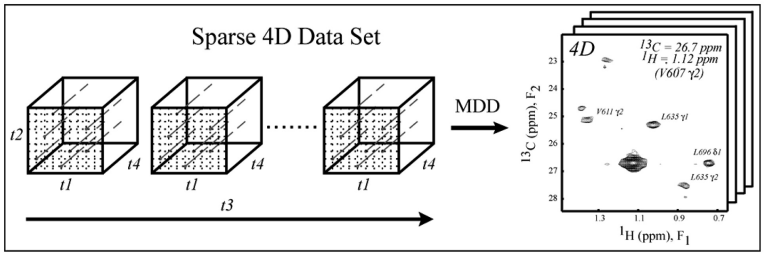
Article

High-Resolution Four-Dimensional H–C NOE Spectroscopy using Methyl-TROSY, Sparse Data Acquisition, and Multidimensional Decomposition

Vitali Tugarinov, Lewis E. Kay, Ilghiz Ibraghimov, and Vladislav Yu. Orekhov

J. Am. Chem. Soc., **2005**, 127 (8), 2767-2775• DOI: 10.1021/ja044032o • Publication Date (Web): 01 February 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 6 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

J|A|C|S
ARTICLES
Published on Web 02/01/2005

High-Resolution Four-Dimensional $^{1}H-^{13}C$ NOE Spectroscopy using Methyl-TROSY, Sparse Data Acquisition, and Multidimensional Decomposition

Vitali Tugarinov,[†] Lewis E. Kay,*[,†] Ilghiz Ibraghimov,[‡] and Vladislav Yu. Orekhov*[,§]

Contribution from the Protein Engineering Network Centers of Excellence and Departments of Medical Genetics, Biochemistry and Chemistry, The University of Toronto, Toronto, Ontario M5S 1A8, Canada, Mathematical Department, Saarbrücken University, 66041 Saarbrücken, Germany, and Swedish NMR Center at Göteborg University, Box 465, 40530 Göteborg, Sweden

Received September 30, 2004; E-mail: kay@pound.med.utoronto.ca; orov@nmr.se

Abstract: An approach for recording four-dimensional (4D) methyl $^{1}H-^{13}C-^{13}C-^{1}H$ NOESY spectra with high resolution and sensitivity is presented and applied to Malate Synthase G (723 residues, 82 kDa). Sensitivity and resolution have been optimized using a highly deuterated, methyl-protonated sample in concert with methyl-TROSY, sparse data sampling in the three indirect dimensions, and 4D spectral reconstruction using multidimensional decomposition (MDD). A sparse data acquisition protocol is introduced that ensures that sufficiently long indirect acquisition times can be employed to exploit the decreased relaxation rates associated with methyl-TROSY, without increasing the duration of the 4D experiment beyond acceptable measurement times. In this manner, only a fraction (~30%) of the experimental data that would normally be needed to achieve a spectrum of high resolution is acquired. The reconstructed 4D spectrum is of similar resolution and sensitivity to three-dimensional (3D) ^{13}C-edited NOE spectra, is straightforward to analyze, and resolves ambiguities that emerge when 3D data sets only are considered.

Introduction

A number of important recent advances have significantly impacted the range of problems that can be addressed by high-resolution multidimensional NMR methods. Included in this group are (i) new isotope labeling schemes for biomolecules,[1] (ii) improved pulse sequence methodologies,[2] (iii) approaches for partial alignment of molecules to introduce anisotropic interactions that would otherwise average to zero,[3,4] and (iv) the development of new hardware. One consequence of these developments is that they have facilitated the study of proteins of increasing size, and a number of systems in the 100 kDa molecular weight range are now investigated.[5,6] These tremendous advances have not yet translated into significant increases in the sizes of molecules for which solution structures have been reported; the majority of structures are still confined to relatively small molecules, with an upper molecular weight bound on the order of 40 kDa.[7−9] An important goal, therefore, is to continue the developments outlined above so that structures of higher molecular weight systems begin to emerge.

A labeling approach that our laboratory has advanced for studies of large proteins is one in which molecules are produced with high levels of deuteration and with protonation confined to selected methyl groups (Ile δ1, Val, and Leu) and to backbone amide positions.[1,10] This level of protonation ensures that spectra recording backbone chemical shifts are of high sensitivity, while retaining a significant number of protons for subsequent studies of structure, through the measurement of amide−amide, amide−methyl, and methyl−methyl NOEs.[11] In this regard, we have found that methyl−methyl NOEs are critical since they connect pairs of helices or helices and strands that are not sufficiently close to be linked via other NOEs, such as those involving pairs of amide protons, for example.[7]

Given the central role that methyl−methyl distances play in establishing global folds of high molecular weight proteins, it is important to optimize both the sensitivity and the resolution of the NOE spectrum from which these contacts are derived. We have recently shown that a methyl-TROSY effect can be realized in HMQC-based experiments, leading to significant improvements in spectral quality,[12,13] and it is straightforward

[†] The University of Toronto.
[‡] Saarbrücken University.
[§] Göteborg University.

(1) Goto, N. K.; Gardner, K. H.; Mueller, G. A.; Willis, R. C.; Kay, L. E. *J. Biomol. NMR* **1999**, *13*, 369−374.
(2) Fernandez, C.; Wider, G. *Curr. Opin. Struct. Biol.* **2003**, *13*, 570−580.
(3) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.
(4) Prestegard, J. H. *Nat. Struct. Biol. NMR* **1998**, *5*, 517−522 (supplement).
(5) Salzmann, M.; Pervushin, K.; Wider, G.; Senn, H.; Wüthrich, K. *J. Am. Chem. Soc.* **2000**, *122*, 7543−7548.
(6) Tugarinov, V.; Muhandiram, R.; Ayed, A.; Kay, L. E. *J. Am. Chem. Soc.* **2002**, *124*, 10025−10035.
(7) Mueller, G. A.; Choy, W. Y.; Yang, D.; Forman-Kay, J. D.; Venters, R. A.; Kay, L. E. *J. Mol. Biol.* **2000**, *300*, 197−212.
(8) Williams, D. C., Jr.; Cai, M.; Clore, G. M. *J. Biol. Chem.* **2004**, *279*, 1449−1457.
(9) Yu, L.; Petros, A. M.; Schnuchel, A.; Zhong, P.; Severin, J. M.; Walter, K.; Holzman, T. F.; Fesik, S. W. *Nat. Struct. Biol.* **1997**, *4*, 483−489.
(10) Gardner, K. H.; Kay, L. E. *J. Am. Chem. Soc.* **1997**, *119*, 7599−7600.
(11) Gardner, K. H.; Rosen, M. K.; Kay, L. E. *Biochemistry* **1997**, *36*, 1389−1401.

to incorporate the HMQC element in 3D and 4D NOE applications. However, the benefits of methyl-TROSY are only fully realized for acquisition times that are sufficiently long to properly exploit the improved relaxation properties of the participating spins. This is particularly a problem with 4D spectroscopy, where acquisition times and subsequently resolution in the indirect dimensions are limited by the time that can be allocated for the experiment. As a consequence, resolution is often so poor in 4D NOE data sets of large proteins that any of the benefits that would normally be associated with the higher dimensionality experiment are very significantly compromised.[14]

An attractive alternative to recording an excessively long 4D spectrum is one where only a relatively small part of the 4D matrix is acquired with the subsequent reconstruction of the complete high-resolution 4D data set. A number of different options are available that can expedite the process of data acquisition in general, including covariance NMR,[15] filter diagonalization,[16] reduced dimensionality (RD) experiments,[17] projection reconstruction (PR) methods,[18] maximum entropy reconstruction,[19] and multidimensional decomposition[20] (MDD). A comparison of the different approaches is beyond the scope of the present work; however, some of the techniques listed above, such as PR and RD, for example, are optimal when sensitivity is not limiting. Unfortunately, the present application is not such a case. Recently, maximum entropy and MDD have been combined with optimal nonuniform sampling[21−23] which can offer significant gains in resolution with little sacrifice in sensitivity. In the work described here, we use MDD[20,24,25] since it has previously been shown that MDD can be used to resolve overlap of resonances in NMR spectra[26] and that MDD combined with nonlinear sampling provides a good alternative for resolution enhancement in NOE spectroscopy.[20,27] In particular, in an application to the 14 kDa protein azurin, Orekhov and co-workers[20] have shown that the original 3D ^{15}N-separated NOESY spectrum can be quantitatively reproduced using only 25% of the FIDs extracted from the complete reference data set.

