

# An Algorithm for the Automated Quantitation of Metabolites in *in Vitro* NMR Signals

Greg Reynolds<sup>1†\*</sup>, Martin Wilson<sup>2†</sup>, Andrew Peet<sup>2</sup> and Theodoros N. Arvanitis<sup>1</sup>

8th June 2006

<sup>1</sup> Department of Electronic, Electrical and Computer Engineering, University of Birmingham, UK

<sup>2</sup> Academic Department of Paediatrics and Child Health, University of Birmingham, UK

<sup>†</sup> Contributed equally to this work as first authors.

\* Correspondence to: Greg Reynolds, Department of Electronic, Electrical and Computer Engineering  
University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK. Email to: gmr001@bham.ac.uk

Word count: 7000 (approx)

## Abstract

The quantitation of metabolite concentrations from *in vitro* NMR spectra is hampered by the sensitivity of peak positions to experimental conditions. The quantitation methods currently available are generally labour intensive and cannot readily be automated. Here, an algorithm is presented for the automatic time domain analysis of high resolution NMR spectra. The TARQUIN algorithm uses a set of basis functions obtained by quantum mechanical simulation using predetermined parameters. Each basis function is optimised by subdividing it into a set of signals from magnetically equivalent spins and varying the simulated chemical shifts of each of these groups to match the signal undergoing analysis. A novel approach to the standard multidimensional minimisation problem is introduced based on evaluating the fit resulting from different permutations of possible chemical shifts, obtained from one-dimensional searches. Results are presented from the analysis of  $^1\text{H}$  proton magic angle spinning spectra of cell lines illustrating the robustness of the method in a typical application. Simulation was used to investigate the biggest peak shifts that can be tolerated.

**Keywords:** quantitation quantum simulation fitting

# 1 Introduction

$^1\text{H}$  nuclear magnetic resonance spectroscopy ( $^1\text{H}$  NMR) provides information on the concentrations of large numbers of metabolites and is increasingly used for the characterisation of biological samples.  $^1\text{H}$  NMR of human plasma provides information on disease status of patients with cardiovascular disease (1) and is under exploration as a biomarker of other diseases including cancer.  $^1\text{H}$  NMR of biofluids also provides a powerful tool for assessing toxicology (2). Cell and tissue samples may be analysed by either processing them to extract water soluble metabolites and/or lipids prior to  $^1\text{H}$  NMR being carried out, or directly by magic angle spinning  $^1\text{H}$  NMR and provide powerful characteristics for diagnosis and prognosis (3). The automated analysis of these spectra is an important goal in realising the clinical potential of the NMR technique.

The progress made in developing  $^1\text{H}$  NMR as a tool for characterising biological samples is due in large part to the use of pattern recognition methods for analysing spectra (4). All molecules containing a proton in an appropriate environment can give rise to a  $^1\text{H}$  NMR signal and so the number of signals present in biological samples is extremely large. The general approach to this problem is to consider the spectrum (frequency domain) as a “fingerprint” (5) and use methods such as principal component analysis combined with cluster analysis to differentiate between different classes of samples (6). However, the biologically relevant information is the concentration of the individual metabolites and automated methods for determining this from *in vitro* spectra do not exist.

Little has been published about the automated analysis of *in vitro* spectra to determine metabolite concentrations; typically spectrometer manufacturers’ software is used to integrate the area under peaks and multiplets. This process is less successful when there are many overlapping peaks in the region of integration. Techniques such as “deconvolution” (7) are user intensive and require complex knowledge of the many metabolite resonances in order to correctly assign peaks.

An alternative to integration are methods in which peaks are selected, fit and removed one at a time (8). The problem with this approach is that it is difficult to determine the exact resonance of a peak (due to the interference from other peaks) which causes cumulative error as more and more peaks are subtracted. Additionally, these methods over-estimate the quantity of a large peak in a crowded region, having a detrimental effect on other components in that region.

The determination of metabolite concentrations from *in vivo*  $^1\text{H}$  MRS has been achieved by using sets of individual metabolite spectra as a basis for least-squares projection. In some methods

these bases are experimentally acquired (9,10) but more recently an approach based on quantum mechanical simulation has been introduced (11). Other fitting techniques are those based on the SVD/Prony's method (12,13) which involve making various assumptions about Lorentzian line shapes and good SNR. Typically the application of these techniques is limited to *in vivo* application due to the computational cost of working with large numbers of points. Another popular method is AMARES (14) based on interactive assignment of singlets and multiplets. Full reviews of both time and frequency domain methods may be found in (15) and (16) respectively.

It is well known that there is a strong dependence between a spin's resonant frequency and its local environment (17); this fact has even been utilised for the exact measurement of pH (18). For most NMR experiments pH and temperature have to be tightly regulated to ensure that peak positions are reproducible across experiments. A problem specific to the quantitation of signals directly from tissue is that unlike biofluids or extracts, it is currently impossible to control the pH environments of metabolites whilst maintaining sample structure.

Metabolite basis sets are an attractive method of quantitation, they incorporate prior knowledge in a powerful manner and make the quantitation of metabolites with many peaks as easy as the quantitation of singlets. This avoids the need to fit each peak, effectively achieving a dimensionality reduction of the problem. However this approach has not yet been used to quantitate *in vitro* spectra perhaps because the shifts in peak position between one experiment and another are often greater than the peak widths making it difficult to generate accurate basis functions.

In this paper a new algorithm TARQUIN (Totally Automatic Robust QUantitation In NMR) is presented for the metabolite quantitation of high resolution NMR data. TARQUIN uses a quantum mechanically simulated time domain basis set which is refined to account for small changes in the metabolite environments and in particular, peak position. Accurate refinement is a considerable problem and most of this paper is devoted to methods of achieving this. The application of TARQUIN to real magic angle spinning (MAS) data from cell lines is demonstrated. The performance on simulated data is used to investigate robustness to chemical shift variations. A MATLAB implementation of the algorithm described in this paper and some results may be downloaded freely from <http://www.lostintheether.net/tarquin/>

## 2 Method

The discrete-time acquired free induction decay signal of  $N$  samples,  $\mathbf{y} \in \mathbb{C}^N$ , can be modeled as a linear combination of the signals from various metabolites (the basis set) and some additive noise. Writing the time-series signals from the  $M$  metabolites as the columns of a matrix  $\mathbf{S} \in \mathbb{C}^{N \times M}$  and the corresponding amplitudes (and phases) as elements of a vector  $\mathbf{a} \in \mathbb{C}^M$ :

$$\mathbf{y} = \mathbf{S}\mathbf{a} + \mathbf{w} \quad [1]$$

where  $\mathbf{w} \in \mathbb{C}^N$  is a vector representing noise and modeling error. The least-squares estimate (19) of the amplitudes can be obtained from:

$$\hat{\mathbf{a}} = \mathbf{S}^+\mathbf{y} = \mathbf{a} + \mathbf{S}^+\mathbf{w}. \quad [2]$$

From [2], when  $\mathbf{w}$  is small, the quantitation problem is reduced to finding the columns of  $\mathbf{S}$ . In the case of an assumed model function, such as the Lorentzian, the problem is to search for the damping and the frequency of each peak, where each peak is a column of  $\mathbf{S}$ . This is the Variable Projection (VARPRO) (20) approach to fitting. The least-squares formulation may also be implemented using a metabolite basis set for the columns of  $\mathbf{S}$ . Existing methods that use this approach seek to modify the basis set in order to match correctly the acquisition by minimising the residual between the acquired signal and a reconstructed model:

$$\min_{\boldsymbol{\theta}} \|\mathbf{y} - \hat{\mathbf{y}}(\boldsymbol{\theta})\| \quad [3]$$

where  $\hat{\mathbf{y}}$  is an estimate of the signal constructed from the basis set and modified by frequency shift, damping and phase parameters in  $\boldsymbol{\theta}$ . There are several ways in which a basis set can be constructed; the two popular methods are experimental acquisition and simulation. Experimentally acquired basis sets possess several features which can make them difficult to use: noise, imperfect lineshape, inter-magnet variation, field-strength dependence and the effort involved in acquisition. An alternative approach is to simulate the basis sets from a theoretical model employing prior knowledge of the parameters from experiments. It is this approach that is used here.

