JOURNAL OF MAGNETIC RESONANCE 83, 205–211 (1989)

Baseline Correction of 2D FT NMR Spectra Using a Simple Linear Prediction Extrapolation of the Time-Domain Data

DOMINIQUE MARION* AND AD BAX

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892

Received January 19, 1989

In recent years, 2D NMR has proven its efficiency in solving the conformation of biological macromolecules in solution (*1*). So far, most structural NMR studies of biological macromolecules rely on qualitative interpretation of 2D NOESY spectra. Quantitative analysis of 2D NOE cross-peak intensities requires the use of very short NOE mixing times and the NOESY spectra obtained at these shorter mixing times are often plagued by serious baseline distortions that are associated with the very high intensity of the diagonal peaks. Consequently, at these short mixing times, the accuracy of NOE cross-peak integration is often determined by the extent of baseline distortion rather than by the true signal-to-noise ratio of the 2D spectrum. If the baseline distortion is linear across the spectrum, correction in the frequency-domain 2D spectrum is straightforward. Often, however, there is also a baseline curvature, which is much harder to correct in the frequency-domain 2D spectrum because the correction procedure must distinguish signals and noise in defining the baseline position. In this Communication, we describe an alternative and efficient time-domain method for improving the baseline features of 2D spectra.

Serious baseline offset can result from the erroneous implementation of the Fourier transformation algorithm, when the first data point is scaled incorrectly (*2*). Nonlinear baseline distortion occurs in FT spectra of time-domain data that have been sampled in a *noncomplex* mode if the required zero-order phase correction is not a multiple of 90° (*3*). In this case, the problem can be solved by correctly setting up the experiment by adjustment of the receiver phase relative to the transmitter RF phase (*3*). The present Communication concentrates on two remaining and closely interrelated sources of distortion that will be referred to as the *first data point(s) problem* and the *delayed acquisition problem.* In the first case, several points are sampled at the correct time, but are altered (usually underestimated) due to the time response of the audiofrequency bandpass filter (*4*). Otting *et al.* (*2*) have proposed to use an empirically determined scaling factor for the first data point to correct for this error. However, if the required scaling factor is much larger than one, amplification of this typically poorly determined first data point may result in an increase in noise in the

*On leave from the Centre de Biophysique Moléculaire, CNRS, 1A, Av. de la Recherche Scientifique, 45071 Orléans Cedex 2, France.

0022-2364/89 $3.00

t_1 dimension. The delayed acquisition problem constitutes a more difficult source of baseline distortion. Finite pulse width or selective excitation may prevent a start of the acquisition at time $t = 0$, leading to a shift of the sampling frame and thus to a frequency-dependent phase error after FT. The use of frequency-dependent phase corrections results invariably in a nonlinear distortion of the baseline (*5*).

The physical origin of any frequency-domain baseline distortion lies in the first few data points. Therefore, time-domain corrections affecting only a few data points should be more suitable than frequency-domain baseline corrections (*6, 7*). Here, we show that linear prediction (LP) techniques can be used efficiently for restoring these missing (or corrupted) points, yielding 2D spectra with excellent baseline properties.

The aim is to compute the first points of the time-domain signal from the correctly sampled data points, acquired at later times without distortion. Polynomial extrapolation, whatever the chosen order is, is poorly suited for this purpose; the result relies mainly on the optimization algorithm used and polynomial fitting of a sum of damped sinusoids is likely to be error prone if the extrapolation extends the signal by more than a cycle of the highest frequency component. A model-free extrapolation of the time-domain data can be obtained using the LP technique. In recent years, maximum entropy (*8–10*) and LP techniques (*11–14*) have been introduced in NMR for the purpose of improving sensitivity and/or resolution, but the present application is quite different.

The success of linear prediction in computing spectral parameters has partially obscured the etymological origin of its name. Its most straightforward use is to predict the future of a time series from the record of its past or *vice versa.* Consider, for example, a signal consisting of k exponentially damped sinusoids sampled as real data, y_n, at regular intervals. According to the LP model, each data point can be expressed as a linear combination of the $2k$ following points,

$$y_n = \sum_{i=1}^{2k} d_i y_{n+i}, \qquad [1]$$

where the LP coefficients, d_i, must be determined from the existing points. In the baseline problem, mentioned above, one or several points at the beginning of the FID are missing or altered and they will be reconstructed from the correctly sampled data points acquired at later times using Eq. [1].

