Spectroscopic and Statistical Techniques for Information Recovery in Metabonomics and Metabolomics

John C. Lindon and Jeremy K. Nicholson

Department of Biomolecular Medicine, Faculty of Medicine, Imperial College London, London SW7 2AZ, United Kingdom; email: j.nicholson@imperial.ac.uk

Annu. Rev. Anal. Chem. 2008. 1:45–69

First published online as a Review in Advance on December 20, 2007

The *Annual Review of Analytical Chemistry* is online at anchem.annualreviews.org

This article's doi:
10.1146/annurev.anchem.1.031207.113026

1936-1327/08/0719-0045$20.00

Key Words

nuclear magnetic resonance, mass spectrometry, chromatography, chemometrics, multivariate statistics, systems biology

Abstract

Methods for generating and interpreting metabolic profiles based on nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), and chemometric analysis methods are summarized and the relative strengths and weaknesses of NMR and chromatography-coupled MS approaches are discussed. Given that all data sets measured to date only probe subsets of complex metabolic profiles, we describe recent developments for enhanced information recovery from the resulting complex data sets, including integration of NMR- and MS-based metabonomic results and combination of metabonomic data with data from proteomics, transcriptomics, and genomics. We summarize the breadth of applications, highlight some current activities, discuss the issues relating to metabonomics, and identify future trends.

1. INTRODUCTION

NMR: nuclear magnetic resonance

MS: mass spectrometry

Metabonomics: the quantitative measurement of the dynamic multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification

Metabolomics: a comprehensive analysis in which all the metabolites of a biological system are identified and quantified

Molecular biology has mainly centered on the determination of multiple gene expression changes, either between subjects or following drug treatment or other interventions (termed transcriptomics). Much effort has also been put into determination of multiple protein expression changes in a cell or tissue (termed proteomics). However, the main problem with interpreting transcriptomic and proteomic data is the difficulty of relating observed gene expression fold changes or protein level (not activity) changes to conventional disease and pharmaceutically relevant endpoints.

At the metabolic (small-molecule) level, similar developments have been taking place. Many years before the development of the various "-omics" approaches, simultaneous analysis of the plethora of metabolites seen in biological fluids had been carried out largely using nuclear magnetic resonance (NMR) spectroscopy (1) and mass spectrometry (MS) (2), and these complex data sets were interpreted in detail using multivariate statistics (3, 4). This approach to understanding the metabolic responses of complex systems to some sort of stimulus was christened metabonomics and was carefully defined with respect to the measurement of biological effects (5). Metabonomics encompasses the comprehensive and simultaneous systematic profiling of metabolite levels and their systematic and temporal changes in whole organisms through effects such as diet, lifestyle, environment, genetic effects, and pharmaceutical effects both beneficial and adverse, and is achieved through the study of biofluids and tissues (5, 6). A parallel approach arose in the late 1990s. Mainly from plant and microbial sciences and originally from the study of in vitro cellular systems, this approach has led to the coining of the similar term metabolomics (7). This term has a broadly analytical definition, but the methods and approaches used for cells, plants, and animals are now highly convergent. The main techniques and applications of both approaches have been reviewed in a recent book (8).

As discussed below, there is a need to integrate information at the transcriptomic, proteomic, and metabonomic levels to provide a full systems biology understanding; this goal has been, at best, only partially fulfilled. For example, environmental and lifestyle effects have a large effect upon all levels of molecular biology. Animals, including humans, can be considered "superorganisms" with an internal ecosystem of diverse symbiotic gut microflora (often with unknown genomes and functional ecologies) whose metabolic processes interact with the host. The complexity of mammalian biological systems and the diverse features that need to be measured to allow -omics data to be fully interpreted have been reviewed recently (9), and novel approaches are required to measure and model such cometabolic processes (10).

Metabonomic studies generally use biofluids or cell or tissue extracts, which are usually readily available. Urine and plasma are obtained essentially noninvasively, and hence can be obtained more easily for use in disease diagnosis and in clinical trials for monitoring drug therapy. However, many other fluids have been studied, including seminal fluids, amniotic fluid, cerebrospinal fluid, synovial fluid, digestive fluids, blister and cyst fluids, lung aspirates, and dialysis fluids (8). In addition, numerous metabonomics studies have analyzed tissue biopsy samples and their lipid and

aqueous extracts, as well as in vitro cell systems such as Caco-2 cells (11), model systems such as yeast (12), tumor cells, and spheroids (13).

UPLC: ultraperformance liquid chromatography

2. ANALYTICAL TECHNOLOGIES FOR METABONOMICS AND METABOLOMICS

2.1. Introduction

The main analytical techniques employed for metabonomic studies are based on NMR spectroscopy and MS. This is for the good reason that both technologies can deliver "high-density" spectroscopic/structural information on a wide range of compound classes and chemistries simultaneously with high analytical precision. The use of MS requires a preseparation of the metabolic components using either gas chromatography (GC) after chemical derivatization or liquid chromatography (LC), with the newer method of ultraperformance LC (UPLC) also being used increasingly. Additionally, the use of capillary electrophoresis (CE) coupled to MS has also shown some promise. Other more specialized techniques, such as Fourier transform infrared spectroscopy and arrayed electrochemical detection, have been used in some cases. The main limitation of the use of these latter techniques is the low level of detailed molecular identification that can be achieved, and, as a result, MS is also employed for metabolite identification.

All metabonomic studies yield complex multivariate data sets; these usually require visualization software and chemometric and bioinformatic methods for interpretation and production of biochemical fingerprints that are of diagnostic or other classification value. The next step is to identify the substances causing the diagnosis or classification, as these are the combination of biomarkers that define the biological or clinical situation.

2.2. Nuclear Magnetic Resonance Spectroscopy

NMR spectroscopy provides detailed information on molecular structure, both for pure compounds and in complex mixtures, but it can also be used to probe metabolite molecular dynamics and mobility through the interpretation of NMR spin relaxation times and by the determination of molecular diffusion coefficients (14). Commercially available instruments up to an observation frequency for ^{1}H of 950 MHz are available (with higher frequencies on the way). For small-molecule studies, the increased sensitivity and dispersion that result from the higher magnetic fields are of great value and there are no disadvantages as there could be in macromolecule studies.

Automatic sample preparation involves simply buffering and adding D_2O as a magnetic field lock signal for the spectrometer; standard NMR spectra typically take only a few minutes to acquire using robotic flow-injection methods. For large-scale studies, bar-coded vials containing the biofluid can be used, and the contents can be transferred into 96-well plates under LIMS system control and prepared for analysis using robotic liquid-handling technology. Currently, the use of these approaches allows well over 100 samples per day to be be measured on one spectrometer.

Alternatively, for more precious samples or for those of limited volume, conventional 5-mm or capillary NMR tubes are usually used, either individually or with a variety of commercially available sample tube changers with automatic data acquisition. The large NMR signal that arises from water in all biofluids is easily eliminated with the use of appropriate standard NMR solvent suppression methods. Absolute concentrations can be obtained if the sample contains an added internal standard of known concentration, if a standard addition of the analyte of interest is added to the sample, or if the concentration of a substance is determined by independent means (e.g., glucose in plasma can be quantified by a conventional biochemical assay).

The ^{1}H NMR spectra of urine show thousands of sharp peaks from predominantly small-molecule metabolites, whereas spectra of blood plasma and serum show broad bands from protein and lipoprotein signals, with sharp peaks from small molecules superimposed thereon (15). Standard NMR pulse sequences, where the observed peak intensities are edited on the basis of molecular diffusion coefficients or NMR relaxation times (such as the Carr-Purcell-Meiboom-Gill spin-echo sequence), can be used to select only the contributions from macromolecules, or alternatively to select only the signals from the small molecule metabolites. A typical 950-MHz ^{1}H NMR spectrum of urine showing the degree of spectral complexity is given in **Figure 1**. Most of the major peaks shown in this image have now been assigned (16).

The development of cryogenic probes wherein the detector coil and preamplifier (but not the samples) are cooled to around 20° K has improved spectral signal-to-noise ratios by up to a factor of five by reducing the thermal noise from the electronics of the spectrometer. Conversely, because the NMR signal-to-noise ratio is proportional to the square root of the number of co-added scans, shorter (by up to a factor of 25) data acquisition times become possible for the same amount of sample. Using NMR spectroscopy of biofluids to detect the much less sensitive ^{13}C nuclei, which have a natural abundance of 1.1%, also becomes possible because of the increase in signal-to-noise ratio (17). This technology also makes the use of tissue-specific microdialysis samples more feasible (18).