Here, we describe a practical realization of nonuniform (sparse) 4D data acquisition with subsequent MDD processing, allowing the reconstruction of a high-resolution 4D $F_2-^{13}C$, $F_3-^{13}C$-edited NOESY data set. The spectrum is recorded on a sample of U-[^{15}N,^{2}H], Ile δ1-[$^{13}CH_3$], Leu,Val-[$^{13}CH_3$,$^{12}CD_3$]-labeled Malate Synthase G (MSG; MW 81.2 kDa, correlation time ≈50 ns in D_2O), a single polypeptide enzyme containing 276 Ile δ1, Leu, and Val methyl groups that has been extensively characterized by NMR over the past several years.[6,28,29] For a molecular system with the complexity of MSG, it is not possible to trade sensitivity for resolution (or vice versa) since both are critical and must, therefore, be optimized simultaneously. This is achieved by a combination of isotope labeling,[30] taking advantage of HMQC-type schemes for optimizing resolution and sensitivity,[12] and finally by nonuniform data sampling. The information content of the new 4D NOESY experiment, acquired in 6.5 days using sparse sampling in the three indirect dimensions and processed using MDD, is compared with conventional ^{13}C-separated 3D HMQC−NOESY[31] and 3D $F_1-^{13}C$, $F_2-^{13}C$ HMQC−NOESY−HMQC[32] data sets. A number of practical aspects involving the application of MDD to 4D methyl-NOE spectroscopy are addressed.

Materials and Methods

NMR Sample. A U-[^{15}N,^{2}H], Ile δ1-[$^{13}CH_3$], Leu,Val-[$^{13}CH_3$,$^{12}CD_3$]-labeled sample of MSG[33] was prepared as described previously[6,30] using U-[^{2}H]-glucose and $^{15}NH_4Cl$ (CIL, Andover, MA) as the main carbon and nitrogen sources. Selective protonation of Ile δ1 methyls and of one of the methyl groups of Leu and Val (Leu,Val-[$^{13}CH_3$,$^{12}CD_3$]) was achieved by addition of 80 mg of 2-keto-3,3-d_2-4-^{13}C-butyrate and 120 mg of 2-keto-3-methyl-d_3-3-d_1-4-^{13}C-butyrate to 1 L of D_2O-based minimal medium 1 h prior to induction.[1,30] Sodium salts of 2-keto-4-^{13}C-butyric and 2-keto-3-methyl-d_3-4-^{13}C-butyric acids were obtained from Isotec (Miamisburg, OH). The sample of MSG was 0.9 mM in protein, 99.9% D_2O, 25 mM sodium phosphate buffer (pH 7.1, uncorrected), 20 mM $MgCl_2$, and 0.05% NaN_3.

Sparse 4D HMQC−NOESY−HMQC Data Acquisition and MDD Data Processing. All NMR experiments were performed on a Varian Inova 800 MHz spectrometer, 37 °C, equipped with a room-temperature pulsed-field gradient triple resonance probe. Sparse acquisition in the three indirect dimensions was implemented in the pulse program (scheme of Figure 1; the pulse-sequence code is available immediately from L.E.K. upon request, and the processing code will be released by summer 2005 or earlier upon contacting V.O.). The values of the parameters controlling the hyper-complex signal acquisition and the incremented delays (t_1, t_2, t_3) spanning acquisition times (t_{1max}, t_{2max}, t_{3max} = 27.0, 21.7, 16.3 ms; spectral widths of F_1(^{1}H), F_2(^{13}C), F_3(^{13}C) = 1.2, 11.0, 11.0 ppm in the indirect dimensions) were taken from a precalculated table that is automatically generated at the start of the experiment by code that is integrated into the pulse scheme (i.e., user intervention is not required beyond indicating the degree of sparsing required). Out of 359 424 possible FIDs corresponding to the complete 4D spectrum, 110 592 (30.8%) were randomly selected for detection, so that "on average", the number of t_i increments for given values of (t_j, t_k), (i, j, k) ∈ (1,2,3) with $i \neq j \neq k$ is reduced to 30% of the number acquired in a conventional data set. The random distribution used for the selection was exponentially biased to match the transverse relaxation times, approximately 40 ms,[30] in the three indirect dimensions, as described briefly in the Appendix. At least two data points were recorded for each column of the three-dimensional (t_1, t_2, t_3) grid. Note that this procedure implies the same average fraction of points

(12) Tugarinov, V.; Hwang, P. M.; Ollerenshaw, J. E.; Kay, L. E. *J. Am. Chem. Soc.* **2003**, *125*, 10420−10428.
(13) Ollerenshaw, J. E.; Tugarinov, V.; Kay, L. E. *Magn. Reson. Chem.* **2003**, *41*, 843−852.
(14) Zwahlen, C.; Gardner, K. H.; Sarma, S. P.; Horita, D. A.; Byrd, R. A.; Kay, L. E. *J. Am. Chem. Soc.* **1998**, *120*, 7617−7625.
(15) Bruschweiler, R.; Zhang, F. *J. Chem. Phys.* **2004**, *120*, 5253−5260.
(16) Chen, J.; De Angelis, A. A.; Mandelshtam, V. A.; Shaka, A. J. *J. Magn. Reson.* **2003**, *162*, 74−89.
(17) Kim, S.; Szyperski, T. *J. Am. Chem. Soc.* **2003**, *125*, 1385−1393.
(18) Kupce, E.; Freeman, R. *J. Am. Chem. Soc.* **2004**, *126*, 6429−6440.
(19) Hoch, J. C.; Stern, A. S. *NMR Data Processing*; Wiley & Sons: New York, 1996.
(20) Orekhov, V. Y.; Ibraghimov, I. V.; Billeter, M. *J. Biomol. NMR* **2001**, *20*, 49−60.
(21) Barna, J. C. J.; Laue, E. D.; Mayger, M. R.; Skilling, J.; Worrall, S. J. P. *J. Magn. Reson.* **1987**, *73*, 69−77.
(22) Schmieder, P.; Stern, A. S.; Wagner, G.; Hoch, J. C. *J. Biomol. NMR* **1993**, *3*, 569−576.
(23) Rovnyak, D.; Frueh, D. P.; Sastry, M.; Sun, Z. Y.; Stern, A. S.; Hoch, J. C.; Wagner, G. *J. Magn. Reson.* **2004**, *170*, 15−21.
(24) Bro, R. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149−171.
(25) Ibraghimov, I. *Numer. Linear Algebra Appl.* **2002**, *9*, 551−565.
(26) Gutmanas, A.; Tu, L.; Orekhov, V. Y.; Billeter, M. *J. Magn. Reson.* **2004**, *167*, 107−113.
(27) Gutmanas, A.; Jarvoll, P.; Orekhov, V. Y.; Billeter, M. *J. Biomol. NMR* **2002**, *24*, 191−201.
(28) Tugarinov, V.; Kay, L. E. *J. Mol. Biol.* **2003**, *327*, 1121−1133.
(29) Tugarinov, V.; Hwang, P. M.; Kay, L. E. *Annu. Rev. Biochem.* **2004**, *73*, 107−146.
(30) Tugarinov, V.; Kay, L. E. *J. Biomol. NMR* **2004**, *28*, 165−172.
(31) Ikura, M.; Kay, L. E.; Tschudin, R.; Bax, A. *J. Magn. Reson.* **1990**, *86*, 204−209.
(32) Ikura, M.; Bax, A.; Clore, G. M.; Gronenborn, A. M. *J. Am. Chem. Soc.* **1990**, *112*, 9020−2022.
(33) Howard, B. R.; Endrizzi, J. A.; Remington, S. J. *Biochemistry* **2000**, *39*, 3156−3168.

