Spectroscopy by Integration of Frequency and Time Domain Information for Fast Acquisition of High-Resolution Dark Spectra

ARTICLE *in* JOURNAL OF THE AMERICAN CHEMICAL SOCIETY · MAY 2009

Impact Factor: 12.11 · DOI: 10.1021/ja807893k · Source: PubMed

3 AUTHORS, INCLUDING:

Yoh Matsuki
Osaka University
39 PUBLICATIONS **691** CITATIONS

SEE PROFILE

Published in final edited form as:
J Am Chem Soc. 2009 April 8; 131(13): 4648–4656. doi:10.1021/ja807893k.

Spectroscopy by Integration of Frequency and Time Domain Information (SIFT) for Fast Acquisition of High Resolution Dark Spectra

Yoh Matsuki[1,3], **Matthew T. Eddy**[2,3], and **Judith Herzfeld**[1,*]

[1]Department of Chemistry, Brandeis University, Waltham, MA 02454, USA

[2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[3]Francis Bitter Magnet Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Abstract

A simple and effective method, SIFT (*S*pectroscopy by *I*ntegrating *F*requency and *T*ime domain information) is introduced for processing non-uniformly sampled multidimensional NMR data. Applying the computationally efficient Gerchberg-Papoulis (G-P) algorithm, used previously in picture processing and medical imaging, SIFT supplements data at non-uniform points in the time domain with the information carried by known “dark” points (i.e. empty regions) in the frequency domain. We demonstrate that this rapid integration not only removes the severe pseudo-noise characteristic of the Fourier transforms of non-uniformly sampled data, but also provides a robust procedure for using frequency information to replace time measurements. The latter can be used to avoid unnecessary sampling in sampling-limited experiments and the former can be used to take advantage of the ability of non-uniformly sampled data to minimize trade-offs between the signal-to-noise ratio and the resolution in sensitivity-limited experiments. Processing 2D and 3D datasets takes about 0.1 and 2 min, respectively, on a personal computer. With these several attractive features, SIFT offers a novel, model-independent, flexible, and user-friendly tool for efficient and accurate processing of multidimensional NMR data.

Introduction

With increasing use of multi-dimensional NMR experiments to resolve the signals of large molecules, there has been growing interest in improving the efficiency of data acquisition [1, 2]. The novel approach described here joins the following two observations.

a. Two very different experimental regimes have similar needs. In the sampling-limited situation, commonly encountered in solution NMR, more points are required for resolution than for signal-to-noise (S/N). Here one wants to take only as many points late in the free induction decay (FID) as needed for resolution. On the other hand, in the sensitivity-limited situation, commonly encountered in solid state NMR, more scans are required for S/N than for resolution. Here one wants to take as many points as possible early in the FID, where the signal is stronger. Thus, non-uniform sampling (NUS), with more points taken early in the FID and just enough taken late in the FID, is desirable in both cases and has received a great deal of attention. The problem is

*Corresponding Author: Judith Herzfeld, E-mail: herzfeld@brandeis.edu, (781) 736-2538.

how to process NUS data without sacrificing the efficiency of the fast Fourier Transform (FFT) and without introducing biasing models and assumptions.

b. FT-NMR made a bargain with the devil in effectively sampling all of frequency space equally when much of it is empty of signals (i.e., "dark"). However, since there is a linear relationship between signals in the time and frequency domains, known information in the frequency domain can in principle replace information from the time domain. Moreover, whereas some schemes for processing NMR data obtained by NUS make use of spectral darkness in an implicit, laborious and model-dependent fashion, it should be possible to use the darkness in an explicit, efficient and model-free manner.

Previously reported approaches to NUS include reduced dimensionality (RD) [3,4] or GFT [5–7], projection reconstruction [8–10], covariance NMR [11–14], filter diagonalization [15], MaxEnt [16–19], multi-dimension decomposition (MDD) [20–24], and non-uniform Fourier transformation (NU-FT) [25–30]. A common issue is the severe pseudo-noise that corresponds to the Fourier transform of steps in the sampling function. In this regard, statistically random NUS seems generally preferable over the radial NUS of the type used for RD because incoherence in the sampling pattern helps to suppress the pseudo-noise [25,31]. Therefore, increasing attention has been drawn to MaxEnt, MDD and NU-FT, which can all process random NUS.

Of these approaches, the conceptually most straightforward is NU-FT, which simply Fourier transforms NUS data without using FFT. However, the residual pseudo-noise in the resulting spectrum is untreated, and often requires a time-consuming post-processing cleaning procedure in the frequency domain [27]. MaxEnt and MDD can actively reduce the pseudo-noise, but they also do so at significant computational expense, especially in the case of MDD. Moreover, the quality of the reconstruction in each case is dependent on several adjustable parameters, with susceptibility to artifacts.[18,20] A more efficient and model-independent procedure is highly desirable.

In general, NMR spectra are naturally relatively "dark", meaning that they include regions where no signals arise (compared, for example, with an image of an object that yields a continuum of pixel intensities). In fact, MDD and MaxEnt both rely on this darkness. [20 32] The darkness in 1D NMR spectra derives from the discrete nature of chemical groups. Additional dimensions, generally increase the darkness, and this is especially so for some short-range correlation experiments, due to the intrinsic correlations between the chemical shifts of directly bonded nuclei. For example, in a 2D ^{1}H-^{13}C HSQC spectrum, the signals tend to cluster on a diagonal, leaving large triangular regions of spectrum bare. And in magic-angle spinning solid-state NMR experiments, the dwell time for the indirect dimension is often rotor-synchronized (to simplify interpretation and gain sensitivity by folding the spinning sidebands onto the corresponding main peak), resulting in a bandwidth that often exceeds the extent of the signal distribution. Known zeroes in the frequency domain constitute concrete and unambiguous spectral information. Furthermore, due to the linear relationship between time and frequency domain intensities, every frequency point with known intensity obviates measurement at one time point. Clearly it is desirable to have a processing scheme that makes systematic use of the information content in known spectral darkness.

In this context, the approach of Gerchberg and Papoulis is promising [33,34]. Extensively used in picture processing and medical imaging [35–37], the Gerchberg-Papoulis (G-P) algorithm iterates alternating Fourier transforms and inverse Fourier transforms, with frequency domain priors (dark points) and time domain data each reimposed in each cycle until convergence is achieved. In this way, information across the domains is integrated without any biasing model or parameters, and the frequency dark points can replace an equal number of time data omitted

by NUS (or simply deleted due to corruption by probearcing, etc). Furthermore, because the time data can be defined on a regular grid, FFT algorithms can be used to enable fast processing. All that is needed is to identify the frequency dark points in advance.

There are various ways to locate the frequency dark points before actually acquiring a full *n*D dataset. The most straightforward and comprehensive is prior experience with similar types of spectra for similar samples. For novel experiments and/or samples, a less comprehensive set of zeroes can be identified conservatively in two ways. The simpler is to scout for zeroes in a low resolution *n*D version of the same experiment. An alternative is to locate empty regions in the (*n*–1)D spectra corresponding to the projections in each of the indirect dimensions in the full *n*D experiment. Of course, it is also possible to combine the two approaches, with low-resolution (*n*–1)D spectra. Dark regions can also be added to spectra by expanding the spectral width. As demonstrated below, even in the least favorable case of an otherwise bright spectrum, the cost of oversampling is more than repaid by the added dark points because the flexibility gained in the choice of time points can improve the S/N without degrading resolution.

Instead of locating the dark points based on *knowledge* provided by experience or scout data or oversampling, dark points can also be *assumed* below a user-defined threshold [38]. Here the definition of the "smallest meaningful spectral intensity" becomes subjective, and caution is required in restoring a spectrum with small signals (e.g., in a NOESY experiment). Moreover, because one does not know the number of dark points in advance, it is difficult to rationally plan data acquisition. The benefits and drawbacks of using *known* vs. *assumed* dark points in processing spectra are demonstrated below.

This article describes the first application of the G-P type algorithm to combine time and frequency information in multi-dimensional NMR. We call the method "*S*pectroscopy by *I*ntegrating *F*requency and *T*ime domain information" (SIFT). To show the power of frequency dark points, we choose the worst case of a bright spectrum in which the only frequency dark points are those produced by oversampling as mentioned above. Specifically, we use a 2D ^{1}H-^{15}N HSQC spectrum of a uniformly ^{15}N labeled 56-residue protein, GB1. We also use synthetic 1D data for further tests. SIFT is shown to be highly user friendly. It is also rapid, typically converging in just 0.1 and 2 minutes for 2D and 3D datasets, respectively. In addition, SIFT has no adjustable parameters and is resistant to misuse, with tell-tale behavior when the intensity at a given frequency is mistakenly set to zero. Furthermore, due to the linearity of the procedure, SIFT quantifies spectral intensities very well. Finally, SIFT is also robust in that the results degrade slowly and smoothly when the number of dark frequency points is not enough to fully replace the omitted time points. These favorable features will make NUS, for efficient acquisition of multidimensional NMR data, accessible to non-expert users. To facilitate adoption, a suite of MATLAB macros implementing SIFT for processing 2D and 3D dataset are available at http://www.brandeis.edu/~herzfeld/SIFT.

Methods

SIFT procedure

The SIFT procedure is described below for a 2D experiment (i.e., one indirect dimension). However, generalization to higher dimensions is straightforward. The heart of SIFTing is the cycle shown in Figure 1. A uniform grid of time points is initially filled with zeroes, corresponding to $f_0(t)$. Zeroes are reasonable starting values due to the oscillating nature of the FID around zero. In the first step of the cycle (top of Figure 1), $f_0(t)$ is obtained by replacing points in $\tilde{f}_0(t)$ according to the experimentally acquired time data. FFT of $f_0(t)$ then generates the frequency spectrum $F_0(\omega)$. At this point, the S/N ratio is calculated for several F1 slices containing NMR signals to provide a metric for convergence of the process. In the third step of the cycle (bottom of Figure 1), $\tilde{F}_0(\omega)$ is obtained by replacing points in $F_0(\omega)$ with zeroes

according to known dark frequencies in the spectrum. Finally, inverse FFT of $\tilde{F}_0(\omega)$ produces $\tilde{f}_1(t)$ and the cycle can begin anew.

Our implementation of this cycle in MATLAB requires input of files that contain (1) the sampling schedule, (2) the corresponding time domain NUS data, and (3) specification of the dark frequency points (or, for application of the thresholding method, a threshold level below which all spectral intensities will be assumed to be zero). The number of cycles can either be preset or the macro can be made to automatically terminate when the S/N does not improve between cycles more than a predefined small amount. The final SIFTed FID, $f_n(t)$, is Fourier transformed and phased as usual, and some additional macros are provided for displaying and inspecting 1D and 2D slices. The spectrum can also be output to Sparky format for further inspection.

Data sets

To form test datasets with varying distributions of t1 samples, data were extracted from a large, uniform master dataset by eliminating t2-FIDs along t1. In the full master dataset, 128 t1 samples were recorded at 250 us intervals (SW = 4 kHz), producing the spectrum shown in Figure 2. In addition, a uniform dataset with half the bandwidth (SW = 2 kHz) was formed by every second t1 sample of the full dataset.

