NMR spectral analysis using prior knowledge

Takuma Kasai[1,2,5], **Kenji Nagata**[3], **Masato Okada**[3] **and Takanori Kigawa**[1,2,4,5]

[1] Laboratory for Biomolecular Structure and Dynamics, Cell Dynamics Research Core, RIKEN Quantitative Biology Center, Yokohama, Japan
[2] JST CREST, Yokohama, Japan
[3] Department of Complexity Science and Engineering, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan
[4] Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, Yokohama, Japan

E-mail: takuma.kasai@riken.jp, kigawa@riken.jp

Abstract. Signal assignment is a fundamental step for analyses of protein structure and dynamics with nuclear magnetic resonance (NMR). Main-chain signal assignment is achieved with a sequential assignment method and/or an amino-acid selective stable isotope labeling (AASIL) method. Combinatorial selective labeling (CSL) methods, as well as our labeling strategy, stable isotope encoding (SiCode), were developed to reduce the required number of labeled samples, since one of the drawbacks of AASIL is that many samples are needed. Signal overlapping in NMR spectra interferes with amino-acid determination by CSL and SiCode. Since spectral deconvolution by peak fitting with a gradient method cannot resolve closely overlapped signals, we developed a new method to perform both peak fitting and amino acid determination simultaneously, with a replica exchange Monte Carlo method, incorporating prior knowledge of stable-isotope labeling ratios and the amino-acid sequence of the protein.

1. Introduction

Nuclear magnetic resonance (NMR) is a widely used method for protein analysis. In NMR spectra, each observable atom in the protein gives a signal at its specific frequency, called a chemical shift, which is determined by the chemical environment of the atom. Moreover, the characteristics of the signals, such as appearances, intensities, and/or line widths, reflect the chemical bonds, distances, mobilities, and/or environments of the atoms, depending on the types of measurements. Therefore, by combining various NMR measurements, the three-dimensional structures and dynamics of the protein can be analyzed.

The determination of the chemical shifts of the atoms, which is called the signal assignment, is usually the first step of the analysis. Sequential assignment with triple resonance measurements is a commonly used technique for the assignment of main chain atoms, by searching neighboring amino acid residues [1]. Additionally or alternatively, amino-acid selective stable isotope labeling (AASIL) is used, especially for challenging targets such as large proteins [2], low-solubility proteins [3] or for protein analyses by in-cell NMR [4], because in such cases, the fast decay of NMR signals, the low

[5] To whom any correspondence should be addressed.

signal-to-noise ratio and/or the signal overlapping may interfere with triple resonance measurements. Naturally abundant ^{12}C and ^{14}N are unobservable with NMR, while ^{13}C and ^{15}N are observable stable isotopes (SIs). For the simple AASIL method, each protein sample corresponds to one amino acid that is selectively ^{15}N labeled, while the others retain ^{14}N. Using these sets of protein samples, one can know the amino acid from which each signal is derived. Furthermore, in the dual selective approach, using a combination of two amino acids in which one is selectively ^{15}N labeled and the other is ^{13}C labeled, one can also know the amino acid of the preceding (N-terminal side) residue by a measurement using the amide nitrogen and the carbonyl carbon [5, 6]. This information greatly narrows down the assignment possibilities, and therefore is useful for main chain assignments. However, because the number of standard amino acids is 20, the simple AASIL method requires a large number of labeled protein samples, and thus generates arduous sample preparation workloads and consumes NMR machine time.

To reduce the number of labeled samples required in AASIL, various combinatorial selective labeling (CSL) methods were proposed [7-18]. In CSL, each amino acid is represented by a combination of multiple samples. We developed a new strategy, named "stable isotope encoding" (SiCode) [19], to further reduce the number of samples, by regarding AASIL as a "encoding-and-decoding" process. Specifically, the amino-acid information is encoded into the SI-labeling ratio pattern of the samples, and decoded from the signal intensity ratio of the NMR spectra. With this strategy, using ternary digits as codewords, 19 amino acids are represented by only 3 labeled samples [19]. However, for both CSL and SiCode, occasional signal overlapping may lead to misinterpretation of the amino acid information. Although the original SiCode procedure successfully decoded some overlapped signals [19], further improvement is needed.

In this paper, we report an improved decoding procedure for SiCode, which utilizes the prior knowledge of the SI labeling pattern and the amino acid sequence of the protein for signal deconvolution. Nagata et al. [20] reported that the Bayesian spectral deconvolution and model selection problems were simultaneously solved, by finding both the fitting parameters and free energy for model selection with the replica exchange Monte Carlo (REMC) method. This method has been applied to SiCode, by regarding the decoding process as a model selection problem.

2. Theory

2.1. The SiCode decoding problem

In the SiCode strategy, the SI labeling pattern is regarded as a codeword table and thus is predefined, as shown in table 1, for example. $\boldsymbol{c}(a) \in [0,1]^S$ and $\boldsymbol{n}(a) \in [n_{\min}, 1]^S$ are S-dimensional vectors that represent the ^{13}C and ^{15}N labeling ratios of amino acid a, respectively, where S is the number of labeled samples, and $n_{\min}$ is a minimum ^{15}N labeling ratio and is set to be greater than zero to avoid loss of information for the ^{13}C labeling ratio, as discussed later. For example, in the labeling pattern shown in table 1, $n_{\min}$ is 0.5. To estimate the signal amplitude, at least one sample is 100% labeled, therefore

$$\forall a, \max \boldsymbol{n}(a) = 1, \max \boldsymbol{c}(a) = 1. \tag{1}$$

^{1}H-^{15}N HSQC (hereafter "HSQC") is a fundamental two-dimensional NMR spectrum. The intensities of the HSQC spectrum are assumed to be

$$\boldsymbol{I}_{\mathrm{HSQC}}(x,y) = \sum_i A^i_{\mathrm{HSQC}}\ \boldsymbol{n}(a_i) \exp\left(-\frac{\left(x - x_0^i\right)^2}{2{\sigma_x^i}^2} - \frac{\left(y - y_0^i\right)^2}{2{\sigma_y^i}^2}\right) \tag{2}$$

where $\boldsymbol{I}_{\mathrm{HSQC}}(x,y)$ is an S-dimensional vector that represents the intensities of HSQC, x and y are the chemical shifts of the ^{1}H and ^{15}N axes, respectively, A^i_{HSQC} is the amplitude of the HSQC signal of residue i, a_i is the amino acid of residue i, x_0^i and y_0^i are the ^{1}H and ^{15}N chemical shifts of the center of the signal of residue i, respectively, and σ_x^i and σ_y^i are the ^{1}H and ^{15}N line widths of the signal of

residue i, respectively. Note that A^i_{HSQC} is 0 if a_i is proline, because of the absence of the amide hydrogen.

