

## Improved Linear Prediction for Truncated Signals of Known Phase

GUANG ZHU AND AD BAX

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

Received June 22, 1990

It has been suggested that increasing the dimensionality of protein NMR spectra may be a more efficient way of simplifying protein NMR spectra than increasing the resolution of conventional 2D NMR spectra (1). Because of measuring time constraints, severe truncation of the NMR signals frequently occurs in at least one and often in two dimensions of 3D and 4D data sets. Fourier transformation of such signals leads to truncation artifacts, and it has been suggested that linear prediction (LP) algorithms are more suitable for treatment of such signals. Linear prediction can be used either for measuring the frequencies, amplitudes, damping factors, and phases of all components of the NMR signal (2-5) or, in a more conservative approach, for simply extending the time-domain data set, prior to Fourier transformation (5-9). In either case, no assumption is made about the shape of the signal apart from the assumption that it can be described by a sum of exponentially damped sinusoids.

In multidimensional NMR experiments the phase of the NMR signal is well known in all the indirectly detected dimensions (10-12). Also, if signals are severely truncated, the small amount of signal decay during the short time domain can be described adequately by a single damping factor. However, apart from simplifying the calculation, this prior knowledge is of no immediate use for conventional linear prediction analysis of the time domain. Here we demonstrate a new method that effectively uses this prior information to obtain a better estimation of the NMR spectrum.

In the case of forward linear prediction, a data point  $x_n$  is expressed as a linear combination of its  $K$  preceding data points:

$$x_n = \sum_{k=1}^K a_k x_{n-k}. \quad [1]$$

In the case where  $N$  complex data points have been sampled,  $N - K$  equations of type [1] can be generated, and the coefficients  $a_k$  can be determined by singular value decomposition (SVD) provided that  $K \leq N/2$ . Note that the maximum number of frequency components that can be determined by LP SVD from such an FID equals  $N/2$ .

In the absence of signal decay, and for the case where the time-domain signal has zero phase at time zero, the time domain can be extended to negative times. If  $x_0$  corresponds to the data point sampled at time 0, and  $x_1$  is the data point sampled

after one dwell time, negative time data points  $x_{-N+1} \cdots x_{-1}$  may be added, with  $x_{-n} = x_n^*$ , where  $x_n^*$  is the complex conjugate of  $x_n$ . This makes the length of the time domain nearly twice as long and allows as many as  $N - 1$  frequency components to be determined. For short durations of the time domain with severe truncation, as is often encountered in 3D and 4D NMR experiments, the decay of the time-domain data is small and can easily be compensated for by multiplying the data with a positive exponential prior to calculating the negative time data points. Provided the duration of the time-domain data is not longer than the average  $T_2$  of the NMR signals and provided the  $T_2$  of these signals does not vary by more than about 50% from one resonance line to another, such a procedure works well in practice.

A second case where the negative time data points can easily be obtained occurs if the sampling is delayed by exactly half a dwell time. Upon Fourier transformation such data would require a linearly frequency-dependent phase correction of exactly  $180^\circ$  across the spectrum (12). As above, the data points at negative time are again the complex conjugates of data recorded at positive time. Note that in this case,  $N$  data points recorded at positive time result in  $2N$  data points when the negative time-domain data are added, allowing the determination of up to  $N$  frequency components. For very short time-domain series, as may be encountered in multidimensional NMR, it is therefore preferable to start sampling with half a dwell time delay rather than starting at time zero.

