Journal of Biomolecular NMR, 5 (1995) 147–153
ESCOM

J-Bio NMR 227

A model-free algorithm for the removal of baseline artifacts

Mark S. Friedrichs

Macromolecular NMR Department, Bristol-Myers Squibb Pharmaceutical Research Institute, P.O. Box 4000, Princeton, NJ 08543-4000, U.S.A.

Received 14 July 1994
Accepted 12 September 1994

Keywords: Baseline-flattening algorithm; Noise

Summary

A novel algorithm for removing baseline distortions in NMR spectra is presented. The algorithm approximates the baseline as the median of the noise extrema. Consequently, the method does not require that NMR peaks be discriminated from noise peaks. In addition, no assumptions regarding the source or functional form of the distortion are made. The algorithm is shown to remove the baseline artifacts present in a particularly distorted NOESY spectrum and to reveal peaks which had been obscured by the artifacts. The parameters and spectral characteristics (signal-to-noise ratio, NMR peak density, peak linewidths) governing the resolution of the calculated baselines are also explored.

Introduction

Baseline distortions often interfere with the interpretation and quantification of NMR data: weak cross peaks may be obscured, and the measurement and calibration of cross-peak volumes can be problematic. In addition, the design of robust automated peak-picking algorithms becomes much more difficult in the presence of baseline artifacts with arbitrary magnitudes and shapes.

Numerous approaches have been developed to correct baseline defects during post-acquisition processing. In some cases the source of the distortions is known, and the artifacts can be avoided or removed during processing. For example, the delayed acquisition of FIDs can lead to errors in the first few datapoints, which in turn introduces baseline roll in the frequency domain. This distortion can be avoided by using backward linear prediction to estimate the first few datapoints (Otting et al., 1986; Marion and Bax, 1989). Often, however, the sources of the artifacts are unknown, and the shapes of the distortions are unpredictable. For these cases, the canonical approach to removing the artifacts is to fit a predefined functional form to a subset of points which lie on or near the baseline. Estimates of the baseline values for the remaining points are then obtained by interpolation and extrapolation. The approximating function serves as a model of the distorted baseline and, when subtracted from the spectrum, yields a flattened baseline.

For the strategy outlined above to be successful, the points to be fit must be reliably identified as baseline points and not NMR signals. From an algorithmic standpoint, the simplest method for discriminating between the two types of points is to have the user designate a set of known baseline points (Barsukov and Arseniev, 1987; Zolnai et al., 1989). While straightforward, this method has the important disadvantage that the selected points are necessarily outside the spectral regions of interest. As a consequence, the interpolated baseline values are often inaccurate, and the distortions lying in these important areas may be only incompletely removed or even made worse. Automated techniques have been developed to discriminate between the baseline and peak points (Pearson, 1977; Dietrich et al., 1991; Güntert and Wüthrich, 1992; Rouh et al., 1993). However, this is an inherently difficult task, given the wide range of possible baseline distortions and peak shapes. While these methods typically work well, examples where the results are unacceptable are readily found.

Once the baseline points have been identified, they are fit to a predefined functional form. Commonly used functions include cubic splines (Zolnai et al., 1989; Rouh et al., 1993), sectionally linear functions (Saffrich et al., 1993), polynomials (Dietrich and Gerhards, 1981; Daubenfeld et al., 1985; Dietrich et al., 1991) and sums of orthogonal polynomials (Pearson, 1977; Henrichs et al., 1986; Barsukov and Arseniev, 1987; Güntert and

Wüthrich, 1992; Rouh et al., 1993). A major drawback to using a fixed function or set of basis functions is that the distortions can vary arbitrarily from spectrum to spectrum and for multidimensional spectra from row to row; this makes the selection of an optimal basis set of functions difficult. If the number of basis functions used in the fit is too small, the range of distortions which can be modeled will be limited. As a result, the fit to the selected baseline points may be poor and the interpolation of the remaining points even poorer. On the other hand, the use of a large number of basis functions will yield an accurate fit, but may also introduce spurious baseline structure in regions containing only a small number of fitted points.

In this report, a novel approach for removing baseline distortions is presented. The algorithm does not require the discrimination of NMR peaks from noise extrema. Moreover, no assumptions are made regarding the source or functional form of the distortion. Thus, the algorithm presented here is 'model-free'. An overview of the method is given, followed by a discussion of its implementation and the critical issues governing its utility and limitations.

