ARTICLE

Automatic maximum entropy spectral reconstruction in NMR

Mehdi Mobli · Mark W. Maciejewski ·
Michael R. Gryk · Jeffrey C. Hoch

Abstract Developments in superconducting magnets, cryogenic probes, isotope labeling strategies, and sophisticated pulse sequences together have enabled the application, in principle, of high-resolution NMR spectroscopy to biomolecular systems approaching 1 megadalton. In practice, however, conventional approaches to NMR that utilize the fast Fourier transform, which require data collected at uniform time intervals, result in prohibitively lengthy data collection times in order to achieve the full resolution afforded by high field magnets. A variety of approaches that involve nonuniform sampling have been proposed, each utilizing a non-Fourier method of spectrum analysis. A very general non-Fourier method that is capable of utilizing data collected using any of the proposed nonuniform sampling strategies is maximum entropy reconstruction. A limiting factor in the adoption of maximum entropy reconstruction in NMR has been the need to specify non-intuitive parameters. Here we describe a fully automated system for maximum entropy reconstruction that requires no user-specified parameters. A web-accessible script generator provides the user interface to the system.

Keywords Nonuniform sampling ·
Maximum entropy reconstruction ·
Reduced dimensionality · Sparse sampling ·
Multidimensional NMR

Mehdi Mobli and Mark W. Maciejewski contributed equally to this work.

M. Mobli · M. W. Maciejewski · M. R. Gryk · J. C. Hoch (✉)
Department of Molecular, Microbial, and Structural Biology,
University of Connecticut Health Center, 263 Farmington Ave.,
Farmington, CT 06030-3305, USA
e-mail: hoch@uchc.edu

Introduction

Development of very high field magnets, cryogenically cooled probes, and sophisticated multidimensional NMR experiments has enabled the application of solution NMR to larger and more complex biomacromolecules. It is now widely recognized, however, that conventional techniques in which the indirect time dimensions are sampled at uniformly-spaced intervals makes it impossible to achieve the potential resolution afforded by high field magnets in a reasonable amount of time (Kupče and Freeman 2003; Malmodin and Billeter 2005; Jaravine et al. 2006). A variety of alternative sparse sampling strategies have been proposed that substantially reduce data collection time, together with special methods for computing frequency spectra of nonuniformly sampled data. Beyond the implications for more efficient use of expensive instrumentation, these methods are indispensable for obtaining sufficient resolution for the study of complex biomolecular systems or samples that are marginally soluble or fleetingly stable.

Recently considerable attention has been directed toward reduced dimensionality (RD) experiments (including G-matrix Fourier Transform (GFT) and Back Projection Reconstruction (BPR)) which all have in common the coupling of evolution periods in the indirect dimensions of multidimensional experiments (Bodenhausen and Ernst 1981; Szyperski et al. 1993; Kupče and Freeman 2003). For three-dimensional experiments, coupling the evolution of the two indirect time dimensions corresponds to sampling along vectors emanating from zero time at various angles, referred to as radial sampling. Fourier transformation of a one-dimensional time series corresponding to one of these vectors produces a spectrum that is equivalent to the projection of the two-dimensional spectral cross-section

(plane) onto a vector at the same angle as the time domain vector. Thus one approach to reconstructing the full spectrum is to employ multiple projections corresponding to different angles together with back projection, by analogy with computer-aided tomography. This approach has strong heuristic appeal, and thus provides a useful introduction to non-uniform sampling (NUS). However, this mode of sampling is only a special case of more general NUS and is not necessarily the ideal one (Mobli et al. 2006). Computing spectra for data collected using more general NUS schemes requires different methods that can accommodate data collected at arbitrary times (Barna et al. 1987; Schmieder et al. 1993). Suitable methods include Bayesian, maximum likelihood (Chylla and Markley 1995), maximum entropy (MaxEnt), and multi-way decomposition (MWD) (Orekhov et al. 2001).

Among the methods of spectrum analysis that can process data collected at arbitrary times, MaxEnt has the advantage that the computation cost does not increase with the complexity of the spectrum (Hoch and Stern 1996). It also makes no assumptions regarding the nature of the signals, in contrast to other approaches. It thus yields accurate spectra for non-Lorentzian signals or extremely noisy data. Even though MaxEnt reconstruction of NUS data has been shown to produce superior spectra compared to both linear prediction (LP) extrapolation (Stern et al. 2002) in conventional processing and BPR for processing of radially-sampled data (Mobli et al. 2006), the method has only been utilized by a comparatively small number of laboratories (Sun et al. 2005; Jordan et al. 2006). One of the potential reasons for this is the need to specify parameters controlling the reconstruction that many users find non-intuitive.

