

Non-negative matrix factorization of two-dimensional NMR spectra: Application to complex mixture analysis

David A. Snyder,^{1,2} Fengli Zhang,² Steven L. Robinette,^{2,3} Lei Bruschiweiler-Li,^{1,2} and Rafael Brüschweiler^{1,2,a)}

¹Department of Chemistry and Biochemistry, Florida State University, Tallahassee, Florida 32306, USA

²National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32310, USA

³HHMI Science for Life Program, University of Florida, Gainesville, Florida 32610, USA

(Received 14 September 2007; accepted 29 October 2007; published online 1 February 2008)

A central problem in the emerging field of metabolomics is how to identify the compounds comprising a chemical mixture of biological origin. NMR spectroscopy can greatly assist in this identification process, by means of multi-dimensional correlation spectroscopy, particularly total correlation spectroscopy (TOCSY). This Communication demonstrates how non-negative matrix factorization (NMF) provides an efficient means of data reduction and clustering of TOCSY spectra for the identification of unique traces representing the NMR spectra of individual compounds. The method is applied to a metabolic mixture whose compounds could be unambiguously identified by peak matching of NMF components against the BMRB metabolomics database. © 2008 American Institute of Physics. [DOI: 10.1063/1.2816782]

The metabolism of even the simplest living organisms is an intricate network of chemical reactions, which biochemistry has only begun to unravel. Just as genomics has provided understanding of the genetic blueprints of life, the emerging science of metabolomics seeks to characterize the complex chemistry in biological systems through the differential analysis of metabolites.¹ NMR, which has a long history in chemistry as a tool for identifying compounds of interest, is emerging as a key technique in metabolomics.² For example, the two-dimensional (2D) total correlation spectroscopy (TOCSY) experiment,³ which monitors nuclear magnetization transfer across spin systems, provides spectroscopic information about a complex mixture that can greatly assist in the identification of its components without requiring hyphenation.^{4,5}

Identification of unique fingerprints of individual compounds in a mixture using a 2D ¹H-¹H TOCSY spectrum can be viewed as a classification or clustering problem. In the absence of spectral overlap and a suitable choice of the mixing time τ_m , each nonzero one-dimensional (1D) cross section (trace) of a TOCSY spectrum contains the resonances of the protons constituting a spin system. The recently introduced DemixC method⁵ is an example of a clustering approach for the elucidation of complex mixtures by TOCSY NMR. It classifies TOCSY traces by the degree to which they are likely to contain resonances from a single spin system only, thereby allowing for the extraction of traces that reflect individual compounds that can be identified in a subsequent analysis step, such as database screening.

2D Fourier transform TOCSY spectra display, with rare exceptions,⁶ positive cross peaks among spins that belong to

the same spin system, which makes them amenable to non-negative matrix factorization (NMF)^{7,8} providing a compact description of trace clustering. In the following, a 2D TOCSY spectrum is represented by a real $N_2 \times N_1$ matrix \mathbf{X} with N_2 points along the direct dimension ω_2 and N_1 points along the indirect dimension ω_1 . NMF decomposes matrix \mathbf{X} according to

$$\mathbf{X} = \mathbf{WH} + \varepsilon, \quad (1)$$

where \mathbf{W} and \mathbf{H} are (generally nonorthogonal) non-negative matrices of sizes $N_2 \times K$ and $K \times N_1$, respectively, and ε is the approximation error. If \mathbf{W} and \mathbf{H} are chosen in such a way that they minimize the residual sum of squares, expressed by $\|\varepsilon\|_F = \|\mathbf{X} - \mathbf{WH}\|_F$, where $\|\mathbf{A}\|_F = \sqrt{\text{trace}(\mathbf{A}^T \mathbf{A})}$ denotes the Frobenius or trace norm of a matrix, the columns of \mathbf{W} are centroids obtainable in a bipartite version of K -means clustering.⁷ K -means clustering⁹ is an efficient, iterative method for obtaining a set of clusters that minimizes the intracluster variance. In both NMF and K -means clustering, the number of components K must be provided as input (*vide infra*). Du *et al.*¹⁰ and Zhao *et al.*¹¹ have previously applied NMF to the analysis of 1D ¹H NMR of rat urine samples for the identification of spectral patterns of toxicity. Xu *et al.* have utilized NMF for the metabolic profiling of type II diabetes using 1D ¹H NMR of blood samples.¹² Matrix decomposition approaches other than NMF have been applied to 2D NOESY and 2D COSY spectra.¹³

The NMF decomposition of 2D TOCSY spectra reported here has the goal to reduce large 2D TOCSY spectra to their essential information, namely, the resonances of individual spin systems, represented by the cluster centroids or columns of \mathbf{W} . In fact, the NMF model represents the i th cross section of the TOCSY spectrum as a weighted sum of one or more components, where the i th column of matrix \mathbf{H} gives the

^{a)} Author to whom correspondence should be addressed. Tel: 850-644-1768. Fax: 850-644-8281. Electronic mail: bruschiweiler@magnet.fsu.edu.