The two-dimensional version of HN(CO) (hereafter "HNCO") is another spectrum used in the SiCode strategy. As in HSQC, the two axes of HNCO correspond to the amide hydrogen and nitrogen. Therefore, their signals should appear at the same chemical shifts as in HSQC, while their intensities are affected by both the ^{15}N labeling ratio of the amide nitrogen of residue i and ^{13}C labeling ratio of the carbonyl carbon of residue $i-1$, as follows:

$$\boldsymbol{I}_{\mathrm{HNCO}}(x,y) = \sum_i A^i_{\mathrm{HNCO}}\ \boldsymbol{n}(a_i) \circ \boldsymbol{c}(a_{i-1}) \exp\left(-\frac{\left(x-x_0^i\right)^2}{2{\sigma_x^i}^2} - \frac{\left(y-y_0^i\right)^2}{2{\sigma_y^i}^2}\right) \quad (3)$$

where $\boldsymbol{I}_{\mathrm{HNCO}}(x,y)$ is an S-dimensional vector that represents the intensities of HNCO, A^i_{HNCO} is the amplitude of the HNCO signal of residue i, and $\circ$ denotes element-wise multiplication. As in HSQC, A^i_{HNCO} is 0 if a_i is proline. Note that if the ^{15}N labeling ratio of one of the samples is a very small value, then we cannot obtain the ^{13}C labeling ratio information for the sample, due to the weak HNCO signal. To avoid this situation, each element of the $\boldsymbol{n}(a)$ value is set to $n_{\min}$ or larger. The decoding problem of SiCode is to determine a_i and a_{i-1} for each signal appearing in the spectra.

Table 1. SI labeling patterns used in this study

amino acid	sample 1		sample 2		sample 3		corresponding codeword
	^{13}C	^{15}N	^{13}C	^{15}N	^{13}C	^{15}N	
G	100%	100%	100%	100%	100%	100%	222
F	100%	100%	100%	100%	50%	75%	221
N	100%	100%	100%	100%	0%	50%	220
L	100%	100%	50%	75%	100%	100%	212
S	100%	100%	50%	75%	50%	75%	211
D	100%	100%	50%	75%	0%	50%	210
M	100%	100%	0%	50%	100%	100%	202
K	100%	100%	0%	50%	50%	75%	201
R	100%	100%	0%	50%	0%	50%	200
A	50%	75%	100%	100%	100%	100%	122
I	50%	75%	100%	100%	50%	75%	121
C	50%	75%	100%	100%	0%	50%	120
V	50%	75%	50%	75%	100%	100%	112
Y	50%	75%	0%	50%	100%	100%	102
Q	0%	50%	100%	100%	100%	100%	022
E	0%	50%	100%	100%	50%	75%	021
H	0%	50%	100%	100%	0%	50%	020
T	0%	50%	50%	75%	100%	100%	012
W	0%	50%	0%	50%	100%	100%	002
P	0%	0%	0%	0%	0%	0%	

2.2. The original version of the decoding procedure (sequential decoding)

In the original version of SiCode, the decoding procedure consisted of three sequential steps: peak fitting, calculation of SI-labeling ratios, and determination of amino acids. Hereafter, we refer to this as "sequential decoding". Since it is difficult to analyze the full region of the spectra at the same time, the spectra are divided into small regions. The least square fitting to the following function is then performed for each region:

$$\begin{pmatrix}\boldsymbol{I}_{\mathrm{HSQC}}(x,y)\\ \boldsymbol{I}_{\mathrm{HNCO}}(x,y)\end{pmatrix}=\sum_{k=1}^{K}\begin{pmatrix}\boldsymbol{B}^{k}_{\mathrm{HSQC}}\\ \boldsymbol{B}^{k}_{\mathrm{HNCO}}\end{pmatrix}\exp\left(-\frac{\left(x-x_0^k\right)^2}{2{\sigma_x^k}^2}-\frac{\left(y-y_0^k\right)^2}{2{\sigma_y^k}^2}\right) \quad (4)$$

where K is the total number of signals in the region and is assumed to be given, and $\boldsymbol{B}^{k}_{\mathrm{HSQC}}$ and $\boldsymbol{B}^{k}_{\mathrm{HNCO}}$ are the HSQC and HNCO intensities of the k-th signal, respectively. $\boldsymbol{B}^{k}_{\mathrm{HSQC}}$, $\boldsymbol{B}^{k}_{\mathrm{HNCO}}$, x_0^k, y_0^k, σ_x^k, and σ_y^k are fitting parameters. As described, at least one sample is 100% labeled in the SiCode labeling pattern for each amino acid. The amplitude of each signal is then estimated by

$$\hat{A}^{k}_{\mathrm{HSQC}}=\max\left(\boldsymbol{B}^{k}_{\mathrm{HSQC}}\right) \quad (5)$$

where $\hat{A}^{k}_{\mathrm{HSQC}}$ is the estimated HSQC amplitude of the k-th signal, assuming that all signals in the region are positive. The ^{15}N labeling ratios are estimated by

$$\hat{\boldsymbol{n}}_k=\boldsymbol{B}^{k}_{\mathrm{HSQC}}/\hat{A}^{k}_{\mathrm{HSQC}} \quad (6)$$

where $\hat{\boldsymbol{n}}_k$ is are the back-calculated SI-labeling ratios (hereafter named "SI-indices") of ^{15}N. Since HNCO intensities are proportional to not only the ^{13}C labeling ratios of residue $i-1$ but also the ^{15}N labeling ratios of residue i, they should be adjusted as follows:

$$\boldsymbol{B'}^{k}_{\mathrm{HNCO}}=\boldsymbol{B}^{k}_{\mathrm{HNCO}}\oslash\hat{\boldsymbol{n}}_k \quad (7)$$

where $\boldsymbol{B'}^{k}_{\mathrm{HNCO}}$ is the adjusted HNCO intensitiy of the k-th signal, and $\oslash$ denotes element-wise division. The HNCO amplitude and the ^{13}C labeling ratios are similarly estimated by

$$\hat{A'}^{k}_{\mathrm{HNCO}}=\max\left(\boldsymbol{B'}^{k}_{\mathrm{HNCO}}\right) \quad (8)$$

$$\hat{\boldsymbol{c}}_k=\boldsymbol{B'}^{k}_{\mathrm{HNCO}}/\hat{A'}^{k}_{\mathrm{HNCO}} \quad (9)$$

where $\hat{A'}^{k}_{\mathrm{HNCO}}$ is the estimated adjusted HNCO amplitude of the k-th signal, and $\hat{\boldsymbol{c}}_k$ is the ^{13}C SI-indices. SI indices can be converted to amino acids, using the given labeling pattern:

$$a_i^k=\underset{a}{\mathrm{argmin}}\,d(\hat{\boldsymbol{n}}_k,\boldsymbol{n}(a)) \quad (10)$$

$$a_{i-1}^k=\underset{a}{\mathrm{argmin}}\,d(\hat{\boldsymbol{c}}_k,\boldsymbol{c}(a)) \quad (11)$$

where a_i^k is the amino acid of the residue from which the k-th signal is derived, a_{i-1}^k is the amino acid of its preceeding (N-terminal side) residue, and d denotes the Euclidean distance. Since the codeword table shown in table 1 uses three ternary digits as a codeword (0 for 0%, 1 for 50%, 2 for 100% labeling of ^{13}C; 0 for 50%, 1 for 75%, 2 for 100% labeling of ^{15}N) and all of the 19 possible codewords are occupied, the simple conversion of the SI indices to the nearest ternary digits (i.e., for $\hat{\boldsymbol{c}}_k$, -0.25 to 0.25 is 0, 0.25 to 0.75 is 1, 0.75 to 1.25 is 2; for $\hat{\boldsymbol{n}}_k$, 0.375 to 0.625 is 0, 0.625 to 0.875 is 1, 0.875 to 1.125 is 2) provides the same result as equations (10) and (11). As the proline residue is not SI-labeled, a small $\hat{A'}^{k}_{\mathrm{HNCO}}$ value suggests that a_{i-1}^k is a proline.

This sequential decoding procedure has two problems. Firstly, it is difficult to find the global optimum of the peak fitting (equation (4)). We used the Levenberg-Marquardt method with the initial parameters determined manually [19]; however, the failure to set appropriate initial parameters leads to trapping in a local optimum. Secondly, since we perform the peak fitting (equation (4)) first and then interpret the resulting amplitude using the codeword table (equations (5) to (11)), peak fitting may fail in cases with close overlapping (as discussed in the Results section).

2.3. Improved decoding procedure with REMC

For a given amino acid set $M=(a_i^k,a_{i-1}^k;\ k=1\ldots K)$ and peak parameter set $\Theta=\left(A^{k}_{\mathrm{HSQC}},A^{k}_{\mathrm{HNCO}},x_0^k,y_0^k,\sigma_x^k,\sigma_y^k;\ k=1\ldots K\right)$, the residual error of each spectrum is calculated as:

$$\boldsymbol{R}_{\mathrm{HSQC}}(x,y;M,\Theta)=\boldsymbol{I}_{\mathrm{HSQC}}(x,y)-\sum_{k=1}^{K}A^{k}_{\mathrm{HSQC}}\,\boldsymbol{n}(a_i^k)\exp\left(-\frac{\left(x-x_0^k\right)^2}{2{\sigma_x^k}^2}-\frac{\left(y-y_0^k\right)^2}{2{\sigma_y^k}^2}\right) \quad (12)$$

$$\begin{aligned}\boldsymbol{R}_{\mathrm{HNCO}}(x,y;M,\Theta) &= \boldsymbol{I}_{\mathrm{HNCO}}(x,y) \\ &- \sum_{k=1}^{K} A^k_{\mathrm{HNCO}}\, \boldsymbol{n}(a_i^k) \circ \boldsymbol{c}(a_{i-1}^k) \exp\left(-\frac{\left(x-x_0^k\right)^2}{2{\sigma_x^k}^2} - \frac{\left(y-y_0^k\right)^2}{2{\sigma_y^k}^2}\right)\end{aligned} \tag{13}$$

where $\boldsymbol{R}_{\mathrm{HSQC}}$ and $\boldsymbol{R}_{\mathrm{HNCO}}$ are the residual errors of HSQC and HNCO, respectively, and A^k_{HSQC} and A^k_{HNCO} are the HSQC and HNCO amplitudes of the k-th signal, respectively. Assuming that the spectral noise is Gaussian distributed and its magnitude in each spectrum of each sample is known, the sum of the normalized squared residual error $E(M,\Theta)$ is defined as:

$$E(M,\Theta) = \frac{1}{2}\sum_{x,y}\left(\left|\boldsymbol{R}_{\mathrm{HSQC}}(x,y;M,\Theta) \oslash \boldsymbol{\sigma}_{\mathrm{HSQC}}\right|^2 + \left|\boldsymbol{R}_{\mathrm{HNCO}}(x,y;M,\Theta) \oslash \boldsymbol{\sigma}_{\mathrm{HNCO}}\right|^2\right) \tag{14}$$

where $\boldsymbol{\sigma}_{\mathrm{HSQC}}$ and $\boldsymbol{\sigma}_{\mathrm{HNCO}}$ are the standard deviations of the HSQC and HNCO noises, respectively. For Bayesian spectral deconvolution and model selection, regarding the amino acid set M as a model, the marginal likelihood, which is the posterior probability of the observed spectra with the given amino acid set M, is calculated as [20]:

$$L(M) = \int \exp\left(-E(M,\Theta)\right)\varphi(\Theta)\mathrm{d}\Theta \tag{15}$$

where $\varphi(\Theta)$ is the prior probability density of Θ, which in this report we assume has a uniform distribution. Since it is difficult to calculate $L(M)$ analytically, we numerically calculate this value with the REMC method, as described [20]. The amino-acid set M should be selected according to the value $-\log L(M)$, which is called the free energy [20]. For every Markov chain Monte Carlo (MCMC) step, a parameter set Θ is sampled and the corresponding $E(M,\Theta)$ is calculated. From these sampling results, we can find the best fit parameter Θ with the smallest $E(M,\Theta)$ value, the posterior distribution of Θ with the low temperature replica, and $L(M)$ from all replicas, as described [20]. The use of REMC avoids trapping in a local optimum.

