analytical chemistry

ARTICLE

pubs.acs.org/ac

Hierarchical Alignment and Full Resolution Pattern Recognition of 2D NMR Spectra: Application to Nematode Chemical Ecology

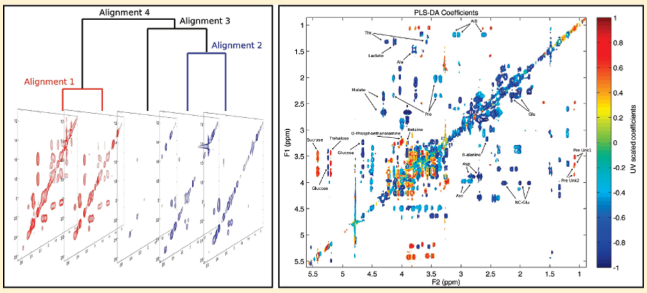
Steven L. Robinette,[†,⊥] Ramadan Ajredini,[†] Hasan Rasheed,[†] Abdulrahman Zeinomar,[†] Frank C. Schroeder,[§] Aaron T. Dossey,[†,¶] and Arthur S. Edison*[,†,‡]

[†]Department of Biochemistry and Molecular Biology and [‡]National High Magnetic Field Laboratory, University of Florida, Gainesville, Florida 32610-0245, United States

[§]Boyce Thompson Institute, Cornell University, Ithaca, New York 14853, United States

S Supporting Information

ABSTRACT: Nuclear magnetic resonance (NMR) is the most widely used nondestructive technique in analytical chemistry. In recent years, it has been applied to metabolic profiling due to its high reproducibility, capacity for relative and absolute quantification, atomic resolution, and ability to detect a broad range of compounds in an untargeted manner. While one-dimensional (1D) ^{1}H NMR experiments are popular in metabolic profiling due to their simplicity and fast acquisition times, two-dimensional (2D) NMR spectra offer increased spectral resolution as well as atomic correlations, which aid in the assignment of known small molecules and the structural elucidation of novel compounds. Given the small number of statistical analysis methods for 2D NMR spectra, we developed a new approach for the analysis, information recovery, and display of 2D NMR spectral data. We present a native 2D peak alignment algorithm we term HATS, for hierarchical alignment of two-dimensional spectra, enabling pattern recognition (PR) using full-resolution spectra. Principle component analysis (PCA) and partial least squares (PLS) regression of full resolution total correlation spectroscopy (TOCSY) spectra greatly aid the assignment and interpretation of statistical pattern recognition results by producing back-scaled loading plots that look like traditional TOCSY spectra but incorporate qualitative and quantitative biological information of the resonances. The HATS-PR methodology is demonstrated here using multiple 2D TOCSY spectra of the exudates from two nematode species: *Pristionchus pacificus* and *Panagrellus redivivus*. We show the utility of this integrated approach with the rapid, semiautomated assignment of small molecules differentiating the two species and the identification of spectral regions suggesting the presence of species-specific compounds. These results demonstrate that the combination of 2D NMR spectra with full-resolution statistical analysis provides a platform for chemical and biological studies in cellular biochemistry, metabolomics, and chemical ecology.



Nuclear magnetic resonance (NMR) is a powerful and almost universal detector due to its ability to analyze essentially all types of molecules at atomic resolution. Recently, it has been applied extensively to metabolic profiling in studies of human and animal biofluids,[1,2] cellular biochemistry,[3,4] and nonmammalian chemical ecology.[5−8] While one-dimensional (1D) ^{1}H NMR spectra have been used in the majority of NMR-based metabolic profiling studies, resonance overlap and lack of structural correlations can limit the utility of the 1D approach. In contrast, powerful multidimensional NMR methods are routinely used in structural biology and natural products studies[9] but have been used much less frequently for metabolomics. Two dimensional (2D) homonuclear (such as total correlation spectroscopy (TOCSY) and correlation spectroscopy (COSY)) and heteronuclear (such as heteronuclear single quantum coherence (HSQC) and heteronuclear multiple-bond correlation (HMBC)) spectra offer increased dispersion of signals in two dimensions that can overcome the problem of signal overlap, which often limits the use of 1D ^{1}H NMR spectra. Additionally, the information provided by these 2D experiments through specific atomic correlations can assist in the identification and assignment of known molecules and the structural elucidation of unknown small molecules that may be significant to the underlying biology. The presence of novel, uncharacterized small molecules in complex biological mixtures is a frequent occurrence, especially in nonmammalian metabolomics where a priori knowledge of the metabolome is often incomplete. Even when complex biological mixtures are composed mainly of known metabolites, the assignment of a large number of spectral peaks is a significant complication for metabolic profiling. The use of 2D correlation

Received: October 14, 2010
Accepted: January 19, 2011
Published: February 11, 2011

spectra such as TOCSY simplify the use of semiautomated assignment platforms such as MetaboMiner[10] and COLMAR.[11−13]

While it is clear that metabolic profiling would benefit from more extensive use of 2D NMR spectra, preprocessing methods such as signal alignment and visualization tools for multiple 2D spectra have lagged behind similar methods for 1D spectra. The importance of signal alignment for statistical analysis of spectral data sets is widely appreciated by the metabolomics community for both NMR[14,15] and mass spectrometry.[16,17] The development of 1D alignment methods has allowed these spectra to be analyzed with increasing accuracy and resolution. However, the relative paucity of 2D spectral alignment[18] and analysis methods has meant that metabolomics studies making use of 2D spectra have relied on low-resolution bins or crosspeak integrals to avoid the problem of peak shift.[19−21] While often quite successful in mitigating the effects of chemical shift variation, the use of integrals or bins for pattern recognition causes significant loss of valuable information from spectral resolution, complicating subsequent model interpretation and assignment of relevant metabolites. Here, we present a method for hierarchical alignment of sets of 2D NMR spectra, thereby enabling pattern recognition using full-resolution data; we call this overall approach HATS-PR for hierarchical alignment of two-dimensional spectra-pattern recognition. We demonstrate that statistical representations of full-resolution spectra such as scaled principle component analysis (PCA) loadings enhance the interpretation of metabolic variation in the data set and assist in the identification and assignment of statistically relevant metabolites.

We have applied HATS-PR to an experimental comparison of the exudates (i.e., exometabolomes) from two nematode species, *Pristionchus pacificus* and *Panagrellus redivivus*. Nematode chemical ecology, especially applied to the model organism *Caenorhabditis elegans*, has become the subject of a number of recent studies in development,[22−25] aggregation behavior,[26] mating behavior,[27,28] aging,[29] chemical ecology,[7,26] and functional genomics.[6,29,30,31] The discovery and structural elucidation of a class of pheromones called ascarosides has expanded the domain of *C. elegans* research well into chemical ecology.[32] In a combination of chemical ecology and genetics, Pungaliya et al. used comparative 2D NMR spectral analysis to discover new ascarosides and their biological roles in *C. elegans*.[27] By applying standard statistical pattern recognition methods to 2D NMR spectra, we demonstrate the utility of HATS-PR to distinguish and characterize chemical differences between two different nematode species.

