IFLAT—A New Automatic Baseline-Correction Method for Multidimensional NMR Spectra with Strong Solvent Signals

CHRISTIAN BARTELS, PETER GÜNTERT, AND KURT WÜTHRICH

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

Received August 31, 1995

Flat baselines are the basis for efficient and reliable analysis of multidimensional NMR spectra, in particular for quantitative evaluation of NOESY spectra in solution structure determinations of biological macromolecules (*1*). During the past two decades, numerous baseline correction algorithms have been developed (e.g., *2–9*), but further improvements are still needed. In particular, strong solvent resonances that have been incompletely suppressed during acquisition constitute a major source of baseline instabilities, which often preclude a quantitative analysis of cross peaks in the vicinity of the solvent signals. Furthermore, a frequently used convolution method (*10*) for removal of residual solvent signals during data processing annihilates all signals near the solvent, but without such pretreatment of the data sets the available algorithms have only limited success in handling baseline distortions caused by solvent signals. In the present Communication, we introduce a new, automatic baseline-correction algorithm, IFLAT (''*i*terative *flat*tening''), capable of correcting simultaneously baseline distortions from strong solvent signals and from other sources without bleaching signals close to the solvent signal frequency.

The key feature of IFLAT is the use of a probabilistic quantity for attributing given data points to a baseline region of the spectrum. This replaces the strict distinction between pure-baseline and signal data points used generally in earlier approaches (*2–8*). IFLAT estimates for each data point the probability that it belongs to a pure-baseline region. These probabilities are then used to weigh the individual data points in an approximate least-squares fit to determine a linear combination of base functions that best represents the baseline distortion. After subtraction of this baseline function from the spectrum, the process is repeated with updated probabilities until convergence (*2*).

The conditional probability that a data point with intensity y corresponds to a pure-baseline point is defined as $p(b \mid y)$. To evaluate this quantity, we assume that each data point belongs to a positive signal, a negative signal, or a pure-baseline region, with a priori probabilities $p(s_+)$, $p(s_-)$, and $p(b) = 1 - p(s_+) - p(s_-)$, respectively. The additional conditional probabilities $p(y \mid b)$, $p(y \mid s_+)$, and $p(y \mid s_-)$ express the likelihood that a data point in a pure-baseline region, a positive signal, or a negative signal, respectively, has intensity y, and they are assumed to be given by normal distributions,

$$p(y \mid b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2\right] \text{ and}$$

$$p(y \mid s_\pm) = \frac{2\Theta_\pm(y)}{\sqrt{2\pi}q\sigma} \exp\left[-\frac{1}{2}\left(\frac{y}{q\sigma}\right)^2\right]. \quad [1]$$

σ is the standard deviation for pure-baseline data points, and the standard deviation for signal data points is assumed to be q times larger. We used $q = 10$ throughout. $\Theta_+(y)$ denotes the Heaviside function, which is equal to 1 for nonnegative arguments and vanishes otherwise, and $\Theta_-(y) = 1 - \Theta_+(y)$. Using these definitions and Bayes' formula, the conditional probability $p(b \mid y)$ can then be calculated as

$$p(b \mid y) = \frac{p(y \mid b)p(b)}{p(y \mid b)p(b) + p(y \mid s_+)p(s_+) + p(y \mid s_-)p(s_-)}$$
$$= \left\{1 + \frac{2}{q}\frac{p(s_+)\Theta_+(y) + p(s_-)\Theta_-(y)}{1 - p(s_+) - p(s_-)} \times \exp\left[\frac{1}{2}\left(\frac{y}{\sigma}\right)^2\left(1 - \frac{1}{q^2}\right)\right]\right\}^{-1}. \quad [2]$$

In its practical implementation, IFLAT corrects each one-dimensional cross section of a multidimensional spectrum separately. For a given cross section consisting of n data points with intensities $y_{1,0}, \ldots, y_{n,0}$ taken from the original, untreated spectrum ($l = 0$), the algorithm starts by setting the parameters $p_0(s_+)$ and $p_0(s_-)$ to 0.05, and σ_0 to a suitable initial value (Fig. 1). The value of σ_0 is obtained by separately fitting each of the orthonormal base functions $f_1, \ldots,$

