JOURNAL OF MAGNETIC RESONANCE 96, 403–407 (1992)

FLATT—A New Procedure for High-Quality Baseline Correction of Multidimensional NMR Spectra

PETER GÜNTERT AND KURT WÜTHRICH

Institut für Molekularbiologie und Biophysik, Eidgenössische Technische Hochschule-Hönggerberg, CH-8093 Zurich, Switzerland

Received July 18, 1991; revised September 10, 1991

Current work on further improvements of the methodology used for obtaining NMR solution structures of biological macromolecules (*1, 2*) is largely focused on three goals: (i) improved quality of the structure determination through direct refinement against the experimental NOESY spectra (*3–5*); (ii) computer-supported interactive or automatic analysis of the NMR spectra (*6–8*); and (iii) work with larger molecules using 3D and 4D NMR (*9, 10*) and 2D ^{1}H NMR experiments with heteronuclear filters (*11*). In practice, the success of these new techniques depends critically on the availability of experimental NMR spectra with high-quality, flat baselines. This Communication introduces a new routine, FLATT, for improved baseline flattening in complex, multidimensional NMR spectra.

A variety of factors causing baseline distortions in NMR spectra have been identified (e.g., *12–16*), and the resulting techniques for the reduction of these experimental artifacts can be classified into three categories, depending on the stage during data accumulation and processing when they are applied: (a) during actual data collection, e.g., optimal phase cycling (*12*), or oversampling during data acquisition (*17*); (b) improvement of the time-domain data, e.g., linear prediction to correct the first few data points of the free induction decay (*16, 18*); and (c) improvements of the frequency-domain data, usually by subtracting a suitable function that fits the baseline distortion. This can be either a polynomial (*19, 20*) or a linear combination of the trigonometric functions that correspond to the first few data points of the time-domain signal (*21, 22*). The presently introduced new baseline correction procedure belongs to this third group and thus has the intrinsic advantage that no modifications of data-acquisition and time-domain data processing are required. In the present implementation of the routine in the program FLATT, the spectrum format of the program EASY (*8*) for the interpretation of 2D NMR spectra is used, but the adaption to other file formats would be straightforward.

The program FLATT improves each individual cross section (row or column) in a 2D NMR spectrum separately. The baseline correction is achieved in two steps: first, the regions of the row representing "pure baseline" are identified and second, a function that represents a correction of the first few time-domain data points is best-fitted to the pure-baseline regions and then subtracted from the complete row. On the basis of the observation that a contiguous piece of a row (this may be small compared to the

0022-2364/92 $3.00

complete length of the row but must be larger than the linewidth of the individual peaks) can be well fitted by a straight line only if it lies in a pure-baseline region, the algorithm used is capable of identifying nearly all regions containing pure baseline. If we consider a row with N data points of intensities $S_1, \ldots, S_N$, a close fit to a straight line is obtained for small values of χ^2,

$$\chi_k^2 = \min_{a,b} \frac{1}{2n+1} \sum_{l=-n}^{n} (S_{k+l} - a - bl)^2, \qquad k = n+1, \ldots, N-n, \qquad [1]$$

where χ_k^2 denotes the average squared deviation for a best fit of a straight line to a stretch of $2n + 1$ data points centered at the data point k. The value of n is selected by the user; typically, in a ^{1}H NMR spectrum n is chosen such that a stretch of $2n + 1$ data points corresponds to about 75 Hz. The minimum in Eq. [1] is achieved for

$$a = \frac{1}{2n+1} \sum_{l=-n}^{n} S_{k+l} \quad \text{and} \quad b = \frac{3}{n(n+1)(2n+1)} \sum_{l=-n}^{n} lS_{k+l}. \qquad [2]$$

In the boundary regions $k \leqslant n$ and $k > N - n$, χ_k^2 is set to the χ^2 value of the nearest data point in the definition range of Eq. [1]. On the basis of the idea that χ^2 values smaller than a cutoff indicate pure baseline, a set B_0 of pure-baseline points is defined for a given row by

$$B_0 = \{k \in \{1, \ldots, N\} \mid \min_{l=-n/3,\ldots,n/3} \chi_{k+l}^2 \leqslant \tau\chi_{\min}^2\}. \qquad [3]$$