The algorithm for generating and matching the basis functions is now described; subsequent sections of the paper will expand on this overview. The first step is to generate a set of signals corresponding to metabolites using known J-coupling and chemical shift values. Each metabolite signal is divided into “groups” of signals from magnetically equivalent spins. These groups are then adjusted to fit their corresponding counterpart in the acquired signal. This adjustment involves

shifting the resonance frequency and computing a damping and phase for each group; a process which is described in detail. An overview of the entire quantitation process is shown in Fig. 1.

## 2.1 Initial Basis Set Simulation

The quantum mechanical simulation of  $^1\text{H}$  NMR signals from small molecules has been extensively described (21). Detailed *a priori* knowledge of metabolite chemistry makes accurate simulation possible and an implementation of the method in (21) was developed as part of this work. Quantum mechanical operators are applied to the density matrix describing the spin system, in order to simulate the time domain signal of each metabolite. Each NMR experiment is modeled by the sequential application of the following operators: thermal equilibrium, RF-pulse, free-evolution and acquisition. For a simple proton one-dimensional homonuclear experiment it is assumed that the spin system begins in a state of thermal equilibrium represented by the the thermal equilibrium operator. Pulse sequences are modeled by sequential application of the RF-pulse and free-evolution operators. The experiment is completed by applying the acquisition operator, ultimately producing sets of frequencies,  $\Omega$ , and amplitudes.

The Hamiltonian directly dictates the free evolution dynamics of a spin system. For a motionally averaged homonuclear spin system, the Hamiltonian can be accurately represented by considering the chemical shift interaction and the J-couplings between spins in the same molecule:

$$\hat{\mathcal{H}}^0 = \sum_{j=1}^n \Omega_j \hat{\mathbf{I}}_{jz} + \sum_{\forall j < k}^n 2\pi J_{jk} \hat{\mathbf{I}}_j \cdot \hat{\mathbf{I}}_k \quad [4]$$

where:

$$\hat{\mathbf{I}}_j = \hat{\mathbf{I}}_{jx} \mathbf{e}_x + \hat{\mathbf{I}}_{jy} \mathbf{e}_y + \hat{\mathbf{I}}_{jz} \mathbf{e}_z \quad [5]$$

where  $\mathbf{I}_{jz}$  is the  $z$ -component of the  $j^{\text{th}}$  angular momentum operator,  $\Omega_j$  is the chemical shift of the  $j^{\text{th}}$  spin in the rotating frame of reference,  $J_{jk}$  is the J-coupling interaction between spins  $j$  and  $k$  and  $n$  is the number of spins in the molecule. All of these parameters can be represented by a lower-triangular matrix  $\mathbf{J}$  and vector  $\mathbf{\Omega}$ . For example, lactate is completely specified for simulation purposes by data taken from (22):

$$\mathbf{\Omega} = \left[ \begin{array}{cccc} 4.0974 & 1.3142 & 1.3142 & 1.3142 \end{array} \right]^T \quad [6]$$

$$\mathbf{J} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 6.933 & 0 & 0 & 0 \\ 6.933 & 0 & 0 & 0 \\ 6.933 & 0 & 0 & 0 \end{bmatrix} \quad [7]$$

## 2.2 Basis Set Refinement

An important property of the J-coupling interaction is that it remains largely invariant to changes in temperature and pH. Chemical shift however varies with temperature and pH so the chemical shift vector needs to be specified precisely for a particular experiment. In a novel extension to the basis set technique, small variations in pH and temperature due to physiological processes are accounted for as part of the fitting process. This is achieved by modeling each metabolite as a combination of separate groups, where each group is a collection of magnetically equivalent spins. At high field strengths, this normally corresponds to breaking a metabolite’s signal up into separate multiplets (groups) well separated by frequency. These groups are formed within the simulation by manipulation of the equilibrium operator. If  $\mathbf{s}_i \in \mathbb{C}^N$  is the complete metabolite and signal, it may be represented as:

$$\mathbf{s}_i[n] = \sum_{k=1}^{G_i} \mathbf{g}_k^i[n] \quad [8]$$

where  $G_i$  is the number of groups in the metabolite,  $\mathbf{g}_k^i \in \mathbb{C}^N$  represents the time-series for each group  $k$  of metabolite  $i$  and  $n$  is the sample index. The time series for each group is the summation of many (undamped) complex exponentials at the frequencies and amplitudes determined by the simulation. Ultimately this means that the lineshape of the basis set is Lorentzian, although the method does not depend on this assumption since the rest of the process is in terms of group signals. For example, the basis set could be simulated with a different model function to take account of poor shimming.

Once a basis set has been simulated it is refined to match better the signal being quantitated. The main part of the algorithm begins by forming the “group matrix”  $\hat{\mathbf{G}} \in \mathbb{C}^{N \times P}$  where  $P = \sum_{i=1}^M G_i$ ,

$$\hat{\mathbf{G}} = \begin{bmatrix} \hat{\mathbf{g}}_1^1 & \dots & \hat{\mathbf{g}}_{G_1}^1 & \hat{\mathbf{g}}_1^2 & \dots & \hat{\mathbf{g}}_{G_2}^2 & \dots & \hat{\mathbf{g}}_1^M & \dots & \hat{\mathbf{g}}_{G_M}^M \end{bmatrix} \quad [9]$$

Each column of  $\hat{\mathbf{G}}$  is synthesised from the frequency and amplitudes determined by simulation and is undamped. The exact resonant frequencies, the dampings and phase of each group vary slightly between each experiment. Prior to forming the metabolite signal  $\mathbf{s}_i$  we need to modify each group

to match the signal currently undergoing analysis:

$$\mathbf{s}_i[n] = \sum_{k=1}^{G_i} \hat{\mathbf{g}}_k^i[n] \exp(j\phi_k^i) \exp((j\Delta\omega_k^i - \lambda_k^i)n\Delta t) \quad [10]$$

where  $\Delta t$  is the sampling interval,  $\phi_k^i$  is the phase,  $\Delta\omega_k^i$  is the shift from the initial frequency and  $\lambda_k^i$  is the damping of each group. It was noticed experimentally that a good model for the  $\Delta\omega_k^i$  due to the unknown pH effect is:

$$\Delta\omega_k^i = 2\pi(\zeta + \xi_k^i) \quad [11]$$

where  $\zeta$  represents an initial offset per experiment and  $\xi_k^i$  is a Gaussian distributed random variable with variance dependent on unknown experimental conditions (in the data presented later,  $\xi_k^i$  has a standard deviation  $\approx 0.001$  ppm). The  $\zeta$  parameter is calculated by solving:

$$\min_{\zeta} \|\mathbf{y} - \hat{\mathbf{g}}_k^i(\zeta)(\hat{\mathbf{g}}_k^i(\zeta))^+ \mathbf{y}\| \quad [12]$$

where  $\hat{\mathbf{g}}_k^i$  is a group  $k$  of some metabolite  $i$  that is found in the spectra being quantitated. The choice of  $\hat{\mathbf{g}}_k^i$  depends on the data being quantitated, typically this is a well separated peak that can be readily assigned and is found in all spectra to be analysed. The argument  $\zeta$  indicates that  $\hat{\mathbf{g}}_k^i$  has been shifted by some frequency  $\zeta$ . Note that in [10] the frequency and damping factors are constrained;  $-2\pi f_{max} \leq \xi_k^i \leq 2\pi f_{max}$  and  $\lambda_{min} \leq \lambda_k^i \leq \lambda_{max}$ . It is also necessary to form the matrix  $\mathbf{G} \in \mathbb{C}^{N \times P}$  at this stage. This matrix represents the basis set as it is updated through the process. Initially it is identical to  $\hat{\mathbf{G}}$  except that a typical damping,  $\lambda_{typ} \approx 6$  Hz, is applied to each column.

