Available online at www.sciencedirect.com
ScienceDirect
Journal of Magnetic Resonance 183 (2006) 145–151
JMR Journal of Magnetic Resonance
www.elsevier.com/locate/jmr

Communication

A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data

J. Carlos Cobas [a,*], Michael A. Bernstein [b], Manuel Martín-Pastor [c], Pablo García Tahoces [d]

[a] *MESTRELAB RESEARCH, Xosé Pasín, 6-5C, 15706, Santiago de Compostela, Spain*
[b] *AstraZeneca R&D Charnwood, Bakewell Rd, Loughborough, LE11 5RH, UK*
[c] *Laboratorio Integral de Dinámica e Estructura de Biomoléculas José R. Carracido, Unidade de Resonancia Magnética, Edificio Cactus, RIAIDT, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain*
[d] *Departamento de Electrónica e Computación, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain*

Received 25 May 2006; revised 5 July 2006
Available online 7 August 2006

Abstract

A new procedure for automatic baseline correction of NMR data sets is presented. It is based on an improved automatic recognition of signal-free regions that uses a Continuous Wavelet transform derivative calculation, followed by a baseline modelling procedure based on the Whittaker smoother algorithm. The method has been proven to automatically flatten 1D and 2D NMR spectra with large baseline distortions arising from different sources, is tolerant to low signal-to-noise ratio spectra, and to signals of varying widths in a single spectrum. Even though this procedure has so far only been applied to NMR spectra, we believe it to also be applicable to other spectroscopies having relatively narrow peaks (e.g., mass spectrometry), and potentially to those with broad peaks (e.g., near infrared or ultraviolet).

Keywords: NMR; Automatic baseline correction; Derivative; Wavelet transform; Penalized least squares; Whittaker smoother

1. Introduction

Baseline distortions can arise from a number of hardware and processing sources and have long been a major problem in FT NMR [1]. Distorted baselines in spectra will result in incorrect integration values-information which can be central to many NMR experiments, e.g., qNMR [2] and the accurate quantification of 2D NOESY spectral cross-peaks. Multicomponent spectra recorded in an LC NMR experiment or with biofluids in the generation of metabonomics data also demand high-quality spectra prior to statistical analysis [3,4]. In addition, peak picking routines will also be adversely affected by weak signals in spectra presenting significant baseline roll, and may not be recognized because the baseline distortions can be significantly larger than the peak intensities. This will be cause for concern with low-concentration sample spectra.

The reasons for baseline distortions are diverse [5] and in many cases they can be removed by adjusting acquisition parameters or Backward Linear Prediction [6]. Modern spectrometer hardware uses oversampling and digital signal processing to improve the baseline [7], but some undesirable broad signals arise from real sources (see below). Thus, a more general solution would employ an efficient post processing baseline correction in the frequency domain. In fact, this is the most common approach found in NMR literature [8–11]. We describe here a new procedure for automatic baseline correction of frequency-domain NMR data sets which we show to be highly effective on 1D and 2D spectral datasets, and preserves the range of component line widths that are present in the sample.

* Corresponding author. Fax: +34 981941079.
E-mail address: carlos@mestrec.com (J. Carlos Cobas).

1090-7807/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2006.07.013

The algorithm consists of two independent processes:

1. Automatic baseline recognition (signal-free regions) based on a Continuous Wavelet Derivative transform (CWT) followed by iterative threshold detection in the Power mode domain.
2. A baseline modelling procedure based on the Whittaker smoother algorithm.

Overall, the algorithm has been designed to afford perfect baselines in NMR data sets spanning a wide range of possible baseline topographies and signal-to-noise ratio (SNR) conditions. In most cases, it can be applied successfully without any operator interaction, but further versatility is assured by two parameters that could be adjusted to guarantee an optimum outcome. The procedure can therefore be "tuned" to achieve more accurate baseline recognition of signal-free regions, or to increase smoothness at the expense of spectral fidelity, or vice versa.