1064-1858/95 $12.00

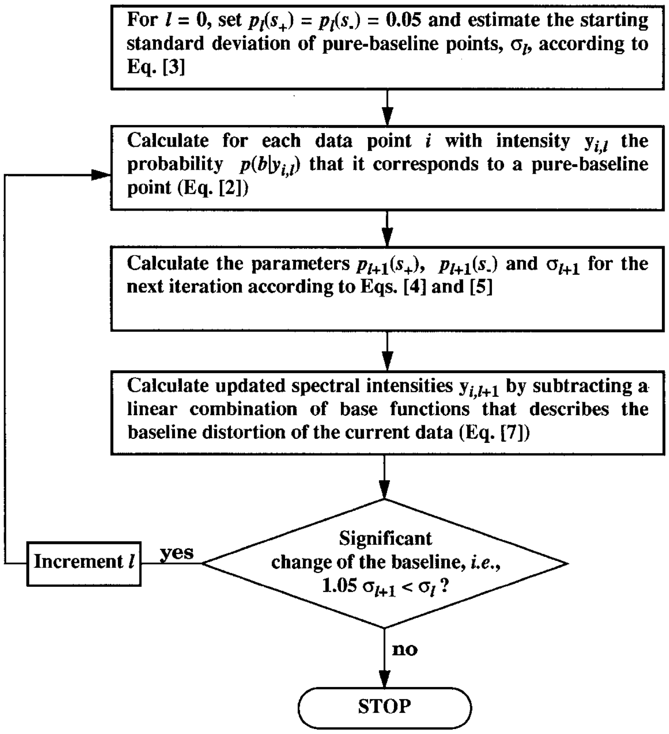
For l = 0, set $p_l(s_+) = p_l(s_-)$ = 0.05 and estimate the starting standard deviation of pure-baseline points, σ_l, according to Eq. [3]
Calculate for each data point i with intensity $y_{i,l}$ the probability $p(b|y_{i,l})$ that it corresponds to a pure-baseline point (Eq. [2])
Calculate the parameters $p_{l+1}(s_+)$, $p_{l+1}(s_-)$ and σ_{l+1} for the next iteration according to Eqs. [4] and [5]
Calculate updated spectral intensities $y_{i,l+1}$ by subtracting a linear combination of base functions that describes the baseline distortion of the current data (Eq. [7])
Significant change of the baseline, i.e., 1.05 $\sigma_{l+1} < \sigma_l$?
yes
Increment l
no
STOP

FIG. 1. Flow chart of the iterative baseline-correction algorithm IFLAT (see also the text). IFLAT is applied separately to each cross section of a spectrum.

f_m (see below) to the cross section; its value is then the sum of the standard deviations of these fitted functions,

$$\sigma_0 = \sum_{k=1}^{m} \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} y_{i,0} f_k(i)\right)^2}. \qquad [3]$$

Using the conditional probabilities $p(b \mid y)$ of Eq. [2] in each iteration, $l = 0, 1, 2, \ldots$, updated values of the parameters $p(s_+)$, $p(s_-)$, and σ are determined according to

$$\begin{aligned} p_{l+1}(s_\pm) &= \frac{1}{n}\sum_{i=1}^{n} p(s_\pm \mid y_{i,l}) \\ &= \frac{1}{n}\sum_{i=1}^{n} [1 - p(b \mid y_{i,l})]\Theta_\pm(y_{i,l}), \qquad [4] \end{aligned}$$

$$\sigma_{l+1} = \sqrt{\sum_{i=1}^{n} p(b \mid y_{i,l}) y_{i,l}^2 / \sum_{i=1}^{n} p(b \mid y_{i,l})}. \qquad [5]$$

The baseline distortion is represented by a linear combination of base functions, $f_1, \ldots, f_m$, and optimal values for the linear combination coefficients $a_1, \ldots, a_m$ are determined by minimizing the linear least-squares expression

$$\chi^2(a_1, \ldots, a_m) = \sum_{i=1}^{n} p(b \mid y_{i,l})(y_{i,l} - \sum_{k=1}^{m} a_k f_k(i))^2. \qquad [6]$$

