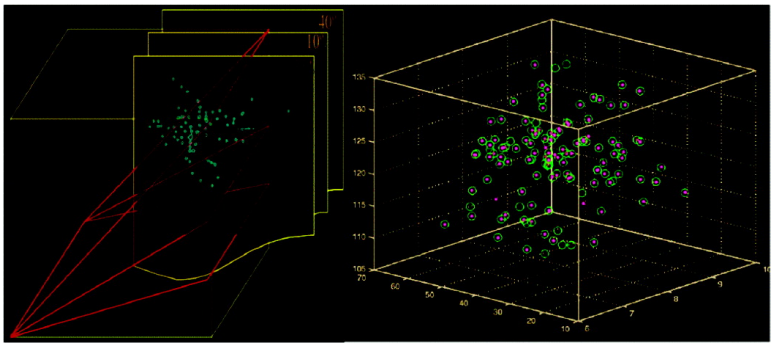
Article

High-Resolution Iterative Frequency Identification for NMR as a General Strategy for Multidimensional Data Collection

Hamid R. Eghbalnia, Arash Bahrami, Marco Tonelli, Klaas Hallenga, and John L. Markley

J. Am. Chem. Soc., **2005**, 127 (36), 12528-12536• DOI: 10.1021/ja052120i • Publication Date (Web): 18 August 2005



J|A|C|S
ARTICLES
Published on Web 08/18/2005

High-Resolution Iterative Frequency Identification for NMR as a General Strategy for Multidimensional Data Collection

Hamid R. Eghbalnia,*[,†,#,^] Arash Bahrami,[†,‡,§] Marco Tonelli,[†,‡,#] Klaas Hallenga,[†,‡,#] and John L. Markley[†,‡,§,#]

Contribution from the National Magnetic Resonance Facility at Madison, Center for Eukaryotic Structural Genomics, Graduate Program in Biophysics, Biochemistry Department, and Mathematics Department, University of Wisconsin—Madison, Madison, Wisconsin 53706

Received January 13, 2005; E-mail: eghbalni@nmrfam.wisc.edu

Abstract: We describe a novel approach to the rapid collection and processing of multidimensional NMR data: "high-resolution iterative frequency identification for NMR" (HIFI−NMR). As with other reduced dimensionality approaches, HIFI−NMR collects *n*-dimensional data as a set of two-dimensional (2D) planes. The HIFI−NMR algorithm incorporates several innovative features. (1) Following the initial collection of two orthogonal 2D planes, tilted planes are selected adaptively, one-by-one. (2) Spectral space is analyzed in a rigorous statistical manner. (3) An online algorithm maintains a model that provides a probabilistic representation of the three-dimensional (3D) peak positions, derives the optimal angle for the next plane to be collected, and stops data collection when the addition of another plane would not improve the data model. (4) A robust statistical algorithm extracts information from the plane projections and is used to drive data collection. (5) Peak lists with associated probabilities are generated directly, without total reconstruction of the 3D spectrum; these are ready for use in subsequent assignment or structure determination steps. As a proof of principle, we have tested the approach with 3D triple-resonance experiments of the kind used to assign protein backbone and side-chain resonances. Peaks extracted automatically by HIFI−NMR, for both small and larger proteins, included ∼98% of real peaks obtained from control experiments in which data were collected by conventional 3D methods. HIFI−NMR required about one-tenth the time for data collection and avoided subsequent data processing and peak-picking. The approach can be implemented on commercial NMR spectrometers and is extensible to higher-dimensional NMR.

Introduction

The acquisition of multidimensional spectra normally is required for NMR investigations of biological macromolecules. To maintain a given level of digital resolution, the number of free induction decays (FIDs) that have to be recorded grows exponentially with the number of dimensions. In addition, relaxation during polarization-transfer steps leads to sensitivity losses as dimensions are added. As a result, data collection times for higher-dimensional spectra can be very long, and acquisitions frequently are limited by the stability of the biological sample. The usual practical compromise it to collect smaller data sets, which leads to lower resolution.[1]

Among the methods that have been proposed for fast data collection, one of the most successful is the reduced dimensionality (RD) approach.[2] Reduced dimensionality alleviates some of the difficulties of *n*-dimensional NMR by combining information from different evolution periods into a single dimension. One approach has been to create $^{15}N-^{13}C$ double- and zero-quantum coherence in a single evolution time.[3] Similarly, two-dimensional (2D) versions of HNCA and HNCO triple-resonance experiments, called MQ−HNCA and MQ−HNCO, have been developed.[4]

In more recent RD experiments, the dimensionality is reduced by monitoring the chemical shift evolution of two or more nuclei simultaneously as single-quantum coherences in a single indirect domain. For example, three-dimensional (3D) experiments can be recorded as 2D planes in which the two indirect chemical shifts are encoded in the second dimension. Although losses resulting from additional polarization transfers during the evolution periods cannot be avoided in RD experiments,[5] resolution and sensitivity gains can be realized.

A number of promising RD techniques have been described recently. These include the G-matrix Fourier transform (GFT)[6,7] and time-proportional phase incrementation (TPPI)[8] methods,

† National Magnetic Resonance Facility at Madison.
‡ Center for Eukaryotic Structural Genomics.
§ Graduate Program in Biophysics.
# Biochemistry Department.
^ Mathematics Department.

(1) Wider, G. *Prog. Nucl. Magn. Reson. Spectrosc.* **1998**, *32*, 193−275.
(2) Szyperski, T.; Yeh, D. C.; Sukumaran, D. K.; Moseley, H. N.; Montelione, G. T. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 8009−8014.
(3) Szyperski, T.; Wider, G.; Bushweller, J. H.; Wuthrich, K. *J. Biomol. NMR* **1993**, *3*, 127−132.
(4) Simorre, J. P.; Brutscher, B.; Caffrey, M. S.; Marion, D. *J. Biomol. NMR* **1994**, *4*, 325−333.
(5) Sattler, M.; Schleucher, J.; Griesinger, C. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *34*, 93−158.
(6) Kim, S.; Szyperski, T. *J. Am. Chem. Soc.* **2003**, *125*, 1385−1393.
(7) Natterer. F. *The Mathematics of Computerized Tomography*; Wiley: New York, 1986.
(8) Ding, K.; Gronenborn, A. M. *J. Magn. Reson.* **2002**, *156*, 262−268.