The fundamental problem of analysing NMR data is the interference between components; each exponentially decaying oscillation has a bandwidth that spans the entire signal. The approach here is to break the signal up into segments; where each segment contains groups that are within a threshold,  $\text{tol}_{\text{seg}}$ , of each other ( $\text{tol}_{\text{seg}} \approx 0.05$  ppm at a field strength of 600 MHz). Each segment then has the corresponding groups in the basis set adjusted until they best fit the segment. The procedure for segmenting the signal works by ordering the groups in increasing order of “group frequency” where group frequency is defined as the frequency of the biggest peak in the Fourier transform of the signal for each group. The Fourier transform is necessary because the output of the simulation specifies a group in terms of many frequencies each with their own amplitude; it would be difficult to directly use these numbers to determine a frequency representative of the group. However, in the case of singlets, a group’s single frequency is used directly from the output of the simulation. The  $P$ -dimensional vector of group frequencies is denoted  $\mathbf{v}$ .

To reduce interference from signals outside the segment, the groups not in the segment are subtracted (in a manner inspired by (8)) using a least-squares estimate to form a residual undergoing analysis. Thus, for a segment  $D$  that contains groups within the tolerance, we form the residual:

$$\mathbf{z} = \mathbf{y} - \mathbf{G}_{-D}\mathbf{G}_{-D}^+\mathbf{y} \quad [13]$$

where  $\mathbf{G}_{-D}$  is the matrix  $\mathbf{G}$  without the columns corresponding to the groups in the set  $D$ . It is the need to create an effective residual that motivated the application of the typical damping,  $\lambda_{typ}$ , above. From this residual the specific adjustment parameters of [10] are determined by solving the following minimisation problem:

$$\min_{\Delta\omega, \lambda} \|\mathbf{z} - \hat{\mathbf{z}}\| \quad [14]$$

where  $\Delta\omega \in \mathbb{R}^b$  and  $\lambda \in \mathbb{R}^b$  and  $b$  is the number of groups in the segment. The estimate  $\hat{\mathbf{z}}$  is constructed in the usual manner, i.e.  $\hat{\mathbf{z}} = \mathbf{G}_D\mathbf{G}_D^+\mathbf{z}$  where  $\mathbf{G}_D$  is a matrix with columns corresponding to the groups in  $D$ . The solution of [14], the ‘‘identification equation’’ is explained in the next section.

### 2.3 Solving the Identification Equation

In its initial form, [14] is a minimisation problem in  $2b$  dimensions. The landscape of this problem possesses a large number of local minima making the solution using a standard multi-dimensional optimisation algorithm (e.g. Levenberg-Marquardt) difficult; traditional methods often present only local solutions. The approach here exploits knowledge of the problem in order to make the solution tractable. The first step is to perform a series of evaluations of a one-dimensional form of [14] over the frequency range with no damping applied; we form the landscape vector:

$$\mathbf{l}_d[u] = \|\mathbf{z} - \hat{\mathbf{z}}_d(u)\| \quad [15]$$

where  $d = 1, 2, \dots, b$  denotes the current group within the segment and  $\hat{\mathbf{z}}_d(u)$ , the estimate of that single group at frequency position proportional to  $u$ , is defined by:

$$\hat{\mathbf{z}}_d[n](u) = \hat{\mathbf{G}}_d[n] \exp(j2\pi(-f_{max} - \psi_d + u\Delta f)n\Delta t) \quad [16]$$

where  $\hat{\mathbf{G}}_d \in \mathbb{C}^N$  is the group of the segment for which this one-dimensional search is being performed,  $\Delta f$  is the step-size of the search and  $u$  is the integer index variable and ranges from zero to the number of points in the search minus one, e.g. for our data our search resolution had 50

points, so  $u = 0, 1, \dots, 49$ .  $\psi_d$  is the “collision offset” and is the remainder of  $v_d/\Delta f$ . Its presence ensures that the landscape vector has an identical numerical value at ambiguous positions of identical groups. For example, in the choline region of the spectrum, glycerophosphorylcholine and phosphorylcholine are close together and a mathematically valid solution would be to swap the position of their singlets around. Later, a method for detecting and correcting for this is described. The collision offset also ensures that if two singlets are adjusted to have identical frequencies then the smallest singular value of  $\mathbf{G}_D$  will be zero. This fact is used to prevent this solution from being considered.

Once  $\mathbf{l}_d$  has been computed for each group in the segment (see Fig. 4), the minima are determined in the obvious way by a numerical differencing method. The  $b$  sets of minima are arranged to form the “permutation matrix”. From the permutation matrix all possible combinations of  $b$ -dimensional “trial vectors” are selected (the Cartesian product of all sets of minima). An evaluation of the identification equation is performed for each trial vector and the best candidate selected as the set of shifts,  $\Delta\omega$ , that will be used. Note that when evaluating the identification equation for each trial vector, as well as applying the frequency shifts, an approximate estimate of the damping must also be applied, e.g.  $\lambda_{typ}$ .

Fig. 3 shows the crowded choline region of the spectrum. In the basis set used here there are five groups in this region with contributions from myo-inositol, taurine, choline, glycerophosphorylcholine and phosphorylcholine. The correct positions of the groups are shown in the figure. Only an exhaustive search for the correct solution will find the best fit. Fig. 4 shows the corresponding landscapes for Fig. 3 produced by frequency shifting each group past the signal. It is from these landscapes that the minima are obtained to construct the permutation matrix.

The best candidate solution, and therefore the obvious choice, is defined as that which has the lowest residual. However in some cases metabolites may have identical groups (e.g. the choline singlets) and so the number of solutions is ambiguous (two trial vectors give the same result). When this analytically-unsolvable ambiguity occurs, the “best” solution is that which has the smallest total shifts, this being the most likely in the general case (pick the trial vector with the smallest  $\|\Delta\omega\|$ ). This is one of the reasons for the collision offset  $\psi_d$  in [16].

Finally the values of  $\lambda$  and fine frequency adjustment are performed simultaneously through the use of a standard global optimisation algorithm (23). This simultaneous adjustment is necessary to take account of the interference between peaks; crucially however the combinatorial approach of different sets of minima ensures that the starting point of the frequency parameters is correct.

It may also be necessary to truncate points at the start of the FID when estimating the damping, in order to allow any lipids to decay first (11). In much of our experimental data 1000 points were discarded to allow the lipid to decay. This loss of points clearly reduces the SNR, and some smaller peaks may no longer be visible as a result of this. Crucially however the frequencies of each peak have largely been determined in the method described above. The columns  $\mathbf{G}$  are updated to reflect the parameters just estimated:

$$\mathbf{G}[n, d] = \hat{\mathbf{G}}[n, d] \exp((j\Delta\omega_d - \lambda_d)n\Delta t) \quad \text{for } d = 1, 2, \dots, b \quad [17]$$

### 2.3.1 Technical Details of Implementation

Each landscape vector will typically have a different number of minima, but if this is assumed to be constant then it is necessary to evaluate the identification equation for  $\approx m^b$  trial sets, where  $m$  is the number of minima. For large complex segments this results in numbers of trial vectors that are too big to be practical. To reduce the number of evaluations of [14] it is necessary to select only a subset of minima from those available. In the experiments described here, the nearest  $\Theta \in \mathbb{N}$  minima were used, where nearest refers to the smallest frequency shift from the original position after simulation. The accuracy of the algorithm can sometimes be improved by permitting more minima to be considered.

