# Chapter 7

# Novel NMR and MS Approaches to Metabolomics

**Ian A. Lewis, Michael R. Shortreed, Adrian D. Hegeman, and John L. Markley**

## Abstract

Identifying and quantifying metabolites in complex biological samples is one of the most challenging aspects of metabolomics. Recently, several important advances in databases, software, instrumentation, and laboratory techniques have greatly simplified the most laborious tasks of metabolite identification and have made quantification more reliable. These technological advances have made bioanalytically oriented studies a feasible alternative to the statistics-based methods commonly used for metabolomics. We discuss the tools that have become most important in our own research and comment on emerging technologies that may play an important role in future studies. In addition, we provide practical guidelines for designing studies and give the step-by-step protocols used in our lab for sample preparation, metabolite identification, and accurate quantification of molecules.

**Key words:** Bioanalytical metabolomics, BMRB (BioMagResBank), extraction methodology, identification, informatics resources, mass spectrometery, MMCD (Madison Metabolomics Consortium Database), NMR spectroscopy, quantification, rNMR, software

## 1. Bioanalytical Metabolomics

All metabolomics studies involve elements of natural products chemistry, analytical chemistry, and statistics. The degree to which each of these elements is weighted influences experimental design and the type of data that is ultimately derived from a study. Currently, most metabolomics investigations emphasize statistics. Spectroscopic data derived from these studies are deciphered by using sophisticated multivariate tools, and potential biomarkers are identified on the basis of their statistical significance (1). Although multivariate analyses are effective for classifying samples, they do not provide a transparent mechanism for identifying and quantifying individual metabolites.

Teresa Whei-Mei Fan et al. (eds.), *The Handbook of Metabolomics*, Methods in Pharmacology and Toxicology, vol. 17,
DOI 10.1007/978-1-61779-618-0_7, 

Bioanalytical metabolomics is an emerging strategy that emphasizes comprehensive metabolite assignment, accurate measures of concentration, and transparent data analyses that minimize the use of statistics. Although these characteristics are similar to traditional metabolism research (2, 3), bioanalytical metabolomics differs from traditional studies in scope. Unfractionated biological samples contain hundreds (NMR) to thousands (MS) of observable signals. Whereas traditional methods restrict analyses to a few predetermined metabolites, bioanalytical metabolomics attempts to identify and quantify all of the observable signals. Until recently, this type of comprehensive metabolite profiling was too labor intensive to be practical in routine analyses. However, the tools and techniques discussed in this chapter have dramatically simplified the laborious aspects of data analysis and have made bioanalytical metabolomics a practical alternative to statistics-based studies.

## 2. Sample Preparation

Consistent sample preparation is an important component of bioanalytical metabolomics. Although a wide variety of techniques are effective (Chap. 2), no protocol is appropriate for all metabolites. Extraction conditions, such as solvent temperature and hydrophobicity, directly affect the molecules that can be identified and quantified in a study. Aqueous solvents extract hydrophilic molecules; nonpolar solvents extract hydrophobic metabolites; harsh conditions (acids, bases, and boiling) promote unwanted chemistry at labile functional groups; gentle conditions (cold methanol-water) are less effective for denaturing proteins. In short, every procedure alters one's perception of in vivo metabolism. The goal is to find a reproducible method that preserves the metabolites relevant to a particular study. In this section, we present some general guidelines for minimizing technical error and provide the sample preparation protocol we use for routine NMR-based metabolomics studies of aqueous metabolites. A more extensive discussion of sample preparation is found in Chap. 2.

### *2.1. Methods for Minimizing Technical Error*

Metabolite concentrations observed in tissue extracts and biological fluids can vary considerably across a dataset. These sample-to-sample differences originate from both natural variability and technical error. Whereas biological variance is essential for interpreting metabolic differences between samples, technical error is simply an obstacle to meaningful data analysis. Although there are many sources of error in metabolomics studies, the main contributors are inconsistent sample preparation and technical shortcomings in analytical equipment (e.g., resonance overlap, ion suppression, and imprecision in peak picking). Careful experimental design can control these sources of error.

*Extraction solution conditions.* Variation in the amount biological material used to prepare each sample is one of the primary sources of technical error in metabolomics. Inconsistency at this level is directly proportional to quantitative error in the final analysis. Sample-to-sample differences also affect solution conditions (osmolarity, pH, etc.) of the extraction buffer and analytical medium. These second-order effects alter the extraction efficiencies of sparingly soluble compounds and complicate data analysis.

A direct method for controlling variation in sample size is to aliquot tissues on the basis of mass. This approach is most effective in studies involving large sample sizes and becomes increasingly error-prone as sample size diminishes. Animal tissues can be weighed directly on an analytical balance whereas plant samples generally require prior lyophilization to standardize water content. An alternative strategy, which is appropriate for cell cultures, is to prepare samples with uniform optical density and aliquot samples on the basis of volume. We have found this method to be less consistent, but considerably more convenient, than centrifuging cell suspensions and aliquotting samples on the basis of mass.

The volume of solvent present in the extraction medium can also contribute to technical error, particularly in extractions involving volatile organic solvents or high temperatures. Solvent loss due to evaporation can be minimized by using sealed reaction vials. We have found 22-mL screw-top vessels to be convenient for parallel extractions; they allow 16 or more samples to be incubated, centrifuged, filtered and lyophilized simultaneously with conventional laboratory instrumentation.

*Preparation of analytical solutions.* Solution conditions in the analytical medium are another major contributor to technical error. Salinity, pH, and the concentrations of metal ions affect the sensitivity of NMR spectrometers and the efficacy bioinformatics-based resonance assignments. The strategy for standardizing these conditions is analogous to the methods used for standardizing extractions; samples need to be prepared with a consistent solute to solvent ratio.

The most reliable method for ensuring consistent analytical conditions is to analyze raw extracts without any additional sample preparation. Unfortunately, metabolites present in unconcentrated extracts are generally too dilute for NMR analyses. Furthermore, extractions typically employ $^{1}$H-containing solvents, which are another potential source of error in NMR analyses (e.g., spectral overlap, radiation damping, and receiver overflow resulting from incomplete solvent suppression). Consequently, NMR-based studies often prepare analytical solutions from dried extracts dissolved at relatively high concentrations in perdeuterated solvents. Though necessary, this strategy introduces some additional complications into sample preparation and data analysis.

We employ two strategies for preparing dried extracts: (1) dissolving all samples in a fixed volume of solvent and (2) preparing solutions on a mass to volume basis. Of the two approaches, the fixed volume method is preferable, because it requires the least amount of sample handling. In studies involving easily standardized extractions, the fixed volume approach yields technical error of less than 5% in our hands. The alternative mass-to-volume approach is appropriate for studies involving large samples (>40 mg extract) with major sample-to-sample differences in salinity. In these cases, we prepare analytical solutions from dried extracts dissolved in 8–17 μL of perdeuterated solvent per mg of sample.

*Titrating pH in analytical solutions.* Rigorous pH control is an essential component of bioanalytical metabolomics. All of the existing tools for bioinformatics-based metabolite identification and resonance assignment require samples to match the solution conditions used in the public databases (4). Minor deviations in pH (0.01) alter NMR chemical shifts and pH-dependent exchange broadening alters the intensities of metabolites that are near their p$K_a$. These effects can have a significant impact on the amount of time required to assign spectra and the reliability of quantities analyses.

In our experience, simply adding a buffer to the analytical medium (up to 25 mM HEPES) is insufficient for controlling pH to the degree required for automated resonance assignment. Each sample must be hand titrated using a pH meter equipped with a small electrode. We adjust sample pH using concentrated acid or base (~1 M DCl or ~1M NaOD) to minimize titration-related volume changes. For a trained technician, titrating 400 μL sample to 7.400 ± 0.004 requires about 3 min. Although titration is tedious, it saves considerable time overall because it allows labor-intensive data analysis to be replaced with automated resonance assignments.

*Internal standards.* The use of internal standards is one of the simplest means of ensuring data quality. Although highly reproducible data can be collected without an internal standard, small changes in NMR line shape, sample dilution, and salt concentration affect the intensities of NMR signals. Standards control for these variables and allow biologically relevant variation to be distinguished from technical error. Furthermore, clever use of internal standards allows one to calculate metabolite concentrations relative to the amount of starting material.

Our strategy is to extract 400 mg of dry weight tissue in 16 mL of water containing 167 μM of an internal standard. This approach allows us to relate metabolite concentrations observed in the NMR tube to the dry weight of the tissue. We typically use HEPES or MES as an internal standard because these compounds have multiple peaks that are well isolated from biological resonances in 2D

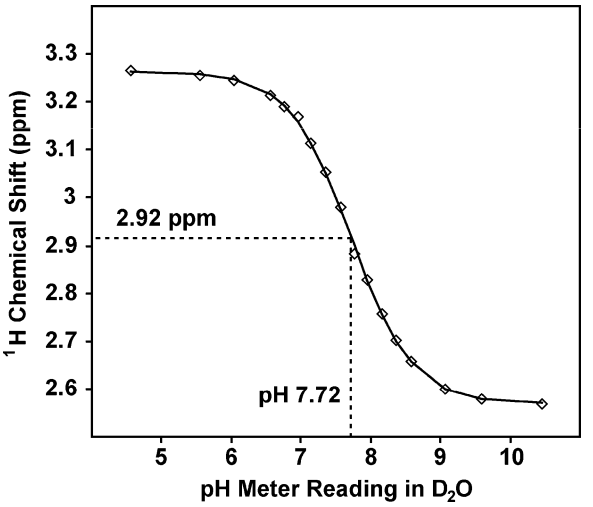


Fig. 1. The $^1$H chemical shifts of several HEPES peaks are pH sensitive and can be used as in internal pH reference. The titration curve was generated from a HEPES buffered saline solution (290 mOsm) in $D_2O$. The *dotted lines* indicate the observed p$K_a$ of HEPES as an uncorrected meter reading in $D_2O$.

$^1$H–$^{13}$C NMR spectra. HEPES is convenient because it acts as an internal pH indicator (Fig. 1), but MES is probably a better concentration reference because of its lower p$K_a$. Neither of these compounds would be suitable for 1D $^1$H or 2D homonuclear $^1$H experiments because their resonances overlap with those of many biological compounds.

Although we have only applied the internal standard normalization strategy to aqueous extractions, the principle should be applicable to other types of extractions. In some cases, biological internal standards may already be present in the data. Plants grown in MES buffer, for example, accumulate MES proportionally to their dry mass (5, 6). Non-metabolized biological internal standards are extremely useful provided that their normal in vivo concentrations can be calculated.