The number of data points, $i_{NUS} < 128$, distinguishes each NUS dataset. Thus, the number of unsampled points on the time grid, or missing data M, is given by $M = 128 - i_{NUS}$. We define *critical sampling* as the case where the number of known dark frequency points D equals the number of missing time data points M ($D = M$). In the absence of noise, this number of dark points would completely determine the spectrum. In contrast, in *sub-critical sampling* the number of dark frequency points does not fully replace the missing time data ($D<M$).

To test the reconstruction of data with a very high dynamic range, or a base-line roll, or an aliased signal, 1D FIDs were synthesized with the same bandwidths, and with 50 Hz-wide Gaussian signals in the same region, and as in the F1 (15N) dimension of the experimental data. The signal intensity was set arbitrarily unless otherwise noted.

Non-uniform sampling schedules

We use “on-grid” random NUS in which a specified number of on-grid samples are chosen quasi-randomly with a Gaussian probability density. To form the on-grid samples, we first generated corresponding off-grid samples using a program available at http://nmr700.chem.uw.edu.pl/ [27]. For the Gaussian sampling probability density, $\exp(-t^2/\sigma^2)$, we chose σ such that the sampling density is halved at the middle of the interferogram and the sample with the longest t1 falls close to the maximum evolution time t_{1max} = 32 ms in our full dataset. Non-uniform samples with significantly larger or smaller σ lead to lower S/N or lower resolution, respectively, in the processed spectrum.

To conform to the grid, we increase the evolution time of each randomly generated point just enough to coincide with the first unoccupied grid point. No sample is lost in this process. An example of an on-grid random NUS schedule generated in this fashion is shown in Figure 3 for 48 samples on our 128-point grid. The probability density closely resembles the Gaussian distribution when less than ~60% of the time points are sampled. With more sampling, the density tends to have a long flat region early in the FID, and then drop relatively steeply, resembling the form $1/(1+[t/(at_{max})]^b)$ with a = 0.6 – 2.0 and b = 4 – 6, where the values of a and b are greater for larger numbers of samples. Of course, the distribution eventually becomes completely flat with a fully occupied grid. Despite this transformation from sampling on a continuum to sampling on a grid, the present NUS shares the fundamental characteristic

with any other common non-uniform distribution of being progressively sparser toward the end of the FID.

Estimation of the S/N

NUS sampling results in pseudo-noise that depends on the positions of missing time points and the intensities that they would have if acquired. This is most readily understood for one-dimensional data, although the behavior generalizes to multidimensional data.

For one dimensional data, the discrete NUS FID, f_0, is related to the discrete master FID, f, by

$$f_0(n/\sigma)=f(n/\sigma)-\sum_{k=1}^{M} f(k/\sigma)\delta_{nk} \tag{1}$$

where $1/\sigma$ is the dwell time determined by the spectral bandwidth, σ, the summation runs over the M missing time data points in the NUS dataset and δ_{nk} is the Kronecker delta function. Because $\mathrm{sinc}(z) = 0$ for integer $z\neq 0$ and $\mathrm{sinc}(0) = 1$, the sinc function forms an orthogonal basis for a discrete FID and we can rewrite Eq. 1 in the continuous form

$$f_0(t)=f(t)-\sum_{k=1}^{M} f(\tau_k)\mathrm{sinc}(\sigma(t-\tau_k)) \tag{2}$$

The Fourier transform, F_0, is then given by

$$F_0(\omega)=F(\omega)-\frac{1}{\sigma}\sum_{k=1}^{M} f(\tau_k)\exp(i\omega\tau_k)P(\omega/\sigma) \tag{3}$$

where the first term is the spectrum that would be given by the full data set and Πx is the boxcar function (1 for $|x| \leq 1$ and 0 otherwise). The coherent oscillations in the second term of Eq. (3) represent the pseudo-noise due to NUS. It is seen that the amplitude of the component pseudo-noise is determined by the time signal intensity $f(\tau_k)$ that would be observed if it were acquired. Thus, in a spectrum empty of signals, there should be no pseudo-noise visible above the thermal noise. More importantly, in a spectrum with high dynamic range, the average pseudo-noise level will be determined primarily by the intensity of large peaks and will mask smaller peaks.

The absence of pseudo-noise in F1 slices empty of signals is illustrated in Figure 4 where the pseudo-noise appears as noise ridges running parallel to the indirect (F1) axis. Due to this localized nature of the pseudo-noise, it is essential to select and register the signal-containing slices to monitor the S/N during the SIFT cycle. The noise level is conveniently measured in a region of interest by the median of the absolute spectral heights taken at relatively small number of (say 500) randomly selected points out of a reduced 2D matrix consisting of a bundle of the F1 slices selected as above. This works well because each slice in the multidimensional spectrum is normally sparse.

Sample preparation

E. coli BL21 (DE3) competent cells were transformed by the T2Q GB1 plasmid (kindly provided by Angela Gronenborn). 2 mL of Luria-Bertani (LB) medium containing 75 ug/mL carbenicillin were innoculated and grown at 37°C for 6–8 hours. A 25 uL aliquot of this culture was used to innoculate 10 mL of M9 minimal media containing 1g/L $^{15}NH_4Cl$ (Cambridge

Isotope Labs) and 8 g/L natural abundance glucose. After growth overnight at 37°C, the 10 mL were transferred into 1L of M9 media of the same composition and grown at 37°C until the cell density reached an OD600 of 0.8. Expression of GB1 was induced with 500 uM isopropyl B-D-thiogalactoside for 3–3.5 hours. Cells were then sedimented at 5000 g for 30 minutes. The cell pellet was homogenized by tip sonication in phosphate buffered saline (200 mM NaCl, 50 mM KH_2PO_4/K_2HPO_4, pH 7), heated to 80°C for 5 minutes, chilled on ice for 15 minutes, and centrifuged at 30,000 g for 30 minutes at 4°C. The resulting supernatant was concentrated using Amicon Ultra-15 3,500 MWCO devices and purified at 4°C by gel filtration chromatography (Sephacryl S-100). Peak fractions were pooled and reconcentrated with Amicon Ultra-15 3,500 MWCO devices. The concentrated protein was then dialyzed three times against 4 L of fresh 50 mM sodium phosphate buffer (pH 5.6) [39]. The final sample concentration was adjusted to approximately 1 mM.

NMR measurements

The master 2D ^1H-^{15}N HSQC spectrum was recorded at 278 K with a gradientenhanced scheme [40] at 591 MHz (^1H Lamor frequency) using a home-built console and software (D. Ruben) and a Z-SPEC 5mm triple-resonance IDTG590-5 probe (NALORAC Co., CA). Four scans were averaged at the recycle delay of 2 sec. The ^{15}N bandwidth was 67 ppm (3984 Hz) sampled with 128 points, and the ^1H bandwidth was 13.6 ppm (8013 Hz), sampled with 1024 complex points. The total acquisition time was 34 min.

Results and Discussions

SIFT is efficient and faithful

As a worst-case example, we demonstrate SIFT using only frequency dark points produced by oversampling in the indirect dimension (Figure 2). The bandwidth for the ^{15}N (F1) dimension was 4 kHz, although the minimum Nyquist bandwidth for the amides in the ^{15}N HSQC is 2 kHz. The populated region in the ^{15}N dimension spans 60 frequency points between 101.2 and 132.0 ppm, i.e. 68 dark frequency points at the spectral edges are available for SIFT, $D = 68$.

Although we will use only the frequency dark points located at the spectral edges, each F1 slice also has varying numbers of dark points between peaks that can be used in spectral processing whenever they are located either by thresholding or by a scouting experiment.

Figure 5 illustrates the effect of SIFT cycles for a representative t1/F1 column at F2 = 8.34 ppm. The top left panel in Figure 5 shows the progress of the S/N (solid lines) and the RMSD between the restored and the original complete interferogram along the indirect dimension (broken lines). Hereafter, we will refer to the interferogram for the indirect evolution as simply the FID. The S/N and RMSD improve steeply in the early iterations of the cycle and quickly settle. In general, the convergence is slower, and the improvement in S/N and RMSD is smaller for datasets with fewer samples. For example, for the datasets with $i_{NUS} = 64$ (black lines) and = 48 (gray lines), the S/N leveled off at 19 SIFT cycles (~ 6 seconds of computation) and 42 SIFT cycles (~ 12 seconds of computation), respectively. We note that SIFT has been proven to converge [41] and there are no parameters to tweak for avoiding local minima etc. It is also important to note that the progress in the S/N and the RMSD occurs together. However, although the RMSD serves here as a direct measure of the restoration of the master dataset, such a measure is not available in a *de novo* application. On the other hand, the S/N can be easily calculated at every cycle, and used as a criterion for terminating processing.

After the SIFT iterations, the SIFTed FID may be processed in the usual manner, as if all of the time points had actually been recorded. The middle and bottom rows of Figure 5 show the representative F1 slice before and after SIFTing. The spectra before SIFTing (left) show the

untreated pseudo-noise characteristic of even random NUS; these spectra correspond to the ones given by NU-FT [26]. Close to or above the critical condition ($D \gtrsim M$, middle row of Figure 5), the noise level in the SIFTed spectrum (middle right) is comparable to that of the full master dataset (top right). At the sub-critical condition, ($D < M$, bottom row of Figure 5), the result is not as good, with greater residual pseudo-noise. However, in spite of the gradual increase in residual pseudo-noise with decreasing numbers of samples, neither spurious peaks nor inaccurate peak shifts are obtained even at deeply sub-critical sampling.

Figure 6 compares a crowded portion of the master 2D spectrum (top) with spectra based on half as many (middle row) and one quarter as many (bottom row) t1 points. The middle column shows that SIFTed NUS datasets preserve the resolution seen in the master spectrum, without reducing the spectral width, even when 75% of data points are missing ($i_{NUS} = 32$). Figure 7 shows that SIFT sustains resolution with decreasing numbers of t1 samples without sacrificing much S/N. The NUS datasets are processed either with SIFT (circles), NU-FT (triangles) or the iterative thresholding (rectangles). While the S/N in the spectra processed by NU-FT (triangles) decreases steeply with the number of samples, due to the untreated pseudo-noise, the S/N in the spectra processed by SIFT decreases much more slowly (circles). This clearly illustrates the power of the information carried by the dark frequency points integrated into the time domain by SIFT.

The power of frequency information is also seen in comparing the SIFT results with the results from uniform sampling with 4 kHz (solid gray line) and 2 kHz (dashed gray line) bandwidths, where points are reduced by truncation (as for the spectra in the left and right columns in Figure 6). We see that the S/N of SIFT-processed spectra (circles) decreases more slowly than that obtained by uniform sampling. This translates into faster data acquisition. For example, the S/N ratio of the SIFT-processed spectrum with $i_{NUS} = 64$ is ~70% higher than that with uniform 64 samples taken at the conventional 2 kHz bandwidth to preserve resolution (broken gray line in Figure 7). Achieving a 70% increase in the S/N ratio would otherwise require $(1.7)^2 = 3$ times as much signal averaging. Furthermore SIFT affords this increased sensitivity without deteriorating the resolution or tightening the bandwidth. Thus, with NUS and SIFT, there is no reason to fold a spectrum and introduce potential complications in the spectral analysis, except in some solid-state experiments with significant spinning side bands.

The rectangles in Figure 7 show that iterative thresholding can boost the S/N of the processed spectrum beyond that realized by SIFT because it can exploit the frequency dark points that exist between peaks. Although iterative thresholding is thus an attractive approach to data reconstruction, its downsides include the difficulty of knowing in advance how sparse the NUS can be and a greater chance of losing small signals during processing (vide infra).

