J Biomol NMR (2012) 54:15–32
DOI 10.1007/s10858-012-9643-4

ARTICLE

Compressed sensing reconstruction of undersampled 3D NOESY spectra: application to large membrane proteins

Mark J. Bostock · Daniel J. Holland · Daniel Nietlispach

Received: 6 April 2012 / Accepted: 30 May 2012 / Published online: 26 July 2012

Abstract Central to structural studies of biomolecules are multidimensional experiments. These are lengthy to record due to the requirement to sample the full Nyquist grid. Time savings can be achieved through undersampling the indirectly-detected dimensions combined with non-Fourier Transform (FT) processing, provided the experimental signal-to-noise ratio is sufficient. Alternatively, resolution and signal-to-noise can be improved within a given experiment time. However, non-FT based reconstruction of undersampled spectra that encompass a wide signal dynamic range is strongly impeded by the non-linear behaviour of many methods, which further compromises the detection of weak peaks. Here we show, through an application to a larger α-helical membrane protein under crowded spectral conditions, the potential use of compressed sensing (CS) l_1-norm minimization to reconstruct undersampled 3D NOESY spectra. Substantial signal overlap and low sensitivity make this a demanding application, which strongly benefits from the improvements in signal-to-noise and resolution per unit time achieved through the undersampling approach. The quality of the reconstructions is assessed under varying conditions. We show that the CS approach is robust to noise and, despite significant spectral overlap, is able to reconstruct high quality spectra from data sets recorded in far less than half the amount of time required for regular sampling.

Keywords Compressed sensing · Nonuniform sampling · NOESY spectroscopy · l_1-norm minimisation · Signal-to-noise ratio · Resolution · NMR spectroscopy

Introduction

Along with X-ray crystallography, NMR spectroscopy is the only atomic resolution technique available to study molecular structure. Developments in the field in recent years have increased the size-limit to which NMR structures can be determined (Pervushin et al. 1997). These developments have allowed NMR spectroscopy to probe increasingly demanding applications, including protein complexes (Fiaux et al. 2002; Sprangers et al. 2007), and the class of larger membrane proteins (Gautier et al. 2010; Hiller et al. 2008; Kim et al. 2009; Nietlispach and Gautier 2011).

Central to such structural studies are multidimensional experiments with three and four dimensions, typically used to assign backbone and side-chain connectivities, and that provide essential distance information via the nuclear Overhauser effect between protons close in space. With the long time required to sample the full Nyquist grid for Fourier-transform processing of multidimensional experiments, the stability of samples is frequently a limiting factor, while low concentrations exacerbate this situation. Furthermore, under conditions where peaks are likely to overlap and be of low intensity as encountered with larger proteins, obtaining conditions that provide both sufficient resolution and signal-to-noise ratio (SNR) can be difficult to achieve.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-012-9643-4) contains supplementary material, which is available to authorized users.

M. J. Bostock · D. Nietlispach (✉)
Department of Biochemistry, University of Cambridge, Cambridge, UK
e-mail: dn206@cam.ac.uk

D. J. Holland
Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK

Hence, data collection coupled to Fourier transformation for multidimensional experiments is frequently a compromise between the number of points acquired in the indirect dimensions and the number of scans acquired per increment. Thus, either the SNR can be increased by increasing the number of scans, at the cost of reducing the number of indirect data points, or the resolution can be improved by recording more data points in the indirect dimensions, at a cost of reducing the number of scans per increment that can be recorded in a given experiment time, thus limiting the SNR. Time-domain undersampling combined with suitable processing methods offers the potential to improve resolution and signal-to-noise.

So-called undersampling, nonuniform or sparse sampling of multidimensional spectra was demonstrated many years ago as an alternative to recording the full Nyquist sampling grid (Barna et al. 1987; Schmieder et al. 1993, 1994). As it is no longer possible to use the discrete Fourier transform to reconstruct such undersampled spectra, alternative reconstruction techniques have been proposed including nonuniform Fourier transform (Coggins and Zhou 2008; Kazimierczuk et al. 2006; Marion 2005, 2006), maximum entropy reconstruction (Barna et al. 1987; Hoch et al. 1990; Rovnyak et al. 2004), multi-dimensional decomposition (Orekhov et al. 2001; Tugarinov et al. 2005), maximum likelihood estimation (Chylla and Markley 1995) and many others (Atreya and Szyperski 2004; Freeman and Kupče 2003; Frydman et al. 2002; Kupce and Freeman 2004; Kupce et al. 2003; Mandelshtam 2000). To date, such methods have largely been used to reconstruct undersampled backbone-assignment experiments (Gautier et al. 2010; Rovnyak et al. 2004), which often show ample signal-to-noise with peak intensities spread over a small dynamic range, or spectra of highly deuterated or selectively-labeled samples, which result in a less crowded spectrum (Hiller et al. 2008; Tugarinov et al. 2005). However, once signal intensities encompass a wider dynamic range, nonparametric methods show difficulties in the correct reconstruction of signal intensities and may bias the detectability of weak peaks. This renders them poorly suited for the crowded spectral environments typical of 3D NOESY experiments. Weak cross peaks often contain the most valuable structural information and they are most likely to be suppressed.

Recently, we and others demonstrated the potential of compressed sensing l_1-norm minimization, to reconstruct highly undersampled NMR spectra (Holland et al. 2011; Kazimierczuk and Orekhov 2011) and Hyberts et al. (2012) showed the applicability of iterative soft thresholding (IST) as one form of l_1-norm minimization to reconstruct 3D and 4D NOESY data demonstrating dramatic time savings. Previously IST was suggested as a version of l_1-norm minimization in the context of extending fully-sampled but truncated data (Stern et al. 2007). While l_1-norm regularization has been available over decades (Logan 1965), in the context of more recent applications the notion 'Compressed Sensing' (CS) was formulated (Donoho 2006; Candes et al. 2006) and CS has become increasingly popular in a number of signal processing fields, including image reconstruction (Holland et al. 2010; Hu et al. 2008; Lustig et al. 2007; Otazo et al. 2010) and was also shown to be suitable for the reconstruction of sparsely-sampled multidimensional NMR spectra (Drori 2007). Here we briefly outline the principle of CS for application to NMR spectroscopy. Consider the system of underdetermined linear equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1}$$

where $\mathbf{A}$ is an $M \times N$ matrix and $\mathbf{x}$ is a vector of length N that is to be recovered from measurements $\mathbf{b}$, where $\mathbf{b}$ is a vector of length M, and $M < N$. In the case of NMR spectroscopy, $\mathbf{x}$ corresponds to the spectrum in the frequency domain, $\mathbf{b}$ to the measurements in the time domain, and $\mathbf{A}$ is the inverse Fourier transform. Since $M < N$, (1) has infinitely many solutions. CS demonstrates that providing the spectrum, $\mathbf{x}$, is sparse, $\mathbf{x}$ can be exactly reconstructed from $O(k)$ random projections (for a k-sparse spectrum i.e. with no more than k nonzero components), by solving

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b} \tag{2}$$

where the l_0-norm is a pseudo-norm defined by:

$$\|\mathbf{x}\|_0 = \sum_i |x_i|^0 \tag{3}$$

with $0^0 = 0$ (Donoho 2006), in other words, finding the sparsest solution, i.e. that with the fewest nonzero elements, which is consistent with the measured data. Above, x_i is the i^{th} element of $\mathbf{x}$. The minimization described by (2) is a combinatorial problem, therefore it has been shown that for realistic spectra, it is not possible to solve (2) computationally (Natarajan 1995). However, where the solution to the l_0-norm is sufficiently sparse, minimising the l_1-norm (4, 5) returns the same solution and can be solved using standard linear programming

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \mathbf{A}\mathbf{x} = \mathbf{b} \tag{4}$$

where

$$\|\mathbf{x}\|_1 = \sum_i |x_i| \tag{5}$$

and which generalizes for the l_p-norm to

$$\|\mathbf{x}\|_p = \left(\sum_i |x_i|^p \right)^{1/p} \tag{6}$$

with $p > 0$. When $\mathbf{b}$ contains noise, or is not sparse, but merely compressible, situations that occur for most

practical applications, the constraint in (4) can be relaxed taking a number of different formalisms in various algorithms such as

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{Ax} - \mathbf{b}\|_2 \leq \delta \quad (7)$$

Compressed sensing requires (i) sparse representation of the desired signal in a particular basis and (ii) incoherent sampling with respect to that basis. An initial estimation of the signal location is not required.

Many NMR spectra are sparse in the frequency domain, with sparsity increasing as the number of dimensions increases. Crucially, CS theory states that the number of measurements required to recover a signal is C$k$log(N), where C is a constant of proportionality. This means the number of measurements is dependent primarily on the number of non-zeros in the signal (i.e. k), and is dependent only weakly on the number of points in the final spectrum, N. Therefore, the dimensionality of the experiment can be increased with only a slight increase in the number of measurements. For example, to add a third dimension to a 2D experiment containing n_2 points in the first indirect dimension and n_3 points in the second indirect dimension, would conventionally require a factor of n_3 additional measurements; in the CS framework only a factor of $(1 + \log(n_3)/\log(n_2))$ additional measurements are required. Assuming say 32 points in each dimension, this would represent a time saving of up to 16 times.

The incoherence of aliasing artifacts due to undersampling in the time-domain is typically increased by randomized nonuniform undersampling. Next to the reconstruction method used, this introduces a further dependence of the obtainable spectral quality on the sampling schedule. Many sampling approaches have been proposed in this ongoing field of research (Barna et al. 1987; Coggins and Zhou 2008; Eddy et al. 2012; Hyberts et al. 2010, 2011; Kazimierczuk et al. 2008; Kupce and Freeman 2003; Mobli et al. 2006; Rovnyak et al. 2004). The degree of randomness and level of sampling artifacts generated by a given sampling scheme can be readily assessed by calculation of the point spread function (psf). This involves computing the Fourier transformation of the binary sampling function, where ones represent sampled points, and all other points are set to zero (Hoch et al. 2008; Kazimierczuk et al. 2010). Growing randomization typically increases the reconstruction quality (Hoch et al. 2008).

Compressed sensing has been demonstrated as suitable for reconstructing highly-undersampled spectra for backbone assignment e.g. HNCA and HN(CO)CA, and, crucially, was shown to successfully reconstruct low-intensity peaks in the presence of stronger peaks (Holland et al. 2011). The apparent linearity of the method over a large dynamic range of signal intensities suggested its suitability for reconstructing 3D NOESY spectra (Holland et al. 2011; Kazimierczuk and Orekhov 2011) as was recently demonstrated (Hyberts et al. 2012). In situations where backbone-assignment experiments are recorded following a sparse sampling protocol, 3D NOESY spectra are likely to become the time-limiting data-recording step. In this paper, we demonstrate with CS the possibility to obtain high quality 3D NOESY spectra from undersampled data recorded on a fully-protonated sample of a large membrane protein. We show that the full range of signal intensities can be accurately reproduced and despite extremely crowded spectra, substantial time savings of 60–70 % are achievable.

We assess the quality of the reconstructions and the use of different undersampling factors through comparison with the Fourier-transformed spectrum of the complete data matrix. We validate the linearity of the reconstruction over the wide range of intensities present in the NOESY spectrum, as well as the fidelity of reconstruction of peaks at the correct chemical shift positions; results are compared using different reconstruction algorithms and sampling schemes. We conclude that the robustness of the reconstructions to noise, the high quality of the produced spectra and the substantial time savings make the methodology a considerable asset for the study of large biomolecular systems.

