Opinion TRENDS in Plant Science Vol.xx No.xx Monthxxxx

Potential of metabolomics as a functional genomics tool

Raoul J. Bino[1,2,3], Robert D. Hall[2,3], Oliver Fiehn[4], Joachim Kopka[4], Kazuki Saito[5], John Draper[6], Basil J. Nikolau[7], Pedro Mendes[8], Ute Roessner-Tunali[9], Michael H. Beale[10], Richard N. Trethewey[11], B. Markus Lange[12], Eve Syrkin Wurtele[13] and Lloyd W. Sumner[14]

[1]Plant Physiology Department, Wageningen University, Arboretumlaan 4, 6703 BD Wageningen, The Netherlands
[2]Plant Research International B.V., POB 16, 6700 AA Wageningen, The Netherlands
[3]Centre for BioSystems Genomics, POB 98, 6700 AB Wageningen, The Netherlands
[4]Max-Planck Institute of Molecular Plant Physiology, 14424 Potsdam, Germany
[5]Graduate School of Pharmaceutical Sciences, Chiba University, Yayoi-cho 1-33, Inage-ku, Chiba 263-8522, Japan
[6]Institute of Biological Science, Edward Llwyd Building, University of Wales, Aberystwyth, Ceredigion, UK SY23 3DA
[7]Center for Designer Crops, 2210 Molecular Biology Building, Iowa State University, Ames, IA 50011, USA
[8]Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, 1880 Pratt Drive, Blacksburg, VA 24061, USA
[9]Australian Centre for Plant Functional Genomics, School of Botany, University of Melbourne, Victoria 3010, Australia
[10]National Centre for Plant and Microbial Metabolomics, Rothamsted Research, Harpenden, Herts, UK AL5 2JQ
[11]Metanomics GmbH & Co KGaA, Tegeler Weg 33, 10589 Berlin, Germany
[12]Institute of Biological Chemistry, Washington State University, Pullman, WA 99164-6340, USA
[13]Department of Genetics, Development, and Cell Biology, 441 Bessey Hall, Iowa State University, Ames, IA 10011, USA
[14]Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

Metabolomics is developing as an important functional genomics tool; however, there is still room for technical improvements in both the large-scale determination of metabolites from complex plant tissues and the dissemination of metabolomics research data. For the continued maturation of metabolomics, the following three objectives need to be achieved: (i) improvement in the comprehensive coverage of the plant metabolome, (ii) facilitation of comparison of results between laboratories and experiments, and (iii) enhancement of the integration of metabolomic data with other functional genomic information. Because these challenges are widely recognized and endorsed, we propose community-based efforts to define common criteria and to initiate concerted actions directed towards the release of standard reference materials, construction of consolidated metabolite libraries, and development of metabolite-specific data-management systems.

Metabolomics (comprehensive analysis in which all the metabolites of an organism are identified and quantified [1]) has emerged as a functional genomics methodology that contributes to our understanding of the complex molecular interactions in biological systems [2]. As such, metabolomics represents the logical progression from large-scale analysis of RNA and proteins at the systems level [3]. In recent years, several reviews have been published [1,4–9] describing the use of metabolomics in functional genomics research. Currently, metabolomics is being applied in many biological studies ranging from carbon–nitrogen interactions in plants [4] to the development of personal metabolomics as the next generation of nutritional assessment in humans [10]. Indeed, the biochemical response of an organism to a conditional perturbation can be characterized by its effect on the differential accumulation of individual metabolites [11]. A better understanding of the correlation between genes and the functional phenotype of an organism is the true goal of all functional genomics strategies. But, how far has metabolomics developed towards meeting this goal and how does it compare with other functional genomics approaches? In this opinion article, we briefly discuss the current status and suggest additional steps that are needed for further maturation and advancement of metabolomics as a productive, complementary functional genomics and systems biology tool.

Corresponding author: Raoul J. Bino (raoul.bino@wur.nl).

Current status

Biological relevance

For a holistic understanding of the biological behavior of a complex system, it is essential to follow, as unambiguously as possible, the response of an organism to a conditional perturbation at the transcriptome, proteome and metabolome levels [12,13]. These three levels of expression profiling provide a complete picture of the RNAs, proteins and metabolites that enable one to: infer relevant associations between macromolecules; identify functional linkages between phenotypic expressions; and construct models that quantitatively describe the dynamics of the biological system. Broad phenotypic analyses are essential if we are to progress from prediction to experimental validation of gene function [4].

