M, Oono Y, Kamei A, Yamaguchi-Shinozaki K, Shinozaki K (2004) RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. J Exp Bot 55: 213-223

Shulaev V (2006) Metaboloics technology and bioinformatics. Brief Bioinform 7: 128-139

Silivo S, Roberto T (2005) To clone or not to clone plant QTLs: present and future challenges. Trends plant science 10: 297-304

Silivo S, Roverto T, Elena C, Marco M, Stanislas V, Leon B, Peter I, Keith E, Ronald LP (2002) Toward positional cloning of Vgt1, a QTL controlling the transition from the vegetative to the reproductive phase in maize. Plant Mol Biol 48: 601-603

Smith CM, Shindyalov IN, Veretnik S, Gribskov M, Taylor SS, Ten Eyck LF, Bourne PE (1997) The protein kinase resource. Trends Biochem Sci 22: 444-446

Snyder EE, Stormo GD (1995) Identification of protein coding regions in genomic DNA. J Mol Biol 248: 1-18.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts (Vol. 34, pp. W435-439)

Stephen R, Heiko S, Klaus M (2005) PlantMarkers-a database of predicted molecular markers from plants. Nucleic Acids Research 33: 628-632

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28: 33-36

Terabe S, Markuszewski MJ, Inoue N, et al. (2001) Capillary electrophoretic techniques toward the metabolome analysis. Pure Appl Chem 73: 1563-1572

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

Viant MR, Rosenblum ES, and Tieerdema RS (2003) NMR-based metabolomics: a powerful approach for characterizing the effects of environmental stressors on organism health. Environ SciTechnol 37: 4982?4989

Walter G, Bussow K, Cahill D, Lueking A,Lehrach H (2000) Protein arrays for gene expression and molecular interaction screening. Curr Opin Microbiol 3: 298-302

Wan XY, Wan JM, Jiang L, Wang JK, Zhai HQ, Wang JF, Wang HL, Lei CL, Wang JL, Zhang X, Cheng ZJ, Guo XP (2006) QTL analysis for rice grain length and fine mapping of an identified QTL with stable and major effects. Theor Appl Genet 112: 1258-1270

Wang J, Delabie J, Aasheim H, Smeland E,Myklebost O (2002) Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. BMC Bioinformatics 3: 36

Watson JD, Crick FHC (1953) Molecular structure of nucleic acids-A structure for deoxyribose nucleic acid. Nature 171: 737-738

Wayne ML, McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. PNAS 99: 14903-14906

Windig W, Phalp JM, Payne AW (1996) A noise and background reduction methods for component detection in liquid chromatography/mass spectrometry. Anal Chem 68: 3602-3606

Wolter DA, Washburn MP, Yates JR 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem 73:5683-5690

Xu Y, Mural R, Shah M, Uberbacher E. (1994) Recognizing exons in genomic sequence using GRAIL II. Genet Eng (N Y) 16: 241-253

Yang Y, Shah J, Klessig DF (1997) Signal perception and transduction in plant defense responses. Genes Dev 11: 1621-1639

Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J,Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L,Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L, Liu J, Qi Q, Liu J, Li L, Li T, Wang X, Lu H, Wu T, Zhu M, Ni P, Han H, Dong W, Ren X, Feng X, Cui P, Li X, Wang H, Xu X, Zhai W, Xu Z, Zhang J, He S, Zhang J, Xu J, Zhang K, Zheng X,Dong J, Zeng W, Tao L, Ye J, Tan J, Ren X, Chen X, He J, Liu D, Tian W, Tian C, Xia H, Bao Q, Li G, Gao H, Cao T, Wang J, Zhao W, Li P, Chen W,Wang X, Zhang Y, Hu J, Wang J, Liu S, Yang J,Zhang G, Xiong Y, Li Z, Mao L, Zhou C, Zhu Z,Chen R, Hao B, Zheng W, Chen S, Guo W, Li G, Liu S, Tao M, Wang J, Zhu L, Yuan L, Yang H (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296: 79-92

Zhang MQ. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci 94: 565-568

Zhang W, Ruan J, Ho T-hD, You Y, Yu T, Quatrano RS. (2005) *cis*-Regulatory element based targeted gene finding: genome-wide identification of abscisic acid- and abiotic stress-responsive genes in *Arabidopsis thaliana.* Bioinformatics 21: 3074-3081

Zhou T, Wang Y, Chen J-Q, Araki H, Jing Z, Jiang K, Shen J, Tian D (2004) Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Gen Genomics 271: 402-415

(접수일자 2006년 10월 17일, 수리일자 2006년 10월 27일)