Metabolomics technology and bioinformatics

Vladimir Shulaev

Submitted: 17th March 2006; Received (in revised form): 7th April 2006

Abstract

Metabolomics is the global analysis of all or a large number of cellular metabolites. Like other functional genomics research, metabolomics generates large amounts of data. Handling, processing and analysis of this data is a clear challenge and requires specialized mathematical, statistical and bioinformatics tools. Metabolomics needs for bioinformatics span through data and information management, raw analytical data processing, metabolomics standards and ontology, statistical analysis and data mining, data integration and mathematical modelling of metabolic networks within a framework of systems biology. The major approaches in metabolomics, along with the modern analytical tools used for data generation, are reviewed in the context of these specific bioinformatics needs.

Keywords: metabolomics; metabolite profiling; mass spectrometry; GC–MS; LC–MS; CE–MS; bioinformatics

INTRODUCTION

Metabolomics is the study of cells by measuring profiles of all, or a large number, of their metabolites. Metabolomics was originally proposed as a method of functional genomics [1] but its utility extends well beyond that—it is useful whenever an assessment of changes in metabolite levels is important. Examples exist for applications in microorganisms, plants and animals including humans. Metabolomics is used for comparing mutants [2], assessing responses to environmental stress [3, 4], studying global effects of genetic manipulation [5], comparing different growth stages [6, 7], toxicology [8], drug discovery [9], nutrition [10–14], cancer [15], diabetes [16] and natural product discovery [17]. Metabolite profiling, whether targeting specific metabolite classes or untargeted, can also be applied as a tool in systems biology [18, 19], where metabolite snapshots [20] are used to study cellular dynamics [21–23] through mathematical models [24, 25]. There are three major approaches used in metabolomics studies: (i) targeted analysis, (ii) metabolite profiling and (iii) metabolic fingerprinting (Table 1) [26, 27].

TARGETED ANALYSIS

Targeted analysis is the most developed analytical approach in metabolomics. It is used to measure the concentration of a limited number of known metabolites precisely. To perform targeted analysis, one must know the structure of the target metabolite and have an analytical method developed to properly measure its concentration in the sample. Targeted analysis is a truly quantitative approach and provides very low limits of detection for known metabolites. It can also be used in a high-throughput mode, depending on the analyte of interest. The major limitation of the targeted analysis for metabolomics is that it requires the compounds of interest to be known a priori, and these must be available in purified form. Currently a large number of metabolites cannot be positively identified in samples using existing analytical techniques, and for many metabolites that can be identified, purified standards are not available. Therefore, this method cannot be used to identify novel metabolic markers or, at present, survey global metabolic changes. In the future, when more metabolites will be identified and

Vladimir Shulaev, Virginia Bioinformatics Institute, Washington Street, Blacksburg, VA 24061, USA. Tel: 540-231-3489; Fax: 540-231-2606; E-mail: vshulaev@vbi.vt.edu

Vladimir Shulaev is an Associate Professor at the Virginia Bioinformatics Institute at the Virginia Polytechnic Institute and State University. He heads the Biochemical Profiling Group, which focuses on developing methods for high throughput metabolite profiling and application of metabolomics to systems biology and to study stress response in microorganisms, plants and animals.
pure compounds will be available to develop quantitative assays, targeted analysis will be increasingly used to survey global metabolic changes since it provides truly quantitative data.

**METABOLIC FINGERPRINTING**

Metabolic fingerprinting does not attempt to identify or precisely quantify all the metabolites in the sample. Rather, it considers a total profile, or fingerprint, as a unique pattern characterizing a snapshot of the metabolism in a particular cell line or tissue [28]. Pattern recognition tools are used to classify the fingerprints and identify the specific features of the profile that are characteristic for each pattern. Metabolic fingerprinting is most useful in biomarker discovery and diagnostics [29, 30]. Fingerprinting is usually performed with spectroscopic techniques such as nuclear magnetic resonance (NMR) [31–33], Fourier transform infrared spectroscopy (FT-IR) [34, 35], Fourier transform ion cyclotron resonance mass spectroscopy (FTICR-MS) or mass spectrometry (MS) [28] by directly acquiring physical spectra without prior separation techniques like chromatography or electrophoresis. The disadvantage of metabolic fingerprinting is that it does not allow direct comparison of the metabolic pathways since compounds in the samples are not readily identifiable. Metabolic fingerprinting has been widely used to discover specific metabolic patterns of diseases [36, 37]. Metabolic fingerprinting is particularly useful in combination with pattern recognition and discriminant analysis techniques [2, 30, 38–42].