1.1. Automatic baseline recognition

Baseline recognition of signal-free regions is performed using an improved version of the Dietrich method [8]. In short, Dietrich's approach for automatic peak recognition consists of calculation of the first numeric derivative of the spectrum to eliminate baseline distortions, followed by conversion to a power spectrum to generate absorptive peaks. From this point, an iterative thresholding algorithm is applied by first defining an initial threshold as the mean plus three times the standard deviation using all the points in the spectrum. Next, a new threshold is calculated in the same manner but this time using only the spectral data points below the first threshold. This iterative process is repeated until no new points exceed the final threshold during an iterative step.

The result of this stage is a binary mask w_i which contains "false" (or zero) values if a point belongs to real peaks and "true" (or one) values otherwise (see [8]). As this mask might contain undesired spikes, a 1D-erosion filter is then applied. This examines each of the "true" points in the mask; if the majority of its neighbors are "false" the point is set to "false", causing the "true" region to shrink. This binary mask will be used in the second stage of this algorithm (baseline modelling) as the vector of weights which the Whittaker smoother algorithm uses for interpolation.

The pivotal operation for the baseline recognition phase is the numeric derivative calculation. Dietrich used a standard numeric derivative algorithm (where each spectral data point is replaced by the difference between one point and the next one). However, it is well known that this calculation has a major drawback in increasing the noise level. In order to improve the SNR of the derivative calculation via conventional numerical differentiation, noise reduction is usually performed before calculating the derivative: Dietrich used a moving average filter. This approach only performs correctly when the SNR is good. In addition, peak heights and widths are usually 'washed out' by adjacent averaging making the algorithm ineffective in spectra containing a combination of broad and sharp peaks. More advanced smoothing routines such as the well-known Savitzky–Golay [12] method has been commonly used, but significantly increases the computational burden. This extra computational effort would not be an issue with most routine 1D NMR spectra but can clearly be seriously limiting in the case of 2D- or higher order multidimensional spectra.

In this work, we propose the utilization of a novel method for the derivative calculation based on the Continuous Wavelet transform (CWT). It has been recently shown that, compared to other methods (such as the conventional numerical differentiation, the Fourier transform method or the Savitzky–Golay method), the proposed CWT method is more efficient in improving the SNR of the derivatives of noisy signals [13,14]. Furthermore, it is very fast and simple to implement as it involves a simple convolution in which both the smoothing and derivative calculations are combined in one single step.

The CWT of a signal $f(x)$ can be represented as follows:

$$Wf(a,b) = \int_{-\infty}^{+\infty} f(x)\psi^*_{a,b}(x)\mathrm{d}x, \tag{1}$$

where the asterisk represents the complex conjugate and $f(x)$ and $\psi_{a,b}(x)$ both have to belong to $L^2(R)$, being $L^2(R)$ the Hilbert space of measurable square-integrable one-dimensional functions, i.e., the space of signals of finite energy, as is the case for NMR.

$\psi_{a,b}(x)$ can be obtained by dilations and translations of a single function $\psi(x)$ called the mother wavelet

$$\psi_{a,b}(x) = \frac{1}{\sqrt{a}}\psi\left(\frac{x-b}{a}\right), \tag{2}$$

where $a \in \Re$, $a > 0$ is the parameter for dilation and $b \in \Re$ is the parameter for translation ($\Re$ denotes real number). Substituting Eq. (2) into Eq. (1), we obtain

$$\begin{aligned} Wf(a,b) &= \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} f(x)\psi^*\left(\frac{x-b}{a}\right)\mathrm{d}x \\ &= f(b) \otimes \psi^*_a(b), \end{aligned} \tag{3}$$

where $\otimes$ denotes the convolution of both functions.

The mother wavelet can be obtained as $\mathrm{d}^n/\mathrm{d}x^n h(x)$, where $h(x)$ is a smoothing function. If the smoothing function is symmetric with regard to a point $x = p$, the position of the maximum peak lies in $x = p$ and its derivatives are indefinitely derivable, being the result of this convolution the derivative of $f(x)$ as it has been demonstrated by Nie et al. [14]. Examples of smoothing functions that have those properties are, among others, the Haar function or the Gaussian function.