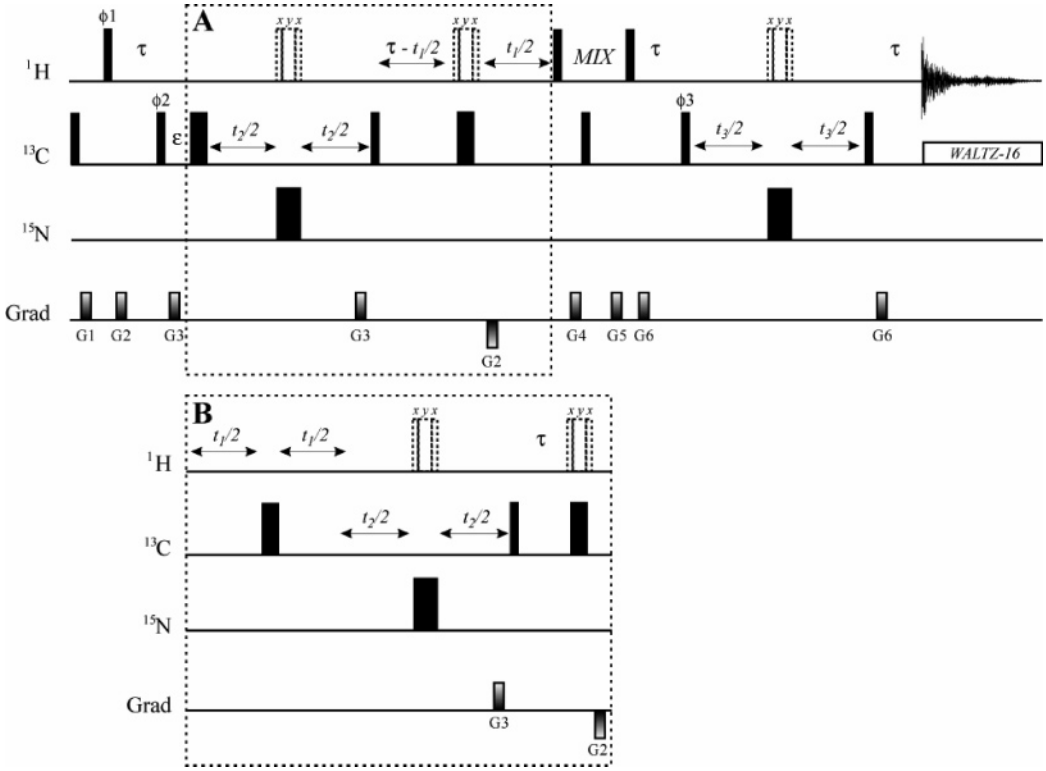


Figure 1. The 4D $^1H-^{13}C-^{13}C-^1H$ HMQC−NOESY−HMQC pulse scheme. A sparse data acquisition protocol and subsequent MDD reconstruction of the 4D data set has been employed, as described in the text. All narrow (wide) rectangular pulses are applied with flip angles of 90° (180°) along the *x*-axis unless indicated otherwise. The ^1H, ^{15}N, and ^{13}C carriers are positioned in the center of the methyl region (1.0 ppm), at 119 ppm, and in the center of the Val, Leu methyl region (23 ppm), respectively. (In the present application, all Ile δ1 methyl resonances are aliased.) All ^1H, ^{13}C, and ^{15}N pulses are applied with the highest available power with ^{13}C WALTZ-16 decoupling[50] achieved using a 2 kHz field. The ^1H pulses shown with dashed lines are of the composite variety,[51] $90°x-180°y-90°x$. The τ delays are set to 3.6 ms, while delay ϵ is set to G3+pwN180, where pwN180 is the duration of the ^{15}N 180° pulse. For $t_1 > 2\tau$, scheme B is substituted for A (see text). The durations and strengths of the pulsed field gradients applied along the *z*-axis are as follows: G1 = 1.0 ms, 7.5 G/cm; G2 = 50 μs, 10.0 G/cm; G3 = 50 μs, 15 G/cm; G4 = 0.5 ms, 12 G/cm; G5 = 1 ms, 12 G/cm; G6 = 0.4 ms, −8 G/cm. The phase cycle is as follows: $\phi1 = x$; $\phi2 = x, -x$; $\phi3 = 2(x), 2(-x)$; rec. $= x, -x, -x, x$. Quadrature detection in t_1, t_2, t_3 is achieved with States-TPPI incrementation of $\phi1$, $\phi2$, and $\phi3$, respectively.[52]

sampled for all indirect dimensions and is more restrictive than required by theory, as described below. A 64 ms acquisition time was used in the direct ^1H dimension, along with 4 transients/FID, a 175 ms NOE mixing time, and a relaxation delay of 1 s for a total acquisition time of 6.5 days.

The sparse 4D data matrix was converted to NMRPipe format[34] and Fourier transformed along the directly detected (methyl ^1H) dimension. Both real and imaginary parts of the signal in the range from −0.5 to 1.5 ppm were extracted and used for subsequent MDD processing. The sparsed data set, S' (see below), was used as input to solve for 350 components in a single MDD calculation (eq 5.1), using a nonlinear least-squares minimization routine that has been described previously.[20,35] The number of components used exceeds somewhat the total number of cross-peaks (272) found in the methyl region of the high-resolution $^{13}C-^1H$ 2D HMQC spectrum of Ile δ1-[$^{13}CH_3$], Leu,Val-[$^{13}CH_3$,$^{12}CD_3$]-labeled MSG.[30] The results of MDD calculations were found to be insensitive to the setting of λ in eq 5.1, in agreement with previous studies.[35] In the following, results are presented for $\lambda = 0.001$. Processing was performed with home-written software, mddNMR, using a parallel Linux cluster of 8 Intel CPUs (2.4 GHz, overall duration of approximately 2 days). Convergence of the minimization routine was established by extending the computation for an additional period with no further decrease in the residual (eq 5.1). The output of MDD calculations produces time domain profiles (denoted as "shapes" in what follows), along with the amplitudes for all components (see below). The shapes associated with individual components were converted to the NMRPipe format,[34] extended 2-fold by forward−backward linear prediction,[36] zero-filled, and Fourier transformed using standard routines in NMRPipe.[34] Finally, the 4D spectrum was reconstructed from the frequency domain shapes and amplitudes obtained for all components using eq 4.

