J Biomol NMR (2011) 49:307–323
DOI 10.1007/s10858-011-9480-x

ARTICLE

Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra

Pascal Mercier · Michael J. Lewis ·
David Chang · David Baker · David S. Wishart

Received: 27 September 2010 / Accepted: 29 November 2010 / Published online: 1 March 2011

Abstract Nuclear magnetic resonance (NMR) and Mass Spectroscopy (MS) are the two most common spectroscopic analytical techniques employed in metabolomics. The large spectral datasets generated by NMR and MS are often analyzed using data reduction techniques like Principal Component Analysis (PCA). Although rapid, these methods are susceptible to solvent and matrix effects, high rates of false positives, lack of reproducibility and limited data transferability from one platform to the next. Given these limitations, a growing trend in both NMR and MS-based metabolomics is towards targeted profiling or "quantitative" metabolomics, wherein compounds are identified and quantified via spectral fitting prior to any statistical analysis. Despite the obvious advantages of this method, targeted profiling is hindered by the time required to perform manual or computer-assisted spectral fitting. In an effort to increase data analysis throughput for NMR-based metabolomics, we have developed an automatic method for identifying and quantifying metabolites in one-dimensional (1D) proton NMR spectra. This new algorithm is capable of using carefully constructed reference spectra and optimizing thousands of variables to reconstruct experimental NMR spectra of biofluids using rules and concepts derived from physical chemistry and NMR theory. The automated profiling program has been tested against spectra of synthetic mixtures as well as biological spectra of urine, serum and cerebral spinal fluid (CSF). Our results indicate that the algorithm can correctly identify compounds with high fidelity in each biofluid sample (except for urine). Furthermore, the metabolite concentrations exhibit a very high correlation with both simulated and manually-detected values.

Keywords Metabolomics · Nuclear magnetic resonance · Targeted profiling · Automated targeted spectral profiling

P. Mercier · M. J. Lewis · D. Chang
Chenomx Inc, Edmonton, AB T5K 2J1, Canada

D. S. Wishart
Department of Computing Science and Biological Sciences,
University of Alberta, Edmonton, AB T6G 2E8, Canada

D. Baker
Pfizer Inc, Groton, CT, USA

D. Chang (✉)
Director of Products and Services, Suite 800, 10050 – 112 St,
Edmonton, AB T5K 2J1, Canada
e-mail: dchang@chenomx.com

Introduction

Metabolomics is the field of omics science concerned with the comprehensive characterization of small molecule metabolites found in cells, tissues, biofluids and organisms. It uses a variety of analytical chemistry techniques to specifically identify metabolites or generate metabolic spectral profiles. Because metabolomics is concerned with looking at the small molecule products of gene, protein and environmental interactions, it provides complementary information to what is normally obtained via genomics, transcriptomics and proteomics. As a consequence, metabolomics has found a niche in a wide variety of research fields, such as nutrition (Cevallos-Cevallos et al. 2009; Wishart 2008b; Scalbert et al. 2009; Hall et al. 2008), oncology (Odunsi 2007; Spratlin et al. 2009; Tiziani et al. 2009; Tainsky 2009; Beckonert et al. 2010; Woo et al. 2009; Kim et al. 2008; Bezabeh et al. 2009; Griffin and Kauppinen 2007; Oakman et al. 2010; Kim et al. 2009; Li et al. 2008), biomarker discovery (Vangala and Tonelli 2007; Fonville et al. 2010; Woo et al. 2009; Kim et al. 2008; Quinones and Kaddurah-Daouk 2009; Wolfender et al.

2009; Serkova and Niemann 2006; Bertram et al. 2009; Kristal et al. 2007; Kim et al. 2009; Weiss et al. 2008), disease diagnosis (Odunsi 2007; Tiziani et al. 2009; Sinclair et al. 2010; Kimura et al. 2009; Schiffmann et al. 2010; Young and Wallace 2009; Waterman et al. 2010; Quinones and Kaddurah-Daouk 2009; Kaddurah-Daouk et al. 2008; Aich et al. 2009; Bezabeh et al. 2009; Sinclair et al. 2010) and drug development (Wishart 2008a; Vangala and Tonelli 2007; Chen et al. 2007a; Kaddurah-Daouk et al. 2008; Keun 2006). Nuclear magnetic resonance (NMR) spectroscopy and Mass Spectroscopy (MS) are the two most common spectroscopic analytical techniques employed in metabolomics. NMR is quantitative, highly reproducible, non-selective and non-destructive, but lacks the sensitivity of MS. On the other hand, MS is not particularly quantitative and tends to be somewhat more selective. For these reasons, NMR and MS have often been used in a complementary fashion (Lindon and Nicholson 2008).

Over the past decade, numerous data processing and statistical treatments have been explored to handle the large amount of data generated by metabolomic studies. Spectral alignment techniques (Veselkov et al. 2009; Staab et al. 2010) followed by multivariate statistical approaches, such as Principal Component Analysis (PCA), or supervised classification methods such as Partial Least Squares Discriminant Analysis (PLS-DA), are commonly used to quickly identify spectral regions or spectral patterns of interest. With these chemometric methods, the compounds of interest are only identified and quantified after the statistical analyses have been performed. In many cases, the compound identification steps (which are manually intensive) often prove too difficult because information about the compound(s) identity has been lost as a consequence of the spectral reduction or spectral alignment process. On the other hand, quantitative or targeted profiling (Weljie et al. 2006; Wishart 2008c) offers an alternative route to spectral reduction techniques such as binning, where experimental spectra are reconstituted at their full resolution from a sum of their underlying components using a reference compound library. In this approach, compounds are identified and quantified prior to performing any kind of multivariate statistical analyses. The challenge in quantitative metabolomics lies in the time and effort needed to identify and quantify compounds in biofluid mixtures. Additional details on the origins, context of use, and practical aspects of each approach (quantitative or targeted vs. chemometric) have been covered in the past (Wishart 2008c; Chang et al. 2007b) and will not be repeated here.

This paper attempts to address one of the bottlenecks that hinder both quantitative and chemometric metabolomics, i.e., compound identification and quantification (Milgram and Nordstrom 2009). Current approaches are almost all manual or at best, semi-automated (Chenomx 2010; Xia et al. 2008; Cui et al. 2008; Markley et al. 2007). In this context, the work presented in this paper focuses strictly on the automated spectral deconvolution of high-resolution one-dimensional ^{1}H-NMR spectra using a reference compound library as prior knowledge. Here we will describe the algorithm (called AutoFit) in detail and assess its performance in analyzing NMR spectra of synthetic mixtures as well as biological spectra of serum, plasma and cerebral spinal fluid (CSF). We will also compare the performance (in term of speed and accuracy) of AutoFit against the results obtained using a more established manual approach.

Experimental procedures

Sample preparation

The lumbar cerebrospinal fluid (CSF) samples utilized in this study were from previously published research (Wishart et al. 2008). The spectra were not reacquired but simply manually refit using a pH-sensitive compound library (see below).

The collection and preparation of the human serum samples used for this work was somewhat more complex and is described here in more detail. Prior to collecting blood samples, all blood donors were approached using approved ethical guidelines and those who agreed to donate were required to sign consent forms. All blood samples were collected via standard overnight fasting, vein-puncture methods and stored in untreated (no heparin or EDTA) tubes. The samples were subsequently spun down for 10 min at 1,600 g at 4°C and the serum decanted into clean glass tubes and frozen to −80°C within 2 h to minimize any possible metabolite degradation. All serum samples were thawed on ice for approximately 2 h before use. To remove large molecular weight proteins and lipid particles, the serum samples were filtered via ultracentrifugation. Prior to filtration, two 0.5 ml, 3 KDa cut-off centrifugal filter units (Millipore Microcon YM-3) were rinsed four times each with 0.5 ml of H_2O and centrifuged at 11,000 rpm for 1 h, to remove residual glycerol bound to the filter membranes. Two 150 µl aliquots of each serum sample were then transferred into the two centrifuge filter devices. The samples were then spun at a rate of 11,000 rpm for 140 min, to remove macromolecules (primarily proteins and lipoproteins) from the sample. The subsequent filtrates were then checked visually for a red tint as an indication that the membrane was compromised. For those samples where the membrane was compromised the filtration process was repeated with a different filter and the filtrate was inspected again. The subsequent filtrates were pooled and the volume was recorded. If the total volume of the sample was under 300 μL an appropriate amount from a 50 mM NaH_2PO_4 buffer (pH 7) was added to the sample until the total volume was 300 μl. Subsequently, 35 μl of D_2O and 15 μl of a standard buffer solution (11.7 mM DSS [disodium-2,2-dimethyl-2-silapentane-5-sulphonate], 730 mM imidazole, and 0.47% NaN_3 in H_2O) was added to the sample. The serum sample (350 μl) was then transferred to a standard Shigemi microcell NMR tube for subsequent spectral analysis.

Urine was collected from mid-stream samples, divided into 1 ml aliquots, and stored in cryovials at −70°C within 30 min of collection. Cryovials were shipped frozen to the NMR laboratory, and stored at −80°C. To prepare the NMR samples, the cryovials were thawed on ice, and urine was mixed in a 9:1 ratio with a Chenomx internal standard solution (300 mM phosphate buffer, 4.3 mM DSS, 5.7 mM DFTMP, and 0.2% w/v sodium azide in D_2O).

NMR spectroscopy

All CSF and serum NMR spectra were acquired at 25°C on a Varian INOVA 500 equipped with a 5 mm triple resonance probe and *z*-axis pulsed field gradients. The urine spectra were collected at 25°C on a Varian INOVA 800 instrument operated using a 5 mm triple resonance cryogenic probe and triple-axis gradients. The same pulse sequence and pulse sequence parameters were used on both instruments (see Fig. 1 and Table 1). All spectra were processed using Version 6.0 of the Chenomx NMR Suite software. Data were zero-filled to twice the number of acquired points, multiplied by an exponential apodization function (0.2–0.5 Hz), and multi-point baseline corrected (Chang et al. 2007a). To remove any line asymmetry and Voigt-shape components, reference deconvolution (Morris et al. 1997) using the methyl DSS resonance peak at 0.00 ppm as the reference was applied to all spectra.