If the amino acid sequence of the protein is not known, then the number of possible M is large; namely, $(19\times 20)^K$, since a_i^k is one of all 19 non-proline amino acids and a_{i-1}^k is one of all 20 amino acids. Excluding an impossible combination of amino acids, using the prior knowledge of the amino acid sequence, may improve the decoding correctness in challenging cases. To achieve this, M should comply with

$$\forall p, C(p,M) \le T(p) \tag{16}$$

where $p = (a_{i-1}, a_i)$ is an adjacent pair of amino acids, $C(p,M)$ is the number of amino-acid pairs p found in M, and $T(p)$ is the number of amino-acid pairs p found in the amino-acid sequence of the protein, assuming that one residue gives no more than one signal in a spectrum. Hereafter, we call this procedure "model-selection decoding".

As discussed, we can obtain $L(M)$ for all possible M by running REMC for each M; however, a combinatorial explosion of the number of possible M occurs as K increases. To find the best fit M and Θ or the posterior distribution of M and Θ, we can alternatively run REMC once, by sampling both the discrete variables M and the continuous variables Θ as parameters. In this report, we assume the uniform distribution of M. Each amino acid pair of M, as well as each parameter of Θ, is updated in MCMC step. This REMC run gives us the best fit M or posterior distribution of M, which can be used as a decoding result for SiCode. Hereafter, we call this procedure "model-optimization decoding".

3. Materials and methods

3.1. Protein preparation and NMR measurement

Three selectively labeled samples of the SH2 domain (residues 291 to 393) of the human BMX protein (UniProt: P51813), with an N-terminal 7-residue cloning artifact (GSSGSSG), were prepared according to the SI-labeling pattern shown in table 1, using an *Escherichia coli* based cell-free protein synthesis system [21-25] with metabolic inhibitors to suppress SI-labeling scrambles [26], as

described [19]. To evaluate the decoding procedures in various signal-to-noise ratios (SNRs), NMR spectra were acquired under three conditions: at a 0.35 mM concentration with 8 scans for HSQC and 32 scans for HNCO (hereafter, high-SNR conditions), at a 0.05 mM concentration with 64 scans for HSQC and 256 scans for HNCO (hereafter, medium-SNR conditions), and at a 0.05 mM concentration with 16 scans for HSQC and 64 scans for HNCO (hereafter, low-SNR conditions), as described [19].

Since it is important for SiCode to compensate for variances in concentrations among samples and/or other reasons that affect intensities, the spectra should be normalized using glycine residues [19]. In this report, prior to the REMC analyses, the spectra acquired for samples 1, 2, and 3 were divided by the values of 0.937, 0.911, and 1.000, respectively, in the high-SNR conditions, and 0.997, 0.896, and 0.946, respectively, in the medium- and low-SNR conditions.

3.2. Model-selection decoding

All possible amino-acid combinations M were generated from the amino-acid sequence of BMX SH2, according to equation (16). For each M, REMC was performed essentially as described [20]. The standard deviations of the spectral noise were evaluated using the region without signals shown in figure 1, for each spectrum and each sample. The spectral regions used for analysis were manually determined, as shown in figure 1. The upper limits of A^k_{HSQC} and A^k_{HNCO} were set to 1.2-fold of the maximum of each spectrum in the region, while their lower limits were set to 3-fold of the average of $\boldsymbol{\sigma}_{\mathrm{HSQC}}$ and $\boldsymbol{\sigma}_{\mathrm{HNCO}}$, respectively. The upper and lower limits of x_0^k and y_0^k were set to be the ends of the spectral region. The upper and lower limits of σ_x^k were 0.1 ppm and 0.001 ppm, respectively. The upper and lower limits of σ_y^k were 0.75 ppm and 0.01 ppm, respectively. The number of temperatures L was 96, and the inverse temperature β^l was

$$\beta^l = \begin{cases} 0, & (\text{if } l = 1) \\ r^{l-L}, & (\text{otherwise}) \end{cases} \tag{17}$$

where $r = 1.15$, $l = 1 \ldots L$. The MCMC steps were set to 5,000 for burn-in and 1,000 for sampling. In each MCMC step, the sequential updates of the parameters A^1_{HSQC}, A^1_{HNCO}, ..., A^K_{HSQC}, A^K_{HNCO}, x_0^1, y_0^1, ..., x_0^K, y_0^K, σ_x^1, σ_y^1, ..., σ_x^K, σ_y^K were performed with the Metropolis algorithm, followed by replica exchange between two adjacent temperatures, as described [20].