An attractive feature of SIFT is its high fidelity in spectral quantification. Figure 6 shows that the peak shifts are identical to the true shifts under both critical and sub-critical conditions, within 0.013 ppm digital resolution in the ^{1}H dimension and 0.065 ppm digital resolution in the ^{15}N dimension. Figure 8 compares the peak intensities observed in uniformly and non-uniformly sampled data. Close to the critical condition, the intensities in the spectra processed by SIFT (circles) and by iterative thresholding (crosses) accurately match the “true” intensities observed in the full uniformly sampled dataset. At sub-critical conditions, the intensities in the SIFTed spectrum became less accurate in the sense that they deviate from the diagonal reference line. Nevertheless, the trend remains highly linear, indicating that the relative peak intensities remain accurate. On the other hand, iterative thresholding produces some serious outliers of the peak intensities.

The fidelity of the SIFT-processing was tested further for a higher dynamic range using synthetic 1D data. Figure 9(a) demonstrates the exquisite accuracy and linearity of the SIFT-

processed peak intensities in a spectrum where signals vary over two orders of magnitude. This validates SIFT for NOESY-type experiments.

When the dynamic range of the spectrum is high, iterative thresholding has a greater chance than SIFT to lose a small signal by *assuming* zero intensity for the frequency points below a user-defined threshold. This is illustrated in Figure 9 (b)–(f). Because the amplitudes of large signals determine the average amplitude of the pseudo-noise seen in the same slice (see Methods), if one rationally sets the threshold above the pseudo-noise level (the dashed line at 40 in Figure 9c), it is actually well above the amplitude of the smaller signals. As a result, iterative thresholding loses the peak, as seen in Figure 9e, while SIFT, which uses only *known* dark points, suppresses the pseudo-noise without losing the small peak, as seen in Figure 9d. The faithful rendering of the small peak close to the noise level illustrates the robustness of SIFT for noisy data. However, iterative thresholding can profitably be used to further improve a SIFT-processed spectrum. When the pseudo-noise level has been suppressed by prior treatment with SIFT, the threshold can be reduced (the dashed line at 12 in Figure 9d). Figure 9f shows the result of this two-step procedure, with SIFT followed by thresholding, where the S/N ratio was significantly enhanced while the small peaks are conserved. Thus, it is always advisable to treat data with SIFT before thresholding.

The fidelity of SIFT also extends to spectra with signals of different signs. Figure 10 shows the SIFT reconstruction of data with aliased and non-aliased signals, where the former at ~110 ppm is phase-inverted with respect to the latter. In the post-SIFT spectrum (middle panel), one can see that the pseudo-noise is almost perfectly removed. Thus, SIFT is generally applicable to a data with peaks of positive and negative intensity, such as that may arise in experiments using constant time evolution.

We have also examined the processing of a uniformly sampled short record, i.e. a truncated dataset. A number of previous reports have been interested in extrapolation of a truncated data for super-resolution. In such applications, slow convergence has been noted as a major downside of the G-P algorithm [36]. We show here that NUS greatly accelerates the convergence. In the Fourier transform of a truncated FID, the pseudo-noise manifests as Gibbs wiggles at the skirt of each peak. Because the frequency dark points produced by oversampling outside the bright region of the spectrum cannot treat this localized noise, iterative thresholding was used for all the reconstruction shown in Figure 11. As shown at the upper left, the S/N converged at 20 cycles for NUS (solid line) but only at 1,000 cycles for the truncated data (dashed line). Slow convergence for a dataset with long stretches of contiguous missing data points has been noted previously [41]. And our patience is not well rewarded. In addition to slower convergence, the truncated data (Figure 11, bottom right) yields noticeably distorted peak shapes: broader lines and remaining wiggles. On the other hand, the NUS dataset yields a nearly perfect result (Figure 11, bottom left). Therefore, NUS is always preferable to zero-filling a truncated uniform record.

SIFT is tolerant and candid

SIFT is not only efficient and faithful but also robust against experimental artifacts and user errors. In Figure 12, one can see by comparing the left and right middle panels (lower trace) that a base-line roll coexisting with a signal does not compromise the pseudo-noise removal by SIFT. This is because the SIFT process is linear just like the Fourier transform, i.e. SIFT treats each signal component independently. Data with a signal and base-line roll can be understood as a superposition of two components: a very short FID decaying to zero in the initial several time points, and a much longer one for the signal. Due to the linearity of SIFT, restoration of the signal is independent of the baseline roll. More actively, one could use SIFT to correct the base-line roll by removing the two offending time data points and letting SIFT fill them in from the frequency domain information. As shown by the upper trace in the middle

right panel of Figure 12, the baseline roll is removed by SIFT, along with the pseudo-noise. This approach is also suitable for restoring other corrupted time data points due, e.g., to RF-arcing.

Although there are no adjustable parameters in SIFT, one possible source of user-dependent error is the mistaken specification of dark frequency points. While the contiguous dark points at the spectral edges used in the current examples are easy to locate, aggressive use of dark points elsewhere requires more care. To test the behavior of SIFT when mistaken frequency zeroes are applied, we deliberately specified an excessively broad dark region on the low field side of our test spectrum. The results for two F1 slices, each with two peaks, are shown on the left and right sides of Figure 13. The arrows show the boundaries of the dark edges in the mistaken SIFT (third row) and the correct SIFT (fourth row). We see that the effects of the mistake are limited and diagnostic. Of the four peaks, only the one that has been incorrectly assigned to the dark region does not gain intensity with SIFT processing. And of the two slices, only the one in which a peak was incorrectly suppressed remains noisy. Thus the independence of SIFT processing for each signal generates tell-tale signs that can be used to locate mistaken assignments of darkness. And it is always easy to modify the assignment and rerun SIFT because the processing is so fast.

Conclusion

We have shown that a Gerchberg-Papoulis type algorithm can be used to integrate frequency and time information in multi-dimensional NMR experiments with great fidelity and efficiency. Unambiguous frequency domain information is available in the form of known dark points. Using them to effectively replace time points, SIFT processing of NUS data can simultaneously achieve significantly higher S/N, resolution and spectral width than is possible via uniform sampling with the same number of data points. The SIFT cycles converge quickly and the results are model-free and robust.

SIFT is also a useful precursor to thresholding. Thresholding can provide access to more dark points than can be readily identified in advance. But it also has the potential to mistake weak points for zeroes. The reduced noise realized by processing first with SIFT enables thresholding with a lower and less dangerous threshold.

The sensitivity (or equivalently time) gain illustrated here with 2D examples is generalizable to higher dimensions and the benefits will be multiplicative. For example, if SIFT enables a three-fold reduction in the number of samples in each indirect dimension of a 3D experiment, the overall sampling requirement is reduced by a factor of nine. An application of SIFT to a 3D NUS dataset is currently underway, and will be reported in a forthcoming publication.

Acknowledgements

We thank Angela Gronenborn for providing the T2Q GB1 plasmid, Christopher Turner for carrying out the ^{15}N HSQC experiment for our master data set, and Mikhail Veshtort, Jeffrey Hoch and Hartmut Oschkinat for helpful discussions. This research was supported by NIH grants EB001035, EB001960 and EB002026. Y.M acknowledges partial financial support from the Naito Foundation.

References

1. Freeman R, Kupce E. J. Biomol. NMR 2003;27:101–113. [PubMed: 12962120]
2. Jaravine V, Ibraghimov I, Orekhov VY. Nat. Methods 2006;3:605–607. [PubMed: 16862134]
3. Bodenhausen G, Ernst RR. J. Magn. Reson 1981;45:367–373.
4. Szyperski T, Yeh DC, Sukumaran DK, Moseley HNB, Montelione GT. Proc. Natl. Acad. Sci. USA 2002;99:8009–8014. [PubMed: 12060747]

5. Atreya HS, Garcia E, Shen Y, Szyperski T. J. Am. Chem. Soc 2007;129:680–692. [PubMed: 17227032]
6. Kim S, Szyperski T. J. Am. Chem. Soc 2003;125:1385–1393. [PubMed: 12553842]
7. Liu GH, Shen Y, Atreya HS, Parish D, Shao Y, Sukumaran DK, Xiao R, Yee A, Lemak A, Bhattacharya A, Acton TA, Arrowsmith CH, Montelione GT, Szyperski T. Proc. Natl. Acad. Sci. USA 2005;102:10487–10492. [PubMed: 16027363]
8. Kupce E, Freeman R. J. Am. Chem. Soc 2004;126:6429–6440. [PubMed: 15149240]
9. Kupce E, Freeman R. J. Biomol. NMR 2004;28:391–395. [PubMed: 14872130]
10. Coggins BE, Venters RA, Zhou P. J. Am. Chem. Soc 2005;127:11562–11563. [PubMed: 16104707]
11. Bruschweiler R, Zhang FL. J. Chem. Phys 2004;120:5253–5260. [PubMed: 15267396]
12. Bruschweiler R. J. Chem. Phys 2004;121:409–414. [PubMed: 15260561]
13. Chen YB, Zhang FL, Bermel W, Bruschweiler R. J. Am. Chem. Soc 2006;128:15564–15565. [PubMed: 17147346]
14. Snyder DA, Xu YQ, Yang DW, Bruschweiler R. J. Am. Chem. Soc 2007;129:14126–14127. [PubMed: 17973386]
15. Hu HT, De Angelis AA, Mandelshtam VA, Shaka AJ. J. Magn. Reson 2000;144:357–366. [PubMed: 10828203]
16. Hyberts SG, Heffron GJ, Tarragona NG, Solanky K, Edmonds KA, Luithardt H, Fejzo J, Chorev M, Aktas H, Colson K, Falchuk KH, Halperin JA, Wagner G. J. Am. Chem. Soc 2007;129:5108–5116. [PubMed: 17388596]
17. Mobli M, Maciejewski MW, Gryk MR, Hoch JC. Nat. Methods 2007;4:467–468. [PubMed: 17538627]
18. Schmieder P, Stern AS, Wagner G, Hoch JC. J. Magn. Reson 1997;125:332–339. [PubMed: 9144266]
19. Stern AS, Li KB, Hoch JC. J. Am. Chem. Soc 2002;124:1982–1993. [PubMed: 11866612]
20. Luan T, Jaravine V, Yee A, Arrowsmith CH, Orekhov VY. J. Biomol. NMR 2005;33:1–14. [PubMed: 16222553]
21. Orekhov VY, Ibraghimov IV, Billeter M. J. Biomol. NMR 2001;20:49–60. [PubMed: 11430755]
22. Tugarinov V, Kay LE, Ibraghimov I, Orekhov VY. J. Am. Chem. Soc 2005;127:2767–2775. [PubMed: 15725035]
23. Orekhov VY, Ibraghimov I, Billeter M. J. Biomol. NMR 2003;27:165–173. [PubMed: 12913413]
24. Jaravine VA, Orekhov VY. J. Am. Chem. Soc 2006;128:13421–13426. [PubMed: 17031954]
25. Kazimierczuk K, Zawadzka A, Kozminski W. J. Magn. Reson 2008;192:123–130. [PubMed: 18308599]
26. Kazimierczuk K, Zawadzka A, Kozminski W, Zhukov I. J. Biomol. NMR 2006;36:157–168. [PubMed: 17031529]
27. Kazimierczuk K, Zawadzka A, Kozminski W, Zhukov I. J. Magn. Reson 2007;188:344–356. [PubMed: 17822933]
28. Kazimierczuk K, Zawadzka A, Kozminski W, Zhukov I. J. Am. Chem. Soc 2008;130:5404–5405. [PubMed: 18376830]
29. Kazimierczuk K, Kozminski W, Zhukov I. J. Magn. Reson 2006;179:323–328. [PubMed: 16488634]
30. Pannetier N, Houben K, Blanchard L, Marion D. J. Magn. Reson 2007;186:142–149. [PubMed: 17293138]
31. Hoch JC, Maciejewski MW, Filipovic B. J. Magn. Reson 2008;193:317–320. [PubMed: 18547850]
32. Donoho DL, Johnstone IM, Hoch JC, Stern AS. J. Royal Statist. Soc. B Methodological 1992;54:41–81.
33. Gerchberg RW. Optica Acta 1974;21:709–720.
34. Papoulis A. IEEE Trans. Circuits and Systems 1975;22:735–742.
35. Plevritis SK, Macovski A. Magn. Reson. Med 1995;34:686–693. [PubMed: 8544688]
36. Plevritis SK, Macovski A. IEEE Trans. Medical Imaging 1995;14:487–497.
37. Stokely EM, Twieg DB. IEEE International Conference, ICIP-94, Image Processing 1994;3:6–10.
38. Stern AS, Donoho DL, Hoch JC. J. Magn. Reson 2007;188:295–300. [PubMed: 17723313]