10.1021/ja052120i CCC: $30.25

which differ in the strategy used for extracting frequencies that evolve simultaneously. In GFT, quadrature detection of all simultaneously evolving signals is carried out. For each increment, multiple FIDs are recorded that encode all the possible combinations of sine and cosine terms for the evolving frequencies. A matrix, called the G-matrix, is then applied to linearly combine the FIDs and extract all the various frequency terms. Traditional Fourier transformation then provides a 2D spectrum for each combination of the evolving chemical shifts. In the TPPI approach to RD, quadrature detection is performed sequentially on the signals of each of the simultaneously evolving types of nuclei. After the two data sets are Fourier transformed, signals appear with positive and negative frequency offsets. These two signals correspond to the two evolving chemical shifts. In TPPI, an artificially large resonance offset for the carrier frequency is introduced to avoid overlap. As a result, the peaks representing different frequency combinations are located in distinct regions of the single 2D spectrum. In the case of a 3D spectrum, GFT provides a series of 2D planes each carrying a different combination of the simultaneously evolving frequencies, whereas the TPPI method locates peaks for different frequency combinations in distinct regions of a single 2D spectrum. The larger number of FIDs needed in GFT to achieve quadrature detection of all the convolved frequencies is offset in the TPPI approach by the larger number of increments collected to achieve the same digital resolution for the wider spectral domain that needs to be covered.[9]

The strength of RD methods lies in the possibility of both reducing the collection time of high-dimensional spectra and increasing their digital resolution. However, at the digital resolution achieved, some 3D peaks may still be overlapped. In this case, additional information would be needed to make assignments. Automated peak assignments require adequate definition of peak positions.[10] Thus, it is important for RD methods to provide maximal resolution.

A different approach for fast multidimensional NMR data collection was recently presented by Freeman and Kupce.[11] In the specific case of 3D to 2D reduction, this method generalizes GFT from a fixed angle to several different angles. Peaks that overlap at a given angle frequently can be resolved at other angles (combinations of chemical shifts). Two-dimensional tilted planes are collected by simultaneously incrementing the evolution times for two indirect chemical shifts, with the projection angle being determined by the ratio of the two incrementation times. This approach has the potential of speeding up the data collection by recovering the peak positions in multidimensional spectra from a limited number of 2D tilted planes—potentially without sacrificing spectral resolution.

We present here a novel method for fast data collection and analysis. HIFI−NMR uses an iterative approach to recover the peaks in 3D spectra from two orthogonal planes plus a minimal number of 2D tilted planes collected one-by-one at optimal angles determined from analysis of the prior data collected. Data collection is terminated when the addition of data from another plane is predicted not to improve peak recovery. By collecting data at multiple angles and by focusing on peak identification instead of complete 3D reconstruction, HIFI−NMR circumvents complications encountered with the processing of data from other reduced dimensionality approaches.[12] We demonstrate here the application of the approach to NMR spectroscopy of proteins with the goal of recovering an optimal number of peaks in a minimal amount of spectrometer time from HIFI−NMR versions of 3D triple-resonance NMR experiments.

The underlying principle behind HIFI−NMR is to combine the high digital resolution afforded by 2D spectra, the ability of tilted-plane data collection to separate overlapped peaks, and the flexibility afforded by real-time analysis of the emerging pattern of peaks from prior data collection so as to adaptively determine the next tilted plane to be collected or, alternatively, to terminate data collection because the model of extracted peaks cannot be improved by the addition of another plane. Peaks are identified from high digital resolution 2D spectra rather than from the inferior resolution of reconstituted 3D (or higher-dimensional) spectra. The collection of variable tilted-plane data allows one to avoid peak overlaps that occur as the result of the well-known correlation between ^{1}H and ^{13}C chemical shifts in $^{1}H-^{13}C$ moieties in proteins. This poses a potential problem with data collected at a 45° tilt plane (the default angle for the GFT−RD or TPPI−RD approaches) but can be overcome by collecting data at multiple angles (Figure 1).

Experimental Section

Rationale. The standard approach in NMR spectroscopy has long been the reconstruction of signals from a sampled time series by appealing to theories for perfect reconstruction. In HIFI−NMR, we have chosen to recover the chemical shift frequencies only (partial reconstruction), rather than to reconstruct the spectra (perfect reconstruction). The reasons for this approach are both practical and theoretical. On the practical side, recovery of the frequencies of peaks with a degree of confidence is sufficient for the purpose of NMR peak assignments. On the theoretical side, perfect reconstruction from sampled data remains an active area of research into carefully devised and application-specific approaches. From our experience, in comparison with standard three-dimensional Fourier transform (3D FT) methods, to determine the frequencies of a comparable number of peaks HIFI−NMR data collection requires on the order of 5−15% of the sampled points. For perfect reconstruction from a comparable sparse set, one may consider using the methods of nonuniform sampling,[13] statistical reconstruction,[14] or tomographic reconstruction.[15] These approaches, in their respective standard settings, have to deal with challenges that arise from smearing, blurring, and false peaks. To our knowledge, no approach has been developed that is suitable for perfect reconstruction of NMR data from the category of sparsely sampled data obtained by tilted planes described here. We discuss further the theory of reconstruction and provide relevant references in the Supporting Information.

A more modest goal is to recover certain important features of the spectra in a way that is most consistent with the data collected. In multidimensional, multinuclear NMR experiments collected for the purposes of assignments, the key features to be recovered are the frequencies. By using an overcomplete dictionary of functions constructed from shifted and scaled copies of the standard *sinc* function, we have devised a signal representation that is sparse in a sense that can be precisely defined. In addition, we have developed an efficient algorithm for the estimation of peak positions and for the refinement of the estimates for this representation through stepwise data collection.

(9) Kozminski, W.; Zhukov, I. *J. Biomol. NMR* **2003**, *26*, 157−166.
(10) Olson, J. B., Jr.; Markley, J. L. *J. Biomol. NMR* **1994**, *4*, 385−410.
(11) Freeman, R.; Kupce, E. *J. Biomol. NMR* **2003**, *27*, 101−113.
(12) Tonelli, M.; Metha, D. P.; Eghbalnia, H.; Westler, W. M.; Markley, J. L. Presented at the 45th Experimental NMR Conference, Asilomar, Pacific Grove, CA, April 18−23, 2004, Abstract 347.
(13) Feichtinger, H. G.; Pandey, S. S. *J. Math. Anal. Appl.* **2003**, *279*, 380−397.
(14) Tenorio, L. *Siam Rev.* **2001**, *43*, 347−366.
(15) Faridani, A.; Ritman, E. L. *Inverse Problems* **2000**, *16*, 635−650.

