Published in final edited form as:

Curr Metabolomics. 2013 ; 1(1): . doi:10.2174/2213235X11301010028.

Databases and Software for NMR-Based Metabolomics

James J. Ellinger, Roger A. Chylla, Eldon L. Ulrich, and John L. Markley*

Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison WI 53706, USA

Abstract

New software and increasingly sophisticated NMR metabolite spectral databases are advancing the unique abilities of NMR spectroscopy to identify and quantify small molecules in solution for studies of metabolite biomarkers and metabolic flux. Public and commercial databases now contain experimental 1D ^1H , ^{13}C and 2D ^1H - ^{13}C spectra and extracted spectral parameters for over a thousand compounds and theoretical data for thousands more. Public databases containing experimental NMR data from complex metabolic studies are emerging. These databases are providing information vital for the construction and testing of new computational algorithms for NMR-based chemometric and quantitative metabolomics studies. In this review we focus on database and software tools that support a quantitative NMR approach to the analysis of 1D and 2D NMR spectra of complex biological mixtures.

Keywords

Chemometrics; databases; quantitative metabolomics; software; nuclear magnetic resonance

Introduction

NMR spectroscopy has for decades been a key technology used successfully by medicinal, natural product, and organic chemists to identify and quantitate both major products and impurities in complex reaction mixtures. More recently, NMR has been used extensively in evaluating and exploring the human metabolome for markers of disease states and the diversity of metabolic pathways in a variety of organisms. As the various 'omics' fields have develop ever growing comprehensive lists of the chemical parts that make up a living organism and its environment, interest is turning to investigations of the temporal flux each component undergoes as the organism goes through life. In turn, these studies are leading to the development of computational simulations of the metabolic evolution of biological systems.

The strengths of NMR spectroscopy in metabolic studies are the variety of compounds that can be detected (charged, neutral, hydrophobic, hydrophilic), often the presence of multiple signals per compound, the ability to quantify the concentrations of individual metabolites in a sample, and the flexibility to handle many different sample types (solvents, buffers, salts, and other excipients). The major limitations of NMR spectroscopy are sensitivity and spectral resolution.

Extensive early studies of metabolic pathways have provided a deep understanding of the compounds involved and their chemical relationships. Many of these compounds can be purchased and NMR data collected on the pure material. By constructing database of the NMR spectra of the pure compounds, issues of spectral resolution can be evaluated, and the richness of each compound spectrum (chemical shifts, peak multiplicity, and peak relative intensity) can be utilized in identifying and quantifying metabolites in metabolic studies.

Over the past ten years a number of NMR spectral techniques have been optimized and applied to metabolic studies including one-dimensional ^1H and ^{13}C and two-dimensional ^1H - ^1H and ^1H - ^{13}C methods. A number of software tools also have been developed to better analyze the NMR spectral data and automate the extraction of quantitative information for spectral feature analysis. In many cases, the software applications make use of the standard NMR metabolite spectral databases.

Two overall approaches exist for analyzing NMR-based metabolomics data: chemometrics and quantitative NMR (see Fig. 1). In the chemometric approach, spectral patterns and intensities are analyzed and compared statistically to yield relevant features that differ and distinguish sample classes. After statistical analysis, metabolites are identified from the spectral patterns of enrichment and depletion. This has the advantage of reducing the amount of work required, because only the important metabolites are identified. In the quantitative approach, all detectable metabolites are formally identified and quantified prior to subsequent analysis. Although the quantitative approach yields more reliable results [1], the chemometric approach is still the most commonly employed mode of analysis, probably because of its speed and ease of use.

Regardless of which method is used, metabolites must eventually be identified within spectra. Traditional methods for identifying a unique compound in a solution sample by NMR spectroscopy are time intensive requiring multiple experiments. For example, about 49 and 45–50 individual metabolites are observable in typical samples of human blood serum [2] and human cerebrospinal fluid [3], respectively. Even a small metabolomics research project will require the analysis of metabolites in twenty or more samples accounting for controls and biological replication. Data analysis is thus a major bottleneck for metabolomics applications [4] where hundreds of samples may be processed daily. To accomplish this in a time-efficient manner, have chosen have chosen to adopt a bioinformatics approach. The ^1H , ^{13}C , ^{31}P and other chemical shifts observed in NMR are reproducible under standardized conditions. As a result, public and commercial databases containing standardized NMR data for individual metabolites have been constructed. The basic compound identification approach is to collect experimental data under conditions that match the standardized conditions used to construct the databases and to compare the experimental chemical shifts to those found in a database. Advanced search and identification algorithms take advantage of metabolite-specific information such as J -coupling, peak multiplicity, and relative peak intensities. Public databases for archiving the experimental data from metabolomics studies are now emerging. These resources will allow the experimental data to be reexamined and validated as well as combined and utilized in new ways over time.

