Published in final edited form as:

J Magn Reson. 2015 November ; 260: 83–88. doi:10.1016/j.jmr.2015.09.007.

Effortless assignment with 4D covariance sequential correlation maps

Bradley J. Harden[†], Subrata H. Mishra[†], and Dominique P. Frueh[†]

[†]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 701 Hunterian 725 N Wolfe St Room, Baltimore, MD 21205

Abstract

Traditional Nuclear Magnetic Resonance (NMR) assignment procedures for proteins rely on preliminary peak-picking to identify and label NMR signals. However, such an approach has severe limitations when signals are erroneously labeled or completely neglected. The consequences are especially grave for proteins with substantial peak overlap, and mistakes can often thwart entire projects. To overcome these limitations, we previously introduced an assignment technique that bypasses traditional pick peaking altogether. Covariance Sequential Correlation Maps (COSCOMs) transform the indirect connectivity information provided by multiple 3D backbone spectra into direct (H, N) to (H, N) correlations. Here, we present an updated method that utilizes a single four-dimensional spectrum rather than a suite of three-dimensional spectra. We demonstrate the advantages of 4D-COSCOMs relative to their 3D counterparts. We introduce improvements accelerating their calculation. We discuss practical considerations affecting their quality. And finally we showcase their utility in the context of a 52 kDa cyclization domain from a non-ribosomal peptide synthetase.

Graphical Abstract

Keywords

backbone assignment; peak picking; covariance sequential correlation maps; COSCOM; 4D

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

Thanks to improvements in cryoprobe technology, pulse sequence development, and isotopic labeling, it is becoming more commonplace to study proteins in the 35–50 kDa range by NMR spectroscopy. Analysis of these overlapped and complex spectra now presents challenges that the NMR community could not have imagined in earlier days. When studying large proteins by NMR, substantial effort is required to assign each resonance to its corresponding nucleus. Traditional protein NMR assignment procedures seek to identify (H, N) systems belonging to sequential residues in the polypeptide chain. This task is accomplished using through-bond correlations to CA, CB, and CO carbon nuclei along the peptide backbone. These (H, N, C) correlations are detected using two classes of 3D experiments. The first provides correlations between (H, N) spin systems and carbon nuclei of both the same and the preceding residue, i.e. correlations of the form (H_i, N_i, C_i) and (H_i, N_i, C_{i-1}) . We refer to this type of experiment as Intra-3D to emphasize the former correlation. The second type, Seq-3D, provides correlations only with carbons of the preceding residue, i.e. (H_i, N_i, C_{i-1}) . Thus, HNCA, HN(CA)CB, and HN(CA)CO are Intra-3D experiments whereas HN(CO)CA, HN(COCA)CB, and HNCO are Seq-3D experiments.

The traditional assignment process relies on users to first identify and label all correlations in each 3D spectrum, a process usually termed “peak-picking.” Next, NMR assignment software sifts through these “picked peaks” to identify pairs of (H, N) correlations in Intra-3D and Seq-3D spectra with common carbon frequencies. Such pairs, of the form (H_i, N_i, C_i) and (H_{i+1}, N_{i+1}, C_i) , may belong to (H, N) spin systems of sequential residues. The results of this search are presented to users in the form of spectrum strips that are taken along the carbon dimensions of the various 3D spectra. Users are then tasked with accepting or rejecting the proposed sequential (H, N) link based on comparisons between strips from up to three different pairs of Intra-3D and Seq-3D experiments. Consequently, so-called “strip-matching” is critically dependent upon accurate and complete peak-picking. In cases where signals have been erroneously picked or entirely neglected, the sequential assignment suggestions offered by software will be incorrect or absent. In the former case, the mis-assignment of residues can have disastrous consequences including erroneous identification of structural constraints or binding sites. If signals have been neglected, the inability to extend a sequence of assigned residues can thwart entire projects. Furthermore, when searching for un-picked and often weak peaks, researchers may waste valuable time investigating potential “signals” that are in fact entirely noise. Because accurate peak-picking is absolutely essential, even small errors by experienced users can have disastrous consequences. The time-tested approach of preliminary peak-picking is more arduous for larger systems, and human error can often stall projects. We endeavor to remove the Achilles’ heel of backbone assignment by using covariance NMR methods to bypass peak-picking altogether.

