Journal of Biomolecular NMR, **12:** 277–297, 1998.
KLUWER/ESCOM

Practical model fitting approaches to the direct extraction of NMR parameters simultaneously from all dimensions of multidimensional NMR spectra

Roger A. Chylla[1], Brian F. Volkman & John L. Markley*
National Magnetic Resonance Facility at Madison, Department of Biochemistry, University of Wisconsin-Madison, 420 Henry Mall, Madison, WI 53706, U.S.A.

Received 20 October 1997; Accepted 13 March 1998

Key words: CHIFIT, maximum likelihood, multidimensional NMR analysis, spectral deconvolution

Abstract

A maximum likelihood (ML)-based approach has been established for the direct extraction of NMR parameters (e.g., frequency, amplitude, phase, and decay rate) simultaneously from all dimensions of a *D*-dimensional NMR spectrum. The approach, referred to here as HTFD-ML (hybrid time frequency domain maximum likelihood), constructs a time-domain model composed of a sum of exponentially-decaying sinusoidal signals. The apodized Fourier transform of this time-domain signal is a model spectrum that represents the 'best fit' to the equivalent frequency-domain data spectrum. The desired amplitude and frequency parameters can be extracted directly from the signal model constructed by the HTFD-ML algorithm. The HTFD-ML approach presented here, as embodied in the software package CHIFIT, is designed to meet the challenges posed by model fitting of *D*-dimensional NMR data sets, where each consists of many data points (10^8 is not uncommon) encoding information about numerous signals (up to 10^5 for a protein of moderate size) that exhibit spectral overlap. The suitability of the approach is demonstrated by its application to the concerted analysis of a series of ten 2D ^{1}H-^{15}N HSQC experiments measuring ^{15}N T_1 relaxation. In addition to demonstrating the practicality of performing maximum likelihood analysis on large, multidimensional NMR spectra, the results demonstrate that this parametric model-fitting approach provides more accurate amplitude and frequency estimates than those obtained from conventional peak-based analysis of the FT spectrum. The improved performance of the model fitting approach derives from its ability to take into account the simultaneous contributions of all signals in a crowded spectral region (deconvolution) as well as to incorporate prior knowledge in constructing models to fit the data.

Introduction

The extraction of primary NMR parameters (e.g., frequency, amplitude, phase, and decay rate) from the acquired free induction decay (FID) is a prerequisite to further NMR data analysis. The traditional first step in this process is Fourier transformation (FT) of the FID into a suitably apodized and phase-corrected absorption spectrum (for a review see Hoch and Stern, 1996). Subsequent to this step, frequencies for desirable signals are typically obtained from interpolation of peaks whose heights exceed a specified threshold. Typically, the signal intensities are estimated either from the peak height or the integral of the peak, depending upon the experimental context.

Accurate estimation of signal intensities from peak height or peak integral measurements requires that nearby signals have negligible overlap in the frequency-domain spectrum. This requirement often is not met in NMR spectra of large molecules, where the signals have relatively broad linewidths. When working with multidimensional spectra (par-

[1]Present address: ADAC/Geometrics Corp., 6510 Grand Teton Place, Madison WI 53719, USA.
*To whom correspondence should be addressed. Tel. (608) 263-9349; fax (608) 262-3453; e-mail markley@nmrfam.wisc.edu.

ticularly 3D and 4D), the problem of spectral overlap is exacerbated by the application of apodization functions along the indirectly-detected dimensions. One approach to remedying the problem of poor frequency resolution involves extrapolating the time-domain data along the indirectly-detected dimensions to yield a resolution-enhanced spectrum upon subsequent Fourier transformation. Reported methods for extrapolation of the time-domain FID include linear prediction serially in one dimension (Gesmar and Led, 1989; Kay et al., 1992; Miller and Greene, 1989), 2D linear prediction (Zhu and Bax, 1992), *n*-dimensional Bayesian (Chylla and Markley, 1993), and *n*-dimensional maximum likelihood (ML) analysis (Chylla and Markley, 1995). Another approach is the direct replacement of the low-resolution data spectrum with one of greater resolution derived from frequency-domain maximum likelihood analysis (Wang et al., 1994) or maximum entropy reconstruction (Sibisi et al., 1984). Although these approaches employ different means of producing a resolution-enhanced absorption spectrum, they all ultimately rely upon peak-based measurements to obtain frequency and amplitude estimates.

Recent efforts have focused upon obtaining NMR parameter estimates from parametric signal models, as opposed to measurements of peak attributes. Model-based approaches that attempt to 'fit' the acquired experimental data to analytical signal functions provide a means for direct parameter extraction from the optimized signal models. The simultaneous contributions of numerous signals can be calculated by such methods, thus allowing one to obtain greater accuracy in parameter estimates when signals overlap in the frequency spectrum. Model-based methods have the additional advantage of providing a more systematic, automated approach to NMR parameter extraction. Two parametric approaches have been reported for estimating NMR parameters from one-dimensional time-domain data: Bayesian probability theory (Bretthorst, 1990; Fitzgerald et al., 1995) and use of the maximum likelihood principle (Miller and Greene, 1989; Umesh and Tufts, 1996).

The practical application of a model-based approach to parameter estimation from large, *D*-dimensional NMR spectra is a non-trivial exercise. Our previous approach (Chylla and Markley, 1995) to applying the maximum likelihood (ML) principle to parameter extraction from *D*-dimensional time-domain data was limited to modeling subsets of a *D*-dimensional experiment. For each data point along the acquisition dimension, a signal model was constructed to fit the corresponding (*D*-1)-dimensional FID. The signal models derived from this analysis were used to extrapolate each of the $(D-1)$-dimensional FIDs with synthetic data points prior to Fourier transformation, but only portions of the entire *D*-dimensional data set could be modeled.

We describe here an approach for applying the maximum likelihood principle to the extraction of NMR parameters *simultaneously* along *all* dimensions of a *D*-dimensional NMR spectrum. In this approach, line shapes present in the frequency-domain NMR spectrum are modeled by curves derived from theoretical time-domain signals (such as exponentially-decaying exponentials) that have been truncated, apodized, and Fourier transformed in the same manner as the time-domain NMR data. The proposed approach, referred to here as a hybrid time frequency-domain maximum likelihood (HTFD-ML) method, meets the challenges posed by model fitting of *D*-dimensional NMR data, i.e., large data sets (up to 10^9 data points) composed of numerous (up to 10^5) partially-overlapped signals. We present the theory and implementation of the HTFD-ML algorithm and illustrate how it can be applied to the concerted analysis of a series of 2D ^{1}H-^{15}N HSQC T_1 relaxation experiments carried out on a medium sized soluble protein (the carbon monoxide ligated form of the monomeric hemoglobin component IV, GMH4CO, from *Glycera dibranchiata*, 147 amino acids). The goals were (1) to test the practicality of the HTFD-ML approach for the analysis of multidimensional NMR data from a macromolecule in a situation where quantitative information about signal frequencies and amplitudes is essential, and (2) to compare the results with those of conventional analysis of the relaxation data.

Materials and Methods

Sample preparation

Glycera dibranchiata component IV monomeric hemoglobin (GMH4) was overexpressed and uniformly ^{15}N-labeled in *Escherichia coli*, purified and reconstituted with b-type hemin as described previously (Alam et al., 1998). The reduced CO-ligated form of GMH4 (GMH4CO) was produced by reduction of pure ferric GMH4 with an excess of sodium dithionite. ^{15}N relaxation measurements were

performed on a sample of uniformly ^{15}N -labeled GMH4CO (3.5 mm), buffered in 100 mm potassium phosphate, 100 mm KCl, pH 5.0 in 90% H_2O/10% D_2O.

NMR spectroscopy

^{15}N T_1 values were measured at 750.13 MHz ^{1}H frequency with a pulse scheme that utilized gradients for sensitivity enhancement and selective pulses for water flip-back (Farrow et al., 1994). All spectra were acquired with 16 scans per FID, 200 complex (200$*$) ^{15}N points and 1024 complex (1024$*$) ^{1}H points. Spectral widths were 10 000 Hz in the ^{1}H dimension and 2500 Hz in the ^{15}N dimension. T_1 values were obtained with relaxation delays $T = 10$, 60, 120, 200, 400, 800, 1400, and 2200 ms. Duplicate spectra were recorded at $T = 10$ and 200 ms for making precision estimates.

Noise added to synthetic spectra

Noise added to synthetic spectra was generated by collecting data without a sample at 500.13 MHz using a single 'hard' proton pulse. The noise was Fourier transformed using the same digital filters and zero-filling properties that were used to transform the corresponding synthetic data. The entire noise data set was then multiplied by a constant before being added to the synthetic frequency-domain spectra. The value of the constant was chosen to produce the desired signal-to-noise ratio of the spectral simulation.

Peak measurements

Measurements of peak position along each dimension were derived from a three-point parabolic interpolation of the extremum and its two adjacent frequency-domain values along each dimension. The quadratic coefficients (a,b,c) of a parabola ($ax^2 + bx + c$) that 'passes through' three equally spaced points, with values of ($p0$,$p1$, $p2$), are given, respectively, by ($\frac{p0+p2-2p1}{2a}$, $\frac{p2-p0}{2}$, $p1$).

The position of the peak was obtained from the position of the maximum of the parabola, χ_{max}, given by ($-b/2a$). The peak height is given by ($ax^2_{max} + bx_{max} + c$).

To minimize systematic overestimation of peak integrals caused by overlap among closely-spaced peaks in the frequency spectrum, all peak integrals were calculated as a simple sum over a fixed number of points centered about the peak maximum. The number of points was chosen to be $3/2$ the average width of a peak at half-height.

Maximum likelihood analysis

Maximum Likelihood (ML) estimation of frequency, amplitude, phase, and decay rate parameters was carried out by a computer program named *CHIFIT* using the approach and algorithm described, respectively, in the Theory and Algorithm sections. CHIFIT is written in C^{++} and runs on Silicon Graphics workstations operating under a version of IRIX 5.3-6.$*$. The program displays its graphics using X-Windows and thus can be run in a client-server configuration. *Chifit* reduces the need for very large amounts of random access memory (RAM) by loading only those portions of the spectrum that are actively being modeled. The RAM requirements depend on the number of signals contained in the data set: 32 MB of RAM suffices for data sets containing fewer than 10^3 signals; 128 MB of RAM is required for 10^3–10^4 signals; larger RAM capacity ($\sim$512 MB) is needed in order to process data sets with very large numbers of signals (10^5). The academic version of the software is available from the National Magnetic Resonance Facility at Madison (http://www.nmrfam.wisc.edu/software.html), and the commercial version can be purchased from Spectrum Research (http://www.specres.com).

Fourier transforms and plotting

Conventional FT processing was performed using the commercial software package Felix95 (Molecular Simulations, Inc., San Diego, CA, USA). Details about window functions, zero-filling, and other processing methods are documented in the appropriate figure legends. The multidimensional data were loaded and processed into Felix matrices using a submatrix format that is accessed directly by the CHIFIT software. The contour plots appearing in the figures were generated by Felix95.

Theory

In previous work (Chylla and Markley, 1995), we presented a theory for using the maximum likelihood principle to perform signal analysis of multidimensional time-domain data. The focus of this section is to present an adaptation of this to HTFD-ML analysis, and to outline the approach used to implement HTFD-ML analysis to perform efficient signal recognition

and parameter estimation on large *D*-dimensional NMR data sets.