Here we describe a method for estimating values of the parameters required for MaxEnt reconstruction, and implementation of an online web-based script generator for automating MaxEnt reconstruction. The method is applicable to data that have been collected using conventional or NUS schemes, and should enable more widespread use of this powerful technique by a broader cross-section of the biomolecular NMR community.

Theory

The theory of MaxEnt reconstruction has been extensively reviewed (Sibisi et al. 1984; Hoch and Stern 1996), and we offer only a brief introduction here. Specific details pertain to the implementation of MaxEnt and automated parameter selection by the Rowland NMR Toolkit (Hoch and Stern 2005). MaxEnt reconstruction computes the spectrum $\mathbf{f}$ that has the highest entropy

$$S(\mathbf{f}) \approx \sum_{n=1}^{N} f_n \log f_n \tag{1}$$

subject to the constraint that the inverse DFT of the spectrum is consistent with the measured FID. Consistency is defined by the inequality

$$C(\mathbf{f}) = \frac{1}{M} \sum_{m=1}^{M} (g_m - d_m)^2 \le C_0 \tag{2}$$

where g_m is an element of the inverse DFT of $\mathbf{f}$, and C_0 is an estimate of the noise level.

The entropy given by Eq. 1, the Shannon entropy, is clearly not applicable to spectra that have both a real and imaginary part or contain negative components. The form we use can be derived either by enforcing continuity of the first derivative or from the quantum-mechanical entropy of an ensemble of spin-1/2 particles (expressed in terms of the spectral components):

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

Here def is a scale factor that in principle is related to the spectrometer sensitivity; in practice we treat it as an adjustable parameter. The expression in Eq. 3 has the property that it is insensitive to the phase of the signal, in contrast to other expressions that have been applied to complex spectra. Note that other software packages (e.g. GIFA Pons et al. 1996, NMRPipe Delaglio et al. 1995) utilize different entropy functionals. We previously showed that these alternatives can produce spectral phase distortions or are not convex, or both (Hoch et al. 1990).

MaxEnt reconstruction amounts to solving the constrained optimization

Maximize $S(\mathbf{f})$ subject to $C(\mathbf{f}) \le C_0$ $\qquad\qquad$ (4)

We convert this constrained optimization into an unconstrained optimization,

Maximize $Q(\mathbf{f}) = S(\mathbf{f}) - \lambda\, C(\mathbf{f})$ $\qquad\qquad$ (5)

where λ is a Lagrange multiplier. The MaxEnt reconstruction corresponds to a critical point of Q, that is a value of $\mathbf{f}$ for which $\bigtriangledown Q$ is zero. Since the entropy and constraint statistic are both convex, there is a unique, global solution to the constrained optimization problem.

Automated parameter determination

From Eqs. 2, 3 and 5 we see that the variables C_0, def and λ are dependent on the underlying data. The C_0 variable (which is related to the *aim* parameter in the Rowland NMR Toolkit) depends on the uncertainty (noise level) in the input spectrum. An estimate of the noise level is therefore of primary importance when reconstructing a spectrum. We consider two ways to automatically obtain reliable estimates of the noise level.

In a non-constant-time multidimensional NMR experiment, in which the evolution times in the indirect dimensions are incremented, the last FID collected will correspond to the longest evolution times and thus provides a signal containing the smallest contribution from the systematic part of the signal (under the assumption that the noise is stationary, and does not vary in magnitude during the course of the experiment). Fourier transformation enables estimation of the noise level in the frequency domain, provided that no apodization is applied, as Parseval's theorem (Hoch and Stern 1996) states that the power in the frequency domain is the same as the power in the time domain. If zero-filling is applied, the noise is scaled by the square root of the ratio of the number of samples to the size of the zero-filled signal,

$$\sigma_t * \sqrt{M} = \sigma_f * \sqrt{N}, \tag{6}$$

where σ_t, σ_f are the standard deviations of noise prior to and following Fourier transformation. [Note Eq. 6 is appropriate for the definition of the DFT employed in the Rowland NMR Toolkit; other software packages may use different normalization conventions in the DFT.] To avoid overestimation due to treating systematic components as noise, a percentage (20% by default) of the largest amplitude values is then rejected together with the central region, which generally contains the water signal. The standard deviation of the remaining data points in the frequency domain is then calculated and used as the estimate of the noise level. Alternatively a signal containing only noise can be collected contemporaneously with the multidimensional data. This is preferably done by collecting an FID with the RF carrier well off-resonance, so that possible contributions such as sample heating and instrument noise, are all present. This approach to automatic estimation of noise level is implemented in the program *noisecalc*, an adjunct to the Rowland NMR Toolkit.