When computers are employed for computation, the signal to be analyzed is discrete. This make necessary to use the discrete form of Eq. (3)

$$\mathrm{CWT}[f(a,iT_s)] = T_s\frac{1}{\sqrt{a}}\sum_n f(nT_s)\Psi^*\left(\frac{(n-i)T_s}{a}\right), \tag{4}$$

where the asterisk represents the complex conjugate, a is a variable used to control the dilation called *scale* parameter, i and n are indexes of the data point of $f(nT_s)$, T_s corresponds to the sampling interval, and $\Psi(t)$ is defined by the mother wavelet function. We have selected the first and second derivative of a Gaussian and Haar function. In practice, we have not found major differences between these functions, so in order to simplify our study, we have selected a wavelet Haar function to compute the derivatives. This function is defined as follows:

$$\Psi(t) = \begin{cases} 1 & 0 \leqslant t < \frac{1}{2}, \\ -1 & \frac{1}{2} \leqslant t < 1, \\ 0 & \text{otherwise.} \end{cases}$$

The CWT only affords approximate derivatives, which are modulated by parameter a, which is the first adjustable parameter of our algorithm. Increasing a reduces noise and therefore improves SNR (Fig. 1a) but with a broadening of the signals (Fig. 1b).

The optimum values of the scale factor for different SNR values have been determined in Ref. [14]. In practice, we have found that the scaling factor may be calculated as a proportional function of the broadest signal in the spectrum. If the scaling factor is set too small, broad peaks will be considered as a baseline distortion rather than a real peak, and removed.

1.2. Baseline modelling

The next step in the algorithm is the building of a baseline model from the previously-detected baseline points. An ideal baseline model should match the baseline distortion (fidelity) whilst being smooth. Different models have been used in the past, such as polynomial functions [15], cubic splines [16], Bernstein polynomials [9] or linear segments [9]. Obviously, the simplest method is the linear segments model, but by definition this is not a smooth function. Polynomial functions are smooth but, in theory, NMR baseline distortions do not match the shape of a polynomial and, in particular, high order polynomials tend to oscillate (Runge's phenomenon). This problem can be circumvented by using spline curves, which are piecewise polynomials. However the spline is forced through the detected baseline points so that, for spectra with a low SNR, the fit can be poor and severely distorted. To circumvent all these problems, in this paper, we propose use of the so-called Whittaker smoother algorithm (WS).

WS was introduced more than 80 years ago [17] and recently revisited by Eilers [18]. This algorithm attempts a balanced combination of the two conflicting goals previously mentioned: (1) *Fidelity* to the data (i.e., the function may stay close enough to the spectral baseline) and (2) *smoothness*.

Fidelity to the data can be expressed as:

$$S = \sum_i (y_i - z_i)^2, \tag{5}$$

where z_i is the desired smoothed vector and y_i is the original raw spectrum.

Smoothness can be expressed, to a first order smoother, by the squared differences between neighbors:

$$R = \sum_i (z_i - z_{i-1})^2 = \sum_i (\Delta z_i)^2. \tag{6}$$

A balanced combination of the two goals is the sum

$$Q = S + \lambda R, \tag{7}$$

where λ is a user-defined parameter. Thus, we have a standard sum of squares problem with penalization, where the goal is to find the series z_i which minimizes Q. Large values of λ will make the R term higher as the smoother effect will be larger, but at the cost of a deterioration of the fit to the data (which is the penalization concept).

From partial derivatives $\frac{\partial Q}{\partial z_i} = 0$ we get a linear system of equations which can be easily solved:

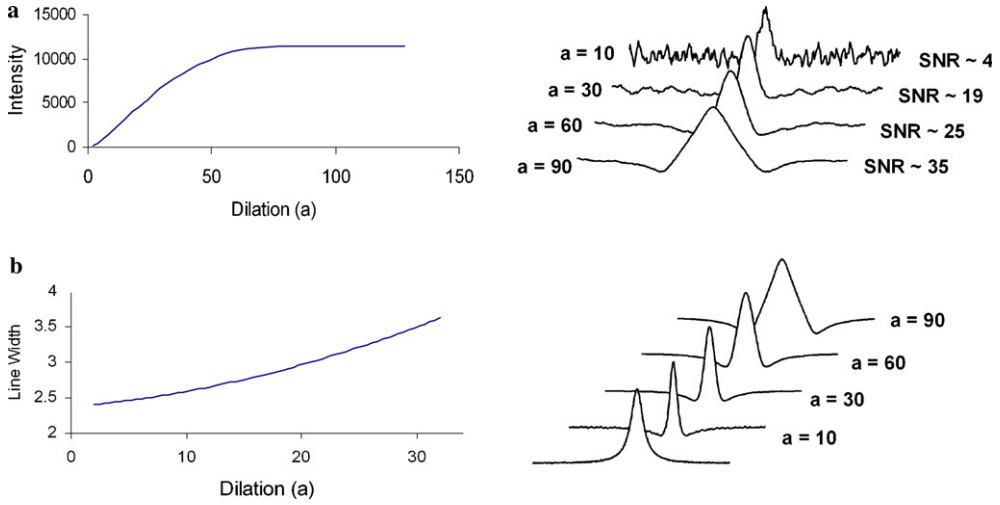

Fig. 1. The influence of dilation, a, on the SNR (a) and line width (b) of the approximate first derivative evaluated by CWT method. Line widths and peaks heights were calculated from the imaginary part of the CWT-derivative spectrum.

$$z = (I + \lambda D^{t}D)/y, \tag{8}$$

where D is the derivative of the identity matrix I. For example, if the number of points in the spectrum is 4, then D would be

$$D = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

At this point we can introduce the binary mask calculated in the first stage which we use as a vector of weights w_i so that at the positions where w_i (and hence y_i) is zero (i.e., peaks positions), z is automatically and smoothly interpolated. The vector of weights is therefore introduced in the fidelity term:

$$S = \sum_i w_i (y_i - z_i)^2. \tag{9}$$

So that the system of equations changes to:

$$(W + \lambda D^{t}D)z = Wy. \tag{10}$$

This is a linear system of equations which can be efficiently solved by using sparse coding. Computation times for standard 1D spectra (e.g., 32 Kb) takes less than 1 s using up-to-date personal computers.

In Eq. (10), λ represents the second adjustable parameter in our algorithm.

2. Results

To evaluate the performance of the algorithm, we have selected spectra having baseline distortions caused by different means.

In Fig. 2a, we show a ^{13}C spectrum with a severe baseline rolling due to improper adjustment of the pre-acquisition delay. This experimental baseline distortion was artificially exacerbated by altering the very first points in the FID. The application of our algorithm (baseline correction parameters were $\lambda = 1000$, $a = 50$) allows us to obtain a spectrum with a flat baseline (Fig. 2b). Note in Fig. 2a that the baseline model (light curve) matches the baseline shape perfectly. In all these examples the vertical scale has been greatly expanded so the baseline quality is clear.

The ^{13}C spectrum in Fig. 3a is of a dilute quinine solution, recorded on a modern Cryo-Probe. This represents the "real world" use of this very sensitive hardware. With simple FT, the baseline is severely distorted, and extracting chemical shift values is difficult in spite of the very adequate SNR. We commonly observe these baseline distortions, and can partially compensate using backward linear prediction of up to 256 data points. But this approach is undesirable if SNR is limiting, and a general solution that corrects the baseline would be of more general utility. The spectrum in Fig. 3c shows the same data after application of our baseline correction algorithm: it is clear that the baseline modelling is excellent (Fig. 3b).

In Fig. 4a a ^{1}H spectrum presenting a broad peak (at low field) in the presence of sharp signals is used to assess the efficiency of the algorithm under these conditions. If the value of a was set too small, we would risk identifying the broad signal as part of the baseline, and as a consequence this broad signal could disappear after carrying out the baseline correction. Fig. 3a, the value $a = 80$ was employed, leading to the wide signal being identified correctly, and therefore, preserved in the resulting spectrum (Fig. 4b).