Methods

Data recording

NMR experiments were recorded as 3D NOESY ^{15}N HSQC experiments (τ_{mix} = 100 ms) on a 0.5 mM sample of ^{15}N-labeled V17C mutant of sensory rhodopsin II (pSRII) (pH 6.0, 63 mM c7-DHPC) on a Bruker DRX800 spectrometer equipped with a 5 mm TXI HCN cryoprobe at 308 K. Experiments used water-flipback and Watergate (Piotto et al. 1992) implementations and were recorded with 512 complex data pairs in the directly detected ^{1}H dimension (acquisition time = 51 ms). A full data matrix (suitable for FT in all dimensions) was recorded consisting of 110 complex pairs in the ^{1}H ($t_{1,max}$ = 13.1 ms) and 39 complex pairs in the ^{15}N ($t_{2,max}$ = 15.5 ms) indirect dimensions. Quadrature detection was achieved following the States-TPPI recipe (Marion et al. 1989) and the first increments were adjusted for (90°, −180°) phase settings. The full data matrix was recorded in sets of 8 scans (50 h), as required for phase cycling, which could further be co-added to result in spectra with varying signal-to-noise ratios. Subsets of this full data matrix were selected to generate the undersampled data sets. The included data

points varied depending on the chosen level of undersampling and sampling scheme applied.

Fourier transform processing

All spectra were initially processed using the software package Azara (W. Boucher, unpublished), where the directly-detected ^{1}H dimension was Fourier-transformed once apodization with a Gaussian window function, a 90° shifted sinebell squared function and zerofilling to 1,024 points had been applied. Finally, every row was baseline-corrected to remove any DC offsets. For the full data matrix of 110* × 39* points (* indicates complex) Fourier transformation in MATLAB was done after application of 90° (^{1}H) and 60° (^{15}N) shifted sinebell window functions and zerofilling to 256 points. The final size of the real 3D data matrix was 430 × 256 × 256 points.

Compressed sensing reconstruction

For the undersampled data sets, following processing of the directly detected dimension in Azara as described above, the full data matrix was subsequently imported into MATLAB. Points in the indirect dimensions were then selected according to the chosen sampling scheme to generate the undersampled data matrices. For each selected point, both cosine and sine modulated components were included, equivalent to undersampling complex data pairs. Data point selections correspond to the following undersampling schemes: exponentially-weighted, on-grid sampling at 40, 30, 20 and 15 % and weighted-Poisson sampling at 40 % (Hyberts et al. 2010, 2011). Exponential sampling was carried out similar to previously published methods (Barna et al. 1987; Rovnyak et al. 2011) using in-house scripts: For a given indirect dimension, an exponential of the form $\exp(-R_2 \times cz/sw)$ was used, where R_2 is the average transverse spin relaxation rate constant of the nucleus in question, sw is the spectral width for the given dimension, z is the number of real points for the given dimension and c is a constant used to scale the exponential to give the appropriate sampling fraction. R_2 values of 66 and 40 Hz (T_2 = 15 and 25 ms) were used for ^{1}H and ^{15}N respectively. The exponentials from both indirect dimensions were combined, forming a probability density function (pdf) spanning the full matrix of indirect dimension points, and the first point (1,1) was set to 1. A previously published Monte-Carlo approach was then used to generate a number of samples equal to the sum of the pdf, and with the lowest transform point spread function (psf) (Lustig et al. 2007). The first point (1,1) was always sampled. The sampling scheme was then replicated such that both sine and cosine points for each time-point were sampled.

Once the undersampled data sets were generated, compressed sensing (CS) reconstruction of both indirect dimensions was carried out with MATLAB using in-house developed scripts or the previously published algorithm, YALL1 (Yang and Zhang 2011) adapted for use with NMR data. Spectra were viewed and analysed in Azara and CCPN Analysis (Vranken et al. 2005) and spectral assignments were carried out in Analysis.

Iterative hard thresholding (IT)

Scripts for this algorithm were based on work by Bredies and Lorenz (2008). An initial estimate of the solution is calculated by replacing the missing data time points with zeroes followed by Fourier transformation. A threshold is subsequently set where data points above it in the frequency domain are stored. The inverse transform of this subset is calculated and the result subtracted from the original data set. The operation is repeated iteratively and can be described as:

$$\mathbf{x}^{[n+1]} = \mathbf{x}^{[n]} + H_{\lambda}(\mathbf{A}^{T}(\mathbf{b} - \mathbf{A}\mathbf{x}^{[n]})) \quad (8)$$

where, $\mathbf{x}^{0} = 0.$ and $H_{\lambda}(\mathbf{a})$is a non-linear operator which sets the elements of $\mathbf{a}$ (a_i), that are below a threshold λ, to zero:

$$H_{\lambda}(\mathbf{a}) = \begin{cases} 0 & |a_i| < \lambda \\ a_i & |a_i| \geq \lambda \end{cases} \quad (9)$$

In the case of NMR data, $\mathbf{A}$ represents the inverse Fourier transform and $\mathbf{A}^{T}$ the forward Fourier transform; $\mathbf{b}$ is the undersampled data and $\mathbf{x}^{n}$ the estimated spectrum at iteration n. For the reconstructions in this paper, λ was set at 90 % of the maximum value in the Fourier transformed data and recalculated for each iteration. Therefore, as n increases, $(\mathbf{A}^{T}(\mathbf{b} - \mathbf{A}\mathbf{x}^{[n]}))$ tends to a spectrum containing only noise. Alternative stopping criteria were used: (i) reconstructions were stopped when the number of points in the reconstruction was equal to a proportion (e.g. 40 % for a 40 % sampling schedule) of the total number of points (assuming full sampling) in the indirect [^{1}H,^{15}N]-plane i.e. 220 × 78; (ii) an estimate of the expected noise in the data set was calculated and the iterations stopped once the standard deviation of the term $(\mathbf{A}^{T}(\mathbf{b} - \mathbf{A}\mathbf{x}^{[n]}))$ approached this noise value. The noise estimate was determined by calculating the standard deviation of each indirect [^{1}H,^{15}N] plane after a zero-filled Fourier transform of the undersampled data. Values for planes with no peaks were used to give the noise estimate; (iii) $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ was calculated at each iteration, where $\mathbf{b}$ is the undersampled data and $\mathbf{A}\mathbf{x}$ represents the inverse Fourier transform of the spectral reconstruction at a given iteration. Iterations continued until $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 < \delta$ was achieved, where δ is user-defined.

Iterative soft thresholding (IST)

In contrast to hard thresholding, IST involves soft thresholding or non-linear shrinkage (Daubechies et al. 2004); data points in the frequency domain below a designated threshold are set to zero, whilst those above are shrunk towards zero. The inverse Fourier transform is calculated and the result removed from the original data. The operation is then repeated iteratively which can be represented as:

$$\mathbf{x}^{[n+1]} = S_{\lambda}(\mathbf{x}^{[n]} + \mathbf{A}^{T}(\mathbf{b} - \mathbf{A}\mathbf{x}^{[n]})) \quad (10)$$

where $S_{\lambda}(\mathbf{a})$ is defined as:

$$S_{\lambda}(\mathbf{a}) = \begin{cases} 0 & |a_i| < \lambda \\ (|a_i| - \omega\lambda) \cdot sign(a_i) & |a_i| \geq \lambda \end{cases} \quad (11)$$

where λ is the threshold parameter, which was set to 10^4 for the reconstructions in this paper and ω is a scaling factor which was set to one for the reconstructions shown here. Reconstructions were stopped either (i) when the gradient of the standard deviation of the residual against iteration number approached zero or, (ii) to achieve lower values of $\|\mathbf{Ax} - \mathbf{b}\|_2$, once the change in $\|\mathbf{Ax} - \mathbf{b}\|_2$ was constant indicating convergence at a given λ, the value of lambda was reduced and the calculation repeated; this was continued (while $\lambda > 1$) until the desired value of $\|\mathbf{Ax} - \mathbf{b}\|_2$ was achieved.

YALL1 (Your ALgorithms for L1 Optimisation)

A method as described by Yang and Zhang (2011) was adapted for use with Fourier transformation and the l_1/l_2 convex minimization equation

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ subject to } \|\mathbf{Ax} - \mathbf{b}\|_2 \leq \delta \quad (12)$$

was solved with δ set to 16,800 (high $\|\mathbf{Ax} - \mathbf{b}\|_2$), or 1 (low $\|\mathbf{Ax} - \mathbf{b}\|_2$) (Table 1). The stopping tolerance was set to 10^{-4}. While no improvement in spectral quality was found by lowering this to 10^{-5}, the reconstruction quality declined on increasing it to 10^{-3}.

For each of the reconstructions the values for $\|\mathbf{Ax} - \mathbf{b}\|_2$ and $\|\mathbf{x}\|_1$ are reported (Table 1). All scripts used for this work will be available implemented in the processing software Azara.

Reconstruction times

Reconstructions were carried out in MATLAB on an AMD Phenom II, Quad Core 3.0 GHz processor, with 8 GB of memory. Reconstruction times varied slightly dependent on machine load but were typically 30 - 60 min for IT reconstruction with high $\|\mathbf{Ax} - \mathbf{b}\|_2$ or ~3 h for a tighter constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$, ~30 min for IST using criteria (i), or 2 h 30 with the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2 = 2$ using stopping criteria (ii) and ~4–5 h for YALL1 reconstruction. For more detail, see Table 1.

Spectral data analysis

All NMR spectra were analysed in the CCPN software suite Analysis (Vranken et al. 2005), which was used to obtain spectral assignments, peak positions, peak intensities and values for spectral noise. A noise level of 8.4×10^4 was determined based on the full FT data, and a contour threshold 1.25 times above this noise level was used as a setting for NOE cross-peak picking and analysis of all the spectra. Accordingly, the same corresponding artifact level (rather than noise level) was used for the nonlinear processed spectra. In this report, in the interest of simplicity, we refer to noise level rather than artifact level, regardless of the processing method employed. Picking of the peaks with intensities above this contour threshold was obtained automatically using the in-built peak-picking algorithm in Analysis. To validate the performance of the peak reconstructions for different signal intensities, peaks in the

Table 1 Comparison of the l_1-norm and $\|\mathbf{Ax} - \mathbf{b}\|_2$ parameters for the different algorithms and sampling factors discussed in the paper **A, x** and **b** represent the inverse Fourier transform, spectrum and recorded data respectively as discussed in the paper. $\boldsymbol{\sigma}$ is the standard deviation

	Algorithm	Sampling factor	$\|\|x\|\|_1$	$\|\|Ax-b\|\|_2$	$\sigma(\|\|Ax-b\|\|_2)$	Processing time
High $\|\|Ax-b\|\|_2$	IT	40%	2.98×10^{11}	2.66×10^{4}	1.10×10^{2}	0.5 h
		30%	3.75×10^{11}	2.58×10^{4}	3.19×10^{2}	0.5 h
		20%	3.01×10^{11}	1.92×10^{4}	5.10×10^{2}	1 h
		15%	2.92×10^{11}	1.74×10^{4}	5.20×10^{2}	1 h
	IST	40%	4.06×10^{11}	1.45×10^{4}	4.49×10^{2}	0.5 h
	YALL1 δ=16800	40%	3.90×10^{11}	1.62×10^{4}	1.77×10^{2}	4.5 h
		30%	3.61×10^{11}	3.10×10^{4}	6.80×10^{3}	5.5 h
Low $\|\|Ax-b\|\|_2$	IT	40%	5.93×10^{11}	1.96×10^{0}	3.2×10^{-2}	3 h
	YALL1 δ=1	40%	5.87×10^{11}	3.11×10^{0}	6.98×10^{0}	3.5 h
		30%	5.50×10^{11}	2.34×10^{0}	4.72×10^{0}	4.5 h
	IST	40%	5.83×10^{11}	2.03×10^{0}	9.70×10^{-3}	2.5 h

original Fourier-transformed spectrum were classified according to their intensities. Accordingly, from the large number of selected peaks, which spread over a wide dynamic range, several subsets (bins) each consisting of 100 peak intensities, were randomly chosen in the following peak intensity ranges (I): $10^5 \leq I < 5 \times 10^5$, $5 \times 10^5 \leq I < 10^6$, $10^6 \leq I < 10^7$, $10^7 \leq I < 10^8$, $I \geq 10^8$ (corresponding to a SNR range from 1.2 to $>$ 1,200). Where less than 100 peaks were present, all peaks in that particular bin were included. Upon closer inspection of the selected peaks a small number, which were clearly recognizable as spurious noise, truncation artifacts, or with their intensities distorted by the presence of neighboring peaks, were excluded from the analysis. Intensities and chemical shift positions of the remaining peaks were selected for a quantitative evaluation of the performance of the CS reconstructions. Root mean squared errors (RMS) (%) for the sums of the peak intensities within any of the above intensity bin ranges were calculated according to

$$\text{RMS } (\%) = \frac{\left(\sum_i (p_i^{FT} - p_i^{CS})^2\right)^{0.5}}{\left(\sum_i (p_i^{FT})^2\right)^{0.5}} \quad (13)$$

where p_i^{FT} and p_i^{CS} are the intensities for the ith peak in a given bin for the fully sampled FT spectra and CS reconstructed undersampled spectra, respectively.