EXPERIMENTAL SECTION

Collection of *P. pacificus* and *P. redivivus* Exudates. *P. pacificus* and *P. redivivus* strains were obtained from the Sternberg laboratory at Caltech.

P. pacificus. Ten nematode growth medium (NGM) plates seeded with *Escherichia coli* (OP50 strain) were inoculated with worms and incubated at room temperature for 3 days. Worms were washed off the plates with 25 mL of S-complete medium and transferred to a 250 mL Erlenmeyer flask containing 3% (0.75 g) *E. coli* (HB101 strain). Worms were incubated at 22 °C for 3 days and then transferred to a 2 L Erlenmeyer flask containing 250 mL of S-complete medium and 3% (7.5 g) HB101. Once the majority of worms were adults with eggs, sucrose flotation was used to remove bacteria and other life-stage worms. Adults with eggs were then incubated for 48 h at 22 °C at 250 rpm without food to allow the adult worms to lay the eggs and the eggs to hatch. J2 *P. pacificus* were separated from adult worms using a 20 μm nylon filter (Sefar Inc.) to obtain a semisynchronous culture. Semisynchronized J2 *P. pacificus* (2.5 million) were grown to adult stage on S-complete medium supplemented with 3% HB101 and incubated in a shaker at 250 rpm and 22 °C. Worms from this culture were then put onto a sucrose gradient to remove bacteria. Clean *P. pacificus* were placed in M9 buffer for 30 min to clear their gut. Worms were washed three times with double-distilled water (ddH_20) and then incubated for 1 h in ddH_20 with a density of ~30 000 worms/mL to collect worm exudates. Worm exudates were filtered with a 0.2 μm nylon syringe filter and stored at −80 °C for further analysis.

P. redivivus. *P. redivivus* were cultivated similarly to *P. pacificus* and transferred to liquid culture. A mixed population of *P. redivivus* with worm density of 10 000 worms/mL was incubated in 250 mL of S-complete medium supplemented with 3% HB101. *P. redivivus* were cultured for 12 days, and 3% HB101 was added every third day. After 12 days of incubation, the worms were separated from bacteria via sucrose flotation. The *P. pacificus* protocol described above was also used to collect *P. redivivus* exudates.

Sample Fractionation. *P. pacificus* and *P. redivivus* exudates were each fractionated by C18 solid phase extraction using a Varian Mega BE C18, which has a 5 mL void volume. A 50 mL syringe was attached to the column using a rubber adapter. Teflon tape was placed around the rubber stopper so that rubber pieces would not contaminate the solution. Approximately constant pressure was applied to the column by gently pressing the syringe. Each C18 column was first washed with 50 mL of 90% methanol, followed by 50 mL of ddH_20 for equilibration. *P. pacificus* and *P. redivivus* exudates were then added to individual columns, and the flow-through was collected. The columns were washed with 25 mL of ddH_20 and then eluted with 25 mL of 50% methanol and 25 mL of 90% methanol. All samples were collected in glass vials with Teflon caps and stored at −80 °C.

NMR Spectroscopy. The flow-through fractions were lyophilized, dissolved in 500 μL of 99.9% D_2O (Cambridge Isotope Laboratories), and put into 5 mm NMR tubes. The 50% MeOH fractions were lyophilized, dissolved in 110 μL of 99.95% methanol-d_4 (Cambridge Isotope Laboratories), and put into 2.5 mm NMR tubes. 1D ^1H and 2D TOCSY spectra were collected on a Bruker Avance II spectrometer operating at 600.23 MHz using a 5 mm TXI cryoprobe. NMR data were collected with spectral widths of 11 ppm at a sample temperature of about 300 K. 1D ^1H spectra were collected using a simple pulse-acquire sequence with presaturation of residual water. TOCSY spectra were collected with the DIPSI-2 mixing sequence[33] with water presaturation and 60 or 90 ms mixing times for the 50% MeOH or flow-through fractions, respectively. TOCSY data were collected with 2048 and 256 complex data points in the direct and indirect dimensions, respectively. TOCSY data were zero filled to 2048 F2 and 1024 F1 data points, Fourier transformed, phased, and baseline corrected in Topspin 2.0 (Bruker) before being transferred to Matlab 2009b for alignment and pattern recognition analysis. As an additional application to a more complex mixture, TOCSY spectra were acquired from human urine (details provided in Supporting Information), and data were processed similarly to the worm exudate spectra.

Hierarchical Alignment of Two-Dimensional Spectra. The HATS alignment method presented here makes use of a guide tree to structure the alignment process. While many existing 1D

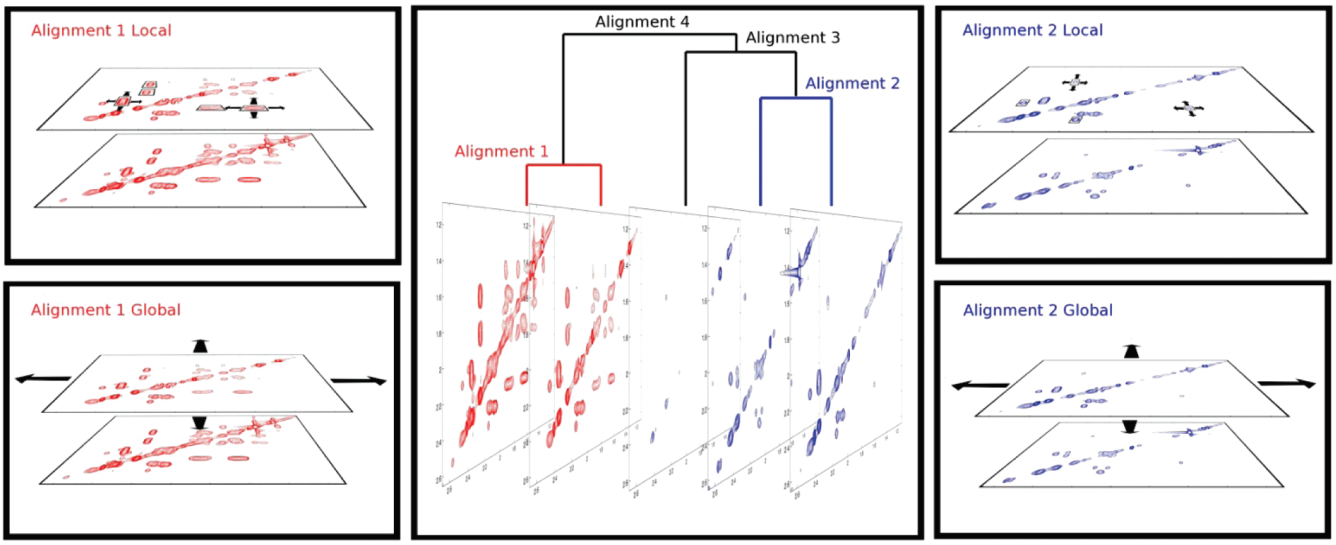


Figure 1. Overall strategy of hierarchical alignment of two-dimensional spectra (HATS). Before alignment, the HATS algorithm produces a guide tree by the UPGMA agglomerative hierarchical clustering algorithm from a matrix of pairwise correlation distances. This guide tree is used to structure the alignment process. At each intersection of branches in the guide tree, the spectra specified by the two branches are aligned globally and locally. For the alignment tree shown here, the two red spectra are aligned first, followed by the two blue spectra. The two blue spectra are then aligned with the black spectrum, and finally, the three blue and black spectra are aligned with the red spectra.

alignment methods have attempted to align each spectrum in a data set to a single consensus spectrum, the quality of these "star alignments" can suffer if there is wide chemical diversity between spectra, as any individual spectrum chosen as a star may not include the full complement of metabolites present in the other data sets. The use of guide trees was a significant advance for multiple sequence alignment in sequence bioinformatics, and methods such as CLUSTAL-W[34] have enjoyed widespread popularity in nucleic acid and protein sequence alignment. We suggest that the use of a guide tree will be especially important in the alignment of 2D spectra as correlation spectra are often collected when significant metabolic diversity is expected and the assumptions underlying star alignments may not be appropriate.

