Deconvolution of Two-Dimensional NMR Spectra by Fast Maximum Likelihood Reconstruction (FMLR): Application to Quantitative Metabolomics

Roger A. Chylla,*,† Kaifeng Hu,† James J. Ellinger,‡ and John L. Markley†,‡

National Magnetic Resonance Facility at Madison and Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, Wisconsin 53706

*To whom correspondence should be addressed. Address: NMRFAM, Department of Biochemistry, University of Wisconsin Madison, 433 Babcock Drive, Madison, WI 54706-1544, USA. Tel:(608) 263-9396. E-mail: rchylla@wisc.edu
†National Magnetic Resonance Facility at Madison.
‡Department of Biochemistry.
Supplemental Method Sections

Preparation of Synthetic Metabolite Mixtures. Mixtures of synthetic compounds, including relevant metabolites were prepared from stock solutions at known concentrations. Alanine, methionine, 3-hydroxybutyrate and glucose were from Sigma. 2-(N-morpholino)ethanesulfonic acid (MES) was from Research Organics. DSS and 99.8% D₂O were from Cambridge Isotope Laboratories. All dry reagents were lyophilized overnight to remove moisture. Alanine, methionine, 3-hydroxybutyrate, glucose and MES were weighed directly into a grade A 50 mL Kimax® volumetric flask to give final concentrations of 32.5, 41.0, 60.0, 50.0, and 90 mM respectively. Each reagent was prepared from weighing to 4 significant figures on an analytical balance of ±0.1 mg precision. A stock solution of 33.3 mM DSS was prepared by initially placing the weighed quantity in a 10 mL volumetric flask and dissolving in D₂O containing 500 µM NaN₃. The DSS solution was then transferred to a 50 mL flask, and D₂O containing 500 µM NaN₃ was added to the mark. The pH of the synthetic mixture was not adjusted.

Preparation of Liver Extracts: Beef liver was frozen, cut into small portions and lyophilized until dry. After lyophilization, small morsels in the range of 500 to 600 mg were prepared by gently breaking the lyophilized liver with a metal spatula. Filters, pre-washed to remove glycerol, were stored at 4 °C until use with a small amount of water to keep the membrane moist. Each morsel was placed in a 50 mL conical tube on dry ice and was homogenized for 30 s using a rounded glass rod. To each sample, 16 mL 95°C MES buffer (250 µM) was added, and each sample was briefly agitated using a vortex mixer and then immersed in a 95°C water bath for 7.5 min. Samples were removed from the hot water bath, briefly agitated using a vortex mixer, and then centrifuged at 8000 RPM for 20 min (4°C) to collect cellular debris. Water was removed from the centrifugal filters, and the supernatant containing liver extract was decanted into the filter. Samples were centrifuged at 4100 RPM in a swinging-bucket rotor for 10.5 h until the solution above the filter was below the dead space volume of 200 µL. Filtered liver extracts were lyophilized and dissolved in 800 µL D₂O containing 1 mM 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) and 500 µM NaN₃ to inhibit microbial growth. The resulting solution was titrated with concentrated DCl or NaOD as needed to achieve an observed pH of 7.400 (±0.015).

Steps A-G: FMLR Algorithm

A. Calculation of Model and Residual Spectra: The model spectrum is calculated by applying the \( \hat{F}_1 \) and \( \hat{F}_2 \) operators to the time-domain model from the previous step (for the first iteration, the model contains no signals). The basis functions and related parameters used to construct the signals are summarized in Table 1 (main, section). A residual spectrum is calculated by subtraction of the data spectrum from the model spectrum.

B. Signal Generation by Peak Picking: A set of most probable signals are identified by peak picking of the residual spectrum at a defined threshold. The coordinates of the peak are used to generate initial frequency values for the new signals. Estimates of decay rates (line widths) are obtained from prototype values based on the experiment type and peak attributes.

