J Biomol NMR (2011) 50:247–262
DOI 10.1007/s10858-011-9512-6

ARTICLE

Knowledge-based nonuniform sampling in multidimensional NMR

Adam D. Schuyler · Mark W. Maciejewski · Haribabu Arthanari · Jeffrey C. Hoch

Received: 28 February 2011 / Accepted: 7 May 2011 / Published online: 29 May 2011

Abstract The full resolution afforded by high-field magnets is rarely realized in the indirect dimensions of multidimensional NMR experiments because of the time cost of uniformly sampling to long evolution times. Emerging methods utilizing nonuniform sampling (NUS) enable high resolution along indirect dimensions by sampling long evolution times without sampling at every multiple of the Nyquist sampling interval. While the earliest NUS approaches matched the decay of sampling density to the decay of the signal envelope, recent approaches based on coupled evolution times attempt to optimize sampling by choosing projection angles that increase the likelihood of resolving closely-spaced resonances. These approaches employ knowledge about chemical shifts to predict optimal projection angles, whereas prior applications of tailored sampling employed only knowledge of the decay rate. In this work we adapt the matched filter approach as a general strategy for knowledge-based nonuniform sampling that can exploit prior knowledge about chemical shifts and is not restricted to sampling projections. Based on several measures of performance, we find that exponentially weighted random sampling (envelope matched sampling) performs better than shift-based sampling (beat matched sampling). While shift-based sampling can yield small advantages in sensitivity, the gains are generally outweighed by diminished robustness. Our observation that more robust sampling schemes are only slightly less sensitive than schemes highly optimized using prior knowledge about chemical shifts has broad implications for any multidimensional NMR study employing NUS. The results derived from simulated data are demonstrated with a sample application to PfPMT, the phosphoethanolamine methyltransferase of the human malaria parasite Plasmodium falciparum.

Keywords Spectrum analysis · Computational methods · Nonuniform sampling · Envelope matched sampling · Maximum entropy · PfPMT

Electronic supplementary material The online version of this article (doi:10.1007/s10858-011-9512-6) contains supplementary material, which is available to authorized users.

A. D. Schuyler · M. W. Maciejewski · J. C. Hoch (✉)
Department of Molecular, Microbial and Structural Biology, University of Connecticut Health Center, Farmington, CT 06030-3305, USA
e-mail: hoch@uchc.edu

H. Arthanari
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA 02115, USA

Introduction

Somewhat paradoxically, obtaining high resolution in the indirect dimensions of multidimensional NMR experiments becomes more difficult as the magnetic field increases. The difficulty results from the need to maintain the same maximum evolution time, while the stronger magnetic field increases spectral dispersion, thus requiring a decreased interval between samples to avoid aliasing. The combined effect is a greater sampling burden which translates to longer experimental acquisition times. A solution to this problem is to collect samples at nonuniform intervals in the indirect dimensions, permitting data to be collected at long evolution times (affording high resolution) without collecting all intervening samples at the Nyquist interval. Using non-Fourier methods of spectrum analysis, this data can then be used to reconstruct the complete frequency

domain spectrum. The success of approaches based on nonuniform sampling (NUS) depends on the distribution of sampled times, the nature of the signal and on the reconstruction method.

One of the earliest applications of NUS in multidimensional NMR was the so-called "accordion" experiment (Bodenhausen and Ernst 1981), in which two indirect time dimensions were sampled in concert. Application of a 2D Fourier transformation results in a 2D projection of the 3D spectrum. The first general approach for reconstructing the full-dimensional spectrum employing NUS was described by Barna et al. (1987). Exponentially-weighted random sampling is employed in the indirect dimension of a 2D NMR experiment and maximum entropy reconstruction (Barna et al. 1986; Hoch and Stern 2001) produces the full spectrum. More recently, GFT (Kim and Szyperski 2003) and back-projection (Kupce and Freeman 2003) approaches emerged and employ coupled evolution periods (as in accordion experiments) to seek high dimensional information from low dimensional experiments. Back projection attempts to reconstruct the fully dimensional spectrum, whereas GFT attempts a parametric decomposition of the spectral features. Back projection proved to have enormous heuristic value for introducing the principles and advantages of NUS to a wider audience. However, it has been demonstrated that the use of coupled evolution periods is not as efficient as less regular NUS approaches when the goal is to reconstruct the fully-dimensional spectrum (Mobli et al. 2006).

Applications of coupled evolution periods include methods in which the aim is not to reconstruct the fully dimensional spectrum, but to tabulate peak positions in various projections and deduce frequencies and correlations through analysis of data tables. The two most prominent methods, HIFI (Eghbalnia et al. 2005) and APSY (Hiller et al. 2005), differ in a number of respects, but they both seek to optimize sampling by choosing projection angles that maximally resolve resonances based on prior knowledge of chemical shifts. More general approaches to NUS that are not restricted to sampling along radial vectors have previously not exploited information about chemical shifts. We show here how the more general approach to NUS can be adapted to exploit chemical shift information.

The notion of the "matched filter" (North 1943; Van Vleck and Middleton 1946) was applied in NMR spectroscopy by Ernst (1966) and has become a bedrock principle for the design and implementation of NMR experiments. In its most prominent application, the principle stipulates that the time domain apodization function that optimizes the sensitivity in the frequency domain is an exponential decay with a rate that matches the decay of the signal envelope (R2). Barna et al. (1987) introduced "matched sampling" by reasoning that the sensitivity of a spectrum obtained from NUS data would be optimized by ensuring that the sampling density decays exponentially at the same rate as the signal envelope.

The relationships between matched filtering, matched sampling and knowledge-based sampling are well illustrated with an example signal containing two exponentially decaying sinusoids of different frequencies (Fig. 1). The composite time domain signal contains a beat with a period corresponding to the difference in frequencies. The decay rates of each sinusoid define the signal envelope (green line) for the composite signal. Matched filtering uses the signal envelope to define an apodization function that scales the time domain data and matched sampling uses the signal envelope to bias the NUS sampling towards the higher signal amplitudes observed at the shortest evolution times. While signal intensity does decay according to the signal envelope, the beat (blue line) introduces signal modulation *within* the signal envelope. Alternative forms of knowledge-based matched sampling make use of this feature and bias the NUS sampling to the beat envelope (blue line) or, depending on the level of a priori knowledge, may even follow the exact signal evolution (black line).

While matched sampling has great heuristic appeal—one samples more frequently when the signal is more intense—there appear to be few formal results on the optimality of matched sampling. The goal of this work is to characterize the performance of various NUS techniques relative to a conventional Fourier Transform (FT) which employs the complete, uniformly sampled data set. This is achieved by constructing and processing synthetic data sets while evaluating the response of sensitivity and resolution over a range of injected noise amplitudes. Here we use maximum entropy (MaxEnt) reconstruction to compute spectra from NUS data sets.

Methods

The synthetic data analysis consists of the following stages: (1) system selection, (2) master data construction, (3) FT processing, (4) NUS schedule generation, (5) MaxEnt processing, (6) peak picking, (7) frequency domain analysis of peaks, (8) time-domain analysis of NUS schedules. A schematic of this procedure is shown in Fig. 2. Each of the stages is discussed in the following sections.

The primary platform for data construction and processing is the Rowland NMR Toolkit (RNMRTK, Hoch and Stern 1996, http://rnmrtk.uchc.edu). The built in functions of the toolkit are integrated with BASH scripts and the computations are performed in parallel across a cluster of modest, off-the-shelf PCs. The spectral data is imported into MATLAB for analysis and plot generation.

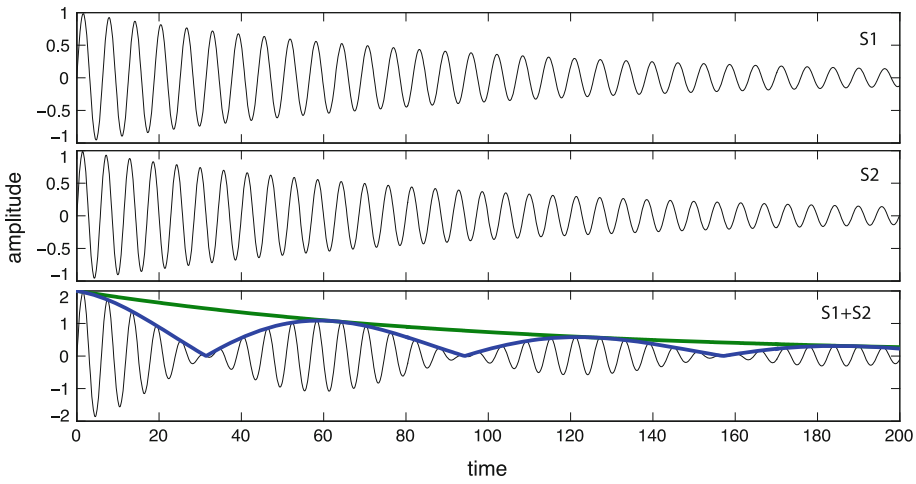

Fig. 1 Matched filter versus matched sample. Sinusoidal decay functions differing in frequency by 10%, labeled S1 and S2. The *bottom panel* shows the composite signal (*black*) produces a beat pattern with embedded sinusoidal modulation. The signal decay envelope (*green*) defines the time domain apodization function employed by matched filtering and defines the decay rate employed by matched sampling. Knowledge-based matched sampling approaches may follow the signal envelope (*green*) or even features within the envelope, such as the beat envelope (*blue*) or the signal itself (*black*)

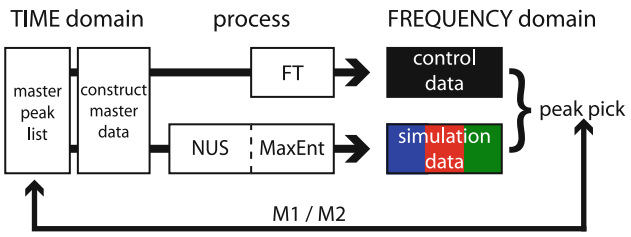

Fig. 2 Processing schematic. The master data is constructed in the time domain based on the master peak list and includes various noise profiles. The master data is processed by the traditional FT to produce a control spectrum in the frequency domain. In parallel, the BMS, EMS and RS NUS approaches are each paired with MaxEnt reconstruction to generate spectra. All spectra are peak picked and the M1/M2 metrics are used to compare the recovered peaks with those in the master peak list

This process is automated and the scripting framework is modular to allow for the incorporation of new components. The main scripts are available in the Supplemental Data online.