A quantitative assessment of the effect of the CS reconstructions on the peak positions within the spectra was made on a per residue basis for the indirect dimensions. Combined nucleus-weighted chemical shift changes between FT and CS processed spectra were determined according to:

$$\Delta\delta = \left((\delta_{^1H^{CS}} - \delta_{^1H^{FT}})^2 + ((\delta_{^{15}N^{CS}} - \delta_{^{15}N^{FT}})/2.5)^2\right)^{0.5} \quad (14)$$

Results and discussion

In this contribution we show, based on 3D NOESY ^{15}N HSQC data, the suitability of compressed sensing methodology to reconstruct extremely crowded NOESY spectra recorded on a detergent-micelle solubilized membrane protein. Due to the large number of signals with intensities covering a considerable dynamic range and the frequent spectral overlap this is a demanding application. We assess the performance of the CS reconstructions qualitatively through direct comparison of the spectra and quantitatively in particular with regard to peak intensities and peak positions. Comparing against the Fourier-transformed spectra as the 'gold standard,' we show and discuss the CS results obtained using a range of algorithms and different conditions. For a reliable assessment of the different CS reconstruction methods, comparisons are made using similar values for the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$. Several cases are tested and the results emphasize the robustness of the CS reconstructions.

Common to many other processing methods, the benefit of CS lies in its powerful ability to reconstruct sparsely sampled data sets, thus allowing experiment times to be considerably reduced if the overall signal-to-noise permits, or alternatively if sample concentrations are limiting to benefit from improved spectral resolution and SNR per unit time. SNR and resolution issues are particularly pronounced in the study of large proteins and/or for biomolecules that have a relatively limited sample lifetime and hence such cases will benefit dramatically from sparse sampling.

Our study shows that CS reconstructions are of very high quality, do not noticeably reduce the detectability limit for the weakest peaks, prove to be linear over the tested wide range of spectral intensities observed in a NOESY spectrum and do not show any systematic deviations in peak positions. Consequently, CS-reconstructed NOE spectra are suitable for obtaining structural information with substantial time savings. The CS technique is potentially better suited when compared to other reconstruction methods, which may struggle under the conditions of the large signal dynamic range encountered here, and may reduce or even suppress the structurally important, weaker cross peaks.

As a typical example of an application to a larger biomolecular entity which gives crowded NOESY spectra, we have recorded 3D NOESY ^{15}N HSQC data on a fully protonated sample of sensory rhodopsin II (pSRII) which forms a 70 kDa protein-detergent complex in c7-DHPC micelles (Gautier et al. 2010). Spectra were recorded at 800 MHz and 308 K, where the rotational correlation time of the protein is 34 ns. Spectra were processed using Fourier transformation as well as CS reconstruction using several algorithms and sampling schemes. A full data matrix with 110 $\times$ 39 complex pairs in the indirect ^{1}H and ^{15}N dimensions was recorded for the Fourier-transformed reference spectrum while undersampled data sets were generated from this full data matrix by selecting only the time points which were in agreement with the chosen sampling scheme. This was then followed by subsequent CS reconstruction of the spectra.

Figure 1 shows a comparison between the fully-sampled, Fourier-transformed reference data and spectra reconstructed by CS, using the iterative hard thresholding algorithm (see "Methods") from 50 % of the original data matrix in both ^{1}H and ^{15}N indirect dimensions. The indirect dimensions were undersampled using an exponential sampling scheme weighted by the average T_2 values of the

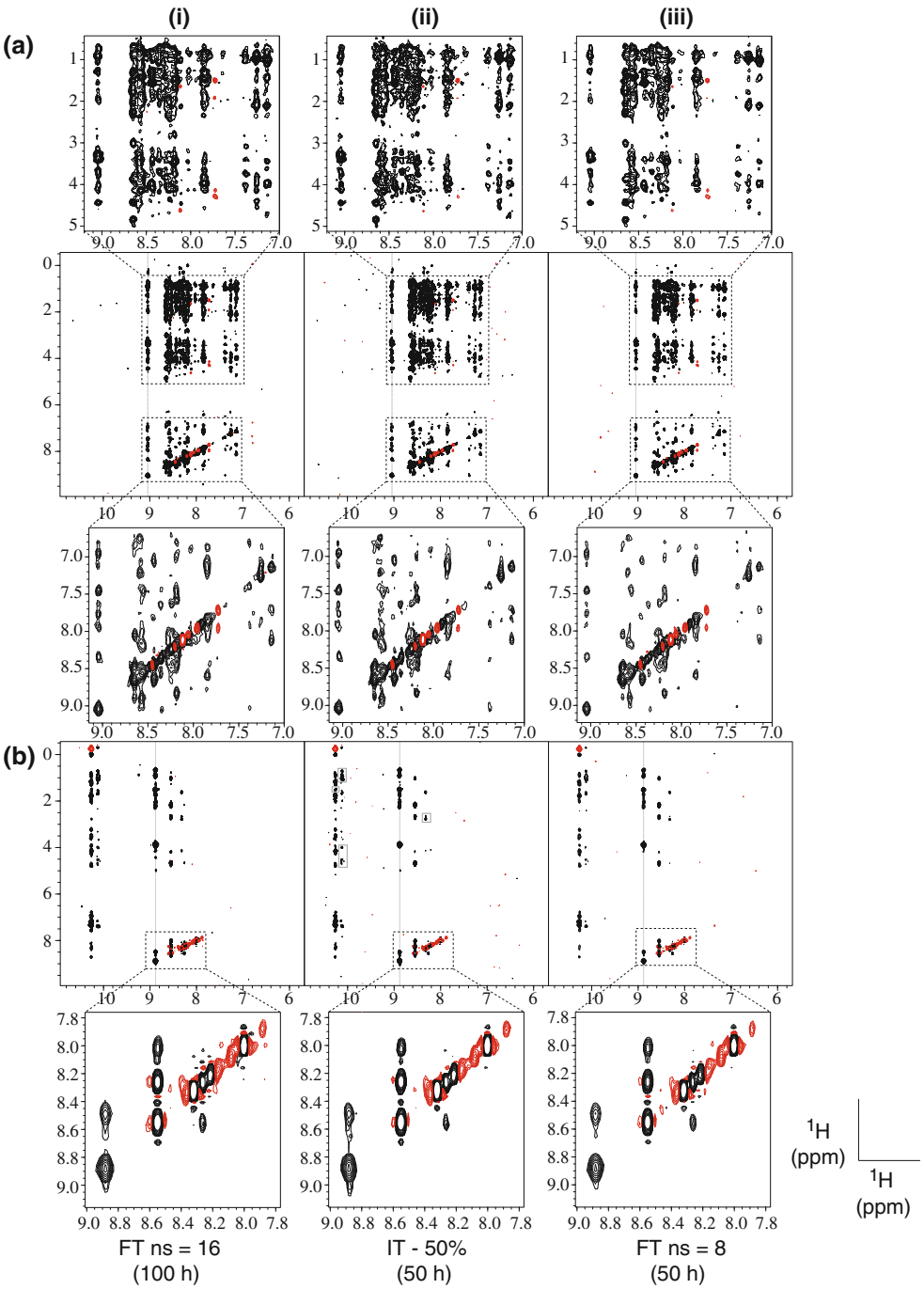

Fig. 1 Comparison of two [^{1}H,^{1}H] planes from a 3D NOESY ^{15}N HSQC spectrum at ^{15}N positions of (**a**) 120.80 ppm (Phe 210) and (**b**) 127.58 ppm (Ala 64), recorded and processed with different methods: (**i**) fully sampled Fourier-transformed (FT) spectrum recorded with 16 scans (100 h); (**ii**) compressed sensing reconstruction using 50 % sampling (equivalent to 50 h) with the iterative hard thresholding algorithm (IT); (**iii**) the fully sampled Fourier-transformed spectrum recorded with 8 scans (50 h). Spectral expansions of diagonal and aliphatic regions are shown for the areas surrounded by a *dashed square*. *Dotted grey* lines indicate the corresponding shift positions of the amides at which the two planes in (**a**) and (**b**) have been centered. In (**b**) regions marked by *grey rectangles* indicate a few examples where the compressed sensing reconstruction shown in (**ii**) is able to detect weak peaks that are not observable in the FT spectrum (**iii**) recorded over the same amount of time (50 h). The negative peak at 10.25 ppm (F1) is a folded diagonal signal

^{1}H (15 ms) and ^{15}N (25 ms) spins in question (see "Methods"). Two representative planes at ^{15}N frequencies of 120.80 ppm (Fig. 1a) and 127.58 ppm (Fig. 1b) are shown, demonstrating the fidelity of the CS reconstruction under very contrasting spectral conditions: in Fig. 1a the cross peak regions are very overlapped and a lot of the peaks are weak compared to the diagonal signals, while in Fig. 1b, overlap is not a significant problem and the signals are also generally more intense. The enlarged areas demonstrate the high quality of the reconstruction for regions both near the intense diagonal of the amide signal and the weaker aliphatic cross peaks. As can be seen, the achieved resolution in the 50 % undersampled CS spectra matches that in the FT-processed spectra.

To illustrate the impact of the CS reconstruction approach on signal-to-noise and resolution, both planes of Fig. 1a and Fig. 1b are shown as (i) processed with FT using the full data matrix recorded over 100 h (16 scans), (ii) the 50 % undersampled CS reconstruction of the data set recorded in only 50 h (16 scans) and (iii) the analogous

FT spectrum as shown in (i) but with only half the number of scans, corresponding to 50 h (8 scans) i.e. the same amount of experiment time as for (ii). All our subsequent observations apply both to Fig. 1a, b: Comparing the spectra in (i) and (ii) shows the fidelity of the CS reconstruction with the FT spectrum that was recorded for twice as long, underlining the potential time-saving achievable using the CS method. Comparing (ii) and (iii) shows the CS reconstruction (16 scans) and full Fourier transform spectrum (8 scans), both recorded for identical times. The contour base levels are scaled such that the viewed noise level is equivalent in both spectra and it is clear that for the same experiment time, CS achieves a higher SNR and is able to detect weak peaks, which are not present in the time-equivalent FT spectrum (iii). Examples of peaks close to the noise limit, which show improved signal-to-noise in the CS reconstruction, are highlighted by grey rectangles in Fig. 1b.

It could be argued that the apparent lower signal-to-noise with the FT method is the result of an unfair comparison, since large amounts of time in the fully-sampled FT spectrum have to be spent to obtain sufficient resolution while sacrificing signal-to-noise. For illustration purposes, the signal-to-noise benefit that underlies exponentially-weighted sampling can be approximated in the fully-sampled experiment by applying in each indirect dimension an additional matched filter with a decay rate equivalent to that used to generate the nonuniform sampling schedules. Supplementary Figure 1 shows a comparison of the FT spectra processed as shown in Fig. 1b, without and with the additional matched filtering. As expected, when processing with the matched filter, improvements in the SNR are observed (and the SNR is now comparable to the CS-processed data); however, the resolution has already dropped to a point that interferes with the spectral interpretation, even if for the region considered overlap is only moderate. In view of the high complexity of the spectra studied here, a performance comparison conducted under the conditions of comparable and higher resolution for the FT spectrum as shown in Fig. 1 seems more appropriate.

