Available online at www.sciencedirect.com
SCIENCE DIRECT®
PERGAMON
Phytochemistry 62 (2003) 817–836
PHYTOCHEMISTRY
www.elsevier.com/locate/phytochem

Review

Plant metabolomics: large-scale phytochemistry in the functional genomics era

Lloyd W. Sumner[a,*], Pedro Mendes[b], Richard A. Dixon[a]

[a] *Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA*
[b] *Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University,1880 Pratt Drive, Blacksburg, VA 24061, USA*

Received 8 October 2002; received in revised form 7 November 2002

Abstract

Metabolomics or the large-scale phytochemical analysis of plants is reviewed in relation to functional genomics and systems biology. A historical account of the introduction and evolution of metabolite profiling into today's modern comprehensive metabolomics approach is provided. Many of the technologies used in metabolomics, including optical spectroscopy, nuclear magnetic resonance, and mass spectrometry are surveyed. The critical role of bioinformatics and various methods of data visualization are summarized and the future role of metabolomics in plant science assessed.

Keywords: Metabolomics; Metabolic profiling; Metabolite profiling; Analytical instrumentation; Mass spectrometry; MS; Bioinformatics; Metabolite databases; *Medicago truncatula*

Contents

* Corresponding author. Tel.: +1-580-224-6710; fax: +1-580-224-6692.
E-mail address: lwsumner@noble.org (L.W. Sumner).

0031-9422/03/$ - see front matter
doi:10.1016/S0031-9422(02)00708-2

1. Introduction

Recent advances in technology have brought about a revolution in the manner in which biological systems are visualized and queried. Advances in genetics and automated nucleotide sequencing have made possible the large scale physical mapping and sequencing of over twenty genomes including *Arabidposis thaliana* (The *Arabidopsis* Initiative, 2000), rice (Goff et al., 2002; Yu et al., 2002), and humans (Venter et al., 2001). Expressed sequence tag (EST) sequencing and mRNA profiling using either microarrays (Kehoe et al., 1999) or serial analysis of gene expression (SAGE) (Velculescu et al., 1995) now allow for the comprehensive analysis of the transcriptome. Advances in mass spectrometry have enabled the analysis of cellular proteins and metabolites (proteome and metabolome respectively) on a scale previously unimaginable. The cumulative utilization of these technologies has advanced the fields of functional genomics (Holtorf et al., 2002; Oliver et al., 2002; Somerville and Somerville, 1999) and systems biology (Ideker et al., 2001; Kitano, 2000). Both fields comprise traditional molecular biology, enzymology and biochemistry; however, the predominant difference from previous approaches is the significantly larger scale upon which they are conducted.

Functional genomics seeks to decipher unknown gene function. The functions of many genes revealed in large scale sequencing projects can be inferred through nucleotide similarity with gene sequences of known function determined through traditional empirical methods. However, there still remains a large number of predicted open reading frames (ORFs) that have no assigned function based on similarity (The *Arabidopsis* Initiative, 2000; The EU *Arabidopsis* Project, 1998; Somerville and Dangl, 2000) and of those that have been functionally annotated, only a small proportion of them have been demonstrated experimentally to have the function assigned (The *Arabidopsis* Initiative, 2000). Thus, empirical methods of functional determination are required. Functional elucidation of genes can be pursued through the systematic perturbation of gene expression followed by quantitative and qualitative analyses of gene expression products including mRNA, protein, and now metabolite levels (see Fig. 1). Genetic perturbations can be achieved by mutations caused by chemicals or ionizing radiation, or by integration of foreign DNA sequences leading to either over- or under-expression of genes in either targeted or random approaches (Weigel et al., 2000; Wesley et al., 2001). Transient alterations in gene expression can also be generated, for example, by the use of viral vectors (Baulcombe, 1999; Burton et al., 2000). Once expression has been altered, expression products are quantified through various profiling approaches and the resultant changes are assessed to infer gene function. Function may also be deciphered through analysis of co-responses. Stephen Oliver's group has coined the term FANCY or *f*unctional *an*alyses by *c*o-response in *y*east (Raamsdonk et al., 2001; Teusink et al., 1998). This method relies on the pair-wise comparisons of metabolite concentration changes obtained following perturbation of known genes (60% of the yeast genome) with those following perturbation of genes with unknown function. If an unknown gene yields a similar response, it is assigned a similar function. An advantage of the FANCY approach is that it assigns cellular rather than molecular "function", e.g. one assigns a gene to be involved in oxidative phosphorylation rather than naming it a kinase. Cellular function is more informative with regards to phenotype and from a systems biology perspective.

Systems biology is similar to functional genomics in its approach, but is slightly different in its objectives. Systems biology encompasses a holistic approach to the study of biology and the objective is to simultaneously monitor all biological processes operating as an integrated system. Through the study of systems, one can begin to visualize how individual pathways or metabolic networks are interconnected. This approach is based on solid theoretical frameworks and uses computer modeling to explain experimental observations. It is envisioned that holistic biology will be of great value for directing metabolic engineering strategies because modifications to the expression of single genes do not always bring about the predicted or desired effects due to cross-talk between pathways (Oliver, 2002).

At the analytical level, both functional genomics and systems biology rely on the comprehensive profiling of large numbers of gene expression products. These

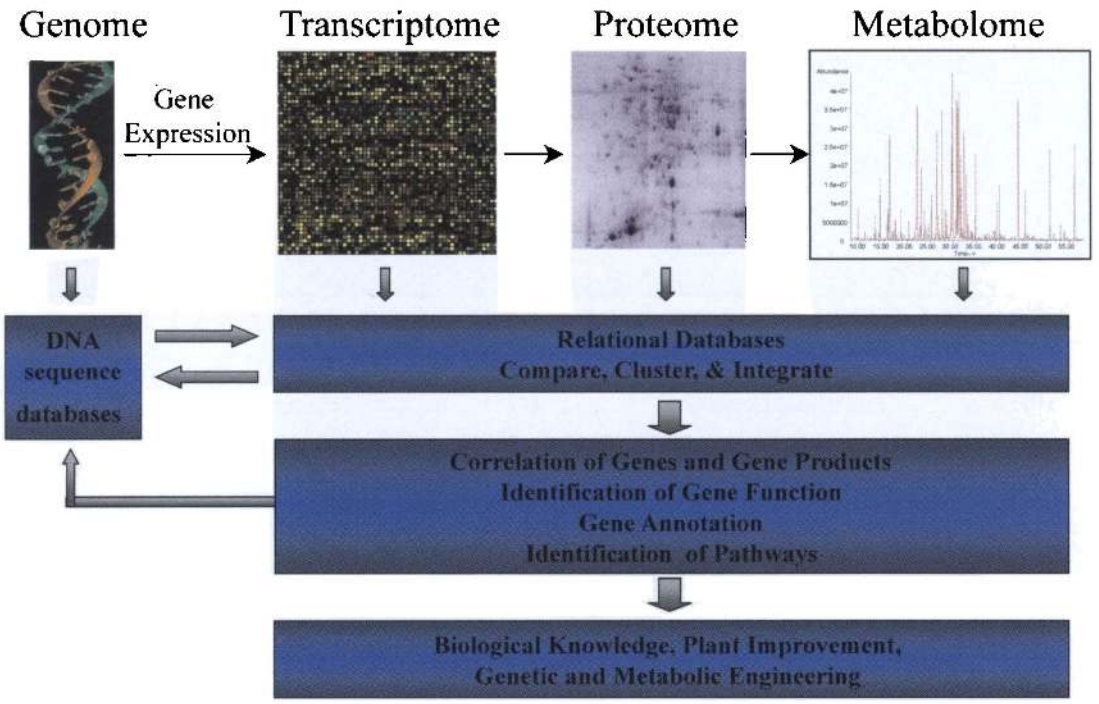


Fig. 1. Integrated functional genomics. The effects of gene perturbations are evaluated at multiple levels including the transcriptome, proteome, and metabolome. Changes in the metabolome occur as a consequence of those changes in the transcriptome that result in changes in the levels or catalytic activities of enzymes. Therefore, metabolome analysis is a valuable tool for inferring gene function.

approaches are commonly referred to as transcriptomics (Holtorf et al., 2002; Oliver et al., 1998), proteomics (Blackstock and Weir, 1999; Thiellement et al., 1999; van Wijk, 2001), and metabolomics (Fiehn et al., 2000; Oliver et al., 1998; Trethewey et al., 1999; Tretheway, 2001). The use of these "omics" technologies in biological research during the last 6 years is summarized in Fig. 2, based on the number of publications per year for each area. It is clear from Fig. 2 that the concept of profiling the metabolome has not been as eagerly engaged as its parallel "omics" counterparts even though the necessary technologies have a much longer history. This is primarily due to the technical complexity of metabolomics, as will be discussed later. Nevertheless, exponential growth is being observed in the use of metabolomics, and the impact of this approach is now being felt in many areas of biology.

There appear to be differences of opinion as to how best to define the comprehensive profiling of the metabolome. Some have chosen "Metabolomics" while others have chosen "Metabonomics". The term metabonomics is believed to have arisen from the root word "*genomics*" and has been largely utilized in the realm of toxicology. Alternatively, it has been suggested that the origins of the different terms are founded in the technology platform chosen for analyses, i.e. metabolomics for mass spectrometry based approaches and metabonomics for NMR based approaches. Regardless of the approach, we believe that the term metabolomics should be used because it is most consistent with the parallel terminology of transcriptomics and proteomics. Cases for the definition and differentiation of the terms target analysis, metabolite/metabolic profiling, metabolomics, and metabolic fingerprinting have been recently made (Fiehn, 2002) with the suggestion that metabonomics has been erroneously used to describe comprehensive analysis of the metabolome, and that a more correct terminology for metabonomics would be metabolic fingerprinting. We have also seen a similar misuse of the term metabolomics for less comprehensive methods such as biomarker analysis. We propose that any technology whose output is processed with pattern recognition software and without differentiation of individual metabolites should be termed metabolic fingerprinting and not metabolomics or metabonomics.

2. The metabolome

The components of the metabolome can be viewed as the end products of gene expression and define the biochemical phenotype of a cell or tissue. Quantitative and qualitative measurements of large numbers of cellular metabolites thus provide a broad view of the biochemical status of an organism that can be used to monitor and assess gene function (Fiehn et al., 2000). Profiling of the transcriptome and proteome has received some criticism due to its inability always to predict gene function.

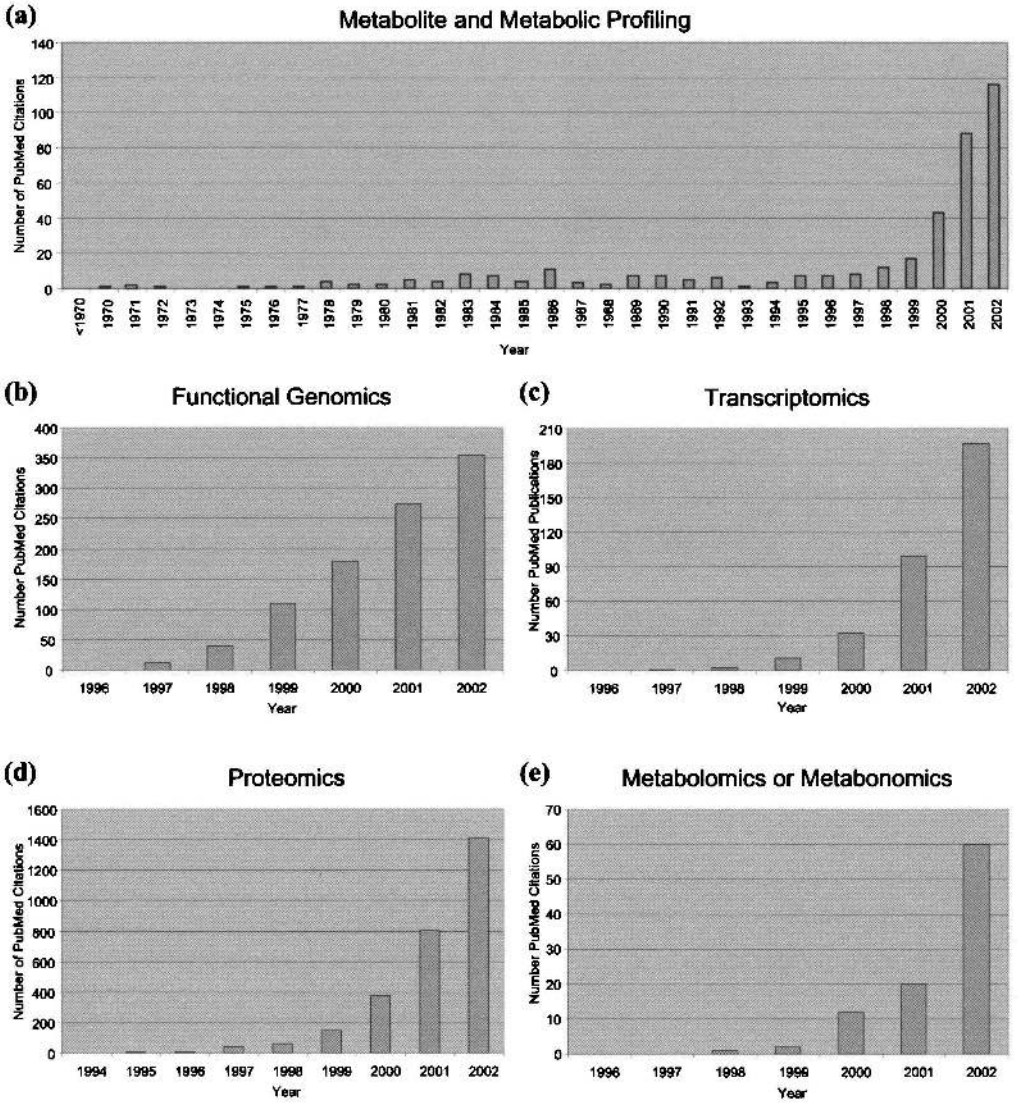


Fig. 2. PubMed literature search results document the continuously growing research areas of, (A) metabolite or metabolic profiling, (B) functional genomics, (C) transcriptomics, (D) proteomics and (E) metabolomics based on numbers of publications. Although metabolite profiling is much the oldest technology, the number of metabolomics publications is still relatively low compared to those describing other "omics" approaches.

