Real-time Acquisition of Three Dimensional NMR Spectra by Non-uniform Sampling and Maximum Entropy Processing

Jun-Goo Jee

Center for Priority Areas, Tokyo Metropolitan University, Tokyo 192-0397, Japan. E-mail: jee-jungoo@tmu.ac.jp
Department of Molecular Biology, Nara Institute of Science and Technology, Nara 630-0101, Japan
Received July 16, 2008

Of the experiments to shorten NMR measuring time by sparse sampling, non-uniform sampling (NUS) is advantageous. NUS miminizes systematic errors which arise due to the lack of samplings by randomization. In this study, I report the real-time acquisition of 3D NMR data using NUS and maximum-entropy (MaxEnt) data processing. The real-time acquisition combined with NUS can reduce NMR measuring time much more. Compared with multidimensional decomposition (MDD) method, which was originally suggested by Jaravine and Orekhov (JACS 2006, 13421-13426), MaxEnt is faster at least several times and more suitable for the real-time acquisition. The designed sampling schedule of current study makes all the spectra during acquisition have the comparable resulting resolutions by MaxEnt. Therefore, one can judge the quality of spectra easily by examining the intensities of peaks. I report two cases of 3D experiments as examples with the simulated sub-dataset from experimental data. In both cases, the spectra having good qualities for data analysis could be obtained only with 3% of original data. Its corresponding NMR measuring time was 8 minutes for 3D HNCO of ubiquitin.

Key Words : Non-uniform sampling, Maximum entropy processing, NMR, Real-time acquisition

Introduction

Most of researchers have employed discrete Fourier transformation (DFT) to process multidimensional NMR data. Here the data must have certain formats consisting of 1D spectra which have the frequency information of indirect dimensions with serial and gradual intervals. For example, a 3D data of HNCO which has 1024* (H) × 64* (C) × 32* (N) time domain points comprises 1D data of 64* × 32* (= 8,192). Maximum points in each time domain impact the resolution of resulting frequency domain, and time domain size must be increased to result in better resolution by DFT. Therefore, a dataset in conventional 3D experiments need hours to several days of machine time, depending on the unit time for 1D. Recent developments in NMR hardware have improved the sensitivity and decreased the time for 1D, making it possible to measure some 3D data within 1 day. Nonetheless, several weeks of NMR machine times are still needed to gain all the 3D data for protein structure calculation.

To shorten the measuring time, several algorithms have been proposed during last several years. They are largely grouped by three ideas; one by introducing new hardware, the other by shortening repetition delay and the last by sparse sampling. Those for sparse sampling include GFT,[1] Projection-Reconstruction (PR),[2] and non-uniform sampling (NUS).[3] NUS means multidimensional data acquisition through random sparse sampling. Instead of 1D spectra at 64* × 32* points (= 8,192), NUS uses some parts of it, for example sparse 325 points (325 × 4 = 1,500 1D spectra). NUS has the advantage that systematic error due to the lack of samplings is minimized by randomization, which is compared with GFT and PR.

So far three methods to process NUS data have been reported and they include maximum entropy (MaxEnt), multidimensional decomposition (MDD), and multidimensional Fourier transformation (MFT).[3-5] Several studies have demonstrated the performances by NUS and these tools.[3,6-8] As long as sensitivity doesn't matter, one can gain 3D spectra for backbone assignment within one day with about 1/5 points compared to original ones for DFT in the proteins of medium size (~15 kDa).

The real-time acquisition can shorten the time by NUS even more. User processes data during acquisition and stops the measurement if the quality is acceptable, minimizing measuring time. Jaravine and Orekhov reported recently to combine the real-time NUS acquisition and MDD processing.[9] There they could reduce the time needed to about 1/20 compared to that of conventional experiment in a case. While MaxEnt algorithm has been known for several ten years, the mathematical framework for MDD is relative new. Although MDD works for the complicate dataset including NOESY, it demands huge computational time. Note that most of MDD calculations have used cluster system. It probably impedes to employ MDD for the real-time detection where acquisition and processing should be coupled quickly.