Linear prediction is especially successful at extrapolating signals which are smooth and oscillatory. The LP extrapolation of a few data points requires orders of magnitude less computation than the extensive derivation of the spectral parameters, as achieved in conventional use of LP. Because the extrapolation extends the time-domain data by only a very small fraction of its total duration, the accuracy requirements are not very stringent in this application, permitting the use of the fast Burg algorithm (*15, 16*), instead of the classical singular value decomposition method (LPSVD) (*17*).

The method developed by Burg (*16*) for autoregressive power spectral density estimation yields the LP coefficients. This recursive algorithm computes the p-order prediction coefficients from those obtained previously for $(p - 1)$ order using the so-called *Levinson* recursion formula (*18*). At each step of the iteration a single new term, d_p, called the reflection coefficient, is derived from the previous backward and

forward linear prediction errors. The whole iteration works exceedingly fast without assuming anything about the availability or nonavailability of data outside those which have been measured (*8*). The aim of the Burg algorithm is only to choose, at each stage, a new reflection coefficient which minimizes the noise variance and consequently the prediction error. A detailed description of this algorithm may be found in Refs. (*15, 18*).

For reliable extrapolation, LP algorithms require the use of a number of time-domain data points that is larger than the number of frequency components present in the spectrum with a significant amplitude. This number can be several hundred or more in a typical protein 2D data set. To avoid lengthy calculations needed for such large numbers of coefficients, an alternative processing strategy is used, defined in the flowchart of Fig. 1. The accuracy of the LP extrapolation in a given dimension (say t_1) is likely to be higher for the same computing payload if the data are previously transformed with respect to the orthogonal dimension (t_2), because fewer resonances contribute then to the t_1 interferogram. In fact, for NOESY spectra recorded with short mixing times, the signal often has only a single major component (the diagonal). Extrapolation of the t_1 time-domain signal is straightforward, and Fourier transfor-

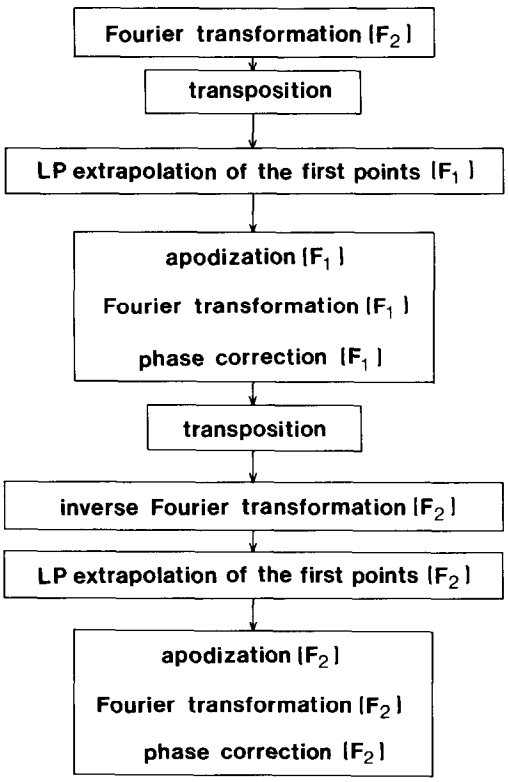

FIG. 1. Flowchart of the algorithm used for correcting the first data points using linear prediction extrapolation. In our work, the standard steps such as apodization, FT, phase correction, and transposition were performed using commercially available software (New Methods Research, Inc., Syracuse, New York).

mation gives F_1 slices free of baseline distortion. FT is a reversible algorithm and t_2 time-domain data can be recovered by means of an inverse FT. Because the F_1 FT has taken place already, only few frequency components are present in the recovered t_2 slices, and LP extrapolation of their first data points is straightforward.

Instabilities occur when the characteristic polynomial

$$z^{2k} - \sum_{i=1}^{2k} d_i z^{2k-i} = 0 \qquad [2]$$

has at least one of its $2k$ complex roots far outside the unit circle defined by $|z| = 1$. Note that these roots are connected to the spectral characteristics (frequencies and decays) of the k damped sinusoids. The instability problem obviously becomes more severe if the linear prediction is extended further. In practice, however, only one or two pairs of data points are altered by the time response of the filter (*first data point(s) problem*) and experimental constraints rarely delay the t_1 or t_2 acquisition by more than a few dwell times. For such short extrapolations, instabilities do not become a problem in signal-containing data sets. However, even in 2D spectra of large biomolecules, some spectral regions do not contain any signal and thus LP applied to these noisy, but signalless, data may lead to erratic results, reflected in large values of $|z|$. A simple but efficient procedure to prevent this first solves the polynomial in [2], and then reflects roots with a magnitude larger than about 1.3, so that they fall inside the unit circle, i.e., z_j becomes $1/z_j$, and these "massaged" LP coefficients are used for the reconstruction (*15*).

