식물생명공학회지 제33권 제3호
J Plant Biotechnol
Vol. 33, No. 3, 161-169 (2006)

식물 대사체 연구의 진보

김석원[1], 정희일[2], 유장렬[3*]
[1]한국생명공학연구원 생물자원센터, [2]한양대학교 화학과, [3]한국생명공학연구원 식물유전체연구센터

Advances in Plant Metabolomics

Sukweon Kim[1], Hoeil Chung[2], and Jang R. Liu[3*]
[1]Biological Resource Center and [3]Plant Genome Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 305-333, Korea
[2]Department of Chemistry, Hanyang University, Seoul 133-171, Korea

ABSTRACT Plant metabolomics is a plant biology field for identifying all of the metabolites found in a certain plant cell, tissue, organ, or whole plant in a given time and conditions and for studying changes in metabolic profiling as time goes or conditions change. Metabolomics is one of the most recently developed omics for holistic approach to biology and is a kind of systems biology. For holistic approach, metabolomics frequently uses chemometrics or multivariate statistical analysis of metabolic profillings. In plant biology, metabolomics is useful to determine functions of genes often in combination with DNA microarrays by analyzing tagged mutants of the model plants Arabidopsis and rice. This review paper attempted to introduce basic concepts of metabolomics and practical uses of multivariate statistical analysis of metabolic profiling obtained by ^{1}H NMR and Fourier transform infrared spectrometry.

서 론

식물 대사체학 (metabolomics)은 식물에 존재하고 있는 모든 대사산물의 시간적, 공간적 변화를 추적 조사하는 연구 분야이다. 유전체학에서 제공되는 풍부한 생물학적 정보와 분석 기기의 발달에 힘입어 기능 유전체학의 연구수단으로 쓰여질 수 있을 뿐만 아니라 유용물질 생산 조절 등 다양한 연구 분야에서 사용됨으로써 폭넓은 관심을 불러일으키고 있다.

식물은 신진대사에 필요한 다양한 화합물을 가지고 있으며 그 중 많은 화합물이 의약, 염료, 향신료 등으로 이용되고 있어서 우리 생활에 매우 중요한 역할을 하고 있다. 식물계 전체에 존재하는 대사산물의 종류는 9만-20만 종류가 존재하는 것으로 추정되고 있다 (Roessner et al. 2001). 유전체 연구 모델식물인 애기장대의 경우 약 5천 종류의 대사산물을 가지고 있는데 이는 특정 미생물과 동물이 가지고 있다고 추정되는 1천5백과 2천5백 종류에 비하면 대단히 많은 숫자이다. 그러나 이와 같은 대사산물의 다양성은 genomics, transcriptomics, proteomics 등에서 다루고 있는 특정 개체의 유전자, transcript, 단백질의 종류에 비하면 매우 낮은 숫자이다. 또한 동일한 대사산물의 경우는 상동의 유전자, transcript, 단백질과는 달리 생물의 종류에 관계없이 동일한 화학구조를 가지므로 대사체 연구는 최근 연구 대상으로 부각되고 있는 여러 omics 가운데 가장 접근하기가 용이하다고 볼 수도 있다. 그러나 대사체 연구에 필수적인 분석 기기의 해상도 및 정밀도 문제는 대사체 연구의 커다란 제약 요인으로 작용하고 있다.

식물분야에서 최근에 대사체 연구에 관심을 가지게 된 직접적 계기는 식물 기능 유전체 연구와 연계되어 있다. 즉, 애기장대 혹은 벼와 같은 모델식물의 게놈 염기서열분석이 완료된 이후 확보된 무수한 유전자들의 기능을 규명하기 위하여 여러 분석 기법들이 활용되고 있다. 이중 대표적인 유전자

*Corresponding author Tel 042-860-4430 Fax 042-860-4608
E-mail: jrliu@kribb.re.kr

기능 결정 방법으로 특정 유전자가 tagged mutant를 제작하여 wild type과 형태적으로 나타나는 차이를 분석함으로써 tagged gene의 기능을 규명하는 방법이 있다 그러나 이와 같은 접근 방법으로는 구조 분석이 완료된 무수한 유전자들의 기능을 일일이 결정하기에는 분명한 한계를 가지고 있으므로 보다 빠르고 간편한 방법으로 유전자의 기능을 유추하기 위한 수단으로 최근 대사체 연구 기법들이 활용되고 있다. 형태적 표현형에 기초한 유전자 기능분석 연구에 있어서 또 하나의 문제점으로 애기장대의 tagged mutant의 90%이상이 형태적 변화를 수반하지 않는다는 점이다. 이런 경우 mutant의 metabolic profiling을 통해서 tagged 유전자의 기능을 인지할 수 있음이 밝혀졌다 (Fiehn et al. 2000). 이런 점에서 식물 대사체학은 인체를 대상으로 하는 대사체학이 질병의 biomarker를 찾는데 일차적 관심을 갖는 것과 현저하게 구별된다.

그러나 신진대사는 여러 생합성 혹은 생분해성 경로들이 복잡하게 연결되어 있으므로 모델식물의 tagged mutant의 metabolic profiling을 통하여 간단하게 해당 유전자가 해당식물의 신진대사에서의 기능을 결정할 수 있도록 허용하지는 않는다. 다만 형태적 phenotype이 관찰되지 않는 tagged mutant를 metabolic profiling을 통하여 wild type과 상호 구별되도록 할 뿐이며 tagged gene의 기능을 유추하기 위해서는 DNA microarray 등 다른 보조 수단을 동원해야 한다. 뿐만 아니라 필요에 따라서는 tagged mutant를 여러 조건에 노출시킨 후 얻은 다양한 metabolic profiling을 Bayesian network과 같은 통계적 방법을 보조적으로 사용해야 하는 경우도 있으며 전체적으로 보아 비록 metabolic profiling을 효과적으로 수행한다고 할지라도 현재의 기술수준으로는 해당 유전자의 기능을 정확하게 유추하는 것이 거의 불가능하다고 할 만큼 아직은 매우 초보적 수준에 있다고 할 수 있다.

식물 대사체학은 또한 식물 이외의 다른 생물을 대상으로 하는 대사체학과 마찬가지로 대사경로를 규명하는 목적을 가지고 있으며 특히 미생물 분야의 대사체학이 그러하듯 시스템 생물학 (systems biology) 입장에서 전체 metabolic flux를 수식으로 표시하고자 하는 방향으로 발전하고 있다. 본 총설은 metabolic profiling의 방법과 해석, functional genomics의 보조 수단으로 사용되는 식물 대사체학 그리고 시스템 생물학적 접근을 간단히 다루었다.