The UPGMA agglomerative hierarchical clustering method[35] was used to construct the initial guide tree for the subsequent pairwise alignment of spectral groups. We have used the squared distance of Pearson's correlation coefficient as a metric for clustering (eqs 1 and 2), and the resulting dendrogram specifies the order in which alignment of the spectra proceeds.

$$r_{mn} = \frac{\sum_{i=1}^{j}(S_{im} - \bar{S}_m)(S_{in} - \bar{S}_n)}{\sigma_{s_m}\sigma_{s_n}} \qquad (1)$$

$$d_{mn} = 1 - r_{mn}{}^2 \qquad (2)$$

Here, r_{mn} is the correlation coefficient between the reshaped vectors ($1 \times \mathrm{F1}\cdot\mathrm{F2}$) of spectra S_m and S_n with mean values $\bar{S}_m$ and $\bar{S}_n$ and standard deviations σ_{s_m} and σ_{s_n} and d_{mn} is a distance version of the correlation value. At each branch point in the dendrogram, the set of spectra specified by the branches are aligned. For the alignment illustrated in Figure 1, the two red spectra are aligned first, followed by the two blue spectra; then, the two blue spectra are aligned with the black spectrum, and finally, the two red spectra are aligned with the three blue and black spectra. If there are multiple spectra in an alignment group, positional adjustments are made equally to all spectra in the group.

The alignment at each branch point of the guide tree consists of two parts: a global alignment in which all spectral data points are shifted equally and a local alignment in which specific crosspeak regions are shifted individually. The global alignment step seeks to correct defects such as miscalibration or referencing, which produce chemical shift differences across the entire spectrum. One of the groups of spectra is passed over the other group as a mask in a point-wise manner in both F1 and F2 dimensions, and at each positional step x and y, the alignment quality is expressed as r_{xy}, the mean correlation between groups (eq 3). Here, a is the number of spectra in group 1, and b is the number in group 2, and r_{mn} is the correlation coefficient between the reshaped vectors of spectrum m in group 1 and spectrum n in group 2. Spectra in the mask group are shifted to the points m and n producing the maximal mean correlation which is identified by a simple gradient ascent algorithm.

$$r_{xy} = \frac{1}{a \cdot b}\sum_{m=1}^{a}\sum_{n=1}^{b} r_{mn} \qquad (3)$$

While global alignment can resolve factors affecting all resonances equally, it is not capable of resolving differences in specific resonances caused by conditions such as pH, temperature, metal ion concentration, and osmolality. To resolve local chemical shift variation, crosspeak regions must be aligned individually. To identify segments of the spectrum corresponding to crosspeak regions, a local noise surface is calculated for each spectrum.[36] Initial spectral segments are identified as regions within 0.04 ppm of a peak, defined as a local maximum with intensity at least 10× the noise value at that point. Each segment is then labeled with a unique integer value, and these initial segments are expanded by iterated minimum and maximum filtering. Min/max filtering has the effect of growing the segments until they encounter another spectral segment or until they reach a maximum frequency range of approximately 0.2 ppm (Figure 1S, Supporting Information). This procedure creates spectral segments that are bounded by

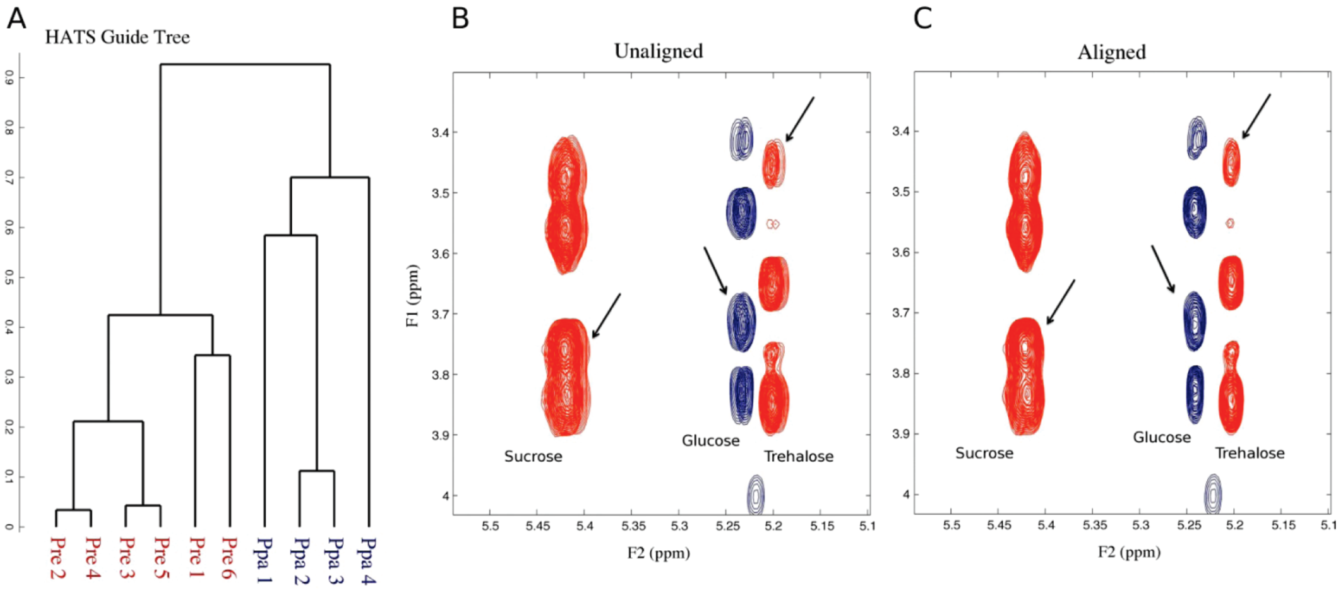


Figure 2. Alignment results for *P. redivivus* (Pre) and *P. pacificus* (Ppa) TOCSY spectra of flow-through fractions from C18 solid phase extractions. (A) The guide tree produced by HATS groups spectra from the same species, resulting in intraspecies alignment before interspecies alignment. The utility of a guide tree-based alignment is exemplified by the anomeric crosspeaks of glucose, sucrose, and trehalose in the TOCSY spectra (B, C). By aligning spectra with similar composition first, false positive alignments of nearby but unrelated crosspeaks, such as the anomeric crosspeaks shown above, are avoided.

nearby resonances yet avoid close cropping of crosspeak lineshapes. Each crosspeak segment is then aligned individually by maximizing the mean cross-correlation of crosspeak segments between alignment groups in the same way as the global alignment but treating each crosspeak segment independently (eq 3, where r_{xy} is the line shape vector for a single crosspeak region rather than the entire spectrum).