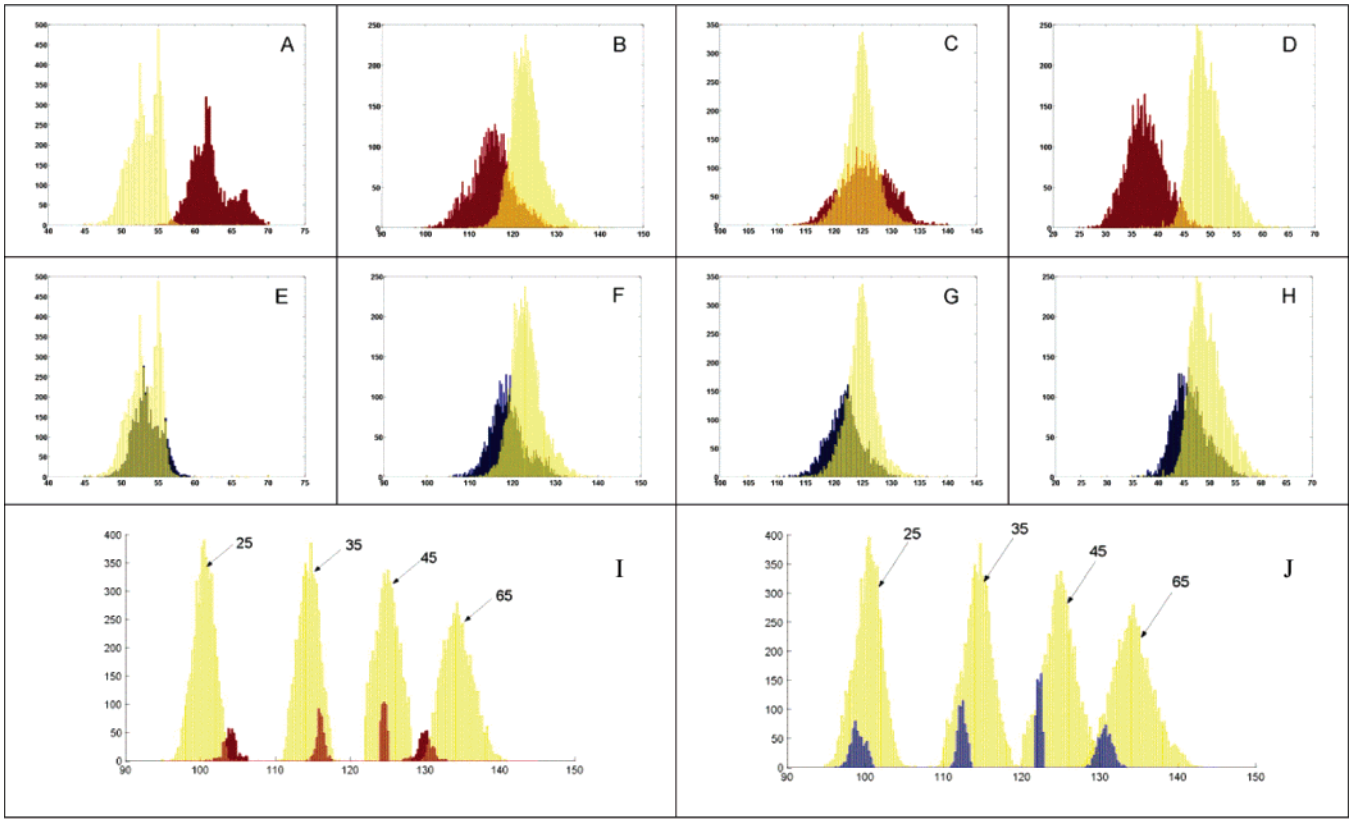


Figure 1. Illustration of the problem of *statistical indistinguishability* and its solution through data collection at multiple tilt angles. Shown are chemical shift distributions for Ala (yellow), Thr (red), and Asn (blue) in proteins (from BMRB, www.bmrb.wisc.edu on 03/10/2004). The first and second rows illustrate the statistics of combined $^{13}C^{\alpha}$ and ^{15}N chemical shift co-evolution corresponding to a plane at 45°. (Row 1) Comparison of the chemical shifts of Ala and Thr (a "best case") shows that they are partially distinguishable. (Row 2) Comparison of the chemical shifts of Ala and Asn (a more "typical" case) shows that they are statistically indistinguishable: (A, E) $^{13}C^{\alpha}$ chemical shifts; (B, F) ^{15}N chemical shifts; (C, G) sums of $^{13}C^{\alpha}$ and ^{15}N chemical shifts; (D, H) differences of $^{13}C^{\alpha}$ and ^{15}N chemical shifts. (Row 3) Effect of tilted-plane data collection (angle shown in the figure) on peak distinguishability for sums of $^{13}C^{\alpha}$ and ^{15}N chemical shifts: (I) full population of Ala chemical shifts compared with a subpopulation of Thr chemical shifts; (J) full population of Ala chemical shifts compared with a subpopulation of Asn chemical shifts. In each case (Thr and Asn), the subpopulation was selected to represent the 20% of chemical shifts closest to the mean. The relative movement of each subpopulation with respect to the Ala distribution illustrates the idea that distinguishability can always be achieved when multiple angles are considered.

We present the details of these issues, including the measure of optimality given the sparsity of coefficients in the representation, in a more mathematical setting in the Supporting Information. The basic idea is to iteratively refine the positions of peaks by optimally selecting planes that offer the "best evidence" for refining the positions of peaks that best explain the observed data. Our approach employs Bayesian methods to refine the peak positions after the collection and peak analysis of each new tilted plane.

For optimal performance, the HIFI−NMR strategy requires a number of components: pulse sequences adapted for tilted-plane data collection, a method for statistical analysis of spectral space, a robust automated peak-picking algorithm, an algorithm for analyzing prior data to choose the next-best plane, an algorithm for deciding when to cease data collection, and an algorithm for statistical, probabilistic analysis of peak positions (frequencies in each dimension) and their validity.

Data Collection. The 2D spectra required for extracting 3D resonances by HIFI−NMR are the two orthogonal planes (F_1−F_3, F_2−F_3), as well as a small number of carefully selected tilted planes (aF_1bF_2−F_3). The ratio between the dwell times of the two simultaneously evolving dimensions determines the angle of the resulting tilted plane. We set the spectral windows for F_1 and F_2 to certain values and multiply the corresponding dwell times for F_1 and F_2, respectively, by the sine and the cosine of the chosen tilt angle. (This is equivalent to dividing the spectral windows by the sine and cosine of the angle.) Whereas this approach has the advantage of presenting only a single recognized parameter for the user to set, it has the consequence of fixing the spectral window of each tilted plane.

