J|A|C|S COMMUNICATIONS

Published on Web 02/02/2010

Poisson-Gap Sampling and Forward Maximum Entropy Reconstruction for Enhancing the Resolution and Sensitivity of Protein NMR Data

Sven G. Hyberts, Koh Takeuchi, and Gerhard Wagner*

Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115

Received September 21, 2009; E-mail: Gerhard_Wagner@hms.harvard.edu

The capabilities of modern NMR spectrometers have recently improved dramatically through the use of stronger magnetic fields. To utilize the potential resolution of these spectrometers in multidimensional experiments, various forms of sparse or nonuniform sampling (NUS) have been proposed.[1,2] For optimal processing of such spectra, we recently developed the forward maximum entropy (FM) reconstruction method,[3] which was improved and combined with a distillation procedure.[4]

The quality of spectra obtained from NUS depends crucially on the sampling schedules. In the past, we examined various forms of random sampling and realized that the quality of data retrieval depended significantly on the choice of the seed number when using standard Unix random number generators (e.g., drand48). We realized that (1) big gaps in the sampling schedule are generally unfavorable and (2) gaps at the beginning or end of the sampling are worse than those in the middle. A third, crucial criterion is that (3) the sampling requires suitably random variation to prevent violation of the Nyquist theorem.

To cope with this, we have evaluated a sinusoidal-weighted Poisson distribution of the gap lengths between sampling points followed by FM reconstruction. To achieve this distribution, we assume an average gap length of λ in the common Fourier grid and a specific gap size k between two acquired data points. Thus, a λ of 0.0 yields uniform sampling, while a λ of 1.0, for example, creates a nonuniform schedule of 50% overall sampling density.

The overall probability f for a specific gap size $k \geq 0$ is assumed to be given by the Poisson distribution: $f(k; \lambda) = (\lambda^k e^{-\lambda})/(k!)$. This would satisfy criteria (1) and (3). Obviously, no integer values of k less than zero are allowed, as no negative gap sizes are realizable.

To satisfy point (2) above, we further optimized the sampling schedule with a sinusoidal variation of λ; we call this sine-weighted Poisson-gap sampling (SPS). Here $\lambda = \Lambda \sin \theta$, where Λ is an adjustment factor used to make the average λ satisfy the targeted sampling density. θ varies linearly from 0 to π through the sampling schedule when no apodization is applied prior to reconstruction. As apodization commonly scales the signal to zero at the end of the evolution time, we restrict the variation of θ from 0 to $\pi/2$ when apodization is intended. The method intrinsically imposes some "order", and sampling points are not chosen fully stochastically. The sinusoidal weighting of gap sizes is equivalent to very dense sampling at the beginning of the time-domain data. This is ideal for exponentially decaying time-domain data. For other data, such as antiphase signals, a different weighting may be optimal. A C program for generating the SPS schedule is provided in the Supporting Information (SI).

To test the performance of Poisson-gap sampling, we generated a free-induction decay for a single resonance, tested different sampling schedules, and examined how accurately the FM algorithm could reconstruct the signals. As an example, 256 out of 1024 time domain data points were extracted with different selection procedures. The data sets were then FM-reconstructed and Fourier transformed without apodization. In particular, we tested how the value of the Unix seed number affects the quality of the result. Figure 1 shows an L^2 norm analysis of the accuracy of the reconstruction for 100 seed numbers. The L^2 values were ordered according to size. The top, middle, and bottom traces show results for plain random sampling, Poisson-gap sampling without modulation, and SPS with $\theta = [0, \pi]$, respectively. Poisson-gap sampling alone is almost 2-fold better than plain random sampling, and SPS is by far the best. Importantly, it is almost completely insensitive to the choice of the Unix seed number. Thus, SPS is our method of choice for generating sampling schedules.