Maximum likelihood principle

A fundamental task in signal processing is to determine what type of model functions should be used to describe the data and what values should be chosen for the free parameters of the model. Given a vector of N discretely sampled data points in time t

$$y(t) = \left[y(t_1), y(t_2), \ldots, y(t_N)\right] \quad 1 \leq i \leq N \quad (1)$$

where each data point $y(t_i)$ is the sum of a systematic signal component $f(t_i)$ and a random noise component $\delta(t_i)$,

$$?y(t_i) = f(t_i) + \delta(t_i) \quad (2)$$

the systematic component can be described using a parametric model $f(t \mid P)$. The maximum likelihood (ML) principle expresses the probability of a given model and set of parameters P in terms of the likelihood $l(f \mid y, P)$. Discarding any terms independent of P, the log-likelihood $\log[l(f \mid y, P)]$ is given by

$$\log\left[l\,(f \mid y, P)\right] \propto \left(\frac{1}{\sigma^2}\right) \sum_{i=1}^{N} y(t_i) f(t_i \mid P) - \left(\frac{1}{2\sigma^2}\right) \sum_{i=1}^{N} \left[f(t_i \mid P)\right]^2 \quad (3)$$

where σ^2 is the variance (a constant that for the time being is assumed to be known) of the random noise $\delta(t)$. Thus, according to the ML principle, the most likely set of parameters P given a model and the data is the P that maximizes $\log[l(f \mid y, P)]$.

Time domain maximum likelihood analysis

The maximum likelihood principle addresses the issue of *parameter estimation*, i.e., what values should be chosen for the set of free parameters P in a specified model function, but does not directly address the question of *model selection*, i.e., what type of model function $f(P)$ should be chosen to describe the data. The choice of appropriate model functions to describe NMR data depends fundamentally upon whether the experimental data are modeled in the time domain (FID) or the frequency domain. A detailed explanation of a *D*-dimensional approach to model fitting NMR data in the time domain has been reported previously (Chylla and Markley, 1995). The salient aspects of that theory are reformulated here.

The systematic portion of the signal at a given data point $f(t_i)$ is described as the linear combination of a set of $1 \leq j \leq J$ signal functions $V(t_i \mid \Omega_j)$ with non-linear parameters Ω_j

$$y(t_i) = \sum_{j=1}^{J} A_j V(t_i \mid \Omega_j). \quad (4)$$

Each multidimensional signal function $V(t_i \mid \Omega_j)$ is the D-dimensional product of a one-dimensional signal function $U(t_{id} \mid \Omega_{jd})$.

$$V(t_i \mid \Omega_j) = \prod_{d=1}^{D} U(t_{id} \mid \Omega_{jd})) \quad (5)$$

where $U(t_{id} \mid \Omega_{jd})$ is given by

$$U(t_{id} \mid \Omega_{jd}) = e^{i\omega_j dt_{id}} e^{i\phi_j dt_{id}} e^{-\alpha_{jd} t_{id}}. \quad (6)$$

The basis function $e^{i\omega_j dt_{id}}$ is a complex sinusoid associated with the *j*th signal along dimension *d*. It has an angular frequency given by ω_{jd}. Analogously, $e^{i\phi_j dt_{id}}$ is a complex phasor with a phase given by ϕ_{jd}, and $e^{-\alpha_{jd} t_{id}}$ is an exponential with a decay rate given by α_{jd}. The symbol t_{id} represents the time along dimension d of the ith data point. Each index i corresponds to a unique D-dimensional time coordinate $[t_{i1}, t_{i2}, \ldots, t_{iD}]$. Ω_{jd} is the set of non-linear parameters $[\omega_{jd}, \phi_{jd}, \alpha_{jd}]$.

If the number of signals J and the values for the non-linear frequency, phase, and decay rate parameters (Ω_j) of each signal are known, then $V(t_i \mid \Omega_j)$ can be calculated and the maximum likelihood values for the amplitudes A_j of the signals are given by

$$\frac{\partial\left(\sum_{i=1}^{N} (y(t_i) - f(t_i))^2\right)}{\partial A_j} = 0. \quad (7)$$

Substitution of the value of $f(t_i)$ contained in Equation 4 yields:

$$\frac{\partial\left(\sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{J} A_k V(t_i \mid \Omega_k)\right)^2\right)}{\partial A_j} = 0 \quad (8)$$

The solution to the vector of amplitudes A_j appearing in Equation 8 is given by the matrix equation

$$A_j e(t)_{jk} = h(t)_j, \quad (9)$$

where $e(t)_{jk}$ is the 'interaction matrix',

$$e(t)_{jk} \equiv \sum_{i=1}^{N} V(t_i \mid \Omega_j) V(t_i \mid \Omega_k), \quad (10)$$

and $h(t)_j$ is the projection of the data upon the signal functions,

$$h(t)_j \equiv \sum_{i=1}^{N} y(t_i) V(t_i \mid \Omega_j). \quad (11)$$

The solution vector is found by multiplication of the projection vector $h(t)_j$ by the matrix inverse of $e(t)_{jk}$,

$$A_j = h(t)_j \left[e(t)_{jk}\right]^{-1}. \quad (12)$$

Equation 12 provides an expression for finding the maximum likelihood amplitudes that are consistent with the data $y(t_i)$, a set of J signal functions $V(t_i \mid \Omega_j)$, and the parameters of each signal, Ω_j. From an initial estimate of J and Ω_j, the maximum likelihood values for A_j can be found by linear least squares analysis (Equation 12), and the sufficient statistic $\hbar^2(t)$ can be calculated according to Equation 13.

$$\hbar^2(t) = \sum_{j=1}^{J} A_j h(t)_j. \quad (13)$$

The sufficient statistic $\hbar^2(t)$ is a value that both maximum likelihood and Bayesian probability theory (Bretthorst, 1990) predict to be an indicator of the likelihood of the set of non-linear parameters Ω (Chylla and Markley, 1995). The most likely Ω is the set of values that maximizes $\hbar^2(t)$. It is more efficient to maximize $\hbar^2(t)$ than to minimize chi squared, because the former can be calculated directly from the dot product given by Equation 13. This equation thus serves as a basis for performing non-linear least squares optimization of Ω to achieve the maximum value of $\hbar^2(t)$.

Frequency domain maximum likelihood analysis

If the experimental data are modeled in the frequency domain, each data point $f(\varpi_i)$ in the data vector now represents a discrete D-dimensional frequency coordinate. The appropriate signal functions in the frequency domain are now Lorentzian functions of the form:

$$U(\varpi_{id} \mid \Omega_{jd}) = A_j \left[\frac{e^{-\phi_{jd}}}{\gamma_{jd}}\right] \left[\frac{1}{1 - \frac{i}{\gamma_{jd}}(\omega_{id} - \omega_{jd})}\right] \quad (14)$$

where $\Omega_{jd} = [\omega_{jd}, \phi_{jd}, \gamma_{jd}$ refers to the respective angular frequency, phase, and linewidth of signal j along dimension d.

In all other respects, frequency-domain ML analysis is analogous to time-domain analysis: i.e., Equations 7–13 are still applicable. As with time-domain analysis, the sufficient statistic is calculated for a set of J signals with known Ω_{jd} and can be used as a basis for non-linear optimization of each $\Omega_{jd} = [\omega_{jd}, \phi_{jd}, \gamma_{jd}]$.

Hybrid time-frequency domain maximum likelihood (HTFD-ML) analysis

In this section we present an approach that involves elements of both time and frequency domain analysis. We briefly evaluate the strengths and limitations of the two methods and discuss why a hybrid approach seems appropriate for the practical application of maximum likelihood analysis to large D-dimensional data sets.

The strengths of time-domain analysis are the appropriateness of its model functions for describing the data and the amenability of the method to the use of weighting functions. Weighting functions are useful for either increasing signal-to-noise ratios or increasing frequency resolution. Figure 1 illustrates the differences in resolution obtained by applying different apodization functions to the same synthetic data set which contains three signals with added random noise. One of the signals is isolated in frequency space, and the other two form a closely spaced pair. The parameters associated with this three-signal model are given in Table 1. The absorption spectrum obtained by processing the synthetic FID with a shifted cosine-squared-bell weighting function (Figure 1B) shows far better frequency resolution than that obtained with a pure cosine-squared-bell weighting function (Figure 1A). Resolution-enhancing apodization functions such as that used in Figure 1B can be applied readily to the data and to the basis functions of Equation 6 when using a time-domain maximum likelihood approach.

The central weakness of time-domain analysis is that its basis functions are delocalized such that an accurate projection of the data upon the basis functions requires calculation of multidimensional dot products over the entire data set. Although this fact does not prevent practical application of time-domain ML analysis to 1D NMR data sets and small D-dimensional NMR data sets with few signals (Chylla and Markley, 1995), the computational cost of these dot products becomes prohibitive when the size of the data set surpasses 10^5 points or when the number of signals surpasses 10^3. Any attempt to reduce the size

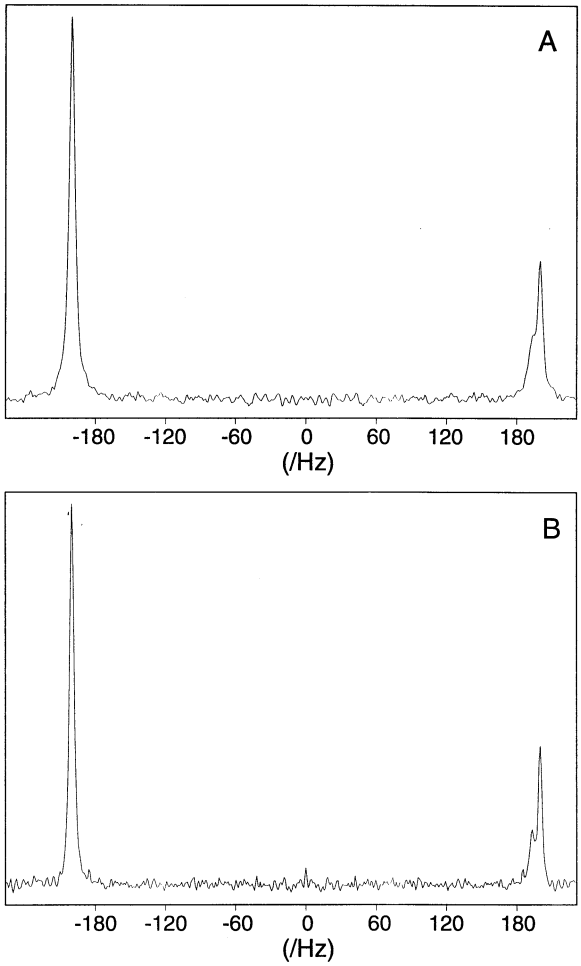

Figure 1. Absorption spectra consisting of three theoretical signals with added noise. A synthetic time-domain FID of length 512 complex points was constructed from the sum of three exponentially-decaying sinusoidal signals as described in Table 1. (A) Spectrum derived from the synthetic FID by application of a 90° shifted sine-squared bell window function (512 complex points) and zero-filling (to a length of 1024 complex data points) prior to Fourier transformation. (B) Spectrum derived in the same manner except that the window function was shifted by 60°. Comparison of the spectra in A and B clearly shows the utility of resolution-enhancing functions in resolving closely spaced signals.

Table 1. Parameters used in producing a synthetic model free induction decay (FID) containing three signals[a]

Signal number	Amplitude (/arbitrary unit)	Frequency (/Hz)	Line width (/Hz)
1	100	−200	17
2	32	200	15
3	16	194	16

[a] A synthetic time-domain FID of length 512 complex points was constructed from the sum of three exponentially-decaying sinusoidal signals using the parameters shown above: random noise was added to the synthetic FID.