To avoid aliased peaks in tilted planes, we note that the size x of a spectral window is given by $x = \Delta t_N/\Delta t_C = sw_C/sw_N$, and that the simultaneous evolution can be written as $\Sigma_i \cos\{(\omega_C \pm x\omega_N)\Delta t_C i\}$, where Δt_N and Δt_C are the sampling or dwell times, sw_C and sw_N are spectral windows, and the subscripts indicate the nitrogen and carbon frequency domains, respectively. The frequency $\omega_C \pm x\omega_N$ is thus sampled at a rate sw_C. For standard values sw_C^{MIN} and sw_N^{MIN}, respectively, $x = sw_C^{MIN}/sw_N^{MIN}$, and the maximum combined frequency can be no larger than twice the highest carbon frequency. In our experiments, the sampling/dwell times are multiplied by cos α and sin α; therefore, it is sufficient to make both spectral windows larger by a factor $2\cos(\pi/4)$. However, by taking advantage of the data in the orthogonal planes, we can find a more efficient spectral window by filtering out outlier peaks, which by definition are uniquely identifiable. As a result, in practice, a factor of 1.2 was found to be sufficient in nearly all cases. An alternative strategy of adjusting spectral windows on the basis of observed data and expected peak positions can be devised in order to further optimize data collection.

To achieve quadrature detection, we use the States method[16] independently for both F_1 and F_2, with the result that four FIDs are recorded for each increment in the spectrum. These four FIDs, which are collected interleaved for each increment, represent all possible combinations of real and imaginary points for the indirect dimensions.

(16) States, D. J.; Haberkorn, R. A.; Ruben, D. J. *J. Magn. Reson.* **1982**, *48*, 286−292.

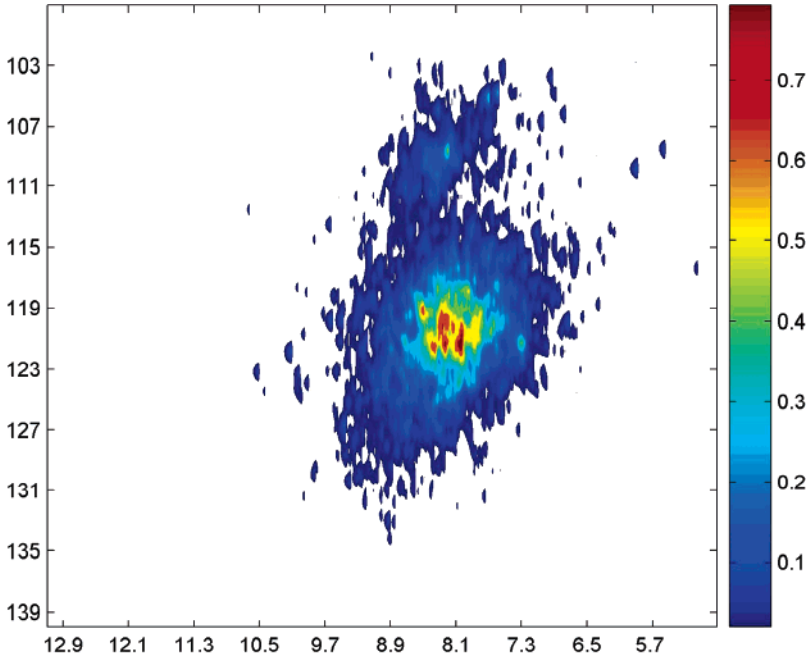


Figure 2. $^{1}H-^{15}N$ probability density plot (chemical shift priors) for the mouse protein Mm202773 generated from the sequence of the protein. The colors correspond to the probability scale on the right. The approach used in constructing chemical shift priors from data in BMRB will be described separately (A. Bahrami et al., to be published).

The States−TPPI modification for shifting axial peaks to the edge of spectra[17] is performed on only one of the simultaneously evolving dimensions.

Data are processed in a way similar to that described by Kozminsky and Zhukov.[9] The four FIDs are separated into two spectra, each containing the real and imaginary points for one of the simultaneously evolving dimensions (e.g. F_1), but only the real or the imaginary points of the second one (F_2). Fourier transformation of these two data sets results in two spectra, 90° out of phase, with duplicated peaks arising from the indistinguishable "+" and "−" F_2 frequencies.

The sum of these two spectra (after the phase of one of them is corrected by 90°) results in a spectrum that contains a single set of peaks with the frequencies of the two simultaneously evolving nuclei added together ("+ tilt" spectrum). The difference between the two spectra yields a spectrum containing the other set of peaks at the position of the subtracted frequencies ("− tilt" spectrum). The signal-to-noise ratio in either case would be equivalent to that of a single spectrum acquired over the sum of the data acquisition times of the two spectra.

In adapting 3D pulse sequences for HIFI−NMR, an important modification is made: the constant-time ^{15}N evolution that characterizes many triple-resonance experiments is converted into a constant−semiconstant-time evolution. This modification enables us to collect any desired number of points in the indirect dimension. This is critical to take full advantage of the ability of HIFI−NMR to resolve peaks by collecting 2D planes with more data points than would be reasonably possible with a full 3D data set. The ^{15}N evolution period behaves like a constant-time evolution, provided that the number of increments required can be contained within the constant-time delay, and it switches into a semiconstant-time evolution when a higher number of increments is required for adequate signal-to-noise. The excess time that cannot be accommodated by the constant-time delay is distributed uniformly throughout the ^{15}N increments. In testing pulse sequences incorporating this modified ^{15}N evolution period on both small (5 kD) and larger proteins (up to 20 kD) studied, we observed no significant distortion of the ^{15}N line shape, even when large numbers of increments were used.

(17) Marion, D.; Ikura, M.; Tschudin, R.; Bax, A. *J. Magn. Reson.* **1989**, *85*, 393−399.

In choosing the spectral window in the indirect dimension of tilted planes, it is important to realize that the simultaneous incrementation of both evolution times connects/interconnects the otherwise independent chemical shift evolutions. These evolutions can now be written as a product with a single time variable, $\cos(\omega_A t)\cos(x\omega_B t)$, where ω_A and ω_B denote the chemical shift evolution for the two simultaneously evolving nuclei, and x is the ratio between their time increments. By using standard trigonometric identities, this product can be rewritten as the sum of two terms, the first containing a weighted sum of the chemical shift evolutions and the second a weighted difference. This situation is similar to the double- and zero-quantum coherences created in the original RD experiments described by Szyperski[3] and Simorre,[4] and explains our observation that peaks in the indirect dimension of tilted planes can be aliased even though the individual simultaneous frequencies are within the range of the spectral window chosen for each of them. Thus, to avoid aliasing, it is necessary to choose a sufficiently wide spectral window in the tilted planes, as discussed above.