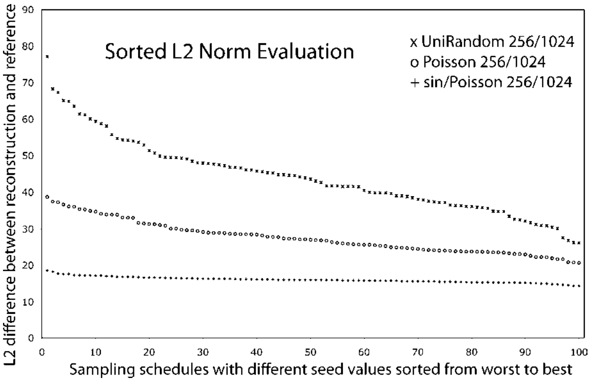

Figure 1. Effect of Poisson-gap sampling on the quality of spectra obtained with FM reconstruction. A synthetic time-domain signal (1024 data points) was created, and 256 sampling points were selected with three different methods. The L^2 values (i.e., the Euclidian norm $L^2 = \|\mathbf{f} - \mathbf{f}^{\mathbf{rec}}\| = [\sum_i (f_i - f_i^{rec})^2]^{1/2}$ or a non-normalized root-mean-square deviation) of the difference between the reconstructed and linearly sampled spectra were calculated for 100 different Unix seed numbers and ordered according to decreasing L^2 values. (top) Plain random sampling. (middle) Poisson-gap sampling without λ variation. (bottom) Poisson-gap sampling with a sinusoidal variation of λ ($\theta = [0, \pi]$). The latter procedure clearly has the lowest L^2 values and is nearly independent of the seed number.

Next we examined experimentally the performance of Poisson-gap sampling and FM reconstruction on a 2D ^{13}Cα-detected NCA experiment on an alternately ^{13}C-labeled B1 domain of protein G (GB1).[5] We compared various sampling schedules and analyzed their effect on spectral quality, signal-to-noise ratio (S/N), recovery of very weak peaks, and fidelity of peak positions. The results are summarized in Figure 2. A total of five NCA experiments (labeled A1−A5) were recorded. Two NCA reference spectra of 256 uniformly sampled increments in which the number of scans per increment (ns) was (A1) 2 and (A5) 8 were measured in 22 min and 1.5 h, respectively. We then recorded three spectra with one-quarter of the increments selected but accumulating eight scans per increment. These required 22 min of measuring time each and were