*Entropy in sample order.* Metabolite signals observed in either NMR or MS spectra can be influenced by the order of sample analysis. In the case of NMR data, systematic changes in shimming over the course of many samples results in different (usually worse) line shape. While these variations can, in theory, be corrected by normalization to an internal standard, data should always be collected in a random order to minimize systematic error. Furthermore, all NMR and MS data should be collected with technical replicates. Although technical replication is routine in the MS world, the NMR community has tended to avoid technical replication on the grounds that NMR yields high technical reproducibility. While NMR analysis can be made extremely reproducible

through careful shimming, we have found that data collected with automated sample changers can have considerable variation in line shape. Technical replication and randomized sample order minimize the chance of systematic errors.

*Errors in error bars.* Careful scrutiny of the various studies that have quantified the speed of light since 1676 indicates that the value of *c* appears to have changed significantly over the last few centuries (7). Although these results could be interpreted as an exciting physical phenomenon, the more realistic explanation is that the error bars given on the light speed estimates were too low. Misleading error estimates are not unique to measurements of universal constants. In a landmark 1984 paper, Stuart Hurlbert showed that about half of the inferential statistics published in ecology between 1960 and 1984 were based on questionable data replication (8). This problem also applies to metabolomics. Confusing technical replicates with biological replicates grossly misrepresents real biological variation and will inevitably lead to erroneous interpretations of statistical tests. Biologically meaningful investigations require careful experimental design with respect to replication. Although Hurlbert's paper was written for ecology, his recommendations for experimental design are directly applicable to the design of metabolomics experiments. Mehta et al. have also published a review on this topic that addresses the problems encountered in high-throughput studies (9).

### 2.2. Protocol for Extracting High-Abundance Aqueous Metabolites

The majority of research topics encountered in our laboratory involve studies of highly abundant water-soluble metabolites. The protocol presented here is a general method for preparing samples that is suited to NMR-based assays of amino acids, sugars, polyamines, polyols, and other thermally stable molecules found at high abundance in tissue extracts. We prefer this protocol because it is easy to parallelize, produces low technical error with relatively high yields, and results in a similar complement of NMR-observable metabolites from a wide variety of samples (6). The disadvantages of the protocol include its specificity to hydrophilic molecules and poor recovery of thermally unstable compounds (e.g., ATP, and other phosphorylated intermediates). However, many of the highly abundant metabolites found in tissue extracts are thermally stable and suitable to this preparation method. Of the amino acids we have observed, only glutamine (which undergoes thermal conversion to pyroglutamate) is significantly degraded.

1. Cryogenically homogenize tissue in a ball mill or mortar and pestle.
2. Lyophilize homogenized samples for 24 h.
3. Aliquot 400 mg dry weight tissue samples into 22-mL screw-top reaction vessels.

4. Add 16 mL of boiling dd$H_2O$ containing 167 μM HEPES or MES. HEPES or MES is used as an internal concentration reference and will have a final concentration of 1–5 mM in the final NMR analysis solution (depending on the volume of $D_2O$ used).
5. Suspend sealed sample vials in a boiling water bath for 7.5 min.
6. Cool samples on ice for 10 min and then centrifuge reaction vials in a swinging bucket centrifuge for 30 min at 1,000 × *g*.
7. Harvest supernatant and pressure filter the mixture through glass wool to remove any remaining particulate matter.
8. Microfilter the metabolite extract with a 3 kDa molecular weight cutoff spin concentrator to remove soluble proteins. Microfilters must be thoroughly washed prior to this step to remove glycerol from the membrane surface. We wash 25-mL spin concentrators by running 100 mL of water through the filters prior to use; even this amount of washing leaves detectable levels of glycerol in the final sample.
9. Lyophilize the metabolite filtrates to a dry powder.
10. Dissolve dried metabolite powder in a fixed volume of NMR solvent ($D_2O$ containing 500 μM DSS and 500 μM $NaN_3$). We use 800 μL for most tissues, but this volume is dependent the salinity of the tissue. The object with this step is to concentrate samples as much as possible without leaving a precipitate or generating excessively high salt concentrations (which will be evident from long 90° NMR pulse lengths). All samples related to a study must be prepared using the same volume of solvent.
11. Titrate samples with concentrated acid or base (~1 M DCl or ~1M NaOD) to a pH of 7.400 ± 0.004.

## 3. Resources for NMR- and MS-Based Metabolomics

Traditional methods for identifying and quantifying molecules rely on visual inspection of data and hand assignment of signals. Although this approach is effective when applied to simple mixtures, it is too labor intensive to be practical for comprehensive analysis of complex biological extracts. Modern bioanalytical metabolomics relies on bioinformatics, databases of metabolite standards, and specialty software, to make comprehensive analyses a more tractable challenge. Recent developments in these tools have dramatically improved the efficiency and reliability of analyses and have made bioanalytical metabolomics an increasingly popular research strategy. In this section, we focus on the tools we have found most effective in our own research; a more complete listing of useful resources can be found in Table 1.

**Table 1**
**List of freely accessible metabolomics resources**

| |
|---|
| Standards Initiative |
| Metabolomics Standards Initiative (http://msi-workgroups.sourceforge.net/) |
| Small molecule databases |
| ChemIDplus (http://chem.sis.nlm.nih.gov/chemidplus) |
| Human Metabolome Database (www.hmdb.ca) |
| Madison Metabolomics Consortium Database (http://mmcd.nmrfam.wisc.edu) |
| Metlin (http://metlin.scripps.edu) |
| PubChem (http://pubchem.ncbi.nlm.nih.gov) |
| Metabolic pathway databases |
| BioCyc (www.biocyc.org) |
| ExPASy (www.expasy.ch/cgi-bin/search-biochem-index) |
| KEGG (www.genome.jp/kegg) |
| Reactome (www.reactome.org) |
| TAIR (www.arabidopsis.org) |
| UM-BBD (http://umbbd.msi.umn.edu) |
| Laboratory Information Management Systems (LIMS) |
| Sesame LIMS (www.sesame.wisc.edu) |
| SetupX (http://fiehnlab.ucdavis.edu/projects/binbase_setupx) |
| NMR and MS databases |
| BioMagResBank (BMRB) (www.bmrb.wisc.edu) |
| Human Metabolome Database (www.hmdb.ca) |
| Madison Metabolomics Consortium Database (http://mmcd.nmrfam.wisc.edu) |
| Mass Spectrometry Database Committee (www.ualberta.ca/~gjones/mslib.htm) |
| NIST Chemistry WebBook (http://webbook.nist.gov/chemistry) |
| NMR metabolomics database of Linkoping (http://www.liu.se/hu/mdl/main) |
| NMRShiftDB (http://www.nmrshiftdb.org) |
| NMR Data Analysis software |
| MetaboMiner (http://wishart.biology.ualberta.ca/metabominer/) |
| rNMR (http://rnmr.nmrfam.wisc.edu/) |

### *3.1. Databases and Bioinformatics*

For NMR-based studies, one of the most important advances in recent years has been the introduction of three libraries of experimental data collected on pure metabolite standards. Although several commercial and public NMR libraries have been in existence for years, the previous databases were either not curated, contained data collected under a variety of conditions (10), or were not focused on biologically relevant molecules (11). Furthermore, all of the earlier resources were restricted to 1D NMR data. In the last few years, the Madison Metabolomics Consortium (MMC), the Human Metabolome Project (HMP), and Bruker have expended considerable resources on collecting NMR data of standard compounds. These libraries are distinct from their predecessors in that they were collected under defined conditions, include a wide variety of NMR experiments and, in the case of the MMC and HMP data,

are freely available. Between the MMC and HMP, spectra of about 1,000 metabolites are freely available over the Web. These data have made it possible to replace many of the time-consuming steps of metabolite identification and resonance assignment with bioinformatics and have paved the way for quantitative NMR-based metabolomics.

For MS-based researchers, the state of experimentally derived data libraries is more complicated. GC-MS researchers enjoy a high degree of standardization with respect to commercial instruments, instrumental settings, retention time reporting, and mass fragmentation. As a result, several large, high-quality libraries (e.g., NIST Standard Reference Database) are available as well as specialized metabolomics libraries (e.g., BinBase from Oliver Fiehn's laboratory). LC-MS has yet to achieve the same level of standardization enjoyed by GC-MS. As a result, experimental libraries of LC-MS-observed retention times, exact masses, and fragmentation patterns are of more limited use. None the less, the HMP has made a considerable effort in standardizing data collection and reporting of LC-MS data and has collected spectra of about 2,000 metabolites.

*BioMagResBank* (*BMRB*). The BMRB has served the NMR community for many years as the world repository of NMR data related to proteins and nucleic acids (12, 13). Recently, the BMRB expanded its archives to include spectra of small molecules collected under standardized conditions (14, 15). The defining characteristics of the BMRB for metabolomics are as follows: the data are freely available, data entries are curated, all of the solvent and NMR spectral parameters are clearly defined, and the raw spectral data (in addition to peak-picked and processed spectra) can be downloaded. Having the raw spectral data at hand is important because it allows spectra of standards to be overlaid over extracts for hand verification of metabolite assignments.

Currently, the BMRB contains more than 7635 NMR spectra of 1070 unique compounds collected by the MMC. These data were tailored to tissue-based metabolomics analyses in the design of their solvent conditions ($D_2O$, pH 7.400 $\pm$ 0.004, 50 mM $NaPO_4$ for water soluble metabolites; $CDCl_3$ for organic soluble metabolites) and have shown to be useful for identifying metabolites in complex $^{1}H$–$^{13}C$ NMR spectra of a variety of extracts (6). Although all of the data currently available from BMRB have been contributed by the MMC, the BMRB and HMP are working toward incorporating the extensive HMP archives into the BMRB.

The BMRB offers several bioinformatics tools to enhance its function as a data repository. In contrast to the tools offered by the Human Metabolome Database (HMDB) and Madison Metabolomics Consortium Database (MMCD), which are best suited to broader queries of the metabolite literature, the BMRB tools are primarily designed to make all of the archived data easily accessible.

An NMR peak query provides a rapid means of locating standards. Similarly, MS tools allow researchers to translate exact masses into molecular formulae or locate records on the basis of monoisotopic masses. In summary, the BMRB provides an extensive collection of freely available, high-quality NMR data coupled with an efficient query system. The BMRB database is available over the Web at www.bmrb.wisc.edu.

