Available online at www.sciencedirect.com
ScienceDirect
Journal of Magnetic Resonance 189 (2007) 190–199
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Removal of t_1 noise from metabolomic 2D ^1H–^{13}C HSQC NMR spectra by Correlated Trace Denoising

Simon Poulding [a,1], Adrian J. Charlton [b,*], James Donarski [b], Julie C. Wilson [c]

[a] *Department of Mathematics, University of York, York, YO10 5DD, UK*
[b] *Department for Environment, Food and Rural Affairs, Central Science Laboratory, Sand Hutton, York, YO41 1LZ, UK*
[c] *York Structural Biology Laboratory, Department of Chemistry, University of York, York, YO10 5DD, UK*

Received 23 May 2007; revised 6 September 2007
Available online 14 September 2007

Abstract

The presence of t_1 noise artefacts in 2D phase-cycled Heteronuclear Single Quantum Coherence (HSQC) spectra constrains the use of this experiment despite its superior sensitivity. This paper proposes a new processing algorithm, working in the frequency-domain, for reducing t_1 noise. The algorithm has been developed for use in contexts, such as metabolomic studies, where existing denoising techniques cannot always be applied. Two test cases are presented that show the algorithm to be effective in improving the SNR of peaks embedded within t_1 noise by a factor of more than 2, while retaining the intensity and shape of genuine peaks.
Crown copyright © 2007 Published by Elsevier Inc. All rights reserved.

Keywords: 2D NMR Spectroscopy; HSQC; t_1 noise; Trace correlation; Metabolomics

1. Introduction

Metabolomics (here used to mean both metabonomics and metabolomics) requires the ability to resolve and identify the metabolites present in biological samples containing a complex mixture of compounds [1,2]. NMR experiments are a very effective analysis technique for metabolomic studies [3,4], and 2D NMR experiments are often used to discriminate metabolites in samples containing a large number of similar compounds by leveraging the additional spread of resonances across two dimensions [5,6].

NMR data acquired from complex mixtures of small molecules contain signals from many compounds that may have a wide range of concentrations. The analysis of complex mixtures by NMR spectroscopy can be hindered by the presence of high intensity peaks in the spectra, which produce low intensity artefacts at a similar intensity to the resonances from low concentration compounds. The effect of artefacts is an increase in the detection limit in the regions of the spectra affected by the artefacts, thus reducing the apparent sensitivity of the NMR experiment.

It is highly desirable to maintain maximum detection sensitivity when using NMR to characterise unknown matrices and the choice of NMR pulse sequence is crucial for optimising the range of compounds that can be detected. This is particularly true for the application of multidimensional NMR techniques as the repetition times can be considerable and thus, for a given experiment time, the apparent sensitivity of the experiments can be low. The Heteronuclear Single Quantum Coherence (HSQC) experiment is particularly useful for the unequivocal determination of the presence of organic compounds within a complex mixture. The distinctive combination of, for example, ^{13}C and ^1H is often sufficient to determine the presence of a specific compound in a mixture. For this reason, the HSQC experiment has been utilised in many metabolomic studies [7–14].

The 2D phase-cycled HSQC is inherently more sensitive than similar two-dimensional experiments. For example, it

[*] Corresponding author. Fax: +44 (0) 1904 462133.
E-mail address: adrian.charlton@csl.gov.uk (A.J. Charlton).
[1] Present address: SEBASE Project, Department of Computer Science, University of York, York, YO10 5DD, UK.

1090-7807/$ - see front matter Crown copyright © 2007 Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2007.09.004

is at least $\sqrt{2}$ times more sensitive than the equivalent gradient-selected HSQC experiment [15]. However, the phase-cycled HSQC experiment does not result in the desired 'artefact free' spectra and in particular can contain t_1 noise. (Fig. 1 shows an example of t_1 noise in a phase-cycled ^{1}H–^{13}C HSQC spectrum of sucrose: the noise is seen as 'ridges' parallel to the F_1 axis at the F_2 frequencies of intense peaks.) These artefacts result from the incomplete cancellation of ^{12}C signals owing to instrumental imperfections and external disturbances [16,17]. The largest t_1 noise ridges can be higher in intensity than genuine peaks associated with low concentration compounds, causing some small peaks to be obscured by t_1 noise. Relatively high intensity t_1 noise ridges can also hinder both manual and automatic peak identification when using, for example, thresholding techniques to pick peaks. The presence of artefacts in 2D NMR data therefore limits the application of fully automated NMR peak assignment programs and thus automated compound identification using databases such as [18–23], which would be highly desirable for metabolomic studies. For example, Fig. 1 in [11] shows t_1 noise in a 2D ^{1}H–^{13}C HSQC spectrum: it would be difficult to distinguish the high intensity t_1 noise peaks in this spectrum from genuine peaks of a similar or lower intensity using automated techniques.

Post-acquisition processing of the spectrum that removes t_1 noise, while retaining low intensity 'genuine' peaks, would enable the high sensitivity of HSQC experiments to be leveraged in metabolomic studies without the disadvantages that arise from t_1 noise artefacts.

A number of highly effective denoising algorithms have been described—such as Reference Deconvolution [24–26] and the Cadzow procedure [27]—that might fulfil this role. However, although these techniques are successfully applied to 2D HSQC experiments in many other contexts, their use in the domain of metabolomics is constrained by specific pre-requisites of the algorithms. For example, Reference Deconvolution requires a strong, unconvoluted reference signal which is not always available in the highly complex spectra typical of metabolomic samples. The Cadzow procedure requires the number of peaks in the sample to be specified, but this information is not known *a priori* in metabolomic studies.

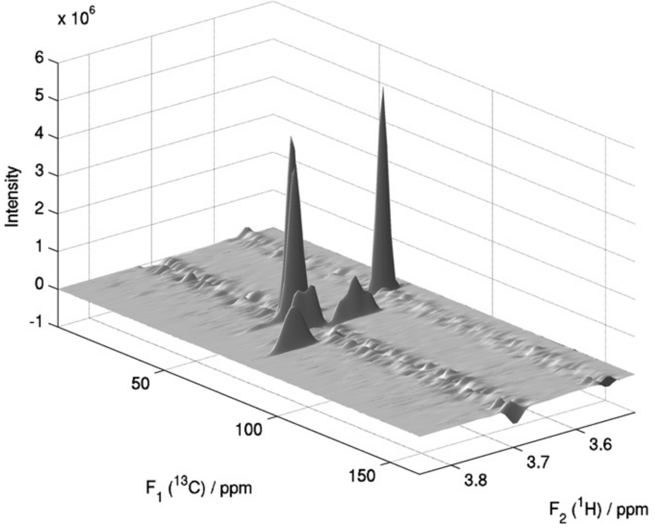

Fig. 1. Example of t_1 noise in a phase-cycled HSQC spectrum of sucrose. (Only a small section of the F_2 range is shown.)

