2D NMR metabonomic analysis: a novel method for automated peak alignment

Ming Zheng¹, Peng Lu¹, Yanzhou Liu², Joseph Pease², Jonathan Usuka¹, Guochun Liao¹,* and Gary Peltz¹

¹Department of Genetics & Genomics and ²NMR Spectroscopy, Roche Palo Alto LLC, Palo Alto, CA 94304, USA

Received on May 17, 2007; revised on August 15, 2007; accepted on August 16, 2007
Advance Access publication September 10, 2007
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Comparative metabolic profiling by nuclear magnetic resonance (NMR) is showing increasing promise for identifying inter-individual differences to drug response. Two dimensional (2D) ¹H—¹³C NMR can reduce spectral overlap, a common problem of 1D ¹H NMR. However, the peak alignment tools for 1D NMR spectra are not well suited for 2D NMR. An automated and statistically robust method for aligning 2D NMR peaks is required to enable comparative metabonomic analysis using 2D NMR.

Results: A novel statistical method was developed to align NMR peaks that represent the same chemical groups across multiple 2D NMR spectra. The degree of local pattern match among peaks in different spectra is assessed using a similarity measure, and a heuristic algorithm maximizes the similarity measure for peaks across the whole spectrum. This peak alignment method was used to align peaks in 2D NMR spectra of endogenous metabolites in liver extracts obtained from four inbred mouse strains in the study of acetaminophen-induced liver toxicity. This automated alignment method was validated by manual examination of the top 50 peaks as ranked by signal intensity. Manual inspection of 1872 peaks in 39 different spectra demonstrated that the automated algorithm correctly aligned 1810 (96.7%) peaks.

Availability: Algorithm is available upon request.

Contact: guochun.liao@roche.com

1 INTRODUCTION

Nuclear magnetic resonance spectroscopy (NMR) is a well-established method for the analysis of complex biological samples. Profiling of endogenous metabolites (metabonomics) by NMR is showing increasing promise for predicting an individual’s response to drugs. For example, comparative metabolic profiling by NMR was used to predict the response to doxorubicin and interleukin-2 treatment (Ewens et al., 2006). Many metabonomic studies to date utilize one dimensional (1D) ¹H NMR analysis because the data is relatively quick and easy to obtain and, analysis is straightforward. However, 1D ¹H NMR spectra of complex biological samples typically have high spectral overlap, which significantly limits the number of metabolites that can be uniquely identified and quantified. Recently, 2D ¹H—¹³C HSQC NMR was used to analyze global metabolic changes in yeast (Lu et al., 2007). Because almost all endogenous metabolites contain carbon, the addition of the second (¹³C) NMR dimension improves the resolution and enables the accurate identification of a large number of metabolites that are not resolvable in a standard 1D ¹H NMR spectrum. An example of the improved resolution is shown in Figure 1.

However, metabonomic analyses require comparison of metabolic profiles obtained from multiple replicates of samples exposed to different experimental conditions. This requires aligning NMR cross-peaks that represent the same chemical groups across multiple NMR spectra. The complexity of 2D ¹H—¹³C HSQC NMR spectra presents a number of challenges for comparative spectral analysis. First, the position of a cross-peak representing the same chemical group across multiple spectra is not fixed on 2D ¹H—¹³C NMR spectra. There is always a slight shift in the position of a cross-peak because the experimental conditions cannot be replicated exactly. Subtle differences in pH, temperature or ionic strength all can cause minor shifts in cross-peak location. Second, shifts in peak position are not systematic. The direction and extent of the shift for each peak is not consistent throughout a spectrum, and occur to a different extent within different areas of a spectrum. Therefore, a global correction for an individual spectrum is not helpful. Finally, not all metabolites are present in all samples analyzed for a given metabonomics study. Therefore, the ability to distinguish a peak with an insignificant signal from a neighboring peak representing a different metabolite is critical for the alignment of 2D NMR spectra.