I suggest applying MaxEnt for the real-time acquisition of 3D experiments instead. MaxEnt is fast and several minutes are sufficient for processing with modern single CPU computer. In this study, I report the procedure and results of the real-time acquisition using MaxEnt with simulated sub-dataset from the experimental 3D data. Results demonstrate the real-time NUS/MaxEnt 3D HNCO of ubiquitin (Ub) and 3D CBCA(CO)NH of GB1[A34F] can be measured within 8 and 40 minutes respectively.

Experimental

Experimental 3D data. Two 3D experiments of 3D HNCO for Ub and 3D CBCA(CO)NH for GB1[A34F][10] proteins were used for this study. Ub consists of 76 residues, and GB1[A34F] is a dimer and a unit comprises 56 residues. Single mutation of Ala-34-Phe converts wild type monomeric GB1 into side-by-side dimer. The concentrations in this study were 1.7 mM and 1.2 mM for Ub and GB1[A34F] respectively. The data of 3D HNCO and CBCA(CO)NH have been gained by cryogenically cooled probe equipped Varian 600 MHz and Bruker 600 MHz machines respectively. The time domains have 3D matrices of 1024* × 64* × 32* (= 8,192 1D data) for Ub and 512* × 50* × 30* (= 6,000 1D data) for GB1[A34F]. The numbers of transients in 3D HNCO and CBCA(CO)NH are 2 and 16 respectively.

Schedules for non-uniform sampling. An in-house program which creates similar sampling to "sampsched2d" program of RNMRTK was made for NUS schedule. The program can generate three sampling profiles in each dimension, which are exponential, Gaussian and equally weighted distributions. Users decide the profile according to the characters of their samples and experiments. The details including their effects on the results will be published elsewhere (J.G. Jee, manuscript in preparation).

Non-uniform sampling data and maximum entropy data processing. Raw Varian data was converted into Bruker format and then the same procedures for 3D HNCO and CBACA(CO)NH were used except the parameters specific to each experiment. All the scripts in this study were written by Tcl, Python, Octave and Fortran. Based on the schedule, NUS data were made from intact full 3D data. Then MaxEnt processing scripts were generated automatically and all the calculations were executed using Mac OSX Intel 2.0 GHz dual-core machine. NMRPipe software package[11] is used for the normal processing of intact dataset, the calculation of baseline noise level and the peak pickings of 2D/3D dataset.

Results

The procedure which is used for the real-time acquisition

Figure 1. Schematic flow chart for the real-time acquisition with NUS and MaxEnt.

of NUS and MaxEnt is shown in Figure 1. First, the sampling schedule that is proper for the real-time acquisition is generated. Second, MaxEnt processing is fulfilled at given intervals of NUS acquisition. Third, the quality is judged with the simple procedure, deciding whether to continue or stop experiment automatically.

Sampling schedule of non-uniform sampling. Figure 2 displays example NUS sampling schedules for 3D HNCO of Ub. Different from original schedule by Wagner and his colleagues, where the points were arranged in ordered way, the current schedule has the shuffled order. The first three points are arranged to be (N_y^1, N_z^1), (N_y^{max}, N_z^1), and (N_y^y, N_z^{max}) (Fig. 2-A) and followed by shuffled other points, where $N_i^{1,max,k}$ mean the first, maximum, and any values in axis i. Here as the original schedule follows exponential distribution in ^{13}C dimension, sub-schedule from shuffled one has similar exponential distribution. For ^{15}N dimension which was obtained with constant time delay, the random
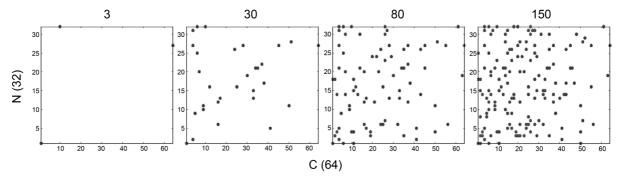
Figure 2. Example NUS sampling schedule. This NUS schedule is used for 3D HNCO of Ub. For the quadrature detection in each ^{13}C and ^{15}N dimension, two 1D spectra are needed respectively. Therefore, one point corresponds to four 1D spectra. The numbers of points are written on the top and the values in parentheses are the maximum sampling numbers in each dimension. As maximum sampling points $(N_C^{max} = 64$ and $N_N^{max} = 32)$ are identical in these schedules, all the resulting spectra by MaxEnt show the similar resolutions.