This paper proposes a new frequency-domain processing algorithm, named *Correlated Trace Denoising*, that has been developed in response to the specific requirements of metabolomics, and that can be used in situations where existing denoising techniques are inappropriate. The algorithm is able to remove much of the t_1 noise while retaining low intensity 'genuine' peaks that may be embedded within the noise.

The correlation between t_1 noise on different traces in 2D spectra is well-established [28] and this is the basis for Correlated Trace Denoising. Here we used the phase-cycled HSQC experiment to demonstrate an application of Denoising for the removal of t_1 noise. The presence of multiple coherence transfer pathways in this experiment is such that we would anticipate more dramatic improvements in the efficacy of the algorithm when applied to experiments where the correlation between t_1 noise on different traces within the 2D spectra is greater (e.g. gradient-selected HSQC). The choice of the HSQC experiment was made due to the usefulness of the data that it produces in the context of metabolomic studies. The choice of the phase-cycled experiment was determined by its inherent sensitivity and therefore its potential for widespread application in the metabolomics field.

The remainder of the paper is organised as follows. The next section describes existing t_1 noise reduction techniques in more detail and explains why they cannot always be used for spectra from metabolomic samples. Section 3 introduces the Correlated Trace Denoising algorithm, and Section 4 provides evidence of its efficacy on two different types of spectrum. Concluding remarks are in Section 5, followed by details of materials and methods.

2. Existing noise reduction techniques for 2D NMR

2.1. Reference Deconvolution

Reference Deconvolution may be used for suppressing t_1 noise [24–26], as well as for resolution enhancement [26,29]. The technique operates in the time-domain using a correcting function derived by comparing the experimental form of an F_1 trace from a strong signal (usually that of the reference compound) with its theoretical form. By using a series of traces through the t_1 noise ridge of the reference signal across a small range of F_2 values, the technique can also account for changes in the t_1 noise in the t_2 direction that correspond to changes during acquisition of the FID.

However, reference Deconvolution may not always be suitable for use with the highly complex spectra typical of metabolomic samples. In such spectra, a strong, unconvoluted reference signal may not be present. This may be due to the absence of a robust chemical shift reference in the sample, such as may be the case when using LC–NMR, or due to signal overlap between potential reference signals and the complex signal patterns obtained from the metabolite mixture. The Correlated Trace Denoising algorithm described in this paper does not require such a reference signal, and is particularly useful for application with datasets from complex mixtures.

2.2. Cadzow procedure

The Cadzow procedure can be used to directly denoise the FIDs acquired in a 2D NMR experiment using a process described by Brissac et al. [27]. It leverages mathematical properties of a matrix derived from the time-domain signal to remove all signals apart from those resulting from a specified number of resonance frequencies.

However, the technique requires the number of genuine peaks in each FID to be specified, but this information is not known *a priori* for metabolomic samples. In [27], the peak count is obtained from a simple threshold-based peak picker, and it is unclear whether small genuine peaks would be identified by such a peak picker were they to be lower in intensity than the t_1 noise. In comparison, the Correlated Trace Denoising algorithm does not require the number of peaks to be known, and it is specifically designed to retain small 'genuine' peaks of intensity lower than nearby t_1 noise.

3. Correlated Trace Denoising algorithm

3.1. Overview

The Correlated Trace Denoising algorithm is applied to the frequency-domain spectrum after Fourier transforms of the acquired signal. It is based on the observation that there is significant similarity in the structure of t_1 noise ridges, even at widely separated F_2 values. This suggests that genuine peaks embedded within t_1 noise can be distinguished by comparison with other t_1 noise traces where the peaks may not be present. However, the t_1 noise ridges change in amplitude and phase as F_2 varies—for example, the phase of the noise can rapidly change by π radians across the centre of a ridge—requiring the algorithm to adjust for both these factors.

The algorithm begins with the 2D complex frequency-domain spectrum and consists of the following high-level steps:

(1) The spectrum is separated into 'peak' and 'noise' component spectra using a thresholding technique. The 'peak' spectrum contains large peaks with amplitude greater than the t_1 noise, while the 'noise' spectrum contains both the t_1 noise and small genuine peaks embedded within the noise.
(2) For each F_1 trace (a 1D spectrum through the 2D spectrum at a fixed F_2 frequency—effectively a slice parallel to the F_1 axis) in the 'noise' spectrum, a masking trace is derived based on the correlation between traces. The criteria for deriving the mask are chosen to remove noise but retain genuine peaks.
(3) The spectrum formed from the masking traces is subtracted from the 'noise' spectrum, leaving a spectrum consisting of the small genuine peaks. This is added to the 'peak' spectrum to create a denoised spectrum.

The following subsections describe these steps in detail. (Note that although the t_1 noise does not normally have a significant amplitude across the entire spectrum, it is found that it is effective to apply the algorithm to all F_1 traces regardless of the amplitude of the t_1 noise in that trace, and so the steps below describe its application to the spectrum as a whole.)

3.2. Separation of peak and noise spectra

The complex spectrum is separated into 'peak' and 'noise' components using a threshold derived from a statistical analysis of the spectrum. The noise separation is applied to each F_1 trace independently since the amplitude of the noise varies with F_2 but is relatively consistent along each F_1 trace.

In the following, the full complex spectrum is denoted as $\Phi(f_1,f_2)$, and the F_1 trace at the F_2 value of f_2 is denoted $\Phi_{f_2}(f_1)$. Each such trace is considered in its complex polar form:

$$\Phi_{f_2}(f_1) = r_{f_2}(f_1)\mathrm{e}^{\mathrm{i}\theta_{f_2}(f_1)} \tag{1}$$

where $r_{f_2}(f_1)$ is the amplitude (or modulus) and $\theta_{f_2}(f_1)$, the phase (or argument). It is the amplitude that is thresholded to achieve the separation of the peak and noise components, leaving the phase unchanged.