39. Schmidt HLF, Sperling LJ, Gao YG, Wylie BJ, Boettcher JM, Wilson SR, Rienstra CA. J. Phys. Chem. B 2007;111:14362–14369. [PubMed: 18052145]

40. Kay LE, Keifer P, Saarinen T. J. Am. Chem. Soc 1992;114:10663–10665.

41. Jorge P, Ferreira SG. IEEE Trans. Signal Processing 1994;42:2596–2606.

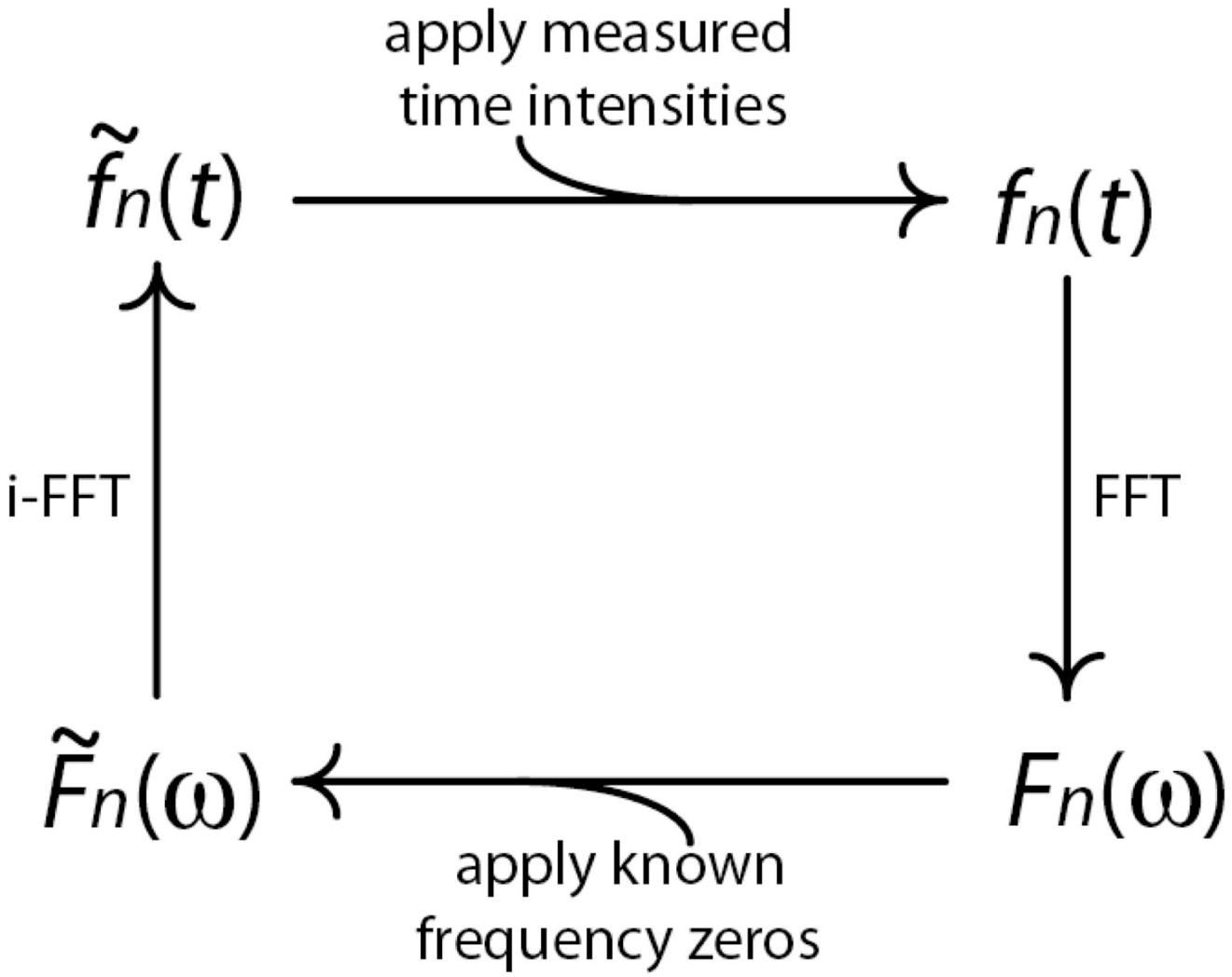

Figure 1.
The SIFT cycle: alternating FFT (right) and inverse-FFT (left) are interleaved with reinstatement of time data (top) and frequency information (bottom).

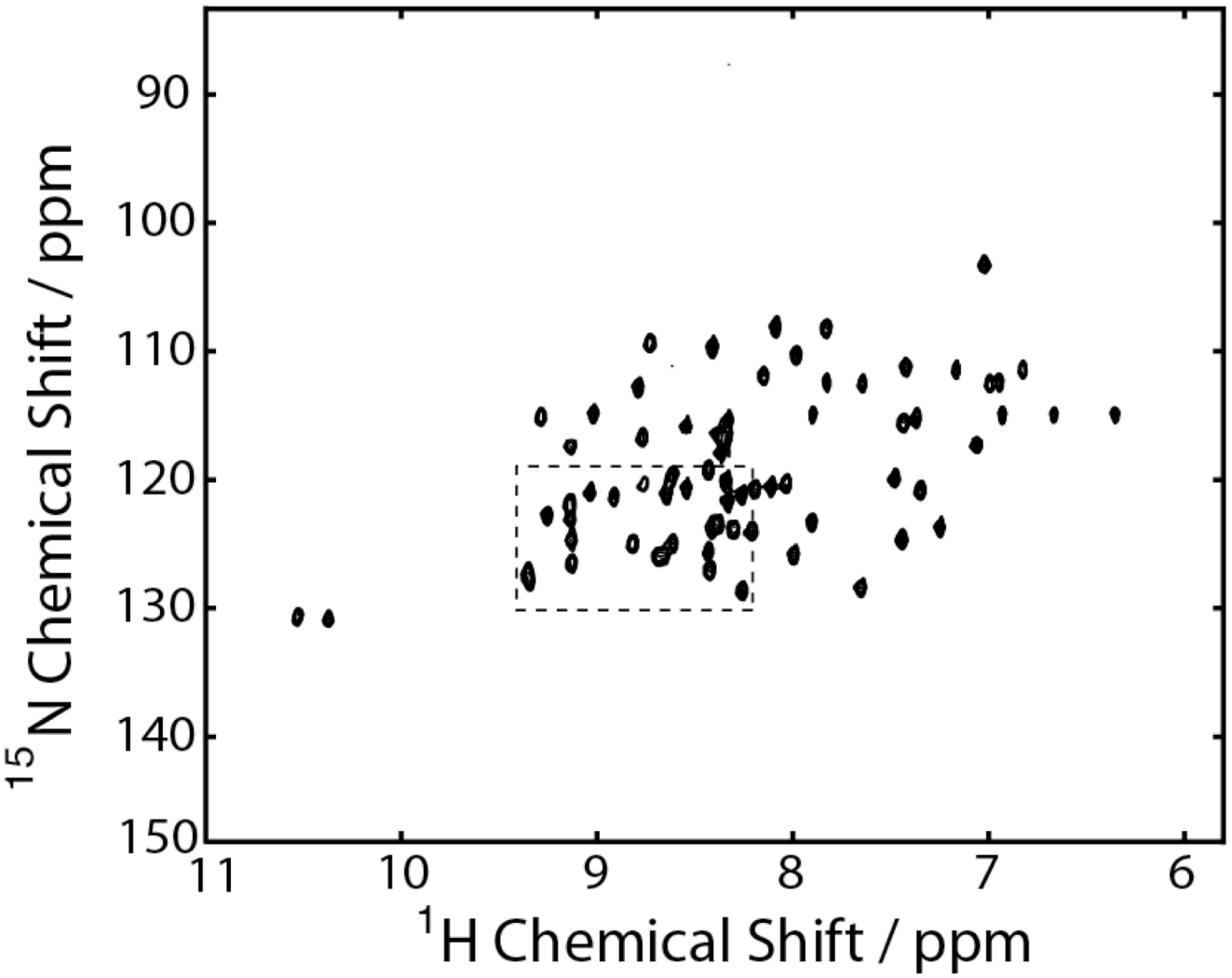

Figure 2.
The ^{15}N HSQC spectrum of uniformly ^{15}N-labeled GB1 derived from the master dataset with 128 linear t1 samples. The entire oversampled ^{15}N (F1) dimension is included. The spectral region enclosed by the dashed rectangle is expanded in Figure 6.