For calculation of the prediction coefficients it is often better to use an overdetermined set of equations (2). For  $2N$  data points, we typically use  $K \approx 2N/3$  coefficients, resulting in  $4N/3$  equations of type [1]. With the prediction coefficients in hand, either we can proceed to calculate the future of the time domain signal (6-9) or we can determine the frequency components encoded in the coefficients,  $a_n$ , by calculating the roots of the polynomial

$$z^K - a_1 z^{K-1} - \cdots - a_K = 0. \quad [2]$$

Each frequency component corresponds to one of the roots,  $z_n$ . For a decaying frequency component, the root will fall within the unit circle; for a stationary signal, the root will be very close to the unit circle. Noise components may fall either within or outside the unit circle. Because in our procedure we multiply the time-domain signal by a positive exponential prior to the SVD routine, the signal roots will be very close to the unit circle. Especially for interferograms ("FIDs" in the indirectly detected time dimensions of a multidimensional NMR experiment) that do not contain any real signals, roots that fall significantly outside the unit circle are often found. If for these signals the future of the time domain were calculated without root determination, this would result in signals that show a rapid increase of noise components in the predicted part of the data. Although this does not become a problem if the data are extended by only a small fraction (9), if one attempts to double or triple the length of the time domain, noise roots outside the unit circle give rise to spikes in the frequency-domain spectrum upon Fourier transformation of the extended time domain. These noise spikes can be eliminated by reflecting all roots outside the unit circle to fall within the unit circle (13); i.e., a root  $z_n$  will be multiplied by  $|z_n|^{-2}$  if  $|z_n| > 1$ . Therefore, although calculating the roots of Eq. [2] is computationally intensive, we find this to be a necessary step in using linear prediction in an optimal manner.

The linear prediction method is illustrated for two cases, a simulated signal to which Gaussian noise is added and a slice of an experimental 3D triple-resonance NMR spectrum. Figure 1 shows the results obtained for the simulated FID, containing signals from a singlet, a doublet, a triplet, and a quartet, all with identical integrated intensities. The simulated FID contains 64 data points with a 1 ms dwell time, starting at time zero. The decay constant,  $T_2$ , was set to 100 ms for all signal components; i.e., the signal had decayed by  $e^{-0.63}$  at the end of the time domain. The  $J$  splitting for each of the multiplets was 10 Hz. Calculations are performed for a high and for a low signal-to-noise FID. Prior to any other operation, the time-domain signal is multiplied by  $e^{t/0.1}$ , to counteract the effect of decay of the signal. Figure 1a shows the Fourier transform of a high  $S/N$  FID, when apodized by a cosine bell and zero-filled to 512 complex data points. Figure 1b shows the spectrum obtained if the 64 data points are extrapolated to 256 data points by linear prediction, prior to cosine-bell filtering, zero filling, and Fourier transformation. Resolved multiplets are observed for the doublet and the triplet, but not for the quartet. Figure 1c shows the result obtained for the case where the negative time-domain data are also used as input for linear prediction. Thus, 63 data points corresponding to negative time were used in combination with the regular 64 data points to forward predict an additional 192 data points, resulting in a 319 data point time domain. The first 63 data points were subsequently removed prior to Fourier transformation and zero filling. The multiplet resolution visible in Fig. 1c is clearly improved relative to that in the spectrum of Fig. 1b. Figures 1d–1f show the analogous spectra, but now obtained for a low  $S/N$  FID. In this case it is also clear that use of the extended time-domain data offers the best resolution.

Figure 2 compares results obtained for an ( $F_1, F_3$ ) slice of the HCA(CO)N triple-resonance 3D spectrum of the protein calmodulin (14, 15). The 3D spectrum contains