Two-dimensional (2D) NMR spectroscopy is useful for increasing signal dispersion and for elucidating the connectivities between signals, thus helping to identify metabolites. These include the ^{1}H-^{1}H 2D J-resolved experiment, which attenuates the peaks from macromolecules and yields information on the multiplicity and coupling patterns of resonances. A projection of such a spectrum onto the chemical shift axis yields a fingerprint of peaks from only the most highly mobile small molecules, with all spin-coupling peak multiplicities removed. Other 2D experiments [e.g., correlation spectroscopy (COSY) and total correlation spectroscopy (TOCSY)] provide ^{1}H-^{1}H spin-spin coupling connectivities, providing information as to which hydrogens in a molecule are close in chemical bond terms. Use of other types of nuclei, such as naturally abundant ^{13}C or ^{15}N, or, where present, ^{31}P, can be important in helping to assign NMR peaks through inverse-detected heteronuclear correlation NMR experiments. In these experiments, the lower-sensitivity or less-abundant nucleus NMR spectrum (such as ^{13}C) is detected indirectly using the more sensitive or abundant nucleus (^{1}H) by utilizing spin-spin interactions such as the one-bond ^{13}C-^{1}H spin-spin coupling. These yield both ^{1}H and ^{13}C NMR chemical shifts of

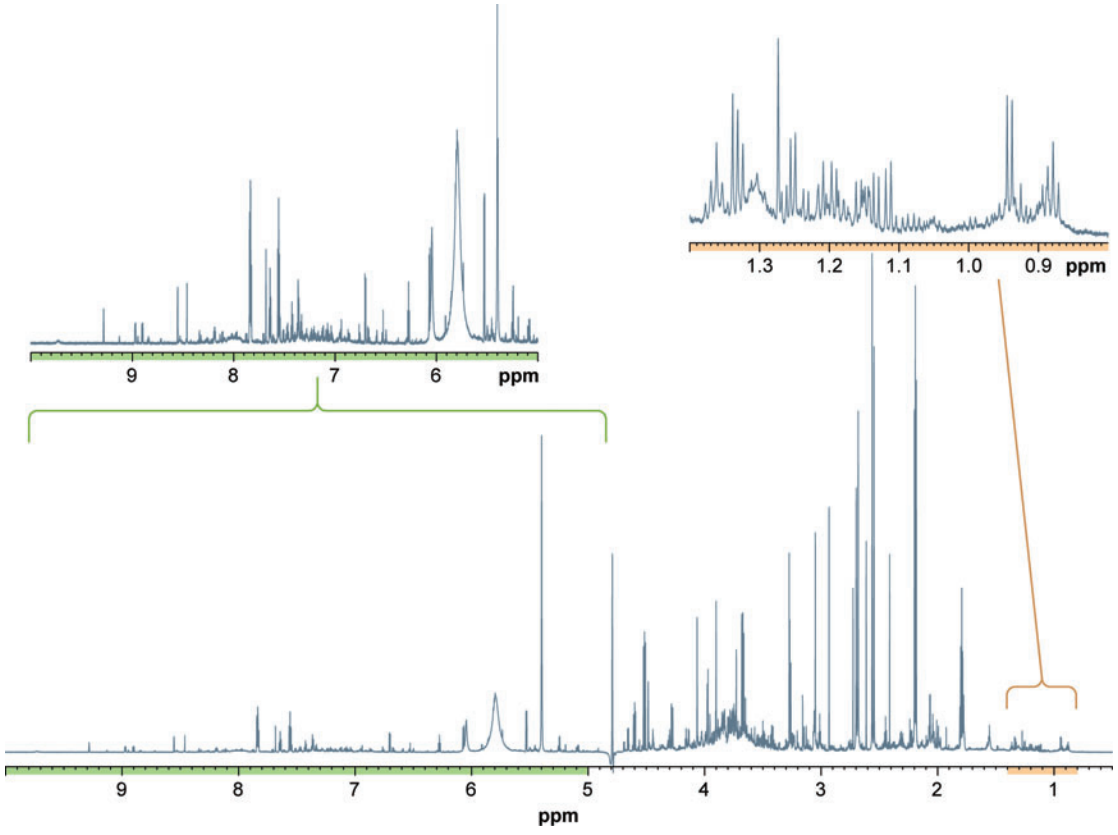

Figure 1

A 950-MHz ^{1}H nuclear magnetic resonance spectrum of human urine with expansions showing the degree of spectral complexity.

CH, CH_2, and CH_3 groups, which are useful for identification purposes (19). There is also a sequence that allows correlation of protons to quaternary carbons based on long-range ^{13}C-^{1}H spin-spin coupling.

Using a technique called high-resolution ^{1}H magic angle spinning (MAS) NMR spectroscopy, it is possible to acquire high-resolution NMR data on small pieces of intact tissues with no pretreatment (8, 20, 21). Rapid spinning of the sample (typically at ~4–6 kHz) at an angle of 54.7° relative to the applied magnetic field serves to reduce the loss of information caused by the line-broadening effects seen in nonliquid samples such as tissues (such effects are caused by sample heterogeneity and residual anisotropic NMR parameters that are normally averaged out in free solution where molecules can tumble isotropically and rapidly). MAS NMR spectroscopy requires straightforward, but manual, sample preparation. NMR spectroscopy on a tissue sample in an MAS experiment is the same as solution-state NMR, and all common-pulse techniques can be employed in order to study metabolic changes and to perform molecular structure elucidation and molecular dynamics studies.

2.3. Mass Spectrometry

HPLC: high-performance liquid chromatography

MS, including tandem MS methods, has also been widely used in metabolic fingerprinting and metabolite identification. Although most studies to date have been on plant extracts and model cell system extracts, the application of MS to mammalian studies is becoming more common. In general, a prior separation of the complex mixture sample using chromatography is required. MS is inherently much more sensitive than NMR spectroscopy, but it is generally necessary to employ different separation techniques (e.g., different LC column packings) for different classes of substances. Analyte quantitation by MS in complex mixtures of highly variable composition can be impaired by variable ionization and ion suppression effects. For plant metabolic studies, most investigations have used chemical derivatization to ensure volatility and analytical reproducibility, followed by GC-MS analysis. Some approaches using MS rely on more targeted studies, for example a detailed analysis of lipids (22).

For the application of metabonomics to biofluids such as urine, a high-performance liquid chromatography (HPLC) chromatogram is generated with MS detection, usually utilizing electrospray ionization, and both positive and negative ion chromatograms. Typically a time-of-flight (TOF) instrument is used, but other studies have used ion cyclotron resonance (also known as Fourier transform) MS (FT-MS). At each sampling point in the chromatogram, there is a full mass spectrum and so the data are three dimensional in nature (i.e., retention time, mass, and intensity).

The recently introduced technique known as UPLC is a combination of a 1.7-μm reversed-phase packing material and a chromatographic system operating at around 12,000 psi. UPLC provides approximately a ten-fold increase in speed and a three- to five-fold increase in sensitivity compared to a conventional stationary phase. Because of the much-improved chromatographic resolution of UPLC, the problem of ion suppression from co-eluting peaks is greatly reduced. A typical UPLC-MS chromatogram from a rat serum extract is shown in **Figure 2**; also shown for comparison is a chromatogram from a conventional HPLC-MS trace.

More recently, capillary electrophoresis coupled to mass spectrometry has also been explored as a suitable technology for metabonomics studies (23). Using this technique, metabolites are first separated by CE based on their charge and size and then selectively detected using MS monitoring. This method has been used to measure 352 metabolic standards and has been employed for the analysis of 1692 metabolites from *Bacillus subtilis* extracts, revealing changes in metabolite levels during the bacterial growth.

2.4. Hyphenated Systems

For biomarker identification, it is also possible to separate out substances of interest on a larger scale from a complex biofluid sample using techniques such as solid-phase-extraction or HPLC. For metabolite identification, directly coupled chromatography–NMR spectroscopy methods can also be used. The most general of these "hyphenated" approaches is HPLC-NMR-MS (24), in which the eluting HPLC peak is split, with parallel analysis by directly coupled NMR and MS techniques.

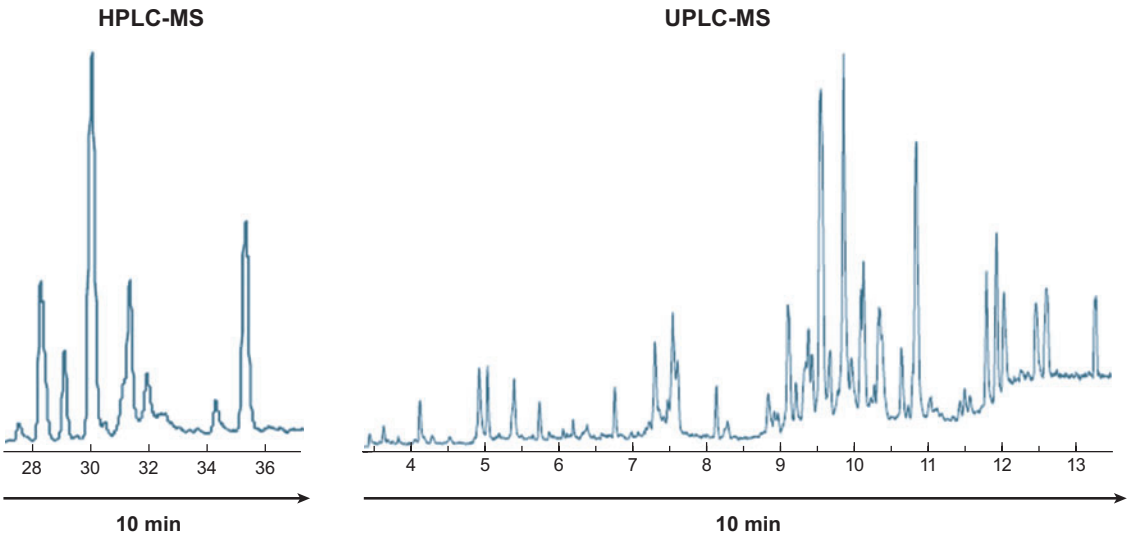

Figure 2

A comparison of high-performance liquid chromatography (HPLC)–mass spectrometry (MS) (*left*) and ultraperformance liquid chromatography (UPLC)–MS (*right*) chromatograms showing a 10-m segment derived from a methanolic extract of rat serum. Reproduced courtesy of E. Want, Imperial College London.