Statistical Organization of Spectral Space. We divide 3D spectral space into discrete voxels on the basis of prior information: the number of expected peaks, the nuclei involved and their expected chemical shift ranges, and the data collection parameters. Next, we obtain an estimate for the prior probability distribution of peak locations in the multidimensional spectral space. For this step, we combine information about the amino acid sequence of the protein and empirical as well as derived analytical distributions that are based on deposited chemical shifts (BMRB, www.bmrb.wisc.edu) by using a Markov Chain Monte Carlo (MCMC) inference method. The resulting joint probability distribution estimated in this step signifies the probability of a peak's existence in a specific voxel of the spectrum. In most cases, this distribution gives "strong evidence" to a few regions of spectral space and a relatively flat probability distribution in other regions (Figure 2). In regions of low probability, peaks generally are dispersed and readily detected. However, in regions of high probability, overlaps are likely, and so the algorithm collects additional planes whenever the model suggests evidence for ambiguity. We denote the prior distribution as $P_{pr}(\mathbf{X})$, where the elements of the vector $\mathbf{X} = (x_1,x_2,x_3)$ represent the voxels in *three*-dimensional space.

Overview of Signal Recovery. The initial step in the HIFI approach is to collect data for the two orthogonal planes (0° and 90° planes). For experiments involving ^{1}H, ^{15}N, and ^{13}C dimensions (for example, HNCO, HNCACB), the 0° (^{13}C−^{1}H) and the 90° (^{15}N−^{1}H) planes are collected and peak-picked. Although any automated peak-picking routine can be used to identify peaks in each 2D plane collected, we use a new algorithm (PEANUT) to enhance the peaks and remove noise so as to improve the quality of peak-picking (H. R. Eghbalnia et al., manuscript in preparation). As in other manual or automated approaches to NMR peak-picking, each plane threshold is selected slightly below the noise level so as to recover as many peaks as possible for the given plane. The final output of HIFI−NMR is an objective probabilistic analysis of all detected peaks and their rankings, which can be combined with similar HIFI−NMR analyses from other experiments in order to provide an effective and unbiased global list of ranked peaks.

Because the ^{1}H dimension is common to the two planes, the possible candidates for peaks in 3D space can be generated by considering all combinatorial possibilities for peaks in the two planes that have the same ^{1}H chemical shift within a given tolerance. This generates many more candidates than actual peaks but will include most, but not all, real peaks. These candidate peaks restrain the prior probability space generated in the earlier step. Voxels that have no positive probability in the 0° and 90° planes are assigned a lower probability. These probabilities may increase later if evidence for these voxels is discovered in subsequent 2D planes. As additional data are obtained, the principle of statistical invariance requires us to recompile and reevaluate the evidence in the sample space. Heuristically, we must base the current prior distribution on all current data, regardless of the order in which they were obtained. The correct application of this idea is critical to the success of our approach. At the end of this stage, we have gained new information about the probability space consisting of candidate peaks; we designate this probability as $P_0(x)$.

Statistical Evaluation of Signals. For the moment, we assume that data for a new tilted plane at the optimal angle have been obtained. The projection of the current set of candidate peaks in 3D space on the tilted plane at angle θ is the map D,

$$\mathbb{R}^3 \xrightarrow{D} \mathbb{R}^2: \quad D(x_1,y_2,z_3) = (x_1,y_2\cos(\theta) + z_3\sin(\theta)) \tag{1}$$

in which $\mathbf{X} = (x_1,x_2,x_3)$ are the Cartesian coordinates of a candidate peak. We first discretize the coordinates and obtain the associated voxel label. We next examine each candidate peak in turn, in light of the data from the new tilted plane and prior data, and update its probability on the basis of Bayes' rule.

To do this, we compare a candidate peak x that arises from the projection $x_\theta = D(x)$ with the set of automatically picked peaks a_1, a_2, a_3, ..., a_k (treated as distributions) in the latest tilted plane θ by defining the statistical event,

$$x\_observed\_in\_\theta \equiv (\exists a_i \ni d(x_\theta,a_i) < \delta_i) \tag{2}$$

in which d represents the distance based on the divergence measure and δ_i is the allowable statistical deviation of peak a_i. In intuitive language, measuring the event $x\_observed\_in\_\theta$ is equivalent to asking whether or not the peak x has been observed in the new tilted plane designated by the angle θ. The use of divergence instead of Euclidean metric is desirable, because ^{1}H and the mixture of ^{15}N and ^{13}C have different tolerance values and the tolerance values vary from plane to plane.

By invoking Bayes' rule, we find that the probability of x being a "real peak", $P_0(x)$, is updated after observation of peaks in tilted plane θ by

$$P_1(x) \equiv P(x|x\_observed\_in\_\theta) = \frac{P(x\_observed\_in\_\theta|x)\,P_0(x)}{P(x\_observed\_in\_\theta)} \tag{3}$$

or

$$P_1(x) \equiv P(x|x\_not\_observed\_in\_\theta) = \frac{P(x\_not\_observed\_in\_\theta|x)\,P_0(x)}{P(x\_not\_observed\_in\_\theta)} \tag{4}$$

in which $x\_not\_observed\_in\_\theta$ is the complement event for $x\_observed\_in\_\theta$. By "event X" we mean that the candidate peak x is statistically distinguishable from noise and therefore is a real peak. Note that $P_0(x)$ is the prior probability that we have calculated in the previous stage. The conditional likelihood $P(x\_observed\_in\_\theta|x)$ represents the probability that the candidate peak x would be observed in tilted plane θ, given that it is a "real peak". The complement of this probability, $P(x\_not\_observed\_in\_\theta|x)$, is defined in the same way. The term in the denominator, $P(x\_observed\_in\_\theta)$ (or $P(x\_not\_observed\_in\_\theta)$), represents the probability of a voxel x in plane θ to be considered a peak (or not a peak), regardless of any other consideration. We consider this probability to be independent of the angle θ and obtainable only from our prior distribution. In cases where x is located in a crowded region, such that evidence for its existence may vary from plane to plane taken at different angles, we use empirical database information to determine an "independent estimate" by isolating the crowded region and by estimating the likelihood that our observation is false. The independent estimate enables us to approximate the probability $P(x\_observed\_in\_\theta)$.

The probabilities $P(x\_observed\_in\_\theta|x)$ and $P(x\_not\_observed\_in\_\theta|x)$ do not experience much angular variation and can be reasonably approximated by the empirical data.