The estimate of the noise level enables determination of the value of C_0. However frequently it is convenient to fix the value of λ, rather than C_0, for example to insure that the nonlinearity of the reconstruction is uniform across rows or planes of a multidimensional spectrum, decrease the intermediate storage required to compute the reconstruction, or

when conducting quantitative comparison of different spectra (Schmieder et al. 1997). The correct value for λ can be determined by choosing representative rows (or planes), and computing the MaxEnt reconstruction with an appropriate value for C_0. The resulting value of λ is then used to compute a row-wise (or plane-wise) reconstruction of the full spectrum; this approach is referred to as the constant-λ algorithm; see Fig. 1 for schematics.

For a two-dimensional spectrum in which MaxEnt reconstruction is performed in the indirect dimension this is done by computing the DFT of the first FID, which is equivalent to the projection of the full spectrum onto the f_2 axis and contains all of the frequency components (in f_2) contained in the two-dimensional spectrum. This one-dimensional spectrum is sorted by amplitude and the 10 largest components (again excluding the central region where the solvent signal is strong) are selected, identifying 10 columns in the indirect dimension likely to contain strong signals. MaxEnt reconstruction is then performed on these 10 columns using values of C_0 calculated from the standard deviation of the noise. The resulting λ values are then averaged to determine the single value of λ used for the full MaxEnt reconstruction. This approach is easily extended to three-dimensional data by using planes instead of columns. Note that this approach is indifferent to the phase of the spectrum, and can be employed with non-uniformly sampled data, provided that N in Eq. 6 is the number of time intervals actually sampled.

The parameter *def* affects the peak intensities in the reconstructed spectrum and to a large extent determines the
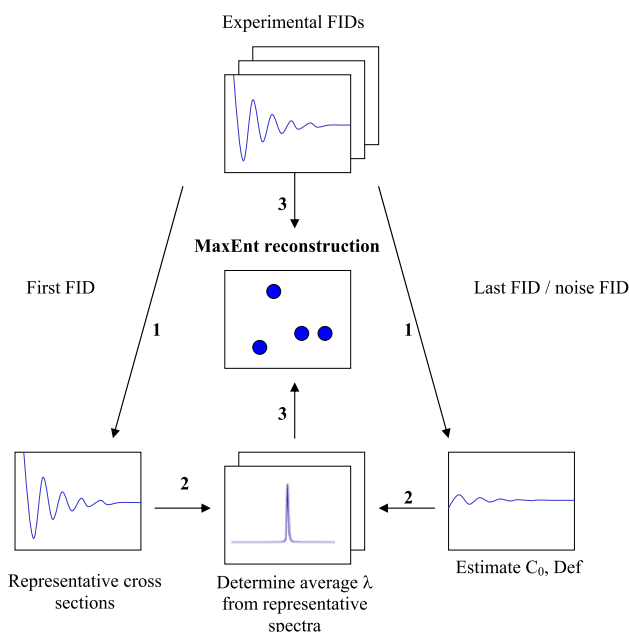

Fig. 1 Flowchart of automated MaxEnt reconstruction. The numbers associated with the arrows indicate the sequential order in which the different tasks are performed

threshold at which the non-linear effects of MaxEnt become significant (Schmieder et al. 1997). It can be formally described as a parameter reflecting the sensitivity of the spectrometer (Daniell and Hore 1989), but we usually treat it as an adjustable parameter. We find that a value close to the standard deviation of the noise gives reasonable results. Increasing the value of *def* gives smoother reconstructions, while values that are too small typically lead to "spiky" noise distributions, and furthermore result in slower convergence. Fortunately the results are not overly sensitive to *def*, and there is considerable latitude in the choice of its value.