While the 50 % sampling case was used for direct comparison with the FT-reconstructed spectra in Fig. 1, high quality reconstructions can also be obtained by sampling only 40 % of the full data matrix, shown in Fig. 2 (i).

Figure 2 also illustrates the impact of the amount of sampling on the ability of CS to produce faithful reconstructions. Three levels of exponentially-weighted sampling are compared side-by-side; all spectra were reconstructed using the iterative hard thresholding algorithm from data with: (i) 40 % sampling, (ii) 30 % sampling and (iii) 20 % sampling of the full data matrix. The fidelity of the reconstruction remains high with 30 % sampling as illustrated by both planes shown in Fig. 2a, b with an experiment time corresponding to 30 h. Comparing with the FT spectrum recorded over 50 h [see Fig. 1(iii)], a benefit in signal-to-noise is still observable. However, the quality of the reconstruction using 30 % of the data is slightly inferior to 40 %, and additional spurious signals are starting to appear in positions where there should not be any peaks (indicated by grey rectangles in Fig. 2b). In the case of the more complex spectra of the membrane protein investigated here, it is believed that the higher sampling factor of 40 % is beneficial (i.e. corresponding to a 60 % time saving over recording the full data matrix required for FT) and that the inferior spectral quality in the 30 % case can lead to misinterpretations. Figure 2(iii) allows assessment of the reconstruction quality obtained with 20 % sampling. It becomes obvious that by now, the quality of the reconstructed spectra has deteriorated quite significantly. This manifests itself in several ways: in the crowded regions of the spectra, the peaks have started to merge as the resolution is starting to break down; a number of intense artifacts are observed and finally, additional false signals are starting to appear that are comparable in intensity to real peaks; again selected examples are indicated by grey rectangles in Fig. 2b. Clearly, any of this is undesirable in spectra used for structure determination. Nevertheless, as emphasized by (iii) in Fig. 2b, the regions with more intense peaks and less crowding are still reconstructed accurately with no apparent distortions in peak position, shape or intensity, suggesting that, for spectra with higher SNR and/or less overlap than in the case demonstrated here, further undersampling than that deemed acceptable for pSRII will be possible.

To further illustrate the quality of the 40 % sampled CS-reconstructed IT spectrum, Fig. 3 shows several [^{1}H,^{1}H] strip plots for a range of different residue types that are mainly located in the overlapped α-helical structured regions of pSRII. The FT (100 h, 16 scans) and CS (40 h, 16 scans) spectra are shown side by side. The CS reconstructions are recognizable by the thicker frames. It can be seen qualitatively that the spectra look very similar in intensity and furthermore that the peaks in the CS spectrum are recovered with accurate positions and no obvious systematic distortions in peak shape. Sequential series of residues—Ala 91-Gly 92-Leu 93 (at the end of helix C and in the loop connecting to helix D) and Thr 204-Lys 205 (in the core of the transmembrane region)—are also displayed showing, via dashed lines, some of the sequential NOE assignments in order to highlight consistency in the cross peak pattern and observed chemical shifts. Various regions in the indirect ^{1}H dimension are shown in order to demonstrate the fidelity of the reconstruction both close to the diagonal as well as in the side-chain regions. Some long-range NOE cross peaks are also indicated. Residue assignments were obtained based on the CS reconstructed

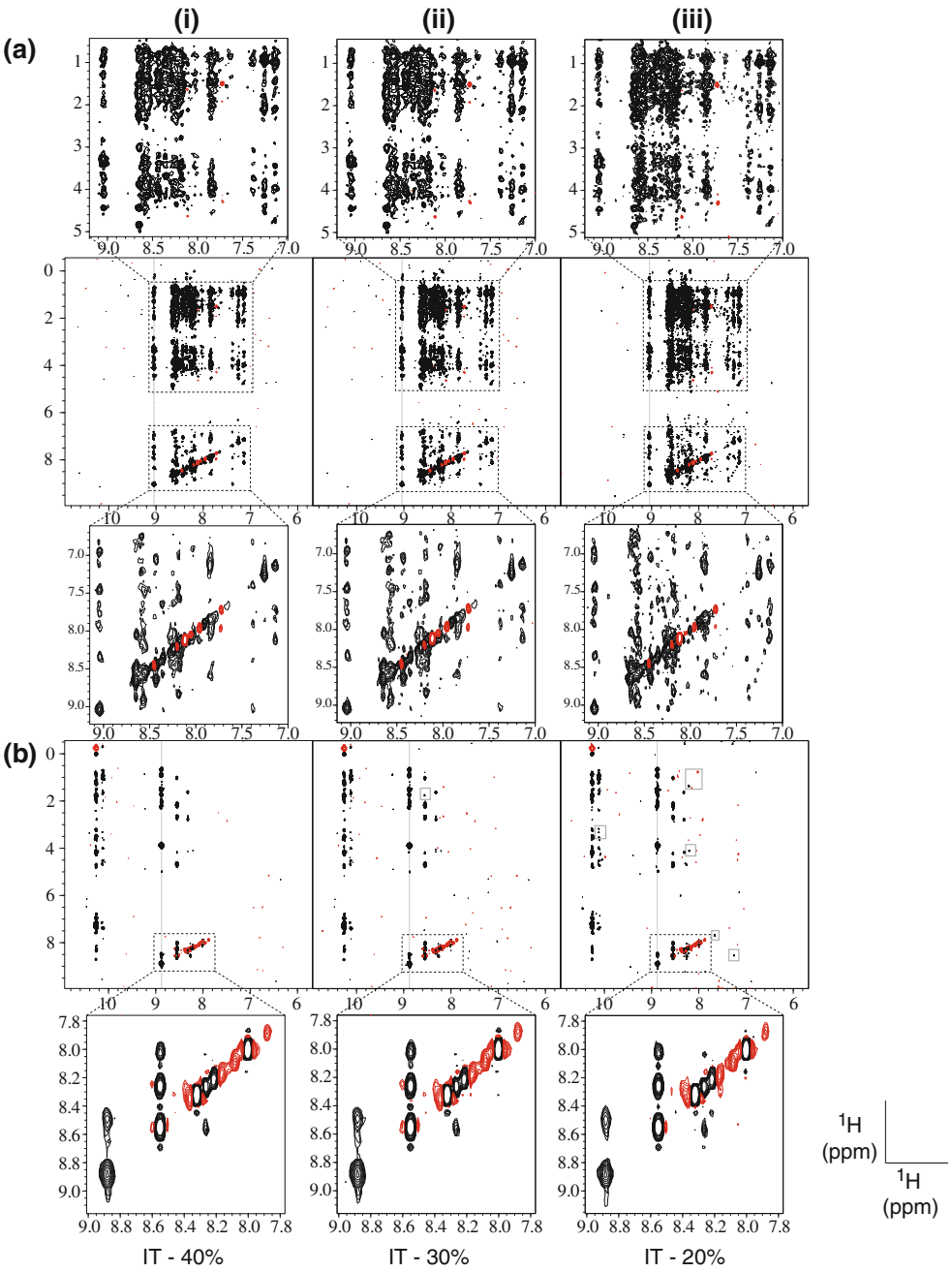

Fig. 2 The 2D [^{1}H,^{1}H] planes extracted at the same positions as shown in Fig. 1, using the iterative hard thresholding algorithm (IT) with different sampling fractions: (**i**) 40 % (40 h); (**ii**) 30 % (30 h); (**iii**) 20 % (20 h). Examples of regions where the reconstructions start to break down are indicated in (**b**) through *grey rectangles*

spectra and successfully verified against the FT processed spectra.

Under the demanding conditions tested here, we show that iterative hard thresholding is able to reliably produce accurate reconstructions of the crowded spectra even from extensively undersampled time-domain data sets. However, a range of other algorithms is also available for the reconstruction of underdetermined systems of linear equations according to (2). Using the 3D NOESY ^{15}N HSQC data sets sampled at 30 and 40 %, respectively, we compare the performance of the iterative soft thresholding (IST) algorithm (see "Methods") and the YALL1 l_1-minimization method (Yang and Zhang 2011) against iterative hard thresholding (IT). The YALL1 solver uses an alternating direction method to solve a range of l_1-norm minimisation models. The algorithm has been reported to reduce the relative error between a fully sampled data set and an undersampled reconstruction faster than other available solvers under conditions of noisy data, to be largely insensitive to the choice of starting point and initial parameters and does not rely on a continuation or line-search technique (Yang and Zhang 2011). Our

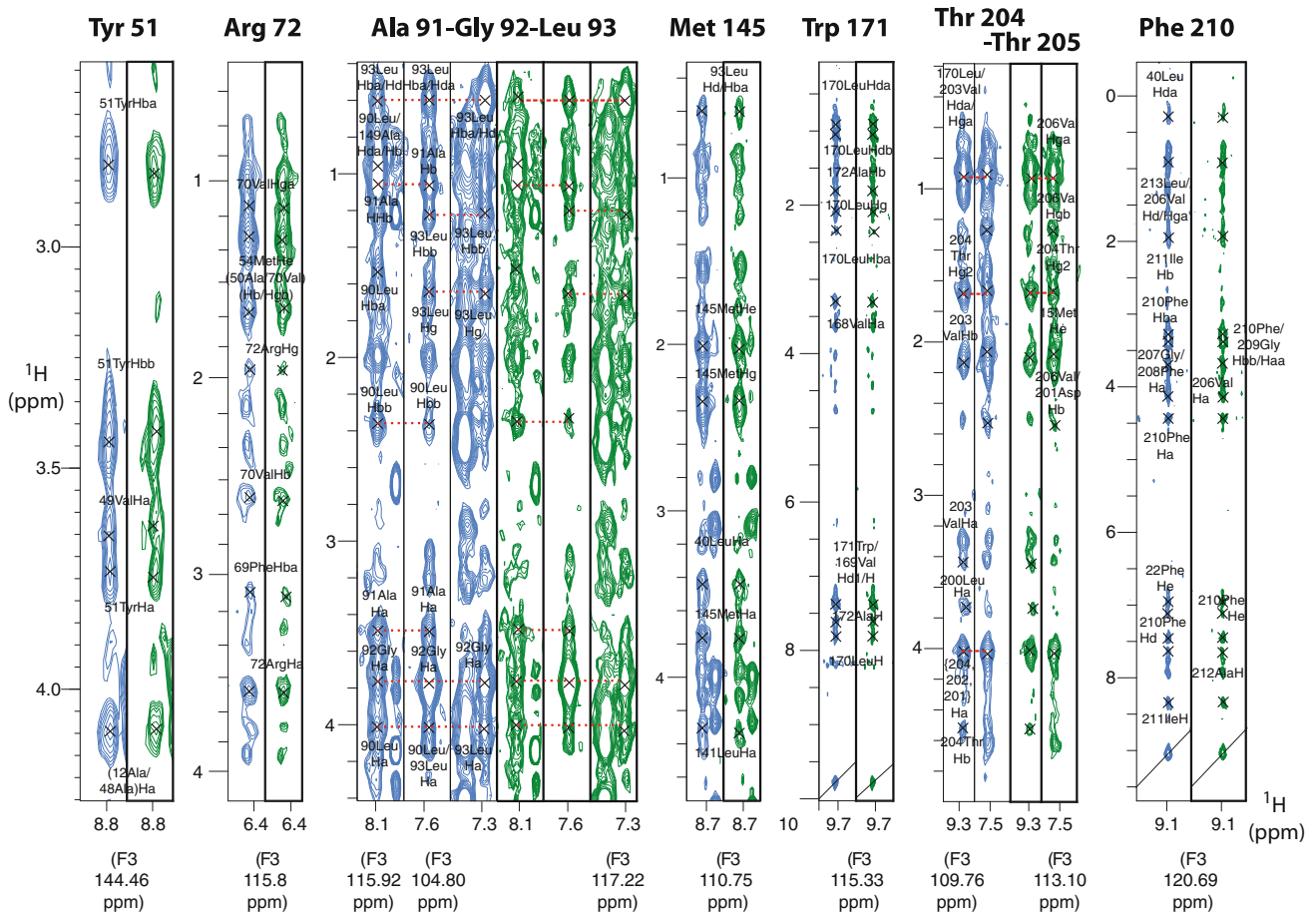

Fig. 3 [^{1}H,^{1}H] (F1,F2) strip plots from a 3D NOESY ^{15}N HSQC recorded on pSRII. The spectra were reconstructed using CS iterative hard thresholding with 40 % sampling (40 h) and Fourier transformation on the fully sampled data (ns = 16, 100 h). The CS reconstructions are recognizable by thicker frames. Cross peak assignments are shown in proximity to the relevant peaks, with *dashed lines* indicating sequential NOEs. *Black diagonal lines* indicate the diagonal peak positions. Labels are not shown on every strip to avoid crowding, however, equivalent labels can be found in adjacent strips

implementation used a model corresponding to (12) (see "Methods").