*HMDB.* The HMDB was officially launched in January, 2007 and currently holds the honorable distinction of being the world's largest repository of NMR and MS data collected under standardized conditions relevant to metabolomics (16). As the name implies, the HMP's main focus is human metabolism. To this end, the Canadian group has amassed a prodigious database related to the biological significance, metabolic pathways, and physical properties of metabolites found in humans. The centerpiece of the HMDB is their collection of experimentally acquired MS and NMR data. Although there is approximately 30% overlap between the data collected by the MMC and HMP, the datasets were collected with different purposes in mind. The HMDB's main focus is biological fluids, whereas the MMCD is aimed at analysis of tissue extracts. These different foci provide flexibility to the metabolomics community with respect to experimental design.

The HMDB offers a number of browsing and bioinformatics tools for accessing their data and searching the literature. One of the most useful HMDB features is their "metabocard," a single Web page containing approximately 90 data fields summarizing all of the HMDB information related a particular compound. This feature provides an efficient mechanism for learning about your recently identified metabolites. Similar to the MMCD, the HMDB provides direct links to other databases containing information relevant to a metabolite of interest. The links lead to a multitude of other resources specific to the metabolite in question. Although the HMDB also offers several bioinformatics tools for querying the database with experimental MS or NMR data, we have found the main advantage of the database to be in browsing the metabolomics literature. The HMDB is unquestionably the most powerful resource available to the metabolomics community for this purpose. The HMDB is available over the Web at http://www.hmdb.ca.

*MMCD.* The MMCD was initially developed as an in-house tool for identifying metabolites. The MMCD was released to the public in late 2006 as a bioinformatics resource for both MS and NMR based metabolomics (4). In its first year of operation the MMCD received 91,000 visitors from around the globe. Public interest in the MMCD stems from its collection of 20,000 small molecules of biological interest gathered from electronic databases and the scientific literature. These data include the following: chemical

formula, names and synonyms, structure, physical properties, NMR and MS data on pure compounds (when available), NMR chemical shifts determined by empirical and theoretical approaches, calculated isotopomer masses, information on the presence of the metabolite in different biological species, and extensive links to other databases.

In contrast to the HMDB, which excels at browsing metabolite data, the main advantage of the MMCD is in its bioinformatics capabilities. For identifying metabolites by name, the text search engine has a large collection of synonyms and automatically allows for fuzzy text matching. Users can also enter database ID numbers from a variety of other public resources (e.g., KEGG and CAS). For structure-based searches, the MMCD allows queries by molecular formula, string representation (e.g., SMILES and INCHI) or common structure files (e.g., .mol and .pdb). Alternatively, the structure can be drawn directly into a molecular graphics window. Users can combine as many as six structural criteria in a logical fashion to further refine the searches.

Perhaps the most powerful attributes of the MMCD are its metabolite assignment tools. Users can upload experimental NMR or mass spectra for bulk queries of the database. The NMR-based searches compensate for differences in magnetic field strength and filter search results on the basis of the overall patterns in the submitted peaks. NMR-based queries average about 95% sensitivity and 4% false discovery when analyzing 2D $^{1}H$–$^{13}C$ HSQC spectra of pure compound mixtures (4). Although the MMCD's performance is diminished by chemical shift variation in real biological extracts, it is still one of the most effective automatic NMR-based metabolite identification tools in the public domain.

For mass-based searches, the MMCD is primarily designed for identifying metabolites by exact mass, although the MMCD can also handle LC-MS, and MS/MS data. Users can specify the ionization mode, mass accuracy, carbon and nitrogen isotopic composition and allow for common adducts. Experimental LC-MS and MS/MS peak lists can be uploaded directly either as flat text files or in JCAMP-DX format for batch queries. Although sensitivity and false discovery rates are harder to estimate for MS based queries, MMCD users can expect between one and three matches for each mono-isotopic mass entry. As with most mass based queries, the efficacy of the search engine is primarily dictated by the mass accuracy and the mass range being queried. MMCD's main advantage for MS queries is that it allows users to restrict mass queries to known metabolites by using the biological filter provided under the miscellaneous search engine.

We recently upgraded the MMCD server and software to match heavy user traffic and will continue to expand the resource as demand increases. Currently, users can expect metabolite assignments on $^{1}H$–$^{13}C$ HSQC spectra in less than 2 s per spectrum. The MMCD is available over the Web at http://mmcd.nmrfam.wisc.edu.

### 3.2. Laboratory Information Management Systems

Laboratory information systems (LIMS) are a convenient means of organizing the large volume of data produced by metabolomics studies. Ideally, a LIMS should track every procedure that has been performed and capture all of the information known about a given sample. In addition to improving record keeping, a well designed LIMS should also make repetitive data entry tasks more efficient. Although LIMS can be helpful in many smaller settings, they are particularly useful for large-scale projects requiring the coordination of multiple investigators. In this section, we describe Sesame, a freely available LIMS system used by our laboratory.

We initially developed Sesame as a versatile Web-based LIMS to support both small and large-scale structural genomics projects (14, 17). In this application, Sesame has been adopted by a number of laboratories around the world and has proved indispensable to our own research. Although older versions of Sesame could be used for metabolomics, the LIMS was not well adapted to the requirements of small molecule studies. In response to this, we developed "Lamp," a new module for Sesame specifically for metabolomics. Lamp is divided into two subparts: one section manages data related to small molecule standards; the other is devoted to tracking information related to biological samples. The standards section captures the source of the standard, the amount of material remaining, the storage location, the investigators who have used the standard to prepare samples, the protocol used to prepare the standard. In addition, Lamp captures NMR spectra, chromatograms, MS data, and other instrumental data. The section of Lamp devoted to biological samples was designed to ensure that the LIMS captured all of the requisite information outlined by the metabolomics standards initiative (18). These entries include metadata (e.g., species, tissue, sex, age, environmental conditions, etc.), extraction protocols, NMR/LC/MS data, and results derived from any given study.

One of the most useful features of Sesame is its flexibility. The LIMS is built on a Java platform, and its Web-based format makes it accessible to any computer with Internet access. Furthermore, users can modify the existing format, design their own data fields, and create lab-specific protocols. Although the platform was designed to give everyone equal access to data, project managers can restrict access according to the needs of their own applications. The flexibility of this system allows pertinent information from databases, bench scientists, laboratory instrumentation, and software packages to be seamlessly integrated in a single interface.

Another useful feature of Sesame is its ability to track and read barcodes. Beyond the obvious application of identifying samples by bar code, the system allows specific actions to be linked to a barcode scanner. This is particularly useful in repetitive tasks such as performing inventories of standards. The Sesame LIMS is freely available from http://www.sesame.wisc.edu.

### 3.3. Software for NMR Data Analysis

Bioanalytical NMR-based metabolomics studies often require more than a thousand resonances assignments. Currently, this is a laborious task that is subject to human error and is difficult to document. Chemical shifts of many metabolites are subject to unpredictable variations resulting from uncontrolled differences in solution chemistry between samples. Consequently, resonances assigned in one spectrum cannot be transferred directly to other spectra. Although there are several effective software tools for identifying metabolites in complex NMR data (e.g., MMCD, metaboMiner, Chenomx, Bruker Amix, and BioRad KnowItAll), none of the existing tools were designed for assigning resonances across multiple spectra. As a result, every spectrum must be assigned individually. Using existing software tools, this task can require weeks of visual data inspection.

To make comprehensive NMR data analysis more feasible, we developed an open source software (rNMR) (19) written for the R statistical software environment. rNMR operates on a fundamentally different principle from existing NMR tools; rather than assigning peaks, rNMR extracts user-defined regions of interest (ROIs) from spectra. Unlike peak lists, which are static summaries containing limited information, ROIs contain all of the NMR data present with a defined set of chemical shift ranges and can be visually inspected. rNMR displays ROIs extracted from hundreds of samples side-by-side, and allows users to dynamically manipulate the size and placement of ROIs while simultaneously visualizing all of the NMR data related to an assignment. This strategy allows thousands of resonances to be visually inspected in a few minutes. Moreover, rNMR allows users to correct assignment errors at any stage of an analysis by simply adjusting the bounds of the affected ROI. Because all resonance assignments are made within the context of a defined chemical shift range, rNMR enforces consistent resonance assignments across hundreds of samples while maintaining flexibility to variations in chemical shift.

In addition to simplifying resonance assignment procedures, rNMR also makes quantitative analyses more transparent. Quantitative algorithms are based directly on the ROI data displayed to users, and the underlying NMR data behind any data point can be examined by simply clicking on the appropriate ROI. Because rNMR generates quantitative data on the fly from raw NMR spectra and a table containing the boundaries of each ROI, any rNMR analysis can be replicated by other researchers.

We initially developed rNMR as an in-house tool to solve practical problems encountered in our own research. Since its initial development, rNMR has expanded to include a broad range of peak picking, data visualization, and metabolite assignment tools that simplify data analysis. In addition, rNMR's architecture and licensing (general public license version 3) give users the freedom to customize and redistribute the program. The rNMR program,

extensive help documentation, instructional videos, compiled standards data from the BMRB, and example datasets are available free of charge from http://rnmr.nmrfam.wisc.edu.

## 4. Mass Spectrometry Methods for Identifying and Quantifying Metabolites

Metabolomics researches fall into two categories: those who use mass spectrometry (MS), and those who wish that NMR had the sensitivity of MS. Although the respective advantages of NMR and MS are well known, it is worth mentioning that despite the clear superiority of MS with respect to detection limit, MS has two fundamental challenges: nonuniform ionization efficiencies and translating identified masses into specific metabolites.

Direct analysis of metabolic extracts by MS has been reported (20, 21), but MS analysis typically requires some form of fractionation to reduce ionization artifacts. Traditionally, this is handled by online coupling of liquid (LC) or gas chromatograph (GC) with the mass spectrometer. GC-MS is used extensively for profiling nonpolar compounds and derivatives of some polar molecules (22, 23). Many of the technical challenges associated with GC-MS based research have been resolved, but LC-MS is becoming increasingly popular for metabolomics analyses because of its compatibility with a wider range of biological compounds (24–30). The topics addressed in this section cover several new LC-MS compatible methods that have made identifying and quantifying metabolites a more tractable problem. A detailed overview of mass spectrometry methods is given in Chap. 4.

*Quantification with selective isotope labeling.* The analytical precision of ESI-MS is primarily limited by two related variables: ionization efficiency and matrix effects (31). Ionization efficiency, or the percent of a molecular species that is ionizable, depends on a number of instrumental factors, molecular characteristics and solution conditions. Instrumental factors, particularly pressure and temperature at the ion source, are difficult to control and can produce significant run-to-run and day-to-day variations. Matrix effects, which occur when ions other than the target compound compete for charge, are also problematic because minor changes in the matrix can have a pronounced effect on ionization efficiency. Although chromatography can reduces these problems, observed peak intensities in metabolomics studies are inevitably influenced by factors other than metabolite concentrations (32).

