NMR Spectroscopy

DOI: 10.1002/anie.200604599

Robust Deconvolution of Complex Mixtures by Covariance TOCSY Spectroscopy**

*Fengli Zhang and Rafael Brüschweiler**

A fundamental problem in many areas of chemistry is the identification of components in chemical mixtures, such as different solutes in a solution. The recent advent of metabolomics has generated a critical demand for powerful analysis methods for fluid mixtures in the food and life sciences.[1–3] While important progress is being made in potentially laborious and costly hyphenated methods,[4] spectroscopic methods have the power to circumvent or reduce the need for hyphenation prior to analysis.[5]

Most compounds contain multiple NMR-active spins that are *J*-coupled and allow the identification of spin–spin coupling networks for discrimination between components, as well as their subsequent identification by screening against a database. Particularly useful in this regard is the 2D NMR ^{1}H–^{1}H TOCSY experiment,[6] which monitors multiple relay transfers of spin magnetization within a spin system to provide a wealth of information on scalar spin–spin coupling connectivity with high sensitivity. Because experimental efficiency is a prerequisite for high-throughput applications, TOCSY is combined here with covariance NMR,[7–9] which produces high-resolution spectra largely independent of the number of increments along the indirect time domain t_1.

A recently proposed unsupervised deconvolution method[10] uses principal-component analysis (PCA) of the covariance TOCSY spectrum of a mixture. In the absence of significant spectral overlap, the dominant PCA eigenmodes approximate well the 1D spectra of the individual components. For increasing degrees of spectral overlap between components, however, "mixed modes" emerge whose assignment to known compounds can pose a significant challenge.[10] The method presented here, which is termed DemixC (C stands for clustering), overcomes this limitation by identifying for each component characteristic traces that are essentially free of overlap and can be identified and assigned with high confidence.

The DemixC method is demonstrated for three samples of differing complexity. Sample I consists of three amino acids (Glu, Lys, Val) dissolved in D_2O. Sample II contains four amino acids (Glu, Leu, Lys, Val) in D_2O. The amino acid concentration of samples I and II is 7 mM. Sample III contains the cyclic decapeptide antamanide[11] [-Val-Pro-Pro-Ala-Phe-Phe-Pro-Pro-Phe-Phe-] dissolved in deuterated chloroform at a concentration of 1 mM. While the dissolved peptide of sample III is not an actual mixture, in terms of its proton NMR properties it behaves like a mixture of 10 amino acids at 1 mM concentration each. The low variability of the amino acid composition (four phenylalanine and four proline residues) leads to significant resonance overlap[12] providing a rigorous test case for the performance of the proposed method.

Two-dimensional TOCSY experiments for samples I and II were performed at 600 MHz with mixing times τ_m of 97 and 62 ms with 2048 complex points in t_2 and 1024 points in t_1 in TPPI mode. The TOCSY spectra of sample III were recorded at 800 MHz with mixing times τ_m of 97 and 76 ms with 2048 complex points in t_2 and 512 complex points in t_1 in TPPI-States mode. The TOCSY mixing sequence is MLEV-17[13] for all three mixtures. All NMR experiments were performed at 298 K.

Covariance processing was performed in the mixed-time frequency domain as described previously.[9] Briefly, for 2D TPPI TOCSY datasets the time-domain data are Fourier-transformed along the direct dimension t_2, phase- and baseline corrected, followed by elimination of the dispersive part, and then subjected to singular-value decomposition (SVD) to determine the matrix square-root of the covariance spectrum. For 2D TPPI-States TOCSY datasets, the cosine- and sine-modulated t_1 parts are first Fourier-transformed along t_2, followed by phase correction and elimination of the dispersive parts. The square-root of the covariance spectrum $\mathbf{C}$ is then computed by applying SVD individually to the cosine- and the sine-modulated parts before they are co-added. As is characteristic for covariance NMR, the resulting spectra are fully symmetric and display the same high spectral resolution along both dimensions. Examples of covariance TOCSY spectra of the three samples are shown in Figure 1.

Next, the similarity or overlap O_{ij} between each row vector $\mathbf{c}_i^T$ and column vector $\mathbf{c}_j$ of $\mathbf{C}$ is determined. The inner product between these vectors (traces) represents a suitable metric of similarity [Eq. (1)]

$$O_{ij} = \mathbf{c}_i^T \cdot \mathbf{c}_j \quad (1)$$

or, in matrix notation [Eq. (2)]

$$\mathbf{O} = \mathbf{C}^T \cdot \mathbf{C} \quad (2)$$

[*] Dr. F. Zhang, Prof. R. Brüschweiler
National High Magnetic Field Laboratory
Department of Chemistry and Biochemistry
Florida State University
Tallahassee, FL 32310 (USA)
Fax: (+1) 850-644-8281
E-mail: bruschweiler@magnet.fsu.edu
Homepage: http://spin.magnet.fsu.edu

[**] We thank Dr. Art S. Edison for stimulating discussions. This work was supported by the National Institutes of Health (grant R01 GM 066041).