The LP extrapolation can be used successfully in both F_1 and F_2 dimensions, provided that the data have been suitably sampled, so that a data point at $t_1 = 0$ ($t_2 = 0$) can be estimated. In contrast to polynomial extrapolation, LP can predict the time-domain signal only at discrete times, separated by an integral number of dwell times (Δ_1 or Δ_2). Computing a point at $t_1 = 0$ or $t_2 = 0$ thus requires sampling the first point at $t_1 = n\Delta_1$ (or $t_2 = n\Delta_2$); if, for instrumental reasons, the data point at $t_1 = 0$ or $t_2 = 0$ cannot be obtained, the first data point should be taken at a value that is the lowest integral multiple of the dwell time. In the t_2 dimension, the signal is delayed by the time response of the audiofilter and sampling at $t_2 = 0$ (the time when the phase of transverse magnetization components is independent of their offset) is often possible. If sampling is started at this time, no frequency-dependent phase correction will be necessary in the F_2 dimension. In the F_1 dimension, there are no audiofilters and sampling a point at $t_1 = 0$ would require infinitely short pulses.

Figure 2 shows the comparison of two NOESY spectra recorded on the small cyclic peptide surfactin (*19*) from *Bacillus subtilis*. This example illustrates the difficulties encountered in NOESY spectra with weak cross peaks. For a fair comparison of the best achievable results with and without LP extrapolation, the spectra originate from two separate data sets recorded under identical conditions except for the initial t_1 increment (see below). In both experiments, the delay before the start of t_2 data acquisition has been carefully adjusted such that no linearly frequency-dependent phase correction of the F_2 data is needed. In contrast, a linear phase correction in F_1 is needed in the regular uncorrected NOESY spectrum because of the finite duration of the 90° pulses flanking the t_1 period. It can be calculated that, to a good approximation, for offsets that are smaller than the strength of the RF field, the first 90° pulse

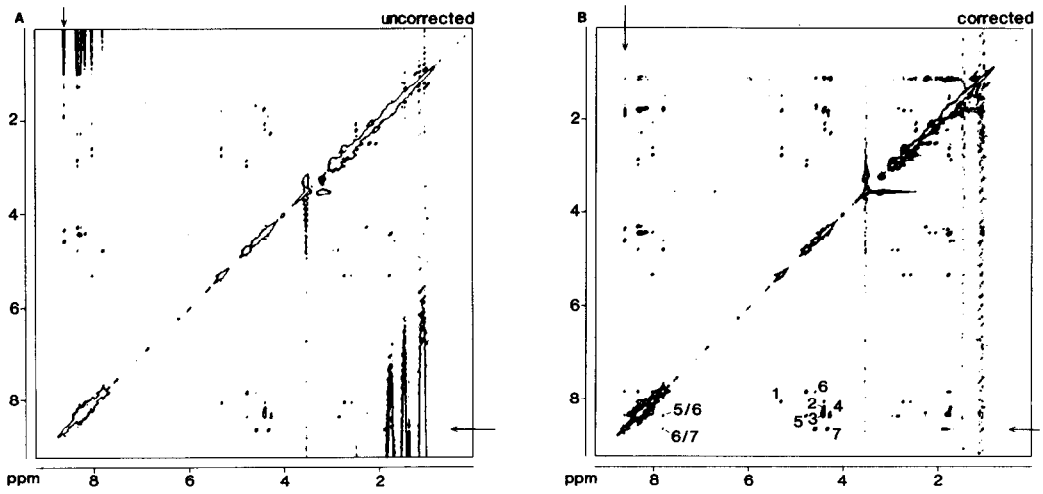

FIG. 2. Comparison of the uncorrected and corrected phase-sensitive 500 MHz NOESY spectra (200 ms mixing) recorded on surfactin in dimethyl sulfoxide at 27°C. Surfactin is a cyclic peptide with the following sequence (*19*):

$$CH_3(CH_2)_{10}-CH(=O)-CH_2-CO-\text{L-Glu}-\text{L-Leu}-\text{D-Leu}-\text{L-Leu}-\text{D-Leu}-\text{L-Asp}-\text{L-Val}$$