In practice, $f_1, \ldots, f_m$ are obtained by orthonormalizing a set of user-specified functions (examples are described below). Since the majority of the n data points $y_{1,l}, \ldots, y_{n,l}$ usually belong to baseline regions, an efficient approximate solution of the least-squares problem [6] can be obtained based on the following: if the base functions are chosen to be orthonormal with respect to the n-dimensional standard scalar product, the orthogonality is to a good approximation preserved after weighing of the data points with their probabilities of belonging to a pure-baseline region. The updated spectral intensities, $y_{i,l}^{(m)}$, are then obtained from the current intensities, $y_{i,l}$, by recursion over all base functions f_k:

$$y_{i,l}^{(k)} = y_{i,l}^{(k-1)} - \frac{\sum_{j=1}^{n} p(b \mid y_{j,l}) f_k(j) y_{j,l}^{(k-1)}}{\sum_{j=1}^{n} p(b \mid y_{j,l}) f_k(j)^2} f_k(i),$$
$$i = 1, \ldots, n;\ k = 1, \ldots, m. \qquad [7]$$

In Fig. 1, $y_{i,l}^{(m)} \equiv y_{i,l+1}$ and $y_{i,l}^{(0)} \equiv y_{i,l}$. In order to obtain correctly updated spectral intensities also in cases where the above approximation is less well fulfilled, the recursive scheme of Eq. [7] is repeated until the changes of the spectral intensities are small compared to the estimated standard deviation of pure-baseline points, σ_l.

One of the IFLAT applications in our laboratory is shown in Figs. 2 and 3. Figure 2A shows a contour plot of a NOESY spectrum of the *Antp(C39S, W56S)* homeodomain (C. Bartels, D. Resendez-Perez, P. Güntert, D. Braun, W. J. Gehring, and K. Wüthrich, unpublished results) before baseline correction. Since the baseline distortions are most pronounced in the direct dimension, the correction was first applied to all cross sections along ω_2. The base function set consisted of the trigonometric functions corresponding to the first four complex time-domain data points and two functions describing contributions from the water signal to the baseline. These were the absorptive and dispersive components of a signal with half-width at half-height of one data point centered at the solvent resonance frequency. After correction in the direct dimension, only minor baseline distortions remained in ω_1, which were removed using as base functions the two trigonometric functions corresponding to the first complex time-domain data point. Figure 2B shows that a flat baseline was thus obtained in the entire spectrum. Figure 3 shows in more detail the ability of IFLAT to correct baseline distortions arising from strong solvent signals. As an illustration, we use the proton spin system of Thr 41 in the *Antp(C39, W56S)* homeodomain, which consists of the amide proton, C$^\alpha$H, C$^\beta$H, and C$^\gamma$H$_3$. In the uncorrected spectrum (Figs. 3A and 3E), strong baseline distortions along ω_2 are due to the water signal and obliterate cross peaks close to the water resonance frequency. For example, although the H$^\beta$ line can