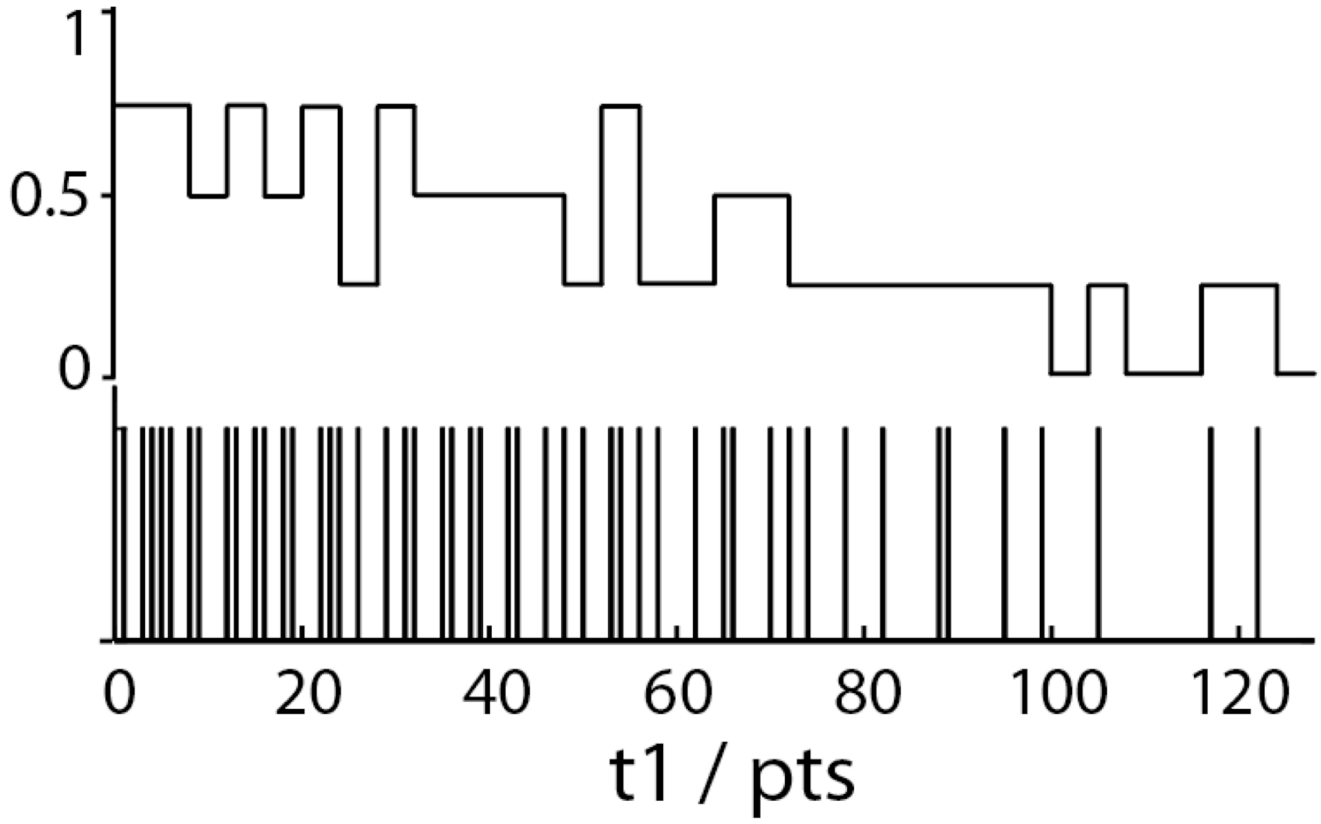

Figure 3.
NUS pattern (bottom) and density (top) for i_{NUS} = 48. The density was calculated using a four-point window.

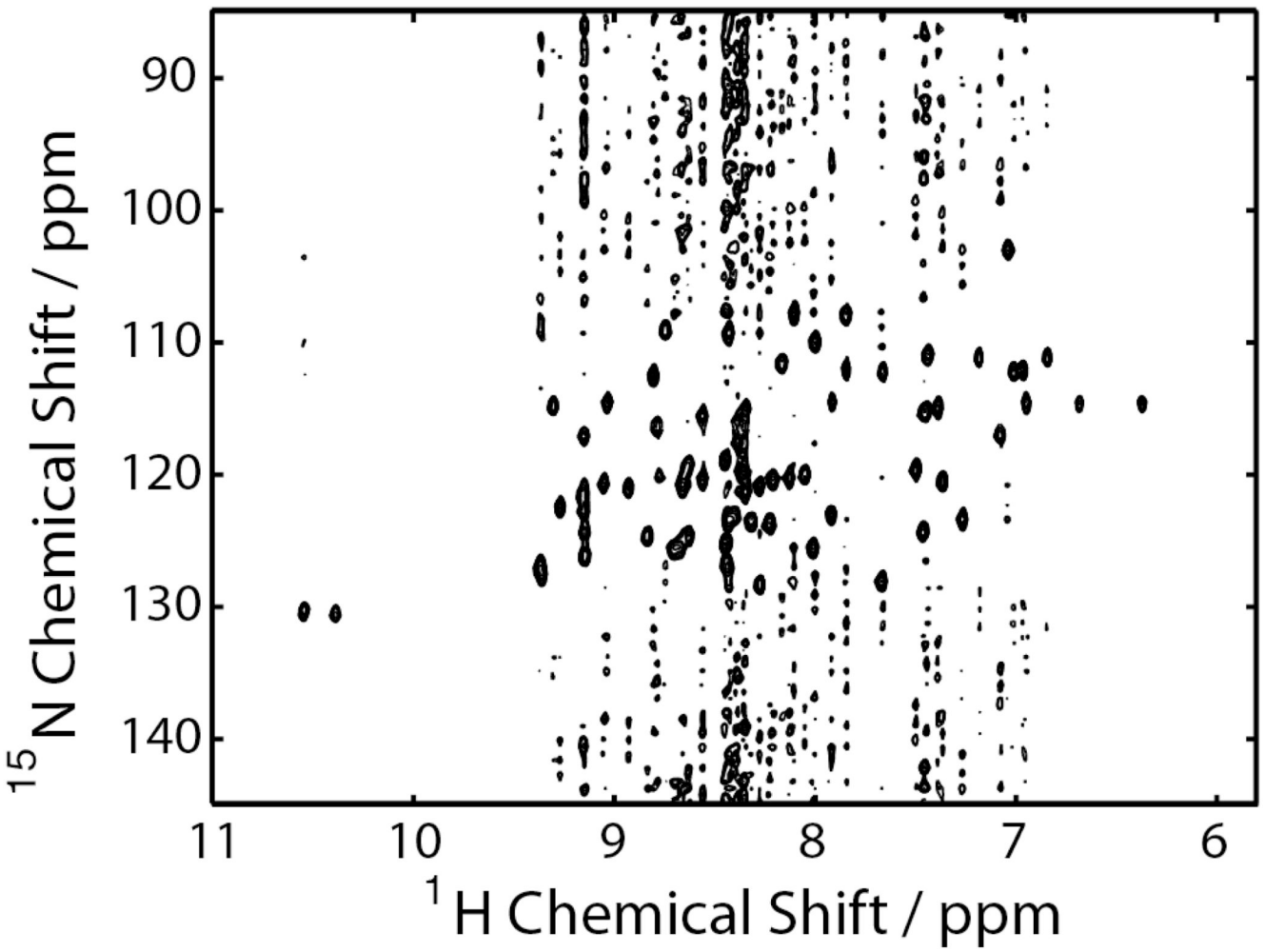

Figure 4.
The same spectral region as in Figure 2, but for NUS prior to SIFT processing.

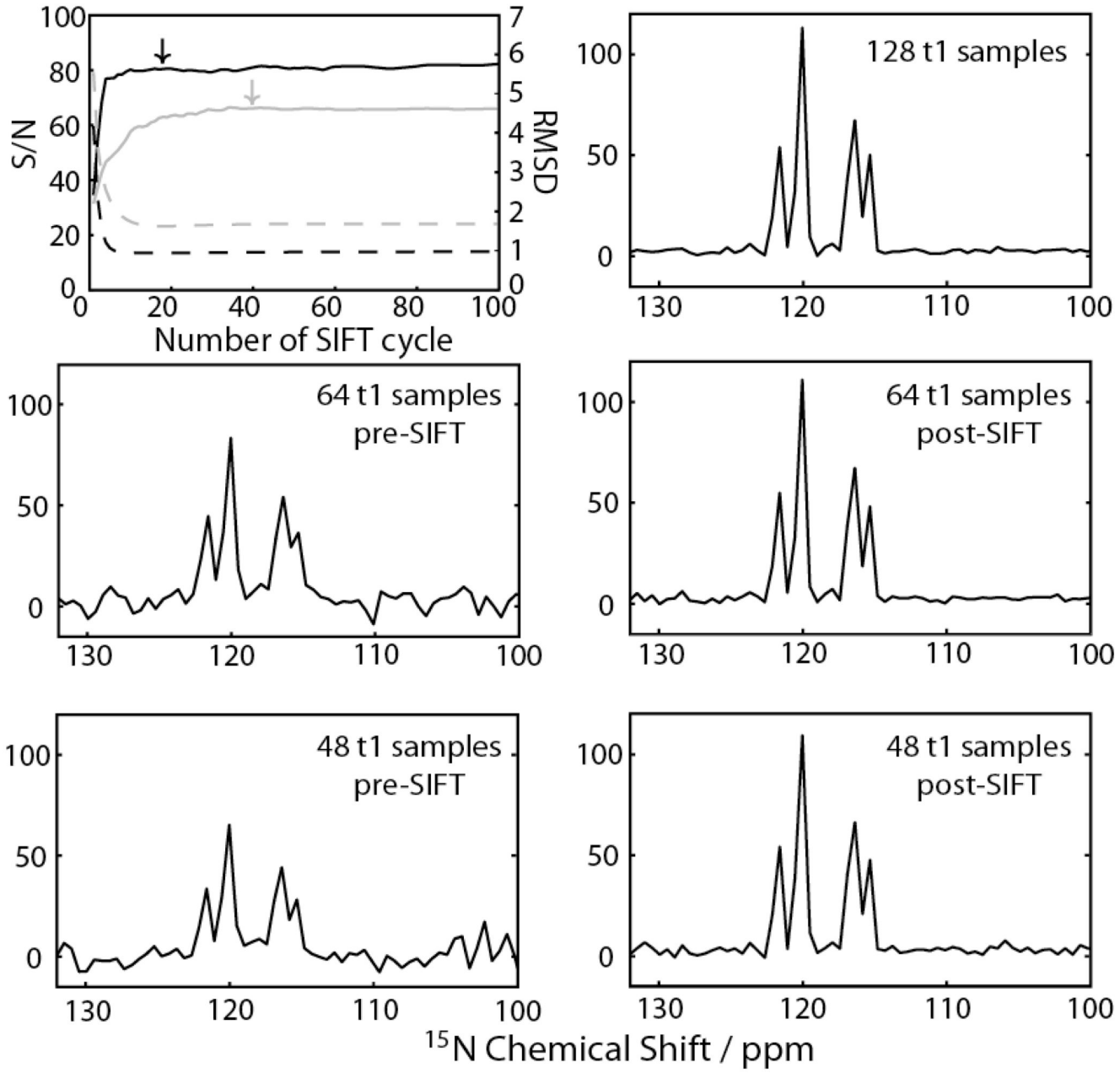

Figure 5.

The effect of SIFT cycles. At the top left, the signal-to-noise ratio (solid lines, left y-axis) and RMSD between the SIFTed NUS FID and the master FID (dashed lines, right y-axis) are shown as a function of the number of SIFT cycles. Black and gray lines plot results for data-sets with i_{NUS} = 64 and = 48, respectively, and arrows show the chosen termination points (19 cycles for i_{NUS} = 64 and 42cycles for i_{NUS} = 48). In the other panels, a representative slice taken at F2 = 8.34 ppm is shown for the master data set (top right), unprocessed NUS data (middle and bottom, left), and SIFTed data (middle and bottom, right).