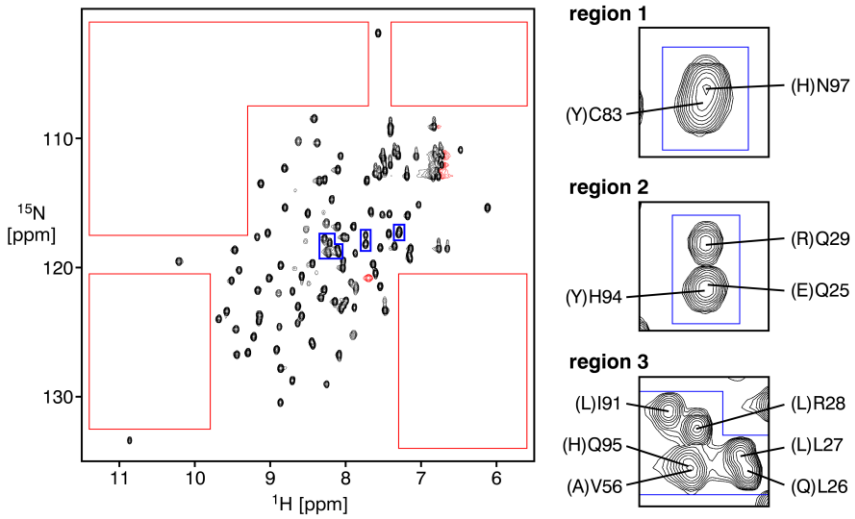

Figure 1. HSQC spectrum of sample 1 of BMX SH2 protein under the high-SNR conditions. Black and red contours represent positive and negative signals, respectively. Regions used for estimation of background noise are shown by red boxes. Regions used in this study for deconvolution of overlapped signals are shown by blue boxes, which are expanded in the right panel. Each peak label consists of a^k_{i-1} (in parentheses), a^k_i, and residue number i.

3.3. Model-optimization decoding

Simultaneous optimizations of both M and Θ were performed essentially as described in the above model-selection decoding, except for the following. In each MCMC step, the amino acid combinations $p_1, \ldots, p_K$ were updated in the sequence before the update of Θ, where the amino-acid pair $p_k = (a_{i-1}^k, a_i^k)$. If M violated equation (16) as a result of the update, this update was rejected, regardless of the change in the residual error. The number of temperatures L was 192 and $r = 1.1$. The number of MCMC steps were set to be 50,000 or 200,000 for burn-in, in the case of $K < 6$ or $K \geq 6$, respectively, and 10,000 for sampling.

4. Results

4.1. Sequential decoding

In this report, we evaluated the decoding procedures with overlapped signals appearing in the spectra. As shown in figure 1, region 1 contains two signals, (Y)C83 and (H)N97, with overlapping that makes the decoding difficult. However, even the sequential decoding successfully decodes these two signals under the high-SNR conditions (figure 2). Decoding the signals in region 2, which contains three signals, (E)Q25, (R)Q29, and (Y)H94, is a more difficult problem because two of them are very closely overlapped (figure 1). The sequential decoding failed to find the correct answer in the analysis of this region under the low-SNR conditions (figure 3).

As discussed in the Theory section, we first perform the peak fitting in the sequential decoding. The resulting amplitude of each signal is then converted to the amino acid information, using the predefined SI-labeling pattern. Therefore, a failure in the peak fitting step leads to the misinterpretation of the amino acid information. As shown in table 1, the SI-labeling ratio of ^{15}N is either 50%, 75%, or 100%; however, SI indices of ^{15}N that differ from these values were obtained; for example, 33.0% and 62.8% in the analysis of region 2 (figure 3). These values clearly indicate the failure of the peak fitting, as the signal deconvolution of closely overlapped signals is difficult. Moreover, one of the decoding results of three signals in region 2 (figure 3), (Y)I, is an amino acid pair that does not appear in the amino-acid sequence of BMX SH2, also indicating failure. Hence, using prior knowledge of the SI-labeling pattern and the amino-acid sequence in the peak fitting step may prevent such failures.

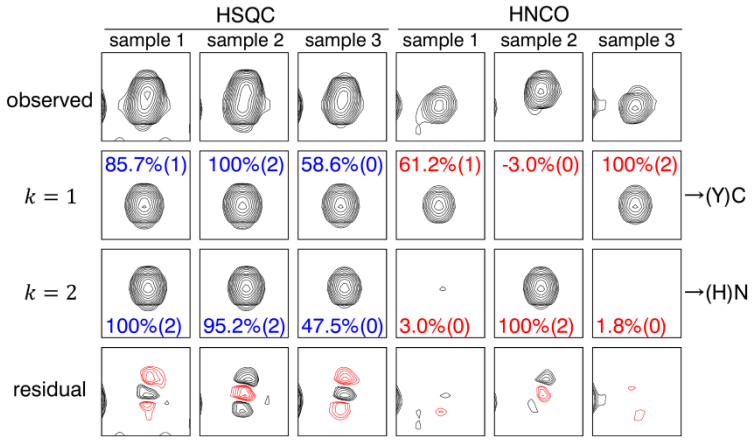

Figure 2. Sequential decoding of region 1 under high-SNR conditions. Observed spectrum ("observed"), deconvoluted signals ($k = 1$ and 2), and residual error ("residual") are shown. SI indices of ^{15}N ($\hat{\boldsymbol{n}}_k$) and ^{13}C ($\hat{\boldsymbol{c}}_k$) are shown with blue and red numbers, respectively. Each corresponding ternary digit is shown in parentheses immediately following the SI index. Decoding results are shown on the right of the spectra.

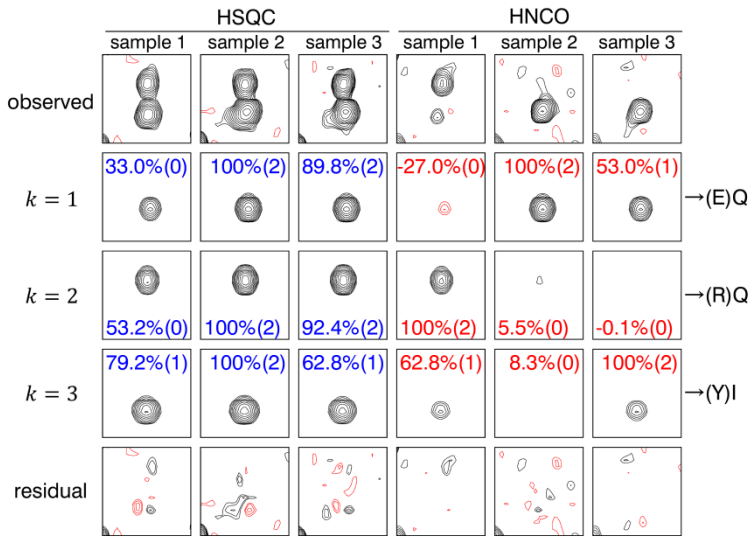

Figure 3. Sequential decoding of region 2 under low-SNR conditions. Spectra are shown as in figure 2.