전체론적 접근 (Holistic Approach)

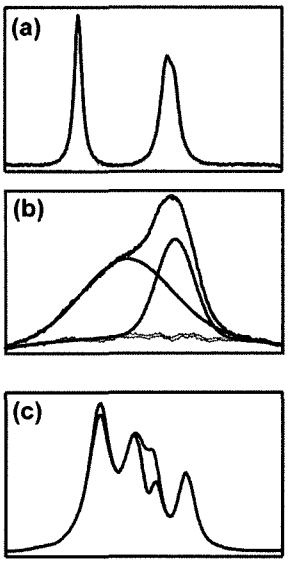
생물의 전 게놈 염기서열 결정이 이루어짐에 따라 생명과학은 기존의 특정 생명현상에 대한 환원주의 (reductionism)적 관점에서 탈피하여 생명현상 전체를 합목적적으로 이해하는 전체론적 접근을 가능케 하였다. Genomics, transcriptomics, proteomics, metabolomics와 같은 omics적 관점은 필연적으로 환원주의에서는 경험할 수 없었던 엄청난 양의 실험 대상군을 대하게 된다. 따라서 개별 실험을 robotics 등의 기법으로 자동화하고 단순화하며 한꺼번에 많은 실험을 동시에 실시하며 단위 실험의 시간을 최대한 단축할 수 있는 즉 High throughput system (HTS)을 갖추어야 하는 것이다. 뿐만 아니라 이로 인해 쏟아져 나오는 천문학적인 데이터를 다룰 수 있는 정보학을 필요로 하게 된다. 생물정보학 (bioinformatics)은 컴퓨터를 이용하여 엄청난 양의 데이터를 통계학적으로 처리하여 필요한 정보를 추출하는 학문이라고 할 수 있다. 이런 holistic approach는 생물을 궁극적으로 수리화 된 시스템으로 이해하게 하므로 시스템 생물학을 가능케 한다.

분석기기의 선택

현재의 기술수준으로는 어떤 한 종류의 분석기기로 metabolome 전체를 커버하는 profiling은 불가능하다 (Dunn et al. 2005). 분석기기의 종류에 따라 분석 가능한 대사산물의 종류가 한정되며 민감도에도 큰 차이가 있다. Functional genomics 입장에서는 다루어야 할 샘플의 수가 많으므로 분석기기의 처리 속도도 중요한 요소가 된다. 또한 대사산물을 정성적으로 파악하기 위해서는 두 가지 종류의 분석기기를 상호 연결하여 사용해야 할 경우가 많다.

일반적으로 식물 재료를 사용할 때는 HPLC를 사용하면 분석할 수 있는 범위가 넓고 정량적 데이터를 얻을 수 있으며 mass spectrometer (MS)와 연결하여 사용하면 authentic sample이 없어도 다양한 대사산물에 대한 정성분석이 가능해진다. 그러나 HPLC는 재현성이 떨어지는 단점이 있다. 이에 비해 gas chromatography (GC)는 HPLC가 분석하기 어려운 지방산 계열의 화합물을 대상으로 할 때 매우 유리하고 데이터의 재현성과 안정성이 HPLC에 비해 월등히 높다. 또한 HPLC와 마찬가지로 MS와 연결하여 사용하는 것이 일반적이다. 다만 GC를 사용할 때는 샘플을 유도체화 시켜야 하는 불편함이 있다.

한편 spectroscopy를 활용한 분석기기로써 NMR (nuclear magnetic resonance spectroscopy) 과 FTIR (Fourier transform infrared spectroscopy) 이 대사산물의 metabolic fingerprinting (지문 분석)에 활용되고 있다. NMR 중 특히 ^{1}H NMR은 LC나 MS와 연결하여 사용하기도 하지만 단독으로도 정량성이 뛰어나고 상당한 정도의 정성분석이 가능하며 샘플 전처리에 어려움이 없고 속도가 빠르다는 이점이 있다. 그러나 NMR은 GC에 비하여 민감도가 매우 낮은 단점을 가지고 있다. FTIR은 화합물의 정성분석적 정보를 별로 제공하지 못하지만 샘플의 전처리, 속도, 민감도 등에서 발군의 성적을 보여준다. 특히 최근에는 한꺼번에 384개의 샘플을 처리할 수 있도록



Simple Analysis - Band Ratios
- area
- width and shape

Rapid and Robust

Peak Fitting - Curve Fit / Band Ratios
- position
- area
- width and shape

Robust and Intelligent

PLS, PCA, PCR, Correlation
- all spectral features together

Sensitive

Figure 1. Three different spectral data sets requiring different methods of spectral analysis; a) band area or band ratio, b) curve fitting and c) multivariate approach.

고안된 기종도 출시되어 다량의 시료분석에 있어서 HTS 분석이 가능하다.

최근에 MS를 중심으로 괄목할 만한 기술적 진보가 이루어지고 있다. 이에 따라 capillary electrophoresis mass spectroscopy (CE-MS)가 대사체학에 광범위하게 사용되고 있는데 이는 비용, 선택성 (selectivity), 민감도, 정량정성 분석의 정도, 속도, 분석기기의 가격 등 다양한 결정 요소를 적절하게 타협적으로 만족시키기 때문이다. FT-MS는 단번에 수백 혹은 일천 종 이상의 대사산물의 정량정성 분석이 가능하다는 이점 때문에 주목을 받는 분석기기이나 현재로서는 고가이므로 일반 연구자들이 사용하기 어려운 것이 현실이다.

Metabolic Profiling과 데이터 전처리 (Data Preprocessing)

Metabolic profiling은 targeted와 nontargeted profiling으로 대별할 수 있다. Targeted profiling은 세포내의 전체 대사산물 중에서 특정한 대사산물의 양적 질적 변화에 대한 분석을 하는 것으로 LC-MS, GC-MS 등 분석기기를 통해 주로 분석이 이루어진다. 이에 반하여 nontargeted profiling은 특정한 대사산물보다는 세포내의 모든 대사산물에 대한 전반적인 패턴 변화를 조사하는 방법으로 metabolic fingerprinting으로 불리기도하며 비교적 간편한 추출과정을 통해 주로 FTIR 및 ^{1}H NMR을 통해 분석이 이루어진다.