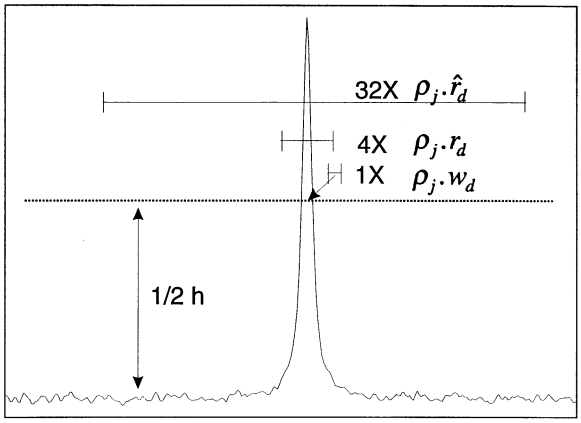

Figure 2. Definition of the 'optimization' and 'signal' regions. The figure shows how the sizes of the signal ($\rho_j \cdot \hat{r}_d$) and optimization ($\rho_j \cdot r_d$) regions along a given dimension d are derived from the width of the corresponding peak j at half peak height ($\rho_j \cdot w_d$).

of the dot products by calculating the projection of the basis functions over just a portion of the data set will result in a loss of frequency resolution. The nature of the time-domain signal is that its frequency information is distributed evenly over the full span of the FID (delocalization).

In contrast to time-domain ML analysis, the information about a particular signal in the frequency-domain absorption spectrum is heavily localized around the frequency coordinate corresponding to the angular frequency of the signal. Consequently, a strength of frequency domain analysis is that the information about the amplitude of a signal can be approximated very accurately by a projection of the Lorentzian basis function over only a small portion of the entire spectral width. The time (t_c) required to compute h_j (Equation 11) is linearly proportional to the size of the data set. The reduction in t_c of frequency-domain analysis vs. time domain analysis is thus given by

$$t_c \propto \prod_{d=1}^{D} \left(r_d/\hat{r}_d\right) \tag{15}$$

where r_d is the 'optimization region' along dimension d, and $\hat{r}_d$ is the full spectral width along dimension d (see Figure 2). The fractional reduction in t_c is experiment specific, but is of the order of 10^{-2}–10^{-4} for $D = 2$ and 10^{-3}–10^{-6} for $D = 3$.

A significant drawback of a pure frequency-domain approach is that the closed-form expression

for a discrete Lorentzian (Hoch and Stern, 1996, p. 30) is only a valid basis function to describe the frequency-domain spectrum if the time-domain data have been sampled uniformly and have been digitally filtered with exponential weighting functions. This limitation is noteworthy in cases where it is desirable to use digital filters that increase frequency resolution. As shown in Figures 1A–1B, the use of such non-exponential weighting functions can be strongly advantageous for processing overlapped spectra. Because the shape of the signal in the frequency domain is determined by the characteristics of the time-domain processing (uniform vs. non-uniform data sampling, type of window function, extent of zero-filling, etc.), the analytical frequency-domain basis functions in a pure frequency-domain approach must be modified to model the signal characteristics (line shapes) produced by each processing procedure.

Taking these considerations into account, it is desirable to adopt an approach that combines strengths of both the pure time domain and pure frequency-domain methods. Since the Fourier transform is a linear operation, the peak height (or integral) of a signal in the frequency domain is linearly proportional to its amplitude in the time domain. The calculation of h_j (Equation 11) using the time domain basis functions $U(t_i \mid \Omega_{jd})$ (Equation 6) can thus be approximated accurately by the process described as follows.

Let $F[s(t)]$ denote a digital operator that transforms a 1D complex time-domain vector $s(t)$ into a phase-sensitive absorption spectrum $s(\varpi)$.

$$F[s(t)] = s(\varpi). \tag{16}$$

This operator can be extended to the D-dimensional case by defining a D-dimensional operator that consists of a series of D separate $F[s(t)]$ operators applied consecutively along each dimension. For the remainder of this section, we will limit our discussion to the one-dimensional case to avoid cumbersome notation.

Let $\hat{F}[s(t)]$ refer to a specific $F[s(t)]$ operator that transforms the discrete time-domain FID $y(t)$ into the discrete frequency domain spectrum $y(\varpi)$:

$$\hat{F}[y(t)] = y(\varpi). \tag{17}$$

The $\hat{F}[s(t)]$ operator can also be used to transform the discrete time-domain basis functions $U(t \mid \Omega_{jd})$ into discrete frequency-domain basis functions $U(\varpi \mid \Omega_{jd})$:

$$\hat{F}\left[U(t \mid \Omega_{jd})\right] = U(\varpi \mid \Omega_{jd}). \tag{18}$$

In the algorithm described in the following section, the $\hat{F}(t)$ operators consist of the following operations:

(1) Multiplying the complex $s(t)$ by an apozidation function $\upsilon(t)$ of length l_d (l_d is usually just the length of $s(t)$).
(2) Zero-filling the data from l_d to a length l_f such that l_f is a power of two and $l_f \geq 2l_d$.
(3) Fourier transforming the data using a Fast Fourier (FFT) algorithm.
(4) Extraction of the real portion of the data to form a vector $s(\omega)$ of length l_f.

As a consequence of the linearity of $\hat{F}(s(t))$, the systematic portion of the frequency domain signal $y(\varpi_i)$ can be described by:

$$y(\varpi_i) = \sum_{j=1}^{J} A_j V(\varpi_i \mid \Omega_j) \tag{19}$$

where

$$V(\varpi_i \mid \Omega_j)_i = \prod_{d=1}^{D} U(\varpi_{id} \mid \Omega_{jd}) \tag{20}$$

and $U(\varpi_{id} \mid \Omega_{jd})$ is given by Equation 18. The maximum likelihood amplitudes A_j can now be found from the solution of the matrix equation:

$$A_j e(\varpi)_{jk} = h_j(\varpi) \tag{21}$$

where

$$e(\varpi)_{jk} \equiv \sum_{i=1}^{N} V(\varpi_i \mid \Omega_j) V(\varpi_i \mid \Omega_k) \tag{22}$$

and

$$h_j(\varpi) \equiv \sum_{i=1}^{N} y(\varpi_i) V(\varpi_i \mid \Omega_j). \tag{23}$$

A new sufficient statistic, $\hbar^2(\varpi)$, can now be formulated.

$$\hbar^2(\varpi) = \sum_{j=1}^{J} A_j h_j(\varpi). \tag{24}$$

The reformulation of the sufficient statistic in Equation 24 allows non-linear optimization of the parameters in the time-domain basis functions $U(t \mid \Omega_{jd})$ (Equation 6) using the frequency-domain projections defined by Equation 23. The advantage of this reformulation becomes apparent if the basis function

of length n_d is truncated to the length r_d. This substantially reduces the time t_c (Equation 15) required to compute the sufficient statistic without a significant loss in signal information. When the data are multidimensional, the reduction in time required to calculate $e(\varpi)_{jk}$ for a set of smaller basis functions $V'(\varpi \mid \Omega_{jd})$ and to project the basis functions over the corresponding reduced data set more than compensates for the additional time required to Fourier transform the basis functions (Equation 18).

Algorithm for HTFD-ML analysis of multidimensional NMR data

A software application named *CHIFIT*, which is written in C^{++} and runs under IRIX 5.3–6.4, implements the algorithm that incorporates the theoretical formulations of HTFD-ML analysis described in the preceding section and in our earlier work (Chylla and Markley, 1995).

(1) A phase sensitive absorption spectrum $y(\varpi)$ is created from conventional Fourier processing of the time-domain data $y(t)$. The $\upsilon(t)$ window functions and zero-filling properties along each dimension define a Fourier operator $\hat{F}[s(t)]$.

(2) To record the general signal characteristics of the spectra, a set of J 'trial signals' is found in the spectrum. Any non-overlapped signal with good signal-to-noise characteristics is a suitable member for the set of trial signals. Peaks corresponding to these signals are located in the spectrum. In the context of the CHIFIT algorithm, a peak p_j is any discrete point in the absorption spectrum which has an extremum whose absolute value (a) is above a required threshold, and (b) is greater than any point contained within the D-dimensional region formed by the center of the peak $p_j \cdot c_d$ and the half-peak width $p_j \cdot w_d$ along each dimension $1 \geq d \geq D$, where $p_j \cdot c_d$ is obtained from a 3-point parabolic interpolation of the points adjacent to the center (along dimension d), and $p_j \cdot w_d$ is derived from the interpolated width of p_j at half-height (along dimension d).

(3) Two regions, an 'optimization' and a 'model' region (r_j and $\hat{r}_j$ respectively), are defined for each signal j associated with peak p_j (Figure 2). The optimization region r_j is the product of all points defined by a segment r_{jd} along each dimension

$$r_j \equiv \prod_{d=1}^{D} r_{jd} \tag{25}$$

where r_{jd} is the set of points given by

$$r_{jd} \equiv \{k_{jd}\} \ (c_{jd} - 2w_{jd}) \leq k_{jd} \leq (c_{jd} + 2w_{jd}). \tag{26}$$

The larger model region, $\hat{r}_j$, is given by

$$\hat{r}_j \equiv \prod_{d=1}^{D} \hat{r}_{jd} \tag{27}$$

$$\hat{r}_{jd} \equiv \{\hat{k}_{jd}\} \ (c_{jd} - 16w_{jd}) \leq \hat{k}_{jd} \leq (c_{jd} + 16w_{jd}). \tag{28}$$

(4) Appropriate sinusoidal basis functions are chosen to describe the signal along each dimension. If the absorption spectrum is well phased along dimension d, then the phasor basis function is omitted from $U(t_{id} \mid \Omega_{jd}$ (Equation 6). Similarly, if the data are acquired with 'constant-time' evolution periods along dimension d, then the exponential decay function may be omitted from $U(t_{id} \mid \Omega_{jd})$.

(5) The frequency, phase (if applicable), and decay rate (if applicable) parameters along each dimension are set to their initial values. The starting frequency values are the angular frequencies corresponding to position c_{jd} in the frequency spectrum. The phases are initialized to zero. The decay rates are initialized to an arbitrary value of -0.01 which is equivalent to a linewidth that is 0.01 of the sweep width along the relevant dimension. Because the trial signals are isolated from other signals in the spectrum, the ability to find the maximum likelihood value of the decay rate parameters is insensitive to the accuracy of the initial decay rate estimates.

(6) Given the $\Omega_{jd} = [\omega_{jd}, \phi_{jd}, \gamma_{jd}]$ obtained from the previous step, the basis functions for trial signal j along dimension $dU(t_{id}|\Omega jd)$ can be calculated according to Equation 6.

(7) The frequency-domain basis functions, $U(\varpi_d \mid \Omega_{jd})$, are obtained from applying the corresponding $\hat{F}$ operators to the time-domain basis functions $U(t_d \mid \Omega_{jd}$. The resulting basis functions are truncated along each dimension d from a length of n_d to the corresponding length r_{jd} (Equation 26) to obtain the truncated D-dimensional basis function, $V'(\varpi_i \mid \Omega_j)$.

(8) The sufficient statistic $\hbar^2$ and the set of maximum likelihood amplitudes A_j consistent with the data and $V'(\varpi_i \mid \Omega_j)$ are obtained from Equations 21–24.

(9) For any set of Ω_j associated with the set of signals, $\hbar^2$ can be calculated according to steps 3–8. The set of truncated basis functions which comprise the current model form a series of signal networks with each network containing a set of one or more overlapping signals. A gradient-search algorithm is applied to find the set of Ω_j in each network which maximizes the local $\hbar^2$. The gradient-search algorithm employs a modified Marquardt approach (Marquardt, 1963) to find the local maximum in the $\hbar^2$ that exists in the vicinity of the parameter space determined by each network's set of Ω_j.

