Available online at www.sciencedirect.com
ScienceDirect
Journal of Magnetic Resonance 183 (2006) 269–277
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Weighted least-squares deconvolution method for discovery of group differences between complex biofluid ^{1}H NMR spectra

Geoffrey T. Gipson [a,b,*], Kay S. Tatsuoka [b], Brian C. Sweatman [c], Susan C. Connor [c]

[a] School of Biomedical Engineering, Science, and Health Systems, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104, USA
[b] GlaxoSmithKline, Bioinformatics Division, 709 Swedeland Road, King of Prussia, PA 19406, USA
[c] Safety Assessment, GlaxoSmithKline, Park Road, Ware, Herts SG12 0DP, UK

Received 16 June 2006; revised 1 September 2006
Available online 29 September 2006

Abstract

Biomarker discovery through analysis of high-throughput NMR data is a challenging, time-consuming process due to the requirement of sophisticated, dataset specific preprocessing techniques and the inherent complexity of the data. Here, we demonstrate the use of weighted, constrained least-squares for fitting a linear mixture of reference standard data to complex urine NMR spectra as an automated way of utilizing current assignment knowledge and the ability to deconvolve confounded spectral regions. Following the least-squares fit, univariate statistics were used to identify metabolites associated with group differences. This method was evaluated through applications on simulated datasets and a murine diabetes dataset. Furthermore, we examined the differential ability of various weighting metrics to correctly identify discriminative markers. Our findings suggest that the weighted least-squares approach is effective for identifying biochemical discriminators of varying physiological states. Additionally, the superiority of specific weighting metrics is demonstrated in particular datasets. An additional strength of this methodology is the ability for individual investigators to couple this analysis with laboratory specific preprocessing techniques.

Keywords: NMR; Weighted least-squares; Biomarkers; Biofluid; Reference standard

1. Introduction

Metabolomics is an area of increasing scientific interest and promise. To date, the most widely utilized data generation technologies for mammalian metabolomics investigations have been either ^{1}H NMR- (NMR) or MS-based [1]. NMR is an important "omics" platform because of its ability to readily and reproducibly assay accessible samples from blood, urine, other fluids, or tissue extracts. This makes it an amenable platform to identify and validate key discriminative markers of disease, drug efficacy, toxicity, or other physiological parameters (e.g. gender, age, metabolic status).

Commonly, NMR datasets are analyzed by applying univariate and multivariate statistical approaches to discrete spectral regions in an attempt to identify regions that are altered by a perturbation (e.g. a group difference arising from genetic modification or xenobiotic treatment). Following identification of regions of interest, metabolites with resonances associated with these regions are investigated more closely via manual visual inspection of spectra and additional analytical assays. The chemical shift position and intensity of all NMR resonances for a particular metabolite, which could be termed its 'NMR signature,' are essential for definitive metabolite identification. Based on the NMR signature, a metabolite assignment can often be confirmed unambiguously by comparison with database information, using standard one- and two-dimensional

* Corresponding author. Fax: +1 610 270 5580.
E-mail address: gtg25@drexel.edu (G.T. Gipson).

1090-7807/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2006.09.003

NMR experiments. However, this process can be very time-consuming to do manually, even for known, well characterized entities. Additionally, peak overlap can make this straightforward NMR identification impossible for some metabolites without partial or complete purification prior to NMR and MS analysis. This is particularly the case for some sugars that contain no clear anomeric proton signal and overlapping fatty acid signals.

Direct (absolute or relative) quantification of compound levels via spectral analysis of NMR data would be of great value to metabolomics investigators, yet there are a number of challenges that must be overcome to achieve this task. Biofluid NMR spectra are the integration of many individual overlapping metabolite spectral features (i.e. peaks). In highly proteinaceous biofluids (e.g. blood plasma or serum), low molecular weight metabolites are often protein bound, rendering them less amenable to reliable quantification by NMR, because of line-broadening and loss of NMR visibility [2]. In urine, however, all metabolites above the detection limit with non-labile protons are observed, which leads to highly complex spectra. Additionally, there is a much larger variability in the physico-chemical parameters (i.e. pH, ionic strength, compound concentrations) of urine compared to more homeostatically controlled biofluids such as serum, which can affect the absolute positioning of corresponding peaks across multiple samples [3]. Several techniques are commonly implemented to reduce the impact of peak shift (e.g. spectral region binning, spectral alignment) and continue to be developed and refined to deal with this inter-individual variation [4,5]. As such, while the global quantitative analysis of NMR spectra derived from biofluids and tissue extracts is challenging, signal quantification in urine samples presents additional difficulties.

A number of attempts have been made to decompose NMR spectra into individual components (e.g. independent component analysis, molecular factor analysis) without any prior knowledge of the underlying data structure [6–9]. The primary disadvantage of these methods continues to be the difficulty in interpreting the results within a biochemical context. In other words, since there is no underlying metabolite data structure built into these methods, the components rarely match known metabolite profiles.