Analysis of the amplitude in traces through t_1 noise ridges suggested an approximate Rayleigh distribution (or, equivalently, a χ distribution with two degrees of freedom), which is consistent with the real and imaginary components of the noise being normally distributed. This distribution is assumed when deriving an appropriate threshold. (The sensitivity of this distribution to experimental settings was not investigated, but it would be straightforward to analyse the distribution for different settings and modify the derivation of the threshold accordingly.)

The median value, $\tilde{\Phi}_{f_2}$, of the trace is calculated, and the denoising threshold, Λ_{f_2}, set to $2.921\tilde{\Phi}_{f_2}$. The constant of 2.921 is chosen since 99.73% of the values lie in the range 0 to $2.921\tilde{\Phi}_{f_2}$ assuming a Rayleigh distribution: this proportion is equivalent to a range of the mean ±3 times the standard deviation for a normal distribution. Note that the median value is used since it is more robust than the mean—in particular, large peaks in the trace influence the median less than the mean—and so it is a better measure of the central tendency of the noise amplitude.

The amplitude is thresholded using the formula:

$$r^{\text{peak}}_{f_2}(f_1) = \begin{cases} 0 & \text{if } r_{f_2}(f_1) < \Lambda_{f_2} \\ r_{f_2}(f_1) - \frac{\Lambda^2_{f_2}}{r_{f_2}(f_1)} & \text{otherwise} \end{cases} \tag{2}$$

This is a compromise between 'hard thresholding' that simply truncates at the threshold value, and 'soft thresholding' that subtracts the threshold from all values larger than the threshold: it avoids the abrupt changes in amplitude of the former and the reduction in peak volume that results from the latter.

The amplitude for the noise component is calculated by subtracting the peak amplitude from the trace amplitude, i.e., $r^{\text{noise}}_{f_2}(f_1) = r_{f_2}(f_1) - r^{\text{peak}}_{f_2}(f_1)$, and so the corresponding noise component of the trace is:

$$\Phi^{\text{noise}}_{f_2}(f_1) = r^{\text{noise}}_{f_2}(f_1)\mathrm{e}^{\mathrm{i}\theta_{f_2}(f_1)} \tag{3}$$

Fig. 2 shows a section of a spectrum after separation into peak and noise components by the thresholding method described above.

Separation using wavelet methods was also investigated, but was not used since it added significant artefacts (pseudo-Gibbs phenomena) to the denoised spectrum, even when using translation invariant wavelet decompositions which have been reported to minimise such artefacts [30].

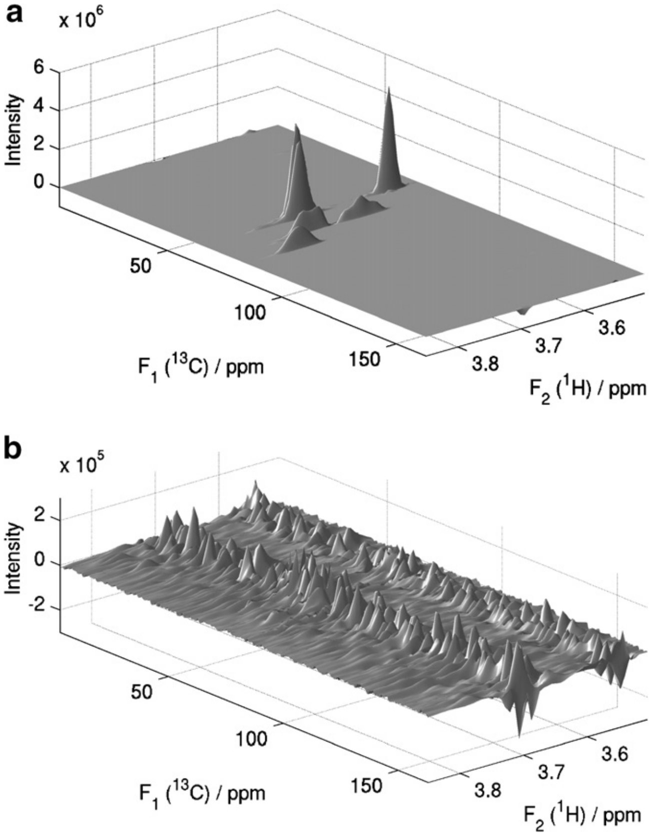

Fig. 2. Example of peak and noise spectrum separation in a phase-cycled HSQC spectrum of sucrose. (Only a small section of the F_2 range is shown.) (a) the peak spectrum; (b) the corresponding noise spectrum (using a different intensity scale).

3.3. Identification of masking traces

The noise component spectrum produced in the previous step is processed to derive a 'masking' spectrum that will be used to remove the t_1 noise.

For each trace, $\Phi^{\text{noise}}_{f_2}(f_1)$, in the noise spectrum, the set of all other noise traces, $\{\Phi^{\text{noise}}_{f'_2} : f'_2 \neq f_2\}$, is considered. From these, a subset—defined by a set M of f'_2 values—is chosen, and this subset of traces contributes to the masking trace. The criteria used to choose M are described in the following sections, and represent a balance between removing as much t_1 noise as possible, while retaining genuine peaks embedded within the noise.

3.3.1. Highly Correlated

The criterion is that the set M contains only f'_2 values for which the correlation between the noise traces at f_2 and f'_2 is above a threshold. The correlation is measured using the complex correlation, $\rho(f_2, f'_2)$, defined as:

$$\frac{\sum_{f_1}\left\{\Phi_{f_2}(f_1) - \overline{\Phi_{f_2}}\right\}\left\{\Phi_{f'_2}(f_1) - \overline{\Phi_{f'_2}}\right\}}{\sqrt{\sum_{f_1}\left\{\Phi_{f_2}(f_1) - \overline{\Phi_{f_2}}\right\}^2}\sqrt{\sum_{f_1}\left\{\Phi_{f'_2}(f_1) - \overline{\Phi_{f'_2}}\right\}^2}} \tag{4}$$

where $\overline{\Phi_{f_2}}$ is the mean value of the noise trace. When the complex value of the correlation is considered in polar form, the modulus, r_ρ, represents the degree of correlation and the argument, θ_ρ, measures the phase difference between the noise traces. Thus, this criterion is expressed as: $r_\rho(f_2, f'_2) \geqslant R_M$, where R_M is a chosen constant. The criterion accounts for the changes in the structure of the t_1 noise as F_2 varies and so ensures that only very similar traces (once phase differences are accounted for) contribute to the mask.