To use $\min_{l=-n/3,\ldots,n/3} \chi_{k+l}^2$ instead of χ_k^2 extends pure-baseline regions in the vicinity of signal peaks and guarantees a minimal width for each pure-baseline region. The cutoff, $\tau\chi_{\min}^2$, consists of the user-adjustable parameter τ (typically, $\tau = 10$), which is independent of the scaling and the noise level of the spectrum, and the minimal expected χ^2 value, $\chi_{\min}^2$, which is defined either as the minimum value of χ^2 in the row considered or as the average of the minima of χ^2 in all or, in practice, in a selection of rows, whatever is larger.[1]

Normally, the baseline correction is applied separately to the parts of a row that are separated by the diagonal or the water resonance. It can then happen that one finds a large gap between the position of the diagonal or the water line and the nearest pure-baseline point found with the criterion of Eq. [3] calibrated for the entire row. The program FLATT imposes a maximum allowed gap width, i.e., either 10% of the width of the region that is currently baseline-corrected or 5% of the complete row, whatever is larger. As long as the width of the gap exceeds this limit, the gap region is separately searched for additional pure baseline using less stringent criteria; i.e., τ is stepwise increased up to 1.5^4 times its original value (the steps are 1.5, 1.5^2, 1.5^3, and 1.5^4). The set of *all* pure-baseline points is denoted B.

[1] The need for this rather involved selection of $\chi_{\min}^2$ becomes apparent from the following considerations. The average of χ^2 over a selection of rows is used because if $\chi_{\min}^2$ were set to the minimum of χ^2 in the row considered, too little pure baseline would be found in the case in which the row contains a small region that can be significantly better fitted by a straight line than the rest of the row. However, if the smallest value of χ^2 in a given row is bigger than the average of the minimal χ^2 values in the selected rows, too little pure baseline will again be found, which could typically happen for rows containing artifactual noise bands.

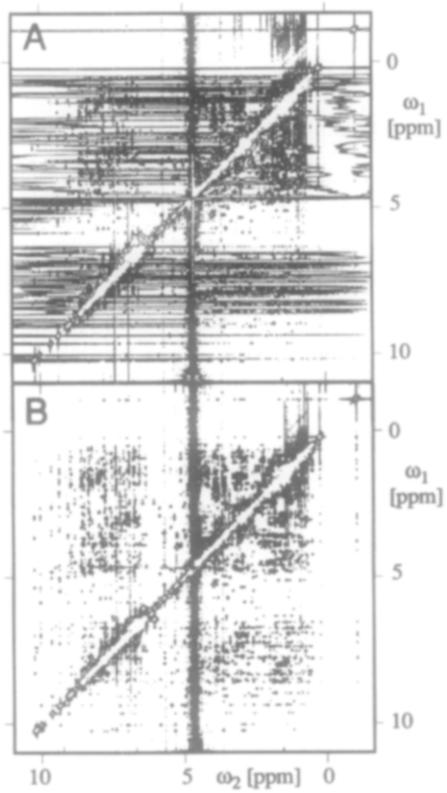

FIG. 1. Contour plots of a NOESY spectrum of Dendrotoxin K (15 m*M* solution of the protein in 95% H_2O/5% D_2O, pH 4.6; temperature, 36°C; proton frequency, 600 MHz; mixing time, 40 ms; recorded time-domain data set with 920 and 2048 data points in the t_1 and t_2 dimensions, respectively; in t_1 the set was zero-filled to 1024 points; cosine filtering in both dimensions; spectral width, 12.62 ppm in both dimensions). Two positive and two negative levels are plotted without distinction; the second contour level is five times higher than the first. (A) Not baseline corrected. (B) Baseline corrected in both dimensions using the program FLATT.