To overcome some of the vulnerabilities of peak-picking, we recently developed a complementary assignment procedure that circumvents it entirely [1]. Instead, the information provided separately by pairs of Intra-3D and Seq-3D spectra are mathematically combined using covariance NMR [2–7] to create a suite of 3D spectra, dubbed 3D-

COSCOMs (for Covariance Sequential Correlation Maps), which feature 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}) . In combination, these four 3D maps identify candidate sequential (H, N) correlations without making any *a priori* assumptions.

Although 3D COSCOMs each harness the combined information of all original 3D spectra, assigning sequential correlations nevertheless requires parallel analysis of two different correlation maps, and much time can be spent finding the chemical shifts of the true sequential residue. To find a sequential candidate, users must jointly inspect proton and nitrogen sequential dimensions (e.g. H_{i+1} and N_{i+1}). In the presence of spectral crowding, as found in large proteins but also in disordered and alpha-helical proteins, there inevitably arise cases in which many correlations occur in these sequential dimensions. Users must then cross-reference all possible combinations of ^1H and ^{15}N signal positions observed in 3D-COSCOMs with those seen in 2D HN correlation maps (HSQCs [8] or TROSYs [9]). In effect, users are being tasked with mentally reconstructing a 2D plane from its corresponding 1D projections. In other words, users must mentally reconstruct a fourth dimension from two 3D spectra. Indeed, as previously mentioned [1], the 3D-COSCOMs correspond to the four possible projections of a 4D spectrum, with correlations of the form $(H_i, N_i, H_{i+1}, N_{i+1})$.

Here, we present a software package to directly generate 4D COSCOMs, bypassing the previously described limitations of 3D COSCOMs. We introduce a number of improvements aimed at accelerating their calculation, and we discuss in detail practical aspects affecting their quality and use. We demonstrate the advantages of 4D-COSCOMs relative to their 3D counterparts, and we showcase their utility in the context of a 52 kDa cyclization domain from the non-ribosomal peptide synthetase (NRPS) Yersiniabactin Synthetase.

Methods

We have designed a script to generate 4D COSCOMs. It is written in the MATLAB programming language and utilizes the covariance NMR toolbox [10]. It has been tested in both MATLAB 8.5 and the open-source alternative Octave 3.8.1, and it is available upon request from the corresponding author. Pairs of original 3D spectra (e.g. HNCA and HN(CO)CA) submitted to the script must satisfy two requirements: they must be in NMRPipe [11] format, and specific dimensions, both within and between different pairs of Intra-3D and Seq-3D spectra, must match each other in a point-by-point fashion. We have previously discussed procedures used to properly match dimensions [1,12] and further instructions are provided in the documentation accompanying our script.

All NMR spectra were recorded on Cy1, a 52 kDa cyclization domain excised from the NRPS Yersiniabactin Synthetase [13] at 25°C on an 800 MHz Varian spectrometer equipped with a Chili-Probe. The Cy1 construct used in our studies (Cy1H6) retained the C-terminal hexahistidine tag used for purification. Two labeling schemes were used during data collection: one with uniform ^2H , ^{15}N and ^{13}C labeling (CDN) and the other with ^1H and ^{13}C labeled methyl side-chains ($\delta 1$ position only) for residues Ile, Leu, and Val in an otherwise

uniform ^2H , ^{15}N and ^{12}C background (ILV). The expression and purification protocols for both isotopic labeling schemes have been discussed previously [12,14]. The CDN and ILV samples were concentrated to 650 and 640 μM respectively in the final NMR buffer: 20 mM sodium phosphate pH 7.0, 10mM NaCl, 1mM EDTA, 5mM DTT and 5% D_2O .

The NMR spectra used for the covariance procedure were acquired with the following parameters. The HNCA and the HN(CO)CA were recorded on the ILV sample while the HNCO and the HN(CA)CO were recorded on the CDN sample. The HNCA was acquired with a 1s recycling delay, 8 scans, and spectral parameters: ^1H (1200 complex points, 4.758 ppm carrier, 16000 Hz spectral width), ^{13}C (60 complex points, 58 ppm carrier, 5530 Hz spectral width), ^{15}N (48 complex points, 118 ppm carrier, 2836 Hz spectral width). The HN(CO)CA was acquired with a 1.1 s recycling delay, 32 scans, and spectral parameters: ^1H (810 complex points, 4.758 ppm carrier, 13500 Hz spectral width), ^{13}C (60 complex points, 58 ppm carrier, 6435 Hz spectral width), ^{15}N (33 complex points, 118 ppm carrier, 2836 Hz spectral width). The HNCO was acquired with a 1 s recycling delay, 8 scans, and spectral parameters: ^1H (825 complex points, 4.758 ppm carrier, 15001 Hz spectral width), ^{13}C (53 complex points, 176.976 ppm carrier, 2816 Hz spectral width), ^{15}N (53 complex points, 118 ppm carrier, 2836 Hz spectral width). The HN(CA)CO was acquired with a 1 s recycling delay, 32 scans, and spectral parameters: ^1H (825 complex points, 4.758 ppm carrier, 15009 Hz spectral width), ^{13}C (41 complex points, 176.976 ppm carrier, 2816 Hz spectral width), ^{15}N (52 complex points, 118 ppm carrier, 2836 Hz spectral width).

