NMR Spectroscopy

International Edition: DOI: 10.1002/anie.201806144
German Edition: DOI: 10.1002/ange.201806144

XLSY: Extra-Large NMR Spectroscopy

*Yulia Pustovalova, Maxim Mayzel, and Vladislav Yu. Orekhov**

Abstract: *NMR studies of intrinsically disordered proteins and other complex biomolecular systems require spectra with the highest resolution and dimensionality. An efficient approach, extra-large NMR spectroscopy, is presented for experimental data collection, reconstruction, and handling of very large NMR spectra by a combination of the radial and non-uniform sampling, a new processing algorithm, and rigorous statistical validation. We demonstrate the first high-quality reconstruction of a full seven-dimensional HNCOCACONH and two five-dimensional HACACONH and HN(CA)CONH experiments for a representative intrinsically disordered protein α-synuclein. XLSY will significantly enhance the NMR toolbox in challenging biomolecular studies.*

With the ever-increasing sizes and complexity of biomolecular systems studied by NMR spectroscopy, the number of peaks and hence signal overlap increases which seriously complicates and compromises the data analysis. The problem is addressed by enhancing spectral resolution and dimensionality with the radial (RS) and non-uniform sampling (NUS).[1] However, the task of reconstructing and handling of very large spectra is still awaiting a good solution.

The RS approach, which is based on direct analysis of planar spectral projections, is an efficient way of signal detection in high-dimensional experiments.[2] However, a set of planar projections is not a fair substitute for a true multidimensional experiment, especially in case of a highly crowded spectrum of a challenging protein system such as an intrinsically disordered protein (IDP).

The best algorithms for reconstructing spectra from NUS[3] data are impractical for large data sets owing to unbearable computational and storage requirements. Existing methods for spectra with five dimensions require a three-dimensional reference spectrum or peak list[4] whereas six- and seven-dimensional spectra are produced only as their reduced dimensionality projections.[4b] A possible solution for the large data sets may be found in the family of parametric algorithms,[5] although no examples of spectra reconstructions with more than four dimensions have been presented so far.

Herein we introduce XLSY–NMR spectroscopy for extra-large datasets. When dealing with a large spectrum, the main problem stems from its size that requires huge amounts of computational power for processing and by far exceeds computer memory. Notably, a multidimensional NMR spectrum is sparse and thus can be presented in a compact form both in time and frequency domains. XLSY is a non-iterative procedure that converts a small number of RS/NUS measurements into a compact, high-quality sparse spectrum without ever dealing with the huge full data representation in either the time or frequency domains.

The XLSY algorithm for spectrum reconstruction consists of three steps: 1) frequency identification, 2) intensity evaluation, and 3) validation.

The frequency identification borrows part of the SFFT algorithm[6] for finding a short list of frequencies in the spectrum that may have significant (that is, higher than noise) intensities. This part is based on the radial sampling and the Fourier projection theorem.[7] For an illustration of the algorithm, let us consider the simplest case of only two spectral dimensions each spanning N points. The two-dimensional spectrum contains $N \times N$ points with frequency coordinates (f_1,f_2). In an experiment, we measure a one-dimensional projection that contains N frequency points enumerated by index f. To distinguish frequency points in a multidimensional spectrum (f_1,f_2) and in a 1D projection, we call the later buckets. Value of each bucket with position f is given by the sum of N points of the 2D, which positions (f_1,f_2) fulfil the relation [Eq. (1a)]:

$$\alpha_1 f_1 + \alpha_2 f_2 = f \bmod N \tag{1a}$$

For a spectrum with many dimensions, the corresponding general relation is [Eq. (1b)]:

$$\sum \alpha_i f_i = f \bmod N \tag{1b}$$

where mod is the modulo operator, α_1 and α_2 are integers, and α_1/α_2 represents a slope of the projection. Since the spectrum is sparse, most buckets contain only noise and are considered empty. A few buckets, the intensities of which exceed a chosen noise threshold, correspond to one or a few non-zero frequencies points (f_1,f_2). To find the exact position of these essential frequencies in the 2D plane, we measure and analyze several projections with different tilt angles. Figure 1 illustrates the cumulative analysis of several projections from a 2D spectrum with only 3 non-zero points. After occupied buckets of the first projection are identified, all N frequencies that contributed to those buckets get a vote. This procedure is

[*] Dr. Y. Pustovalova, Prof. Dr. V. Y. Orekhov
Department of Chemistry and Molecular Biology
University of Gothenburg
P.O. Box 465, Gothenburg 405 30 (Sweden)
E-mail: vladislav.orekhov@nmr.gu.se
Dr. M. Mayzel, Prof. Dr. V. Y. Orekhov
Swedish NMR Centre, University of Gothenburg
P.O. Box 465, Gothenburg 405 30 (Sweden)

Supporting information and the ORCID identification number(s) for the author(s) of this article can be found under: https://doi.org/10.1002/anie.201806144.