System selection

The simulations are performed across two indirect time dimensions (a direct dimension is not necessary). The first dimension is defined as ^{15}N and spans 64 data points with a spectral width of 1525.0 Hz, carrier frequency of 60.8 MHz and a 0 Hz reference value of 117.0 ppm. The second dimension is ^{13}C and spans 128 data points with a spectral width of 9650.0 Hz, a carrier frequency of 150.9 MHz and a 0 Hz reference value of 42.0 ppm. These values are typical for a 20 KDa protein at a field strength of 600 MHz.

Master data construction

The master peak list, referred to as the set $\mathcal{M}$, is composed of 82 peak locations, generated to reflect ^{13}C and ^{15}N shift distributions for a hypothetical protein. Each peak in $\mathcal{M}$ is referred to as p_m, and is defined by its frequency, line width and magnitude. These parameters are discussed below.

Extending the sampling evolution time of an NMR experiment continues to improve resolution until the point at which the signal has diminished and contributions from noise dominate. Zero filling and linear prediction are frequently used to extend the maximum time increment to avoid reducing signal to noise ratio (S/N). Zero filling and linear prediction are not employed for the synthetic data utilized in this analysis. Instead, the R2 decay rates are simply defined in each dimension such that the FID decays down to 10% magnitude at the final time point. This relationship is given by the expression

$$R2 = \frac{-SW \cdot \ln p}{m - 1} \quad (1)$$

where SW is the dimension's spectral width and p is the percentage of the FID signal remaining at sample point m. The constraints are met for a ^{15}N R2 decay rate chosen at 55.7370 Hz and the ^{13}C R2 decay rate chosen at 174.9602 Hz. These R2 rates correspond to line widths of 17.7417 and 55.6916 Hz, respectively, which are used to define the peaks in the master data set.

The relative intensities of the peaks, especially those in close proximity, have a strong impact on peak detection and accurate peak location recovery. This is addressed by fixing the peak positions and line widths (as defined above), but randomly varying the peak intensities over the

range 0.1–20 in order to generate 10 peak lists. Each of the 10 peak lists is injected into the time domain using the RNMRTK inject function. These time domain data sets represent the raw data.

There are several characteristics of the synthetic data that require further discussion. First, the inject function generates cosine-modulated data. This choice does not render the results irrelevant to sine-modulated time domains, which are, for example, utilized by the popular COSY experiments. Biasing a sample schedule to the signal envelop of sine-modulated data is qualitatively no different than biasing the sampling to the signal envelope of cosine-modulated data. Additional discussion on NUS sampling of sine-modulated data is available in Schmieder et al. (1993).

Second, the inject function generates Lorentzian line shapes. Experimentally collected data is often non-Lorentzian and may include distorted peaks as a result of magnet inhomogeneity. In the present study, for the purpose of most clearly identifying NUS performance trends, a "pure" synthetic data set is employed with ideal Lorentzian line shapes. In addition, the analysis steps, in particular the peak picker used to quantify the spectra, make no assumptions about the line shapes. The fidelity of our peak picker should be very similar for Lorentzian peaks or peaks with less than ideal line shapes.

Third, noise is naturally occurring in real data, but must be artificially injected on top of the synthetic raw data with the RNMRTK noise function. The noise is defined by its root mean square deviation (RMSD) and is taken at 25 values: from 0 to 7.5 in 0.5 step increments and from 10 to 50 in 5 step increments. Noise profiles are not deterministic; accordingly, 25 repetitions are made at each RMSD value of injected noise.

FT processing

The control data is computed with the conventional FT function from RNMRTK. This approach to signal processing utilizes the complete, uniformly sampled, 64 $\times$ 128 grid of time domain data. FT processing is performed on the 10 master data sets at the 25 noise levels with 25 repetitions at each noise level, for a total of 6,250 spectra.

NUS schedule generation

The first NUS approach, envelope matched sampling (EMS), biases random sampling with an exponential decay (green line from Fig. 1) and follows the principles introduced by Barna and colleagues. The sample schedules are generated with the standalone program, ScheduleTool (Maciejewski 2011). In the case of this synthetic data analysis, the decay rate of the EMS schedule can be fixed at the decay rate of the injected peaks, which is known and uniform across all peaks. However, in application, the decay rates of peaks are not necessarily uniform and may not be known accurately in advance. Furthermore, it is far from clear that matched sampling is optimal in the context of NUS and non-Fourier spectrum reconstruction. In order to explore these possibilities EMS sample schedules are generated for a range of decay rates, which are defined relative to the actual decay rates of the injected peaks. The set of scaling factors is: 0.5×, 1×, 1.5×, 2× and 5×. This range of scaling probes the robustness of EMS and helps identify optimal NUS approaches with respect to the resolution and sensitivity metrics.

The second NUS approach, beat matched sampling (BMS), collects data only at the greatest time domain signal intensities (black line from composite signal of Fig. 1). This approach is an extension of the matched sampling principle proposed by Barna and colleagues and aims to construct a sample schedule based on features *within* the signal envelope. In order to match these features, BMS schedule generation requires prior knowledge of peak locations. In the case of this synthetic analysis, the master peak list is provided as input, thus allowing the optimal performance of BMS to be tested. In practice, scout experiments may be used to determine peak locations prior to NUS multi-dimensional experiments.

The third NUS approach, random sampling (RS), is not matched—it is a random distribution of sample points across the time domain. This approach serves as a control to assess the effectiveness of the sampling bias introduced by the EMS and BMS approaches.

A primary parameter for all NUS schedules is sample coverage, which defines the percentage of time domain data values that are collected, relative to the uniform Nyquist grid with the same maximum evolution times. Coverage is a zeroth order description of a given sample schedule. Higher order descriptions quantify the distribution of sample points across the indirect dimensions. The coverage values used in this study are 3, 6, 12 and 24%. The BMS approach is deterministic and a single sample schedule is generated at each sample coverage value for a total of 4 schedules. The EMS approach has a stochastic component and 5 schedules are generated at each combination of sample coverage value and decay rate scaling for a total of 100 schedules. The RS approach is stochastic and 5 schedules are generated at each sample coverage value for a total of 20 schedules. A summary schematic of the NUS schedule types, including graphical representations of typical schedules, is shown in Fig. 3.

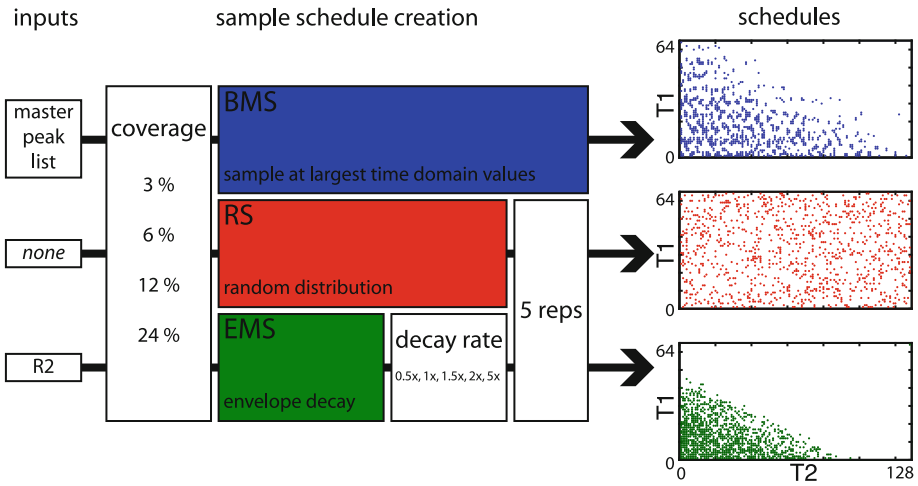

Fig. 3 Sample schedule generation. Schematic showing the input parameters that define each of the 3 NUS sample schedule types. The RS and EMS schedules incorporate a random component, which is not dependent upon the system (and thus not listed as an input). The random component is addressed by generating 5 repetitions of each RS and EMS schedule. The schedules shown at the *right* are taken at a sample coverage of 12% and the EMS schedule is at a 1× decay rate scaling factor (i.e. the decay rate of the sample schedule is matched to the decay rate of the signal envelope)

MaxEnt processing

MaxEnt reconstruction paired with each of the 124 sample schedules generated in the previous section is performed on each of the 10 master data sets and repeated 25 times at each of the 25 noise levels for a total of 775,000 spectra. A brief overview of the MaxEnt method is given below; additional details are available elsewhere (Hoch et al. 1990; Donoho et al. 1990; Schmieder et al. 1993).