Although the transcriptome represents the delivery mechanism of a translational code to the cellular machinery for protein synthesis, increases in mRNA levels do not always correlate with increases in protein levels (Gygi et al., 1999). Furthermore, once translated a protein may or may not be enzymatically active. Due to these factors, changes in the transcriptome or the proteome do not always correspond to alterations in biochemical (i.e. metabolic) phenotypes. Another consideration when profiling the transcriptome and proteome is that most modern techniques identify mRNA and protein through sequence similarity or database matching; thus, identification is based primarily on the quality of the match and is therefore indirect. In the absence of existing database information, transcript or protein profiling often yield only limited information. Based on the above limitations, profiling the metabolome may actually provide the most "functional" information of the "omics" technologies. There are, nevertheless, many instances in which transcriptome and proteome profiling have successfully pointed the observer to functional information, and therefore an integrated approach is preferred when resources permit.

The comprehensive quantitative and qualitative analysis of all metabolites within a cell, tissue or organism is a very ambitious goal and is still far from a reality for any system, although substantial progress is being made. Many responses involving altered gene expression, particularly those of plants to environmental stimuli, result in qualitative changes in metabolite pools, and therefore qualitative identification of the metabolites will be critical. At the same time, some genetic modifications or environmental responses may result only in temporal or spatial changes in metabolite con-

centrations. Thus, accurate and reproducible quantitative methods are also necessary to differentiate samples at a level that can provide an understanding of functional relationships between genomes and metabolomes.

Central to the metabolomic process are a variety of chemical profiling technologies. These technologies are used to compare different metabolic states resulting from differences in gene expression. Variations in metabolic states are manifested in "differential display" of metabolites or "discovery events" in the metabolome data set. To interpret the discovery event, the differentially expressed metabolites must be chemically identified. Single metabolites are identified first. Correlations are then sought between sets of differentially displayed metabolites. The individual as well as the correlated metabolites are then used to identify metabolic pathways or networks that have been affected. These pathways are then used to determine the broader biological significance of the response or to assign gene function. This can be done in a systematic way using the method of clique-metabolite matrices, developed in Fiehn's group (Kose et al., 2001).

"Omics" technologies are based on comprehensive biochemical and molecular characterizations of an organism, tissue or cell type. However, what constitutes a comprehensive analysis? A truly comprehensive analysis of the metabolome is currently not feasible and the number of primary and secondary metabolites in any given plant species is still uncertain. We suggest that in practice, a comprehensive analysis for plants should cover multiple metabolic pathways in both primary and secondary metabolism. We suggest at minimum inclusion of carbohydrates, amino acids, organic acids, lipids/fatty acids, vitamins and various classes of natural products such as phenylpropanoids, terpenoids, alkaloids, and glucosinolates, the latter classes depending upon the taxonomy of the species under consideration.

2.1. *The development of metabolomics*

Metabolomics originates from metabolite profiling. The earliest metabolite profiling publications originated from the Baylor College of Medicine in the early 1970s (Devaux et al., 1971; Horning and Horning, 1970, 1971a,b). These authors illustrated their concept through the multicomponent analyses of steroids, acids, and neutral and acidic urinary drug metabolites using GC/MS. They are also credited with coining the term "metabolite profiling" to refer to qualitative and quantitative analyses of complex mixtures of physiological origin. Soon afterwards, the concept of using metabolite profiles to screen, diagnose, and assess health began to spread (Cunnick et al., 1972; Mroczek, 1972). Thompson and Markey expanded on the quantitative aspects of using GC/MS for metabolite profiling in 1975 (Thompson and Markey, 1975) and by the late 1970s the methodology had attained enough interest to support a review article (Gates and Sweeley, 1978). Publications on the automation (Vrbanac et al., 1982) and expansion of GC based methods to other chemical classes soon followed (Niwa, 1986). During the early 1980s, results from the application of HPLC and NMR for metabolite profiling (Bales et al., 1984, 1988; Nicholson et al., 1984) began appearing in the literature. Interest in this area continued and resulted in a special edition of The Journal of Chromatography in 1986 that focused on metabolite profiling (Deyl et al., 1986; Fridland and Desiderio, 1986; Holland et al., 1986; Liebich, 1986; Niwa, 1986). Metabolic profiling research reached a steady state during the 1980s and into the early 1990s with approximately 10–15 publications per year. Many of these reports became more targeted and began to deviate from the original broad scope approach; however they still utilized the familiar metabolite profiling terminology. An example would be the pharmaceutical determination of the metabolic fate of drugs (Gerding et al., 1990; Woolf et al., 1992). In the early 1990s, Sauter and colleagues from BASF reported comprehensive GC/MS metabolic profiling as a diagnostic technique for determining the mode of action of various herbicides on barley plants (Sauter et al., 1991). This report established the principles and approaches that would be used by many soon to follow.

During the turn of the century, multiple genome and EST sequencing projects were underway or nearing completion and were fueling the "genomics era" (Goff et al., 2002; The *Arabidopsis* Initiative, 2000; Yu et al., 2002). It soon became clear that a large number of predicted genes, revealed by high throughput sequencing projects, could not be assigned a function based on sequence information alone, and proposals to assess gene function using large-scale analyses at the transcriptome level initiated the "functional genomics" era. It then became apparent that proteomics might yield a better or at least parallel means to monitor the results of gene expression. Oddly, the continuation of this thought process did not rapidly trickle down to consideration of the metabolome. It is believed that Oliver was the first to make this connection (Oliver, 1997) based on the perceived need for quantitative and qualitative measurement of phenotype to assess genetic function and redundancy in yeast. His group estimated the number of yeast metabolites to be approximately 600 and proposed the concept of metabolomics. This approach was then pioneered for plants by researchers at the Max Planck Institute (Trethewey, Willmitzer, Fiehn, Fernie at Golm, Germany) based on the analytical approach described by Sauter and coworkers (Sauter et al., 1991). Other plant groups were soon to follow, including The Samuel Roberts Noble Foundation that selected *Medicago truncatula* for in-depth analysis (Sumner et al., 2002), the Genomic Arabidopsis Resource Network

(GARNet), Iowa State University, and others listed in Table 1. Commercial entities such as Metanomics, Paradigm Genetics, and Phenomenome Discoveries were also quick to capitalize on the utility of metabolic profiling in plants. The number of academic and commercial groups using and entering this field is growing exponentially.

2.2. Limitations of metabolomics

The major limitation of metabolomics is its current inability to comprehensively profile all of the metabolome. This inability is directly related to the chemical complexity of the metabolome, the biological variance inherent in most living organisms, and the dynamic range limitations of most instrumental approaches. In many ways, this is similar to the situation of the Human Genome Project in 1990, when the technological means to sequence genomes were not yet available.

The genome and transcriptome consist of linear polymers of four nucleotides with highly similar chemical properties, facilitating high throughput analytical approaches. The proteome is substantially more complex, but is still based on a limited set of 22 primary amino acids. The chemistry of these biopolymers are nevertheless well defined and two-dimensional polyacrylamide gel electrophoresis (2-DE) can readily differentiate a large number of proteins in a single analysis, with several thousand being routine and 10,000 representing the upper boundary (Klose and Kobalz, 1995). When one surveys the metabolome, the chemical complexity is significantly greater. The chemical properties of metabolites range from ionic inorganic species to hydrophilic carbohydrates, hydrophobic lipids, and complex natural products. The chemical diversity and complexity of the metabolome make it extremely challenging to profile ALL of the metabolome simultaneously. Currently, no single analytical technique provides the ability to profile all of the metabolome. This obstacle is being circumvented through the use of selective extraction and parallel analyses using a combination of technologies to obtain the most comprehensive visualization of the metabolome (Sumner et al., 2002).

Analytical variance is defined as the coefficient of variance or relative standard deviation that is directly

Table 1
Plant metabolomics programs accessible via the internet

Academic/non profit	
2nd International Meeting on Plant Metabolomics	http://www.metabolomics-2003.mpg.de
Max Planck Institute	http://www.mpimp-golm.mpg.de/fiehn/index-e.html
The Noble Foundation	http://www.noble.org/plantbio/MS/index.htm
GARNet	http://www.york.ac.uk/res/garnet/beale.htm
Iowa State University	http://www.plantsciences.iastate.edu/ http://www.bb.iastate.edu/faculty/dimmas/index.html http://www.public.iastate.edu/~botany/wurtele.html
Platform for Plant Metabolomics	http://www.dpw.wau.nl/pf/PPM/indexppm.html
Plant Research International	http://www.plant.wageningen-ur.nl
John Innes Centre	http://www.jic.bbsrc.ac.uk/corporate/Facilities/metabolomics.html
Norwich Research Park	http://www.metabolomics-nrp.org.uk/nrp.html
Chiba University, Japan	http://www.p.chiba-u.ac.jp/lab/idenshi/index-e.html
Leibniz Institute of Plant Biochemistry, Germany	http://www.ipb-halle.de/english/institute/institute.htm http://www.ipb-halle.de/english/institute/research.htm
Wageningen University, The Netherlands	http://www.wau.nl/welcome.html
Leiden University	http://www.fwn.leidenuniv.nl/gs/bio_pharmaceutical_sciences/staff/Verpoorte.htm
Imperial College, London	http://www.med.ic.ac.uk/divisions/1/nicholson.asp
Michigan State University	http://www.bch.msu.edu/faculty/dellapenna.htm
Institute of Biological Sciences, University of Wales, Aberystwyth	http://www.aber.ac.uk/biology/ http://gepasi.dbs.aber.ac.uk/dbk/metabol.htm
Center for Novel Agricultural Products (CNAP), University of York	http://www.york.ac.uk/org/cnap/01_research/01c_labB/01c6_plant/01c6_plant.htm
Commercial	
Metanomics	http://www.metanomics.de/
Paradigm Genetics	http://www.paradigmgenetics.com/default.asp
Phenomenome Disoveries	http://www.phenomenome.com/
The Netherlands Organization for Applied Scientific Research (TNO)	http://www.voeding.tno.nl/biotechnology
Pioneer Hybrid	http://www.pioneer.com
Syngenta	http://www.syngenta.com/en/index_flash.asp
Unigen, Korea Crop Design, Belgium	http://www.cropdesign.com
Exelixis Plant Sciences	http://www.exelixis.com/discovery/plant_biotech
Large Scale Biology, USA	http://www.lsbc.com/index.php
Unilever	http://research.unilever.com
Numico	http://www.numico-research.com/splashpage.html

related to the experimental approach. This variance differs in accordance with the technology platform being used and is indeterminate in origin. Biological variance is also indeterminate in origin and arises from quantitative variations in metabolite levels between plants of the same species grown under identical or as near as possible identical conditions. Biological variations typically exceed analytical variations. Recently, Roessner and coworkers reported that the biological variability exceeded the analytical variability of GC/MS by a factor of ten (Roessner et al., 2000). Our data suggest that the average biological variance for *Medicago truncatula* is approximately 50% (unpublished data). These large biological variations represent the major limitations of the "resolution" of the metabolomics approach. One way to reduce biological variance is to pool samples, either by analyzing different tissues of the plant within a single sample, or by pooling multiple replicate plants. This helps minimize random variations through statistical averaging; however, many variations in metabolite levels often have biological significance and result from functional differentiation of tissues. Pooling tissue can, therefore, result in undesirable dilution of site or tissue specific up/down-regulated metabolites. An alternative, if relevant to the goals of the experiment, is to start with homogeneous tissue such as cell cultures, but this has obvious restrictions since the synthesis of some plant metabolites, particularly natural products, may be linked to cellular differentiation. Plant growth stage, environmental parameters, and sampling are critical. Therefore, strategies need to be incorporated to minimize variations.