Method

The algorithm circumvents the problem of distinguishing between NMR and noise peaks by exploiting the ubiquity of noise in experimental spectra. For a spectrum of *only* noise, the baseline can be defined by the criterion that the area above the line is equal to the area below it, i.e., the average value of the spectrum. This is obviously not a viable technique in the presence of NMR resonances, since the area under these peaks would usually dominate the area under the noise peaks. Instead, as a proxy measure for the area, the number of local maxima and minima is used. Hence, the baseline is defined here by the criterion that the number of local extrema above the line equals the number of local extrema below the line. Employing this measure, an NMR peak is just another local maximum or minimum, and therefore the classification of peaks into noise or NMR resonances is unnecessary. From a statistical perspective, the baseline is approximated by the median of the noise peak heights instead of the average area. This approximation is valid provided the density of NMR peaks is small relative to the density of noise peaks; this point is discussed further below.

The drawback of forcing the calculated baseline to follow a user-specified form under the standard approach is eliminated here by using a free-form method. The local baseline structure is tracked by finding the median of the extrema within a small region or window centered about each point in the spectrum. An extremum is taken to be a point i such that the intensity at i, I(i), is either strictly greater or less than both I(i − 1) and I(i + 1). The approach is illustrated in Fig. 1: the window size, W, is 34 points and each window has typically 18–24 extrema;

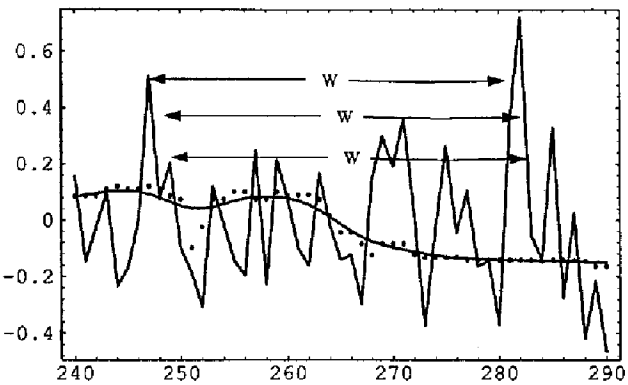

Fig. 1. An expanded region of noise from the middle of a 1D spectrum. The medians are calculated from the extrema which fall within 17 points of each point (W = 34). For example, 24 extrema fall within the topmost window for point 264. The median values are represented by the dots, and the solid interpolating line is the final result after the convolution of the median values with the Gaussian function.

the median values are represented by the solid dots. The array of median values provides a preliminary model for the distorted baseline. Since no assumptions regarding the functional form of the artifact are made, the shape of the distortion that can be handled is arbitrary.

The windows employed for points within W/2 points of the spectrum boundaries must be modified to take into account the ends of the spectrum. If I(1) ≈ I(N), where N is the number of points in the spectrum, then the windows are wrapped across the spectrum boundaries and will include points from both ends of the spectrum. For example, the window associated with point N would comprise the first and last W/2 points. The continuity condition, I(1) ≈ I(N), will in general be satisfied for spectra in which the linear phase correction is a multiple of 360°.

For spectra in which the continuity condition is not satisfied, for example due to an inability to phase the spectrum properly or a pathological distortion, the windows for points within W/2 points of the spectrum boundaries are not wrapped. Instead the windows for these points only include either the first or last W points. For instance, the window associated with point 1 would span points 1 through W, and the window associated with point N would span points N − W + 1 to N. As a result, the windows for the boundary points are not centered at the point for which the median is being calculated. In addition, the baseline estimates for the first and last W/2 points are constant. A discussion of how the program determines if the spectrum is continuous across the boundaries is given in the Appendix.

A Gaussian function is convoluted with the median values to smooth any sharp discontinuities. The final estimated baseline value at point i, B(i), is given by:

$$B(i) = \sum_{j=i-(W/2)+1}^{i+(W/2)} M(j)\,G(i-j)$$

where M(j) is the median value associated with point j. The Gaussian function, G(k), is centered about zero and normalized such that

$$\sum_{k=-(W/2)}^{(W/2)-1} G(k) = 1$$

By decreasing the standard deviation of the Gaussian, δ, from $\delta \gg W$ to $\delta \ll 1$, the weighting of the median values can be adjusted from a uniform to a delta-function weighting. In Fig. 1 the convoluted result is represented by the solid line interpolating between the dots. This smoothing operation is not essential in most cases – the difference between using smoothed and unsmoothed baselines is usually barely discernable to the user on the scale of the NMR resonances. The convoluted result is then subtracted from the original spectrum to produce a flattened baseline.