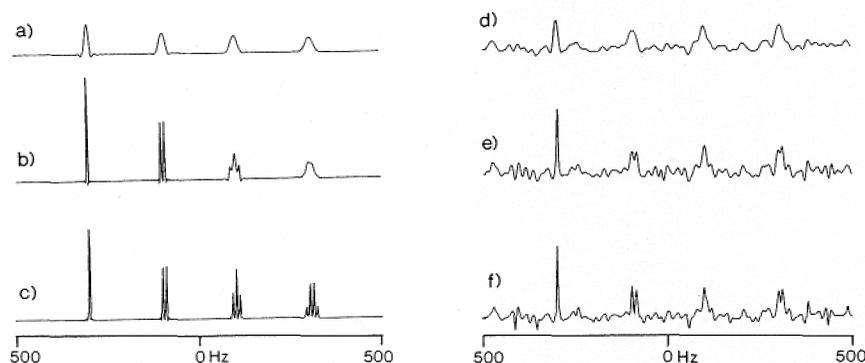


FIG. 1. Spectra obtained from a simulated truncated time-domain signal containing a singlet, a doublet, a triplet, and a quartet, each with  $J$  splittings of 10 Hz. The length of the time domain was 63 ms, the dwell time was 1 ms, and the decay constant,  $t_2$ , was 100 ms. Spectra a–c are obtained for high  $S/N$  data and spectra d–f for data with 130 times lower  $S/N$ . (a, d) Fourier transforms of the FID, after apodization with a cosine bell, 3 Hz exponential line narrowing, and zero filling to 512 complex data points. (b, e) Fourier transforms of the FID after 3 Hz exponential filtering of the original FID, linear prediction to 256 data points using root reflection, cosine-bell apodization, and zero filling to 512 data points. (c, f) Processed like b, e, but using the negative plus positive time-domain data points (127 complex data) as input for the forward linear prediction.