Several fitting methods utilizing combinations of empirically derived or modeled reference spectra exist [10–12]. A previous study examining a longitudinal NMR dataset suggested the use of weighted principal components analysis (PCA) to provide an alternative view of the data versus unweighted PCA [13]. However, differentially weighting spectral regions in the process of deconvolving NMR spectra into individual metabolite levels has not previously been described.

Here, we propose the use of a weighted, constrained least-squares algorithm for the estimation and comparison of relative metabolite levels (referenced to control values of the same metabolite) across groups of divergent physiological states. Our aim is to demonstrate that deconvolving complex spectra with the incorporation of a non-uniform weighting scheme, will lead to the identification of metabolites of biological interest that would be missed otherwise. In order to efficiently deconvolve the spectra into individual component spectra, it is often necessary to account for heterogeneous interference. In other words, the signal of certain metabolites of interest may be deeply buried in certain spectral regions, but easily distinguished in others. Additionally, incorporating statistical information about the signal of interest into the deconvolution algorithm can be useful. Previous methods of linear deconvolution (i.e. LCModel) place equal weight on all spectral regions when fitting additive models [10,14]. The novelty of our approach for deconvolving complex NMR spectra lies in the application of a weighted, constrained least-squares method for identifying metabolites that may be discriminative markers of biological effect based on the relative quantitative estimate in context of scaled, control intensities.

2. Experimental

2.1. Spectral decomposition and metabolite detection

The digitization of NMR spectral data is the fine-scale discretization of a continuous phenomenon. Often, investigators find it useful to analyze NMR data at a coarser resolution due to inter-individual peak alignment issues. The process of integrating a spectral region into larger discrete representations is commonly referred to as bucketing or binning. Here, we refer to all discrete spectral representations as "bins". However, it should be noted that the algorithm described here can be applied to discrete spectral data of any resolution, including raw digitized spectra.

An NMR spectrum is the summation of the intensities of multiple, individual metabolite spectra. Though it is unreasonable to assume that an investigator will have a complete (i.e. all compounds present in a given biofluid) set of reference standards, all available, characterized metabolites should be incorporated into the analysis. Eq. (1) expresses the relationship between the observed intensity at bin l of subject j (d_{j1}), the unknown intensity of metabolite k of subject j (m_{jk}), and the relative intensity of known metabolite k in bin l (i_{k1}).

$$d_{j1} = \sum_{k=1}^{n} m_{jk} i_{k1} \tag{1}$$

Since many metabolites are simultaneously detected during a single NMR data acquisition, and the intensity level of individual bins may be a result of contributions from several metabolites, the identification and quantification of individual metabolites measured via NMR is a challenging task. In order to attribute the NMR spectra to individual metabolites, a linear model (Eq. (2)) was used to describe the system and allow for the decomposition of the NMR signal into a series of metabolite signals. An important inherent property of NMR that makes this a

reasonable approach is the linear relationship between concentration and signal intensity and hence the additivity of spectral intensities.

$$D = MI \tag{2}$$

Eq. (2) represents the linear relationship between the matrix of intensity vectors across all individuals (D), the matrix of metabolite intensities across all individuals (M), and the matrix of bin-specific relative intensities across all metabolites (I). Since actual metabolite levels can only have non-negative values, we would like to solve this linear system subject to the constraint that all elements in matrix M are greater than or equal to zero. In order to solve the linear system subject to the inequality constraints, we implemented the penalized constrained least-squares fitting (pcls) function within the mgcv library (version 1.3-1) of R [15–18]. The pcls algorithm finds the minimum sum of squares, subject to the non-negativity criteria (Eq. (3)) through quadratic programming. Although this function has the capability of fitting non-linear, penalized regression splines, for our purposes, we are interested in a weighted, constrained linear fit. As such, the use of penalties is unnecessary since we are interested in a linear response and not higher order models. We iteratively run the pcls function to estimate the M matrix piecewise (M_{calc}) by minimizing a function of the weighting vector (w), individual metabolite vectors (m_j), and individual data vectors (d_j), for each individual in the dataset.

$$\min \|\sqrt{w}(m_j I - d_j)\|^2 \text{ subject to } m_j I > \overline{0} \tag{3}$$

The pcls method requires that the I matrix be of full column rank. Prior to implementing the pcls function, the rank of the I matrix is verified via QR decomposition, and all rank deficiencies are eliminated. Since the I matrix is strictly non-negative, the estimated metabolite intensity levels are constrained from taking negative values. M_{calc} contains information regarding the relative quantities of the characterized metabolites across the individuals in the dataset.

In addition to providing inter-metabolite relative quantities for an individual, M_{calc} can also provide insight into metabolite production between individuals or groups of individuals. For example, the fold change of an individual metabolite k between two groups or the correlation between two metabolites can be calculated using the estimated metabolite levels.