^{19}F NMR spectra are typically recorded with large spectral windows at high frequency, and present a number of technical challenges (see above). A further difficulty lies in contamination of the spectrum with broad signals derived from materials used in the probe construction. An example of a ^{19}F NMR spectrum showing a poor baseline and broad, contaminating peaks is depicted in Fig. 5a. Application of our algorithm (baseline correction parameters: $\lambda = 50000$, $a = 80$) allows us to obtain a perfectly flat

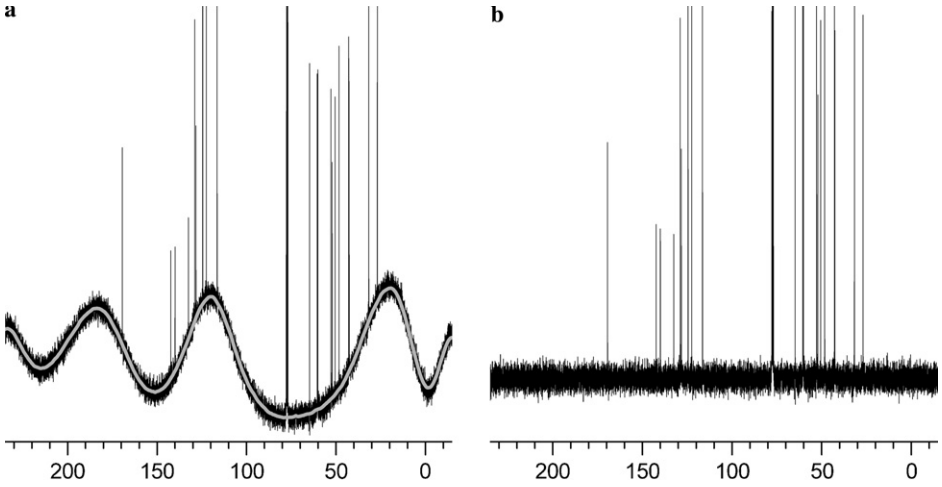

Fig. 2. ^{1}H-decoupled ^{13}C spectrum of strychnine in $CDCl_3$ acquired on a Varian Inity Inova 400 NMR spectrometer using a sweep width of 25,157 Hz. The FID was processed with an exponential function (line broadening = 1 Hz) and a FT of 64 Kb (a) and corrected spectrum (b) after applying the proposed algorithm ($\lambda = 1000$, $a = 50$).