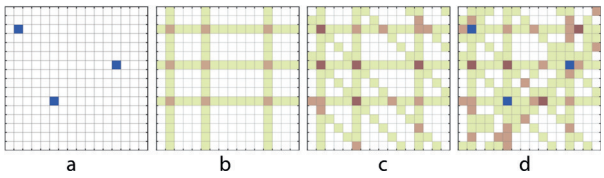

Figure 1. An illustration of the voting procedure using Eq. (1): a) positions of three signals, b) voting with two orthogonal projections, c) addition of the first diagonal projection, and d) voting with two orthogonal and two diagonal projections. Pixel colour in (b)–(d) from light to dark indicates the number of votes from one to four. See the text for more explanations.

repeated for different projections until a consistent short list of essential frequencies is discriminated by maximum number of votes. Points with frequencies that accumulated none or too few votes are considered to have exactly zero intensities in the spectrum and are omitted from further consideration and storage. To ensure picking up of low-intensity peaks, the threshold in the projections can be as low as 2σ noise and the voting cut-off is set on the level of 70–80% of the maximum defined by the number of used projections. For example, in Figure 1d, three correct frequencies collect a maximal possible number of four votes each (blue). There are also four points that collect 3 votes (the darkest brown), which are artefacts of the radial sampling. These points may be kept in the short list of frequencies and will be eliminated at the final validation step.

Evaluation of intensities for the frequencies shortlisted at the identification step is performed by solving a system of linear equations [Eq. (2a)]:

$$\boldsymbol{A}\boldsymbol{s} = \boldsymbol{t} \tag{2a}$$

where vector $\boldsymbol{s}$ consists of N_f unknown spectral intensities. Vector $\boldsymbol{t}$ is composed of N_t experimental complex time-domain data points. $\boldsymbol{A}$ is a $N_t \times N_f$ complex matrix obtained from the matrix of discrete inverse d-dimensional Fourier transform by retaining only columns and rows corresponding to the shortlisted frequencies and available experimental points, respectively. Matrix elements $A_{n,k}$ are calculated as [Eq. (2a)]

$$A_{n,k} = \frac{1}{N^d} e^{2\pi i \hat{n}\cdot\hat{k}/N} \tag{2b}$$

where d is the number of indirect dimensions spanning N points each (in our case the same for all dimensions); $\hat{k}$ is a d-dimensional vector of coordinates of the k-th point in the frequency domain corresponding to intensity s_k; and $\hat{n}$ is a d-dimensional vector of coordinates of the n-th point in the time domain corresponding to the measured value t_n.

To obtain a unique and reliable solution, matrix $\boldsymbol{A}$ in the system in Equation (2a) must be skinny, that is, number of unknown spectral intensities N_f must be lower than the number of linear equations N_t. Besides, to obtain well-conditioned matrix $\boldsymbol{A}$, it is essential to augment the RS data by NUS measurements. The possibility to use NUS along with RS data is a key feature of the new method that for the first time allowed ambiguities to be resolved and spurious aliasing peaks that are inherent in RS data to be avoided.[1b,3c] To further stabilize the solution of the linear system in Equation (2a) for the most crowded spectral regions, we use a mild Tikhonov regularization.[8]

At the final validation step, which is also unique for the XLSY algorithm, the bootstrap approach[9] is used to estimate individual uncertainties for every calculated point in the spectrum. This defines a local noise level that may vary significantly from one spectral region to another depending on the signal density. The local noise estimate also sets an upper intensity boundary for the weak peaks that might be lost in the spectrum reconstruction. Furthermore, the uncertainties play an important role in the XLSY algorithm for detecting true weak peaks and discarding a sizable fraction (20–80%) of points in the frequency shortlist that originate from the radial sampling artefacts. Since the NUS data are used for solving Equation (2), these frequencies are easily spotted as having statistically insignificant intensities.

Additional comment should be made about sensitivity of the XLSY method. Although at the evaluation and validations steps of the algorithm, all the RS and NUS experimental data can be used together, the identification step implies signal thresholding in the individual projections, which are recorded in a fraction of total experiment time. Similar to the APSY approach,[2d,e] the corresponding loss of sensitivity is largely offset by the combined analysis of the projections in the voting procedure. For our spectra, the signal threshold in the projections was set to $2–3\sigma$ noise level, which is well below the threshold commonly used for peak detection in individual NMR spectra. Note that 5% (1%) of noise intensities exceed the 2 (3)σ noise threshold just by chance.