equal distribution is used. In this study, N_C^{max} and N_N^{max} are 64 and 32 for Ub, and 50 and 30 for GB1^{A34F} respectively. The processed spectra by MaxEnt from these schedules will result in comparable resolutions, since each schedule and resulting spectra have the identical maximum numbers. On the other hand, sensitivities vary dependent on the number of points. The detailed studies on the effects by different sampling schedule will be reported elsewhere.

Maximum entropy data processing. For the data processing by MaxEnt, the software package of Rowland NMR Toolkit (RNMRTK)[12] was employed. Other software may be a little bit different in some detailed parts of algorithm. MaxEnt uses two constraints *S* and *C* to process data. Target frequency domain data of *f*, which consists of *N* frequency points, can be gained from experimental time domain data *d*, of *M* time points, by maximizing *Q*,

$$Q(f,d) = S(f) - \lambda C(f,d)$$

Where

$$C(f,d) = \sum_{i=0}^{M-1} \left| IDFT(f_i) - d_i \right|^2$$

$$S(f) = -\sum_{n=0}^{N-1} \frac{|f_n|}{def} \log\left(\frac{\frac{|f_n|}{def} + \sqrt{4+\frac{|f_n|^2}{def^2}}}{2} \right) - \sqrt{4+\frac{|f_n|^2}{def^2}}$$

and IDFT means inverse discrete Fourier transformation. Here user has to adjust two parameters of λ and *def*. But in most cases there is no big difference in resulting data when user keeps the default values of λ and *def* respectively.[3] As

the purpose of this study is to check the quality of spectra in the middle of acquisition, the default values of 1.0 and 10.0 for λ and *def* were used respectively.

Sub-spectra by the real-time acquisition of NUS and MaxEnt. Figure 3 and 4 display the resulting NUS/MaxEnt spectra of Ub and GB1^{A34F} respectively at 1, 5, and 10% sampling points of original data. (A) contains 2D overlay by ^{13}C-^{1}H plane and (B) by ^{15}N-^{1}H. They demonstrate the robustness of NUS and MaxEnt well. Only with 10% points of original schedule, 200 for Ub and 150 for GB1^{A34F}, all the expected peaks were observed. The peaks positions are identical to those found in normal 3D data, as shown in overlaying Figures (Supplementary Figure 1, 2). It should be noted that as expected by the schedule, the resolutions during the acquisition are nearly identical in all the sub-spectra. Only signal-to-noise ratio of peaks varies, and it gets better by increasing the number of points. However, one cannot predict at which point the resulting spectrum reaches to the enough quality for the analysis. For instance, it is likely that most of the peaks exist with 5% points and we can find some peaks even with 1% points. If one checks all the peaks directly and manually at every step, it will be straightforward to judge the qualities. However, it is not proper for the real-time acquisition aiming at automation. Instead, the peaks in each NUS/MaxEnt spectrum are compared with the peaks of reference, which is 2D ^{15}N-^{1}H HSQC for 3D HNCO and 3D CBCA(CO)NH.