10.1021/ja908004w © 2010 American Chemical Society

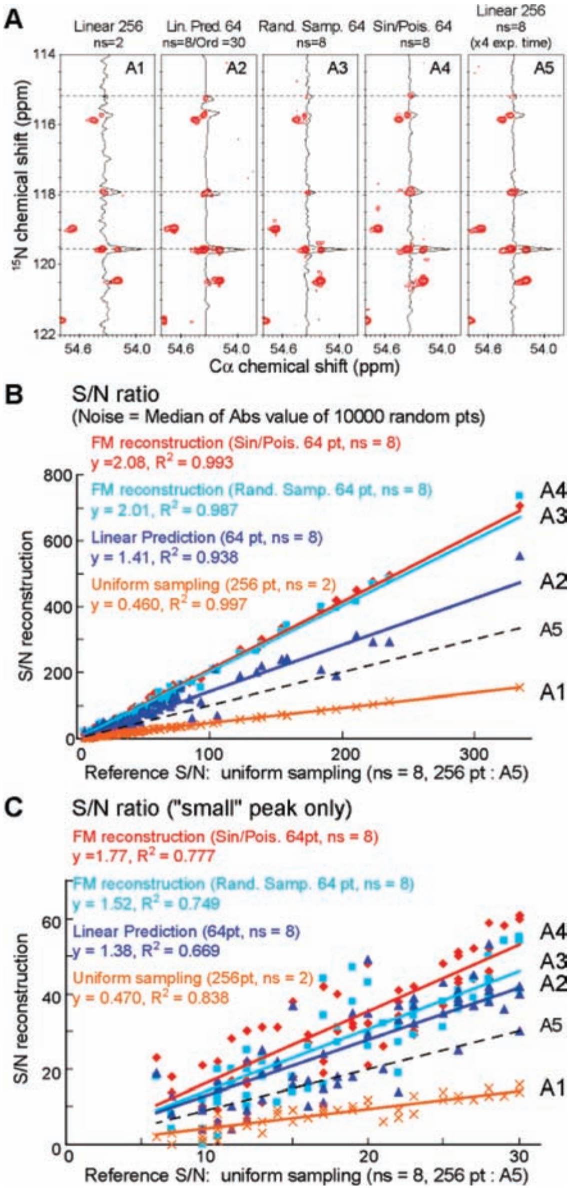

Figure 2. Comparison of nonuniform sampling schedules with uniform acquisition and linear prediction. (A) Representative strips from the NCA spectrum[5] of alternately ^{13}C-labeled protein GB1. The four strips at the left represent experiments with the same total measuring time (512 total scans). Strip 5 was recorded with 256 linear increments and eight scans per increment (ns = 8), requiring a 4-fold longer experiment. For panel A2, the first 64 linear increments were extended with linear prediction using an order parameter of 30. Panels A3 and A4 were obtained with random sampling (64 points) and SPS (64 points), respectively. (B, C) Plots of the S/N values of (B) all and (C) small peaks. The S/N values of the selected peaks obtained with procedures A1−A4 are plotted against the A5 S/N.

sampled as follows: (A2) 64 first of 256 uniformly sampled increments, (A3) 64 out of 256 increments randomly selected with uniform sampling density, and (A4) 64 out of 256 selected by SPS with $\theta = [0, \pi/2]$, where Λ was adjusted by the schedule generator to create the requested number of sampling points (see the SI). The nonuniformly sampled spectra (A3 and A4) were subsequently FM-reconstructed. The spectrum A2 that sampled the 64 first increments was linearly predicted to 256 time points using an order parameter of 30 (see the SI for the choice of this order parameter). The effective maximum nitrogen evolution time for all experiments was 84 ms, which is ~30% of the ^{15}N T_2. A representative 2D strip with a cross section through the Cα position of residue 15 is used in Figure 2A for comparison. The panels are labeled A1 to A5 to indicate the acquisition methods described above.

Figure 2B,C plots the S/N values of all and only weak peaks, respectively. As the signal we used the peak height, and the median of the absolute values of 10 000 randomly picked spectral points was used as the noise. The S/N values of A1−A4 were plotted against that of spectrum A5 recorded with the 4-fold longer measuring time and fitted using linear regression. The linear coefficients shown in Figure 2B,C represent the S/N relative to that of conditions A5, and the R^2 terms report the fidelities of reconstruction.

The data show that the S/N of A1 is approximately half that of A5 (0.46 and 0.47 in panels B and C, respectively), since only one-quarter of the scans were recorded. Plain random sampling with FM reconstruction (A3) yields S/N values relative to A5 of 2.01 and 1.52 for strong and weak peaks, respectively. SPS (A4) results in S/N values relative to A5 of 2.08 and 1.77 for strong and weak peaks, respectively. Thus, SPS combined with FM reconstruction performs best. Interestingly, the S/N of A4 is ~4 times higher than that of A1 although it was obtained in the same total measuring time, and it is ~2 times higher than that of A5 although it was obtained in one-quarter of the measuring time. Thus, the procedures described result in a significant gain in S/N per measuring time.

We next asked whether the procedure suffers from false positives. Indeed, there are some weak peaks (at 118.0 and 118.5 ppm) in panel A4 and also a few weak peaks in panel A3. To explore this, we selected 10 areas of the spectrum that did not contain peaks and measured the signal to peak noise. The results are shown in Figure S1 in the SI and indicate that Poisson-gap sampling also has superior signal to peak noise but to a smaller extent, particularly for small peaks.