An alternative minima selection scheme would be to pick only a few of the biggest minima in each landscape, rather than those representing the smallest shift. This may make the algorithm more robust to real data where there are many very small minima throughout the landscape due to the noise.

## 2.4 Iterative Improvement of Basis Set

Following the adjustment of all segments, the basis set is re-simulated and synthesised at the new frequencies just determined. This is to allow for strong-coupling interactions between the groups to be taken into account within the simulation and, as an additional benefit, allows for the effects of receiver dead-time. The algorithm then repeats the previously described process, using the new basis set, for as many iterations as deemed necessary by the user. Typically, for weakly-coupled signals the parameters converge within only one or two iterations; for the more strongly-coupled such as glutamate typically five or six iterations are needed before convergence is achieved.

Iterations improve both the damping and frequency estimates of all groups. On the first itera-

tion, if no groups are assumed to be in the correct position, the cumulative effect of the interference may be significant. Once groups have been adjusted for each segment, their subtraction from the signal on subsequent iterations will lead to an improved optimisation landscape for the identification equation.

In order to increase speed, the frequency range on all subsequent iterations can be reduced, since the frequency of the groups is assumed to be approximately correct after the first iteration. Note that no re-simulation takes place on the final iteration before quantitation. This is because the values produced from the optimisation process are intended to match the current basis set.

## 2.5 Final Quantitation

One of the advantages of considering each metabolite to be composed of groups is that those parts of the signal which did not fit well can be discarded from the quantitation. This may be the case if one of the groups is in a particularly crowded segment of the spectrum. The simple criterion used here is whether the value of  $\lambda$  for a particular group is  $\lambda_{max}$ ; if so then since this is an unlikely value, the group signal is set to zero before being used in the metabolite. This procedure also means that those metabolites present in the basis set but not in the signal will be removed, since in the absence of a component to fit to, their damping attains the maximum value. After the columns of  $\mathbf{G}$  have been damped and shifted to match the signal it is necessary to adjust the phase of each group before summing them as in [8] to form the metabolite signal. The fact that the phase is computed from each group means that any frequency dependent phase shift present in the original signal will also be present in the basis. In the ideal case with all parameters perfectly estimated and no dead-time at the start of the signal all the phases would be zero; but this is seldom the case for real-world data. The phase term is computed from the least-squares estimate of the group amplitudes,  $\hat{\mathbf{a}} = \mathbf{G}^+ \mathbf{y}$ :

$$\phi_k^i = \arctan \left( \frac{\text{Im}\{\hat{a}_k^i\}}{\text{Re}\{\hat{a}_k^i\}} \right) \quad [18]$$

and the metabolite signal is formed from the adjusted basis set as in [10]. Estimates of metabolite amplitudes are computed using [2].

## 2.6 Experimental Methods

### 2.7 Application to MAS $^1\text{H}$ NMR of Cell Lines

To assess the effects of baseline, noise and lineshape imperfections the algorithm was run on a set of 54 1D pulse and acquire MAS spectra taken on 14 neuroblastoma cell lines each with three or more repeats. Cells were grown to confluence in single 75 cm<sup>2</sup> flasks and were harvested by removing the medium and washing the cells three times whilst still adherent to the flask with 3 ml of ice cold phosphate buffer solution. The cells were then removed from the flask using a manual scraper and centrifuged at 250g for 6 minutes to form a pellet which was snap frozen in liquid nitrogen and stored at  $-80^\circ\text{C}$ . Just prior to MAS  $^1\text{H}$  NMR the cells were defrosted and 36 $\mu\text{l}$  pipetted into a wide mouthed Zirconium sample tube (Varian Inc, USA) and 4 $\mu\text{l}$  of 10mM TMSP dissolved in D<sub>2</sub>O was added as a chemical shift standard.

Spectra were acquired on a Varian 600 MHz spectrometer fitted with an Inova console and a 4mm gHX nanoprobe. A simple  $^1\text{H}$  pulse sequence was used consisting of a one-second pre-saturation pulse for water suppression and then a  $90^\circ$  pulse followed by acquisition of 16384 complex points at a sampling frequency of 7198.19 Hz corresponding to 12 ppm. 256 transients were acquired to provide good SNR giving a total experiment time of 14 minutes. The SNR (in dB) is defined as ten times the log to base ten of the ratio of the square of the biggest peak's height in the magnitude spectrum, to the variance of the noise in an unoccupied part of the magnitude spectrum. The sample was spun at 2500 Hz and the probe head temperature was regulated at  $0.1^\circ\text{C}$ . Baring air was cooled to  $-10^\circ\text{C}$  and drive air was supplied at room temperature. These conditions were found to give a consistent methanol calibrated temperature of  $6.7^\circ\text{C}$  inside the rotor.

The basis set was constructed to have frequencies relative to the frequency of the TMSP peak (defined to be at 0 ppm). Prior to fitting, the lipid and macromolecule content of the signal was suppressed by truncating the first 1000 points of the FID (step one of the method as in (11)). The fit parameters used were:  $f_{max} = 0.003$  ppm,  $\lambda_{min} = 0$  Hz,  $\lambda_{typ} = 6$  Hz,  $\lambda_{max} = 20$  Hz,  $\Theta = 10$ ,  $\text{tol}_{seg} = 0.05$  ppm. The algorithm typically ran for ten minutes per sample on a SUN Microsystems V240 dual 1.28GHz CPU computer.

### 2.8 Application to Simulated Data

In order to test in a controlled manner the ability to tolerate chemical shift perturbation, the method was tested by fitting to a simulated spectrum constructed from 18 metabolites. Each

group of each metabolite in this spectrum was shifted by a Gaussian distributed random number with lower and upper limits of  $-0.02$  ppm and  $+0.02$  ppm, concordant with the typical size of a “bucket” in spectral partitioning techniques (e.g. (24)). Our experiments indicate that these limits are of the same order as those seen in cell-line spectra. Different trials were made for different variances of the perturbation.

The signal was then analysed using exactly the same basis set that was used to construct it, except that no knowledge of the frequency perturbation was available. With the exception of phosphorylcholine all the metabolites had unit amplitude. The amplitude of phosphorylcholine was increased significantly so that the simulated spectrum exhibited one of the most troublesome features of real spectra; a large peak sitting in the middle of many smaller peaks. This problem is also demonstrated by the presence of glutamate, glutamine and NAA. Results are shown in Table 1.

The simulated spectra were calculated at 16384 points, with a transmitter frequency of 599.8 MHz and a sampling frequency of 7198.19 Hz, using a pulse and acquire sequence. All groups were damped using  $\lambda = 4$  Hz. The basis set was segmented using  $\text{tol}_{\text{seg}} = 0.05$  ppm which resulted in 13 segments. All segments used  $\Theta = 5$  when computing candidate solutions; except for the troublesome segment in the choline region ( $\Theta = 9$ ) and the obviously complicated glutamate/glutamine region ( $\Theta = 11$ ). The other fit parameters used were:  $f_{\text{max}} = 0.003$  ppm,  $\lambda_{\text{min}} = 0$  Hz,  $\lambda_{\text{typ}} = 3$  Hz,  $\lambda_{\text{max}} = 20$  Hz.