One method for improving the quantitative performance of MS is to relate observed signals to isotopically labeled ($^{2}H$, $^{13}C$, $^{15}N$, or $^{18}O$-substituted) internal standards for each of the target molecules. Some care should be taken with $^{2}H$ labeled compounds

to ensure that labels are limited to nonlabile atoms and that the perdeuterated positions do not interact with the chromatographic columns (perdeuteration can lead to chromatographic shifts) (33, 34). The main advantage of internal isotopic standards is that they coelute with their unlabeled counterparts and thus can be used to normalize variation in ionization efficiency and matrix effects. In practice, this entails dividing peak intensities (or volumes) of unlabeled compounds by the intensities of the corresponding labeled standards. Accurate concentrations of the unlabeled compounds can then be calculated by multiplying the normalized signals by the known concentrations of the standards. The primary disadvantage of this approach is that an isotopically labeled standard is needed for every compound of interest, which becomes difficult in comprehensive studies because of the price (~$100/mg) and potential unavailability of labeled standards.

There are several alternatives for larger-scale projects in which absolute quantification can be replaced by relative abundance. One method involves in vivo isotopic labeling of a control sample using an economical substrate (e.g., *Escherichia coli* grown on U–$^{13}$C glucose or acetate). A fixed amount of the labeled mixture produced in vivo is added to each of the test samples, and the relative abundances of metabolites are computed by comparing the signal from labeled molecules to their corresponding unlabeled counterparts (24, 29, 35, 36). This strategy works well for most small free living organisms (yeast, bacteria, and tissue cultures) and is applicable to some whole plants. However, the approach is limited in mammals because of the difficulty in achieving uniform isotopic labeling and the prohibitive expense.

An alternative strategy for calculating relative abundances of metabolites is to use selective chemistry to isotopically label molecules containing a particular functional group (Fig. 2) (37–41). The general strategy for the selective chemistry approach is similar to the in vivo approach in that an isotopically-labeled control mixture is used as a concentration reference sample for a series of test samples. Quantification based on selective chemistry requires each test mixture to be derivatized in the same manner as the isotope-labeled control. To distinguish between molecules originating from the test and control mixture, test samples are derivatized with unlabeled reagents (natural abundance levels of isotopes), whereas control mixtures are derivatized with isotopically labeled reagents. The “heavy” control mixtures are then mixed with “light” test samples creating a composite mixture. When analyzed, the heavy and light derivatives coelute from the LC-column and appear in the mass spectrum as pairs of peaks with a mass-shift equal to the difference in mass of the two isotopic labels (Fig. 3). The ratio of peak areas or intensities for each pair can then be used to compute the relative metabolite abundances in each of the test samples.

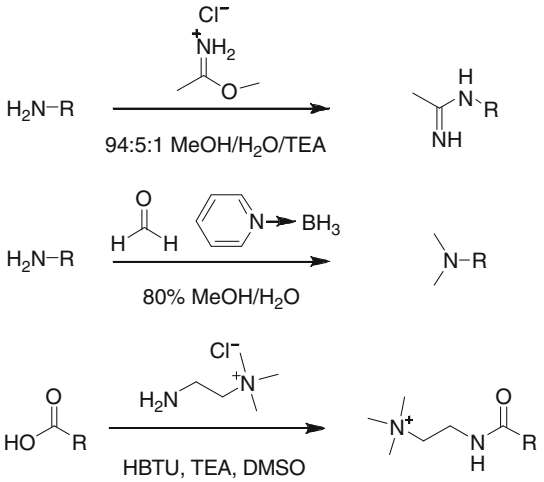


Fig. 2. Isotopic labeling chemistry under different reaction conditions. Isotopic shifts resulting from differential labeling of amines with (*top*) methylacetimidate, where a 2 Da shift is produced from the two $^{13}C$ atoms; (*middle*) formalin, where primary amines acquire two $^{13}C$ methyl groups to produce a 2 Da shift, and (*bottom*) cholamine, where $^{2}H$ on all three methyl groups produce a 9 Da shift. (Abbreviations: MeOH, methanol; TEA, triethanolamine; HBTU, 2-(1H-benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate; and DMSO, dimethyl sulfoxide).

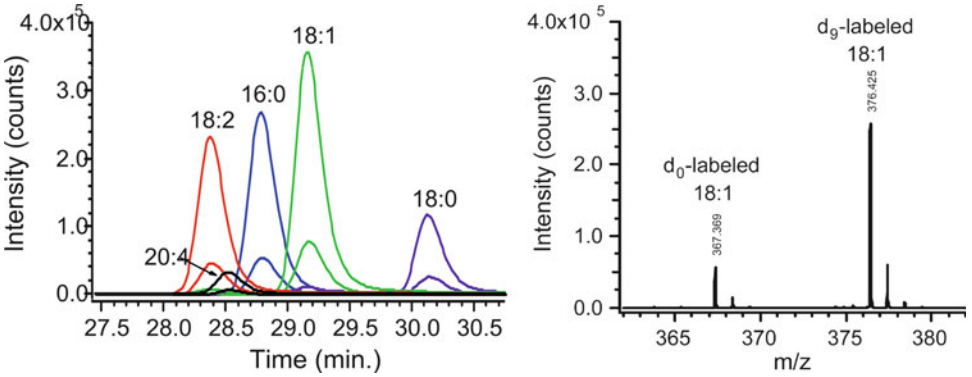


Fig. 3. Representative extracted ion chromatograms and a mass spectrum for cholamine-labeled fatty acids (1). Light and heavy labeled fatty acids coelute from reverse phase LC (*left*) and are easily distinguished by MS by their characteristic 9 Da shift (*right*). Amines labeled as shown in Fig. 2 yield analogous results.

The selective chemistry approach to quantification offers a number of advantages to the metabolomics community. Most importantly, it improves quantitative precision by normalizing variations in detection sensitivity resulting from ionization suppression (42–44) and variability in retention time between runs (34). This is possible because the "light" and "heavy" metabolites coelute within a single run and therefore have identical retention times and are electrosprayed from identical solution conditions. A second benefit of using a derivatization reagent is that it can help identify a metabolite by indicating the presence of a certain functional group.

By employing a number of labeling strategies, one can target various classes of compounds. Furthermore, well-designed labeling reagents can improve chromatographic separation, enhance detection sensitivity and yield low coefficients of variation (37–39, 45). Although it is too early to judge the efficacy of this approach in metabolomics settings, selective chemistry promises to be a powerful, cost-effective, tool in the MS metabolomics arsenal.

*Isotope constrained formula assignments.* High-resolution MS spectra of tissue extracts contain hundreds to thousands of peaks. Assigning identities to each of these features is far from trivial. Although common metabolites can be identified by GC-MS using the well developed commercial libraries, these libraries are of little use for novel compounds. Furthermore, existing LC-MS/MS literature (primarily from the HMDB) is of limited utility because of the large platform-dependent variability present in LC-MS systems. Currently, LC-MS analysis and novel compound identification require more extensive analytical techniques than those used for GC-MS.

One analytical strategy that can be used for identifying metabolites is to calculate elemental compositions of mass peaks obtained from high accuracy mass measurements. This approach is only feasible for compounds of low molecular weight and requires very high mass accuracy estimates, such as those obtained via FT-ICR MS (Chap. 4). Unique molecular formula assignments of compounds less than 250 amu typically require a mass accuracy of 3 ppm. As mass increases, or mass accuracy decreases, the number of matching formulas balloons exponentially.

Computational and experimental constraints can be used to reduce the number of possible formulas for higher molecular weight species. Both natural abundance isotopic distribution (46) and heuristically-derived limits on elemental composition (47) are effective means of constraining molecular assignments. A more experimental approach is to create mixtures of uniformly substituted isotopomers and measure the mass shift associated with isotopic labeling (48). For example, a spectrum can be collected for a mixture of unlabeled, fully $^{13}C$ labeled, and fully $^{15}N/^{13}C$ labeled versions of a molecule. Mass shifts observed in the spectrum of this mixture can then be used to calculate the number of carbon and nitrogen atoms present in a compound, and these, in turn, can be used to restrict the number of possible formula assignments (49). Such empirically-determined constraints greatly simplify molecular formula calculations and allow unique formula assignments to be made at much higher mass values (Fig. 4).

Metabolomics applications of the isotopically-constrained formula assignment method require uniform isotopic labeling. As mentioned previously, in vivo $^{15}N$ and/or $^{13}C$ isotopic labeling is relatively straightforward and cost-effective in small free

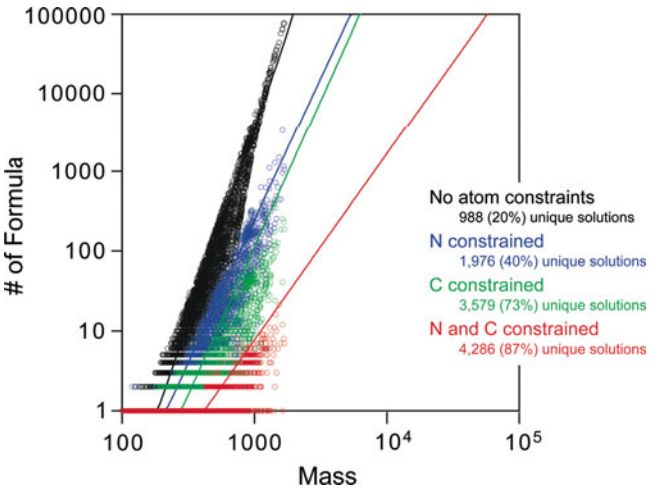


Fig. 4. Number of calculated formulae with and without nitrogen and carbon constraints for 4,918 unique formulae derived from the BMRB database assuming mass accuracy of ±3 ppm.

living organisms or tissue cultures (50). However, some care should be taken to ensure that metabolites are uniformly labeled. Partial labeling results in excessive spectral complexity resulting from the various partially substituted isotopomers. We have recently added several computational resources to the BMRB as a tool for researchers who are interested in using isotope constrained assignments (49).

## 5. NMR Methods for Identifying and Quantifying Metabolites

Although NMR is a relatively insensitive technique when compared to MS, NMR-based analyses enjoy several advantages over MS. Specifically, NMR peak intensities scale proportionally with concentration, NMR is sensitive to a wide variety of metabolite structures, and NMR analyses require little to no sample preparation. Although NMR has become a popular tool for statistics-based metabolomics, NMR-based bioanalytical studies are relatively rare because of practical challenges in data analysis. However, the bioinformatics and software tools discussed in Sect. 3 have largely eliminated these practical constraints.