A number of methods have been used for ¹H 1D NMR spectrum alignment, including spectral binning (Anthony et al., 1994), kernel smoothing (Smith et al., 2006) and targeted profiling (Weljie et al., 2006). In the spectral binning method, the entire 1D spectrum is divided into small bins of fixed or variable size, and then the total intensity within each bin is further analyzed (Anthony et al., 1994). Although spectral binning can be applied directly to 2D NMR data, determination of meaningful bin boundaries is very difficult due to heterogeneity within different regions of spectrum. The kernel...
smoothing method estimates the overall distribution of all peaks, and then overlapping intervals are used for peak matching (Smith et al., 2006). Recently, targeted profiling was proposed to identify and quantify unknown metabolic peaks within uncharacterized spectra by searching a database of pure compound spectra (Weljie et al., 2006). Because of the increased complexity and insufficient prior knowledge of individual NMR resonances within 2D NMR spectra, kernel smoothing and targeted profiling methods have limited applications for analysis of 2D NMR spectra. The problem of aligning 2D NMR spectra is actually similar to the ‘point match problem’ encountered in the field of computer vision. A robust point-matching method was developed to align features in two visual images, which resembles the peaks within 2D NMR spectra. This method uses a distance similarity measure (Chui and Rangarajan, 2000; Geiger and Girosi, 1991; Gold and Rangarajan, 1996; Rangarajan et al., 1997) to align the visual feature. However, this method cannot be applied directly to 2D NMR spectrum alignment. This method works well when there are less than 100 features in each image; but 2D NMR spectra usually contain 300-500 peaks. In addition, this method can only analyze two different images, and modifications are required to enable alignment of multiple images.

Here, we present a simple and effective statistical method for alignment of peaks within 2D NMR spectra. Since peaks representing different chemical groups often exhibit a similar
shift pattern within a small region of a spectrum, we designed a similarity measure that assesses the degree of local pattern match. Then, we developed a heuristic algorithm to maximize the similarity measure across the whole spectrum, which enables alignment of peaks representing the same chemical groups in different spectra. This method was applied to 1H–13C 2D NMR spectra that were produced as part of a research effort to identify genetic factors regulating susceptibility to acetaminophen-induced liver toxicity.

2 DATA AND METHODS

Overview of metabonomic analysis of acetaminophen-induced liver toxicity. Acetaminophen is a safe and effective drug when administered appropriately; however, an overdose can cause liver damage by inducing localized centrilobular cell death (Bessems and Vermeulen, 2001; James et al., 2003). Inbred mouse strains showed differential susceptibility to acetaminophen-induced liver damage. Identifying the factors that contribute to the resistance to acetaminophen-induced liver toxicity can lead to novel prevention or treatment strategies. The abundance of soluble metabolites in liver extracts from four inbred mouse strains was compared using 1H–13C 2D NMR analysis to identify metabolites that have a unique abundance pattern in the strain (SJL) that was resistant relative to the three strains (DBA/2, C57BL6 and SMJ) that were sensitive to the drug-induced liver toxicity.

Animal husbandry and drug treatment. All animal experiments were performed in the Laboratory Animal Facilities at Roche Palo Alto using protocols that were approved by Roche Institutional Animal Care and Use Committee. Seven to eight weeks old male mice of the following strains were obtained from Jackson Laboratory (Bar Harbor, ME, USA) and acclimatized for an additional week before use: SJL, DBA/2J, C57BL/6J and SM/J. The mice were housed under pathogen-free environment and provided food and water ad libitum with a 12h:12h light: dark cycle until experimental use. Before each experiment, food was withheld from the animals overnight (>16 h) to uniformly deplete hepatic glutathione stores (Bartolome et al., 1987). All mice were administered a single 300 mg/kg intraperitoneal dose of freshly prepared acetaminophen (Voigt Global Distribution INC, www.VGDLLC.com) suspended in PBS (pH 7.4), and were allowed free access to food and water after treatment. The mice were euthanized by CO2 inhalation to collect blood samples and liver tissues at 0, 3 and 6 h after dosing.

Metabolite extraction and NMR data acquisition. Frozen liver tissues (500 mg) were pulverized with liquid nitrogen and immediately plunged into 15 ml solution of 67% MeOH/33% H2O at −20 °C. Tissues were lysed by freezing and thawing three times, thoroughly mixed, and then centrifuged at 12000 g for 30 min at 4 °C. The supernatants were dried by speed-vacuum and re-suspended in 500 μl D2O (Cambridge Isotope Laboratories, Inc.). Samples were then centrifugally filtered through 10-kDa cutoff filters (Microcon YM-10, Millipore) to remove precipitated proteins. The filtrate was lyophilized and dissolved in 200 μl D2O containing 1 mM sodium-3-(tri-methylsilyl)-2,2,3,3-tetramethylurea (TSP; Sigma-Aldrich), which was an internal standard for NMR analysis.

