J|A|C|S
ARTICLES
Published on Web 03/28/2007

Ultrahigh-Resolution $^{1}H-^{13}C$ HSQC Spectra of Metabolite Mixtures Using Nonlinear Sampling and Forward Maximum Entropy Reconstruction

Sven G. Hyberts,[†] Gregory J. Heffron,[†] Nestor G. Tarragona,[‡,§] Kirty Solanky,[†,||] Katherine A. Edmonds,[†] Harry Luithardt,[⊥] Jasna Fejzo,[#] Michael Chorev,[‡,§] Huseyin Aktas,[‡,§] Kimberly Colson,[||] Kenneth H. Falchuk,[‡,§] Jose A. Halperin,[‡,§] and Gerhard Wagner*[,†]

Contribution from the Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology, 240 Longwood Avenue, Boston, Massachusetts 02115, Harvard Medical School, Laboratory for Translational Research, One Kendall Square, Building 600, Cambridge, Massachusetts 02139, Brigham and Women's Hospital, Department of Medicine, 75 Francis Street, Boston, Massachusetts 02115, Bruker Biospin, 15 Fortune Drive, Billerica, Massachusetts 01821, Delta Search Labs, 400 Technology Square, Cambridge, Massachusetts 02139, Novartis Institutes for Biomedical Research, 250 Mass Avenue, Cambridge, Massachusetts 02139

Received November 28, 2006; E-mail: gerhard_wagner@hms.harvard.edu

Abstract: To obtain a comprehensive assessment of metabolite levels from extracts of leukocytes, we have recorded ultrahigh-resolution $^{1}H-^{13}C$ HSQC NMR spectra of cell extracts, which exhibit spectral signatures of numerous small molecules. However, conventional acquisition of such spectra is time-consuming and hampers measurements on multiple samples, which would be needed for statistical analysis of metabolite concentrations. Here we show that the measurement time can be dramatically reduced without loss of spectral quality when using nonlinear sampling (NLS) and a new high-fidelity forward maximum-entropy (FM) reconstruction algorithm. This FM reconstruction conserves all measured time-domain data points and guesses the missing data points by an iterative process. This consists of discrete Fourier transformation of the sparse time-domain data set, computation of the spectral entropy, determination of a multidimensional entropy gradient, and calculation of new values for the missing time-domain data points with a conjugate gradient approach. Since this procedure does not alter measured data points, it reproduces signal intensities with high fidelity and does not suffer from a dynamic range problem. As an example we measured a natural abundance $^{1}H-^{13}C$ HSQC spectrum of metabolites from granulocyte cell extracts. We show that a high-resolution $^{1}H-^{13}C$ HSQC spectrum with 4k complex increments recorded linearly within 3.7 days can be reconstructed from one-seventh of the increments with nearly identical spectral appearance, indistinguishable signal intensities, and comparable or even lower root-mean-square (rms) and peak noise patterns measured in signal-free areas. Thus, this approach allows recording of ultrahigh resolution $^{1}H-^{13}C$ HSQC spectra in a fraction of the time needed for recording linearly sampled spectra.

Introduction

Comprehensive measurements of the concentrations of large numbers of metabolites can provide detailed insights into the state of cells.[1,2] This has the potential of being used to diagnose disease, to follow the effect of drug treatment, or to study toxicity. Comparison of metabolite samples from different groups, such as healthy and sick individuals, or normal and transformed cells, may lead to identification of biomarkers that can be invaluable for targeted disease diagnosis or for understanding metabolic pathways and mechanisms of disease. Metabolic profiling of cell lines may lead to new insights into metabolic pathways and their alteration in disease.

Assessment of metabolite levels in metabolomics studies has primarily relied on mass spectroscopy, NMR spectroscopy, chromatographic separation techniques, and various multivariate data analysis techniques.[1−3] Among the NMR methods, one-dimensional (1D) ^{1}H spectroscopy is most commonly used, where the spectrum is divided into a limited number of buckets. The signal intensities of the buckets are either directly compared between multiple samples using principle component analysis

[†] Harvard Medical School, Department of Biological Chemistry and Molecular Pharmacology.
[‡] Harvard Medical School, Laboratory for Translational Research.
[§] Brigham and Women's Hospital.
[||] Bruker Biospin.
[⊥] Delta Search Labs.
[#] Novartis Institutes for Biomedical Research.

(1) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Curr. Opin. Mol. Ther.* **2004**, *6*, 265−272.
(2) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181−1189.
(3) El-Deredy, W. *NMR Biomed.* **1997**, *10*, 99−124.

10.1021/ja068541x CCC: $37.00

(PCA) or, if a prior model is assumed, partial least-square discriminant analysis (PLS-DA), or similar statistical approaches are used in order to separate groups of samples. All of these are done in an effort to identify the metabolites (biomarkers) that are most different between the groups.[4] Alternatively, 1D NMR spectra have been directly fitted to databases of metabolite reference spectra to obtain concentrations of the most abundant molecules in a metabolite mixture.[5] Occasionally, two-dimensional (2D) NMR spectra have been used to enhance the spectroscopic resolution for more detailed metabolite identification.[6−10] Resonances of metabolites typically have very long transverse relaxation times so that 2D NMR spectra, such as $^{1}H-^{13}C$ HSQC or $^{1}H-^{1}H$ TOCSY, can principally be recorded at very high resolution and can resolve nearly all metabolite signals. However, this requires a large number of increments, which makes data acquisition very time consuming and impractical for recording spectra from multiple samples as is necessary for statistical analysis.

It has been proposed and experimentally verified in the past that the measuring times of 2D NMR spectra can be reduced with nonlinear sampling in the indirect dimensions.[11,12] Barna and co-workers sampled 2D NMR data with an exponential schedule[13] and processed the data with a maximum entropy method (MEM) relying on the Burg algorithm.[14] Subsequently, nonlinear sampling was seriously pursued by several other groups.[15−19] Processing software was developed that relied on an alternative maximum entropy (MaxEnt) reconstruction algorithm that could handle phase-sensitive data and included adjustable parameters to tune the outcome of the data reconstruction. This software is available through the Rowland NMR Toolkit (RNMRTK)[15] and has been successfully applied for processing 2D nonlinearly sampled COSY spectra,[20] constant-time evolution periods of triple-resonance data,[21] or quantification of HSQC spectra for relaxation experiments.[22] Development of a two-dimensional MaxEnt reconstruction procedure designed to run in parallel on workstation clusters[23] has stimulated several applications to explore optimum evolution times,[24] to rapidly acquire complete sets of triple resonance experiments for sequential assignments,[25] to facilitate side-chain assignments,[26] and to enable high-resolution triple resonance experiments.[27,28]

The principle advantages of nonlinear sampling are increasingly recognized.[29] Besides MaxEn reconstruction, other methods are used for processing nonlinearly recorded spectra, such as the maximum likelihood method (MLM),[30] a Fourier transformation of nonlinearly spaced data using the Dutt−Rokhlin algorithm,[31] or multidimensional decomposition (MDD).[32−35]

A special case for reducing the time to record 3D or higher-dimensional data is used with sampling along various angles of the indirect sampling space,[36,37] and previously developed projection reconstruction procedures[38] are used for processing the data. Another strategy for minimizing sampling times is the reduced dimensionality approach,[39] which has been further developed into the GFT method[40] and applied to proteins up to 21 kDa.[41] However, to our knowledge, the projection reconstruction method and the GFT approach have primarily been applied to rather small systems where sensitivity is less of an issue. Unlike MaxEnt reconstruction, this class of approaches is only suited for three- and higher-dimensional experiments but not applicable to shortening acquisition of 2D experiments.