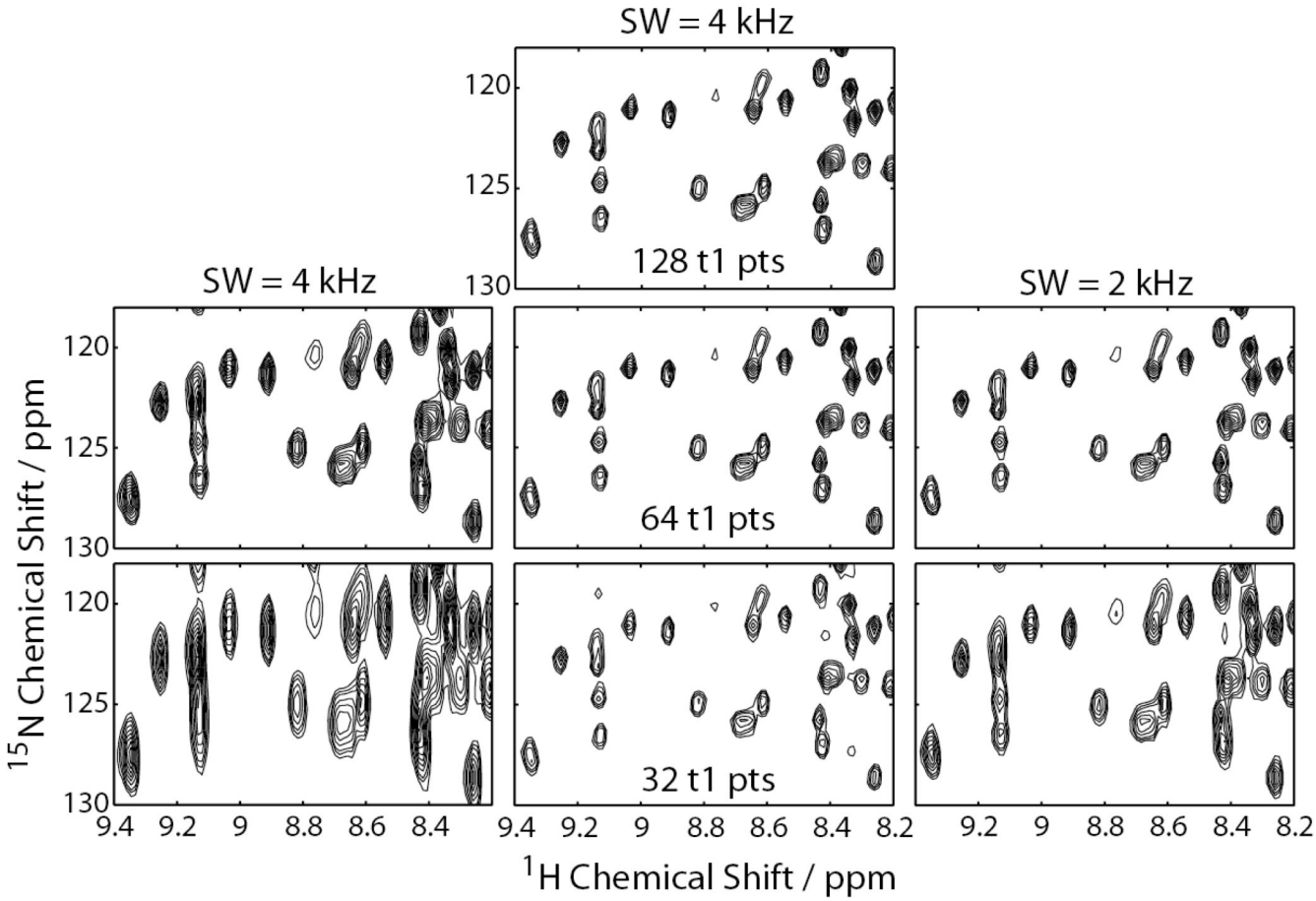

Figure 6.

A crowded region of the 2D spectra obtained with 128 (top row), 64 (middle row) and 32 (bottom row) t1 points, distributed in uniform (left and right) and non-uniform (middle) fashion. The i_{NUS} = 64 data were processed with 15 SIFT cycles and the i_{NUS} = 32 data were processed with 25 SIFT cycles. All datasets were multiplied by a squared-sine weighting function prior to Fourier transformation. The lowest contour line corresponds to 10% of the tallest peak in each spectrum.

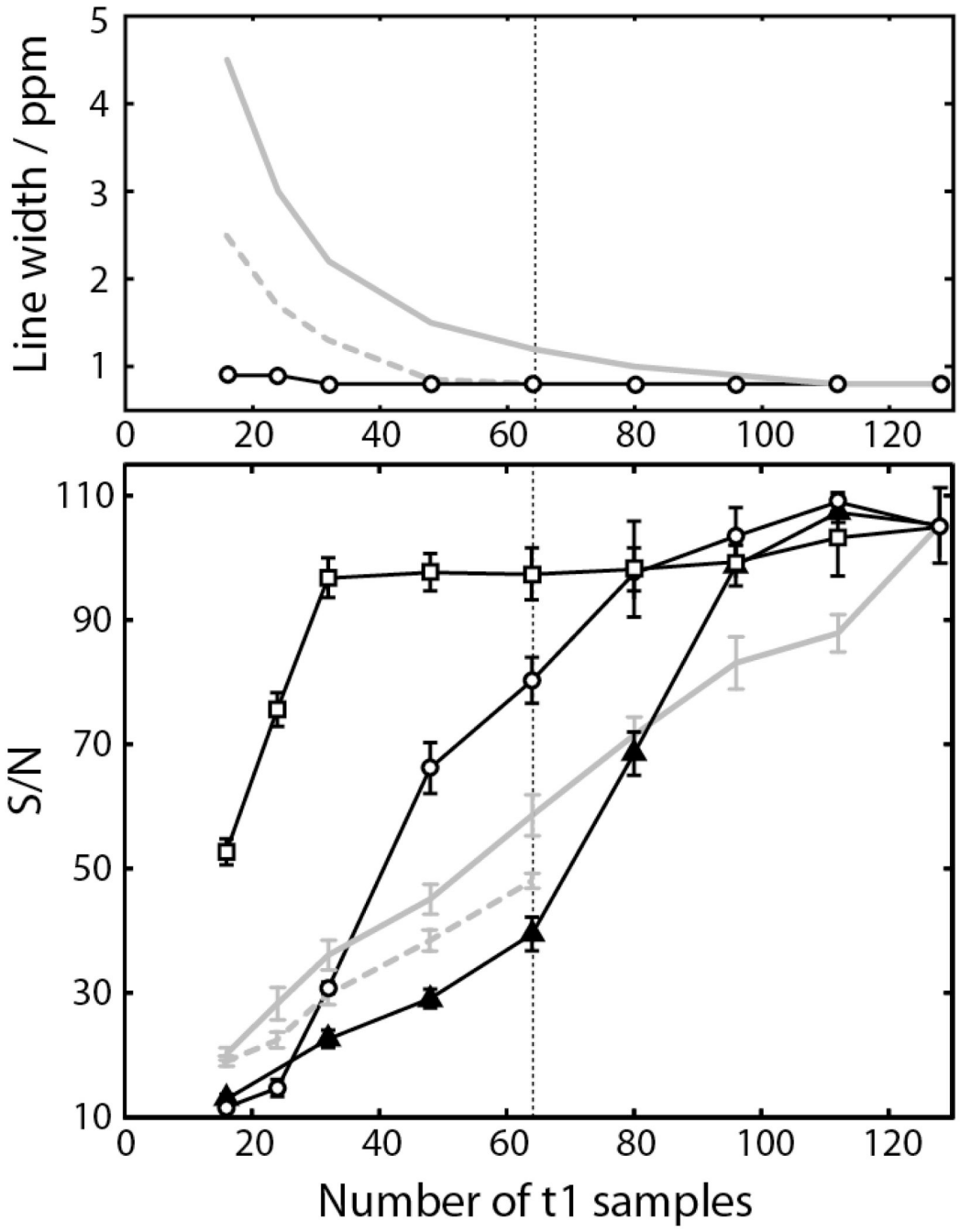

Figure 7.

Resolution (top) and S/N (bottom) observed in the spectra of NUS dataset processed by SIFT (circles), NU-FT [26] (triangles) and the iterative thresholding (squares), and of uniformly sampled data with the bandwidth of 4 (solid gray line) and 2 kHz (dashed gray line) at various number of t1 samples. The rightmost data point corresponds to the full master dataset. The error bars represent five repetitions of processing with different randomly selected points for S/N evaluation (see Methods). The average number of cycles used was: 1.0, 5.0, 10.2, 15.2, 25.4, 25.4, 7.4 and 1.0 for SIFT, and 1.7, 2.0, 5.8, 9.4, 21.0, 78.4, 67.0 and 59.6 for iterative thresholding, on datasets with i_{NUS} = 112, 96, 80, 64, 48, 32, 24 and 16, respectively. The

number of useful cycles is maximal for moderate size datasets. For large datasets, few cycles are needed to get good spectra. For small datasets, cycling offers less gain.

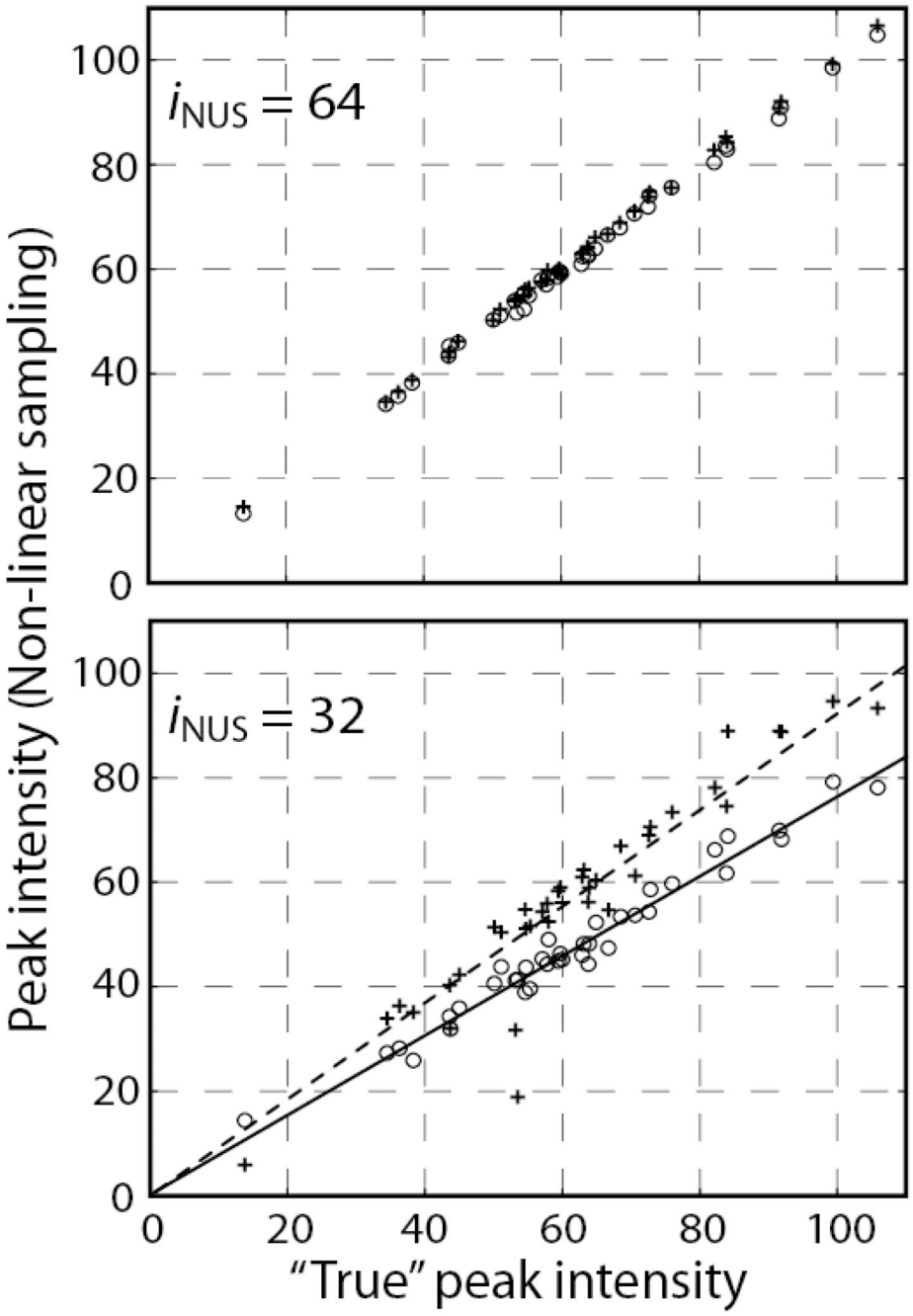

Figure 8.
Comparison of peak intensities observed in the master (abscissa) and non-uniformly sampled (ordinate) datasets. The datasets with i_{NUS} = 64 (top) and i_{NUS} = 32 (bottom) were restored by SIFT (circles) or iterative thresholding (crosses).