HPLC-NMR-MS can be operated in on-flow, stopped-flow, and loop-storage modes and thus can provide the full array of NMR- and MS-based molecular identification tools. These include 2D NMR spectroscopy as well as MS-MS for identification of fragment ions and FT-MS or TOF-MS for accurate mass measurement and hence derivation of molecular empirical formulae.

2.5. Chemometric Methods

To achieve optimal characterization of the samples and enable efficient biomarker detection, GC-MS and HPLC-MS data sets particularly, but also NMR data sets generally, require several steps of data preprocessing prior to any multivariate analysis. Considerable effort has gone into the development of algorithms to align signals from the same compound in data acquired from different samples (25, 26). This problem has been partially solved for NMR spectroscopic data (which can be subject to peak position variation problems) by dividing the spectrum into frequency windows and integrating signal intensity within these segments (27, 28), or by means of advanced principal components analysis (PCA) techniques (29). Analogous approaches have been applied to MS data. In one study, a series of prominent marker peaks (30) was used to divide chromatograms into time windows in which the retention time scale was then adjusted piecewise by an interpolation scheme. Another approach (31) divided chromatograms into time windows separated by regions showing a baseline response in all samples, constructed the equivalent direct-injection mass spectra,

and then, using alternating regression, extracted the significantly varying *m/z* values. For NMR spectroscopy, the need to segment the data in this way has largely been eliminated by the use of correlation techniques (see Section 4 below).

The data point intensities in an NMR spectrum or in an LC-MS data set can be considered as a multidimensional graph of metabolic coordinates; the spectrum is therefore a point in a multidimensional metabolic hyperspace. The initial objective in metabonomics is primarily to classify a sample based on identification of its detailed spectral patterns and then to identify those metabolic features responsible for the classification. The approach can also be used for reducing the dimensionality of complex data sets, for example by 2D or 3D mapping procedures, to enable easy visualization of any clustering of the various samples. Alternatively, in what are known as supervised methods, multiparametric data sets can be modeled so that the class of separate samples (known as a validation set) can be predicted based on a series of mathematical models derived from the original data (the training set).

One simple technique used extensively in metabonomics is PCA (32). Conversion of the data matrix to principal components results in two matrices known as scores and loadings. Scores, the linear combinations of the original variables, represent the coordinates for the samples in the established model and may be regarded as new latent variables. In a scores plot, each point represents a single sample spectrum. The PC loadings define the way in which the old variables are linearly combined to form the new variables and indicate those variables carrying the greatest weight in transforming the position of the original samples from the data matrix into their new position in the scores matrix. In the loadings plot, each point represents a different spectral intensity. Thus, the cause of any spectral clustering observed in a PC scores plot is interpreted by examination of the loadings that cause any cluster separation. In addition, there are many other visualization (i.e., unsupervised) methods, such as nonlinear mapping and hierarchical cluster analysis.

One widely used supervised method (i.e., using a training set of data with known endpoints) is partial least squares (PLS) (32). This method relates a data matrix containing independent variables from samples, such as spectral intensity values (an X matrix) to a matrix containing dependent variables, such as measurements of response, for those samples (a Y matrix). PLS can also be combined with discriminant analysis (DA) to establish the optimal position in which to place the discriminant surface that best separates classes. It is possible to use such supervised models to provide classification probabilities and quantitative response factors for a wide range of sample types, but given the strong possibility of chance correlations when the number of descriptors is large, it is important to build and test such chemometric models using independent training data and validation data sets. Orthogonal signal correction (OSC) can be used to remove irrelevant parts of the data that are uncorrelated with the endpoints; OSC has been integrated into the PLS algorithm for optimum use (33).

PCA and PLS use linear combinations of parameters for dimension reduction or classification, but other methods do exist. For example, in neural network analysis a training set of data is used to develop algorithms, which "learn" the structure of the data and can cope with complex functions. Recently, probabilistic neural networks, which represent an extension to the approach, have shown promise for metabonomics

applications in toxicity (34). Other approaches currently being tested include genetic algorithms (35), machine learning (36), and Bayesian modeling (37).

STOCSY: statistical total correlation spectroscopy

3. THE STRENGTHS AND WEAKNESSES OF NUCLEAR MAGNETIC RESONANCE AND MASS SPECTROMETRY FOR METABOLIC PROFILING

Although both NMR spectroscopy and mass spectrometry have been widely used in metabolic profiling studies, each has its champions and often the real benefits of each are not wholly appreciated. **Table 1** presents a comparison of these two highly complementary approaches.

4. THE CONCEPT OF STATISTICAL SPECTROSCOPY AND ITS IMPLEMENTATION

A recent development has been the implementation of the statistical total correlation spectroscopy (STOCSY) analysis method for aiding the identification of potential biomarker molecules in metabonomic studies based on NMR spectroscopic data (38). STOCSY takes advantage of the multicollinearity of the intensity variables in a set of spectra (e.g., ^{1}H NMR spectra) to generate a pseudo-two-dimensional NMR spectrum that displays the correlation among the intensities of the various peaks across the whole sample. This method is not limited to the usual connectivities that are deducible from more standard 2D NMR spectroscopic methods, such as TOCSY. Moreover, two or more molecules involved in the same pathway can also present high intermolecular correlations because of biological covariance due to common pathway control or can even be anticorrelated for the same reason.

A combination of STOCSY with supervised pattern recognition, particularly orthogonal projection on latent structure–discriminant analysis (OPLS-DA), offers a powerful new framework for the analysis of metabonomic data. First, OPLS-DA extracts the parts of NMR spectra related to discrimination. This information is then cross-combined with the STOCSY results to help identify the molecules responsible for the metabolic variation. This method has been applied to ^{1}H NMR spectra of urine from a metabonomic study of a model of insulin resistance based on the administration of a carbohydrate diet to three different mice strains, in which a series of metabolites of biological importance was conclusively assigned and identified (38).

The background to this method was introduced by Sasic et al. (39) and is based on another method, proposed by Noda, for generalized 2D correlation spectroscopy (40, 41, 42, 43). Successful previous applications of this correlative approach include infrared, Raman, near-infrared, and fluorescence spectroscopies (41, 42), with which correlations between different spectral features were identified.

STOCSY is based on the properties of the correlation matrix **C**, computed from a set of sample spectra according to where **X**1 and **X**2 denote the autoscaled experimental matrices of $n \times v1$ and $n \times v2$, respectively; n is the number of spectra (one for each sample) and $v1$ and $v2$ are the number of variables in the spectra for each matrix. **C** is therefore a matrix of $v1 \times v2$, where each value is a correlation

Table 1 The relative strengths and weaknesses of nuclear magnetic resonance and mass spectrometry for metabolic profiling[a]

	NMR	MS
Detection limits	Low-micromolar at typical observation frequencies (600 MHz), but nanomolar using cryoprobes	Picomolar with standard techniques, but can be much lower with special techniques
Universality of metabolite detection	If metabolite contains hydrogens it will be detected, assuming the concentration is sufficient or protein binding does not cause marked line broadening	Usually needs a more targeted approach. There can be problems with poor chromatographic separation; with the loss of metabolites in void volumes; with ion suppression (but this is reduced when using UPLC); lack of ionization; ability to run both +ve and −ve ion detection gives extra information
Sample handling	Whole sample analyzed in one measurement	Different LC packings and conditions for different classes of metabolite; usually samples have to be extracted into a suitable solvent; samples have to be aliquoted but some recent studies have avoided the need for chromatography
Amount of sample used	Typically 200–400 μL, but much less for microcoil probes, down to 5–10 μL	Low μL range
Sample recovery	Technique is nondestructive	Technique is destructive but only small amounts used
Analytical reproducibility	Very high	Fair
Sample prepreparation	Minimal: addition of buffer, D_2O and chemical shift reference (not always required)	Can be substantial; often needs different LC columns and protein precipitation
Ease of molecular identification	High, both from databases of authentic material and by self-consistent analysis of 1D and 2D spectra	Difficult, often only the molecular ion is available; this needs extra experiments, such as routine tandem MS; GC-MS is generally better with accurate retention times and comprehensive databases of spectra
Time to collect basic data	5 min for 1D ^{1}H NMR	10 min for UPLC-MS run
Quantitation	1–5%	5% intraday and interday is now common with or without prior chromatography
Robustness of instruments	High	Low
Molecular dynamics information	Yes, from T1, T2 relaxation time and diffusion coefficient measurements	No
Analysis of tissue samples	Yes, using MAS NMR	No
Availability of databases	Not yet comprehensive but increasing; several are available freely on the web; some commercial products also exist	Comprehensive databases for electron impact MS allow spectral comparisons; For electrospray ionization, as is usual in LC-MS, only mass values can be compared

[a]Abbreviations: GC, gas chromatography; LC, liquid chromatography; MAS, magic angle spinning; MS, mass spectrometry; NMR, nuclear magnetic resonance; UPLC; ultraperformance liquid chromatography.

coefficient between two variables of the matrices **X**1 and **X**2. The simplest case is the autocorrelation analysis where **X**1 = **X**2. Because the different resonance intensities from a single molecule will always have the same ratio, if the spectrometer conditions are kept identical between samples the relative intensities will theoretically be totally correlated (correlation coefficient $r = 1$). In real samples of biofluids, r will always be <1 because of spectral noise or peak overlaps from other molecules. In practice, however, the correlation matrix from a set of spectra containing different amounts of the same molecule shows very high correlations between the variables corresponding to the resonances of the same molecule. Plotting the correlation matrix provides a graphic representation of the multi-sample spectroscopic data set comparable to that of a 2D correlation NMR experiment conducted on one sample containing all the molecules of all the samples.