In the past, we have used the MaxEnt reconstruction procedure of the Rowland NMR Toolkit (RNMRTK).[12] It is very efficient and ideally suited for handling nonlinearly sampled data with a low dynamic range, such as triple-resonance spectra where all peaks have similar intensities. However, processing of data with a large variation of peak intensities, such as in NOESYs, TOCSYs, mixtures of metabolites, spectra with diagonals, or peaks close to the noise level seems to suffer from effects of nonlinearity of peak intensities. This has previously been recognized, and remedies for correcting intensities have been suggested for MaxEnt reconstructions of relaxation data.[42] However, there remains the problem of losing weak peaks that

(4) Gavaghan, C. L.; Wilson, I. D.; Nicholson, J. K. *FEBS Lett.* **2002**, *530*, 191−196.
(5) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78*, 4430−4442.
(6) Liang, Y. S.; Kim, H. K.; Lefeber, A. W.; Erkelens, C.; Choi, Y. H.; Verpoorte, R. *J. Chromatogr., A* **2006**, *1112*, 148−155.
(7) Keating, K. A.; McConnell, O.; Zhang, Y.; Shen, L.; Demaio, W.; Mallis, L.; Elmarakby, S.; Chandrasekaran, A. *Drug Metab. Dispos.* **2006**, *34*, 1283−1287.
(8) Dunn, W. B.; Bailey, N. J.; Johnson, H. E. *Analyst* **2005**, *130*, 606−625.
(9) Rhee, I. K.; Appels, N.; Hofte, B.; Karabatak, B.; Erkelens, C.; Stark, L. M.; Flippin, L. A.; Verpoorte, R. *Biol. Pharm. Bull.* **2004**, *27*, 1804−1809.
(10) Frederich, M.; Cristino, A.; Choi, Y. H.; Verpoorte, R.; Tits, M.; Angenot, L.; Prost, E.; Nuzillard, J. M.; Zeches-Hanrot, M. *Planta Med.* **2004**, *70*, 72−76.
(11) Barna, J. C. J.; Laue, E. D. *J. Magn. Reson.* **1987**, *75*, 384−389.
(12) Hoch, J. C.; Stern, A. S.; Donoho, D. L.; Johnstone, I. M. *J. Magn. Reson.* **1990**, *86*, 236−246.
(13) Barna, J. C. J.; Laue, E. D.; Mayger, M. R.; Skilling, J.; Worrall, S. J. P. *J. Magn. Reson.* **1987**, *73*, 69−77.
(14) Burg, J. In *Maximum Entropy Spectral Analysis*, 37th Meeting of the Society for Exploratory Geophysics, 1967; Childers, D. G., Ed.; IEEE Press: New York, 1978.
(15) Hoch, J. C. *Methods Enzymol.* **1989**, *176*, 216−241.
(16) Delsuc, M. A. A New Maximum Entropy Processing Algorithm, with Applications to Nuclear Magnetic Resonance Experiments. In *Maximum Entropy and Bayesian Methods*; Skilling, J., Ed.; Kluwer Academic: Amsterdam, 1989.
(17) Robin, M.; Delsuc, M. A.; Guittet, E.; Lallemand, J. Y. *J. Magn. Reson.* **1991**, *92*, 645−650.
(18) Jones, J.; Hore, P. *J. Magn. Reson.* **1991**, *92*, 276−292.
(19) Jones, J.; Hore, P. *J. Magn. Reson.* **1991**, *92*, 363−376.
(20) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Biomol. NMR* **1993**, *3*, 569−576.
(21) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Biomol. NMR* **1994**, *4*, 483−490.
(22) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Magn. Reson.* **1997**, *125*, 332−339.
(23) Li, K. B.; Stern, A. S.; Hoch, J. C. *J. Magn. Reson.* **1998**, *134*, 161−163.
(24) Rovnyak, D.; Hoch, J. C.; Stern, A. S.; Wagner, G. *J. Biomol. NMR* **2004**, *30*, 1−10.
(25) Rovnyak, D.; Frueh, D. P.; Sastry, M.; Sun, Z. Y.; Stern, A. S.; Hoch, J. C.; Wagner, G. *J. Magn. Reson.* **2004**, *170*, 15−21.
(26) Sun, Z. J.; Hyberts, S. G.; Rovnyak, D.; Park, S.; Stern, A. S.; Hoch, J. C.; Wagner, G. *J. Biomol. NMR* **2005**, *32*, 55−60.
(27) Sun, Z. Y.; Frueh, D. P.; Selenko, P.; Hoch, J. C.; Wagner, G. *J. Biomol. NMR* **2005**, *33*, 43−50.
(28) Frueh, D. P.; Sun, Z. Y.; Vosburg, D. A.; Walsh, C. T.; Hoch, J. C.; Wagner, G. *J. Am. Chem. Soc.* **2006**, *128*, 5757−5763.
(29) Tugarinov, V.; Kay, L. E.; Ibraghimov, I.; Orekhov, V. Y. *J. Am. Chem. Soc.* **2005**, *127*, 2767−2775.
(30) Chylla, R. A.; Markley, J. L. *J. Biomol. NMR* **1995**, *5*, 245−258.
(31) Marion, D. *J. Biomol. NMR* **2005**, *32*, 141−150.
(32) Korzhneva, D. M.; Ibraghimov, I. V.; Billeter, M.; Orekhov, V. Y. *J. Biomol. NMR* **2001**, *21*, 263−268.
(33) Orekhov, V. Y.; Ibraghimov, I. V.; Billeter, M. *J. Biomol. NMR* **2001**, *20*, 49−60.
(34) Orekhov, V. Y.; Ibraghimov, I.; Billeter, M. *J. Biomol. NMR* **2003**, *27*, 165−173.
(35) Gutmanas, A.; Jarvoll, P.; Orekhov, V. Y.; Billeter, M. *J. Biomol. NMR* **2002**, *24*, 191−201.
(36) Kupce, E.; Freeman, R. *J. Biomol. NMR* **2004**, *28*, 391−295.
(37) Kupce, E.; Freeman, R. *J. Am. Chem. Soc.* **2004**, *126*, 6429−6440.
(38) Hounsfield, G. N. *Br. J. Radiol.* **1973**, *46*, 1016.
(39) Szyperski, T.; Wider, G.; J.H., B.; Wuthrich, K. *J. Am. Chem. Soc.* **1993**, *115*, 9307−9308.
(40) Kim, S.; Szyperski, T. *J. Am. Chem. Soc.* **2003**, *125*, 1385−1393.
(41) Liu, G.; Aramini, J.; Atreya, H. S.; Eletsky, A.; Xiao, R.; Acton, T.; Ma, L.; Montelione, G. T.; Szyperski, T. *J. Biomol. NMR* **2005**, *32*, 261.
(42) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Magn. Reson.* **1997**, *125*, 332−339.

are barely above noise level, which is a significant problem in spectra of metabolite mixtures with large variations of peak intensities.