The calculation outlined here is similar in several respects to the low-frequency deconvolution method introduced by Marion et al. (1989) to remove the zero-frequency component of an FID. In both cases estimates of the quantity to be calculated (zero-frequency component or baseline) are made over localized regions. These estimates are smoothed by convolution with a weighting function and then subtracted from the original FID or spectrum. Besides the very different problems being addressed, the main difference between the two calculations is that in the low-frequency deconvolution method an average instead of the median is computed.

The critical parameter governing the success of the outlined method is the size of the window, W, used in calculating the median. W must be large enough so that the number of local extrema arising from the noise dominates the median statistic. For a given window, if the number of NMR peaks is comparable to the number of noise peaks, then the calculated baseline will be biased upwards for positive NMR peaks; this bias will lead to a reduction in the intensities and volumes of the NMR peaks when the calculated baseline is subtracted from the spectrum. On the other hand, the resolution to which the artifacts can be monitored is inversely proportional to the window size. If W is too large, the median estimates will be influenced by distant points and will not accurately reflect the local baseline structure. Thus far, simultaneously satisfying these two constraints has been readily achieved for all applications of the algorithm in our laboratory. For example, window sizes of 50–80 points have been empirically found to be good compromises for proton spectra with digital resolutions of approximately 8 Hz/point. With these window sizes, the number of local extrema is between 20 and 50. Consequently, only in very crowded regions of the spectrum will the density of NMR peaks be high enough to noticeably skew the baseline. These considerations imply that the best results will be obtained when the NMR peak dispersion is maximal; hence, the algorithm should be applied only after all spectral dimensions have been processed.

The computation of the medians is the most time-consuming step in the execution of the algorithm. The medians are calculated by sorting the intensities of the local extrema within each window. The median is the value of the midpoint of the ordered array if the number of local extrema in the window is odd. If the number is even, the median is taken as the average value of the two middle points of the sorted array. The CPU time required

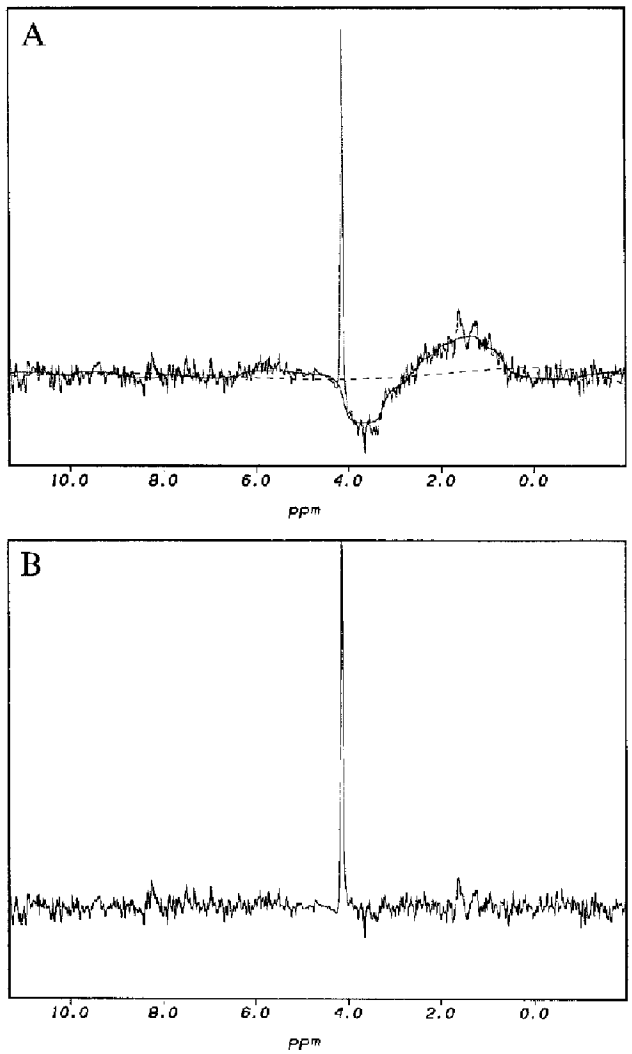

Fig. 2. (A) A particularly distorted slice of a 2D gradient-enhanced, half-reversed filtered, 200 ms ^{1}H,^{1}H NOESY spectrum with suppression of ^{13}C- and ^{15}N-attached protons (Wittekind et al., 1994). The experiment was performed on a 1.2 mM sample of a 15-residue peptide complexed with the N-terminal SH3 domain of a Grb2 protein. The original spectrum is shown with the calculated baselines using the method introduced here (solid line) and the method proposed by Dietrich et al. (1991) (dotted line). For the method set forth here, the window size was set to 70 points, and the standard deviation of the Gaussian smoothing function was set to 5.0 points. In the approach of Dietrich et al. the size of the moving average filter was 4, the default value. (B) The final, baseline-corrected spectrum after the application of the method introduced here.