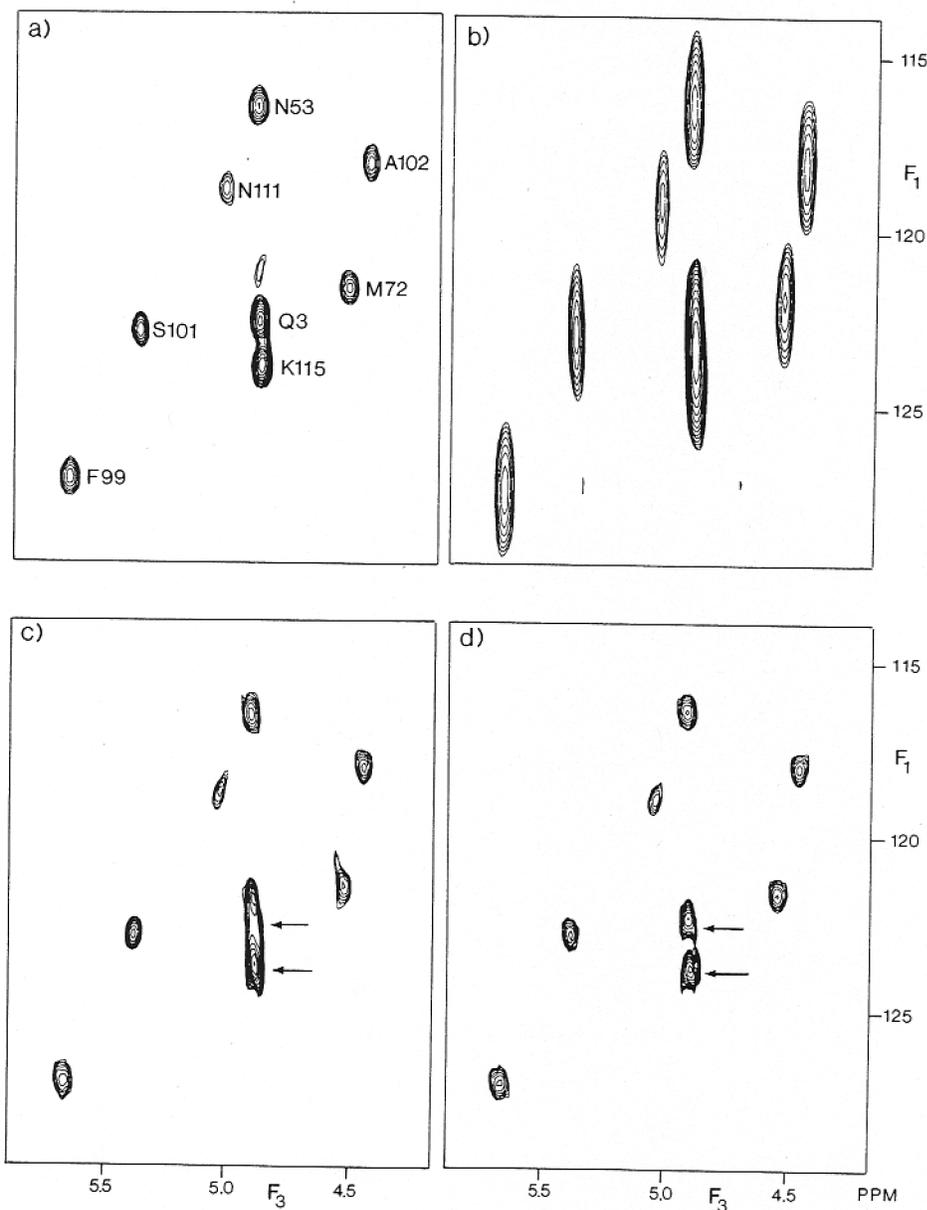


FIG. 2. ( $F_1, F_3$ ) slice of the triple-resonance 3D HCA(CO)N relay spectrum as originally shown in Ref. (15). Each of the panels shown has been processed identically in the  $F_3$  dimension. (a) Fourier transform obtained after 9 Hz exponential line narrowing followed by cosine<sup>2</sup>-bell apodization of 32 complex  $t_1$  time-domain data points and zero filling to 128. (b) Same as a, but using only the first 8  $t_1$  time-domain data points. (c) Spectrum obtained by 9 Hz exponential line narrowing, linear prediction of the first 8 data points out to 32 followed by cosine<sup>2</sup>-bell apodization, zero filling, and Fourier transformation. (d) Processed like c, but using the negative plus positive time-domain data points (16 complex data) for linear prediction of the additional 24 data points. The arrows in c and d mark the true  $F_1$  coordinate of the Q3 and K115 resonances as observed in a.

correlations between  $H\alpha$  ( $F_3$ ) and  $C\alpha$  ( $F_2$ ) of one residue and the  $^{15}\text{N}$  resonance ( $F_1$ ) of the following residue. Figure 2a shows the results if all 32 complex  $t_1$  domain data points are used as input for the Fourier transformation. Prior to  $F_1$  Fourier transformation, exponential line narrowing (9 Hz) and  $\cos^2$ -bell apodization are used. Figure 2b shows the result if only the first 8  $t_1$  data points are Fourier transformed, again with 9 Hz exponential line narrowing,  $\cos^2$ -bell apodization, and zero filling. Figure 2c shows the result obtained by linear prediction of the first 8  $t_1$  data points (after 9 Hz exponential line narrowing) to 32 data points, followed by  $\cos^2$ -bell apodization, zero filling, and Fourier transformation. Finally, Fig. 2d shows the result obtained by including the negative time domain, where 8 reflected data points are added at the beginning of the time domain prior to linear prediction. Note that in this experiment  $t_1$  sampling was delayed by exactly half a dwell time. The spectrum of Fig. 2d is clearly superior in quality to the spectrum obtained with regular linear prediction (Fig. 2c). Note that especially for the two partly overlapping resonances of K115 and Q3, the distortion is significantly less in Fig. 2d than in Fig. 2c. It should also be noted, however, that the peak positions as observed for Q3 and K115 are slightly shifted in the  $F_1$  dimension, relative to the spectrum of Fig. 2a, where 32 time-domain data points are used for processing. Although significantly less for Fig. 2d, such small shifts in peak position can have serious consequences for using the spectral information in an automated manner. Predicting the time-domain data to four times its original length, as demonstrated in Fig. 2, slightly exceeds the upper limit of what can be accomplished in practice for data sets with limited signal-to-noise ratios, even when using the reflected data at negative time. Predicting the time-domain data even further results in artificial resolution where peaks can be shifted from their true position by more than the linewidth.