2.2. Penalized, constrained least-squares estimates

Although there are an enormous number of possible weighting vectors to utilize in the least-squares analysis, we have chosen here to examine two non-uniform vectors in addition to a uniform weighting vector. In order to demonstrate the utility of "clear" spectral regions, we will closely examine the case in which a relatively low intensity metabolite, which is found in areas of both high and low interference, is altered between two groups. Incorporation of information regarding the relative interference of the different spectral regions was achieved through using the inverse of the number of observed metabolites in a given spectral region as the weighting vector. Constrained least-squares (CLS) will be used to estimate the underlying metabolite intensity levels both with the inverse metabolite count weighting vector (mCLS) and with a uniform, or non-weighted, vector (nwCLS).

Additionally, we implemented a weighting vector that incorporated the binwise group variance to extract the underlying metabolites of interest (vCLS). More specifically, the weight of each bin was calculated as the inverse of the square-root of the product of the variances ($1/\sqrt{\sigma_1\sigma_2}$) of the bin intensities of the two groups of interest. The mCLS and vCLS weighting factors were implemented with the specific aim of algorithmically placing more emphasis on fitting bins that were less confounded and more consistent across biological replicates, respectively. The algorithm described here was fully developed in R and will be made available upon request.

2.3. Simulations

The generated datasets were simulated in such a way as to closely approximate real NMR spectra, integrated to create sequential bins of width 0.02 ppm. A typical range of NMR data spans about 10 ppm, which reduces to 500 bins, 60% of which we assume to contain metabolite peaks. Additionally, though there are thousands of metabolites that could potentially be measured in biofluids, it is likely that much fewer make up the vast majority of the NMR signal. Here, we assume that the majority of the signal is produced by no more than 300 metabolites and any other metabolites are at or below the limits of NMR detection. While we feel that these assumptions fairly represent a real dataset, the actual number of metabolites making up an NMR signal will be dependent on the sensitivity of the instrumentation being used (e.g. cryo versus non-cryo probe, field strength).

All simulations consisted of 300 metabolites (150 of which were randomly assigned as known, i.e. contained information in the intensity matrix), 300 spectral bins, and 10 subjects (five from each group). An intensity matrix (I matrix) was randomly generated for all 300 metabolites (300 metabolites $\times$ 300 bins) with relative intensity values ($U[0,1]$) for an average of approximately 5 bins per metabolite (drawn from the empirical distribution of our reference standard assignment database) and distributed amongst the bins with probability according to a function of the geometric distribution ($G[p = 0.2] + 1$), yielding an average of approximately five metabolites per bin. The data matrix (D) was then calculated as the matrix product of the simulated underlying metabolite intensity level matrix (M_{init}) and the relative intensity matrix (I), followed by the addition of a baseline (shared across individuals) and simulated instrumental variability (specific to individuals), with intensity values ranging from 0% to 40% and 0% to 10% of the mean metabolite intensity level, respectively. Biological

variation was simulated via sampling individual metabolite levels from a normal distribution when generating M_{init}. Once the D matrix was generated (10 individuals × 300 bins), 150 of the metabolites were randomly withheld from the I matrix in order to simulate the reality of incomplete metabolite information in metabolomics studies.

2.4. Spectral regions with a single metabolite resonance (clear spectral regions)

Non-weighted linear deconvolution methods may miss biologically important compounds when there is a high level of interference in spectral regions and the compound of interest is present in relatively low quantities. To demonstrate this point, we simulated an NMR metabolomics dataset in which the concentrations of an individual metabolite, with peaks in areas of both high and low interference, were significantly different between two groups of subjects (10 individuals per group). M_{init} (20 individuals × 300 metabolites) for this investigation contains one metabolite that is altered in one of the two groups and 299 that have no group difference. The unaltered metabolite intensity levels were sampled from normal distributions with means ranging from 1 to 10 ($U[1,10]$) and standard deviations equal to half the mean intensity value. Altered group intensity levels were sampled from normal distributions with means deviating by a random factor ($U[1.2,5]$) from their baseline counterparts and the same standard deviations. The direction of change of altered group intensity levels could be either positive or negative. Since this method of metabolite level simulation does not strictly preclude the generation of negative values, and negative metabolite levels have no biological meaning in this context, all generated negative values were replaced by zeros.

The I matrix was generated as described previously, with the exception that the number of randomly populated bins was restricted to 299. Following the random generation of the 299 bin I matrix, an additional bin was added in which only the significant metabolite was present.

Univariate statistics ($\alpha = 0.05$) were performed on the metabolite intensity levels estimated via nwCLS, mCLS and vCLS, and M_{init} values for the significantly altered metabolite. Following classification of metabolites as having group differences or not, a receiver operating characteristics (ROC) analysis was performed to compare the sensitivity/specificity profiles of the various weighting methods. The area under the curve (AUC) of the ROC curves was calculated via Somers' rank correlation. Pairwise comparisons (Bonferroni adjusted, paired t-test) were made on 200 simulations to determine if the various methods differed in their ability to successfully identify the simulated metabolite level difference.