### 3 Results and Discussion

#### 3.1 Application to MAS $^1\text{H}$ NMR of Cell Lines

Fig. 2 and Fig. 3 show results of a fit to one of the neuroblastoma cell lines; the in-phase and visually acceptable nature of the fit are good indicators of the success. For the data analysed, SNR values were in the range 50dB to 75dB and this appeared to have little effect on the success of the overall fit. Groups with a SNR of as low as 42dB were seen to fit. Of the 54 spectra quantitated, 51 had a clearly discernible, well separated creatine peak and this was chosen to determine the initial shift  $\zeta$ . In the remaining three spectra phosphorylcholine was used to compute  $\zeta$  as creatine was not present. Of the 18 metabolites in the basis set, an average of 93% were fit correctly per spectrum, where “correct fit” means that metabolites present were well fit and metabolites not present were correctly discarded. Based on visual examination of the spectra and their fits it appears that the reason for the 7% failure is due to deviations from the  $\zeta + \chi_k^i$  model of [11]. The success rate of

93% indicates that [11] is clearly a good approximation and is satisfactory for this application.

A minor problem for the fitting of strongly coupled chemicals is that temperature dependent J-coupling variations are more pronounced. This did not cause significant problems for the data used here, but it may be necessary to obtain J-coupling data for certain metabolites at temperatures significantly different from those presented in (22).

Fig. 5 demonstrates the general validity of the shift model presented in this work. The initial offset shift is present in all the shifted peaks, as can be seen in the figure. On top of this shift, there is also a much smaller shift (not so visible in the figure) that is unique to each peak.

Fig. 3 shows one of the significant benefits of the combinatorial approach initialisation of the shifts passed to the optimiser. It would be difficult to quantify correctly the groups of myo-inositol and taurine shown (at approximately 3.26 ppm). The combinatorial method ensures that the optimiser gets the shift positions corresponding to the correct peak positions, so that the damping can correctly be computed. Fig. 3 also shows that the method is tolerant to an overcomplete basis set. In the basis set used to analyse this spectrum, both phosphocreatine and creatine were present. In the fit, phosphocreatine has successfully been removed (by the damping criteria), despite the fact that it is close to creatine.

Fig. 2 shows the ability of the method to fit to strongly coupled signals even in the presence of broad overlapping signals. The re-simulation of the glutamate and glutamine basis vectors at each iteration of the simulation ensures that the heavily chemical-shift-dependent multiplets match the signal. Fig. 2 also shows that the algorithm is robust to the presence of small components of the signal that are not present in the basis set. In the case of bigger peaks that are not in the basis set and not well separated from other components of the spectrum, the fit will almost certainly not work in that segment. For example, the removal of phosphorylcholine from the basis set used in Fig. 3 would result in that peak being assigned to glycerophosphorylcholine and the true glycerophosphorylcholine peak being unassigned. Additionally, the taurine group would have an increased amplitude due to the unaccounted for parts of the signal. A solution to this is to reduce  $f_{max}$  so that groups cannot move to occupy an incorrect position, however this reduces the maximum possible value of  $\xi_k^i$  and hence the applicability of the method to spectra with large chemical shift perturbations. Any choice of  $f_{max}$  must be a compromise between the completeness of the basis set and the maximum value of  $\xi_k^i$  present in the data.

### 3.2 Application to Simulated Data

The results shown in Table 1 are promising; they indicate that the method is robust to frequency perturbations. Clearly though, as the shifts get bigger the variance of the estimates increases and the accuracy is reduced for a given  $\Theta$ . Given more minima, the accuracy can be increased and even bigger shifts tolerated, but at the cost of computation time. One way of improving the speed of the algorithm would be to parallelise the evaluations of the identification equation for sets of the trial vectors. This is possible since each operation does not depend on the result of the previous and would clearly lead to a speed improvement due to the exponential complexity of the problem. This is well suited to most types of computer clusters since the amounts of data to be transferred are minimal.

The algorithm in the form presented makes the assumption that the basis set is nearly-complete, i.e. most of the signal can be accounted for in the basis set. In the case when the basis set is undercomplete it is proposed that the basis set should be augmented with singlets obtained by one of a number of “peak-picking” algorithms, e.g. (8,12) for the initial analysis. This is necessary to ensure that the interference from the unknown components does not have an adverse effect on the known compounds. The basis set could then be expanded by determining the parameters of the next metabolite and adding it for subsequent analyses.

In the case when the basis set is overcomplete, i.e. there are more metabolites in the basis set than there are in the signal, it is proposed that the backward greedy optimal basis selection algorithm (25,26) should be used in between iterations to prune unwanted components. This algorithm could also be used to rank metabolites in terms of how much signal they account for, as an approximate guide in interpretation.

Unlike algorithms with an experimentally acquired basis, the simulation parameters can be optimised for each experiment. A potential future improvement could be to alter the model function of the basis set to account for poor shimming, lineshape deviations, etc.

## 4 Conclusion

In conclusion, an algorithm was presented, which described the generation of a basis set with sufficient degrees of freedom for it be matched to a measured signal. This extra freedom was obtained by dividing each metabolite signal into groups of magnetically equivalent spins to form a new basis. A novel method of adjusting the basis to match the signal undergoing analysis was

described, based on converting the standard continuous multi-dimensional optimisation problem into a combinatorial one. Results were presented that show the method gives a reliable fit for real data with variable levels of noise, baseline imperfections and incompleteness of the basis.

## 5 Acknowledgments

Greg Reynolds has a Ph. D. studentship funded by the EPSRC. Martin Wilson has a Ph. D. studentship funded by the European Union Framework 6 project eTUMOUR. Andrew Peet holds a Department of Health National Clinician Scientist Award. The authors would like to thank Dr Carmel McConville for her work on the neuroblastoma cell lines. We would also like to thank the anonymous reviewers whose comments greatly improved the resubmitted version of this paper.

## References

1. Brindle JT, Antti H, Holmes E, Tranter G, Nicholson JK. Rapid and non-invasive diagnosis of the presence of coronary heart disease using  $^1\text{H}$  NMR based metabonomics. *Nature Med* 2002;8:1439-1444.
2. Griffin JL, Bollard ME. Metabonomics: its potential as a tool in toxicology for safety assessment and data integration. *Curr Drug Metab* 2005;280:7530-9.
3. Griffin JL, Shockcor JP. Metabolic profiles of cancer cells. *Nat Rev Cancer* 2004;4:1128-34.
4. El-Dereby W. Pattern recognition approaches in biomedical and clinical magnetic resonance spectroscopy: A Review. *NMR Biomed* 1997;10:99-124.
5. Fiehn O. Metabolomics-the link between genotypes and phenotypes. *Plant Mol Biol* 2002;48:155-71.
6. Stoyanova R, Brown TR. NMR spectral quantitation by principle component analysis. *NMR Biomed* 2001;14:271-277.
7. Swansons GM, Zektzer AS, Tabatabai L, Simko J, Jarso S, Keshari K, Schmitt L, Carroll P, Shinohara K, Vigneron D, Kurhanewicz J. Quantitative Analysis of Prostate Metabolites using  $^1\text{H}$  HR-MAS Spectroscopy. *Magn Reson Med* 2006;55:1257-1264.