 1360-1385/$ - see front matter doi:10.1016/j.tplants.2004.07.004

Metabolites in plants function in many resistance and stress responses and contribute to the color, taste, aroma and scent of flowers and fruits. The biochemical phenotype of an organism is the final result of interactions between the genotype and the environment (G×E), but it is also modulated by sub-cellular physiological fluctuations that are part of homeostasis [3]. Thus, the simultaneous identification and quantification of metabolites is necessary to study the dynamics of the metabolome, to analyze fluxes in metabolic pathways and to decipher the role of each metabolite following various stimuli [14]. The challenge of metabolomics is to find changes in the metabolic network that are functionally correlated with the physiological and developmental phenotype of a cell, tissue or organism [15]. Linkage of functional metabolomic information to mRNA and protein expression data makes it possible to visualize the functional genomic repertoire of an organism. This knowledge has great potential for application, for example, the efficient engineering of crops that combine an attractive appearance and taste with improved levels of phytonutrients such as flavonoids and carotenoids.

Comprehensiveness

The enormous biochemical diversity displayed in the plant kingdom is estimated to exceed 200 000 different metabolites [16]. It is therefore in plants that large-scale, comprehensive metabolite profiling meets its greatest challenge – a challenge that provides the impetus for cutting-edge technological developments. Various experimental approaches are currently being pursued to profile and determine the chemical identity of plant metabolites (Box 1). The need for multiple technologies reflects the technical difficulty of measuring metabolites owing to large variations in their relative concentrations and chemical complexities [4,6]. The analytical approaches deployed vary in their ability to provide unambiguous identification of individual metabolites in complex samples [17]. Currently plant metabolomics is still limited in its comprehensiveness. Of the estimated 5000 different primary and secondary metabolites anticipated in a typical *Arabidopsis* leaf, we estimate that ~10% have been annotated using current technologies. Typically, metabolites are identified through spectral comparisons with authentic compounds [18] contained within spectral libraries such as the NIST (http://www.nist.gov/), Wiley (http://www.wileyregistry.com/) or Sigma–Adrich (http://www.sigmaaldrich.com/Area_of_Interest/Equip___Supplies_Home/Spectral_Viewer/FT_NMR_Library.html) libraries. Although these libraries contain >350 000 entries, most of these are non-biological compounds and lack information on chromatographic behavior, which is essential, particularly for the identification of isomers [18]. Thus, custom databases are often used to enhance identification confidence and success rates, but up to 70% of peaks in a typical gas chromatography mass spectrometric (GC/MS) analysis of a plant extract remain unidentified. To complicate the story further, liquid chromatography (LC/MS) mass spectral libraries are, in most cases, instrument dependent and therefore standard reference LC/MS libraries are unavailable for general use, making the identification efficiency even lower for non-volatile metabolites.

Box 1. Metabolomics technologies

Large-scale, comprehensive metabolite profiling was first approached by Ute Roessner *et al.* [44–46] who detected 150 compounds simultaneously within a potato (*Solanum tuberosum*) tuber using gas chromatographic mass spectrometry (GC/MS). Of these, 77 could be chemically identified as amino acids, organic acids or sugars. Using a similar approach, Oliver Fiehn *et al.* [1] found 326 components in *Arabidopsis thaliana* leaf extracts and could assign a chemical identity to about half of them. Liquid chromatographic mass spectrometry (LC/MS) was used by Vladimir Tolstikov and Oliver Fiehn [47] to detect sugars, amino acids and some glycosides in phloem exudates of *Cucurbita maxima*, and by David Huhman and Lloyd Sumner [48] to identify 27 saponins in *Medicago truncatula*. Both groups used a multidimensional approach in which chromatography was combined with tandem mass spectrometry to obtain detailed metabolite structural information based upon the compound specificity of the collision-induced cleavage of chemical bonds and the observed fragmentation patterns. More recently, capillary LC/MS using monolithic columns have been applied to metabolome profiling of *Arabidopsis* [39]. The authors reported detecting several hundred chromatographic peaks and were able to deconvolute the data to reveal >700 unique ions. Using high-resolution Fourier transform mass spectrometry (FTMS), Asaph Aharoni *et al.* [49] reported 5844 different masses [based on mass-to-charge (*m/z*) values] from strawberry (*Fragaria ananassa*) fruit tissues and assigned putative empirical chemical formulae to more than half of them. The chemical formulae were achieved by determining the most probable elemental composition of the metabolites based upon the accurate mass measured. The results showed variations in both primary metabolites (i.e. amino acids, fatty acids and carbohydrates) as well as secondary metabolites (i.e. flavonoids and terpenoids) in the various strawberry tissues [49]. Recently, Edda Von Roepenack-Lahaye *et al.* [33] detected >1400 components (based on *m/z* values) from *Arabidopsis* leaf extracts using a quadrupole time-of-flight (QTOF) mass spectrometer. In addition to MS-based approaches, nuclear magnetic resonance (NMR) is also being used in metabolomic analyses [50–52]. NMR generates high-throughput fingerprints, it is quantitative and non destructive. However, NMR is generally of lower sensitivity than MS and suffers from overlapping signals, leading to smaller numbers of absolute identifications. The future might involve a combination of LC, NMR and MS systems that could increase the numbers of quantifiable and identifiable metabolites.