3.3.2. Best Correlated

For some traces, Φ_{f_2}, many other traces, $\Phi_{f'_2}$, are sufficiently well-correlated to meet the Highly Correlated criterion. However, it was found empirically that a more effective masking spectrum was derived when only the best of these highly correlated traces were used, rather than all of them. This is the motivation for the Best Correlated criterion: the $r_\rho(f_2, f'_2)$ values are ranked in order and only the f'_2 values for the best N_M of these correlations are included in the set M.

3.3.3. Outside Peak Width

The criterion is: f'_2 is outside the range $f_2 \pm F_M$ where $F_M > 0$ is a chosen constant. This criterion avoids the removal of genuine peaks embedded within the noise by excluding from the mask nearby traces which might also include some signal from the same peak. Hence, the parameter F_M is set according to the F_2 resolution of the experiment so that the range excludes much of the width of a peak in the F_2 direction.

3.3.4. Phase Balanced

This criterion balances the number of traces contributing to the mask that are (relatively close to being) in-phase with the trace at f_2 with the number of those that are out-of-phase, and is designed to retain small peaks embedded within the t_1 noise. The criterion is that the number of f_2' values in M for which the argument (phase) of the complex correlation is in the range $-\pi/2$ to $+\pi/2$ radians is the same as the number of f_2' values whose correlation argument lies outside this range.

Typically the Highly Correlated traces forming the masking spectrum are from a range of nearby traces (often just outside the immediate neighbourhood that is excluded by the Outside Peak Width criterion). The masking spectrum should ideally consist of only the t_1 noise signal, but if these nearby traces forming the mask cover the F_2 width of another peak, then without the Phase Balanced criterion, the mask may also contain much of this peak signal. (The peak signal in the traces may be a residual signal after separation of the noise and peak spectrum (Section 3.2), or may be from a small peak embedded in the noise that has insufficient intensity to be separated from the noise by the thresholding method.) When this peak signal in the masking spectrum coincides with a peak at a similar F_1 frequency in the original trace at f_2, the peak in the original trace may be incorrectly attenuated by the mask. Experiments on spectra containing sets of small adjacent peaks at similar F_1 frequencies confirmed this type of attenuation: *without* the Phase Balanced criterion, the intensity of some of these genuine peaks is reduced.

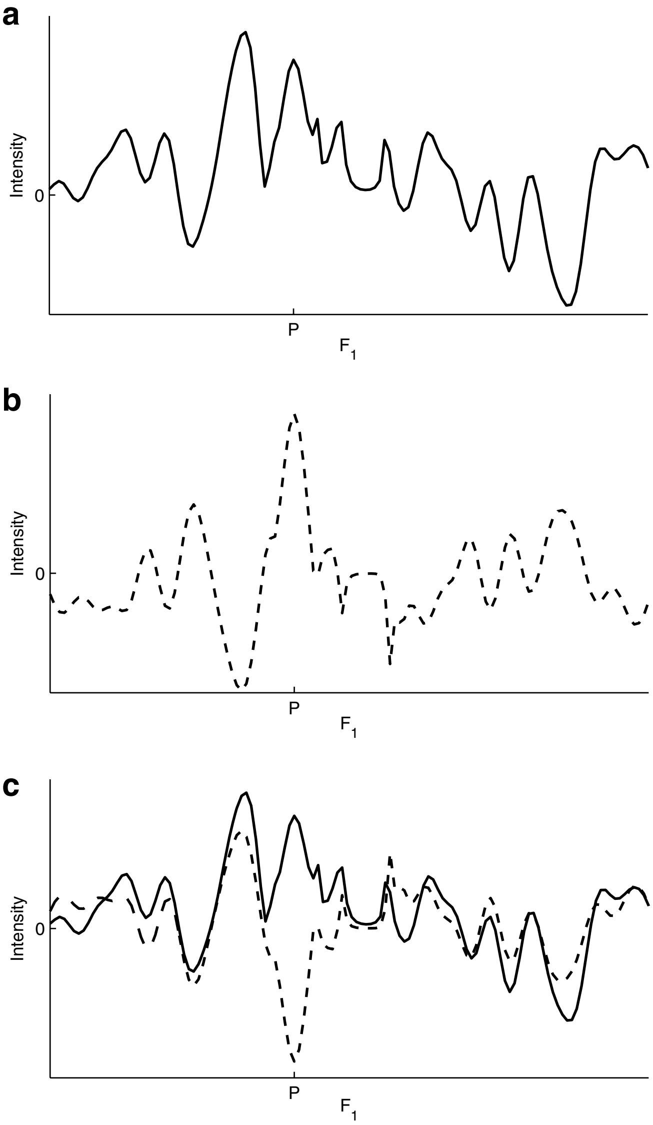

Fig. 3. An illustration of the effect of the Phase Balanced Criterion. (a) and (b) Two F_1 traces that might contribute to the masking spectrum; both contain a signal from a small peak at P. (c) The same two traces after the phases have been adjusted.

Fig. 3 illustrates the effect of the Phase Balanced criterion. The real part of two F_1 traces that might contribute to a masking trace are shown in (a) and (b). Trace (a) is approximately in-phase with the trace at f_2 (the latter is not shown), and hence the argument of its complex correlation with Φ_{f_2} is in the range $-\pi/2$ to $+\pi/2$ radians, while trace (b) is approximately out-of-phase and its correlation argument is outside the range. Both traces contain a signal from a peak at the F_1 frequency P. When these two traces are combined to form a masking trace (see Section 3.4 below), the phases of the traces are adjusted so that the t_1 noise in each trace has the same phase. Part (c) shows the two traces after the phase adjustment. The t_1 noise will be reinforced in the masking trace since the noise signals in the phase-adjusted traces are in-phase. However, the peak signals are now out-of-phase so they will cancel out one another when the masking trace is constructed, and this will avoid attenuation of genuine peaks at the same F_1 frequency as P.