1. NMR Metabolomics Databases

We survey here the major open-access resources available to aid researchers in identifying metabolites present in biological samples. We focus on those offering NMR data (Table 1), although some of these databases contain information from other analytical techniques such as mass spectrometry (MS). In general, data from a given resource have been collected under uniform conditions (i.e. temperature, pH, buffer concentration, etc.), although these

may differ between resources. For example, data in the Biological Magnetic Resonance Data Bank (BMRB) [5], and consequently the Madison-Qingdao Metabolomics Consortium Database (MQMCD) [6], were collected in 99.9% D₂O, 50 mM phosphate buffer, 298 K and at a nominal pH of 7.4, whereas data in the Human Metabolome Database (HMDB) [7] were collected in H₂O, 298 K and at a pH of 7.0. Spectra of metabolites may depend to some extent on the metabolite concentration, pH (those with ionizable groups), and on the magnetic field strength (first order J coupling). The BMRB has an ongoing program of collecting data for the “top 100” metabolites at three different concentrations (0.5, 2, 100 mM) and two magnetic field strengths (500 MHz and 600 MHz for ¹H).

The various databases offer “value-added” material and content in addition to annotated standard NMR spectra. For example, the BMRB, HMDB, MQMCD and Platform for RIKEN Metabolomics (PRIME) [8] provide links for each metabolite to external pathway databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [9-10]. With this information, users can identify relevant pathways for metabolites identified in their samples. Furthermore, the HMDB links metabolites to a MetaboCard, a resource that contains, among other items, specific information acquired from the literature and other databases such as tissue location, average concentrations and associated disorders (for an example of this compilation of information for L-alanine see <http://www.hmdb.ca/metabolites/HMDB00161>).

The Birmingham Metabolite Library (BML-NMR) [11] offers standard spectra collected under two types of water suppression, multiple excitation angles, multiple interscan relaxation delays and a range of pH values (6.6, 7.0 and 7.4). The use of multiple interscan relaxation delays is particularly useful for quantitative NMR in that it may provide a basis for estimating *T*₁ values.

Beyond the curated experimental data derived from pure standards, the BMRB, HMDB, and MQMCD also provide chemical shift data derived from empirical calculation (HMDB and MQMCD) or ab initio quantum calculation (BMRB). The MQMCD also draws in spectra for over 1000 metabolites found in the literature. Furthermore, the BMRB accepts user-submitted standard spectra from raw data (all nuclei observable by NMR are accepted).

The MQMCD provides a useful Java-based applet that links a Jmol structure to its corresponding NMR spectrum (1D ¹H and ¹³C only). By using this tool one can select a peak in the NMR spectrum and see the corresponding atoms highlighted in the Jmol structure (Fig. 2), and vice-versa. This interactive chemical structure and NMR spectrum provides a convenient way for students to learn these relationships.

Searching for metabolites within a database is a straightforward affair that includes multiple options. First, these resources provide simple text-based searches, allowing users to locate metabolites by name, chemical formula or ID numbers (i.e. CAS or PubChem ID). Advanced search features offered by the BMRB, HMDB, and MQMCD include structure-based search such as SMILES, InChI or sketching structures with built-in java web applets. The MetaboHunter package (<http://www.nrcbioinformatics.ca/metabohunter/>) is a web-based resource for identifying metabolites from both the HMDB and MQMCD databases by uploading of either peak lists or entire spectra.

Text- and structure-based queries are useful for extracting metabolite information when specific metabolites have been targeted for study. However, when the majority of a sample's composition is unknown, the true power of these databases lies in the ability to identify and quantify metabolites automatically using peak lists extracted from experimental data. The BMRB, HMDB, MQMCD, and PRIME allow users to upload peak lists generated from experimental spectra. Each database allows the user to select the type of spectrum that was

used to derive the peak list as well as determine chemical shift tolerances. Results are returned as a table with an index indicating likelihood of positive matches. The MQMCD implements a search algorithm that makes use of the entire patterns of the peaks [6], which, when used in conjunction with high resolution, multidimensional NMR spectra provides excellent results.

Some commercial software packages, such as the Chenomx NMR Suite and KnowItAll Metabolomics Edition (discussed below), provide standard spectral libraries and proprietary databases. In the case of both Chenomx (HMDB) and the KnowItAll Metabolomics edition (BMRB), these proprietary databases started with a publicly-available database as a nucleus. Chenomx's spectral library spans metabolite data collected at multiple NMR field strengths and multiple pH values. In general, software with an embedded database does not require peak picking to generate candidate metabolites.

Recently, two resources focusing on specific biofluids were derived and developed from the HMDB. The Human Cerebrospinal Fluid Metabolome (<http://www.csfmetabolome.ca>) [3] and the Human Serum Metabolome (<http://www.serummetabolome.ca>) [12]. These resources have the explicit goal of establishing comprehensive, electronically accessible databases of all the detectable metabolites in human serum and CSF.

Metabolites known to be associated with human serum and CSF were identified by searching the literature. Values for average concentration and associated disease states were also extracted from the literature and included in the databases. To validate the metabolites and concentrations found in the literature, various techniques were used to investigate the metabolite profiles of human serum and CSF samples. Using NMR, 49 and 53 unique metabolites were identified and validated in human serum and CSF, respectively. These tools should provide easy, almost automatic, assignment of NMR spectra of human serum and CSF. The major shortcoming is the apparent inability to download the raw data from the validation NMR experiments. Furthermore, conducting a search from an NMR peak list is done directly through the HMDB NMR search query and does not provide a filter for limiting the results to those metabolites observed in human serum or CSF.