2.5. General simulation

In order to evaluate the relative sensitivity/specificity, and to identify any discriminating features of the metabolites identified by different weighting factors, a number of simulations were performed and concurrently analyzed with and without weighting factors. In this investigation 5% (15 of 300) of the initial metabolites in M_{init} (20 individuals × 300 metabolites) were generated to have group specific differences in intensity level.

The metabolite intensity levels were generated in the same way as in the clear spectral region analysis. Since 50% of the metabolite profiles were removed from the I matrix prior to analysis, on average 7–8 metabolites with simulated alterations were available for discovery. Through the use of the CLS methods coupled with univariate statistics, true and false positives were identified. These simulations were replicated 200 times and the sensitivity and specificity of the CLS methods were then compared both to each other as well as univariate statistics on M_{init}, which represents the maximum possible information content.

2.6. Diabetes dataset

A large dataset of Carr–Purcell–Meiboom–Gill (CPMG) NMR spectra from urine samples across diabetic (db/db) and non-diabetic (db/+) mice was analyzed via CLS methods. Male diabetic and control mice (8 weeks of age) were obtained from The Jackson Laboratory (Bar Harbor, ME). Urine samples of 0.5% methylcellulose treated animals were collected over ice twice, one week apart, from mice individually housed in metabolism cages. In urine samples, where there may be a wide range of 'normal' sample ionic strengths and pHs, it may be expected that differences in shift and shape may also occur for resonances experiencing second order coupling (e.g. lysine, ornithine). We have tried to circumvent this issue with the use of buffered samples, including an excess of phosphate buffer. NMR spectral processing consisted of automated adjustment of the chemical shift of TSP to $\delta\text{H} = 0$ ppm, application of a semi-automated phase correction, automated baseline adjustment using an automated 0–2nd order polynomial and reduction to histogram representations by binning using the method by Forshed et al. [19]. A bin width of 0.02 ppm was chosen with a 50% tolerance either side of the bin boundary. Data were scaled using median-difference scaling of the binned data. Further details concerning the experimental protocol and discriminative marker validation can be found in Connor et al. [20].

The nwCLS and vCLS methods were each used to deconvolve the NMR spectra into constituent compound intensity levels and followed by univariate statistical analyses. Putative discriminative markers for disease were identified through a series of Student's t-tests ($\alpha = 0.05$) comparing diseased and control mice. Specifically, a metabolite was considered a putative discriminative marker if an estimated metabolite level was significant in at least 1 of the 2 days of data. We independently proposed putative discriminative markers based on uncorrected and Bonferroni corrected p-values in order to verify that differences

between the CLS methods ability to accurately identify discriminative markers were robust to varying thresholds of discriminative marker inclusion. The results of the CLS method analyses were then compared to results from a previous study in which univariate and multivariate binwise analyses were utilized to identify spectral regions of interest, with subsequent metabolite assignment and independent validation via partial fractionation, LC–MS and 2D NMR. Direct comparisons were made between the validated discriminative marker assignments from the traditional analysis and the putative discriminative markers suggested via CLS methods.

3. Results

3.1. Spectral regions with a single metabolite resonance (clear spectral regions)

In order to evaluate the different linear deconvolution methods, 200 simulations were performed in which one metabolite was altered and clear regions were strictly provided for the significantly altered metabolites. The average AUC for nwCLS, mCLS, vCLS, and univariate analysis of M_{init} were 0.92, 0.95, 0.97, and 0.97, respectively (Fig. 1). Note that univariate analysis of M_{init} yielded an AUC that was less than 1.0 due to the simulated biological variability. Pairwise paired *t*-tests (Bonferroni corrected) were performed on the AUC estimates of each of the CLS methods and univariate statistics on M_{init}. The results of these analyses indicate that all pairwise differences except for vCLS versus M_{init} were significant (nwCLS versus mCLS, $p < 0.05$; all other pairs, $p < 0.005$). The non-significant difference between the vCLS method and univariate statistics on the true underlying metabolite levels indicates that the variance weighting factor has achieved maximal performance in this scenario.

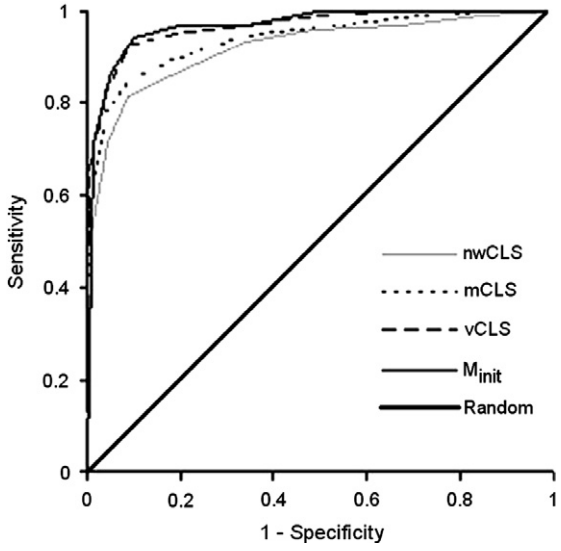

Fig. 1. ROC curves comparing the performance of the nwCLS, mCLS, and vCLS methods and univariate analysis on M_{init} when at least one bin associated with the altered metabolite is uniquely occupied.