C. Signal Generation: Signals are created for each of the peaks. A signal is constructed from assignment of initial values and constraints for each parameter.
D. Calculation of Basis Functions: Model time-domain functions are constructed from the current set of frequency and decay rate parameters for all signals (Table 1, main section). The corresponding frequency domain basis functions are calculated by applying the $\hat{F}$ operators to the time-domain basis functions along dimensions 1 and 2, respectively. During model optimization, the basis functions (and associated derivatives) for any signal are recalculated upon any relevant parameter value change.

E. Amplitude Calculation from Linear Least Squares Analysis: The amplitudes for a set of signals in a signal group with known basis functions are calculated from linear least squares analysis. The linear system solved is the set of “normal equations” where the matrix is the $m \times m$ “interaction matrix” of $m$ basis functions and the vector is the projection of the data against those basis functions. It is worth noting that determination of the least squares amplitudes by solution of the normal equations generates exactly one solution; iterative techniques are not thus required to derive amplitude estimates for any given fixed set of frequency and decay rate parameters.

F. Calculation of Cost Function: The cost function is calculated independently for each signal group from a scalar quantity $\nu^2$, dubbed the “sufficient statistic”. The sufficient statistic is a simplified form of the least squares statistic $\chi^2$ with the term representing the square of the data values removed.¹

G. Gradient Optimization of Signal Parameters: Step G is an optimization that invokes steps D-F within an inner loop to calculate $\nu^2$ for each signal group given a set of non-linear parameters i.e. the frequency and decay rate parameters along each dimension for a fixed set of signals. Using derivatives of this cost function with respect to each non-linear parameter, a simple gradient search algorithm is used to determine the most likely set of frequency and decay rate parameters for a fixed number of signals i.e. the set of parameter values for a signal group that minimizes $\nu^2$ in the local parameter space around the initial estimates of those parameters. This is the final step of the outer iteration loop as the optimized model resulting from this step is used as input to calculate the model and residual spectra in Step A.

As mentioned above, the Newton algorithm is designed to apply the approach of spectral deconvolution specifically to analysis of related multidimensional NMR data sets. The sections below describe aspects of FMLR that were implemented by Newton to address issues pertaining to high-throughput NMR analysis of related data sets, more specifically in the context of metabolite quantification of natural abundance 2D $^1$H-$^{13}$C HSQC spectra. In general, these features of Newton distinguish it from the Chifit application reported earlier.¹

Cluster and ROI Analysis. Quantitative metabolomics is an “identify” and “quantify” approach. The FMLR algorithm breaks the peak identification phase into two sequential steps (B1-B2) and an optional third step (B3):

B1. Peak Identification: Identify peaks (transitions corresponding to new signals) in the residual spectrum of the data ensemble (see Figure 2, main section).

B2. Cluster and Species Assignment: Assemble clusters of peaks and identify species in each data set of the data ensemble. A “species” as used in this context refers to the recognition of corresponding transitions in each data set. Each peak is assigned a species ID after this step.
B3. **ROI Assignment (Optional)**: If an assignment table is provided that maps an atomic species (e.g. methyl proton of alinine) to a spectral region or region center, this information can be used to assign a set of peaks to a region of interest (ROI). This step is not required for spectral deconvolution but may be useful for assigning parameter constraints and further analysis. When this action is performed, each peak is assigned an ROI ID (region of interest ID). For example, the set of peak transitions associated with a methylene proton will be assigned the same ROI ID (as in Figure 2 of main paper).