Two other databases designed in a similar fashion are the Yeast Metabolome Database (<http://www.ymdb.ca>) [13] and the *E. coli* Metabolome Database (<http://www.ecmdb.ca>)¹. All standard NMR data found in these databases were obtained from the HMDB or BMRB. A significant drawback is the lack of experimental data confirming information extracted from the literature.

The recently established MetaboLights database (<http://www.ebi.ac.uk/metabolights/>) [14] is a comprehensive, cross-species and cross-technique resource. While primarily functioning as a repository for user-submitted experimental data collected by NMR or MS, it also aims to provide detailed information regarding metabolite locations and concentrations. MetaboLights also captures metadata that describes experimental conditions and assays.

Experimental data are uploaded through a web-based submission, which also contains the option to update an existing study. Identifying available datasets is accomplished through a standard text-based search bar or through browsing the collection. For browsing, options to filter by technology (NMR or MS) and organism are available. Finally, the study description, protocols and all data associated with an experiment can be downloaded as a single, compressed (.zip) file.

---

¹ECMDB: The *E. coli* Metabolome Database. Guo AC, Jewison T, Djoumbou Y and Wishart DS. *Nucleic Acids Res. Submission in Progress*

The current collection of experimental datasets (only 9 exist in the database) contains submissions from a variety of model organisms including *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Homo sapiens*, *Saccharomyces cerevisiae*, and *Triticum aestivum*. MetaboLights provides a resource that establishes a pipeline for collaborations within the metabolomics community as well as a platform to initiate a set of standards for data collection and representation. Metabolomics applications cover areas such as methods development, diabetes, and growth control of eukaryotic cells. However, there does not currently appear to be a mechanism for data validation or quality control, a problem for which a solution may not be reached until a set of standards is agreed upon.

2. Metabolomics Analysis Software

In an NMR-based metabolomics investigation, the principal analysis task is to detect and measure changes in levels of relevant biological compounds (or some set of relevant “features” associated with those changes) across a series of related spectra. Ideally, the related spectra are of the same experiment type collected with identical acquisition parameters and processed in an identical manner. If the metabolomics study is a group-based study (*e.g.* transgenic study), the related spectra consist of groups of sample replicates, where each group is associated with a known sample attribute (*e.g.* genotype). In a metabolic flux study, the related spectra correspond to elapsed time points after some perturbation to the sample (*e.g.* the time after addition of ^{13}C -glucose). In this review we refer to any set of related spectra as an “ensemble”.

As discussed above, the two popular approaches to NMR metabolomics analysis are spectral profiling (chemometrics) [15] and spectral identification and quantification [4]. The NMR software packages for metabolomics analysis (see Table 2 for complete list of software cited in this review) are generally oriented toward just one of these approaches. Fig. (1) summarizes the associated workflow.

Chemometrics is the most extensively used approach in NMR-based metabolomics. Chemometrics uses multivariate statistical analysis such as principal component analysis (PCA) [16] to detect variances of features within a spectral ensemble. The term “feature” here is used specifically to refer to a scalar value that represents the intensity of some relevant entity within a spectrum. The abstraction of a “feature” from the data set is an attractive aspect of the chemometric approach because spectral interpretation (*e.g.* assigning a peak or a spectral region to a set of NMR transitions) is not required in order to identify relevant changes within the ensemble. The term “untargeted profiling” can be associated with chemometrics. In untargeted profiling, features are detected without knowledge of the chemical shift entities giving rise to those patterns. Software packages containing statistical tools useful for untargeted metabolite profiling and chemometrics are summarized in Table 3. Although not specific to metabolomics, a number of packages for the R statistical computing language are also available for chemometric analysis of NMR data (<http://cran.at.r-project.org/web/views/ChemPhys.html>).

An additional set of NMR software tools (summarized in Table 4) are designed for spectral identification and quantification, *i.e.* the specific identification and quantification of chemical shift entities (CSE) and the modulation in levels of those CSEs within a spectral ensemble. The identification portion is essentially an assignment problem where the type of NMR experiment (1D ^1H , 2D ^1H - ^{13}C HSQC, 2D ^1H - ^1H TOCSY, etc.), set of observed peak characteristics (chemical shift region, peak pattern, intensity), and set of sample conditions (field strength, pH, temperature, etc.) are matched against the available known transitions of standard compounds. These known transitions are obtained from a metabolomics database (such as those discussed above) populated with measurements of standards under controlled

sample conditions, but “known” transitions can also be derived from theoretically calculated chemical shifts. The popular commercial NMR software packages for identification are thus often tightly coupled to a dedicated database for that software (listed in Table 3). The quantification portion of targeted profiling consists of measurement of the intensity of each signal and its profile across the ensemble. Although the “identify” task is usually thought of as a prerequisite to the “quantify” task, a modeling approach such as deconvolution [17-19] may quantify signals and modulations of their intensity *prior to* the signals being assigned to an NMR transition.