3D NOE Spectroscopy. Nonsparsed ^{13}C-separated 3D HMQC−NOESY ($^1H[F_1]$, $^{13}C[F_2]$, $^1H[F_3]$) and 3D $F_1-^{13}C$, $F_2-^{13}C$ HMQC−NOESY−HMQC ($^{13}C[F_1]$, $^{13}C[F_2]$, $^1H[F_3]$) data sets were collected using acquisition times of 41, 36, and 64 ms and 36, 36, and 64 ms, respectively, in the (t_1, t_2, t_3) dimensions and spectral widths of 11 ppm (1.2 ppm) for the indirect $^{13}C(^1H)$ dimensions. The NOE mixing delay immediately precedes the acquisition period in the case of the ^{13}C-separated 3D HMQC−NOESY. The total experimental time for each 3D data set was 4.5 days (relaxation delay 1 s, NOE mixing period 175 ms). The optimal NOE mixing time was determined from the build-up of cross-peak intensities in a series of short 3D spectra with mixing times ranging from 40 to 250 ms. Three-dimensional NOESY spectra were processed using conventional processing schemes with NMRPipe/NMRDraw software.[37] The 3D and the reconstructed 4D NOESY data sets were analyzed simultaneously using the NMRView[38] program along with tcl/tk scripts for spectral visualization written in house.

(34) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.
(35) Orekhov, V. Y.; Ibraghimov, I.; Billeter, M. *J. Biomol. NMR* **2003**, *27*, 165−173.
(36) Zhu, G.; Bax, A. *J. Magn. Reson.* **1992**, *98*, 192−199.
(37) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.
(38) Johnson, B. A.; Blevins, R. A. *J. Biomol. NMR* **1994**, *4*, 603−614.

Results and Discussion

Methyl−Methyl 4D HMQC−NOESY−HMQC Pulse Scheme. The advantages of HMQC relative to HSQC spectroscopy of methyl groups attached to high molecular weight proteins have been described in some detail in the literature.[12] The benefits derive from a TROSY effect which depends on the cancellation of relaxation interactions due to the interference between intra-methyl dipolar fields. Like other types of TROSY enhancements,[39] protons external to the spin system in question (in this case a methyl group) lead to a decreased TROSY enhancement. It is, therefore, important to employ a labeling scheme which minimizes contributions from external spins, and in the present study, MSG has been prepared with protonation confined to only one of the two methyl groups of Leu and Val and to the δ-methyl position of Ile.[30]

Figure 1 shows the pulse sequence that has been designed to take advantage of both the methyl-TROSY effect and the labeling scheme that has been employed. The sequence is similar to that used to record other $F_2-^{13}C$, $F_3-^{13}C$-edited 4D NOE data sets,[40,41] with the exception that HMQC elements are employed. Notably, ^1H evolution proceeds initially (for $t_1 \leq 2\tau \sim 7$ ms) during an interval in which ^1H magnetization (*single* quantum) is refocused with respect to the one-bond $^1H-^{13}C$ coupling (Figure 1A). In this manner, the length of the pulse sequence does not increase (due to t_1 evolution) until $t_1 > 2\tau$. Subsequently, the ^1H chemical shift is recorded when the operative coherence is of the multiple-quantum variety, with refocusing of methyl ^{13}C chemical shifts, as shown in Figure 1B. This approach exploits the slower relaxation times of methyl multiple-quantum coherences, on average by 15%, relative to the decay of ^1H single-quantum magnetization for the Ile δ1-[$^{13}CH_3$], Leu,Val-[$^{13}CH_3$,$^{12}CD_3$]-labeled preparation of MSG in D_2O used in this work.[30] Indeed, sensitivity gains on the order of 15−20% were realized in a comparison of test-3D data sets where ^1H chemical shifts were recorded using the scheme of Figure 1 relative to an approach in which single-quantum coherences exclusively evolve during t_1.

Multidimensional Decomposition (MDD). In what follows, a brief discussion of MDD as applied to 4D NOE spectroscopy is presented. A description of the minimization procedure that is employed to extract the "shapes" that are subsequently used to reconstruct the spectrum (see below) is given elsewhere.[24,25]

Consider a time domain signal, S, that spans four evolution times (t_1, t_2, t_3, and t_4) in a 4D NOE time domain data matrix

$$S(t_1, t_2, t_3, t_4) = \sum_{i,j} a_{i,j} \cos(\omega_i^H t_1) \cos(\omega_i^C t_2) \cos(\omega_j^C t_3) \cos(\omega_j^H t_4) \quad (1)$$

where the proton and carbon frequencies are denoted by ω^H and ω^C, respectively; the effects of relaxation during the mixing time and magnetization transfer efficiencies are subsumed in the coefficients $a_{i,j}$, and the indices, i and j, distinguish the sites of origination (methyl i) and destination (methyl j) of magnetization. For simplicity, relaxation during each of t_1-t_4 has been neglected, but it is implicitly taken into account by the MDD procedure (eqs 4 and 5 below). In practice, the NMR time domain signal is comprised of both cosine- and sine-modulated components, facilitating quadrature detection. The essential features of the MDD approach can be illustrated, however, by focusing on the cosine signal above (eq 1). As usual, the evolution times are defined on a grid with regular intervals, Δt

$$t_1 = (p-1)\Delta t_1,\ t_2 = (q-1)\Delta t_2,\ t_3 = (m-1)\Delta t_3,\ t_4 = (n-1)\Delta t_4 \quad (2)$$

where indexes p, q, m, and n run from 1 to the maximal values P, Q, M, and N defined by the maximum acquisition times in dimensions 1−4, respectively.

Substituting, $F1^j(t_1,t_2) = \Sigma_i a_{i,j} \cos(\omega_i^H t_1) \cos(\omega_i^C t_2)$, $F2^j(t_3) = \cos(\omega_j^C t_3)$, $F3^j(t_4) = \cos(\omega_j^H t_4)$, eq 1 takes the form

$$S(t_1, t_2, t_3, t_4) = \sum_j F1^j(t_1,t_2) F2^j(t_3) F3^j(t_4) \quad (3)$$

In eq 3, the factors $F2^j(t_3)$ and $F3^j(t_4)$ correspond to time domain ^{13}C ($F2$) and ^1H ($F3$) profiles, or shapes, associated with methyl group j. The shape $F1^j(t_1,t_2)$ can be thought of as a two-dimensional $^1H-^{13}C$ time domain correlation spectrum, where signal intensities depend on relaxation during the mixing time of all methyl protons i from which magnetization originates and of all protons of the destination group j. The Fourier transform of this shape would contain a diagonal peak and cross-peaks from all methyl protons i that are proximal to methyl j.

To provide a clearer picture of the relation between the "model" spectrum, S_M, and the shapes that define it, it is convenient to write eq 3 as

$$S_M = \sum_j^R \alpha^j F1^j \otimes F2^j \otimes F3^j \quad (4)$$

Here, the modeled signal, S_M, is "decomposed" into R components that for the particular case of the 4D methyl NOESY spectrum correspond to R observable methyl groups in the protein and $\otimes$ denotes tensor product. For each component, there are three one-dimensional (vector) normalized shapes, $F1$, $F2$, and $F3$, and a coefficient, α^j. Equation 4 holds for data in the time domain (as has been discussed above), in the frequency domain, or for a mixture of the two (time domain in the indirect dimensions and frequency data in the direct detect dimension, for example). The only assumption made by the model is that the signal can be represented by a tensor product of three one-dimensional shapes. No specific assumption about the shapes is needed to ensure the uniqueness of the solution in most practical situations.[20,42] Indeed, for the 4D (time domain) spectrum considered here, the 1D shape, $F1^j$, describes a complex two-dimensional time domain object rather than an analytical *cosine* function, for example, while for a 3D (time domain) data set, $F1^j$ would denote a 1D time domain response.