The MaxEnt reconstruction of the spectrum of a complex-valued time series $\boldsymbol{d}$ is the spectrum $\boldsymbol{f}$ which maximizes the entropy $S(\boldsymbol{f})$, subject to the constraint that the inverse discrete Fourier transform (IDFT) of the spectrum, is consistent with the time series $\boldsymbol{d}$. The MaxEnt solution is found by maximizing the objective function

$$O = S(\boldsymbol{f}) - \lambda C(\boldsymbol{f}, \boldsymbol{d}) \tag{2}$$

where $S(\boldsymbol{f})$ is the entropy of a complex spectrum and $C(\boldsymbol{f}, \boldsymbol{d})$ is the consistency condition.

The entropy component of Eq. (2) is defined as

$$S(\boldsymbol{f}) = -\sum_{n=0}^{N-1} \frac{|f_n|}{def} \log\left(\frac{1}{2}\left(\frac{|f_n|}{def} + \sqrt{4 + \frac{|f_n|^2}{def^2}}\right)\right) - \sqrt{4 + \frac{|f_n|^2}{def^2}} \tag{3}$$

The *def* term acts as an adjustable parameter which determines the scale at which the nonlinearity of MaxEnt becomes pronounced.

The consistency condition of Eq. (2) is defined by

$$C(\boldsymbol{f}, \boldsymbol{d}) \leq C_0 \tag{4}$$

where C_0 is an estimate of the noise level and $C(\boldsymbol{f}, \boldsymbol{d})$ is the unweighted chi-squared statistic

$$C(\boldsymbol{f}, \boldsymbol{d}) = \sum_{i=0}^{M_1} |IDFT(\boldsymbol{f})_i - d_i|^2 \tag{5}$$

The value of the Lagrange multiplier λ in Eq. (2) is adjusted to obtain $C = C_0$. This depends on the values of the data and the parameters *def* and C_0. Practical guidelines for choosing the values of *def* and C_0 are described elsewhere (Hoch and Stern 2001), and an automated procedure for determining their values has been described (Mobli et al. 2007a, b).

The equations above generalize to higher dimensional data. However, for larger data sets where it may not be desirable to perform MaxEnt reconstruction across the entire data set due to computational resource limitations, the reconstruction may be performed on individual rows (or planes or any other subset of the data set) with the following consideration. The extent of the nonlinearity of the reconstructions can vary from row to row, resulting in distorted peak shapes. One way to avoid this is to use a fixed value of λ rather than allowing the reconstruction of each row to produce its own value of λ (Schmieder et al. 1997). The universal value for λ can be determined by choosing a representative row, then computing the MaxEnt reconstruction on each row with an appropriate value for C_0.

MaxEnt reconstruction is nonlinear and must be the last time-to-frequency transformation applied. The nature of

the nonlinearity of MaxEnt reconstructions is fairly well understood. In general, MaxEnt reconstruction tends to scale intensities down, compared to the FT, with small amplitudes scaled down more than large amplitudes (Donoho et al. 1990). The extent of the nonlinearity can be diminished by increasing λ (i.e. placing more weight on the consistency condition).

The MaxEnt reconstructions performed on the synthetic data sets use a fixed value of $def = 0.1$. The last FID of each data set is used to evaluate the noise level and set the C_0 parameter. The MaxEnt reconstruction converges on a λ value and generates the spectrum. In the limit of large λ, the nonlinearity is minimized and MaxEnt behaves similarly to so-called "forward MaxEnt" (FM, Hyberts et al. 2007).

Peak picking

MaxEnt spectrum reconstruction usually introduces nonlinear scaling to the frequency domain, thereby complicating the use of the l_2 norm. As an alternative, the metrics employed here are based on peaks identified by a threshold-based peak picker, rather than an element-wise comparison to a reference spectrum. We use the following peak-picking algorithm:

1. Create an empty set ($\mathcal{R}$) to collect all recovered peak locations.
2. Isolate the real-real data quadrature component from the 2D hypercomplex spectrum.
3. Create a list of grid locations ($\mathcal{P}$) whose data values are in the top 5% of all positive data values. These locations are all considered to be in peaks. The remaining procedure partitions this set into individual peaks, each comprised of a peak tip and a surrounding neighborhood.
4. While $\mathcal{P}$ is not empty, repeat the following:
   (a) Select the grid location ($\boldsymbol{p}$) from $\mathcal{P}$ with the largest value; this location defines the tip of the current peak. Remove $\boldsymbol{p}$ from $\mathcal{P}$ and add $\boldsymbol{p}$ to $\mathcal{R}$.
   (b) Create a neighborhood set ($\mathcal{N}$) and initialize it to contain the peak tip ($\boldsymbol{p}$).
   (c) Create an edge set ($\mathcal{E}$) and initialize it to contain the peak tip ($\boldsymbol{p}$).
   (d) While the edge set is not empty, repeat the following procedure:
      i. Remove the first entry from the edge set, call it $\boldsymbol{e}$.
      ii. By moving in the $+1$ and -1 directions along each of the 2 grid dimensions, $\boldsymbol{e}$ has up to 4 neighbors. For each neighbor that is listed in $\mathcal{P}$ *and* has a data value no larger than that of $\boldsymbol{e}$, add the neighbor to the edge set ($\mathcal{E}$) and add it to the neighborhood set ($\mathcal{N}$).

At each completion of step 4d, the edge set ($\mathcal{E}$) is emptied and the neighborhood set ($\mathcal{N}$) is populated with all grid locations in $\mathcal{P}$ that form a non-increasing neighborhood around the current peak tip ($\boldsymbol{p}$). The peak tip and its neighborhood are recorded and the algorithm continues from step 4a until $\mathcal{P}$ is exhausted. The quantities $\boldsymbol{p}$, $\mathcal{N}$ and $\mathcal{E}$ are labeled on the 1D illustration in Supplemental Figure S1.

Analysis of peaks

The metrics defined in this section quantify the similarity between the master peak list ($\mathcal{M}$) and a recovered peak list ($\mathcal{R}$) from peak picking the spectrum of a processed synthetic data set. Prior to defining the resolution and sensitivity metrics, it is critical to note that, by design, the master peak list is *not* constrained to be on-grid, while the recovered peaks *are* inherently constrained to be on-grid. Rather than assigning a peak to the location of the local maximum value (as done presently), it is possible to compute an off-grid "center-of-mass" location based on the grid intensities of the entire peak neighborhood ($\mathcal{N}$). This approach is not taken here for two critical reasons. First, without a deconvolution based peak picker, asymmetric truncation will skew peak locations. Second, the nonlinear scaling introduced by MaxEnt may additionally skew overlapping peak locations (Kubat et al. 2007). Therefore, the metrics must be developed with the understanding that on-grid ($\mathcal{R}$) and off-grid ($\mathcal{M}$) locations are directly compared.

Metric M1: peak location recovery

Resolution is the ability to distinguish signal components that have similar frequencies. Accurate determination of the frequencies of partially overlapping resonances is a challenging problem and typically requires modeling or deconvolution. We previously showed that the accuracy with which the frequencies of isolated resonances with known frequencies can be recovered is an excellent proxy for resolution (Stern et al. 1982). In multidimensional NMR experiments where analysis requires correlation of resonances that share frequencies in a particular dimension, the accuracy of frequency recovery is a vital aspect of resolution.

We define a metric M1, which computes the "distance" in ppm values between the peaks in the master peak list $\mathcal{M}$ and the peaks in a given recovered peak list $\mathcal{R}$. These peak sets do

not have a correspondence mapping and, in fact, will likely have a different number of peaks. In order to compute the distance between two such sets, a variant of the symmetric Hausdorff metric (Huttenlocher et al. 1993) is employed.

The first component of M1 is given by the RMSD from pairing each peak in $\mathcal{M}$ with its closest peak in $\mathcal{R}$

$$d(\mathcal{M},\mathcal{R}) = \sqrt{\frac{1}{|\mathcal{M}|}\sum_{p_m\in\mathcal{M}} \min_{p_r\in\mathcal{R}}\{||p_m - p_r||\}^2} \tag{6}$$

The second component of M1 is given by the RMSD from pairing each peak in $\mathcal{R}$ with its closest peak in $\mathcal{M}$

$$d(\mathcal{R},\mathcal{M}) = \sqrt{\frac{1}{|\mathcal{R}|}\sum_{p_r\in\mathcal{R}} \min_{p_m\in\mathcal{M}}\{||p_r - p_m||\}^2} \tag{7}$$

These components are averaged to produce a single (and symmetric) measure of peak location recovery

$$\mathrm{M1}(\mathcal{M},\mathcal{R}) = \frac{1}{2}[d(\mathcal{M},\mathcal{R}) + d(\mathcal{R},\mathcal{M})] \tag{8}$$

For two sets, X and Y, the conventional max-min Hausdorff is based on identifying the point x in X whose nearest neighbor in Y is the farthest away. The value of this metric is thus a measurement between two points. The variant used to define M1 incorporates a root-mean-square component so that all pairs of points contribute to the metric.

The traditional Hausdorff metric is more sensitive to outliers and missing peaks than the M1 metric. An outlier or missing peak creates a situation where a peak from one set does not have a corresponding peak in the second set. The traditional Hausdorff metric is defined by the single largest outlier or missing peak, whereas the M1 metric is less sensitive and more accurately provides a global measure of the inter-set distance.