In this section, we discuss several alternative strategies for collecting quantitatively reliable NMR data that can be used for bioanalytical metabolomics. A unifying theme of the techniques presented here is that they provide a mechanism for separating overlapped NMR signals. Overlapped signals, such as those found in 1D $^1$H spectra of biological extracts, scale proportionally to the total overlapped spectral density and can neither be assigned nor quantified (6).

Although 1D $^{1}$H NMR is a reproducible chemometrics tool (51), bioanalytical studies require well-defined spin systems for resonance assignments and isolated peaks for quantification (6). Currently, there are three general strategies for producing the sufficiently sparse spectra required for quantitative studies: mathematical deconvolution of spectra, multidimensional NMR, and selective pulse sequences (see Chap. 6 for a detailed description of NMR principles and methodologies).

### 5.1. Mathematical and Statistical Methods

As a rule, signals correlate very well with themselves. Not surprisingly, various peaks from the same compound are highly covariant across multiple spectra. The Nicholson laboratory recognized this several years ago and formalized the use of covariance matrices to group resonances of various compounds (52). This approach has the advantage of allowing investigators to use 1D $^{1}$H NMR spectra, the fastest and most sensitive NMR experiment, and still provide a mechanism for dispersing resonances from individual compounds. However, the covariance approach does not provide a mechanism for quantification. As a result, quantitative estimates of peaks identified through covariance are still rooted in the dubious metric of overlapped 1D $^{1}$H signals.

An alternative approach developed by Weljie et al. is to fit overlapped signals with modeled peaks and base quantitative estimates on the modeled data (53). Resonance deconvolution has been used for decades in a wide variety of traditional NMR studies and is a well established method for separating overlapped peaks. The curve fitting method introduced by Weljie, A.M. et al. is probably the most reliable approach for quantifying metabolites from overlapped 1D $^{1}$H NMR data and has become the preferred strategy for many bioanalytically oriented researchers. However, resonance deconvolution is a finicky hand-manipulated process whose performance is affected by the skill of the person operating the software. A commercial implementation of Weljie's method, Chenomx, removes some of this uncertainty. However, Chenomx is dependent upon matching metabolites with one of the standards present in their commercial library and thus can be ineffective for some compounds. In addition, Chenomx is expensive and only supports 1D NMR analyses. Despite these disadvantages, resonance deconvolution is a viable method for deriving quantitative information from overlapped NMR spectra and may be an attractive alternative to researchers who are constrained to using 1D NMR.

### 5.2. Multidimensional NMR Methods

One of the most effective methods for mitigating problems associated with resonance overlap is to use one of the myriad of multidimensional NMR experiments that have been developed by bimolecular NMR spectroscopists (Chap. 6). In contrast to other methods for separating overlapped resonances discussed here, multidimensional NMR has the significant advantage of contributing

empirically determined structural information about the observed resonances. This is of no small consequence to bioanalytical metabolomics studies, which must differentiate between many structurally similar metabolites.

Several groups have applied multidimensional NMR to metabolomics (5, 54–56), but practical challenges have made this technique unpopular for routine studies. 2D pulse sequences require longer acquisition times, are less quantitatively robust, are prone to data artifacts, and require more NMR expertise than traditional 1D spectroscopy. In this section, we provide guidelines for mitigating these problems and discuss the trade-offs of various multidimensional NMR strategies.

*Resolution versus sensitivity.* Two categories of 2D NMR have proven effective in metabolomics: $^{1}H$–$^{1}H$ homonuclear and $^{1}H$–$^{13}C$ heteronuclear experiments. These categories have several inherent trade-offs that must be considered in the experimental design phase of any bioanalytical metabolomics study. Homonuclear proton experiments are more sensitive because of the 100% natural abundance of $^{1}H$ and its favorable magnetogyric ratio. However, metabolite signals in $^{1}H$–$^{1}H$ experiments occupy a narrow bandwidth (roughly 10 ppm), and most $^{1}H$–$^{1}H$ pulse sequences produce multiple signals from each resonance (i.e., symmetrical cross-peaks, and diagonal peaks). The low bandwidth and signal redundancy result in resonance overlap problems. In contrast, $^{1}H$–$^{13}C$ sequences offer superior separation of metabolite signals because of carbon's larger bandwidth (roughly 180 ppm) and because most $^{1}H$–$^{13}C$ pulse sequences produce fewer signals per metabolite. Although modern $^{1}H$-detected $^{13}C$ experiments are significantly more sensitive than the traditional direct detection methods used in 1D–$^{13}C$ experiments, the low natural abundance of carbon (1.1%) limits metabolite investigation in unlabeled samples. The lower sensitivity of $^{1}H$–$^{13}C$ experiments can be partially mitigated by concentrating NMR samples (we prepare samples at up to 10 $\times$ higher concentrations than found in vivo) or by isotopically enriching samples. However, sample limitation and biological constraints often make concentration or isotopic enrichment unfeasible. As a result, many NMR-based studies must either choose high sensitivity experiments at the cost of increased spectral overlap, or better signal separation at the cost of lower sensitivity. The correct choice is governed by the biological goals of a study, the amount of material that can be obtained, the concentrations of target metabolites, and the complexity of the mixture being investigated.

Additional practical considerations may influence the decision between 2D homonuclear $^{1}H$ and heteronuclear $^{1}H$–$^{13}C$ experiments. Studies involving 2D $^{1}H$–$^{1}H$ experiments require considerably longer NMR acquisition times to adequately resolve metabolite signals. In addition, chemical shift variations are more problematic in $^{1}H$–$^{1}H$ spectra because $^{1}H$ shifts are

more affected by solution conditions than $^{13}C$ shifts. Consequently, bioinformatics-based assignments of $^{1}H$–$^{1}H$ data are less reliable, and metabolite identifications are more ambiguous. On the other hand, $^{1}H$–$^{13}C$ analyses at natural abundance $^{13}C$ levels requires 40 mg of metabolites per NMR sample (roughly 400 mg starting material). If sufficient starting material can be obtained for a single representative sample, then we recommend using $^{1}H$–$^{13}C$ analysis of a concentrated sample for metabolite identification purposes. These assignments can then be transferred to $^{1}H$–$^{1}H$ data for analyses of more dilute test samples. For investigators who are new to 2D NMR, we recommend learning the metabolite identification and quantification process using a biological model that allows all samples to be analyzed via $^{1}H$–$^{13}C$ HSQC (heteronuclear single quantum coherence) or HMQC (heteronuclear multiple quantum coherence). This recommendation is based on the relative ease of assigning and quantifying $^{1}H$–$^{13}C$ data.

*Metabolite identification*. The introduction of the BMRB, MMCD, and HMDB databases has dramatically reduced the length of time required to assign NMR spectra. These resources allow researchers to submit peak lists from experimental data and return a list of possible metabolite identifications. Currently, the results of these queries must be verified by overlaying spectra of standards (available from www.bmrb.wisc.edu) onto a representative extract. Step-by-step instructions for validating resonance assignments are given below. Although our methods may be adapted to some 1D NMR applications, the procedures described here primarily refers to the analysis of 2D $^{1}H$–$^{13}C$ or $^{1}H$–$^{1}H$ NMR spectra. As mentioned above, we have found that 2D $^{1}H$–$^{13}C$ HSQC data are easier to assign than $^{1}H$–$^{1}H$ experiments and recommend that new investigators learn the assignment process with $^{1}H$–$^{13}C$ data.

*5.2.1. Protocol for Metabolite Identification*

1. Collect a high-resolution sensitivity enhanced $^{1}H$–$^{13}C$ HSQC spectrum (e.g., Varian pulse sequence gHSQC) of a representative sample. This will likely require 512–2,048 increments in the indirect dimension, four scans, and as long an acquisition time as the decoupling strategy allows. The goal of the initial spectrum is to produce one high-quality, unambiguous dataset with minimal peak overlap to be used for metabolite identification purposes. It is virtually impossible to resolve all of the signals, but most aliphatic signals are dispersed in 1,028 indirect increments. It is important to match the sample's solvent conditions to those used by either the MMC or HMDB because bioinformatics-based assignments require predictable peak locations.
2. After a high quality spectrum has been collected, process the data with the minimum appropriate window function (excessive line broadening will mask J-coupling), reference the

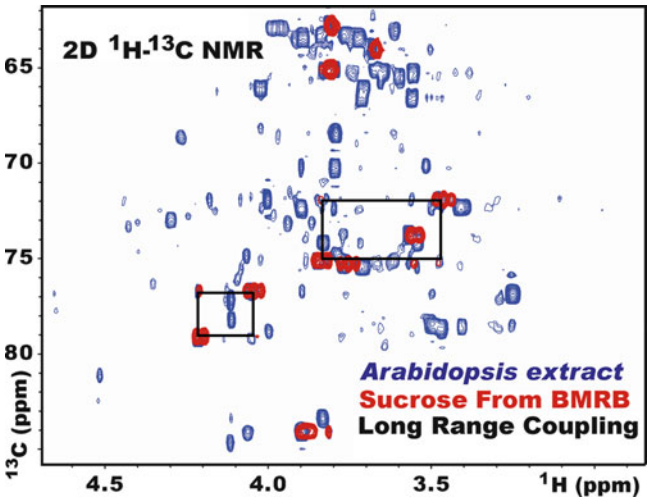


Fig. 5. Two-dimensional $^{1}$H–$^{13}$C HSQC NMR spectrum of sucrose from the BMRB (*red*) overlaid onto an aqueous whole-plant extract from *Arabidopsis thaliana* (*blue*). *Black boxes* indicate long-range proton carbon couplings used to validate the assignment.

chemical shifts, peak-pick the data, and submit the peak list to the MMCD, HMDB, or BMRB. Accurate chemical shift referencing is critical for bioinformatics-based assignments.

3. Bring the experimental data into rNMR (19) for analysis (rNMR supports conversions of data in Bruker, Varian, or NMRpipe formats to the Sparky format used by rNMR) and download spectra of the possible matches from either the HMDB or BMRB. We have converted most of the BMRB $^{1}$H–$^{13}$C HSQC and $^{1}$H–$^{1}$H TOCSY standards to Sparky format, and these data can be downloaded in bulk from http://rnmr.nmrfam.wisc.edu.
4. Overlay the spectra of each potential metabolite onto your high resolution spectrum of the representative extract. Reliable metabolite assignments must have all of the correct $^{1}$H and $^{13}$C chemical shifts, correct peak multiplicities, and intense peaks should show long-range $^{1}$H–$^{13}$C coupling consistent with the standard (Fig. 5). Some tolerance can be given for chemical shift variation (± ~0.025 ppm), provided that all of the other criteria are fulfilled. A few resonances are more variable than others. Specifically, malate, citrate, and a number of aromatic resonances can have considerable variation despite careful pH titration. Ambiguous metabolite assignments can be checked by adding pure standards to the extract; correct assignments will show increases in peak intensity proportional to the amount of standard added.