For successful calculation of covariance matrices, particular dimensions of the spectra must match each other in a point-by-point manner. To accomplish this, spectra were processed to eliminate the differences in spectral widths, number of points, and carrier position by appropriately zero-filling and extracting regions of the spectra. The following dimensions were required to be matched correctly: ^{13}C between HNCA and HN(CO)CA, ^{13}C between HN(CA)CO and HNCO, both ^1H and ^{15}N between HNCA and HN(CA)CO, and finally both ^1H and ^{15}N between HNCO and HN(CO)CA. The ^{13}C time dimensions of the HNCA and HN(CO)CA were doubled by linear prediction and zero filled to 220 and 256 points, respectively. After Fourier transformation 220 points were extracted between 42.5 – 69.7 ppm. The ^{13}C dimension of the HNCO was truncated from 53 points to 41 points to match that of the HN(CA)CO. Subsequently, both were doubled by linear prediction and zero filled to 256 points. The ^1H dimensions of the HNCA, HN(CO)CA, HNCO, and HN(CA)CO were zero filled to 1600, 1350, 1500, and 1500 points respectively in the time domain followed by extraction of 313 points (6.1–10 ppm) after Fourier transformation. The ^{15}N time domains of all four spectra were linear predicted to a final size of 64 complex points, and were zero filled to 100 points, followed by extraction of 89 points (103.28–134 ppm) after Fourier transformation. All dimensions were apodized with a cosine-squared bell function.

Covariance NMR

The calculation of a 4D-COSCOM has three distinct steps. The first entails the application of a derivative along the carbon dimension of Intra-3D and Seq-3D spectra. The next involves calculation of a 4D covariance spectrum from these 3D spectra. The final step combines 4D-COSCOMs from multiple pairs of Intra-3D and Seq-3D spectra into a single

correlation map. Below we describe the practical considerations and improvements at each step that make 4D-COSCOMs a robust tool for the assignment of protein backbone resonances.

A single 4D-COSCOM is calculated from two conventional 3D spectra (e.g. HNCA and HN(CO)CA) using the equation

$$HNH_s N_s(i, j, l, m) = \sum_{k=1}^K Intra3D'(i, j, k) Seq3D'(l, m, k). \quad (1)$$

where indices i and l represent 1H dimensions, j and m represent ^{15}N dimensions, and k represents a common ^{13}C dimension with a total of K points. The amplitude at position (i, j, l, m) in the 4D-COSCOM ($HNH_s N_s$) represents the extent to which a carbon slice from the Intra-3D spectrum at (H, N) position (i, j) correlates to a carbon slice from the Seq-3D spectrum at (H, N) position (l, m) . The amplitude of this covariance peak is dependent on the degree to which the two slices share signals at the same ^{13}C frequencies. If the two slices contain a signal at the exact same frequency, they will produce a strong covariance peak that is indicative of sequential connectivity. If they contain completely disparate signals, no covariance peak will result. However, if the two slices possess nearly degenerate signals, they may produce a covariance peak with reduced intensity depending on the amplitudes of the signals at the common frequency in each slice. Such peaks do not indicate sequential connectivity, yet these near degeneracy artifacts are not immediately identifiable as such. We have previously shown that taking the derivative of each spectrum along the subsumed carbon dimension prior to covariance (denoted by an apostrophe in equation 1) reduces or eliminates near degeneracy artifacts. This procedure was critical to the success of 3D-COSCOMs, which would otherwise be replete with spurious correlations and render them impractical. In what follows, we develop a framework to better understand how ^{13}C digital resolution influences the quality of 4D-COSCOMs, and we offer a rule of thumb for anticipating the frequency separation at which near degeneracy artifacts are completely eliminated.