각 시료의 분석을 통해 얻어지는 metabolic profiling 데이터는 매우 방대한 규모로 다변량 통계분석의 여러 도구들을 활용하여 의미 있는 화학적 생물학적 정보를 추출하게 된다. 그러나 이와 같은 metabolic profiling 데이터는 baseline 보정, normalization 등의 전처리 과정 등 데이터 전처리 과정을 거친 다음 통계분석이 이루어진다. 그러나 아직까지 초기 metabolic profiling 데이터의 전처리 과정에 관한 국제적인 공통 기준이 제시되어 있지 않은 상태이다. Metabolic profiling 데이터의 분석 결과에 대한 재현성을 보다 높이기 위하여 현재 metabolic profiling 데이터의 분석결과 제출시 사용한 분석 기장비명, 데이터 전처리에 사용한 알고리즘, 다변량 분석에 사용한 알고리즘 등 제반 요인에 대한 보다 상세한 기술과 데이터 처리과정의 규격화가 국제적으로 진행되고 있으며 향후 국제 공통 기준에 따른 metabolic profiling 데이터의 분석이 이루어질 것으로 예상된다.

생물정보학과 계량화학 (Chemometrics)

대사체학에서 주로 사용되는 생물정보학 기법으로는 계량화학 (chemometrics)에서 주로 다루는 다변량 분석법 (multivariate analysis)을 들 수 있다. 다변량 분석에는 principal component analysis (PCA, 주성분분석법), partial least squares (PLS, 부분최소자승법) hierarchical cluster analysis (HCA) 등이 주로 사용되고 있으며 데이터의 구조나 목적에 따라 적당한 알고리즘을 선택하여 사용한다.

일반적인 분석화학 데이터를 크게 구분하면 Figure 1과 같이 3가지 경우로 분류 할 수 있다. Figure 1(a) 처럼 시료에 2개의 순수 물질이 혼합 되어있고 얻어진 스펙트럼에서 각각의 피크가 명확하게 분리 되는 경우 (또는 크로마토그램), 각각 물질의 정량분석은 피크 높이나 피크면적을 이용하여 쉽게 정량분석이 가능하다. Figure 1(b) 처럼 2개의 피크가 겹쳐

있더라도, 2개의 순수물질이 존재한다는 사실이 명확하면 curve fitting 방법을 통하여 2개의 피크로 분리가 가능하며, 분리된 각각의 피크가 정량/정성분석 목적으로 사용 될 수 있다. Figure 1(c)에서 보는 바와 같이 시료의 조성이 복잡하고 또한 몇 개의 순수물질이 섞여있는지 추정이 쉽지 않을 경우 (예: 천연물), 얻어진 스펙트럼을 정성 및 정량 목적으로 사용하기가 매우 어려워진다. 위와 같이 많은 변수를 (스펙트럼의 경우 많은 수의 파장 또는 전체 파장) 이용하여 스펙트럼과 조성/물성간의 상호관계를 만들고자 할 때 사용 할 수 있는 방법이 계량화학이다. 계량화학은 데이터 자체의 변수의 수가 많거나, 또는 이론적인 해석이 쉽지 않아 기존의 일반 회귀분석으로는 상관관계를 찾기 어려울 때 사용하게 된다. 데이터의 종류는 분광스펙트럼, 크로마토그램, 볼타모그램 등 다양한 분석데이터에 적용이 가능하다.

Principal Component Analysis (PCA, 주성분분석법)

PCA의 주 목적은 다변량의 데이터로부터 가능한 한 적은 변수로 원래 data set에 있는 변이량 정보를 파악해 집약 시킴으로써 궁극적으로 변수를 감소시켜 데이터를 쉽게 이해하고자 하는 것이다. 예를 들어 고등학교 3학년 1반에 속한 학생들이 국어, 영어, 수학, 과학 등 네 과목에 대해 각각 100점 만점의 기말시험을 보았다고 하자. 일반적인 성적순은 개별 학생이 네 과목에서 얻은 점수의 합의 순서에 따라 석차를 매기는 것이다. 그러나 이 방법은 국어의 1점과 수학의 1점을 동일시하였다는 모순이 있다. 따라서 정확히 석차를 내려면 개별학생의 성적은 네 과목을 각각 독립된 축으로 하는 4차원 공간에서 벡터 값의 합으로 정해지는 공간의 한 점으로 표시하는 것이 될 것이다. 그러나 이 방법 역시 4차원이라는 우리의 공간적 인식을 뛰어 넘는 차원에 표시된 것이므로 전체적 이해에 큰 한계를 드러낸다. 후자의 한계를 벗어나는 방법으로는 각 변수들 간의 상관관계를 조사하여 x축을 언어적 영역으로 보고 국어와 영어 과목의 합으로 표시하고 y축을 수리적 영역으로 간주하고 수학과 과학 점수의 합을 표시하여 2차원 공간에서 학생들의 석차를 표시하는 방법을 생각해 볼 수 있다. 이렇게 되면 단순히 네 과목의 합으로 석차를 정하는 것보다 언어적 또는 수리적 영역에서의 학생들의 성적을 조망할 수 있어서 보다 풍부한 학습지도에 대한 정보를 추출할 수 있게 된다.

이와 마찬가지로 예컨대 담뱃잎의 전체 추출물을 ^{1}H NMR로 분석하여 얻은 profile은 excel sheet에 표시하면 32,000 data point를 넘기게 되는데 이를 다변량 분석법 중 가장 대표적인 PCA를 통하여 2차원 공간에 한 개의 점으로 표시할 수 있다. 다만 이때 x와 y축은 학생들의 성적 순의 예와 달리 언어적 혹은 수리적 영역이라는 특별한 의미를 가지지 않으며 데이터 상호간에 correlation이 존재하지 않는 독립된 두 개의 영역일 뿐이며 구체적 의미는 없다. 또한 PCA의 x와 y축 값을 각각 PC1 및 PC2라고 하고 그 값을 백분율로 표시하는데 이 두 개의 백분율을 합한 값이 가령 91%라고 하면 수만 개의 축으로 표시해야 할 본래의 profile 데이터를 2차원 공간에 표시함으로써 발생하는 데이터의 정확성의 손실이 9%라는 뜻이 된다. 실제로 PCA 분석결과는 스코어와 loading의 산포도를 통해 데이터의 특징을 해석함으로써 유용한 정보를 추출할 수 있다. 스코어는 주성분 공간에서 각 시료의 좌표인데, 이들을 플롯한 산포도는 시료 간의 관계를 나타낸다. 즉 산포도상에서 인접하여 위치할수록 시료 간에 비슷한 성질을 나타내고 있는 것으로 추정할 수 있다. 또한 산포도상의 공간적 위치에 따라 시료 간에는 몇 개의 그룹으로 구분이 가능하다. 로딩은 주성분 축 (고유 벡터)의 계수로서 각 주요 성분에 대해 각 변수들이 기여한 정도를 파악함으로써 어느 변수가 중요하며 불필요한지를 알 수 있게 된다. 즉 PCA를 통한 패턴인식을 해 보면 여러 개의 서로 다른 샘플들이 2차원 공간에 표시되면서 샘플들 간의 특정 공통점에 의해 클러스터링이 이루어짐을 알 수 있다. (혹은 공통점이 없어서 클러스터링이 되지 않는다.) 이때 클러스터 상호간 혹은 클러스터링 내의 샘플간의 근원관계는 HCA 분석에 의해 dendrogram으로 보다 용이하게 인식되도록 표시할 수 있다.