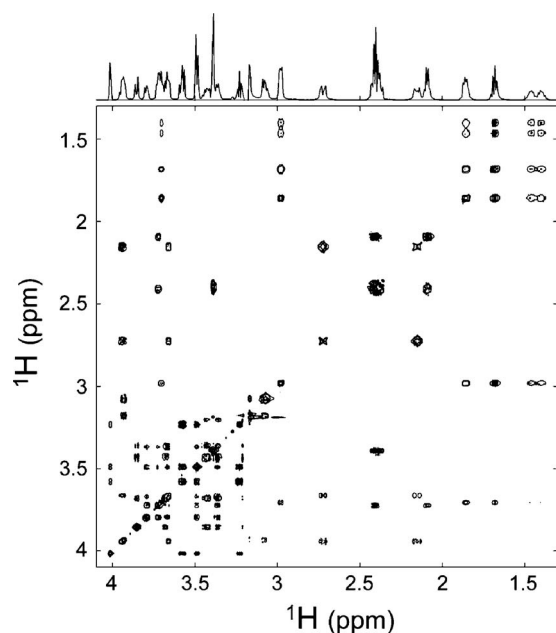


FIG. 1. Region of a 2D NMR TOCSY spectrum of the metabolic mixture subjected to non-negative matrix factorization (NMF). Above the 2D spectrum is the projection onto the direct dimension.

weighting for each component (columns of \mathbf{W}). Thus, the NMF model mirrors the mathematical structure of a TOCSY spectrum, where the slices of the spectrum correspond to the superposition of 1D spectra of one or more spin systems.

NMF is applied to the analysis of a TOCSY spectrum (Fig. 1) of a mixture of the seven common metabolites D-carnitine, D-glucose, L-glutamine, L-histidine, L-lysine, myo-inositol, and shikimic acid, at concentrations of 1 mM each, dissolved in D_2O . The spectrum was recorded at 298 K with a mixing time $\tau_m=90$ ms using the MLEV-17 mixing sequence¹⁴ with 2048 complex points in t_2 and 1024 complex points in t_1 . Prior to NMF analysis, the diagonal signals were rescaled to be no more intense than the most intense

off-diagonal signals in the corresponding cross section along ω_2 and a (minor) t_1 -noise artifact around 3.20 ppm was removed. Without rescaling, the diagonal peaks tend to dominate the NMF results and traces that belong to the same spin system are not necessarily recognized as such.

Our NMR metabolomics query server¹⁵ identifies traces resulting from NMF analysis by automated comparison with the BioMagResBank (BMRB) metabolomics database.¹⁶ For each trace, all local maxima with intensity greater than 1/8th of the global maximum were picked and the most intense local maximum within a 0.1 ppm range was defined as a unique peak. Local maxima isolated from other signals were also included in peak lists even if their intensity, relative to the global maximum for the trace, was less than 1/8th. NMF was performed using a MATLAB program available in the public domain.¹⁷ Using an initial guess for \mathbf{W} and \mathbf{H} with random elements in the interval $[0,1]$, NMF was iteratively performed to ensure that the lowest least-squares result $\|\epsilon\|_F$ is the global minimum.

The NMF results obtained from the TOCSY spectrum using $K=7$ are shown in Fig. 2. Screening the seven NMF components (columns of \mathbf{W}) against the BMRB (Ref. 16) using the NMR metabolomics query server¹⁵ correctly identifies the NMF components.

Several factors account for some observable differences between the NMF components (spectra on the left of Fig. 2) and the BMRB 1H reference spectra (on the right). Due to differences in water suppression and baseline correction of the 1D versus 2D data, resonances near the water line, e.g., the downfield anomeric proton lines of D-glucose, are present in the BMRB 1H spectrum [Fig. 2(M)] but absent in the corresponding NMF component [Fig. 2(F)]. Furthermore, since NMF preferentially clusters complete spin systems, resonances without coupling partners have lower weight and may not show up as NMF components. For example, the aromatic protons in L-histidine [Fig. 2(N)] have no coupling partners and thus do not appear as NMF components. Differ-