Current limitations of metabolomics and comparison with other functional genomics approaches

Metabolomics could benefit from a more comprehensive coverage. This is also true for other functional genomics technologies. For example, Heiko Schoof *et al.* proposed that the *Arabidopsis* genome contains ~28 000 genes, of which 50% have been successfully annotated [19], a percentage likely to increase rapidly [20]. The comprehensiveness of genomic databases facilitates the functional characterization of mRNAs in a transcriptomic approach. But the accuracy and completeness of nucleotide databases are equally important to proteomics because techniques such as homology sequence comparisons, peptide mass fingerprinting, and MS-based sequence tag technologies using tandem mass spectrometry (MS/MS) rely on information contained in these databases for accurate protein identification [21]. The successful identification of proteins also depends upon the efficacy of protein separations; current gel-based proteomic assays of

Arabidopsis can profile ~500–2500 proteins [22] (depending on the particular tissue) of the conservatively estimated 25 000 proteins that might be present in *Arabidopsis*. Apparently, the current degree of comprehensiveness of proteomic approaches in plants is in the same range as metabolomics. Future improvements of proteomics essentially depend upon more effective separation and identification of individual proteins. Similarly, the advancement of metabolomics will also depend upon increases in separation efficiencies and identification of individual metabolites. Notwithstanding that, an important difference is that unlike mRNAs and proteins, it is difficult or impossible to establish a direct link between individual metabolites and genes. Functions have been proven for many plant metabolites or can be inferred from our knowledge of other organisms [23,24]. However, the same metabolite can be a member of several different pathways and also have regulatory effects on multiple biological processes. Therefore, individual metabolites cannot, in most cases, be unambiguously linked to a single genomic sequence [25].

Future directions for metabolomics

Our goal is to promote plant metabolomics as a valuable functional genomics tool that provides a comprehensive characterization of the biochemical phenotype of a plant. The realization of this goal will require improved technology for the determination of metabolites in complex plant tissues and the integration and dissemination of metabolomics research data. To integrate and disseminate metabolomics research data, a metabolomics information standard should help to ensure that metabolite data and metadata (detailed experimental information such as sample preparation, instrument settings and analysis conditions) can be easily interpreted and that results can be independently verified outside the original source laboratory [8]. Such standardized information systems have been suggested for the proteome [26–28] and transcriptome, and in accordance with MIAME (Minimal Information About a Microarray Experiment) [29], we suggest Minimum Information About a METabolomics experiment (MIAMET) in Box 2. However, elevating technical performance to enable the broader capture of metabolomic data with the required throughput and accuracy of identification is even more challenging. Because these challenges are widely recognized and endorsed [2,30], this encourages a community-based effort to define common criteria and to initiate several concerted actions. In response to the acknowledged challenges, we propose the following three steps.

(i) Improved comprehensive coverage of the metabolome

Inference of biological context from metabolomics data ultimately relies on the accurate identification of metabolites. The minimum information acceptable for the identification of novel organic compounds or metabolites has been traditionally defined by the scientific literature criteria and often includes elemental analysis, NMR and MS spectral data for the isolated compound. These data are necessary to ensure accurate metabolic identifications, but do not necessarily need to be repeated in each metabolomics experiments because the majority of metabolites have been previously characterized at this level of analytical rigor in the published literature. Conversely, a single chemical shift or mass value is insufficient to provide confident metabolite identification. Therefore, we suggest the definition of a minimum quality standard for metabolite identifications in metabolic profiling experiments. These criteria should not be as stringent as those for novel compounds (identified for the first time) and we propose a minimum that includes two orthogonal dimensions of chemical characterization relative to an authentic compound. For example, retention time (or retention index) and exact molecular mass or mass spectral fragmentation pattern in a GC/MS or LC/MS experiment. Although we suggest this as a minimum, we believe that more rigorous identification schemes are necessary for compounds without commercially available authentic standards, more complex molecules and for absolute stereochemical elucidation. Hyphenated techniques that couple chromatography to mass spectrometry and/or to NMR, such as LC/MS/NMR analysis, might offer the greatest confidence in sample identifications [31,32] but represent a large expense that might be prohibitive to many laboratories. In many cases, a single spectrometric determination gives insufficient detail for confident metabolite identification. For example, direct-injection (without chromatographic LC separation) into a mass analyzer might allow visualization of many components, but does not enable simple differentiation between isomeric configurations (i.e. glucose or galactose) and is more sensitive to matrix effects. The best analytical approaches for large-scale screening and preliminary identification of unknowns appear to be two-dimensional instrumental techniques (based on each combination of GC/MS, LC/MS, GC/MS/MS, LC/MS/MS or LC/NMR/MS), which enable both comparative profiling and structural elucidation. For example, LC/QTOFMS/MS (liquid chromatographic quadrupole tandem time-of-flight mass spectroscopy) has the potential to provide accurate mass and product-ion information of chromatographically separated metabolites [33]. Experimental mass data can then be used for calculating an elemental composition and be compared with available mass information in, for example, the NIST or KEGG databases; product-ion information from tandem MS can be used to determine or confirm structure.