PCA는 샘플에 대한 어떤 정보도 사전에 주어지지 않고 모든 샘플을 대등하게 다룬다. 이런 통계분석법을 unsupervised learning method라고 한다. 그러나 샘플이라 해도 어떤 것은 서로 다른 처리구에 속하는 것이 있고, 또 어떤 샘플은 특정 처리구의 반복으로 사용되는 것이 있다. 따라서 일반 샘플과 달리 특정 처리구의 replicate로 사용되는 샘플들은 기본적으로 동일 혹은 대단히 유사한 값을 가질 것을 실험자는 기대하고 있다. 이에 반해 처리구 상호간에는 상이점이 극대화되어 나타나기를 기대한다. 이런 기대치가 반영되도록 하는 기법을 supervised learning method라고 하며 discriminant function analysis 등이 대표적으로 이에 속한다.

PCA는 원래의 스펙트럼을 그대로 사용하는 것이 아니라, 스펙트럼을 어떤 기준스펙트럼과 scaling factor를 이용하여 재설정하여 분석하는 방법이다. 개념적인 이해를 돕기 위하여 PCA의 개략도를 Figure 2에 나타내었다. A는 원래의 스펙트럼 데이터 세트이며, n개의 스펙트럼과 p개의 파장으로 이루어져 있다. F는 principal component (PC) 메트릭스로 p개의 파장으로 이루어진다. PC는 또한 eigenvector, spectral loading, loading vector, factor (일반적으로 가장 많이 사용)라고도 한다.

F의 factor들은 원래 스펙트럼들의 변화를 가장 잘 설명하는 모양으로 설정되며 결국 eigenvector에 해당된다. Factor를

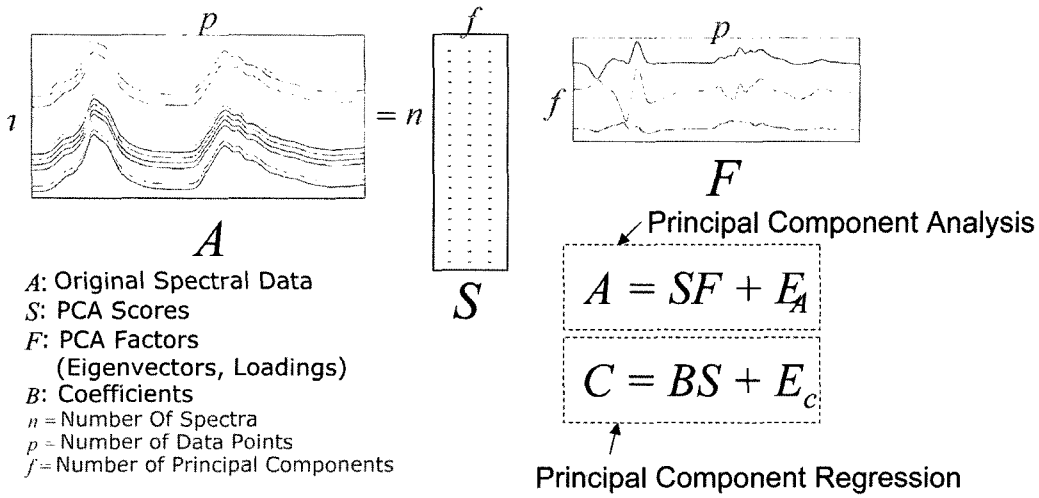


Figure 2. Schematic description of overall procedure of principal component analysis (PCA).

계산하는 프로그램은 아주 쉽게 구할 수 있으며 여기에서는 계산과정에 대한 설명은 생략하기로 한다. 결국 스펙트럼 변화를 잘 설명할 수 있도록 factor를 구한 다음 각각의 원래 스펙트럼을 factor에 투영 (projection)하면 투영되는 정도에 따라 -1과 1사이의 스코어를 가지게 된다. 결국 스코어는 투영된 스펙트럼과 factor (eigenvector)와 cosine값을 의미한다. 스코어가 1이면 투영된 스펙트럼이 factor (eigenvector)와 정확하게 같은 방향이며 (0^{o}), -1은 정확하게 반대방향 (180^{o})임을 의미한다. 스코어가 0이면 직각 (90^{o})임을 의미한다. 따라서 스펙트럼들의 모양이 조금이라도 다르면 다른 스코어를 가지게 된다. 일반적으로 첫 번째 factor가 데이터세트 내에서 가장 큰 변화를 설명하고 (가장 큰 eigenvalue), 두 번째 Factor는 첫 번째 factor에서 설명하지 못한 나머지 변화를 설명하다. 즉 나중에 있는 Factor일수록 데이터세트의 변화를 설명하는 정도가 점점 작아지게 된다 (eigenvalue 값이 작아진다). 위의 Figure 2의 경우는 3개의 Factor를 이용한 경우이며, 3개의 factor를 사용해서 스펙트럼을 각각 3번씩 투영하였기 때문에, 스펙트럼마다 3개의 스코어를 가지게 된다. 결국 메트릭스 A의 스펙트럼들의 변화가 단 3개의 스코어로 표현이 가능하게 되는 것이다. 위와 같이 스펙트럼 전체의 변화를 3개의 변수 (스코어)로 간단하게 표현 할 수 있기 때문에 스펙트럼 데이터 양이 아주 많을 때 전체적으로 변화를 쉽게, 그리고 효과적으로 나타낼 수 있다.

Figure 3은 스펙트럼을 PCA 스코어로 재표현한 예를 보여주고 있다. 수입산 (주로 중국) 및 국산 당귀들을 각각 163, 407개 수집하여 확산 반사를 이용하여 근적외선 스펙트럼을 측정하였고, 또한 그림에서 수입 및 국산 당귀 스펙트럼의 평균을 보여주고 있다. 생물학적으로 보면 국산 당귀와 중국 당귀는 종이 다르기 때문에 포함 된 대사 물질이 약간 다르며,

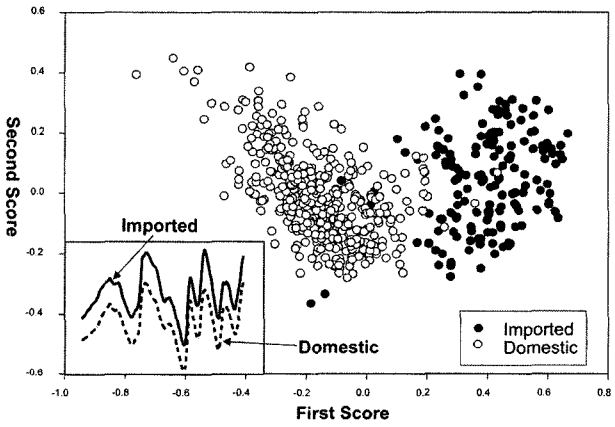


Figure 3. Representation of spectral data (Korean and Chinese herbal medicine) into two dimensional score domain using first and second scores.