4.2. Model-selection decoding

Using prior knowledge of the amino acid sequence is achieved by regarding decoding as a model selection problem. All of the possible models M; i.e., combinations of amino acid pairs, can be listed according to equation (16), provided the total signal number K is given. The sampling results of REMC for each model allow us to calculate the free energy, which shows the probability of each model. For each REMC run, the amino acids a_i^k and a_{i-1}^k are given, and as a consequence the SI-labeling ratios $\boldsymbol{n}(a_i^k)$ and $\boldsymbol{c}(a_{i-1}^k)$ are also given. Therefore, the model-selection decoding uses the SI-labeling pattern as prior knowledge for the peak fitting.

We applied this method to analyze region 1, under high-SNR conditions. The number of possible models is 90 or 4,021, for $K = 1$ or $K = 2$, respectively. We assumed that K is either 1 or 2; therefore, 4,111 models in total were tested. Numerical calculations of free energies by REMC for all of the models revealed that the model (H)N+(Y)C had the smallest free energy (figure 4), and was the correct model.

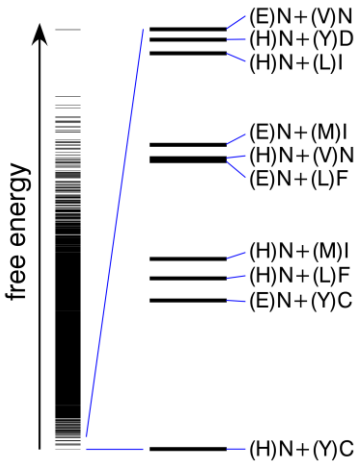

Figure 4. Free energies of each model (combination of amino acid pair). Signals in region 1 were analyzed under high-SNR conditions. The calculated free energy of each model is shown with a black line. All 4,111 models are shown on the left, while 10 models with the lowest free energies are expanded on the right.

4.3. Model-optimization decoding

In the previous section, the model M is given in each REMC run. In case of BMX SH2, the number of possible models is 90, 4,021, or 118,905, for $K = 1$, 2, or 3, respectively. In general, the number of possible models increases exponentially with K. Since model-selection decoding requires REMC runs for all of possible models, it is not feasible in case of $K \geq 3$. Alternatively, the model M and fitting parameters Θ can both be optimized in a single REMC run, to reduce the computation time. This method allows us to obtain the posterior probability distributions of M and Θ, as well as the best fitting parameters M and Θ.

We applied this method to decode the signals in region 2 under the low-SNR conditions, for which the sequential decoding failed to obtain the correct answer (figure 2). As shown in figure 5, the model-optimization decoding successfully obtained the correct answer.

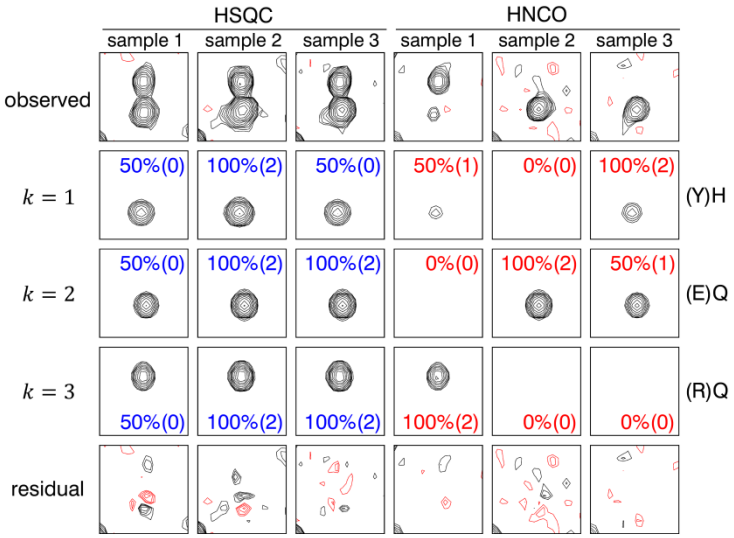

Figure 5. Model-optimization decoding of region 2 under low-SNR conditions. Observed spectrum ("observed"), generated spectra with the best fit parameters M and Θ for each signal ($k = 1$ to 3), and residual error ("residual") are shown. The amino acid pair of each signal is shown on the right, and the corresponding ^{15}N and ^{13}C SI labeling ratios are shown with blue and red numbers, respectively. Each corresponding ternary digit is shown in parentheses immediately following the SI labeling ratio.

The frequency of the appearance of M at the lowest temperature replica corresponds to the estimated posterior probability distribution of M. By model-optimization decoding of the signals in region 1 under the low-SNR conditions, only one model, (H)N+Y(C), was sampled, indicating that the posterior probabilities of the other models were too low to be detected. In contrast, by the same analysis without the prior knowledge of the amino acid sequence; i.e., without limiting models M by equation (16), four models, (H)N+(Y)C, (H)N+(Y)F, (H)N+(Y)N, and (H)N+(M)F, were sampled. Their sampled frequencies; namely, their estimated posterior probabilities, were 51.0%, 24.9%, 24.1%, and 0.02%, respectively. As amino acid pairs (Y)F, (Y)N, and (M)F do not appear in the amino-acid sequence, the models with these pairs are eliminated in the analysis with equation (16). These results show that the prior knowledge of not only the SI-labeling pattern but also the amino-acid sequence certainly facilitates correct decoding.

For the above analyses, the number of signals K is assumed to be given. In the real situation, K is not known, especially in closely overlapped cases. We analyzed region 3 (figure 1), which contains 6 signals, under the medium-SNR conditions with $K = 5, 6, \text{or } 7$ (figure 6). The amino acids were correctly decoded in the $K = 6$ case. However, in the case of $K = 5$, the closely overlapped (A)V56

and (H)Q95 were considered to be a single broad signal (I)T ($k = 4$), and in the case of $K = 7$, an additional very broad signal ($k = 4$) appeared, while the other 6 signals revealed the correct answers. The appearance of the broad signal is due to the wider range of the line width parameter than its actual range in our REMC settings, and thus may be suppressed by narrowing the range. However, in some cases, the appearance of a broad signal may help us to become aware that K is incorrect.