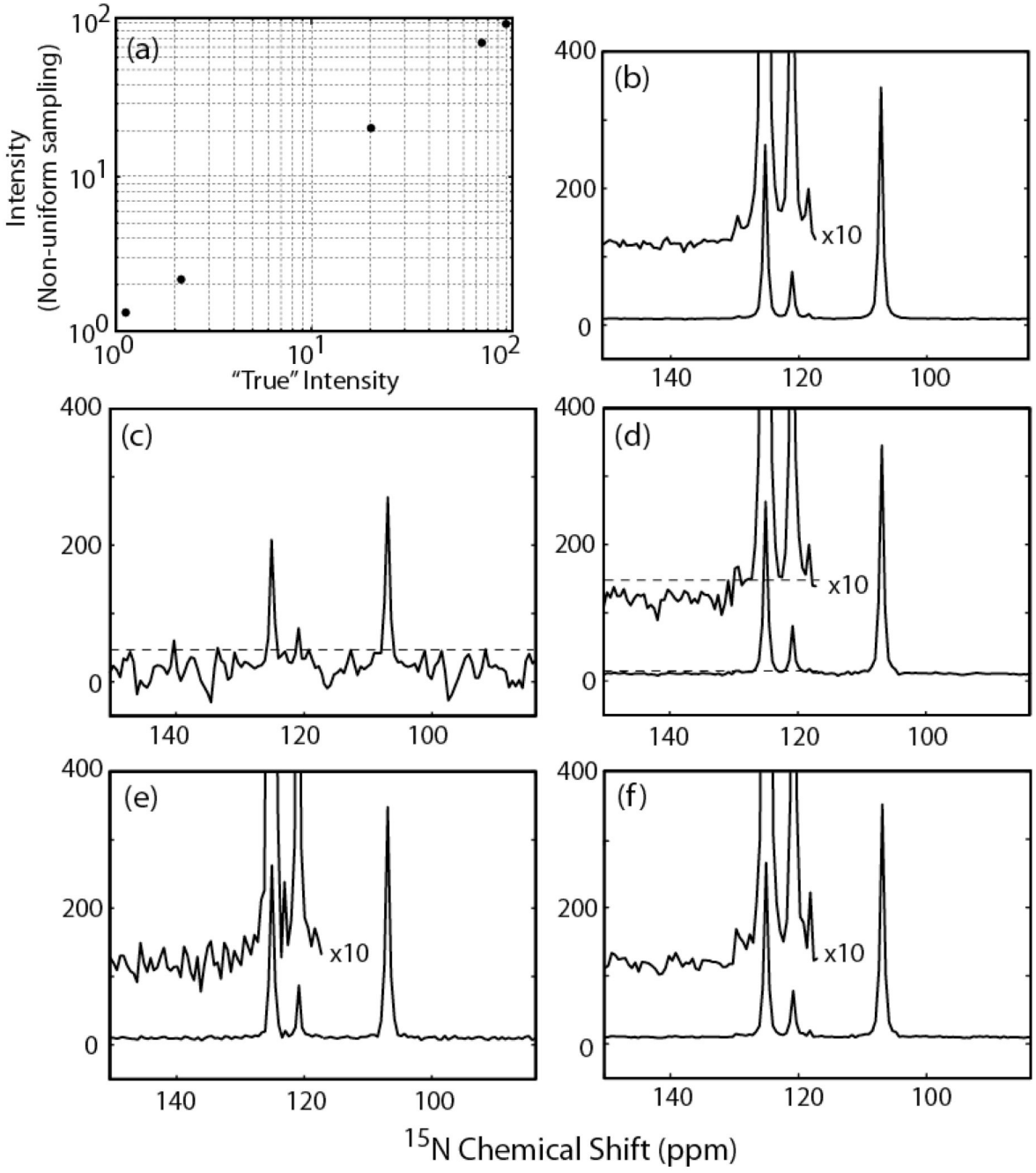

Figure 9.

Reconstruction of synthetic 1D data with high dynamic range. (a) Double-log plot of the normalized peak intensities observed in the transform of the full data (b) and SIFT-processed NUS data (i_{NUS} = 64) (d). Also shown are the transforms of NUS data (c), NUS data restored by the iterative thresholding (e), or by SIFT followed by thresholding (f). In the insets of (b) (d) (e) and (f), the region between 118 and 150 ppm is vertically expanded by a factor of 10. The full data contains peaks at 106.9, 125.0, 120.8, 118.2 and 129.3 ppm, whose relative intensity is set to 100, 75, 20, 2.0 and 1.0, respectively. A threshold at 40 and 12, shown by a dashed line in (c) and (d) was used to yield the spectra in (e) and (f), respectively. The number of SIFT and thresholding cycles was 10 and 50, respectively, for the spectra in (d), (e) and (f).

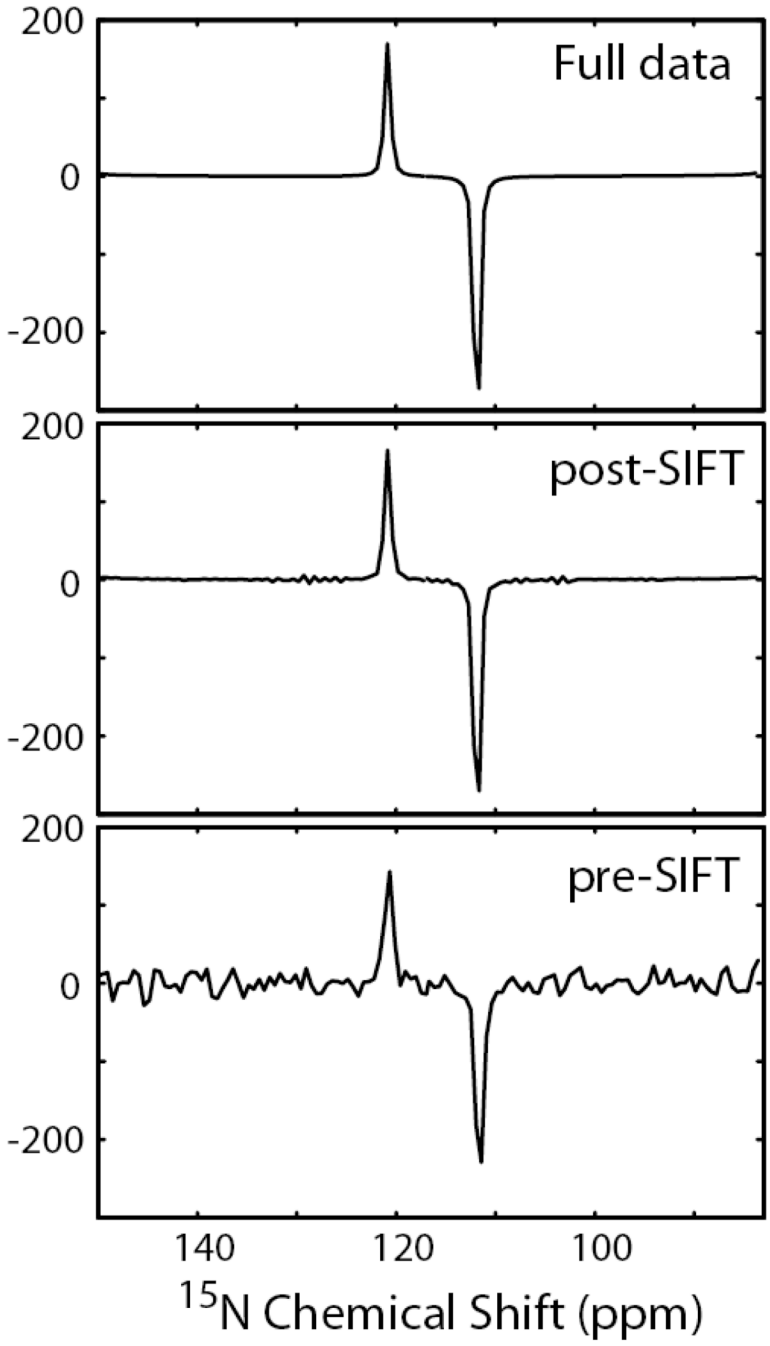

Figure 10.

Reconstruction of synthetic 1D data with negative and positive signals. The panels show, from the top to the bottom, the transform of the full data set, and the post- and pre-SIFT NUS data (i_{NUS} = 64). The phase-inverted peak at ~110 ppm is due to the spectral aliasing, the initial sampling delay equal to half the dwell time, and an appropriate phase correction to yield absorption signals (as happens for a real case). The number of SIFT cycles is 10.