NMR spectra were recorded at 300 K on a Bruker Avance 600 MHz spectrometer operating at 1H frequency of 599.99 MHz and 13C frequency of 150.87 MHz using a 3-mm Nalorac microprobe with z-axis pulsed-field gradient. Resonances were assigned using 1D proton spectra, 2D proton correlated spectroscopy COSY, and 2D 1H–13C single-bond correlated HSQC spectra. Peaks used in the quantitative metabolite analysis were picked from 2D 1H–13C HSQC spectra acquired using z-axis pulsed field gradients for coherence selection. Spectra were acquired using 16–64 scans per FID, 1024 and 256 points for the 1H and 13C dimensions, and total spectrum acquisition times of 2.75–11 h. NMR assignments were confirmed by acquiring the spectra of samples spiked with purified compounds. 2D-NMR spectra were processed with Bruker Topspin software and interpreted using the Sparky Assignment and Integration Software package (http://www.cgl.ucsf.edu/home/sparky).

Description of peak alignment method. In order to align peaks within 2D NMR spectra, we designed an object function to access the similarity between peaks from different spectra, and then used a greedy algorithm to maximize the objective function. First, let us define that a peak is a local maximum within a predefined range (e.g. 0.03 p.p.m. in 1H dimension and 0.35 p.p.m. in 13C) in a spectrum. An alignment proposal is a proposal to assign correspondence to every possible peak pair for a set of spectra: if a pair of peaks is proposed to represent the same chemical group, this pair of peaks is assigned ‘corresponding’. An alignment proposal follows two rules:

(1) ’Corresponding’ assignment is transitive: if peak 1 and peak 2 are corresponding and peak 1 and peak 3 are corresponding, then peak 2 and peak 3 are corresponding.

(2) If a pair of peaks consists of two different peaks from the same spectrum, then they are assigned non-corresponding, because within a single spectrum different peaks are considered to represent different chemical groups.

The rules (1) and (2) imply that any peak in a spectrum cannot be assigned ‘corresponding’ to two peaks in another spectrum. It is clear that a peak is always corresponding to itself and if peak 1 is corresponding to peak 2, then peak 2 is corresponding to peak 1. Therefore, mathematically this corresponding assignment can be viewed as an equivalence assignment.

Let us define a peak group as the set of all peaks assigned to the same chemical group by an alignment proposal. For example, for the four spectra shown in Figure 3A, if an alignment proposal assigns peak #1 of each spectrum to a same chemical group, then these four peaks, each from a different spectrum, make one peak group. A peak group satisfies following:

(1) If two peaks are assigned ‘corresponding’, they are in the same peak group.

(2) Any two peaks in the same peak group are assigned ‘corresponding’.

2.1 Objective function and similarity measure

As mentioned in the Introduction section, shifts in peak position in 2D NMR spectra are not systematic. However, our observation indicates that within a small area of 2D NMR spectrum, shifts in peak position are similar: the peaks within a small area usually shift towards similar directions with similar extent. As a result, the relative positions of peaks within a small area are not changed by peak shifting. Therefore, the local spatial pattern of peaks across different spectra are usually matched. We proposed a similarity measure to measure how well the local patterns are matched and a new algorithm to maximize the similarity measure. By doing these, we are able to assign ‘corresponding’ to peaks across multiple spectra by local pattern match.

We designed an objective function for each alignment proposal as

\[ f(\text{alignment proposal}) = \sum_{\text{all peak groups in the proposal}} f(\text{peak group}), \]

where

\[ f(\text{peak group}) = \sum_{\text{all pairs } \sim_{i,j} \text{ in the group}} \text{sim}(\tilde{p}_i, \tilde{p}_j). \]
This objective function is essentially the total sum of similarity measure between the peak pairs for all peak groups. Here, \( \tilde{\beta}_i \) and \( \tilde{\beta}_j \) are the \( i \)th peak and the \( j \)th peak, respectively, \( \text{sim}(\tilde{\beta}_i, \tilde{\beta}_j) \) is the similarity measure between \( \tilde{\beta}_i \) and \( \tilde{\beta}_j \), and \( \tilde{\beta}_i = (C_i, H_i) \), where \( C_i \) and \( H_i \) are the coordinates of the \( i \)th peak in the carbon and proton dimension, respectively; and \( f(\text{peakgroup}) \) is the total similarity measure for a peak group.