We emphasize that the putatively identified metabolites be further validated by comparison to chromatographic, chemical or spectral characteristics of authentic standards. Standards can also be used to 'spike' extracts, calculate recovery rates and facilitate the quantification of particular metabolites. Although metabolite standards are important, in many cases they are difficult to obtain. To alleviate this situation, we suggest the initiation of a system to facilitate the exchange of purified or synthetic reference compounds (the reference material being a purified fraction from a plant extract, authenticated by NMR and MS) between research centers applying complementary technologies. In this manner, metabolic standards can be analyzed by various techniques to yield

Box 2. MIAMET (Minimum Information About a METabolomics experiment)

Metabolomics is becoming a widely used technology to evaluate global metabolite dynamics. Although many significant results have been derived from metabolomics studies, the lack of standards for presenting and exchanging such data limit the widespread access of metabolomics data to the broader research community. One reason for the difficult exchange of information is that metabolic data are complex; different platforms and experimental designs produce data in various formats and units. For the understanding of the experiment and the interpretation of resulting data, it is imperative that authors provide full information regarding the experimental design, sample preparation, analytical methodology and data analysis. However, it is difficult to present such details in a format that is both inclusive and practical. A similar challenge was faced recently by laboratories using microarray technology for global gene expression data profiling. The recognition of the necessity to establish standards for microarray data annotation and exchange led to the formation of the Microarray Gene Expression Data (MGED) Society, which outlined the minimum information that should be reported about a microarray experiment (MIAME: Minimal Information About a Microarray Experiment [29]; http://www.mged.org/Workgroups/MIAME/miame.html). Here, we propose MIAMET (Minimum Information About a METabolomics experiment) as an analogous standard, but adapted to the specifics of metabolomics.

The following MIAMET is a suggestion to the community in the hopes of soliciting further input to refine this evolving concept.

Experimental design

- Experimental type: for example, is it a comparison of normal versus diseased tissue, a time course, or is it designed to study the effects of a gene knockout?
- Experimental factors: the parameters or conditions tested, such as time, dose or genetic variation.
- Experimental description: a description of the comparisons made in each experiment, whether to a standard reference sample, or between experimental samples. An accompanying diagram or table might be useful.
- Quality control steps taken: for example, biological and/or analytical replicates, blanks, positive and negative controls.
- Protocols, materials and methods should be submitted as separate text documents supporting the experimentation. It is envisioned that they will be incorporated into developing databases.
- URL of any supplemental websites or database accession numbers.

Sampling, preparation, metabolite extraction and derivatization

- The origin of the biological sample (e.g. species name, variety and the provider of the sample) and its characteristics.
- Manipulation of biological samples and protocols used: growth conditions, light–dark photoperiods, light intensity, treatments, specific tissue sampled.
- Details of any sample treatment, such as biotic and/or abiotic perturbations (e.g. pathogen, exogenous elicitors and nutrient deficiencies).
- Protocol for preparing the metabolite extract: such as the extraction, enrichment and/or purification protocols.
- Derivatization protocol (for GC/MS).
- External controls (spikes) added to the samples.
- Special handling information.

Metabolic profiling design

- General metabolic profiling design, including instrumental platform (e.g. NMR, LC/NMR, GC/MS, LC/MS, FTICR-MS and FTIR); model number and name, and performance specifications. Other nontraditional platforms should be described and appropriate references validating the technology provided.
- Instrumental parameters [NMR probe type, field strength and sample temperature, mass spectrometer type (TOF, quadrupole, ion-trap), operational mode such as positive-ion ESI (electro spray ionization) or APCI (atmospheric pressure chemical ionization)].
- Separation conditions for hyphenated chromatography methods (including column, stationary phase composition, flow-rates and split ratios).
- Instrument performance validation (e.g. sensitivity, resolution and mass accuracy).
- Collected data in both raw and processed form should be considered for database submission and public access via the Internet. Use of universal data formats, such as NetCDF for MS data, JCAMP for NMR (http://my.unidata.ucar.edu/content/software/netcdf/index.html) or.txt formatted files is advised.

Metabolite measurement and specifications

- The eventual output of metabolic profiling experiments is expected to be a list of metabolite identifiers, both known and unknown, and a corresponding relative or absolute quantification value.
- The metabolite identifiers should consist of descriptive codes, such as the International Union of Pure and Applied Chemistry (IUPAC) name, chemical abstract number (CAS#), or emerging EBI notation (European Bioinformatics Institute) for known compounds, and according to nomenclature described in Box 5 for unknown compounds.
- Additional supporting information qualifying the identification (software used, *m/z* measured, retention index, spectral database used for comparative identification and matching score, chemical shifts and absorption frequency).
- Description of data processing:

(i) Data normalization.
(ii) Software used for calculating relative or absolute quantifications.
(iii) Formula used for calibration curve equation.
(iv) Formula used for relative quantification, based on internal standard or spike.
(v) Formula for data transformation.

custom reference databases or libraries. These libraries can then be combined to generate consolidated spectral metabolite reference libraries that will be made publicly available. Crucial to the success of this scheme will be the openness of collaborators and the free exchange of information within the public arena.