**METABOLITE PROFILING**

Metabolite profiling is the measurement of the levels of a metabolite set in a sample. Metabolite profiling is a well-established activity in biomedical sciences, used routinely on biological fluids as a form of diagnosis that aids in characterizing the health state of a patient. Those profiles are typically composed of a small number of metabolites (and maybe proteins), but this is now being extended to much larger numbers [19]. Recently, researchers have been paying more attention to metabolite profiling as an extension of functional genomics. It has been postulated that metabolite profiles of the internal state of cells could aid in the identification of the function of genes, especially when mutants in that gene have no apparent phenotype [1]. The rationale is that, while the mutation would have caused an effect, the regulatory mechanisms of the cell acted to counteract that effect resulting in no clear macroscopic observable. However, those regulatory mechanisms would have changed the concentrations of metabolites in the metabolic network, and if one could measure the levels of those metabolites, it could lead one to the function of the mutated gene [1, 2, 21, 43]. The first implication here is that, not knowing a priori which metabolites are expected to be altered, the profiles then need to be as comprehensible as possible; the second implication is that studying the metabolite profiles of a number of mutants might lead to the discovery of the underlying metabolic network [44–46]. This application of metabolite profiling in functional genomics is then similar to transcript and protein profiling, and like these, it will be useful to establish the complete composition of the cell in terms of metabolites—the metabolome.

**Table I: Major approaches in metabolomics**

<table>
<thead>
<tr>
<th>Approach</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Targeted analysis</td>
<td>Quantitative</td>
<td>Limited number of compounds can be targeted</td>
</tr>
<tr>
<td></td>
<td>Low limit of detection</td>
<td>Does not detect compounds that were not targeted</td>
</tr>
<tr>
<td></td>
<td>High throughput</td>
<td>Targeted compounds must be available purified for calibration</td>
</tr>
<tr>
<td>Metabolite profiling</td>
<td>Global (not targeted)</td>
<td>Semi-quantitative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Majority of peaks are not identifiable</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Difficult informatics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Medium throughput</td>
</tr>
<tr>
<td>Metabolic fingerprinting</td>
<td>Global (not targeted)</td>
<td>No compound identification</td>
</tr>
<tr>
<td></td>
<td>Directly applicable to pattern recognition</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Highest throughput</td>
<td></td>
</tr>
</tbody>
</table>

**ANALYTICAL APPROACHES**

There are multiple techniques that can be used for metabolite profiling (reviewed in [47]).
Each technique has associated advantages and drawbacks (Table 2). Thus, a combination of different analytical technologies must be used to gain a broad perspective of the metabolome of a tissue. Analytical techniques that are most often used for metabolite profiling include NMR [8], gas chromatography–mass spectrometry (GC–MS) [48], liquid chromatography–mass spectrometry (LC–MS) [49–51] and capillary electrophoresis–mass spectrometry (CE–MS) [52–54]. Clearly, no single analytical methodology is ideal for all of the many thousands of metabolites within a cell; instead, the combination of techniques is needed for analysing majority of metabolites in different polarity and molecular weight range.

| Table 2: Common analytical techniques used in metabolomics |
|---|---|---|
| Analytical method | Advantage | Disadvantage |
| NMR | • Rapid analysis | • Low sensitivity |
| | • High resolution | • Convoluted spectra |
| | • No derivatization needed | • More than one peak per component |
| | • Non-destructive | • Libraries of limited use due to complex matrix |
| | • High resolution | • Slow |
| | • Robust | • Often requires derivatization |
| GC–MS | • Large linear range | • Many analytes thermally-unstable or too large for analysis |
| | • Large commercial and public libraries | |
| LC–MS | • No derivatization required (usually) | |
| | • Many modes of separation available | |
| | • Large sample capacity | |
| CE–MS | • High separation power | • Slow |
| | • Small sample requirements | • Limited commercial libraries |
| | • Rapid analysis | |
| | • Can analyse neutrals, anions and cations in single run | |
| | • No derivatization (usually) | |
| FTIR | • Rapid analysis | • Extremely convoluted spectra |
| | • Complete fingerprint of sample chemical composition | • More than one peak per component |
| | • No derivatization needed | • Metabolite identification nearly impossible |
| | | • Requires sample drying |