Given an analytical expression for the ^{13}C line shape, we can derive an equation for the amplitude of near degeneracy artifacts as a function of frequency separation. For example, assuming a Lorentzian line shape, the amplitude of the covariance peak between two signals in Intra-3D and Seq-3D spectra can be calculated as follows. Let the function L represent a Lorentzian line shape parameterized by an amplitude A , a relaxation rate R , and a frequency offset f_0 :

$$L(A, R, f_0) = \frac{A \cdot R}{(R^2 + 4\pi^2(f - f_0)^2)} \quad (2)$$

Taking the derivative of L as a function of frequency f gives

$$\frac{dL(A, R, f_0)}{df} = \frac{-8\pi^2 f \cdot A \cdot R}{(R^2 + 4\pi^2(f - f_0)^2)^2} \quad (3)$$

Letting the labels I and S represent peaks from Intra-3D and Seq-3D spectra respectively, the continuous frequency analog of equation 1 with Lorentzian line shapes is

$$C = \int_{-\infty}^{\infty} \frac{dL(A_I, R_I, f_I)}{df} \cdot \frac{dL(A_S, R_S, f_S)}{df} df \quad (4)$$

where C denotes the amplitude of a correlation in the 4D COSCOM at a particular position and corresponds to $\text{HNH}_S\text{N}_S(i, j, l, m)$ in equation 1. Without loss of generality, we can set the frequency of the Intra-3D peak equal to zero and parameterize the Seq-3D peak in terms of its offset relative to the Intra-3D peak, Δf . With this simplification equation 4 reduces to

$$C = \int_{-\infty}^{\infty} \frac{dL(A_I, R_I, 0)}{df} \cdot \frac{dL(A_S, R_S, \Delta f)}{df} df \quad (5)$$

From this equation, the amplitude of the covariance peak is calculated as

$$C = \frac{4\pi^2 A_I A_S (R_I + R_S)((R_I + R_S)^2 - 12\pi^2 \Delta f^2)}{((R_I + R_S)^2 + 4\pi^2 \Delta f^2)^3} \quad (6)$$

Figure 1a plots C as a function of Δf . The signal is maximum when both peaks are found at the same ^{13}C frequency (i.e. $\Delta f = 0$), and the amplitude is

$$C = \frac{4\pi^2 A_I A_S}{(R_I + R_S)^3} \quad (7)$$

As Δf grows, the covariance peak shrinks until it becomes zero at

$$\Delta f_{\text{crit}} = \frac{1}{\sqrt{3}} \frac{(R_I + R_S)}{2\pi} = \frac{1}{\sqrt{3}} FWHM_{\text{av}} \quad (8)$$

where $FWHM_{\text{av}}$ is the average of the two peaks' full width at half max. Frequency differences less than Δf_{crit} result in positive covariance peaks, whereas differences greater than Δf_{crit} produce negative artifacts in the covariance spectrum. Once the signals are completely resolved, no spurious correlation is observed. Because negative artifacts can easily be identified and ignored, for example by displaying only positive contours during analysis, the derivative procedure effectively eliminates all artifacts that originate from degeneracies greater than or equal to Δf_{crit} . This fact is exploited in a novel procedure described below.

Incidentally, f_{crit} is also the frequency offset at which the left inflection point of one Lorentzian line shape aligns with the right inflection point of the other. Because this condition is easily identified by visual inspection (Figure 1b), we suggest this criterion as a rule of thumb for anticipating when two signals will no longer produce a positive peak in the covariance spectrum. Lorentzian line-shapes correspond to Fourier transforms of free induction decays (FIDs) that have fully relaxed. However, this is most often not the case in the indirect dimensions of 3D spectra, and FIDs must be apodized before Fourier transformation. For line shapes less heavy-tailed than the Lorentzian distribution, we anticipate that artifact elimination will occur at frequency differences less than those that are required to align the inflection points. By repeating the derivation described for Lorentzian line shapes with Gaussian line-shapes parameterized by amplitude A , line width parameter σ , and frequency offset f_0 ,

$$G(A, \sigma, f_0) = Ae^{-\frac{4\pi^2(f-f_0)^2}{2\sigma^2}} \quad (9)$$

one obtains

$$\Delta f_{\text{crit}} = \frac{1}{2\pi} \sqrt{\sigma_I^2 + \sigma_S^2} \quad (10)$$

The inflection points of Gaussian line shapes align at

$$\Delta f = \frac{\sigma_I + \sigma_S}{2\pi} \quad (11)$$

Thus, because $\sqrt{\sigma_I^2 + \sigma_S^2}$ is strictly less than $\sigma_I + \sigma_S$ for positive σ , the zero crossing of partial degeneracy artifacts must occur prior to the alignment of the inflection points.