In another strategy to cope with the dynamic range problem we examine here whether this issue can be eliminated by using a new maximum-entropy-related approach that operates on a different principle. In the classic MaxEnt algorithm, such as that used by Hoch and Stern,[43] the linearity of the reconstruction depends of the parameter lambda (λ), a Lagrange multiplier that is required to create the objective function $Q(f) = S(f) - \lambda C(f,d)$, where f and d are the vectors of reconstructed frequency domain and measured time domain data points, respectively.

The Shannon entropy is

$$S(f) = -\sum_{k=1}^{N} f_k \log f_k$$

(which has the same form as Gibbs' entropy, yet applies to information content), and $C(f,d)$ is the constraining function between the measured time-domain data points and those calculated from the inverse Fourier transformed iteratively "guessed" mock spectra. By the nature of altering the spectra in the frequency domain, $C(f,d)$ is always >0, except in very rare cases. In general, keeping it at zero does not allow for alteration of data points in the frequency domain.

Here we present a new procedure termed **F**orward **M**aximum entropy (FM) reconstruction for processing nonlinearly sampled 2D NMR data. Similarly to the MaxEnt reconstruction of Hoch and Stern,[12] it aims to minimize a target function that contains the negative entropy. However, in contrast to the MaxEnt reconstruction procedure,[43] which allows a variation of the measured data points by maximizing the target function $Q(f) = S(f) - \lambda C(f,d)$, the FM reconstruction only optimizes the time-domain data points that have not been measured and does not allow for variation of the acquired time-domain data points. We claim that this approach assures high fidelity of peak intensity reconstruction. Thus, all experimentally measured data points are strictly conserved, which we claim enforces correct relative intensities. In contrast, a procedure that allows variation of measured data points to minimize the target function $Q(f)$ is tempted to do this at the expense of weak peaks. By altering the method such that the "guessing" occurs in the time domain, the constraining function $C(f,d)$ can easily be eliminated, and the requirement of the Lagrange multiplier thus vanishes. The objective function is now reduced to $Q(f) = S(f)$. The implementation hence requires the "mock" data to be constantly "guessed" in the time domain, then forward-Fourier transformed, where they are scored by only calculating the implemented entropy function. This approach does not seriously bias against weak signals. We validate this approach with a ^{1}H$-^{13}$C HSQC spectrum of metabolites from cell extracts recorded with 8k real increments. Acquisition of the linearly sampled reference spectrum required 3.7 days on a 600 MHz spectrometer. We show that the same quality of spectrum can be obtained within fourteen hours using nonlinear sampling and the FM reconstruction. This makes possible recording of multiple ultrahigh-resolution ^{1}H$-^{13}$C HSQC spectra needed for statistical analysis of metabolomics data.

Materials and Methods

Principles of the Forward Maximum Entropy (FM) Reconstruction. We pursued reconstruction of a complex time-domain data set $F(d) = \{d_i;\ i = 1, ..., N\}$ where a subset of points $\{d_k;\ k = 1, ..., M\}$ ($M < N$) has been measured experimentally but all other points are unknown. We guessed the unknown data points in the time domain. Initially, all unknown points are set to complex zero or 0 + i0. Subsequently we processed the spectrum with the fast Fourier transform (FFT) algorithm and calculated its entropy:

$$S(f) = -\sum_{k=1}^{N} f_k \log f_k$$

Because many of the data points are fixed and given by the experimental data set d, we used a simplistic approach of the entropy $S(f)$ for complex data points f_k, namely

$$S(f) = -\sum_{k=1}^{N} |f_k| \log |f_k|$$

We pursued to maximize $Q(f) = S(f)$ iteratively using the forward maximum entropy (FM) reconstruction method proposed here that maintains the experimental time-domain data points unaltered. Since this approach is equivalent to minimizing the negative entropy, we defined the target function $T(f) = -S(f)$, which we aimed to minimize. $T(f)$ is related to the norm of a spectrum, and minimizing it reduces the total signal and noise subject to maintaining the measured time-domain data points.

The FM reconstruction program described here starts by setting the missing time-domain data points to zero. The data are then transformed with the FFT algorithm, and the target function of the resulting spectrum is calculated. Next, the guessed time-domain data points are changed by $\pm\delta$, followed by FFT and calculation of the target functions. From the resulting values a multidimensional gradient of $T(f)$, $\nabla\vec{T}(f)$, is calculated with respect to the missing data points d_j. Using this gradient $T(f)$ is minimized with the Polak$-$Ribiere conjugate gradient method[44] to gradually change the guesses of the missing time-domain data points d_j. In the minimization, the real and the imaginary parts of the time-domain data points d_j are treated as independent variables. This creates a total of $2 \times (N - M)$ free vectors of$\nabla\vec{T}(f)$, where N is the number of grid (final) points, and M is the number of measured time points (see above), and the difference of N and M is multiplied by two, because each point is of complex nature, and the real and imaginary components are minimized independently of each other.

Minimization is iterated until a cutoff criterion is reached, that is the value of the target function $T(f)$ does not decrease by more than the cutoff parameter, the gradient $\nabla\vec{T}(f)$ is sufficiently small, or a maximum number of desired iterations have been reached.

The final result of the FM reconstruction procedure is a time-domain data set where the missing data points are filled in with the optimum values obtained by minimizing $T(f)$. This time-domain data set can now be transformed with any available data processing software and can be manipulated with window functions, zero filling, and/or linear prediction.

Practical Implementation of FM Reconstruction. The package FFTW (http://www.fftw.org), version 3.0.1, was used as a C library for computing discrete Fourier transforms. FFTW is distributed under the GNU software licenses and is free software. The Polak$-$Ribiere conjugate gradient method, gsl_multimin_fdsminimizer_conjugate_pr of the package GSL, vers 1.5 (GNU Scientific Library, http://www.gnu.org/software/gsl), was used for minimization purposes. For data handling, we used the open NMRPipe data format for describing

(43) Hoch, J. C.; Stern, A. S. *Methods Enzymol.* **2001**, *338*, 159−178.

(44) Polak, E.; Ribiere, G. *Rev. Fr. Inf. Rech. Oper.* **1969**, *3*, 35−43.

NMR data. See http://spin.niddk.nih.gov/bax/software/NMRPipe for a description of NMRPipe. The resulting C program was initially compiled using gcc and the i686 instruction set on a Dell dual Xenon 3.0 GHz computer operating under Fedora Core 1 Linux OS. The program has since been compiled and run on other RedHat Linux computers. This includes use of 64-bit Opteron computers. A distributable version of the program will be developed and made available.

