J Biomol NMR (2015) 63:213–221
DOI 10.1007/s10858-015-9978-8

ARTICLE

Stable isotope labeling strategy based on coding theory

Takuma Kasai[1,2] · Seizo Koshiba[1,3] · Jun Yokoyama[1,4,5] · Takanori Kigawa[1,2,4,6]

Received: 10 June 2015 / Accepted: 12 August 2015 / Published online: 21 August 2015

Abstract We describe a strategy for stable isotope-aided protein nuclear magnetic resonance (NMR) analysis, called stable isotope encoding. The basic idea of this strategy is that amino-acid selective labeling can be considered as "encoding and decoding" processes, in which the information of amino acid type is encoded by the stable isotope labeling ratio of the corresponding residue and it is decoded by analyzing NMR spectra. According to the idea, the strategy can diminish the required number of labelled samples by increasing information content per sample, enabling discrimination of 19 kinds of non-proline amino acids with only three labeled samples. The idea also enables this strategy to combine with information technologies, such as error detection by check digit, to improve the robustness of analyses with low quality data. Stable isotope encoding will facilitate NMR analyses of proteins under non-ideal conditions, such as those in large complex systems, with low-solubility, and in living cells.

Keywords Amino-acid selective stable isotope labeling · Cell-free protein synthesis · Coding theory · Combinatorial selective labeling · Signal assignment

Introduction

Stable-isotope (SI) labeling of proteins is an essential technique to investigate three-dimensional structures, ligand interactions or dynamics of proteins by nuclear magnetic resonance (NMR) spectroscopy. The assignment of the main-chain signals, which is generally the first step in these analyses, is usually achieved by a sequential assignment method based on a combination of triple resonance experiments on proteins uniformly labeled with ^{15}N and ^{13}C (Grzesiek and Bax 1993). Amino-acid selective SI labeling (AASIL) helps to discriminate the amino-acid type of each signal, independently of the triple resonance experiment-based sequential assignment. Therefore, it is especially useful for the signal assignment of difficult targets, such as large complex systems (Bertelsen et al. 2009), low-solubility proteins (Cervantes et al. 2013), and proteins in living cells (Hembram et al. 2013). The dual selective labeling method, which utilizes both amide nitrogen and carbonyl carbon labeling, narrows down the assignment possibilities even further (Kainosho and Tsuji 1982), and as a consequence leads to the assignment, without the need for triple resonance experiments, of amino-acid pairs

Electronic supplementary material The online version of this article (doi:10.1007/s10858-015-9978-8) contains supplementary material, which is available to authorized users.

✉ Takanori Kigawa
kigawa@riken.jp

[1] Laboratory for Biomolecular Structure and Dynamics, RIKEN Quantitative Biology Center (QBiC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[2] JST CREST, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[3] Tohoku Medical Megabank Organization, Tohoku University, 2-1, Seiryo-cho, Aoba-ku, Sendai, Miyagi 980-8573, Japan

[4] Cell-Free Technology Application Laboratory, RIKEN Innovation Center (RInC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[5] SI Innovation Center, Taiyo Nippon Sanso Corporation, 2008-2 Wada, Tama-shi, Tokyo 206-0001, Japan

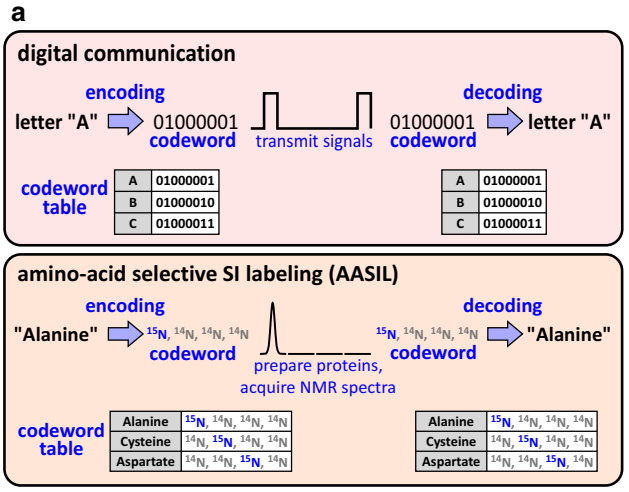
[6] Department of Computational Intelligence and Systems Science, Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology, 4259 Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa 226-8503, Japan