3.4. Derivation of masking spectrum

Using the set of chosen f_2' values in the set M, an unnormalised masking trace, $\Phi_{f_2}^{\text{mask}*}(f_1)$, is derived using the formula:

$$\sum_{f_2' \in M} w\big(r_\rho(f_2, f_2')\big)\, \frac{\Phi_{f_2'}^{\text{noise}}(f_1)}{\text{median}(|\Phi_{f_2'}^{\text{noise}}|)}\, \mathrm{e}^{-\mathrm{i}\theta_\rho(f_2, f_2')} \tag{5}$$

The denominator, median($|\Phi_{f_2'}^{\text{noise}}|$), robustly normalises each trace before contribution to the mask to account for differences in amplitude. The factor $\mathrm{e}^{-\mathrm{i}\theta_\rho(f_2, f_2')}$ adjusts the phase of each f_2' trace to that of f_2. $w(\cdot)$ is a weighting function that scales the contribution of a trace to the mask depending on the correlation, and for the results in this paper was chosen to be the modulus of the correlation itself.

The masking trace is adjusted so that its median modulus (a measure of its signal amplitude) is the same as that of the noise trace, using:

$$\Phi_{f_2}^{\text{mask}}(f_1) = \frac{\text{median}\left(|\Phi_{f_2}^{\text{noise}}|\right)}{\text{median}\left(|\Phi_{f_2}^{\text{mask}*}|\right)} \Phi_{f_2}^{\text{mask}*}(f_1) \tag{6}$$

3.5. Construction of the denoised spectrum

The denoised spectrum is formed by subtracting the masking spectrum (formed by the masking traces derived in the previous step) from noise spectrum, and adding it back to the peak spectrum:

$$\Phi^{\text{denoised}} = \Phi^{\text{peak}} + \Phi^{\text{noise}} - \Phi^{\text{mask}} \tag{7}$$

3.6. Wavelet-based mask derivation

An alternative method of deriving the masking trace was investigated. In this method, each noise trace is separated using wavelet decomposition into signals at different scales. The mask derivation method described above was then applied to each level of wavelet decomposition separately: for a given noise trace signal at a given decomposition level, the mask was formed from the signals of other traces at the *same* level of decomposition. This method would accommodate changes in t_1 noise across the spectrum that were dependent on the 'scale' of the noise signal. However, it was found that the mask derived was usually no better than the method described above with the disadvantage that significantly more processing was required.

4. Experimental investigation

4.1. Objectives

To demonstrate the efficacy of the technique, the Correlated Trace Denoising algorithm was applied to two test cases.

The first test case was a simple spectrum in which a significant peak was embedded within a t_1 noise ridge. The sample was a solution of sucrose (250 mM) and glycine (2 mM). The peak from glycine is the correlation between the $C_\alpha H$ and C_α and coincides with a t_1 noise streak related to a high intensity peak in the sucrose spectrum. The relative concentrations of sucrose and glycine were chosen so that the amplitude of the glycine peak was similar to that of the t_1 noise ridge within which it was embedded.

The second test case was a more complicated spectrum obtained from a sample of a soft beverage. It was designed to test whether the algorithm was effective when the larger number of genuine peaks potentially obscured the correlation between t_1 noise traces that the algorithm leverages.

4.2. Algorithm implementation

The algorithm was implemented using MATLAB, version R14 SP2, from The MathWorks, Inc. The algorithm code is available for download from www.csl.gov.uk/downloads.

The spectrum is passed to the algorithm as MATLAB matrices and the denoised spectrum is returned in the same manner. In this way, the algorithm implementation is not specific to the data file format of any particular NMR acquisition and processing software: it is only necessary to upload the data into MATLAB to create matrices representing the complex spectrum.

For the experiments described below, Bruker Topspin software was used to acquire and process the spectra. Data were passed to the correlated trace denoising algorithm after Fourier transforms, baseline correction and phase correction processing had been applied in Bruker Topspin. Small MATLAB MEX functions (also available for download) were used to import the Bruker Topspin data files into MATLAB as matrices, and export the denoised spectrum back to Bruker Topspin data file format.

As an indication of the speed of the algorithm, each spectrum described here was denoised in approximately 30 s using a PC with a 1.86 GHz Intel Pentium 4 processor and 1 GB of RAM.

4.3. Algorithm parameters

The algorithm parameters used to control the derivation of the masking spectrum are a balance between removing as much t_1 noise as possible while retaining small genuine peaks embedded within t_1 noise ridges. The optimal balance, and therefore the parameter settings, depends partly on the relative importance to the spectroscopist of these two factors. (It is expected that the optimal parameters will also depend on the nature of the NMR experiment itself and the type of post-acquisition processing.)

For the results presented here, the following algorithm parameters were used:

- minimum correlation modulus (R_M): 0.5
- maximum number of traces (N_M): 3 × the number of data points in F_2 quarter-height peak width
- minimum distance from trace (F_M): 1 × the F_2 quarter-height peak width

These parameter settings had been found, by experimentation on similar test spectra, to give consistently good results. The F_2 quarter-height peak width was estimated from the largest peaks in the original spectrum.

4.4. Results

4.4.1. Sucrose–glycine mixture

Fig. 4 shows part of the spectra from the sucrose and glycine mixture before and after denoising. Before denoising, the intensity of the glycine peak is smaller than some nearby t_1 noise artefacts. After denoising, the significant reduction in noise surrounding the peak can be seen clearly, while the

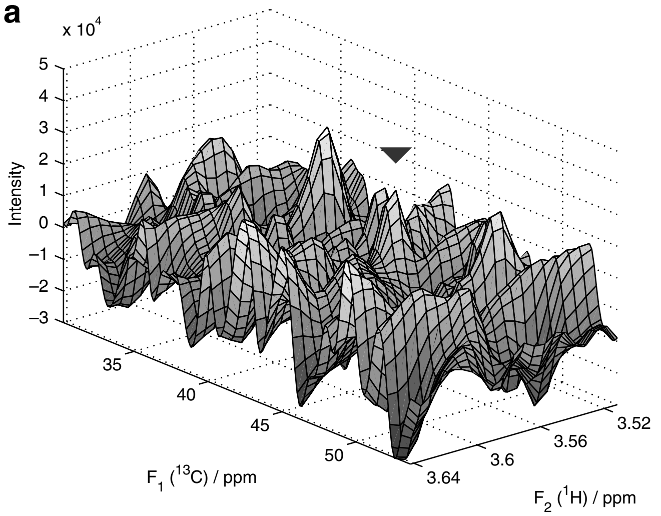

Fig. 4. A section of a phase-cycled HSQC spectrum from a mixture of sucrose (250 mM) and glycine (2 mM). (a) before and (b) after denoising. The arrow indicates the glycine peak at approximately 3.56 ppm ^{1}H, 44.2 ppm ^{13}C.

intensity and shape of the glycine peak is retained (the peak intensity is slightly larger in the denoised spectrum).