The second step in generating 4D COSCOMS can be tailored to and optimized for four-dimensional arrays. Previously, 3D-COSCOMs were calculated by submitting individual $^1\text{H}/^{13}\text{C}$, Intra-3D and Seq-3D planes to the covariance NMR toolbox [10], while looping over the remaining ^{15}N dimensions. The covariance NMR toolbox uses singular value decomposition (SVD) as an optimized means [15] to calculate an arbitrary power of the covariance matrix. Matrix powers permit researchers to optimize the covariance matrix as a function of its parent spectra. For covariance matrices involving a single spectrum, it has been shown that the square-root of the covariance matrix is more closely related to its corresponding 2D Fourier transform than a non-rooted matrix [2], and SVD was introduced to optimize the square root calculation [15]. Indeed, we have found empirically that the quality of our maps improves when applying a matrix power. Our implementation allows users to adjust this parameter in a manner akin to traditional apodization of FIDs. The 4D maps presented in this work use a matrix power of $1/2$, as performed throughout the work of Rafael Brüschweiler and co-workers [16,17]. Unfortunately, calculating covariance matrices in a plane-by-plane fashion comes with a large cost in computational time, since $J \times M$ SVD calculations are required, where J and M are the number of points in the Intra-3D and

Seq-3D ^{15}N dimensions respectively. This renders rapid comparisons of maps obtained with different power arguments impractical. By re-implementing our calculation according to the principles outlined for higher dimensional covariance spectra [18], we now perform only one SVD per pair of Intra-3D and Seq-3D spectra while also applying the specified matrix power uniformly across all dimensions of the 4D-COSCOM. The updated procedure still calculates and stores the 4D-COSCOM one plane at a time to reduce its total memory footprint, but it now does so following a single, whole-spectrum SVD rather than between multiple, plane-wise SVDs. As a result, what was previously a 45-minute calculation [1] has now been reduced to only 60 seconds on identical hardware.

We have also improved the third and final step used in generating COSCOMs, which combines the sequential information obtained through different carbon nuclei into a single spectrum. When multiple pairs of Intra-3D and Seq-3D spectra are available, each sharing a different carbon nucleus (CA, CB, or CO), taking the element-wise product of their corresponding 4D-COSCOMs eliminates many of the spurious peaks found in any individual COSCOM [1]. For example, if the CA of residue i is degenerate with that of residue j , then the CA-based 4D-COSCOM will contain peaks at both $(H_i, N_i, H_{i+1}, N_{i+1})$ and $(H_i, N_i, H_{j+1}, N_{j+1})$. However, if the CO positions of residues i and j are distinct, then there will be no such $j+1$ correlation in the CO-based 4D-COSCOM. Thus in the combined spectrum, the positive peaks at position $(H_i, N_i, H_{i+1}, N_{i+1})$ will multiply constructively, whereas the lack of signal at position $(H_i, N_i, H_{j+1}, N_{j+1})$ in the CO-based COSCOM will tend to eliminate the artifact. Combined with the derivative-based artifact suppression technique, however, this procedure can have deleterious consequences. When partial frequency degeneracies occur between both the CA and CO of two residues, the CA-based and CO-based COSCOMs will each feature a negative peak at the same position. In our previous implementation, the two negative peaks would produce a positive peak after element-wise multiplication of the spectra. That is, the two strategies that separately remove artifacts can in combination reintroduce them. To overcome this effect, we now set all negative values equal to zero in each individual COSCOM prior to their multiplication, further eliminating spurious covariance artifacts when compared to our previous implementation.

Utility of 4D-COSCOMs

The availability of 4D-COSCOMs drastically simplifies the use of covariance correlation maps and enhances their utility when assigning the backbone of large proteins. In what follows we discuss the assignment of backbone resonances of three sequential residues from Cyl, a 52 kDa cyclization domain from the NRPS Yersiniabactin Synthetase. This stretch exemplifies the advantages of 4D-COSCOMs relative to their 3D counterparts and to traditional strip matching.