특히 지표 물질인 decursin의 함량의 차이가 있는 것으로 알려져 있다. 당귀와 같은 천연물들은 조성이 매우 복잡하기 때문에 스펙트럼 모양이 전체적으로 매우 비슷하며, 위와 같이 스펙트럼의 개수가 많아지면 스펙트럼들을 일일이 비교 한다는 것은 현실적으로 매우 힘든 일이다. 따라서 원래의 스펙트럼 정보를 왜곡하지 않으면서 쉽게 재 표현하는 방법이 필요하게 되며, 이런 경우에 PCA가 효과적으로 사용 될 수 있다. 아래 그림은 전체 스펙트럼을 첫 번째 및 두 번째 스코어만을 이용하여 표현 한 것이다. 보는 바와 같이 2개의 군이 약간은 겹쳐 있지만 명확하게 구분되는 것을 볼 수 있다.

일반적으로 PCA를 포함한 계량화학 기법들은 어떠한 이론적인 상관관계에 근본을 둔 방법이 아니라 많은 양을 데이터를 이용한 통계적 및 경험적인 모델링 기법이기 때문에 검증 (validation) 과정을 반드시 거쳐야 한다. 다시 말해서 독립적인 시료를 이용해서 얻어진 PCA 모델이 Figure 3과 같은 정도로 2대의 그룹으로 나누어 지고 있는 것을 독립적인 데

이터를 이용해서 검증하여야 한다는 뜻이다. 독립적인 시료가 어느 그룹에 속하는지를 결정하는 알고리즘들은 많이 개발 되어 있으나 일반적으로 화학분야에서는 Mahalanobis distance (MD) 방법을 많이 쓴다. MD는 어떠한 데이터 그룹의 경계선을 설정하여 주는 방법으로 Figure 4에 개념도가 나타나 있다. 일반적으로 Euclidian distance 방법으로 평균점에서 같은 거리에서 경계선을 설정하면 데이터 분포에 따라 많은 오차를 유발하게 되지만, Figure 3과 같이 MD는 데이터 분포를 감안하여 표준편차가 큰 쪽으로 weight를 주어 경계선을 설정하는 방법으로 경계선의 정확도를 높일 수 있다. 물론 그림과 같이 상대적으로 단순한 분포를 가지지 않고 복잡한 분포를 가질 경우, 다른 방법들을 사용해야 하지만, 여기에서는 논외로 한다. 결국 새로운 시료의 스펙트럼을 PCA로 분석하며 얻어진 스코어들이 95% 신뢰수준에서 그룹 경계선내에 있으면 같은 군으로, 그리고 바깥쪽에 위치하면 다른 군으로 예측하게 된다.

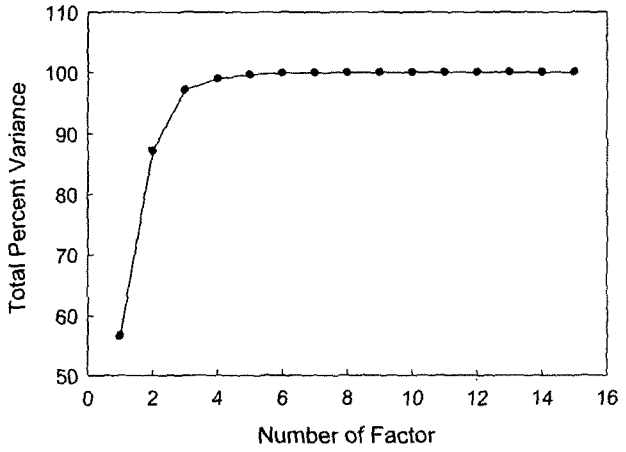


Figure 4. Variation of TPV (total percent variation) as a function of the number of principal component (PC).

또한 정확한 PCA 분석을 위해서는 최적의 factor 개수를 정해야 한다. Figure 5는 최적의 factor 개수를 정하는 과정을 나타내었다. x축은 factor의 개수이며 y축은 TPV (total percent variation)로 표현 되었다. Eigenvalue 값의 합은 1이며 각각의 eigenvalue를 전체에 대한 퍼센트로 표현한 값이 TPV이다. Figure 4에서 첫 번째 factor에 해당되는 TPV값이 57% (eigenvalue 값은 0.57)이며 데이터세트 내 스펙트럼 변화 중 57%를 첫 번째 factor로 설명이 가능하다는 뜻이다. 두 번째 factor에 해당되는 TPV값은 88%로 주어진 첫 번째 및 두 번째 factor를 누적하며 사용하면 전체 스펙트럼 변화의 88%를 설명하고 있다는 뜻이다. 두 번째 factor를 사용하여 전체변화 중 추가로 21%를 설명하였다. 점점 추가되는 factor들에서 스펙트럼 변화를 설명하는 정도는 점점 줄어 들게 된다. 따라서 적당한 factor수를 선정해야 하며, 일반적으로 전체 변화 중 99.5% 또는 95%까지 설명 할 때까지 factor를 누적하여 사용하거나, t-test등을 통하여 추가 된 factor가 데이터를 설명하는데 통계적으로 크게 향상이 되었는지를 판단하여 결정 할 수도 있다.

Partial Least Squares (PLS, 부분최소자승법)

계량화학 알고리즘 중에서 정량목적으로 가장 많이 사용하는 방법이 부분최소자승법이다. PLS는 스펙트럼 데이터뿐만 아니라 시료의 농도까지 동시에 이용하여 농도변화를 가장 설명 할 수 있는 factor들을 설정하는 방법이다.