We then asked whether the small false-positive peaks seen for A3 and A4 in Figure 2 are systematic or random artifacts. We recorded 10 spectra with the conditions of Figure 2A1 (256 linear increments, ns = 2), and 10 spectra with the conditions of Figure 2A4 (64 SPS increments, ns = 8). The results are shown in Figure 3. None of the false peaks show up reproducibly and are thus random artifacts; these are of minor concern, since peaks are typically not trusted unless they are seen reproducibly.

Next we asked whether this NUS/FM reconstruction approach could enhance the sensitivity. Here we define sensitivity as the ability to distinguish weak peaks from noise. As an example, the spectrum shown in Figure 2A contains a weak peak at the nitrogen position of 115.1 ppm (marked near the top with a dotted line). This peak could not be observed with linear sampling (Figure 2A1); however, it is obvious and strong in the spectrum using SPS (A4), although both data sets used the same experimental time, and it is better defined than in the spectrum A5, although it was recorded in only a quarter of the time. Thus, the SPS procedure described here, together with FM reconstruction, seems to increase the sensitivity in addition to the S/N.

Is this sensitivity gain reproducible? To answer this, we compared the uniformly and nonuniformly sampled spectra of Figure 3. The figure shows cross sections and contour plots. The latter are drawn with two different noise levels. The noise level for the spectra recorded with ns = 8 was set to be twice as high as that for the spectra recorded with ns = 2, since the maximum noise value for the ns = 8 spectra (both uniformly and NUS sampled) was twice as high as in spectra with ns = 2. Although it was somewhat subjective to decide whether a peak was there above the noise, the comparison of the spectra clearly shows that procedure A4 (SPS) consistently identifies the weak peak (in the bottom panels of Figure

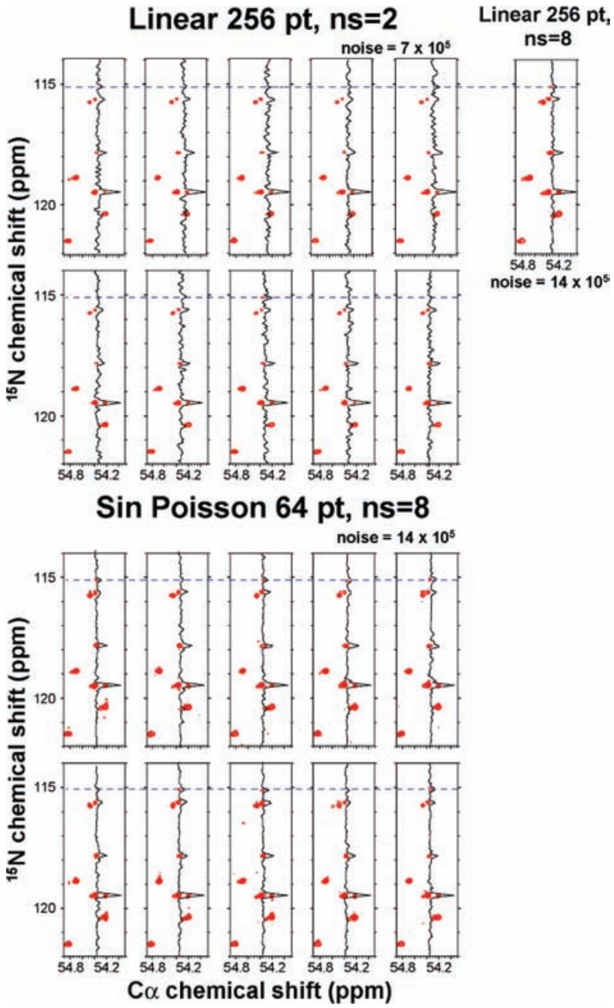

Figure 3. Reliability of small-peak detection. (top left) The same NCA spectrum was recorded 10 times linearly with 256 increments and ns = 2. The same strip as in Figure 2 is shown. (bottom) The same NCA spectrum was recorded 10 times using NUS with 64 increments and ns = 8. The measuring time was 22 min for each of the experiments. (top right) For comparison, the spectrum was also recorded linearly with 256 increments and ns = 8 (total measuring time 1.5 h). The weak peak at the ^{15}N chemical shift of 115.1 ppm, which is clearly manifested in the long linear experiment (1.5 h) is also clearly seen in the NUS spectra recorded in one-fourth the time (22 min). On the other hand, it is barely visible in only a few of the short linear experiments of 22 min.