As a quantitative measure of the improvement after denoising, the signal-to-noise ratio (SNR) for the glycine peak was calculated. The method used was to estimate the root mean square of the noise in the F_1 trace containing the peak as a multiple of the interquartile range—a measure which is relatively robust to real peaks embedded within the trace—and assuming an approximately normal distribution of the noise. In the original spectrum, the SNR of the glycine peak was 2.67; after denoising, the SNR was 6.26. This improvement by a factor of 2.4 was largely as a result of the reduction in the t_1 noise rather than the slight increase in peak intensity after denoising.

4.4.2. Soft beverage

To assess the effect of the algorithm on the genuine peaks in the spectrum, the original and denoised spectra from the soft beverage sample were analysed by an experienced spectroscopist. It was found that 39 peaks were resolvable in the original spectra, and that all these peaks were also resolvable in the denoised spectrum. Specifically, the algorithm had removed no genuine peaks from the spectrum.

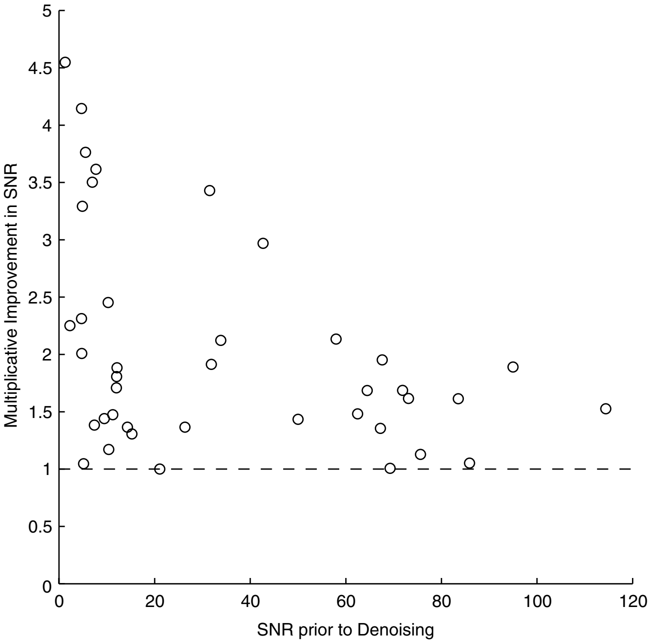

Fig. 5. The ratio of the SNR after denoising to the original SNR for each peak in the soft beverage spectra, plotted against the peak intensity in the original spectrum. The ratio is the multiplicative improvement in SNR after denoising: a ratio of 1 (the dashed horizontal line) indicates no change and values above the line are improvements.

The SNRs for these peaks were calculated using the method described above. Fig. 5 plots the ratio of each peak's SNR after denoising to its original SNR (i.e. the multiplicative improvement in SNR) against the peak's intensity in the original spectrum. Although the intensity of some peaks decreased after denoising (while others increased), the results show that the SNR of *all* peaks is improved—or, at the very least, unchanged—by the denoising algorithm.

In addition, a peak was identified in the denoised spectrum that could not be resolved by the spectroscopist in the original spectrum. This peak arises from benzoic acid and is the coupling between the carbon and the proton in the para position to the carboxylic acid group. Fig. 6 shows a section of spectrum containing this peak in both the original and denoised spectra. The improvement in intensity of the peak and the reduction in surrounding noise can be seen.

5. Conclusion

The rapid and routine deconvolution of complex mixtures requires the application of the most sensitive pulse sequences. The presence of t_1 noise in the data derived from these experiments will thus impede the growth of the

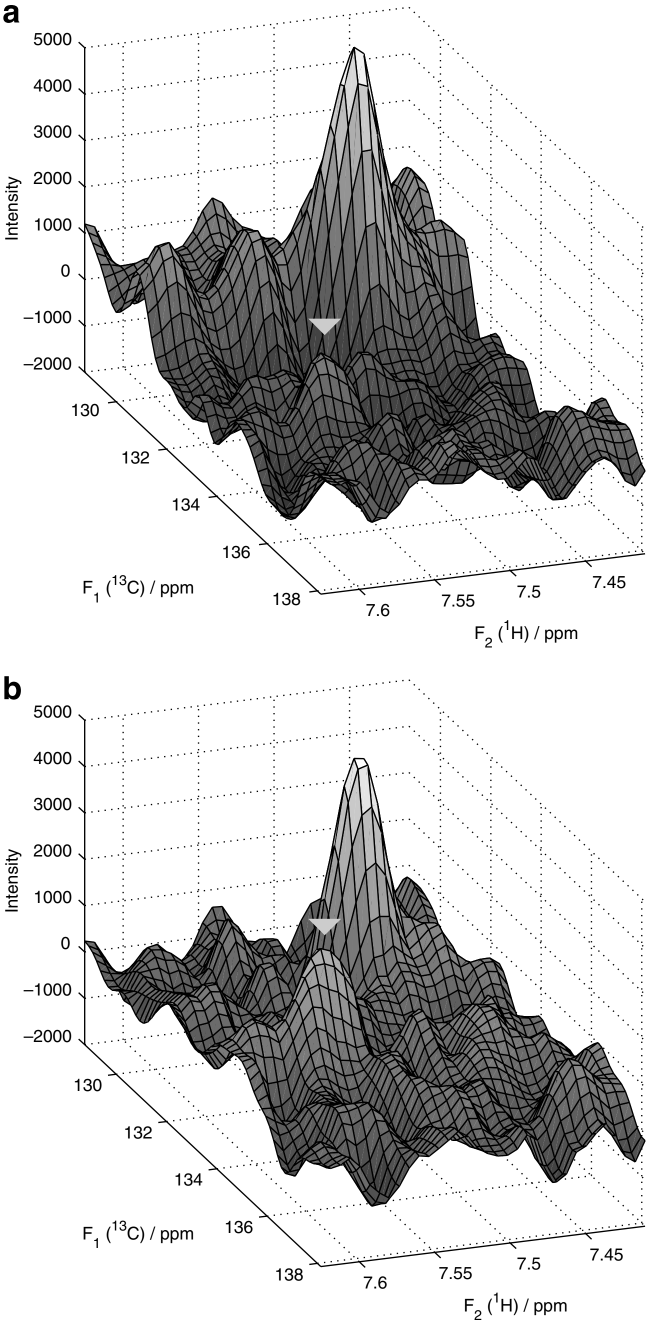

Fig. 6. A section of a phase-cycled HSQC spectrum from a sample of a soft beverage. (a) before and (b) after denoising. The arrow indicates a peak arising from benzoic acid at approximately 7.56 ppm ^{1}H, 134 ppm ^{13}C.