The spectra of the reconstructions using the different algorithms are compared in Supplementary Figure 2, where the same two planes displayed in Fig. 1 and 2 are shown. An initial qualitative inspection shows the results to be quite similar and underlines the fidelity of the reconstructions using different methods for l_1-norm minimization. All of the spectra shown in Fig. 1, 2 and Supplementary Figure 2 correspond to reconstructions using a relaxed data-matching constraint, i.e. high $\|\mathbf{Ax} - \mathbf{b}\|_2$. When compared on this basis the high similarity of the spectra in Supplementary Figure 2 is apparent and the only significant difference between these algorithms is the reconstruction time, which was <1 h for both IT and IST and in excess of 4 h for the YALL1 algorithm.

It is important to evaluate the impact of varying the constraint ($\|\mathbf{Ax} - \mathbf{b}\|_2$) on the reconstructions. To this purpose in a second set of calculations all three algorithms were used to reconstruct the spectrum with a much tighter constraint, i.e. low $\|\mathbf{Ax} - \mathbf{b}\|_2$ (Table 1). Reconstructions with high and low $\|\mathbf{Ax} - \mathbf{b}\|_2$ are compared in Supplementary Figure 3. The results show that enforcing greater data consistency introduces significant artifacts into the reconstructed spectrum, as would be expected, since the tighter constraint will cause the reconstruction method to fit the noise. This is consistent with an increase in the term $\|\mathbf{x}\|_1$ when data matching is enforced, as shown in Table 1. While IST is not shown in Supplementary Figure 3, IST and IT with low $\|\mathbf{Ax} - \mathbf{b}\|_2$ are comparable. Generally, the difference between the three different reconstruction algorithms is minimal when compared at similar values of $\|\mathbf{Ax} - \mathbf{b}\|_2$.

Reconstruction times for the different algorithms varied from approximately 30 min for the iterative soft and hard thresholding algorithms to approximately 4–5 h for the YALL1 approach (see "Methods" and Table 1 for full details). Based on our experience, the main influence

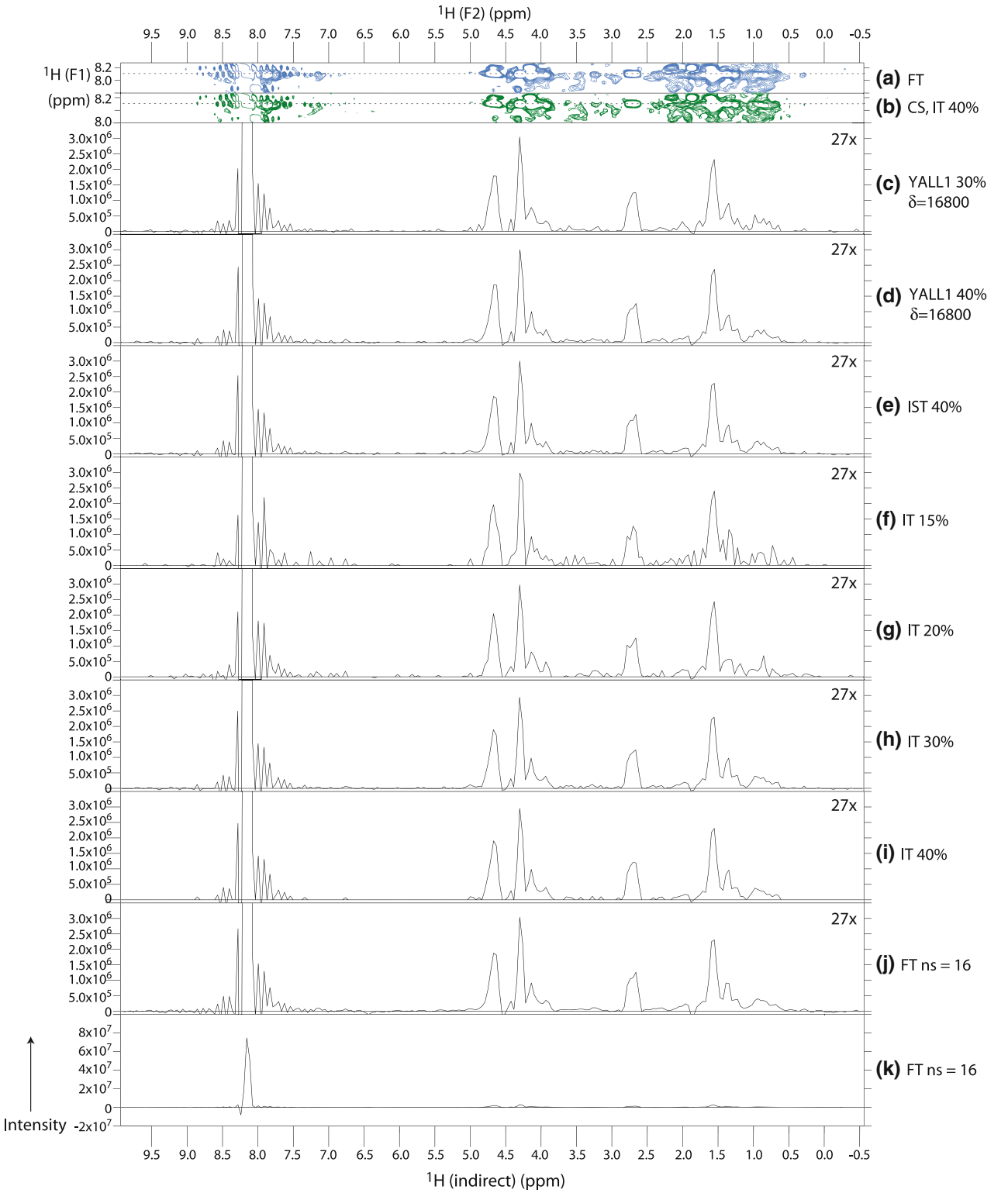

Fig. 4 Intensity comparison of representative 1D slices taken parallel to the indirect (F2) proton dimension, at the amide position of residue Asp 239 (8.153 ppm ^{1}H and 118.727 ppm ^{15}N). In (**a**) and (**b**) the corresponding 2D [F1,F2] strips of the FT and CS (IT 40 %) reconstructions are are displayed, respectively, with the *dashed line* marking the position of the 1D slices. 1D slices: Comparison of different CS algorithms and sampling factors in **(c)** - **(i)** with the Fourier-transformed version (16 scans) shown in **(j)** and **(k)**. The algorithm and sampling factors used are indicated beside each 1D trace. In all cases, the CS reconstructions were carried out using a high value of the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ ($\sim 2 \times 10^4$). The intensity in **(k)** is adjusted to show the diagonal signal, while all the other 1D traces have been scaled by 27 × to emphasize the large dynamic range between diagonal peak and cross peaks

Table 2 Number of observed peaks used to compare the spectral reconstructions: peaks are divided into five categories based on their intensities

	FT	IT 40	IST 40		YALL1 40		IT 30	YALL1 30		IT 20	IT 15
Intensity		(*l*) & (*h*)	(*l*)	(*h*)	(*l*)	(*h*)	(*h*)	(*l*)	(*h*)	(*h*)	(*h*)
$1 \times 10^5 \leq I < 5 \times 10^5$	81	81	79	81	77	78	77	76	73	74	67
	(100 %)	(100 %)	(98 %)	(100 %)	(95 %)	(96 %)	(95 %)	(94 %)	(90 %)	(91 %)	(83 %)
$5 \times 10^5 \leq I < 1 \times 10^6$	88	88	88	88	88	88	88	87	88	88	88
$1 \times 10^6 \leq I < 1 \times 10^7$	80	80	80	80	80	80	80	78	80	80	80
$1 \times 10^7 \leq I < 1 \times 10^8$	26	26	26	26	26	26	26	26	26	26	26
$I \geq 1 \times 10^8$	13	13	13	13	13	13	13	13	13	13	13
Overall	288	288	286	288	284	285	284	280	280	281	274

The numbers of peaks in the lowest intensity category are a measure of the peak detectability. Some of the reconstructions were repeated for low (*l*) and high (*h*) constraint values $\|\mathbf{Ax} - \mathbf{b}\|_2$ (for exact values of the constraint, see Table 1); peak numbers are shown side-by-side

affecting the length of the reconstruction times for the iterative hard thresholding approach is the choice of the stopping criteria (see "Methods"). Initially a reconstruction was performed based on the assumption, derived from CS theory, that the sparsity of the spectrum is related to the amount of undersampling possible. This criterion was found to be consistent with criterion (ii) which was based on the apparent noise level of the residual. Note that evaluation of the noise-based stopping criterion in the frequency domain removes any bias introduced through window functions applied in the time domain. Enforcing greater data matching significantly lengthened the reconstruction time for iterative hard thresholding, such that to reach the lower value of $\|\mathbf{Ax} - \mathbf{b}\|_2 < 2$ took $\sim$3 h, compared to 30 min with the looser constraint. The IST algorithm took $\sim$0.5 h to converge on a constant value, corresponding to $\|\mathbf{Ax} - \mathbf{b}\|_2 \sim 2 \times 10^4$. For the IST implementation, reducing the value of the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ with constant λ significantly lengthens the reconstruction. Therefore, to reconstruct the spectrum with a tight data matching constraint, the algorithm was repeated with decreasing values of λ, resulting in longer reconstruction times for lower values ($\sim$2.5 h for $\|\mathbf{Ax} - \mathbf{b}\|_2 \sim 2$). Further time-savings for the IST reconstructions may also be achieved using more advanced IST algorithms such as FISTA (Beck and Teboulle 2009). For YALL1, varying the weighting of the $\|\mathbf{Ax} - \mathbf{b}\|_2$ term did not affect the reconstruction time significantly, taking $\sim$4–5 h in each case. Further attempts to reduce the reconstruction time to be comparable with the iterative thresholding methods were not made.

A fundamental problem of all non-linear methods is the correct recovery of peak intensities when the dynamic range is large. Figure 4 shows representative 1D traces taken along the F2 dimension (indirect ^{1}H) from the various CS reconstructions, compared with the full FT reconstruction (100 h), at the F1, F3 amide position of Asp 239. The extent of the large dynamic range reproduced by the reconstructions is obvious. The 1D spectra (c–j) show a 27-times amplified version relative to (k), to illustrate the intensity of the cross peaks in the reconstructions. The faithfulness of the reconstructions and ability to recreate peaks accurately over the large dynamic range is apparent, although as noted previously, the quality declines for lower sampling factors.