The parameters of the model can be defined by a least-squares minimization procedure implemented in MDD that is equivalent for both 3D and 4D data sets

$$\min \sum_{k,m,n} G_{k,m,n} |S'_{k,m,n} - \sum_{j=1}^R \alpha^j \times F1_k^j \times F2_m^j \times F3_n^j|^2 + \lambda \sum_{j=1}^R (\alpha^j)^2 \quad (5.1)$$

(39) Pervushin, K.; Riek, R.; Wider, G.; Wüthrich, K. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 12366−12371.
(40) Clore, G. M.; Kay, L. E.; Bax, A.; Gronenborn, A. M. *Biochemistry* **1991**, *30*, 12−18.
(41) Zuiderweg, E. R. P.; Petros, A. M.; Fesik, S. W.; Olejniczak, E. T. *J. Am. Chem. Soc.* **1991**, *113*, 370−371.

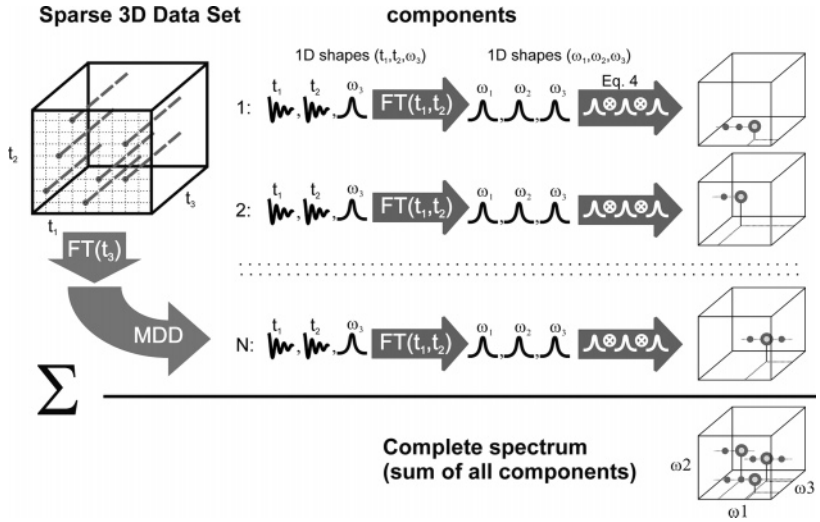


Figure 2. Schematic of the procedure by which the frequency domain spectrum of a sparsely sampled multidimensional data set is reconstructed. See text for details.

where $S'_{k,m,n}$ is the (k,m,n) experimental data point. An index, k, has been introduced that runs over $K = P \times Q$ values, with k an element of (p,q) (see eq 2). It is noteworthy that the only difference between the implementation for 3D and 4D data sets is the size of the index k, with k extending from 1 to P ($P \times Q$) for 3D (4D) data. Thus, shape $F1^j$ is given by the vector $\vec{v}_{3D} = [F1^j_1,F1^j_2,...,F1^j_P]$ or $\vec{v}_{4D} = [F1^j_1,F1^j_2,...,F1^j_{P*Q}]$ for 3D and 4D data, respectively, where $\vec{v}_{4D}$ can be recast as

$$[F^j_{1,1},...,F^j_{1,P},F^j_{2,1},...,F^j_{2,P},...,F^j_{Q,1},...,F^j_{Q,P}] \leftrightarrow \begin{pmatrix} F^j_{1,1} & . & . & F^j_{1,P} \\ . & & . & . \\ . & & . & . \\ F^j_{Q,1} & . & . & F^j_{Q,P} \end{pmatrix} \quad (5.2)$$

so that the relation between $F1^j$ and a two-dimensional matrix is clear. The algorithm that is used to derive the individual shapes is, however, identical for both 3D and 4D data sets, and no changes in the code were required for the present work. The advantage of this "pseudo" 4D MDD procedure over one which fits the data to four separate shapes will be summarized below. In eq 5.1, a total of $R(K + M + N - 2)$ parameters are adjusted in the fit, that is, normalized shapes, $F1$, $F2$, $F3$, and amplitudes, α, for all R components; in the present example, roughly $KMN/3$ experimental data points are acquired by the sparsing procedure (see Materials and Methods) so that the system is very much over-determined. To account for the fact that the matrix S' has "holes" (i.e., points for which there is no data), $G_{k,m,n}$ is set to 0 for any point (k,m,n) where data has not been acquired and to 1, otherwise. The value λ is a Tikhonov regularization factor[43] that improves the convergence of the MDD algorithm.[25] As has been discussed previously,[35] the number of components, R, should be somewhat larger than the total number of diagonal signals in the spectrum to ensure that the most important features of the spectrum can be described by the model (eq 4). In essence, the minimization protocol above (eq 5.1) involves a 3D deconvolution or, in the case of 4D data, a pseudo-4D deconvolution in which one of the shapes ($F1$) is a two-dimensional object. With the development of efficient MDD algorithms for solving the minimization problem of eq 5.1 with large fractions of missing data,[25] it has become possible to use the predictive power of the model defined by eq 4. Such a model describes a complete multidimensional spectrum in the time domain. Consequently, when a large fraction of the points in the experimental data set is omitted, a complete time domain spectrum can be reconstructed using eq 4. The only (non-restraining) requirement is that in the multidimensional experimental data matrix, there be no rows or columns completely lacking acquired data. Although operationally similar to maximum entropy reconstruction methods,[19,23,44,45] the MDD model is especially well suited for resolution enhancement of (crowded) NOE data sets,[27,35] where cross-peaks are associated with strong diagonal signals.

It is worth noting that we have chosen to model the data as the sum over tensor products of three shapes where $F1$ is a composite two-dimensional object. In this way, the number of components is roughly equal to the number of cross-peaks in a 2D $^{13}C-^{1}H$ correlation map, as described above. Since the $F1$ shape contains diagonal and cross-peaks, *only a single pair of shapes (F2,F3) is fit for each set of NOE correlations linking residues i and j*. In contrast, the signal could, of course, also be written as the sum of products of four shapes, each of which would correspond to a one-dimensional time domain response. In this case, however, the number of components would approximate the number of all peaks in the 4D spectrum (that is, cross-peaks and diagonal-peaks). Moreover, an ($F2$,$F3$) pair would be separately fit for each peak, which is much more demanding of the algorithm. In our experience, many cross-peaks are missing in the reconstructed spectrum when a 4D deconvolution procedure (4 shapes) is employed.

The basic features of MDD can be illustrated schematically (Figure 2). Here, a 3D data set is used as an example since the steps involved in processing are easier to "visualize" than those

(42) Kruskal, J. B. In *Multiway Data Analysis*; Coppi, R., Bolasco, S., Eds.; North-Holland Elsevier Science Pub.: Amsterdam, New York, 1989.
(43) Tikhonov, A. N.; Samarskij, A. A. In *Equations of Mathematical Physics*; Dover: New York, 1990.
(44) Hoch, J. C.; Stern, A. S. In *Encyclopedia of Nuclear Magnetic Resonance*; Grant, D. M., Harris, R. K., Eds.; John Wiley: London, 1996; pp 2980−2988.
(45) Hoch, J. C.; Stern, A. S. In *Nuclear Magnetic Resonance of Biological Macromolecules, Part A*; Academic Press Inc.: San Diego, CA, 2001; Vol. 338, pp 159−178.