A major technological challenge encountered in metabolomics is dynamic range. Dynamic range defines the concentration boundaries of an analytical determination over which the instrumental response as a function of analyte concentration is linear. The dynamic range of many techniques can be severely limited by the sample matrix or the presence of interfering and competing compounds. This is one of the most difficult issues to address in metabolomics. Most analytical mass spectrometric methods have dynamic ranges of 10^4–10^6 for individual components; however, this range is commonly and significantly reduced by the presence of other chemical components. In other words, the presence of some excessive metabolites can cause significant or severe chemical interferences that limit the range in which other metabolites may be successfully profiled. For example, high levels of primary metabolites such as sugars often interfere with the ability to profile secondary metabolites such as flavonoids. The positive aspect of this dilemma is that many of the highly expressed metabolites are often unique and can provide exclusive bases for the differentiation of cell states, organs, tissues, varieties and organisms. These exclusive compounds are often referred to as biomarkers. Selective profiling of these biomarkers is very useful in high throughput diagnosis of specific disorders such as diabetes (i.e. glucose monitoring) or cancer, but should not be classified as metabolomics due to the highly targeted nature of the profiling (Fiehn, 2002).

Interfering or competing analytes that may not necessarily be present in excess can nevertheless often lower performance and/or bias MS profiling techniques. For example, it is difficult to profile oligosaccharides by LC/MS in the presence of peptides or amino acids. The reason is that amino acids have greater proton affinities than oligosaccharides and, therefore, yield higher abundances of the charged species necessary for mass measurement. Another problem in electrospray ionization mass spectometry (ESI/MS) is salts. Low levels (i.e. submillimolar) of ionic species are known to reduce the ionization efficiency in ESI/MS and significantly interfere with profiling all species (Smith et al., 1991). Different analytical approaches have been developed to improve dynamic range and to minimize complications, and are discussed later.

2.3. Metabolome technologies

It is generally accepted that a single analytical technique will not provide sufficient visualization of the metabolome and, therefore, multiple technologies are needed for a comprehensive view (Hall et al., 2002; Sumner et al., 2002). Accordingly, many analytical technologies have been enlisted to profile the metabolome. Methods based on infrared spectroscopy (IR) (Oliver et al., 1998), nuclear magnetic resonance (NMR) (Bligny and Douce, 2001; Ratcliffe and Shachar-Hill, 2001; Roberts, 2000), thin layer chromatography (TLC) (Tweeddale et al., 1998), HPLC with ultraviolet and photodiode array detection (LC/UV/PDA) (Fraser et al., 2000), capillary electrophoresis coupled to ultraviolet absorbance detection (CE/UV) (Baggett et al., 2002), capillary electrophoresis coupled to laser induced fluorescence detection (CE/LIF) (Arlt et al., 2001), capillary electrophoresis coupled to mass spectrometry (CE/MS) (Soga et al., 2002), gas chromatography-mass spectrometry (GC/MS), liquid chromatography-mass spectrometry (LC/MS) (Huhman and Sumner, 2002), liquid chromatography tandem mass spectrometry (LC/MS/MS) (Huhman and Sumner, 2002), Fourier transform ion cyclotron mass spectrometry (FTMS) (Aharoni et al., 2002), HPLC coupled with both mass spectrometry and nuclear magnetic resonance detection (LC/NMR/MS) (Bailey et al., 2000a), and LC/NMR/MS/MS (Bailey et al., 2000b) have all been used.

The selection of the most suitable technology is generally a compromise between speed, selectivity and sensitivity. The sensitivities of various techniques are illustrated in Fig. 3. Tools such as NMR are rapid and selective, but have relatively low sensitivity. Other

methods such as capillary electrophoresis coupled to laser induced fluorescence (CE/LIF) detection are highly sensitive, but lack selectivity. Hyphenated mass spectrometry methods such as GC/MS and LC/MS offer good sensitivity and selectivity, but relatively longer analysis times.

2.3.1. Thin layer chromatography

Two-dimensional thin layer chromagtography (2D-TLC) has been used to follow changes in the 70 most abundant ^{14}C-glucose labeled compounds in *E. coli* under varying culture conditions (Tweeddale et al., 1998). This is a relatively simple and low resolution tool that will have difficulties accommodating complex mixtures.

2.3.2. Optical spectroscopic methods

Optical spectroscopic methods using IR and UV are rapid and provide metabolic fingerprints that can be processed with pattern recognition to determine similarities or differences. Oliver and coworkers used Fourier transform infrared spectroscopy (FTIR) to differentiate the metabolic complement of yeast respiratory mutants from that of wildtype FY23 yeast (Oliver et al., 1998). This tool offers rapid assessment of similarities and differences; however, other more selective tools would be necessary in plants to identify specific metabolites responsible for the similarities or differences.

2.3.3. Nuclear magnetic resonance

NMR methods provide metabolic fingerprints with good chemical specificity for compounds containing elements with non-zero magnetic moments such as ^{1}H, ^{13}C, ^{15}N, and ^{32}P that are commonly found in most biological metabolites (Bligny and Douce, 2001). Increased specificity is further realized with the use of high magnetic fields that provide greater resolution and separation of signature chemical shifts. These non-destructive methods can be highly automated to achieve very high sample throughput. Most NMR based programs appear to be focused on biomarker analyses or the pursuit of specific chemical signatures related to a specific metabolic process that is indicative of disease or mode of action, and not on the comprehensive analyses of a large number of metabolic pathways. The Nicholson group at Imperial College, London has provided many examples of the use of NMR in metabolomic approaches (Bailey et al., 2000a,b; Bales et al., 1984, 1988; Bundy et al., 2002; Nicholson et al., 1984, 1999, 2002). Recently, NMR has been interfaced with HPLC and simultaneous mass spectrometry to yield very informative multidimensional data (Bailey et al., 2000a, 2000b). This multidimensional approach appears very promising. The disadvantages of HPLC/NMR would be lower duty cycles of the NMR and elevated expenses due to the need for deuterated mobile phases.

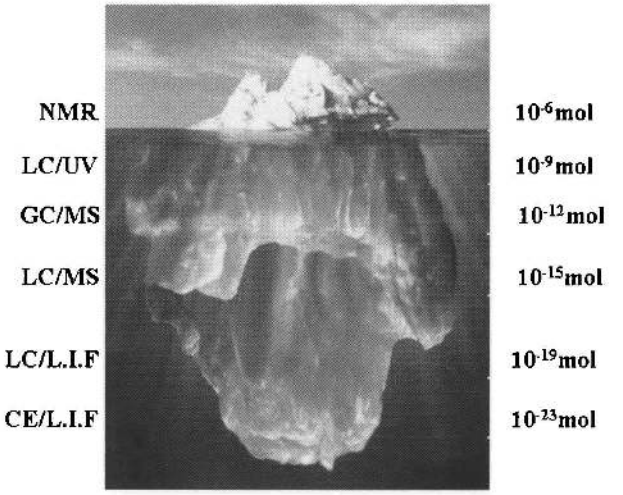


Fig. 3. A comparison of the relative sensitivities of various metabolomic tools. NMR has rapid analysis times but suffers from lower sensitivity thus allowing visualization only of the more concentrated metabolites (i.e. the tip of the iceberg). GC/MS and HPLC/MS provide good selectivity and sensitivity. CE/LIF (laser induced fluorescence) provides very high sensitivity but lower selectivity.

2.3.4. Mass spectrometry

Flow-injection mass spectral analyses have been used for metabolic fingerprinting. For example, flow-injection ESI/MS metabolic fingerprints of cell free extracts have been used for bacterial identification (Vaidyanathan et al., 2001). Similarly, multiple ionization techniques coupled to Fourier transform mass spectrometry (FTMS) were used to identify metabolites specifically associated with the development and ripening of strawberry fruit (Aharoni et al., 2002). FTMS has the added capability of high resolution and high mass accuracy. High resolution allows for the separation and differentiation of very complex mixtures and high mass accuracies allow for the calculation of elemental compositions to aid in structural differentiation and characterization. Unfortunately, this approach cannot differentiate chemical isomers such as those of common hexoses since they have the same exact mass. Differentiation of isomers is commonly achieved by separation technologies such as GC and HPLC, but this adds additional analysis time. The duty cycle of a FTMS operated in high resolution and broad band (wide m/z window) is limited and typically does not provide statistical sampling across most chromatographic peak widths. In addition, the current high cost of FTMS instrumentation is prohibitory for widespread utilization.

GC/MS is a relatively low cost alternative that provides high separation efficiencies that can resolve complex biological mixtures; however, it requires that samples be volatile. This requirement is readily accomplished by chemical derivatization, but at the cost of additional time, processing, and variance. Typically, GC/MS is performed with affordable single quadrupole

mass analyzers for the separation and analysis of complex mixtures. The utilization of an automated mass spectral deconvolution and identification system (AMDIS) enhances the ability to deconvolute and successfully identify overlapping chromatographic peaks (Halket et al., 1999). Newer GC/TOFMS systems incorporating time-of-flight mass analyzers offer an attractive alternative to quadrupoles and provide greater m/z accuracies. These instruments also provide detectors with high scan speeds supporting ultrafast GC/MS (Davis et al., 1999) and the potential to profile increasingly complex mixtures.

HPLC coupled to on-line photodiode array detection (HPLC/UV/PDA) is a good choice for compounds containing chromophores. Recently HPLC/UV/PDA has been utilized in the metabolic profiling of plant isoprenoids (Fraser et al., 2000). HPLC coupled to MS is a powerful alternative that offers high selectivity and good sensitivity. Utilization of a liquid introduction system allows for the analysis of nonvolatile and labile species without the need for derivatization. HPLC can also be simultaneously coupled to both UV/PDA and MS to provide multiple levels of information useful in chemical structure elucidation. Recently, reversed-phase LC/MS has been used for the metabolic profiling of saponins in legumes (Huhman and Sumner, 2002). Hydrophilic interaction liquid chromatography coupled to mass spectrometry (HILIC/MS) has also been used to analyze highly polar plant extracts in *Cucurbita maxima* (Tolstikov and Fiehn, 2002).

2.3.5. *Phenotype microarrays*

Although not strictly a metabolic profiling method, an interesting alternative method of assessing metabolism entitled "Phenotype Microarrays", has recently been demonstrated in *Escherichia coli* (Bochner et al., 2001). This array-based (96 well-plate) colorimetric assay quantifies color changes induced in a tetrazolium dye based on cellular respiration. Metabolism of various nutrient sources results in NADH production and electron flow at membranes and mitochrondria. This electron flow results in quantifiable color change in the assay. The array is composed of up to 700 various nutrient sources including sugars, amino acids, or chemical compounds such as kanamycin and penicillin. The authors demonstrate this technology with various *E. coli* mutants such as the *xylA* mutant that has lost its ability to metabolize maltose and maltotriose. This approach could be used for biochemical phenotype assessment in plant cells in relation to traits such as herbicide or salt tolerance.

3. Bioinformatics

It is obvious that all "omic" approaches will rely heavily upon bioinformatics for the storage, retrieval, and analysis of large datasets; and metabolomics is no exception. Unfortunately, metabolomics is still in an infant state and many of the necessary tools are not available. We describe below the basic tools currently being used and those on the horizon. These tools serve to align, visualize, and differentiate, components in large datasets. Individual components then need to be correlated and placed in metabolic networks or pathways. This information, together with quantitative kinetic indices, can be used to model and simulate pathways that ultimately lead to a better understanding of biological and biochemical phenomena.