Choice of the "Best Tilted Plane". The best tilted plane angle θ is selected according to the current probability distribution $P_0(x)$ and a mathematical method designed to obtain maximum information. We describe the method briefly here but point out that important technical steps (described in the Supporting Information) must be adhered to in order to obtain the optimal angle. The next-best tilted plane is the one that maximizes information about the positions of the peaks. To find this plane, we assume that the candidate peaks have been observed on a plane at angle θ and evaluate the information theoretic impact of this observation on the probability of the peaks. The measure of this influence is the divergence S_θ between the initial probability $P_0(x)$ (probability at the initial step, or step zero) and the predicted probability after the impact of the projection at angle θ has been taken into account, $Q_\theta(x)$:

$$S_\theta = -\sum Q_\theta \ln\frac{Q_\theta}{P_0} \tag{5}$$

where the sum is taken over all voxels in the spectral space. The choice for our next plane optimizes this measure:

$$\theta_{\text{optimal}} = \arg\max_\theta (-S_\theta) \tag{6}$$

Intuitively, by assuming a uniform prior over all plane selections, the maximum information method tends to choose as the next plane that with the highest dispersion of projected peaks. Two details (discussed in the Supporting Information) are worth noting here: (a) our above discussion of probabilities is in reference to the parameters of the model describing the spectra, and (b) a unique optimal angle may not exist.

After the first plane is chosen and the probabilities updated to $P_i(x)$ (probability at the ith step) by use of eqs 3 and 4, the next-best plane is again selected as described above. The probability updating rule after processing the ith plane θ_i is

$$P_i(x) \equiv P(x|x\_observed\_in\_\theta_i) = \frac{P(x\_observed\_in\_\theta_i|x)\,P_{i-1}(x)}{P(x\_observed\_in\_\theta_i)} \tag{7}$$

or

$$P_i(x) \equiv P(x|x\_not\_observed\_in\_\theta_i) = \frac{P(x\_not\_observed\_in\_\theta_i|x)\, P_{i-1}(x)}{P(x\_not\_observed\_in\_\theta_i)} \quad (8)$$

This process is continued until no further information (or negligible information) can be obtained (i.e., the divergence is below a threshold). At this stage, n peaks with probabilities above a threshold will have been reported as the peak list. The value of n is estimated automatically from the number of amino acid residues in the protein, the type of experiment, and the probability distribution $P(x)$. The value of n can be adjusted in a subsequent postprocessing step to include additional expert information regarding the observed data.

Results and Discussion

Test of the HIFI−NMR Approach with 3D Spectra of Proteins. Prior to implementing the adaptive tilted-plane algorithm on an NMR spectrometer, we tested HIFI−NMR in a nonautomated environment. We collected orthogonal-plane and tilted-plane data (at either 5° or 10° intervals between 0° and 90°) according to the Freeman and Kupce RD method[11] for 3D triple-resonance experiments and then used the adaptive tilted-plane algorithm to select which of these in turn would be the next tilted plane for the peak-picking and analysis engine of HIFI−NMR. The proteins we used as test cases were ones that had been studied thoroughly by conventional methods. This enabled us to compare peaks "identified" by HIFI−NMR with those "identified" by manual peak-picking of conventional 3D data sets and to classify these peaks as "correct" (i.e., in agreement with all available data) or as "noise" (identified peaks that did not correspond to an assignment).

Data Sets Collected. To test the efficacy and accuracy of our approach, we colleted a total of eight data sets from five proteins: brazzein (54 residues), ubiquitin (76 residues), mouse protein Mm202773 (101 residues), *Anabaena variabilis* flavodoxin (179 residues), and Prp24_12 (166 residues). All proteins were labeled uniformly with carbon-13 and nitrogen-15.

We show results here for three experiments (HNCO, HNCACB, and CBCA(CO)NH) typically used in determining protein backbone assignments. As controls, we collected parallel data sets by the standard 3D data collection approaches, processed these by 3D FT, and handpicked the peaks for comparative analysis. For proteins other than Prp24_12, peak lists corresponding to solved NMR structures were used to validate peaks as "real"; in the case of Prp24_12, which had no solved structure, the only basis of comparison was handpicked 3D FT data.

Typically, for each 2D plane, four transients were accumulated for each FID (with the exception of the HNCACB planes of the mouse protein, in which eight scans were needed to observe all the intra- and interresidue cross-peaks). Given our modified ^{15}N evolution period (vide supra), 80 and 128 increments were collected in the indirect dimension of each plane (orthogonal or tilted) for HNCO and HNCACB experiments, respectively. However, in the CBCACONH pulse sequence, the constant-time evolution in the ^{13}C dimension limited the number of increments that could be recorded (for an 80 ppm spectral window) to 71. The orthogonal planes were recorded first, with two FIDs for each increment, to allow for quadrature detection in the indirect dimension, according to the States method. As described above, for each tilted plane, each FID was collected four times in order achieve quadrature detection for both simultaneously evolving nuclei. After processing, two planes were obtained for each chosen angle: those corresponding to the "+" and "−" tilt geometries.

In each experiment, we phased the initially collected plane and used the phasing parameters to phase all subsequent planes automatically. We found that specification of an approximate value for the noise level (as derived from the phasing step) was helpful in improving the efficiency of the plane selection algorithm but did not alter its predictions.

Comparison of Peaks Identified by HIFI−NMR to Those from the Controls. Ubiquitin proved not to be a challenging protein in that its peaks are nearly ideally distributed. For example, HIFI−NMR achieved the identification of all correct peaks in the 3D HNCO spectrum after the selection of a single extra plane beyond the two orthogonal planes. In fact, by experimentation we found that peak recovery was not particularly sensitive to the choice of angle for this third plane. For this reason, we do not discuss the ubiquitin data further here.

In the case of CBCA(CO)NH HIFI−NMR data collected for the protein brazzein (Table 1, panel A), we recovered the full 3D data with two orthogonal planes plus two tilted planes chosen sequentially by the HIFI−NMR engine. The first "best tilted plane" was predicted as 41°, and that actually used was the pre-collected plane at 40°. At that stage, the signal recognition module identified 103 peaks (including 82 of 87 correct peaks). The next-best plane was predicted as 34°, and that actually used was the pre-collected plane at 35°. Following analysis of these data, 116 peaks were recognized (including all 87 of the correct peaks).

The algorithm identified this as the stopping place, but as a test, a series of five additional best tilted planes were collected (at 50°, 20°, 25°, 60°, and 70°), and the results were analyzed sequentially (Table 1, panel A). The addition of these tilted planes yielded no more correct peaks and only served to increase the number of noise peaks. By comparison, the 3D FT control yielded 111 handpicked peaks (including all 87 correct). With a random selection of the pre-selected tilted planes (no next-best plane selection), four tilted planes were required on average to recover all 87 peaks (results not shown).

In the case of the 3D CBCA(CO)NH HIFI−NMR data collected for mouse protein Mm202773 (Table 1, panel B), the ideal first tilted plane was predicted as 51° (50° used), and three additional planes (55°, 40°, and 70°) were needed to reach the identified stopping point at which 241 peaks were detected (including 177 correct). The control 3D FT experiment identified 223 peaks (including 178 correct) (Figure 3). In this case, the addition of the next-best tilted plane beyond the predicted stopping point served to decrease the number of noise peaks from 64 to 45, but additional tilted planes increased the number of noise peaks without improving the number of correct peaks.