The metabolite identification protocol outlined here will provide reliable assignments in most cases, but this method of analysis is not unequivocal. Important assignments (i.e., those that are

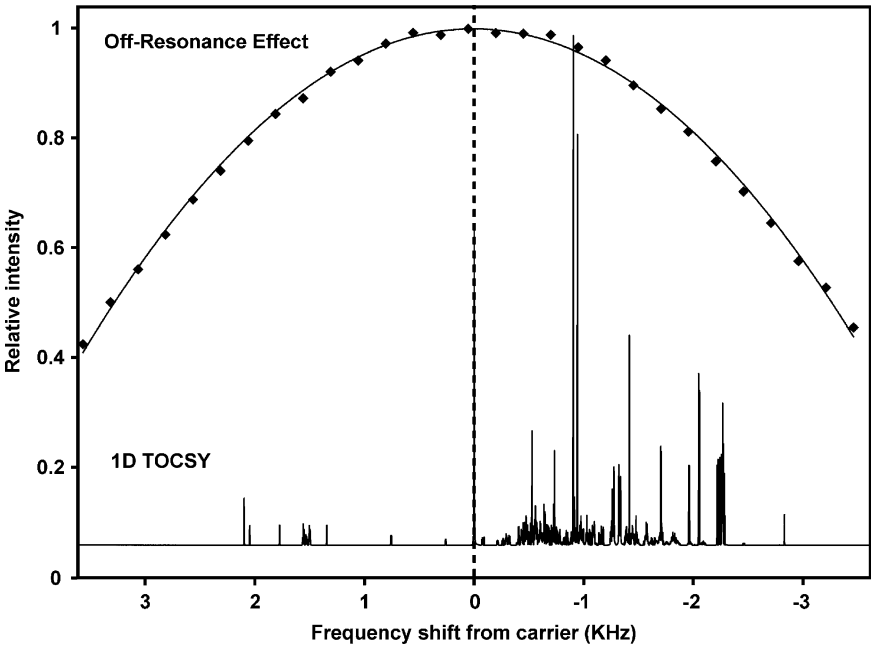


Fig. 6. Off-resonance effects are one of the many factors that can influence peak intensities observed in NMR experiments. In this example, the intensities in $^{1}H$–$^{1}H$ TOCSY were measured as a function of $^{1}H$ frequency offset. The peak intensity is highest when the peak is closest to the transmitter frequency and lowers as the transmitter is tuned to higher or lower frequencies.

shown to have significant changes between groups) should be validated by an independent analytical technique.

*Calculating molar concentrations from 2D NMR spectra.* Multidimensional pulse sequences have significant quantitative problems resulting from off-resonance effects, incomplete resonance transfer, complex relaxation pathways, decoupling artifacts and mixing times. These variables make peak intensities (and volumes) an unreliable metric for quantification (Fig. 6). Several laboratories are developing more quantitatively reliable versions of common bimolecular NMR pulse sequences (57–59). Unfortunately, these efforts have yet to produce experiments that are sufficiently robust for bioanalytical metabolomics. However, we have developed a practical approach to metabolite quantification that allows researchers to use any NMR pulse sequence. The basic feature of our method is that we do not control quantification through pulse sequence design. Instead, we relate observed peak intensities to those of mixtures of pure standards of known concentration. Differences in relaxation rates and shimming are controlled by normalizing observed metabolite signals to an internal standard. The metabolite quantification strategy described here and the fast data collection method described in the next section are the components of our fast metabolite quantification (FMQ by NMR) approach, which is the basis of all of our laboratory's routine metabolomics work.

*5.2.2. Metabolite Quantification Protocol*

1. Identify all metabolites present in an extract (see *Protocol for metabolite identification*).
2. Prepare three mixtures containing all of the identified metabolites at 2, 5, and 10 mM. Although standardized metabolite intensities are linear well beyond the range of these standards (±20-fold), the concentrations of standards should be adjusted to match the approximate range of expected concentrations. Mixtures should be prepared at high volume to minimize weighing errors. All standards and extracts should be titrated to pH 7.400 ± 0.004.
3. Include 5 mM of an internal standard in all of the standards mixtures and extracts to serve as an internal concentration reference. For $^{1}H$–$^{13}C$ NMR studies, we recommend HEPES or MES because both compounds have multiple isolated peaks that do not overlap with biological compounds. HEPES is convenient because it acts as an internal pH indicator, but MES is probably a more reliable concentration reference because of its lower p$K_a$.
4. Collect spectra of extracts and concentration reference samples under identical NMR acquisition conditions at the same time using the same instrument. Every sample, including the concentration reference samples, should be collected twice to produce two technical replicates for each sample. The sample order should be randomized. Each test sample should have at least three (preferably many more) independent biological replicates. From a statistical perspective, it is much better to have many independent biological replicates than to analyze many metabolites (large alpha corrections, such as Bonferroni correction, must be made in studies that use multiple comparisons).
5. Measure the peak intensities (area for 1D, peak height for 2D) of nonoverlapped peaks from extracts and the concentration reference samples. Although peak area is a robust metric in 1D NMR, we have found 2D peak volumes to be considerably less reliable than peak heights when used in this procedure (regardless of the NMR analysis software).
6. Normalize signals observed in each spectrum to the average signal of the dispersed HEPES (MES) peaks. Raw intensities can be used without normalization, but any variations in salt concentration (or paramagnetic relaxation agents) between samples will be erroneously interpreted as differences in metabolite concentrations. If samples and standards are osmotically identical, then normalization is undesirable because it introduces some technical error. However, many NMR probes are highly salt sensitive, and even the standards may show strong nonlinear effects of salt. If in doubt, it is better to normalize to the internal standard and accept a small increase in technical error.

7. Average normalized peak intensities across technical replicates.
8. Regress normalized peak intensities of the standards to produce a concentration versus peak intensity equation for each dispersed signal.
9. Calculate the observed concentrations for each normalized peak in the test samples using the equations derived from the standards samples.
10. Average concentration estimates across all dispersed peaks from each molecule.

The protocol described here produces concentration estimates with as little as 2.7% technical error from complex 2D NMR spectra (Fig. 7.) (6). The main disadvantage of this approach is its dependence on standards, many of which are unavailable or are

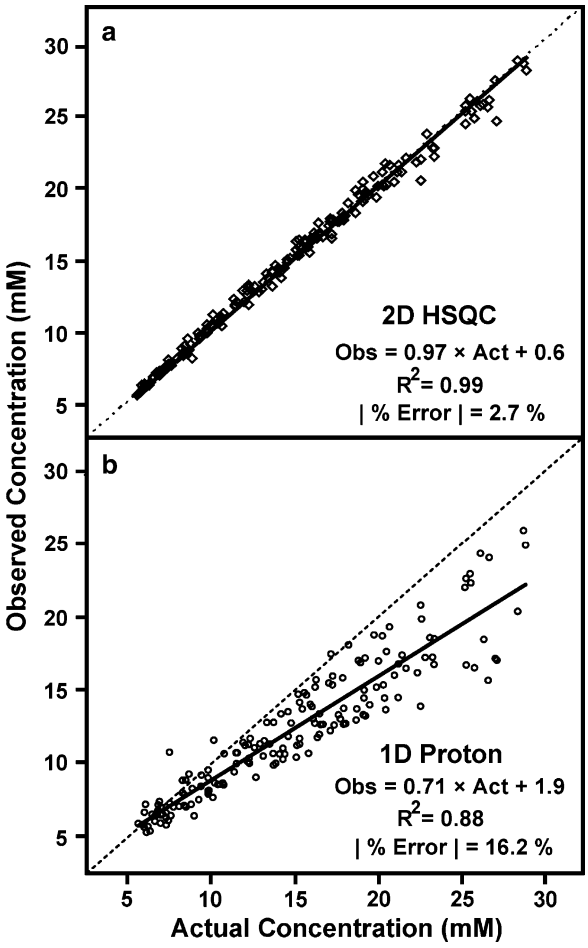


Fig. 7. (**a**) Concentration estimates ($N = 168$) based on two-dimensional $^{1}H$–$^{13}C$ HSQC and (**b**) one-dimensional $^{1}H$ NMR data. Estimates are plotted as a function of the known concentration of metabolites in synthetic mixtures. *Dotted lines* indicate the ideal regression (slope 1), and the *solid lines* indicate the best-fit regression.

prohibitively expensive to use in the quantities required for this procedure. If the requisite standards can be obtained, however, then this strategy produces reliable quantitative information from most of the existing NMR pulse sequences. Moreover, the internal concentration references we use to control for differences in longitudinal ($T_1$) relaxation, and other variations between samples, can be introduced early in the sample preparation process to relate observed signal intensities to biologically relevant concentrations.

*Fast multidimensional experiments.* Although multidimensional NMR has seen qualitative applications to metabolomics for many years (5, 54–56), very few studies have used the technique for quantitative purposes. One reason for this is that multidimensional NMR spectra take longer to collect than 1D $^{1}$H NMR spectra, although the long acquisition times of 2D experiments have been somewhat exaggerated in the metabolomics community. Over the years, protein NMR spectroscopists have developed a wealth of techniques for reducing the requisite acquisition times of multidimensional experiments. These techniques range from mathematical approaches for nonlinear sampling of data (60), to pulse sequences that encode the indirect dimensions using gradients (61) and optimization of pulse angles (62). All of these techniques are applicable to metabolomics (Chap. 6), but the easiest way to save time is to be judicious in setting up the NMR acquisition parameters.

We have shown that a carefully adjusted 2D $^{1}$H–$^{13}$C HSQC spectrum allows metabolites with concentrations over ~500 μM in the NMR tube (30–40 metabolites using our sample preparation methods) to be quantified in about 10 min (6). The secret to our approach is in the prior identification of compounds. As discussed above, a single high-resolution HSQC of a representative sample can be used to identify the abundant metabolites present in an extract. After the metabolites have been identified and the peaks have been assigned, one can capitalize on the assignments by reducing the number of indirect points and time to the minimum required to resolve two signals from each target molecule. These time savings can be achieved while maintaining high quantitative precession (technical error ~3%) (6).