knowledge base in the relatively new field of metabolomics, due to the necessity for painstaking manual spectral assignment or the omission of useful data from subsequent analyses. This paper demonstrates that Correlated Trace Denoising can permit sensitive pulse sequences to be used for metabolomics studies by reducing the intensity of t_1 noise.

Both test cases show that the algorithm significantly improves the SNR of small genuine peaks embedded within t_1 noise, enabling the identification of low intensity correlations such as those arising from low concentration metabolites. At the same time, the algorithm maintains the shape and intensity of genuine peaks in the spectrum.

In the context of metabolomic studies, Correlated Trace Denoising has potential advantages compared to existing noise reduction techniques: it avoids the need for the strong unconvoluted reference signal used by reference Deconvolution and, unlike the technique based on the Cadzow procedure, does not need to estimate the number of genuine peaks.

A particular motivation for a denoising technique, such as Correlated Trace Denoising, that is suitable for metabolomic studies is to enable an automated procedure for the identification of the chemical shifts of all of the peaks in a 2D NMR spectrum or those selected by multivariate analysis [31]. These chemical shifts would then be used as database search queries to determine the identity of the compounds from which they arose. Such an approach would permit the composition of a mixture of metabolites to be automatically determined with minimum user intervention, but would be unable to distinguish high intensity t_1 noise from genuine peaks associated with low concentration metabolites. The application of automated chemical shift assignment tools in the biomolecular NMR field is well-established [32,33] and these methods are highly applicable to metabolomics data, in the absence of spectral artefacts.

Future work may investigate whether refinements to the wavelet-based method of constructing the masking spectrum—described above, but not used for the results in this paper—may lead to further improvements in t_1 noise reduction by allowing independent adaptation to the structure of the noise at different scales.

6. Experimental

6.1. Materials

All chemicals used were of a purity of ⩾ 99%. Deuterium oxide ($^{2}D_2O$) was supplied by Goss Scientific Instruments Ltd. (UK), 3-trimethylsilyl[2,2,3,3-D$_4$] propionic acid (TSP) was supplied by Avocado Research Chemicals Ltd., di-potassium hydrogen phosphate and di-hydrogen potassium phosphate were supplied by BDH Chemicals Ltd., and sucrose and glycine were supplied by Sigma Aldrich UK. Ultrapure water was provided from an Elga Option 2 water purifier.

6.2. Sample preparation

The sucrose and glycine solution was prepared by dilution of stock solutions of sucrose (1 M dissolved in $^{2}D_2O$), glycine (50 mM dissolved in $^{2}D_2O$) and phosphate buffer (1.0 M, 10 mM TSP, pH 7.0 dissolved in $^{2}D_2O$) with $^{2}D_2O$. The final concentration of the sucrose and glycine solution was 250 mM sucrose, 2 mM glycine, 100 mM phosphate buffer pH 7.0, 1 mM TSP.

Soft beverage samples were degassed prior to sample preparation. Degassing was carried out by sonication for 10 minutes. About 480 μL of the degassed solutions were added to labeled 5-mm NMR tubes containing 60 μL of buffer prepared in $^{2}D_{2}O$ (1.2 M, 10 mM TSP, pH 8.5) and 60 μL of sodium azide (10 mM dissolved in $^{2}D_{2}O$).

6.3. Methods

Sample temperature: 300 K.
Reference compound: internal standard of 3-trimethylsilyl[2,2,3,3-D_4] propionic acid, sodium salt (TSP) (0 ppm for both F_1 and F_2).
Spectrometer: Bruker ARX 500 NMR spectrometer tuned to ^{1}H signal at 500.1323506 MHz (^{13}C at 125.7678506 MHz).
2D Phase program:

- HSQC correlation via double INEPT transfer
- 90° pulse lengths: 9.2 μs for ^{1}H; 16.5 μs for ^{13}C
- carbon coupling constant: 145 Hz
- acquisition was recorded with decoupling of ^{13}C via composite pulse decoupling (CPD) with a garp sequence

Spectral width: F_2 6.666 kHz; F_1 20.12 kHz.
Acquisition data points: F_2 1536; F_1 384.
Window function: QSINE.
Baseline correction: automatic (‘quad’ mode).
Post-Fourier transform data points: F_2 2048; F_1 1024 (complex data points in both directions).
Phase correction: manual.

References

[1] J.K. Nicholson, J.C. Lindon, E. Holmes, ‘Metabonomics’: Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica (1999) 1181–1189.

[2] O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R.N. Trethewey, L. Willmitzer, Metabolite profiling for plant functional genomics, Nat. Biotechnol. 18 (2000) 1157–1161.

[3] M. Kime, R.G. Ratcliffe, B.C. Loughman, The application of ^{31}P nuclear magnetic resonance to higher plant tissue II. Detection of intracellular changes, J. Exp. Bot. 33 (1982) 670–681.

[4] J.K. Nicholson, M.J. Buckingham, P.J. Sadler, High-resolution ^{1}H NMR studies of vertebrate blood and plasma, Biochem. J. 211 (1983) 605–615.

[5] W.M.T. Fan, Metabolite profiling by one- and two-dimensional NMR analysis of complex mixtures, Prog. Nucl. Magn. Reson. Spectrosc. 28 (1996) 161–219.

[6] A.J. Charlton, W.H.H. Farrington, P. Brereton, Application of ^{1}H NMR and multivariate statistics for screening complex mixtures: Quality control and authenticity of instant coffee, J. Agric. Food Chem. (2002) 3098–3103.