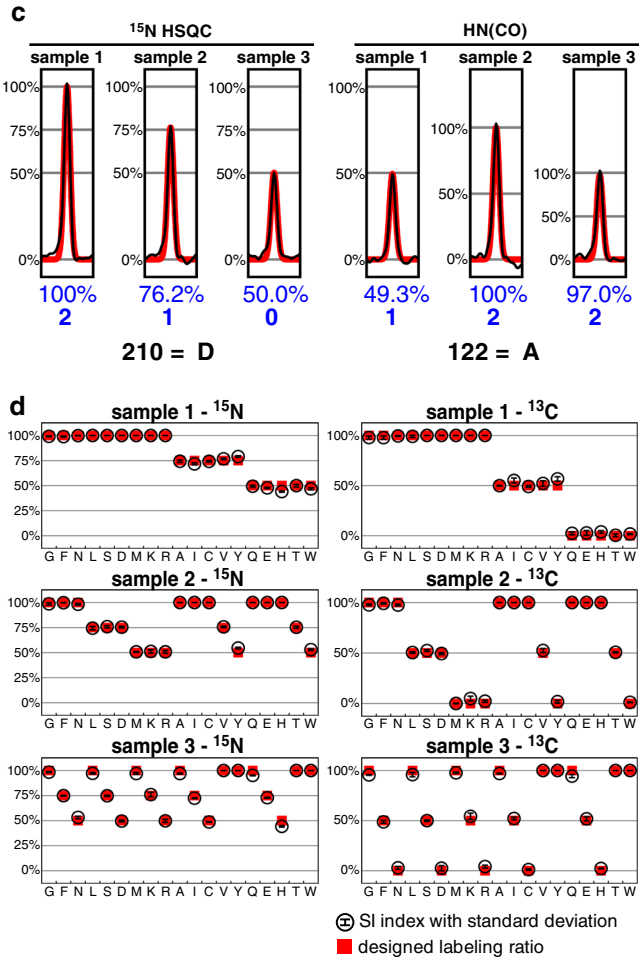
occurring only once in the sequence. However, for the discrimination of all amino-acids, these simple AASIL schemes require a large number of samples, which are typically the same as the number of amino acids (19 for nitrogen or 20 for carbon). For this reason, various combinatorial selective labeling (CSL) schemes (Parker et al. 2004; Shi et al. 2004; Trbovic et al. 2005; Staunton et al. 2006; Wu et al. 2006; Maslennikov et al. 2010; Sobhanifar et al. 2010; Hefke et al. 2011; Krishnarjuna et al. 2011; Jaipuria et al. 2012; Löhr et al. 2012; Maslennikov and Choe 2013) were developed to reduce required number of samples, by representing amino acids as combination of SI labeled samples rather than simply assigning one amino acid to one SI labeled sample. For example, a CSL scheme developed by Parker et al. (2004), which is based on the dual selective approach, can discriminate 16 amino-acids with one uniformly ^{13}C and ^{15}N-labeled reference and four selectively (100 or 0 % for ^{13}C and 100 or 50 % for ^{15}N, respectively) labeled samples. The use of labeling ratio of 100 or 50 % for ^{15}N, rather than that of 100 or 0 %, ensures obtaining ^{13}C labeling information from HN(CO) spectrum, irrespective of ^{15}N labeling ratio. Otting and colleagues reported simpler CSL scheme using five samples (Wu et al. 2006), based on the single selective ^{15}N-labeling approach, in which spectral overlaps were diminished by labeling one amino acid with high occurrence and at most three ones with low occurrence in each sample. Dötche and colleagues developed focused CSL (Trbovic et al. 2005; Sobhanifar et al. 2010), in which 6 or 7 amino acids frequently appearing in transmembrane regions of membrane proteins were labeled with ^{15}N or 1-^{13}C. They further improved the CSL to discriminate up to 20 amino-acids with a number of samples labeled with ^{15}N and/or ^{13}C by using dual selective approach (Hefke et al. 2011), or to discriminate 12 amino-acids with only three samples by introducing triple selective approach (Löhr et al. 2012), in which the samples were labeled with the combination of ^{15}N, 1-^{13}C, and ^{13}C/^{15}N. Choe and colleagues also improved membrane protein-focused CSL (Maslennikov et al. 2010; Maslennikov and Choe 2013) to discriminate up to 19 amino-acids except for glutamate with six samples simply labeled with ^{13}C and/or ^{15}N. A couple of computational methods for designing labeling patterns for CSL were employed (Maslennikov et al. 2010; Hefke et al. 2011; Maslennikov and Choe 2013) in order to maximize assignable residues by using dual selective approaches.

We realized that AASIL can be considered as an "encoding-and-decoding" process. Digital communication frequently involves the "encoding-and-decoding" process, in which a sender converts a letter to a code word, according to a predefined table associating each letter with a codeword (Fig. 1a). The converted codeword, consisting of binary digits, is transmitted through a communication channel, and then the receiver converts it back to the letter according to the table. Similarly, in the AASIL process, if the SI labeling pattern is regarded as a codeword table, the information of the amino acid type of each residue is converted to a codeword according to the table, and then it is retrieved from the observed NMR spectra according to the table. Based on this consideration, we propose a novel SI-labeling strategy based on coding theory, which we call stable isotope encoding (SiCode). Using this strategy, here we report a labeling scheme to discriminate all of 19 non-proline amino-acids with as few as three selectively labeled samples and a method to further improve noise-tolerance with error detection system.

Results

From the point of view of SiCode, the simplest AASIL scheme, in which 19 (for nitrogen) or 20 (for carbon) samples labeled with only a single amino acid are prepared, can be considered as a system in which each amino-acid is assigned to a specific 19 or 20-digit binary codeword with only a single '1' digit (Fig. S1a). One CSL scheme proposed by Parker et al. (2004), which utilizes the combination of one uniformly labeled sample and four selectively labeled samples in order to discriminate 16 amino acids, can be considered as a system in which the information is assigned to a 4-digit binary codeword (Fig. S1b). As long as binary digits are used as codewords, each sample can contain an information amount of one bit, thus limiting the number of discriminable amino acids to 2^n, where n is the number of selectively labeled samples. In AASIL, it is better to minimize the number of labeled samples, in terms of costs and sample preparation workload, as well as NMR machine time. Based on the SiCode concept, such minimization can be achieved by increasing the information content per sample by using three or more discrete SI-labeling levels, while the abovementioned CSL schemes utilize no more than two levels. As the simplest case of this idea, we have designed a novel scheme to use ternary digits as codewords, and an example of a codeword table based on a dual selective approach is shown in Fig. 1b (see "Materials and methods" section for details). In this scheme, the ternary digits, "0", "1", and "2", are represented by SI-labeling levels of 50, 75, and 100 % (for ^{15}N) or 0, 50, and 100 % (for ^{13}C), respectively. Moreover, by using only the codewords with at least one "2", the sample with the largest intensity for each signal can be used as a fully-labeled reference. The number of assignable codewords based on this scheme is 19, which is the exact number required for representing non-proline amino acids. Thus, we can discriminate 19 kinds of non-proline amino acids with only three labeled samples, by omitting the