The primary factor that contributes to lengthy 2D experiments is the number of increments collected in the indirect dimension. Time savings are proportional to the number of increments that can be eliminated from the acquisition. Because resolution in the indirect dimension is a function of the number of increments and the indirect spectral-width, achieving adequate resolution in the minimal amount of time usually involves trimming both the spectral-width and the number of increments. If the indirect spectral-width is cut in half, then the number of indirect points can be halved without affecting resolution. The spectral-width can be reduced well beyond the point where resonances are no longer contained within the

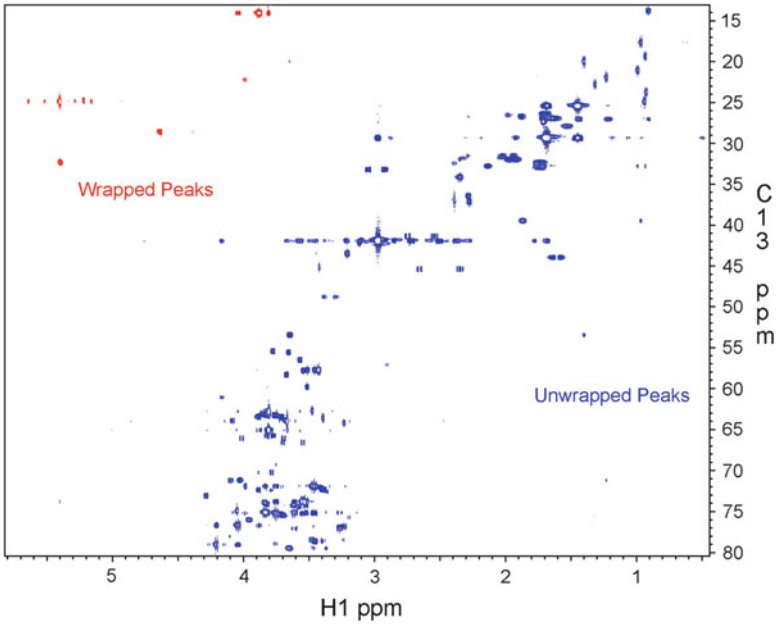


Fig. 8. Quantitative $^{1}H$–$^{13}C$ HSQC NMR spectrum of a synthetic mixture of 26 metabolites. The spectral width in the indirect dimension was reduced to allow for shorter acquisition times. *Blue peaks* are in their correct locations, whereas *red peaks* have been wrapped into the top of the spectrum from their normal downfield positions.

spectral window. Resonances with signals outside of the spectral-width will wrap back into the spectral window with a chemical shift that is equal to their original shift plus or minus a multiple of the spectral-width (Fig. 8). In more extreme cases, resonances may be wrapped multiple times. If the original chemical shifts and the spectral-width are known, then a heavily wrapped spectrum can be unwrapped by using simple arithmetic. Spectral folding may result in undesirable overlap between analytically important signals, but overlap can often be alleviated with small adjustments of the indirect carrier frequency. As discussed above, off resonance effects play a major role in observed peak intensities. After acceptable settings for the spectral-width, number of increments, and the transmitter offset have been found, it is critical that these values be kept constant across all of the test samples and standards.

### 5.3. Selective NMR Methods

Selective NMR is a back to the future idea revived from the early days of protein NMR. Although selective experiments have been largely outdated in protein NMR by modern multidimensional pulse sequences, the technique is appropriate for targeted metabolic studies. Selective NMR works by carefully sculpting the excitation pulse to cover a narrow bandwidth. This, in combination with other sequences such as TOCSY, allows individual spin systems to be isolated from amazingly complex mixtures (Fig. 9).

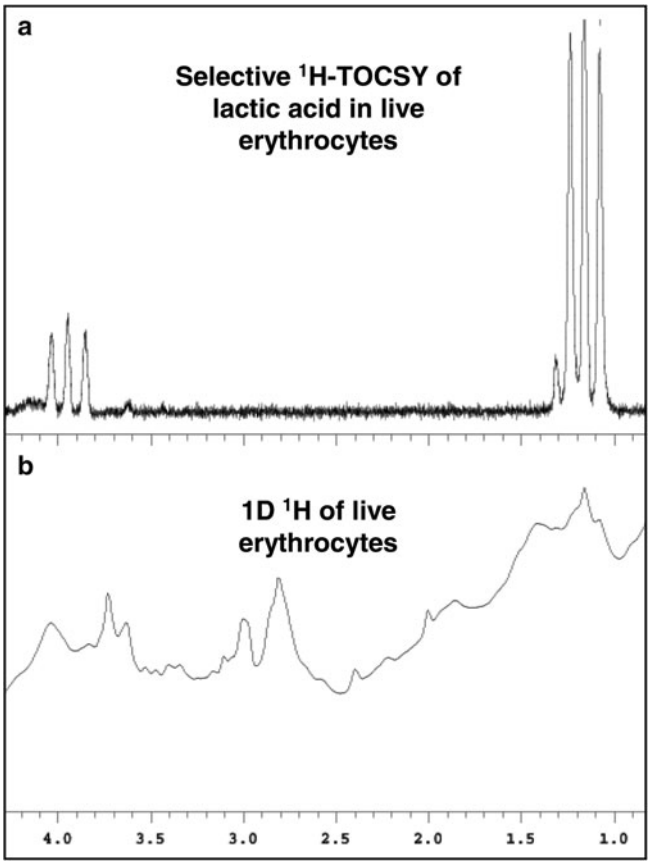


Fig. 9. (**a**) Selective 1D $^{1}$H–$^{1}$H TOCSY for lactate in live red blood cells. (**b**) Standard 1D $^{1}$H NMR spectrum of the same sample. Cells were labeled with [U–$^{13}$C]-glucose, the triplet-like splitting observed in the TOCSY arises from [U–$^{12}$C]-lactate (center peak) and fully labeled [U–$^{13}$C]-lactate (two satellite peaks).

Similarly, the technique can be used to selectively remove overpowering resonances from a spectrum. Dan Raftery's group first introduced the idea to the metabolomics community by showing that signals from minor components of honey could be accurately isolated and quantified without being influenced by the large glucose and fructose signals (63). One of the most powerful aspects of the selective TOCSY is that data can be collected very quickly (~1 min per spectrum). For studies requiring accurate quantification of a few metabolites in complex mixtures, this is one of the fastest NMR techniques available.

One must keep a few practical considerations in mind when using selective pulse sequences. Selective experiments require hand tuning of the excitation pulse, and this tuning may change from sample to sample if the target metabolite is subject to chemical shift variation. Secondly, selective experiments lose their time advantage

in studies involving multiple metabolites. After the number of target metabolites reaches about ten, then a full 2D $^{1}H$–$^{1}H$ TOCSY or $^{1}H$–$^{13}C$ HSQC is more efficient and does not require hand manipulation of the pulses between experiments. However, if a study only calls for analysis of a few molecules, then selective TOCSY is one of the best tools for the job.

## 6. Future Prospects

Technological advances of recent years have dramatically increased the efficiency with which metabolites can be identified and accurately quantified. High quality empirical libraries, bioinformatics-based spectral assignment tools, improved analytical software, and practical methods for identifying and quantifying NMR and MS signals in complex spectra have taken the field a step closer to the automation enjoyed by mainstream bioanalytical methods. Despite these advances, bioanalytical metabolomics is still in its early development and is far from capitalizing fully on state-of-the-art NMR and MS technology. A significant proportion of signals in every study go unassigned, bioinformatics tools suffer from unacceptable false discovery rates, quantification requires ad hoc correction of quantitatively unreliable data, and specialized expertise is required to collect, analyze, and interpret data. Finding practical solutions to these problems is paramount to bioanalytical metabolomics and is a promising area for future technology development.

## Acknowledgments

This work was supported by the National Center for Research Resources of the National Institutes of Health under grant P41 RR02301; I.A.L. was the recipient of a fellowship from the NHGRI 1T32HG002760.

## Glossary

| | |
|---|---|
| 1D | One-dimensional |
| 2D | Two-dimensional |
| Bioanalytical metabolomics | Comprehensive quantitative analysis of metabolites in complex biological samples |

| | |
|---|---|
| HSQC | Heteronuclear single quantum correlation |
| MS | Mass spectrometry |
| NMR | Nuclear magnetic resonance |

## References


1. Lindon JC, Holmes E, Nicholson JK. Metabonomics in pharmaceutical R&D. FEBS J. 2007;274:1140–51.
2. Radda GK, Seeley PJ. Recent studies on cellular metabolism by nuclear magnetic resonance. Annu Rev Physiol. 1979;41:749–69.
3. Shulman RG, Brown TR, Ugurbil K, Ogawa S, Cohen SM, den Hollander JA. Cellular applications of 31P and 13C nuclear magnetic resonance. Science. 1979;205:160–6.
4. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol. 2008;26:162–4.
5. Fan TW, Lane AN, Shenker M, Bartley JP, Crowley D, Higashi RM. Comprehensive chemical profiling of gramineous plant root exudates using high-resolution NMR and MS. Phytochemistry. 2001;57:209–21.
6. Lewis IA, Schommer SC, Hodis B, Robb KA, Tonelli M, Westler WM, Sussman MR, Markley JL. Method for determining molar concentrations of metabolites in complex solutions from two-dimensional $^{1}H$–$^{13}C$ NMR spectra. Anal Chem. 2007;79:9385–90.
7. Brown LM, Pais A, Pippard AB. Twentieth century physics. New York: American Institute of Physics Press; 1995.
8. Hurlbert SH. Pseudoreplication and the design of ecological field experiments. Ecol Monogr. 1984;54:187–211.
9. Mehta T, Tanik M, Allison DB. Towards sound epistemological foundations of statistical methods for high-dimensional biology. Nat Genet. 2004;36:943–7.
10. Steinbeck C, Krause S, Kuhn S. NMRShiftDB-constructing a free chemical information system with open-source components. J Chem Inf Comput Sci. 2003;43:1733–9.
11. Sadtler P. Sadtler Standard NMR Spectra N.M.R. Chemical Shift Index. Philadelphia: Sadtler Research Laboratories, PA (proton and carbon-13 spectra); 1967.
12. Seavey BR, Farr EA, Westler WM, Markley JL. A relational database for sequence-specific protein NMR data. J Biomol NMR. 1991; 1:217–36.
13. Doreleijers JF, Mading S, Maziuk D, Sojourner K, Yin L, Zhu J, Markley JL, Ulrich EL. BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. J Biomol NMR. 2003;26:139–46.
14. Markley JL, Anderson ME, Cui Q, Eghbalnia HR, Lewis IA, Hegeman AD, Li J, Schulte CF, Sussman MR, Westler WM, Ulrich EL, Zolnai Z. New bioinformatics resources for metabolomics. Pac Symp Biocomput. 2007;12:157–68, 17990487.
15. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. BioMagResBank. Nucleic Acids Res. 2008;36:D402–8.
16. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, Cheng D, Jewell K, Arndt D, Sawhney S, Fung C, Nikolai L, Lewis M, Coutouly MA, Forsythe I, Tang P, Shrivastava S, Jeroncic K, Stothard P, Amegbey G, Block D, Hau DD, Wagner J, Miniaci J, Clements M, Gebremedhin M, Guo N, Zhang Y, Duggan GE, Macinnis GD, Weljie AM, Dowlatabadi R, Bamforth F, Clive D, Greiner R, Li L, Marrie T, Sykes BD, Vogel HJ, Querengesser L. HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007;35:D521–6.
17. Zolnai Z, Lee PT, Li J, Chapman MR, Newman CS, Phillips Jr GN, Rayment I, Ulrich EL, Volkman BF, Markley JL. Project management system for structural and functional proteomics: sesame. J Struct Funct Genomics. 2003;4:11–23.
18. Sansone SA, Fan T, Goodacre R, Griffin JL, Hardy NW, Kaddurah-Daouk R, Kristal BS, Lindon J, Mendes P, Morrison N, Nikolau B, Robertson D, Sumner LW, Taylor C, van der Werf M, van Ommen B, Fiehn O. The metabolomics standards initiative. Nat Biotechnol. 2007;25:846–8.
19. Lewis IA, Schommer SC, Markley JL. rNMR: open source software for high-throughput NMR-based bioanalytical chemistry. Magn Reson Chem. 2009;47:S123–6.
20. Vaidyanathan S, Rowland JJ, Kell DB, Goodacre R. Discrimination of aerobic