Spectra result from $2 \times 256 \times 1024$ data matrices recorded on an NT-500 spectrometer. After zero-filling, the digital resolution is 4.6 Hz (F_2) and 9.2 Hz (F_1). Only positive levels are plotted. Spectra A and B result from two identical experiments except for the initial t_1 increment (see text). One point in the F_1 dimension and two points in the F_2 dimension were estimated by LP extrapolation in the corrected spectrum (B). LP extrapolation in both dimensions required about 8 min on a Sun 3/110. For both data sets, 60°-shifted sine-bell functions were applied before FT and in both dimensions, the first data point was divided by a factor of 2 prior to FT. A linear phase correction of 53° was applied in F_1 for the uncorrected spectrum (A) but no such correction was needed for spectrum B. Arrows mark the positions of cross sections, shown in Fig. 3.

may be replaced by a δ pulse, followed by a delay $2\tau_{90°}/\pi$, where $\tau_{90°}$ is the duration of the 90° pulse (*20, 21*). Similarly, the second 90° pulse may be replaced by a δ pulse preceded by $2\tau_{90°}/\pi$. Effectively, the first t_1 interval of a 2D NOESY experiment corresponding to a time, T, is then given by

$$T = 4\tau_{90°}/\pi + t_1(0), \qquad [3]$$

where $t_1(0)$ is the initial (*programmed*) value of the variable evolution period. For the regular NOESY spectrum, $t_1(0)$ is set to its minimum value (1 μs). For the second data set, $t_1(0)$ is adjusted such that $T = \Delta_1$, in order to be able to reconstruct with LP a data point at $t_1 = 0$ ($t_1(0) = \Delta_1 - 4\tau_{90°}/\pi$). In both dimensions, the real and imaginary parts of the complex signal were restored independently using the following 30 points. In F_1, one missing pair was computed (10 poles were estimated by the Burg method); in F_2, two altered pairs were reconstructed (15 estimated poles). As expected, no linear phase correction in F_1 was needed for the spectrum of Fig. 2B.

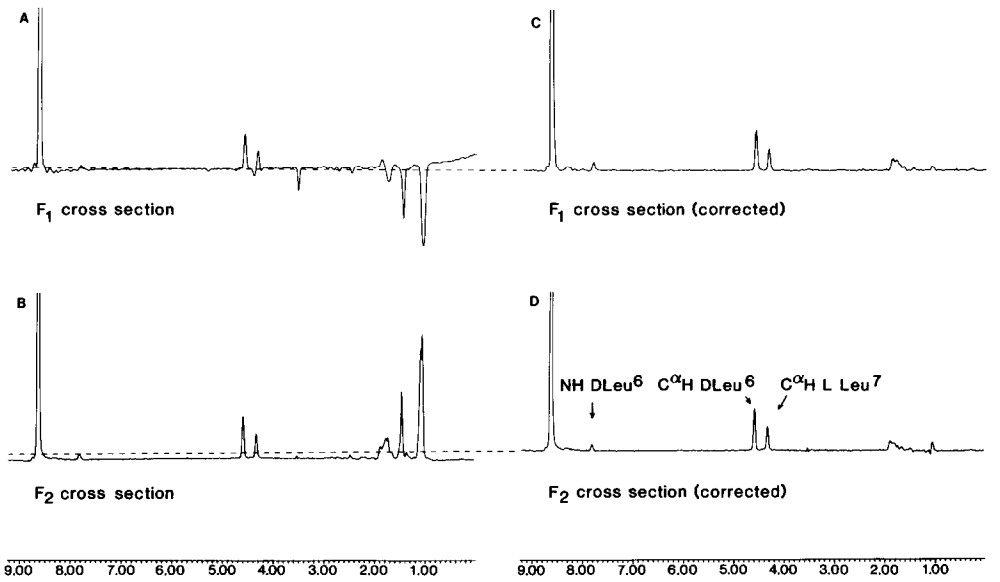

FIG. 3. F_1 and F_2 cross sections of Fig. 2A (A, B) and Fig. 2B (C, D), taken at the diagonal position of L-Leu^7 NH. The zero level is indicated by broken lines. The negative offset and the small curvature (not visible at the vertical scale shown) in B originate from the distorted values of the first two complex t_2 data points, caused by the time response of the filter; the curvature in A is caused by the linear phase correction in F_1 associated with the delayed acquisition. The large artifacts near 1.5 ppm in A and B result from baseline distortions on the sections that carry the intense methyl signals. In contrast to spectra A and B, the F_1 and F_2 sections (C and D) of the corrected 2D spectrum show nearly identical artifact-free profiles.