We demonstrate the XLSY by the first high-quality reconstruction of three very large spectra consisting of 10^{10}–10^{15} points. Figure 2 illustrates XLSY reconstructions of 7D HNCOCACONH,[10] 5D HACACONH,[11] and 5D HN(CA)CONH[12] spectra for a representative 14 kDa IDP α-synuclein (Table 1). Figure 2b–d highlights the genuine resolution of the 7D spectrum. Peaks of E105 and E131 from EEG sequence repeats, which are resolved in the 7D HNCOCACONH spectrum, are fully overlapped in 5D spectra (insets of Figure 2e) and cannot be resolved in any radially sampled planar projections of the spectrum.

The XLSY spectra can be easily handled and analyzed. They have compact sparse-matrix representation and contain only statistically validated intensities. For visualization and detailed analysis, any spectral slice or projection can be obtained including those that are very difficult or impossible to obtain in experiments with lower dimensions. Figure 2a shows example of a unique orthogonal projection C^{O}_{i-1}/N_{i+1} of the 7D HNCOCACONH spectrum. Figures 2e,f illustrate quality of the 5D HACACONH and 5D HN(CA)CONH spectra with two planar projections N_i/C^{α}_{i-1} and N_i/N_{i-1}. Figure 2f shows a partial sequential assignment walk performed in the 5D HN(CA)CONH spectrum for the stretch of amino acids from A69 to T64.

In the multidimensional spectra, peaks are well-resolved and semi-automatic signal detection is straightforward. In the 5D HACACONH and 5D HN(CA)CONH spectra, we found all peaks expected for α-synuclein with exception of five

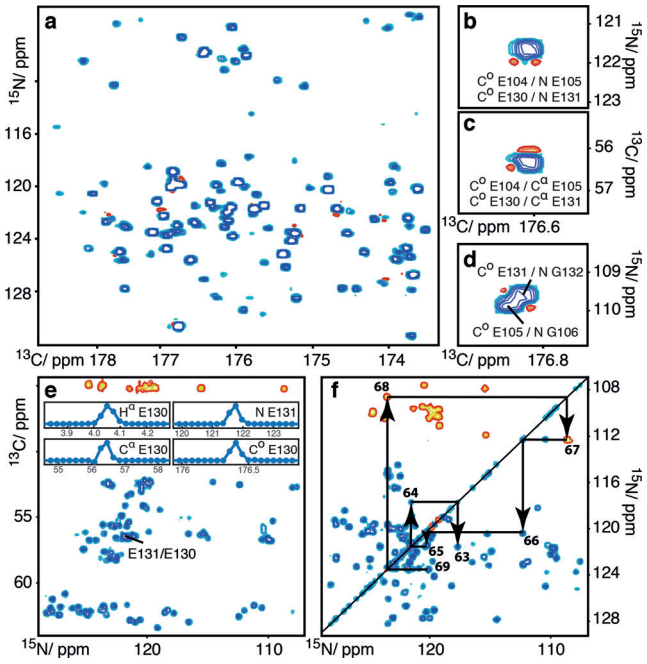

Figure 2. XLSY reconstructions of α-synuclein spectra. a) C^{O}_{i-1}/N_{i+1} projection of 7D HNCOCACONH spectrum. b)–d) Slices of the 7D spectrum through peaks for E105 and E131. e) N_i/C^{α}_{i-1} projection of 5D HACACONH spectrum. Insets show 1D cross-sections taken through the cross peak E131/E130 (overlapped with E105/E104). f) An example of a sequential assignment walk in the N_{i+1}/N_i projection from 5D HN(CA)CONH spectrum.

Table 1: Parameters of NMR experiments and XLSY reconstructions.

	5D	7D
Experimental time RS/NUS/total, hours	23.5/3.5/27	49/68/117
Time-domain projections	19	14
Number of NUS points	75	310
Number of shortlisted frequencies	8×10^4/ 6×10^4 [a]	3×10^6
Final size of the reconstruction after validation, pts	3.9×10^4/ 3.5×10^4 [a]	4.6×10^4

[a] 5D HACACONH/ 5D HN(CA)CONH spectra, respectively.

prolines and two residues at the N-terminus. In the 7D HNCOCACONH we found all peaks that were present in the orthogonal projections of the experiment. The peak lists together with the corresponding backbone assignment of α-synuclein are deposited in BMRB (27586). The assignment is in line with the published assignment (BMRB No. 6968), which was obtained at different sample temperature.

In conclusion, by demonstrating the first high quality reconstruction of complete 7D and 5D spectra of a representative IDP we introduce the XLSY method that removes the limits on spectrum dimensionality and resolution imposed by the existing signal acquisition and processing approaches. We envisage that the method will be most useful in studies of IDPs and in automatized high-throughput characterization of small and medium size globular protein systems, where experiments with high dimensionality and resolution are in the highest demand.[13]

Acknowledgements

The work was supported by the Swedish Research Council (Research Grant 2015–04614); the Swedish NMR Centre is acknowledged for spectrometer time.