The cluster analysis in step B2 is robust with respect to peak “drift” between data sets in the data ensemble. The cluster analysis algorithm uses a grid correlation metric $\nabla$. The grid correlation metric is simply the dot product of two vectors $v_i(\delta)$ and $v_{i+1}(\delta + \epsilon)$ (each of which represent the set of values in a multidimensional grid).

$$\nabla = v_i(\delta)v_{i+1}(\delta + \epsilon)$$

The grid correlation metric $\nabla$ captures the likelihood that a given frequency shift in each dimension of a grid (an ordered pair for $n = 2$ dimensions) maximizes the correlation between two regions of the grid. The metric is used to determine the shift required along each dimension $d$ in data set $i$ to maximize its correlation to an adjacent data set $(i + 1)$. Stated differently, the metric is proportional to the relative probability that a cluster of resonances with center $\delta$ in one data set is located in an adjacent data set at $(\delta + \epsilon)$.

**Reference Spectrum**: In many series of NMR experiments, the concept of a “reference spectrum” can be exploited. A reference spectrum is one which is known to contain all resonances at a detectable level. When configured with a reference spectrum, Newton performs signal recognition (peak picking) on the reference spectrum only. When deconvolution is complete, the signals from the reference spectrum are propagated to other spectra for global optimization.

Depending upon the configuration of the experiment, the propagated signals may share the same basis function parameters or the parameters may be independent (e.g. peak shifts between spectra can thus be accounted for). Scalar amplitudes are always independent. The benefits of the use of a reference spectrum are:

- Amplitudes can be more easily estimated at lower concentrations because the frequency and linewidth characteristics are better defined in the related spectra containing the signal at higher concentrations.
- The incorporation of prior information leads to fewer degrees of freedom in the model and produces more consistent results (Bayesian approach).
- The resulting calculations are more efficient than performing independent signal deconvolution on all spectra.

**Parameter Constraints.** The gradient search algorithm described in Step G is a simple algorithm that uses a parameter gradient of the cost function to find the “local minimum” of the cost function. Because NMR signals are in general partially resolved, this search method is stable and converges with a reasonable number of iterations. For extremely overlapped regions, however, we found it beneficial to place bounds on the parameters, particularly the linewidths of the signals, to keep the search from finding minima that are irrelevant (e.g. the algorithm attempting to fit a region of very overlapped signals with one broad signal with a linewidth 10-times larger than any expected linewidth). A constraint we found to be robust is a geometric one.
which constrains any linewidth parameter to be within a factor \( f = 4 \) of the “expectation” linewidth. For example, a model of an experiment with an expectation linewidth of \( \alpha = 10 \) Hz for the proton dimension would have the linewidth for each signal constrained to the range of \( \frac{\alpha}{f} \leq \alpha \leq f\alpha \) Hz or 2.5-40 Hz for \( f = 4 \). This constraint is only active for portions of the deconvolution when a signal-dense region is being modeled or along dimensions of heavily truncated data where the decay rates are poorly defined by the data.

**Overlap Statistic.** For data sets with a high dynamic range in signal intensities, particularly in signal dense regions with a high degree of overlap, we formulated an overlap statistic that was used to filter peaks. The overlap statistic is simply the ratio of the peak height in the residual spectrum to its corresponding height in the original data spectrum. Peaks with an overlap statistic greater than a threshold (default value of 0.10) are assumed to “pass” the overlap filter test. A peak with a value less than this threshold is ignored and no signal is added to the model in subsequent steps. The overlap filter prevents modeling of statistically insignificant signals in signal dense regions.

**“Shared” Parameters.** The Newton software is designed to incorporate the concept of “shared” parameters, i.e., a parameter appearing more than once in a model. These relationships are modeled through implementation of an abstraction referred to here as a “parameter context”. A parameter context defines a tuple consisting of a parameter type, dimension, and signal ID. A set of parameter contexts is mapped to a parameter ID such that the set of parameter contexts defines all appearances of a given parameter in the model. The Newton program calculates the derivative of the cost function with respect to any parameter context given the current values of all parameters. The derivative of the cost function with respect to any corresponding parameter therefore becomes the simple sum of derivatives of all associated parameter contexts. The parameter context abstraction supports the optimization of parameter values such that they represent global maximum likelihood values across multiple signals and/or multiple data sets.