Supporting information for this article is available on the WWW under http://www.angewandte.org or from the author.

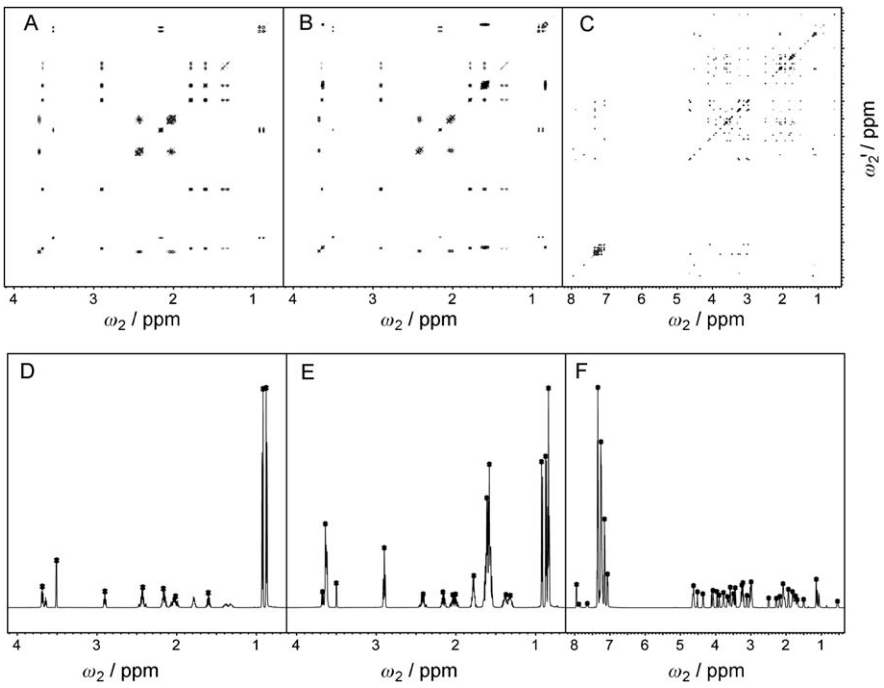

Figure 1. Covariance NMR TOCSY spectra of A) a mixture containing the three amino acids Glu, Lys, Val (sample I), B) a mixture containing the four amino acids Glu, Leu, Lys, Val (sample II), and C) the cyclic decapeptide antamanide (sample III). The importance-index vector **P**, as defined in Equation (3), is shown for sample I (D), sample II (E), and sample III (F). The amino acid mixtures contain amino acids at a concentration of 7 mM in D_2O buffer. The antamanide concentration is 1 mM in $CDCl_3$. The mixing times of the three TOCSY experiments were 97, 62, and 97 ms, respectively, with the MLEV-17 mixing sequence. The experiments were conducted at magnetic field strengths of 600 (samples I and II) and 800 MHz (sample III) at 298 K.

where **O** is the "overlap matrix" with elements O_{ij}. The larger O_{ij} is, the higher is the overlap and thereby the similarity of the covariance TOCSY traces represented by vectors $\mathbf{c}_i$ and $\mathbf{c}_j$. Because diagonal peaks tend to have disproportionately large amplitudes that dominate those of the inner products, prior to the overlap calculation each diagonal peak is replaced by a Gaussian peak with the amplitude of the largest nondiagonal peak in the same column or row. This leads to a modified overlap matrix **O′** for which the influence of the diagonal of the covariance spectrum is diminished.

Next, the elements of each column of **O′** are coadded to form a vector **P** with elements P_j termed importance index [Eq. (3)].

$$P_j = \sum_i O'_{ij} \qquad (3)$$

The importance index P_j is a quantitative measure for the cumulative overlap between the TOCSY trace (column or row) at frequency ω_j with all other traces (columns or rows). A large component P_j indicates that the covariance TOCSY column j has strong overlaps with other rows, whereas a low P_j value reflects little overlap. Overlaps stem from rows belonging to other spins of the same spin system, as well as rows of other spin systems whose resonances overlap with the resonances of row j. Vectors that belong to the same spin system have resonances at the same positions provided that the distribution of magnetization via isotropic mixing during the TOCSY experiment is sufficiently uniform among the spins.