Changes in metabolite levels may be dramatic or subtle. The dramatic changes will be easily recognized; however, subtle changes will require statistical processing to determine whether or not the observed changes are significant (Miller and Miller, 2000; Koosis, 1997). Statistical approaches require careful experimental design including replicate sampling, replicate analyses and application of statistical tests. General statistical tests such as Student's t-test should first be performed to eliminate erroneous data. Means and standard deviations should then be calculated. F-ratios can then be used to determine whether or not a change is significant at a given confidence level. Most often used methods assume that the data follow a normal distribution; however, it may be more prudent to use non-parametric methods, since the normality assumption may not be correct. It is interesting to note that enzyme kinetic data appear to be leptokurtic (Cornish-Bowden and Eisenthal, 1974), i.e. with a higher frequency of "outliers" than the Gaussian distribution. This should guide our expectations for metabolomic data since the transformations of metabolites are indeed dependent on enzyme function.

Statistical analyses must be performed to ensure good analytical rigor, but unfortunately are often a burden when working with large datasets. Tools are therefore needed for high throughput statistical analyses of all components in a dataset to provide sound evidence for the relevance of changes in a metabolite level contained in the raw data. Computer based applications are required that can differentiate whether or not samples are statistically similar or different and what the exact differences/similarities are. Ideally, this would be performed in a fully automated manner. For example, a system should be able to compare the UV, NMR, GC/MS, LC/MS, or CE/MS profiles of a sample set automatically and direct an investigator to the component(s) that are statistically different. The chemical identity of these components could then suggest gene function or the biological response of the system.

A single GC/MS metabolite profile can yield 300–500 distinct components. This provides a wealth of information to be interpreted and leads to significant challenges in processing the data. To simplify the task,

many researchers have used techniques to reduce the dimensionality of the data set and to visualize the data. The most popular approaches include unsupervised methods such as principal component analysis (PCA), hierarchical clustering (HCA) and K-means clustering; however, the utility of machine-learning methods such as self-organizing maps (SOM, also known as Kohonen neural netoworks) (Kohonen, 1995) are promising.

3.1. Principal component analysis (PCA)

Principal component analysis is one of the oldest and most widely used multivariate techniques (Hotellin, 1933). The concept behind PCA is to describe the variance in a set of multivariate data in terms of a set of underlying orthogonal variables (principal components). The original variables (metabolite concentrations) can be expressed as a particular linear combination of the principal components. PCA is a linear additive model, in the sense that each principal component (PC) accounts for a portion of the total variance of the data set. Often, a small set of principal components (2 or 3) account for over 90% of the total variance, and in such circumstances, one can resynthesize the data from those few PCs and thus reduce the dimension of the data set. Plotting the data in the space defined by the two or three largest PCs provides a rapid means of visualizing similarities or differences in the data set, perhaps allowing for improved discrimination of samples.

3.2. Hierarchical cluster analysis (HCA) and K-means clustering

Hierarchical cluster analysis (HCA) is a method of grouping samples in a data set by their similarity. HCA involves a progressive pair-wise grouping of samples by distance. Several distance measures can be used in HCA, such as Euclidean distance, Manhattan distance, or correlation. Results vary according to which distance metric is used. The result of hierarchical clustering is usually visualized as a dendogram or a tree. Branch lengths can be made proportional to the distances between groups. This can provide an easy visualization of the similarities of samples within data sets.

K-means clustering is another method of grouping data, which uses a fixed number (K) of groups. The principal is similar to HCA, in which one must define a distance metric which will govern the clustering, but the way in which the grouping is made is different. Several other clustering algorithms exist, which are variations on the same theme, and it is unclear which ones are best for a specific problem, given that they usually produce different results. Indeed, even using different metrics with the same clustering algorithm (e.g. HCA with Euclidean distances versus HCA with correlations) usually produces different results. Clustering is most useful to classify samples in groups. It is often applied to the data after transformation with PCA, in which case it becomes a means of identifying groups in the reduced dimension data space.

3.3. Self-organizing maps (SOMs)

Self-organizing maps (SOMs) (Kohonen, 1982, 1995) are artificial intelligence methods that are designed to group data. They are similar to K-means clustering in the sense that one predefines the number of groups that data will be classified within (in this case that number must be a power of two). SOMs are gaining popularity due to their enhanced ability to differentiate and visualize data relative to PCA (Kohonen, 1995; Törönen et al., 1999). Recently, SOMs have been applied to the correlation of GC/MS data to compare the morphology of 88 species of ants (Nikiforow et al., 2001).

All of the above methods can be classified as *unsupervised* because they require no other information than the original data set. A different set of methods, called *supervised*, create a calibration using a "training" data set, i.e. a set of observations that have been classified by independent means. An example of a supervised method is the use of standards to calibrate a protein concentration assay. Supervised methods can thus only be carried out if one is able to provide known examples. Despite this drawback, supervised methods are usually more powerful than unsupervised methods. Raamsdonk and coworkers (Raamsdonk et al., 2001) have used discriminant function analysis (Lachenbruch, 1975) for this purpose. Other supervised methods that could be used are feed-forward neural networks (Cowan and Sharp, 1988), support vector machines (Cristianini and Shawe-Taylor, 2000), genetic algorithms (Goldberg, 1989), and genetic programming (Koza, 1992). Kell, Goodacre and coworkers have pioneered the application of supervised methods to metabolomic data (Goodacre and Kell, 1996; McGovern et al., 2002; Oliver et al., 1998; Shaw et al., 2000). A particular supervised method of analysis is the proposed FANCY approach (Raamsdonk et al., 2001; Teusink et al., 1998). FANCY is based on co-response analysis (Hofmeyr et al., 1993; Hofmeyr and Cornish-Bowden, 1996), a branch of metabolic control analysis. Co-response is a measurement of how two metabolite concentrations (or fluxes) respond to a common perturbation. It can be estimated by the ratio of change in one metabolite concentration to the change in the other metabolite concentration, both following the same perturbation. The FANCY approach has been applied to the case of yeast null mutants, where one classifies mutants of genes of known function according to their patterns of metabolite co-responses (measured with metabolomic methods). Then the mutants of genes of unknown function are classified by comparison of

their co-response pattern to the set of known genes. This approach assumes that genes with functions in related metabolic sections produce similar patterns of co-response, i.e. similar changes in metabolite concentrations.

It is also very important to develop means of visualizing large amounts of multivariate or metabolomic data (Tabachnick and Fidell, 1983). Data visualization is the activity of displaying data sets in such a way as to allow direct visual identification of properties of the data sets. Several methods exist for visualization of generic multivariate data sets (Meyer and Cook, 2000), including some that use the analyses described above, which are readily applicable to metabolomics. There are, however, specific properties of metabolomic data that can be capitalized upon to produce visualizations specific for this data type. We know that metabolites are related by their molecular structure and by the fact that they are put together by alteration/combination of other metabolites. Diagrams representing the networks of reactions linking metabolites have been used for a long time as a powerful tool for visualizing relationships between metabolites; see for example (Michal, 1999; Umbreit, 1952). Such metabolic network diagrams are also useful to visualize metabolomic data (Mendes, 2002) and indeed to combine it with gene expression (Wolf et al., 2000) or proteomics data. This form of visualization is done in the context of the underlying biochemistry and has different objectives from other forms of multivariate data visualization, such as the Grand Tour (Asimov, 1985) or exploratory projection pursuit (Swayne et al., 1998). While the latter techniques display the data in a space of variables (each axis is a metabolite) attempting to find the most "informative" projections onto two dimensions, metabolic networks attempt to directly identify the chain of causality that has led to the observations. If one metabolite concentration has increased, it can be because the reactions producing it increased flux, or that the reactions consuming it decreased flux (or a combination of these effects). Indeed, such visualizations may be instrumental in identifying the functions of mutated genes directly, by visual inspection. However, for this to be possible, it is important that the diagrams be complete, i.e. that they display all the reactions that surround the metabolites in question. The traditional metabolic pathway diagrams, such as the ones in the popular database KEGG (Kanehisa et al., 2002; see this and other resources listed in Table 2), are not useful for such analysis, since each diagram omits several "side" reactions. To visualize side reactions, diagrams are needed that display the neighborhood of a metabolite of interest, consisting of all the reactions in which the metabolite enters either as substrate or product. One of us (Mendes) is indeed producing such diagrams and software to visualize metabolomic data. Fig. 4 depicts metabolomic data from *Medicago truncatula* development, focusing on the neighborhood of fumarate.

The above bioinformatic tools provide methods of determining differences or similarities in datasets. The next step is to incorporate metabolomic data with other expression information including mRNA and proteins to infer gene function. To accomplish this, metabolomic data sets must be integrated and correlated in a global manner with genetic and enzymatic data, pathways assembled into systems, and literature references incorporated as learning tools to annotate existing data to yield in silico biological information (Palsson, 2000). Approaches and tools are now available for modeling metabolic systems and are very vital in understanding metabolism (Mendes and Kell, 1998). The challenge of the future is to integrate these approaches and obtain complete integrated functional genomic systems to better understand and visualize systems biology (Mendes, 2001; Voit and Radivoyevithc, 2000).

3.4. Databases

Perhaps the biggest challenge of metabolomics for bioinformatics comes from the current lack of appropriate databases and data exchange formats. The situation is in many ways similar to gene expression a few years ago, but is complicated by the very weak role of sequence analysis here. There is a need for biochemical ontologies that clearly specify what each entity (metabolite) is and how it relates to others and for databases that can store metabolomic data in a way to facilitate relevant queries. It is extremely important that such databases can interoperate with gene expression databases.

3.4.1. Reference biochemical databases

As mentioned above, it is quite useful to use a biochemical context for visualization and interpretation of metabolomic data. Such context is formed by the network of reactions that exist in the organism of interest. To be useful, the biological context should also include the known enzyme activities that catalyze each reaction, the proteins that carry such activities and the genes that code for them. In brief, we need databases that describe the known biochemistry. Such databases already exist (see Table 2), although as we will argue, none is really appropriate for wide use in metabolomics. Perhaps the most popular one, and certainly the most easily accessible, is KEGG (Kanehisa et al., 2002). KEGG also includes the LIGAND (Goto et al., 2002) database of molecules and reactions. KEGG is based on two major resources: information about genes from GenBank/EMBL DDJB, and information about reactions and enzymes from the Enzyme Nomenclature of the IUBMB, accessible from the ENZYME database (Bairoch, 2000). It is important to realize that KEGG

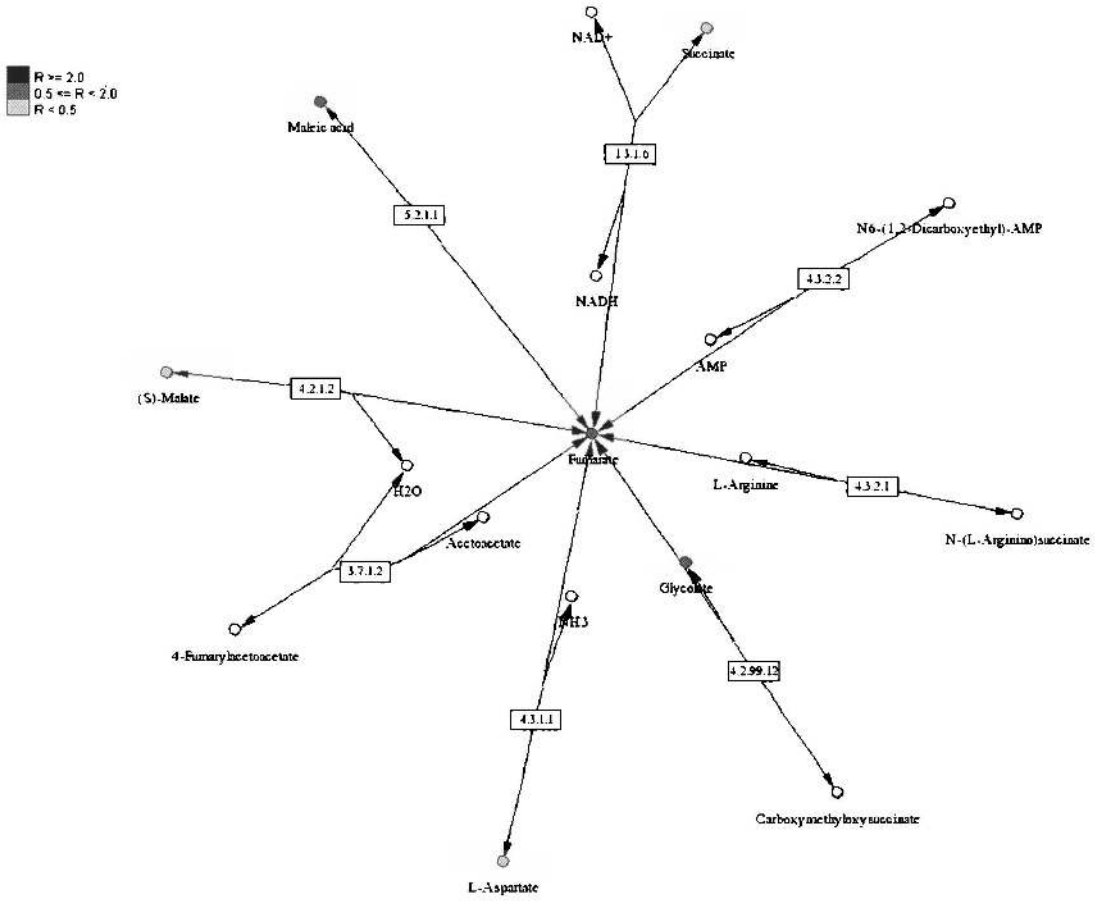


Fig. 4. Ratio of metabolite levels between two samples of *Medicago truncatula* displayed on the metabolite neighborhood of fumarate. Ratios of metabolite levels are indicated by the color that fills the circles representing individual metabolites (in this example a grayscale is used rather than colors).