Fig. (2) displays an overview of the analysis steps involved in NMR metabolomics. Although untargeted approaches and targeted methods are usually thought of as alternative approaches, the abstraction of a “feature” as shown in the figure allows the integration of these approaches such that identification and quantification can be viewed as analysis steps that create a specific feature set (*i.e.* concentration levels of CSE's). Any such feature set is then amenable to further analysis by multivariate methods. The remainder of this review discusses the workflow of metabolomics data analysis as specific analysis steps and lists the NMR software packages with functionality relevant to those steps.

2.1. Spectral Processing

The transformation of the time-domain free induction decay (FID) into the frequency domain spectrum is the initial step of all NMR analysis in metabolomics. These steps are straightforward and generally consist of apodization, zero-filling, Fourier transformation, and phase correction. The software packages provided by the spectrometer vendors as well as the freely available NMRPipe [20] software package all contain myriad spectral processing tools that support both 1D and 2D processing. Most of the commercial software packages dedicated to metabolomics also have built-in spectral processing for 1D experiments. For 2D processing, an additional capability that is commonly used is linear prediction to extend a truncated FID along the indirectly detected dimension. An alternative, is to use non-linear sampling and signal reconstruction [21].

2.2. Post Processing: Baseline Correction

Post-processing steps may be necessary to remove certain artifacts or perform spectral enhancement. The most common post-processing method is baseline correction, which involves the removal of low frequency artifacts from the absorption spectrum. Methods of baseline correction involve fitting a polynomial or other suitable analytical function to a set of points associated with signal-free regions of the spectrum and then subtracting the interpolated function from the processed spectrum. Removal of such artifacts is critical to yield accurate results (*e.g.* error of 1.5% or less) by any subsequent method of quantitative analysis [22]. All of the spectral processing programs (Table 3) contain implementations of baseline correction.

2.3. Post-Processing: Reference Deconvolution

In the technique of “reference deconvolution” [23], line shape enhancement is achieved by extracting a signal of a known reference material from an experimental spectrum and using a comparison between the experimental reference signal and that predicted by theory to correct the full experimental spectrum. The commercial metabolomics package Chenomx uses reference deconvolution extensively to provide a more constant line shape to all peaks. This is done so that the shape of a reference signal, such as an internal chemical shift reference, can be used as a prototype line shape for modeling peak transitions of all other signals.

2.4. Normalization

Normalization refers to a set of techniques that are designed to scale or adjust a spectrum or derived features to make more meaningful comparisons between samples. The need for “row-based” normalization is primarily driven by the need to scale each feature according to some measure of the sample material. The need for “column-based” normalization is to compare different features which may have dynamic range differences on the same scale. In general, one can perform normalization by the amount of sample material either by spectral-based methods prior to feature extraction (usually associated with spectral identification/quantification methods) or by row-based methods after feature extraction (usually associated with chemometric approaches). The various normalization methods and their proper usage are summarized in Table 5. Normalization is one area where care should be exercised in determining the appropriate usage. For example, “log-scaling” is a popular technique within bioinformatics for reducing dynamic range and stabilizing variances within a data set. It is well justified in cases where the amount of noise in a sample grows larger when the signal grows larger. This is not usually the case for NMR and thus indiscriminate log-scaling leads to a distortion of the signal/noise.

2.5. Feature Generation

The term “feature” as defined above refers to a scalar value that represents the intensity of some relevant entity within a spectrum. In the framework of multivariate analysis, the results of a multi-spectrum NMR metabolomics experiment can be represented by a matrix where each cell holds the intensity, the column index encodes the feature identity (e.g. index of a peak), and the row encodes the index of the spectrum. Any row vector from the matrix thus corresponds to an array of feature values for a single spectrum and any column vector yields the profile of a feature across spectra.

The various features used in NMR metabolomics analysis include peaks, bins, integrals, signal amplitudes, region of interest-based (ROI-based) integrals, ROI-based amplitudes, and concentration of chemical entity (see Table 6). In some sense, all of the features can be considered derived values for the concentration of the underlying chemical entity. The choice of a particular spectral feature depends on considerations such as the ease of calculation, speed of calculation, ease of interpretation, and robustness to complicating experimental conditions including signal overlap, signal/noise, dynamic range, acquisition artifacts, and variance in chemical shifts and line widths within the ensemble.

- **Peak:** A peak refers to any intensity value within a frequency spectrum that satisfies a threshold requirement and corresponds to a local extremum. The inherent sharpness of NMR signals makes peak analysis useful. The strengths of peak analysis reflect the ease and speed of calculation as peak detection and filtering algorithms are ubiquitous to all NMR software processing packages. The weakness of peak analysis is that signal overlap can obscure local extrema and peak heights are not proportional to signal intensity across signals of changing line width and line shape.
- **Spectral bin:** A bin is the sum of intensity values for a fixed number of consecutive points. Binning is a common practice in chemometric approaches [24] for the following considerations: i) the size of the feature set is reduced; ii) chemical shift “drift” is accounted for; and iii) the bin calculation is easy and rapid. Typically, a binned spectrum is subjected to a threshold test to remove bins below a given threshold. Pitfalls of binning include loss of spectral resolution, susceptibility to noise artifacts, and artifacts produced when the boundary of a bin happens to lie on the center of a peak. The chemometric packages that offer specialized features for

spectral binning such as spectral alignment include the Bruker Amix package and the MetaboAnalyst package.