In a next step, a subset of rows of interest is identified based on their importance index by applying standard peak picking to **P**. This involves the determination of local maxima above a given threshold. This threshold should be higher than the noise floor and can be adjusted to exclude weak traces that are not of interest. This yields a list of rows of the covariance TOCSY spectrum representing a small subset of all traces. The members of this list are then clustered on the basis of mutual overlaps of the normalized rows of **C**, $O'_{N,ij}$, to identify a unique set of spin systems and compounds. These are then displayed as the corresponding traces of the covariance matrix with the original diagonal peak scaled such that it is identical to the maximal off-diagonal peak in the same trace. The final set of magnitude traces represents the individual components that can be identified and assigned, for example, by screening against a spectral database.

The DemixC method is first demonstrated for sample I, which contains amino acids E, K, V. The covariance TOCSY spectrum is shown in Figure 1 A. The importance index vector **P** is constructed from **O′** followed by peak picking (Figure 1 D). In this way, nine cross sections (rows) in the covariance spectrum **C** are identified (peak positions marked by filled circles in Figure 1 D) and plotted in Figure 2 A. The mutual overlaps

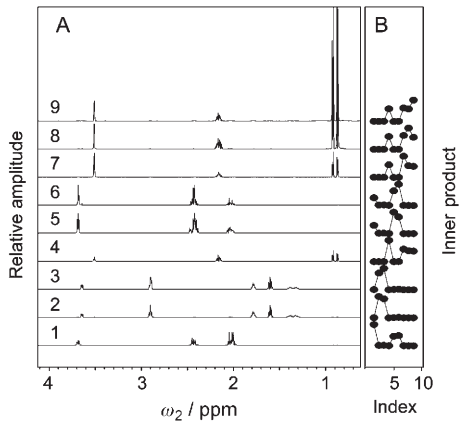

Figure 2. A) Nine traces of the covariance TOCSY spectrum of the three-amino-acid mixture (Figure 1 A) picked according to importance index of Figure 1 D. The traces are sorted according to the intensities in Figure 1 D, with trace 1 being the weakest. B) Normalized inner products among all nine traces of Panel A.

$O'_{N,ij}$ between the nine rows are shown in Figure 2B. The higher a $O'_{N,ij}$ value, the more similar are the corresponding rows i and j.

Basic clustering of the overlaps immediately reveals that rows 1, 5, and 6 represent the same compound (or spin system), rows 2 and 3 represent a second compound, and rows 4, 7, 8, and 9 represent a third compound. Because all nine rows can be assigned to one of the three clusters, it follows that the TOCSY spectrum of sample I contains no other detectable compound. The three clusters are represented by the trace spectra 1, 2, and 4.

In a next step, the compounds underlying the selected cross sections are identified by comparison with 1D spectra contained in an NMR databank. Here we chose the metabolomics/metabonomics part of the Biological Magnetic Resonance Data Bank (BMRB, http://www.bmrb.wisc.edu/metabolomics/), which is worldwide in the public domain. The proton 1D spectra contained in the BMRB for the three amino acids E, K, V are shown in Figure 3B and compared

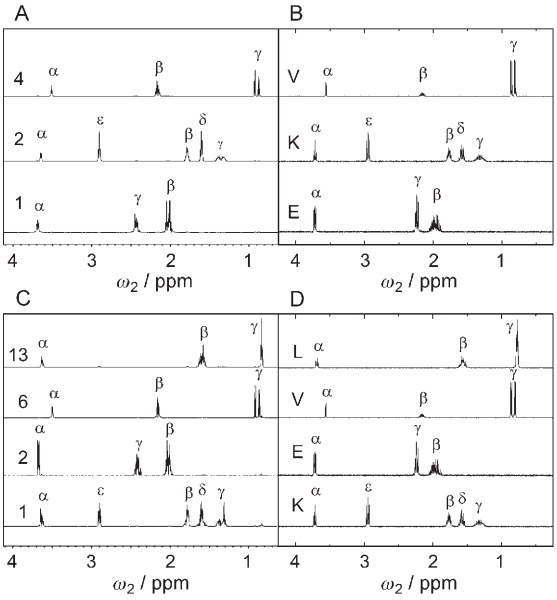

Figure 3. The results of the DemixC deconvolution method for the three-amino-acid mixture (A) in comparison with 1D NMR spectra of the individual amino acids taken from the BMRB data bank (B). The trace numbers correspond to the traces in Figure 2A. The results for the four-amino-acid mixture (C) in comparison with 1D NMR spectra of the individual amino acids taken from the BMRB data bank (D). The trace numbers correspond to the traces picked from the importance index vector of Figure 1E by using a cluster analysis analogous to Figure 2.

with their cluster representatives (rows 1, 2, and 4 of Figure 3A). The correspondence between the covariance traces and the BMRB spectra is very good. Even the peak multiplets show good agreement. Relative peak intensity differences stem from nonuniform TOCSY transfer, differential relaxation effects, and from the scaling of the diagonal part of the covariance TOCSY traces.