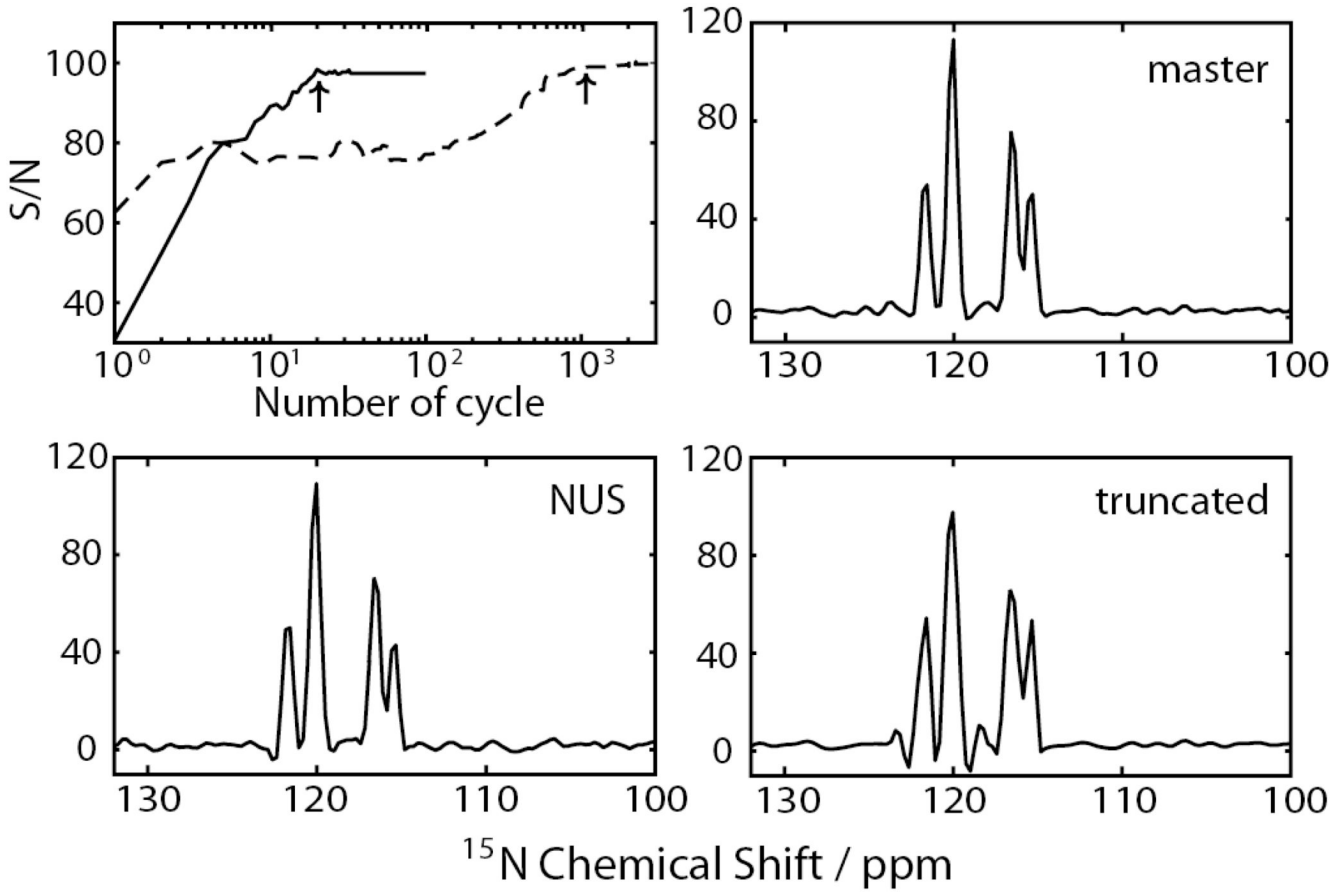

Figure 11.

Comparison of iterative thresholding results for a 48-point NUS dataset and a 48-point uniformly sampled short record. The top left panel shows the S/N along the iteration for NUS data (solid line) and truncated uniformly sampled data (broken line). Arrows indicate the points of cycle termination. The computation took 6 sec for the NUS data and 5 min for the truncated data. The other panels show a slice taken at F2 = 8.34 ppm from the spectrum of the full master dataset (top right), the processed NUS data (bottom left) and processed truncated data (bottom right).

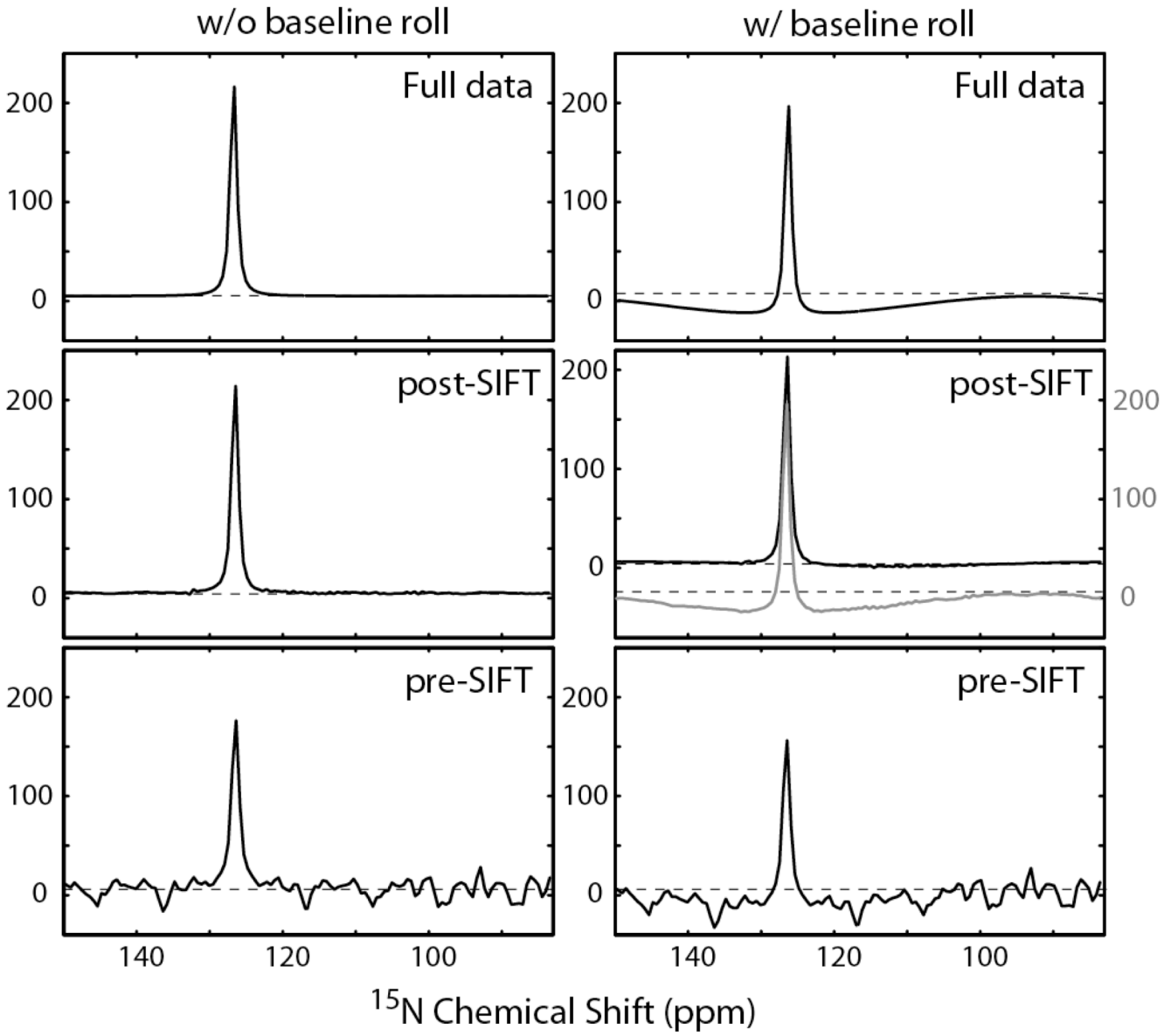

Figure 12.

Reconstruction of a synthetic 1D data with (right column) and without (left column) a base-line roll. The base-line roll was manufactured by intentionally corrupting the initial two time data points (by setting them to zero). The panels in each column show, from the top to the bottom, the transform of the full data, and post- and pre-SIFT NUS data (i_{NUS} = 64). In the middle right panel, the upper trace is a result of SIFT restoring the two corrupted initial time data points, as well as the missing points due to NUS. The number of SIFT cycles was 100 for the upper trace, and 10 otherwise. The black and gray traces refer to the left and right y-axes, respectively.

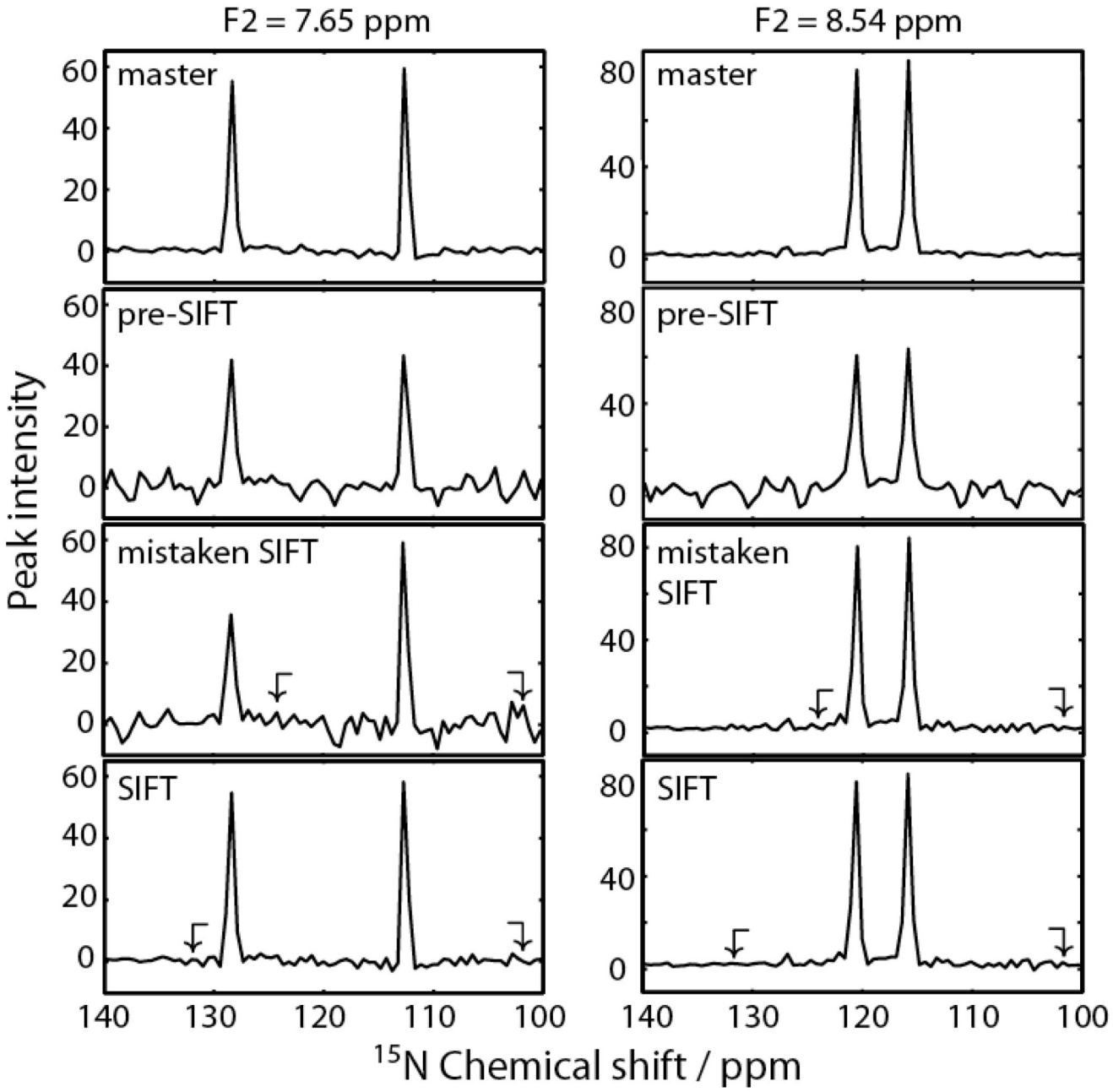

Figure 13.

Two representative slices taken at F2 = 7.65 (left) and = 8.54 (right) ppm from a SIFT-processed spectrum with i_{NUS} = 64. The panels in each column show, from the top to the bottom, the slice taken from the full master dataset, the NUS dataset before SIFT, the NUS dataset after mistaken SIFT, and the NUS dataset after correct SIFT. Arrows indicate the boundaries of the dark region. For the mistaken SIFT, the dark region was deliberately miss-specified to include 128 ppm.