Generation of Sampling Schedules. Three different sampling schedules, S1, S2, and S3, were used to pick increments from the total data set. For S1, increments were picked randomly but with constant density along t_1. S2 and S3 also sample randomly, but the sampling density is not constant. S2 uses an exponentially decreasing sampling density, and S3 applies a linearly decreasing pick rate. The random sampling schedules are created based on Unix tools with the following procedure: *First*, a vector is created with the length of the final number of complex slots N (4096 in our case) with associated ordinal number, starting at 0. *Second*, in each slot, a number is placed that represents the sampling density. For the constant density of schedule S1 this is 1.0 for each slot. For the exponentially decaying density of schedule S2 this is 1.0 for slot 0 and $\exp\{-i/(N-1)\}$ for slot $i = 1$ to $N-1$. For the linear ramp of schedule S3 this is 1.0 for slot 0 and $1.0 - i/N$ for slot $i = 1$ to $N-1$. *Third*, the sum SU(Sk) of all slot values is calculated for each of the schedules Sk (S1, S2, S3). *Fourth*, a random number is generated with the function *drand48*, which yields a 48-bit random number in the range from 0.0 through 1.0. This random value is multiplied by the sum SU(Sk) for each of the three schedules. *Fifth*, the values in the respective slots are cumulatively added to the point where the sum is larger than the value created in the fourth point above. *Sixth*, the slot number is now registered as a point to be sampled in the nonlinear acquisition, and the slot value is now set to zero, which reduces the value of the sum SU(Sk) for the next round. *Seventh*, points, 3, 4, 5, and 6, are repeated until the desired number of sampling points of the nonlinear sampling schedule is reached, and the numbers are then sorted.

To initialize the random function *drand48*, a seed number is required. Either a rationally chosen seed number may be given, or a seed number can be created from the internal clock which registers the number of seconds since New Years Day, 1970, 0.00.00. The latter is the reference point for UNIX. While both approaches yield good results, we have noticed that certain seed numbers yield slightly better reconstructions and speedier convergence than others. Thus, choosing "favorite" or "rationally chosen" seed numbers may be preferable to optimize the obtainable results. For further information on selecting random numbers and the use of *drand48*, we refer to the UNIX manuals.

Preparation of Cell Extracts from Mouse Baf3 Cells. Mouse BaF3 cells were obtained from Dr. James Griffin at the Dana Farber Cancer Institute. BaF3 is a murine blood cell line dependent on IL-3 for survival and differentiation. Cells were cultured in the presence of 5% CO_2 at 37 °C in RPMI 1640 medium supplemented with 10% fetal calf serum, and penicillin/streptomycin. Media for BaF3 was additionally supplemented with WEHI 3B conditioned medium as a source of IL-3. About 2 billion cells (~3 mL) were lysed by the sequential addition of 4 mL of methanol, 4 mL of chloroform, and 4 mL of water. The sample was vortexed vigorously after the addition of each solvent, and the final mixture was stored at −20 °C overnight for phase separation. Complete separation of phases was achieved by centrifugation at 10000*g* for 40 min. Only the aqueous phase was used here. It was lyophilized and dissolved in ^2H_2O for the NMR experiments. Extracts from 2×10^9 cells were used for the sample.

NMR Spectroscopy. NMR spectra were recorded on a Bruker Avance 600 spectrometer equipped with a cryogenic triple-resonance probe. A set of seven $^1H-^{13}C$ HSQC spectra were recorded with 4k increments (complex) and four scans per increment. A relaxation delay of 1.2 s was used between scans. Each of the seven spectra was recorded in 12.7 h; the total measurement time for all seven 2D spectra was 3.7 days.

Results

Recording an Ultrahigh Resolution (UHR) $^1H-^{13}C$ HSQC Spectrum. As a first step toward identifying metabolites in BaF3 cells we have recorded a 1D spectrum of the aqueous phase of cell extracts in ^2H_2O, which is shown in Figure 1A. Identification and measurement of concentrations of metabolites is severely hampered by spectral overlap. Similarly, 2D NMR spectra recorded with the typically 100−200 increments suffer from low resolution in the indirect dimension. Thus, we recorded an ultrahigh resolution (UHR) $^1H-^{13}C$ HSQC spectrum with 4k increments (complex data points) (Figure 1B). The signal separation obtained with 4k complex increments resolves essentially all overlap, as is shown with the expansion of the most crowded region in Figure 1D. The dispersion of the UHR spectrum promises to provide a tool for assigning nearly all metabolites that are present in sufficiently high concentrations and for measuring concentrations of the individual metabolites.

The 4k complex increments result in a maximum t_1 value of 0.16 s. This is rather long compared to typically recorded 2D $^1H-^{13}C$ HSQC experiments; however, it should be compared with the transverse relaxation times for metabolite carbons, which are in the order of one or several seconds. As stated previously, it is desirable to sample evolution times close to T_2 to obtain optimal resolution and sensitivity,[24] which is still far off with the conditions used here. To experimentally examine the optimal number of increments for the sample used here, we transformed the data set using 4k, 2k, 1k, and 512 complex t_1 values. Figure 2 shows the effect on a small crowded region that is indicated with a box in Figure 1D. To have comparable scaling, all data sets were multiplied with a cosine window and zero-filled to 8k points. This ensures identical scaling with the number of points in the FIDs. The 4k and 2k data clearly resolve all peaks, but the transformations of only 1024 and 512 increments do not. A cross section drawn through the strongest peak demonstrates that increased resolution also results in larger peak height. Thus, the relative height of the tallest peak is 3.9 : 3.0 : 2.0 : 1.0 in the 4k, 2k, 1k, and 512 point transforms, respectively (Figure 2). Although the apparent resolution does not improve much by going from 2k to 4k complex data points, in t_1 the peak height increases by approximately 30%, which is close to $\sqrt{2}$ (41% increase) and consistent with the expectation (Figure 2). Obviously, doubling the number of t_1 values from 2k to 4k doubles the measuring time, and the spectrum shown here required a total of 3.7 days of instrument time. This is undesirably long if one wants to measure multiple samples for determining statistically significant differences of metabolite concentrations between cell types or different populations of the same cells. It is indeed the long-term goal of our research to identify metabolites with concentrations that differ between cell samples. This includes comparison of normal and malignant cells, cells before and after treatment with drugs, or any other pairs of distinct cell types. Measurement of multiple samples for each cell sample would be difficult with measurement times of 3.7 days as was used for the spectrum in Figure 1.

To reduce the measurement time we have explored the nonlinear sampling approach with FM reconstruction for data processing. We hypothesized that this approach allows recording