b

amino acid	sample 1 ^{13}C	sample 1 ^{15}N	sample 2 ^{13}C	sample 2 ^{15}N	sample 3 ^{13}C	sample 3 ^{15}N	corresponding codeword
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

100% ^{15}N 75% ^{15}N 50% ^{15}N
100% ^{13}C 50% ^{13}C 0% ^{13}C

Fig. 1 Concept and application of the SiCode strategy. **a** Consideration of AASIL as an "encoding-and-decoding" process, by analogy to digital communication. See text for details. **b** One of the labeling patterns based on the SiCode strategy. For each amino acid and each sample, the ^{13}C and ^{15}N labeling ratios are shown as percentages. The corresponding 3-digit ternary codewords are indicated in the *rightmost column*. **c** Cross peaks of residue D73 of the Smoothelin CH domain. Raw data are shown with *black lines*, while fitted Gaussian functions are shown with *red lines*. The SI indices and the corresponding ternary digits are shown in *blue letters*. Codewords and judged amino acids are shown in *black bold letters*. **d** SI index for each amino acid type of all 89 isolated main-chain signals of the Smoothelin CH domain. *Open black circles* indicate the averaged SI index, accompanied with its standard deviation as an error bar. *Red squares* indicate the designed labeling ratio

additional uniformly labeled reference sample used in the abovementioned CSL (Parker et al. 2004). As a proof of concept, we have applied this scheme to the 116-amino-acid CH domain of Smoothelin protein (BMRB ID: 11572), as described in detail in the Materials and Methods and Supplementary Information. We used an *Escherichia coli*-based cell-free protein synthesis system (Kigawa et al. 1999, 2004; Matsuda et al. 2007; Seki et al. 2008) supplemented with metabolic inhibitors (Yokoyama et al. 2011) in order to achieve the accurate SI-labeling ratios we designed (Fig. 1b) for preparing the three kinds of labeled samples. A pair of 2D ^{15}N-HSQC and 2D HN(CO) spectra were acquired for each of the three samples so that six spectra in total were obtained (full spectra are shown in Fig. S2), and accurate signal intensities on each spectrum were obtained by fitting the signal to a two-dimensional Gaussian function. The ^{15}N-labeling ratios of the corresponding residue (residue i) were calculated from the HSQC signal intensities, and the ^{13}C-labeling ratios of the preceding residue (residue i − 1) were calculated similarly, using both the HSQC and HNCO signal intensities. These back-calculated SI labeling ratios are referred to as "SI indices".

Figure 1c shows the cross peaks of six spectra of residue D73, which is preceded by residue A72. The SI indices of ^{15}N of the residue i were 100, 76.2, and 50.0 %, respectively, which correspond to the codeword "210", indicating that this signal is from an aspartate residue. The SI

indices of ^{13}C of the residue i − 1 were 49.3, 100, and 97.0 %, respectively, which correspond to the codeword "122", revealing that the preceding residue is alanine. To investigate the decoding performance of SiCode, we analyzed 89 isolated (i.e. non-overlapping) main-chain signals out of 111 observable (non-proline) main-chain signals. The amino-acid types of all 89 isolated main-chain signals were correctly discriminated except for two preceding proline residues, as they are not SI-labeled in this scheme (see more examples in Fig. S4). The SI indices of these signals are accurate and precise enough to distinguish among the three levels, thus demonstrating that our strategy can be performed with sufficient accuracy and precision (Fig. 1d, see Supplemental Note S1 for the influence of the signal-to-noise ratio on the SI indices).

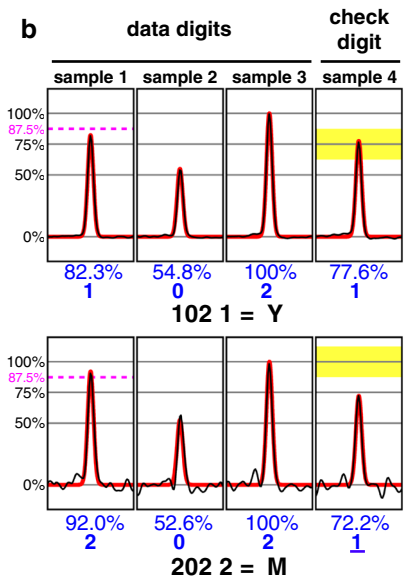
Difficult targets under more challenging conditions, such as higher molecular size, lower solubility, lower signal intensity, and/or limited NMR machine time, usually suffer from poor spectrum quality, which would contribute to errors in decoding. The robustness of SiCode can be further improved by applying information technologies, especially under these severe conditions. Check digit is a widely used error detection technique enabling reliable data transmission over an unreliable data path. We have designed a more robust scheme by implementing check digit (Fig. 2a), where the 1st–3rd samples act as data digits, as described above, while the 4th sample acts as check digit for error detection (see Supplementary Information for details). We have applied this scheme to the 110-amino-acid SH2 domain of the BMX protein (BMRB ID: 11573, full spectra are shown in Fig. S3), and the signals of residue Y77 are shown in Fig. 2b. Based on the high signal-to-noise ratio spectra, the combination of the codeword "102" and the check digit "1"; i.e., tyrosine, was deduced from the ^{15}N SI indices of the four samples, as expected (upper panel of Fig. 2b). In contrast, based on the spectra with nearly 5-fold lower signal-to-noise ratios (lower panel of Fig. 2b), the calculated check digit "2" from the observed codeword "202" did not coincide with the observed check digit "1", because the ^{15}N SI index of the first sample (92.0 %; i.e., digit "2") mistakenly exceeded the threshold of digit "1" (87.5 %). This result demonstrated that the robustness was certainly improved by introducing the error detection mechanism using the check digit (see Supplementary Note S2 for further discussion).