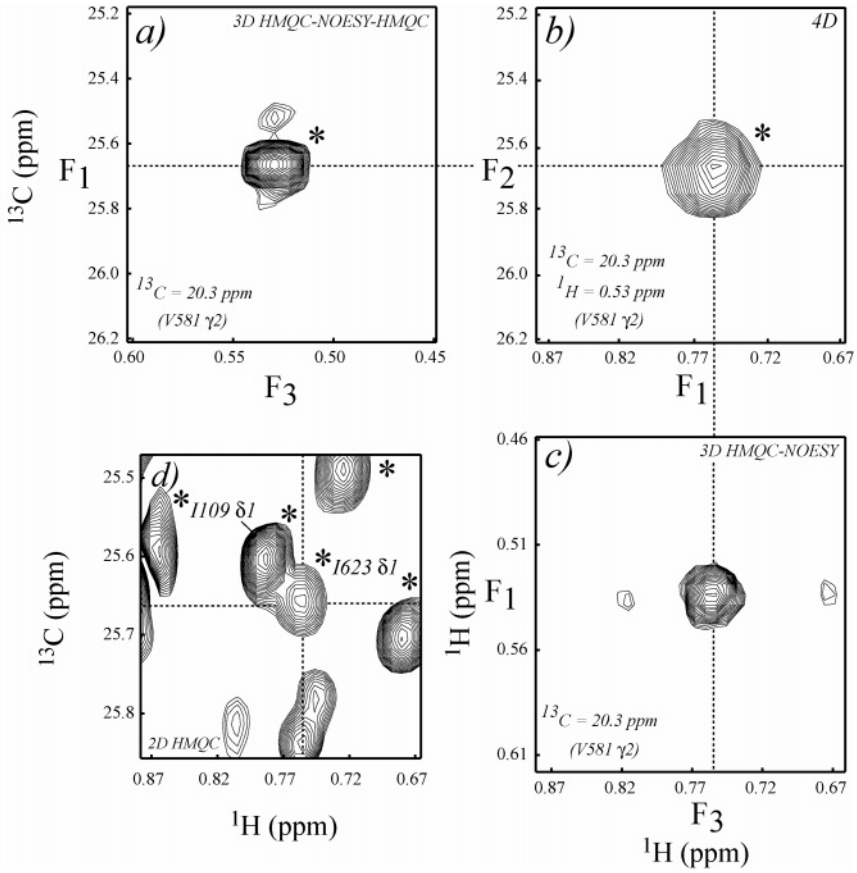


Figure 3. Illustration of the resolution in the 4D reconstructed NOE data set. Cross-peaks from (a) a portion of an $F_1(^{13}C)-F_3(^{1}H)$ plane of the 3D HMQC−NOESY−HMQC, at the $F_2(^{13}C)$ frequency of V581 γ2 methyl (20.3 ppm); (b) a region from the F_1-F_2 plane of the reconstructed 4D spectrum, $F_3(^{13}C)$ = 20.3 ppm, $F_4(^{1}H)$ = 0.53 ppm; (c) a section from an $F_1(^{1}H)-F_3(^{1}H)$ plane of the 3D HMQC−NOESY ($F_2(^{13}C)$ frequency of 20.3 ppm); and (d) a region of the high-resolution $^{1}H-^{13}C$ HMQC correlation map of MSG. The resolution in the 4D data set is sufficient to establish an NOE between V581 γ2 and I623 δ1 and not V581 γ2−I109 δ1. Data sets were recorded at 800 MHz. The correlations in (a) and (b) are aliased in F_1 and F_2, respectively, while those that are indicated by $*$ are aliased in F_1 in (d).

in the case of a 4D matrix, although the approach used for either types of data is identical. Consider, for the sake of discussion, a sparsed 3D ^{13}C-edited NOESY−HMQC data matrix, S', where for simplicity, each of the cross-peaks is resolved in the 2D $^{13}C-^{1}H$ correlation map (although this is not a requirement of the approach). The data set is first Fourier transformed in the acquisition dimension and subsequently subjected to MDD (eq 5.1). For each component, the minimization produces a set of three shapes, $F1$, $F2$, $F3$, where $F1$ and $F2$ ($F3$) are time (frequency) domain objects. Shapes $F1$ and $F2$ can be manipulated as any linear array of time domain data (i.e., the data can be linear-predicted, zero-filled, windowed, and Fourier transformed) to produce frequency domain shapes. For each component, the tensor products of (frequency domain) shapes 2 and 3 define a 2D $^{13}C-^{1}H$ correlation map comprised of a single peak. The frequency domain profile, $F1(\omega)$, in turn, corresponds to a 1D spectrum containing the diagonal peak and all associated NOE correlations; the tensor product of the three frequency domain shapes produces, therefore, a 3D NOE data set for each residue (component). Clearly, the sum of all the data sets so produced reproduces the 3D spectrum.

The distribution of noise in the constructed data set can be quite different than that in the case of a data set that is acquired and processed in the "normal" way. The minimization process of eq 5.1, above, implies that a significant fraction of the noise in the spectrum, which is not taken into account by the model, ends up in the residuals. This has an apparent "denoising" effect on the spectrum. Some noise, however, is present in the shapes, and because the spectrum is reconstructed as a tensor product of shapes, the noise will not be uniformly distributed but rather will be significantly higher in planes where there are cross-peaks. The distribution of noise also depends on the amount of noise in each shape. Consider the example of Figure 2. Here, there is a 1:1 correspondence between time domains t_1, t_2, and t_3 during which data are acquired in the NMR experiment and the time domain shapes $F1(t_1)$, $F2(t_2)$, and $F3(t_3)$, respectively. How well a given shape is defined, say $F1$, depends on the number of points sampled in t_2 and t_3; in general, the more points sampled in t_2, t_3 the lower the noise floor associated with $F1$. If the number of points sampled in t_1 greatly exceeds that in t_2, for example, the noise in the frequency shape $F1(\omega_1)$ will be correspondingly higher than that in $F2(\omega_2)$. This creates a noise-ridge parallel to the F_1 dimension of the reconstructed 3D data set at F_2,F_3 frequencies corresponding to the position of each diagonal peak.

