Facilitated Assignment of Large Protein NMR Signals with Covariance Sequential Spectra Using Spectral Derivatives

Bradley J. Harden,[‡] Scott R. Nichols,[‡] and Dominique P. Frueh*

Department of Biophysics & Biophysical Chemistry, Johns Hopkins University School of Medicine, 701 Hunterian, 725 North Wolfe Street, Baltimore, Maryland 21205, United States

S *Supporting Information*

ABSTRACT: Nuclear magnetic resonance (NMR) studies of larger proteins are hampered by difficulties in assigning NMR resonances. Human intervention is typically required to identify NMR signals in 3D spectra, and subsequent procedures depend on the accuracy of this so-called peak picking. We present a method that provides sequential connectivities through correlation maps constructed with covariance NMR, bypassing the need for preliminary peak picking. We introduce two novel techniques to minimize false correlations and merge the information from all original 3D spectra. First, we take spectral derivatives prior to performing covariance to emphasize coincident peak maxima. Second, we multiply covariance maps calculated with different 3D spectra to destroy erroneous sequential correlations. The maps are easy to use and can readily be generated from conventional triple-resonance experiments. Advantages of the method are demonstrated on a 37 kDa nonribosomal peptide synthetase domain subject to spectral overlap.

Nuclear magnetic resonance (NMR) is a primary tool for structural, dynamic, kinetic, and thermodynamic studies of proteins. However, to harness the full potential of the method, resonances in NMR spectra must be assigned. This task is hindered by frequency degeneracies and signal overlap, as occur in large proteins, disordered proteins, or in some α-helical proteins. This limitation is due in large part to traditional sequential assignment procedures, which require parallel analysis of multiple 3D spectra, early human intervention to identify signals (peak picking), and consequently, constant scrutiny. HN correlation maps are the principal tools in NMR studies of proteins, as each (H,N) correlation reports on an individual amino acid in the protein. Assignment of NMR resonances relies on identifying (H,N) correlations that belong to sequential residues. Two distinct types of 3D spectra convey this information. In the first type, an additional dimension encodes carbon chemical shifts of both the same and the preceding residue (Intra-3D). The second type reports only carbon chemical shifts of preceding residues (Seq-3D). The assignment procedure consists of identifying correlations (H(i+1),N(i+1),C(i)) for residue i+1 in the Seq-3D that feature carbon shifts matching that of a correlation (H(i),N(i),C(i)) found in the Intra-3D. This process is performed using C^α (with HNCA for Intra-3D and HN(CO)CA for Seq-3D), CO (HN(CA)CO and HNCO), and when possible, C^β (HN(CA)-CB and HN(COCA)CB) chemical shifts. The procedure comprises a series of steps. First, (H,N,C) correlations are identified by peak picking. Next, H/C (or N/C) strips are generated for each peak in each spectrum. The strip of a target residue is selected in Intra-3D, and a software package sorts all strips of Seq-3D according to the difference in carbon frequencies as determined by peak picking (strip matching). The procedure requires simultaneous analysis of different carbons (C^α, CO, and C^β) to identify true sequential residues and eliminate accidental degeneracies in carbon frequencies. Clearly, the procedure relies on the accuracy of peak picking, which greatly deteriorates in the presence of frequency degeneracies. Unpicked correlations will not be represented during strip matching. Carbon frequencies of different spectra can be mispaired with (H,N) correlations that overlap; for example, the C^α of residue i could be paired with the CO of residue j. Strip matching will either be unsuccessful or, worse, erroneous. To overcome the limitations of preliminary peak picking, we have designed spectral manipulations that replace this convoluted assignment procedure with a simple inspection of four 3D correlation maps. Each map reports on the combined sequential information contained within *all* pairs of Intra-3D and Seq-3D spectra. The four correlation maps provide correlations of the form (H(i),N(i),H(i+1)), (H(i),N-(i),N(i+1)), (H(i),N(i),H(i−1)), and (H(i),N(i),N(i−1)) and permit direct identification of sequential residues in (H,N) correlation maps. The method employs covariance,[1−9] albeit with spectra suitably modified to minimize artifacts. Covariance and related methods were suggested as tools to help protein assignment by creating novel correlations,[10−13] but artifacts have limited applications to small proteins where such artifacts can be identified. Another elegant solution was tailored to sequential assignment,[14] but it required peak picking and is hence vulnerable to its associated limitations. Overall, covariance methods have not been widely adopted for resonance assignment. Here, sequential correlation maps with minimal artifacts are obtained by (i) taking a spectral derivative prior to covariance between pairs of Intra and Seq spectra and (ii) multiplying the resulting covariance correlation maps to combine the information provided separately by different carbon dimensions into a single spectrum. The advantages of using our covariance sequential correlation maps (COSCOMs) are illustrated with the 37 kDa E_A domain of the nonribosomal peptide synthetase protein HMWP2.