Full Resolution TOCSY Pattern Recognition. Following alignment, HATS-PR makes use of principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) for full resolution spectral pattern recognition. PCA is an unsupervised dimensionality reduction technique that is widely used in metabolomics. PCA transforms a data set of N observations with P features into a set of principal components that are orthogonal linear combinations of features accounting for as much of the variation in the original data set as possible. Each principal component consists of N scores and P loadings, where scores identify relationships between observations (here, TOCSY spectra) and loadings represent the pattern of features (here, 2D NMR chemical shifts) underlying the relationships identified by scores. We used the NIPALS algorithm[37] to identify the first five principal components of the mean-centered and univariate scaled flow-though (N = 10) and 50% MeOH (N = 8) TOCSY spectra. While PCA avoids the problem of overfitting by taking an unsupervised approach to pattern recognition, it is not optimized for class separation. Partial least-squares methods such as PLS-DA are often the supervised method of choice for the analysis of metabolic differentials. We apply SIMPLS[38] PLS-DA to the data set to help clarify the metabolic differential identified by PCA while avoiding overfitting by the use of 4-fold cross-validation and reporting performance statistics for the PLS-DA model.

Pattern recognition of NMR spectra was done in Matlab 2009b following Probabilistic Quotient Normalization[39] to account for random dilution effects and alignment using the HATS method for the TOCSY spectra and the RSPA[15] algorithm for the 1D ^{1}H spectra. Each of the N 2048 × 1024 TOCSY data matrices was reshaped to a 1 × 2 097 152 vector before the application of PCA or PLS. This is a necessary step because it creates an N × P matrix, where $P = F1 \cdot F2$ data points on which pattern recognition algorithms can operate. To visualize the resulting 1 × P loadings, the loadings vectors for each principal component were reshaped back into a 2048 × 1024 data matrix. Contours were defined by multiplying the univariate scaled loadings by the standard deviation along the third dimension of the original TOCSY stack, a procedure known as back-scaling. Contour colors were then defined by the loading values themselves divided by the maximum loading coefficient to produce a range of −1 to +1. This strategy produces loading spectra which have crosspeak lineshapes similar to traditional TOCSY spectra but include sign and intensity information related to the relationships between spectra specified by the corresponding PCA scores. PLS-DA results are presented in a similar manner, but instead of specifying scores and loadings, only the predictive feature vector is shown, along with performance statistics for the model. This 2D NMR unfolding and refolding approach has recently been applied successfully to HSQC spectra and provides a simple methodology to apply statistical pattern recognition approaches to 2D NMR spectra.[40,41]

Semi-Automated Assignment of TOCSY Loadings. To assign the compounds represented in our back-scaled TOCSY loadings plots, we applied COLMAR query[11] to correlated chemical shifts present in the loadings spectra. When COLMAR query identified a match to a compound reference spectrum present in the BMRB,[42,43] HMDB,[44] or MMCD[45] databases, the reference TOCSY spectrum for the suggested compound was downloaded from the database web servers, processed in Topspin 2.0, and overlaid onto an experimental spectrum for confirmation. We considered the compound to be assigned if all reference chemical shifts and experimental chemical shifts differ by less than 0.02 ppm and all through-bond correlations (TOCSY crosspeaks) in the

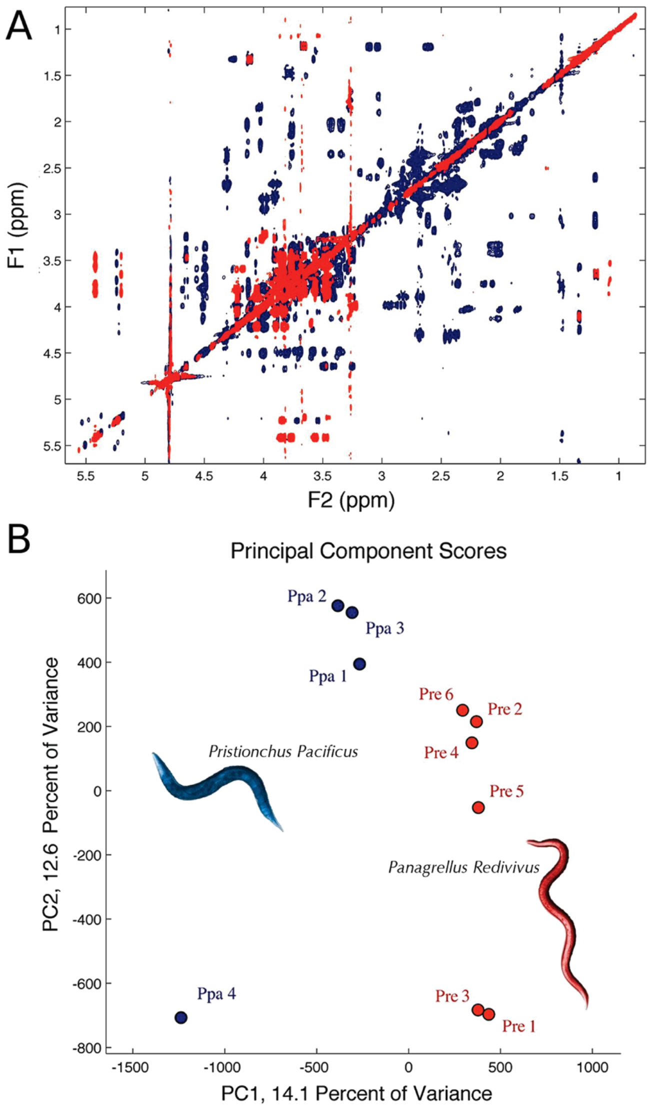


Figure 3. (A) Overlay of *P. redivivus* (Pre) and *P. pacificus* (Ppa) flow-through TOCSY spectra and (B) scores plot from PCA of the flow-through TOCSY data set. *P. redivivus* and *P. pacificus* spectra are clearly differentiated by PC1, which allows the loadings of PC1 to be interpreted as relative quantitative spectral differences between the species.

experimental spin-system were accounted for by the reference spectrum. It should be emphasized that for this study we did not verify assignments by spiking natural samples with authentic commercial or synthetic standards, a common practice for final verification of compound identity by NMR.

RESULTS

Previous work from our group and others has shown that *C. elegans* releases a large number of ascarosides that have been shown to have multiple functions.[22−25,27,28,32] These ascarosides typically partition into the 50% MeOH fraction of a C18 solid phase extraction. We have also shown that *C. elegans* releases an abundance of small polar molecules including amino acids, sugars, and organic acids, many of which elute from a C18 column in the hydrophilic flow-through fraction.[7] Here, we analyzed 2D TOCSY NMR data of the flow-through and 50% MeOH C18 elution fractions from *P. redivivus* and *P. pacificus* exudates in order to demonstrate our new methods and to show the utility of this approach in metabolomics and natural product comparative studies. In order to demonstrate the general applicability of our approach, we have also applied it to the alignment of 2D TOCSY spectra from human urine (Figure 1S, Supporting Information).