Automatic judging process. To quantitize the quality of each spectrum at certain sampling point, calculation of baseline noise level, peak picking of NUS/MaxEnt data, and matching of two peak lists are combined in automatic manner. First, automatic peak picking with "pk.tcl" script of
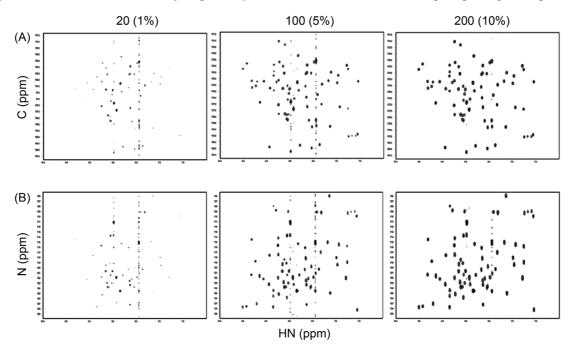

Figure 3. MaxEnt data of the real-time NUS sampling for Ub. 2D ^{13}C-^{1}H overlay (A) and ^{15}N-^{1}H overlay (B). The numbers of sampling points are drawn on the top. The percentage values in parentheses mean the portions of sampling points when those for full data in intact normal 3D are 100%. All the spectra were drawn with the thresholds whose positions are just above the baseline noise levels for the comparison of sensitivity. There is no difference in the peaks' positions compared to those in normal 3D data (Supplementary Figure 1).
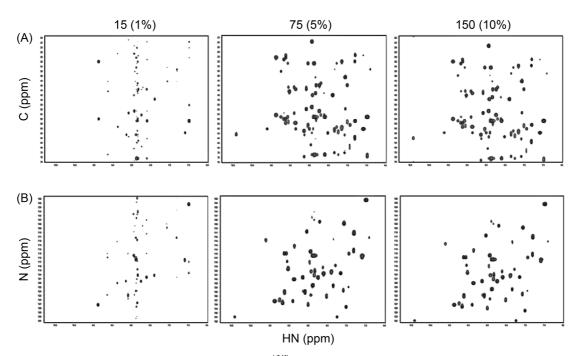
Figure 4. MaxEnt data of the real-time NUS sampling for GB1^{A34F}. All the values have the same notations to those in Figure 3. As 3D CBCA(CO)NH produces twice more peaks than 3D HNCO, 2D ^{13}C-^{1}H overlay of GB1^{A34F} contains more peaks than that in Figure 3. Also it is found that the peaks' positions are almost the same in NUS/MaxEnt and normal 3D data (Supplementary Figure 2).
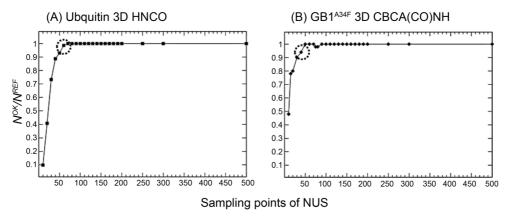

Figure 5. Change of spectral qualities during the real-time acquisition and NUS/MaxEnt. N^{REF} is the number of total peaks in reference, 2D ^{15}N-^{1}H HSQC. N^{OK} is the number of reference peaks whose corresponding peaks are found in 3D NUS/MaxEnt. The value of N^{OK}/N^{REF} indicates how many reference peaks are found in NUS/MaxEnt dataset. In this study, when N^{OK}/N^{REF} is higher than 0.95, the result is judged to have enough quality. The samplings at dashed circle satisfy the criteria and which are 60 and 40 points for 3D HNCO of Ub and 3D CBCA(CO)NH GB1^{A34F} respectively. For the details, see the text.