Covariance NMR can be used to provide a spectral representation of the sequential assignment procedure; however, preliminary treatment of the original spectra is needed to minimize artifacts. To identify and overcome shortfalls of covariance NMR in the presence of near degenerate frequencies, we first reformulate the sequential assignment procedure in a context that over-represents overlap: "Amongst all (H,C^α) correlations in HN(CO)CA, find the one that possesses a C^α frequency matching the observed C^α in HNCA for an (H,N) correlation" and likewise for all pairs of spectra. The mathematical formulation of this procedure consists of calculating the covariance matrix between the H/C projection of HN(CO)CA, referred to as 2D-H(NCO)CA, and each H/C plane of HNCA (for all nitrogen indices). Using the formalism of Brüschweiler and co-workers,[6,7,15] the following 3D array can be constructed:

$$\mathrm{HNH_s}ca(a, b, c) = \frac{1}{D-1} \sum_{d=1}^{D} \mathrm{HN\widetilde{CA}}(a, b, d) \cdot$$
$$\mathrm{H(NCO)\widetilde{CA}}(c, d) \qquad (1)$$

The symbol "~" indicates that the means along the carbon dimensions have been subtracted from each data point in all spectra.[8] Indices a and c represent the HNCA and 2D-H(NCO)CA ^{1}H dimensions, respectively; b is the index along the HNCA ^{15}N dimension, and d is the common index along the ^{13}C dimensions of both spectra (each with D points). The resultant 3D spectrum, $\mathrm{HNH_s}ca$, correlates (H,N) correlations of HNCA with sequential $\mathrm{H_s}$ resonances of 2D-H(NCO)CA. $\mathrm{HNH_s}ca$ provides correlations (H(i),N(i),H($i+1$)) and is an array of covariance matrices $\mathrm{HH_s}$ dispersed along a nitrogen dimension. Unfortunately, false correlations would appear in such a correlation map. To identify the origin of these artifacts and to design a solution, we reformulate the mathematics of covariance NMR into two distinct steps: the element-wise product of two C^α vectors and subsequent summation over the elements of the resulting vector. First we define a vector

$$\vec{v}_{a,c}(d) = \vec{ca}_a(d) \odot \vec{ca}_{-1_c}(d) \qquad (2)$$

where $\vec{ca}_a$ and $\vec{ca}_{-1_c}$ are vectors representing 1D C^α traces at ^{1}H frequencies defined by the index a in HNCA and c in 2D-H(NCO)CA, respectively. Here $\odot$ denotes the element-wise product, and the symbol "~" has been omitted for clarity. Each point (a,c) in the plane $\mathrm{HH_s}$ is proportional to the sum of the elements of the vector $\vec{v}_{a,c}$:

$$\mathrm{HH_s}(a, c) = \frac{1}{D-1} \sum_{d=1}^{D} \vec{v}_{a,c}(d) \qquad (3)$$

By observing the individual C^α vectors and their associated element-wise products $\vec{v}$ prior to summation, we can discern the origin of artifacts in $\mathrm{HH_s}$ that have plagued related applications of covariance NMR thus far.