As noted previously, the peaks in $\mathcal{M}$ are not constrained on-grid, while the peak locations in $\mathcal{R}$ are on-grid. This introduces a lower bound limitation on M1, which is on the order of the distance between grid locations. Formally, for a grid with spacing of s_1 along dimension 1 and s_2 along dimension 2, the average error is given by the quantity

$$\bar{e}(s_1,s_2) = \frac{1}{s_1 s_2}\int_{x=-s_1/2}^{s_1/2}\int_{y=-s_2/2}^{s_2/2}\sqrt{x^2+y^2}\,dx\,dy \tag{9}$$

The quantities involved in Eq. (9), are illustrated in Supplemental Figure S2, and the closed form solution to the integral is

$$\begin{aligned}\bar{e}(s_1,s_2) = \frac{1}{24 s_1 s_2}\big[&4 s_1 s_2 s + s_1^3\log(s+s_2) - s_1^3\log(s-s_2)\\ &+ s_2^3\log(s+s_1) - s_2^3\log(s-s_1)\big]\end{aligned} \tag{10}$$

where $s = \sqrt{s_1^2 + s_2^2}$.

Metric M2: peak sensitivity

Sensitivity is the ability to distinguish signal from noise. We previously showed that the common measure of sensitivity, signal-to-noise ratio (S/N), is not a reliable indicator of sensitivity when employed with nonlinear methods of spectrum analysis (Donoho et al. 1990). Accordingly we define a set of metrics that quantifies the extent to which the set of recovered peaks ($\mathcal{R}$) overlaps the master peak set ($\mathcal{M}$).

The inherent on-grid ($\mathcal{R}$) / off-grid ($\mathcal{M}$) disparity is addressed by first rounding each peak location in $\mathcal{M}$ onto the nearest grid location. This potentially introduces an error of up to one-half a grid spacing along each dimension. This is accounted for by constructing a 3 × 3 section of grid centered on each rounded peak location from $\mathcal{M}$, thus assuring that the original off-grid peak location is completely surrounded by the expanded 3 × 3 peak region. The set of grid locations covered by this expanded master peak set is referred to as $\hat{\mathcal{M}}$. Peak sensitivity is addressed by considering the overlap of the $\hat{\mathcal{M}}$ and $\mathcal{R}$ peak sets. The quantities involved in this comparison are illustrated in the Venn diagram of frequency domain space in Supplemental Figure S3 and are defined as

$$TP = |\hat{\mathcal{M}} \cap \mathcal{R}| \tag{11}$$

$$FP = |\sim\hat{\mathcal{M}} \cap \mathcal{R}| \tag{12}$$

$$FN = |\hat{\mathcal{M}} \cap \sim\mathcal{R}| \tag{13}$$

$$TN = |\sim\hat{\mathcal{M}} \cap \sim\mathcal{R}| \tag{14}$$

The M2 metric values correspond to true positive (TP), false positive (FP), false negative (FN) and true negative (TN) and are defined using set notation: a tilde ($\sim$) preceding a set name indicates the selection of all possible values *not* in the set; the intersect symbol ($\cap$) is a binary operator and selects all set elements in common to the sets preceding and following the operator; the absolute value bars ($|\cdot|$) placed around a set evaluate the number of elements contained by the set. For example, the definition for FP counts the number of grid locations that are not in the expanded master peak set ($\hat{\mathcal{M}}$) and are in the recovered peak set ($\mathcal{R}$).

For a master peak list of a given size and for given indirect dimensions, the full set of M2 quantities is redundant. It is sufficient to only specify the TP and FP values. The other values are recovered as: $FN = |\hat{\mathcal{M}}| - TP$ and $TN = (\text{size of dim 1})\cdot(\text{size of dim 2}) - FN - TP - FP$.

Analysis of NUS schedules

The M1 and M2 metric analyses are designed to evaluate resolution and sensitivity relative to a master peak list. That approach is the primary tool employed in the characterization of NUS/MaxEnt reconstruction. In a secondary analysis, we use mean evolution times and theoretical relative sensitivities (defined below) to characterize sample schedule properties.

Mean evolution time

For a time domain with n_1 steps along dimension 1 and n_2 steps along dimension 2, a sample schedule, K, is represented as an $n_1 \times n_2$ matrix where a value of 1 indicates the collection of the corresponding sample point and a value of 0 is given otherwise. The mean evolution time along dimension 1, is defined as

$$m_1(K) = \frac{\sum_{i=1}^{n_1}\left(i \times \sum_{j=1}^{n_2} K_{i,j}\right)}{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{i,j}} \tag{15}$$

A similar equation gives the mean evolution time along dimension 2.

Relative sensitivity

Rearranging Eq. (1), produces

$$p = \exp\left(\frac{-R2(m-1)}{SW}\right) \tag{16}$$

which defines the magnitude of the time-domain signal envelope as a function of sample point location (m), given decay rate ($R2$) and spectral width (SW). This function is generalized for the 2-dimensional synthetic data as

$$p(i,j) = \exp\left\{-\left(\frac{R2_1(i-1)}{SW_1} + \frac{R2_2(j-1)}{SW_2}\right)\right\} \tag{17}$$

where $p(i, j)$ is the signal envelope magnitude at the grid point located i steps along dimension 1 and j steps along dimension 2. The relative sensitivity of a given sample schedule is thus determined by computing the percentage of the signal envelope that it captures

$$r(K) = \frac{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} K_{i,j}p(i,j)}{\sum_{i=1}^{d_1}\sum_{j=1}^{d_2} p(i,j)} \tag{18}$$

It should be noted that the M2 sensitivity metric is different than the relative sensitivity defined above. Relative sensitivity assumes perfect exponential decay of signal and quantifies how much of the intensity is captured by a given sample schedule. The M2 metric quantifies how well the peaks of a given spectrum are detected, while subject to peak overlap and other artifacts of the reconstruction process.

Results and discussion

The M1 and M2 metrics are first applied to the peaks recovered in the frequency domain to reveal trends in NUS/MaxEnt processing performance. Following these results, attention is shifted to the NUS schedules in an effort to correlate the resolution and sensitivity results with underlying properties of the NUS schedules. Real experimental data for PfPMT is considered in the final section to illustrate the performance gains achievable through optimizing knowledge-based NUS.

Analysis of peaks

Each of the 10 master data sets are processed by the conventional FT, which employs the complete, uniformly sampled data set, and by MaxEnt paired with each of the 3 NUS approaches (EMS, BMS, RS). For each combination of processing parameters (NUS approach / coverage / sampling density decay rate, if applicable), the M1 and M2 metrics are computed across all data sets and the values are averaged across the 10 master data sets and over the 25 repetitions made at each noise level. Under all spectral processing techniques, the M1 and M2 response profiles reach steady rates of degradation beyond an RMSD noise value of $\sim$40. This indicates that the injected noise is sufficient to overwhelm the processing techniques. This important observation demonstrates that the noise regime covered in the simulated processing is sufficient to test the limits of NUS.

The data from FT processing and NUS-based methods are shown on the same plots. Sample schedule coverage labels are only applicable to the NUS-based methods. The FT approach utilizes the complete time domain data.

Metric M1: peak location recovery

The 3 NUS approaches (EMS, BMS, RS) and the FT approach are compared over a range of noise intensities. A plot is constructed for each sample coverage value and the EMS decay rate scaling factor is set to 1×; the series is shown in Fig. 4. Smaller M1 values indicate better peak resolution.

There are four components that contribute to the M1 values in Fig. 4. The first two are peak picker error and grid rounding error, both of which are universally applicable to all methods. The third component is the lower bound performance of a given processing method (i.e. without noise, M1 does not reach a zero value). The fourth

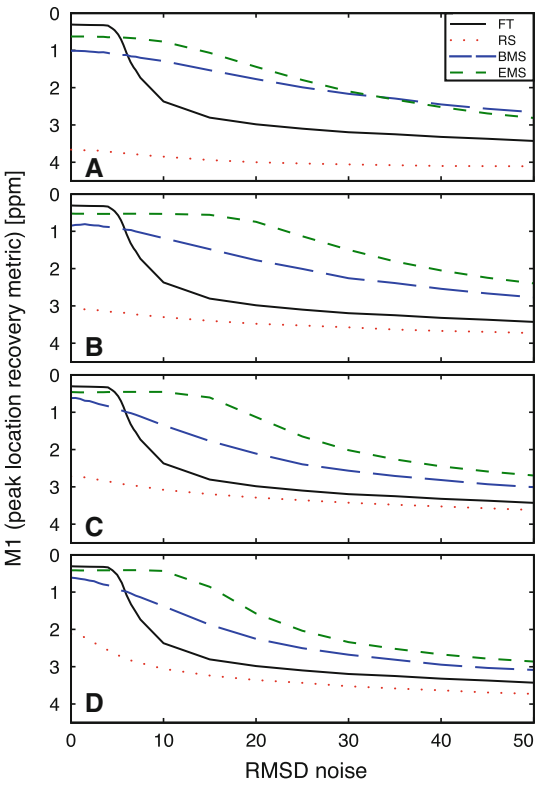

Fig. 4 Metric M1: Method comparison. The FT approach utilizes all time domain data samples. The NUS approaches subsample the time domain data at sample coverage values of 3% (*A*), 6% (*B*), 12% (*C*) and 24% (*D*). All methods are applied across a range of injected noise intensities. The M1 metric is computed and shows each method's ability to recover peak location as sample coverage and noise vary. The FT performs the best at the lowest noise values, but sharply degrades in performance, while the NUS approaches, especially EMS, maintain very high performance levels over an extended noise regime

component captures a method's response to signal noise. These contributions can be inferred from the analysis of the plots in Fig. 4, and are discussed below.

The first component—grid error, as defined in Eq. (10)—is solely a function of the frequency domain grid spacing. The synthetic data sets have grid spacing values of $s_1 = 0.3918$ and $s_2 = 0.4996$ ppm, which yield a grid rounding error of $\bar{e} = 0.1713$ ppm.