Because the main baseline distortions arise from errors in the measurement of the first few time-domain data points (*16–18*), a linear superposition of the corresponding frequency-domain functions is better suited to fit baseline distortions than the often-used polynomials (*19, 20*). Presently, we use the function F to fit the baseline distortion,

$$F_k = \alpha_1 + \sum_{j=1}^{m} \alpha_{2j}\cos\left(\frac{\pi j(k-1)}{N}\right) + \alpha_{2j+1}\sin\left(\frac{\pi j(k-1)}{N}\right), \qquad [4]$$

$k = 1, \ldots, N$, where N is the number of data points in the row, and $\alpha_1, \ldots, \alpha_{2m+1}$ are adjustable parameters. The value of m is selected by the user; typically, $m = 3$. Equation [4] is valid for spectra that extend over the full recorded spectral width and were calculated by sine and cosine Fourier transformation of the time-domain data; otherwise Eq. [4] must be modified. The parameters $\alpha_1, \ldots, \alpha_{2m+1}$ are adjusted for

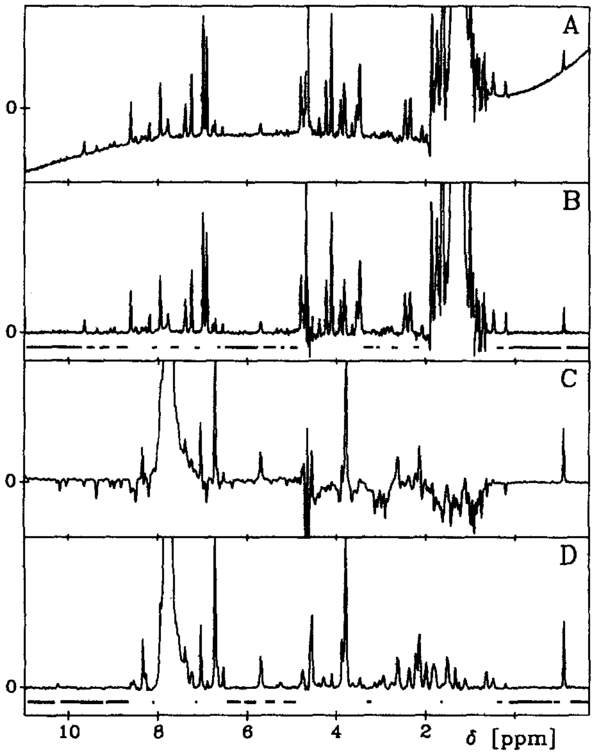

FIG. 2. Cross sections from the NOESY spectra of Fig. 1. In all cases the full spectral width of 12.62 ppm is shown. (A) Row at ω_1 = 1.34 ppm before baseline correction. (B) Same row as in A after baseline correction in both dimensions. (C) Column at ω_2 = 7.76 ppm before baseline correction. There are only small baseline distortions in the column, but large artifacts are caused by baseline distortions of the rows, in particular in the region ω_1 = 0.5–3.5 ppm. (D) Same column as in C after baseline correction in both dimensions. The improvement over C is primarily due to the baseline correction of the rows perpendicular to this column. In B and D the pure-baseline regions identified by the program FLATT are indicated by horizontal lines below the spectrum.

optimal least-squares fit of the function F to the spectrum S in the pure-baseline regions considered. The latter consist of all points in the set B if the baseline correction is made for the entire row, or to a subset of B if the spectral regions separated by the diagonal or the water line are treated separately. The linear least-squares problem is solved by standard techniques (*23*) using singular-value decomposition in order to avoid instabilities in the case of an almost singular problem. The baseline-corrected row, $\tilde{S}$, is then obtained as $\tilde{S}_k = S_k - F_k$, for $k = 1, \ldots, N$. Obviously, the same treatment can be applied to columns in a 2D spectrum, and for baseline corrections in higher-dimensional spectra.

The program FLATT has been in use in our laboratory for some time. In general, better results are obtained with this program than with other available baseline correction routines. As an illustration, Figs. 1 and 2 show a baseline-corrected NOESY spectrum of Dendrotoxin K from *Dendroaspis polylepis polylepis*, a protein with 57 amino acid residues (K. D. Berndt, P. Güntert, and K. Wüthrich, unpublished). The