Metabolite extraction
The first requirement for the extraction protocol is comprehensiveness, i.e. the extract should represent as large number of the cellular metabolites as possible. The other prime requirement is that enzymatic activity be halted for the duration of the extraction process to prevent possible degradation or inter-conversion of metabolites. To this end, high acid/alkali and high organic solvent must be present in the extraction medium. No single extraction methodology is ideal for all metabolites within a cell or tissue. Typically extraction protocol is tailored toward specific compound classes and/or analytical technique. Metabolomics poses a significant challenge for extraction methodology due to the required comprehensiveness of the extract. Extraction conditions that favour preservation of one metabolite class can inevitably destroy other metabolite species, making it extremely difficult to find the right balance between comprehensiveness and metabolite stability. Recently, several studies have been performed to assess the applicability of different extraction methodology to metabolomics [55, 56]. Ultimately, extraction methodology should be adapted toward required metabolite range and the specific analytical technique used. Extraction solutions are usually spiked with one or more internal standards for subsequent quantification of metabolites and to assess their recovery from the sample.

Nuclear magnetic resonance (NMR)
NMR is well-suited to metabolomics studies as it can uniquely identify and simultaneously quantify a wide range of organic compounds in the micromolar range [3, 57–59]. Unlike MS, NMR is non-destructive, so samples can continue for further analysis. Sample preparation for NMR is straightforward and largely automated. However, the analysis of NMR spectra of complex mixtures has traditionally been fraught with difficulty. NMR has been extensively used for metabolite fingerprinting, profiling and metabolic flux analysis [3, 32, 36, 58, 60]. The major limitation of NMR for comprehensive metabolite profiling is its relatively...
low sensitivity, making it inappropriate for the analysis of large number of low-abundance metabolites.

**Mass spectrometry**

Due to its high sensitivity and wide range of covered metabolites, MS has become the technique of choice in many metabolomics studies (reviewed by Glinski et al. [27, 61]). MS can be used to analyse biological samples either directly via direct-injection MS or following chromatographic or electrophoretic separation. Recent developments of new mass analysers as well as improvements in mass accuracy dramatically expanded the range of metabolites that can be analysed by MS and improved the accuracy of compound identification. Direct-injection MS, especially when using a high-resolution mass spectrometer, provides a very rapid technique to analyse large number of metabolites, and therefore is extensively used for metabolic fingerprinting and metabolite profiling. However, direct infusion also has drawbacks, mostly due to a phenomenon known as co-suppression, where the signal of many compounds to be analysed (analytes) that have low ionization efficiencies can be lost at the mass spectrometer interface. To avoid these problems and to decrease the complexity of the sample, MS is often used as a hyphenated technique, i.e. it is coupled with gas chromatography (GC), liquid chromatography (LC) or capillary electrophoresis (CE). In this case, the sample mixture is separated by chromatography or electrophoresis, and individual compounds or less complex mixtures of compounds (which cannot be separated due to similar properties) that are eluted from the column or capillary are analysed by MS.

**Gas chromatography–mass spectrometry (GC–MS)**

Currently the most mature technology for rapid metabolite profiling is gas chromatography coupled with electron impact (EI) quadrupole or time-of-flight (TOF) MS (GC–MS) [48, 62]. The concept of automated GC–MS metabolic profiling was developed over 25 years ago [63–65] and was later adopted as a major technology for metabolomics [48]. Using this approach, it is possible to simultaneously profile several hundred chemically diverse compounds including organic acids, most amino acids, sugars, sugar alcohols, aromatic amines and fatty acids [48]. Volatile metabolites can be separated and quantified by GC–MS, directly. For others, chemical derivatization is required to make them amenable for GC–MS analysis. Although chemical derivatization provides significant improvement in the GC separation of many compounds, it also can introduce artifacts due to the derivatization process itself [27, 66].