Alignment of TOCSY Spectra. Despite the high reproducibility of NMR spectra, resonance variation, sometimes referred to as positional noise, resulting from slight differences in sample conditions hinder comparative studies and complicate the interpretation of statistical analyses such as univariate regression and significance testing, principal component analysis (PCA), and partial least-squares (PLS). Previous research with 1D ^{1}H NMR spectra has shown that peak shift can result in both false positives and negatives when attempting to identify metabolite biomarkers.[46] While binning is often used to reduce positional variation, loss of high-resolution information complicates metabolite assignment and limits the added value of 2D experiments. The HATS alignment methodology presented here compensates for the effects of resonance variation, thereby increasing the accuracy and interpretability of comparative 2D NMR analyses.

The HATS-PR approach to eliminating positional variation is to first cluster all of the spectra through a guide tree, as shown in Figure 2A. The problem of positional variation in our nematode TOCSY spectra is easily seen in the expansion of the anomeric region of the sugars from the flow-through fractions (Figure 2B, C). In the unaligned data (Figure 2B), local differences in resonance frequency of the anomeric protons affect multiple spectra of both *P. redivivus* and *P. pacificus* samples and result in a general "blurriness" that significantly impacts subsequent statistical analyses. Removing this variation is clearly desirable, but the close proximity (∼0.03 ppm) and similar patterns and lineshapes of the *P. pacificus* glucose and *P. redivivus* trehalose crosspeaks raise the possibility of inappropriate alignment of the glucose to the trehalose crosspeaks. The use of the HATS-PR guide tree (Figure 2A) helps avoid this possibility by aligning the most similar before the less similar spectra. Using this approach, the glucose and trehalose resonances are not compared for alignment until they have been well aligned separately. As shown in Figure 2C, this results in the correction of positional variation within both glucose and trehalose crosspeaks and rejection of the inappropriate alignment of the two sugars.

***P. redivivus* and *P. pacificus* Release Different Polar Molecules.** Next, using the aligned TOCSY spectra (Figure 3A) of C18 flow-through fractions, we conducted PCA and PLS analysis. The PCA scores show clear, unambiguous separation of the *P. redivivus* spectra from the *P. pacificus* spectra in the first principal component (Figure 3B). This separation allows the PC1 loadings to be interpreted as a pattern of TOCSY crosspeaks differentiating the two species (Figure 4). The back-scaling procedure outlined above produces contours reflecting the crosspeak lineshapes in the original spectra but colored according to the loading coefficients. As the *P. redivivus* and *P. pacificus* PC1 scores have opposite signs, we represent features with greater intensity in the *P. redivivus* spectra in red (loadings between 0 and 1) and with greater *P. pacificus* intensity in blue (loadings between −1 and 0). The PLS predictors also agree with the PC1 loadings (Figure 2S, Supporting Information). The assignment of the crosspeaks in the PC1 loadings spectrum enables the identification of compounds unique to each species as well as quantitative differences between compound levels.

To interpret this spectral information in a chemical context, it is necessary to assign NMR resonances to their molecules (Table 1S, Supporting Information). While assignment has traditionally

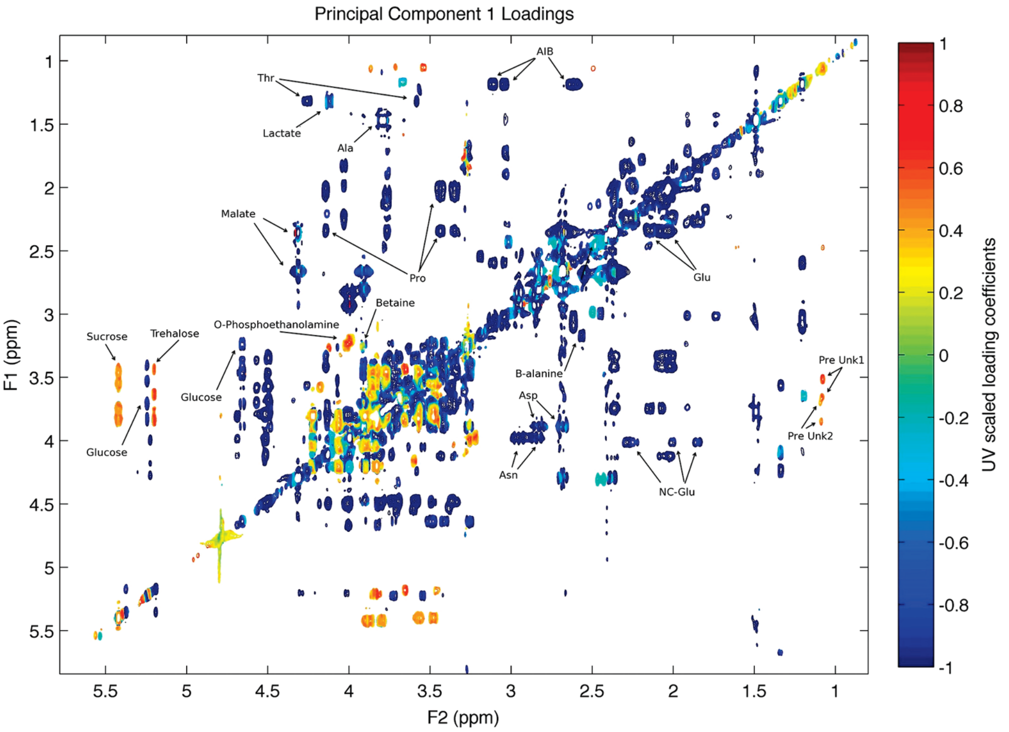


Figure 4. Spectral back-projection of PC1 loadings. Here, contours are defined by the back-projected intensities, while colors are defined by the unit variance-scaled loading coefficients. Crosspeaks with positive loading coefficients (red) represent compounds overexpressed by *P. redivivus* relative to *P. pacificus*. Blue crosspeaks represent negative loading coefficients indicating *P. pacificus* overexpressed compounds. Assignment of crosspeaks in the back-projected loadings places the chemical information contained in the TOCSY spectra in an immediate biological context. Abbreviations: AIB, 3-aminoisobutyrate; Ala, alanine; Asn, asparagine; Asp, aspartate; Glu, glutamate; NC-Glu, *N*-carbamylglutamate; Pre Unk1, *P. redivivus* unknown 1; Pre Unk2, *P. redivivus* unknown 2; Pro, L-proline; Thr, threonine.

been one of the more difficult steps in metabolomic analysis due to the large chemical complexity in most crude biological matrices, the establishment of reference spectra databases for common metabolites such as the BMRB,[41−43] HMDB,[44] and MMCD[45] has facilitated the development of database searching algorithms to match sets of correlated chemical shifts to a compound ID. For example, PRIMe SpinAssign,[47,48] NMRshifDB,[49] MetaboMiner,[10] and COLMAR[12] are freely available web utilities that offer automated database matching capabilities to aid the assignment process. The assignments shown in Figure 4 were generated by a combination of COLMAR query[11] to identify possible matches and manual overlay of reference spectra from the BMRB to confirm identifications, as described in the Experimental Section. While correlated chemical shifts were identified manually for input into COLMAR query, the back-scaled loadings matrix produced by HATS-PR should be compatible with spectral decomposition methods such as DemixC[13] and local peak clustering.[50]