(10) The set of trial signals and their associated Ω are used to derive a set of trial model characteristics that are used to analyze the remaining signals contained in the data set.

(11) A threshold value τ_{min} (the minimum absolute value that a peak must have in the absorption spectrum for its corresponding signal to be modeled) is selected by the user on the basis of visual inspection of the absorption spectrum.

(12) A *D*-dimensional signal model is constructed from the current set of model signals. Initially, the set of model signals will be just the trial signals obtained from steps 1–10. The signal model is formed by the linear combination of the amplitude-weighted, truncated frequency-domain basis functions associated with each model signal. The region associated with each signal, however, is not the 'optimization region' (Equations 25–26) but the larger 'model region' (Equations 27–28). The signal model is subtracted from the frequency-domain data to yield a residual spectrum.

(13) All peaks that satisfy the threshold and overlap criteria defined in step 2 are located in the residual spectrum. The peaks must have an extremum whose absolute value is greater than τ_{min} and is also greater than the value of any points within the region defined by the center of the peak and the trial peak width $\bar{\rho} \cdot w_d$ (see Table 2) along each dimension *d*.

(14) A signal *j* is constructed for each peak obtained from step 13. The signal is assigned an optimization region (r_j) and a model region ($\hat{r}_j$) on the basis of the peak center $\rho_j \cdot c$ and the trial peak width $\bar{\rho} \cdot w$. The form of the basis functions for the nascent signals are the same as those chosen for the set of trial signals. The frequency coordinate of the signal is initialized to the angular frequencies corresponding to $\rho_j \cdot c$. Any phase and decay rate parameters (as well as their constraints) are initialized to the trial phase and decay rate parameter values that are obtained from optimization of the trial signals.

(15) Steps 6–9 are used to perform linear optimization of the amplitudes and non-linear optimization of the parameters associated with these signals and all other signals currently contained in the signal model.

(16) The algorithm loops back to step 12 and cycles through steps 12–15 until no additional valid peaks are obtained in step 13.

Results

Application to a 1D synthetic spectrum

The synthetic 1D model spectra, whose stepwise analysis by application of the HTFD-ML algorithm is depicted in Figure 3, contain a spacing of 4 Hz between signals 2 and 3 (the spectra in Figure 1 have a spacing of 6 Hz between peaks 2 and 3, but otherwise are identical). The upper right panel is a plot of the initial CHIFIT model, which contains no signals. The residual corresponding to this initial condition is simply the input spectrum (upper left panel).

The first stage of peak picking finds evidence for two signals. The frequency estimates for these signals are obtained from parabolic interpolation of the peak maxima. The initial decay rates for the two signals are assigned an arbitrary value of -0.01. Amplitude estimates consistent with these frequency and decay rate values are obtained from steps 6–8 of the algorithm described in the previous section. Holding the frequency estimates constant, the maximum likelihood values for the decay rates for this two-signal model are obtained from optimization of the sufficient statistic as described in step 9. The two-signal model, with optimized amplitudes and decay rates, is shown in the middle right frame of Figure 3. The corresponding residual is shown in the middle left frame. This residual spectrum contains evidence for an additional signal which is added to the model. The three-signal model (not shown) obtained after optimization of the decay rates of all three signals yields a residual that does not contain evidence for another signal. A final optimization of both the decay rates *and* the frequencies for all three signals is performed and yields the optimized model shown in the lower right frame of

Table 2. Analysis of errors in (top) frequency and (bottom) amplitude parameters extracted from a one-dimensional synthetic data set by HTFD-ML analysis and by conventional peak analysis[a]

Frequency parameters

Symbol	Description	Error (/Hz) from ML Analysis	Error (/Hz) from Peak Analysis
A-F1	Accuracy of frequency 1 for all spectra	0.007	0.016
A-F2	Accuracy of frequency 2 for all spectra	0.086	0.082
A-F3	Accuracy of frequency 3 for all spectra	0.187	0.268
A-LRF3	Accuracy of frequency 3 for all low resolution spectra	0.364	0.603
A-LSF3	Accuracy of frequency 3 for all spectra with low sensitivity	0.398	0.448
P-F1	Precision of frequency 1 for all spectra	0.007	0.009
P-F2	Precision of frequency 2 for all spectra	0.070	0.068
P-F3	Precision of frequency 3 for all spectra	0.151	0.206
P-LRF3	Precision of frequency 3 for all low resolution spectra	0.130	0.147
P-LSF3	Precision of frequency 3 for all spectra with low sensitivity	0.378	0.442

Amplitude parameters

Symbol	Description	Error from ML Analysis (/%)	Error from Peak Analysis (/%)
A-A2	Accuracy of A2/A1 for all spectra	7.9	16.3
A-A3	Accuracy of A3/A1 for all spectra	8.0	22.4
A-LRA3	Accuracy of A3/A1 for all low resolution spectra	20.1	102.8
A-LSA3	Accuracy of A3 /A1for all spectra with low sensitivity	13.1	23.4
P-A2	Precision of A2/A1 for all spectra	4.2	2.2
P-A3	Precision of A3/A1 for all spectra	4.7	3.7
P-LRA3	Precision of A3/A1 for all low resolution spectra	4.9	2.9
P-LSA3	Precision of A3/A1 for all spectra with low sensitivity	11.0	9.8

[a] A set of 256 different synthetic data sets was constructed as described in the text. To each of the 256 data sets, a series of 8 different noise spectra was added to create a total of 2048 data sets. HTFD-ML analysis was conducted on each of the 2048 data sets.

[b] The term ‘accuracy’ as used here in the context of frequency estimates represents the standard deviation of the estimated frequencies (obtained for each of the 8 data sets) from the known frequency value. The term ‘precision’ as used in this context represents the standard deviation of the estimated frequencies (obtained for each of the 8 data sets) from the mean frequency value.

[c] The amplitude estimates of signals *2* and *3* obtained from ML analysis were derived from the ratio of the respective amplitudes to the amplitude of signal *1*. The amplitude estimates of signals *2* and *3* obtained from peak analysis were derived from the ratio of the respective peak areas to the peak area associated with signal *1*. The term ‘accuracy’ as used here in the context of amplitude estimates represents the standard deviation of the estimated amplitude ratio (obtained for each of the 8 data sets) from the known amplitude ratio. The term ‘precision’ as used in this context represents the standard deviation of the estimated amplitude ratio (obtained for each of the 8 data sets) from the mean amplitude ratio.

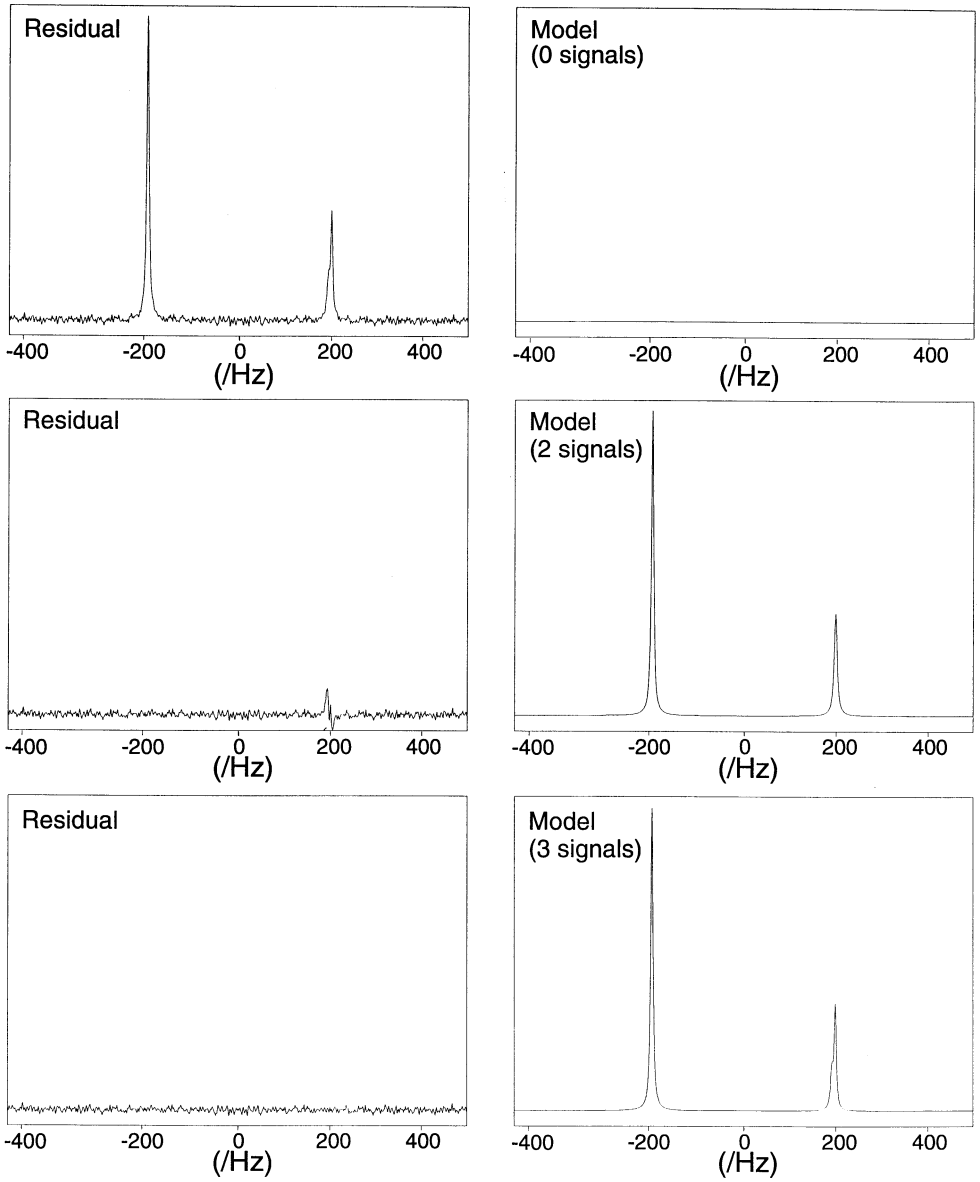

Figure 3. HTFD-ML analysis of a 1D synthetic data set. The figure shows three pairs of residual and model spectra associated with the analysis of the synthetic data set described in Table 1. (Top pair) At the starting point of the analysis, the model contains no signals, and the residual spectrum is simply equivalent to the absorption spectrum of the synthetic data. (Middle pair) The signal model (right) and residual (left) derived from the first round of HTFD-ML analysis. The first round of peak analysis yields two resolved signals. Using the algorithm described in the text, the amplitude and decay rate parameters of the signals derived from these peaks are optimized to fit the data. The residual contains evidence for another peak; the signal derived from this peak is added to the signal model in the next round. (Bottom pair) The signal model and residual at the end of the second round of HTFD-ML analysis; the amplitude, frequency, and decay rate parameters of the three signals in the have been optimized to minimize the residual.