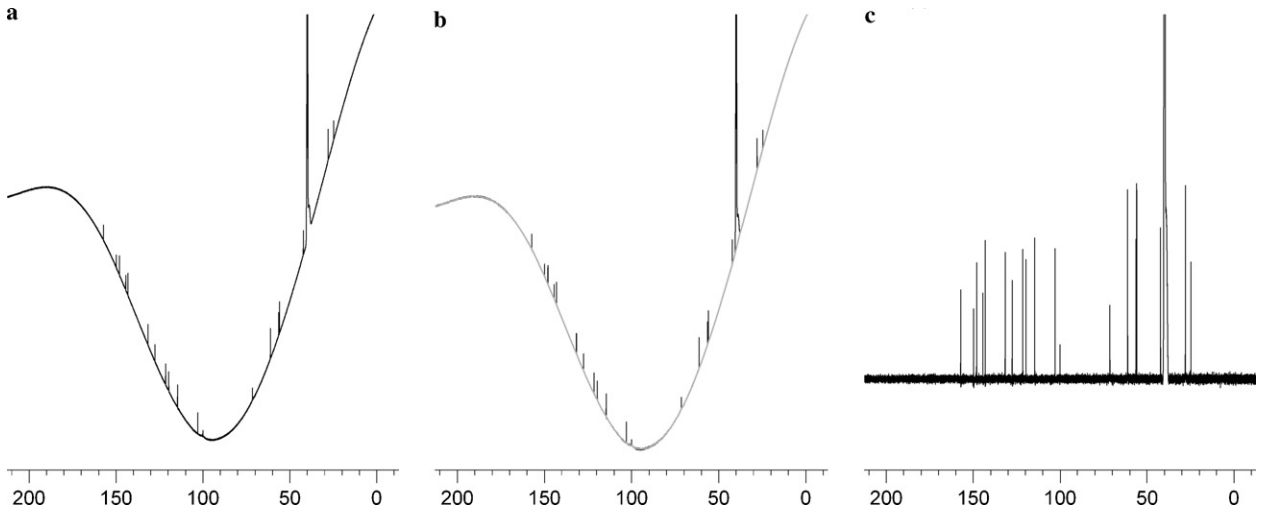

Fig. 3. ^{1}H-decoupled ^{13}C spectrum of quinine in DMSO-d_6 (1.0 mg/0.13 mL) acquired using a Bruker DRX 600 MHz NMR spectrometer fitted with a 3 mm Cryo-Probe. The sweep width was 35,971 Hz. The FID was processed with an exponential function (line broadening = 1 Hz) and a FT of 64 kb (a) and corrected spectrum (c) after applying the proposed algorithm ($\lambda = 10{,}000$, $a = 200$). Note that the baseline model (b) is excellent.

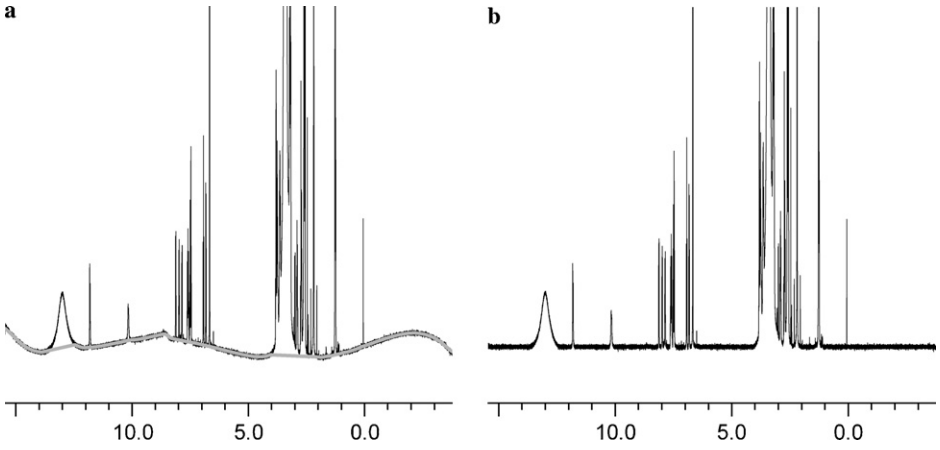

Fig. 4. ^{1}H NMR spectrum of a typical chemist's qNMR sample in DMSO-d_6, containing mesaconic acid as an internal standard. The 500 MHz spectrum was recorded on a Varian UnityInova 500 using a sweep width of 9612 Hz. Processing used an exponential line broadening of 0.3 Hz and FT with 32 kb. ($\lambda = 10{,}000$, $a = 80$).

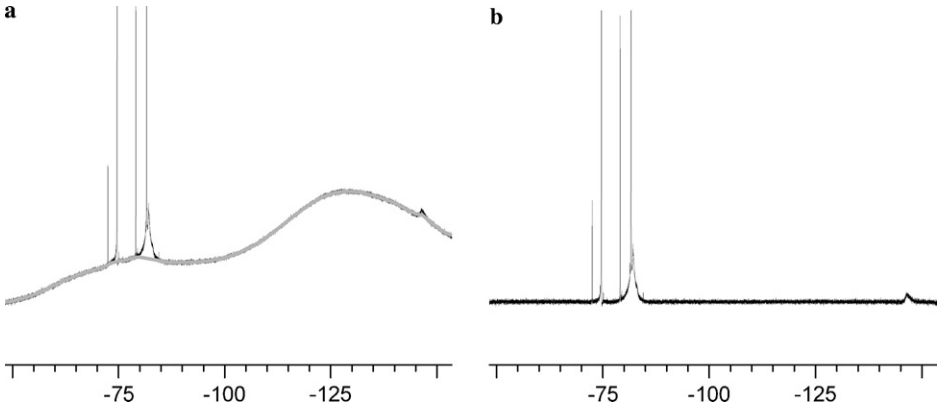

Fig. 5. ^{19}F NMR spectrum of an impure chemist's sample containing a 2,6-disubstituted aromatic ring recorded on a Varian UnityInova 400 NMR spectrometer using a sweep-width of 39,604 Hz. The spectrum shows a bad, rolling baseline and broad, contaminating peaks (a) and the corrected spectrum (b) after baseline correction ($\lambda = 50{,}000$, $a = 80$).