3.2. General simulation

In order to further evaluate the capacities of the linear deconvolution methods, 200 simulations were performed in which 5% of the total number of metabolites were altered and clear regions were not strictly provided for the significantly altered metabolites. The average AUC for nwCLS, mCLS, vCLS, and univariate analysis of M_{init} were 0.74, 0.75, 0.80, and 0.97, respectively. Pairwise paired *t*-tests (Bonferroni corrected) were performed on the AUC estimates of each of the CLS methods and univariate statistics on M_{init}. The results of these analyses indicate that all pairwise differences were highly significant ($p < 0.005$), with the exception of nwCLS versus mCLS. These results indicate that each of the CLS methods performed well in accurately discovering group differences, however, the variance weighting factor performed best.

It is not surprising that the significant difference observed in the clear spectral region analysis between nwCLS and mCLS was not also observed in the general simulation analysis. The mCLS method is highly dependent on the quality of the reference spectra (i.e. I matrix). Real spectral libraries produced by laboratories are likely to have prior experience implicitly incorporated through the inclusion of ''expected'' metabolites. In the general simulation, the random nature of the metabolite spectral properties (peak location, intensity, coincidence with other metabolites) and the random population of the I matrix leads to a situation in which any concept of prior experience is not modeled. The advantage of the mCLS method in the clear spectral region simulation was that significantly altered metabolites were exclusively associated with a minimum of one clear bin (maximum weight), thereby providing the least-squares fit with *a priori* information concerning the quality of clear bins. In other words, the prior probability distribution that significant metabolites are associated with clear bins is not uninformed. However, the prior distribution of the general simulation is uninformed, and therefore the specific advantage of the mCLS method is lost. Through evaluating the variance of the bins, the vCLS method captures information concerning the clarity of the individual bins, yet is independent of specific prior knowledge. Although the two weights are similar in that spectral regions with fewer observed metabolites have lower variance, vCLS has the added value of giving additional weight to regions with fewer observed and unobserved metabolites. For this reason, and the superior performance of vCLS compared to mCLS in both simulation analyses, we chose to focus solely on the vCLS method in our analysis of the diabetes dataset.

3.3. Diabetes dataset

An investigation of the ability of the nwCLS and vCLS methods to identify the 46 previously identified and independently validated (LC–MS, 2D NMR, etc.) discriminative markers [20] further demonstrates the utility of using

weighting factors when deconvolving metabolomics datasets. Analysis of the diabetes dataset with nwCLS and vCLS followed by univariate statistics ($\alpha = 0.05$, p-values unadjusted) recovered 38 and 40 of the 46 metabolites, respectively (Table 1). Adjusting the p-values for multiple comparisons led to the discovery of 26 and 27 of the 46 metabolites via nwCLS, and vCLS, respectively.

In addition to the 46 previously confirmed discriminative markers, all methods predicted ''significant'' metabolites from our reference standard database (137 metabolites) that have not been validated (Table 2). Additional putative metabolites beyond the validated 46 may be confirmed as discriminative markers in the future, but were not followed up during the original confirmation process. Since it is not appropriate to designate these putative discriminative markers as false positives, it is not possible to conduct a formal sensitivity/specificity analysis. Instead, we investigated the performance of randomly selecting a number of ''significant'' metabolites, equal to the number of putative discriminative markers proposed by each method, from our reference standard database and calculating how many of these intersect with our confirmed list of 46. We then cal-

Table 1
Confirmed discriminative markers of diabetes and prediction via CLS methods