NMRPipe is tried with enough threshold not to pick noises. Empirically the value of 30 is chosen as multiplying factor to baseline noise level. The peak picking with this threshold makes all the peaks have higher intensities than baseline at least 30-folds. Among the picked peaks, second, the N^{NUS} peaks of the highest intensities are extracted. Here N^{NUS} is the number of the peaks in 3D data, and is predicted by users according to the experiments. For example, in the ideal case, the number of peaks in 3D HNCO will be the same to that of 2D ^{15}N-^{1}H HSQC, and the peaks in 3D CBCA(CO)NH will be twice more. To take into account the erroneous peaks which may happen by imperfectness of experiments, N^{NUS} is multiplied by 1.5. Thus if 100 peaks are expected, actually

150 are considered as N^{NUS}. Third, a script compares the peaks N^{NUS} with the reference peaks (N^{REF}) and calculate how many peaks of reference are found in N^{NUS} with shared dimensions between NUS/MaxEnt and 2D reference data, setting it as N^{OK}. When the ratio of N^{OK}/N^{REF} reaches to 0.95, the current protocol judges the experiment has enough quality to analyze. The value of 0.95 is determined empirically and it can be 1.0 if user wants better qualities. Figure-5 contains the change of N^{OK}/N^{REF} ratios in Ub and GB1^{A34F} by this protocol. At 60 points of Ub and 40 points of GB1^{A34F}, they start to satisfy the criteria.

Real-time and NUS/MaxEnt 3D data acquisition within 10 minutes. Figure 6 has the representative strip plots from
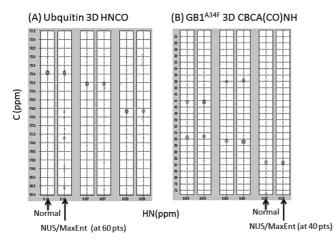
Figure 6. Comparison of strip plots of normal 3D DFT and final data by NUS/MaxEnt. Representatively, three peaks of 2D ^{15}N-^{1}H HSQC were taken and their ^{13}C-^{1}H planes were drawn. Left-side strip plots were taken from normal DFT 3D data of Ub 3D HNCO and GB1^{A34F} 3D CBCA(CO)NH, which were processed by NMRPipe[11]. Right-side plots came from NUS/MaxEnt in this study and their corresponding sampling points are 60 and 40 for Ub and GB1^{A34F} respectively.

3D HNCO of Ub and 3D CBCA(CO)NH of GB1^{A34F} in normal and NUS/MaxEnt 3D experiments, and shows really the resulting spectra have the good qualities for the analysis. Nearly all the peaks in the normal 3D could be found at NUS/MaxEnt data with points of 60 (Ub) and 40 (GB1^{A34F}). Though several small noises are found in 3D HNCO of Ub, they are not problematic for interpretation (*i.e.* backbone assignment). These points of 60 and 40 correspond to only 2.9 and 2.7% of original Ub and GB1^{A34F} dataset and the acquisition times of about 8 and 40 minutes respectively. The sensitivity of 3D CBCA(CO)NH is half compared with that of 3D HNCO, meaning 4 times longer acquisition time to have the same signal-to-noise ratio. Since the sample conditions are different, direct comparison between 3D HNCO and 3D CBCA(CO)NH is difficult, but current result is consistent with the prediction.

Discussion

It is noteworthy that to process the data with MaxEnt, a personal computer of Mac OSX dual core 2.0 GHz is used and all the processing were done within 3 minutes. Considering that most of computers to handle NMR operation are single CPU machines, my approach can be incorporated readily. Jaravine and Orekhov discussed that MDD can use just part of reference spectra and by comparing the most crowd region they could shorten the time needed to process sub-data and judge the quality of the real-time data. However, one cannot predict the efficiency of magnetization transfer in 3D experiments *a priori*. If some peaks undergo worse magnetization transfer, they can be missing during acquisition, although those in the certain regions (*e.g.* most crowd regions) are all detectable. Therefore, instead of comparing parts, to check all the regions will be more reliable. In

addition, it should be emphasized that by locating minimum and maximum values at the first three points of the sampling schedule, only sensitivity varies in current algorithm and it makes the comparison of resulting spectra much simpler. In conventional DFT 3D data, both of sensitivity and resolution change and it is not easy to judge the quality in automatic manner.