We have demonstrated here that for severely truncated data of known phase, improved results can be obtained by linear prediction of a time domain to which data at negative time have been added. As discussed in detail elsewhere (10-12), by correctly adjusting experimental parameters it can be ensured that data are delayed by either zero or half a dwell time in the dimension where one wishes to use linear prediction, resulting in signals of known phase. In 3D NMR spectra, data are often severely truncated in the indirectly detected dimensions. If they are not, a substantial amount of data measuring time could be saved by allowing truncation and linear prediction of the time domain. For example, as demonstrated in Fig. 2, use of only 8 complex  $t_1$  data points gives a spectral quality that is comparable to that obtained by regular processing of 32 data points, thus allowing a factor of four savings in data recording time. Moreover, rapid recording of 3D spectra is a necessity for less stable proteins, which cannot withstand prolonged exposure to room temperature or higher. The penalty one has to pay for linear prediction is a substantial amount of computer time, especially when working with the extended time domain. At present we do not find this to be a major problem for most data sets; on a SUN Sparc-1 workstation, linear prediction in one dimension of a typical triple-resonance 3D spectrum (14, 15) requires between 8 and 32 h of data processing time. Considering that the cost of such a workstation is orders of magnitude lower than the cost of a high-field NMR spectrometer it appears reasonable to devote substantial computational power to optimal processing of the NMR data.

Because the new time-domain signal, to which data at negative time have been added, is exactly symmetric about time zero, there is no difference between forward and backward linear prediction. Therefore, it is not possible to obtain even better results by using our method in combination with the forward/backward linear prediction technique of Porat and Friedlander (16, 13). However, we believe use of negative time-domain data should be directly applicable to total least-squares linear prediction (17), a method that appears more robust for low  $S/N$  data. This application is currently under investigation.

## ACKNOWLEDGMENTS

We thank Drs. Lewis E. Kay and Mitsuhiro Ikura for assistance and Dr. Edwin D. Becker for useful suggestions during the preparation of this manuscript. This work was supported by the Intramural AIDS Targeted Antiviral Program of the Office of the Director of the National Institutes of Health.

## REFERENCES

1. L. E. KAY, G. M. CLORE, A. BAX, AND A. M. GRONENBORN, *Science*, in press.
2. H. BARKHUIJSEN, R. DE BEER, W. M. M. J. BOVEE, AND D. VAN ORMONDT, *J. Magn. Reson.* **61**, 465 (1985).
3. H. GESMAR AND J. J. LED, *J. Magn. Reson.* **76**, 183 (1988).
4. H. GESMAR AND J. J. LED, *J. Magn. Reson.* **76**, 575 (1988).
5. J. HOCH, in "Methods in Enzymology" (N. Oppenheimer and T. L. James, Eds.), Vol. 176, p. 216, Academic Press, San Diego, 1989.
6. R. R. ERNST, G. BODENHAUSEN, AND A. WOKAUN, "Principles of Nuclear Magnetic Resonance in One and Two Dimensions," p. 106. Clarendon Press, Oxford, 1987.
7. J. TANG AND J. R. NORRIS, *J. Magn. Reson.* **78**, 23 (1988).
8. Y. ZENG, J. TANG, C. A. BUSH, AND J. R. NORRIS, *J. Magn. Reson.* **83**, 473 (1989).
9. E. T. OLEJNICZAK AND H. L. EATON, *J. Magn. Reson.* **87**, 628 (1990).
10. A. BAX AND D. MARION, *J. Magn. Reson.* **78**, 186 (1988).
11. L. E. KAY, D. MARION, AND A. BAX, *J. Magn. Reson.* **84**, 72 (1989).
12. A. BAX, M. IKURA, L. E. KAY, AND G. ZHU, *J. Magn. Reson.*, in press.
13. M. A. DELSUC, F. NI, AND G. C. LEVY, *J. Magn. Reson.* **73**, 548 (1987).
14. M. IKURA, L. E. KAY, AND A. BAX, *Biochemistry* **29**, 4659 (1990).
15. L. E. KAY, M. IKURA, AND A. BAX, *J. Magn. Reson.*, in press.
16. B. PORAT AND B. FRIEDLANDER, *IEEE Trans. Acoust. Speech Signal Process.* **ASSP-34**, 1336 (1986).
17. C. F. TIRENDI AND J. F. MARTIN, *J. Magn. Reson.* **85**, 162 (1989).