A web-based script generator

The RNMRTK Script Generator (version 3.9, http://sbtools.uchc.edu/nmr/) has two main components: a web-accessible front-end and a server-based backend. The front-end interface (for three-dimensional data) is shown in Fig. 2. The front-end is a standard HTML form page, although the actual implementation uses Active Server Pages (Microsoft) to dynamically generate the HTML form page using a combination of HTML and VBScript. Data provided by the user is encoded as one of the many available form objects: Boolean checkboxes, mutually exclusive radio-control buttons, pull-down menus, and text boxes. All user-provided data is gathered into one form and sent to the backend for processing using the HTML post method.

User guidance is provided in several ways. (1) Hyperlinks to extensive help documentation are provided for each user-defined parameter. (2) Options available for each form selection are tailored (and sometimes restricted) to support logical and practical decisions. For instance, the options available for the 'Output File Size', which is the number of points following spectral analysis, are restricted to integers of the form $2n$, where n is an integer. (3) The extensive use of JavaScript to assist the user when the standard setting of one parameter is mathematically linked to the setting of another. For instance, the indirect referencing of heteronuclei employs JavaScript along with published Ξ ratios (Wishart et al. 1995) to calculate the heteronuclear 'Spectrometer Frequency' from the ^{1}H 'Spectrometer Frequency' when a given 'Nucleus' is selected. Similarly, when the number of data points is set for any given dimension, JavaScript is used to automatically adjust the size of the apodization window, the linear prediction parameters (*predict*, *coef*, *points*, and *nextrap*), and the MaxEnt parameter(s) *nuse* to appropriate values. All parameters computed by JavaScript can be overridden by the user. (4) Storage of user-defined parameters in the client browser as a cookie. A cookie is stored whenever a script is generated and can also be stored manually by the

user. Upon returning to the webpage, or by pressing the 'Load Cookie' button, the parameters are reloaded. (5) Extensive validation of the parameters and the logical association between various RNMRTK functions is performed on the server-based backend before the script is generated. Values entered by the user that would cause the RNMRTK program to fail or values that would cause the spectrum to be corrupted are flagged by the script generator and lead to a warning rather than the creation of an invalid script. The warning gives the user information on how to correct the problem. (6) Additional error checking that can only be performed when the script is executed has been built directly into the scripts themselves. For example, when non-uniform sampling is used the maximum time delay values in the sampling schedule file are checked for consistency with the processing parameters. Likewise, the log files from MaxEnt reconstructions are analyzed to check that the calculations completed without errors and that convergence occurred within the specified number of loops. The scripts abort if errors are detected during execution. (7) Formatting of RNMRTK input parameters into the correct format, either text, integers, or floating point values, irrespective of how the value was entered. (8) Lastly a "frequently asked questions" (FAQ) page, highlighting common problems that can arise, is provided.

The server-side backend employs Active Server Pages coded in VBScript. The backend performs two essential roles. The first is the aforementioned role of parameter and logic validation prior to generating the RNMRTK script. The scope of the validation includes verifying that mandatory parameters are provided (such as the input filename), verifying proper data types, verifying mathematical relationships (for instance, the apodization window cannot be larger than the number of complex data points), and rejecting ill-advised settings. The second essential role is the conversion of the user's form selections into an executable RNMRTK shell script. The shell script is a complete processing script including the creation of the shared memory section, format conversion and importing data in RNMRTK, processing along all dimensions, saving data and format conversion supporting several popular NMR data analysis software packages, and the removal of the shared memory section and temporary files.

The RNMRTK script generator supports a wide range of workflows involving 2D/3D data and FT or MaxEnt spectral analysis for each dimension (13 basic workflows, 23 workflows counting different methods for MaxEnt reconstruction (constant aim, constant lambda, and auto), 57 workflows counting J-coupling and line-width deconvolution, and 71 different workflows encompassing both uniform and non-uniform data collection. Optional processing operations such as solvent suppression, zero-filling, apodization, drift correction, linear prediction of initial

Fig. 2 Screen snapshot of the web interface to the script generator

points, phasing, removal of imaginary components, spectrum reversal, and the shrinking of data sets are all available. The script generator applies each of these optional operations at the appropriate place in the script without user intervention. The script generator, together with the *noisecalc* routine, implement completely automated MaxEnt reconstruction.