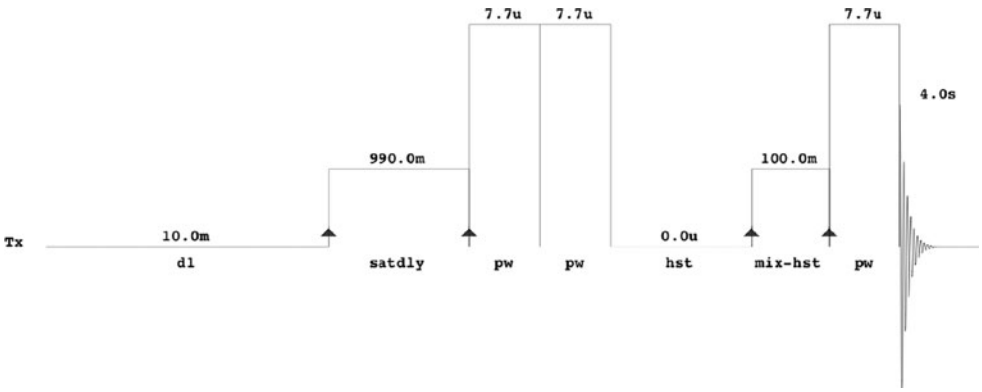


Fig. 1 Chenomx standard NOE-based pulse sequence (Varian version) adapted from the two-dimensional tnnoesy Varian pulse sequence. Parameters: recycling delay = 10 ms, water presaturation delay = 990 ms (power = 6 dB at 500 MHz), proton 90° pulse width ≈ 8 μs (power = 58 dB at 500 MHz), NOE mixing time = 100 ms, acquisition time = 4 s, sweep width = 12 ppm

Table 1 Acquisition parameters and sample experimental conditions for all types of biological samples used for automatic fitting

	CSF	Serum	Urine
Number of spectra	34	35	35
Temperature (°C)	25	25	25
Spectrometer frequency (MHz)	500	500	800
Sweep width (ppm)	12.0	12.0	12.0
Acquisition time (s)	4.0	4.0	4.0
Recycling delay (s)	1.0	1.0	1.0
Sample pH conditions			
Range	[7.19, 7.82]	[6.93, 7.45]	[5.98, 6.87]
Mean	7.46	7.15	6.40
Standard deviation	0.13	0.14	0.29
Autofit parameters			
Library size (number of compounds)	48	44	67
Δ pH	0.5	0.5	0.5
Minimum cluster transform windows (ppm)	±0.02	±0.02	±0.02

Creation of pH-sensitive libraries

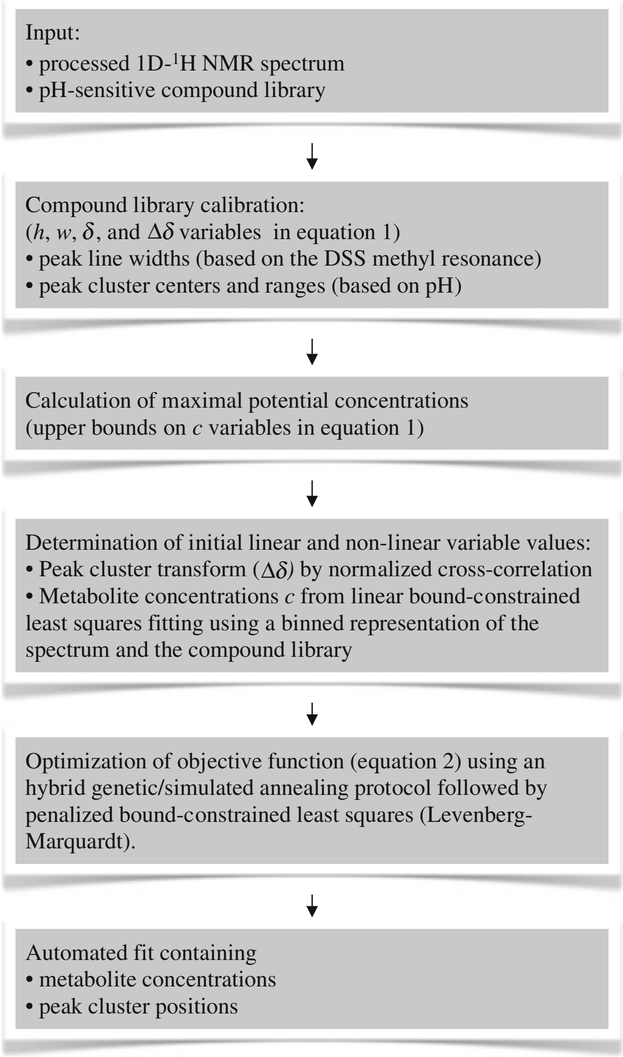
The location and overall shape of NMR peaks are influenced by a number of factors, including experimental sample conditions such as temperature, solvent, pH and ionic strength, as well as NMR acquisition parameters including the nature of the pulse sequence, the relaxation recovery delays and the solvent elimination scheme. While most NMR spectrometer parameters and temperature are usually kept constant during the course of a given study, the pH conditions of a series of samples can intrinsically change from one sample to another. This pH variation usually represents one of the largest sources of peak location variability across samples. In order for any kind of spectral fitting to perform optimally, a reference spectral library built using the same experimental conditions as those employed during sample data collection should be employed. For the current study, the pH-sensitive compound libraries from version 6.0 of the Chenomx NMR Suite software were used. One key feature of the Chenomx library is at the level of compound definition, where the experimental resonance pattern of every individual proton of a given metabolite is modeled and fit using the expected network of J-couplings from the molecular structure. All of the observed peak multiplets are entered/defined as separated entities called “peak clusters”, which can later be manipulated independently from one another in Chenomx’s software.

In the Chenomx compound reference library, the pH-dependence of the chemical shifts of all peak clusters was modeled mathematically using experimental data collected for each individual metabolite between pH 4.0 and 9.0 at internals of 0.5 pH units. The mathematical model used to simulate the chemical shift dependence on pH was derived from standard acid–base equilibrium chemistry and properly reflects the expected behavior of chemical shifts between two or more species in fast exchange on the NMR timescale (Frassineti et al. 1995).

In addition to the well-known pH-dependent changes in resonance frequencies, individual peak intensities and linewidths are also known to vary with pH conditions. These can arise due to exchange with the solvent, modifications of their T_1 and T_2 relaxation properties, and changes in coupling constant patterns. However, the current model does not take these effects into account. In the pH-sensitive compound library, the location of the peak clusters is defined by pH equations but the shape of all clusters is kept constant (peak intensities, line widths and distance to their respective cluster centers) relative to that observed at neutral pH. The current model cannot account for possible T_1 relaxation issues due to discrepancies between the reference and experimental pulse sequences or pulse sequence parameters. For maximum accuracy with AutoFit, the reference compound library and the experimental spectra should be acquired under the same conditions.

Automated profiling algorithm

A high-level description of the automated spectral fitting algorithm is given in Fig. 2. To start the process, the frequency domain trace of a processed NMR spectrum (the query spectrum) is read, along with a reference spectral library. The reference peak of the internal standard (DSS, TSP or formate) is then automatically located and fit. The reference spectral library is then calibrated based on the position, intensity, and linewidth of the reference peak. Information about the pH of the query spectrum and the concentration of the reference compound must be provided by the user (available directly in standard Chenomx cnx file



Input:
- processed 1D-^{1}H NMR spectrum
- pH-sensitive compound library

Compound library calibration:
(h, w, δ, and $\Delta\delta$ variables in equation 1)
- peak line widths (based on the DSS methyl resonance)
- peak cluster centers and ranges (based on pH)

Calculation of maximal potential concentrations
(upper bounds on c variables in equation 1)

Determination of initial linear and non-linear variable values:
- Peak cluster transform ($\Delta\delta$) by normalized cross-correlation
- Metabolite concentrations c from linear bound-constrained least squares fitting using a binned representation of the spectrum and the compound library

Optimization of objective function (equation 2) using an hybrid genetic/simulated annealing protocol followed by penalized bound-constrained least squares (Levenberg-Marquardt).

Automated fit containing
- metabolite concentrations
- peak cluster positions

Fig. 2 High-level schematic description of the automated profiling algorithm

format). This information is used to assist the fitting process and to calibrate the calculated concentrations to the concentration of the reference compound. Once these preliminary data are input and the preliminary reference compound fit is complete, the query spectrum I is reconstructed in the frequency domain using a linear combination of each of the reference compound spectra following a Lorentzian model:

$$\bar{I}(x) = \sum_{i=1}^{\text{nmetabolites}} c_i \sum_{j=1}^{\text{nclusters}} \sum_{k=1}^{\text{npeaks}} \frac{h_{i,j,k} \bullet w_{i,j,k}^2}{4\left(x - \left(\delta_{i,j,k} + \Delta\delta_{i,j}\right)\right)^2 + w_{i,j,k}^2} \tag{1}$$

where $\bar{I}(x)$ is the predicted spectral intensity at frequency x, c_{i} the concentration of metabolite i (linear variables), $\delta_{i,j,k}$ the resonance frequency of peak k of cluster j of metabolite i, $\Delta\delta_{i,j}$ the cluster center offset (called "transform") of peak cluster j of metabolite i (non linear variables), and $h_{i,j,k}$ and $w_{i,j,k}$ the intensity and line width of peak k in cluster j of metabolite i, respectively. The parameters $h_{i,j,k}$ and $w_{i,j,k}$ are constants once the compound library has been calibrated to the provided spectrum. The automated fitting procedure consists of finding the c and $\Delta\delta$ values that minimize the 2-norm target function

$$\chi^2 = \|I - \bar{I}\|^2 \tag{2}$$

This is ultimately achieved using a combination of a hybrid simulated annealing algorithm and a genetic algorithm (Chen et al. 2007b) as well as a Levenberg–Marquardt constrained non-linear least square minimization protocol. Because of the presence of non-linear variables, the success of the method is mainly dependent on the initial value of each variable and on the restriction imposed on the peak cluster transform windows to limit the variable dimension search space, to which pH-based compound libraries largely contribute. The next section describes the strategies employed to calculate starting values and bounds for all variables to be optimized.