Metabolite	nwCLS		vCLS	
	Sig.	p[a]	Sig.	p[a]
2-Oxoglutarate	0	0.262	2	<0.001
2-Hydroxyisobutyrate	2	<0.001	2	<0.001
2-Oxoadipate	1	0.017	1	0.008
3-Ureidopropanoate	0	0.052	2	<0.001
Alanine	2	<0.001	2	<0.001
Allantoin	2	0.001	0	0.085
Citrate	1	0.002	1	0.009
Citrulline	0	0.423	2	0.001
Creatine	2	<0.001	1	0.039
Creatinine	2	<0.001	0	0.090
Formate	2	0.001	2	0.001
Fumarate	2	<0.001	2	<0.001
Glucose	2	<0.001	1	0.007
Glutarate	2	<0.001	2	0.001
Glycine	2	<0.001	2	<0.001
Glycolate	2	<0.001	2	<0.001
Guanidinoacetate	2	<0.001	2	<0.001
Hippurate	1	0.003	2	<0.001
Indoxyl sulphate	1	<0.001	2	<0.001
Isobutyrate	2	<0.001	2	<0.001
Isocaproate	2	<0.001	1	<0.001
Isovalerate	0	0.077	0	1.0
Lactate	0	1.0	0	1.0
Lysine	2	<0.001	2	<0.001
Malate	0	1.0	2	<0.001
Malonate	1	0.011	2	<0.001
Methionine	2	<0.001	1	0.001
Methylamine	1	0.002	1	0.002
*N*1-Methyl-2-pyridone-5-carboxamide	0	0.22	2	<0.001
*N*1-Methyl-4-pyridone-3-carboxamide	2	<0.001	0	1.0
*N*1-Methylnicotinamide	2	<0.001	2	<0.001
*N*1-Methylnicotinic acid	1	0.010	2	<0.001
N-Caproylglycine	2	<0.001	2	<0.001
N-Butyrylglycine	2	<0.001	2	<0.001
N-Isobutyrylglycine	2	0.001	0	0.293
N-Isovalerylglycine	1	0.040	2	<0.001
N-Valerylglycine	2	<0.001	2	0.001
Nicotinamide *N*-oxide	2	<0.001	2	<0.001
Orotate	1	<0.001	1	<0.001
Pantothenate	2	<0.001	1	0.010
Phenylacetylglycine	0	0.112	2	<0.001
Sucrose	2	<0.001	2	<0.001
Taurine	2	<0.001	2	<0.001
Threonine	2	0.001	2	<0.001
Trimethylamine	1	<0.001	2	0.006
Valine	2	<0.001	2	<0.001

[a] Value reported is the minimum unadjusted p-value.

Table 2
Discriminative marker prediction performance

Method	Non-adjusted threshold		Adjusted threshold	
	Confirmed/ predicted	p	Confirmed/ predicted	p
nwCLS	38/105	0.168	26/73	0.360
vCLS	40/106	0.042*	27/59	0.007*

* Significant ($\alpha = 0.05$).

culated the probability that the performance observed by the CLS methods could be matched or surpassed through such a process (Table 2). This analysis ($\alpha = 0.05$) revealed that nwCLS was not significantly different from random selection, but vCLS was significantly different. This evidence further supports the idea that using weighting factors can increase the quality of information gained through least-squares analysis of NMR spectra.

Fig. 2 depicts the binned spectral intensities, fitted intensities (vCLS), and the residual intensities for a representative control (db/+) subject. A calculation of the positive (under-explained) and negative (over-explained) residuals reveals that for this individual, 19% of the spectra remains unexplained and the over-explained area is 7% of the original spectra. This same individual, and a representative diabetic (db/db) subject, were evaluated at a higher level of detail to illustrate the capacity of the CLS methods to identify altered metabolites in crowded spectral regions (Fig. 3). Note that both the spectral regions and the underlying metabolite levels are decreased in the db/db spectra. These changes are reflected in the accurate identification of significant decreases in *N*-caproylglycine, *N*-butyrylglycine, and *N*-valerylglycine via both the nwCLS and vCLS methods.

4. Discussion

The results from the simulation analyses demonstrate the utility of incorporating specific domain knowledge into the biomarker discovery process. The ability of CLS methods to accurately identify metabolites associated with group differences is evidenced by the fact that AUC values for all three CLS methods evaluated in this study were significantly increased above the null model. Additionally, the significant increase in the AUC values attained via incorporation of weighting factors indicates that weighted methods can provide a significant improvement in discriminative marker discovery versus non-weighted least-squares.

Weighting factors that maximize the importance of "clear" spectral regions will be increasingly useful as spectral alignment algorithms improve and bin sizes decrease or become altogether unnecessary. While binning spectral regions is a useful tool in dealing with inter-individual alignment variability, it also masks spectral features that can serve to discriminate between metabolites in a given region [21]. Furthermore, since our algorithm is flexible and can deal with heterogeneous bin sizes, regions less affected by alignment problems can be evaluated at a high resolution, while more problematic regions can be grouped in arbitrary bin sizes, thereby maximizing the information gained.

Experimentation with various parameter settings of the simulated datasets (data not shown) revealed the importance of the specific dataset in quantitatively evaluating the various weighting factors. Therefore, an individual weighting factor will have varying strengths and weaknesses depending on the particular dataset in question. Despite the fact that no two NMR datasets are alike, we attempted to simulate what could be considered a typical NMR dataset. It should be mentioned, however, that in the analysis of data from real samples different underlying biological processes will produce different data configurations and therefore are likely to require attention to different details in the data structure. This fact further supports the concept that the use of specific weighting factors can help investigators to analyze their data more effectively.