using this approach is reduced by a factor of approximately two by taking advantage of the fact that the sorted arrays for two adjacent points differ by at most two extrema, as seen in Fig. 1. Hence, the ordered array for the window centered at point i+1 can be obtained from the ordered array for point i in two steps. First, the point located at the trailing window edge, i−W/2, is deleted from the array if it is an extremum. Likewise, if the point at the leading window edge, i+1+W/2, is an extremum, it is inserted so that the array remains sorted.

For multidimensional spectra, the algorithm is currently applied to 1D slices along each dimension. However, it can be generalized in an obvious fashion to two or more dimensions. For two dimensions, the median would be calculated using rectangular windows of extrema centered about each point; in three or four dimensions, the windows would be cubes or hypercubes centered about the points. The aspect ratio of the rectangles could be set equal to the ratio of the number of points in each dimension. The main benefit from generalizing the algorithm in this way should be an increase in the resolution of the model baseline. For higher dimensional spectra, the density of NMR peaks relative to noise extrema decreases, thereby allowing the window size to be reduced. The ability to use a reduced window size may be important, for example, in crowded 2D ^{1}H,^{1}H NOESY spectra. Also, this generalization would allow dimensions with relatively low digital resolution to be included in the baseline correction. Multidimensional windows have not been implemented here, since the in-house, parallelized version of FELIX 1.0 (Hare Research, Bothell, WA) in which the algorithm has been incorporated is currently only designed for 1D processing. To date, no real applications have been encountered in which the density of peaks has been high enough to justify the extra programming effort.

Results

The algorithm is not a major bottleneck in the processing of spectra. For a 2D spectrum of 1024 × 1024 points, the algorithm required 30 CPU seconds for both dimensions on a Silicon Graphics 4D440 workstation (four 40 MHz processors), and approximately 2000 CPU seconds for the two proton dimensions (F1 and F3) of a 3D spectrum of 512 × 128 × 512 points. In both cases, the window size was set to 70 points.

Figure 2A shows the baseline estimates obtained from the algorithm introduced here (solid line) and the algorithm proposed by Dietrich et al. (1991) (dashed line) for a particularly distorted vector from a 2D gradient-enhanced, half-reversed filtered ^{1}H,^{1}H NOESY spectrum. The baseline obtained by the method of Dietrich et al. does not follow the large dip and hump in the spectrum and would not yield a flattened baseline. The baseline calculated by the algorithm introduced here, on the other hand, tracks the distortions extremely well. The shapes and areas of the peaks are close to their initial values, as seen for example in the two small peaks located at the top of the hump (around 1.8 ppm). The only region of the baseline which is not accurately tracked is the bottom of the major dip near the center of the spectrum. Here, the rapidly sloping sides bias the estimate slightly upwards. Nevertheless, the resulting error in the baseline for this region is within the noise level and is certainly small enough for most purposes. The final, baseline-corrected spectrum using the approach presented here is shown in Fig. 2B.