In the case of the BMX SH2 domain, 7 suspicious overlapped pairs of signals were identified by visual inspection, and 6 of them were detected by the error detection (Fig. S5). By fitting each overlapped pair to two two-dimensional Gaussian functions, 5 pairs of main-chain

a

amino acid	code-word	check digit	sample 4 ^{13}C	sample 4 ^{15}N
G	222	1	50	75
F	221	0	0	50
N	220	2	100	100
L	212	0	0	50
S	211	2	100	100
D	210	1	50	75
M	202	2	100	100
K	201	1	50	75
R	200	0	0	50
A	122	0	0	50
I	121	2	100	100
C	120	1	50	75
V	112	2	100	100
Y	102	1	50	75
Q	022	2	100	100
E	021	1	50	75
H	020	0	0	50
T	012	1	50	75
W	002	0	0	0

100% ^{15}N 75% ^{15}N 50% ^{15}N
100% ^{13}C 50% ^{13}C 0% ^{13}C



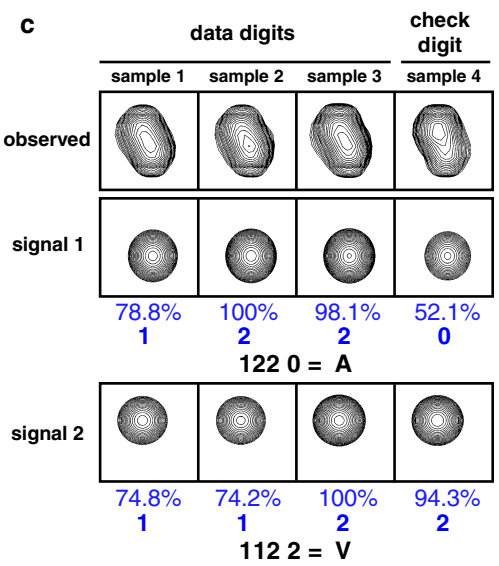


Fig. 2 Error detection and peak deconvolution with the SiCode strategy. **a** Expanded labeling pattern with check digits calculated with the checksum algorithm, based on the pattern shown in Fig. 1b. The corresponding SI-labeling ratios for the 4th sample are shown. **b** Cross peaks of residue Y77 of the BMX SH2 domain, representing data digits (samples 1–3) and check digit (sample 4). All four ^{15}N-HSQC spectra are shown. The high signal-to-noise ratio data obtained with 0.35 mM protein with 8 scans (*upper panel*) and the low signal-to-noise ratio data obtained with 0.05 mM protein with 16 scans (*lower panel*) are shown. The estimated peak height of the check digit spectrum from the data digits is shown in the *yellow area*. SI indices and converted ternary digits are shown in *blue letters*. The *underlined check digit* (i.e., sample 4) indicates a detected error. The codeword, the proper check digits and the judged amino acid are shown in *black bold letters*. **c** Deconvolution for discriminating overlapped peaks for residues A35 and V109 of the BMX SH2 domain. The observed spectra are shown in the *upper panel*, while the two Gaussian functions fitted to the observed spectra are separately shown in the *two lower panels*. SI indices, codewords, and judged amino acids are shown similarly

signals and 1 pair of side-chain signals were correctly discriminated (Fig. S5). For example, the overlapped pair signals of residues A35 and V109, shown in Fig. 2c, were successfully discriminated as two codeword and check digit pairs, "122 and 0" and "112 and 2", corresponding to alanine and valine, respectively. These results indicate that signal overlapping, which can impair correct discrimination in the conventional CSL, was managed by the peak deconvolution implemented as Gaussian peak fitting in the decoding process of SiCode.

Discussion

As described above, conventional CSL schemes generally uses two SI-labeling levels, enabling that 1 bit information is contained in each labeled sample (Parker et al. 2004; Shi et al. 2004; Trbovic et al. 2005; Staunton et al. 2006; Wu et al. 2006; Maslennikov et al. 2010; Sobhanifar et al. 2010; Hefke et al. 2011; Krishnarjuna et al. 2011; Jaipuria et al. 2012; Maslennikov and Choe 2013). In the previously mentioned triple selective CSL approach (Löhr et al. 2012), three SI-labeling types can be discriminated for both residues i and $i-1$: unlabeled, ^{15}N-labeled, or 2-^{13}C/^{15}N-labeled for the residue i and unlabeled, 1-^{13}C-labeled, or 1,2-^{13}C-labeled for the residue $i-1$, respectively, being considered that 1 trit (ternary digit) information is contained in each sample. In this report, we demonstrated a version of SiCode, in which each labeled sample contains 1 trit information by using three SI-labeling levels rather than by increasing number of labeling types. Our approach can discriminate nearly all amino-acids by using simpler combination of ^{15}N and ^{13}C/^{15}N-lablings, and introduce additional information like check digits for robust data analysis.