First, the upper and lower bound values for all cluster transforms ($\Delta\delta$) are calculated based on the pH equations in the reference spectral library and the user-provided pH range. To account for potential changes in chemical shift behaviors in complex mixtures due to ionic strength variations and compound interactions (matrix effects) versus the standard conditions from which the pH equations were derived, a minimal value for the transform windows can be specified. For this study, a minimum value of $\Delta\delta = \pm 0.020$ ppm was used. The initial position of each cluster is set via a normalized cross-correlation process (Stein and Scott 1994). Each peak cluster is transformed over its entire spectral range defined by its lower and upper bound transform values, cross-correlation is calculated at each position, and the initial value for the peak cluster transform ($\Delta\delta$) is set where the normalized cross-correlation value reaches its maximum.

For metabolite concentrations, a non-negativity restraint is imposed on all c variables since compound concentrations cannot be negative ($c \geq 0$). The concentration upper bounds are set according to a "maximum potential concentration" method, by which the trace of a particular compound is not allowed to exceed the query spectrum's spectral intensities, considering all the possible combinations of cluster locations. More precisely, each cluster of a given compound is translated within its transform window range. At each increment, a maximum concentration is calculated. A potential concentration is calculated for each cluster, and the maximum potential concentration of the compound is set to the minimum values found for each cluster. Clusters located close to the water resonance in the [4.40, 5.50] ppm range are skipped during this process. Next, the initial concentration values c are calculated using a non-negative constrained linear fit ($Ac = B$, $c \geq 0$) on a binned representation of the spectrum (B). The matrix A is obtained by binning each compound trace as calculated using the cluster positions determined earlier by normalized cross-correlation. For this study, a bin size of 0.005 ppm was employed.

At this stage the algorithm has starting values and bounds for all variables. The query spectrum is then reconstructed (fit) via an iterative manner, in which the spectrum is decomposed into smaller sub-systems. The solutions found at each step are used as starting values for subsequent iterations. Ultimately, in the last minimization stage, the entire spectrum is fit with the complete set of variables.

To partially account for the presence of larger molecular species (proteins, lipids, etc.) and other resonances whose presence cannot be modeled and accounted for from the compounds in the library, the algorithm also optionally incorporates two types of baseline correction/estimation. The first is derived from a parametric smoothing model (Xi and Rocke 2008) and is applied before any fitting takes place, while the second is based on penalized B-splines (Sima and Van Huffel 2006) and is searched and optimized at each iteration of the Levenberg–Marquardt procedure.

The AutoFit program was entirely written in C, with the exception of a single Fortran routine. The typical time required for fitting a single spectrum on a dual 2.66 GHz core2-duo CPU equipped Mac Pro is approximately 10 min for CSF, 10 min for serum, and 45 min for urine. The amount of time required to perform a single fit depends largely on the number of spectral points (128 K for 800 MHz urine data compared to 64 K for 500 MHz CSF and serum spectra), the total number of variables (number of metabolites and peak clusters) and the application of simulated annealing to sample the variable space, which is largely dictated by the transform windows of the peak clusters.

Consequences of DSS binding on peak linewidth calibration

The engineering behind the Chenomx library allows for calculating (with high accuracy) the expected linewidths of all peak clusters of each individual metabolite based on the experimental linewidth of the DSS peak. The linewidth calibration procedure holds for any experimental shimming condition/quality, provided that medium or strong DSS binding/interactions with other species are absent.

For the CSF samples used in our study, weak DSS binding was observed, causing a widening of the experimental DSS peaks and a slight overestimation of the predicted metabolite peak linewidths. The experimental linewidths of the alanine methyl peaks ($\sim$1.47 ppm) or lactate doublets ($\sim$1.32 ppm) were found to be unaffected by DSS binding and used to back calculate what the DSS linewidth would have been in the absence of DSS binding. No such adjustments were needed with the urine and serum samples.

Special case—citrate

The calcium-content of biological solution (urine, CSF, serum, etc.) significantly affects the chemical shifts and J-couplings values of citrate (Moore and Sillerud 1994; Van Der Graaf and Heerschap 1996). As a consequence, special care must be taken for this particularly common metabolite. Should both citrate clusters dominate the spectrum in their respective potential frequency range, the automatic fitting program can find the citrate doublets and adjust their coupling constants to match the experimental spectrum.

Manual and automatic profiling of metabolites

To compare the performance of the automated fitting algorithm with the results obtained manually, the following procedures were adopted. Because manual spectral fitting can be tedious and prone to some errors, we used a two-person analysis strategy. A single individual initially profiled all NMR spectra manually with version 6.0 of the Chenomx NMR Suite software (using its associated pH-sensitive libraries). A second individual then verified the manual assignments. Ambiguities in the assignments for biological samples (for serum and CSF in particular) were clarified via sample spiking, in which a small amount of the suspected metabolite was added to the sample to confirm its presence. Corrections were made until an agreement was reached. Lower confidence metabolites were removed from the final manual profiles. These "consensus" spectral assignments were then used as "gold standards" for the comparison with the automated fits generated by AutoFit.

For each type of sample (computer-generated mixture, serum, urine, CSF), a targeted compound library formed of the union of all the manually profiled metabolites was constructed and used by the automatic profiling program. As such, both the manual and automatic profiling procedure used the exact same subset of compounds for each category of samples. See Table 2 for a detailed list of the identity of the metabolites used for profiling each category of the biological samples.

Comparison between manual and automatic metabolite profiling

The performance of the automated algorithm was quantitatively measured on the basis of the level of agreement between the automated and manual metabolite profiles, namely from the recovery rate of cluster positions and metabolite concentrations. For a given metabolite, not all peak clusters are systematically reliable for identification or quantification. Version 6.0 of the Chenomx NMR Suite software introduced the notion of "valid clusters", defined as a binary quality indicator of the goodness of fit and trust for a given peak cluster. In particular, a peak cluster is "valid" if it mathematically satisfies the following inequality:

$$\frac{\text{standard deviation}(|\Delta_x|)}{\frac{1}{n}\sum_{x=1}^{n} max(x)} < 0.2 \tag{3}$$

where $|\Delta_x|$ is a vector of length n, containing the absolute difference between the experimental spectrum and the cluster intensities at the current metabolite concentration at spectral point index x, and max(x) is the maximum intensity between the experimental intensity and the peak cluster intensity at point index x. Note that the relationship is not linear and will have a lower tolerance for relatively low intensity peaks (noise). Figure 3 graphically summarizes the concept of valid clusters.

The manually and automatically determined metabolites concentrations were compared including error ranges (confidence levels) associated with both sets. The expected errors on the reported manually profiled concentrations were calculated on a per compound basis in each spectrum from twice the spectral noise level. The values calculated in this manner encapsulate the individual compound intensity response with concentrations and represent the minimal expected concentration errors. The confidence intervals on the concentrations determined by the automated program were derived from the square root of the element on the diagonal of the covariance matrix of the fits (i.e., the standard deviations).

Table 2 List of metabolites used for automated profiling of CSF, serum and urine spectra

CSF	Serum	Urine	Urine(cont)
2-Hydroxybutyrate	2-Hydroxybutyrate	1-Methylnicotinamide	Quinolinate
2-Hydroxyisovalerate	3-Hydroxybutyrate	2-Aminobutyrate	Serine
2-Oxoglutarate	Acetate	2-Hydroxyisobutyrate	Succinate
2-Oxoisovalerate	Acetone	3-Aminoisobutyrate	Sucrose
3-Hydroxybutyrate	Adipate	3-Hydroxybutyrate	Taurine
3-Hydroxyisovalerate	Alanine	3-Hydroxyisovalerate	Threonine
Acetaminophen	Arginine	3-Indoxylsulfate	Trigonelline
Acetate	Asparagine	4-Hydroxyphenylacetate	Trimethylamine
Acetoacetate	Betaine	Acetaminophen	Trimethylamine N-oxide
Acetone	Butyrate	Acetate	Tryptophan
Alanine	Carnitine	Acetoacetate	Tyrosine
Choline	Choline	Acetone	Uracil
Citrate	Citrate	Adipate	Urea
Creatine	Creatine	Alanine	Valine
Creatinine	Creatinine	Asparagine	Vanillate
Dimethyl sulfone (DMSO2)	DSS	Betaine	Xylose
DSS	Ethanol	Carnitine	cis-Aconitate
Dimethylamine	Formate	Citrate	trans-Aconitate
Formate	Glucose	Creatine	π-Methylhistidine
Fructose	Glutamate	Creatinine	τ-Methylhistidine
Glucose	Glutamine	DSS	
Glutamate	Glycerol	Dimethylamine	
Glutamine	Glycine	Ethanolamine	
Glycerol	Histidine	Formate	
Histidine	Imidazole	Fucose	
Imidazole	Isobutyrate	Fumarate	
Isoleucine	Isoleucine	Glucose	
Isopropanol	Lactate	Glutamine	
Lactate	Leucine	Glycine	
Leucine	Lysine	Glycolate	
Lysine	Malonate	Guanidoacetate	
Mannose	Methanol	Hippurate	
Methanol	Methionine	Histidine	
Methionine	Ornithine	Hypoxanthine	
Phenylalanine	Phenylalanine	Isoleucine	
Propylene glycol	Proline	Lactate	
Pyroglutamate	Propylene glycol	Leucine	
Pyruvate	Pyruvate	Lysine	
Serine	Serine	Malonate	
Succinate	Succinate	Methanol	
Threonine	Taurine	N,N-Dimethylglycine	
Tryptophan	Threonine	O-Acetylcarnitine	
Tyrosine	Tyrosine	Phenylacetylglycine	
Urea	Urea	Phenylalanine	
Valine	Valine	Propylene glycol	
Xanthine		Pyroglutamate	
b-Hydroxyisobutyric acid		Pyruvate	
myo-Inositol			