lists enzymes in a specific organism only if a sequence exists in GenBank that is annotated as coding for such an enzyme. Unfortunately, many enzymes are known to exist in organisms but their genes have not yet been identified, even in well-characterized genomes such as *E. coli*, and certainly not in those less well characterized such as *A. thaliana* and other plants. On the other hand, the overwhelming majority of genes in GenBank are annotated by sequence similarity and have no experimental backing. This means that KEGG has a considerably incomplete set of data when compared with the knowledge obtainable from the literature, and it also includes information whose value is dubious. Another problem is that the enzyme nomenclature does not cover well all known enzyme activities and is ambiguous in its coverage of isozymes (which become very important to understand actual metabolite levels, though perhaps not for classification of enzymes).

Another database that can be useful for the purpose of reference is EcoCyc (Karp et al., 2002), which collects genetic and biochemical information about the *E. coli* bacterium recovered from the literature by human curation. The data set contained in EcoCyc is thus more complete than KEGG for *E. coli* because it does not rely exclusively on genome annotation. Although EcoCyc is being extended to other organisms (including *Arabidopsis* by the TAIR project, http://www.tair.org) it is not clear how much information is really being recovered from the literature (i.e. experiments), as opposed to being recovered from genome annotation (i.e. BLAST results). The danger with this effort is that the biochemistry of other organisms is being described by analogy with that of *E. coli*. This may be appropriate for Enterobacteria, but certainly not for plants. Other databases that could be used as references are EMP (Selkov et al., 1996), PathDB (Mendes et al., 2000), UM-BBD (Ellis et al., 2001), and BRENDA (Schomburg et al., 2002). All these databases contain features that make them unique, but none of them alone fulfills all the requirements for a good reference for metabolomics (Mendes, 2002; Wittig and De Beuckelaer, 2001). It is most important that such a reference lists specific molecules (not classes thereof), that it distinguishes between different isomers and isozymes, that it lists reactions independently of the enzymes that catalyze them, that it captures the complexity of relations between genes, proteins and meta-

bolites (which are many-to-many relationships), and finally, that it identifies the evidence that was used to infer the existence of each particular molecule. Because no such database exists for *M. truncatula*, we are constructing such a reference for this organism.

A particularly important role of reference biochemical databases is to list metabolomes. A metabolome is the complete set of all metabolites that exist in a biological species. No complete metabolome is yet known; however, it is conceivable that metabolomics may possibly achieve this objective. Reference databases would then be suitable to store metabolomes, and should strive to list the set of all known metabolites of model organisms.

3.4.2. *Metabolite profile databases*

Metabolite profiles are composed of measurements of metabolites at a specific state of a biological system. Experiments may consist of various metabolite profiles whether they be time courses, comparisons of different mutants under similar conditions, or comparisons of the same biological system subjected to different environmental challenges. These data are equivalent to microarray results or protein profiles. Like those data types they should be stored in appropriate databases that will become the means of publishing such data (because they are too extensive to be printed in journal articles). No such database of metabolite or protein profiles exist at this time. Microarray gene expression data can already be archived in databases such as ArrayExpress (Brazma et al., 2002) and GEO (Edgar et al., 2002). Recently the journal Nature has established a policy that requires data to be deposited in either of these two databases prior to publishing in their journal. At some point in time we envision the same will happen with metabolomic as well as proteomic data. It will then be important to relate metabolomic, proteomic and transcriptomic data, since these are just different aspects of the same: the behavior of the biological system in question. Systems biology requires that these data be combined. Currently it is most urgent that metabolic profile lab databases be able to export data in common data formats. It is important that the community, small as it is at this time, agrees on such data formats. This would avoid much confusion and wasted time in data conversion between labs. It would also greatly facilitate the ability to repeat each others' results.

3.5. *Modeling and simulations*

There is value in using metabolomic data to construct computer models of metabolism. Models are useful to summarize large amounts of disparate data and to prove that they are consistent. Often, the exercise of modeling reveals voids in knowledge that must be filled for a full understanding of observed phenomena. If this were the minimum objective, it would be immensely useful. It is also possible to achieve a level in which a model does indeed show that all the known facts put together are consistent, and then the model can be used for prediction and generation of hypotheses. A recent example is the model of yeast glycolysis by Teusink and coworkers (Teusink et al., 2000) and its improvement by Pritchard and Kell (2002). Traditionally, such models have been constructed from data obtained *in vitro* with purified enzymes. This has several problems, not least of which is that purified enzymes often do not function as they do in intact cells (Srere and Ovádi, 1990). It is thus an objective of computational biochemistry and systems biology (Mendes, 2001) to be able to construct models of metabolism directly from metabolomic data, i.e. observations of the working system. The methods to carry out such an ambitious endeavor are not all in place yet, though it is clear that global optimization methods are going to be an essential piece (Mendes, 2001; Mendes and Kell, 1998). Such modeling is likely to result in proposition of novel metabolic networks. This approach will be especially important in plant secondary metabolism, because the way in which the majority of natural products are synthesized is not yet fully characterized.

4. Applications of metabolomics to plant systems

Metabolomics offers the unbiased ability to differentiate genotypes based on metabolite levels that may or may not produce visible phenotypes (Raamsdonk et al., 2001; Roessner et al., 2001). Furthermore, in those instances in which mutations or expression of transgenes lead to measurable phenotypic changes (Boyes et al., 2001), metabolomic approaches can be used to decipher the biochemical cause or consequence of the observed phenotypes.

Metabolomics is at its most powerful when performed on a large scale and integrated with corresponding data on the transcriptome and proteome. However, few examples exist of this approach in plants, although work in this area is in progress (May, 2002). More selective metabolic profiling has, however, been used in a number of areas to provide biological information beyond the simple identification of plant constituents. These areas include:

- Fingerprinting of species, genotypes or ecotypes for taxonomic or biochemical (gene discovery) purposes (Gorinstein et al., 1995; Huhman and Sumner, 2002; Stashenko et al., 2000).
- Monitoring the behavior of specific classes of metabolites in relation to applied exogenous chemical and/or physical stimuli (Bednarek et al., 2001; Chong et al., 2001; Lois, 1994; von Röpenack et al., 1998).

- Studying developmental processes such as establishment of symbiotic associations (Harrison and Dixon, 1993; Maier et al., 1999) or fruit ripening (Aharoni et al., 2002).
- Comparing and contrasting the metabolite content of mutant or transgenic plants with that of their wild-type counterparts (see later).

In each of these cases, application of metabolite profiling can be coupled with other "omics" technologies to provide an integrated picture encompassing all aspects of information flow from genome to metabolome and resulting phenotype.

4.1. Metabolic profiling of transgenic plants

To date, there are few reports of exhaustive, unbiased metabolic profiling of transgenic plants. Roessner et al. used GC/MS-based metabolic profiling to study transgenic potato plants over-expressing invertase specifically in the tubers (Roessner et al., 2000). The data rapidly confirmed that the reduction in starch accumulation resulted from partitioning of carbon flux into glycolysis. In subsequent studies, statistical data mining tools were used to determine the major biochemical phenotypes of transgenic potato lines overexpressing invertase, glucokinase or sucrose phosphorylase (Roessner et al., 2001). This was one of the first reports to reveal the power of metabolite profiling coupled to statistical data analysis for plant phenotyping.

Technologies have been developed over the past several years for the detailed profiling of various classes of metabolites peculiar to, or characteristic of, plants. These include the use of HPLC methods with diode array or mass detection for profiling flavonoids and their conjugates (Graham, 1991; Lin et al., 2000; Sumner et al., 1996) or carotenoids and related compounds such as tocopherols and plastoquinones (Fraser et al., 2000). Ion-paired reversed phase HPLC of fluorescent *etheno*-substituted acyl CoA esters is, for the first time, allowing detailed profiling of acyl CoAs in plant extracts (Larson and Graham, 2001). This technology is being applied to plant lines modified in their fatty acid composition. GC methods are still popular, and are being optimized for obtaining maximum levels of information from extracts from different plants and tissue types (Katona et al., 1999). The desorption-concentration-induction (DCI) technique coupled to GC has even made it possible to profile monoterpenes from single peltate trichomes from the leaves of mint plants (Voirin and Bayet, 1996).

Two examples of metabolic engineering in the phenylpropanoid pathway further highlight the value of the metabolic profiling approach for discovery and or confirmation of metabolic mechanisms. There is considerable interest in engineering plants to contain isoflavone phytoestrogens in view of their perceived chemopreventive activities against hormone-dependent cancers, cardiovascular disease, and post-menopausal ailments. Most non-legumes do not contain isoflavonoids, but the pathway can be introduced by genetic transformation with an isoflavone synthase (IFS) gene from a legume such as soybean. In such plants, the ubiquitous flavanone intermediate naringenin is converted, through the action of IFS, to glycoconjugates of the isoflavone genistein (Jung et al., 2000; Liu et al., 2002). However, attempts to date have only resulted in very low levels of genistein accumulation in non-legumes (Yu et al., 2000). In an attempt to increase isoflavone production, a cDNA encoding alfalfa chalcone isomerase, the enzyme step prior to naringenin formation, was over-expressed in transgenic *Arabidopsis* already expressing soybean IFS. Profiling of flavonoid compounds by LC/MS indicated that over-expression of CHI alone led to an approximately 3-fold increase in the levels of the major glycoconjugates of the flavonols quercetin and kaempferol, and additional types of conjugates also appeared (Liu et al., 2002). A similar profiling approach had previously been used to show increased flavonol production in the peel of tomato fruit over-expressing CHI (Muir et al., 2001). Expressing IFS in the CHI over-expressing background did not lead to elevated genistein levels in Arabidopsis as had been hoped. However, IFS expression strongly reduced the increased flavonol levels resulting from CHI expression in a manner disproportionate to the small amounts of genistein produced (Liu et al., 2002). Such cross-talk between endogenous and introduced pathways may be a feature of metabolically engineered transgenic plants, and more in-depth profiling in the present case may have revealed further unexpected consequences of transgene expression.

The exact nature of the metabolic pathways leading to lignin formation in vivo are still a matter of much debate (Humphreys and Chapple, 2002), in spite of the demonstration of all necessary enzymatic activities in vitro. LC/MS profiling revealed an unexpected accumulation of caffeic acid glucoside in the soluble phenolic fraction in both alfalfa (Guo et al., 2000) and poplar (Meyermans et al., 2000) following down-regulation of caffeoyl CoA 3-*O*-methyltransferase (CCOMT), whereas this compound was not detected following down-regulation of caffeic acid *O*-methyltransferase. As caffeic acid glucoside most probably arises from glucosylation of free caffeic acid (released from caffeoyl CoA by thioesterase activity), these data are consistent with recently proposed models in which caffeoyl CoA is the in vivo substrate for CCOMT, but caffeic acid is not an in vivo substrate for COMT (Humphreys and Chapple, 2002). Several enzymes of plant natural product biosynthesis have relatively promiscuous activities in vitro, and metabolic profiling of transgenic plants in which

expression of these enzymes is modified is a powerful new approach for ascribing in vivo function.

HPLC profiling of isoprenoid compounds in tomato expressing a bacterial phytoene desaturase confirmed increased accumulation of β-carotene, lutein and cyclic carotenoids (Fraser et al., 2000). Application of the same profiling method to *Arabidopsis* plants treated with different types of bleaching herbicides demonstrated different profiles diagnostic of the site of inhibition, indicating the value of metabolic profiling for determining sites of action of new agrochemicals (Fraser et al., 2000).