It is known that MaxEnt is not proper for the processing of the data like 3D HCCH-TOCSY and 3D NOESY-HSQC due to their large dynamic ranges of peak intensities in these experiments. However, the purpose of this experiment is to judge whether the quality is enough or not. Thus, alternatively, one can stop acquisition, when there is no improvement in the number of peaks compared to the result in the previous sampling points. And the final processing can be done with time-consuming but sophisticated method like MDD. However, in the case of triples resonances for the backbone assignment purpose, there is no big differences in the peaks qualities between MaxEnt and MDD (J.G. Jee, manuscript in preparation).

The time reduction of 3D experiments from the real-time acquisition by NUS and MaxEnt will be improved by introducing new hardware or shortening repetition delay, since they are independent each other and can be used together. Schandra *et al.* reported that they could gain a triple resonance within 10 minutes with SOFAST methods by reducing repetition delay to 50 ms and having number of scan of 1.[13] Its performance is comparable to that by NUS/MaxEnt in this study and the combination of SOFAST and NUS/MaxEnt is expected to reduce the time even more, as long as sensitivity is not issue.

The applicable areas of fast-NMR by the real-time acquisition and MaxEnt/NUS will be wide. Most of all, one can apply it for the study of proteins having short life time. There are many proteins which are unstable *in vitro* and at the concentration for NMR study. As long as the sensitivity suffices, one can gain all the spectra for the backbone assignment within several hours and gain structural information before the protein is damaged. Also researchers can apply it for the high-throughput screening like SAR by NMR. When it is difficult to chase the change of peaks due to its complexity in 2D, one can use 3D instead without demanding huge measuring times. Finally, I expect that the researchers who don't have NMR machine and suffer from insufficient machine times can have the opportunities to do structural study with modern NMR machines by employing the real-time acquisition.

Supporting Information Available. Additional figures to compare normal (100%) DFT and NUS (10%) and MaxEnt.

This material is available free of charge via the Internet at http://newjournal.kcsnet.or.kr or on request from the author.

References

1. Kim, S.; Szyperski, T. *J. Am. Chem. Soc.* **2003**, *125*, 1385-1393.
2. Kupce, E.; Freeman, R. *J. Am. Chem. Soc*. **2003**, *125*, 13958-13959.
3. Rovnyak, D.; Frueh, D. P.; Sastry, M.; Sun, Z. Y.; Stern, A. S.; Hoch, J. C.; Wagner, G. *J. Magn. Reson.* **2004**, *170*, 15-21.
4. Kazimierczuk, K.; Kozminski, W.; Zhukov, I. *J. Magn. Reson.* **2006**, *179*, 323-328.
5. Orekhov, V. Y.; Ibraghimov, I.; Billeter, M. *J. Biomol. NMR* **2003**, *27*, 165-173.
6. Kazimierczuk, K.; Zawadzka, A.; Kozminski, W.; Zhukov, I. *J. Am. Chem. Soc.* **2008**, *130*, 5404-5405.
7. Jaravine, V.; Ibraghimov, I.; Orekhov, V. Y. *Nat. Methods* **2006**, *3*, 605-607.
8. Tugarinov, V.; Kay, L. E.; Ibraghimov, I.; Orekhov, V. Y. *J. Am. Chem. Soc.* **2005**, *127*, 2767-2775.
9. Jaravine, V. A.; Orekhov, V. Y. *J. Am. Chem. Soc.* **2006**, *128*, 13421-13426.
10. Jee, J.; Byeon, I. J.; Louis, J. M.; Gronenborn, A. M. *Proteins* **2008**, *71*, 1420-1431.
11. Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277-293.
12. Hoch, J. C.; Stern, A. S. *Methods Enzymol* **2001**, *338*, 159-178.
13. Schanda, P.; Van Melckebeke, H.; Brutscher, B. *J. Am. Chem. Soc.* **2006**, *128*, 9042-9043.