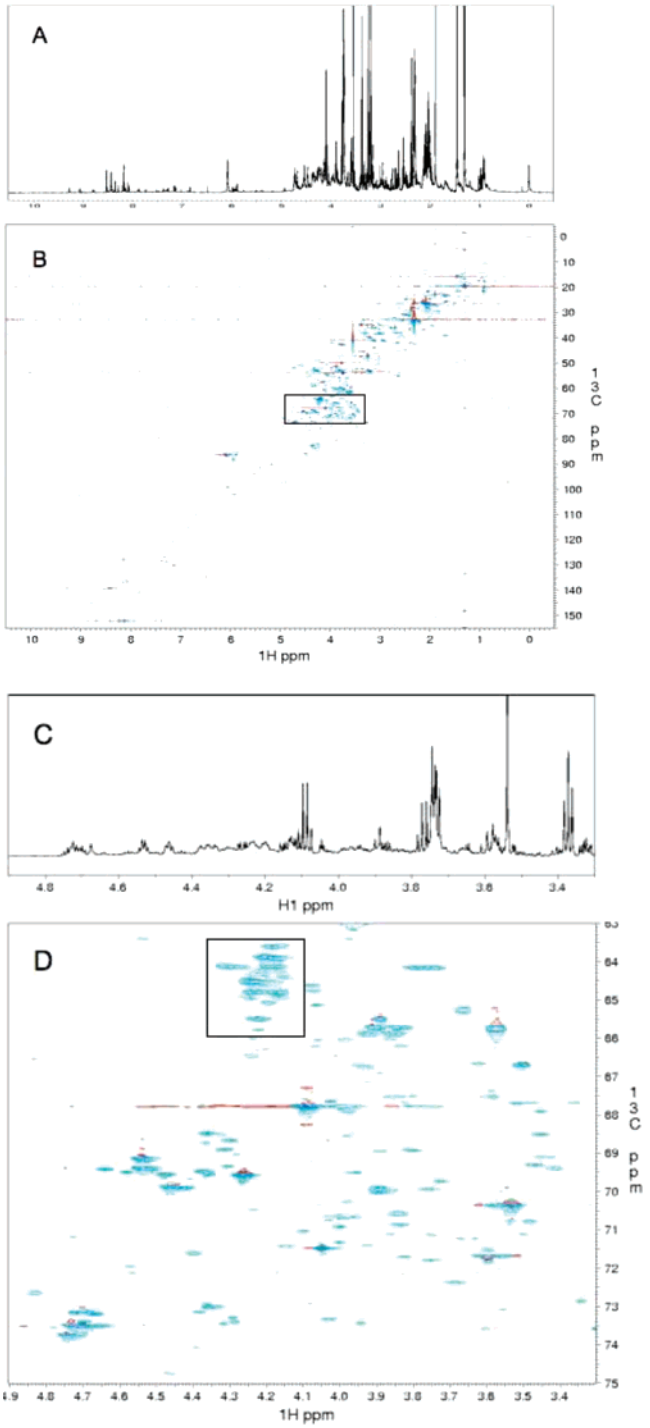

Figure 1. NMR spectra of the aqueous fraction of cell extracts from mouse BaF3 cells in ^2H_2O, pH 6.5, 25 °C. (A) 1D ^1H NMR spectrum. (B) $^1H-^{13}C$ HSQC spectrum recorded with 4k complex points in the indirect dimension. (C and D) Expansion of the section indicated with the box in B and corresponding 1D spectrum. Note that the 2D spectrum of C is a small portion of the entire spectrum. The 1D spectrum of C also contains signals from outside the ^{13}C region of Figure 2D (compare Figure 1A and 1B).

of high-resolution 2D spectra within a reasonably short time and developed the FM reconstruction method with the goal to avoid bias against weak peaks.

To test this approach we have recorded seven identical linearly sampled high resolution $^1H-^{13}C$ HSQC spectra acquired over a period of 12.7 h each. Addition of the seven spectra and standard FFT yields the spectrum shown in Figure 1. Recording all seven spectra required 3.7 days of instrument time. The high resolution achieved is demonstrated by showing an expansion of a small portion of the most crowded spectral region (Figure 1D). Numerous $^1H-^{13}C$ cross-peaks are visible, and all are very well resolved.

Impact of Nonlinear Sampling with FM Reconstruction on Resolution and S/N. To avoid impractically long measuring times and to enable measurements on several samples for statistical analysis, we tested whether nonlinear sampling and a suitable processing routine would allow shortening the measurement time. Since spectra of mixtures of metabolites have a large variation of intensities, we employed the FM reconstruction approach outlined above. To test this we have recorded the data set described above. To get the subsets of nonlinearly sampled data we (1) added all seven data sets together, (2) we selected a nonlinear sampled regime from the combined data set using one-seventh of the increments. This allowed us to compare spectra recorded linearly within 12.7 h with data recorded nonlinearly within the same amount of time, with only one-seventh of the increments but 7 times the number of scans per increment.

Figure 3 shows different versions of a representative cross section along the carbon dimension of the $^1H-^{13}C$ HSQC spectrum of Figure 1. Figure 3A provides transformations of all linearly acquired data points. The lower three traces are from spectra 2, 4, and 6 out of the seven spectra recorded for 12.7 h each, using 16 scans per increment. (Note that the peak at 57.35 ppm decreases with time due to metabolic changes ongoing in the sample; it is only visible in spectra 1−3 and disappears in later spectra. Although we tried to stabilize the metabolite samples, some peaks change over a period of 3.7 days, and the individual spectra are not entirely identical. It was indeed a motivation for developing the fast method described here to cope with long-term changes in the metabolite samples.) The top spectrum is the sum of all seven spectra and represents a measuring time of 3.7 days with 7×16 scans per increment. It demonstrates the gain in signal-to-noise ratio (S/N) by a factor of $\sqrt{7}$ compared to the individual spectra, such as 2, 4, or 6 shown in the lower part of the figure. The top spectrum of Figure 3B is the same as in Figure 3A. The lower three traces, however, are FM reconstructions using only one-seventh of the increments but 7×16 scans per increment. Thus, each of the lower three traces represents the same total measuring time of 12.7 h, equal to each of the lower traces of Figure 3A. Three different sampling schedules, S1, S2, and S3, were used to pick increments from the total data set. For S1, increments are picked randomly but with constant density along t_1. S2 and S3 also sample randomly, but the sampling density is not constant. S2 uses an exponentially decreasing sampling density, and S3 applies a linearly decreasing pick rate. As can be seen, the quality of the three nonlinearly sampled spectra, transformed with FM, is comparable to that of the top trace, which represents a 7-fold longer measurement time. There is no obvious bias in favor of strong peaks or against weak peaks. The three sampling schedules exhibit similar results, but the schedule with exponentially decreasing weight (S2) seems to have the best S/N with a small margin. A more systematic examination of sampling schedules will be required, however, to optimize this procedure.