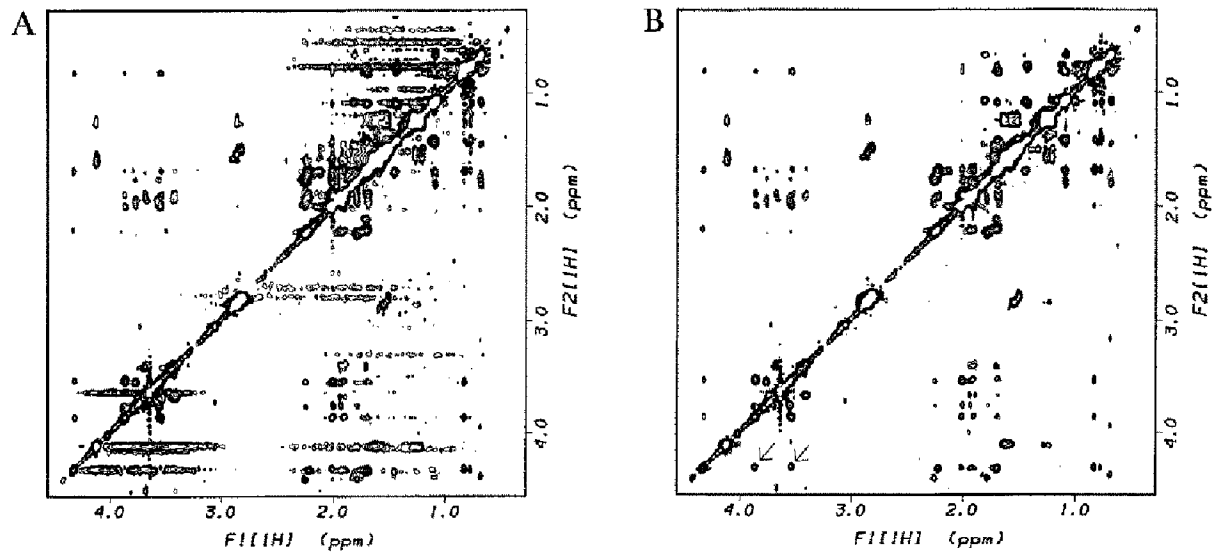

Fig. 3. A region of the same 2D spectrum as described in Fig. 2. (A) and (B) show the region before and after the application of the algorithm. The window size used was 70 points, and the standard deviation of the Gaussian smoothing function was set to 5.0 points. The arrows in (B) point to cross peaks that were obscured by an artifact prior to the application of the algorithm.

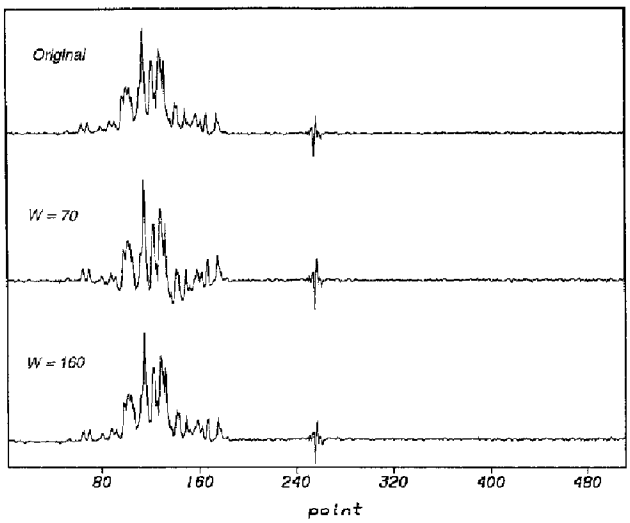

Fig. 4. The $t_1 = 0$ slice from a 2D $^1H,^{15}N$ HSQC spectrum which has only been processed in t_2. The top graph is the original spectrum, the middle graph is the result obtained from the application of the algorithm with a window size of 70 points and the bottom graph is the result obtained with a window size of 160 points. The standard deviations of the Gaussian smoothing function were set to 5.0 and 11.43 points for the applications with window sizes of 70 and 160, respectively.

Figure 3 shows a region of the same 2D spectrum before and after the application of the algorithm to both dimensions. The baseline-corrected spectrum is much cleaner, with all of the streaks either removed or greatly reduced in intensity. Importantly, peaks obscured by streaks in the original spectrum become visible after the application of the algorithm. For instance, the peaks at (3.87,4.31) and (3.54,4.31) ppm (indicated by the arrows near the bottom left-hand corner of Fig. 3B) are hidden before the application of the algorithm, but are evident after the artifact is removed.

To date the algorithm has been applied to a variety of spectra in our laboratory and in all cases it has removed or greatly attenuated the baseline artifacts without introducing any noticeable distortions. To explore the limitations of the approach proposed thus far, two artificial cases are examined; both examples are atypical and would not arise on a regular basis in the work of a protein NMR spectroscopist. An enhancement to the algorithm is then presented which can better treat these cases, albeit with an increase in the computational requirements.