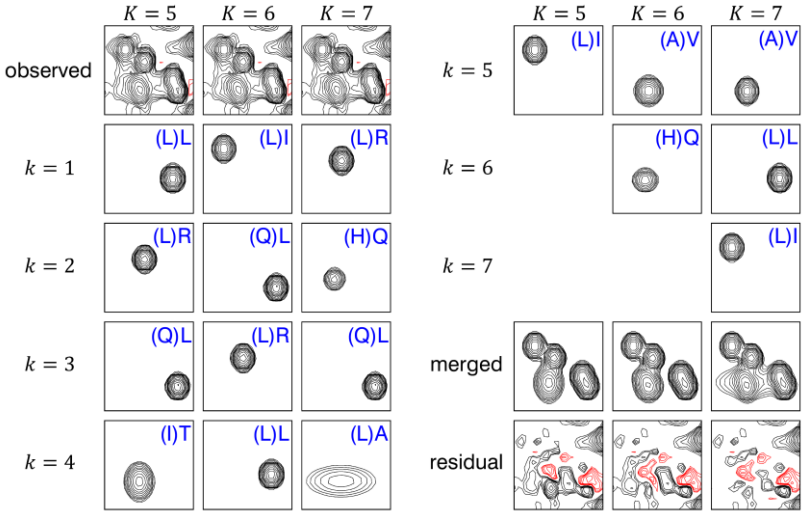

Figure 6. Model-optimization decoding of region 3 under the medium SNR-conditions with various given numbers of signals. HSQC spectra of sample 1 of observed spectra ("observed"), generated spectra with the best fit parameters M and Θ for each signal ($k =$ 1 to 7), sum of the generated spectra ("merged"), and residual error ("residual") are shown. The amino acid pair of each signal is shown with blue letters.

5. Discussion

In this report, we showed that spectral deconvolution with prior knowledge, such as the SI-labeling ratios and the amino acid sequence, improves the decoding reliability in the case of signal overlapping, which interferes with obtaining amino-acid information in SiCode and other CSLs. The previously reported [19] sequential decoding was a kind of heuristics, in which the peak fitting result was assumed to be correct in the following steps. Such heuristics worked well in relatively easier cases. However, in cases with a low SNR and/or close overlapping, the intensities of each signal may contain errors due to inaccurate signal deconvolution. In such cases, using prior knowledge in the early stage of analysis improves the result.

We previously proposed different SI-labeling patterns optimized for various numbers of samples and amino acids [19]. Some of them are more complicated than that shown in table 1, and use more than three SI-labeling levels. The proposed methods; namely, the model-selection decoding and the model-optimization decoding, can also deal with such patterns. Moreover, the NMR spectra of other CSLs can be analyzed with the proposed methods, to improve the tolerance to signal overlapping.

The proposed methods are based on assumptions, such as the correctness of the SI-labeling ratio of the prepared protein samples, the same chemical shifts and line widths among spectra and samples, the two-dimensional Gaussian line shapes, the Gaussian background noise, the absence of minor conformers, and the prior uniform distribution of parameters. Some of these assumptions are actually incorrect to some extent, and result in residual errors of peak fitting. Careful modifications of these assumptions, according to the real phenomena, may further improve the correctness and reliability of the methods. However, as discussed above, even with these assumptions, the proposed methods certainly improved the decoding accuracy.

As shown in equation (16), the prior knowledge of the amino-acid sequence is used as the upper limits of the frequency of amino acid pairs. Using it as the lower limits may further improve the decoding, if the number of signals derived from minor conformers is limited. However, to use it as the lower limits, the full region of the spectra must be analyzed simultaneously. The computation times in this study were 3 min 26 sec for region 1 ($K = 2$), 3 min 15 sec for region 2 ($K = 3$), and 62 min 12 sec for region 3 ($K = 6$), with a computer equipped with an Intel Core i7 4930K CPU (6 cores, 3.4 GHz). The computation time depends largely on the number of signals (K), the number of data points in the region, and the number of MCMC steps. Although the number of MCMC steps in this study is excessive, and thus can be reduced, applying the method to the full region of the spectra ($K > 100$) is not feasible so far, and therefore further improvement is expected.

For small proteins, most of signals are not overlapped, and thus they can be analyzed with the previous sequential decoding. However, close signal overlapping is often observed in large proteins, helical proteins, and/or intrinsically disordered proteins. In addition, a low SNR, caused by low solubility, high molecular weight of the protein and/or challenging situations, such as in-cell NMR, makes the deconvolution difficult. Failures in the signal deconvolution and the amino acid determination lead to mistakes in the assignment, which impedes the NMR analysis of the structures and dynamics of the protein. Spectral analyses using prior knowledge with the help of computation may promote NMR research of difficult proteins in challenging situations.

Acknowledgements
We thank the lab members at RIKEN for their help in sample preparation and data acquisition. We also thank Professors Toshiyuki Tanaka, Shiro Ikeda, and Koji Hukushima for valuable discussions. This work was supported in part by Grants-in-Aid for Scientific Research on Innovative Areas (Grant Nos. 25120003 and 25120009) and a Grant-in-Aid for Challenging Exploratory Research (Grant No. 26650027) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan and the Japan Society for the Promotion of Science (JSPS).