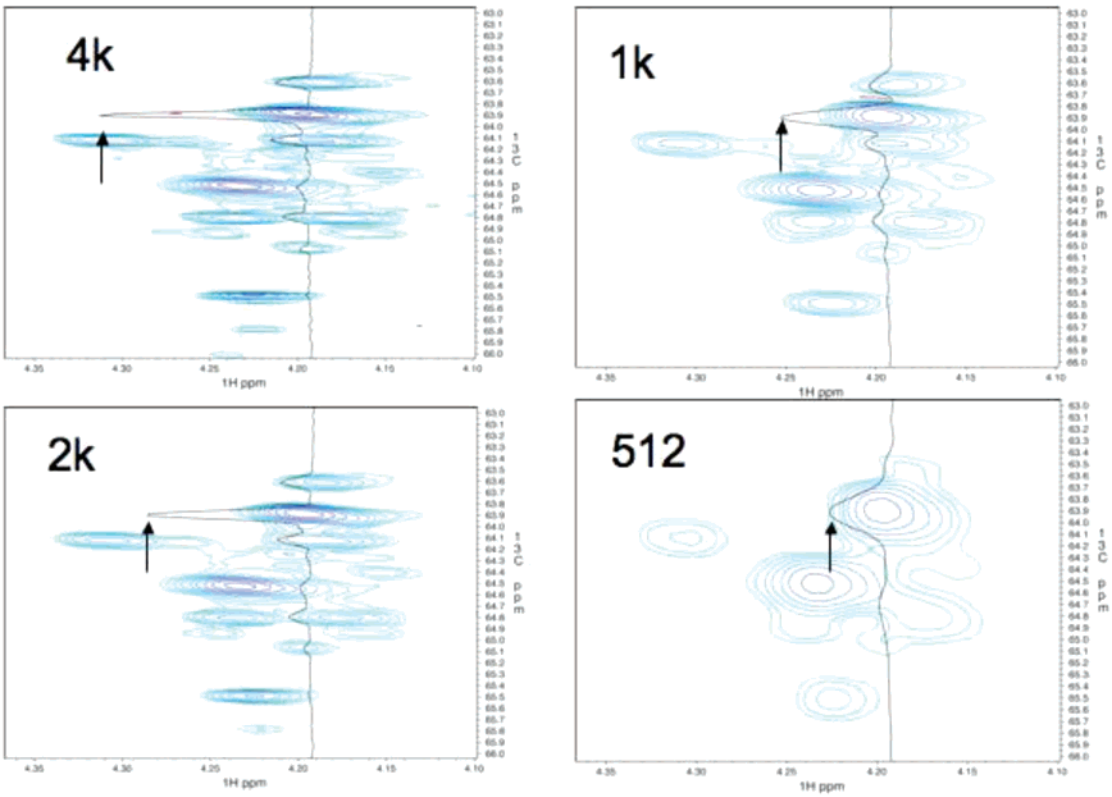

Figure 2. Comparison of a small spectral region indicated with a box in Figure 1D transformed with different numbers of increments. Clearly, only 2k and 4k complex points can resolve all peaks. While 2k complex increments seem to resolve all peaks, going to 4k complex points sharpens the peaks and increases the peak height by approximately 30% (see arrows). To obtain equal scaling in all four cases, the measured data points were multiplied with a cosine bell and zero-filled to 8k real points.

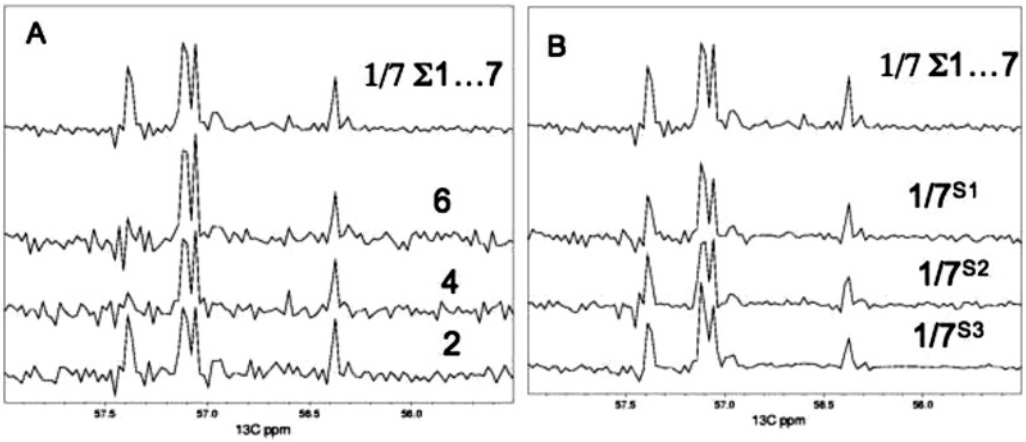

Figure 3. Comparison of representative cross sections along the ^{13}C direction at a proton frequency of 3.76 ppm. (A) Use of the full linearly sampled data. The bottom three traces are from the full linearly sampled data sets 2, 4, and 6. The top trace is from the average of all seven linearly sampled data sets (sum of all seven spectra divided by 7). (B) The top trace is the same as in A. The three traces at the bottom, however, are obtained by selecting one-seventh of the increments of the averaged data set and by transforming with FM reconstruction. In the sampling schedule, S1, one-seventh of the increments were picked randomly from the averaged data set with equal density along t_1. In schedule S2, one-seventh of the increments were picked with exponentially decreasing density, and in schedule S3, one-seventh of the increments were picked with linearly decreasing density. Note that the peak at 57.35 ppm disappears with time and is only visible in spectra 1−3. Although we tried to stabilize the metabolite samples, some peaks change over a period of 3.7 days. Thus, the spectra 2, 4, and 6 are not entirely identical.

To further assess the quality of the NLS spectra processed with FM reconstruction, we analyzed the apparent noise in the linearly and nonlinearly sampled spectra. Figure 4 shows a small representative section of the $^{1}H-^{13}C$ HSQC spectrum that contains strong and weak peaks. Cross sections through the strongest peak along the carbon dimensions were plotted on top of the contour plots. The traces from the linearly sampled spectrum (4 days measuring time) and the nonlinearly sampled spectrum (12.7 h measuring time) are essentially indistinguishable.

For further comparison of the quality of the two spectra, the apparent noise was measured in a region outside the range that contains signals (Figure 5). Both the root-mean-square (rms) noise and the peak noise were measured for the three linearly

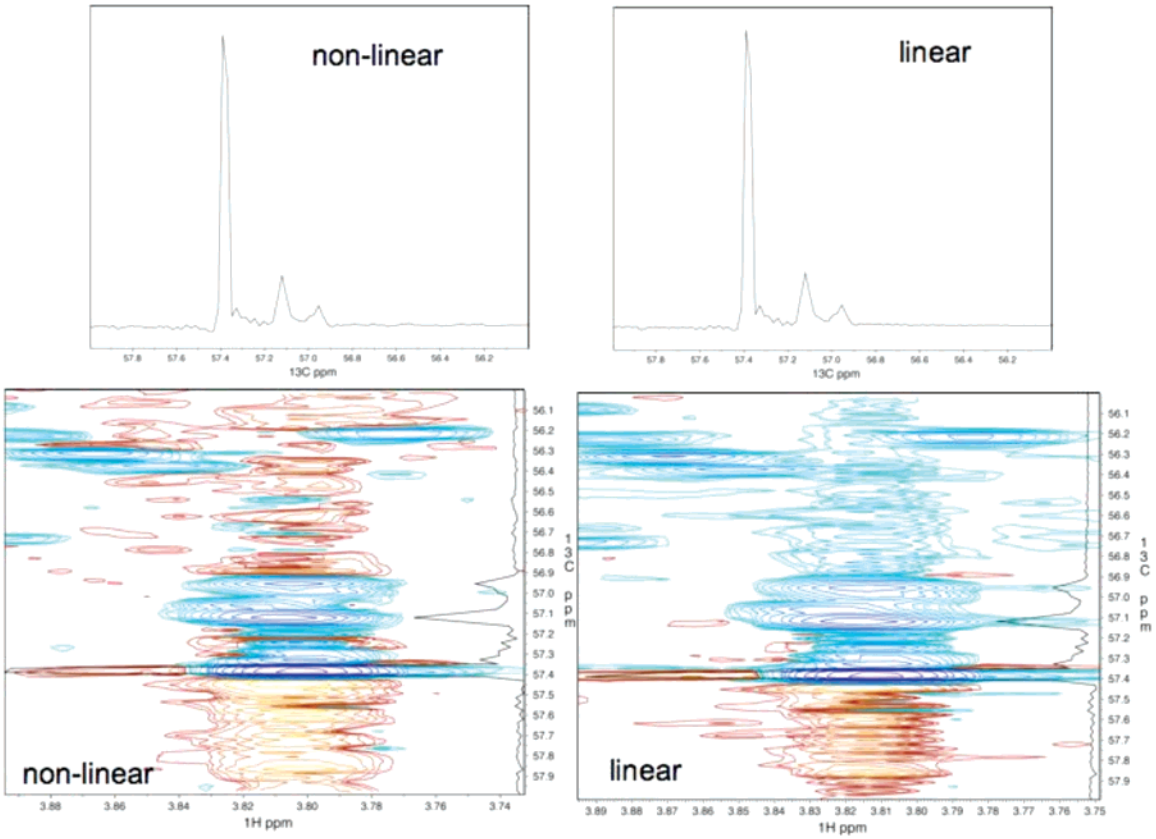

Figure 4. Comparison between the linearly sampled 3.7-day experiment and the NLS/FM reconstruction of a 12.7-h subset of increments. Both the 2D plot of a small portion of the spectrum and the cross section along the ^{13}C direction at the position of the strongest peak are nearly indistinguishable. This demonstrates the high fidelity of the FM reconstruction.