(ii) Reference materials and facilitation of comparative results

Standard reference materials would allow comparison of the experimental and instrumental efficacy between laboratories and technologies. Because most metabolomic approaches use different technology platforms (e.g. Fourier Transform/MS, Time-Of-Flight/MS, ion-trap and NMR) that vary in their range of measured metabolites, accuracy, resolution, dynamic range and sensitivity [17], reference materials would allow validation of technical performance and a mechanism for comparative performance evaluation. To facilitate such cross-platform comparisons, we will select and make available standard mixtures of authenticated compounds and plant reference materials of known chemical composition and established metabolic phenotype. Currently, we are in the process of composing such a set of reference materials for *Lycopersicon esculentum* (Plant Research International, The Netherlands), *Arabidopsis thaliana* (The National Centre for Plant and Microbial Metabolomics, Rothamsted Research, UK) and *Medicago truncatula* (Noble Foundation, USA) that will be made available through a simple procedure to the community. The reference materials will

consist of lyophilized plant material of standard mixtures of extracts from homozygous plant lines. The plant material will also be made available as seeds that can be sown by each individual researcher to analyze the biological variation typifying the local conditions. The reference material can be used to test the efficacy of the metabolomic platform used. The composition of the reference mixtures and extracts will be characterized and analysis reports provided by the institute of origin, for example, the standard GC/EI (electron ionization)/MS platform established at the Max-Planck Institute of Molecular Plant Physiology, Germany (Box 3).

(iii) Integration of metabolomics with other functional genomics data

The development, establishment and integration of metabolomics databases will bridge the barriers between metabolomics and other functional genomics approaches (i.e. transcriptomics and proteomics) and will allow the development of biological systems networks by integrating transcriptome, proteome, metabolome and flux data [12,34]. Possibly the most advanced genomics database for plants is The *Arabidopsis* Information Resource (TAIR) [35]. Other plant resources are available as well (e.g. http://www.york.ac.uk/res/garnet/garnet.htm, http://www.genome.ad.jp/kegg/pathway.html and http://www.maizegdb.org). Integrated biological networks of known interactions in plants are beginning to be assembled (Box 4). As these pathway-oriented networks are emerging, it becomes even more essential to develop comprehensive metabolomic datasets.

It is our opinion that the maturation of metabolomics as the next cornerstone of functional genomics ultimately depends on establishing metabolomics relational databases that store, compare, integrate and enable the determination of causal relationships between genes, transcripts, proteins and metabolites. All functional genomics approaches are 'information-rich', but each method is also vulnerable to various statistical caveats because the data originate from a few samples, yet each sample is characterized by several thousand features [e.g. genes, or m/z values (mass-to-charge ratios of metabolites or metabolite fragments) and spectral intensities] that might lead to difficulties in the interpretation

Box 3. Mass spectral libraries for identification of metabolites in complex samples

Each laboratory involved in metabolite profiling is challenged by identification and characterization of the hundreds to thousands of metabolites. The crucial step of metabolite identification is currently solved by a large-scale series of time-consuming additional experiments that catalog the mass spectra of standard metabolites and chromatographic retention indices. This information is then used to generate custom spectral libraries that help to further identify the unknown metabolites. Currently, no platform exists that facilitates exchange of metabolite identification or information on as yet unidentified components. To help elevate this situation, the Max-Planck Institute of Molecular Plant Physiology has made available a collection of mass spectral libraries (MSRI) that have been compiled from authentic reference compounds and a range of plants and non-plant reference samples (http://csbdb.mpimp-golm.mpg.de/csbdb/dbma/msri.htm). The composition of the reference mixtures and extracts has been characterized by gas chromatographic electron-ionization mass spectroscopy (GC/EI/MS). The libraries include 2000 mass spectra and are available to the scientific community to screen samples for known metabolites. The libraries are continually updated and query forms are generated for each new biological sample and new application. This approach makes it possible to compare results using standardized GC/MS settings (e.g. carrier flow, temperature ramp and capillary columns) but is limited in respect to the volatility of compounds. Therefore, additional reference libraries still need to be established for other technologies such as LC/MS and NMR.

Box 4. Current and evolving plant metabolomics databases and data analysis tools

AraCyc (http://www.Arabidopsis.org/tools/aracyc/)
A tool to visualize biochemical pathways of *Arabidopsis* [53]. The software allows querying and the graphical representation of biochemical pathways and expression data.