We collected two additional HIFI−NMR data sets (HNCACB, HNCO) for mouse protein Mm202773. The results (Table 1, panels C and D) clearly show that the number of tilted planes needed for optimal signal identification and their angles are dependent on the experiment type. Seven tilted planes were required to reach the stopping point in the HNCACB experiment, whereas only three were required in the HNCO experiment. For the HNCO experiment, the three sequentially chosen ideal next tilted planes were at 49°, 37°, and 9° (50°, 35°, 10°

Table 1. Comparison of the Number of Peaks Extracted Manually from Three Triple-Resonance Experiments (Used as Control) with Those Extracted Automatically by HIFI−NMR after the Collection of the Number of Tilted Planes Specified (1−7)[a]

Panel A CBCA(CO)NH – brazzein – 54 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		40°	35°	50°	20°	25°	60°	70°
Peaks identified	111	103	116	117	118	120	119	118
Correct peaks	87	82	87	87	87	87	87	87
Noise peaks	24	21	29	30	31	33	32	31
Panel B CBCA(CO)NH – mouse protein Mm202773 – 101 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		50°	55°	40°	70°	30°	65°	20°
Peaks identified	223	215	201	225	241	222	232	241
Correct peaks	178	171	174	176	177	177	177	177
Noise peaks	45	44	27	49	64	45	55	64
Panel C HNCACB – mouse protein Mm202773 – 101 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		20°	10°	30°	40°	50°	60°	70°
Peaks identified	387	425	371	403	427	439	419	472
Correct peaks	335	303	308	318	321	324	325	329
Noise peaks	52	122	63	85	106	115	94	143
Panel D HNCO – mouse protein Mm202773 – 101 residues								
	Manual	Plane 1	Plane 2	Plane 3	Plane 4	Plane 5	Plane 6	Plane 7
Tilted Plane		50°	35°	10°	70°	20°	25°	45°
Peaks identified	110	115	109	117	113	115	116	115
Correct peaks	91	90	91	92	91	91	91	91
Noise peaks	19	25	18	25	22	24	25	24
Panel E HNCO – Prp24_12 protein – 166 residues								
	Manual	Plane 1	Plane 2	Plane 3[a]	Plane 4	Plane 5		
Tilted Plane		24°	38°	51°	71°	60°		
Peaks identified	221	182	180	166	175	185		
Correct peaks	136	117	122	133	134	135		
Noise peaks	85	65	58	33	41	50		

[a] Data were collected at 600 MHz on a Varian/Inova NMR spectrometer equipped with a cold probe. HIFI stops plane collection at columns marked with heavier lines and gray background.

used). By comparison, the first three optimal planes for HNCACB were located at 21°, 12°, and 32° (20°, 10°, 30° used). These results demonstrate that the optimal selection of planes plays an important role in the rapid identification of peaks.

The resolution of peaks in the optimal first plane for the HNCACB HIFI−NMR experiment (20°) was found to differ greatly from that of a plane acquired at 45°, the fixed tilt angle of the GFT−RD or TPPI−RD experiments (Figure 4). At 45°, many peaks canceled out, owing to overlaps of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ signals, whereas the plane at 20° showed much clearer peak separation.

Conclusions

Our results show that the HIFI−NMR approach is capable of automatically identifying 98−101% of the peaks shown to be correct in handpicked data from conventional 3D FT spectra (Table 1). The positions of the peaks picked automatically by HIFI−NMR were statistically indistinguishable from those determined manually (<0.04 ppm in ^{13}C and ^{15}N; <0.003 ppm in ^{1}H). However, as a percentage of total peaks, HIFI−NMR returned 4−14% more noise peaks than were handpicked from 3D FT data sets. The required time for HIFI−NMR data collection was typically on the order of one-tenth (never less than one-fourth) that for conventional 3D FT (Table 2). By providing probabilistic peak lists as output (peaks with corresponding signal likelihood and uncertainties in frequencies), HIFI−NMR obviates the need for lengthy postprocessing and peak-picking as needed for 3D FT or other RD approaches.

The mathematical technique and the algorithm introduced in HIFI−NMR represent important steps toward achieving more complete automation of biomolecular NMR spectroscopy. The nonautomated results shown here involved the processing of pre-collected tilted-plane data, which was made available to the algorithm one plane at a time according to the angle selected. Our results (Table 1) show that the angle of the ideal first plane depends on both the experiment and the protein and is rarely the same from one case to the next. Thus, adaptive sampling provides an efficient approach to data collection.

We recently have succeeded in integrating these tools with a commercial Varian NMR spectrometer so as to enable real-time collection by the HIFI−NMR approach. This has enabled us to achieve the predicted gains (Table 2) afforded by real-time adaptive tilted plane selection and to collect data at the actual tilt angle selected. Upon repeating HIFI−NMR data collection of the HNCO data for protein Mm202773, the tilted planes selected and used were 54°, 35°, and 71°. When requested to select an additional plane beyond the determined stopping point, a fourth plane was added at 8°. Interestingly, these angles are similar to those selected (49°, 37°, 9°, 70°) and used (50°, 35°, 10°, 70°) in the nonautomated trial, but they were chosen in a different order. In both cases, the algorithm called for the same number of tilted planes, and the numbers of correct peaks and noise peaks returned were identical.

With this integrated tool, we have collected data sets and extracted 3D peaks from two larger proteins (19−20 kDa). For