NMR measurements were carried out on a Bruker AM 600 spectrometer. Because the most severe baseline distortions occur in the rows, these were corrected first, using $n = 20$ in Eq. [1]. Subsequently, the baseline in the columns was corrected, using $n = 10$ (after zero-filling, the total number of data points along ω_2 is twice that along ω_1; see caption to Fig. 1). In both dimemsions, $\tau = 10$ and $m = 3$ were used in Eqs. [3] and [4], respectively, and the baseline correction was done separately for the regions on the left and the right of the diagonal. The complete baseline correction took 12.5 min of CPU time on a Sun SPARC station 2. Figure 1 affords a comparison of the complete spectra before and after baseline correction, which shows that the baseline distortions could successfully be removed. The high quality of the baseline correction can be seen more clearly in Fig. 2, which shows cross sections through the spectra of Fig. 1. The pure-baseline regions found by FLATT are indicated in Figs. 2B and 2D; in both the row and the column, the algorithm was able to distinguish reliably between pure-baseline and peak-containing regions. After baseline correction (Figs. 2B and 2D), the spectrum has zero intensity (within the noise uncertainty) in almost all pure-baseline regions.

The program FLATT is available for use in other laboratories; please contact the authors (do not send a tape).

ACKNOWLEDGMENTS

We thank Dr. K. D. Berndt for the use of unpublished spectra of Dendrotoxin K. Financial support by the Schweizerischer Nationalfonds, Project 31.25174.88, is gratefully acknowledged.

REFERENCES

1. K. Wüthrich, "NMR of Proteins and Nucleic Acids," Wiley, New York, 1986.

2. K. Wüthrich, *Science* **243,** 45 (1989).

3. B. A. Borgias and T. L. James, *J. Magn. Reson.* **79,** 493 (1988).

4. P. Yip and D. A. Case, *J. Magn. Reson.* **83,** 643 (1989).

5. J. E. Mertz, P. Güntert, W. Braun, and K. Wüthrich, *J. Biomol. NMR* **1,** 257 (1991).

6. M. Billeter, V. J. Basus, and I. D. Kuntz, *J. Magn. Reson.* **76,** 400 (1988).

7. G. J. Kleywegt, R. Boelens, M. Cox, M. Llinás, and R. Kaptein, *J. Biomol. NMR* **1,** 23 (1991).

8. C. Eccles, P. Güntert, M. Billeter, and K. Wüthrich, *J. Biomol. NMR* **1,** 111 (1991).

9. S. W. Fesik and E. R. P. Zuiderweg, *Q. Rev. Biophys.* **23,** 97 (1990).

10. G. M. Clore, L. E. Kay, A. Bax, and A. M. Gronenborn, *Biochemistry* **30,** 12 (1991).

11. G. Otting and K. Wüthrich, *Q. Rev. Biophys.* **23,** 39 (1990).

12. R. R. Ernst, G. Bodenhausen, and A. Wokaun, "Principles of Nuclear Magnetic Resonance in One and Two Dimensions," Clarendon Press, Oxford, 1987.

13. R. Freeman, "A Handbook of Nuclear Magnetic Resonance," pp. 11–13, Longman Scientific & Technical, Harlow, United Kingdom, 1988.

14. E. O. Stejskal and J. Schaefer, *J. Magn. Reson.* **14,** 160 (1974).

15. D. I. Hoult, C.-N. Chen, H. Eden, and M. Eden, *J. Magn. Reson.* **51,** 110 (1983).

16. G. Otting, H. Widmer, G. Wagner, and K. Wüthrich, *J. Magn. Reson.* **66,** 187 (1986).

17. G. Wider, *J. Magn. Reson.* **89,** 406 (1990).

18. D. Marion and A. Bax, *J. Magn. Reson.* **83,** 205 (1989).

19. Bruker Analytische Messtechnik GmbH, "Uxnmr User's Guide," Rheinstetten, Germany, 1990.

20. W. Dietrich, C. H. Rüdel, and M. Neumann, *J. Magn. Reson.* **91,** 1 (1991).

21. P. M. Henrichs, J. M. Hewitt, and R. H. Young, *J. Magn. Reson.* **69,** 460 (1986).

22. J. Cavanagh and M. Rance, *J. Magn. Reson.* **88,** 72 (1990).

23. W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, "Numerical Recipes: The Art of Scientific Computing," pp. 515–519, Cambridge Univ. Press, Cambridge, 1986.