PLS방법의 개념도를 Figure 5에 나타내었다. 스펙트럼의 변화는 항상 주어진 성분의 농도 변화에 따라 가장 크게 나타나지는 않을 수 도 있다. 실질적으로 기기변화에 따른 바탕선 변화, 잡음 (noise), 시료 상태에 따른 산란 등이 존재할 때 스펙트럼은 농도의 변화보다는 이런 현상들에 의해 더욱 크게

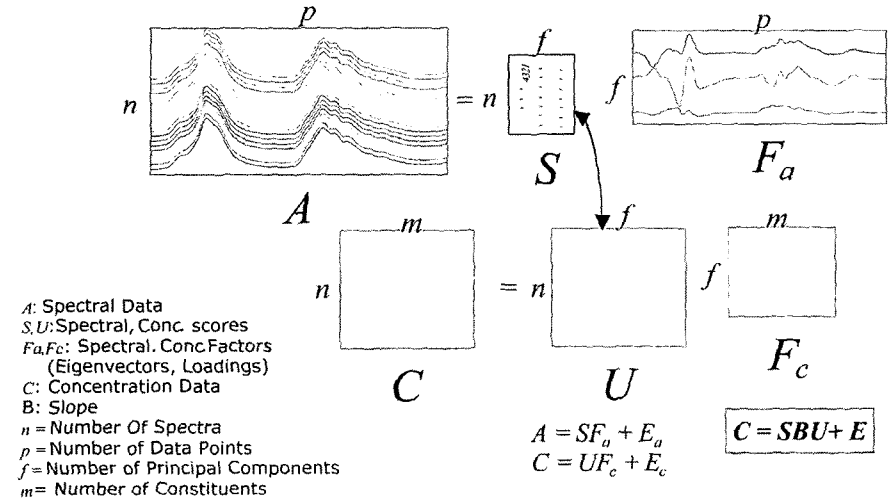


Figure 5. Schematic description of overall procedure of partial least squares (PLS) regression.

변화 할 수 있다. 그러나 PLS를 사용하면 factor 설정 시 농도 정보가 이용되기 때문에, 측정성분과 관계없는 다른 스펙트럼 변화가 존재 할 때 농도변화를 더 설명을 잘 할 수 있는 factor가 설정 될 수 있고, 이에 따라 분석 성능도 향상될 수 있다.

검량세트 (calibration set)의 스펙트럼을 이용하여 PLS를 포함하는 다변량 회귀분석법을 이용하여 정량 할 경우, 실제농도와 계산된 농도간의 오차가 최소화 되도록 검량식이 작성되기 때문에, 잡음이 많은 영역이나 측정성분과 관계없는 정보를 이용하여 마치 측정성분의 농도 변화를 잘 설명하는 것처럼 보일 수 있다. 그러나 다른 시료들을 예측해보면 예측결과가 만족스럽지 못한 경우가 발생한다. 이는 다변량회귀분석법의 단점으로, 측정 물질의 농도변화와 관계없는 다른 스펙트럼정보를 이용해 측정물질의 농도변화와 상관관계를 만들 수 있는 위험성도 내포하고 있다. 이런 over-fitting을 방지하기 위해서 독립적인 예측세트를 (prediction set) 이용해서 개발한 검량식을 반드시 검증해야 한다. 일반적으로 검량식을 만들 때 cross validation (CV)을 사용하여 검량세트 내에서 우선 내부 검증을 한다. CV방법은 검량세트 내 시료 중 일부분을 제외하고 나머지 시료를 이용해서 검량식을 만들고, 제외된 나머지 시료를 예측하여 검량식을 검증한다. 예를 들면 50개 시료가 있으면 45개를 이용해 검량식을 만들고 나머지 5개를 예측하며, 모든 시료가 예측될 수 있도록 10번 이 과정을 반복한다. 만약 1개의 시료만 예측한다면 50번 반복되어야 한다. CV후 SECV (standard error of cross validation)을 계산하며, factor (loading) 개수 변화에 따른 SECV 변화를 관찰하여 최적의 factor수를 정한다. Figure 6은 factor 개수가 증가함에 따른 전형적인 SECV 변화를 보여주고 있다. 첫 번째 factor부터 초기의 몇 개의 factor까지는 SECV가 급격하게 떨어지며 그 후 완만하게 떨어진다. 이는 factor수를 증가함에 따라 농도와 관계된 정보가 검량식에 계속 반영되기 때문이다. 그러나 5번째 factor부터는 SECV가 커지며 더 이상 크게 향상되지 않는다. 이는 잡음이나 농도와 관계없는 필요치 않은 정보를 검량식에 반영시키기 때문에, 결국 오차는 줄어들지 않는다. 이 경우 오차가 최저인 4개 factor를 선정하는 것이 일반적이다. 이렇게 검량식을 작성 한 후 별도의 예측세트를 예측하여 표준예측오차 (standard error of prediction, SEP)를 확인 및 검증해야 한다.

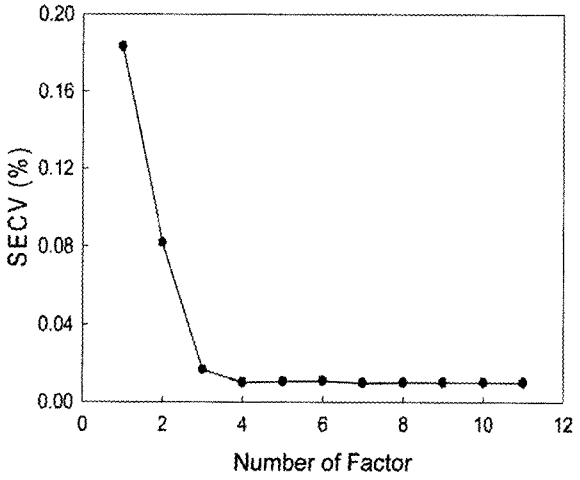


Figure 6. Variation of SECV (Standard Error of Cross Validation) as a function of the number of principal component (PC, or factor).

PLS대 대한 이해를 돕기 위해서 사용례를 들어 보기로 하자. 예컨대 HPLC와 GC를 이용하여 식물의 추출물에 함유된 해당 화합물의 정량정성 분석이 가능하지만 이 방법은 상대적으로 시간이 많이 소요되므로 분석해야 할 시료가 수천 가지가 된다면 HTS으로 사용할 수 있는 대안을 생각해 보아야 할 것이다. 이때 PLS를 이용해 볼 수 있다. HPLC를 이용하여 벼 낟알에 함유된 sucrose, glucose, fructose 함량을 정량분석하며, GC로 linoleic acid, oleic acid, stearic acid 등 지방산의 함량을 정량 분석한 다음 동일 시료에 대한 FTIR 스펙트럼 데이터를 확보하여 얻어진 FTIR 스펙트럼 데이터와 HPLC와 GC로 해당 화합물의 정량 데이터를 연계하여 각 화합물에 대한 regression analysis equation을 결정한다 (낟알 20개 정도의 개별 낟알에 대한 HPLC와 GC의 정량 데이터로 regression equation이 성립될 수 있다 (Figure 7). 따라서 FTIR은 비파괴적으로 시료를 다루므로 각 낟알에 대해 FTIR로 profile을 얻은 후 동일한 낟알에 대한 HPLC 혹은 GC 데이터를 확보할 수 있다. 일단 regression analysis equation이 확보되면 그 후에는 HPLC나 GC를 사용하지 않고 FTIR 분석만으로 해당 화합물의 정량분석을 빠르게 진행시킬 수 있다.