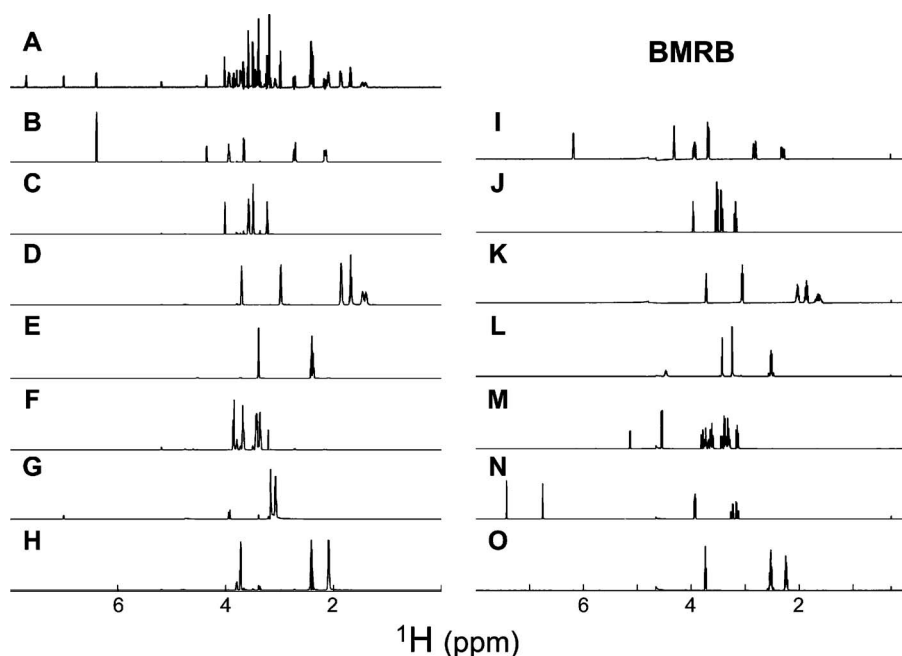


FIG. 2. Comparison of components obtained by NMF with 1D 1H NMR spectra of the compounds constituting a metabolic mixture. (A) 1D 1H spectrum of the metabolic mixture dissolved in D_2O . [(B)–(H)] All seven NMF components ($K=7$) of the TOCSY spectrum of the mixture of Fig. 1. [(I)–(O)] Individual reference 1D 1H spectra, taken from the BMRB (Ref. 16), of the metabolites in the mixture: (I) shikimic acid, (J) myo-inositol, (K) L-lysine, (L) D-carnitine, (M) D-glucose, (N) L-histidine, and (O) L-glutamine.

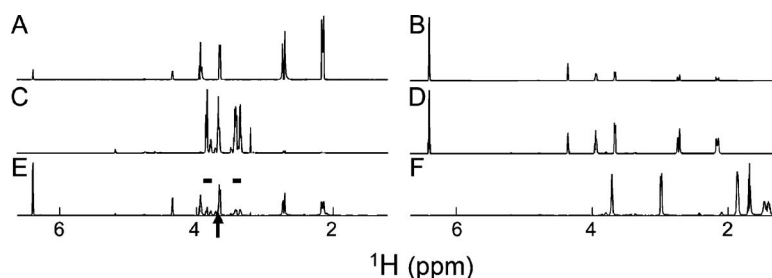


FIG. 3. Example traces obtained from NMF decomposition of the TOCSY spectrum of Fig. 1 using a variable number of components. [(A) and (B)] Pair of duplicate components with different relative peak intensities obtained with $K=20$ belonging to shikimic acid. [(C) and (D)] Using $K=7$, NMF yields for each compound of the mixture exactly one component, including (C) D-glucose and (D) shikimic acid (cf. Fig. 2). [(E) and (F)] Two NMF components with $K=5$. Component E mixes two partially overlapping spin systems, namely, D-glucose (indicated by the bars) and shikimic acid (the overlapping resonance is indicated by the arrow). Component F represents a single compound only (lysine).

ences in processing can also affect the behavior of NMF analysis with respect to resonances without coupling partners. This is the case for the methyl protons of D-carnitine [middle peak in Fig. 2(L)], where NMF fails to represent the methyl protons in a component: their diagonal peak in the TOCSY spectrum lies under the t_1 -noise ridge, and, lacking coupling partners, they have no cross peaks linking them to any other NMF component. Despite these differences, the database query is capable of correctly assigning the NMF component to each of the compounds comprising the mixture under investigation.

When analyzing a mixture one may know, at least approximately, the number of components (compounds) present. This number then provides a starting point for the number of components K used in NMF to decompose the TOCSY spectrum. Alternatively, it has been proposed to estimate K by principal component analysis (PCA).¹⁰ Due to the orthogonality property of principal components, they are not well suited to separate overlapping spin systems into modes that can be uniquely attributed to the different spin systems.⁴ However, PCA still gives an estimate of the minimal number of modes required to explain the dominant features of the spectrum. For the present TOCSY spectrum, the singular values reach a plateau at $K=13$ providing an upper bound for the range of K .

In addition, it is useful to vary K and monitor its effect on the behavior of the components (Fig. 3). Ideally, K is chosen to correspond to the effective number of spin systems K_{spin} in the mixture ($K_{\text{spin}}=7$ in the present case; see above). If $K < K_{\text{spin}}$, it is found that spin systems that (partially) overlap in the spectrum are represented by a single component. If $K > K_{\text{spin}}$, on the other hand, duplicate components occur with closely related peak patterns that represent the same spin system. Figure 3 illustrates these effects by comparing representative NMF components for $K=5, 7$, and 20 .