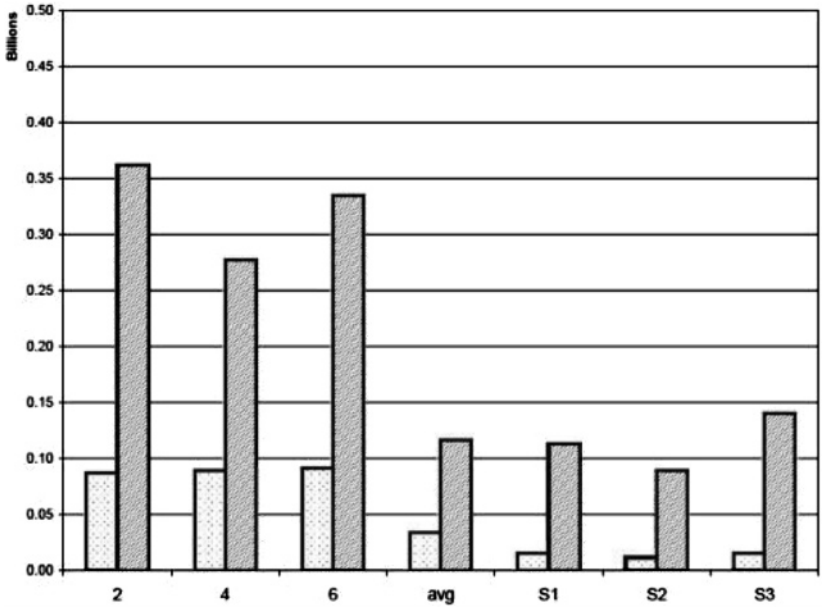

Figure 5. Comparison of rms noise (light) and peak noise (dark) of the cross sections along the ^{13}C dimension. Both rms and peak noise are measured outside of the region that contains signals. The columns labeled 2, 4, and 6 represent three of the seven linearly sampled spectra of 12.7-h duration. The column labeled "avg" shows rms and peak noise for the average of the seven linearly sampled spectra and corresponds to 3.7 days of data acquisition. As expected, the noise levels decrease by $\sqrt{7}$. S1, S2, and S3 show the measured noise values for three sampling schedules of the averaged spectra but use only one-seventh of the increments. S1 is a randomly distributed schedule, S2 is sampled with exponentially decreasing sampling frequency, and S3 corresponds to a linearly decreasing ramp.

sampled spectra 2, 4, and 6 as well as the average of all seven linearly recorded spectra. As expected, both the rms and peak noise are approximately $\sqrt{7}$ lower when averaging the seven linearly sampled spectra (left and middle columns). Importantly, the FM reconstruction of the NLS data picked from the averaged data set have roughly the same peak noise as the full averaged

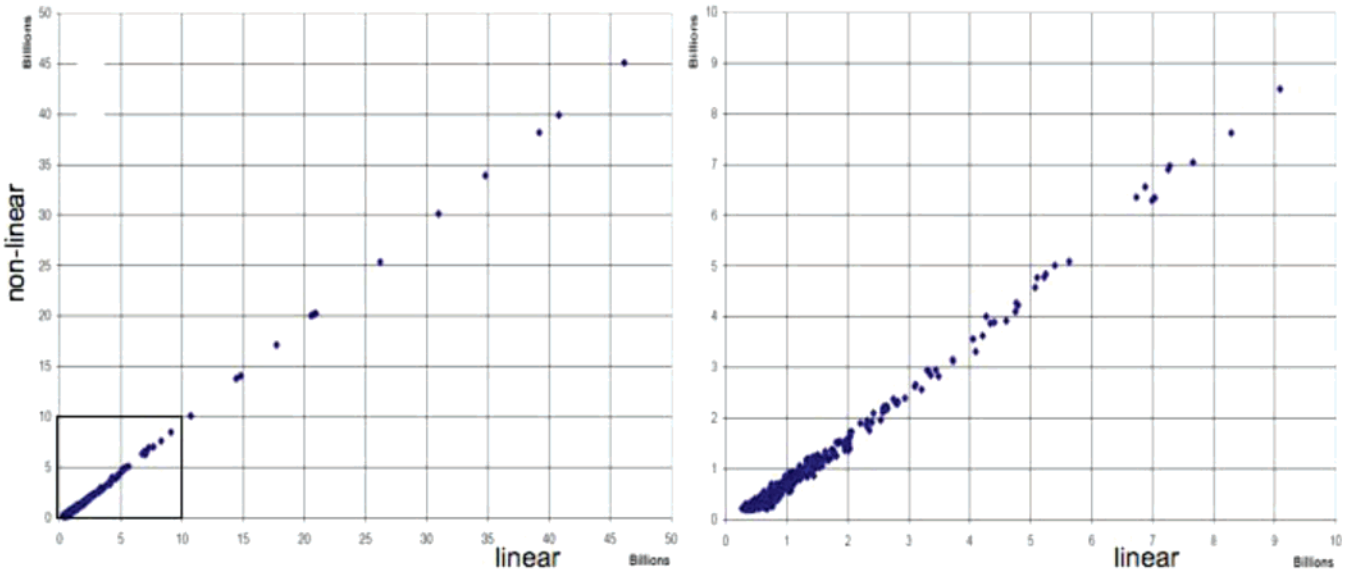

Figure 6. Comparison of peak heights between the linearly sampled 3.7-day experiment (horizontal) and the randomly sampled 12.7-h experiment (vertical). The latter provides a high-fidelity reproduction of the peak heights in the full linearly sampled spectrum. (Left) Total range of peak heights of peaks. The region shown in the expansion on the right is indicated with a box. (Right) Expansion of the section containing weak peaks. Peaks were measured if they were larger than approximately 0.2×10^9, which is approximately twice the amount of peak noise (see Figure 5). Note that the values for the nonlinearly sampled data are lower than those of the linearly sampled spectra by a constant offset of approximately the value of the peak noise.

data; it is lowest for the exponential schedule S2. For all three sampling schedules, the rms noise is approximately 2-fold lower than in the transform of the full linear averaged data set. Thus, the FM processing of the NLS data sets, which can be acquired in one-seventh of the measuring time, yield high-quality spectra comparable in quality to that of the DFT of the full linearly sampled averaged data set (Figure 5, middle and right columns).