Figure 1 uses simulated data to demonstrate the source of artifacts in covariance NMR spectra. Figure 1a,b displays the same vector $\vec{ca}_a$ at an index H(i) = a in ^{1}H of HNCA. Figure 1c displays a vector $\vec{ca}_{-1_c*}^{*}$ that contains the true sequential peak at index H($i+1$) = $c*$ in 2D-H(NCO)CA, while Figure 1d displays $\vec{ca}_{-1_c X}^{X}$ containing a nearly degenerate C^α peak at index H($i+1$) = c^X. The element-wise products of $\vec{ca}_a$ with $\vec{ca}_{-1_c*}^{*}$
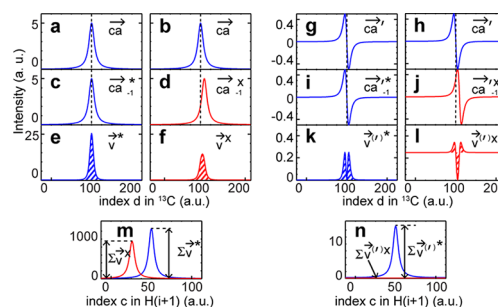

Figure 1. Spectral derivatives suppress spurious correlations in covariance NMR spectra. The $*$ and X indicate true and erroneous correlations, respectively: (a,b) $\vec{ca}$ at an index H(i) = a (see eqs 2 and 3); (c) $\vec{ca}_{-1}^{*}$ at an index H($i+1$) = $c*$; (d) $\vec{ca}_{-1}^{X}$ for an erroneous correlation at H($i+1$) = c^X. (e,f) Element-wise products of $\vec{ca}$ with $\vec{ca}_{-1}^{*}$ ($\vec{v}*$) and $\vec{ca}$ with $\vec{ca}_{-1}^{X}$ ($\vec{v}^X$). (g–j) Derivatives ($\vec{ca}'$) of $\vec{ca}$ vectors in a–d, respectively. (k,l) Element-wise products of $\vec{ca}'$ with $\vec{ca}_{-1}^{'*}$ ($\vec{v}^{(')}*$) and $\vec{ca}'$ with $\vec{ca}_{-1}^{'X}$ ($\vec{v}^{(')X}$). $\vec{v}^{(')}*$ and $\vec{v}^{(')X}$ denote the products of the derivatives and not the derivatives of the products. (m) H($i+1$) trace in $\mathrm{HH_s}$ at index H(i) = a, *without* derivatives. (n) Corresponding H($i+1$) trace *with* derivatives.

($\vec{v}*$) and $\vec{ca}_a$ with $\vec{ca}_{-1_c X}^{X}$ ($\vec{v}^X$) are shown in Figure 1e and f, respectively. Summing the vectors $\vec{v}*$ and $\vec{v}^X$ provides the amplitudes of $\mathrm{HH_s}$ at indices (a,$c*$) and (a,c^X) in Figure 1m. We can see a false correlation resulting from partial overlap in the C^α dimension. This artifact can be reduced by taking the derivative along the C^α dimensions prior to covariance (Figure 1g–j). In this case, $\vec{v}^{(')}*$ now contains only positive elements (Figure 1k), while $\vec{v}^{(')X}$ contains both positive and negative elements due to the mismatched inflection points in $\vec{ca}_a'$ and $\vec{ca}_{-1_c X}^{'X}$ (Figure 1l). Summing $\vec{v}^{(')}*$ results in a positive correlation at index (a,$c*$) in Figure 1n, whereas the sum of $\vec{v}^{(')X}$ gives zero amplitude at index (a,c^X). Here, the degree of C^α frequency degeneracy was chosen to completely suppress artifacts when using spectral derivatives. Stronger degeneracy would result in positive yet reduced artifacts, while weaker degeneracy would create negative artifacts that can safely be ignored.