In the absence of overlaps between resonances belonging to different spin systems, as is the case for sample I, TOCSY traces for spins of the same spin system reflect the 1D spectrum of the spin system, and therefore they contain equivalent information. In the presence of overlap, the situation changes, as is seen for sample II, which contains Leu (L) as a fourth amino acid. For this mixture, significant peak overlap occurs, particularly between the Leu and Lys spin systems (Figure 1B). Still, the protocol produces clusters of traces which can be assigned to individual components (Figure 3C). Importantly, the representative trace for each cluster is chosen to have a minimal importance index (Figure 1E). This ensures selection of those traces that have low overlap with other spin systems. From Figure 3 it is evident that the selected traces (Panel C) agree well with the BMRB spectra of these components (Panel D).

Application of the algorithm to the cyclic decapeptide antamanide provides a stringent test of the deconvolution method. The ten amino acids lead to the rich covariance spectrum shown in Figure 1C, which exhibits substantial peak overlaps. Peak picking of the importance vector (Figure 1F) yields 33 trace vectors together with their mutual overlap matrix (see Supporting Information). Inspection of the traces reveals numerous regions with strong overlap. Cluster analysis yields the 11 representative traces depicted in Figure 4.

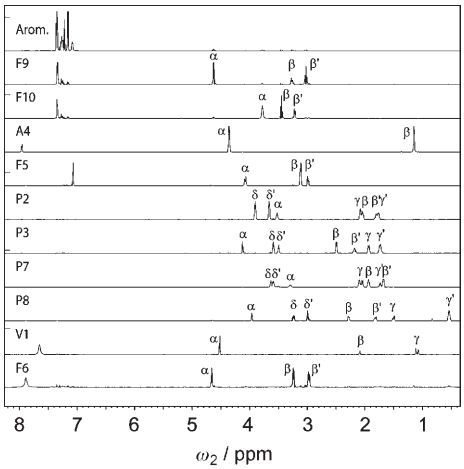

Figure 4. The result of the DemixC method for antamanide. The spin systems of the aliphatic protons of all ten amino acid residues are correctly identified by the method. In addition a trace is identified that corresponds to the overlapping aromatic ring protons of the phenylalanine residues (top trace). The H^N signals for V1, A4, F5, F6, F9, and F10 are between 7 and 8 ppm.

The bottom 10 traces correspond to the amide and aliphatic proton resonances of the 10 amino acids, while trace 11 (top trace in Figure 4) represents the strongly overlapping aromatic resonances of the phenylalanine rings. The amino acid traces of Figure 4 are fully consistent with the assignments of antamanide.[12] The traces of Ala and Val are as easily identified as the traces of samples I and II. The four Phe and four Pro residues show significant variability in their chemical shifts due to structural and dynamic differences.[14–18] The

residue that overlaps most severely is F9: its α, β, and β′ proton signals fully or partially overlap with those of F6, and its H^N proton signal overlaps with that of F10. Nonetheless, the DemixC protocol succeeds in finding a representative trace of this residue.

The analysis of mixtures by the DemixC method is based on abundant spin connectivities in total correlation spectroscopy and provides an efficient means for the spectral identification of spin systems and their compounds. The covariant nature of the spectrum ensures high resolution along both frequency dimensions, which is critical for the success of the method. Covariance TOCSY fundamentally differs from STOCSY[19,20] in that it uses covariances over 1D spectra with different t_1 evolution times of the same sample, whereas in STOCSY covariances are computed over 1D spectra of different samples. A previous method based on principal-component analysis (PCA)[10] requires a series of TOCSY spectra recorded with different mixing times. For the method presented here, this is not a requirement provided that the chosen mixing time is long enough to allow sufficient magnetization transfer throughout the whole spin system. For the mixtures used here, mixing times between 60 and 100 ms work well. Longer mixing times are feasible, although relaxation effects will lower the signal-to-noise ratio.

For compounds that contain multiple spin systems (i.e., spin systems that are disconnected from each other), as is the case for the individual amino acids of antamanide, each spin system yields an independent trace as if it belonged to an individual molecule. When identifying compounds in mixtures from the covariance TOCSY traces, this property of the TOCSY experiment must be taken into account.