- **Integral:** An integral refers to the area under a given peak, which is proportional to signal intensity even across signals of different line shapes. As with peak analysis, the calculation of integrals is rapid and straightforward by methods available in all NMR processing software. The primary drawback to integral analysis is the ambiguity of determining the boundaries (footprint) of an integral and its limitation to the accurate calculation of well resolved signals or the sum of well resolved signals. The ambiguity in determining a footprint usually means that ROI-integrals are used more prevalently in NMR metabolomics particularly for 2D experiments.
- **ROI-integral:** An ROI-based integral is simply the sum of all intensities existing with a fixed chemical shift region. The strength of ROI-based integrals is that the center of the signal can drift and still be captured within the ROI footprint. ROI's can be defined once and reused across samples of the same experiment type. The existence of multiple overlapping resonances in high resolution ^1H NMR renders ROI-integrals much more useful in 2D metabolomic studies. The NMR assignment package rNMR [25] uses ROI's as a foundation for implementing graphical tools for ROI based assignment (see Fig. 3).
- **Signal Amplitude:** A signal refers to the construction of a mathematical model where the line shape is modeled by a basis function and the intensity is given by its scalar coefficient (amplitude). As with integral analysis, a signal amplitude can be used to accurately compare the intensity of signals with different line shapes. Methods of spectral deconvolution can also yield relatively accurate values for signal intensities even in the presence of significant signal overlap. The prevalence of signal analysis in NMR spectroscopy has been limited by the inherent complexity and need for specialized software for NMR spectral deconvolution. NMR software packages that perform mathematical modeling of spectra for NMR-based metabolomics include most of the packages listed in Table 4.
- **ROI-Amplitude:** An ROI-based amplitude is simply the sum of signal amplitudes that occur within a given chemical shift region. ROI-based amplitudes are used to associate signals with spectral transitions (assignments) and group the intensity of signals belonging to the same NMR transition. As with ROI-based integrals, ROI-based amplitudes are much more useful for 2D than 1D analysis of complex biological mixtures. The Newton software package for fast maximum likelihood reconstruction of 1D and 2D NMR spectra [19], for example (Fig. 4), makes extensive use of ROI-based amplitudes in performing quantitative analysis of “time-zero” extrapolated 2D HSQC spectra [26].
- **Molar concentration of species:** The molar concentration of the species is the true metric of interest in a metabolomics investigation. The primary difficulty in using molar concentration is that it requires accurate interpretation of numerous features of a complex spectrum. The most common method of deriving molar concentrations is the “landmark peak” method where a peak/signal can be unambiguously assigned to a single transition of a chemical entity. Because some metabolites do not contain a landmark peak with the 1D spectrum of a complex biological mixture, this method by itself cannot be used to quantify all observable metabolites. Identification and quantification of additional chemical entities requires a combination of mathematical modeling and a detailed spectral library as found in dedicated commercial packages such as Chenomx NMR Suite and KnowItAll Metabolomics Edition (BioRad Corporation).

2.5. Multivariate Statistics

Once a set of features has been extracted for an ensemble of NMR data, numerous statistical packages both specific and non-specific to NMR are suitable for performing various types of statistical analysis, visualizing the results, and performing tests of significance. In addition to the software packages specific to NMR statistical analysis (Table 3), the statistical computing language R (<http://www.r-project.org>) and Matlab (<http://www.mathworks.com/products/matlab/>) contain tools for plotting and visualization. R is freely available and is distributed under the GNU GPL version 2 and 3 licenses. Matlab is commercial software that offers licenses for academic and student use.

The most fully featured package for statistical analysis related to metabolomics is the web-based MetaboAnalyst [27]. Data analysis functionality within the package (Fig. 5) includes multi-group data analysis, two-factor analysis and time-series data analysis. The package has additional related modules: (i) a quality-control module that allows users to evaluate their data quality before conducting any analysis, (ii) a functional enrichment analysis module that allows users to identify biologically meaningful patterns, and (iii) a metabolic pathway analysis module that allows users to perform pathway analysis and visualization for 15 different model organisms.

Acknowledgments

Supported by National Institutes of Health Grants P41 LM05799 and P41 GM103399.