Both spectra shown in Fig. 2 are plotted with identical positive contour levels. The lowest is drawn at approximately 1.5 times the thermal noise level. Many of the NOE cross peaks not visible in Fig. 2A are clearly visible on both sides of the diagonal in Fig. 2B. The difference between the 2D spectra can be seen more clearly in Fig. 3, which compares F_1 and F_2 cross sections taken at the positions indicated by arrows in Fig. 2. The regular F_1 cross section (Fig. 3A) shows a curvature of the baseline due to the first-order phase correction (53° across the spectrum); the distortion of the F_2 cross section (Fig. 3B) consists mainly of a vertical offset related to the underestimation of the first data point. The artifacts in these spectra near 1.5 ppm originate from baseline distortion in the orthogonal dimension amplified by the large magnitude of the corresponding diagonal. After the LP reconstruction of the first data points, both the F_1 curvature and the F_2 offset have been removed completely.

These spectra provide evidence of the power of linear prediction for the suppression (before FT) of artifacts related to the early beginning of the 2D time-domain data. This permits the development of improved 2D and 3D phase-sensitive NMR experiments because the constraint of sampling at $t_1 = 0$ and $t_2 = 0$ is now removed. Applications related to solvent suppression and 2D and 3D schemes that utilize selective excitation are currently under investigation.

ACKNOWLEDGMENTS

D.M. acknowledges financial support from the Centre National de la Recherche Scientifique (France) and from a NATO fellowship. We thank Dr. Edwin D. Becker for useful suggestions during the preparation of the manuscript and Professor Marius Ptak (Orléans, France) and Georges Michel (Lyons, France) for the gift of surfactin used in this study.

REFERENCES

1. K. WÜTHRICH, "NMR of Proteins and Nucleic Acids," Wiley, New York, 1986.
2. G. OTTING, H. WIDMER, G. WAGNER, AND K. WÜTHRICH, *J. Magn. Reson.* **66,** 187 (1986).
3. D. MARION AND A. BAX, *J. Magn. Reson.* **79,** 352 (1988).
4. D. I. HOULT, C.-N. CHEN, H. EDEN, AND M. EDEN, *J. Magn. Reson.* **51,** 110 (1983).
5. P. PLATEAU, C. DUMAS, AND M. GUÉRON, *J. Magn. Reson.* **54,** 46 (1983).
6. R. E. KLEVIT, *J. Magn. Reson.* **62,** 551 (1985).
7. I. L. BARSUKOV AND A. S. ARSENIEV, *J. Magn. Reson.* **73,** 148 (1987).
8. D. S. STEPHENSON, *Prog. NMR Spectrosc.* **20,** 515 (1988).
9. S. SIBISI, *Nature (London)* **301,** 134 (1983).
10. E. D. LAUE, J. SKILLING, AND J. STAUNTON, *J. Magn. Reson.* **63,** 418 (1985).
11. A. E. SCHUSSHEIM AND D. COWBURN, *J. Magn. Reson.* **71,** 371 (1987).
12. M. A. DELSUC, F. NI, AND G. C. LEVY, *J. Magn. Reson.* **73,** 548 (1988).
13. H. GESMAR AND J. J. LED, *J. Magn. Reson.* **76,** 183 (1988).
14. H. GESMAR AND J. J. LED, *J. Magn. Reson.* **76,** 575 (1988).
15. W. H. PRESS, B. P. FLANNERY, S. A. TEUKOLSKY, AND W. T. VETTERLING, "Numerical Recipes: The Art of Scientific Computing," Chap. 12, Cambridge Univ. Press, Cambridge, 1986.
16. J. P. BURG, *in* "Proceedings, 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City, Oklahoma, October 31, 1967."
17. R. KUMARESAN AND D. W. TUFTS, *IEEE Trans. Acoust. Speech Signal Process.* **30,** 833 (1982).
18. S. M. KAY AND S. L. MARPLE, JR., *Proc. IEEE* **69,** 1380 (1981).
19. K. ARIMA, A. KAKIMURA, AND G. TAMURA, *Biochem. Biophys. Res. Commun.* **31,** 488 (1968).
20. R. R. ERNST, G. BODENHAUSEN, AND A. WOKAUN, "Principle of Nuclear Magnetic Resonance in One and Two Dimensions," pp. 119–124, Clarendon Press, Oxford, 1987.
21. A. BAX AND D. MARION, *J. Magn. Reson.* **78,** 186 (1988).