So far, TOCSY traces along ω_2 were interpreted in terms of superpositions of the columns of \mathbf{W} , which has the advantage that NMF components have the same high spectral resolution as the direct TOCSY dimension ω_2 . Transposition of Eq. (1) yields $\mathbf{X}^T = \mathbf{H}^T \mathbf{W}^T + \varepsilon^T$, which means that one can interpret the rows of \mathbf{H} as centroids⁷ for the reconstruction of the rows of \mathbf{X} along the indirect dimension ω_1 . Due to the near symmetry of a homonuclear TOCSY spectrum, both types of NMF components should yield comparable results

and thereby provide a cross validation for the identification of spin systems. When such an analysis is carried out for either $K=5$ or $K=7$, each row of \mathbf{H} corresponds to a column of \mathbf{W} . However, when for example $K=20$, five rows of \mathbf{H} cannot be assigned to columns of \mathbf{W} . The lack of cross validation by the rows of \mathbf{H} for the spin systems obtained as columns of \mathbf{W} indicates that the actual number of components K_{spin} is significantly smaller than 20 , which is consistent with the other K estimators and our knowledge of the actual composition of the mixture.

In summary, NMF provides a compact matrix-based formalism for a reduced representation of multidimensional NMR spectra. In the case of TOCSY, the spin-connectivity information is clustered yielding component centroids that are related to 1D spectra of individual spin systems. The centroids allow identification of the compounds in the mixture under investigation by screening with such tools as the NMR peaks query at the BMRB website,¹⁶ the human metabolome database,¹⁸ and the NMR metabolomics query server.¹⁵ Interpretation of 2D spectra other than TOCSY, such as NOESY, absolute value COSY, and heteronuclear correlation spectra may also benefit from this approach. The NMF method has the potential to be a useful tool for data reduction and the (semi-)automated analysis of multidimensional spectra obtained both by liquid and solid-state NMR. It contributes to the arsenal of methods applicable to metabolomics studies, in which NMR together with other analytical techniques plays an essential role to elucidate the complex chemical composition of life.

This work was supported by NIH Grant No. GM 066041.

¹O. Fiehn, *Plant Mol. Biol.* **48**, 155 (2002); R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, *Trends Biotechnol.* **22**, 245 (2004); J. K. Nicholson and I. D. Wilson, *Nat. Rev. Drug Discovery* **2**, 668 (2003).

²E. M. Lenz and I. D. Wilson, *Journal of Proteome Research* **6**, 443 (2007).

³L. Braunschweiler and R. R. Ernst, *J. Magn. Reson.* **53**, 521 (1983).

⁴F. Zhang and R. Brüscheiler, *ChemPhysChem* **5**, 794 (2004).

⁵F. Zhang, A. T. Dossey, C. Zachariah, A. S. Edison, and R. Brüscheiler, *Anal. Chem.* **79**, 7748 (2007); F. L. Zhang and R. Brüscheiler, *Angew. Chem., Int. Ed.* **46**, 2639 (2007).

⁶M. Rance, *Chem. Phys. Lett.* **154**, 242 (1989).

⁷C. Ding, X. He, and H. Simon, *Proceedings of the SIAM International Conference on Data Mining*, 2005 (unpublished).

- ⁸D. D. Lee and H. S. Seung, *Nature (London)* **401**, 788 (1999).
- ⁹J. B. MacQueen, *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1967 (unpublished).
- ¹⁰S. Du, P. Sajda, R. Stoyanova, and T. Brown, *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **5**, 4731 (2005).
- ¹¹Q. Zhao, R. Stoyanova, S. Du, P. Sajda, and T. R. Brown, *Bioinformatics* **22**, 2562 (2006).
- ¹²L. Xu, J. Dong, Z. Chen, and X. Dai, *First International Conference on Bioinformatics and Biomedical Engineering*, 2007 (unpublished).
- ¹³T. F. Havel, I. Najfeld, and J. X. Yang, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 7962 (1994).
- ¹⁴A. Bax and D. G. Davis, *J. Magn. Reson.* **65**, 355 (1985).
- ¹⁵S. Robinette, F. Zhang, R. Brüschweiler, <http://spinportal.magnet.fsu.edu>.
- ¹⁶B. R. Seavey, E. A. Farr, W. M. Westler, and J. L. Markley, *J. Biomol. NMR* **1**, 217 (1991).
- ¹⁷C.-J. Lin, *Neural Comput.* **19**, 2756 (2007).
- ¹⁸D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M. A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser, *Nucleic Acids Res.* **35**, D521 (2007).