More complicated labeling patterns than that using three labeling levels can be easily achieved by using the cell-free protein synthesis system without SI scrambling (Yokoyama et al. 2011). Assuming that thermal noise of the observed data is the main reason for distorting the SI index, the labeling pattern should be designed in order to maximize minimum Euclidean distance between amino acids to achieve the best noise tolerance (see Supplementary Note S3 for the detailed discussion). Based on this strategy, we can easily design labeling patterns for the given number of samples and the given number of amino-acids, such as the pattern to discriminate 20 amino acids including proline. Since noise tolerance in discrimination between two specific amino acids depends on their information distance, specialized labeling patterns would be useful in some cases. For example, in the sequential assignment, amino acid pairs with similar Cα and Cβ chemical shifts could be easily discriminated with the help of SiCode specially designed so that such amino acid pairs have long information distance. When SI scrambling in the protein expression system is not strictly suppressed, the distance around the scrambling-prone amino acids should be increased. As mentioned above, noise tolerance of the selective labeling method can be evaluated based on the information distance. SiCode is the noise-tolerant method compared to the other selective labeling methods under the given total measurement time (see Supplementary Note S3 for the detail).

One of the major motivations to use AASIL is simplifying NMR spectrum by reducing the number of signals. Signal overlapping usually impede amino acid discrimination especially in CSL, therefore, the number of labeled amino-acids is reduced according to its occurrence in some CSL (Trbovic et al. 2005; Wu et al. 2006; Sobhanifar et al. 2010; Löhr et al. 2012). In the present study, we labeled all of the non-proline 19 amino acids, however, quantitative peak fitting used for decoding information in SiCode solved the signal overlapping issue as demonstrated. In addition to noise-tolerance based on coding theory, this feature will be especially crucial for analyzing difficult targets.

Almost all of the CSL studies, including this work, have so far used the cell-free synthesis system for protein expression in order to achieve the accurate SI-labeling by avoiding SI scrambling and dilution. SiCode can be performed using the protein expression system with manageable level of SI scrambling (see Sample Preparation section in the Supplementary Information). In addition, customized labeling pattern suitable for specific expression system could be designed by evaluating its SI scrambling profile based on information distance between amino acids (see Supplementary Note S3). Therefore, SiCode could also be achieved by in vivo expression system, for example, by the combination of the single protein production system (Suzuki et al. 2007; Schneider et al. 2010) and amino-acid auxotroph *E. coli* strains.

SiCode introduces a new concept into AASIL, by enabling the combination with information techniques that have rarely used in NMR field, such as detection of errors in signal intensities. From the standpoint of information science, its performance will be further improved, for instance, by optimizing the labeling pattern according to the amino acid content or sequence of the target, or by increasing the number of labeling levels or samples to expand the codeword's space, which will enable the implementation of redundant messages for error detection and correction.

Materials and methods

Designing the codeword table

As mentioned in the text, the number of codewords consisting of three ternary digits with at least one "2" is 19.

Therefore, 19 kinds of non-proline amino acid types can be discriminated with only three kinds of labeled samples. We designed the codeword tables shown in Figs. 1b, 2a based on the following considerations (see Supplementary Note S3 with respect to the SI-labeling ratio).

First, the signal intensity of each sample is disturbed by protein concentration differences and/or other technical reasons, such as magnetic field inhomogeneity (hereafter called "intensity disturbance"). The intensity disturbance has to be compensated, because the accuracy and precision of SI indices are critical for decoding amino acid information in the SiCode strategy. For this compensation, the signal of a fully labeled amino acid in all samples, namely that mapped to the codeword "222", was used as described below (see Supplementary Note S4 for the result of the compensation). We have assigned "222" to glycine, as shown in Fig. 1b, because its signal rarely overlapped with other signals in ^{1}H-^{15}N HSQC-type spectrum and thus it can be easily distinguished.

Second, as described in the text, the *E. coli*-based cell-free system supplemented with metabolic inhibitors to suppress isotopic scrambling (Yokoyama et al. 2011) was used in order to achieve accurate and precise SI labeling, which is quite crucial for SiCode. However, asparagine to aspartate conversion could not be fully suppressed because 5-diazo-4-oxo-L-norvaline, which can achieve this suppression (Yokoyama et al. 2011), was not used in the present study because it was unavailable. In order to overcome this scrambling issue, we assigned asparagine and aspartate to "220" and "210", respectively. As both amino acids were designed to have the same digit "2" or "0", namely the same SI-labeling ratios, for samples 1 and 3, the scrambling does not matter. In addition, we intentionally lowered the SI-labeling ratio of aspartate in the protein production processes of samples 2 and 4, whereas asparagine was fully labeled, so that the SI-labeling ratio of aspartate would be within the range for digit "1" (between 25 and 75 % for ^{13}C and between 62.5 and 87.5 % for ^{15}N, as described below) even if it became increased by asparagine to aspartate conversion. For the same reasons, 6-diazo-5-oxo-L-norleucine, which is responsible for suppressing glutamine to glutamate conversion (Yokoyama et al. 2011), was not used in the present study. In order to overcome this scrambling, glutamine and glutamate were assigned to "022" and "021", respectively, and the SI-labeling ratio of glutamine for the preparation of samples 3 and 4 was intentionally lowered, as described in Supplementary Information.