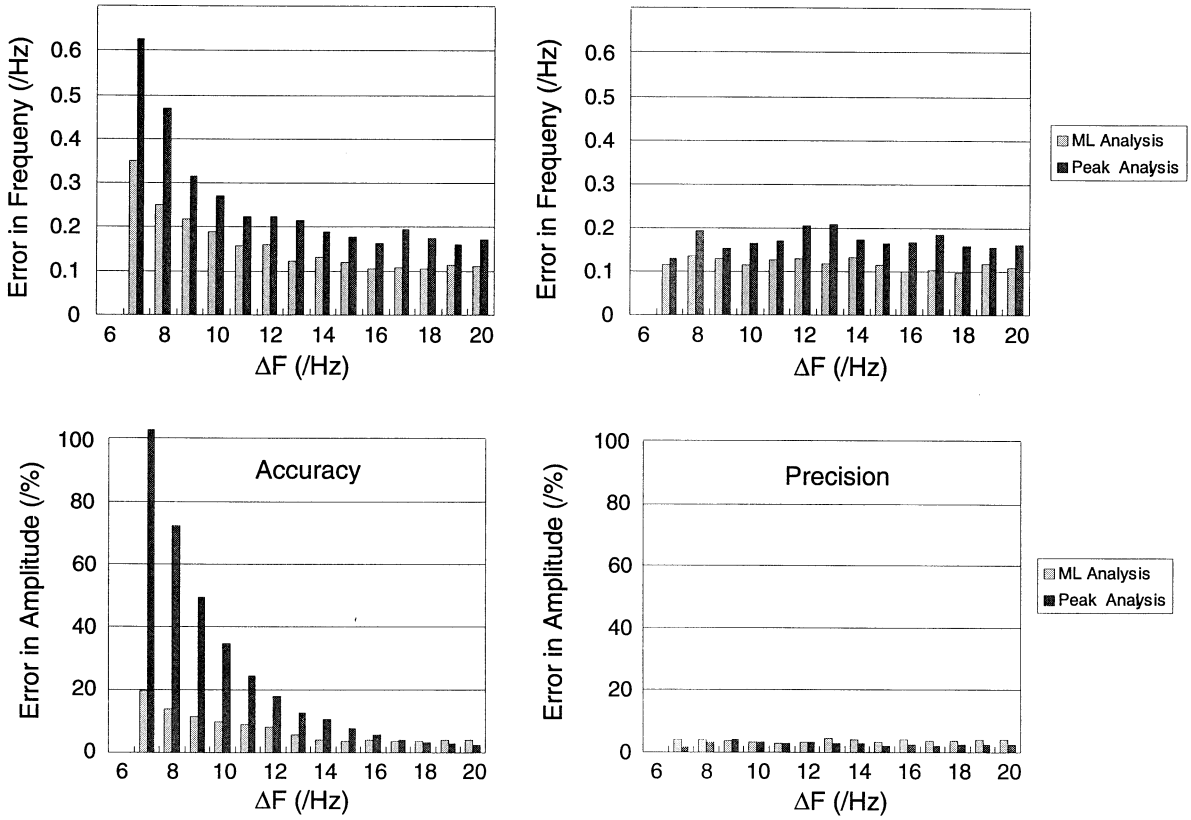

Figure 4. Error analysis of (top) frequency estimates and (bottom) amplitude estimates obtained from HTFD-ML analysis (light shaded bars) and peak analysis (dark shaded bars) of a series of 1D synthetic data sets (Table 2).

Figure 3. The corresponding residual is shown in the lower left frame.

To obtain a quantitative comparison of the precision and accuracy of both the frequency and amplitude estimates obtained from HTFD-ML analysis versus peak analysis, a matrix of $16 \times 16 = 256$ synthetic 1D spectra was created, in which each spectrum was similar to the three-signal spectrum whose parameters are displayed in Table 1. The parameters of signals *2* and *3* were chosen to provide a range of 16 different frequency resolutions (difference in frequency between signals *2* and *3*) and 16 signal sensitivities (ratio of the amplitudes of signals *2* and 3 versus the noise). Over a range of 16 discrete values, the amplitudes of signals *2* and *3* were decreased linearly from respective values of 32 and 16 to respective values of 2 and 1 (arbitrary units). The standard error of the noise was kept constant over the range thus producing a set of sixteen steadily decreasing signal-to-noise ratios. In some of the spectra, the presence of signal *3* was not clearly discernible from the noise. These spectra are referred to as 'low sensitivity' spectra. Along with changes in sensitivity, the frequency of signal *2* was decreased linearly from a value of 211 Hz to 196 Hz thus yielding a frequency separation between signals *2* and *3* that ranged from 19 Hz to 4 Hz. In some of these spectra, signal *3* was not clearly discernible from signal *2* (see Figure 3). These spectra are referred to as 'low resolution' spectra. The matrix of $16 \times 16 = 256$ synthetic spectra thus contained a full range of sensitivity and resolution combinations. For each of these 256 synthetic models, eight different spectra were created by adding eight different sets of random noise to the model. To ensure that the random noise accurately reflected the characteristics of spectrometer noise, the noise added to the synthetic data was obtained from processed spectra of FID's recorded on an NMR spectrometer (see Methods). The addition of eight data sets with different noise components and identical systematic components yielded a final matrix of $16 \times 16 \times 8 = 2048$ data sets.

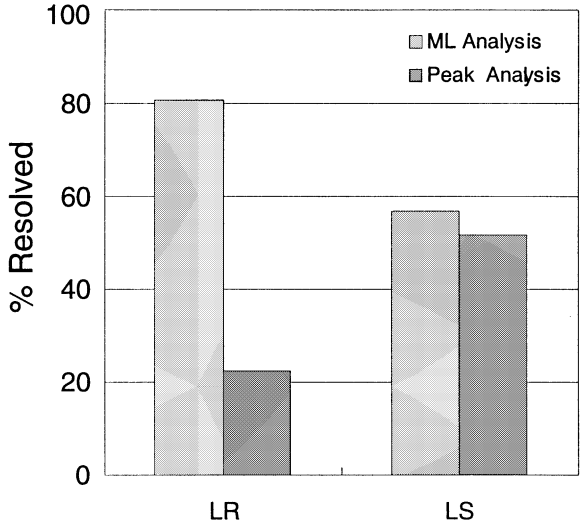

Figure 5. Percentage of signals resolved from HTFD-ML analysis (light shaded bars) and peak analysis (dark shaded bars) of a series of 1D synthetic data sets. A set of 256 different synthetic data sets were constructed as described in Table 2. A portion of the data sets contained very poor resolution (LR) between signals *2* and *3*. Another portion of the data had a low signal-to-noise ratio (low sensitivity, LS) associated with signal *3*. The bar graphs displayed above represent the percentage of low resolution (LR) and low sensitivity (LS) spectra in which signals *1–3* were resolved. In the context of HTDF-ML analysis, signal *3* was considered to be resolved if the residual contained sufficient evidence for a signal after a two-signal model had been constructed to model the data. In the context of peak analysis, signal *3* was considered to be resolved from signal *2* if two extrema were present in the corresponding region of the spectrum (see Figure 1).

Each of the 2048 data sets was analyzed by both HTFD-ML and by peak analysis. A summary of the error analysis of the frequency estimates is shown in Table 2 (top) and graphed in Figure 4 (top). The data give a comparison between the accuracy and the precision of the frequency estimates for HTFD-ML analysis and peak analysis. The term 'accuracy' as used in this context represents the standard deviation of the estimated frequencies (obtained for each of the eight data sets) from the known frequency value. The term 'precision' as used in this context represents the standard deviation of the estimated frequencies (obtained for each of the eight data sets) from the mean frequency value.The error analysis of the frequency estimates for signals *1–3* is shown averaged over all data sets as well as over subsets of the data containing signals of low resolution and low sensitivity (see Table 2 and Figure 4). Several observations can be made about the results of the frequency estimates.

(1) The overall error in both the precision and accuracy of the frequency estimates for the two methods of analysis was comparable for signals *1* and *2*.
(2) Averaged over all data sets and over all low-sensitivity data sets, both the accuracy and the precision of the signal *3* frequency estimates derived from HTFD-ML analysis were slightly improved relative to peak analysis.
(3) Averaged over all data sets and over all low-sensitivity data sets, the precision of the frequency estimates was a reliable indicator of the accuracy of the frequency estimates.
(4) For low-resolution data sets, the precision of the frequency estimate was a poor indicator of its accuracy.
(5) For low-resolution data sets, the accuracy of the frequency estimates derived from HTFD-ML analysis were significantly improved relative to peak analysis.

A similar analysis was made of the amplitude estimates: Table 2 (bottom) and Figure 4 (bottom). Because there was no direct way to measure the accuracy of the amplitudes derived from peak integration, the amplitude estimates of signals *2* and *3* were expressed as ratios relative to the amplitude of signal *1*. The following observations can be made concerning the results of the amplitude estimates.
(1) Both for averages over all data sets and for averages over low-sensitivity data sets, the overall errors in both the precision and accuracy of the amplitude estimates derived from HTFD-ML analysis were about half as large as those derived from peak analysis.
(2) For all data sets, the precision of the amplitude error was a poor estimator of the accuracy of the amplitude error.
(3) For low-resolution data sets, the overall error in the accuracy of the amplitude estimates derived from HTFD-ML analysis was substantially less than the equivalent error estimates derived from peak analysis, despite the fact that the precision estimates from the two methods were comparable.

The relative performance of HTFD-ML analysis relative to that of peak analysis could be measured, not only in terms of the accuracy of the frequency and amplitude results (parameter estimation), but also in terms of the number of signals that could be detected (signal recognition). Figure 5 displays a bar graph showing the percentage of signals *3* that were resolved from signals *2* averaged over all the low-resolution (LR) and the low-sensitivity (LS) data sets. Although the percentage of signals resolved by the two approaches was comparable for the low-sensitivity data

sets, four times as many signals could be detected by HTFD-ML analysis than by peak analysis when the signals were poorly resolved.

It can be concluded from Figures 4 and 5 that the relative accuracy of both the frequency and amplitude estimates afforded by HTFD-ML analysis, as compared to peak analysis, is weakly affected by the signal-to-noise ratio of the data but strongly influenced by the frequency resolution of the data. Figure 6 illustrates how the errors in frequency and amplitude estimates change as a function of the separation in frequency between signals *2* and *3*. Each bar represents the sum over all error estimates obtained for signals of different signal-to-noise ratios. For the error estimates of the frequencies (Figure 6, top), the precision is a reliable indicator of the accuracy when the difference between signals *2* and *3* is greater than 9 Hz (about half the linewidth of the two signals). For the error estimates of the amplitudes (Figure 6, bottom), the precision is a reliable indicator of the accuracy only when the difference between signals *2* and *3* is greater than 15 Hz (approximately the linewidth of the two signals). The data clearly show that the error of amplitude estimates derived from peak integration is much more sensitive to the problem of spectral overlap than the error of frequency estimates derived from peak interpolation. The data also show that the largest gain obtained from HTFD-ML analysis occurs when spectral overlap is significant.

Application to a series of 2D ^1H–^{15}N T_1 relaxation data

This section presents an illustrative application of the HTFD-ML approach to the concerted analysis of a series of 2D ^{1}H–^{15}N HSQC T_1 relaxation experiments carried out on a medium sized soluble protein (the carbon monoxide ligated form of the monomeric hemoglobin Component IV, GMH4CO, from *Glycera dibranchiata*, 147 amino acids). A contour plot of a 2D ^{1}H–^{15}N HSQC ^{15}N T_1 spectrum measured at relaxation delay, $T = 10$ ms, is shown in Figure 7. The spectrum was one of a series of ten experiments collected at eight T values ($T = 10$ (twice), 60, 120, 200 (twice), 400, 800, and 1400 ms) in order to measure ^{15}N T_1 relaxation. The spectrum in Figure 7 contains over 200 signals. These arise from the expected ^{1}H–^{15}N pairs in the non-proline backbone and the side chain amide residues of the major form of the protein in solution and also, in part, from a minor species ($\sim$ 10%) with the heme inserted in an alternative orientation.