A significant advantage of GC–MS with EI ionization is the availability of many searchable mass spectral libraries. The largest commercial EI libraries are the 2005 NIST/EPA/NIH Mass Spectral Library (http://www.nist.gov/srd/nist1.htm) containing 190 825 spectra and 7th edition of the Wiley Registry of Mass Spectral Data (http://www.wiley.com), which contains 338 000 EI mass spectra (or over 460 000 EI mass spectra combined with NIST library). In addition to commercially available EI libraries, a number of public libraries are also available. A list of available GC–MS libraries can be found in [27]. Unfortunately, most mass spectral libraries are tailored toward the chemical industry, drug studies or natural product discovery, and therefore do not represent a large number of naturally occurring metabolites and their intermediates. This limits their applicability to metabolomics studies.

**Liquid chromatography–mass spectrometry (LC–MS)**

LC–MS is being increasingly used in metabolomics applications due to its high sensitivity and a range in analyte polarity and molecular mass wider than GC–MS. Liquid chromatography, and specifically high-performance liquid chromatography (HPLC), is a mature technique that combines high resolution and analytical flexibility. It can be tailored for the analysis of a specific metabolite or class of compounds, or it can be used for the analysis of a broad range of compound classes. Numerous applications for a broad range of metabolites already exist or can be converted for using LC in combination with MS (reviewed in [67–69]). LC–MS has one advantage over GC–MS, in that, there is largely no need for chemical derivatization of metabolites (which is required for analysis of non-volatile compounds by GC–MS). During the last decade, LC–MS techniques have developed, which employ soft ionization methods like electrospray (ESI), making mass spectrometers more sophisticated and robust for daily use. More recently, achievements in separation science using monolithic
Capillary columns [51] and ultra-performance liquid chromatography (UPLC) [70–81] provide much better solutions for separation of complex biological mixtures than previously attainable. A substantial drawback for the LC–MS as a non-targeted profiling tool is the lack of transferable mass spectral libraries. On the other hand, LC–MS can be successfully used to elucidate the structure of unknown compounds. Modern ion-trap mass spectrometers can carry out multiple sequential steps of MS providing significant aid in structure elucidation [82].

Capillary electrophoresis–mass spectrometry (CE–MS)

Despite the fact that CE–MS is a relatively recent technique compared with GC–MS and LC–MS, it is becoming one of the major analytical techniques for metabolomics. CE–MS provides several important advantages over other separation techniques. It has very high resolving power with plate numbers of 100 000–1 000 000, very small sample requirement with average injection volume ranging from 1 to 20 nl, and short analysis time. CE has been used for both targeted and non-targeted analysis of metabolites [83–86], including analysis of inorganic ions [87], organic acids [88, 89], amino acids [90], nucleotides and nucleosides [91], vitamins [92], thiols [93], carbohydrates [94] and peptides [95]. One of the significant advantages of the CE–MS is the ability to separate cations, anions and uncharged molecules in a single analytical run, and therefore CE can be used for simultaneous profiling of many different metabolite classes [52, 53, 96]. This feature makes it a very attractive and promising analytical technique for high-throughput non-targeted metabolomics.

Other technologies

Although NMR and MS are most often used for large-scale analysis, metabolomics is not limited to these techniques. Other alternatives include thin-layer chromatography [6, 97], HPLC with UV/visible absorbance, photodiode array (PDA) [49] or electrochemical detectors [98–100], FT-IR [34, 35, 101], ‘phenotype microarrays’ [102] and variety of other enzymatic assays. Combined use of multiple techniques or multiple detectors in online or parallel analysis can significantly increase the metabolite coverage, increase quantification limits and improve identification of metabolites from a single biological sample. Using a combination of mass spectrometric and electrochemical array detection, for example, Gamache and colleagues [98] achieved broader metabolite coverage and an improved concentration range of endogenous rat urinary metabolites following xenobiotic toxin exposure.