References

1. Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM. Targeted profiling: quantitative analysis of 1H NMR metabolomics data. *Anal Chem*. 2006; 78(13):4430–4442. [PubMed: 16808451]
2. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, Sinelnikov I, Krishnamurthy R, Eisner R, Gautam B, Young N, Xia J, Knox C, Dong E, Huang P, Hollander Z, Pedersen TL, Smith SR, Bamforth F, Greiner R, McManus B, Newman JW, Goodfriend T, Wishart DS. The Human Serum Metabolome. *Plos One*. 2011; 6(2):e16957–e16957. [PubMed: 21359215]
3. Wishart DS, Lewis MJ, Morrissey JA, Flegel MD, Jeroncic K, Xiong Y, Cheng D, Eisner R, Gautam B, Tzur D, Sawhney S, Bamforth F, Greiner R, Li L. The human cerebrospinal fluid metabolome. *Journal of Chromatography B*. 2008; 871(2):164–173.
4. Wishart DS. Quantitative metabolomics using NMR. *Trends Anal Chem*. 2008; 27(3):228–237.
5. Markley JL, Anderson ME, Cui Q, Eghbalian 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*. 2007:157–168. [PubMed: 17990489]
6. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalian HR, Sussman MR, Markley JL. Metabolite identification via the Madison Metabolomics Consortium Database. *Nat Biotechnol*. 2008; 26(2):162–164. [PubMed: 18259166]
7. 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(Database issue):D521–526. [PubMed: 17202168]
8. Akiyama K, Chikayama E, Yuasa H, Shimada Y, Tohge T, Shinozaki K, Hirai MY, Sakurai T, Kikuchi J, Saito K. PRIME: A Web Site That Assembles Tools for Metabolomics and Transcriptomics. *In Silico Biology*. 2008; 8(3):339–345. [PubMed: 19032166]
9. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 1999; 27(1):29–34. [PubMed: 9847135]

10. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* 2006; 34(90001):D354–D357. [PubMed: 16381885]
11. Ludwig C, Easton J, Lodi A, Tiziani S, Manzoor S, Southam A, Byrne J, Bishop L, He S, Arvanitis T, Günther U, Viant M. Birmingham Metabolite Library: a publicly accessible database of 1-D ¹H and 2-D ¹H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). *Metabolomics.* 2012; 8(1):8–18.
12. Belyaeva EA, Dymkowska D, Wieckowski MR, Wojtczak L. Mitochondria as an important target in heavy metal toxicity in rat hepatoma AS-30D cells. *Toxicol Appl Pharmacol.* 2008; 231(1):34–42. [PubMed: 18501399]
13. Jewison T, Knox C, Neveu V, Djoumbou Y, Guo AC, Lee J, Liu P, Mandal R, Krishnamurthy R, Sinelnikov I, Wilson M, Wishart DS. YMDB: the Yeast Metabolome Database. *Nucleic Acids Res.* 2011; 40(D1):D815–D820. [PubMed: 22064855]
14. Steinbeck C, Kuhn S, Jayaseelan K, Moreno P. Computational metabolomics – a field at the boundaries of cheminformatics and bioinformatics. *Journal of Cheminformatics.* 2011; 3(0):1–1. [PubMed: 21214931]
15. Lindon JC, Nicholson JK, Holmes E, Everett JR. Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids. *Concepts in Magnetic Resonance.* 2000; 12(5):289–320.
16. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol.* 1933; 24(6):417–441.
17. Miller MI, Greene AS. Maximum-Likelihood Estimation for Nuclear Magnetic Resonance Spectroscopy. *J Magn Reson.* 1989; 83:525–548.
18. Chylla RA, Markley JL. Theory and application of the maximum likelihood principle to NMR parameter estimation of multidimensional NMR data. *J Biomol NMR.* 1995; 5(3):245–258. [PubMed: 7787422]
19. Chylla RA, Hu K, Ellinger JJ, Markley JL. Deconvolution of two-dimensional NMR spectra by fast maximum likelihood reconstruction: application to quantitative metabolomics. *Anal Chem.* 2011; 83(12):4871–4880. [PubMed: 21526800]
20. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPIPE - A Multidimensional Spectral Processing System Based on UNIX Pipes. *J Biomol NMR.* 1995; 6(3):277–293. [PubMed: 8520220]
21. 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(16):5108–5116. [PubMed: 17388596]
22. Malz F, Jancke H. Validation of quantitative NMR. *J Pharm Biomed Anal.* 2005; 38(5):813–823. [PubMed: 15893442]
23. Morris GA. Compensation of instrumental imperfections by deconvolution using an internal reference signal. *Journal of Magnetic Resonance (1969).* 1988; 80(3):547–552.
24. Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Anal Chem.* 2006; 78(7):2262–2267. [PubMed: 16579606]
25. Lewis IA, Schommer SC, Markley JL. rNMR: open source software for identifying and quantifying metabolites in NMR spectra. *Magn Reson Chem.* 2009; 47:s123–s126. [PubMed: 19821464]
26. Hu K, Ellinger JJ, Chylla RA, Markley JL. Measurement of Absolute Concentrations of Individual Compounds in Metabolite Mixtures by Gradient-Selective Time-Zero ¹H-¹³C HSQC with Two Concentration References and Fast Maximum Likelihood Reconstruction Analysis. *Anal Chem.* 2011; 83(24):9352–9360. [PubMed: 22029275]
27. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0—A Comprehensive Server for Metabolomic Data Analysis. *Nucleic Acids Res.* 2012; 40(14):W1.
28. Bingol K, Zhang F, Bruschweiler-Li L, Bruschweiler R. TOCCATA: A Customized Carbon Total Correlation Spectroscopy NMR Metabolomics Database. *Anal Chem.* 2012

29. Lundberg, P.; Vogel, T.; Malusek, A.; Lundquist, P.; Cohen, L.; Dahlqvist, O. MDL - The Magnetic Resonance Metabolomics Database (mdl.imv.liu.se). ESMRMB 2005 Congress; Basel, Switzerland. 2005.
30. Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ¹H NMR Metabonomics. Anal Chem. 2006; 78(13):4281–4290. [PubMed: 16808434]

Fig. (1).