**Treatment of \( t_1 \) noise ridges:** Noise filters can be added to prevent the selection of peaks in the residual that lie along \( t_1 \) ridges. These filters specify a spectral range upon which to calculate a separate noise threshold. Picks from the residual that exist within this spectral range are subjected to an additional noise threshold. The effect of \( t_1 \) noise on quantification of peaks is minimized by the multidimensional nature of FMLR. The summation of projections in both dimensions works to cancel out intensities that appear as peaks in only one of the dimensions.

**Termination Conditions for Signal Recognition.** The outer iteration loop (A–G) of Figure 1 (main body) requires a termination condition. A number of different statistics are possible for deciding when to stop adding signals to the model. For the studies in this paper, we used a simple threshold test in locating peaks in the residual. The threshold is either scaled to the noise estimate of the data (e.g. 10 \( \times \) the noise variance) or assigned an absolute value based on visual inspection of the data. In either case, the deconvolution is terminated when the residual no longer contains any peaks that satisfy the threshold test. A number of more statistically rigorous techniques can potentially be used to determine how many signals and thus how many degrees of freedom are appropriate for a given increase in maximum likelihood but unfortunately, the abundant present of artifacts in NMR experiments renders such techniques ineffective. NMR experiments consist of three entities: desirable signals, undesirable signals (artifacts such as \( t_1 \) ridges), and noise.
General frameworks for distinguishing signals from noise do not take into account the abundant presence of systematic but undesirable signals in NMR experiments.

References

(1) Chylla, R. A.; Markley, J. L. *J. Biomol. NMR* 1995, 5, 245-258.
(2) Lewis, I. A.; Schommer, S. C.; Hodis, B.; Robb, K. A.; Tonelli, M.; Westler, W. M.; Sussman, M. R.; Markley, J. L. *Anal. Chem.* 2007, 79, 9385-9390.
(3) Hu, K.; Westler, W. M.; Markley, J. L. *J. Am. Chem. Soc.* 2011, 133, 1662-1665.
(4) Markley, J. L.; Anderson, M. E.; Cui, Q.; Eghbalnia, H. R.; Lewis, I. A.; Hegeman, A. D.; Li, J.; Schulte, C. F.; Sussman, M. R.; Westler, W. M.; Ulrich, E. L.; Zolnai, Z. *Pac. Symp. Biocomputing* 2007, 157-169.
(5) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* 2007, 35, D521-D526.
Table 1. NMR Experimental Parameters and References

| NMR experiment                              | Spectrometer | Field strength (MHz) | Key acquisition parameters                                                                 | References |
|---------------------------------------------|--------------|----------------------|-------------------------------------------------------------------------------------------|------------|
| Sensitivity enhanced 2D $^1$H-$^1^3$C HSQC  | Varian       | 600                  | • Probe: $^1$H-$^1^3$C-$^1^5$N, $^2$H lock                                                | 2          |
|                                             |              |                      | • $^1^3$C sweep width: 70 ppm                                                             |            |
|                                             |              |                      | • Acquisition Time: 300 ms.                                                                |            |
|                                             |              |                      | • Number of scans: 4 (32 silent)                                                           |            |
|                                             |              |                      | • Collection Time: 12 min.                                                                 |            |
| 2D Extrapolated time-zero $^1$H-$^1^3$C HSQC| Bruker       | 700                  | • Probe: QCI                                                                              | 3          |
|                                             |              |                      | • $^1^3$C sweep width: 80 ppm                                                             |            |
|                                             |              |                      | • Acquisition Time: 183 ms.                                                                |            |
|                                             |              |                      | • Number of scans: 16 (DS 16)                                                              |            |
|                                             |              |                      | • Collection time: 90 min × 3.                                                             |            |
### Table 2. Web Locations of Newton Runtime Environment, Documentation, and Resources