The second component—peak picker error—involves the accuracy in the assignment of peak locations. The simple approach employed in this study assigns peak locations to the locally maximal data points. The error associated with this approach is not additive to the grid error, but rather masked by it and thus undetectable. In addition, the peak picker error is uniformly distributed and is independent of processing technique or noise level. This factor does not impact the trends observed in M1 performance.

The third component—method performance without noise—is assessed by considering the zero noise level in the plots of Fig. 4. Once the contributions from grid error are accounted for, the M1 metric still does not achieve a zero value. For example, the FT approach has a no noise M1 value of 0.3062 ppm. The grid error accounts for roughly 55% of this, which leaves ~45% of the no noise M1 value as a result of the FT procedure. The no noise M1 values for the NUS/MaxEnt approaches are higher, but the EMS and BMS methods quickly approach FT performance as sample coverage increases. The RS method also improves with increasing sample coverage, but never attains the high performance observed in the other approaches.

The fourth, and most significant, component contributing to the M1 value is the response of each method to increasing signal noise. The previous three components are noise independent, so this component is readily observed in the shape of each M1 plot. The most striking feature of the plots in Fig. 4, is the initial flat response of the FT method to noise, followed by an abrupt falloff. The threshold for this behavior is at an RMSD noise level of ~4.0. The NUS/MaxEnt methods also show an initial flat response to noise, but then, in contrast to the FT results, show a continuous response to increasing signal noise. This difference is a fundamental feature of MaxEnt reconstruction.

The above discussion of peak resolution focusses on the comparison between FT and NUS/MaxEnt performance. The RS and BMS approaches to NUS illustrate MaxEnt performance features, but are limited in their applicability: BMS requires prior knowledge of peak frequencies and RS results in very poor performance for decaying signals. The derivation by Ernst showing that the matched filter window function optimizes S/N is based on linear (FT) processing. It is not clear whether matching sampling density decay to the signal envelope similarly yields optimal sensitivity, and so we also investigated sampling density decay rates above and below the signal envelope decay rate. In practice, the decay rate of the peaks is not (precisely) known in advance and will vary. The range of scaling factors covers an order of magnitude, which is sufficient to reveal performance trends. The series of plots in Fig. 5 are based on EMS schedules whose decay rates have been scaled relative to the decay rate of the injected peaks. As before, the analysis also spans a series of sample coverage values applied over a range of injected noise values. In the lowest noise regime, lower decay rate scaling factors (i.e. sample schedules with higher mean evolution times) perform better than higher decay rate scaling factors, as expected, but only marginally. At noise levels above 7.5 RMSD, higher decay rate scaling factors actually perform significantly better than lower decay rate scaling factors. The plateau region of optimal performance is extended over a wider noise regime.

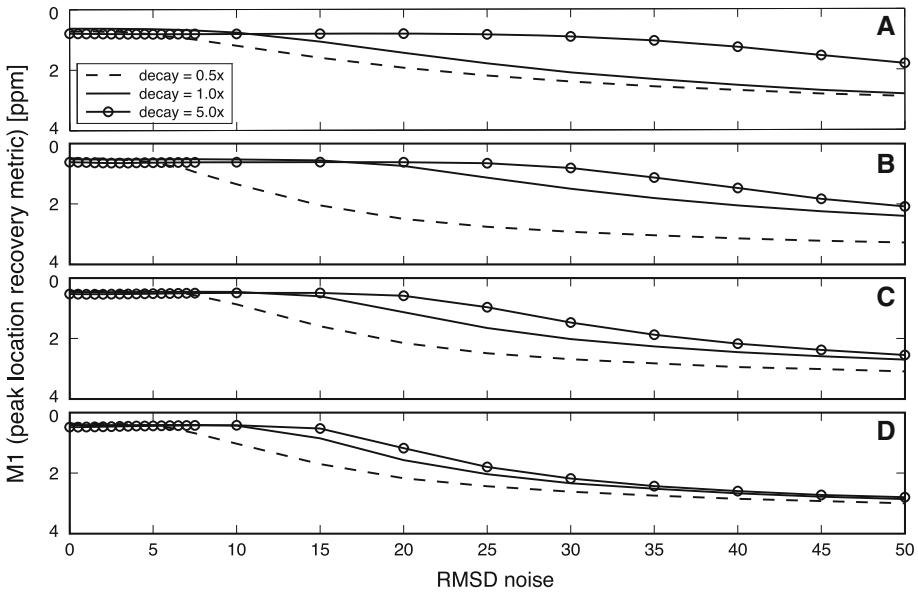

Fig. 5 Metric M1: EMS with decay rate scaling. This series of plots follows the same axes as Fig. 4. The series isolates the EMS approach and demonstrates how scaling the decay rate parameter impacts the M1 performance across the sample coverage values of 3% (*A*), 6% (*B*), 12% (*C*) and 24% (*D*). At noise levels below ~7.5 RMSD, lower decay rate scaling performs marginally better than higher decay rate scaling. At higher noise levels, increased decay rate scaling significantly enhances performance. This trend is seen across all sample coverage values

Metric M2: peak sensitivity

As for the M1 analysis, the M2 values for the three NUS approaches (EMS, BMS, RS) and the FT approach are compared over a range of noise intensities. A plot is constructed for each sample coverage; the series of plots for the true positive (TP) metric is shown in Fig. 6, and the series for the false positive (FP) metric is shown in Fig. 7. The plots are oriented with better results (i.e. more TP/less FP) towards the top of the vertical axes. The true negative (TN) and false negative (FN) quantities are not shown, as they may be derived from the TP and FP values (given the known size of $\hat{\mathcal{M}}$).

First, consider the TP plots (Fig. 6). The FT approach achieves the highest value, but quickly falls off as the noise increases. The RS approach is terrible (unsurprisingly) and modestly improves with increased sample coverage. The BMS approach does not achieve the same level of performance as the FT at zero noise, but it does have a broad plateau of maximal performance before it slowly degrades. The EMS approach shows the most interesting profile—performance actually *improves* as noise *increases*, before reaching a maximum and gradually falling off. The profile maximum is sample coverage dependent and shifts to higher noise values as coverage increases from 3 to 6%. Further increasing the sample coverage, shifts the maximum performance to the lower noise regime. The counterintuitive response of the EMS approach is in part due to peaks whose intensities are just below the peak picking threshold being "pushed" over the threshold by the superposition of positive valued noise. This effect may also be responsible for the low noise regime plateau response from the other NUS-based approaches.

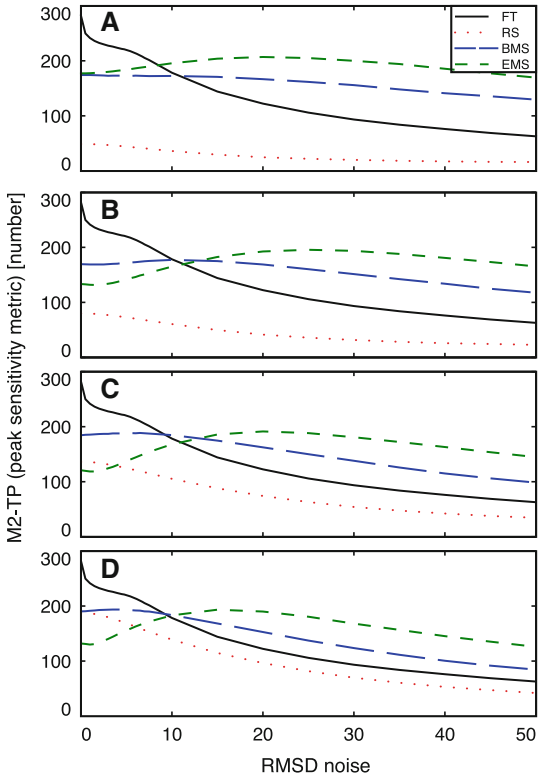

Fig. 6 Metric M2-TP: Method comparison. The FT approach utilizes all time domain data samples. The NUS approaches subsample the time domain data at sample coverage values of 3% (*A*), 6% (*B*), 12% (*C*) and 24% (*D*). All methods are applied across a range of injected noise intensities. The M2-TP metric is computed and shows each method's ability to recover peaks as sample coverage and noise vary. The FT method is the clear winner at the lowest noise levels, but falls off quickly. Beyond the threshold value of noise ~10 RMSD, the EMS method outperforms all others

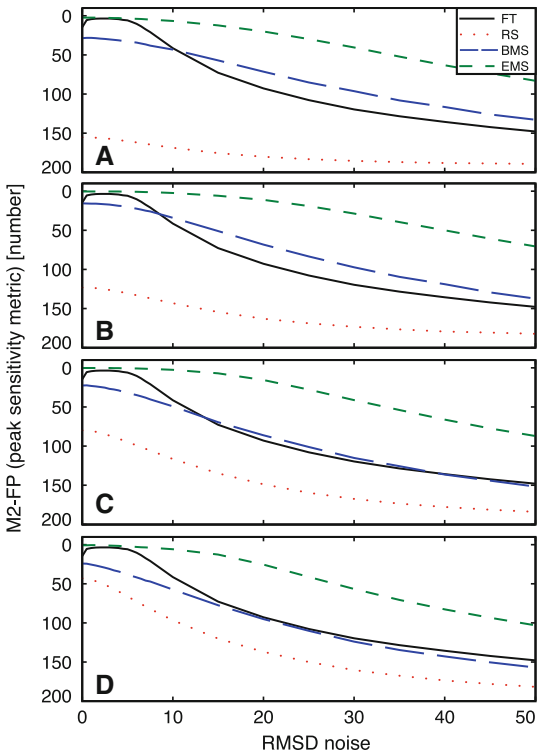

Fig. 7 Metric M2-FP: Method comparison. The FT approach utilizes all time domain data samples. The NUS approaches subsample the time domain data at sample coverage values of 3% (*A*), 6% (*B*), 12% (*C*) and 24% (*D*). All methods are applied across a range of injected noise intensities. The M2-FP metric is computed and shows how many peaks are incorrectly identified, as sample coverage and noise vary. The EMS approach outperforms all other methods, including FT, at all sample coverage values and over all noise levels

Now consider the FP plots (Fig. 7). The FT and all NUS methods perform very well up to an RMSD noise level of ~7.5. Beyond this threshold, the FT results degrade the fastest, while the EMS approach slowly degrades and outperforms all methods.