endospore-forming bacteria via electrospray-lonization mass spectrometry of whole cell suspensions. Anal Chem. 2001;73:4134–44.

21. Aharoni A, Ric de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino R, Goodenowe DB. Nontargeted metabolome analysis by use of Fourier transform Ion cyclotron mass spectrometry. OMICS. 2002;6:217–34.
22. Koek MM, Muilwijk B, van der Werf MJ, Hankemeier T. Microbial metabolomics with gas chromatography/mass spectrometry. Anal Chem. 2006;78:1272–81.
23. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L. Metabolite profiling for plant functional genomics. Nat Biotechnol. 2000;18:1157–61.
24. Wu L, Mashego MR, van Dam JC, Proell AM, Vinke JL, Ras C, van Winden WA, van Gulik WM, Heijnen JJ. Quantitative analysis of the microbial metabolome by isotope dilution mass spectrometry using uniformly $^{13}$C-labeled cell extracts as internal standards. Anal Biochem. 2005;336:164–71.
25. Want EJ, O'Maille G, Smith CA, Brandon TR, Uritboonthai W, Qin C, Trauger SA, Siuzdak G. Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. Anal Chem. 2006;78:743–52.
26. Wang W, Zhou H, Lin H, Roy S, Shaler TA, Hill LR, Norton S, Kumar P, Anderle M, Becker CH. Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. Anal Chem. 2003;75:4818–26.
27. von Roepenack-Lahaye E, Degenkolb T, Zerjeski M, Franz M, Roth U, Wessjohann L, Schmidt J, Scheel D, Clemens S. Profiling of Arabidopsis secondary metabolites by capillary liquid chromatography coupled to electrospray ionization quadrupole time-of-flight mass spectrometry. Plant Physiol. 2004;134:548–59.
28. Tolstikov VV, Fiehn O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. Anal Biochem. 2002;301:298–307.
29. Lafaye A, Labarre J, Tabet JC, Ezan E, Junot C. Liquid chromatography-mass spectrometry and 15N metabolic labeling for quantitative metabolic profiling. Anal Chem. 2005;77:2026–33.
30. Dalluge JJ, Smith S, Sanchez-Riera F, McGuire C, Hobson R. Potential of fermentation profiling via rapid measurement of amino acid metabolism by liquid chromatography-tandem mass spectrometry. J Chromatogr A. 2004;1043:3–7.
31. Stokvis E, Rosing H, Beijnen JH. Stable isotopically labeled internal standards in quantitative bioanalysis using liquid chromatography/mass spectrometry: necessity or not? Rapid Commun Mass Spectrom. 2005;19:401–7.
32. Roy SM, Anderle M, Lin H, Becker CH. Differential expression profiling of serum proteins and metabolites for biomarker discovery. Int J Mass Spectrom. 2004;238:163–71.
33. Yang WC, Mirzaei H, Liu XP, Regnier FE. Enhancement of amino acid detection and quantification by electrospray ionization mass spectrometry. Anal Chem. 2006;78:4702–8.
34. Pan CL, Kora G, Tabb DL, Pelletier DA, McDonald WH, Hurst GB, Hettich RL, Samatova NF. Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. Anal Chem. 2006;78: 7110–20.
35. Birkemeyer C, Luedemann A, Wagner C, Erban A, Kopka J. Metabolome analysis: the potential of in vivo labeling with stable isotopes for metabolite profiling. Trends Biotechnol. 2005;23:28–33.
36. Mashego MR, Wu L, Van Dam JC, Ras C, Vinke JL, Van Winden WA, Van Gulik WM, Heijnen JJ. MIRACLE: mass isotopomer ratio analysis of U–13C-labeled extracts. A new method for accurate quantification of changes in concentrations of intracellular metabolites. Biotechnol Bioeng. 2004;85:620–8.
37. Lamos SM, Shortreed MR, Frey BL, Belshaw PJ, Smith LM. Relative quantification of carboxylic acid metabolites by liquid chromatography-mass spectrometry using isotopic variants of cholamine. Anal Chem. 2007;79:5143–9.
38. Shortreed MR, Lamos SM, Frey BL, Phillips MF, Patel M, Belshaw PJ, Smith LM. Ionizable isotopic labeling reagent for relative quantification of amine metabolites by mass spectrometry. Anal Chem. 2006;78:6398–403.
39. Yang WC, Adamec J, Regnier FE. Enhancement of the LC/MS analysis of fatty acids through derivatization and stable isotope coding. Anal Chem. 2007;79:5150–7.
40. Regnier FE, Julka S. Primary amine coding as a path to comparative proteomics. Proteomics. 2006;6:3968–79.
41. Berry KA, Murphy RC. Analysis of cell membrane aminophospholipids as isotope-tagged derivatives. J Lipid Res. 2005;46:1038–46.
42. Sterner JL, Johnston MV, Nicol GR, Ridge DP. Signal suppression in electrospray ionization Fourier transform mass spectrometry of multi-component samples. J Mass Spectrom. 2000;35:385–91.

43. Constantopoulos TL, Jackson GS, Enke CG. Effects of salt concentration on analyte response using electrospray ionization mass spectrometry. J Am Soc Mass Spectrom. 1999;10:625–34.
44. Annesley TM. Ion suppression in mass spectrometry. Clin Chem. 2003;49:1041–4.
45. Guo K, Ji C, Li L. Stable-isotope dimethylation labeling combined with LC-ESI MS for quantification of amine-containing metabolites in biological samples. Anal Chem. 2007; 79:8631–8.
46. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. BMC Bioinforma. 2006;7:234–44.
47. Kind T, Fiehn O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. BMC Bioinforma. 2007;8:105.
48. Rodgers RP, Blumer EN, Hendrickson CL, Marshall AG. Stable isotope incorporation triples the upper mass limit for determination of elemental composition by accurate mass measurement. J Am Soc Mass Spectrom. 2000;11:835–40.
49. Hegeman AD, Schulte CF, Cui Q, Lewis IA, Huttlin EL, Eghbalnia H, Harms AC, Ulrich EL, Markley JL, Sussman MR. Stable isotope assisted assignment of elemental compositions for metabolomics. Anal Chem. 2007; 79:6912–21.
50. Beynon RJ, Pratt JM. Metabolic labeling of proteins for proteomics. Mol Cell Proteomics. 2005;4:857–72.
51. Dumas ME, Maibaum EC, Teague C, Ueshima H, Zhou B, Lindon JC, Nicholson JK, Stamler J, Elliott P, Chan Q, Holmes E. Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. Anal Chem. 2006;78:2199–208.
52. Holmes E, Cloarec O, Nicholson JK. Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: application to HgCl2 toxicity. J Proteome Res. 2006;5:1313–20.
53. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Anal Chem. 2006;78:4430–42.
54. Fan TW. Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures. Prog Nucl Magn Reson Spectrosc. 1996;28:161–219.
55. Viant MR. Improved methods for the acquisition and interpretation of NMR metabolomic data. Biochem Biophys Res Commun. 2003;310:943–8.
56. Kikuchi J, Shinozaki K, Hirayama T. Stable isotope labeling of *Arabidopsis thaliana* for an NMR-based metabolomics approach. Plant Cell Physiol. 2004;45:1099–104.
57. Peterson DJ, Loening NM. QQ-HSQC: a quick, quantitative heteronuclear correlation experiment for NMR spectroscopy. Magn Reson Chem. 2007;45:937–41.
58. Koskela H, Kilpelainen I, Heikkinen S. Some aspects of quantitative 2D NMR. J Magn Reson. 2005;174:237–44.
59. Heikkinen S, Toikka MM, Karhunen PT, Kilpelainen IA. Quantitative 2D HSQC (Q-HSQC) via suppression of J-dependence of polarization transfer in NMR spectroscopy: application to wood lignin. J Am Chem Soc. 2003;125:4362–7.
60. Hyberts SG, Heffron GJ, Tarragona NG, Solanky K, Edmonds KA, Luithardt H, Fejzo J, Chorev M, Aktas H, Colson K, Falchuk KH, Halperin JA, Wagner G. Ultrahigh-resolution (1)H-(13)C HSQC spectra of metabolite mixtures using nonlinear sampling and forward maximum entropy reconstruction. J Am Chem Soc. 2007;129:5108–16.
61. Shrot Y, Frydman L. Spatially resolved multidimensional NMR spectroscopy within a single scan. J Magn Reson. 2004;167:42–8.
62. Ross A, Salzmann M, Senn H. Fast-HMQC using Ernst angle pulses: an efficient tool for screening of ligand binding to target proteins. J Biomol NMR. 1997;10:389–96.
63. Sandusky P, Raftery D. Use of selective TOCSY NMR experiments for quantifying minor components in complex mixtures: application to the metabonomics of amino acids in honey. Anal Chem. 2005;77:2455–63.