ArMet (http://www.armet.org/)
A framework for the description of plant metabolomics experiments and their results. ArMet encompasses the entire timeline of a plant metabolomics experiment. The design and code allow the detailed description of each step in the experiment and has the ability to define detailed sub-components. ArMet and MIAMET (Box 2) together make it possible to describe all relevant metabolic information and to communicate results in a standard way to the scientific community.

DOME (a Database of OMEs; http://medicago.vbi.vt.edu/dome.html)
Composed of various sub-sections that contain metadata, raw data, analysis results and an ontology describing the known molecular biology of the plant species of interest. Results are processed using multiple statistical tools and visualized using a BRowser for OMEs (BROME).

MetaCyc (http://metacyc.org/)
A metabolic pathway database that contains pathways from >150 different organisms. MetaCyc describes metabolic pathways, reactions, enzymes and substrate compounds that were gathered from a variety of literature and on-line sources and contains citations to the source of each pathway.

MapMan (http://gabi.rzpd.de/projects/MapMan/)
A user-driven tool that displays large datasets onto diagrams of metabolic pathways or other processes [54]. It is composed of multiple modules for hierarchical grouping of transcript and metabolite data that can be visualized using a separate user-guided module. Editing existing modules and creation of new categories or modules is possible and provides flexibility.

MetNet http://www.public.iastate.edu/~mash/MetNet/homepage.html
Contains a suite of open-source software tools for systems biology and is designed to provide a framework for the formulation of testable hypotheses regarding the function of specific genes. It currently contains four tools. (i) MetNetDB is a metabolic and regulatory network map that contains a growing map of *Arabidopsis* entities (genes, RNAs, polypeptides, protein complexes and metabolites) and the catalytic and regulatory interactions between them. (ii) GeneGobi analyzes datasets statistically and visually. (iii) FCModeler is a tool to graph and model data allowing submission of organelle-specific data and visualization of sub-cellular compartmentalized metabolites. (iv) MetNetVR is a biological network in virtual reality that makes it possible to visualize experimental datasets interactively in combination with the metabolic and regulatory network from MetNetDB in 3-dimensional-space [55].

and validation of resultant data [36]. The generation of database tools to query and/or comprehensively mine metabolomic data depends on the availability of metabolite databases that can be trusted and for which the source of data and its history are maintained and made publicly accessible. In each data repository, expert assessment and data curation are important to assure the uniformity and quality of the information [37,38]. Data acquisition, transformation, validation and annotation are all aspects of curation. Although several bioinformatic tools to unravel 'deconvolute' and process metabolomic data have been developed [39,40] (for pre-processing of MS-data, see www.metalign.nl), no generally adopted procedure to transform and annotate metabolite data has yet been proposed that can be used independently from its technological platform. There is a great need for validated data models that define suitable approaches for generation, pre-processing and storage of metabolomic data. To facilitate data model development, we will encourage cooperative research that integrates datasets. As a first step, we foresee the need to name each known and unknown compound uniquely; Figure I in Box 5 depicts a preliminary scheme to name unknowns in a developing database that collects data from three distinct laboratories. We propose to use such a scheme to generate a consensus peak list from a commonly used matrix that can be used as a basis for the development of data-processing models that accept raw data from different instruments.

Implementation of future steps

Many of the details for the suggested common actions, outlined in sections (i–iii) above, still need to be established for effective implementation. For example, generation of online libraries with retention indices not only requires the willingness of individual laboratories to contribute, but also requires mutual agreement on the technical details of such libraries, including chromatographic columns and the development of retention time indices standards. Detailed input for the common datasets need to be established: these might include the tissue extraction protocols, separation column specifics and information about the chromatographic behavior of each peak. The International Committee on Plant Metabolomics, of which authors of this article are members (http://www.metabolomics.nl/), represents a platform to accomplish the proposed actions, to establish the required details, to interact with other microbial and animal metabolomics groups, and to facilitate integration of metabolomics into systems biology.

Future steps will use the validated and curated metabolomic information to study the dynamics of the metabolome, to analyze fluxes in metabolic pathways and to decipher the biological relevance of each metabolite. As the comprehensiveness increases and bioinformatic tools mature, functional metabolomic information can be linked to transcriptome and proteome datasets to enable a better understanding of plant biology [41,42]. Obviously, new issues and challenges will emerge, such as those associated with spatially resolved technologies in which materials are isolated via laser-capture micro-dissection [43]. Such methods require the preservation of sufficient anatomical detail following sectioning and necessitate maximum recovery of the molecules of interest from the isolated cells. If considered necessary, we will define new common actions to tackle these emerging problems because our goal is to pursue the exciting opportunities of the comprehensive and nontargeted profiling of metabolites in plants and to further advocate and document

Box 5. Naming the unknowns

We propose a scheme for uniquely naming unknown compounds or hypothetical compounds (Figure Ia) and show an example (Figure Ib). Note that the first identifier (laboratory) is the only one that requires coordination by the community, the following qualifiers are particular to that laboratory. The last qualifiers (FEATURE2 onwards) are optional and are designed to allow specifying characteristic ions resulting from, for example, fragmentation (MS-MS). Using this scheme, each compound will have a unique name until it is possible to name it according to its chemical nature.