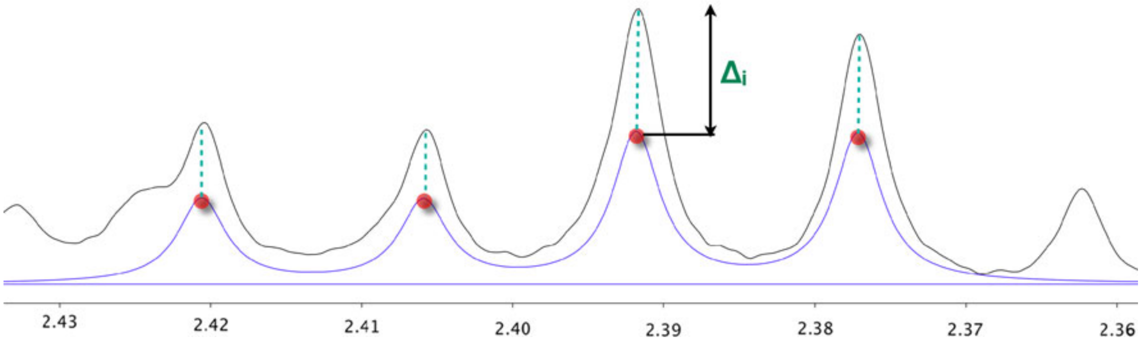


Fig. 3 Illustration of the concept of "valid cluster" in Chenomx NMR Suite 6.0 with a hypothetical peak cluster composed of 4 peaks. The absolute difference Δ_i between the experimental spectrum and the 4 characteristic peaks is measured at the resonance frequency of each peak (*filled circles*), leading to a vector Δ of length $n = 4$ elements. The standard deviation of Δ and the average of the sum of the maximum intensities at the 4 locations are then computed. The cluster is 'valid' if the ratio of both quantities satisfies Eq. 3

Preparation of computer-generated mixtures

A set of 100 computer-generated mixtures of CSF (at 500 MHz) and urine samples (at 800 MHz) were generated using a reference compound library composed of the same compounds profiled manually in their corresponding datasets (see Table 2). No serum simulated mixtures were generated due to the similar number and type of compounds in the serum and CSF compound libraries (see Table 1). The simulated CSF and urine spectra were generated from compound libraries calibrated at pH conditions matching the observed mean and standard deviations listed in Table 1. The metabolite concentrations were randomly chosen from a continuous uniform distribution, where the lower and upper bound concentrations values were set based on the minimum and maximum observed in the manually-fit experimental spectra. White noise was also added to the simulated mixtures to reproduce the level of noise found in the respective experimental datasets. The minimum transform window for all peak clusters was set to ±0.02 ppm.

Results and discussion

The spectral reconstruction methodology for automated spectral deconvolution presented in this paper shares some common features with MRI-related spectral fitting, for which mathematical solutions and dedicated software have been widely documented (Poullet et al. 2007; Poullet et al. 2008). However, due to the large number of spectral points and the number of metabolites to deconvolute, the complexity/dimensionality of the type of problem tackled here is much higher. As a result, a much wider variable space has to be explored. In addition, due to the nature of the reference compound library, spectral fitting is performed in the frequency domain instead of the time domain. Fitting in the frequency domain allows for the application of different weightings to different spectral regions. This potential feature was not directly exploited in the current study, except for the elimination of the spectral region associated with the solvent (water).

Two kinds of performance assessments were conducted with two different kinds of data. The first assessment was based on characterizing precisely known mixtures (computer-generated or "simulated" mixtures). The second assessment was based on characterizing biological mixtures in which the exact content and concentrations are not precisely known, but for which we have a relatively fairly good degree of confidence through independent analyses. The intent of the first assessment was to test the program's performance with "perfect" or "near-perfect" input data and "perfect" knowledge of the correct answer. The intent of the second assessment was to test the program's performance with real biological samples where matrix effects, spectral overlap and level of uncertainty in the mixture composition made the fitting task far more challenging. More specifically with the latter case, we compared the results to those obtained by people experienced with computer-assisted or manual fitting.

Two types of performance measures were applied on the simulated and experimental datasets. One was based on the specificity/sensitivity (Altman and Bland 1994) with respect to pure compound identification, independent of their concentrations. Compounds with at least one valid cluster (Eq. 3) were counted as positive hits (either as true or false positives). Oppositely, compounds with no valid clusters were classified as true or false negatives. The second performance assessment was based on the level of agreement between the automated and computer-generated or manual metabolite profiles. This was calculated from the recovery rate of "valid" cluster positions and metabolite concentrations. More specifically, we were particularly interested in quantifying the recovery rate of metabolites concentrations and the position of meaningful clusters (those reported as ‘valid’) in automated profiles relative to solutions from the simulated or manual profiles.

Two sub-analyses are also presented, the first focusing exclusively on peak cluster positions, whereas the second focuses on metabolite concentrations. Equation 1 shows that the automatically profiled metabolite concentrations (linear variables) are tightly coupled to the peak cluster locations, which form the set of non-linear variables of the optimization function. If the latter were known, the optimization problem would reduce to solving a simple set of linear equations. It is therefore of great importance for the algorithm to properly position peak clusters, as they basically drive the solution of the linear coefficients. In the first analysis, the capacity of the automated algorithm to position peak clusters at the same simulated or manually determined locations is evaluated. This assessment was performed on a per cluster basis, independently of their parent metabolite and the metabolite concentrations.

Computer-generated mixtures (spectra)

All the statistics relevant to the analysis of the computer-generated mixtures are presented in Table 3. At the level of compound identification using the presence of at least one valid cluster as a binary classifier, no false positives or false negatives were detected in either the simulated CSF or urine datasets. In other words, AutoFit performs perfectly when given perfect data. This explains the perfect sensitivity and specificity scores. Figure 4 shows a comparison of the location of the peak clusters found in the computer-simulated profiles and the AutoFit-generated profiles. The number of reported valid peak clusters in the manual and automated profiles, the number of reported valid peak clusters common to both, and how many of the latter set are positioned within 0.2 Hz relative to one another are presented. A relatively large cluster position recovery rate (~90%) for valid clusters in both datasets is obtained, despite the relatively large transform windows for some clusters, such as acetaminophen (0.07 ppm), dimethyl sulfone (0.08 ppm), xanthine (0.12 ppm), 2-hydroxybutyrate (0.13 ppm), 2-oxoisovalerate (0.2 ppm), beta-hydroxyisobutyric acid (0.3 ppm),

Table 3 Automated profiles statistics on simulated and experimental spectra for the different biofluid samples

	Simulated		Experimental[a]		
	CSF	Urine	CSF	Serum	Urine
Number of compounds in the compound library	48	67	48	44	67
Number of clusters in the compound library	188	270	188	164	270
Average number of clusters per compound	3.9	4.0	3.9	3.7	4.0
Sensitivity/specificity analysis					
Sensitivity (recall rate)	1.0	1.0	1.0	1.0	0.64 ± 0.06
Specificity (true negative rate)	–	–	–	–	0.88 ± 0.05
Positive predictive value (precision)	1.0	1.0	1.0	1.0	0.88 ± 0.04
Negative predictive value	–	–	–	–	0.63 ± 0.05
Accuracy	1.0	1.0	1.0	1.0	0.74 ± 0.04
Peak cluster analysis					
Average number of valid peak clusters in manual profiles	52 ± 4 (28% of 188)	112 ± 5 (41% of 270)	39 ± 6 (20% of 188)	36 ± 5 (22% of 164)	63 ± 5 (23% of 270)
Average number of valid peak clusters in automated profiles	50 ± 4 (27% of 188)	103 ± 5 (38% of 270)	38 ± 5 (20% of 188)	35 ± 6 (21% of 164)	48 ± 5 (18% of 270)
Percentage of valid peaks clusters in manual profiles recovered within ± 0.2 Hz in automated profiles	90 ± 5	87 ± 5	82 ± 5	81 ± 6	60 ± 8%
Metabolite concentrations analysis					
Average number of eligible compounds for analysis[b]	26 ± 2	53 ± 3	23 ± 3	21 ± 3	33 ± 4
Average number of compounds with concentrations compatible with simulated or manually determined values	26 ± 3	39 ± 4	14 ± 4	13 ± 4	21 ± 6
Average number of compounds with manually matching concentrations and cluster positions for all valid clusters	25 ± 3	37 ± 4	12 ± 4	10 ± 3	17 ± 4

[a] The manual metabolomic profiles are used as gold standards

[b] Only compounds with at least one valid peak cluster in the automated profiles were kept for analysis

histidine (0.43 ppm), imidazole (0.46 ppm), π-methylhistidine (0.51 ppm) and τ-methylhistidine (0.55 ppm). The existence of these large transform windows is typically due to either lack of pH-sensitive information or large pH-sensitive chemical shift dependence in the experimental pH conditions range.

For the second part of the analysis, which focused on extracting compound concentrations, we discarded all the AutoFit profiled compounds with no reported valid clusters. The concentrations of the remaining metabolites were then compared quantitatively to their respective simulated values. Metabolites whose AutoFit-determined concentrations (including a 95% confidence interval or 1.96 standard deviations) intersected with their associated manually reported concentrations were kept. On average, 100% of the CSF metabolites eligible for analysis (having at least one valid cluster) have concentrations within the simulated values (see panel A of Fig. 5 and the bottom part of Table 3). For simulated urine samples, this percentage drops to ~70%. This is without a doubt due to the significantly greater complexity and spectral overlap seen in urine NMR spectra relative to CSF or serum NMR spectra.