Third, from an economic viewpoint, we designed the table in order to limit the total consumption of relatively expensive SI-labeled amino acids; for example, tryptophan is assigned to "002".

NMR spectral analysis for amino-acid discrimination of the Smoothelin CH domain

All NMR spectra were recorded on an AVANCE 700 spectrometer equipped with a CryoProbe (Bruker Biospin, Germany) at 295 K, and processed with the program NMRPipe (Delaglio et al. 1995). Acquisition and processing parameters are shown in Table S1. For the ^{1}H-^{15}N HSQC spectrum of sample 1, cross peaks were picked with the program NMRview (Johnson and Blevins 1994). The peaks were grouped so that a pair of peaks for which the chemical shift difference was less than 0.1 ppm in proton and 0.8 ppm in nitrogen was in the same group.

Discrimination of amino acids was achieved as follows. At first, to obtain the precise peak intensity of each cross peak, least square fitting to the two-dimensional Gaussian functions shown in Eq. 1 was performed for every peak group, using the program R (http://www.r-project.org/index.html) with the minpack.lm package (http://CRAN.R-project.org/package=minpack.lm):

$$\begin{pmatrix} I_{\mathrm{HSQC1}}(x,y) \\ I_{\mathrm{HSQC2}}(x,y) \\ I_{\mathrm{HSQC3}}(x,y) \\ I_{\mathrm{HNCO1}}(x,y) \\ I_{\mathrm{HNCO2}}(x,y) \\ I_{\mathrm{HNCO3}}(x,y) \end{pmatrix} = \sum_{k=1}^{n} \begin{pmatrix} a^{k}_{\mathrm{HSQC1}} \\ a^{k}_{\mathrm{HSQC2}} \\ a^{k}_{\mathrm{HSQC3}} \\ a^{k}_{\mathrm{HNCO1}} \\ a^{k}_{\mathrm{HNCO2}} \\ a^{k}_{\mathrm{HNCO3}} \end{pmatrix} \times \exp\left(-\frac{\left(x - x^{k}_{0}\right)^{2}}{2\sigma^{k2}_{x}} - \frac{\left(y - y^{k}_{0}\right)^{2}}{2\sigma^{k2}_{y}} \right) \quad (1)$$

where x and y are the chemical shifts of the ^{1}H and ^{15}N axes, respectively, $I_{\mathrm{HSQC1}}(x,y)$ is the intensity at position $(x,\ y)$ of HSQC of sample 1 (so are $I_{\mathrm{HSQC2}}(x,y)$ to $I_{\mathrm{HNCO3}}(x,y)$), n is the number of peaks in the group, and x^{k}_{0}, y^{k}_{0}, σ^{k}_{x}, σ^{k}_{y}, and a^{k}_{HSQC1} to a^{k}_{HNCO3} are variables. Initial values for fitting were: chemical shifts of picked peak as described for x^{k}_{0} and y^{k}_{0}, intensity of each spectrum at position (x^{k}_{0}, y^{k}_{0}) for a^{k}_{HSQC1} to a^{k}_{HNCO3}, 0.02 ppm for σ^{k}_{x}, and 0.15 ppm for σ^{k}_{y}. The distribution of line width variables σ_x and σ_y is shown in Fig. S6.

Secondly, to discriminate the amino acid type of the residue i of each amide signal, ^{15}N SI indices were calculated by

$$a_{\mathrm{HSQC}} = \mathrm{absmax}\{a_{\mathrm{HSQC1}}, a_{\mathrm{HSQC2}}, a_{\mathrm{HSQC3}}\} \quad (2)$$

$$\begin{pmatrix} r_{\mathrm{N1}} \\ r_{\mathrm{N2}} \\ r_{\mathrm{N3}} \end{pmatrix} = \frac{1}{a_{\mathrm{HSQC}}} \begin{pmatrix} a_{\mathrm{HSQC1}} \\ a_{\mathrm{HSQC2}} \\ a_{\mathrm{HSQC3}} \end{pmatrix} \quad (3)$$

where a_{HSQC} is the HSQC intensity when the sample is 100 % ^{15}N labeled, r_{N1}, r_{N2}, and r_{N3} are the ^{15}N SI index of each sample. Distribution of a_{HSQC} values is shown in Fig. S6. The function absmax was defined by

$$\mathrm{absmax}\{\mathbf{a}\} \equiv \begin{cases} \max\{\mathbf{a}\}, & \text{if } |\max\{\mathbf{a}\}| \geq |\min\{\mathbf{a}\}| \\ \min\{\mathbf{a}\}, & \text{if } |\max\{\mathbf{a}\}| < |\min\{\mathbf{a}\}| \end{cases} \quad (4)$$

Since a_{HSQC} is the HSQC intensity of 100 % ^{15}N labeled sample, a_{HSQC} value of negative signals such as aliased peaks should be negative. Therefore, the absmax function should be used instead of simple max function to process both positive and negative signals. Each SI index was converted to the ternary digit by

$$d_{\mathrm{N}i} = \begin{cases} 2, & \text{if } 87.5\,\% \leq r_{\mathrm{N}i} \leq 100\,\% \\ 1, & \text{if } 62.5\,\% \leq r_{\mathrm{N}i} < 87.5\,\% \\ 0, & \text{if } 37.5\,\% \leq r_{\mathrm{N}i} < 62.5\,\% \end{cases} \quad (5)$$

The amino acid type was judged from the combination of converted digits, based on the labeling pattern shown in Fig. 1b.