Figures 2a, b & d display three planes taken from a combined CA/CO 4D-COSCOM of Cyl. The planes display the sequential H_{i+1} and N_{i+1} dimensions of the 4D taken at three different (H_i, N_i) positions: A415, L416, and V417. The first plane, taken at the (H_i, N_i) position of A415, demonstrates the overall fitness of 4D-COSCOMs (Figure 2a). Here we see only two strong peaks: that of the truly sequential residue, L416, and the “auto”

correlation at A415. The latter peak occurs when sequential correlations are present in Intra-3D spectra. Such correlations could be eliminated by the use of intraresidual-HNCA and intraresidual-HN(CA)CO experiments [19–22], but in our hands the associated sensitivity losses have precluded their use. The second H_{i+1}/N_{i+1} plane features three strong signals (Figure 2b). The “auto” correlation at L416 and the true sequential correlation at V417 are present. However, there is also a strong peak at the position of K321. Examination of this correlation reveals that its predecessor CA_{i-1} and CO_{i-1} chemical shifts do indeed match the intra CA_i and CO_i signals of V416. Indeed, the correct successor to V416 could only be identified through careful comparison of the original 3D data sets along with cross-validation using NOESY spectra. This example highlights an advantage of COSCOMs relative to strip matching. Here, the two possibilities are realized directly and immediately, whereas when using strip-matching, the same conclusion could only be reached after the comparison of multiple strips. This situation is identical to that showcased when we presented 3D-COSCOMs. However, here, the resonances of V417 and K321 are immediately identified. With 3D-COSCOMs, we would have to investigate which combinations of 1H and ^{15}N chemical shifts exist in an HN-TROSY before reaching the same conclusion. The third and final H_{i+1}/N_{i+1} plane (Figure 2d), taken at the position of V417, contains an “auto” correlation as well as correlations to both E418 and Q430. This phenomenon is the result of 1H and ^{15}N overlap between residues V417 and N429 in the original 3D data. Indeed, Figure 2c contains a 2D $^1H/^{15}N$ projection of the Cy1 HNCO, and its inset demonstrates that V417 overlaps strongly with N429. Finally, the additional correlation to F442 seen in Figure 2d highlights a vulnerability of COSCOMs. This correlation is another consequence of sequential correlations in Intra-3D experiments. While the CO_{i-1} of F442 does indeed match the CO_i of V417, the CA_{i-1} of F442 matches the sequential CA_{i-1} peak of N429 rather than either of the intra CA_i peaks present in the slice. In a similar fashion, lingering sequential correlations in Intra-3D spectra do occasionally create erroneous peaks other than the expected “auto” correlations. Therefore we emphasize again that 4D-COSCOMs are intended to supplement rather than supplant traditional assignment procedures.

The final H_{i+1}/N_{i+1} plane shown in Figure 2d emphasizes the advantages of 4D-COSCOMs relative to their 3D counterparts. This plane contains four distinct peaks, yet the strips from the corresponding 3D projections, shown below and to the right of the plane (Figure 2d), contain only three resonances each. Because of E418’s simultaneous 1H degeneracy with V417 and ^{15}N degeneracy with Q430, its presence is masked in the 3D-COSCOMs. Ergo, one may implicitly assume that only three resonances are present, one of which corresponds to the auto-correlation. In this case, E418 would be completely neglected.

Figure 2e contrasts the traditional strip matching approach with that of the 4D-COSCOMs shown above. With strip matching, it is often the case that several candidate resonances must be investigated side-by-side while displaying many carbon dimensions before correctly identifying a residue’s successor. In the example we discussed two out of three assignments may have been hindered when using strip matching. First, if the CA_{i-1} signal of V417 had either been erroneously picked or not picked, L416 may have been linked to K321 instead of V417. Second, the CA of V417 may have been paired with the CA_{i-1} of N429 (and similarly for CO resonances) thereby producing an artificial sequence of residues. A careful

investigator would have had to probe all possible CA_i/CA_{i-1} and CO_i/CO_{i-1} combinations to investigate which of them lead to successors. The very nature of COSCOMs overcomes these issues effortlessly; all possible sequential candidates are displayed in an H/N plane. Only those sequential candidates that simultaneously match both CA_i and CO_i resonances are shown. There is no need to investigate multiple CA_i/CA_{i-1} and CO_i/CO_{i-1} combinations for V417 and N429; Q430 and E418 are simply seen. Nevertheless, COSCOMs are still vulnerable to erroneous correlations involving sequential signals in Intra-3D experiments, and 4D-COSCOMs must be inspected together with the original data. Overall, 4D-COSCOMs offer an intangible benefit in that they present data from multiple spectra in a more simple and intuitive way than strip matching, resulting in faster and more efficient assignment. Because of these advantages and because COSCOMs do not require any additional data beyond that which is already acquired for traditional assignment, we believe they provide a rather useful tool in facilitating, proof-reading, and editing NMR resonance assignments.