A strength of any parametric approach to NMR data analysis is that prior information about the data can be used to set constraints on the analysis. The optimum approach to analyzing the T_1 relaxation data is to analyze all of the spectra in concert making use of prior information about the experiment. Prior information about the ^{15}N T_1 relaxation data suggests adoption of the following constraints:

(1) Each of the spectra should have the same number of signals. A given ^{1}H-^{15}N signal in the 10 ms spectrum should thus exist at *approximately* the same position in the other nine spectra. The set of ten such signals will be referred to as 'corresponding' signals.

(2) The decay rates of corresponding signals should be equal along both dimensions, i.e., corresponding signals should have the same shape.

(3) Of the four classes of spectral parameters in the model (frequency, phase, decay rate, and amplitude), only the amplitude parameters are expected to change significantly within a set of corresponding signals.

To implement these constraints, the ten data sets were loaded into a single 3D matrix, in which each plane of the matrix contained a single 2D spectrum. The spectrum shown in Figure 7 was used as a 'reference' spectrum on which signal recognition by HTFD-ML analysis was applied. Figure 8 displays the results for a portion of the spectrum (that corresponding to the dashed-line box in Figure 7). The data, the model, and the residual for a crowded region of the spectrum are shown as contour plots in the upper row and as stacked plots in the lower row (Figure 8). The model derived from the reference spectrum was used as the starting point for constructing an initial model to fit to the corresponding signals from the remaining nine data sets. The amplitudes, frequencies, and decay rates of the signals were allowed to vary from their initial values obtained from the reference spectrum. After this optimization step was complete, the weighted average of the decay rates was calculated for each set of ten corresponding signals. The decay rates of all ten signals were then fixed at the weighted average. A final optimization step was performed in which the decay amplitudes and frequencies of all signals in the data sets (including the ones in the reference spectrum) were allowed to vary while the decay rates were held fixed at their previously optimized values. The time required to perform the complete analysis of the ten data sets containing a total of 2090 signals was about

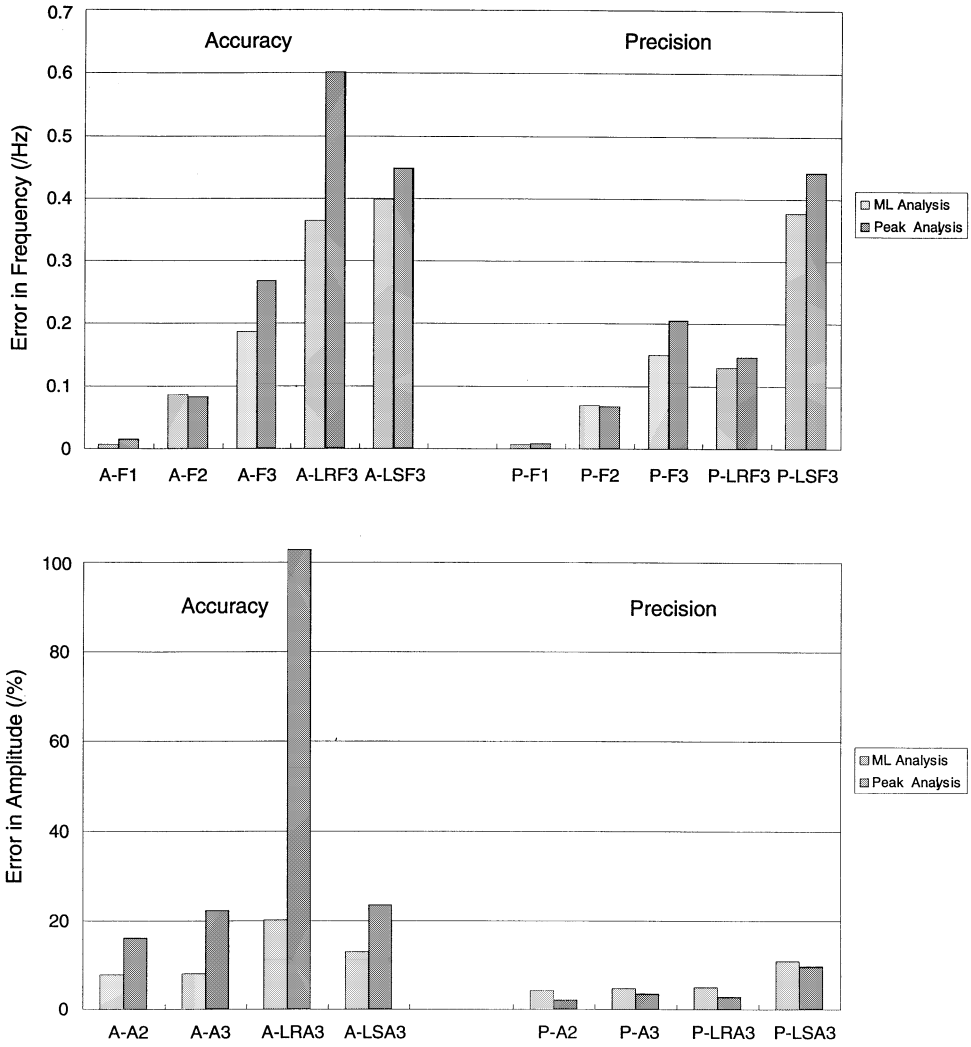

Figure 6. Error analysis of (top) frequency and (bottom) amplitude estimates obtained from HTFD-ML analysis and peak interpolation vs. the resolution between signals *2* and *3*. A set of 256 different synthetic data sets (2048 total data sets) were constructed as described in the legend of Table 2. HTFD-ML analysis and peak interpolation were conducted on each of the data sets, and a summary of the errors of the frequency estimates are displayed in the above figure. The errors in accuracy (left) and precision (right) of the estimates for signal *3* are plotted as a function of the separation between signals *2* and *3*. Each bar represents the sum over all error estimates obtained for signals of different signal-to-noise ratio.

3.4 hours on an SGI Indigo^2 R8000 running a version of IRIX 5.3.

Since it is known that the shapes of corresponding peaks do not change as a function of the relaxation delay, measurements of peak height are preferable to measurements of peak integrals as a method for extracting relaxation information. The left and right portions of Figure 9 represent the decay of the normalized amplitude of the signal assigned to lysine-33 (K33) as a function of the relaxation time derived, respectively, from HTFD-ML analysis and peak height analysis. The smooth curve drawn through the data points is a theoretical least-squares fit of the data to a mono-exponential decay function. Assuming that the amplitudes do decay exponentially with relaxation, the root mean distance of the data points from the theoretical curve can be used as a measure of the precision of the amplitudes. Visual inspection of the amplitude versus relaxation time profiles of Figure 9 suggests that the levels of precision obtained from HTFD-ML analysis and peak height analysis are equivalent. Indeed, the root mean distance of the relaxation profiles

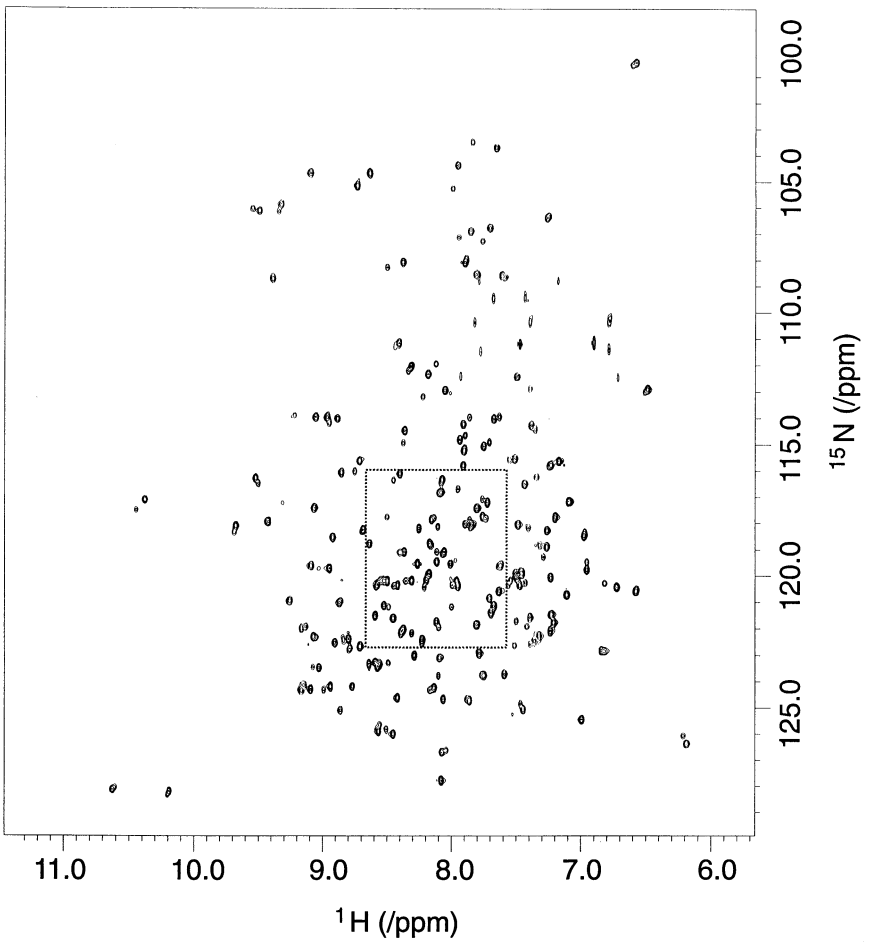

Figure 7. Contour plot of a 2D ^{1}H–^{15}N HSQC experiment conducted upon *Glycera dibranchiata* component IV monomeric hemoglobin (GMH4). The data were measured at 750.13 MHz ^{1}H frequency with a pulse scheme that utilized gradients for sensitivity enhancement. The spectrum was acquired with 200$*$ ^{15}N points in t_1 and 1024$*$ ^{1}H points in t_2. Further details of the experiment are contained in the methods section. The spectrum was one of a series of experiments collected in order to measure ^{15}N T_1 relaxation. For the spectrum shown in the above figure, $T = 10$ ms. The time-domain data were digitally filtered along t_1 and t_2 with a sine-squared bell window function shifted 40° and 70° respectively. After apodization, the data were zero-filled to a length of 512 points along the ^{15}N dimension and 2048 points along the ^{1}H dimension. The portion of the ^{1}H spectrum downfield from the water signal was discarded to yield a final spectral resolution of 1024 $\times$ 512 data points. The dashed box in the figure shows the portion of the 2D spectrum which is displayed in Figure 8.

averaged over all signals in the data set are statistically equivalent (results not shown).

The results obtained from the analysis of the model 1D spectra indicated that the precision of amplitude estimates is not always a reliable indicator of their accuracy. Since the actual amplitude profiles were not known for the experimental data shown in Figures 7–9, the accuracy of the amplitude estimates obtained from HTFD-ML analysis and peak analysis were compared by an analysis of synthetic data sets with systematic properties equivalent to those of the experimental data set. The values for the amplitude, frequencies, and decay rates obtained from HTFD-ML analysis of the measured ^{15}N T_1 relaxation data were used to construct a reference spectrum and related spectra with amplitude versus relaxation time profiles analogous to the acquired spectrum. From this series of ten synthetic 2D ^{1}H–^{15}N spectra with known frequency parameters and known ^{15}N T_1 relaxation decay rates, a series of eight data sets were constructed by adding spectrometer noise to the spectra (see Methods section). The eight data sets represent eight analogous sets of ten synthetic spectra where each of the eight data sets are identical in their systematic component and differ only in their random (noise) component.