In the upper inset of Fig. 4, the $t_1 = 0$ slice from a 2D $^1H,^{15}N$ HSQC spectrum which has been processed along t_2, but not t_1, is shown. Normally, the algorithm would only be applied after the t_1 dimension had also been transformed. The closely spaced peaks in the amide region, however, test the ability of the algorithm to handle spectra with a high density of NMR resonances. As seen in the middle inset of Fig. 4, if the window size is 70 points, then the baseline is distorted by the algorithm. For this window size, the extrema statistics are dominated by the NMR peaks. However, if the window size is increased to 160 points then the baseline is not deformed, as illustrated in the bottom inset of Fig. 4. For this window size, the noise extrema near the window's edges make a large enough contribution to the extrema statistics for the distortion introduced by the algorithm to be relatively small. The price paid for increasing the window size is a reduction in the resolution to which the actual baseline is mapped.

A second problematic but atypical case is illustrated in Fig. 5 for a synthetic spectrum containing a very broad peak (linewidth = 150 Hz in a spectrum with a spectral width of 3000 Hz). Here, the signal-to-noise ratio has been set low enough that the noise superimposed on the peak leads to a substantial number of local extrema on the peak itself. Because these extrema are included in the calculation of the median, the estimated baseline is biased upwards at the peak's base, as can be seen in Fig. 5. Consequently, the peak's intensity and volume are reduced after the application of the algorithm. For this artificial example, Fig. 6A shows the magnitude of the distortions introduced by the algorithm as a function of the window size and signal-to-noise ratio. The abscissa in Fig. 6 gives the window size, while the ordinate represents the distortion introduced by the algorithm. The distortion is computed as the ratio of the maximum value of the calculated baseline to the peak height; ideally, this ratio should be zero. The three lines from top to bottom represent the results for spectra with increasing signal-to-noise ratios. For spectra with high signal-to-noise ratios, fewer local extrema are present on the peak, and therefore the upward bias in the calculated baseline is smaller. From Fig. 6A, the distortion is unacceptable for window sizes

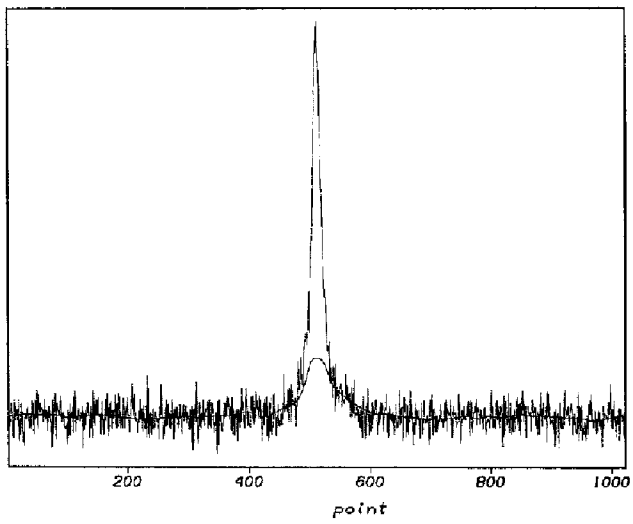

Fig. 5. A synthetic spectrum with a spectral width of 3000 Hz (1024 points) and a signal-to-noise ratio of approximately 16. The linewidth of the large peak is 150 Hz. The smooth line underneath the peak is the baseline obtained from the application of the algorithm to the spectrum with a window size of 70 points. The standard deviation of the Gaussian smoothing function was set to 5.0 points.