***P. redivivus* and *P. pacificus* Appear to Produce Different Ascarosides.** *C. elegans* releases at least 10 different ascarosides important in the nematode's chemical ecology,[32] and methods to efficiently compare other species for ascaroside-like compounds would be useful. The most diagnostic NMR signatures for ascarosides are typically resonance peaks from anomeric, methyl, and methylene protons (Figure 5) present in the eluate of C18 SPE chromatography matrix with 50% MeOH C18. As with the flow-through fraction, PC1 separates the *P. redivivus* and *P. pacificus* 50% MeOH fraction spectra (Figure 3S, Supporting Information). The PC1 loadings from the 50% MeOH fractions show numerous crosspeaks in the ascaroside signature region. These peaks are difficult to resolve using 1D ^{1}H NMR, and loadings from pattern recognition on the 1D spectra are difficult to interpret (Figure 4S, Supporting Information). Interestingly, both *P. redivivus* and *P. pacificus* show signals representing ascaroside-like compounds; however, differences in chemical shift and signal intensity suggest different sets of ascarosides produced by these species. We tried database matching the 50% MeOH, as we did with the flow-through fractions, but we were unsuccessful in finding any matches. Identification of these compounds is in progress using traditional approaches and will be reported elsewhere.

DISCUSSION AND CONCLUSIONS

We have presented an efficient method to align, statistically analyze, and identify compounds in 2D NMR spectra. Although we demonstrated the approach with TOCSY spectra, all of the algorithms used for alignment, PCA, and PLS-DA are general and should work without modification for other types of 2D NMR data sets. In addition to problems with alignment, which we have addressed in this study, the increased acquisition time per sample relative to 1D experiments remains a barrier to adoption of 2D NMR methods in metabolic profiling. However, many groups are developing rapid methods for 2D data acquisition[51−53] and designing more sensitive probes,[54−56] making 2D NMR analysis of relatively large numbers of samples increasingly feasible.

The additional structural information provided by 2D NMR methods is often helpful for structural characterization of known metabolites and is necessary for assignment of unknown or novel compounds. Our use of COLMAR query for assignment of compounds present in multiple TOCSY spectra constitutes a

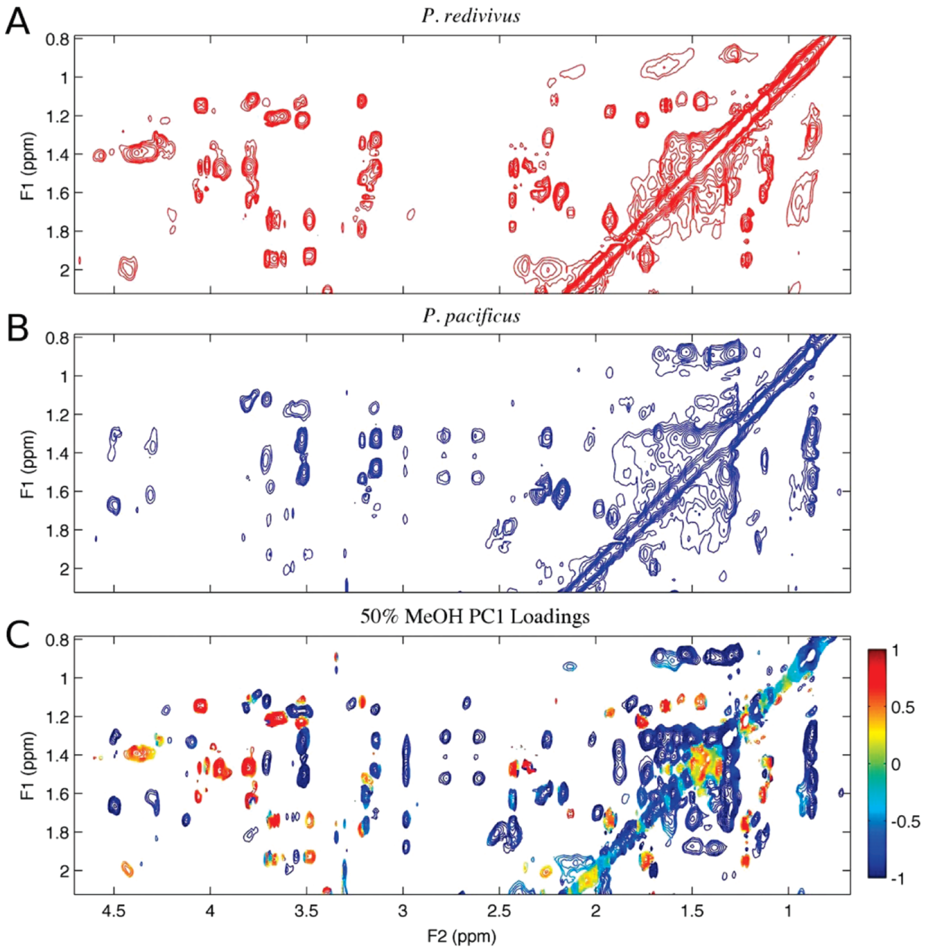


Figure 5. Representative spectra and PC1 loadings for *P. redivivus* and *P. pacificus* 50% MeOH fractions. The region selected is useful for ascaroside differentiation, as crosspeaks here indicate correlations from the methylene protons on the ascaroside side chain to protons near the terminal functional groups. PC1 loadings suggest that *P. redivivus* and *P. pacificus* produce different mixtures of ascaroside-like compounds.

proof-of-principle demonstration of how metabolomic analysis can be streamlined by integrating multi-2D statistical pattern recognition with informatics tools for compound assignment. While not applied in this study, we suggest that the application of automated spectral deconvolution methods, which isolate signals arising from single molecules, such as local clustering[50] and DemixC,[13] to back-scaled 2D NMR loadings matrices will result in further streamlining of metabolomic data analysis. While automated assignment is currently restricted to compounds for which reference data is present in spectral databases, progress is being made on methods to identify probable chemical structures directly from chemical shift calculations.[57] The polar metabolites present in the flow-through fraction of the nematode exudate C18 SPE extractions were particularly well suited for semiautomated assignment by COLMAR query as the reference databases contain large numbers of sugars, amino acids, small organic acids, and other common biological metabolites that tend to partition in the aqueous phase.

It is common practice with 2D NMR to overlay spectra with different colors, as shown for example in Figure 3A. There are several advantages that HATS-PR has oversimple spectral overlay. First, with spectral overlay, it is very difficult or impossible to see quantitative changes in a particular compound. Using HATS-PR, quantitative information is clearly represented by changes in the color of the crosspeaks in the loadings plots. Second, using overlays is not particularly amenable to evaluating statistical replicates and comparing multiple conditions. However, with HATS-PR, like any PCA or PLS analysis, statistical replicates enhance the analysis and are easily visualized through the corresponding scores plots and regression statistics. Additionally, quantitative covariance between 2D NMR crosspeaks may provide a means by which to identify connectivity in metabolites with multiple spin systems.