8. Umesh S, Tufts DW. Estimation of Parameters of Exponentially Damped Sinusoids Using Fast Maximum Likelihood Estimation with Application to NMR Spectroscopy Data. *IEEE Trans Signal Process* 1996;44:2245-2259.
9. Provencher SW. Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magn Reson Med* 1993;30:672-679.
10. Provencher SW. Automatic quantitation of localized in vivo  $^1\text{H}$  spectra with LCModel. *NMR Biomed* 2001;14:260-264.
11. Ratiney H, Sdika M, Coenradie Y, Cavassila S, Ormond D, Graveron-Demilly D. Time-domain semi-parametric estimation based on a metabolite basis sets. *NMR Biomed* 2004;18:1-13.
12. Kumaresan R, Tufts DW. Estimating the parameters of exponentially damped sinusoids and pole-zero modelling in noise. *IEEE Trans Accoustic Speech Signal Process*, 1982;30:833-840.
13. Laudadio T, Selen Y, Vanhamme L, Stoica P, Van Hecke P, Van Huffel S. Subspace-based MRS data quantitation of multiplets using prior knowledge. *J Magn Res*, 2004;168:53-65.
14. Vanhamme L, van den Boogart A, Van Huffel S. Improved Method for Accurate and Efficient Quantification of MRS Data with Use of Prior Knowledge. *Magn Reson Med*, 1997;129:35-43.
15. Vanhamme L, Sundin T, Van Hecke P, Van Huffel S. MR Spectroscopy quantitation: a review of time-domain methods. *NMR Biomed*, 2001;14:233-246.
16. Mierisova S, Ala-Korpela M. MR Spectroscopy quantitation: a review of frequency-domain methods. *NMR Biomed*, 2001;14:247-259.
17. Gadian G. *NMR and its applications to living systems*. Oxford Science Publications; 1995.
18. Pan J W, Hamm J R, Rothman D L, Shulman R G. Intracellular pH in human skeletal muscle by  $^1\text{H}$  NMR. *Proc Natl Acad Sci USA*, 1988;85:7836-7839.
19. Scharf LL. *Statistical Signal Processing*. Addison-Wesley; 1991. p 359-415.
20. Veen JWC, Beer R, Luyten PR, Ormond D. Accurate Quantification of in vivo  $^{31}\text{P}$  NMR Signals using the Variable Projection Method and Prior Knowledge. *Magn Reson Med* 1988;6:92-98.
21. Levitt MH. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Wiley; 2001.

22. Govindaraju V, Young K, Maudsley AA. Proton NMR chemical shifts and coupling constants for brain metabolites. *NMR Biomed* 2000;13:129-153.
23. Coleman TF, Li Y, An Interior Trust Region Approach for Nonlinear Minimization Subject to Bounds. *SIAM J Optimization* 1996;6:418-445.
24. Waters NJ, Holmes E, Williams A, Waterfield CJ, Farrant R.D, Nicholson JK. NMR and Pattern Recognition Studies on the Time-Related Metabolic Effects of  $\alpha$ -Naphthylisothiocyanate on Liver, Urine, and Plasma in the Rat: An Integrative Metabonomic Approach. *Chem Res Tox* 2001;14:1401-1412.
25. Couvreur C, Bresler Y. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM J Mat Anal Appl* 2000;21:797-808.
26. Reeves SJ. An efficient implementation of the backward greedy algorithm for sparse signal reconstruction. *IEEE Signal Process Lett* 1999;6:266-269.

Table 1: Amplitude of each metabolite expressed as a mean value determined from 10 runs per variance ( $\sigma^2$ ) of the frequency perturbation. The correct amplitude of each metabolite should be 1.00, with the exception of phosphorylcholine which was 8.00. Note that in order to achieve results in a feasible time the number of groups in both glutamate and glutamine was reduced from five to three by combining groups that were very close. The “var. ” column shows the variance of the amplitude estimation over the 10 runs.

Metabolite	$\sigma^2 = 1$ Hz		$\sigma^2 = 2$ Hz		$\sigma^2 = 3$ Hz		$\sigma^2 = 4$ Hz		$\sigma^2 = 5$ Hz	
	mean	var.	mean	var.	mean	var.	mean	var.	mean	var.
glutamate	1.00	0.00	1.00	0.00	1.01	0.00	1.04	0.02	1.02	0.06
glutamine	1.00	0.00	1.02	0.00	0.94	0.02	1.05	0.02	1.01	0.04
acetate	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
alanine	1.00	0.00	1.00	0.00	1.00	0.00	0.99	0.00	1.00	0.00
aspartate	1.00	0.00	0.99	0.00	1.01	0.01	1.05	0.02	1.02	0.01
choline	1.00	0.00	1.00	0.00	1.63	3.89	2.07	5.33	2.22	4.32
creatine	1.00	0.00	1.01	0.00	1.02	0.00	1.07	0.02	1.08	0.07
glycerophosphorylcholine	1.00	0.00	1.70	4.90	3.15	11.26	2.37	4.94	4.29	14.13
glycine	1.00	0.00	1.05	0.06	1.23	0.34	1.52	0.45	1.46	0.43
lactate	1.00	0.00	1.00	0.00	1.00	0.00	1.08	0.05	1.23	0.17
myo-inositol	1.00	0.00	1.02	0.00	1.01	0.01	1.06	0.01	0.97	0.01
NAA	1.00	0.00	1.07	0.05	1.00	0.00	1.23	0.15	1.07	0.02
phosphocreatine	1.00	0.00	1.04	0.02	1.02	0.00	1.16	0.31	1.15	0.08
phosphorylcholine	7.99	0.00	7.29	4.94	5.30	11.87	6.34	9.13	5.39	9.86
succinate	1.00	0.00	1.05	0.03	1.00	0.00	1.05	0.03	1.14	0.18
taurine	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.01	1.06	0.01
threonine	1.00	0.00	1.00	0.00	1.11	0.05	1.12	0.10	1.16	0.11
TMSP	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00

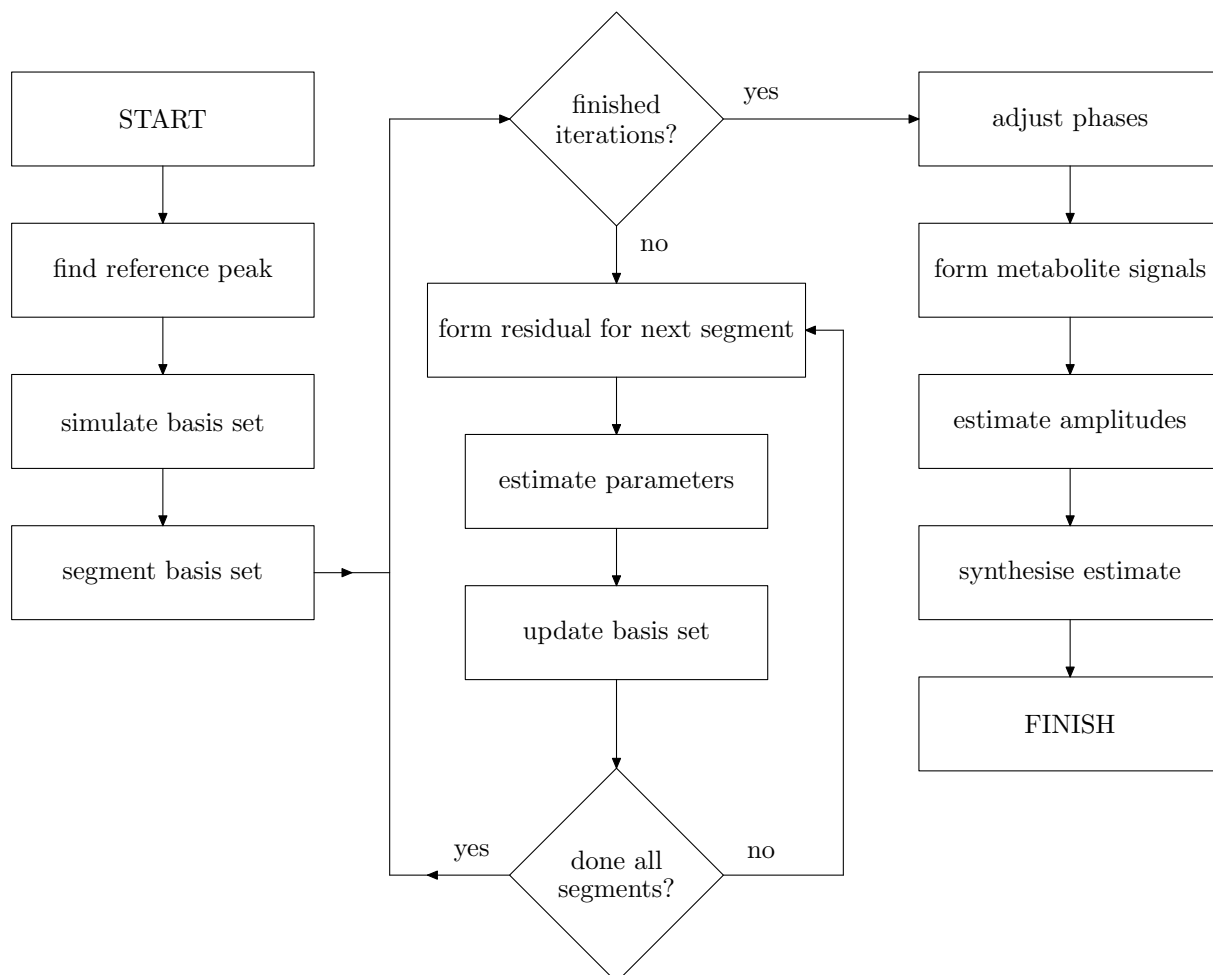


FIG. 1: The stages of the TARQUIN algorithm.