Figures 2 and 3 illustrate experimentally the effectiveness of artifact suppression in covariance matrices when using derivatives of original spectra. Figure 2 shows $\vec{ca}$ and $\vec{ca}_{-1}$ vectors as well as their element-wise products $\vec{v}$, and Figure 3a,b shows traces from the covariance matrix $\mathrm{HH_s}$. Although the vector $\vec{ca}$ shown in Figure 2a should only correlate with $\vec{ca}_{-1}^{*}$ (Figure 2b), it also correlates, among others, with $\vec{ca}_{-1}^{X}$ (Figure 2c). Both vectors $\vec{v}*$ and $\vec{v}^X$ (Figure 2d,e) have only positive elements that, after summation, give rise to the signals labeled $*$ and X in Figure 3a. Results are improved if the derivatives of the vectors $\vec{ca}$ and $\vec{ca}_{-1}$ are used for covariance analysis (Figure 2f–h). After element-wise multiplication (Figure 2i,j) and summation, the amplitude of the artifact is either reduced or becomes negative in the covariance matrix (Figure 3b, signal labeled X). Thus, true sequential correlations can be distinguished to a large extent from contributions of residues with carbons of nearly identical frequencies.

A single COSCOM conveys information obtained with four separate spectra. The traditional sequential assignment procedure requires that C^α and CO strips, for example, be analyzed in parallel to distinguish accidental frequency degeneracies from true sequential correlations. The COSCOM
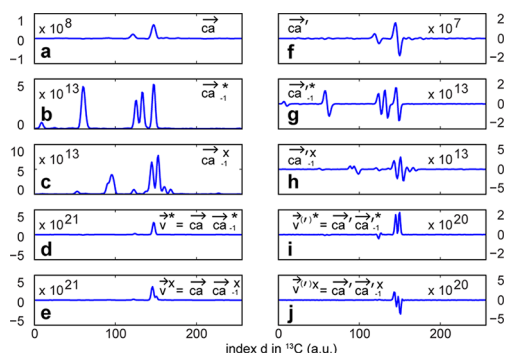
Figure 2. Differentiating between true sequential correlations (*) and erroneous correlations due to partially overlapping signals (X): (a) $\vec{ca}$ (C^α 1D trace) from HNCA at H(i) = 7.558 ppm and N(i) = 120.023 ppm; (b) $\vec{ca}^*_{-1}$ from 2D-H(NCO)CA at H($i+1$) = 7.608 ppm; (c) $\vec{ca}^X_{-1}$ from 2D-H(NCO)CA at H($i+1$) = 8.602 ppm. (d) Element-wise product of $\vec{ca}$ with $\vec{ca}^*_{-1}$ ($\vec{v}^*$). (e) Element-wise product of $\vec{ca}$ with $\vec{ca}^X_{-1}$ ($\vec{v}^X$). (f–h) Derivatives of $\vec{ca}$ vectors in a–c, respectively. (i,j) Element-wise products of $\vec{ca}'$ with $\vec{ca}'^*_{-1}$ ($\vec{v}^{(')*}$) and $\vec{ca}'$ with $\vec{ca}'^X_{-1}$ ($\vec{v}^{(')X}$), respectively. The normalized sum of the elements of $\vec{v}^*$, $\vec{v}^{(')*}$, $\vec{v}^X$, and $\vec{v}^{(')X}$ lead to correlations that are highlighted by the symbols * and X in Figure 3a,b. Data collected with the 37 kDa E_A domain.
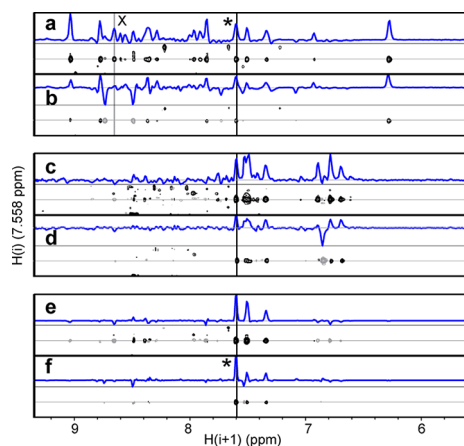