Real biological mixtures

As reported in Table 3, and similar to the situation observed with the simulated datasets, the real CSF and serum datasets present perfect sensitivity/specificity rates. However, for urine samples, a number of false positives and false negatives are observed, leading to an average specificity rate of ~88% and an average sensitivity of ~64%. Overall, the accuracy for AutoFit for urine spectra drops to ~75%.

The same cluster position and concentration recovery rate analyses discussed in the previous section were also performed on the experimental biological spectra. The results are presented in Figs. 6 and 7. The CSF and the serum datasets exhibit similar statistics (see Table 3), with over 80% of all the valid peak clusters in the manual profiles located within 0.2 Hz in the automated profiles. The remaining 20% of the peak clusters are either not reported as valid in the automated profiles, or not located within 0.2 Hz in the automated profiles. For the urine samples, the average recovery rate of matching peak cluster positions drops to 60 ± 8%. If we measure the average recovery of metabolite concentrations for all three

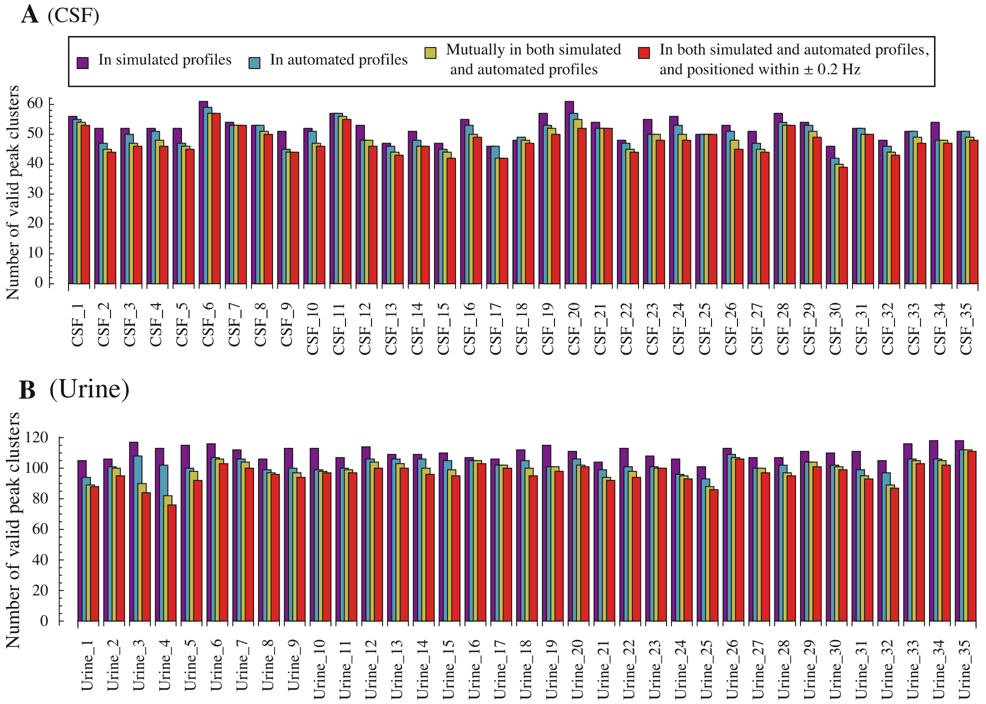


Fig. 4 Comparison of computer-generated and automated metabolomic profiling with bar charts of the total number of valid peak clusters present: in the simulated (*purple*) and automated (*blue*) profiles separately, mutually in both the computer-simulated and automated profiles (*gold*), and mutually in both the simulated and automated profiles while having their valid peak clusters positioned within 0.2 Hz of each other (*red*) for CSF (**A**) and urine samples (**B**). The compound libraries used for automated profiling were composed of a total of 188 and 270 peak clusters for the CSF and urine samples, respectively. Only the results of the first 35 simulated spectra of both datasets are shown

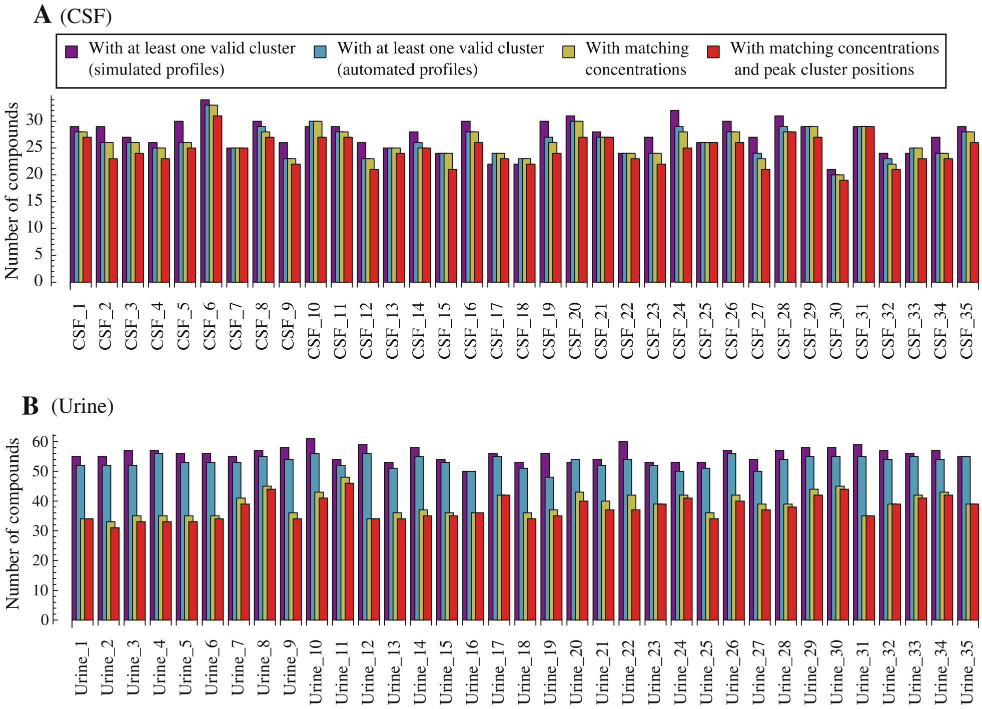


Fig. 5 Comparison of computer-generated and automated metabolomic profiling with bar charts of the total number of compounds having: at least one reported valid cluster in the simulated (*purple*) and automated (*blue*) profiles, matching concentrations (*gold*), matching concentrations and positioned peak clusters within 0.2 Hz of each other (*red*) for (**A**) CSF and (**B**) urine samples. The compound libraries used for automated profiling were composed of a total of 44 and 67 compounds for CSF and urine samples, respectively. Only the results of the first 35 simulated spectra of both datasets are shown

datasets (urine, serum and CSF), only half of the metabolites eligible for analysis (having at least one valid cluster in the automated profiles) end up with concentrations within the manually profiled values (see Fig. 7 and the bottom part of Table 3).

Sources of error

To understand these results, it is important to look at the different sources of errors that may affect the outcome of an individual automated profiling, and the general challenges that automated profiling faces. Fundamentally, the automated profiling approach uses a set of individually collected metabolite signatures as prior knowledge. Thus, the quality/suitability of the compound library used to build the "model matrix" for spectral reconstruction is of pivotal importance. The strength of the Chenomx compound library resides in the fact that the observed resonance pattern (multiplet) of each individual proton of a given metabolite form separate entities (called peak clusters), which can then be manipulated independently from one another. Not only is this separation process powerful, it is also necessary to comply to the NMR behavior of peak resonances with changing pH, where peak clusters associated with a given compound can move in different directions (upfield/downfield) and by different amounts. The reference compound library constitutes the primary framework to the automated fitting program, and it should be optimized for the current experimental conditions. The Chenomx library, unlike other spectral reference libraries, allows for the calculation of expected peak cluster centers and their range of possible values (called cluster transforms) simply based on the pH of the experimental samples and the allowed pH error (set to ±0.5 units of pH for all samples in this study). Another key advantage of the Chenomx compound library is the possibility to dynamically calibrate all peak line widths based on the observed experimental line width of the DSS methyl peak (DSS is also used as internal standards for concentrations). The compound library can thereby be used under any experimental shimming conditions, which is, to our knowledge, a unique feature of the Chenomx compound library.