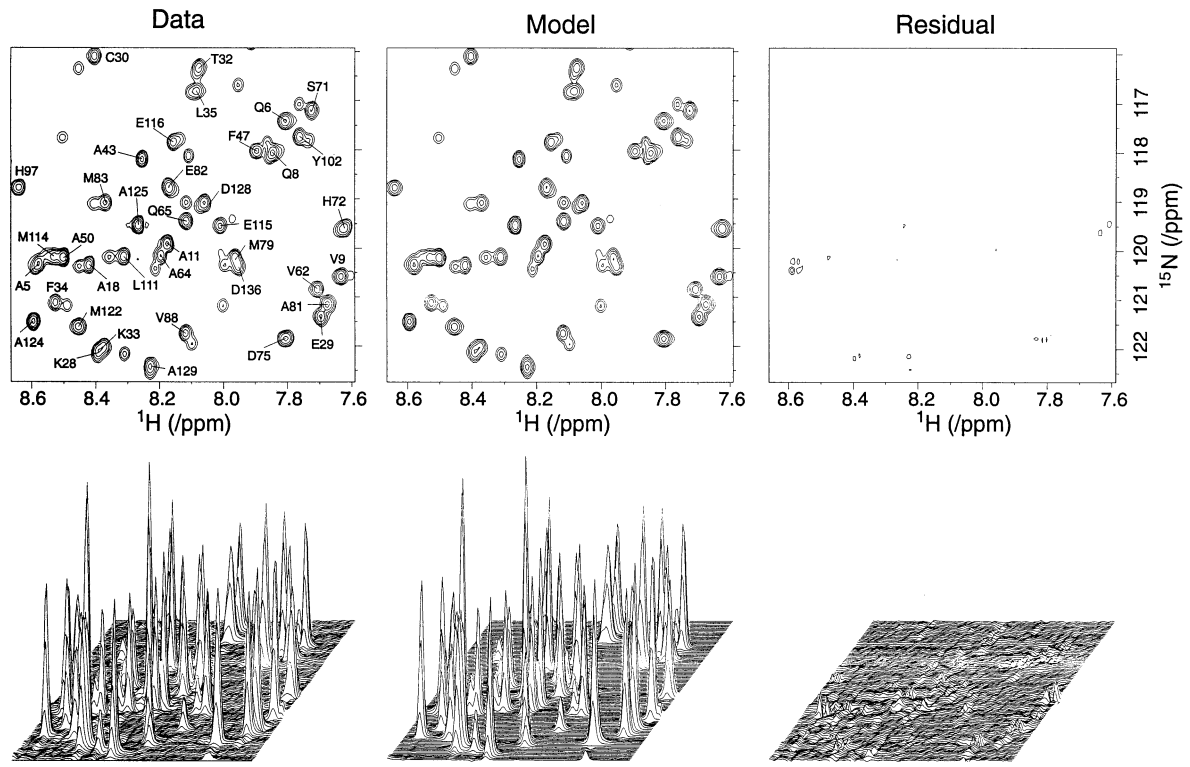

Figure 8. Contour plots (top row) and stacked plots (bottom row) of a zoomed region of the data, model, and residual spectra associated with HTFD-ML analysis of a 2D ^{1}H–^{15}N spectrum of *Glycera dibranchiata* component IV monomeric hemoglobin (GMH4) acquired and processed as explained in the legend of Figure 7. A zoomed (Zolnai et al., 1996) region of this spectrum is shown in the left panel (Data). The 2D spectrum was subjected to HTFD-ML analysis using the algorithm described in the text. The corresponding region of the model spectrum obtained from HTFD-ML analysis is shown in the middle panel (Model). The difference between the data and model spectrum is shown in the right panel (Residual). The entire signal model constructed to fit the data consisted of 209 signals. The spectrum shown above was analyzed simultaneously with nine other data sets as described in the text. The ten data sets collected to measure ^{15}N T_1 relaxation corresponded to relaxation delay values T of 10 (twice), 60, 120, 200 (twice), 400, 800, 1400 ms. The steps involved in the analysis of the entire relaxation series consisted of:
(a) Complete analysis including optimization of the frequency, decay rate, and phase parameters) of the reference data set shown in Figure 7 (209 signals).
(b) Optimization of the frequency, decay rate, and amplitude parameters for the other nine data sets ($9*209 = 1881$ signals) using the frequency and decay rate values derived from step (a) as initial estimates.
(c) Optimization of the phase and amplitude parameters for all 2090 signals holding the frequency and decayrate estimates constant at the values obtained from step (b).
(d) Calculation of the phase parameters along the acquisition dimension for each reference signal according to the weighted average of the phase values obtained from all 10 data sets.
(e) Calculation of the decay rate parameters for each reference signal according to the weighted average of the decay rates obtained from all 10 data sets.
(f) Final optimization of the frequency and amplitude estimates for all 2090 signals holding the phase parameters constant at their values obtained from step (d) and the decay rate parameters held constant at their values obtained from step (e).
The time required to perform the sequence of steps listed above was 3.4 hours on a Silicon Graphics R4000 processor running under IRIX 6.2 (128 MB of RAM).

The eight synthetic data sets (each of which contained ten 2D ^{1}H–^{15}N T_1 relaxation spectra) were analyzed by HTFDL-ML analysis in a manner identical to the analysis of the experimental data. ^{15}N T_1 relaxation rates were then extracted from the normalized amplitude (or peak height) versus relaxation time profiles associated with the ^{1}H–^{15}N signals assigned to the backbone atoms of GMH4CO. Figure 10 shows the standard deviation of the measured ^{15}N T_1 relaxation rates from their known values. The data show that the average error in the accuracy of the HTFD-ML derived ^{15}N T_1 relaxation rates was about three times less than the equivalent error of the peak height-derived relaxation rates.

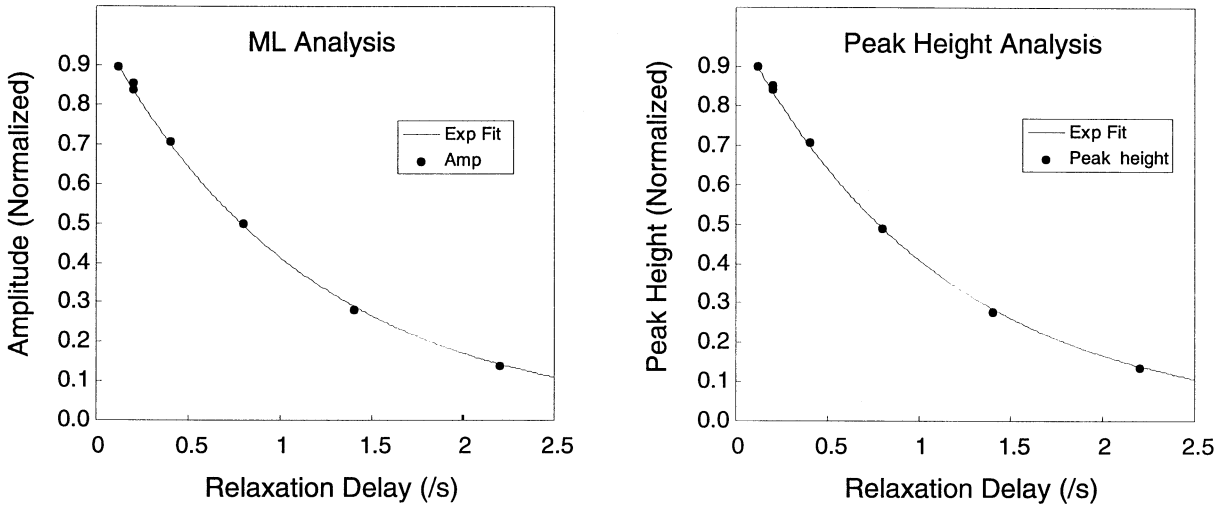

Figure 9. Plots of the normalized HTFD-ML derived amplitude (left) and normalized peak height (right) of the signal assigned to lysine 33 (K33) of *Glycera dibranchiata* component IV monomeric hemoglobin (GMH4) as a function of the relaxation delay (10 data points with 8 different relaxation delays). A series of ^{15}N T_1 relaxation spectra were acquired, processed, and analyzed by HTFD-ML methods as described in the legend of Figure 10. The smooth curve drawn through the points in each plot represents the non-linear least squares exponential fit to the data points. An amplitude and peak-height profile of this kind was used in determining the ^{15}N relaxation time for each signal, present in the ^{1}H–^{15}N spectrum. The amplitude and peak-height profiles from the two sets of plots exhibited comparable precision.

Discussion

The results presented here demonstrate that the introduction of a parametric model-fitting approach, such as maximum likelihood analysis, can lead to a significant improvement in the quality of NMR parameters extracted from the data. Moreover, the approach lends itself to automation and to statistical analysis of the signals in multidimensional spectra. The strengths of conventional Fourier analysis are that it is fast, easy to implement, and produces a spectrum in which the frequency and amplitude information are localized. The localized nature of the absorption spectrum produced by Fourier procedures is essential to its ability to process large multidimensional data sets containing numerous signals. The HTFD-ML algorithm described here complements the strengths of Fourier analysis because of its ability to take into account the simultaneous contributions of all signals in a crowded spectral region. Conventional Fourier analysis assumes that the amplitude associated with a signal can be measured from the height (or integral) of a signal at its maximum point in the absorption spectrum. This assumption is equivalent to saying that the interactions matrix of the signals in an NMR data set is diagonally dominant: i.e., the interactions between signals as measured by the off-diagonal elements of the amplitude matrix are negligible in comparison to the diagonal elements of the matrix. Because the HTFD-ML approach constructs a theoretical model to fit the data, it can construct and solve the entire interaction matrix associated with a given number of signals and their non-linear parameter values. The HTFD-ML approach can thus take into account the non-negligible contributions of the off-diagonal elements of the interaction matrix.

The HTFD-ML approach is essentially a series of approximations that make it practical to apply model fitting methods to large multidimensional matrices containing numerous signals. A rigorous approach to model fitting of NMR data simultaneously considers the contributions of all signals in the model and calculates the projections of the model basis functions over the entire data set. The large number of signals and the large size of NMR data make this approach impractical for *D*-dimensional NMR analysis, even when the advanced computing power of commercially available workstations is taken into account. By transforming the time-domain models into the frequency domain and applying a digital cutoff filter (see Figure 2), the ML approach is able to drastically reduce both the size of the dot products and the size of the interaction matrices required to perform spectral deconvolution. Only those signals that overlap appreciably in the spectrum are optimized simultaneously. The overwhelming task of solving simultaneously an