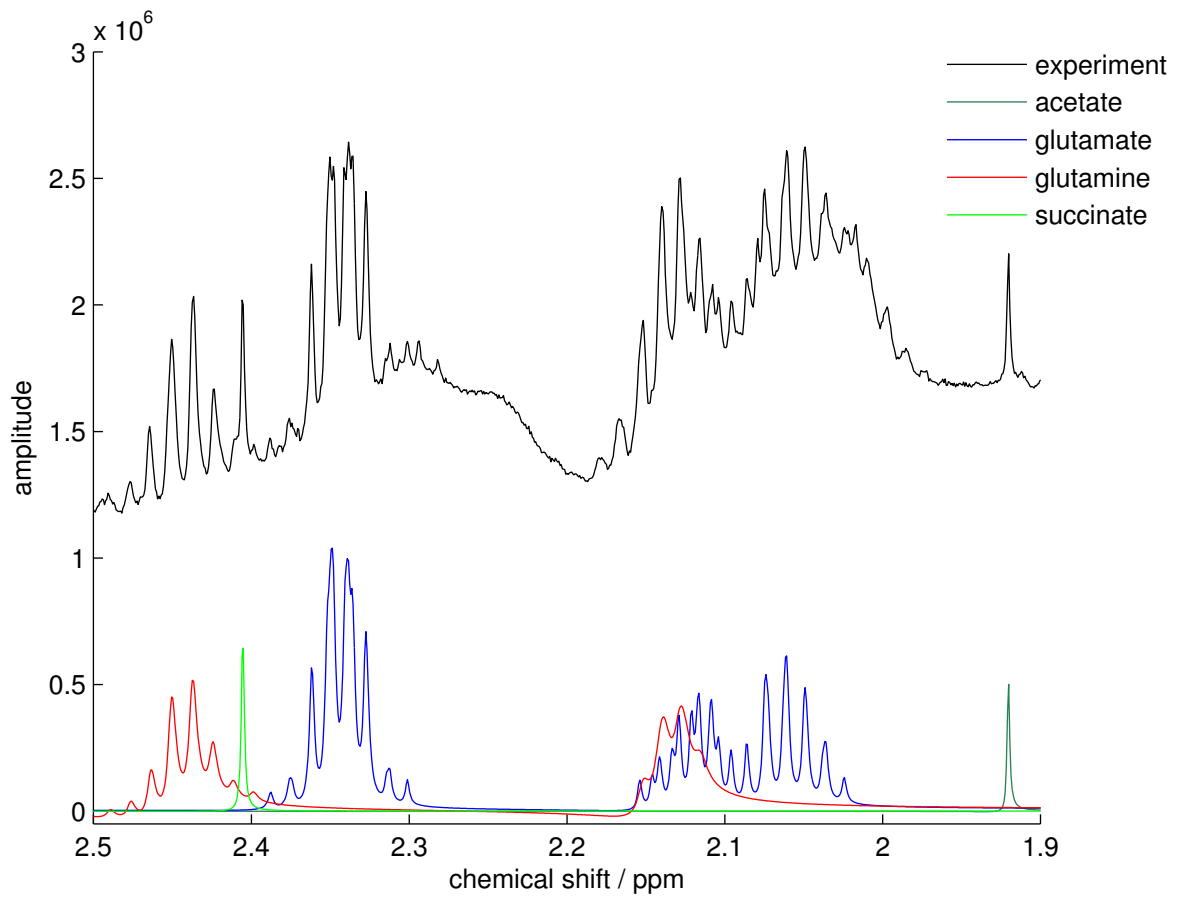


FIG. 2: The fit to a spectrum from a neuroblastoma cell line, in the region 2.5 ppm to 1.9 ppm. The glutamate and glutamine metabolites are particularly difficult to quantitate by hand due to the complexity of the multiplets. The height of the experimental spectrum's baseline is due to a combination of poor initial points and lipids and macro-molecules.

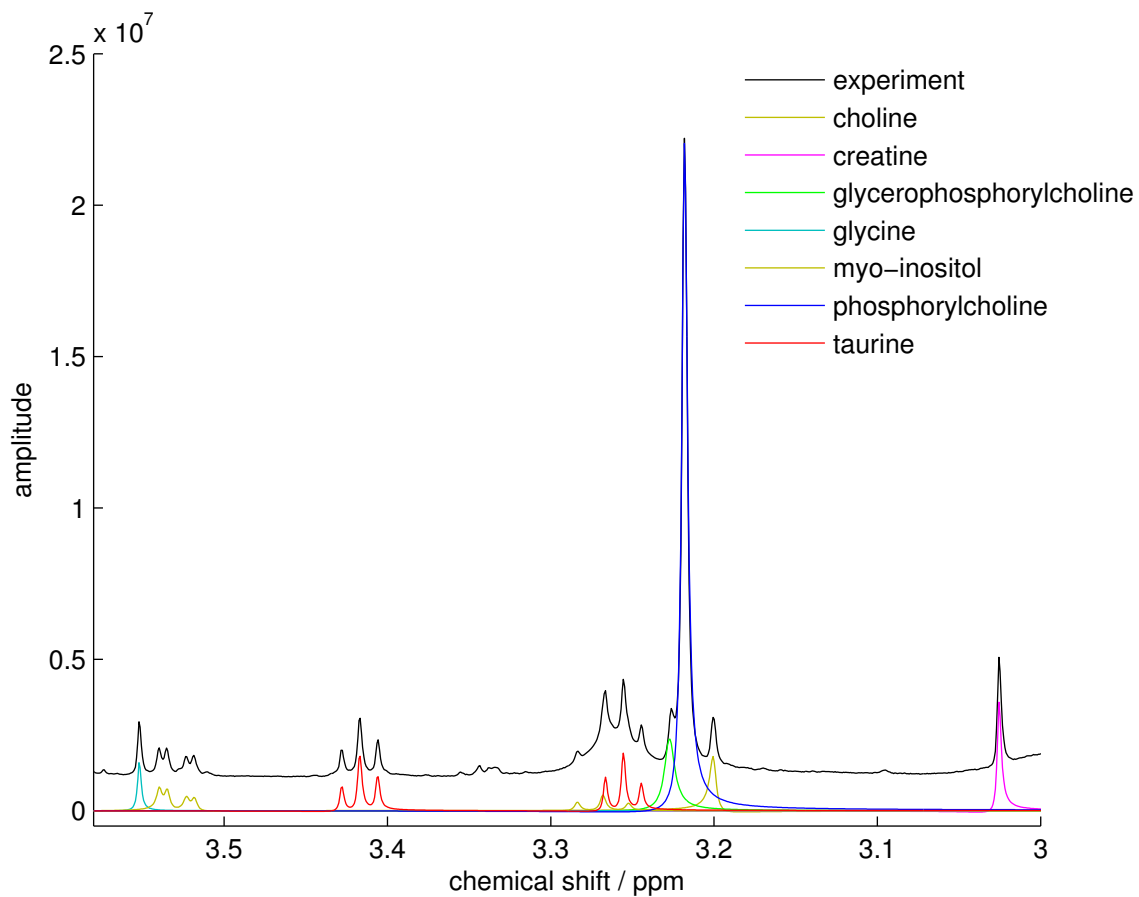


FIG. 3: The 3.6 ppm to 3.0 ppm region of the spectrum in Fig. 2. The three choline-based metabolite singlets in their correct position demonstrate the use of the smallest-shift criteria when generating the starting point  $\Delta\omega$  for the optimiser.