Furthermore, since there continues to be a great deal of active research in the field of data preprocessing, we have implemented our algorithm within a framework that accommodates such inquiries. This model performs the least-squares fitting at a user defined level of spectral precision that need not be homogeneous within an individual subject. Reference spectra data input is extremely flexible and can be derived from modeled data, spike-in analyses,

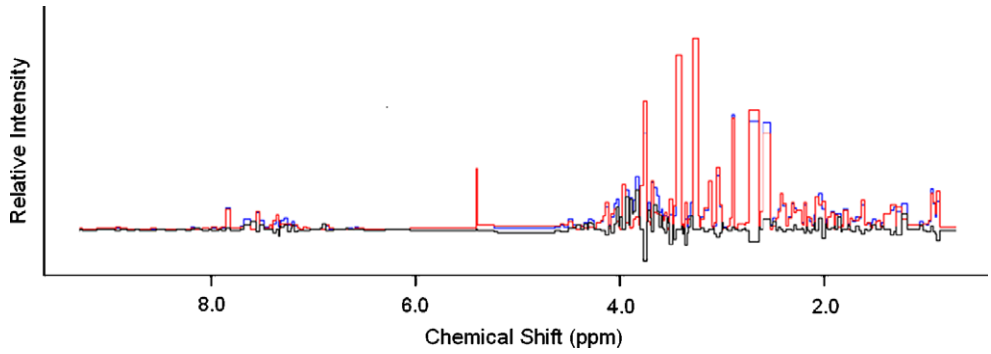

Fig. 2. Binned spectra (blue), fitted vCLS intensities (red), and residual intensities (black) for a representative control (db/+) subject from the first time point.

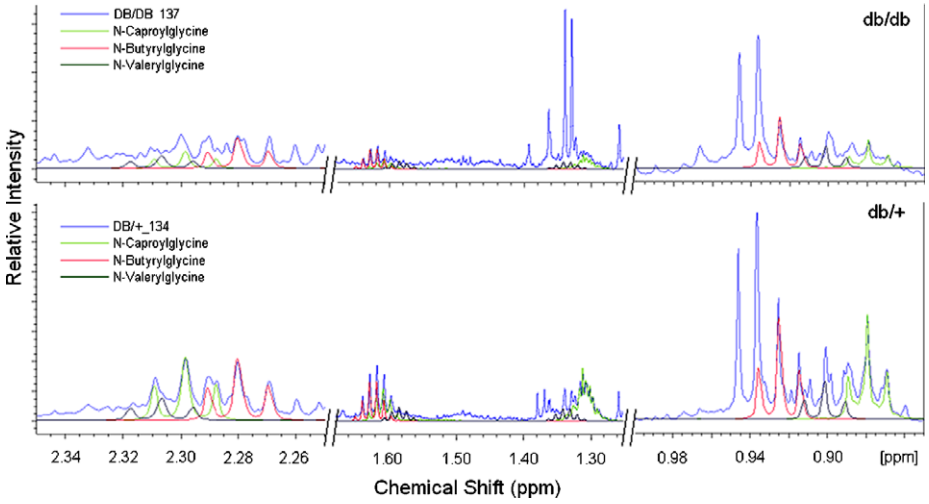

Fig. 3. Diabetic (top) and control (bottom) spectra manually fit with reference spectra. Relative intensity values (y-axis) have been scaled to allow for comparisons between the two individuals.

or literature sources. This can be an important consideration for metabolites that show strong pH dependence to peak shape and position. Though the model has been seen to be robust in the absence of baseline estimates, if desired (e.g. when estimating the protein contribution to the baseline of male mouse urine NMR data [20], externally derived estimates of baseline can also easily be integrated into the reference spectra.

In addition to the default non-negativity constraint, users can also choose to constrain the model to the upper limit of the data matrix (i.e. the model is prevented from over-explaining the data). Typically in NMR datasets, there will be an unequal assignment confidence throughout the spectra, depending on prior knowledge, peak overlap and the degree of analytical confirmation of each component (2D homonuclear and heteronuclear NMR, fractionation, LC–MS confirmation). The described method allows users to experiment with the weighting factor used in the least-squares fit. Here, we critically evaluated weighting factors that were a function of the number of compounds populating a particular spectral region or a function of the group variances. However, there are likely many other weighting factors that will prove useful. For example, it has been seen (data not shown) that a function of binwise correlations can also serve as an effective weighting factor. Binwise correlations are of increasing interest in the field of metabolomics [22,23]. Furthermore, model fits can be restricted to subsets of the spectra either through manipulation of the input dataset or the spectral weighting.