Metabolic Fingerprinting과 Genetic Programming

서로 다른 품종의 농산물을 metabolic profiling 만으로 판별하려고 한다면 굳이 각 품종의 구성 화합물을 정성적으로 결정할 필요는 없다. PCA와 같은 기법을 이용하여 metabolic profile의 패턴인식법을 이용하면 된다. 이 때 자주 사용되는 분석기기는 FTIR, ^{1}H NMR 등이다. Figure 8에서 보는 바와 같이 동일 공간과 기간 동안 재배된 여러 종류의 식물의 잎을 샘플로 하여 FTIR로 profiling을 한 후 이를 PCA로 분석하면 클러스터링이 나타나고 이들 샘플 상호간의 관계를 HCA를 이용하여 dendrogram으로 표시하면 이들 식물간의 계통분류학적 관계와 일치하게 된다 (Kim et al. 2004). 이때 사용한 방법은 식물 전추출물의 개개의 구성물질에 대해서는 외면한 채 단지 profile의 패턴만을 이용한 것이며 이를 metabolic fingerprinting이라고 한다. 이 방법은 유전적으로 동일한 농산물이라 하더라도 재배 원산지에 따른 차이도 밝혀 낼 수 있으

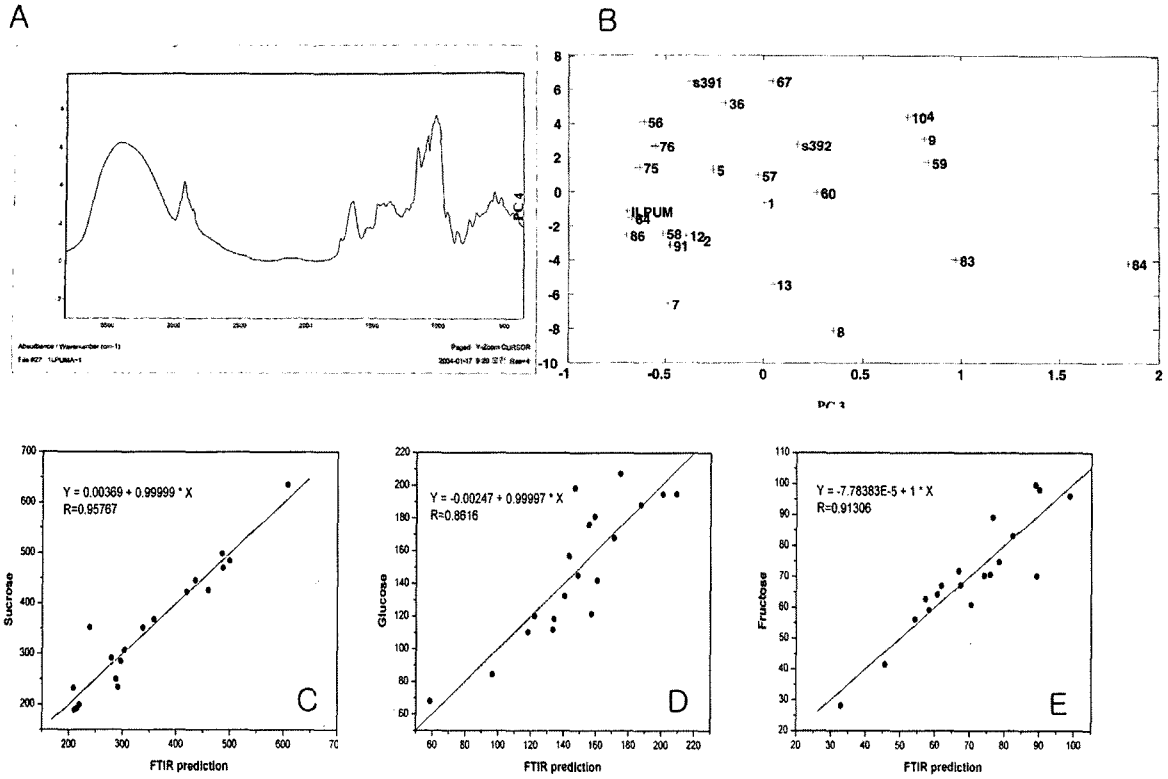


Figure 7. A: A representative FTIR spectrum of a rice grain; B: PCA plot of FTIR spectrum from rice grains; C, D, E: Linear regression of sucrose, glucose, and fructose derived from PLS of rice grains, respectively.

므로 농산물 원산지를 확인하는데 사용될 수 있다. 또한 인삼과 같이 다년생 농산물의 연령을 구별하는데도 사용될 수 있다.

아울러 metabolic fingerprinting으로 일일초의 여러 품종을 ^{1}H NMR로 판별할 수 있는데 이때 genetic program (Kell et al. 2001)의 도움을 받으면 복잡한 profile에서 어느 부분에 의해 품종간의 차이가 발생하는지를 쉽게 결정할 수 있다 (Kim et al. 2006).

식물 대사체학과 Functional Genomics

FANCY (functional analysis by co-responses in yeast): 효모 (*Saccharomyces cervisae*)의 knockout mutant를 개별적으로 metabolic profile한 후 이를 PCA와 DF로 표시하면 유전자의 기능에 따른 클러스터링이 나타나게 된다 (Raamsdonk et al. 2001). 이를 FANCY라고 하는데 이 방법을 이용하여 식물의 tagged mutant의 metabolic profiling을 통하여 기지의 유전자에 tagged mutant와 미지의 유전자에 tag된 mutant가 함께 클러스터링이 되면 미지의 유전자의 기능을 기지의 유전자의 기능으로 유추해 볼 수 있을 것으로 전망된다.

Metabolome과 transcriptome의 연계: Arabidopsis의 영양 및 이차대사 과정에 관련된 transcriptome 데이터와 FT-MS 등의 metabolome 데이터와 연계를 통하여 식물의 대사과정을 총체적으로 해석 접근하는 시도가 이루어지고 있다 (Hirai et al. 2004). 이는 tagged mutant 만으로 결정할 수 없는 많은 유전자의 기능을 효과적으로 밝히기 위한 방법으로 애기장대의 특정 조건과 식물부위에서의 FT-MS를 이용한 nontargeted metabolic profiling과 DNA microarray를 사용하여 해당 mutant를 유발하는 유전자의 기능을 보다 효과적으로 유추할 수 있게 된다.