[7] J.-P. Grivet, A.-M. Delort, J.-C. Portais, NMR and microbiology: from physiology to metabolomics, Biochimie 85 (2003) 823–840.

[8] N. Haroune, B. Combourieu, P. Besse, M. Sancelme, T. Reemtsma, A. Kloepfer, A. Diab, J.S. Knapp, S. Baumberg, A.-M. Delort, Benzothiazole degradation by *Rhodococcus pyridinovorans* strain PA: Evidence of a catechol 1,2-dioxygenase activity, Appl. Environ. Microbiol. 68 (12) (2002) 6114–6120.

[9] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Metabonomics: Metabolic processes studied by NMR spectroscopy of biofluids, Concepts Magn. Reson. 12 (5) (2000) 289–320.

[10] A.K. Mattoo, A.P. Sobolev, A. Neelam, R.K. Goyal, A.K. Handa, A.L. Segre, NMR spectroscopy based metabolite profiling of transgenic tomato fruit engineered to accumulate spermidine and spermine reveals enhanced anabolic and nitrogen–carbon interactions, Plant Physiol. 142 (4) (2006) 1759–1770.

[11] N. Shanaiah, M.A. Desilva, G.A.N. Gowda, M.A. Raftery, B.E. Hainline, D. Raftery, Class selection of amino acid metabolites in body fluids using chemical derivatization and their enhanced ^{13}C NMR, Proc. Natl. Acad. Sci. USA 104 (28) (2007) 11540–11544.

[12] M.R. Viant, E.S. Rosenblum, R.S. Tjeerdema, NMR-based metabolomics: A powerful approach for characterizing the effects of environmental stressors on organism health, Environ. Sci. Technol. 37 (2003) 4982–4989.

[13] J.L. Ward, J.M. Baker, M.H. Beale, Recent applications of NMR spectroscopy in plant metabolomics, FEBS J. 274 (2007) 1126–1131.

[14] C. Yang, A.D. Richardson, J.W. Smith, A. Osterman, Comparative metabolomics of breast cancer, Pacific Symposium on Biocomputing 12 (2007) 181–192.

[15] W.F. Reynolds, R.G. Enriquez, Choosing the best pulse sequences, acquisition parameters, postacquisition processing strategies, and probes for natural product elucidation by NMR spectroscopy, J. Nat. Prod. 65 (2002) 221–244.

[16] A.F. Mehlkopf, D. Korbee, T.T. A, R. Freeman, Sources of t_1 noise in two-dimensional NMR, J. Magn. Reson. 58 (1984) 315–323.

[17] W.F. Reynolds, R.G. Enriquez, Gradient-selected versus phase-cycled HMBC and HSQC: pros and cons, Magn. Reson. Chem. 39 (2001) 531–538.

[18] B.R. Seavey, E.A. Farr, W.M. Westler, J.L. Markley, A relational database for sequence-specific protein NMR data, J. Biomolecular NMR 1 (1991) 217–236, BMRB metabolomics database: <www.bmrb.wisc.edu/metabolomics> (accessed 2 September 2007).

[19] J.L. Markley, M.E. Anderson, Q. Cui, H.R. Eghbalnia, I.A. Lewis, A.D. Hegeman, J. Li, C.F. Schulte, M.R. Sussman, W.M. Westler, E.L. Ulrich, Z. Zolnai, New bioinformatics resources for metabolomics, Pacific Symposium on Biocomputing 12 (2007) 157–168.

[20] P. Lundberg, T. Vogel, A. Malusek, P.-O. Lundquist, L. Cohen, O. Dahlqvist, MDL - the magnetic resonance metabolomics database, ESMRMB, Basel, Switzerland, <mdl.imv.liu.se> (accessed 2 September 2007).

[21] D.S. Wishart, et al., HMDB: The human metabolome database, Nucleic Acids Res. 35 (Database Issue) (2007) D521–6, <www.hmdb.ca> (accessed 2 September 2007).

[22] R. Stones, A. Charlton, J. Godward, NMR manager: Metabolomics database application for interpreting NMR spectra, Central Science Laboratory, <bioinformatics.csl.gov.uk/posters/NMR.pdf> (accessed 2 September 2007).

[23] Madison Metabolomics Consortium Database, <mmcd.nmrfam.wisc.edu> (accessed 2 September 2007).

[24] A. Gibbs, G.A. Morris, A.G. Swanson, D. Cowburn, Suppression of t_1 noise in 2D NMR spectroscopy by reference deconvolution, J. Magn. Reson. 101 (1993) 351–356.

[25] T.J. Horne, G.A. Morris, Combined use of gradient-enhanced techniques and reference deconvolution for ultralow t_1 noise in 2D NMR spectroscopy, J. Magn. Reson., Ser A (1996) 246–252.

[26] G.A. Morris, H. Barjat, T.J. Horne, Reference deconvolution methods, Prog. Nucl. Magn. Reson. Spectrosc. 31 (1997) 197–257.

[27] C. Brissac, T.E. Malliavin, M.A. Delsuc, Use of the Cadzow procedure in 2D NMR for the reduction of t_1 noise, J. Biomol. NMR 6 (1995) 361–365.

[28] G. Bodenhausen, P.H. Bolton, Elimination of flip-angle effects in two-dimensional NMR spectroscopy. application to cyclic nucleotides, J. Magn. Reson. 39 (1980) 399–412.

[29] K.R. Metz, M.M. Lam, A.G. Webb, Reference deconvolution: A simple and effective method for resolution enhancement in nuclear magnetic resonance spectroscopy, Concepts Magn. Reson. 12 (1) (2000) 21–42.

[30] C. Perrin, B. Walczak, D.L. Massart, The use of wavelets for signal denoising in capillary electrophoresis, Anal. Chem. 73 (2001) 4903–4917.

[31] C. Antz, K.P. Neidig, K.H. R, A general Bayesian method for an automated signal class recognition in 2D NMR spectra combined with a multivariate discriminant analysis, J. Biomol. NMR 5 (1995) 287–296.

[32] D.E. Zimmerman, G.T. Montelione, Automated analysis of nuclear magnetic resonance assignments for proteins, Curr. Opin. Struct. Biol. 5 (1995) 664–673.

[33] R. Koradi, M. Billeter, M. Engeli, et al., Automated peak picking and peak integration in macromolecular NMR spectra using AUTOPSY, J. Magn. Reson. 135 (1998) 288–297.