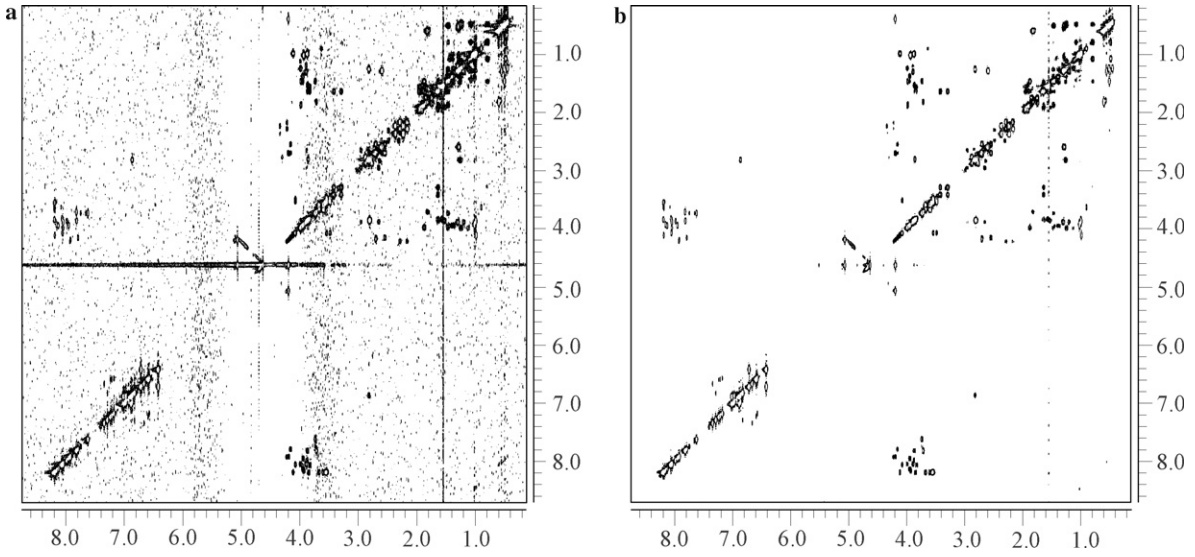

Fig. 6. 2D WET-g-COSY spectrum of a sample of human obestatine (a) Spectral dimensions after double FT are 2048×2048 (see text for more details) and (b) resulting baseline corrected spectrum ($\lambda = 100$, $a = 2$).

baseline (Fig. 5b). Note that the broad peak at high-field end (probably an impurity) is maintained in the corrected spectrum. By simply decreasing parameter a, this peak can be selectively removed if required.

The computational performance of this algorithm is well represented by its application to the correction of 2D NMR data sets. Fig. 6 exemplifies the higher quality that can be achieved with 2D data sets by the use of the new base-line correction algorithm. In both spectra of Fig. 6 the data set is the same 2D WET-g-COSY spectrum of a sample of human obestatine, a 23 amino acids peptide, dissolved in H_2O/D_2O 90:10 and PBS buffer. The WET module provides a high degree of solvent suppression of the strong water resonance. Besides, pulse-field-gradients are used to select only one of the two possible echo/antiecho coherence pathways during the evolution of the t1 dimension, which results in a COSY magnitude spectrum. The spectrum was acquired with 2048×256 data points. The spectrum was processed with a high-pass filter along the t2 dimension to suppress the residual solvent signal. A sinebell apodization function was applied in both dimensions to give, after double FT, a final matrix of 2048×2048 real points which was represented in the magnitude mode. The two spectra of Fig. 6 correspond to the same data set processed in this way. The only difference is that the spectrum of Fig. 6 right was subsequently treated with the new base-line correction along all rows and columns, a process that took ca. 2.3 s in a Pentium IV computer running at 1.6 GHz under Windows XP.