BIOINFORMATICS FOR METABOLOMICS

Like other functional genomics research, metabolomics generates large amounts of data. Handling, processing and analysing this data is a clear challenge for researchers and requires specialized mathematical, statistical and bioinformatical tools (reviewed in [103]). Functional genomics data sets are extensive and multi-dimensional, making their organization in properly designed databases a necessity rather than an option. The analysis of these data sets is equally challenging, and by no means is it an established procedure—useful algorithm selections are still being established. Metabolomics has unique bioinformatics needs in addition to others common in microarray or proteomics data. Major areas where developments in bioinformatics are crucial for further progress of metabolomics include: data and information management, raw analytical data processing, metabolomics standards and ontology, statistical analysis and data mining, data integration and mathematical modelling of metabolic networks within the framework of systems biology.

Metabolomics raw data processing is probably the most challenging and time consuming step in data analysis. Generally, processing a set of raw chromatograms involves noise reduction, spectrum deconvolution, peak detection and integration, chromatogram alignment, compound identification and quantification. By its nature metabolomics requires automated data processing solutions. AMDIS (automated mass spectral deconvolution and identification system, http://chemdata.nist.gov/mass-spc/amdis/) software that utilizes well-described algorithms, has proven to be extremely useful in processing of the GC–MS data [104], but its applicability to LC–MS or CE–MS is somewhat limited. ESI–LC–MS data can be processed using component detection algorithm (CODA)[105] or ‘windowed mass selection method’ (WMSM) [106]. Although a series of commercial and public tools exist [105, 107–111], none of them provides a comprehensive solution to meet the challenges of metabolomics. Future developments in this area
are vital for progress. Availability of a vendor-independent data-processing software pipeline that is modular and flexible enough to incorporate new algorithms could significantly boost the progress in metabolomics. Recent development of the MZmine (http://mzmine.sourceforge.net/index.shtml), a platform-independent software for processing of the LC–MS data in metabolomics and proteomics applications, provides an important step in this direction [112]. The software employs a modular infrastructure with the ability to integrate new algorithms and applications. Other important feature of the software is that it is expandable to other types of mass spectral data like GC–MS and CE–MS and is vendor independent.

Metabolomics data have been analyzed with a wide range of statistical and machine-learning algorithms. These can be classified in two major classes: unsupervised and supervised algorithms [25, 47]. Examples of unsupervised methods that have been routinely used in analysing metabolomics data are hierarchical clustering [113], principal component analysis (PCA) [113, 114] and self-organizing maps [113, 115]. Supervised methods include ANOVA [116], partial least squares (PLS) [117] and discriminant function analysis (DFA) [2].

Metabolomics data sets, like other ‘omic’ data sets, are largely underdetermined, meaning they contain many more variables than samples [118]. In a typical ‘omic’ experiment an average of several hundreds to tens of thousands of variables are measured (i.e. all the genes in microarray experiment, or hundreds of metabolites in metabolomics study), but only a relatively small number of samples are collected to examine this high-dimensional space. For statistical analysis of these data, it is important to reduce the number of variables in order to obtain uncorrelated features in the data. This can be best achieved either through significance methods in ANOVA and t-tests, through linear combinations of variables in PCA or by using evolutionary algorithms such as genetic algorithms or genetic programming. Evolutionary algorithms are usually carried out in combination with a second analysis algorithm (e.g. PLS or DFA) that search for combinations of variables most effective in the secondary algorithm, and are guided by principles of evolution and selection of species (reviewed in [119]). Evolutionary algorithms have been successfully applied to metabolomics data [41, 120, 121]. Other challenges that today’s bioinformatics faces is the problem of integrating data from different ‘omics’ platforms (for a recent review, see [103]). This is becoming more evident as ‘omics’ research is moving toward modeling biochemical networks through systems biology.