The method is not restricted to ^{1}H-^{1}H NMR correlation, but can be applied to different nuclei. If these include different NMR-active nuclei (^{13}C-^{13}C, ^{1}H-^{13}C, ^{13}C-^{31}P, etc.), then heteronuclear correlation is also possible and will yield novel molecular connectivity information using both types of nuclear spin properties. We have recently applied this approach to correlation of ^{1}H and ^{31}P MAS NMR spectra of tissues (44). An illustration of the correlation between ^{1}H and ^{31}P MAS NMR spectra of rat liver tissues, following administration of a model liver toxin galactosamine, is shown in **Figure 3** (44). The one-dimensional ^{1}H and ^{31}P correlation plots constructed from the two-dimensional STOCSY cross-peak at heteronuclear chemical shifts of δ_{31P} 12.772 and δ_{1H} 5.9822 are given in **Figure 3*a,b,c***. The two high-field ^{31}P resonances correlate to a host of ^{1}H resonances and the ^{31}P spectral metabolites have been assigned to UDP-galNAc and UDP-glcNAc. These STOCSY-led assignments have been confirmed via spiking of standard compounds into an aqueous extract of liver.

Finally, it should be noted that STOCSY can be used to derive NMR spectral splittings and J couplings with the same theoretical precision as the 1D spectral properties from which the 2D data set was derived; this derivation is not limited by the generally lower resolution in the $F1$ domain of most correlation 2D experiments, provided, of course, that any physicochemical environment variation between samples does not induce variation of the peak positions.

An approach to enhancing information recovery from cryogenic probe on-flow LC-NMR spectroscopic analyses of complex biological mixtures has been demonstrated using a variation on the STOCSY method (45). Cryoflow probe technology enables more sensitive and hence faster NMR detection of metabolites using on-flow HPLC-NMR or UPLC-NMR, and the rapid spectral scanning allows multiple spectra to be collected over unresolved chromatographic peaks that can contain several species with similar, but nonidentical, retention times. This enables the identification of ^{1}H NMR signal connectivities between close-eluting metabolites, resulting in a "virtual" chromatographic resolution enhancement visualized directly in the NMR spectral projection. This approach is of wide general applicability to any complex mixture analysis problem involving chromatographic peak overlap, with particular application in metabolomics and metabonomics for identifying both endogenous and xenobiotic metabolites.

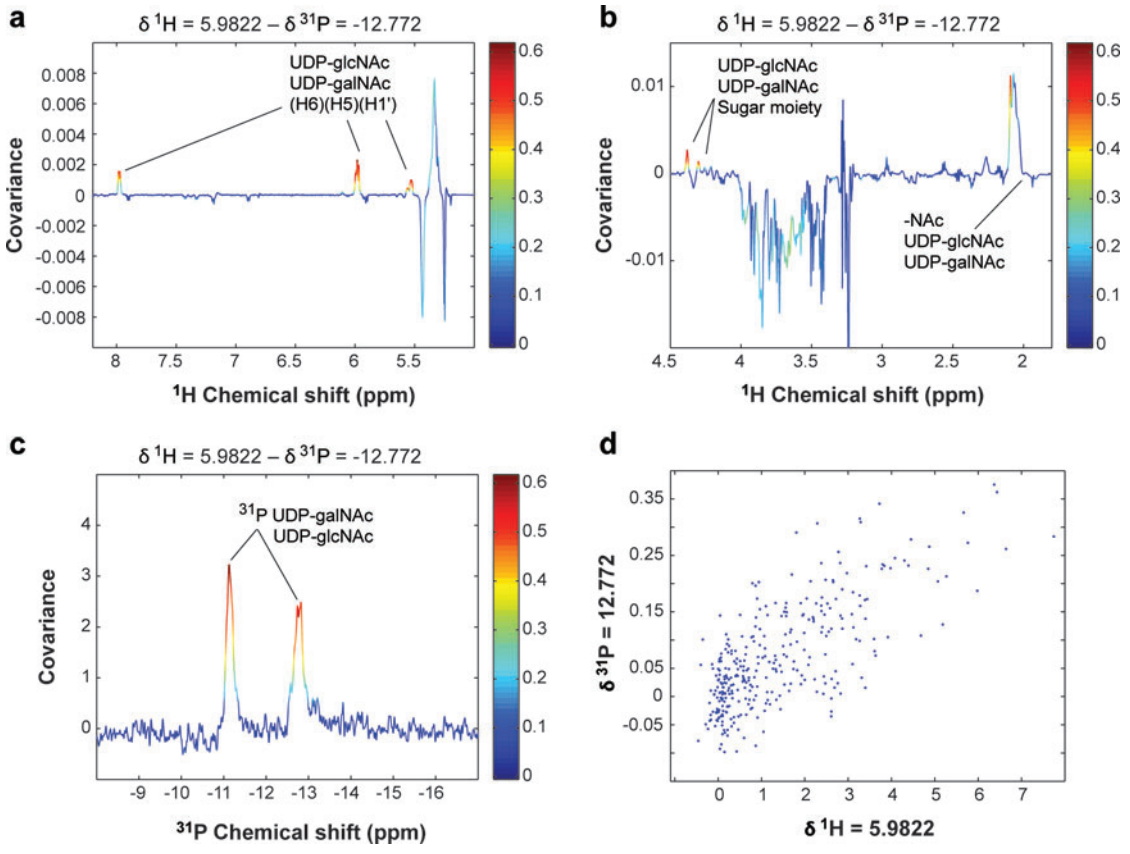

Figure 3

The use of statistical total correlation spectroscopy (STOCSY) to aid identification of linked nuclear magnetic resonance (NMR) peaks from ^{1}H and ^{31}P magic-angle-spinning NMR spectra of rat liver tissue. The ^{1}H and ^{31}P NMR chemical shifts for the anomeric proton and one of the two phosphorus atoms, respectively, are around 6 ppm and −12 ppm for both UDP-GlcNac and UDP-GalNac. The correlations from the peak pair at 5.982 and −12.77 ppm to aromatic protons are shown in panel *a*, and the correlation to aliphatic protons is shown in panel *b*. Panel *c* indicates that the intensities of the two high-field ^{31}P NMR peaks are highly correlated as expected, as both UDP-GlcNac and UDP-GalNac have two phosphorus atoms and the ^{31}P chemical shifts of these are at the same positions in both substances. The correlation of the intensities of the peaks at 5.98 and −12 ppm is given in panel *d*.

The complementarity of NMR and MS methods as structural tools has encouraged their parallel use in structure elucidation studies for natural product research, drug metabolite analysis, and other complex mixture analysis problems for many years. Typically, NMR and MS spectra of various types are examined together, and structural parameters such as chemical shifts and coupling constants (NMR) and exact molecular mass and fragmentation patterns (MS) are compared, often for a single

sample, to generate structural assignment information that is consistent with the outputs of both technologies. However, in many metabonomic studies, multiple samples with a wide range of biochemical variation are available for both NMR and MS analysis, creating the opportunity for statistical analysis of signal amplitude covariation between technologies and direct cross-correlation of data for assignment purposes.

SHY: statistical heterospectroscopy

Statistical heterospectroscopy (SHY) is an extension of STOCSY for the coanalysis of multispectroscopic data sets acquired on multiple samples (46). This method also operates through the analysis of the intrinsic covariance between signal intensities in the same and related molecules measured by different techniques across cohorts of samples. The potential of SHY has been illustrated using both 600 MHz ^{1}H NMR and UPLC-TOF-MS data obtained from control rat urine samples and from a corresponding group treated with hydrazine, a model liver toxin (46). We have shown that direct cross-correlation of spectral parameters, e.g., chemical shifts from NMR and *m/z* data from MS, is readily achievable for a variety of metabolites, leading to improved efficiency of molecular biomarker identification. In addition to structural information, higher-level biological information can be obtained on metabolic pathway activity and connectivities by examination of different levels of the NMR-to-MS correlation and anticorrelation matrixes. The SHY approach can be used if two or more independent spectroscopic data sets are available for any sample cohort. This approach is of wide applicability and can be extended beyond NMR and LC-MS to include any spectroscopic, electrochemical, or other multivariate analytical measurements where multiple samples are measured by more than one technology. An example of the use of SHY to correlate NMR and MS data is given in **Figure 4** (46).