In order to further assess the quantitative aspects of the reconstructions with respect to intensities, weak-peak detectability and positions of peaks, a comprehensive peak list was produced using the Fourier-transformed 3D NOESY ^{15}N HSQC spectrum of the full time domain matrix. The peak-picking algorithm implemented in the software package CCPN Analysis (Vranken et al. 2005) was used to pick all peaks above a threshold set to 1.25 $\times$ the average noise level (8.4×10^4) corresponding to a contour level display of 2D planes which was found suitable for manual spectral assignment. To allow a more comprehensive representative statistical analysis of the reconstructions based on many different peak intensities, the resulting master list of peaks was divided into five intensity bins, with each bin containing peaks with intensities normalized with respect to the noise level in the range I: $1.2 \leq I < 6$, $6 \leq I < 12$, $12 \leq I < 120$, $120 \leq I < 1{,}200$, $I \geq 1{,}200$. Within each bin, 100 peaks were randomly chosen, or if less were available, all peaks were included. The selected peaks were then briefly manually inspected. Obvious noise signals or peaks where the intensity was strongly distorted by the presence of a neighboring peak were removed from the list. The final filtered set of intensity bins in the FT spectrum consisted of a total of 288 peaks, which were used to identify the corresponding peaks in the CS reconstructions. Total peak numbers for the different reconstructions are given in Table 2. For each peak the true maximum in the CS spectra was adjusted following a grid search for the local intensity maximum within a few points of the original

value. The final coordinates and intensities were then compared against the FT standard.

The intensities in the various CS reconstructions compared with the fully-sampled FT spectrum (ns = 16) are displayed in Fig. 5. Each point corresponds to an individual peak; the lines indicate equal intensities in both the FT standard and CS reconstruction. For clarity, the different correlations are offset along the y axis and intensities are also expressed as a function of the SNR. It is clearly visible that all the reconstructions are very linear and that the intensity accuracy of all the reconstructions improves with increasing S/N until the different methods start to produce very similar results. At S/N values below five, differences between the methods are starting to show, however, a linear relationship is still maintained regardless of the algorithm used. While for many of the methods the intensity differences are relatively small, it is noticeable that the spread in intensities increases as the amount of sampling is reduced, again reflecting the prior qualitative observations. Nevertheless, reconstructions based on 30 % sampling are largely able to preserve the true intensities well, but lower sampling factors quickly become unreliable (20 and 15 %) and struggle to give accurate reproductions over an increasingly wide range of data, although the accuracy of the more intense peaks still remains high.

Figure 6a emphasizes the trend where reducing the sampling factor increases the lower limit for which peak intensities can still be faithfully reconstructed. All comparisons are made for similar high constraint values and under comparable regularization (Table 1). The percentage root mean squared (RMS) deviation between the FT

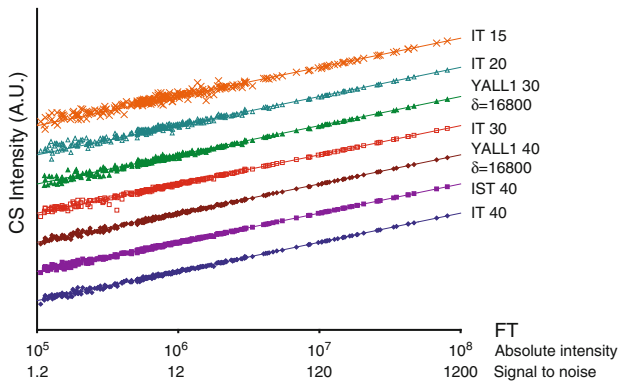

Fig. 5 Peak intensity comparisons as a function of the signal-to-noise ratio, between the Fourier-transformed spectrum and the CS reconstructed spectra using different reconstruction algorithms and sampling factors. All reconstructions used a high value of the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ ($\sim 2 \times 10^4$). Every data point represents **a** experimentally determined peak intensity (see Table 2). *Lines* indicate where the two compared methods would give equal intensities. For clarity, the different comparisons are offset along the *y* axis of the plot. Comparisons are given as a function of absolute peak intensity and signal-to-noise ratio

intensities and the CS reconstructed intensities based on (13) (see "Methods") indicates the measure of precision that has been achieved when considering all the peaks within a given intensity bin. With 40 % sampling the RMS values in the lowest intensity bin remains around 10 %, which is very respectable when considering that the S/N ratio for most of these peaks is significantly below five. With increasing intensities, the RMS deviations drop rapidly for all sampling regimes, with the exception of the two lowest sampling regimes IT 20 and IT 15, which only become comparable with the other sampling factors for intensities above 1×10^6 or 1×10^7 respectively. Figure 6a illustrates also that whilst the YALL1 algorithm shows comparable performance to the iterative thresholding methods in the lowest intensity bin, the error in the reconstruction does not decrease quite as rapidly for the higher intensity bins compared to e.g. the IT method. Furthermore, at 30 % sampling, the YALL1 reconstruction remains higher over the first three intensity bins. The IST method performs well over all intensity bins and whilst the differences between the methods are small, with the biggest effect coming from the sampling schedule, we tentatively suggest, based on extensive visual inspection as well as the quantitative analysis that the iterative thresholding methods result in slightly higher-quality reconstructions. To illustrate that this trend is not simply a consequence of the choice of constraint and regularization values (see Table 1), IT, IST and YALL1 reconstructions are shown in Fig. 6b using comparable values for two cases with either a low or a high constraint. The trends observed in Fig. 6a are largely reproduced and it becomes clear that although introducing a tight constraint on $\|\mathbf{Ax} - \mathbf{b}\|_2$ does increase the artifact level in the reconstructed spectra (see Supplementary Figure 3), it exercises a relatively small influence on the RMS deviation in intensity for the true peaks. The results emphasize that the reconstructions show an inherent robustness towards the choice of constraint.

The mean errors in the chemical shift positions of peaks for the indirect dimensions between the FT and CS reconstructions using 40 % and 20 % sampling are shown in Fig. 7a, as determined for each of the five intensity bins. Other reconstructions are omitted for clarity. The differences in the chemical shift positions were calculated on a peak-by-peak basis as given in (14) (see "Methods"). A dependence on the amount of sampling is very small. The measured chemical shift differences are of a similar size to the digital resolution of the individual indirect dimensions (0.04 ppm for ^{1}H and 0.1 ppm for ^{15}N) in the spectra, confirming the high quality of recovery of peak positions. Chemical shift deviations were of the same order for all the investigated reconstructions. Figure 7b illustrates the effect of varying the data matching term, $\|\mathbf{Ax} - \mathbf{b}\|_2$ for the individual methods at 30 and 40 % sampling. Similar to

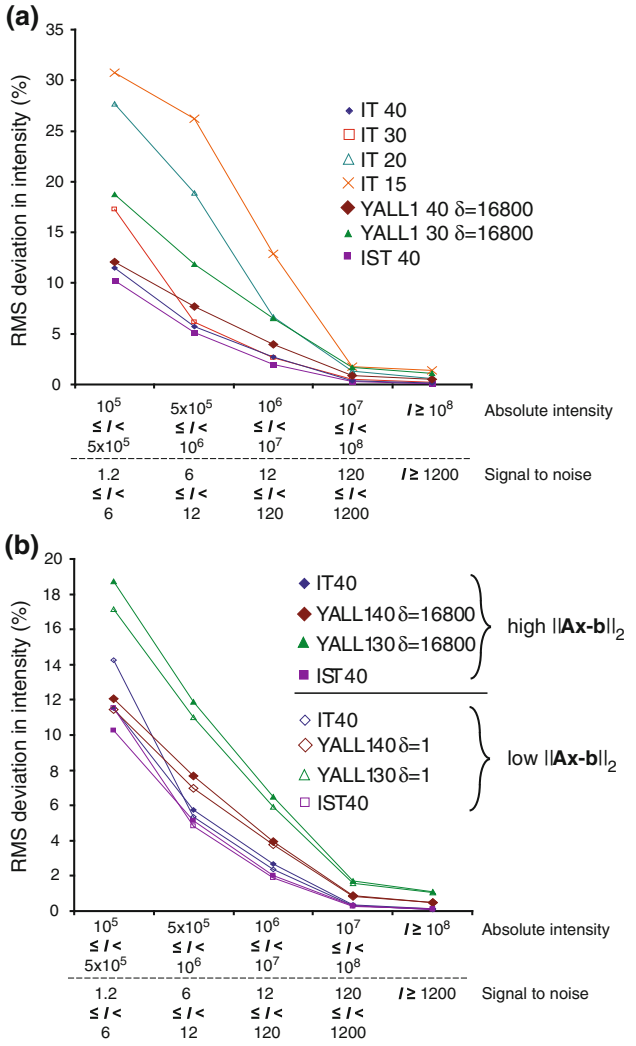

Fig. 6 Peak intensity root mean squared (RMS) deviation in percentage for five different intensity ranges (bins). The CS reconstructions are compared with the Fourier-transformed spectrum. **a** Shows different algorithms and sampling factors using a high value of the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ ($\sim 2 \times 10^4$). **b** Compares the results on varying the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ (for values see Table 1). For details of the formula used for the calculation of the RMS deviation see "Methods"

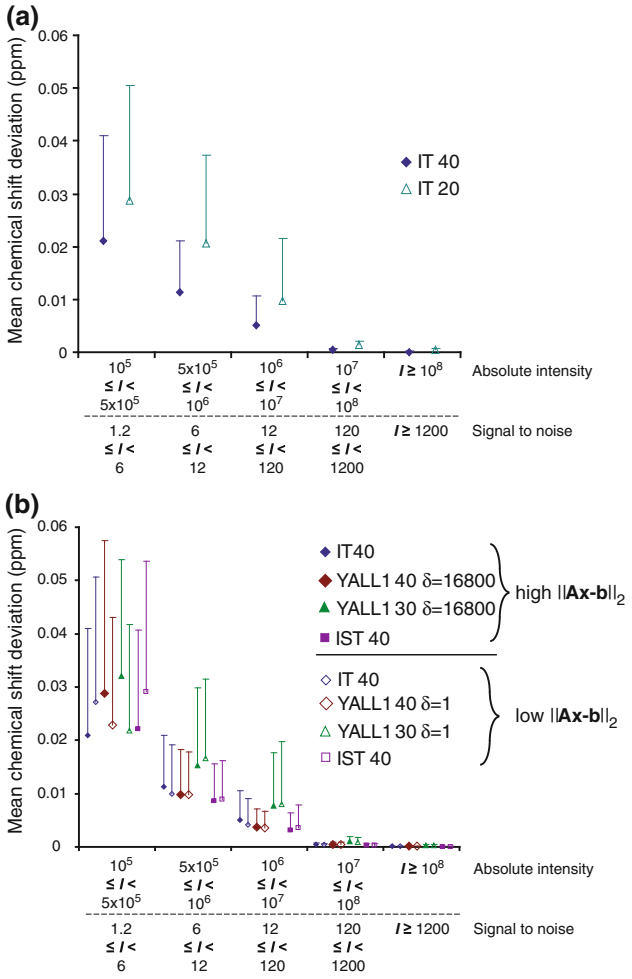

Fig. 7 Mean chemical shift deviation in peak positions as a function of the five considered intensity ranges (intensity bins). **a** Compares IT reconstructions at different sampling factors (other reconstructions are omitted for clarity), using a high value of the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ ($\sim 2 \times 10^4$). **b** Compares the results on varying the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ (for values see Table 1). Weighted averages of the two indirect dimensions in the CS reconstructions are compared with the Fourier transformed spectrum. Error bars indicate the standard deviation of the mean. For further details, see "Methods"

our prior observations for the intensitiy behaviour, these results confirm that even though a tight constraint on $\|\mathbf{Ax} - \mathbf{b}\|_2$ increases the artifact level in the reconstructed spectra, it exercises a relatively small influence on the chemical shift deviation.