As with the M1 analysis, we now focus on the impact of scaling the sample schedule decay rate in the EMS approach. The series of plots showing the M2-TP response is in Fig. 8, and the series of plots showing the M2-FP response is in Fig. 9. The M2-TP plots show that at all sample coverage values, decreasing (increasing) the decay rate scaling factor improves (degrades) the M2-TP values at RMSD noise levels below ~20 and degrades (improves) the M2-TP performance above the RMSD noise level of ~20. The noise level transition point moves slightly to lower values as sample coverage increases. The M2-FP plots show that increasing the decay rate, significantly improves the method at all sample coverage values. False positives are nearly eliminated up to an RMSD noise level of ~7.5, above which a slow degradation in performance occurs.

Analysis of NUS schedules

The M1 and M2 metric analyses reveal resolution (M1) and sensitivity (M2) performance trends as observed in the reconstructed spectra of the synthetic data. These results characterize the response of MaxEnt reconstruction to variations in sample coverage, signal noise and sample schedule parameters. In a real experimental application, there is no master peak least and thus no way of judging the quality of the recovered spectra relative to some control. Expected performance can, however, be determined in advance through the analysis of the sample schedule. Mean evolution time (defined in Eq. 15) and relative sensitivity (defined in Eq. 18) are employed.

Mean evolution time

The mean evolution times are shown in Supplemental Figure S4, for the NUS schedules generated by RS, BMS and EMS (at decay rate scaling of 1×). Three features are evident in this plot. First, the values all fall along the diagonal (dashed line). This is as expected for the RS and EMS approaches which, by definition, treat the sample dimensions equally. It is interesting to note that the BMS schedules also fall along the diagonal. This indicates that the time domain data, when sorted from largest magnitude to smallest magnitude, is distributed equally along each of the dimensions. Second, all RS sample schedules (red symbols), regardless of sample coverage, fall around T1 = 32/T2 = 64. This location is the center of the sample grid and indicates that the random distribution is correctly positioned. Third, BMS and EMS schedules are both biased towards shorter evolution times. Therefore, as the sample coverage increases, the schedules are forced to expand by selecting sample points at increasing evolution times.

Supplemental Figure S5, shows the mean evolution for the EMS schedules at each sample coverage value, over the range of decay rate scaling factors. Each of the sample coverage subplots shows that increasing the decay rate scaling factor, decreases the mean evolution times. The tight clustering of like symbols indicates the strength of this relationship.

Supplemental Figures S4 and S5, show that all sample schedules follow the T1/T2 diagonal line in mean evolution time. It is therefore appropriate, and more convenient, to introduce a single combined metric, by taking the average of the normalized mean evolution times

$$m(K) = \frac{1}{2}\left(\frac{m_1(K)}{n_1} + \frac{m_2(K)}{n_2}\right) \tag{19}$$

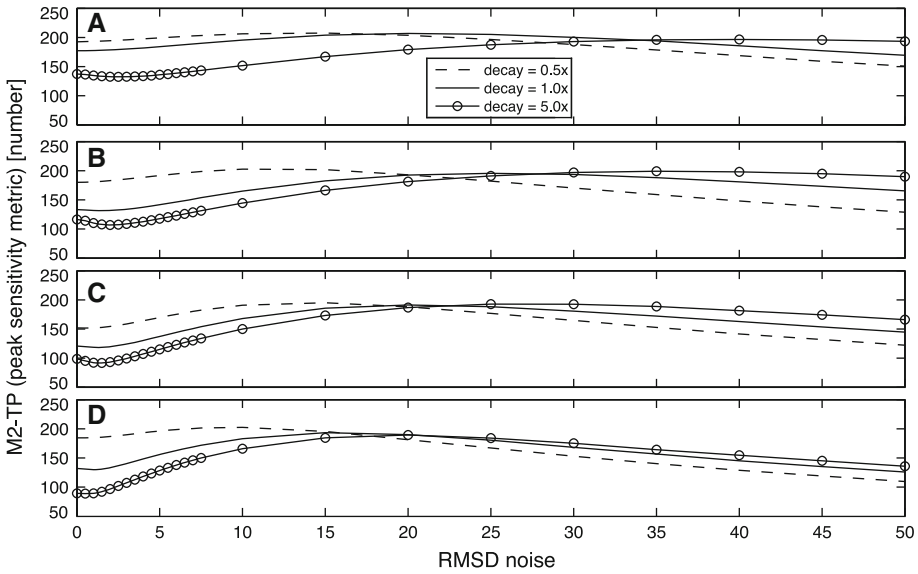

Fig. 8 Metric M2-TP: EMS with decay rate scaling. This series of plots follows the same axes as Fig. 6. The series isolates the EMS approach across the sample coverage values of 3% (*A*), 6% (*B*), 12% (*C*) and 24% (*D*). *Each panel* shows the effect the decay rate scaling parameter has on M2-TP performance. At each sample coverage value there is a threshold noise level ~20 RMSD, below (above) which, decreasing (increasing) the decay scaling results in optimal performance

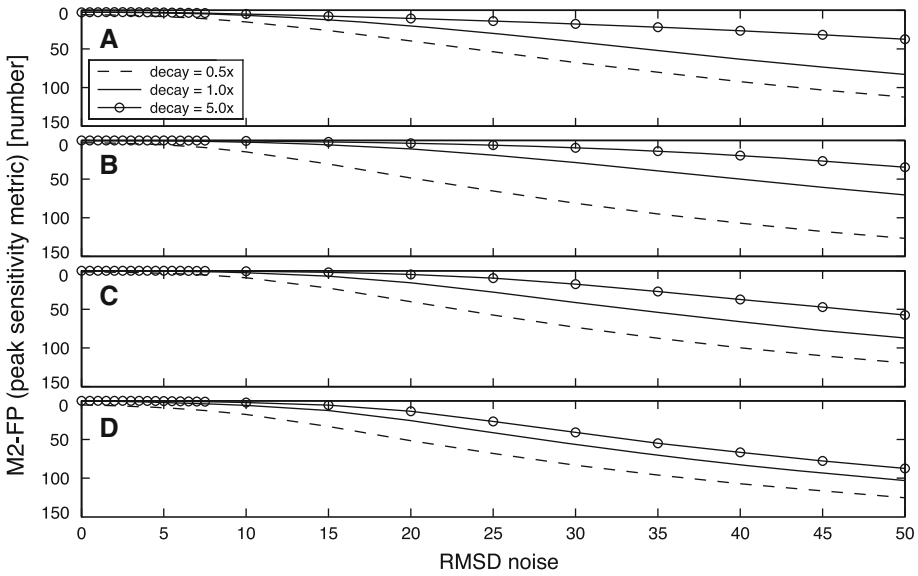

Fig. 9 Metric M2-FP: EMS with decay rate scaling. This series of plots follows the same axes as Fig. 7. The series isolates the EMS approach across the sample coverage values of 3% (*A*), 6% (*B*), 12% (*C*) and 24% (*D*). *Each panel* shows the effect the decay rate scaling parameter has on M2-FP performance. Increasing the decay scaling improves performance at all sample coverage values and over all noise levels. In addition, performance degrades with increasing sample coverage

Relative sensitivity

There are two ways to improve relative sensitivity, given the relationship in Eq. (18): including more sample points and including sample points of higher signal intensity. The first is achieved through higher sample coverage and the second is achieved through biasing the sample points to lower evolution times. These relationships are observed in Fig. 10, which shows the relative sensitivity ($r(K)$) as a function of the average of the normalized mean evolution times ($m(K)$). Two sets of lines are superimposed (by visual inspection) over the data to demonstrate various correlations. The first set of lines (negative slope and shown in black) run along data points of the same sample coverage; a dashed segment is used to connect the knowledge based methods (BMS and EMS) to the RS data points. The second set of lines (positive slope and shown in color) run along data points generated for each reconstruction method.