5. A DATA SPACE REDUCTION APPROACH FOR IMPROVING THE SEARCH FOR BIOMARKERS

One of the major problems with interpreting metabonomic UPLC-MS data sets is the sheer number of variables to consider. A strategy for biomarker recovery from such data sets from biofluids has been presented and exemplified using a study on hydrazine-induced liver toxicity (47). A key step in this strategy involves a novel procedure for reducing the spectroscopic search space by differential analysis of cohorts of normal and pathological samples using orthogonal projection-to-latent structures discriminant analysis (O-PLS-DA). This efficiently sorts principal discriminators of toxicity from the background of thousands of metabolic features commonly observed in the data sets generated by UPLC-MS analysis of biological fluids; it is thus a powerful tool for biomarker discovery.

The advent of UPLC has led to the detection of many thousands of metabolic features by mass spectrometry. Paradoxically, this creates a new statistical problem related to biomarker discovery: it becomes difficult to recover key discriminating biomarker information in the background of the thousands of detectable metabolic variables, the majority of which do not actually contribute to class separation or diagnosis. We have approached the processing of UPLC-MS data acquired on urine samples from an exemplar toxicology study by applying O-PLS-DA to matrices formed by binning the raw data over windows in both retention time and *m/z*. The sample

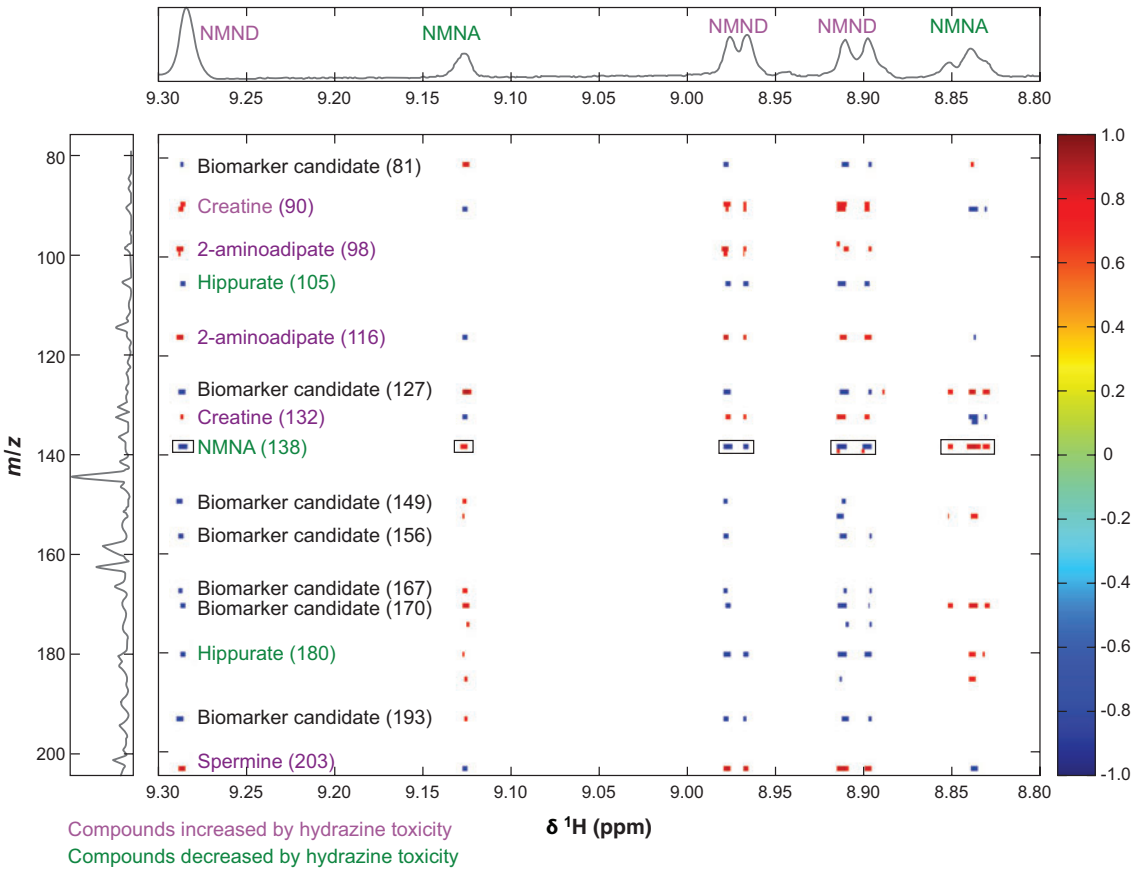

Figure 4

The use of statistical heterospectroscopy to correlate ^{1}H nuclear magnetic resonance (NMR) spectral data with mass spectrometry (MS) *m/z* values. NMR-MS correlation for urine data from a group of rats that had been treated with the liver toxin hydrazine, expanded to show an N-methylnicotinic acid (NMNA)/N-methylnicotinamide (NMND) region of the ^{1}H NMR spectrum. The insets show mean NMR and mass spectra. All identified ions (with their *m/z* values in paretheses) show directions of correlation consistent with known effects of hydrazine. For example, NMNA and NMND are negatively correlated, as they are transaminase related (inhibited by hydrazine). Also, NMNA correlates positively with itself, as must be the case. These NMNA/NMND correlations are shown boxed. The newly identified spermine ion is shown to correlate positively with toxicity. All unidentified ions (also with their *m/z* values in paretheses) are also candidate biomarkers.

set analyzed contained urine specimens from both control animals and from those exhibiting a strong response to administration of the model toxin hydrazine, which elicits a steatotic effect in the liver (48). The aim of the study was to demonstrate the stratification of the many UPLC-MS metabolite peaks according to their relevance to toxicity (and, in principle, to any other pathological or disease state). In this way, subsequent metabolite identification and quantitation efforts were focused on the

most important peaks. Use of this approach efficiently avoids the vexed problems of deconvolution or peak alignment between samples and the calculation of relative concentrations before applying the multivariate analysis, and allows significant peaks that are close to the detection limit to be recovered statistically without time-consuming inspection of individual spectra. We have shown that, given a suitable choice of window size, useful peak selection can be obtained rapidly and the peaks can be treated as biomarker candidates for further analysis and structure characterization.

The use of a cross-validated method like O-PLS-DA provides greater confidence than simple correlation analysis of variable against class. In addition, the method detects and models separately any variation that is orthogonal to class, so this can be separately analyzed and does not affect the class dependency results. The analytical strategy employed is summarized in **Figure 5**. Following the binning of data comes the selection step, which consists of filtration of variables by weight in an O-PLS-DA model; this yields a measure of correlation between peak intensity and dosage. The weights, color coded by value, are displayed in 2D diagrams labeled by retention time bin and *m*/*z* bin, which are expanded as required. Once the discriminating peaks are found, a library search is conducted to identify the possible compounds involved. Given that a single bin may contain more than one peak and that accurate masses and retention times are required, it is still necessary to interrogate the

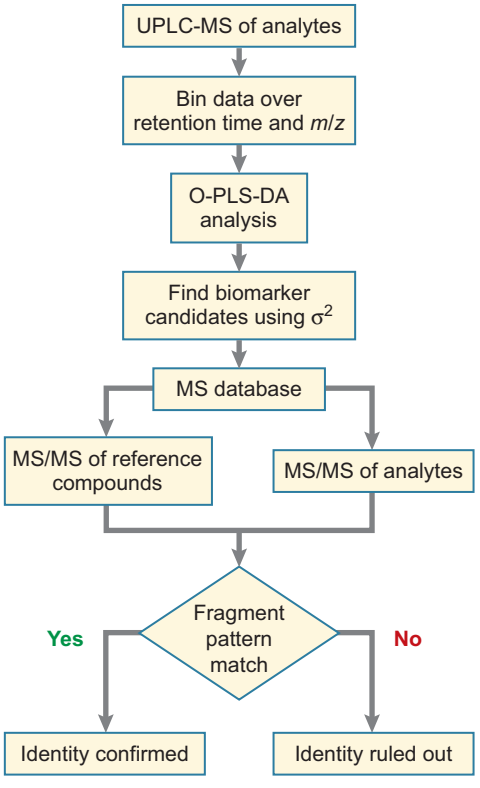

Figure 5

The procedure for data space reduction in the search for biomarkers separating two classes of sample.

original UPLC-MS spectra to find the exact values for the biomarker candidates. In this study (47), the library search against accurate *m/z* was done manually for the list of peaks of interest; however, this search may be automated. The final step compares MS/MS fragment patterns of pure compounds with urine samples from the study to either confirm or exclude candidate compounds highlighted by the search procedure.

6. RECENT MAMMALIAN APPLICATIONS AND TRENDS

Numerous studies have used metabonomics to characterize normal metabolic variation, caused by a range of inherent and external factors, in experimental animals such as mice and rats (49 and references therein). Such differences may help explain differential toxicity of drugs between strains and interanimal variation within a study. Many effects can be distinguished using NMR- and MS-based metabonomics, including male/female differences, wild-type and genetically modified animal models, age-related changes, estrus cycle effects in females, diet, diurnal effects, and interspecies differences and similarities (8, 49). Analogous studies have also been undertaken in humans (50, 51, 52). The importance of the symbiotic relationship between mammals and their gut microfloral populations has been recognized (9) and studied extensively using axenic animals (53), different animal colonies (54), parasitic infections in animals (55), and probiotics (56).