While computational speeds have risen dramatically, so that processing even large 3D data sets with MaxEnt reconstruction is feasible using personal computers, it is sometimes advantageous to speed up computation times. To this end the RNMRTK script generator supports multiple CPU systems and loosely-coupled clusters. The cluster option calls two additional scripts, available for download, and a text file listing domain names of the computers in the cluster. The scripts perform load balancing and only require that each computer be accessible via ssh and the ability to access a common folder where the data is located.

Results and discussion

Results of the script generator for automatic MaxEnt reconstruction of two-dimensional ^{15}N–^1H HSQC data for

cofilin (Gorbatyuk et al. 2006) are compared with conventional processing in Fig. 3. Panel A shows the results obtained using 128 samples in t_1, extrapolated to 256 points using linear prediction, and sinebell apodization. Panel B & C illustrates the results obtained using MaxEnt reconstruction with the same data, without and with deconvolving a lineshape kernel, respectively. Note that apodization was applied to the acquisition dimension in panel A, but no convolution was applied in panel B, which accounts for the broader lines in the proton dimension in panel B.

One quarter of the original dataset is used for the reconstruction shown in Fig. 4. Panel A illustrates the results obtained using 32 nonuniformly sampled points from the 128-point data set, processed using automatic MaxEnt reconstruction. The subset used in Panel B is the first quarter of the data set (32 t_1 samples) followed by linear prediction, and DFT. The lowest contour levels for all panels were set to ten times the standard deviation of the

noise. While we do not claim that the values of the MaxEnt parameters determined by the script generator are optimal, it is clear from Figs. 3 & 4 that they represent reasonably useful values.

Although MaxEnt reconstruction is more robust than LP extrapolation, and was developed contemporaneously, it has not enjoyed similar widespread adoption. Obstacles such as the computational cost have largely been overcome by developments in computing hardware, but the non-intuitive nature of the adjustable parameters remained an impediment. The development of an automated procedure for estimating the values of the parameters in MaxEnt reconstruction should make the method accessible to a much broader cross section of the biomolecular NMR community, facilitating such applications as nonuniform sampling to reduce data acquisition time for multidimensional NMR experiments (Schmieder et al. 1993, 1994; Mobli et al. 2006) and virtual decoupling via MaxEnt deconvolution (Shimba et al. 2003, 2004; Jordan et al. 2006).
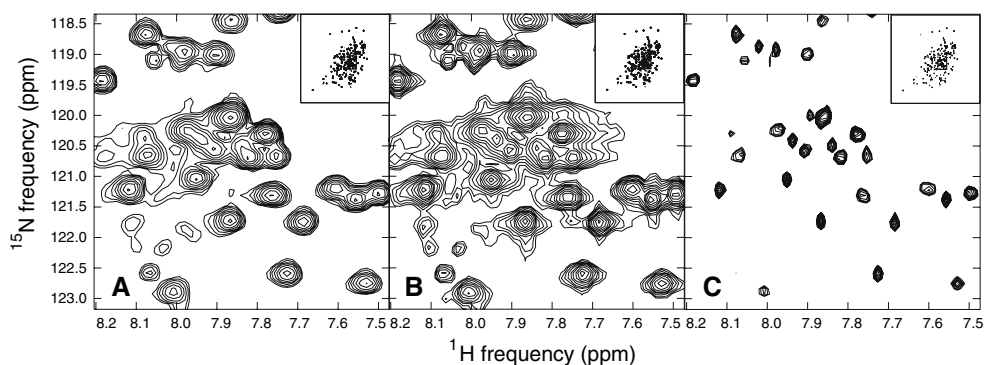

Fig. 3 ^{15}N–^{1}H HSQC spectra for cofilin, processed via conventional and automated MaxEnt reconstruction. (**A**) Conventional LP/DFT processing, using 128 samples in t_1 (**B**) Automated MaxEnt reconstruction, without deconvolution, using 128 samples in t_1 (**C**) Automated MaxEnt reconstruction using 128 samples in t_1, and linewidth deconvolution (10 and 20 Hz in the ^{13}C and ^{15}N dimensions, respectively). The contour levels were chosen at multiples of 1.4 starting with 10 times the estimated rms noise level