In-depth metabolic profiling is set to become a critical technology in the scientific and political battle for acceptance of genetically modified organisms. The gradual realization by the non-scientific community that plants contain both beneficial and harmful chemicals whose levels can be altered by genetic manipulation has fueled public awareness of issues relating to transgenic foodstuffs, and has been accompanied by extensive efforts to define and thereby regulate pharmaceuticals, nutraceuticals, functional foods, food additives and food supplements (Kleter et al., 2001).

In spite of some remarkable arguments about the dangers of genetically modified foods in the popular press (Longman, 1999), the transgene itself is "chemically neutral", being composed of the same four nucleosoide triphosphates as any other gene. Other than cases where the transgene protein itself might be either toxic or cause allergic reactions, the concept of "substantial equivalence" between transgenic and the corresponding non-transgenic plants will be based on the biochemical phenotype that can be determined by metabolic profiling. Although it is likely that greater statistical differences exist in metabolite or transcript profiles between different cultivars or ecotypes than between a transgenic plant and its corresponding wild-type control, it is nevertheless possible that the goal posts for substantial equivalence may keep shifting to more and more rigorous positions as the technology for metabolite profiling becomes increasingly sophisticated.

4.2. *Spatially resolved metabolomics*

Most current metabolomic efforts utilize pooled tissue samples; however, our understanding and knowledge of metabolism will greatly improve with improvements in the ability to spatially resolve the metabolome. Differentiation of the metabolomes of individual tissues, single cells, and subcellular organelles will remove dilution effect and allow visualization of detailed metabolic differences of various cell types and subcellular organelles. The concept of high resolution spatial analysis of plants has recently been reviewed (Kehr, 2001) whereas Korolev and coworkers recently reported on the spatial and temporal distribution of a small number of sugars and ions in carrot taproot (Korolev et al., 2000). Farré and coworkers recently used a nonaqueous extraction method (Stitt et al., 1989) and GC/MS to study the compartmentalization of glycoltyic intermediates, nucleotides, sugars, organic acids, amino acids and sugar alcohols in the amyloplast, cytosol, and vacuole of potato tubers (Farré et al., 2001). These exciting results offer great hope for successful subcellular metabolomics; however, the approaches are still very challenging.

5. Future perspectives

The interest in metabolomics as a large-scale assessment of gene expression is greatly accelerating. We expect that the number of successful projects using metabolomics will continue to grow exponentially; however for this to take place, continued advancement of the analytical technologies will be necessary to enable the visualization of a greater proportion of the metabolome at greater speeds. To accomplish this, it will be essential that multiple and parallel approaches for comprehensive analyses be incorporated. It would be desirable if these comprehensive analyses would expand to encompass many of the chemical classes of molecules that are currently being overlooked such as peptides (i.e.

Table 2
Bioinformatic resources accessible via the internet

KEGG	http://www.genome.ad.jp/kegg/kegg2.html
BRENDA	http://www.brenda.uni-koeln.de/
The EMP Project	http://www.empproject.com
Institute of Biological Sciences, University of Wales, Aberystwyth	http://www.aber.ac.uk/biology/research/abml.html
Douglas Kell's Group	http://qbab.aber.ac.uk/home.html
Virginia Bioinformatics Institute	http://www.vbi.vt.edu
IUBMB Enzyme Nomenclature	http://www.chem.qmul.ac.uk/iubmb/enzyme
Dr. Duke's Phytochemical and Ethnobotanical Databases	http://www.chem.qmul.ac.uk/iubmb/enzyme
The University of Arizona Natural Products Database	http://npd.chem.arizona.edu/about.asp
Iowa State University	http://www.public.iastate.edu/~botany/wurtele.html
Platform Plant Metabolomics	http://www.metabolomics.nl
EcoCyc	http://biocyc.org:1555/ECOLI/class-subs-instances?object=Pathways

the "peptidome"). To extract the biological information contained within these large metabolomic datasets, bioinformatic tools capable of managing the massive experimental data sets and processing them to yield biological knowledge will be necessary. These tools are needed now but are unfortunately only starting to emerge. On the horizon are computer models of cell biochemistry incorporating all levels of gene expression. It is envisioned that such "virtual cells" will help explain and illustrate many of the more challenging details of molecular cellular systems such as metabolite channeling, compartmentalization and transport. Construction of such computer models depends on the availability of extensive quantitative metabolite profiles and improved algorithms. We are hopeful that these tools will soon be in the hands of phytochemists. These tools will have a major impact on the ability to engineer the productive and nutritious crops necessary to feed tomorrow's generations. They will also have a dramatic impact on the ability to use plants as bioreactors producing tomorrow's medicines and chemical stocks.

Acknowledgements

Work in the authors' laboratories is financially supported by The Samuel Roberts Noble Foundation, The Virginia Bioinformatics Institute, The National Science Foundation Grant #DBI-0109732, and The Oklahoma Center for the Advancement of Science and Technology Award #HR02-040.

References

Aharoni, A., Ric de Vos, C.H., Verhoeven, H.A., Maliepaard, C.A., Kruppa, G., Bino, R., Goodenowe, D., 2002. Non-targeted metabolomic profiling using Fourier transform ion cyclotron mass spectrometry (FTMS). OMICS: A Journal of Integrative Biology 6, 217–234.

Arlt, K., Brandt, S., Kehr, J., 2001. Amino acid analysis in five pooled single plant cell samples using capillary electrophoresis coupled to laser-induced fluorescence detection. Journal of Chromatography A 926, 319–325.

Asimov, D., 1985. The grand tour: a tool for viewing multidimensional data. SIAM Journal on Scientific and Statistical Computing 6, 128–143.

Baggett, B.R., Cooper, J.D., Hogan, E.T., Carper, J., Paiva, N.L., Smith, J.T., 2002. Profiling isoflavonoids found in legume root extracts using capillary electrophoresis. Electrophoresis 23, 1642–1651.

Bailey, N.J., Stanley, P.D., Hadfield, S.T., Lindon, J.C., Nicholson, J.K., 2000a. Mass spectrometrically detected directly coupled high performance liquid chromatography nuclear magnetic resonance spectroscopy mass spectrometry for the identification of xenobiotic metabolites in maize plants. Rapid Communications in Mass Spectrometry 14, 679–684.

Bailey, N.J., Cooper, P., Hadfield, S.T., Lenz, E.M., Lindon, J.C., Nicholson, J.K., Stanley, P.D., Wilson, I.D., Wright, B., Taylor, S.D., 2000b. Application of directly coupled HPLC-NMR-MS MS to the identification of metabolites of 5-trifluoromethylpyridone (2-hydroxy-5-trifluoromethylpyridine) in hydroponically grown plants. Journal of Agriculture and Food Chemistry 48, 42–46.

Bairoch, A., 2000. The ENZYME database in 2000. Nucleic Acids Research 28, 304–305.

Bales, J.R., Bell, J.D., Nicholson, J.K., Sadler, P.J., Timbrell, J.A., Hughes, R.D., Bennett, P.N., Williams, R., 1988. Metabolic profiling of body fluids by proton NMR: self-poisoning episodes with paracetamol (acetominophen). Magnetic Resonance in Medicine 6, 300–306.

Bales, J.R., Higham, D.P., Howe, I., Nicholson, J.K., Sadler, P.J., 1984. Use of high resoution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine. Clinical Chemistry 30, 426–432.

Baulcombe, D.C., 1999. Fast forward genetics based on virus-induced gene silencing. Current Opinion in Plant Biology 2, 109–113.

Bednarek, P., Franski, R., Kerhoas, L., Einhorn, J., Wojtaszek, P., Stobiecki, M., 2001. Profiling changes in metabolism of isoflavonoids and their conjugates in *Lupinus albus* treated with biotic elicitor. Phytochemistry 56, 77–85.

Blackstock, W.P., Weir, M.P., 1999. Proteomics: quantitative and physical mapping of cellular protiens. Trends in Biotechnology 17, 121–127.

Bligny, R., Douce, R., 2001. NMR and plant metabolism. Current Opinion in Plant Biology 4, 191–196.

Bochner, B.R., Gadzinski, P., Panomitros, E., 2001. Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. Genome Research 11, 1246–1255.

Boyes, D.C., Zayed, A.M., Ascenzi, R., McCaskill, A.J., Hoffman, N.E., Davis, K.R., 2001. Growth stage-based phenotypic analysis of *Arabidopsis*: a model for high throughput functional genomics in plants. Plant Cell 13, 1499–1510.

Brazma, A., Sarkans, U., Robinson, A., Vilo, J., Vingron, M., Hoheisel, J., Fellenberg, K., 2002. Microarray data representation, annotation and storage. Advances in Biochemical Engineering and Biotechnology 77, 113–139.

Bundy, J.B., Spurgeon, D.J., Svendsen, C., Hankard, P.K., Osborn, D., Lindon, J.C., Nicholson, J.K., 2002. Earthworm species of the genus *Eisenia* can be phenotypically differentiated by metabolic profiling. FEBS Letters 521, 115–120.

Burton, R.A.D.M.G., Bacic, A., Findlay, K., Roberts, K., Hamilton, A., Baulcombe, D.C., Fincher, G.B., 2000. Virus-induced silencing of a plant cellulose synthase gene. Plant Cell 12, 691–705.

Chong, J., Pierrel, M.A., Atanassova, R., WerckReithhart, D., Fritig, B., Saindrenan, P., 2001. Free and conjugated benzoic acid in tobacco plants and cell cultures. Induced accumulation upon elicitation of defense responses and role as salicylic acid precursors. Plant Physiology 125, 318–328.

Cornish-Bowden, A., Eisenthal, R., 1974. Statistical considerations in the estimation of enzyme kinetic parameters by the direct linear plot and other methods. Biochemal Journal 139, 721–730.

Cowan, J.D., Sharp, D.H., 1988. Neural nets. Quarterly Review of Biophysics 21, 365–427.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, Cambridge.

Cunnick, W.R., Cromie, J.B., Cortell, R., Wright, B., Beach, E., Seltzer, F., Miller, S., 1972. Value of biochemical profiling in a periodic health examination program: analysis of 1,000 cases. Bulletin of New York Academy of Medicine 18, 5–22.

Davis, S.C., Makarov, A.A., Hughes, J.D., 1999. Ultrafast gas chromtography using time-of-flight mass spectrometer. Rapid Communications in Mass Spectrometry 13, 237–241.

Devaux, P.G., Horning, M.G., Horning, E.C., 1971. Benyzl-oxime derivative of steroids; a new metabolic profile procedure for human urinary steroids. Analytical Letters 4, 151.

Deyl, Z., Hyanek, J., Jorakova, M., 1986. Profiling of amino acids in body fluids and tissues by means of liquid chromatography. Journal of Chromatography 379, 177–250.

Edgar, R., Domrachev, M., Lash, A.E., 2002. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30, 207–210.

Ellis, L.B., Hershberger, C.D., Bryan, E.M., Wackett, L.P., 2001. The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. Nucleic Acids Research 29, 340–343.

Farré, E.M., Tiesssen, A., Roessner, U., Geigenberger, P., Trethewey, R.N., Willmitzer, L., 2001. Analysis of the compartmentation of glycolytic intermediates, nucleotides, sugars, organic acids, amino acids, and sugar alcohols in potato tubers using a nonaqueous fractionation method. Plant Physiology 127, 685–700.

Fiehn, O., 2002. Metabolomics –the link between genotypes and phenotypes. Plant Molecular Biology 48, 155–171.

Fiehn, O., Kopka, J., Dormann, P., Altmann, T., Trethewey, R.N., Willmitzer, L., 2000. Metabolite profiling for plant functional genomics. Nature Biotechnology 18, 1157–1161.

Fraser, P.D., Pinto, M.E., Holloway, D.E., Bramley, P.M., 2000. Application of high-performance liquid chromatography with photodiode array detection to the metabolic profiling of plant isoprenoids. Plant Journal 24, 551–558.

Fridland, G.H., Desiderio, D.M., 1986. Profiling of neuropeptides using gradient reversed-phase high-performance liquid chromatography with novel detection methodologies. Journal of Chromatography 379, 251–268.

Gates, S.C., Sweeley, C.C., 1978. Quantitative metabolic profiling based on gas chromatography. Clinical Chemistry 24, 1663–1673.