In principle, it is possible to do PCA analysis on 2D NMR data without alignment. However, alignment of the spectra aids in the interpretation of individual compounds by improving the quality of the pattern recognition analysis. Pattern recognition on unaligned 1D spectra often results in line shapes resembling first derivatives rather than Lorentzian line shapes.[15] In 2D NMR spectra, misalignment often results in the duplication of peaks (Figure 5S, Supporting Information) in the back-scaled loadings or attenuation of loading coefficients (Figure 6S, Supporting Information) due to peak position differences and can result in compound misidentification. Despite making small changes in the chemical shifts of the crosspeaks, alignment of crosspeak lineshapes to their maximum correlation results in more interpretable loadings for both the flow-though and 50% MeOH fraction, as shown by supplemental Figures S4 and S5, Supporting Information. HATS-PR represents, to our knowledge, the first usage of a guide tree in a spectral alignment method, which removes the requirement that all compounds be present in a single spectrum selected as the alignment reference. We believe guide trees will also be generally useful for peak alignment in 1D ^{1}H NMR, chromatographic profiles, and LC-MS spectra as well.

HATS-PR is a relatively simple way to chemically discriminate two (or more) biological states, such as treated vs untreated, diseased vs normal, one species vs another, etc. Using existing database search algorithms and rapidly growing NMR metabolomic databases, it is straightforward using HATS-PR not only to compare biological states but also to identify and assign the relevant compounds that differentiate these states. In this study, we compared exudates of two nematode species. The samples were prepared under identical conditions, yet revealed surprisingly large differences, particularly in the polar flow-through fractions. In previous work, we found that *C. elegans* produces at least 40 common, mostly polar metabolites,[7] and we expected to find a similar group of compounds in *P. redivivus* and *P. pacificus*. HATS-PR very efficiently showed us that the two species produce significantly different sets of polar small molecules, including sugars. *P. redivivus* in general releases a less diverse set of polar molecules but produces large amounts of the disaccharide trehalose and their exudates also contain a large quantity of sucrose, yet no glucose was observed. While *P. pacificus* releases a very diverse set of polar small molecules including glucose, the two disaccharides sucrose and trehalose were not detected. All of the nematodes for this study were separated from their bacterial food using a sucrose gradient, so it is possible that some or all of the sucrose detected in *P. redivivus* originates from the sample preparation. However, we find a comparable amount of trehalose in the same samples, and more importantly, we find no sucrose in *P. pacificus*, which were prepared with an identical sucrose gradient.

Despite the fact that spectra collected on the flow-through fractions of *P. redivivus* were much simpler than those from *P. pacificus*, the TOCSY spectra of 50% MeOH fractions from the same worm preparations exhibited similar levels of complexity but also some clear differences. Although, on the basis of the data presented here, we are unable to determine unambiguously whether these fractions contain ascarosides, the spectral signatures of the 50% MeOH fractions for both species suggest the presence of complex ascaroside mixtures, similar to those found in *C. elegans*. The identification and assignment of these compounds requires additional work and additional NMR spectroscopic and mass spectrometric analyses and is beyond the scope of this study. However, it is clear that HATS-PR is a powerful method to compare small molecules from one organism to another and as such provides a significant bridge between metabolomics and natural products analysis. As our use of HATS-PR for alignment of a set of urine spectra demonstrates, the method is broadly applicable and can successfully align arrays of 2D spectra even in situations where sample variation leads to significant chemical shift differences between spectra. We expect that this work will contribute significantly to informatics approaches to metabolic profiling as well as to biological and structural characterization of metabolic phenotypes in both mammalian and nonmammalian models.

ASSOCIATED CONTENT

S Supporting Information. Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: aedison@ufl.edu.

Present Addresses

⊥Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London, UK SW7 2AZ.

¶United States Department of Agriculture, Agricultural Research Service.

ACKNOWLEDGMENT

S.L.R. acknowledges financial support from the Barry M. Goldwater Foundation and from the Howard Hughes Medical Institute Science for Life program at UF. We thank James R. Rocca, Paul Sternberg, and Andrea Choe for helpful discussions. Burt Singer and Greg Stupp kindly read the manuscript and provided us with stimulating discussions. Nematode NMR data were collected in the University of Florida AMRIS Facility, which is part of the NSF-funded National High Magnetic Field Laboratory. Human urine NMR data were collected at Imperial College, and we are grateful to Jeremy Nicholson for the use of the NMR spectrometer. This work was supported by NIH 1R01GM085285-01and 3R01GM085285-01A1S1 to A.S.E.

REFERENCES

(1) Nicholson, J. K.; Wilson, I. D. *Prog. Nucl. Magn. Reson. Spectrosc.* **1989**, *21*, 449–501.

(2) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153–161.

(3) Behrends, V.; Ebbels, T. M.; Williams, H. D.; Bundy, J. G. *Appl. Environ. Microbiol.* **2009**, *75*, 2453–2463.

(4) Schroeder, F. C.; Gibson, D. M.; Churchill, A. C.; Sojikul, P.; Wursthorn, E. J.; Krasnoff, S. B.; Clardy, J. *Angew. Chem., Int. Ed. Engl.* **2007**, *46*, 901–904.

(5) Bundy, J. G.; Sidhu, J. K.; Rana, F.; Spurgeon, D. J.; Svendsen, C.; Wren, J. F.; Sturzenbaum, S. R.; Morgan, A. J.; Kille, P. *BMC Biol.* **2008**, *6*, 25.

(6) Blaise, B. J.; Giacomotto, J.; Elena, B.; Dumas, M. E.; Toulhoat, P.; Segalat, L.; Emsley, L. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 19808–19812.

(7) Kaplan, F.; Badri, D. V.; Zachariah, C.; Ajredini, R.; Sandoval, F. J.; Roje, S.; Levine, L. H.; Zhang, F.; Robinette, S. L.; Alborn, H. T.; Zhao, W.; Stadler, M.; Nimalendran, R.; Dossey, A. T.; Bruschweiler, R.; Vivanco, J. M.; Edison, A. S. *J. Chem. Ecol.* **2009**, *35*, 878–892.

(8) Bundy, J. G.; Spurgeon, D. J.; Svendsen, C.; Hankard, P. K.; Osborn, D.; Lindon, J. C.; Nicholson, J. K. *FEBS Lett.* **2002**, *521*, 115–120.

(9) Rae, R.; Iatsenko, I.; Witte, H.; Sommer, R. J. *Environ. Microbiol.* **2010**, *12*, 3007–3021.

(10) Xia, J.; Bjorndahl, T. C.; Tang, P.; Wishart, D. S. *BMC Bioinformatics* **2008**, *9*, 507.

(11) Robinette, S. L.; Zhang, F.; Bruschweiler-Li, L.; Bruschweiler, R. *Anal. Chem.* **2008**, *80*, 3606–3611.

(12) Zhang, F.; Robinette, S. L.; Bruschweiler-Li, L.; Bruschweiler, R. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S118–122.

(13) Zhang, F.; Dossey, A. T.; Zachariah, C.; Edison, A. S.; Bruschweiler, R. *Anal. Chem.* **2007**, *79*, 7748–7752.

(14) Lee, G. C.; Woodruff, D. L. *Anal. Chim. Acta* **2004**, *513*, 413–416.