Thirdly, as the labeling ratios for glycine were defined as 100 % for all samples (Fig. 1b), the intensity disturbance was corrected based on the average labeling ratios of the signals judged as those of glycine residues in the second step (see "Designing the codeword table" section for details), as follows:

$$\begin{pmatrix} I'_{\mathrm{HSQC1}}(x,y) \\ I'_{\mathrm{HSQC2}}(x,y) \\ I'_{\mathrm{HSQC3}}(x,y) \\ I'_{\mathrm{HNCO1}}(x,y) \\ I'_{\mathrm{HNCO2}}(x,y) \\ I'_{\mathrm{HNCO3}}(x,y) \end{pmatrix} = \begin{pmatrix} I_{\mathrm{HSQC1}}(x,y) \Big/ \overline{r^{\mathrm{G}}_{\mathrm{N1}}} \\ I_{\mathrm{HSQC2}}(x,y) \Big/ \overline{r^{\mathrm{G}}_{\mathrm{N2}}} \\ I_{\mathrm{HSQC3}}(x,y) \Big/ \overline{r^{\mathrm{G}}_{\mathrm{N3}}} \\ I_{\mathrm{HNCO1}}(x,y) \Big/ \overline{r^{\mathrm{G}}_{\mathrm{N1}}} \\ I_{\mathrm{HNCO2}}(x,y) \Big/ \overline{r^{\mathrm{G}}_{\mathrm{N2}}} \\ I_{\mathrm{HNCO3}}(x,y) \Big/ \overline{r^{\mathrm{G}}_{\mathrm{N3}}} \end{pmatrix} \quad (6)$$

where $I'_{\mathrm{HSQC1}}(x,y)$ to $I'_{\mathrm{HNCO3}}(x,y)$ are the corrected intensities, $I_{\mathrm{HSQC1}}(x,y)$ to $I_{\mathrm{HNCO3}}(x,y)$ are the raw intensities, $\overline{r^{\mathrm{G}}_{\mathrm{N1}}}$ is the average of the r_{N1} values of residues judged as glycine, and so are $\overline{r^{\mathrm{G}}_{\mathrm{N2}}}$ and $\overline{r^{\mathrm{G}}_{\mathrm{N3}}}$. After this compensation for the intensity disturbance, the first and second steps were repeated using the corrected intensities.

Finally, the ^{13}C SI indices of the residue i − 1 of each amide signal were calculated. As the signal intensity of HNCO was proportional to both the ^{15}N labeling ratio of the residue i and ^{13}C labeling ratio of the residue i − 1 (Ikura et al. 1990), the labeling ratio of ^{13}C was calculated by

$$\begin{pmatrix} a'_{\mathrm{HNCO1}} \\ a'_{\mathrm{HNCO2}} \\ a'_{\mathrm{HNCO3}} \end{pmatrix} = \begin{pmatrix} a_{\mathrm{HNCO1}} / r_{\mathrm{N1}} \\ a_{\mathrm{HNCO2}} / r_{\mathrm{N2}} \\ a_{\mathrm{HNCO3}} / r_{\mathrm{N3}} \end{pmatrix} \quad (7)$$

$$a'_{\mathrm{HNCO}} = \mathrm{absmax}\{a'_{\mathrm{HNCO1}}, a'_{\mathrm{HNCO2}}, a'_{\mathrm{HNCO3}}\} \quad (8)$$

$$\begin{pmatrix} r_{\mathrm{C1}} \\ r_{\mathrm{C2}} \\ r_{\mathrm{C3}} \end{pmatrix} = \frac{1}{a'_{\mathrm{HNCO}}} \begin{pmatrix} a'_{\mathrm{HNCO1}} \\ a'_{\mathrm{HNCO2}} \\ a'_{\mathrm{HNCO3}} \end{pmatrix} \quad (9)$$

where r_{C1}, r_{C2}, and r_{C3} are the ^{13}C SI index of the residue i − 1 of each sample. Each SI index was converted to the ternary digit by

$$d_{\mathrm{C}i} = \begin{cases} 2, & \text{if } 75\,\% \leq r_{\mathrm{C}i} \leq 100\,\% \\ 1, & \text{if } 25\,\% \leq r_{\mathrm{C}i} < 75\,\% \\ 0, & \text{if } -25\,\% \leq r_{\mathrm{C}i} < 25\,\% \end{cases} \quad (10)$$

Note that the SI index could be negative, because it was back-calculated from signal intensities that might be corrupted by noise. The amino acid type of the residue i − 1 was judged from the combination of the converted digits in a similar manner to the type of the residue i, as described above.