Figure 3. Identification of unique proton sequential correlations when using spectral derivatives and when multiplying COSCOMs. (a,b) HNH$_s$ca, (c,d) HNH$_s$co, and (e,f) HNH$_s$caco obtained by multiplying a and c and b and d, respectively. (a,c,e) Correlations obtained without derivatives in the carbon dimensions. (b,d,f) Correlations obtained with derivatives. Covariance was performed with the MATLAB[16] covariance NMR toolbox.[6] The amplitudes of signals labeled * are $\Sigma\vec{v}^*$ and $\Sigma\vec{v}^{(')*}$ in a and b, respectively, while those labeled X are $\Sigma\vec{v}^X$ and $\Sigma\vec{v}^{(')*X}$, with the vectors $\vec{v}$ as defined in Figure 2. The * denotes the true correlation. Data collected with the 37 kDa E_A domain.

procedure applied to C^α in the previous paragraph can also be applied to HN(CA)CO and HNCO to produce HNH$_s$co spectra (Figure 3c,d). Because HNH$_s$ca and HNH$_s$co provide sequential correlations along a common proton dimension, placing the spectra side-by-side (or overlaying them) readily identifies common sequential correlations. Alternatively, the sequential information contained in each COSCOM spectrum can be combined via element-wise multiplication, permitting further reduction in artifacts due to the destructive interference of erroneous correlations. Indeed, Figure 3e,f shows that multiplication of HNH$_s$ca and HNH$_s$co to produce HNH$_s$caco removes a majority of the erroneous correlations that resulted from accidental degeneracies in C^α and CO carbon frequencies.

Without using spectral derivatives, three sequential proton candidates remain in HNH$_s$caco (Figure 3e). However, when taking the derivative prior to covariance, only a single correlation remains. The other two correlations are severely damped, since they originate from partial overlap in ^{13}C signals, and the true sequential correlation is identified (Figure 3f). In the end, rather than analyzing four carbon dimensions in four 3D spectra, the sequential correlation is unambiguously identified with the single ^1H trace of HNH$_s$caco of Figure 3f.

Optimal COSCOMs are obtained when all dimensions of the original spectra are probed. So far, we have investigated the quality of covariance maps in a situation that exacerbates the effect of spectral crowding, namely, by using a 2D projection of the 3D-HN(CO)CA. However, in practice, two 3D spectra are available, and the sequential assignment procedure can be reformulated as "find which (H,N) correlations in HN(CO)CA possess C^α frequencies matching those observed for (H,N) correlations in HNCA." This sentence translates to:

$$HNH_sN_sca(a, b, c, e)$$
$$= \frac{1}{D-1}\sum_{d=1}^{D} \widetilde{HNCA}(a, b, d)\cdot\widetilde{HN(CO)CA}(c, e, d)$$

(4)