결 론

Metabolome은 유전자 발현의 최종산물로 궁극적으로 표현형 변화의 주요한 원인이 된다. 즉 metabolome은 생물체의 genotype과 phenotype을 연계하는 가교 역할을 하고 있는 것으로 이해된다. 식물에서는 다양한 tagged mutant를 생산해 내는 것이 용이하며 모델 식물 (애기장대 및 벼)에 대해서는 DNA microarray가 판매되고 있으므로 이를 이용한 transcriptome 데이터를 확보할 수 있다. 또한 FT-MS 등의 분석기기를 이용하여 대량의 metabolome 데이터가 얻어지므로 이들 양자의 데이터를 상호 연계하기 위한 생물정보학적 기법이 향후 개발되어야 한다. 이와 같은 유전체 데이터와 대사체 데이터의 연관분석 체계는 유전자 기능 분석 및 정의를 한층 더 가속화 시킬 수 있을 것으로 예상된다. 아울러 Bayesian network 등의 통계방법을 이용하여 관련 대사경로와 관련 유전자를 규명할 수 있을 것으로 전망된다. 이와 같은 여러 omics 분석의 연계는 생

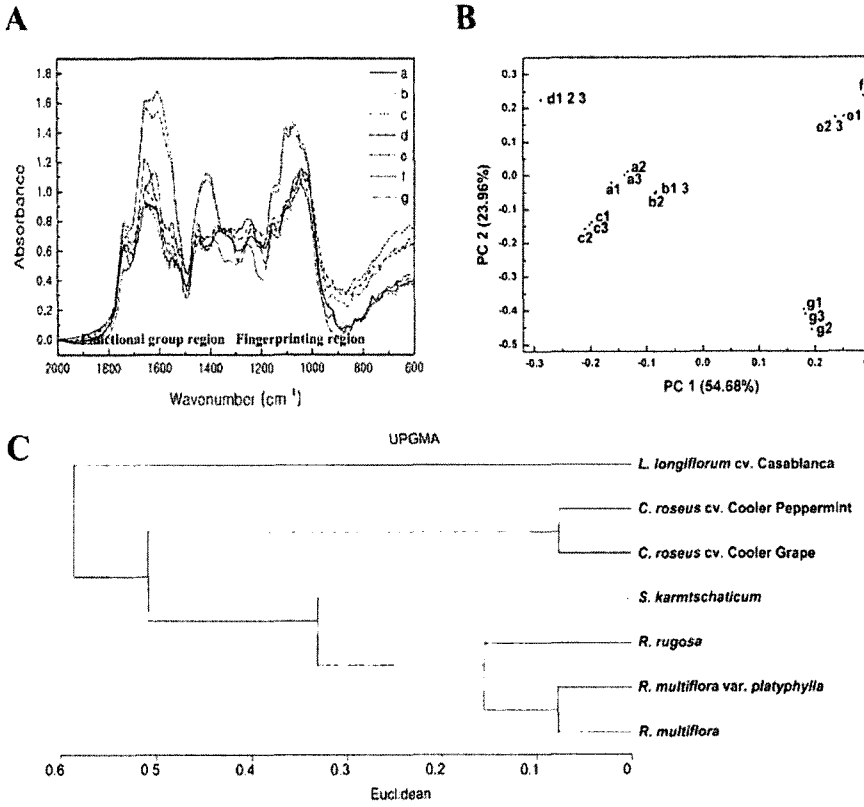


Figure 8. FTIR spectra, PCA of FTIR data, and a dendrogram based on PCA of FTIR data from seven plants. A: Representative FTIR spectra of seven plants; B: PCA of FTIR data from seven plants; C: A dendrogram based on PCA of FTIR data from seven plants.

명현상을 유전자 하나가 아닌 총체적 접근 즉 시스템 생물학을 통한 해석이 가능해질 것으로 전망된다.

적 요

식물대사체학은 식물세포, 조직, 기관, 혹은 개체수준에서 주어진 시간과 조건에서 발견되는 모든 대사물질을 밝히고, 시간의 경과 혹은 조건의 변화에 따른 metabolic profiling의 변화를 연구하는 식물학 분야이다. 식물대사체학은 생물에 대한 전체론적 접근을 위한 가장 최근에 발달된 omics 분야의 하나로서 일종의 시스템 생물학이다. 전체론적 접근과 이해를 위해서 대사체학은 metabolic profiling의 계량화학 혹은 다변량분석 방법을 자주 사용한다. 식물학 분야에서 대사체학은 애기장대나 벼와 같은 모델식물에 tag를 도입하여 형질전환시킨 돌연변이체에 대해 DNA microarray와 함께 사용하여 유전자의 기능을 밝히는데 유용하게 사용된다. 본 총설에서는 식물대사체학의 기본 개념과 1H NMR 혹은 FTIR으로 얻은 metabolic profiling의 다변량분석에 대한 실용적인 사용법을 소개하고자 하였다.

사 사

본 논문은 과학기술부의 21세기 프론티어 프로그램 작물유전체기능연구사업단, 바이오그린 21사업단, 마린바이오 21사업단의 해양극한생물 분자유전체연구단, 과학재단 SRC의 경희대 식물대사연구센터, 한국생명공학연구원 기본사업의 분자표적핵심기반기술개발사업의 연구비 지원으로 이루어졌음.

인용문헌

Dunn W B, Bailey NJ C, Johnson HE (2005) Measuring the metabolome: current analytical technologies. Analyst 130: 606-625

Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18: 1157-1161

Hirai MY, Yano M, Goodenowe DB, Kanaya S, Kimura T, Awazuhara M, Arita M, Fujiwara T, Saito K (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. Proc Natl Acad Sci USA 101: 10205-10210

Kell DB, Darby RM, Draper J (2001) Genomic computing: explanatory analysis of plant expression profiling data using machine learning. Plant Physiology 126: 943-949

Kim SW, Ban SH, Chung H, Cho SH, Chung HJ, Choi PS,

Yoo OJ, Liu JR (2004) Taxonomic discrimination of higher plants by multivariate analysis of Fourier transform infrared spectroscopy data. Plant Cell Rep 23: 246-250

Kim SW, Ban SH, Jeong SC, Chung HJ, Ko S, Yoo OJ, Liu JR (2006) Genetic discrimination between *Catharanthus roseus* cultivars by metabolite fingerprinting using ^{1}H NMR spectra of aromatic compounds. Plant Cell Rep (accepted)

Raamsdonk LM, Teusink B, Broadhurst D, Zhang N, Hayes A, Walsh MC, Berden JA, Brindle KM, Kell DB, Rowland JJ, Westerhoff HV, van Dam K, Oliver SG (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nature Biotech 19: 45-50

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie A (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. Plant Cell 13: 11-29

(접수일자 2006년 11월 1일, 수리일자 2006년 11월 11일)