Minimizing the occurrence of drug adverse effects is one of the most important aims of pharmaceutical research and development, and the pharmaceutical industry is now embracing metabonomics for evaluating the adverse effects of candidate drugs (e.g., through the COMET project; see below). Metabonomics classification of the target organ or region of toxicity, the biochemical mechanism of that toxin, the identification of combination biomarkers of toxic effect, and evaluation of the time-course of the effect (e.g., the onset, evolution, and regression of toxicity) can all be determined (57, 58). The usefulness of metabonomics for the evaluation of xenobiotic toxicity effects has recently been comprehensively explored by the Consortium for Metabonomic Toxicology (COMET). COMET was formed among five pharmaceutical companies and Imperial College London (59) with the aim of developing methodologies for the acquisition and evaluation of metabonomic data generated using ^{1}H NMR spectroscopy of rat and mouse urine and blood serum for preclinical toxicological screening of candidate drugs. New methodologies for analyzing and classifying the complex data sets were developed. For example, because the predictive expert system takes into account the metabolic trajectory over time, a new way of comparing and scaling these multivariate trajectories was developed (60). Additionally, a novel classification method, termed Classification Of Unknowns by Density Superposition (CLOUDS), has been generated in order to identify the class of toxicity based on all of the NMR data for a given study (61).

COMET showed that it is possible to construct predictive and informative models of toxicity using NMR-based metabonomic data, delineating the whole time course of toxicity. This successful outcome is evidenced by the generated databases of spectral and conventional results for a wide range of model toxins (147 in total) that served as

the basis for computer-based expert systems for toxicity prediction (62). The project's goals, namely the generation of comprehensive metabonomic databases (now containing approximately 35,000 × 600 MHz ^{1}H NMR spectra) and the creation of successful, robust multivariate statistical models (expert systems) for prediction of toxicity—initially for liver and kidney toxicity in the rat and mouse—have now been achieved, and the predictive systems and databases have been transferred to the sponsoring companies (63).

Many examples exist in the literature (reviewed in 8) of the use of NMR-based metabolic profiling to aid human disease diagnosis. A promising use of NMR spectroscopy of urine and plasma, as evidenced by the number of publications on the subject (8, chapter 14), is in the diagnosis of inborn errors of metabolism in children. In addition, tissues themselves can be studied by metabonomics through the MAS technique and there are many published examples (8, chapter 4, and references therein). In this field, MS is also a well-established technique and many diseases are routinely diagnosed using MS-based methods, but investigation of rare and unexpected diseases often benefits from an exploratory NMR spectroscopic approach.

The value of obtaining multiple NMR spectroscopic (or indeed other types of analysis) data sets from various biofluid samples and tissues of the same animals collected at different time-points has been demonstrated. This procedure, termed integrated metabonomics (6), can be used to describe the changes in metabolism in different body compartments affected by exposure to, for example, toxic drugs (64, 65). Such timed profiles in multiple compartments are themselves characteristic of particular types and mechanisms of pathology and can be used to give a more complete description of the biochemical consequences than can be obtained from one fluid or tissue alone.

Integration of metabonomic data with data from multivariate techniques in molecular biology, such as gene array experiments or proteomics, is also feasible. Thus, it is also possible to integrate data from transcriptomics and metabonomics in order to find, for example after acetaminophen administration to mice, common metabolic pathways implicated by both gene expression changes and changes in metabolism (66). It has also been found that evaluation of both transcriptomic and metabolic changes following administration of the toxin bromobenzene provides a more sensitive approach for detecting the effects of the toxin (67). Similarly, changes in gene expression detected in microarray experiments can lead to the identification of changed enzyme activity; this can also be achieved by analysis of metabolic perturbations (68).

A novel method of integrating proteomic and metabonomic data has been developed and applied to a human tumor xenograft mouse model of prostate cancer (69). Parallel 2D–difference gel electrophoresis proteomic data and ^{1}H NMR metabolic profile data were collected on blood plasma from mice implanted with a prostate cancer xenograft and from matched control animals. To interpret the xenograft-induced differences in proteomic and metabonomic plasma profiles, multivariate statistical algorithms, including OPLS, were applied to generate models characterizing the disease profile and to elucidate metabolite concentrations and protein abundances that could be directly related to the disease model.

Clearly, characterizing the relationships between genomic and phenotypic variation is an essential step in understanding disease processes. To this end, the first real transcriptomic-metabonomic cross-correlations have been achieved (70). Data sets derived from pathophysiological, proteomic, and transcriptomic profiling were used to map so-called quantitative trait loci (QTLs). Metabolic traits, as used in plant QTL studies, can be used to define phenotypes in mammalian genetics to characterize disease biomarkers. Untargeted plasma metabolic fingerprints derived from NMR spectroscopic analysis were mapped to chromosomes found in a rat strain derived from crossing diabetic and control animals. Identifying such metabotypic QTLs appears to be a practical approach to understanding genome-phenotype relationships in mammals and might uncover deeper biological complexity, such as extended genome (microbiome) perturbations that could affect disease processes through transgenomic effects.

7. THE FUTURE

Developments in NMR spectroscopy will continue to provide improvements in sensitivity from the use of cryogenic probes and increased magnetic field strengths. Improved identification of metabolites will come from inspection of databases of NMR spectra of standard substances. An increasing use of MS-based analyses in mammalian systems is expected and improvements in detection, sensitivity, and reliability will undoubtedly be achieved. Given the need for data sets to provide information on molecular identity, it is unlikely that techniques other than NMR and MS will find widespread use.

The main pharmaceutical areas in which metabonomics is already having an impact include validation of animal models of disease (as in genetically modified animals); preclinical evaluation of candidate drugs in safety studies; assessment of safety in humans in clinical trials, and, after product launch, quantitation or ranking of the beneficial effects of pharmaceuticals; improved understanding of the causes of highly sporadic idiosyncratic toxicity of marketed drugs; and patient stratification for clinical trials and drug treatment. In addition, in terms of disease studies, metabonomics is playing a role in improved, differential diagnosis and prognosis of human diseases, particularly for chronic and degenerative diseases and for diseases caused by genetic effects. A better understanding of large-scale human population differences through epidemiological studies is also being achieved. Other applications where major expansion is expected are nutritional studies; sports medicine and lifestyle studies, including the effects of diet, exercise, and stress; and evaluation of the effects of interactions among drugs and between drugs and diet.

If personalized healthcare is to become a reality, an individual's drug treatments must be tailored so as to achieve maximal efficacy and avoid adverse drug reactions. One of the long-term goals of pharmacogenomics is to understand the genetic makeup of different individuals (their genetic polymorphisms) and how well they are able to handle pharmaceuticals—this is important for identifying both beneficial and adverse effects. An alternative approach to understanding intersubject variability in response to drug treatment involves using a combination of multivariate metabolic profiling

and chemometrics to predict the metabolism and toxicity of a dosed substance, based solely on the analysis and modeling of a predose metabolic profile (71). Unlike pharmacogenomics, this approach, termed pharmacometabonomics, is sensitive to both the genetic and environmental influences that determine the basal metabolic fingerprint of an individual, as these will also influence the outcome of a chemical intervention. This new approach has been illustrated with studies of the toxicity and metabolism of compounds, administered to rats, with very different modes of action. The next challenge is to adapt pharmacometabonomic approaches to humans, and although we are more metabolically variable than rats, so are our range of responses, meaning that pharmacometabonomics is likely to be successful in predictive drug metabolism and toxicity for at least some compound classes. Analogous to this application to pharmaceuticals, it has very recently been shown to be possible to distinguish individuals with different food preferences from their basal metabolic profiles of urine and blood plasma (72).

A major initiative is under way to make consensual recommendations for standardizing reporting arrangements for metabonomics studies. This research was initiated by the Standard Metabolic Reporting Structures group (**http://www.smrsgroup.org**), which produced a draft policy document covering all of the aspects of a metabolic study that are recommended for recording, including the origin of a biological sample, the analysis of material from that sample and chemometric and statistical approaches, and retrieval of information from the sample data; a summary publication has appeared (73). The various levels of (and consequent detail for) reporting needs, including journal submissions, public databases, and regulatory submissions, have also been addressed. In parallel, a scheme called ArMet for capturing data and metadata from metabolic studies has been proposed and developed (74). This has been followed up with a workshop and discussion meeting sponsored by the U.S. National Institutes of Health (75), from which the Metabolomics Standards Initiative was born (see **http://msi-workgroups.sourceforge.net**). This organization has now produced draft reports on many aspects of standardization and reporting in the subject and a number of papers have been published covering characterization of sample-related metadata, technical standards and related data, metadata and QC matters for the analytical instrumentation, data transfer methodologies, schema for the implementation of such activities, and development of standard vocabularies to enable transparent exchange of data (76).