There are three components in the proposed similarity measure. The first component is the weighted Euclidean distance between peaks, which is used to access local pattern match. Because the scales of the carbon and proton dimensions are different, the two coordinates of peaks need to be rescaled to balance their contribution to the similarity measure. We found that the peaks corresponding to the same chemical group in different spectra often had similar 2D shapes. Therefore, for each peak we defined a neighborhood region and used the neighborhood correlation as the second component to measure the shape similarity between a pair of peaks. It is common that replicate spectra are measured in many studies, so we considered the similarity of signal intensity between replicates as well and used it as the third component. Therefore, we designed the following similarity measure for alignment:

\[
\text{sim}(\tilde{\beta}_i, \tilde{\beta}_j) = \begin{cases} 
-\infty, & \text{if } \tilde{\beta}_i \text{ and } \tilde{\beta}_j \text{ is far away} \\
\frac{1}{2} \sum_{s_1} \sum_{s_2} s_1^2 + s_2^2, & \text{otherwise}
\end{cases}
\]

where \( s_1 = -(\text{dist}(\tilde{\beta}_i, \tilde{\beta}_j))^2 \), \( s_2 = \gamma \cdot (\text{correlation} - 1) \), \( \gamma = \frac{1}{\sqrt{\text{SD of log-intensity of all peaks}}} \), \( \text{SD} \) is the standard deviation of the log-intensity of all peaks in a spectrum, \( \text{correlation} \) is the weighted Pearson’s correlation between the two peaks, \( \gamma \) is the weight for the correlation term and \( \eta \) is the weight for the log-intensity difference term. Although peaks in a 2D NMR spectrum can shift, the extent of shift is always limited. Furthermore, we defined upper bounds for the coordinate differences between peak pairs in the same peak group. When the distance between a pair of peaks exceeds the upper bound, this pair of peaks is considered as ‘far away’ and the alignment proposal will not be accepted by assigning a negative infinity score to the similarity measure. The upper bounds for carbon and proton dimensions were estimated by manually aligning two reference peak groups representing two known metabolites that are next to each other in 2D NMR spectrum. We assumed that the coordinates of carbon dimension for the peaks in one peak group were independently drawn from a uniform distribution \( U[-B_c, B_c] \), where \( B_c \) is the maximum range that a peak can shift in the carbon dimension, which can be estimated unbiasedly from the reference data. The upper bound \( B_c \) for the carbon dimension is then defined as \( 2R_c \). The upper bound \( B_{pi} \) for the proton dimension can be calculated in the similar way.

The normalization constant \( \lambda \) in Equation (3) is estimated as \( \lambda = \frac{B_{pi}}{B_{pj}} \).

\( \gamma \) is determined by the user to balance the position difference and the shape difference. Empirically, we suggest to choose a \( \gamma \) to make the scales of the correlation term and the square distance term to be the same. From the spectra we want to align, we can randomly pick a large number (e.g. 10 000) of peak pairs such that the two peaks in the pair are from different spectra and that their position differences in both carbon dimension and proton dimension are within the permitted ranges, which are \( B_c \) and \( B_{pi} \), respectively. Then, we use the ratio of the mean of the weighted square distances to the mean of one minus the correlations as a reasonable estimation for \( \gamma \), so that the estimation of \( \gamma \) is always positive. \( \eta \) is defined as the ratio of the SD of the position shifts in carbon dimension and the SD of the log-intensity of all peaks when we align replicated spectra, \( \eta \) is zero otherwise. The position shifts in carbon dimension is assumed to follow a uniform distribution \([-B_c, B_c]\), therefore the SD is \( B_c / \sqrt{12} \).

Peak with insignificant signal or missing peak was handled by adding an imaginary ‘missing peak’ to a peak group. The similarity measure between a real peak and a missing peak is defined as:

\[
\text{sim}(\tilde{\beta}_i, \text{missing peak}) = \text{sim}(\text{missing peak}, \text{missing peak}) = \alpha.
\]

Here, \( \alpha \) is the penalty term for missing peaks, and \( \tilde{\beta}_i \) is any real peak in the same peak group. Therefore, there should always be \( \text{S}(S - 1)/2 \) terms in the summation, where \( S \) is the number of spectra in the summation.