Overview of NMR metabolomics analysis workflow: The figure integrates the workflow involved in NMR metabolomics data analysis by both a chemometrics and “identify and quantify” approach. In a typical chemometrics approach, the workflow proceeds from data collection, processing, binning, and then multivariate analysis to identify spectral patterns of change. Metabolite identification, if performed at all, will occur after multivariate analysis. In an “identify and quantify” approach, the step of spectral identification occurs much earlier in the workflow, either after spectral processing or region of interest (ROI) segmentation. Model fitting is a technique that can be used in either workflow to create amplitudes to be used in the feature matrix. All of the features in the feature matrix (see Table 6) are actually derived values from the metabolite concentration.

Fig. (2). L-alanine as an example of an interactive ^1H -NMR session in the MQMCD. The doublet at 1.47 ppm was selected in the NMR spectrum, and the corresponding methyl protons were highlighted in the Jmol structure (yellow atoms).

Fig. (3). Spectrum from an *E. coli* extract (top) demonstrating use of regions of interest (ROIs) with rNMR [25]. Array of ROIs (red boxes) for multiple metabolites from a series of *E. coli* extracts (bottom) from rNMR.

Fig. (4).

Fast maximum likelihood reconstruction (FMLR) of a 2D ^1H - ^{13}C spectrum of liver extract performed by the Newton software package [19]. (Left) contour plot of a region of the ^1H - ^{13}C HSQC spectrum. (Middle) The FMLR reconstruction of the region. (Right) The corresponding residual. Annotations on the spectrum denote the centers of signals that were identified by FMLR. Signals from glucose (A) are much higher than those nearby from proline (B), and fructose (C,D,E). The volume of the observable residual in region A is less than 3% of the volume of the peaks. The amplitudes of the signals in the reconstruction are proportional to the concentration of the underlying species.

Fig. (5).

Overview of functionality and modules in the MetaboAnalyst 2.0 Package. Figure reproduced from [27].

Table 1
Publicly Available Databases and Resources that Provide a Useful Starting Point for NMR-based Metabolomics Data Analysis. Among the Databases, Search Methods Include, But are not Limited to, Text Query, Chemical Formula, and Chemical Shift Peak Lists

Database	URL	Number of compounds*	Types of Spectra	Predicted spectra?	Can download raw data?
Biological Magnetic Resonance Bank (BMRB) [5]	http://www.bmrwisc.edu/metabolomics	906	¹ H, ¹³ C, DEPT90, DEPT 135, ¹ H J-resolved, ¹ H- ¹³ C HSQC, ¹ H- ¹³ C HMBC, ¹ H- ¹ H TOCSY, ¹ H- ¹ H COSY	Yes	Yes
Human Metabolome Database (HMDB) [7]	http://www.hmdb.ca	916	¹ H, ¹³ C HSQC	Yes	Yes
Madison-Qindao Metabolomics Consortium Database (MQMCD) [6]	http://mmcd.nmr.fam.wisc.edu [†]	794	¹ H, ¹³ C, DEPT90, DEPT 135, ¹ H- ¹³ C HSQC, ¹ H- ¹³ C HMBC, ¹ H- ¹ H TOCSY, ¹ H- ¹ H COSY	Yes	Yes (via BMRB)
Birmingham Metabolite Library (BML-NMR) [11]	http://www.bml-nmr.org	208	¹ H, ¹ H J-resolved	No	Yes
Platform for RIKEN Metabolomics via SpinAssign (PRIME) [8]	http://prime.psc.riken.jp	80	¹ H, ¹³ C, ¹ H- ¹³ C HSQC	No	No
TOCSY Customized Carbon Trace Archive (TOCCATA) [28]	http://spinportal.magnet.fsu.edu/toccatat/webquery.html	463	¹³ C- ¹³ C TOCSY	No	No
Magnetic Resonance Metabolomics Database (MRMD) [29]	http://www.liu.se/hu/mdl/main/	350	¹ H, HMQC, HMQC-TOCSY, DQF-COSY, TOCSY	No	No
NMRShiftDB	http://nmrshiftdb.nmr.uni-koeln.de/	>40,000	¹ H, ¹³ C	No	Yes

* Represents only the number of compounds available from standard experimental NMR spectra (some of the databases also include MS data).