8. CHALLENGES AND OPPORTUNITIES

In summary, it is clear that metabonomics will have an increasing and major impact. The analytical procedures used are stable and robust and have a high degree of reproducibility. Although advances will obviously be made in the future, current data will always be readable and interpretable. In contrast to other -omics, metabonomics enjoys a good level of biological reproducibility and the cost per sample and per analyte is low. It has the advantage of not having to preselect analytes, and through use of biofluids it is minimally invasive, creating the possibility of hypothesis-generation studies. Metabolic biomarkers are identifiable with real

biological endpoints and provide a global systems interpretation of biological effects, including interactions between multiple genomes (such as those of humans and their gut microflora). One major potential strength of metabonomics is the possibility that metabolic biomarkers will be more easily used across species than transcriptomic or proteomic biomarkers; this flexibility should be important for pharmaceutical studies.

On the other hand, metabonomics suffers from the application of multiple analytical technologies, there are questions of the sensitivity and dynamic range of the technologies used, and the data sets are complex. Through the use of chemometrics, it is possible to overinterpret the data, but this is easily avoided by applying correct statistical rigor. Currently, the groups using metabonomics are moving towards defining standards for data and operations; a good start has been made, but there remains a need for the regulatory agencies to be trained in the data interpretation and for better-trained practitioners.

The reality of the complexity of disease and drug effects means that the use of biomarker combinations will increase; thus, there are many opportunities for metabonomics that have yet to be explored, such as its use in environmental toxicity studies, in directing the timing of transcriptomic and proteomic experiments, and for deriving theranostic biomarkers. Metabonomics will surely be an integral part of any multiomics study where all the data sets are combined in order to derive an optimum set of biomarkers.

The ultimate goal of systems biology is the integration of data acquired from living organisms at the gene, protein, and metabolite levels. Hopefully, through the combination of transcriptomics, proteomics, and metabonomics, an improved understanding of an organism's total biology will result; subsequently, better understanding of the causes and progression of human diseases will ensue.

DISCLOSURE STATEMENT

The authors are not aware of any biases that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. Nicholson JK, Wilson ID. 1989. High-resolution proton magnetic resonance spectroscopy of biological fluids. *Prog. NMR Spectrosc.* 21:449–501
2. van der Greef J, Tas AC, Bouwman J, Ten Noever de Brauw MC, Schreurs WHP. 1983. Evaluation of field-desorption and fast atom–bombardment mass-spectrometric profiles by pattern-recognition techniques. *Anal. Chim. Acta* 150:45–52
3. Gartland KPR, Sanins SM, Nicholson JK, Sweatman BC, Beddell CR, Lindon JC. 1990. Pattern recognition analysis of high resolution ^{1}H NMR spectra of urine: a nonlinear mapping approach to the classification of toxicological data. *NMR Biomed.* 3:166–72

4. Gartland KPR, Beddell CR, Lindon JC, Nicholson JK. 1991. The application of pattern recognition methods to the analysis and classification of toxicological data derived from proton NMR spectroscopy of urine. *Mol. Pharmacol.* 39:629–42
5. Nicholson JK, Lindon JC, Holmes E. 1999. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 29:1181–89
6. Nicholson JK, Connelly J, Lindon JC, Holmes E. 2002. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Disc.* 1:153–62
7. Fiehn O. 2002. Metabolomics: the link between genotypes and phenotypes. *Plant Mol. Biol.* 48:155–71
8. Lindon JC, Nicholson JK, Holmes E, eds. 2007. *The Handbook of Metabonomics and Metabolomics*. Amsterdam: Elsevier
9. Nicholson JK, Wilson ID. 2003. Understanding 'global' systems biology: metabonomics and the continuum of metabolism. *Nat. Rev. Drug Disc.* 2:668–76
10. Nicholson JK, Holmes E, Lindon JC, Wilson ID. 2004. The challenges of modeling mammalian biocomplexity. *Nat. Biotech.* 22:1268–74
11. Lamers RJAN, Wessels ECHH, van der Sandt JJM, Venema K, Schaafsma G, et al. 2003. A pilot study to investigate effects of inulin on Caco-2 cells through in vitro metabolic fingerprinting. *J. Nutr.* 133:3080–84
12. Villas-Boas SG, Hojer-Pedersen J, Akesson M, Smedsgaard J, Nielsen J. 2005. Global metabolite analysis of yeast: evaluation of sample preparation methods. *Yeast* 22:1155–69
13. Bollard ME, Xu JS, Purcell W, Griffin JL, Quirk C, et al. 2002. Metabolic profiling of the effects of D-galactosamine in liver spheroids using ^{1}H NMR and MAS-NMR spectroscopy. *Chem. Res. Toxicol.* 15:1351–59
14. Liu M, Nicholson JK, Lindon JC. 1996. High resolution diffusion- and relaxation-edited one- and two-dimensional ^{1}H NMR spectroscopy of biological fluids. *Anal. Chem.* 68:3370–76
15. Nicholson JK, Foxall PJD, Spraul M, Farrant RD, Lindon JC. 1995. 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma. *Anal. Chem.* 67:793–811
16. Lindon JC, Nicholson JK, Everett JR. 1999. NMR spectroscopy of biofluids. In *Annual Reports on NMR Spectroscopy*, volume 38, ed. GA Webb, pp. 1–88
17. Keun HC, Beckonert O, Griffin JL, Richter C, Moskau D, et al. 2002. Cryogenic probe ^{13}C NMR spectroscopy of urine for metabonomic studies. *Anal. Chem.* 74:4588–93
18. Price KE, Vandaveer SS, Lunte CE, Larive CK. 2005. Tissue targeted metabonomics: metabolic profiling by microdialysis sampling and microcoil NMR. *J. Pharmaceut. Biomed. Anal.* 38:904–9
19. Claridge TDW. 2004. *High Resolution NMR Techniques in Organic Chemistry*. Amsterdam, Netherlands: Elsevier
20. Garrod SL, Humpfer E, Spraul M, Connor SC, Polley S, et al. 1999. High resolution magic angle spinning ^{1}H NMR spectroscopic studies on intact rat renal cortex and medulla. *Magn. Reson. Med.* 41:1108–18

21. Cheng LL, Chang IW, Louis DN, Gonzalez RG. 1998. Correlation of high-resolution magic angle spinning proton magnetic resonance spectroscopy with histopathology of intact human brain tumor specimens. *Cancer Res.* 58:1825–32
22. Morris M, Watkins SM. 2005. Focused metabolomic profiling in the drug development process: advances from lipid profiling. *Curr. Opin. Chem. Biol.* 9:407–12
23. Soga T, Ohashi Y, Ueno Y, Nasaoka H, Tomita M, Nishioka T. 2003. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J. Proteome Res.* 2:488–94
24. Lindon JC, Nicholson JK, Wilson ID. 2000. Directly-coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development. *J. Chromatog. B* 748:233–58
25. Jonsson P, Johansson AI, Gullberg J, Trygg J, AJ, Grung B, et al. 2005. High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal. Chem.* 77:5635–42
26. Idborg-Bjorkman H, Edlund P-O, Kvalheim OM, Schuppe-Koistinen I, Jacobsson SP. 2003. Screening of biomarkers in rat urine using LC/electrospray ionization-MS and two-way data analysis. *Anal. Chem.* 75:4784–92
27. Farrant RD, Lindon JC, Rahr E, Sweatman BC. 1992. An automatic data reduction and transfer method to aid pattern-recognition analysis and classification of NMR spectra. *J. Pharm. Biomed. Anal.* 10:141–44
28. Holmes E, Foxall PJ, Nicholson JK, Neild GH, Brown SM, et al. 1994. Automatic data reduction and pattern-recognition methods for analysis of H-1 nuclear magnetic resonance spectra of human urine from normal and pathological states. *Anal. Biochem.* 220:284–96
29. Stoyanova R, Nicholls AW, Nicholson JK, Lindon JC, Brown TR. 2004. Automatic alignment of individual peaks in large high-resolution spectral data sets. *J. Magn. Reson.* 170:329–35
30. Yang J, Xu G, Zheng Y, Kong H, Wang C, et al. 2005. Strategy for metabonomics research based on high-performance liquid chromatography and liquid chromatography coupled with tandem mass spectrometry. *J. Chromatogr. A* 1084:214–21
31. Jonsson P, Bruce SJ, Moritz T, Trygg J, Sjostrom M, et al. 2005. Extraction, interpretation and validation of information for comparing samples in metabolic LC/MS data sets. *Analyst* 130:701–7
32. Lindon JC, Holmes E, Nicholson JK. 2001. Pattern recognition methods and applications in biomedical magnetic resonance. *Prog. NMR Spectrosc.* 39:1–40
33. Trygg J, Wold S. 2002. Orthogonal projections to latent structures (O-PLS). *J. Chemomet.* 16:119–28
34. Holmes E, Nicholson JK, Tranter G. 2001. Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks. *Chem. Res. Toxicol.* 14:182–91
35. Kell DB. 2002. Metabolomics and machine learning: explanatory analysis of complex metabolome data using genetic programming to produce simple, robust rules. *Mol. Biol. Rep.* 29:237–41
36. Kell DB. 2006. Metabolomics, modelling and machine learning in systems biology: towards an understanding of the languages of cells. *FEBS J.* 273:873–94