The resultant 4D spectrum is the HNH$_s$N$_s$ca featuring correlations (H(i),N(i),H($i+1$),N($i+1$)). The index e spans the HN(CO)CA nitrogen dimension. However, the computational implementation of eq 4 is problematic, as the 4D spectrum rapidly exceeds memory capacities. Instead, all four 3D projections of the 4D spectrum are calculated on the fly. Covariance spectra originating from different carbon correlations are also multiplied on the fly, resulting in computational time and disk-space savings. In the end, our MATLAB[16] processing script (available upon request) produces four 3D COSCOMs: HNH$_s$caco providing (H(i),N(i),H($i+1$)), HNN$_s$caco providing (H(i),N(i),N($i+1$)), H$_s$N$_s$Hcaco providing (H(i),N(i),H($i-1$)), and H$_s$N$_s$Ncaco providing (H(i),N(i),N-($i-1$)). These COSCOMs are renamed HNH$_{suc}$, HNN$_{suc}$, HNH$_{pre}$, and HNN$_{pre}$, respectively. Tests performed on the well-known protein ubiquitin demonstrate successful suppression of false correlations, and only two pairs of residues (out of 70) could not be linked with COSCOMs (Supporting Information Figure S1). The H, C^α, and CO chemical shifts of G47 and G75 are nearly degenerate. These residues are nevertheless assigned by identifying correlations for surrounding residues (e.g., A46 identifies G47). Alternatively, the correct assignment is also revealed by close inspection of the original 3D spectra. The latter observation highlights that COSCOMs provide a means to rapidly assign residues and overcome the limitations of peak picking, but it is nevertheless a method to supplement rather than supplant conventional protocols.

The advantages of sequential covariance spectra over traditional methods are exemplified with a 37 kDa monomeric protein. We used COSCOMs with a 37 kDa protein for which backbone assignment had been in progress for about 6 months with conventional methods. Figure 4 showcases both the ease of use of COSCOMs and their ability to overcome the limitations of peak picking. Four COSCOMs were used to scan the unassigned HN-TROSY of the protein. The backbone signals of L189−Q196 were simultaneously picked and assigned within only 30 min (Figure S2). HNCA, HN(CA)CO, and HN(CA)CB were used for residue type assignment. In contrast, only A194, G195, and Q196 had been assigned with
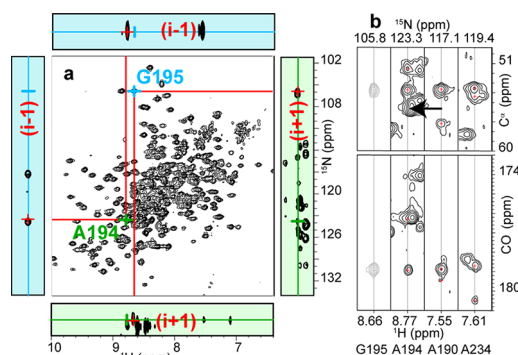
13108

dx.doi.org/10.1021/ja5058407 | *J. Am. Chem. Soc.* 2014, 136, 13106–13109

Figure 4. Scanning HN-TROSY with COSCOMs overcomes shortfalls of strip matching. (a) HN-TROSY of the 37 kDa E_A with strips of HNH_{pre} (left) and HNN_{pre} (top) at the (H,N) coordinates of G195 (cyan), as well as strips of HNH_{suc} (bottom) and HNN_{suc} (right) at the coordinates of A194 (green). (b) Strip matching for the predecessor of G195. A194 was initially missing; its C^α was erroneously picked at the position indicated by the arrow. Correlations to A190 and A234 (very weak) are seen in HNH_{pre} and HNN_{pre} (unlabeled).

strip matching. Several mistakes had impeded proper assignment of this segment of residues. First, the signals of A190 were erroneously assigned to A194 as all ^{13}C sequential correlations in G195 (C^α, C^β, and CO) had frequencies matching those of A190. Second, A194 had not been identified by strip matching because its C^α had been mis-assigned. Finally, the signals of L189 had not been picked. When scanning G195 with HNH_{pre} and HNN_{pre} (Figure 4a top and left), A194 (labeled with a red "+" in Figure 4a) and A190 (unlabeled) were identified. NOESY-HN-TROSY identified which of the signals of A190 and A194 belonged to the predecessor of G195. Sequential residues were rapidly identified with COSCOMs down to A190, previously erroneously assigned to A194. Weak correlations in HNH_{pre} and HNN_{pre} identified a new (H,N) correlation for the predecessor of A190, L189. L189 had previously escaped peak picking because its weak (H,N) correlation overlaps partially with that of a very intense signal. The low amplitudes of L189 signals prevented further assignment. The complete sequence of residues L189−Q196 was assigned in a matter of minutes by simple scanning of HN-TROSY with COSCOMs, whereas strip matching only provided the correct assignment for two of these residues. The comparison between assignments provided by COSCOMs and those obtained with traditional methods was carried out for 2 weeks. Another three mistakes were corrected, and eight new links were found. In the end, 70% of the backbone resonances were assigned. Absence of correlations in COSCOMs (as in L189) demonstrates that signals are missing in the original spectra and more sensitive data must be recorded to complete assignment. Without COSCOMs, significant time would be wasted seeking signals of sequential residues that may not exist.