| URL                                      | Description                                                                 |
|------------------------------------------|-----------------------------------------------------------------------------|
| [http://newton.nmrfam.wisc.edu](http://newton.nmrfam.wisc.edu) | Homepage for Newton on NMRFAM web site. Links to all resources are published on this page. |
| [http://sourceforge.net/projects/newton-nmr/develop](http://sourceforge.net/projects/newton-nmr/develop) | Developer web page for Newton located on the open source portal SourceForge. This web page contains information on how to download the source code using a Subversion client. |
| [http://www.oracle.com/technetwork/java/javase/downloads/index.html](http://www.oracle.com/technetwork/java/javase/downloads/index.html) | Download site for Java Runtime Environment (JRE) for supported platforms. A JRE v 1.6+ is required to run the Newton program. The Newton binaries files can, in theory, be run on any platform with an installed JRE 1.6+. The program has been tested on three platforms: Windows, Linux (Ubuntu), and Macintosh. |
Table 3. Metabolites in Synthetic Mixture I

| Metabolite                        | Sym. | Conc.  | ROI assignment$^a$ |
|-----------------------------------|------|--------|--------------------|
| Alanine                           | Ala  | 66.4 mM|                    |
| 3-Hydroxybutyrate                 | HB   | 119 mM |                    |
| 4,4-dimethyl-4-silapentane-1-sulfonic acid | DSS  | *10.7 mM|                    |
| 2-(N-morpholino)ethanesulfonic acid | MES  | 179 mM |                    |
| Methionine                        | Met  | 82.2 mM|                    |
| Glucose                           | Glu  | 101 mM |                    |
The Rx atom assignments denote the corresponding regions of interest in the time zero $^1$H-$^{13}$C HSQC spectra (see Figure 3). The concentrations were in the high millimolar range to test the accuracy of the quantification. The accuracy reflects the limits of resolution and the analysis method rather than any error introduced by signal to noise or limits of detection. DSS was used as a chemical shift (but not concentration) standard. NMR assignments were extracted manually from the BioMagResBank$^4$ and Human Metabolome Database (HMDB)$^5$.

*Denotes a caveat that the gravimetric determination of DSS concentration is subject to greater error by its hygroscopic and undesirable physical crystal properties (non-granular) which complicates efforts to reliably measure its mass.
Figure 1. Log-linear regression of amplitudes. Plots of log $A$ vs. $t_i$ for the ROIs in (A) Figure 2A and (B) Figure 2B in the main text. The term $A$ is the ROI amplitude (sum of signal amplitudes belonging to the ROI) and $t_i$ is the number of repetition times in the HSQC series. The semi-log linear regression is used to calculate the unattenuated amplitude at $t = 0$ that is linearly proportional to the concentration of the species giving rise to the ROI. All of the correlations shown in the figure have a regression coefficient $\geq 0.9999$. The lowest correlation coefficient for the non-displayed ROIs was 0.9996.
Figure 2. Reference $^1$H-$^{13}$C HSQC spectrum for FMQ series with identified regions of interest. (A) Contour plot of the 10 mM reference spectrum from the $^1$H-$^{13}$C FMQ series described in the text. (B) Contour plot of the outlined region with overlaid boxes that represent ROIs identified by Newton across all data sets. The ROIs were constructed from the results of spectral convolution and an input assignment table of ROIs from known metabolites. The resonances for the input assignment table were assembled manually using data from BioMagResBank (BMRB)$^4$ and the Human Metabolite Data Base (HMDB)$^5$. The footprint of a model signal is determined by the point moving away from the center at which the signal profile dips below a noise threshold. The boundary of the encapsulated ROI is the union of all such boundaries of constituent signals. Accordingly, the boundaries of the ROI shown in the graphic are sometimes not centered or sometimes have large margins around the visible resonances in the reference spectrum. These characteristics are the result of the corresponding peaks of the ROI in the related spectra (not shown) being shifted with respect to the reference spectrum.