37. Vehtari A, Makinen V-P, Soininen P, Ingman P, Makela S, et al. 2007. A novel Bayesian approach to quantify clinical variables and to determine their spectroscopic counterparts in H-1 NMR metabonomic data. *BMC Bioinformatics* 8(Suppl. 2):S8
38. Cloarec O, Dumas ME, Craig A, Barton RH, Trygg J, et al. 2005. Statistical total correlation spectroscopy (STOCSY): a new approach for individual biomarker identification from metabonomic NMR datasets. *Anal. Chem.* 77:1282–89
39. Sasic S, Muzynski A, Ozaki Y. 2000. A new possibility of the generalized two-dimensional correlation spectroscopy. 1. Sample-sample correlation spectroscopy. *J. Phys. Chem. A* 104:6380–87
40. Noda I, Dowrey AE, Marcott C, Story GM, Ozaki Y. 2000. Generalized two-dimensional correlation spectroscopy. *Appl. Spectrosc.* 54:236A–248A
41. Noda I. 1989. Two-dimensional infrared-spectroscopy. *J. Am. Chem. Soc.* 111:8116–18
42. Noda I. 1993. Generalized 2-dimensional correlation method applicable to infrared, raman, and other types of spectroscopy. *Appl. Spectrosc.* 47:1329–36
43. Noda I. 1990. 2-dimensional infrared (2D IR) spectroscopy: theory and applications. *Appl. Spectrosc.* 44:550–61
44. Coen M, Hong Y-S, Cloarec O, Rhode C, Reily MD, et al. 2007. Magic angle spinning ^{31}P-^{1}H hetero-nuclear statistical total correlation spectroscopy (STOCSY). *J. Proteome Res.* 6:2711–19
45. Cloarec O, Campbell A, Tseng L-H, Braumann U, Spraul M, et al. 2007. Virtual chromatographic resolution enhancement in cryoflow LC-NMR experiments via statistical total correlation spectroscopy. *Anal. Chem.* 79:3304–11
46. Crockford DJ, Holmes E, Lindon JC, Plumb RS, Zirah S, et al. 2006. Statistical HeterospectroscopY (SHY), a new approach to the integrated analysis of NMR and UPLC-MS datasets: application in metabonomic toxicology studies. *Anal. Chem.* 78:363–71
47. Crockford DJ, Lindon JC, Cloarec O, Plumb RS, Bruce SJ, et al. 2006. Statistical search space reduction and 2-dimensional data display approaches for UPLC-MS in biomarker discovery and pathway analysis. *Anal. Chem.* 78:4398–408
48. Jenner AM, Timbrell JA. 1994. Effect of acute and repeated exposure to low-doses of hydrazine on hepatic-microsomal enzymes and biochemical parameters in-vivo. *Arch. Toxicol.* 68:240–45
49. Bollard ME, Stanley EG, Lindon JC, Nicholson JK, Holmes E. 2005. NMR-based metabonomics approaches for evaluating physiological influences on biofluid composition. *NMR Biomed.* 18:143–62
50. Dumas M-E, Maibaum EC, Teague C, Ueshima H, Zhou B, et al. 2006. Assessment of analytical reproducibility of ^{1}H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Anal. Chem.* 78:2199–208
51. Lenz EM, Bright J, Wilson ID, Hughes A, Morrison J, et al. 2004. Metabonomics, dietary influences and cultural differences: a H-1 NMR-based study of urine samples obtained from healthy British and Swedish subjects. *J. Pharmaceut. Biomed. Anal.* 36:841–49

52. Lenz EM, Bright J, Wilson ID, Morgan SR, Nash AFP. 2003. A H-1 NMR-based metabonomic study of urine and plasma samples obtained from healthy human subjects. *J. Pharmaceut. Biomed. Anal.* 33:1103–15
53. Nicholls AW, Mortishire-Smith RJ, Nicholson JK. 2003. NMR spectroscopic-based metabonomic studies of urinary metabolite variation in acclimatizing germ-free rats. *Chem. Res. Toxicol.* 16:1395–404
54. Robosky LC, Wells DF, Egnash LA, Manning ML, Reily MD, Robertson DG. 2005. Metabonomic identification of two distinct phenotypes in Sprague-Dawley (Crl: CD(SD)) rats. *Toxicol. Sci.* 87:277–84
55. Wang Y, Holmes E, Nicholson JK, Cloarec O, Chollet J, et al. 2004. Metabonomic investigations in mice infected with *Schistosoma* mansoni: an approach for biomarker identification. *Proc. Nat. Acad. Sci. USA* 101:12676–81
56. Martin FPJ, Wang Y, Sprenger N, Holmes E, Lindon JC, et al. 2007. Effects of probiotic *Lactobacillus paracasei* treatment on the host gut tissue metabolic profiles probed via magic-angle-spinning NMR spectroscopy. *J. Proteome Res.* 4:1471–81
57. Lindon JC, Holmes E, Bollard ME, Stanley EG, Nicholson JK. 2004. Metabonomics technologies and their applications in physiological monitoring, drug safety evaluation and disease diagnosis. *Biomarkers* 9:1–31
58. Lindon JC, Holmes E, Nicholson JK. 2004. Toxicological applications of magnetic resonance. *Prog. NMR Spectrosc.* 45:109–43
59. Lindon JC, Nicholson JK, Holmes E, Antti H, Bollard ME, et al. 2003. Contemporary issues in toxicology—the role of metabonomics in toxicology and its evaluation by the COMET project. *Toxicol. Appl. Pharmacol.* 187:137–46
60. Keun HC, Ebbels TMD, Bollard ME, Beckonert O, Antti H, et al. 2004. Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles. *Chem. Res. Toxicol.* 17:579–87
61. Ebbels T, Keun H, Beckonert O, Antti H, Bollard M, et al. 2003. Toxicity classification from metabonomic data using a density superposition approach: 'CLOUDS.' *Anal. Chim. Acta* 490:109–22
62. Ebbels TMD, Keun HC, Beckonert OP, Bollard ME, Lindon JC, et al 2003. Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The Consortium on Metabonomic Toxicology screening approach. *J. Proteome Res.* 6:3944–51
63. Lindon JC, Keun HC, Ebbels TMD, Pearce JMT, Holmes E, Nicholson JK. 2005. The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements. *Pharmacogenomics* 6:691–99
64. Waters NJ, Holmes E, Williams A, Waterfield CJ, Farrant RD, Nicholson JK. 2001. NMR and pattern recognition studies on the time-related metabolic effects of α-naphthylisothiocyanate on liver, urine, and plasma in the rat: an integrative metabonomic approach. *Chem. Res. Toxicol.* 14:1401–12
65. Coen M, Lenz EM, Nicholson JK, Wilson ID, Pognan F, Lindon JC. 2003. An integrated metabonomic investigation of acetaminophen toxicity in the mouse using NMR spectroscopy. *Chem. Res. Toxicol.* 16:295–303
66. Coen M, Ruepp SU, Lindon JC, Nicholson JK, Pognan F, et al. 2004. Integrated application of transcriptomics and metabonomics yields new insight into the

toxicity due to paracetamol in the mouse. *J. Pharmaceut. Biomed. Anal.* 35:93–105

67. Heijne WHM, Lamers RJAN, van Bladeren PJ, Groten JP, van Nesselrooij JHJ, Van Ommen B. 2005. Profiles of metabolites and gene expression in rats with chemically induced hepatic necrosis. *Toxicol. Pathol.* 33:425–33
68. Griffin JL, Bonney SA, Mann C, Hebbachi AM, Gibbons GF, et al. 2004. An integrated reverse functional genomic and metabolic approach to understanding orotic acid–induced fatty liver. *Physiol. Genomics* 17:140–49
69. Rantaleinen M, Cloarec O, Beckonert O, Wilson ID, Jackson D, et al. 2006. Statistically integrated metabonomic-proteomic studies on a human prostate cancer xenograft model in mice. *J. Proteome Res.* 5:2642–55
70. Dumas ME, Wilder SP, Bihoreau MT, Barton RH, Fearnside JF, et al. 2007. Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nat. Genetics* 39:666–72
71. Clayton TA, Lindon JC, Antti H, Charuel C, Hanton G, et al. 2006. Pharmacometabonomic phenotyping and personalised drug treatment. *Nature* 440:1073–77
72. Rezzi S, Ramadan Z, Martin F-PJ, Fay LB, van Bladeren P, et al. 2007. Human metabolic phenotypes link directly to specific dietary preferences in healthy individuals. *J. Proteome Res.* 6:4469–77
73. Lindon JC, Nicholson JK, Holmes E, Keun HC, Craig A, et al. 2005. Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotech.* 23:833–38
74. Jenkins H, Hardy N, Beckmann M, Draper J, Smith AR, et al. 2004. A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotech.* 22:1601–6
75. Castle AL, Fiehn O, Kaddurah-Daouk R, Lindon JC. 2006. Metabolomics standards workshop and the development of international standards for reporting metabolomics experimental results. *Briefing Bioinformatics* 7:159–65
76. Sumner LW, Amberg A, Barrett D, Beale MH, Daykin C, et al. 2007. Proposed minimum reporting standards for chemical analysis. *Metabolomics* 3:211–21