Gerding, T.K., Drenth, B.F., Rossenstein, H.J., de Zeeuw, R.A., Tepper, P.G., Horn, A.S., 1990. The metabolic fate of the dopamine agonist 2-(*N*-propyl-*N*-2-thienylethylamino)-5-hydroxytetralin in rats after intravenous and oral administration. I. Disposition and metabolic profiling. Xenobioctica 20, 515–524.

Goff, S.A., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296, 92–100.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, Mass.

Goodacre, R., Kell, D.B., 1996. Pyrolysis mass spectrometry and its applications in biotechnology. Current Opinion in Biotechnology 7, 20–28.

Gorinstein, S., Zemser, M., Vargasalbores, F., Ochoa, J.L., 1995. Classification of seven species of cactaceae based on their chemical and biochemical properties. Bioscience, Biotechnology, and Biochemistry 59, 2022–2027.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., Kanehisa, M., 2002. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Research 30, 402–404.

Graham, T.L., 1991. A rapid, high resolution high performance liquid chromatography profiling procedure for plant and microbial aromatic secondary metabolites. Plant Physiology 95, 584–593.

Guo, D., Chen, F., Inoue, K., Blount, J.W., Dixon, R.A., 2000. Down-regulation of caffeic acid 3-*O*-methyltransferase and caffeoyl CoA 3-*O*-methyltransferase in transgenic alfalfa (*Medicago sativa* L.): impacts on lignin structure and implications for the biosynthesis of G and S lignin. Plant Cell 13, 73–88.

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between protein and mRNA abundance in yeast. Molecular and Cell Biology 1720–1730.

Halket, J.M., Przyborowska, A., Stein, S.E., Mallard, W.G., Down, S., Chalmers, R.A., 1999. Deconvolution gas chromatography/mass spectrometry of urinary organic acids – potential for pattern recognition and automated identification of metabolic disorders. Rapid Communications in Mass Spectrometry 13, 279–284.

Hall, R., Beale, M., Fiehn, O., Hardy, N., Sumner, L., Bino, R., 2002. Plant metabolomics as the missing link in functional genomics strategies. Plant Cell 14, 1437–1440.

Harrison, M.J., Dixon, R.A., 1993. Isoflavonoid accumulation and expression of defense gene transcripts during the establishment of vesicular arbuscular mycorrhizal associations in roots of *Medicago truncatula*. Molecular Plant-Microbe Interactions 6, 643–654.

Hofmeyr, J.H.S., Cornish-Bowden, A., Rohwer, J.M., 1993. Taking enzyme kinetics out of control–putting control into regulation. European Journal of Biochemistry 212, 833–837.

Hofmeyr, J.H., Cornish-Bowden, A., 1996. Co-response analysis: a new experimental strategy for metabolic control analysis. Journal of Theoretical Biology 182, 371–380.

Holland, J.F., Leary, J.J., Sweeley, C.C., 1986. Advanced instrumentation and strategies for metabolic profiling. Journal Chromatography 1279, 313–345.

Holtorf, H., Guitton, M.-C., Reski, R., 2002. Plant functional genomics. Naturwissenschaften 89, 235–249.

Horning, E.C., Horning, M.G., 1970. Metabolic profiles: chromatographic methods for isolation and characterization of a variety of metabolites in man. In: Olson, R.E. (Ed.), Methods in Medical Research. Year Book Medical Publishers, Chicago, p. 369.

Horning, E.C., Horning, M.G., 1971a. Human metabolic profiles obtained by GC and GC/MS. Journal of Chromatographic Science 9, 129–140.

Horning, E.C., Horning, M.G., 1971b. Metabolic profiles: gas-phase methods for analysis of metabolites. Clinical Chemistry 17, 802–809.

Hotellin, H., 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology 24, 417–441.

Huhman, D.V., Sumner, L.W., 2002. Metabolioc profiling of saponins in *Medicago sativa* and *Medicago truncatula* using HPLC coupled to an electrospray ion-trap mass spectometer. Phytochemistry 59, 347–360.

Humphreys, J.M., Chapple, C., 2002. Rewriting the lignin roadmap. Current Opinion in Plant Biology 5, 224–229.

Ideker, T., Galitski, T., Hood, L., 2001. A new approach to decoding life: sytems biology. Annual Review of Genomics and Human Genetics 2, 343–372.

Jung, W., Yu, O., Lau, S.-M.C., O'Keefe, D.P., Odell, J., Fader, G., McGonigle, B., 2000. Identification and expression of isoflavone synthase, the key enzyme for biosynthesis of isoflavones in legumes. Nature Biotechnology 18, 208–212.

Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A., 2002. The KEGG databases at GenomeNet. Nucleic Acids Research 30, 42–46.

Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collaado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., Gama-Castro, S., 2002. The EcoCyc Database. Nucleic Acids Research 30, 56–58.

Katona, Z.F., Sass, P., Molnár-Perl, I., 1999. Simultaneous determination of sugars, sugar alcohols, acids and amino acids in apricots by gas chromatography-mass spectrometry. Journal of Chromatography 847, 91–102.

Kehoe, D.M., Villand, P., Somerville, S., 1999. DNA microarrays for studies of higher plants and other photosynthetic organisms. Plant Science 4, 38–41.

Kehr, J., 2001. High resolution spatial analysis of plant systems. Current Opinion in Plant Biology 4, 197–201.

Kitano, H., 2000. Perspectives on systems biology. New Generation Computing 18, 199–216.

Kleter, G.A., van der Krieken, W.M., Kok, E.J., Bosch, D., Jordi, W., Gilissen, L.J.W.L., 2001. Regulation and exploitation of genetically modified crops. Nature Biotechnology 19, 1105–1110.

Klose, J., Kobalz, U., 1995. Two-dimensional electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome. Electrophoresis 16, 1034–1059.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69.

Kohonen, T. In: Kohonen, T., Schroeder, M.R., Huang, T.S. (Eds.), 1995. In Self-Organizing Maps, 3rd Ed. Springer-Verlag, New York.

Korolev, A.V., Tomos, A.D., Bowtell, R., Farras, J.F., 2000. Spatial and temporal distribution of solutes in the developing carrot taproot measured at single-cell resolution. Journal of Experimental Botany 51, 567–577.

Kose, F., Weckwerth, W., Linke, T., Fiehn, O., 2001. Visualizing plant metabolomic correlation networks using clique-metabolite matrices. Bioinformatics 17, 1198–1208.

Koosis, D.J., 1997. Statistics A Self-Teaching Guide, 4th Ed. John Wiley & Sons, New York.

Koza, J.R., 1992. Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge Mass.

Lachenbruch, P.A., 1975. Discriminant Analysis. Hafner Press, New York.

Larson, T.R., Graham, I.A., 2001. A novel technique for the sensitive quantification of acyl CoA esters from plant tissues. The Plant Journal 25, 115–125.

Liebich, H.M., 1986. Gas chromatographic profiling of ketone bodies and organic acids in diabetes. Journal of Chromatography 379, 347–366.

Lin, L.Z., He, X.G., Lindenmaier, M., Yang, J., Cleary, M., Qiu, S.X., Cordell, G.A., 2000. LC-ESI-MS study of the flavonoid glycoside malonates of red clover (*Trifolium pratense*). Journal of Agricultural and Food Chemistry 48, 354–365.

Liu, C.-J., Blount, J.W., Steele, C.L., Dixon, R.A., 2002. Bottlenecks for metabolic engineering of isoflavone glycoconjugates in *Arabidopsis*. Proceedings of the National Academy of Sciences, USA 99, 14578–14583.

Lois, R., 1994. Accumulation of UV-absorbing flavonoids induced by UV-B radiation in *Arabidopsis thaliana* L. Planta 194, 498–503.

Longman, P.J., 1999. The curse of Frankenfood. Genetically modified crops stir up controversy at home and abroad. U.S. News & World Report July 16, 39–41.

Maier, W., Schmidt, J., Wray, V., Walter, M.H., Strack, D., 1999. The arbuscular mycorrhizal fungus, *Glomus intraradices*, induces the accumulation of cyclohexenone derivatives in tobacco roots. Planta 207, 620–623.

May, G.D., 2002. An integrated approach to *Medicago* functional genomics. In: Romeo, J.T., Dixon, R.A. (Eds.), Recent Advances in Phytochemistry, Vol. 36. Pergamon, Oxford, UK, pp. 179–195.

McGovern, A.C., Broadhurst, D., Taylor, J., Kaderbhai, N., Winson, M.K., Small, D.A., Rowland, J.J., Kell, D.B., Goodacre, R., 2002. Monitoring of complex industrial bioprocesses for metabolite concentrations using modern spectroscopies and machine learning: application to gibberellic acid production. Biotechnology and Bioengineering 78, 527–538.

Mendes, P., 2002. Emerging bioinformatics for the metabolome. Briefings in Bioinformatics 3, 134–145.

Mendes, P., 2001. Modeling large scale biological systems from functional genomic data: parameter estimation. In: Kitano, H. (Ed.), Foundations of Systems Biology. MIT Press, Cambridge, MA, pp. 163–186.

Mendes, P., Bulmore, D.L., Farmer, A.D., Steadman, P.A., Waugh, M.E., Wlodek, S.T., 2000. PathDB: a second generation metabolic database. In: Hofmeyr, J.-H.S., Rohwer, J.M., Snoep, J.L. (Eds.), Animating the Cellular Map. Stellenbosch University Press, Stellenbosch, pp. 207–212.

Mendes, P., Kell, D.B., 1998. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics 14, 869–883.

Meyer, R.D., Cook, D., 2000. Visualization of data. Current Opinion in Biotechnology 11, 89–96.

Meyermans, H., Morrell, K., Lapierre, C., Pollet, B., De Bruyn, A., Busson, R., Herdewijn, P., Devresse, B., Van Beeumen, J., Marita, J.M., Ralph, J., Chen, C., Burggraeve, B., Van Montagu, M., Messens, E., Boerjan, W., 2000. Modifications in lignin and accumulation of phenolic glycosides in poplar xylem upon down-regulation of caffeoyl coenzyme A *O*-methyltransferase, an enzyme involved in lignin biosynthesis. Journal of Biological Chemistry 275, 36899–36909.

Michal, G., 1999. Biochemical Pathways. An Atlas of Biochemistry and Molecular Biology. John Wiley & Sons, New York.

Miller, J.N., Miller, J.C., 2000. Statistic and Chemometrics for Analytical Chemistry, 4th Ed. Prentice Hall, New York.

Mroczek, W.J., 1972. Biochemical profiling and the natural history of hypertensive diseases. Circulation 45, 1332–1333.

Muir, S.R., Collins, G.J., Robinson, S., Hughes, S., Bovy, A., De Vos, C.H.R., van Tunen, A.J., Verhoeyen, M.E., 2001. Overexpression of petunia chalcone isomerase in tomato results in fruit containing increased levels of flavonols. Nature Biotechnology 19, 470–474.

Nicholson, J.K., Connelly, J., Lindon, J.C., Names, E., 2002. Metabonomics: a platform for studying drug toxicity and gene function. Nature Reviews Drug Discovery 1, 153-161.

Nicholson, J.K., Lindon, J.C., Holmes, E., 1999. "Metabonomics" :understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. Xenobiotica 29, 1181–1189.

Nicholson, J.K., O'Flynn, M., Sadler, P.J., Macleod, A., Juul, S.M., Sonksen, P.H., 1984. Proton NMR studies of serum, plasma and urine from fasting normal, and diabetic subjects. Biochemical Journal 217, 275–365.

Nikiforow, A., Schlick-Steiner, B., Steiner, F., Kalb, R., and Mistrik, R., 2001. Classification of GC-MS data of epicuticular hydrocarbon from *Tetramorium* ants by self-organizing maps for morphological determinations. In Proceedings of the 49th ASMS Conference on Mass Spectrometry and Allied Topics. Chicago, IL, A011287.

Niwa, T., 1986. Metabolic profiling with gas chromatography mass spectrometry and its application to clinical medicine. Journal of Chromatography 379, 3–26.

Oliver, D.J., Nikolau, B., Wurtele, E.S., 2002. Functional genomics: high-throughput mRNA, protein, and metabolite analyses. Metabolic Engineering 4, 98–106.

Oliver, S., 2002. Demand management in cells. Nature 418, 33–34.

Oliver, S., 1997. Yeast as a navigational aid in genome analysis. Microbiololgy 143, 1483–1487.

Oliver, S.G., Winson, M.K., Kell, D.B., Baganz, F., 1998. Systematic functional analysis of the yeast genome. Trends in Biotechnology 16, 373–378.