Conclusion

We have presented novel four-dimensional covariance correlation maps that harness the information of conventional 3D assignment spectra and combine them into a single 4D spectrum. This represents a major improvement to our previous approach [1], replacing the analysis of multiple 3D spectra with the simple inspection of a single 4D spectrum, as demonstrated with a large and challenging 52 kDa protein. We have derived a relationship between the frequency separation of signals in the subsumed carbon dimension and the amplitude of their corresponding near degeneracy artifacts. We presented simple rules of thumb to help users better relate the quality of input spectra with that of output covariance spectra. Finally, we introduced improvements to the calculation algorithm in terms of both execution efficiency and artifact elimination. Specifically, performing single-value-decomposition on the entire array rather than on each individual plane of the array accelerated the calculation 45-fold, and eliminating negative artifacts from individual COSCOMs before combining them prevents the reintroduction of spurious peaks. Overall, we believe 4D COSCOMs to be an important addition to the procedures employed during NMR resonance assignment that should help promulgate studies of systems with increased spectroscopic complexity.

Acknowledgments

We thank Rafael Brüschweiler for fruitful discussions on the quality of 4D-COSCOMs as a function of matrix power. We also thank Scott Nichols for his helpful comments during the preparation of this manuscript.

This work was supported by the National Institutes of Health (Grant No. RO1GM104257).

References

1. Harden B, Nichols S, Frueh D. Facilitated Assignment of Large Protein NMR Signals with Covariance Sequential Spectra Using Spectral Derivatives. *J Am Chem Soc.* 2014; 136:13106–13109. <http://pubs.acs.org/doi/abs/10.1021/ja5058407>. [PubMed: 25226241]
2. Brüschweiler R. Theory of covariance nuclear magnetic resonance spectroscopy. *J Chem Phys.* 2004; 121:409–414.10.1063/1.1755652 [PubMed: 15260561]