The DemixC method readily identifies the best candidates for individual spin-system traces based on their importance index determined by the sum of the overlaps with all other candidate traces. Essential for the success of the method is the recognition that traces with a low to medium importance index are more likely to represent individual spin systems, whereas traces with a large importance index are more likely to be prone to overlap. By contrast, the PCA method tends to represent overlapping spin systems by some of the largest modes in case they explain together a larger fraction of the TOCSY spectrum.

Extreme resonance overlap imposes natural restrictions: if all resonances of a certain compound overlap with resonances of other systems, there is no guarantee that the deconvolution method will succeed in identifying the compound. The Phe-9 residue of antamanide represents such a case. Although the deconvolution procedure produces the correct result, generally the trace selection tends to become ambiguous when the number of overlaps of a component is very large.

In conclusion, the DemixC deconvolution approach introduced here takes full advantage of the high spectral resolution and redundant connectivity information of covariance TOCSY spectra. The trace analysis based on the importance index and subsequent clustering is highly efficient, remarkably robust, and provides individual 1D spectral information on the underlying spin systems. The method is directly applicable to the semi-automated side-chain assignment of peptides and small proteins. As small-molecule NMR databases are rapidly growing, such as the BMRB metabolomics databank, traces identified in covariance TOCSY spectra can be automatically screened against these databases to identify and quantify the TOCSY traces. This provides a path for the deconvolution of complex biological mixtures that is both efficient and reliable.

Received: November 10, 2006
Revised: December 14, 2006
Published online: March 5, 2007

Keywords: amino acids · analytical methods · correlation spectroscopy · NMR spectroscopy · peptides

[1] J. K. Nicholson, I. D. Wilson, *Nat. Rev. Drug Discovery* **2003**, *2*, 668.
[2] J. van der Greef, P. Stroobant, R. van der Heijden, *Curr. Opin. Chem. Biol.* **2004**, *8*, 559.
[3] A. T. Dossey, S. S. Walse, J. R. Rocca, A. S. Edison, *Chem. Biol.* **2006**, *1*, 511.
[4] S. P. Dixon, I. D. Pitfield, D. Perrett, *Biomed. Chromatogr.* **2006**, *20*, 508.
[5] S. Christophoridou, P. Dais, L. H. Tseng, M. Spraul, *J. Agric. Food Chem.* **2005**, *53*, 4667.
[6] L. Braunschweiler, R. R. Ernst, *J. Magn. Reson.* **1983**, *53*, 521.
[7] R. Brüschweiler, F. Zhang, *J. Chem. Phys.* **2004**, *120*, 5253.
[8] R. Brüschweiler, *J. Chem. Phys.* **2004**, *121*, 409.
[9] N. Trbovic, S. Smirnov, F. Zhang, R. Brüschweiler, *J. Magn. Reson.* **2004**, *171*, 277.
[10] F. Zhang, R. Brüschweiler, *ChemPhysChem* **2004**, *5*, 794.
[11] T. Wieland, H. Faulstich, *Crit. Rev. Biochem.* **1978**, *5*, 185.
[12] H. Kessler, A. Müller, K. H. Pook, *Liebigs Ann. Chem.* **1989**, 903.
[13] A. Bax, D. G. Davis, *J. Magn. Reson.* **1985**, *65*, 355.
[14] T. Bremi, R. Brüschweiler, R. R. Ernst, *J. Am. Chem. Soc.* **1997**, *119*, 4272.
[15] H. Kessler, C. Griesinger, J. Lautz, A. Müller, W. F. van Gunsteren, H. J. C. Berendsen, *J. Am. Chem. Soc.* **1988**, *110*, 3393.
[16] Z. L. Madi, C. Griesinger, R. R. Ernst, *J. Am. Chem. Soc.* **1990**, *112*, 2908.
[17] M. J. Blackledge, R. Brüschweiler, C. Griesinger, J. M. Schmidt, P. Xu, R. R. Ernst, *Biochemistry* **1993**, *32*, 10960.
[18] J. M. Schmidt, R. Brüschweiler, R. R. Ernst, R. L. Dunbrack, D. Joseph, M. Karplus, *J. Am. Chem. Soc.* **1993**, *115*, 8747.
[19] E. Holmes, O. Cloarec, J. K. Nicholson, *J. Proteome Res.* **2006**, *5*, 1313.
[20] O. Cloarec, M. E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes, J. Nicholson, *Anal. Chem.* **2005**, *77*, 1282.