Despite these pH-sensitive capabilities, the Chenomx library does not encapsulate all sources of potential spectral variability. For instance, relative cluster peak height ratios and observed coupling constants in the individual spectral metabolite signatures are affected by the salt composition, ionic strength, and content (matrix) of the experimental samples. The current compound library does not model the spectral consequences these effects may have. For instance,

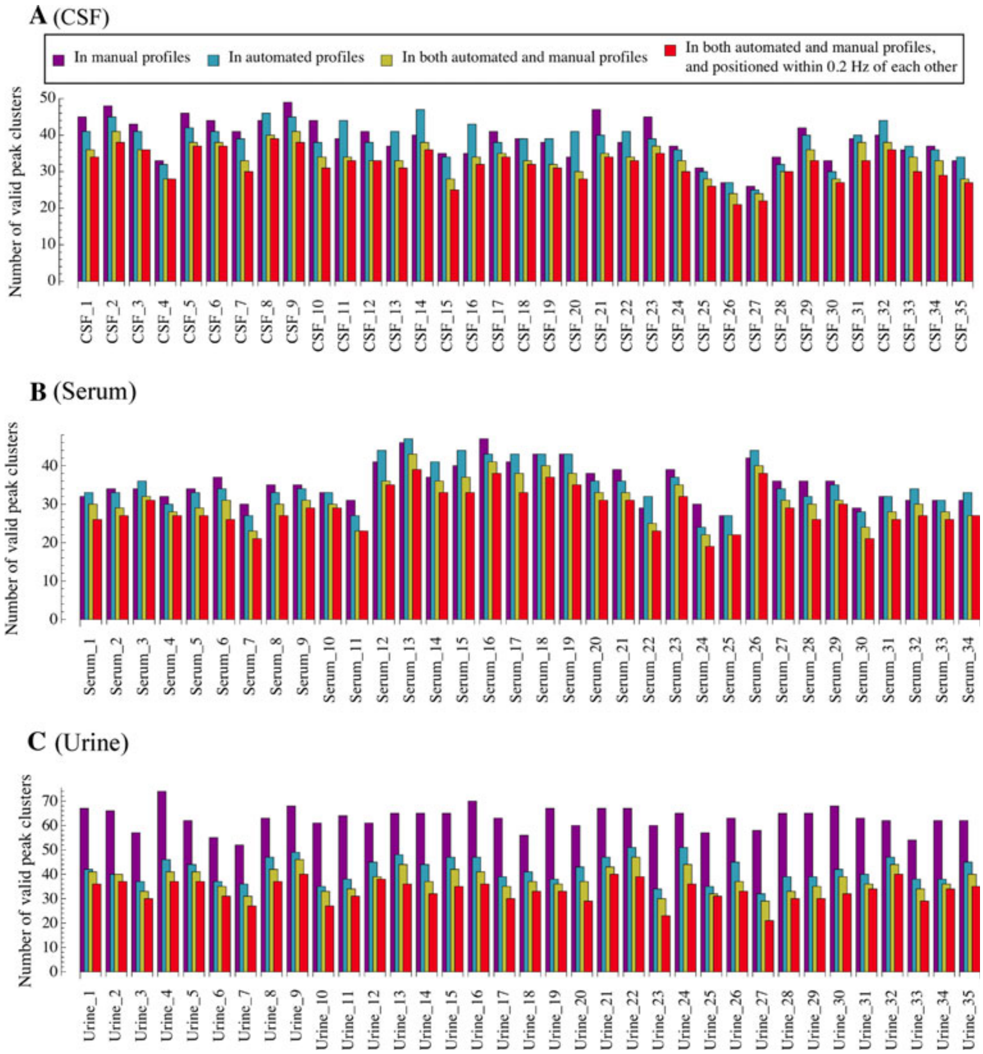


Fig. 6 Comparison of manual and automated metabolomic profiling with bar charts of the total number of valid peak clusters present: in the manual (*purple*) and automated (*blue*) profiles separately, mutually in both the manual and automated profiles (*gold*), and mutually in both the manual and automated profiles while having their valid peak clusters positioned within 0.2 Hz of each other (*red*) for CSF (**A**), serum (**B**) and urine samples (**C**). The compound libraries used for automated profiling were composed of a total of 188, 164 and 270 peak clusters for the CSF, serum and urine samples, respectively

citrate is particularly sensitive to calcium content (Lindon et al. 2007), while the spectral signatures of other metabolites like glutamine, and aromatic amino acids like phenylalanine and tyrosine are known to be affected by pH conditions and matrix effects. Except for citrate where the J-coupling of the two characteristic doublets can be optimized to match the experimental spectrum, the shape of all other compound spectra is that of the observed one at neutral pH. Additionally, some compounds, like sugars, also undergo fast chemical exchange (on the NMR time-scale) triggered by different enantiomeric forms. To better cover the matrix and ionic strength effects, the upper and lower bounds calculated for each individual peak cluster center are extended to a minimum of ±0.02 ppm. However, this extended range widens the searchable variable space, increasing the complexity of the problem, the length of the automated process, and the likelihood of convergence in local minima (Eq. 2).

Although the pH conditions of all samples used in this study were controlled using chemical buffers, the impossibility of constructing a perfect model matrix constitutes a potential source of error. Possible peak height and peak shape discrepancies (due to different J-couplings) between those seen experimentally and the ones modeled in the library impact the reported concentrations at two levels. During the earliest stage of automated profiling, the AutoFit algorithm evaluates a maximal potential concentration for each metabolite. This corresponds to the largest concentration a compound can be set to without exceeding the experimental spectrum. These values are then used to set upper bounds on concentration variables and, whereas those values are only suggestive during manual fitting and can be overridden, the automated algorithm strictly conforms to them. In the current context of spectral reconstruction based on a least squares approach using an incomplete compound library (relative to the real number of compounds present in the biological samples), the upper bound concentrations play a critical role in preventing over fitting of some spectral regions to better fill the spectral areas in other parts of the spectrum.

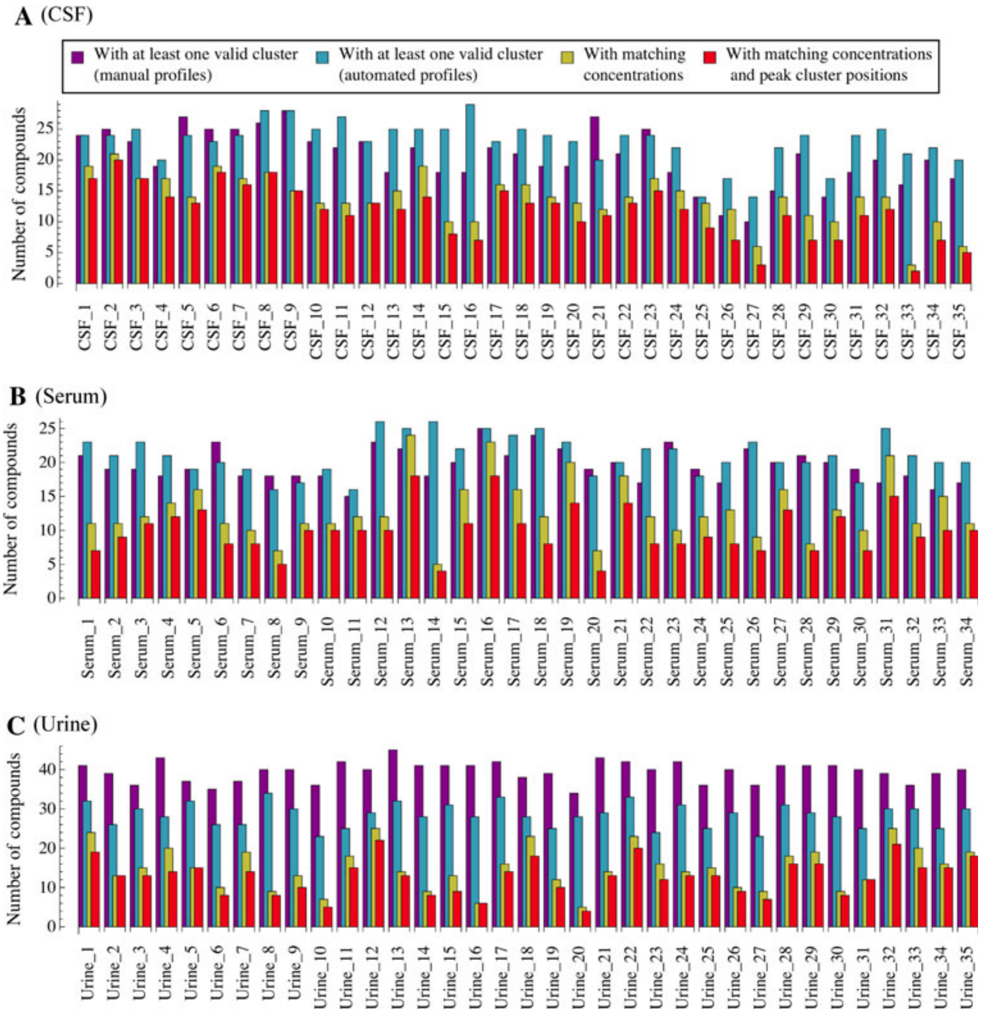


Fig. 7 Comparison of manual and automated metabolomic profiling with bar charts of the total number of compounds having: at least one reported valid cluster in the manual (*purple*) and automated (*blue*) profiles, matching concentrations (*gold*), matching concentrations and positioned peak clusters within 0.2 Hz of each other (*red*) for (**A**) CSF, (**B**) serum, and (**C**) urine samples. The compound libraries used for automated profiling were composed of a total of 44, 48 and 67 compounds for CSF, serum and urine samples, respectively

Another source or error or variability is the quality of water suppression. Failure to properly calibrate the water presaturation pulse will generate a different attenuation profile than the one used during library creation, therefore affecting the spectral fit and the calculated concentrations. Incomplete water suppression can also distort the experimental spectrum, and the application of a subjective baseline correction may be required prior to spectral profiling. Imperfections associated with baseline correction may also alter the concentration determination.

Manual versus automated profiling

There is a fundamental difference between how manual and automated spectral fitting is performed. The concentration of some compounds during the manual profiling process may be systematically determined from relatively well-isolated or overlap-free clusters, ignoring the contribution of other clusters. On the other hand, the automated fitting algorithm uses a penalized and bounded least square approach that does not favor any specific spectral region or cluster over another one. For this reason, the automated solution may be viewed as more “global” than “local”. For example, the methyl region (between 0.5 and 1.5 ppm) is usually information-rich, well resolved, and particularly important for the quantification for amino acids such as alanine, valine, leucine, isoleucine, threonine, and other methyl moiety containing metabolites. This relatively overlap-free spectral region offers a series of “cluster drivers” serving as guides during the manual profiling of a good proportion of the compounds. However, the methyl region is also affected by the presence of larger molecular species like proteins or lipids. Manual profiling of this region usually necessitates a visual estimation of the underlying lipid/protein signal, and the metabolite concentrations are adjusted as to optimize the smoothness of the residuals. The human eye usually performs this process better than can be modeled computationally. Our automated algorithm attempts to mimic the manual approach with the inclusion of a non-parametric baseline based on P-splines (Sima and Van Huffel 2006). The metabolite concentrations are adjusted as to optimize the reconstitution of the experimental spectrum and minimize the sum of the second derivative of the residuals (weighted with a penalty term, which is also optimized by the algorithm). In general, the regulated/penalized methodology used in the automated procedure does improve the overall agreement between the manual and automated metabolite concentrations. This is especially true when the experimental samples contain extraneous material not modeled by the compound library (urine samples in particular). However, even with a large number of spline knots, the resulting baseline may not fully agree with the human estimate, which will undeniably lead to differences in calculated or estimated concentrations. Additionally, it is possible that in some cases the algorithm may simply chose not to include a compound (by setting its concentration to zero) in favor of a smoother subtraction line.