In the case of the 4D data set, the processing scheme used effectively combines the t_1, t_2 time domains to produce a single 2D composite shape, $F1$ (see eq 4 above). Because the combined number of data points sampled in t_1, t_2 exceeds that in t_3, in the frequency domain the noise floor in $F2(\omega_3)$ is significantly lower than that in $F1(\omega_1,\omega_2)$. The effect is that noise ridges run parallel to F_1 and F_2, with the noise level varying significantly within

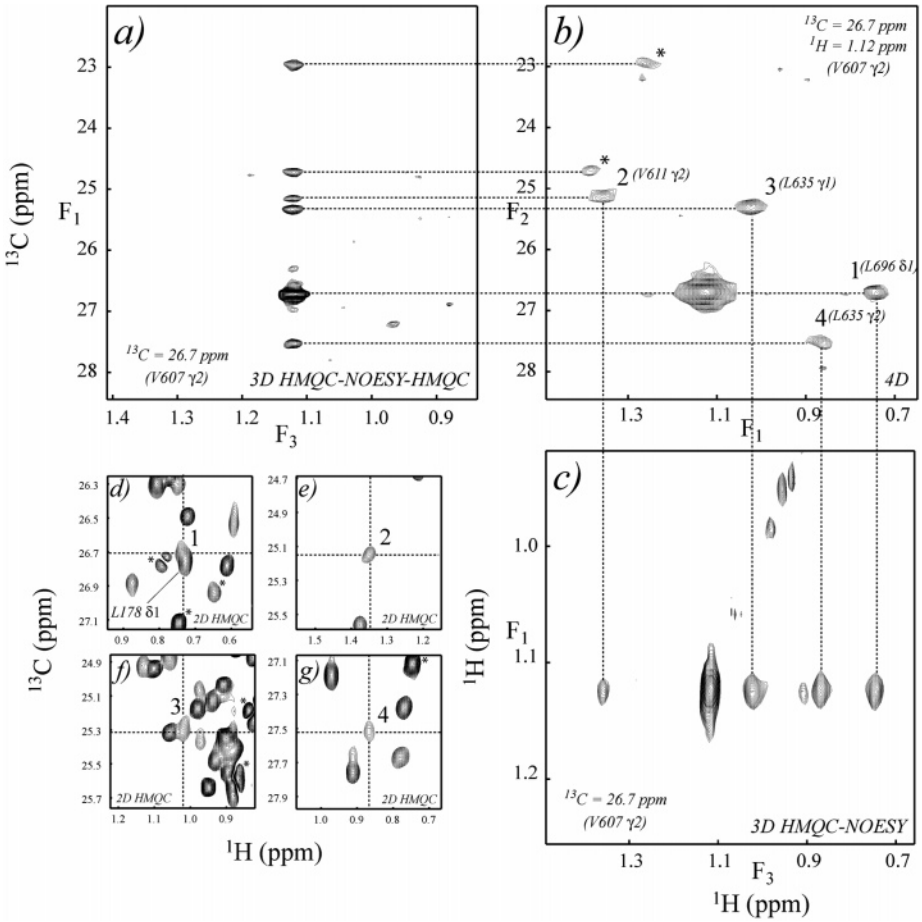


Figure 4. Comparison of selected slices from 3D and 4D NOE data sets recorded on MSG, 37 °C, 800 MHz. (a) $F_1(^{13}C)-F_3(^1H)$ plane of the 3D HMQC−NOESY−HMQC, at an $F_2(^{13}C)$ frequency of the V607 γ2 methyl (26.7 ppm); (b) $F_2(^{13}C)-F_1(^1H)$ plane from the reconstructed 4D spectrum drawn at $F_3(^{13}C) = 26.7$ ppm, $F_4(^1H) = 1.12$ ppm (peaks marked with asterisks are aliased in the $F_1(^1H)$ dimension); (c) $F_1(^1H)-F_3(^1H)$ plane of the 3D HMQC−NOESY, $F_2(^{13}C)$ frequency of 26.7 ppm. Panels (d−g) show correlations from the 2D $^1H-^{13}C$ HMQC, with the cursors indicating the positions of methyls that are the origination sites of magnetization transfer in the 4D NOE slice shown in panel (b). Aliased cross-peaks in F_1 are indicated by ∗.

each F_3-F_4 plane. In contrast, within each F_1-F_2 plane, the distribution of noise is relatively uniform, although the noise level increases significantly in planes with cross-peaks. It is convenient, therefore, to examine the spectrum as a set of F_1-F_2 planes, where the noise distribution poses no problem for analysis.

Some Aspects of Resolution and Sensitivity of the Reconstructed 4D NOESY. The resolution obtained in the reconstructed 4D NOESY spectrum of MSG is illustrated in Figure 3. Here, sections from (a) an F_1-F_3 plane of the 3D HMQC−NOESY−HMQC and from (c) the corresponding F_1-F_3 plane of the 3D HMQC−NOESY, both at the $F_2(^{13}C)$ frequency of the V581γ2 methyl (20.3 ppm), are compared with a region from the $F_1(^1H)-F_2(^{13}C)$ plane of the reconstructed 4D spectrum, $F_3(^{13}C) = 20.3$ ppm and $F_4(^1H) = 0.53$ ppm (b). While the F_2 resolution in the 4D data set (b) is not as good as that in the corresponding dimension of the 3D matrix (a), it is nonetheless sufficient to establish that I623 δ1 and not I109 δ1 is proximal to V581γ2, despite the fact that the two Ile methyls are partially overlapped even in the high-resolution 2D HMQC map (d). This distinction would not be possible using conventional 4D spectroscopy with short acquisition times in the indirectly detected domains. A further example of the utility of the 4D data is illustrated in Figure 4. Again, we compare the appropriate 4D slice (b; V607γ2 as the destination methyl) with slices from the 3D HMQC−NOESY−HMQC (a) and 3D HMQC−NOESY (c) data sets. In the analysis of the 3D data sets, the identities of cross-peaks are established by symmetry-related correlations and/or by matching ^{13}C (F_1) and ^1H (F_1) chemical shifts of cross-peaks in corresponding slices in HMQC−NOESY−HMQC and HMQC−NOESY maps (same $F_2(^{13}C)$ frequency, for example slices of Figure 4a,c) with cross-peaks in 2D $^1H-^{13}C$ spectra (Figure 4d−g). In practice, however, we prefer the 4D data set for assignment since ^{13}C and ^1H frequencies are correlated immediately from peak positions in the F_1-F_2 planes, shown for peaks numbered 1−4 in Figure 4b.

The high resolution of the reconstructed 4D data set is made clear by noting that L696 δ1 and L178 δ1 are nearly overlapped in the 2D correlation map (Figure 4d), yet it is still possible to unambiguously assign peak 1 to an NOE between V607 γ2 and L696 δ1.

Addition of each new dimension to a multidimensional NMR spectrum leads to a $\sqrt{2}$ loss in sensitivity.[46] Therefore, all other factors equal, 3D spectra are $\sqrt{2}$ more sensitive than their 4D counterparts. This sensitivity advantage disappears, however, if two 3D spectra are to be recorded to replace a single 4D data

(46) Ernst, R. R.; Bodenhausen, G.; Wokaun, A. *Principles of Nuclear Magnetic Resonance in One and Two Dimensions*; Oxford University Press: Oxford, 1987.

set. In practice, the relative sensitivities of 3D and 4D spectra depend on a combination of several factors, including differences in acquisition times in the indirect dimensions of the data sets and additional magnetization transfer steps in the 4D experiments. Some of the sensitivity losses associated with the 4D data set can be recouped by the nonlinear sampling approach, whereby data acquisition can be matched to the signal decay.[23] Clearly, the most relevant and direct way of assessing the quality (see below) of the reconstructed 4D experiment presented here is through direct comparison with the two conventionally acquired 3D spectra, ^{13}C-separated 3D HMQC−NOESY, and 3D F_1-^{13}C, F_2-^{13}C HMQC−NOESY−HMQC that might be recorded in its place. It is worth emphasizing that it is difficult to rigorously compare the signal-to-noise ratio of the reconstructed 4D data set with 3D matrices recorded and processed in the conventional manner because of the uneven distribution of noise in the 4D, as discussed above. Here, we define very loosely the quality of data in terms of the numbers of NOE correlations that are observed in the data set. A systematic analysis of the 4D spectrum performed in the same manner as illustrated in Figure 4 (i.e., comparing F_1-F_2 planes with the corresponding planes in the 3D data sets) shows similar numbers of correlations (recorded in 6.5 days) relative to the 3D spectra that were each recorded in 4.5 days. Of the 385 NOEs that were selected conservatively in a combined analysis of the 3D and 4D data, 350 of the correlations were observed in (and could be assigned from) the 4D data, while approximately 10% of the cross-peaks could not be uniquely assigned from the 3D spectra alone. In structural studies of high molecular weight proteins, such as MSG, where only a small number of NOEs/residue are measured, even an additional 10% of restraints is significant. That this is the case is made clear in our recent global fold determination of MSG, based in very large part on the NOE data presented here, where, on average, one long-range NOE/residue (NOEs between amino acids i and j with $|i - j| > 3$) was obtained.[47]