Conflict of interest

The authors declare no conflict of interest.

Keywords: intrinsically disordered protein · NMR spectroscopy · non-uniform sampling · XLSY

How to cite: *Angew. Chem. Int. Ed.* **2018**, *57*, 14043–14045
Angew. Chem. **2018**, *130*, 14239–14241

[1] a) J. C. Hoch, *J. Magn. Reson.* **2017**, *283*, 117–123; b) K. Kazimierczuk, V. Orekhov, *Magn. Reson. Chem.* **2015**, *53*, 921–926.
[2] a) B. E. Coggins, R. A. Venters, P. Zhou, *Prog. Nucl. Magn. Reson. Spectrosc.* **2010**, *57*, 381–419; b) H. R. Eghbalnia, A. Bahrami, M. Tonelli, K. Hallenga, J. L. Markley, *J. Am. Chem. Soc.* **2005**, *127*, 12528–12536; c) R. Freeman, E. Kupce, *Conc. Magn. Reson.* **2004**, *23A*, 63–75; d) S. Hiller, F. Fiorito, K. Wüthrich, G. Wider, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 10876–10881; e) R. L. Narayanan, U. H. Dürr, S. Bibow, J. Biernat, E. Mandelkow, M. Zweckstetter, *J. Am. Chem. Soc.* **2010**, *132*, 11906–11907.
[3] a) S. G. Hyberts, A. G. Milbradt, A. B. Wagner, H. Arthanari, G. Wagner, *J. Biomol. NMR* **2012**, *52*, 315–327; b) K. Kazimierczuk, V. Y. Orekhov, *Angew. Chem. Int. Ed.* **2011**, *50*, 5556–5559; *Angew. Chem.* **2011**, *123*, 5670–5673; c) M. Mobli, A. S. Stern, J. C. Hoch, *J. Magn. Reson.* **2006**, *182*, 96–105; d) X. B. Qu, M. Mayzel, J. F. Cai, Z. Chen, V. Orekhov, *Angew. Chem. Int. Ed.* **2015**, *54*, 852–854; *Angew. Chem.* **2015**, *127*, 866–868; e) S. J. Sun, M. Gill, Y. F. Li, M. Huang, R. A. Byrd, *J. Biomol. NMR* **2015**, *62*, 105–117; f) J. Ying, F. Delaglio, D. A. Torchia, A. Bax, *J. Biomol. NMR* **2016**, 1–18.
[4] a) K. Kosiński, J. Stanek, M. J. Górka, S. Żerko, W. Koźmiński, *J. Biomol. NMR* **2017**, *68*, 129–138; b) S. Żerko, W. Koźmiński, *J. Biomol. NMR* **2015**, *63*, 283–290.
[5] a) A. L. Hansen, D. Li, C. Wang, R. Bruschweiler, *Angew. Chem. Int. Ed.* **2017**, *56*, 8149–8152; *Angew. Chem.* **2017**, *129*, 8261–8264; b) V. Jaravine, I. Ibraghimov, V. Y. Orekhov, *Nat. Methods* **2006**, *3*, 605–607.
[6] H. Hassanieh, M. Mayzel, L. Shi, D. Katabi, V. Y. Orekhov, *J. Biomol. NMR* **2015**, *63*, 9–19.
[7] R. N. Bracewell, *Aust. J. Phys.* **1956**, *9*, 198–217.
[8] A. N. Tikhonov, A. A. Samarskii, *Equations of Mathematical Phisics*, Dover Publications, New York, **2011**.
[9] B. T. Efron, *An Introduction to the Bootstrap.*, Chapman & Hall, London, **1993**.
[10] F. Fiorito, S. Hiller, G. Wider, K. Wuthrich, *J. Biomol. NMR* **2006**, *35*, 27–37.
[11] S. Hiller, G. Wider, K. Wüthrich, *J. Biomol. NMR* **2008**, *42*, 179–195.
[12] V. Motáčková, J. Nováček, A. Zawadzka-Kazimierczuk, K. Kazimierczuk, L. Žídek, H. Šanderová, L. Krásný, W. Koźmiński, V. Sklenář, *J. Biomol. NMR* **2010**, *48*, 169–177.
[13] The XLSY matlab scripts for sampling generation and spectra processing as well as guidelines and a test data are available upon request from the authors.

Manuscript received: May 28, 2018
Revised manuscript received: August 24, 2018
Accepted manuscript online: September 2, 2018
Version of record online: October 1, 2018