As a major drawback of many nonparametric regularization methods, the weakest signals are the ones most prone to be reduced into the noise and hence to become undetectable. In our analysis the number of peaks found in the lowest intensity category in Table 2 can be considered as a measure of detectability or completeness for each of the reconstructions. It can be seen that for 40 % sampling, the two iterative thresholding methods recover all the expected peaks endorsing the reliability of these reconstructions. YALL1 recovers slightly less, 96 % for the same amount of sampling (and high $\|\mathbf{Ax} - \mathbf{b}\|_2$) while the number of peaks varies only slightly with the δ value. More peaks are lost at the lower sampling factors, where 91 and 83 % of the peaks are observed for IT 20 and IT 15, respectively. Recovering 95 % of the peaks, the IT method is still competitive at 30 % sampling. These values seem largely unaffected by changes in the l_1-norm and constraint parameter from $\|\mathbf{Ax} - \mathbf{b}\|_2 < 2$ to 2×10^4 (Table 1), with the largest effects coming from variations in sampling level. At 30 % sampling the YALL1 reconstruction is generally of lower quality than the iterative thresholding

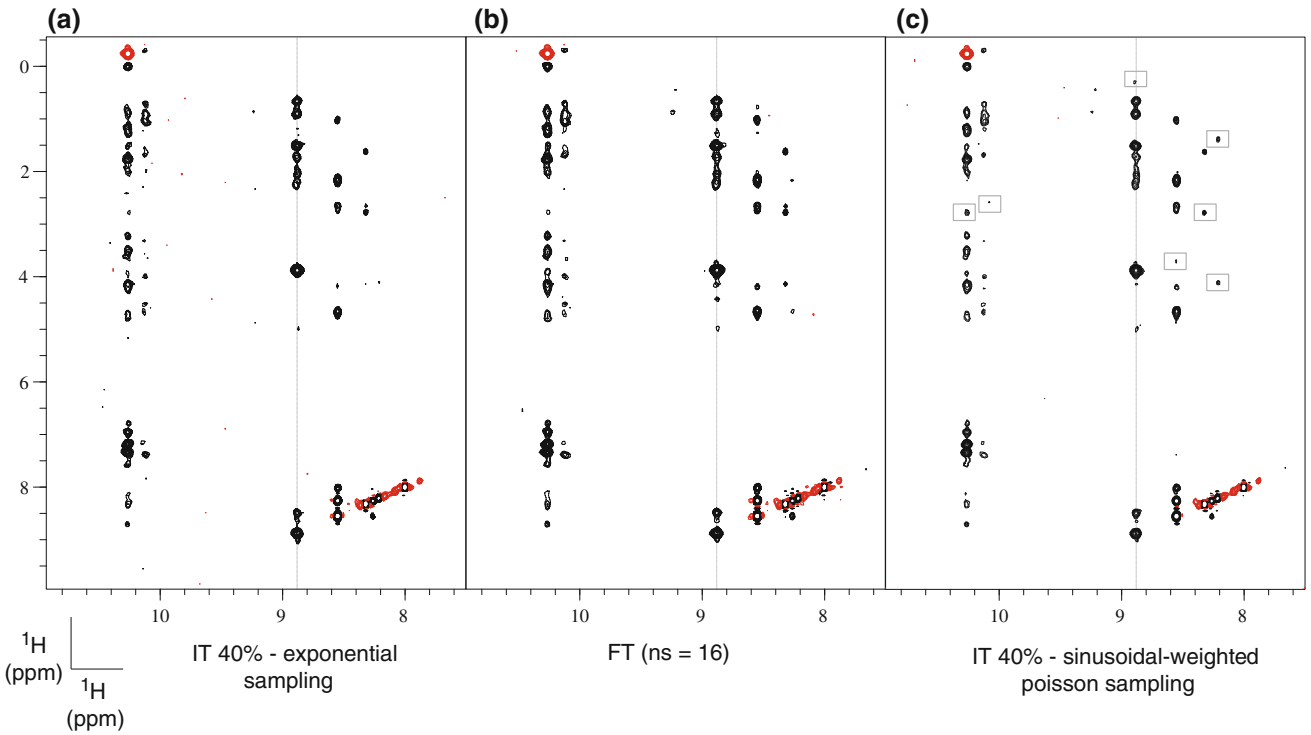

Fig. 8 Influence of the sampling scheme on the quality of the spectral reconstruction: comparing the results of **a** exponential weighting, **b** full sampling, and **c** sinusoidal-weighted Poisson sampling. Data sets in (**a**) and (**c**) were sampled to 40 % and reconstructed with iterative hard thresholding, while (**b**) was Fourier-transformed. Differences in the Poisson-sampled reconstruction are highlighted by *grey squares*. The figure shows the F1, F2 [^{1}H,^{1}H] plane taken at the ^{15}N position 127.58 ppm (Ala 64), corresponding to part (**b**) in Figs. 1, 2 and Supplementary Figure 2 and 3. Both reconstructions used a high value for the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ ($\sim 2 \times 10^4$)

reconstruction (see also Fig. 6). However, overall the differences are relatively small and should not be overinterpreted.

Variation of the constraint value in the range tested shows relatively little significant effect on intensity and shift positions of peaks (Figs. 6b and 7b) confirming the inherent robustness of the reconstructions. The major effect observed is the increase in artifact level already discussed (see also Supplementary Figures 2 and 3); consequently we favour the choice of a higher value for the constraint $\|\mathbf{Ax} - \mathbf{b}\|_2$ which results in a lower l_1-norm.

The choice of sampling pattern can have a dramatic effect on the outcome of non-FT reconstructions but a systematic investigation is beyond the scope of this work. In this study, NOESY data obtained through exponentially-weighted random sampling (see "Methods") leads to very reliable and robust reconstructions as illustrated. Recently, sinusoidal-weighted Poisson sampling has been reported to result in high quality reconstructions (Hyberts et al. 2011; Hyberts et al. 2010). We compared the psf-optimized exponentially-weighted sampling scheme used in this work with weighted Poisson sampling using the iterative hard thresholding CS reconstruction with 40 % sampling of the data (IT 40 %). The results of a direct comparison of the two sampling types are shown in Fig. 8. Although similar in overall appearance, in our case weighted-Poisson sampling resulted in reconstructions with slightly more artifacts when using IT. Notable spurious deviations in the reconstruction of the Poisson data are highlighted with grey rectangles. Although in our hands this would seem to favor exponential weighting, a more detailed comparison will be required for thorough assessment.

To demonstrate the robustness of CS reconstruction to noise, the iterative hard thresholding method was used to reconstruct data sets recorded with only 8 scans. With sampling levels of 40 and 50 %, this corresponds to experiment times of 20 h and 25 h, respectively. A comparison with the corresponding Fourier-transformed spectra (8 and 16 scans) is given in Fig. 9, where the three sequential residues Ala 91-Gly 92-Leu 93 are shown representative of a crowded spectral environment. The same region based on a 16 scan reconstruction was shown in Fig. 3. 1D spectra in Fig. 9 illustrate the basic signal-to-noise range of the experiments recorded with 8 and 16 scans, resulting in experiment times of 50 h and 100 h for FT processing. It is clear that at this lower signal-to-noise ratio, the recovery is not as high quality as in Fig. 3. However, as a test, automatic peak picking of the planes was performed, again without any manual interference: The picked peaks and dashed lines indicate a substantial

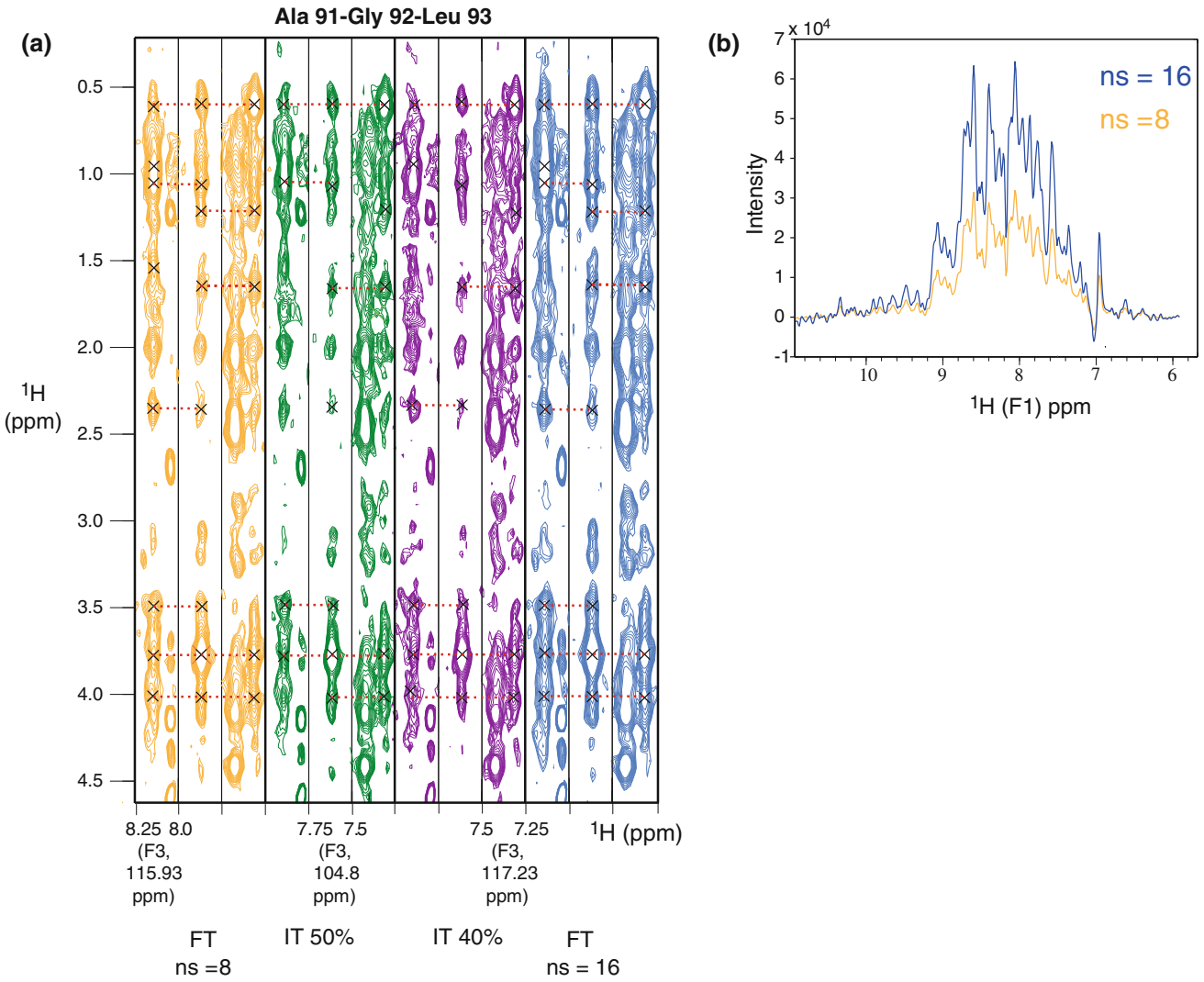

Fig. 9 Robustness to noise: **a** 2D [^{1}H,^{1}H] strip plots of the sequential residues Ala 91-Gly-92-Leu 93 are shown for the iterative hard thresholding reconstructions of the 8 scan data with 50 or 40 % sampling, compared with the 8 scan and 16 scan Fourier transformed data. Peaks identified by a cross were automatically picked by the peak-picking routine in CCPN Analysis. Assignments are omitted for clarity but can be found in Fig. 3. *Dashed lines* indicate sequential NOE assignments. **b** 1D projections illustrate the basic sensitivity of the experiments recorded with 8 and 16 scans, respectively, which would correspond to experiment times of 50 and 100 h for a fully sampled data matrix

number of sequential NOE assignments and illustrate that the reconstructed spectra are still of quality suitable for assignment.