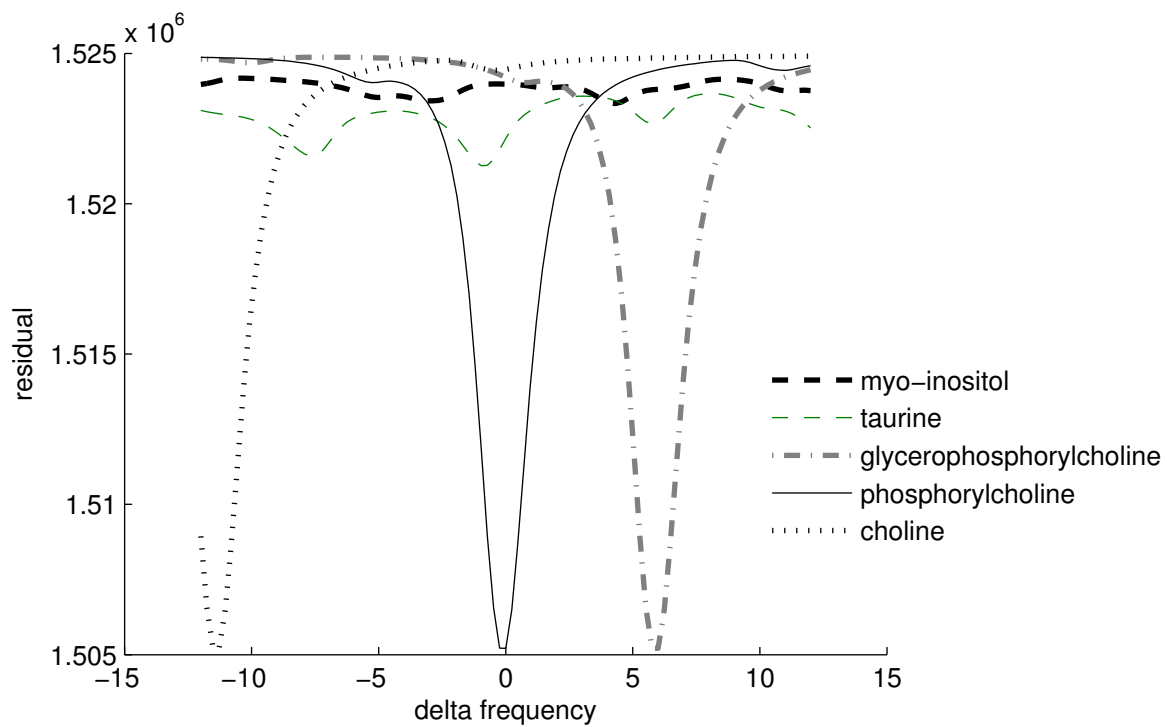


FIG. 4: The landscapes used to generate the values of  $\Delta\omega$  passed to the optimiser. Note the three larger minima in each of the cholines' landscapes corresponding to the location of each of the three singlets with the large phosphorylcholine peak.

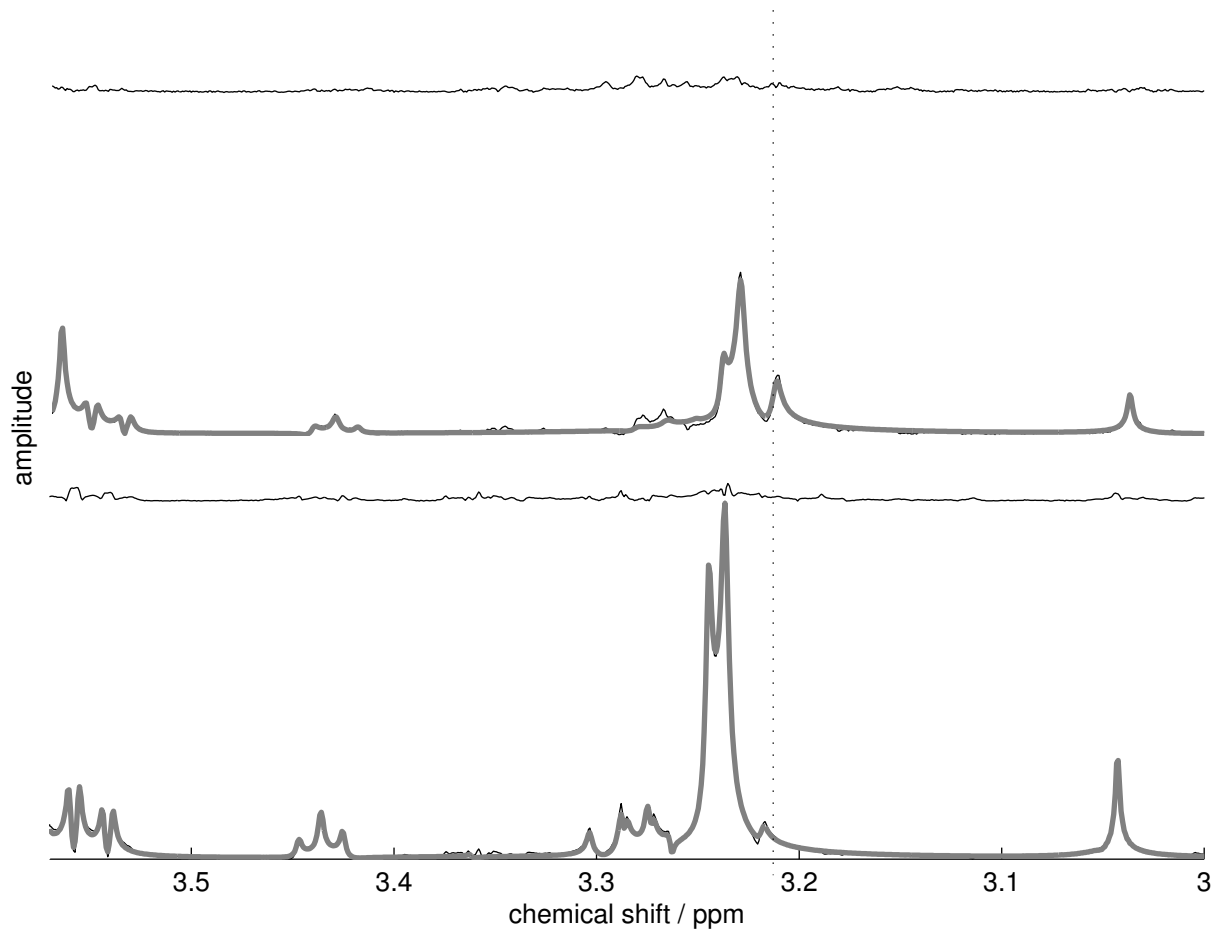


FIG. 5: Two fits of the same part of the spectrum for two different neuroblastoma cell line spectra demonstrating the shifting of groups. The dotted line indicates that same chemical shift in both spectra. The fit is shown in thick dark grey, the original signal in black and the residual is plotted above the signal and fit in both cases.