Finally, in order to compare objective function \( f(\text{alignment proposal}) \) between proposals with different number of peak groups, we added imaginary peak groups composed of all missing peaks to the proposal with fewer peak groups to make the total number of peak groups in each proposal equal.

### 2.2 Maximizing the objective function

A greedy algorithm was used to maximize the objective function. This greedy algorithm has two main steps: generating initial alignment proposal and refining the alignment to maximize the objective function (Fig. 2).

Generating the initial alignment proposal is done by creating peak groups from unprocessed peaks in a greedy manner iteratively. As the starting point, all peaks in all spectra are unprocessed. First a peak, denoted as \( \tilde{\beta}_i \), is randomly selected from a spectrum to start a peak group containing this peak only. The best possible members for this peak group are then identified from the remaining spectra. Suppose that currently \( n - 1 \) spectra have been processed, therefore, the peak group has \( n - 1 \) members, either real peaks or ‘missing’ peaks: \( \tilde{\beta}_i, \tilde{\beta}_j, \tilde{\beta}_k \). Then, the unprocessed peak in the \( n \)th spectrum that is most similar to \( \tilde{\beta}_i \) is identified and let us denote it as \( \tilde{\beta}_n \) (for any peak \( \tilde{\beta}_i \) in the \( n \)th spectrum, \( \text{sim}(\tilde{\beta}_i, \tilde{\beta}_n) = \sum_{i=1}^{n-1} \text{sim}(\tilde{\beta}_i, \tilde{\beta}_n) \)).

If \( \tilde{\beta}_n \) and \( \tilde{\beta}_i \) are compatible, i.e. \( |C_n - C_i| < B_c \) and \( |H_n - H_i| < B_{pi} \) for \( i = 1, 2, \ldots, n - 1 \), where \( C_i \) and \( H_i \) are the coordinates of \( \tilde{\beta}_i \) in the carbon and proton dimension, respectively, then \( \tilde{\beta}_n \) is added to the peak group and marked as processed (\( \tilde{\beta}_n \) is automatically marked upon selection), or a ‘missing’ peak is added to the peak group for this spectrum and \( \tilde{\beta}_n \) remains unprocessed. If \( \tilde{\beta}_n \) does not exist, i.e. all peaks in the \( n \)th spectrum have been processed, then a ‘missing peak’ is added to the current peak group. This is repeated until the current peak group contains exactly one peak from each spectrum. Next, another unprocessed peak is selected to start another peak group. This process is repeated until all peaks in all spectra have been processed. The selection of unprocessed peaks can be in a stochastic order or a deterministic order. The application shown in this article used a deterministic order for peak selection: spectra were randomly ordered; peaks within each spectrum were ordered first by their \( ^1\text{C} \) coordinate and then by their \( ^1\text{H} \) coordinate and the selection of unprocessed peaks was based on this order until all peaks were processed.

Refining alignment proposal is achieved by re-assigning peaks to peak groups to maximize the total similarity score. First, three peak groups are selected. Then, a spectrum is selected; each selected peak group has one peak in the selected spectrum, either a real peak or a ‘missing’ peak. Then, six combinations of assigning three peaks to three peak groups are examined and the one with the highest similarity score is kept in the alignment proposal (Fig. 2). All spectra are evaluated iteratively until no further improvement can be achieved for that triplet. This process is repeated for all possible triplets of peak groups until there is no further increase in the objective function.
2.3 Computational complexity

In generating initial alignment proposal, to construct a peak group, \( S \) peaks need to be added to the group, where \( S \) is the number of spectra. Adding a peak of a spectrum to the current peak group takes \( O(s \cdot g) \), where \( s \) is the current number of peaks in the group and \( g \) is the number of peaks in the spectrum that remains unprocessed. \( s \) and \( g \) are always less than or equal to \( S \) and \( G \), respectively, where \( G \) is the number of peak groups in the final initial-alignment proposal. Therefore, generating a peak group takes \( O(S^2) \). Because there are \( G \) peak groups needed to be generated for the final initial-alignment proposal, the computational complexity for in generating initial alignment proposal step is \( O(S^2 \cdot G) \). In refining alignment proposal, there are \( O(G^3) \) possible triplet peak groups; for each triplet, there are \( O(S) \) calculations of similarity measure for each of the \( S \) target spectra. Therefore, the computational complexity for refinement step is \( O(S^2 \cdot G^3) \). Therefore, the total computational complexity for this method is controlled at \( O(S^2 \cdot G^3) \). This algorithm was implemented in C on a Redhat Linux platform with a 3 GHz CPU, and it took <1 min to process 30 spectra.