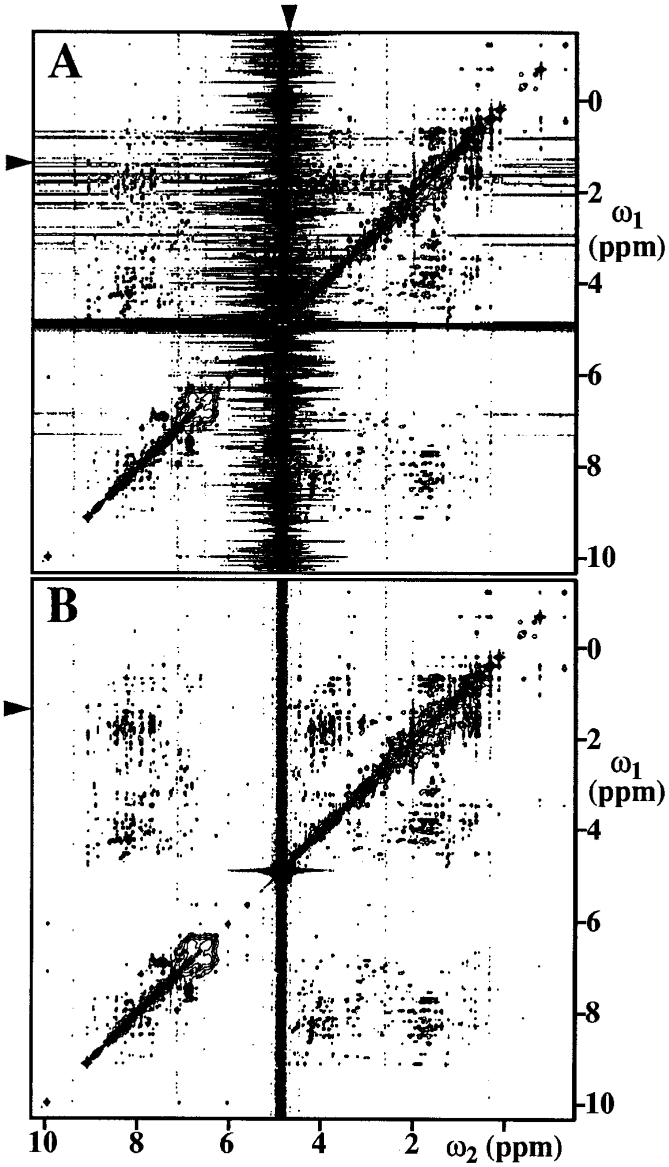

FIG. 2. Contour plots of a NOESY data set (*11*) of the *Antp*(*C39S, W56S*) homeodomain (2 m*M* solution of the protein in 95% H_2O/5% D_2O, pH 4.3, *T* = 10°C, ^1H frequency = 500 MHz, mixing time = 80 ms). The solvent was suppressed using continuous-wave preirradiation during the relaxation delay and the mixing period with a B_2 field of approximately 10 Hz (*12*). The time-domain data set was recorded with 350 and 1024 complex data points in t_1 and t_2, respectively. In t_1, it was zero-filled to 1024 complex points and cosine filtered, and, in t_2, it was zero-filled to 2048 complex points and filtered with a sine window shifted by 75° (*13*). The total sweep width was thus 13.2 ppm in ω_1 and 18.2 ppm in ω_2, but border regions containing no peaks are not shown. Positive and negative contour levels are plotted without distinction, and there is a factor of two between successive levels. The arrowheads indicate the positions at which the cross sections of Fig. 3 were taken. (A) Before baseline correction. (B) Baseline corrected in both dimensions using IFLAT implemented in the program PROSA (*7*).

fortuitously be observed (but hardly integrated) in Fig. 3A, the resonances of Thr 41 cannot be identified in the ω_1 trace of Fig. 3E, which runs parallel to the water line at a distance of 90 Hz. Using IFLAT, the distortions of the spectrum could be removed without bleaching of the signals near to the water line, and after the baseline correction the signals of Thr 41 in the ω_1 cross section of Fig. 3F can clearly be identified.