In conclusion, we have presented a method that enables sequential assignment of NMR resonances upon simple inspection of correlation maps bypassing preliminary peak picking and associated limitations. We have shown that using spectral derivatives in the dimension to which covariance is applied either removes artifacts or clearly identifies them by a change of sign. Further improvements were obtained by multiplying covariance spectra that convey the same sequential information. The resulting sequential correlations allow rapid and reliable assignment of backbone resonances. Human error

is minimized since the information provided by the original 3D spectra is combined mathematically before any user interaction is required. The method does not require data other than those traditionally used for assignment, and it is readily applicable to projects that may have stalled due to errors in peak picking. In the end, we have developed a tool that should greatly facilitate resonance assignment, which is often a bottleneck in NMR investigations of biological macromolecules. As such, CO-SCOMs should be an asset in widening the range of proteins for which NMR can be used.

■ ASSOCIATED CONTENT

ⓈSupporting Information

Sample preparation, NMR data acquisition, COSCOM strips for ubiquitin, detailed assignment of L189−Q196 for E_A. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author
dfrueh@jhmi.edu

Author Contributions
‡These authors contributed equally.

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Blinov, K. A.; Larin, N. I.; Williams, A. J.; Zell, M.; Martin, G. E. *Magn. Reson. Chem.* **2006**, *44*, 107.
(2) Chen, Y.; Zhang, F.; Snyder, D.; Gan, Z.; Brüschweiler-Li, L.; Brüschweiler, R. *J. Biomol. NMR* **2007**, *38*, 73.
(3) Bingol, K.; Salinas, R. K.; Brüschweiler, R. *J. Phys. Chem. Lett.* **2010**, *1*, 1086.
(4) Blinov, K. A.; Larin, N. I.; Kvasha, M. P.; Moser, A.; Williams, A. J.; Martin, G. E. *Magn. Reson. Chem.* **2005**, *43*, 999.
(5) Snyder, D. A.; Brüschweiler, R. *J. Phys. Chem. A* **2009**, *113*, 12898.
(6) Short, T.; Alzapiedi, L.; Brüschweiler, R.; Snyder, D. *J. Magn. Reson.* **2011**, *209*, 75.
(7) Zhang, F.; Brüschweiler, R. *J. Am. Chem. Soc.* **2004**, *126*, 13180.
(8) Trbovic, N.; Smirnov, S.; Zhang, F.; Brüschweiler, R. *J. Magn. Reson.* **2004**, *171*, 277.
(9) Brüschweiler, R.; Zhang, F. *J. Chem. Phys.* **2004**, *120*, 5253.
(10) Bartels, C.; Wuthrich, K. *J. Biomol. NMR* **1994**, *4*, 775.
(11) Kupce, E.; Freeman, R. *J. Am. Chem. Soc.* **2006**, *128*, 6020.
(12) Benison, G.; Berkholz, D. S.; Barbar, E. *J. Magn. Reson.* **2007**, *189*, 173.
(13) Chen, K.; Delaglio, F.; Tjandra, N. *J. Magn. Reson.* **2010**, *203*, 208.
(14) Lescop, E.; Brutscher, B. *J. Am. Chem. Soc.* **2007**, *129*, 11916.
(15) Brüschweiler, R. *J. Chem. Phys.* **2004**, *121*, 409.
(16) *Matlab Reference Guide*, Natick, MA, 1992.