3. Brüschweiler R, Zhang F. Covariance nuclear magnetic resonance spectroscopy. *J Chem Phys.* 2004; 120:5253–5260.10.1063/1.1647054 [PubMed: 15267396]
4. Zhang F, Brüschweiler R. Indirect covariance NMR spectroscopy. *J Am Chem Soc.* 2004; 126:13180–13181.10.1021/ja047241h [PubMed: 15479045]
5. Benison G, Berkholz DS, Barbar E. Protein assignments without peak lists using higher-order spectra. *J Magn Reson.* 2007; 189:173–181.10.1016/j.jmr.2007.09.009 [PubMed: 17919953]
6. Snyder DA, Ghosh A, Zhang F, Szyperski T, Brüschweiler R. Z-matrix formalism for quantitative noise assessment of covariance nuclear magnetic resonance spectra. *J Chem Phys.* 2008; 129:1–9.10.1063/1.2975206
7. Snyder DA, Brüschweiler R. Generalized indirect covariance NMR formalism for establishment of multidimensional spin correlations. *J Phys Chem A.* 2009; 113:12898–12903.10.1021/jp9070168 [PubMed: 19810742]
8. Bodenhausen G, Ruben DJ. Natural abundance nitrogen-15 NMR by enhanced heteronuclear spectroscopy. *Chem Phys Lett.* 1980; 69:185–189.10.1016/0009-2614(80)80041-8
9. Pervushin K, Riek R, Wider G, Wüthrich K. Attenuated T2 relaxation by mutual cancellation of dipole-dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci U S A.* 1997; 94:12366–71. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=24947&tool=pmcentrez&rendertype=abstract>. [PubMed: 9356455]
10. Short T, Alzapiedi L, Brüschweiler R, Snyder D. A covariance NMR toolbox for MATLAB and OCTAVE. *J Magn Reson.* 2011; 209:75–78.10.1016/j.jmr.2010.11.018 [PubMed: 21215669]
11. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR.* 1995; 6:277–93. <http://www.ncbi.nlm.nih.gov/pubmed/8520220>. [PubMed: 8520220]
12. Mishra SH, Frueh D. Assignment of Methyl NMR Resonances of a 52 kDa Protein with Residue-specific 4D Correlation Maps. *J Biomol NMR.* 201510.1007/s10858-015-9943-6
13. Keating TA, Miller DA, Walsh CT. Expression, purification, and characterization of HMWP2, a 229 kDa, six domain protein subunit of Yersiniabactin synthetase. *Biochemistry.* 2000; 39:4729–39. <http://www.ncbi.nlm.nih.gov/pubmed/10769129>. [PubMed: 10769129]
14. Mishra S, Harden B, Frueh D. A 3D time-shared NOESY experiment designed to provide optimal resolution for accurate assignment of NMR distance restraints in large proteins. *J Biomol NMR.* 2014; 60:265–274.10.1007/s10858-014-9873-8 [PubMed: 25381567]
15. Trbovic N, Smirnov S, Zhang F, Brüschweiler R. Covariance NMR spectroscopy by singular value decomposition. *J Magn Reson.* 2004; 171:277–283.10.1016/j.jmr.2004.08.007 [PubMed: 15546754]
16. Snyder DA, Xu Y, Yang D, Brüschweiler R. Resolution-enhanced 4D ¹⁵N/¹³C NOESY protein NMR spectroscopy by application of the covariance transform. *J Am Chem Soc.* 2007; 129:14126–14127.10.1021/ja075533n [PubMed: 17973386]
17. Chen Y, Zhang F, Snyder D, Gan Z, Brüschweiler-Li L, Brüschweiler R. Quantitative covariance NMR by regularization. *J Biomol NMR.* 2007; 38:73–77.10.1007/s10858-007-9148-8 [PubMed: 17351818]
18. Snyder DA, Zhang F, Brüschweiler R. Covariance NMR in higher dimensions: Application to 4D NOESY spectroscopy of proteins. *J Biomol NMR.* 2007; 39:165–175.10.1007/s10858-007-9187-1 [PubMed: 17876709]
19. Nietlispach D, Ito Y, Laue ED. A novel approach for the sequential backbone assignment of larger proteins: Selective intra-HNCA and DQ-HNCA. *J Am Chem Soc.* 2002; 124:11199–11207.10.1021/ja025865m [PubMed: 12224968]
20. Permi, P. [accessed August 27, 2014] Intraresidual HNCA: An experiment for correlating only intraresidual backbone resonances; *J Biomol NMR.* 2002. p. 201-209.<http://link.springer.com/article/10.1023/A:1019819514298>
21. Brutscher B. Intraresidue HNCA and COHNCA experiments for protein backbone resonance assignment. *J Magn Reson.* 2002; 159:155–159.10.1006/jmre.2002.2546 [PubMed: 12081454]

22. Nietlispach D. A selective intra-HN (CA)CO experiment for the backbone assignment of deuterated proteins. *J Biomol NMR*. 2004; 28:131–136.10.1023/B:JNMR.0000013829.17620.39 [PubMed: 14755157]

Highlights

- Traditional NMR protein assignment relies on peak-peaking
- Peak picking has severe limitations
- Covariance Sequential Correlation Maps (COSCOMs) bypass peak-picking
- 4D-COSCOMs overcome the limitations of 3D-COSCOMs
- 4D-COSCOMs were used successfully to assign a 53 kDa protein

Figure 1.

Partial degeneracy artifacts (a) Covariance peak amplitude C as a function of frequency offset Δf . Here $A_I = A_S = 1$ and $R_I = R_S = 8 \text{ s}^{-1}$. Δf_{crit} marks the zero crossing of the covariance peak (b) An illustration of two Lorentzian line shapes with their inflection points in alignment. For Lorentzians, this condition is equivalent to a frequency separation of Δf_{crit}

Figure 2.

4D-COSCOMs of Cy1 (a) H_{i+1}/N_{i+1} plane from the 4D at the (H_i, N_i) position of A415 (b) H_{i+1}/N_{i+1} plane at L416 (c) $^1\text{H}/^{15}\text{N}$ projection of the HNCO. The inset shows a zoomed region around the overlapped residues V417 and N429. (d) H_{i+1}/N_{i+1} plane at V417. The green and blue strips below and to the right of the plane are taken along the H_{i+1} and N_{i+1} dimensions at the (H_i, N_i) coordinates of 3D-HNH_s and 3D-HNN_s, respectively. Both strips are taken at the (H_i, N_i) position of V417 (e) Traditional strip matching of Cy1. Intra-3D spectra are shown in red and Seq-3D spectra are shown in blue. CA spectra are shown to the left and CO spectra are shown to the right.