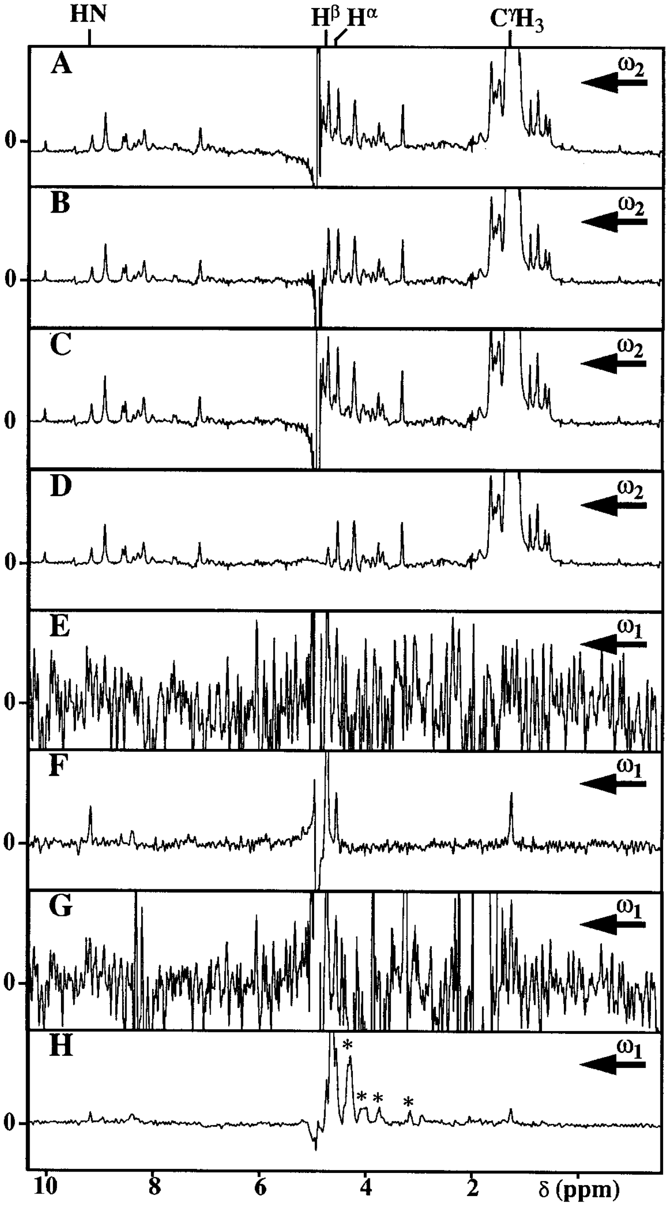

FIG. 3. Cross sections from the NOESY spectrum of Fig. 2 before baseline correction (A, E), and after baseline correction in both dimensions using IFLAT (B, F). For comparison (see text), traces obtained after baseline correction with the program FLATT (*6*) (C, G), and after removal of the residual water signal with the convolution method (*10*) followed by baseline correction with FLATT (D, H) are also shown. (A to D) Row along ω_2 at the chemical shift of $C^{\gamma}H_3$ of Thr 41 (ω_1 = 1.22 ppm). (E to H) Column at the chemical shift of H^{β} of Thr 41 (ω_2 = 4.72 ppm). The chemical shifts of the four resonance lines of Thr 41 are indicated at the top. In trace (H), the peaks identified by asterisks are artifacts resulting from the application of the convolution routine (*10*).

The principal motivation for developing IFLAT was the limitations of the precursor program FLATT (*6*), in dealing with distortions arising from solvent signals (Figs. 3C and 3G) (we choose FLATT to represent ''conventional'' baseline routines (*2–8*) in this respect). In practice, FLATT was therefore always used in combination with a routine to remove the solvent line during data processing, usually the convolution method (*10*). Although this combined treatment results in a very nice baseline and virtually complete water suppression (Figs. 3D and 3H), the signals close to the water resonance line are reduced in intensity and some of them are covered by artifacts (Figs. 3D and 3H).

To obtain an estimate for the width of the spectral region covered by the residual water signal after baseline correction with IFLAT, we analyzed cross peaks close to the water resonance in NOESY spectra of the *Antp*(*C39, W56S*) homeodomain, the killer toxin from *Williopsis mrakii* (unpublished results), and the pheromone E*r*-22 from the ciliated protozoan *Euplotes raikovi* (unpublished results). The NOESY spectra were recorded in mixed solvents of 90% H_2O/10% D_2O on Bruker AMX 500, Bruker AMX 600, and Varian Unity+ spectrometers, respectively, using preirradiation for water suppression. In all cases, peaks at a distance of 100 Hz or more from the water resonance (0.2 ppm at 500 MHz, 0.13 ppm at 750 MHz) were well resolved and undistorted, and, in some cases, peaks as close to the water resonance as 40 Hz could readily be identified and integrated.

IFLAT has been implemented as a new command in the program PROSA (*7*), which is available from the authors upon request. The cpu times used for the baseline correction are comparable to those needed for the more conventional approach with the program FLATT (*6, 7*). For example, the baseline correction shown in Fig. 2 used 59 s on an IBM 6000-590 workstation, and 1.8 s on a NEC SX3 supercomputer.

ACKNOWLEDGMENTS

The use of the NEC SX3 supercomputer of the Centro Svizzero di Calcolo Scientifico and of the C4 workstation cluster of the ETH Zürich are gratefully acknowledged. Financial support was obtained from the Kommission zur Förderung der wissenschaftlichen Forschung (Project 2223.1). We thank R. Marani for the careful processing of the manuscript.

REFERENCES

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

2. G. A. Pearson, *J. Magn. Reson.* **27,** 265 (1977).

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

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

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

6. P. Güntert and K. Wüthrich, *J. Magn. Reson.* **96,** 403 (1992).

7. P. Güntert, V. Dötsch, G. Wider, and K. Wüthrich, *J. Biomol. NMR* **2,** 619 (1992).

8. A. Rouh, M. A. Delsuc, G. Bertrand, and J. Y. Lallemand, *J. Magn. Reson. A* **102,** 357 (1993).

9. M. S. Friedrichs, *J. Biomol. NMR* **5,** 147 (1995).

10. D. Marion, M. Ikura, and A. Bax, *J. Magn. Reson.* **84,** 425 (1989).

11. Anil Kumar, R. R. Ernst, and K. Wüthrich, *Biochem. Biophys. Res. Commun.* **95,** 1 (1980).

12. G. Wider, R. V. Hosur, and K. Wüthrich, *J. Magn. Reson.* **52,** 13 (1983).

13. A. DeMarco and K. Wüthrich, *J. Magn. Reson.* **24,** 201 (1976).