First, consider the black lines, which mark the linear correlation between decreasing mean evolution times and increasing relative sensitivity. The strength of this correlation (i.e. the slope of the lines) is sample coverage dependent—stronger relationships are observed at higher sample coverage values. At each sample coverage, the EMS schedules (green symbols) have the lowest mean evolution times and show the highest relative sensitivity, followed by the BMS schedule (blue symbol). The data from the RS schedules do not fall along the direction of the

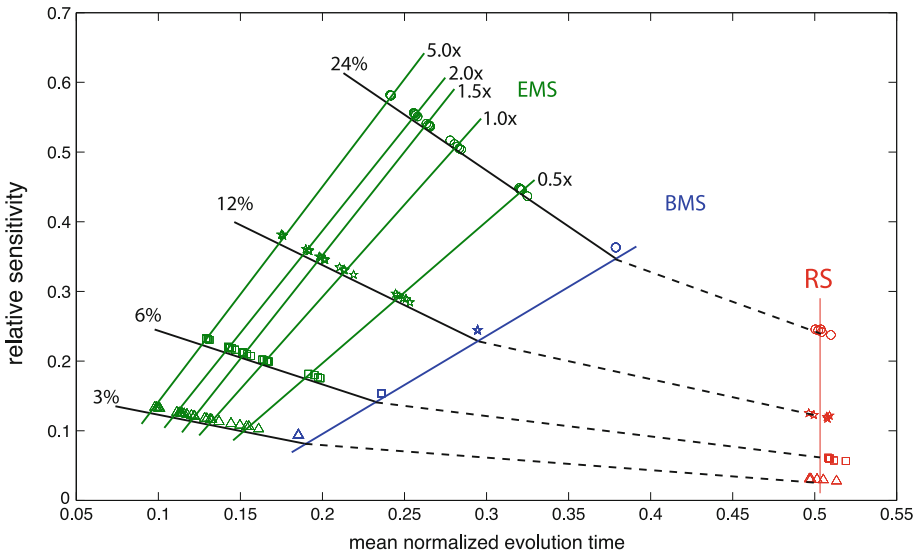

Fig. 10 Schedule relative sensitivity. A plot showing the relative sensitivity (Eq. 18) as a function of mean normalized evolution time (Eq. 15) for all NUS sample schedules. All data points are color coded by sampling method (RS *red*, BMS *blue*, EMS *green*), with *lines of matching color* running along the data points of each method (*separate lines* are included for each EMS decay rate scaling factor). *Symbol shapes* indicate sample coverage (3% *triangle*, 6% *square*, 12% *star* and 24% *circle*). The EMS and BMS data points at each sample coverage value are collinear and marked by *black lines*. The RS data points are not collinear and are thus connected by *dashed lines* to the *black lines* of corresponding sample coverage values. The *black* and *colored lines* form a distorted grid over the plot space. The varying slopes of these *lines* indicate the strength of correlations between relative sensitivity and mean normalized evolution time. Note: the superimposed lines are placed by visual inspection to allow maximum visibility of the data points

EMS/BMS lines. They have significantly higher mean evolution times and significantly lower relative sensitivities at each sample coverage value.

Now consider the colored lines, which mark the extent to which increasing sample coverage contributes to enhanced resolution and/or sensitivity. Each line corresponds to a specific NUS approach; there is one line for each scaling factor employed in EMS. The RS line is vertical and the symbols along the line are linearly spaced according to sample coverage. This indicates that additional sampling in RS is not able to improve resolution, but will linearly improve relative sensitivity. This is as expected because the RS sample points are chosen randomly and each point, on average, contributes the same time domain signal intensity, but does not alter the mean evolution time. In contrast the BMS and EMS schedules bias the selection of their sample points to capture more points from the low evolution times. The EMS lines (green) and BMS line (blue) have positive slopes that indicate resolution and sensitivity improvements are made with increased sample coverage. The EMS lines are steeper, which indicates that for a given sample coverage increase it is expected that EMS will show a larger relative gain in sensitivity, whereas BMS will show a larger relative gain in resolution.

Example application: PfPMT

Conventional, uniformly sampled data was collected in a 3D HCCH-TOCSY on a 600 MHz spectrometer for PfPMT, the phosphoethanolamine methyltransferase of the human malaria parasite Plasmodium falciparum (Ben Mamoun et al. 2010; Reynolds et al. 2008). The first indirect dimension is ^{1}H and spans 75 data points with a spectral width of 6720.0 Hz, a carrier frequency of 599.1 MHz and a 0 Hz reference value of 4.772 ppm. The second indirect dimension is ^{13}C and spans 90 data points with a spectral width of 9791.9 Hz, a carrier frequency of 150.7 MHz and a 0 Hz reference value of 42.50 ppm. The direct dimension is ^{1}H and spans 1024 data points with a spectral width of 9058.0 Hz, a carrier frequency of 599.1 MHz and a 0 Hz reference value of 4.772 ppm.

In the synthetic data analysis, the EMS schedules are produced with envelope decay rates scaled relative to the uniform (and predefined) decay rates of the injected peaks. In the PfPMT data, the peak decay rates vary and are known only approximately in advance. It is possible to estimate the decay rate based on molecular size (Maciejewski et al. 2000; Cavanagh et al. 2007) or via a scout experiment. In the present analysis, the full time domain data set is utilized to ensure the most accurate decay rate

and demonstrate the effects of decay rate scaling. There are 4 planes which contain artifact peaks. The spectral power in each of these planes is several orders of magnitude larger than the mean spectral power seen across all planes. There are 53 planes clustered around F3 = 4.7 ppm, with each showing a spectral power several orders of magnitude lower than the mean—these result from the solvent suppression algorithm. These 57 planes are excluded from the following R2 recovery procedure. The time domain data is averaged across T1 and fit with an exponential decay, revealing a decay rate of R2 = 52.8 Hz. The same procedure is applied across T2 and produces a decay rate of R2 = 78.4 Hz.

The EMS schedules are generated across the following parameters: 4 sample coverage values (3, 6, 12, 24%) and 5 decay rate scaling factors (0.5×, 1.0×, 1.5×, 2.0×, 5.0×). The 20 combinations are each repeated 5 times for an EMS set of 100 schedules. MaxEnt reconstruction is performed with each schedule and a control FT is also computed. The processing script is available in the Supplemental Data online.

Figure 11, shows a representative set of spectra computed with MaxEnt employing EMS sample schedules at 3 and 24% coverage with envelope decay rates at 0.5×/1.0×/5.0×. These NUS-based spectra are shown along with the spectrum computed by the traditional FT, employing the complete, uniformly sampled data set. The F1–F2 planes are taken at F3 = 2.749 ppm. The superimposed 1D slices are taken at F1 = 2.7727 ppm. This slice runs through the center of a peak at F2 = 38.8559 ppm. Of particular interest is the adjacent, lower intensity, double peak at F2 = 41/42 ppm. A careful analysis of this subregion reveals the impact that sample coverage and decay rate scaling have on spectral quality.

At each of the decay rate scaling factors, increasing the EMS coverage rate (i.e. moving across the rows of the top 6 panels) produces sharper peaks and a greater level of detail, especially around the double peak at F2 = 41/42 ppm. Panels B and E correspond to EMS schedules whose envelope decay rates are matched to the decay rate of the average peak. At 3% coverage (panel B) the double peak appears as a single broadened peak that overlaps the neighboring peak at F2 = 39 ppm, whereas at 24% coverage (panel E), the double peak is clearly resolved and independent from the neighboring peak.

At each sample coverage rate, increasing the decay rate scaling factor (i.e. moving down the columns of the top 6 panels) reveals the proper peak shapes. In addition, the relative peak intensities are greatly improved. At 24% coverage and 0.5× decay rate scaling (panel D) the components of the double peak are resolved, but their relative intensities are skewed. In addition, the base of the large neighboring peak at F2 = 38.8559 ppm shows a small peak on each of its sides.

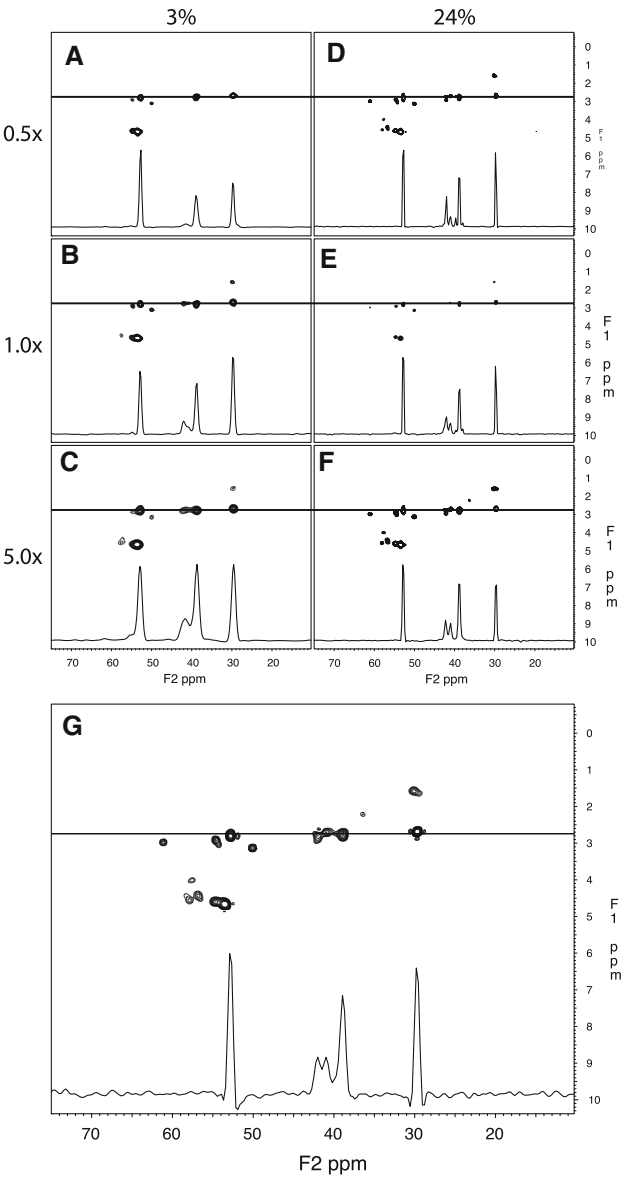

Fig. 11 EMS versus FT application to PfPMT. All panels are F1–F2 planes taken at F3 = 2.749 ppm. The *top six panels* are spectra computed with EMS/MaxEnt. Spectra in the *left column* (*A,B,C*) are based on schedules with 3% coverage and spectra in the *right column* (*D,E,F*) are based on schedules with 24% coverage. Spectra in the first row (*A,D*) are based on schedules with envelope decay rates scaled at 0.5×, the second row (*B,E*) is at 1.0× and the third row (*C,F*) is at 5.0×. The bottom panel (*G*) is the conventional FT performed with the complete, uniformly sampled data set. The horizontal lines run through F1 = 2.7727 ppm, which is the center of the peak located at F2 = 38.8559. Contour plots use the same levels across all spectra. The 1-D cross sections shown in the bottom of each panel are taken along the *horizontal lines* and are individually scaled so that the largest intensity peak appears at a uniform height. *Panel F* demonstrates a 4× reduction in measuring time with no loss of resolution or sensitivity, using a sampling density decay rate five times greater than the signal envelope decay rate

Increasing the decay rate scaling to 5.0× (panel F) fixes both artifacts and produces a spectrum functionally equivalent to that produced by the FT (panel G).