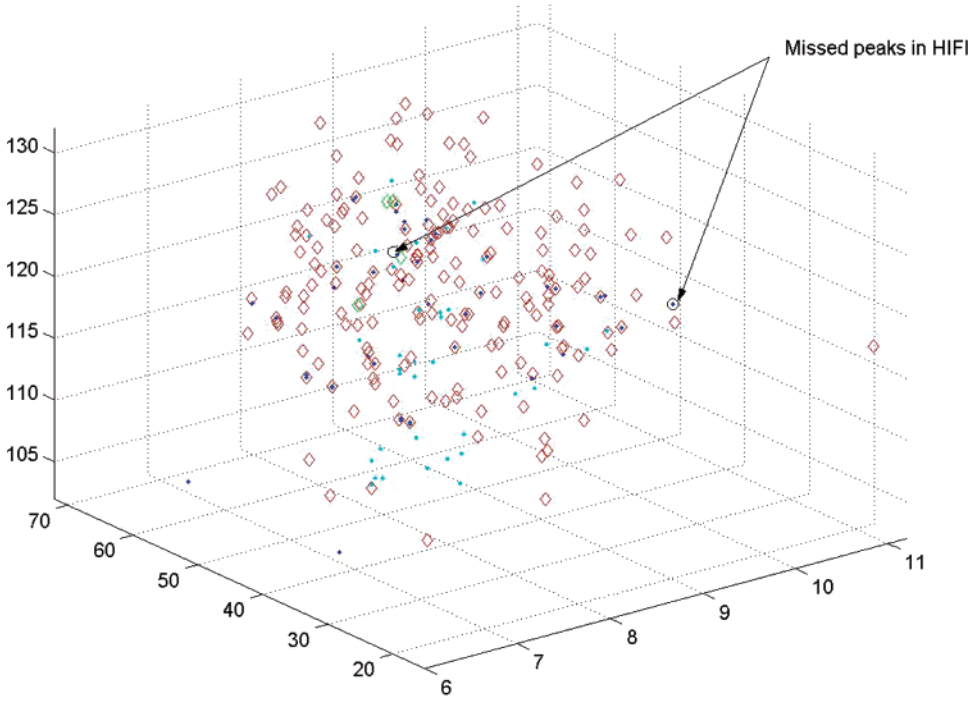


Figure 3. Schematic results from a HIFI−NMR version of the CBCA(CO)NH experiment. Displayed here are results generated from the two orthogonal planes plus three adaptively selected planes. The protein sample was mouse protein Mm202773. Red diamond symbols correspond to peaks identified by both HIFI−NMR and manual analysis of conventional 3D that correspond to real signals; green diamond symbols correspond to peaks identified by both HIFI−NMR and manual 3D that correspond to noise; blue dots correspond to peaks identified by manual 3D analysis but not by HIFI−NMR that correspond to real signals.

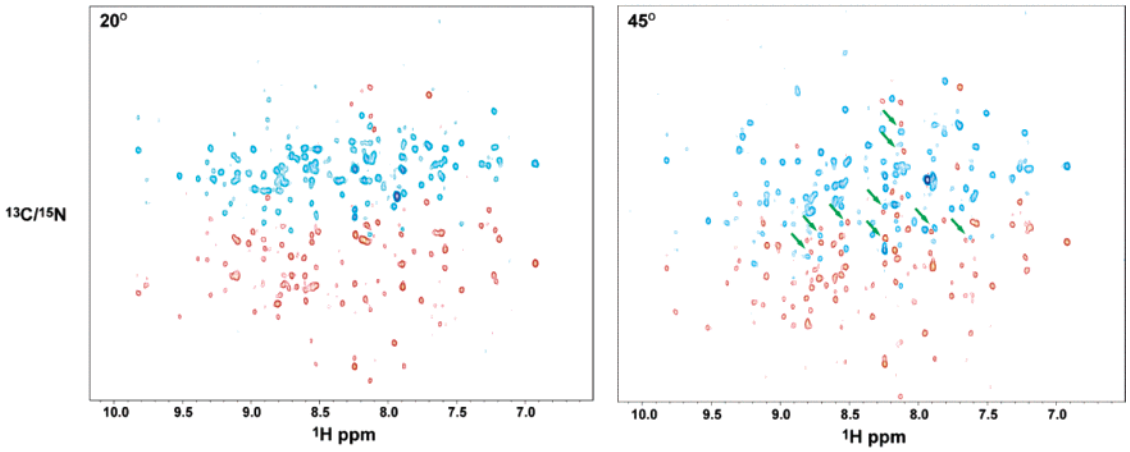


Figure 4. Two 2D tilted planes from the HNCACB HIFI−NMR spectrum of the mouse protein collected as described in the text. The 2D spectrum for the optimal first plane predicted by our algorithm at 20° is compared with the 2D spectrum for 45°. The arrows indicate the positions of signals that have been canceled because of spectral overlaps.

example, HNCO data from a perdeuterated sample of Prp24_12 (166 residues, 19.2 kDa) were collected and automatically processed from five planes in approximately 6.2 h to generate a 3D peak list with 135 entries. By comparison, 3D FT data collection for the same sample required 22 h (plus time for manual peak-picking) and resulted in 136 peaks, including all 135 identified by HIFI−NMR (Table 1, panel E). For flavodoxin (179 residues, 20 kDa), only four planes were needed to recover the peaks from HNCO or HNCOCA experiments. These results demonstrate that the current version of HIFI−NMR, when used for 3D NMR experiments of the type used in protein backbone assignments, can automatically recover more than 98% of the peaks found by conventional 3D FT in a fraction of the time.

The mathematical and computational tools proposed are sufficiently general to allow the combination of information from various sources into a single model. This will make it possible to improve the overall process through the integration of data

Table 2. Comparison of the Performance of HIFI−NMR with Conventional (Manual) Methods for Data Collection, Analysis, and Peak-Picking of Data from [U-^{13}C,U-^{15}N]-Mm2022773 (101-Residue Mouse Protein) Collected at 600 MHz on a Varian/Inova NMR Spectrometer Equipped with a Cold Probe

Experiment	CBCA(CO)NH		HNCO	
Data collected	HIFI	3D FT	HIFI	3D FT
Number of identified peaks:	215	223	117	110
Number of correct peaks	177	178	92	91
Total time required for data collection.	2 h [a]	22 h	1 h [a]	12 h

[a] Time for HIFI includes automatic generation of the peak list; this process is performed separately in 3D FT.

from multiple experiments. For example, by combining data from 3D experiments with common data planes (for example, CBCA(CO)NH, HNCACB, and HNCO), it should be possible to improve the discrimination between signal and noise. Most importantly, the direct output from HIFI−NMR is a peak list that can be used as input to automated assignment software packages.

The mathematical methods and associated tools developed here are part of a broader effort to achieve highly efficient and streamlined approaches for NMR structure determination and validation. A subset of these automation tools is available for public use at bija.nmrfam.wisc.edu.

Acknowledgment. This research was supported by Biomedical Research Technology Program, National Center for Research Resources, through NIH grant P41 RR02301, which supports the National Magnetic Resonance Facility at Madison. A.B. received partial support from the National Institute of General Medical Science's Protein Structure Initiative through NIH grant 1 P50 GM64598, which supports the Center for Eukaryotic Structural Genomics. During part of this work, H.E. was supported as a postdoctoral trainee by the National Library of Medicine under grant 5T15LM005359. We thank Eldon L. Ulrich and William M. Westler for advice and encouragement. This work made extensive use of the BioMagResBank and the Protein Data Bank.

Supporting Information Available: Mathematical details for selecting the next tilted plane and overcomplete representation, and an example illustrating the use of distributions on the parameter space for predictive density estimation of a normal model and how the parameters for intermediate densities need an enlarged space. This material is available free of charge via the Internet at http://pubs.acs.org.

JA052120I