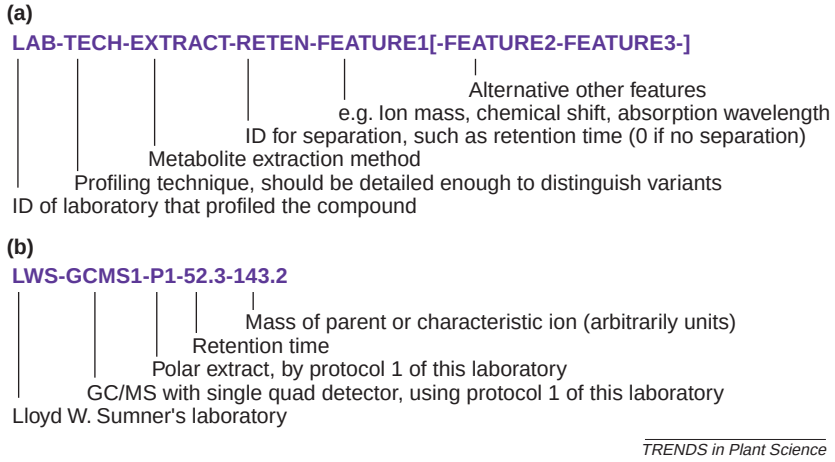

Figure I.

metabolomics as a discovery tool for functional genomics and systems biology in plant sciences.

Acknowledgements
We thank Richard A. Dixon, Maarten Koornneef, Sander van der Krol and Corey Broeckling for reading the text and valuable suggestions. The Samuel Roberts Noble Foundation is acknowledged for its financial support (R.J.B. and L.W.S.).