Concluding remarks

The traditional, uniformly sampled, FT approach to NMR spectrum computations is appealing because it is well understood. This is primarily due to the fact that the FT is a linear operation and there is a large body of supporting theory. In contrast, the nonlinearity of MaxEnt, while critical to achieving high spectral quality at low sample coverage, obscures the fundamental response characteristics of the method. In addition, the performance of any spectrum reconstruction method is heavily dependent on NUS schedule selection. The interdependence of the response of the method used for spectrum analysis and the sampling schedule further complicates optimization. In the present work, we address these issues by constructing a synthetic data set and utilizing two metrics to evaluate the performance of MaxEnt reconstruction paired with several NUS approaches. We are primarily interested in exploring the question of whether prior knowledge of chemical shifts can enable the design of optimal sampling schedules.

While we find that prior knowledge of chemical shifts (i.e. BMS) can improve the efficiency of NUS in some noise regimes, in general we find that use of only the knowledge of the signal decay rates (i.e. EMS), and not chemical shifts, leads to more robust sampling schemes. This finding is partially explained by comparing the point spread functions (PSFs) (Maciejewski et al. 2011). The central component of the EMS PSF is more focussed and shows very low intensity side lobes compared to the BMS PSF. This translates to reconstructed spectra whose peaks may be more accurately recovered and spectra with enhanced sensitivity.

Spectral analysis almost invariably displays a tradeoff between resolution and sensitivity. NUS schedules biased towards lower mean evolution times enhance sensitivity and schedules biased towards higher mean evolution times enhance resolution. The EMS analysis surprisingly reveals that sampling density decay rate can be controlled to optimize both sensitivity *and* resolution; a finding that is confirmed in recent work by Rovnyak et al. (2011). In low noise regimes, slower sampling density decay rates (i.e. bias towards higher evolution times) are best. At higher noise levels, faster sampling density decay rates (i.e. bias towards lower evolution times) is best. This optimization approach displays the greatest improvement at low sample coverage, thereby making demanding multidimensional NMR studies more accessible.

These observations derived from analysis of synthetic data are mirrored in the application to experimental 3D HCCH-TOCSY data. In particular, at low sampling coverage, the faster sampling density decay rate aids in the detection of a "hidden" double peak. At higher sampling coverage, the faster sampling density decay rate refines the shape of the double peak and suppresses adjacent artifacts. Overall, the EMS approach with sampling density decay faster than the envelope decay rate produces excellent results and demonstrates that NUS approaches to multidimensional NMR can be both robust and vastly more efficient than conventional uniform sampling.

In order to produce an optimized NUS/MaxEnt approach prior to running an experiment, the trends observed in the M1 and M2 metrics must be converted to optimization rules for NUS sample schedule construction. As a first step towards this ultimate goal, we consider the relationships between mean evolution times and relative sensitivity. Strong correlations are apparent, but are insufficient to fully capture the more complex trends observed in M1 and M2 performance. This indicates that a more sensitive a priori metric for sampling schedule performance is necessary, ideally one that is able to capture the effects of the pseudo-random component of a sample schedule. Efforts to develop a unified theory for optimal, a priori sampling schedule construction which afford deterministic performance will require this metric.

Acknowledgments We thank Drs. Alan Stern and Mehdi Mobli for useful discussions. We thank Dr. Choukri Ben Mamoun for providing the PfPMT construct used to prepare the sample. Financial support from the US National Institutes of Health (GM047467 and RR020125) is gratefully acknowledged.

References

Barna J, Laue E, Mayger M, Skilling J, Worrall S (1986) Reconstruction of phase-sensitive two-dimensional nuclear-magneticresonance spectra using maximum entropy. Biochem Soc Trans 14:1262–1263

Barna J, Laue E, Mayger M, Skilling J, Worrall S (1987) Exponential sampling, an alternative method for sampling in two-dimensional NMR experiments. J Magn Resonance 73:69–77

Ben Mamoun C, Prigge ST, Vial H (2010) Targeting the lipid metabolic pathways for the treatment of malaria. Drug Dev Res 71:44–55

Bodenhausen G, Ernst R (1981) The accordion experiment, a simple approach to three-dimensional NMR spectroscopy. J Magn Resonance 45:367–373

Cavanagh J, Fairbrother WJ, Palmer AG III, Skelton NJ (2007) Protein NMR spectroscopy: principles and practice. Academic Press, London

Donoho DL, Johnstone IM, Stern AS, Hoch JC (1990) Does the maximum entropy method improve sensitivity?. Proc Natl Acad Sci USA 87:5066–5068

Eghbalnia HR, Bahrami A, Tonelli M, Hallenga K, Markley JL (2005) High-resolution iterative frequency identification for NMR as a general strategy for multidimensional data collection. J Am Chem Soc 127:12528–12536

Ernst RR (1966) Advances in magnetic resonance, vol 2, Chapter Sensitivity Enhancement in Magnetic Resonance. Academic Press, London, pp 1–135

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). Proc Natl Acad Sci 102:10876

Hoch JC, Stern AS (1996) NMR data processing. Wiley, New York

Hoch JC, Stern AS (2001) Maximum entropy reconstruction, spectrum analysis and deconvolution in multidimensional nuclear magnetic resonance. Methods Enzymol 338:159–178

Hoch JC, Stern AS, Donoho DL, Johnstone IM (1990) Maximum entropy reconstruction of complex (phase-sensitive) spectra. J Magn Resonance 86:236–246

Huttenlocher D, Klanderman G, Rucklidge W (1993) Comparing images using the hausdorff distance. IEEE Trans Pattern Anal Mach Intell, pp 850–863

Hyberts SG, Heffron GJ, Tarragona NG, Solanky K, Edmonds KA et al (2007) Ultrahigh-resolution (1)h–(13)c hsqc spectra of metabolite mixtures using nonlinear sampling and forward maximum entropy reconstruction. J Am Chem Soc 129:5108–5116

Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125:1385–1393

Kubat JA, Chou JJ, Rovnyak D (2007) Nonuniform sampling and maximum entropy reconstruction applied to the accurate measurement of residual dipolar couplings. J Magn Reson 186:201–211

Kupce E, Freeman R (2003) Projection?reconstruction of three-dimensional NMR spectra. J Am Chem Soc 125:13958–13959

Maciejewski MW. Scheduletool. http://sbtools.uchc.edu/nmr/sample_scheduler

Maciejewski MW, Liu D, Prasad R, Wilson SH, Mullen GP (2000) Backbone dynamics and refined solution structure of the N-terminal domain of DNA polymerase beta. Correlation with DNA binding and dRP lyase activity. J Mol Biol 296:229–253

Maciejewski MW, Mobli M, Schuyler AD, Stern AS, Hoch JC (2011) Data sampling in multidimensional NMR: fundamentals and strategies. In: Orekhov V, Billeter M (eds) Novel sampling approaches in higher dimensional NMR. Springer (in press)

Mobli M, Stern AS, Hoch JC (2006) Spectral reconstruction methods in fast NMR: reduced dimensionality, random sampling and maximum entropy. J Magn Reson 182:96–105

Mobli M, Maciejewski MW, Gryk MR, Hoch JC (2007) Automatic maximum entropy spectral reconstruction in NMR. J Biomol NMR 39:133–139

Mobli M, Maciejewski MW, Gryk MR, Hoch JC (2007) An automated tool for maximum entropy reconstruction of biomolecular NMR spectra. Nat Methods 4:467–468

North D (1943) An analysis of the factors which determine signal/noise discrimination in pulsed carrier systems. RCA Labs Technical Report PTR6C

Reynolds JM, Takebe S, Choi JY, El Bissati K, Witola WH et al (2008) Biochemical and genetic analysis of the phosphoethanolamine methyltransferase of the human malaria parasite plasmodium falciparum. J Biol Chem 283:7894–7900

Rovnyak D, Sarcone M, Jiang Z (2011) Sensitivity enhancement for maximally resolved two dimensional NMR by nonuniform sampling. Magn Reson Chem (in press)

Schmieder P, Stern AS, Wagner G, Hoch JC (1993) Application of nonlinear sampling schemes to cosy-type spectra. J Biomol NMR 3:569–576

Schmieder P, Stern AS, Wagner G, Hoch JC (1997) Quantification of maximum-entropy spectrum reconstructions. J Magn Reson 125: 332–339

Stern AS, Li KB, Hoch JC (2002) Modern spectrum analysis in multidimensional NMR spectroscopy: comparison of linear-prediction extrapolation and maximum-entropy reconstruction. J Am Chem Soc 124:1982–1993

Van Vleck J, Middleton D (1946) A theoretical comparison of the visual, aural, and meter reception of pulsed signals in the presence of noise. J Appl Phys 17:940–971