Structural Information from Methyl−Methyl NOEs in MSG. Distance restraints derived from NOE cross-peaks are a critical source of structural information for the ab initio determination of solution global folds of large proteins.[7,11,48,49] In studies with a highly deuterated, methyl-protonated sample of maltose-binding protein, for example, methyl−methyl NOEs accounted for greater than 30% of the measured long-range contacts,[7] while the corresponding number is approximately 60% for MSG. Figure 5 shows the distribution of methyl−methyl NOEs in apo-MSG obtained from the reconstructed 4D NOESY spectrum, as a function of the distance between the geometrical centers of methyl protons calculated from the X-ray structure of the glyoxylate-bound enzyme.[33] Although the fraction of potential NOE contacts that are observed falls off sharply for distances exceeding 6 Å, some NOEs between methyls that are up to ∼10 Å apart are measured. Methyl NOEs are particularly

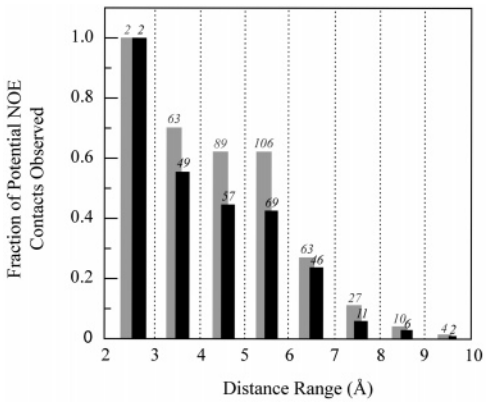


Figure 5. Fraction of potential methyl−methyl NOEs obtained from a preliminary analysis of the reconstructed 4D NOESY spectrum as a function of distance, calculated from the X-ray structure of glyoxalate-bound MSG (PDB access code[33] 1d8c). The distances were calculated between pseudo-atoms representing geometric averages of the three methyl proton positions. The number of detected cross-peaks is indicated on top of the vertical bars corresponding to each distance range. The total number of NOEs (long and short range) are indicated by the gray bars, while NOEs of the long-range variety ($|i - j| > 3$) are tabulated in black.

critical in the case of MSG, where over 50% of the protein and where approximately 75% of the regular secondary structure is helical. In these regions, long-range backbone HN−HN contacts are scarce or missing completely.

Over 50 structurally important NOE contacts between the molecular core and peripheral domains of MSG could be identified from the combined analysis of 3D and 4D methyl−methyl NOE spectra. Methyl−methyl, methyl−HN, and HN−HN NOEs, along with orientational and dihedral angle restraints, have been used to define the global fold of apo-MSG de novo,[47] without input from the available X-ray structures of the enzyme.[33] The methyl−methyl NOEs proved crucial in this regard, despite the fact that the great majority were interpreted only assuming a distance range of 1.8−8 Å to account for the difficulty in quantification due to a variety of factors, including, principally, differences in relaxation times among methyls and spin diffusion effects. Notably, while dipolar couplings define the relative orientation of the domains, the positions of the domains cannot be established without distance restraints.

In summary, a simple and sensitive pulse scheme has been presented for recording methyl−methyl NOEs with high resolution. Optimal performance of the experiment is achieved in applications involving highly deuterated, methyl-protonated proteins, as described for MSG above. To achieve the high resolution needed to confidently assign NOEs in a large system, such as MSG, a nonuniform (sparse) data acquisition procedure has been implemented and the experimental data matrix subsequently processed using multidimensional decomposition. The resolution achieved in the reconstructed 4D data set is comparable to that obtained in conventional 3D NOE spectroscopy. The methyl−methyl distance restraints obtained in the present study are critical to the success of structural studies of MSG and, in combination with other important classes of restraints, have led to the de novo determination of the global fold of this 723 residue enzyme.

Appendix

Description of the Algorithm used for the Selection of the Indirect−Detect Points in the Sparsed 4D Data Set. The basic

(47) Tugarinov, V.; Choy, W. Y.; Orekhov, V.; Kay, L. E. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 622−627.
(48) Venters, R. A.; Metzler, W. J.; Spicer, L. D.; Mueller, L.; Farmer, B. T. *J. Am. Chem. Soc.* **1995**, *117*, 9592−9593.
(49) Metzler, W. J.; Wittekind, M.; Goldfarb, V.; Mueller, L.; Farmer, B. T. *J. Am. Chem. Soc.* **1996**, *118*, 6800−6801.
(50) Shaka, A. J.; Keeler, J.; Frenkiel, T.; Freeman, R. *J. Magn. Reson.* **1983**, *52*, 335−338.
(51) Levitt, M.; Freeman, R. *J. Magn. Reson.* **1978**, *33*, 473−476.
(52) Marion, D.; Ikura, M.; Tschudin, R.; Bax, A. *J. Magn. Reson.* **1989**, *85*, 393−399.

approach is given below with additional details available from the authors upon request. For each value of (t_1, t_2, t_3, $\phi 1$, $\phi 2$, $\phi 3$), we calculate an initial selection probability, P_S, as

$$P_S(t_1,t_2,t_3,\phi 1,\phi 2,\phi 3) = \exp(-R_{2,F1}t_1 - R_{2,F2}t_2 - R_{2,F3}t_3) \quad \text{(A1)}$$

In eq A1, t_1, t_2, and t_3 are acquisition times in the indirectly detected dimensions of the 4D data set; ϕi is either 0 (phase x) or 1 (phase y) to generate cosine- or sine-modulated signals in dimension i, respectively (see Figure 1), and $R_{2,Fi}$ ($i = 1-3$) is the average transverse relaxation rate of the coherence of interest in t_i (here, we set $R_{2,F1} = R_{2,F2} = R_{2,F3} = 25\ s^{-1}$, as described previously[30]). P_S is then compared with a random number, RN, between 0 and 1; if $P_S >$ RN, the point is added to a table of acquisition times and phases to be used in recording the 4D data set. If at the end of the process a fraction, $f1$, of the potential data points has been selected, while a fraction, $f2$, is desired ($f1 > f2$), the process is repeated using a prefactor in front of the exponent in eq A1 of approximately $f2/f1$.

Acknowledgment. This work was supported by grants from the Canadian Institutes of Health Research (CIHR) to L.E.K., and from the Swedish Research Council (624-2004-31) and the Swedish National Allocation Committee (SNIC 3/04-44) to V.O. The authors thank Dr. D. M. Korzhnev for helpful discussions. V.T. acknowledges the support of the Human Frontiers Science Program. L.E.K. holds a Canada Research Chair in Biochemistry.

JA044032O