It is clear from Fig. 7 that the level of agreement between the reported metabolite concentrations for the manual and automated profiles varies from one sample to another. At first sight, poor agreement is observed for CSF samples 15, 16, 27, 30, 33 and 35, in particular. Closer inspection of the automated solutions for those samples actually reveal that the automated profiling algorithm did not fail, but rather managed to better reconstruct the experimental spectrum. A typical case is presented in Fig. 8, with an overlay of the experimental spectrum of CSF_33 with the manual and automated spectral reconstruction, for which the disparity is the greatest (see Fig. 7). The intrinsic ability of the algorithm to juggle different combinations of variables has clearly exceeded the human procedure in this case in reconstituting with higher fidelity to the query spectrum. It is important to remind the reader that the solution is only mathematically better and may not necessarily constitute a better biological or biochemical result. However, a low level of agreement between the manual and automated solutions does not necessarily indicate poor performance of the algorithm.

All the results presented so far depend largely on two factors: the "valid" cluster to decide whether or not a reported metabolite/concentration can be trusted and classified as being present, and the manually-determined concentration error range based on the experimental spectral signal-to-noise level. To test whether the latter methodology is appropriate and/or too restrictive we compared the simulated (via synthetic biofluid spectra) or manually derived (via real biofluid spectra) concentrations to the AutoFit-determined concentrations. Figure 9 depicts a weighted linear regression (without a constant basis, i.e., $y = a \cdot x$) between the two sets of values for the two types of simulated data and three kinds of biological samples. The weights on the automated concentrations were obtained directly from the inverse of the reported fit covariance matrix of each spectrum. For the 5 regressions, a total number of 1,575 data points were used for the CSF samples, 1,428 for the serum samples and 2,275 for the urine samples. High linearity and coefficients of determination R^2 are observed for all datasets when the entire range of concentrations is considered. For biological samples, the fitted slope is lower than unity, which suggests that globally, the automated algorithm tends to report

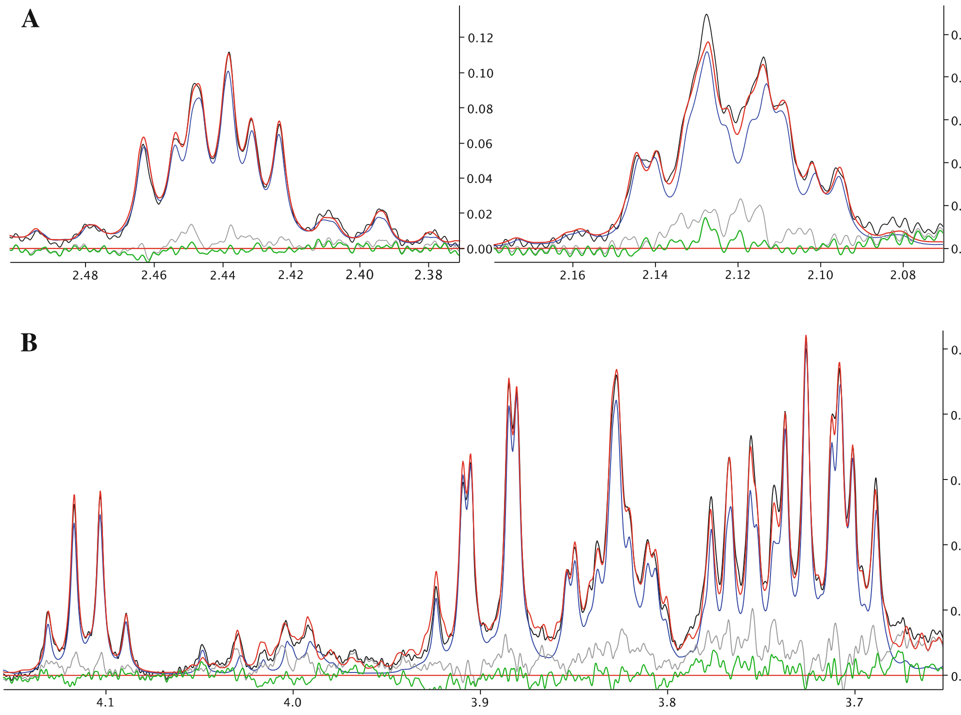


Fig. 8 Comparison of manual and automated profiling of a CSF sample (CSF_33) for **A** two glutamine peak clusters and **B** the aliphatic region (dominated by glucose). The experimental spectrum is shown in *black*, the automated deconvolution in *red*, the manual reconstitution in *blue*, the residuals with the automated deconvolution in *green*, and the residuals with the manual reconstitution in *gray*

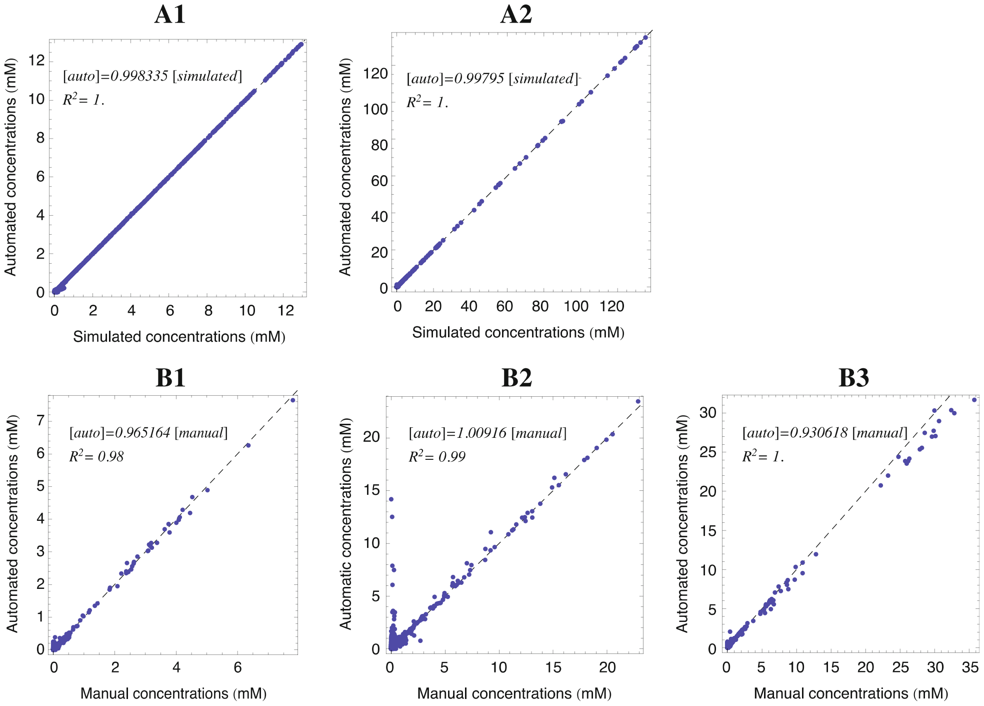


Fig. 9 Correlation between all of the simulated (*panels A*) or manually (*panels B*) concentrations and automated determined compound concentrations for all of the CSF (*A1*, *B1*), urine (*A2*, *B2*) and serum (*B3*) samples. In all *B panels*, urea, imidazole and lactate points were removed from the analysis due height or peak shape discrepancies between the experimental data and the compound database. The *dashed line* represents the expected locations of the data points

concentrations that are slightly lower than the ones obtained manually. Overall, a high level of linearly and agreement is observed between the automated concentrations with the expected or manually determined values.

Conclusion

The main challenges of automated metabolite profiling of NMR spectra lie in: a) the imperfect reconstruction of a model matrix due to discrepancies (despite the use of sophisticated compound libraries) between the reference compound library and the experimental spectra (overall peak height and shape differences, change in coupling constants, effects of those on concentration upper bounds, etc.), b) the incompleteness of the reference compound library and the presence of endogenous solution components, c) the very large variable space needed to be explored for the optimizer and the possibility of singularity (same reconstructed spectrum from different combinations of variable values), and d) a chi-square-based global mathematical solution versus individual compound/cluster tuning during manual profiling.

Different strategies were put in place with these limitations in mind and the algorithm was tested against simulated and real spectra of biological fluids. The AutoFit program performed with high accuracy against all simulated datasets, at the level of both compound identification and concentration determination. Since the exact composition of the real biological fluids is unknown, the performance of the AutoFit program could only be assessed on a relative basis against the manually-determined profiles. Perfect sensitivity and specificity scores were obtained with the automated algorithm against experimental CSF and serum datasets. However, for the experimental urine spectra, due to an incomplete reference compound library and general spectral complexity, the accuracy dropped to ~75% using the manually-determined metabolite profiles as gold standards.

The compound concentrations obtained by AutoFit showed a very high level of correlation with those from the computer-simulated set and those obtained manually on real biological samples. We believe that in its current state, the AutoFit algorithm can be confidently used on biofluid mixtures containing ~50 detectable compounds (serum, CSF, saliva, cell extracts) with the caveat that the majority of the experimental resonances are also present in the reference compound library used to perform the automated spectral reconstruction.