Conclusion

We show the suitability of compressed sensing as a methodology to reconstruct crowded 3D NOESY ^{15}N HSQC spectra recorded on a large protein. In combination with the use of sparse data sampling, this can lead to substantial time gains. Conservative approaches can result in a 60 % time saving, achieved by recording only 40 % of the indirect data points sampled randomly with an exponential T_2 weighted bias. Recording 40 % of the data matrix, we obtained high quality spectra of a large membrane protein solubilized in detergent-micelles, despite considerable overlap; consequently this approach enables improvements in signal-to-noise and resolution, typically limiting factors. The same approach can equally be applied to other experiments such as e.g. 3D ^{13}C-separated NOESY making the combination of sparse sampling and compressed sensing reconstruction a powerful tool for the study of large biomolecules. Sampling as little as 20 % also produced acceptable results; when applied to less demanding cases i.e. smaller proteins, much larger time-savings are obtainable. Several l_1-minimization algorithms were tested which produced comparable results, possibly

with iterative thresholding being slightly superior and providing shorter reconstruction times. Over the range tested, the influence of the regularization ($\|\mathbf{x}\|_1$) and constraint ($\|\mathbf{Ax} - \mathbf{b}\|_2$) parameters on the final reconstruction quality was found to be relatively minor, although enforcing too tight a constraint on the data matching term ($\|\mathbf{Ax} - \mathbf{b}\|_2$) does increase the level of artifacts in the spectrum. The CS approach is demonstrated to be robust to noise and less demanding on signal-to-noise than corresponding experiments recorded for FT processing. Overall the robustness to noise, high quality of the reconstructed spectra and the substantial time savings obtainable make the CS methodology a considerable asset for the study of large biomolecular systems.

References

Atreya HS, Szyperski T (2004) G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. Proc Natl Acad Sci USA 101:9642–9647. doi:10.1073/pnas.0403529101

Barna JCJ, Laue ED, Mayger MR et al (1987) Exponential sampling, an alternative method for sampling in two-dimensional NMR experiments. J Magn Reson 73:69–77

Beck A, Teboulle M (2009) A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J Imag Sci 2:183–202. doi:10.1137/080716542

Bredies K, Lorenz DA (2008) Iterated hard shrinkage for minimization problems with sparsity constraints. SIAM J Sci Comput 30:657–683. doi:10.1137/060663556

Candes EJ, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inf Theory 52:489–509. doi:10.1109/TIT.2005.862083

Chylla RA, Markley JL (1995) Theory and application of the maximum likelihood principle to NMR parameter estimation of multidimensional NMR data. J Biomol NMR 5:245–258. doi:10.1007/BF00211752

Coggins BE, Zhou P (2008) High resolution 4-D spectroscopy with sparse concentric shell sampling and FFT-CLEAN. J Biomol NMR 42:225–239. doi:10.1007/s10858-008-9275-x

Daubechies I, Defrise M, De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun Pure Appl Math 57:1413–1457. doi:10.1002/cpa.20042

Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory 52:1289–1306. doi:10.1109/TIT.2006.871582

Drori I (2007) Fast l_1 minimization by iterative thresholding for multidimensional NMR Spectroscopy. EURASIP J Adv Sig Pr 2007:1–11. doi:10.1155/2007/20248

Eddy MT, Ruben D, Griffin RG, Herzfeld J (2012) Deterministic schedules for robust and reproducible non-uniform sampling in multidimensional NMR. J Magn Reson 214:296–301

Fiaux J, Bertelsen EB, Horwich AL, Wüthrich K (2002) NMR analysis of a 900 K GroEL-GroES complex. Nature 418:207–211. doi:10.1038/nature00860

Freeman R, Kupče E (2003) New methods for fast multidimensional NMR. J Biomol NMR 27:101–114. doi:10.1023/A:1024960302926

Frydman L, Scherf T, Lupulescu A (2002) The acquisition of multidimensional NMR spectra within a single scan. Proc Natl Acad Sci USA 99:15858–15862. doi:10.1073/pnas.252644399

Gautier A, Mott HR, Bostock MJ et al (2010) Structure determination of the seven-helix transmembrane receptor sensory rhodopsin II by solution NMR spectroscopy. Nat Struct Mol Biol 17:768–774. doi:10.1038/nsmb.1807

Hiller S, Garces RG, Malia TJ et al (2008) Solution structure of the integral human membrane protein VDAC-1 in detergent micelles. Science 321:1206–1210. doi:10.1126/science.1161302

Hoch JC, Stern AS, Donoho DL, Johnstone IM (1990) Maximum entropy reconstruction of complex (phase-sensitive) spectra. J Magn Reson 86:236–246. doi:10.1016/j.jmr.2007.07.008

Hoch JC, Maciejewski MW, Filipovic B (2008) Randomization improves sparse sampling in multidimensional NMR. J Magn Reson 193:317–320. doi:10.1016/j.jmr.2008.05.011

Holland DJ, Malioutov DM, Blake A et al (2010) Reducing data acquisition times in phase-encoded velocity imaging using compressed sensing. J Magn Reson 203:236–246. doi:10.1016/j.jmr.2010.01.001

Holland DJ, Bostock MJ, Gladden LF, Nietlispach D (2011) Fast multidimensional NMR spectroscopy using compressed sensing. Angew Chem Int Ed 50:6548–6551. doi:10.1002/anie.201100440

Hu S, Lustig M, Chen AP et al (2008) Compressed sensing for resolution enhancement of hyperpolarized ^{13}C flyback 3D-MRSI. J Magn Reson 192:258–264. doi:10.1016/j.jmr.2008.03.003

Hyberts SG, Takeuchi K, Wagner G (2010) Poisson-gap sampling and forward maximum entropy reconstruction for enhancing the resolution and sensitivity of protein NMR data. J Am Chem Soc 132:2145–2147. doi:10.1021/ja908004w

Hyberts SG, Arthanari H, Wagner G (2011) Applications of non-uniform sampling and processing. Top Curr Chem. doi:10.1007/128

Hyberts SG, Milbradt AG, Wagner AB et al (2012) Application of iterative soft thresholding for fast reconstruction of NMR data non-uniformly sampled with multidimensional Poisson Gap scheduling. J Biomol NMR 52:315–327. doi:10.1007/s10858-012-9611-z

Kazimierczuk K, Orekhov VY (2011) Accelerated NMR spectroscopy by using compressed sensing. Angew Chem Int Ed 50:5556–5559. doi:10.1002/anie.201100370

Kazimierczuk K, Koźmiński W, Zhukov I (2006) Two-dimensional Fourier transform of arbitrarily sampled NMR data sets. J Magn Reson 179:323–328. doi:10.1016/j.jmr.2006.02.001

Kazimierczuk K, Zawadzka A, Koźmiński W (2008) Optimization of random time domain sampling in multidimensional NMR. J Magn Reson 192:123–130. doi:10.1016/j.jmr.2008.02.003

Kazimierczuk K, Stanek J, Zawadzka-Kazimierczuk A, Koźmiński W (2010) Random sampling in multidimensional NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 57:420–434. doi:10.1016/j.pnmrs.2010.07.002

Kim HJ, Howell SC, Van Horn WD et al (2009) Recent advances in the application of solution NMR spectroscopy to multi-span integral membrane proteins. Prog Nucl Magn Reson Spectrosc 55:335–360. doi:10.1016/j.pnmrs.2009.07.002

Kupce E, Freeman R (2003) Projection-reconstruction of three-dimensional NMR spectra. J Am Chem Soc 125:13958–13959. doi:10.1021/ja038297z

Kupce E, Freeman R (2004) Projection-reconstruction technique for speeding up multidimensional NMR spectroscopy. J Am Chem Soc 126:6429–6440. doi:10.1021/ja049432q

Kupce E, Nishida T, Freeman R (2003) Hadamard NMR spectroscopy. Prog Nucl Magn Reson Spectrosc 42:95–122. doi:10.1016/S0079-6565(03)00022-0

Logan BF (1965) Properties of high-pass signals. Ph.D. Thesis, Columbia University, New York

Lustig M, Donoho DL, Pauly JM (2007) Sparse MRI: the application of compressed sensing for rapid MR imaging. Magn Reson Med 58:1182–1195. doi:10.1002/mrm.21391

Mandelshtam VA (2000) The multidimensional filter diagonalization method: I. Theory and numerical implementation. J Magn Reson 144:343–356. doi:10.1006/jmre.2000.2023

Marion D (2005) Fast acquisition of NMR spectra using Fourier transform of non-equispaced data. J Biomol NMR 32:141–150. doi:10.1007/s10858-005-5977-5

Marion D (2006) Processing of ND NMR spectra sampled in polar coordinates: a simple Fourier transform instead of a reconstruction. J Biomol NMR 36:45–54. doi:10.1007/s10858-006-9066-1

Marion D, Ikura M, Tschudin R, Bax A (1989) Rapid recording of 2D NMR spectra without phase cycling. Application to the study of hydrogen exchange in proteins. J Magn Reson 85:393–399

Mobli M, Stern AS, Hoch JC (2006) Spectral reconstruction methods in fast NMR: reduced dimensionality, random sampling and maximum entropy. J Magn Reson 182:96–105. doi:10.1016/j.jmr.2006.06.007

Natarajan BK (1995) Sparse approximate solutions to linear systems. SIAM J Comput 24:227–234. doi:10.1137/S0097539792240406

Nietlispach D, Gautier A (2011) Solution NMR studies of polytopic α-helical membrane proteins. Curr Opin Struct Biol 21:497–508. doi:10.1016/j.sbi.2011.06.009

Orekhov VY, Ibraghimov IV, Billeter M (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60. doi:10.1023/A:1011234126930

Otazo R, Kim D, Axel L, Sodickson DK (2010) Combination of compressed sensing and parallel imaging for highly accelerated first-pass cardiac perfusion MRI. Magn Reson Med 64:767–776. doi:10.1002/mrm.22463

Pervushin K, Riek R, Wider G, Wüthrich K (1997) Attenuated T-2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci USA 94:12366–12371

Piotto M, Saudek V, Sklenář V (1992) Gradient-tailored excitation for single-quantum NMR spectroscopy of aqueous solutions. J Biomol NMR 2:661–665. doi:10.1007/BF02192855

Rovnyak D, Frueh DP, Sastry M et al (2004) Accelerated acquisition of high resolution triple-resonance spectra using non-uniform sampling and maximum entropy reconstruction. J Magn Reson 170:15–21. doi:10.1016/j.jmr.2004.05.016

Rovnyak D, Sarcone M, Jiang Z (2011) Sensitivity enhancement for maximally resolved two-dimensional NMR by nonuniform sampling. Magn Reson Chem 49:483–491. doi:10.1002/mrc.2775

Schmieder P, Stern A, Wagner G, Hoch J (1993) Application of nonlinear sampling schemes to COSY-type spectra. J Biomol NMR 3:569–576. doi:10.1007/BF00174610

Schmieder P, Stern AS, Wagner G, Hoch JC (1994) Improved resolution in triple-resonance spectra by nonlinear sampling in the constant-time domain. J Biomol NMR 4:483–490. doi:10.1007/BF00156615

Sprangers R, Velyvis A, Kay LE (2007) Solution NMR of supramolecular complexes: providing new insights into function. Nat Methods 4:697–703. doi:10.1038/nmeth1080

Stern AS, Donoho DL, Hoch JC (2007) NMR data processing using iterative thresholding and minimum l_1-norm reconstruction. J Magn Reson 188:295–300. doi:10.1016/j.jmr.2007.07.008

Tugarinov V, Kay LE, Ibraghimov IV, Orekhov VY (2005) High-resolution four-dimensional ^{1}H–^{13}C NOE spectroscopy using methyl-TROSY, sparse data acquisition, and multidimensional decomposition. J Am Chem Soc 127:2767–2775. doi:10.1021/ja044032o

Vranken WF, Boucher W, Stevens TJ et al (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59:687–696. doi:10.1002/prot.20449

Yang J, Zhang Y (2011) Alternating direction algorithms for l_1-problems in compressive sensing. SIAM J Sci Comput 33:250–278. doi:10.1137/090777761