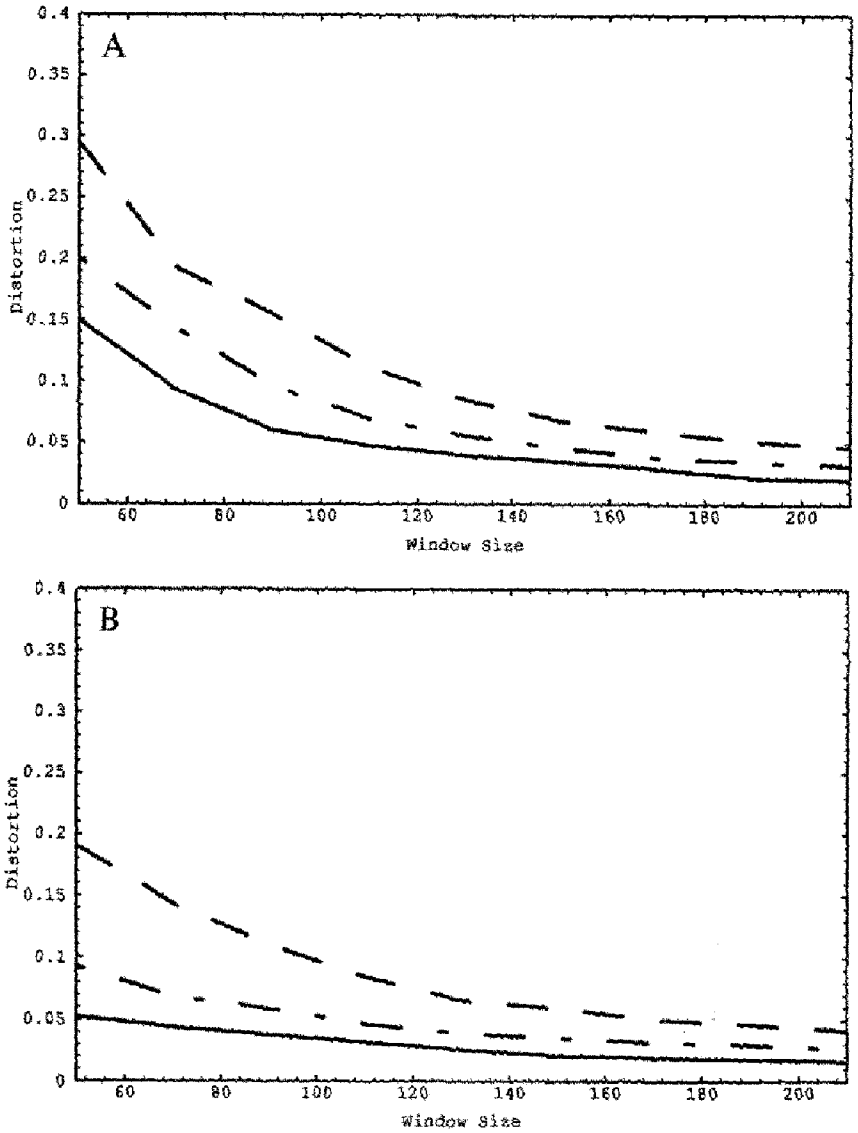

Fig. 6. (A) A plot of the variation of the distortion introduced by the algorithm as a function of the window size and signal-to-noise ratio of the spectrum. The spectral and peak parameters were identical to those used in Fig. 5. The abscissa is the window size in points. The ordinate is the ratio of the maximum value of the calculated baseline to the peak height; it is a measure of the distortion introduced by the algorithm. From top to bottom, the signal-to-noise ratios are 10, 20 and 60. The signal-to-noise ratios were computed as: (peak height/2*standard deviation of the first 300 points). (B) is similar to (A), but with the window size automatically adjusted, as discussed in the text. The abscissa here is the minimum window size used. For all lines, the standard deviation of the Gaussian smoothing function was set to 5.0 points.

which are approximately equal to the linewidth at all signal-to-noise ratios. For higher signal-to-noise ratios (60), window sizes which are twice the linewidth are reasonable, while for lower signal-to-noise ratios (10–20) window sizes which are at least two to three times larger than the linewidth of the peak are required. Of course for narrower and hence more realistic peaks, the distortions would be smaller for the same window sizes and signal-to-noise ratios.

The algorithm can be made more robust by allowing the window size to vary for each point. In spectral regions with a high density of NMR peaks, the density of extrema is lower than elsewhere in the spectrum, due to the smoothing effect of the NMR peaks. Therefore, the program can identify problematic regions, such as those in Figs. 4 and 5, by their relatively low densities of extrema. By subsequently increasing the window size outside these areas, the extrema statistics are more likely to be dominated by the noise peaks, and hence the distortions introduced by the algorithm will be reduced. While the detailed implementation of this idea can take several forms, the following approach gave the best results. For a fixed, user-specified window size, the number of extrema, E(i), within the window of each point i is found; the maximum, X, of the E(i) over all points i within the 1D slice is then determined. The actual window size used for each point is then set so that the number of extrema encompassed by the window is equal to X.

The main advantage to using elastic window sizes is that the resolution to which the regions with ample noise extrema are modeled is maximized, while the magnitude of the distortion introduced in regions with poor noise extrema statistics is minimized. For example, the window size for the spectrum in Fig. 4 could be reduced from 160 to 120 points; in this context, the '120 points' refers to the initial window size used to calculate the number of extrema to be enclosed by each window. For most points, the actual window sizes used were in the range 120–130 points; only in the bank of peaks is the window size increased significantly – up to 174 points. Figure 6B plots the distortions produced when variable window sizes are used for the spectrum displayed in Fig. 5. When Figs. 6A and 6B are compared, it is evident that the distortions are reduced for the same window size. Finally, it should be noted that the results shown in Figs. 2 and 3 were essentially unchanged when the window sizes were allowed to vary.