Fig. 4 ^{15}N–^{1}H HSQC spectra for cofilin using short data collection. (**A**) Automated MaxEnt reconstruction along f_1, without deconvolution, using 32 samples collected nonuniformly in t_1 (**B**) Conventional LP/DFT processing, using 32 samples in t_1. The contour levels were chosen as in Fig. 3
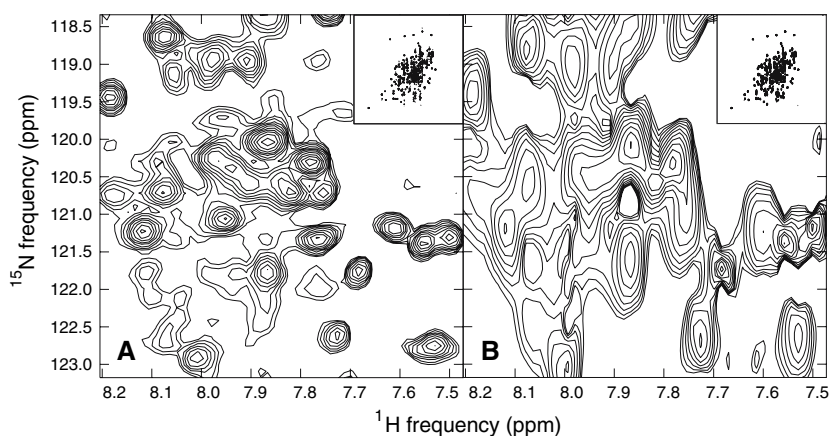
References

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

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

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

Daniell GJ, Hore PJ (1989) Maximum entropy and NMR—A new approach. J Magn Reson 84:515–536

Delaglio F, Grzesiek S et al (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293

Gorbatyuk VY, Nosworthy NJ et al (2006) Mapping the phospho-inositide-binding site on chick cofilin explains how PIP2 regulates the cofilin-actin interaction. Mol Cell 24:511–522

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

Hoch JC, Stern AS (2005) RNMR Toolkit, Version 3

Hoch JC, Stern AS et al (1990) Maximum entropy reconstruction of complex (phase-sensitive) spectra. J Magn Reson 86:236–246

Jaravine V, Ibragimov I et al (2006) Removal of a time barrier for high-resolution multidimensional NMR spectroscopy. Nat Methods 3:605–607

Jordan JB, Kovacs H et al (2006) Three-dimensional 13C-detected CH3-TOCSY using selectively protonated proteins: facile methyl resonance assignment and protein structure determination. J Am Chem Soc 128:9119–9128

Kupče Ē, Freeman R (2003) New methods for fast multidimensional NMR. J Biomol NMR 27:101–113

Malmodin D, Billeter M (2005) High-throughput analysis of protein NMR spectra. Prog Nuc Mag Res Spect 46:109–129

Mobli M, Stern AS et al (2006) Spectral reconstruction methods in fast NMR: reduced dimensionality, random sampling and maximum entropy. J Magn Reson 182:95–105

Orekhov VY, Ibraghimov IV et al (2001) MUNIN: a new approach to multi-dimensional NMR spectra interpretation. J Biomol NMR 20:49–60

Pons JL, Malliavin TE et al (1996) Gifa V4: a complete package for NMR data-set processing. J Biomol NMR 8:445–452

Schmieder P, Stern AS et al (1993) Application of nonlinear sampling schemes to COSY-type spectra. J Biomol NMR 3:569–576

Schmieder P, Stern AS et al (1994) Improved resolution in triple-resonance spectra by nonlinear sampling in the constant-time domain. J Biomol NMR 4:483–490

Schmieder P, Stern AS et al (1997) Quantification of maximum entropy spectrum reconstructions. J Magn Reson 125:332–339

Shimba N, Kovacs H et al (2004) Optimization of 13C direct detection NMR methods. J Biomol NMR 30:175–179

Shimba N, Stern AS et al (2003) Elimination of 13Calpha splitting in protein NMR spectra by deconvolution with maximum entropy reconstruction. J Am Chem Soc 125:2382–2383

Sibisi S, Skilling J et al (1984) Maximum entropy signal processing in practical NMR spectroscopy. Nature 311:446–447

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

Sun ZJ, Hyberts SG et al (2005) High-resolution aliphatic side-chain assignments in 3D HCcoNH experiments with joint H–C evolution and non-uniform sampling. J Biomol NMR 32:55–60

Szyperski T, Wider G et al (1993) Reduced dimensionality in triple resonance NMR experiments. J Am Chem Soc 115:9307–9308

Wishart DS, Bigam CG et al (1995) 1H, 13C and 15N chemical shift referencing in biomolecular NMR. J Biomol NMR 6:135–140