3, the peaks are obvious in 7 of 10 contour plots, and the cross sections are always positive). However, with procedure A1, the peaks are obvious only in 2 of 10 contour plots, and the cross sections are even negative in some cases (Figure 3, top panels). This clearly indicates better ability to distinguish signals from noise in the reconstructed spectra. Thus, the Poisson-gap NUS and FM reconstruction enhances the sensitivity, i.e., the ability to detect weak peaks above noise.

This sensitivity enhancement seems counterintuitive. However, one should keep in mind that the FM reconstruction is not a transformation but a minimization procedure. On the basis of the design of the approach,[3,4] the only experimental noise originates from the measured time-domain data points. Since we can measure here 4 times more scans than in the uniformly sampled data, the signal-to-noise of the measured time-domain data points is 2-fold higher. There is no experimental noise in the reconstructed data points. However, there is reconstruction noise due to the point-spread function (sampling schedule). We minimize this reconstruction noise by optimizing the sampling schedule and by the conjugant gradient optimization. Thus, it is possible to increase the signal-to-noise ratio and the ability to enhance the detection of weak signals above noise, which we consider to be an enhancement of the sensitivity.

We also compared the Poisson-gap sampling and FM reconstruction with linear prediction based on the first 64 linear increments (Figure 2A2). The S/N values relative to A5 are 1.41 and 1.38 for strong and weak peaks, respectively. This is significantly better than A1 and even A5. However, we had to use a large order parameter of 30 in the nmrPipe program.[6] The benefits of using large order parameters in linear prediction are discussed in detail in the SI and in Figure S2. However, as previously reported,[7] linear prediction can cause small changes in peak positions, and this is clearly seen for the peak at the nitrogen position of 115.1 ppm (top dotted line in Figure 2 and Figure S2). Such chemical shift changes are not observed in the SPS/FM reconstruction approach. Furthermore, linear prediction clearly suffers from significantly lower resolution, which matters for crowded spectral regions, as shown in Figure S3.

Currently, the Poisson-gap sampling is implemented only in a single NUS dimension. However, implementation in multiple dimensions is possible and is currently being pursued.

Acknowledgment. This research was supported by the National Institutes of Health (Grants GM47467 and AI37581).

Supporting Information Available: Comparison of signal to peak noise ratios, effect of order parameters in linear prediction, comparison of resolution using linear prediction and Poisson-gap sampling, graphical representation of the SPS schedule used, and source code for a C program generating the SPS schedule. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Mobli, M.; Stern, A. S.; Hoch, J. C. *J. Magn. Reson.* **2006**, *182*, 96.
(2) Kazimierczuk, K.; Zawadzka, A.; Kozminski, W. *J. Magn. Reson.* **2008**, *192*, 123.
(3) Hyberts, S. G.; Heffron, G. J.; Tarragona, N. G.; Solanky, K.; Edmonds, K. A.; Luithardt, H.; Fejzo, J.; Chorev, M.; Aktas, H.; Colson, K.; Falchuk, K. H.; Halperin, J. A.; Wagner, G. *J. Am. Chem. Soc.* **2007**, *129*, 5108.
(4) Hyberts, S. G.; Frueh, D. P.; Arthanari, H.; Wagner, G. *J. Biomol. NMR* **2009**, *45*, 283.
(5) Takeuchi, K.; Sun, Z. Y.; Wagner, G. *J. Am. Chem. Soc.* **2008**, *130*, 17210.
(6) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277.
(7) Stern, A. S.; Li, K. B.; Hoch, J. C. *J. Am. Chem. Soc.* **2002**, *124*, 1982.

JA908004W