The drawback to using elastic window sizes is an increase in the computational time by a factor of approximately three. Because the windows of points i and i + 1 no longer have a simple relationship, the method discussed above for finding the median is not used; instead, a full sort is performed for each window of extrema. In addition, the calculation of the window sizes adds to the CPU time. Currently, the default option is to have a fixed window size since it is faster and capable of handling most experimental data.

Conclusions

In summary, a novel approach for removing baseline artifacts has been described. The algorithm is fully automated. It makes no assumptions about the shape or functional form of the distortions, and does not require that NMR peaks be distinguished from noise peaks. Instead, the algorithm exploits the omnipresence of noise extrema to trace the baseline. The only major requirement is that the density of NMR peaks be small relative to that of the noise peaks in the spectrum, a condition satisfied by the majority of multidimensional NMR experiments. The software is available from the author upon request.

Acknowledgements

The author thanks Mike Wittekind, Luciano Mueller, Keith L. Constantine, Bennett Farmer II and William J. Metzler for critical reading of the manuscript and the Editor for a helpful comment.

References

Barsukov, I.L. and Arseniev, A.S. (1987) *J. Magn. Reson.*, **73**, 148–149.
Daubenfeld, J.M., Boubel, J.C. and Delpuech, J.J. (1985) *J. Magn. Reson.*, **62**, 195–208.
Dietrich, W., Rüdel, C.H. and Neumann, M. (1991) *J. Magn. Reson.*, **91**, 1–11.
Dietrich, W. and Gerhards, R. (1981) *J. Magn. Reson.*, **44**, 229–237.
Güntert, P. and Wüthrich, K. (1992) *J. Magn. Reson.*, **96**, 403–407.
Henrichs, P.M., Hewitt, J.M. and Young, R.H. (1986) *J. Magn. Reson.*, **69**, 460–466.
Marion, D. and Bax, A. (1989) *J. Magn. Reson.*, **83**, 205–211.
Marion, D., Ikura, M. and Bax, A. (1989) *J. Magn. Reson.*, **84**, 425–430.
Otting, G., Widmer, H., Wagner, G. and Wüthrich, K. (1986) *J. Magn. Reson.*, **66**, 187–193.
Pearson, G.A. (1977) *J. Magn. Reson.*, **27**, 265–272.
Rouh, A., Delsuc, M.A., Bertrand, G. and Lallemand, J.Y. (1993) *J. Magn. Reson. Ser. A*, **102**, 357–359.
Saffrich, R., Wolfgang, B., Neidig, K.P. and Kalbitzer, H.R. (1993) *J. Magn. Reson. Ser. B*, **101**, 304–308.
Wittekind, M.W., Farmer II, B.T. and Mueller, L., manuscript in preparation.
Zolnai, Z., Macura, S. and Markley, J.L. (1989) *J. Magn. Reson.*, **82**, 496–504.

Appendix

The decision to wrap or not wrap the windows near the spectrum boundaries is based on whether the condition $I(1) \approx I(N)$ is satisfied. To quantitate this condition, the medians M_1 and M_N of the extrema within the first and last W points of the spectrum, respectively, are compared. A scale for the comparison is established by calculating the standard deviations, σ_1 and σ_N, of a subset of the extrema within the two regions. A subset of the full set of extrema is used to exclude NMR peaks from the calculation of the σs. If these peaks were included in the calculations, the standard deviations could be sharply increased. The subset is chosen as the first two quartiles of the absolute values of the extrema contained within each window. For example, if there are L extrema, E_i, within the first W points, and they are sorted and labeled such that $|E_1| \leq |E_2| \leq \ldots \leq |E_L|$, then σ_1 is the standard deviation of the set $\{ |E_1|, |E_2|, \ldots, |E_{L/2}| \}$. By not including the third and last quartile, any NMR peaks in the two end regions should be excluded from the calculations. Once σ_1 and σ_N are found, the minimum of the two is taken and set to σ; the minimum is used since a single scale for the comparisons is required and by choosing the smaller of the two, the condition for continuity is made more stringent. Finally, if the following condition is satisfied:

$$|M_1 - M_N| < 2.0\sigma$$

the spectrum is taken to be continuous across the boundaries and the windows are wrapped.