3 RESULTS AND DISCUSSION

We wanted to evaluate the abundance of endogenous soluble metabolites in liver extracts from mouse strains that are sensitive (SM/J, DBA/2, C57BL/6) or resistant (SJL) to acetaminophen-induced liver toxicity. To do this, liver extracts were prepared from these four strains at 0, 3 and 6h after treatment with a single dose of 300 mg/kg of acetaminophen; the metabolites were extracted and \(^1\)H–\(^13\)C 2D NMR spectra were prepared as described. We found that analysis of the 500 individual peaks in each of the 39 spectra (3 or 4 replicates from each of the 4 strains at 3 time points) was an overwhelming task that would require several weeks to complete, even when performed by an investigator (P.L.) that had several years of experience with analysis of NMR spectra. Therefore, we developed and evaluated an automated method for peak alignment in 2D NMR spectra.

Before applying the alignment algorithm, the spectra were normalized so that the total absolute intensity of each spectrum equaled 1. Two reference metabolites-glutathione and glutamine with well-known 2D NMR coordinates (33–35 p.p.m. of \(^13\)C dimension and 2.3–2.5 p.p.m. of \(^1\)H dimension) were manually aligned across all spectra. This enabled the upper and lower bounds for the coordinate differences between peak pairs within the same peak group to be determined. The shape similarity term (\( \gamma \)) was set to 0.5 empirically (the estimated value for \( \gamma \) to balance the contribution of position difference and shape difference was 0.36 from the data, but we decided to give slightly more weight to the shape and therefore selected 0.5); all other parameters were calculated from the data as described in the Methods section. The raw NMR data was processed to identify the metabolite peaks and then the alignment algorithm was applied.
Since three or four independent replicate experiments were performed for each of the 12 conditions analyzed (4 strains by 3 time points after dosing) to produce a total of 39 spectra, the spectral alignment was performed in two steps. First, the replicate spectra from each individual condition were aligned. Then, the average intensities and coordinates of the aligned peak groups were calculated to generate a representative spectrum for each strain at a particular time point. A peak group was included in the representative spectrum only if the corresponding peak appeared in more than three replicates, or if the corresponding peak appeared in two replicates with a normalized signal intensity that was above the user-defined signal threshold of 0.00001. The representative spectra for 12 conditions were then aligned to generate the final alignment with 514 peak groups. Because a gold standard for assessing the 2D NMR spectral alignment does not exist, the automated alignments were manually examined by an experienced investigator to validate the alignments generated. First, all peak groups were ranked according to the maximal intensity of the individual peaks within each group. Then, the top 50 peak groups (Table 1) were manually inspected. The manual inspection indicated that we could not unambiguously align two peak groups without performing spiking experiments to determine the actual identity of the metabolites. Therefore, these two peak groups were excluded from the validation process. However, manual inspection also indicated that all 39 peaks within each of 33 different peak groups were perfectly aligned by the automated method. There was one peak group that was missing one peak in one spectrum. The peak was missing because it was no longer a local maximum due to the effect of a nearby larger peak. The manual inspection also indicated that seven peak groups had minor errors: incorrect peaks were included in one or two of the 39 spectra analyzed. Seven other peak groups had major misalignments: the local patterns were disrupted by the mixture effect caused by peak shifting and/or missing/extra peaks, especially the phenomenon that twin peaks merged to a single peak in certain spectra. Examples of regions with properly aligned or misaligned peaks are shown in Figure 3. Thus, manual inspection of 1872 peaks (39 peaks within each of 48 peak groups analyzed) indicated that the algorithm performed quite well; 1810 (96.7%) peaks were correctly aligned by the automated algorithm.

We describe an automated method for peak alignment in 2D-NMR spectra that enables assessment of the relative abundance of individual metabolites. It uses a greedy algorithm, which continuously refines peak alignments until a local maximum of the object function is reached. The heuristic analysis identifies a reasonable local maximum that is not far from the true global maximum. The final result is dependent upon the order in which spectra are added into the initial alignment. By assessing multiple orders of spectral addition and selecting the best result, the probability of identifying an optimum local or even a global maximum is increased. Furthermore, this method only interchanges peaks within a triplet of peak groups. It is possible to swap peaks within a higher order of peak groups, such as quartets or quintets. However, this requires a longer processing time.