To analyze whether NLS with FM reconstruction affects the relative intensities of signals, we measured peak heights of all detectable signals in the carbon cross section for 16 ^{13}C traces. In Figure 6, the peak heights in the linear averaged data set are plotted against the corresponding values in the NLS (random sampling schedule S1) data set processed with FM reconstruction. The left-hand side of Figure 6 shows the entire range of peak heights up to 45×10^9. The right-hand side displays an expansion of the range up to 10×10^9. The smallest peaks measured are around 2×10^8, which is about twice the level of peak noise (see Figure 5). Clearly, there is an excellent correlation, and there is no bias against weak peaks. Thus, perfect linearity is found over a dynamic range of 200, the largest found in the spectra analyzed here. However, the values in the nonlinearly sampled spectrum are smaller by a constant offset of approximately the value of the peak noise. This may be related to the fact that there is noise at the top of peaks as well, and if noise is reduced by the reconstruction procedure throughout the spectrum and at the tip of the peaks, the apparent peak heights decrease by a value in the order of the noise level. However, a more quantitative analysis of this effect has to be pursued. Of interest is that a similar empirical observation has been reported by Hore and co-workers for a different maximum entropy reconstruction approach.[18,19]

Discussion

NMR spectra of mixtures of metabolites as obtained from cell extracts or other metabolomics samples contain a large number of signals, and high-resolution is needed in 2D spectra if one wants to resolve all signals. Since the metabolites have all narrow line widths, the individual signals are resolvable in 2D NMR spectra but only when the experiments are sampled to long evolution times in the indirect dimension. Previously, we have argued signals should be recorded to about T_2^* in order to obtain good resolution without significantly deteriorating signal-to-noise.[24] For spectra of metabolites with narrow line widths this requires sampling to long evolution times and needs very long measuring times. We have recorded a ^{1}H$-^{13}$C HSQC spectrum with 8k by 2k (real) data points in the t_1 and t_2 dimensions, respectively. Estimating that the spectrum contains as many as 1000 cross-peaks, each peak is defined by 32k data points in the average. Thus, the information content of linearly sampled data is large compared to the number of parameters needed to characterize the spectrum. Thus, it seems feasible to extract spectral parameters from a reduced data set, such as obtained with NLS.

The maximum t_1 value in the spectrum recorded here is 0.16 s, which is still much less than the T_2 of ^{13}C signals of most metabolites ($>$1 s). Spectral folding in the indirect dimension would allow reaching longer evolution times at reasonable overall duration of the experiments. Attempts toward this goal are in progress but have not been used initially for convenience of comparison with database spectra of metabolites. Spectral folding raises the question of whether the FM method can reproduce positive and negative signals. The procedure can handle this very well. Indeed we have obtained positive and negative peaks, and the reconstructed FIDs are complex. We phase correct the final spectrum after FFT of the reconstructed time-domain data set.

We have developed a simple new algorithm that estimates missing time-domain data points of nonlinearly sampled data by conjugate gradient minimization of the target function $T(\mathbf{f})$, which is constructed from the negative Shannon entropy of the

frequency spectrum, $S(f)$. Here, the negative entropy is used as a convenient convex function that is efficient for minimizing the target function. It is essentially a norm of the frequency spectrum. The final spectrum reaches the minimum of the norm being consistent with the experimentally measured time-domain data points. We do not normalize the data points of the frequency spectrum, f_k, since they should not be considered probabilities. Thus, our approach differs from other MaxEnt methods for NMR spectrum reconstruction,[17] which pursue such normalization. Obviously, the Shannon entropy used here should not be confused with the thermodynamic or statistical entropy. Here we could use and have explored other convex functions to minimize the norm of the spectrum.

We have used zeros for the initial guesses of the missing data points. One could consider using other starting values. However, due to the oscillating nature of the free-induction decays, zero is in the center of the distribution of the possible expected values. We have explored using other starting values, such as using the values of adjacent measured points. This did not alter the outcome but typically extended the time to reach convergence.

We realize that experimental scientists often face situations where false minima are found in optimization procedures. We have examined different sampling schedules, which yielded almost identical results differing essentially only slightly in the noise level and the time it takes to reach convergence. Thus, we think that there is little danger of being trapped in local minima. However, this is likely to depend on how many data points are sampled in relation to the complexity of the spectra, and further investigations of this aspect are to be pursued.

It is worth asking whether the FM reconstruction alters peak shapes as has been reported for other methods of reconstructing NLS data. This does not directly apply to the metabolite data presented here, however. Because of the very long carbon transverse relaxation times, peaks are defined by one or two data points only. Thus, no distortion has or can be seen in the type of data shown here. However, we have started to apply this method to protein NOESY spectra and do not see significant distortions of peak shapes as long as we have enough data points in the indirect dimensions to define all spectral parameters. This leads to the question of what is the minimum number of nonlinearly sampled data points to faithfully reconstruct a spectrum? This depends on the type of spectrum and the density of signals. A NOESY of a large protein with a long mixing time and many cross-peaks will require more points than one of a small protein with a short mixing time.

By principle of design, the FM reconstruction algorithm presented here does not allow for variation of measured time-domain data points. Our hypothesis was that this feature would avoid de-emphasizing weak signals as long as they are represented in the measured data points. The results shown here confirm that this is indeed the case. FM reconstruction does not require setting of parameters for the reconstruction. The only operator decision to be made is when to terminate the iterative minimization of the target function $T(f)$. On the other hand, our FM reconstruction is not suitable for and cannot be applied to linearly sampled data sets. However, it could be used for correcting erroneous or lost data points. The final FM reconstruction result is a time-domain data set that can be transformed with any of the available processing programs, manipulated with apodization functions, or extended with linear prediction. Here we have used the NMRPipe software package[45] for all processing.

While we have applied window functions to the final reconstructed time-domain data one could also consider doing this to the initial nonlinearly sampled data. As the reconstruction method used here is deemed to be of nonlinear nature, there is no warranty that the reconstruction yields exactly the same result in these two cases. However, because we are aiming for linear response of signals, the FM reconstruction has to be at least of near-linear nature. This means that the FM reconstruction method should be nearly independent of whether window functions are applied before or after apodization. In other words, the result would be extremely similar, within the scope of setting stop criteria of the minimization routine, and in either case, apodization does not interfere with the core of the FM method.

Thus far, we have used FM reconstruction for processing spectra with only one nonlinearly sampled indirect dimension. However, this approach should be applicable to nD data that are sampled nonlinearly in more than one indirect dimension. Here we have used a standard conjugate gradient minimizer, and processing is relatively slow as reconstruction of one 4k complex time-domain data set with six-sevenths of the data points missing takes about 3−5 h on an Opteron computer. Obviously, the processing time depends on the number of missing points, and shorter time-domain data can be processed significantly faster. Processing of 2D spectra benefits from farming out the reconstruction of FIDs to processors of PC clusters. However, the current FM reconstruction is significantly slower than the MaxEnt routine used in the Rowland NMR Toolkit (RNMRTK), which uses an analytically calculated gradient for minimizing the target function.

In principle, it is possible to apply FM reconstruction to higher-dimensional spectra with more than one nonlinearly sampled dimension, and efforts toward this aim are in progress.

Acknowledgment. This research was supported by NIH (Grants GM47467, DK020299, and EB002026). We thank Drs. Jeffrey Hoch and Werner Maas for stimulating discussions on the topic of this publication.

JA068541X

(45) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.