While the two spectra of Fig. 6 are represented at the same contour level and provide eventually the same intensity for all the cross-peaks, the spectrum of Fig. 6 right shows an evident reduction of t1-noise occurring at $\sim$1.3 ppm, and the suppression of the artefacts caused by the residual solvent signal at $\sim$4.7 ppm along the F2 dimension that were not completely removed by the high-pass filter. Clearly, the use of our new base-line correction algorithm in the spectrum of Fig. 6 right enhances the possibilities for the detection of small cross peaks close to the t1-noise or to the residual solvent line.

3. Conclusions

While spectrometer electronics and probe construction will continue to work towards high-quality, sensitive spectra, baseline distortions still are common in NMR spectra. These are a nuisance at the least, but in many cases the distortions must be eliminated for the data to be useful. Existing approaches have limitations in their effectiveness and universal application. We have shown that the CWT algorithm in combination with the penalized least squares (Whittaker) effectively describes baseline regions of a spectrum without loss in SNR or smoothing artifacts.

Together this affords an efficient process for eliminating undesirable baseline effects which we have demonstrated here for a sample of 1D and a 2D spectrum. A further communication will elaborate on this utility and the degree to which "default" parameters can be used for unsupervised processing. In more demanding situations these parameters can be modified to best afford the desired effect. We believe that the approach to baseline correction described here is well suited to a number of NMR spectral conditions, but will also find utility in light- and vibrational spectroscopies.

Acknowledgment

The authors would like to thank Dr. Paul Eilers for his support on the Whittaker Smoother.

References

[1] J.C. Hoch, A.S. Stern, NMR Data Processing, John Wiley, Chichester, 1998.

[2] F.P. Guido, U.J. Birgit, C.L. David, Quantitative 1H NMR: development and potential of a method for natural products análisis, J. Nat. Prod. 68 (2005) 133–149.
[3] I. Pelczer, Curr. Opin. Drug Discov. Devel. 8 (2005) 127–133.
[4] J.C. Lindon, E. Holmes, J.K. Nicholson, So what's the deal with metabonomics? Anal. Chem. 74 (2003) 384A–391A.
[5] C. Tang, An analysis of baseline distortion and offset in NMR spectra, J. Magn. Reson. (1994) 232–240.
[6] D. Marion, A. Bax, Baseline correction of 2D FT NMR spectra using a simple linear prediction extrapolation of the time-domain data, J. Magn. Reson. 83 (1989) 205.
[7] D. Moskau, Application of real time digital filters in NMR spectroscopy, Concepts Magn. Reson. 15 (2002) 164–176.
[8] W. Dietrich, C.H. Rüdel, M. Neumann, Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra, J. Magn. Reson. 91 (1991) 1–11.
[9] D.E. Brown, Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials, J. Magn. Reson. 114 (1995) 268–270.
[10] S. Golotvin, A. Williams, Improved baseline recognition and modelling of FT NMR spectra, J. Magn. Reson. 146 (2000) 122–125.
[11] G. Schulze, A. Jirasek, M.M.L. Yu, A. Lim, R.F.B. Turner, W.B. Michael, Investigation of selected baseline removal techniques as candidates for automated implementation, App. Spectrosc. 59 (2005) 545–574.
[12] A. Savitzky, M.J.E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. 36 (1964) 1627–1639.
[13] A.K.M. Leung, F.T. Chau, J.B. Gao, Wavelet Transform: A Method for Derivative Calculation in Analytical Chemistry, Anal. Chem. 70 (1998) 5222–5229.
[14] X. Shao, C. Pang, Q. Su, A novel method to calculate the approximate derivative photoacoustic spectrum using continuous wavelet transform Fresenius, J. Anal. Chem. 367 (2000) 525–529.
[15] P. Güntert, K. Wüthrich, FLATT—a new procedure for high-quality baseline correction of multidimensional NMR spectra, J. Magn. Reson. 96 (1991) 403–407.
[16] Z. Zolnai, S. Macura, J.L. Markley, Spline method for correcting baseplane distortions in two-dimensional NMR spectra, J. Magn. Reson. 82 (1989) 496–504.
[17] E.T. Whittaker, On new method of graduation, Proc. Edinburgh Math. Soc. 41 (1923) 63–75.
[18] P.H. Eilers, A perfect smoother, Anal. Chem. 75 (2003) 3631–3636.