[†] Spectral data derived from BMRB

Table 2
NMR Software Packages (Alphabetical Order) Cited in this Review

Package	Source	URL
ACD	ACD Labs	http://www.acdlabs.com/resources/freeware/nmr_proc/
AMIX tool-kit	Bruker Topspin	http://www.bruker-biospin.com/amix.html
Automics	Softpedia	http://www.softpedia.com/get/Science-CAD/Automics.shtml
CCPN Metabolomics	CCPN	http://www.ccpn.ac.uk/collaborations/metabolomics
Chenomx NMR Suite	Chenomx	http://www.chenomx.com/software/
COLMAR	Florida State University	http://spinportal.magnet.fsu.edu/
dataChord Spectrum Miner	One Moon Scientific	http://www.onemoonscientific.com/dcsn/summary.html
KnowItAll Metabolomics	BioRad Corporation	http://www.bio-rad.com/.../KnowItAll-Metabolomics-Edition-Software
MetaboAnalyst	University of Alberta	http://www.metaboanalyst.ca/
MetaboHunter	Natl Res Council of CA	http://www.nrcbioinformatics.ca/metabohunter/
MetaboLab	University of Alberta	http://beregond.bham.ac.uk/nmrlab1d/metabolab.html
MetaboMiner	University of Alberta	http://wishart.biology.ualberta.ca/metabominer/
MNOVA	Mestrelab	http://mestrelab.com/software/mnova-nmr/
Newton	NMRFAM	http://newton.nmrfam.wisc.edu
NMR-Pipe	National Institutes of Health (NIH)	http://spin.niddk.nih.gov/NMRPipe/
NUTS	Acorn Software	http://www.acornnmr.com/nuts.htm
PRIME: SpinAssign	Platform for Riken Metabolomics	http://prime.psc.riken.jp/?action=standard_index
rNMR	NMRFAM	http://rnmr.nmrfam.wisc.edu
TopSpin	Bruker Topspin	http://www.bruker-biospin.com/topspin3-dir.html
vNMRJ	Agilent	http://www.chem.agilent.com/en-US/products-services/Software-Informatics/VnmrJ-30

Table 3
Metabolomics Software with Statistical Approach (Chemometrics)

Software Package	Deployment	Commercial	Free	1D	2D	Spectrum Grouping	Spectral Binning	Multivariate	Statistical Tests	Pathway Analysis
AMIX tool-kit	Client	✓		✓			✓	✓	✓	
dataChord Spectrum Miner	Client	✓		✓		✓	✓	✓	✓	
MetaboLab	Matlab		✓	✓			✓			
Automics	Client		✓	✓		✓	✓	✓	✓	
MetaboAnalyst 2.0	Web		✓	✓			✓	✓	✓	✓
KnowItAll Metabolomics Edition	Client	✓		✓			✓	✓	✓	

Table 4
Metabolomics Software for Spectral Identification & Quantification

Software Package	Deployment	Commercial	Free	1D	2D	Spectral Fitting	Embedded Database	Automated Identification	Quantification
Chenomx NMR Suite	Client	✓		✓		✓	✓	✓	✓
"Know It All" Metabolomics	Client	✓		✓		✓	✓	✓	✓
rNMR	Client		✓	✓	✓				✓
Newton	Client		✓	✓	✓				✓
MetaboMiner	Client		✓		✓		✓	✓	
CCPN Metabolomics	Client		✓	✓	✓		✓	✓	
*COLMAR	Web		✓		✓	✓		✓	✓

* Specialized package for 2D covariance spectroscopy

Table 5
Methods of Data Normalization and Desired Usage

Type	Method	Spectral Equivalent	Usage
Row-based	Normalization by sum: normalize by the sum of all values in the row	Normalize each spectrum by its own integral	Express each value as percentage of the total
	Normalize by a reference sample	Normalize each spectrum by the ratio of the integral of the spectrum to the integral of a reference spectrum	Normalize by amount of sample material
	Normalize by a reference feature	Normalize by the integral of a reference signal (<i>e.g.</i> DSS)	Normalize according to concentration of internal standard
	Normalize by probabilistic quotient [30]	Normalize by the median quotient relative to a reference	Used to account for dilution effects when integral is not representative of sample dilution.
Column-based	Log-based	Replace each value by the log of its value	Compresses dynamic range: Not generally appropriate for NMR where noise does not increase with signal
	Auto-scaling: Mean centered and divided by some metric related to the range or standard deviation of the feature		Normalize each feature change relative to the variance so as to measure statistical significance of the change (t-score)

Table 6
Features Used to Describe Intensity of NMR Signals

Feature	Description	Usage	Related Software
Peak	Intensity of a local maximum	Useful for comparing intensities of signals with similar line widths	All spectral processing software
Spectral bin	Sum of intensities within bins of constant width	Useful for chemometric approaches	Chemometric packages (see Table 4)
Integral	Sum of intensities within a specified footprint	Useful for comparing intensities of well-resolved signals with different line widths	All spectral processing software
ROI Intensity	Integral where the footprint is a rectangular region	Useful when ROI assignment is unambiguous (e.g. 2D NMR)	rNMR, TopSpin, Newton
Signal Amplitude	Scalar amplitude of signal found from model fitting	Used with model fitting methods when signals overlap	Any package with model fitting
ROI Amplitude	Sum of signal amplitudes within an ROI	Useful with model fitting methods when ROI assignment is unambiguous	Newton, rNMR
Molar concentration	Molar concentration of a molecular species	Best metric if known	NA