Palsson, B., 2000. The challenges of *in silico* biology. Nature Biotechnology 18, 1147–1150.

Pritchard, L., Kell, D.B., 2002. Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis. European Journal of Biochemistry 269, 3894–3904.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M.C., Berden, J.A., Brindle, K.M., Kell, D.B., Rowland, J.J., Westerhoff, H.V., van Dam, K., Oliver, S.G., 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nature Biotechnology 19, 45–50.

Ratcliffe, R.G., Shachar-Hill, Y., 2001. Probing plant metabolism with NMR. Annual Review of Plant Physiology and Plant Molecular Biology 52, 499–526.

Roberts, J.K.M., 2000. NMR adventures in the metabolic labyrinth within plants. Trends in Plant Science 5, 30–34.

Roessner, U., Luedemann, A., Brust, D., Fiehn, O., Linke, T., Willmitzeer Fernie, A.R., 2001. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. The Plant Cell 13, 11–29.

Roessner, U., Wagner, C., Kopka, J., Trethewey, R.N., Willmitzer, L., 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography mass spectrometry. The Plant Journal 23, 131–142.

Sauter, H., Lauer, M., Fritsch, H., 1991. Metabolic profiling of plants a new diagnostic technique. In: Baker, D.R., Fenyes, J.G., Moberg, W.K. (Eds.), American Chemical Society Symposium Series No. 443. American Chemical Society, Washington DC, pp. 288–299.

Schomburg, I., Chang, A., Schomburg, D., 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Research 30, 47–49.

Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov Jr., E., Yunus, I., 1996. The metabolic pathway collection from EMP: the enzymes and metabolic pathways database. Nucleic Acids Research 24, 26–28.

Shaw, A.D., Winson, M.K., Woodward, A.M., McGovern, A.C., Davey, H.M., Kaderbhai, N., Broadhurst, D., Gilbert, R.J., Taylor, J., Timmins, E.M., Goodacre, R., Kell, D.B., Alsberg, B.K., Rowland, J.J., 2000. Rapid analysis of high-dimensional bioprocesses using multivariate spectroscopies and advanced chemometrics. Advances in Biochemical Engineering/Biotechnology 66, 83–113.

Smith, R.D., Loo, J.A., Ogorzalek-Loo, R.R., Busman, M., Udseth, H.R., 1991. Principles and practice of electrospray ionization-mass spectrometry for large polypeptides and proteins. Mass Spectrometry Reviews 31, 472–485.

Soga, Y., Ueno, Y., Naraoka, H., Ohashi, Y., Tomita, M., Nishioka, T., 2002. Simultaneous determination of anionic intermediates for *Bcillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. Analytical Chemistry 74, 2233–2239.

Somerville, C., Dangl, J., 2000. Plant biology in 2010. Science 290, 2077–2078.

Somerville, C., Somerville, S., 1999. Plant functional genomics. Science 285, 380–383.

Srere, P.A., Ovádi, J., 1990. Enzyme-enzyme interactions and their metabolic role. FEBS Letters 268, 360–364.

Stashenko, E.E., Acosta, R., Martinez, J.R., 2000. High-resolution gas-chromatographic analysis of the secondary metabolites obtained by subcritical-fluid extraction from Colombian rue (*Ruta graveolens* L.). Journal of Biochemical and Biophysical Methods 43, 379–390.

Stitt, M., Lilley, R.M., Gerhardt, R., Heldt, H.W., 1989. Metabolite levels in specific cells and subcellular comparments of plant leaves. Methods in Enzymology 174, 518–550.

Sumner, L.W., Duran, A.L., Huhman, D.H., Smith, J.T., 2002. Metabolomics: a developing and integral component in functional genomic studies of *Medicago truncatula*. In: Romeo, J.T., Dixon, R.A. (Eds.), Recent Advances in Phytochemistry, Vol. 36. Pergamon, Oxford, UK, pp. 31–61.

Sumner, L.W., Paiva, N.L., Dixon, R.A., Geno, P.W., 1996. HPLC-Continuous-flow liquid secondary ion mass spectrometry of flavonoid glycosides in leguminous plant extracts. Journal of Mass Spectrometry 31, 472–485.

Swayne, D.F., Cook, D., Buja, A., 1998. XGobi: interactive dynamic data visualization in the X Window System. Journal of Computational and Graphical Statistics 7, 113–130.

Tabachnick, B. G., Fidell, L. S., 1983. Using Multivariate Statistics. Harper & Row, New York, pp. 509.

Teusink, B., Baganz, F., Westerhoff, H.V., Oliver, S.G., 1998. Metabolic control analysis as a tool in the elucidation of the function of novel genes. Methods in Microbiology 26, 297–336.

Teusink, B., Passarge, J., Reijenga, C.A., Esgalhado, E., van der Weijden, C.C., Schepper, M., Walsh, M.C., Bakker, B.M., van Dam, K., Westerhoff, H.V., Snoep, J.L., 2000. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. European Journal of Biochemistry 267, 5313–5329.

The *Arabidopsis* Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

The EU *Arabidopsis* Genome Project, 1998. Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature 391, 485–488.

Thiellement, H., Bahrman, N., Damerval, C., Plomion, C., Rossignol, M., Santoni, V., de Vienne, D., Zivy, M., 1999. Proteomics for genetic and physiological studies in plants. Electrophoresis 20, 2013–2026.

Thompson, J.A., Markey, S.P., 1975. Quantitative metabolic profiling of urinary organic acids by gas chromatography-mass spectrometry: comparison of isolation methods. Analytical Chemistry 47, 1313–1321.

Tolstikov, V.V., Fiehn, O., 2002. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. Analytical Biochemistry 301, 298–307.

Törönen, P., Kolehmainen, M., Wong, G., Castrén, E., 1999. Analysis of gene expression data using self-organizing maps. FEBS Letters 451, 142–146.

Trethewey, R.N., Krotzky, A.J., Willmitzer, L., 1999. Metabolic profiling: a Rosetta Stone for genomics? Current Opinion in Plant Biology 2, 83–85.

Trethewey, R.N., 2001. Gene discovery via metabolic profiling. Current Opinions in Biotechnology 12, 135–138.

Tweeddale, H., Notley-McRobb, L., Ferenci, T., 1998. Effect of slow growth on metabolism of *Escherichia coli*, as revealed by global metabolite pool ("Metabolome") analysis. Journal of Bacteriology 180, 5109–5116.

Umbreit, W.W., 1952. Metabolic Maps. Burgess Publishing Company, Minneapolis, MN.

Vaidyanathan, S., Kell, D.B., Goodacre, R., 2001. Flow-injection electrospray ionization mass spectrometry of crude cell extracts for high-throughput bacterial identification. Journal of the American Society for Mass Spectrometry 13, 118–128.

van Wijk, K.J., 2001. Update on plant proteomics. Challenges and prospects of plant proteomics. Plant Physiology 126, 501–508.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzlet, K.W., 1995. Serial analysis of gene expression. Science 270, 484–487.

Venter, J.C., et al., 2001. The Human Genome. Science 291, 1304–1351.

Voirin, B., Bayet, C., 1996. Developmental changes in the monoterpene composition of *Mentha×piperita* leaves from individual peltate trichomes. Phytochemistry 43, 573–580.

Voit, E.O., Radivoyevithc, T., 2000. Biochemical systems analysis of genome-wide expression data. Bioinformatics 16, 1023–1037.

von Röpenack, E., Parr, A., Schulze-Lefert, P., 1998. Structural analyses and dynamics of soluble and cell wall-bound phenolics in a broad spectrum resistance to the powdery mildew fungus in barley. Journal of Biological Chemistry 273, 9013–9022.

Vrbanac, J.J., Braselton, W.E.J., Holland, J.F., Sweeley, C.C., 1982. Automated qualitative and quantitative metabolic profiling analysis of urinary steroids by gas chromatography-mass spectrometry-data system. Journal of Chromatography 239, 265–276.

Weigel, D., Ahn, J.H., Blazquez, M.A., Borewitz, J., Christensen, S.K., Fankhauser, C., Ferrandiz, C., Kardailsky, I., Malanchaurvil, E.J., Neff, M.M., Nguyen, J.T., Sato, S., Wang, Z.-H., Xia, Y., Dixon, R.A., Harrison, M.J., Lamb, C.J., Yanofsky, M.F., Chory, J., 2000. Activation tagging in *Arabidopsis*. Plant Physiology 122, 1003–1013.

Wesley, V.S., Helliwell, C.A., Smith, N.A., Wang, M.B., Rouse, D.T., Liu, Q., Gooding, P.S.S.P.S., Abbott, D., Stoutjesdijk, P.A., Robinson, S.P., Gleave, A.P., Green, A.G., Waterhouse, P.M., 2001. Construct design for efficient, effective and high-throughput gene silencing in plants. The Plant Journal 27, 581–590.

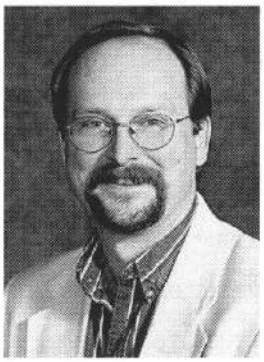
Wittig, U., De Beuckelaer, A., 2001. Analysis and comparison of metabolic pathway databases. Briefings in Bioinformatics 2, 126–142.

Wolf, D., Gray, C.P., de Saizieu, A., 2000. Visualising gene expression in its metabolic context. Briefings in Bioinformatics 1, 297–304.

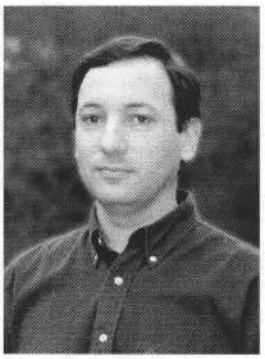
Woolf, T.F., Black, A., Sedman, A., Chang, T., 1992. Metabolic disposition of the non-steroidal anti-inflammatory agent isoxicam in man. European Journal of Drug Metabolism and Pharmacokinetics 17, 21–27.

Yu, J., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296, 79–92.

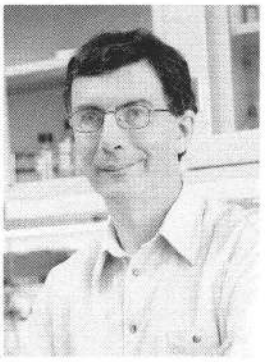
Yu, O., Jung, W., Shi, J., Croes, R.A., Fader, G.M., McGonigle, B., Odell, J.T., 2000. Production of the isoflavones genistein and daidzein in non-legume dicot and monocot tissues. Plant Physiology 124, 781–794.



Lloyd W. Sumner is Head of Biological Mass Spectrometry and Assistant Scientist in the Plant Biology Division, Samuel Roberts Noble Foundation. He obtained his PhD in Analytical Chemistry from Oklahoma State University. He then joined Texas A&M as a staff scientist and later an adjunct faculty member managing The Chemistry Department's mass spectrometry core facility and The Laboratory for Biological Mass Spectrometry with David Russell. Dr Sumner's group is focused on proteomic and metabolomic studies of the model legume, *Medicago trucatula*, and incorporation of these studies into integrated functional genomic and systems biology strategies. His research interests also seek to better understand the fundamental issues associated with mass spectrometry and exploitation of this understanding to better address biological issues.



Pedro Mendes is a Research Assistant Professor at the Virginia Bioinformatics Institute at Virginia Tech. He heads the Biochemical Networks Modeling Group, which carries out research in the areas of systems biology, and bioinformatics applied to functional genomics. Prior to joining VBI, Dr. Mendes was the Program Leader for Pathways at the National Center for Genome Resources (Santa Fe, NM). He graduated as a Biochemistry Major from the University of Lisbon, and obtained a PhD in 1994 from the University of Wales Aberystwyth, under the supervision of Douglas Kell. Dr. Mendes is the author of the popular biochemical simulation software Gepasi (http://www.gepasi.org), widely used for research and education.



Richard A. Dixon is Director of the Plant Biology Division, Samuel Roberts Noble Foundation, and also holds Adjunct Professorships at four US state universities. He received his Bachelor's and Doctoral degrees in Biochemistry from Oxford University, and postdoctoral training with Derek Bendall at Cambridge University. His research interests center on molecular biology and metabolic engineering of natural product pathways. Dr. Dixon is a fellow of the American Association for the Advancement of Science, a member of the editorial boards of Phytochemistry and five other international journals, and was recently named by The Institute for Scientific Information as one of the 15 most cited authors in the plant and animal sciences, 1991–2001.