References

1 Fiehn, O. (2002) Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48, 155–171
2 Hall, R. *et al*. (2002) Plant metabolomics: the missing link in functional genomics strategies. *Plant Cell* 14, 1437–1440
3 Weckwerth, W. (2003) Metabolomics in systems biology. *Annu. Rev. Plant Biol.* 54, 669–689
4 Stitt, M. and Fernie, A.R. (2003) From measurement of metabolites to metabolomics an 'on the fly' perspective illustrated by recent studies of carbon–nitrogen interactions. *Curr. Opin. Biotechnol.* 14, 136–144
5 Fernie, A.R. (2003) Metabolome characterization in plant system analysis. *Funct Plant Biol.* 30, 111–120
6 Sumner, L.W. *et al*. (2003) Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62, 817–836
7 Trethewey, R.N. (2004) Metabolite profiling as an aid to metabolic engineering in plants. *Curr. Opin. Plant Biol.* 7, 196–201
8 Mendes, P. (2002) Emerging bioinformatics for the metabolome. *Brief. Bioinform.* 3, 134–145
9 Kopka, J. *et al*. (2004) Metabolite profiling in plant biology: platforms and destinations. *Genome Biol.* 5, 109–117
10 German, J.B. *et al*. (2003) Personal metabolomics as a next generation nutritional assessment. *J. Nutr.* 133, 4260–4266
11 Raamsdonk, L.M. *et al*. (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat. Biotechnol.* 19, 45–50
12 Oliver, D. *et al*. (2002) Functional genomics: high throughput mRNA, protein, and metabolite analyses. *Metab. Eng.* 4, 98–108
13 Sweetlove, L. *et al*. (2003) Predictive metabolic engineering: a goal for systems biology. *Plant Physiol.* 132, 420–425
14 Fiehn, O. and Weckwerth, W. (2003) Deciphering metabolic networks. *Eur. J. Biochem.* 270, 579–588
15 Weckwerth, W. *et al*. (2004) Metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U. S. A.* 101, 7809–7814
16 Pichersky, E. and Gang, D. (2000) Genetics and biochemistry of secondary metabolites: an evolutionary perspective. *Trends Plant Sci.* 5, 439–445
17 Cristoni, S. and Bernardi, L.R. (2003) Development of new methodologies for the mass spectrometry study of bioorganic macromolecules. *Mass Spectrom. Rev.* 22, 369–406
18 Wagner, C. *et al*. (2003) Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* 62, 887–900
19 Schoof, H. *et al*. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.* 32, D373–D376
20 Yamada, K. *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302, 842–846
21 Yates, J.R., III. (1998) Mass spectroscopy and the age of the proteome. *J. Mass Spectrom.* 33, 1–19
22 Heazlewood, J.L. and Millar, A.H. (2003) Integrated plant proteomics – putting the green genomes to work. *Funct Plant Biol.* 30, 471–482
23 Nikiforova, V. *et al*. (2003) Transcriptome analysis of sulfur depletion in *Arabidopsis thaliana*: interlacing of biosynthetic pathways provides response specificity. *Plant J.*, 633–650
24 Hirai, M.Y. *et al*. (2003) Global expression profiling of sulfur-starved *Arabidopsis* by DNA macroarray reveals the role of *O*-acetyl-L-serine as a general regulator of gene expression in response to sulfur nutrition. *Plant J.* 33, 651–663
25 Schwab, W. (2003) Metabolome diversity: too few genes, too many metabolites? *Phytochemistry* 62, 837–849
26 Westbrook, J. *et al*. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31, 489–491
27 Hermjakob, H. *et al*. (2004) The HUPO PSI's molecular interaction format – a community standard for the representation of protein interaction data. *Nat. Biotechnol.* 22, 177–183
28 Taylor, C. *et al*. (2003) A systematic approach to modeling, capturing, and disseminating proteomics experimental data. *Nat. Biotechnol.* 21, 247–254
29 Brazma, A. *et al*. (2001) Minimum information about a microarray experiment (MIAME) – toward standards for microarray data. *Nat. Genet.* 29, 365–371
30 Mazur, B. (2003) Plant metabolic profiling *en route* to destination. *Nat. Biotechnol.* 21, 875–876
31 Lenz, E. *et al*. (2002) Flow injection analysis with multiple on-line spectroscopic analysis (UV, IR, 1H-NMR and MS). *J. Pharm. Biomed. Anal.* 27, 191–200
32 Grivet, J-P. *et al*. (2003) NMR and microbiology: from physiology to metabolomics. *Biochimie* 85, 823–840
33 Von Roepenack-Lahaye, E. *et al*. (2004) Profiling of *Arabidopsis* secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. *Plant Physiol.* 134, 548–559
34 Ge, H. *et al*. (2003) Integrating 'omic' information: a bridge between genomics and systems biology. *Trends Genet.* 19, 551–560
35 Rhee, S.Y. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* 31, 224–228
36 Somorjai, R. *et al*. (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 19, 1484–1491
37 Reiser, L. *et al*. (2002) Surviving in a sea of data: a survey of plant genome data resources and issues in building data management systems. *Plant Mol. Biol.* 48, 59–74
38 Kell, D.B. *et al*. (2001) Genomic computing. Explanatory analysis of plant expression profiling data using machine learning. *Plant Physiol.* 126, 943–951
39 Tolstikov, V. *et al*. (2003) Monolithic silica-based capillary reversed-phase liquid chromatography/electrospray mass spectrometry for plant metabolomics. *Anal. Chem.* 75, 6737–6740
40 Duran, A. *et al*. (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* 19, 2283–2293
41 Sweetlove, L. *et al*. (2003) Predictive metabolic engineering: a goal for systems biology. *Plant Physiol.* 132, 420–425
42 Hirai, M.Y. *et al*. (2004) Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in *Arabidopsis thaliana*. *Proc. Natl. Acad. Sci. U. S. A.* 101, 10205–10210
43 Asano, T. *et al*. (2002) Construction of a specialized cDNA library from plant cells isolated by laser capture microdissection: toward comprehensive analysis of the genes expressed in the rice phloem. *Plant J.* 32, 401–408
44 Roessner, U. *et al*. (2001) High-resolution metabolic phenotyping of genetically and environmentally diverse potato tuber systems. Identification of phenocopies. *Plant Physiol.* 127, 749–764
45 Roessner, U. *et al*. (2000) Simultaneous analysis of metabolites in potato tuber by gas chromatography-mass spectrometry. *Plant J.* 23, 131–142
46 Roessner, U. *et al*. (2001) Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell* 13, 11–29
47 Tolstikov, V.V. and Fiehn, O. (2002) Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and elctrospray ion mass trap spectrometry. *Anal. Biochem.* 301, 298–307
48 Huhman, D.V. and Sumner, L.W. (2002) Metabolic profiling of saponins in *Medicago sativa* and *Medicago truncatula* using HPLC coupled to an electrospray ion-trap mass spectrometer. *Phytochemistry* 59, 347–360
49 Aharoni, A. *et al*. (2002) Nontargeted metabolome analysis by use of Fourier transform ion cyclotron mass spectrometry. *OMICS* 6, 217–234
50 Nicholson, J. *et al*. (2002) Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discov.* 1, 153–161

51 Nicholson, J. and Wilson, I. (2003) Opinion: understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* 2, 668–676

52 Ward, J.L. *et al*. (2003) Assessment of ^{1}H NMR spectroscopy and multivariate analysis as a technique for metabolite fingerprinting of *Arabidopsis thaliana*. *Phytochemistry*, 949–957

53 Mueller, L. *et al*. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Physiol.* 132, 453–460

54 Thimm, O. *et al*. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* 37, 914–939

55 Wurtele, E. *et al*. (2003) MetNet: software to build and model the biogenetic lattice of Arabidopsis. *Comp. Funct. Genomics* 4, 239–245