(15) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 56–66.

(16) Wong, J. W. H.; Cagney, G.; Cartwright, H. M. *Bioinf.* **2005**, *21*, 2088–2090.

(17) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779–787.

(18) Zheng, M.; Lu, P.; Liu, Y.; Pease, J.; Usuka, J.; Liao, G.; Peltz, G. *Bioinformatics* **2007**, *23*, 2926–2933.

(19) Lewis, I. A.; Schommer, S. C.; Markley, J. L. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S123–126.

(20) Ludwig, C.; Ward, D. G.; Martin, A.; Viant, M. R.; Ismail, T.; Johnson, P. J.; Wakelam, M. J.; Gunther, U. L. *Magn. Reson. Chem.* **2009**, *47* (Suppl 1), S68–73.

(21) Van, Q. N.; Issaq, H. J.; Jiang, Q.; Li, Q.; Muschik, G. M.; Waybright, T. J.; Lou, H.; Dean, M.; Uitto, J.; Veenstra, T. D. *J Proteome Res.* **2008**, *7*, 630–639.

(22) Butcher, R. A.; Ragains, J. R.; Clardy, J. *Org. Lett.* **2009**, *11*, 3100–3103.

(23) Butcher, R. A.; Ragains, J. R.; Kim, E.; Clardy, J. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 14288–14292.

(24) Butcher, R. A.; Fujita, M.; Schroeder, F. C.; Clardy, J. *Nat. Chem. Biol.* **2007**, *3*, 420–422.

(25) Jeong, P. Y.; Jung, M.; Yim, Y. H.; Kim, H.; Park, M.; Hong, E.; Lee, W.; Kim, Y. H.; Kim, K.; Paik, Y. K. *Nature* **2005**, *433*, 541–545.

(26) Macosko, E. Z.; Pokala, N.; Feinberg, E. H.; Chalasani, S. H.; Butcher, R. A.; Clardy, J.; Bargmann, C. I. *Nature* **2009**, *458*, 1171–1175.

(27) Pungaliya, C.; Srinivasan, J.; Fox, B. W.; Malik, R. U.; Ludewig, A. H.; Sternberg, P. W.; Schroeder, F. C. *Proc. Natl. Acad. Sci. U.S.A.* **2009**, *106*, 7708–7713.

(28) Srinivasan, J.; Kaplan, F.; Ajredini, R.; Zachariah, C.; Alborn, H. T.; Teal, P. E.; Malik, R. U.; Edison, A. S.; Sternberg, P. W.; Schroeder, F. C. *Nature* **2008**, *454*, 1115–1118.

(29) Fuchs, S.; Bundy, J. G.; Davies, S. K.; Viney, J. M.; Swire, J. S.; Leroi, A. M. *BMC Biol.* **2010**, *8*, 14.

(30) Blaise, B. J.; Giacomotto, J.; Triba, M. N.; Toulhoat, P.; Piotto, M.; Emsley, L.; Segalat, L.; Dumas, M. E.; Elena, B. *J. Proteome Res.* **2009**, *8*, 2542–2550.

(31) Atherton, H. J.; Jones, O. A.; Malik, S.; Miska, E. A.; Griffin, J. L. *FEBS Lett.* **2008**, *582*, 1661–1666.

(32) Edison, A. S. *Curr. Opin. Neurobiol.* **2009**, *19*, 378–388.

(33) Shaka, A. J.; Lee, C. J.; Pines, A. *J. Magn. Reson.* **1988**, *77*, 274–293.

(34) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.

(35) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; Freeman: San Francisco, CA, 1973.

(36) Koradi, R.; Billeter, M.; Engeli, M.; Guntert, P.; Wuthrich, K. *J. Magn. Reson.* **1998**, *135*, 288–297.

(37) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

(38) Dejong, S. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251–263.

(39) Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281–4290.

(40) Hedenstrom, M.; Wiklund-Lindstrom, S.; Oman, T.; Lu, F. C.; Gerber, L.; Schatz, P.; Sundberg, B.; Ralph, J. *Mol. Plant* **2009**, *2*, 933–942.

(41) Hedenstrom, M.; Wiklund, S.; Sundberg, B.; Edlund, U. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 110–117.

(42) 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. Biocomput.* **2007**, 157–168.

(43) Ulrich, E. L.; Akutsu, H.; Doreleijers, J. F.; Harano, Y.; Ioannidis, Y. E.; Lin, J.; Livny, M.; Mading, S.; Maziuk, D.; Miller, Z.; Nakatani, E.; Schulte, C. F.; Tolmie, D. E.; Kent Wenger, R.; Yao, H.; Markley, J. L. *Nucleic Acids Res.* **2008**, *36*, D402–408.

(44) 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*, D521–526.

(45) Cui, Q.; Lewis, I. A.; Hegeman, A. D.; Anderson, M. E.; Li, J.; Schulte, C. F.; Westler, W. M.; Eghbalnia, H. R.; Sussman, M. R.; Markley, J. L. *Nat. Biotechnol.* **2008**, *26*, 162–164.

(46) Dumas, M. E.; Maibaum, E. C.; Teague, C.; Ueshima, H.; Zhou, B. F.; Lindon, J. C.; Nicholson, J. K.; Stamler, J.; Elliott, P.; Chan, Q.; Holmes, E. *Anal. Chem.* **2006**, *78*, 2199–2208.

(47) Akiyama, K.; Chikayama, E.; Yuasa, H.; Shimada, Y.; Tohge, T.; Shinozaki, K.; Hirai, M. Y.; Sakurai, T.; Kikuchi, J.; Saito, K. *In Silico Biol.* **2008**, *8*, 339–345.

(48) Chikayama, E.; Sekiyama, Y.; Okamoto, M.; Nakanishi, Y.; Tsuboi, Y.; Akiyama, K.; Saito, K.; Shinozaki, K.; Kikuchi, J. *Anal. Chem.* **2010**, *82*, 1653–1658.

(49) Steinbeck, C.; Kuhn, S. *Phytochemistry* **2004**, *65*, 2711–2717.

(50) Robinette, S. L.; Veselkov, K. A.; Bohus, E.; Coen, M.; Keun, H. C.; Ebbels, T. M.; Beckonert, O.; Holmes, E. C.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 6581–6589.

(51) Shrot, Y.; Frydman, L. *J. Chem. Phys.* **2008**, *128*, 052209.

(52) Sommer, R. J.; Tautz, D. *Nature* **1993**, *361*, 448–450.

(53) Freeman, R.; Kupce, E. *J. Biomol. NMR* **2003**, *27*, 101–113.

(54) Zhang, F. L.; Bruschweiler, R. *J. Am. Chem. Soc.* **2004**, *126*, 13180–13181.

(55) Brey, W. W.; Edison, A. S.; Nast, R. E.; Rocca, J. R.; Saha, S.; Withers, R. S. *J. Magn. Reson.* **2006**, *179*, 290–293.

(56) Olson, D. L.; Peck, T. L.; Webb, A. G.; Magin, R. L.; Sweedler, J. V. *Science* **1995**, *270*, 1967–1970.

(57) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. *BMC Bioinf.* **2008**, *9*, 400.