References

[1] Grzesiek S and Bax A 1993 Amino acid type determination in the sequential assignment procedure of uniformly $^{13}C/^{15}N$-enriched proteins *J. Biomol. NMR* **3** 185-204
[2] Bertelsen E B, Chang L, Gestwicki J E and Zuiderweg E R 2009 Solution conformation of wild-type *E. coli* Hsp70 (DnaK) chaperone complexed with ADP and substrate *Proc. Natl. Acad. Sci. U. S. A.* **106** 8471-6
[3] Cervantes C F, Handley L D, Sue S C, Dyson H J and Komives E A 2013 Long-range effects and functional consequences of stabilizing mutations in the ankyrin repeat domain of IκBα *J. Mol. Biol.* **425** 902-13
[4] Hembram D S, Haremaki T, Hamatsu J, Inoue J, Kamoshida H, Ikeya T, Mishima M, Mikawa T, Hayashi N, Shirakawa M and Ito Y 2013 An in-cell NMR study of monitoring stress-induced increase of cytosolic Ca^{2+} concentration in HeLa cells *Biochem. Biophys. Res. Commun.* **438** 653-9
[5] Kainosho M and Tsuji T 1982 Assignment of the three methionyl carbonyl carbon resonances in *Streptomyces* subtilisin inhibitor by a carbon-13 and nitrogen-15 double-labeling technique. A new strategy for structural studies of proteins in solution *Biochemistry* **21** 6273-9
[6] Yabuki T, Kigawa T, Dohmae N, Takio K, Terada T, Ito Y, Laue E D, Cooper J A, Kainosho M and Yokoyama S 1998 Dual amino acid-selective and site-directed stable-isotope labeling of the human c-Ha-Ras protein by cell-free synthesis *J. Biomol. NMR* **11** 295-306
[7] Hefke F, Bagaria A, Reckel S, Ullrich S J, Dötsch V, Glaubitz C and Güntert P 2011 Optimization of amino acid type-specific ^{13}C and ^{15}N labeling for the backbone assignment of membrane proteins by solution- and solid-state NMR with the UPLABEL algorithm *J. Biomol. NMR* **49** 75-84
[8] Jaipuria G, Krishnarjuna B, Mondal S, Dubey A and Atreya H S 2012 Amino acid selective

labeling and unlabeling for protein resonance assignments *Adv. Exp. Med. Biol.* **992** 95-118
[9] Krishnarjuna B, Jaipuria G, Thakur A, D'Silva P and Atreya H S 2011 Amino acid selective unlabeling for sequence specific resonance assignments in proteins *J. Biomol. NMR* **49** 39-51
[10] Löhr F, Reckel S, Karbyshev M, Connolly P J, Abdul-Manan N, Bernhard F, Moore J M and Dötsch V 2012 Combinatorial triple-selective labeling as a tool to assist membrane protein backbone resonance assignment *J. Biomol. NMR* **52** 197-210
[11] Maslennikov I and Choe S 2013 Advances in NMR structures of integral membrane proteins *Curr. Opin. Struct. Biol.* **23** 555-62
[12] Maslennikov I, Klammt C, Hwang E, Kefala G, Okamura M, Esquivies L, Mörs K, Glaubitz C, Kwiatkowski W, Jeon Y H and Choe S 2010 Membrane domain structures of three classes of histidine kinase receptors by cell-free expression and rapid NMR analysis *Proc. Natl. Acad. Sci. U. S. A.* **107** 10902-7
[13] Parker M J, Aulton-Jones M, Hounslow A M and Craven C J 2004 A combinatorial selective labeling method for the assignment of backbone amide NMR resonances *J. Am. Chem. Soc.* **126** 5020-1
[14] Shi J, Pelton J G, Cho H S and Wemmer D E 2004 Protein signal assignments using specific labeling and cell-free synthesis *J. Biomol. NMR* **28** 235-47
[15] Sobhanifar S, Reckel S, Junge F, Schwarz D, Kai L, Karbyshev M, Löhr F, Bernhard F and Dötsch V 2010 Cell-free expression and stable isotope labelling strategies for membrane proteins *J. Biomol. NMR* **46** 33-43
[16] Staunton D, Schlinkert R, Zanetti G, Colebrook S A and Campbell I D 2006 Cell-free expression and selective isotope labelling in protein NMR *Magn. Reson. Chem.* **44** S2-9
[17] Trbovic N, Klammt C, Koglin A, Löhr F, Bernhard F and Dötsch V 2005 Efficient strategy for the rapid backbone assignment of membrane proteins *J. Am. Chem. Soc.* **127** 13504-5
[18] Wu P S, Ozawa K, Jergic S, Su X C, Dixon N E and Otting G 2006 Amino-acid type identification in ^{15}N-HSQC spectra by combinatorial selective ^{15}N-labelling *J. Biomol. NMR* **34** 13-21
[19] Kasai T, Koshiba S, Yokoyama J and Kigawa T 2015 Stable isotope labeling strategy based on coding theory *J. Biomol. NMR* **63** 213-21
[20] Nagata K, Sugita S and Okada M 2012 Bayesian spectral deconvolution with the exchange Monte Carlo method *Neural Netw* **28** 82-9
[21] Kigawa T, Yabuki T, Matsuda N, Matsuda T, Nakajima R, Tanaka A and Yokoyama S 2004 Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression *J. Struct. Funct. Genomics* **5** 63-8
[22] Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T and Yokoyama S 1999 Cell-free production and stable-isotope labeling of milligram quantities of proteins *FEBS Lett.* **442** 15-9
[23] Matsuda T, Koshiba S, Tochio N, Seki E, Iwasaki N, Yabuki T, Inoue M, Yokoyama S and Kigawa T 2007 Improving cell-free protein synthesis for stable-isotope labeling *J. Biomol. NMR* **37** 225-9
[24] Seki E, Matsuda N, Yokoyama S and Kigawa T 2008 Cell-free protein synthesis system from *Escherichia coli* cells cultured at decreased temperatures improves productivity by decreasing DNA template degradation *Anal. Biochem.* **377** 156-61
[25] Yabuki T, Motoda Y, Hanada K, Nunokawa E, Saito M, Seki E, Inoue M, Kigawa T and Yokoyama S 2007 A robust two-step PCR method of template DNA production for high-throughput cell-free protein synthesis *J. Struct. Funct. Genomics* **8** 173-91
[26] Yokoyama J, Matsuda T, Koshiba S, Tochio N and Kigawa T 2011 A practical method for cell-free protein synthesis to avoid stable isotope scrambling and dilution *Anal. Biochem.* **411** 223-9