Database management systems for metabolomics are required to collect both metadata, raw and processed experimental data. Storing metadata, covering experimental design, the nature of the samples and their treatment prior to the analysis and information about the analytical technique and data-processing details are important to be able to reproduce the experimental conditions and compare results obtained in different laboratories. This is, in part, similar to the requirements of the MIAME protocol for microarray data [122], but has a number of extra requirements beyond MIAME. Equally important is storing and organizing raw data coming from the analytical instrument, as well as subsequently processed and statistically analysed data. In many cases, data for a single biological sample contains several parallel streams from different instruments obtained using different analytical techniques. This creates an additional challenge for storing and archiving metabolomics data and emphasizes the need for properly designed databases. Also greatly needed are reference databases that collect the list of metabolites observed in each species, e.g. human, Arabidopsis, Drosophila, yeast, etc. Ideally, the metabolomics database should be comprehensive and flexible enough to incorporate new data types owing to novel developments in technology as well as to accommodate corresponding data from parallel ‘omics’ platforms including transcriptomics and proteomics that are often collected in the same experiment and share metadata.

A number of databases, data management, analysis and visualization tools are currently publicly available. These include, among others, metabolic pathway databases and pathway viewers KEGG (http://www.genome.ad.jp/kegg/), MetaCyc (http://metacyc.org/), AraCyc (http://www.Arabidopsis.org/tools/aracyc/), MapMan (http://gabi.rzpd.de/projects/MapMan/) and KaPPA-View (http://kpv.kazusa.or.jp/kappa-view/), the data model for plant metabolomics experiments ArMet (http://www.armet.org/), functional genomics databases MetNet (http://metnet.vrac.iastate.edu/) and DOME (http://medicago.vbi.vt.edu). DOME, a database developed by the Mendes group at the Virginia Bioinformatics Institute [25, 45, 103],
provides an example of a comprehensive data management system for metabolomics as well as for other genomics data. DOME is intended for the management of metabolic, proteomic and transcriptomic data. The system includes a subset, B-Net, which acts as a reference database. The B-Net schema is intended to capture the background information about the molecular biology of a specific species that exists in the literature. DOME also incorporates a suite of tools for statistical data analysis and data visualization (e.g. BROME, standing for BROWSER for OMEs).

Metabolomics as a research area will ultimately require standards of data management, analysis and reporting to be adopted by community [123–125]. At this time there are no agreed upon standards for metabolomics, although the recently created Metabolomics Society (http://www.metabolomics society.org/) has a working group and discussion forum focused on the development of such standards (http://www.metabolomics society.org/nstandards.html). Furthermore, in August 2005 the Metabolomics Standards Workshop sponsored by the NIH and organized with the help of the Metabolomics Society was held at the NIH to discuss ongoing standards efforts and to outline the needs, challenges and approaches in developing a Metabolomics Standards Initiative (MSI) (discussed in detail by Castle et al. [126] in this issue).

CONCLUSIONS
Metabolomics is a relatively young discipline, but it has already developed as an important integrative part of genomics and systems biology research. Further developments in this area require improvements in both analytical science and bioinformatics. Development of new analytical techniques is largely focused on increasing resolution and comprehensiveness, increasing speed and throughput of analytical assays and equipment miniaturization. Examples of new promising technologies for metabolomics include micro-fluidic devices [127–130], UPLC [70–81], multidimensional chromatography with various detectors (i.e. GC × GC × TOF–MS [131–133], GC × GC × Quadropole MS [134] or LC–LC–MS or LC–CE–MS [135–139], LC–NMR–MS [140–147]) and metabolite arrays.

To meet growing challenges in metabolomics, a concerted community–wide effort is required. Several recent community papers [123, 124], as well as activities of the recently created Metabolomics Society, are major steps in this direction and pave the way for further steady progress of the discipline.

Key Points
- There are three major approaches used in metabolomics studies: targeted analysis, metabolite profiling and metabolic fingerprinting.
- The combination of analytical techniques is needed for analysing the majority of metabolites within a cell.
- MS is the technique of choice in metabolomics because of its high sensitivity and wide range of covered metabolites.
- Metabolomics requires specialized mathematical, statistical and bioinformatics tools.
- Further developments in metabolomics require improvements in both analytical science and bioinformatics as well as standards of data management, analysis and reporting.

Acknowledgements
The research in the author’s lab is financially supported by the NIGMS grant R01 GM068947–01 and NSF grants MCB-03128257 and MCB-0520140. The author thanks Jim Walke for critically reading the manuscript.

References


41. Kell DB. Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic


Metabolomics technology and bioinformatics.


108. Windig W, Marchcin TF, Meyer GN. Determining sample differences in a sample series of complex liquid