References

Aich P, Potter AA, Griebel PJ (2009) Modern approaches to understanding stress and disease susceptibility: a review with special emphasis on respiratory disease. Int J Gen Med 2:19–32

Altman DG, Bland JM (1994) Diagnostic tests 1: sensitivity and specificity. BMJ 308:1552

Beckonert O, Coen M, Keun HC, Wang Y, Ebbels TM, Holmes E, Lindon JC, Nicholson JK (2010) High-resolution magic-angle-spinning NMR spectroscopy for metabolic profiling of intact tissues. Nat Protoc 5:1019–1032

Bertram HC, Eggers N, Eller N (2009) Potential of human saliva for nuclear magnetic resonance-based metabolomics and for health-related biomarker identification. Anal Chem 81:9188–9193

Bezabeh T, Somorjai RL, Smith ICP (2009) MR metabolomics of fecal extracts: applications in the study of bowel diseases. Magn Reson Chem 47:S54–S61

Cevallos-Cevallos JM, Reyes-De-Corcuera JI, Etxeberria E, Danyluk MD, Rodrick GE (2009) Metabolomic analysis in food science: a review. Trends Food Sci Tech 20:557–566

Chang D, Banack CD, Shah SL (2007a) Robust baseline correction algorithm for signal dense NMR spectra. J Magn Reson 187: 288–292

Chang D, Weljie A, Newton J (2007b) Leveraging latent information in NMR spectra for robust predictive models. Pac Symp Biocomput 115–126

Chen C, Gonzalez FJ, Idle JR (2007a) LC-MS-based metabolomics in drug metabolism. Drug Metab Rev 39:581–597

Chen DJ, Lee CY, Park CH, Mendes P (2007b) Parallelizing simulated annealing algorithms based on high-performance computer. J Global Optim 39:261–289

Chenomx Nmr Suite (2010) Chenomx Inc., Edmonton, AB, Canada. http://www.chenomx.com

Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, Westler WM, Eghbalnia HR, Sussman MR, Markley JL (2008) Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol 26:162–164

Fonville JM, Maher AD, Coen M, Holmes E, Lindon JC, Nicholson JK (2010) Evaluation of full-resolution J-resolved 1H NMR projections of biofluids for metabonomics information retrieval and biomarker identification. Anal Chem 82:1811–1821

Frassineti C, Ghelli S, Gans P, Sabatini A, Moruzzi MS, Vacca A (1995) Nuclear magnetic resonance as a tool for determining protonation constants of natural polyprotic bases in solution. Anal Biochem 231:374–382

Griffin JL, Kauppinen RA (2007) Tumour metabolomics in animal models of human cancer. J Proteome Res 6:498–505

Hall RD, Brouwer ID, Fitzgerald MA (2008) Plant metabolomics and its potential application for human nutrition. Physiol Plant 132:162–175

Kaddurah-Daouk R, Kristal BS, Weinshilboum RM (2008) Metabolomics: a global biochemical approach to drug response and disease. Annu Rev Pharmacol Toxicol 48:653–683

Keun HC (2006) Metabonomic modeling of drug toxicity. Pharmacol Ther 109:92–106

Kim YS, Maruvada P, Milner JA (2008) Metabolomics in biomarker discovery: future uses for cancer prevention. Futur Oncol 4: 93–102

Kim K, Aronov P, Zakharkin SO, Anderson D, Perroud B, Thompson IM, Weiss RH (2009) Urine metabolomics analysis for kidney cancer detection and biomarker discovery. Mol Cell Proteomics 8:558–570

Kimura T, Noguchi Y, Shikata N, Takahashi M (2009) Plasma amino acid analysis for diagnosis and amino acid-based metabolic networks. Curr Opin Clin Nutr Metab Care 12:49–53

Kristal BS, Shurubor Y, Marur V (2007) Projection-based informatics approaches to serum/plasma metabolomics data: applications to biomarkers for caloric intake in rats. Faseb J 21:A310–A310

Li H, Jiang Y, He FC (2008) Recent development of metabonomics and its applications in clinical research. Yi Chuan 30:389–399

Lindon JC, Nicholson JK (2008) Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. Annu Rev Anal Chem 1:45–69

Lindon JC, Nicholson JK, Holmes E (2007) The handbook of metabonomics and metabolomics, 1st edn. Elsevier, Amsterdam

Markley JL, Anderson ME, Cui Q, Eghbalnia HR, Lewis IA, Hegeman AD, Li J, Schulte CF, Sussman MR, Westler WM, Ulrich EL, Zolnai Z (2007) New bioinformatics resources for metabolomics. Pac Symp Biocomput 157–168

Milgram E, Nordstrom A (2009) Asms metabolomics workshop survey. http://fiehnlab.ucdavis.edu/staff/kind/Metabolomics-Survey-2009/

Moore JG, Sillerud OL (1994) The pH dependence of chemical shift and spin-spin coupling for citrate. J Magn Res B 103:87–88

Morris GA, Barjat H, Home TJ (1997) Reference deconvolution methods. Prog Nucl Mag Res Sp 31:197–257

Oakman C, Tenori L, Biganzoli L, Santarpia L, Cappadona S, Luchinat C, Di Leo A (2010) Uncovering the metabolomic fingerprint of breast cancer. Int J Biochem Cell Biol

Odunsi K (2007) Cancer diagnostics using 1H-NMR-based metabonomics. Ernst Schering Found Symp Proc 4:205–226

Poullet JB, Sima DM, Simonetti AW, De Neuter B, Vanhamme L, Lemmerling P, Van Huffel S (2007) An automated quantitation of short echo time MRS spectra in an open source software environment: AQSES. NMR Biomed 20:493–504

Poullet JB, Sima DM, Van Huffel S (2008) MRS signal quantitation: a review of time- and frequency-domain methods. J Magn Reson 195:134–144

Quinones MP, Kaddurah-Daouk R (2009) Metabolomics tools for identifying biomarkers for neuropsychiatric diseases. Neurobiol Dis 35:165–176

Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, Pujos-Guillot E, Verheij E, Wishart D, Wopereis S (2009) Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. Metabolomics 5:435–458

Schiffmann R, Waldek S, Benigni A, Auray-Blais C (2010) Biomarkers of Fabry disease nephropathy. Clin J Am Soc Nephrol 5:360–364

Serkova NJ, Niemann CU (2006) Pattern recognition and biomarker validation using quantitative 1H-NMR-based metabolomics. Expert Rev Mol Diagn 6:717–731

Sima DM, Van Huffel S (2006) Regularized semiparametric model identification with application to nuclear magnetic resonance signal quantification with unknown macromolecular base-line. J Roy Stat Soc B 68:383–409

Sinclair AJ, Viant MR, Ball AK, Burdon MA, Walker EA, Stewart PM, Rauz S, Young SP (2010) NMR-based metabolomic analysis of cerebrospinal fluid and serum in neurological diseases—a diagnostic tool? NMR Biomed 23:123–132

Spratlin JL, Serkova NJ, Eckhardt SG (2009) Clinical applications of metabolomics in oncology: a review. Clin Cancer Res 15: 431–440

Staab JM, O'Connell TM, Gomez SM (2010) Enhancing metabolomic data analysis with Progressive Consensus Alignment of NMR Spectra (PCANS). BMC Bioinform 11:123

Stein SE, Scott DR (1994) Optimization and testing of mass-spectral library search algorithms for compound identification. J Am Soc Mass Spectr 5:859–866

Tainsky MA (2009) Genomic and proteomic biomarkers for cancer: a multitude of opportunities. Biochim Biophys Acta 1796: 176–193

Tiziani S, Lopes V, Gunther UL (2009) Early stage diagnosis of oral cancer using 1H NMR-based metabolomics. Neoplasia 11:269–276

Van Der Graaf M, Heerschap A (1996) Effect of cation binding on the proton chemical shifts and the spin-spin coupling constant of citrate. J Magn Res B 112:58–62

Vangala S, Tonelli A (2007) Biomarkers, metabonomics, and drug development: can inborn errors of metabolism help in understanding drug toxicity? AAPS J 9:E284–E297

Veselkov KA, Lindon JC, Ebbels TMD, Crockford D, Volynkin VV, Holmes E, Davies DB, Nicholson JK (2009) Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. Anal Chem 81:56–66

Waterman CL, Kian-Kai C, Griffin JL (2010) Metabolomic strategies to study lipotoxicity in cardiovascular disease. Biochim Biophys Acta 1801:230–234

Weiss RH, Kim K, Tolstikov V (2008) Use of urinary metabolomics to identify biomarkers for kidney cancer. Cancer Biomark 4:167–167

Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM (2006) Targeted profiling: quantitative analysis of 1H NMR metabolomics data. Anal Chem 78:4430–4442

Wishart DS (2008a) Applications of metabolomics in drug discovery and development. Drugs R&d 9:307–322

Wishart DS (2008b) Metabolomics: applications to food science and nutrition research. Trends Food Sci Tech 19:482–493

Wishart DS (2008c) Quantitative metabolomics using NMR. Tractrend Anal Chem 27:228–237

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 (2008) The human cerebrospinal fluid metabolome. J Chromatogr B Analyt Technol Biomed Life Sci 871:164–173

Wolfender JL, Glauser G, Boccard J, Rudaz S (2009) MS-based plant metabolomic approaches for biomarker discovery. Nat Prod Commun 4:1417–1430

Woo HM, Kim KM, Choi MH, Jung BH, Lee J, Kong G, Nam SJ, Kim S, Bai SW, Chung BC (2009) Mass spectrometry based metabolomic approaches in urinary biomarker study of women's cancers. Clin Chim Acta 400:63–69

Xi Y, Rocke DM (2008) Baseline correction for NMR spectroscopic metabolomics data analysis. BMC Bioinform 9:324

Xia JG, Bjorndahl TC, Tang P, Wishart DS (2008) MetaboMiner—semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. Bmc Bioinform 9:ARTN 507

Young SP, Wallace GR (2009) Metabolomic analysis of human disease and its application to the eye. J Ocul Biol Dis Infor 2: 235–242