This work attempts to provide tools for the detection and assignment of group differences within a flexible, robust framework for metabolomics investigators to explore and analyze NMR data. While traditional methods of NMR spectral analysis are extremely time-consuming, using the method described here, an investigator can perform a complete analysis in a matter of minutes. Additionally, a successful analytical technique should provide investigators a broad scope of inference. Since different datasets will have different structures, investigators are not limited to a predefined suite of weighting parameters. The sole data input for LCModel is time-domain *in vivo* data and there is no user interaction in the data processing. While we agree with the necessity of inter-laboratory comparability, NMR data preprocessing is still an active area of research and we feel that it is more appropriate for investigators to work within a well-defined, yet less stringent, framework of inquiry. Furthermore, we agree with the conclusions of Jansen et al. [13], though in a different context, that the use of a weighting factor can provide an additional, more focused view of the data. In addition, it is clear that when working with some datasets, it may make the difference between successfully identifying a discriminative marker and missing it altogether. Our method provides a robust, flexible framework for compound level estimation.

Acknowledgments

The authors thank Michal Magid-Slav for comments and David Searls, Mike Lutz, Mike Luther, Terry Ryan, John Haselden and Pankaj Agarwal for their support in preparation of this manuscript.

References

[1] W.B. Dunn, D.E. Ellis, Metabolomics: current analytical platforms and methodologies, Trend. Anal. Chem. 24 (2005) 285–294.
[2] J.K. Nicholson, P.J.D. Foxall, M. Spraul, R.D. Farrant, J.C. Lindon, 750 MHz 1H and 1H–13C NMR spectroscopy of human blood plasma, Anal. Chem. 67 (1995) 793–811.
[3] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids, Concepts Magn. Reson. 12 (2000) 289–320.
[4] N. Trbovic, F. Dancea, T. Langer, U. Gunther, Using wavelet denoised spectra in NMR screening, J. Magn. Reson. 173 (2005) 280–287.

[5] B. Lefebvre. Technical note: intelligent bucketing for metabonomics. <http://www.acdlabs.com/download/technotes/80/nmr/intelli_bucket.pdf/>.
[6] C. Ladroue, F.A. Howe, J.R. Griffiths, A.R. Tate, Independent component analysis for automated decomposition of in vivo magnetic resonance spectra, Magn. Reson. Med. 50 (2003) 697–703.
[7] C.D. Eads, C.M. Furnish, I. Noda, K.D. Juhlin, D.A. Cooper, S.W. Morrall, Molecular factor analysis applied to collections of NMR spectra, Anal. Chem. 76 (2004) 1982–1990.
[8] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, J. Selbig, Metabolite fingerprinting: detecting biological features by independent component analysis, Bioinformatics 20 (2004) 2447–2454.
[9] R. Stoyanova, J.K. Nicholson, J.C. Lindon, T.R. Brown, Sample classification based on Bayesian spectral decomposition of metabonomic NMR data sets, Anal. Chem. 76 (2004) 3666–3674.
[10] S.W. Provencher, Estimation of metabolite concentrations from localized in vivo proton NMR spectra, Magn. Reson. Med. 30 (1993) 672–679.
[11] D.J. Crockford, H.C. Keun, L.M. Smith, E. Holmes, J.K. Nicholson, Curve-fitting method for direct quantitation of compounds in complex biological mixtures using ^{1}H NMR: application in metabonomic toxicology studies, Anal. Chem. 77 (2005) 4556–4562.
[12] Chenomx. <http://www.chenomx.com/>.
[13] J.J. Jansen, H.C.J. Hoefsloot, H.F.M. Boelens, J. van der Greef, A.K. Smilde, Analysis of longitudinal metabolomics data, Bioinformatics 20 (2004) 2438–2446.
[14] S.W. Provencher, Automatic quantitation of localized in vivo ^{1}H spectra with LCModel, NMR Biomed. 14 (2001) 260–264.
[15] S.N. Wood, Monotonic smoothing splines fitted by cross validation, SIAM J. Sci. Comput. 15 (1994) 1126–1133.
[16] S.N. Wood, Modelling and smoothing parameter estimation with multiple quadratic penalties, J.R. Statist. Soc. B 62 (2000) 413–428.
[17] S.N. Wood, Stable and efficient multiple smoothing parameter estimation for generalized additive models, J. Am. Stat. Assoc. 99 (2004) 673–686.
[18] R Development Core Team. R foundation for statistical computing, Vienna, Austria, 2005.
[19] J. Forshed, F.O. Andersson, S.P. Jacobsson, NMR and Bayesian regularized neural network regression for impurity determination of 4-aminophenol, J. Pharmaceut. Biomed. 29 (2002) 495–505.
[20] S.C. Connor et al. (in preparation).
[21] R. Stoyanova, A.W. Nicholls, J.K. Nicholson, J.C. Lindon, T.R. Brown, Automatic alignment of individual peaks in large high-resolution spectral data sets, J. Magn. Res. 170 (2004) 329–335.
[22] O. Cloarec, M. Dumas, A. Craig, R.H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J.C. Lindon, E. Holmes, J. Nicholson, Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets, Anal. Chem. 77 (2005) 1282–1289.
[23] P. Sandusky, D. Raftery, Use of semiselective TOCSY and the Pearson correlation for the metabonomic analysis of biofluid mixtures: application to urine, Anal. Chem. 77 (2005) 7717–7723.