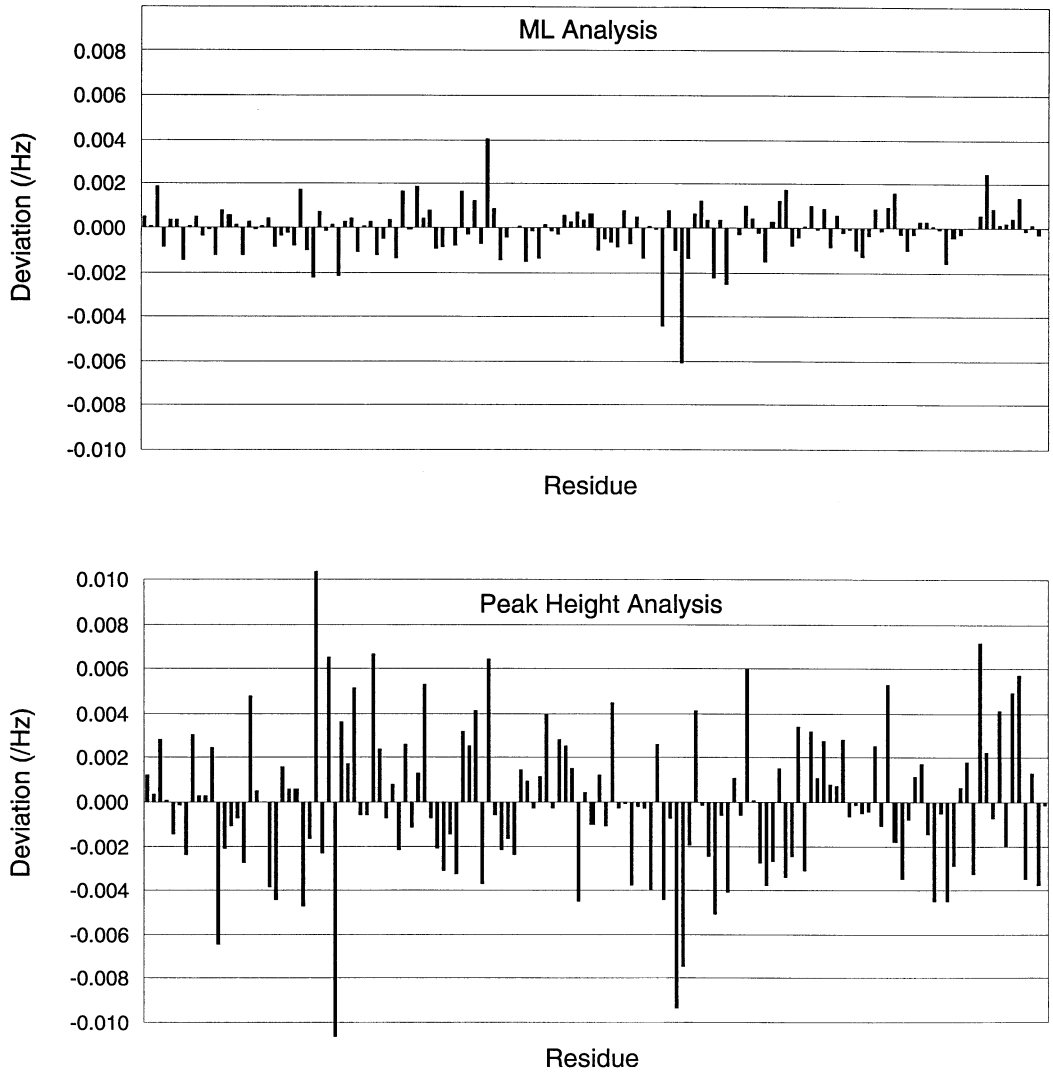

Figure 10. Standard deviations of measured ^{15}N T_1 relaxation rates from their known values determined at each residue position by (top) HTFD-ML analysis and (bottom) by peak-height analysis of a synthetic data set of eighty spectra. The synthetic data set was constructed as follows (see Methods section for additional details). Values for the amplitude, frequencies, and decay rates obtained from HTFD-ML analysis of the experimental ^{15}N T_1 relaxation data set for *Glycera dibranchiata* component IV monomeric hemoglobin (GMH4) were used to construct a synthetic data set with analogous amplitude versus relaxation time profiles. To this series of ten synthetic, partially relaxed, 2D ^{1}H–^{15}N spectra, with known frequency parameters and known ^{15}N T_1 relaxation decay rates, eight levels of spectrometer noise were added to generate the 80 spectra used in this analysis.

enormous matrix of all signals in the model is thus reduced to the manageable operation of solving a series of small signal networks sequentially. Results from the previous section demonstrate that the HTFD-ML approach implemented on a conventional desktop computer can be used to analyze a series of 1024×512 2D NMR matrices containing more than 2000 signals in a period of only several hours.

The results presented here show that amplitude estimates are much more sensitive than frequency estimates to the effects of spectral overlap. This observation can be rationalized by noting that the 'center' of a peak (measured by peak interpolation) is less affected by the presence of nearby signals than is the 'tail' of a peak (measured during peak integration). This fact in part explains why, when comparing signals with equivalent line shapes, peak height measurements are

more precise measurements of signal amplitude than are peak integral measurements. Another explanation for the greater precision of peak height measurements comes from the fact that the peak-height method takes into account prior information about the signals: i.e., information that they have identical line shapes. Peak height measurements are thus affected only by the imprecision in measuring signal intensity. Peak integrals, on the other hand, do not take this information into account and thus are affected by the imprecision of measuring both intensities and line widths (decay rates). The fixing of the decay rates of corresponding signals along both dimensions was thus a necessary constraint to obtaining precise amplitude estimates from the concerted analysis of the 2D ^{15}N T_1 relaxation data sets shown in Figures 7–9.

Results from analysis of the 1D model spectra (Figures 3–6) indicated that improvement in frequency and amplitude estimates from the incorporation of model-fitting methods are modest for spectra containing well separated signals. With all other factors held constant, HTFD-ML analysis showed a significant advantage over peak-height analysis in accurately estimating the amplitudes of noisy signals only when the separation between the signals in the spectrum approached the linewidth of those signals. Similarly, HTFD-ML analysis showed substantial improvements in frequency estimates when separations between the signals were approximately half of the linewidth of the overlapping signals in the spectrum.

The results of the concerted analysis of the ^{15}N T_1 relaxation data demonstrate the practicality of applying the HTFD-ML method to extract quantitative information about signal frequencies and amplitudes from series of *D*-dimensional NMR spectra. In addition to showing the greater accuracy that can be achieved from use of the HTFD-ML algorithm, the results suggest that attempts to quantify the accuracy of amplitude measurements from repeated measurements of relaxation times should be interpreted with caution. The error estimates obtained from measuring the standard deviation of amplitudes at the same mixing time was found to provide a good *lower estimate* of the actual accuracy of the amplitude estimate. Such a measurement yields the *random error* associated with a peak height measurement. This random error accurately estimates the total error involved in the peak height only for signals that are well separated. The actual error involved in a peak height measurement for an overlapped signal may be substantially greater than this precision estimate owing to the presence of systematic error introduced by contributions to the peak intensity from nearby signals.

The analysis of the ^{15}N T_1 relaxation data also highlight the importance that weighting functions play in the processing of NMR spectra. The spectrum shown in Figure 7 was produced by the application of resolution enhancing window functions (see legend) along both the t_1 (^{15}N) and t_2 (^{1}H) dimensions. The equivalent spectrum produced by application of exponential weighting functions prior to Fourier transformation display a much greater degree of overlap (data not shown). As a result of the application of non-exponential digital filters, however, the line shapes of the signals in the ^{15}N T_1 relaxation spectrum (Figure 7) are markedly non-Lorentzian. This would pose a challenge to model-fitting using a pure frequency-domain approach, since a basis function or sum of basis functions would have to be found that accurately models the non-Lorentzian peak shapes exhibited in the spectrum shown in Figure 7. In contrast, the HTFD-ML approach employs a model that is completely independent of the digital filters applied during Fourier transformation. This fact should allow the HTFD-ML approach to be easily adapted to the future analysis of NMR data where the data are sampled non-uniformly along one or more of the indirectly-detected dimensions.

The HTFD-ML algorithm described in this manuscript approximated a NMR spectrum with increasing number of signals until a residual criterion was satisfied (see steps 13 and 16 in the algorithm section). In earlier work (Chylla and Markley, 1995), we implemented a non-linear least squares approach that used a more statistically rigorous termination criteria based on the minimum description length (MDL) statistic. The MDL is a model selection statistic that takes into account not only the 'goodness of fit' of a model to the data (chi square) but also the number of degrees of freedom present in the model. In our earlier work, an optimized model containing $(n + x)$ signals was judged to be a more 'likely' model than a model containing n signals if and only if the reduction in chi square brought about by introduction of the x additional signals was great enough to balance the penalty incurred by any additional free parameters associated with the added signals. The MDL statistic is appropriate when all of the systematic components present in the data are accounted for by the model. In multidimensional NMR spectra, however, this condition frequently does not hold due to the presence of weak signals produced by incomplete cancellation of un-

desired magnetization pathways and other non-ideal conditions. The MDL statistic used as a termination criteria in spectra with these nuisance signals present will result in the time-consuming optimization of very weak signals that are of no interest. It is worth noting, however, that for many experiments the use of the MDL statistic is an appropriate and more statistically rigorous termination criterion.

The algorithm presented here has been found to be applicable to a wide range of spectra. Its most serious limitation stems from its parametric nature: the models employed must fully describe all of the NMR transitions that give rise to the signal. The algorithm thus has success in analyzing spectra whose signals are modeled accurately by a digitally filtered and transformed exponentially decaying sinusoid. Such spectra include 1D spectra, *D*-dimensional HMQC and HSQC spectra, and NOESY spectra of large molecules. The HTFD-ML approach currently is less suited to modeling TOCSY spectra and NOESY spectra of small molecules, where the peak shapes reflect the contributions of numerous pathways of magnetization transfer. To accurately model these spectra in the future, it will be necessary to employ more sophisticated analytical models. This should not pose great difficulties, however. Even in its present form, the HTFD-ML algorithm represents a significant advance toward the goal of completely automating the pathway from the acquisition of NMR data to the extraction and analysis of primary NMR parameters.

Acknowledgements

The authors thank Drs. Steve L. Alam and James D. Satterlee who provided the labeled sample of component IV *Glycera dibranchiata* monomeric hemoglobin-CO used in this study. This work was supported by a grant (RR02307) from the Biomedical Research Technology Program of the National Center for Research Resources, National Institutes of Health.

References

Alam, S.L., Volkman, B.F., Markley, J.L. and Satterlee, J.D. (1998) *J. Biomol. NMR*, **11**, 119.
Bretthorst, G.L. (1990) *J. Magn. Reson.*, **88**, 533.
Chylla, R.A. and Markley, J.L. (1993) *J. Biomol. NMR*, **3**, 515.
Chylla, R.A. and Markley, J.L. (1995) *J. Biomol. NMR*, **5**, 245.
Farrow, N.A., Muhandiram, R., Singer, A.U., Pascal, S.M., Kay, C.M., Gish, G., Shoelson, S.E., Pawson, T., Forman-Kay, J.D. and Kay, L.E. (1994) *Biochemistry*, **33**, 5984.
Fitzgerald, W.J., Nuzillard, D. and Nuzillard, J.-M. (1995) *J. Magn. Reson. A*, **117**, 285.
Gesmar, G. and Led, J.J. (1989) *J. Magn. Reson.*, **83**, 53.
Gesmar, H., Led, J.J. and Abildgaard, F. (1990) *Progress in NMR Spectroscopy*, **22**, 255.
Hoch, J. and Stern, A. (1996) *NMR Data Processing*, John Wiley and Sons.
Kay, L.E., Wittekind, M., McCoy, M.A., Freidrichs, M.S. and Mueller, L. (1992) *J. Magn. Reson.*, **98**, 443.
Marquardt, D.W. (1963) *J. Soc. Ind. Appl. Math.*, **11**, 431.
Miller, M.I. and Greene, A.S. (1989) *J. Magn. Reson.*, **83**, 525.
Sibisi, S., Skilling, J., Brereton, R.G., Laue, E.D. and Staunton, J. (1984) *Nature*, **311**, 446.
Umesh, S. and Tufts, D.W. (1996) *IEEE Transactions on Signal Processing*, **44**, 2245.
Wang, S., Pelczer, I., Borer, P.N. and Levy, G.C. (1994) *J. Magn. Reson. A*, **108**, 171.
Zhu, G. and Bax, A. (1992) *J. Magn. Reson.*, **98**, 192.
Zolnai, Zs., Juranić, N., Markley, J.L. and Macura, S. (1996) *J. Magn. Reson. A*, **119**, 53.