The similarity measure described here represents our initial attempt to assess the degree of pattern matching among peaks. There are a number of opportunities that could improve this method. Clearly, other measures of similarity can subsequently be assessed, including those that more properly restrict local shape matching. The Pearson’s correlation of local neighborhood term in Equation (3) is a raw approximation for shape matching. A shape model could be established and fitted to each peak to provide more precise shape information for matching. The overall distribution of neighboring peaks can also provide information that can be used for peak alignment; this will probably solve the misalignment shown in Figure 3B and C. Since the alignment algorithm is independent to the method used for peak identification, kernel smoothing (Smith et al., 2006) or wavelet transform (Du et al., 2006) could be applied to detect peaks before aligning the spectra. Additionally, modifications to the point-matching method can be introduced into this algorithm to improve the ability to maximize the total similarity measure, which could also improve the final alignment. Irrespective of the modifications that are subsequently utilized, this automated method represents a significant first step that provides a substantial improvement in our ability to analyze 2D NMR metabonomic data.

### Table 1. Comparison of automated and manual alignment of peaks on 2D NMR spectra

<table>
<thead>
<tr>
<th>Number of peak groups</th>
<th>Number of peaks aligned correctly</th>
<th>Number of peaks analyzed</th>
<th>Comparison result</th>
<th>Error type</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>1287</td>
<td>1287</td>
<td>Perfectly aligned</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>38</td>
<td>39</td>
<td>Correctly aligned, missing peaks</td>
<td>Missing peaks due to peak missing/merging</td>
</tr>
<tr>
<td>7</td>
<td>266</td>
<td>273</td>
<td>Minor error</td>
<td>Incorrect peaks in 1–2 of 39 spectra</td>
</tr>
<tr>
<td>7</td>
<td>219</td>
<td>273</td>
<td>Major error</td>
<td>Local patterns disrupted by peak shifting and missing/extra peaks</td>
</tr>
</tbody>
</table>

For this comparison, peaks representing 48 different chemical groups across 39 2D NMR spectra were analyzed. All peak groups were ranked by their maximum peak intensity; then, the top 50 peak groups were manually inspected. Two peak groups were excluded because manual alignment could not be accomplished. The manual alignment results were then compared to the output of the automated alignment algorithm. Overall, manual inspection of 1872 peaks indicated that 96.7% (n = 1810) of the peaks were aligned correctly by the automated alignment algorithm.
Fig 3. Examples of peaks that were properly aligned (A) had minor errors (B) or misaligned (C) by the automated algorithm. Each image in a column shows the same region of a 2D NMR spectrum of metabolites in liver obtained from each of the four indicated mouse strains. The crossed lines in the spectra in each column indicate the exact same $^{13}$C and $^1$H coordinates. The numbers inside and outside of the parenthesis are the chemical group identities assigned to the peaks by the automated or manual alignment methods, respectively. For a given peak, the different numbers inside and outside of the parenthesis indicate a peak misalignment. (A) Example of peaks that are properly aligned by the automated method. The identity of chemical group 1 was determined to be Glutathione according to HSQC, COSY and TOCSY spectra. This assignment was subsequently confirmed by acquiring the spectra of samples that were spiked with purified Glutathione. (B) Example of automated peak alignments with minor errors. A peak located near the cross had a relatively large shift in the SMJ spectrum. This caused the automated alignment algorithm to select a nearby peak causing a misalignment of this peak. (C) Example of a significant error by the automated peak alignment algorithm. The peaks representing three chemical groups (3, 4 and 5) within the area near the cross are shown. In the SJL and C57BL6 spectra, all three peaks are present and correctly aligned. However, in the DBA and SMJ spectra, the peaks representing chemical group 4 are missing. This caused the algorithm to align the peaks representing chemical group 5 in the DBA and SMJ spectra with the peaks representing chemical group 4 in the SJL and C57BL6 spectra.
ACKNOWLEDGEMENTS

We thank Erin Farrell for the help on the animal study and sample preparation. P.L. was supported by a grant (1 R01 GM068885-01A1) from the NIGMS awarded to G.P.

Conflict of Interest: All the authors are employees of Roche Palo Alto LLC, and there is a patent being filed by Roche for the method.

REFERENCES