NMR spectral analysis with error detection by check digit for the BMX SH2 domain

The NMR spectral measurement and analysis for the BMX SH2 domain were performed by essentially the same procedure as described above for the Smoothelin CH domain, except for the following. Firstly, NMR spectra were acquired at 298 K. Secondly, peak fitting was simultaneously performed for all four of the samples, including sample 4 for check digit, by:

$$\begin{pmatrix} I_{\mathrm{HSQC1}}(x,y) \\ I_{\mathrm{HSQC2}}(x,y) \\ I_{\mathrm{HSQC3}}(x,y) \\ I_{\mathrm{HSQC4}}(x,y) \\ I_{\mathrm{HNCO1}}(x,y) \\ I_{\mathrm{HNCO2}}(x,y) \\ I_{\mathrm{HNCO3}}(x,y) \\ I_{\mathrm{HNCO4}}(x,y) \end{pmatrix} = \sum_{k=1}^{n} \begin{pmatrix} a^k_{\mathrm{HSQC1}} \\ a^k_{\mathrm{HSQC2}} \\ a^k_{\mathrm{HSQC3}} \\ a^k_{\mathrm{HSQC4}} \\ a^k_{\mathrm{HNCO1}} \\ a^k_{\mathrm{HNCO2}} \\ a^k_{\mathrm{HNCO3}} \\ a^k_{\mathrm{HNCO4}} \end{pmatrix} \times \exp\left(-\frac{\left(x - x^k_0\right)^2}{2\sigma^{k2}_x} - \frac{\left(y - y^k_0\right)^2}{2\sigma^{k2}_y} \right) \quad (1')$$

Then, the ^{15}N SI indices were calculated by

$$a_{\mathrm{HSQC}} = \mathrm{absmax}\{a_{\mathrm{HSQC1}}, a_{\mathrm{HSQC2}}, a_{\mathrm{HSQC3}}\} \quad (2)$$

$$\begin{pmatrix} r_{\mathrm{N1}} \\ r_{\mathrm{N2}} \\ r_{\mathrm{N3}} \\ r_{\mathrm{N4}} \end{pmatrix} = \frac{1}{a_{\mathrm{HSQC}}} \begin{pmatrix} a_{\mathrm{HSQC1}} \\ a_{\mathrm{HSQC2}} \\ a_{\mathrm{HSQC3}} \\ a_{\mathrm{HSQC4}} \end{pmatrix} \quad (3')$$

Note that since the reference intensity a_{HSQC} was evaluated from the first three samples; i.e., the data digits, the SI index of the 4th sample, r_{N4}, might exceed 100 %. Each SI index was converted to the ternary digit by

$$d_{\mathrm{N}i} = \begin{cases} 2, & \text{if } 87.5\,\% \le r_{\mathrm{N}i} \le 112.5\,\% \\ 1, & \text{if } 62.5\,\% \le r_{\mathrm{N}i} < 87.5\,\% \\ 0, & \text{if } 37.5\,\% \le r_{\mathrm{N}i} < 62.5\,\% \end{cases} \quad (5')$$

The ^{13}C SI indices and their corresponding ternary digits for the residue i − 1 were calculated by

$$\begin{pmatrix} a'_{\mathrm{HNCO1}} \\ a'_{\mathrm{HNCO2}} \\ a'_{\mathrm{HNCO3}} \\ a'_{\mathrm{HNCO4}} \end{pmatrix} = \begin{pmatrix} a_{\mathrm{HNCO1}}/r_{\mathrm{N1}} \\ a_{\mathrm{HNCO2}}/r_{\mathrm{N2}} \\ a_{\mathrm{HNCO3}}/r_{\mathrm{N3}} \\ a_{\mathrm{HNCO4}}/r_{\mathrm{N4}} \end{pmatrix} \quad (7')$$

$$a'_{\mathrm{HNCO}} = \mathrm{absmax}\left\{a'_{\mathrm{HNCO1}}, a'_{\mathrm{HNCO2}}, a'_{\mathrm{HNCO3}}\right\} \quad (8)$$

$$\begin{pmatrix} r_{\mathrm{C1}} \\ r_{\mathrm{C2}} \\ r_{\mathrm{C3}} \\ r_{\mathrm{C4}} \end{pmatrix} = \frac{1}{a'_{\mathrm{HNCO}}} \begin{pmatrix} a'_{\mathrm{HNCO1}} \\ a'_{\mathrm{HNCO2}} \\ a'_{\mathrm{HNCO3}} \\ a'_{\mathrm{HNCO4}} \end{pmatrix} \quad (9')$$

$$d_{\mathrm{C}i} = \begin{cases} 2, & \text{if } 75\% \le r_{\mathrm{C}i} \le 125\% \\ 1, & \text{if } 25\% \le r_{\mathrm{C}i} < 75\% \\ 0, & \text{if } -25\% \le r_{\mathrm{C}i} < 25\% \end{cases} \quad (10')$$

Thirdly, if the converted check digit from the labeling ratio of sample 4 was inconsistent with the digit generated based on the defined codeword, as in Fig. 2a (see Supplementary Note S5 for the detail of check digit calculation), the judged amino acid type was considered to be incorrect.

Acknowledgments We thank the lab members at RIKEN QBiC and RInC, particularly S. Watanabe, N. Matsuda, and Y. Kasai for their kind help in preparing the materials and S. Yasuda for secretarial assistance. This work was supported in part by a Grant-in-Aid for Scientific Research on Innovative Areas (Grant No. 25120003), a Grant-in-Aid for Challenging Exploratory Research (Grant No. 26650027), and a Grant-in-Aid for Young Scientists (B) (Grant No. 24770108) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan and the Japan Society for the Promotion of Science (JSPS).

Compliance with ethical standards

Conflict of interest T. Kasai and T. Kigawa are co-inventors on a patent application (JP 2013-82543) related in part to the material presented here. J. Yokoyama is a salaried employee of Taiyo Nippon Sanso Corp., a company that has commercial interests in the cell-free protein synthesis system. Cell-Free Technology Application Laboratory was jointly funded by RIKEN and Taiyo Nippon Sanso Corp. S. Koshiba declares no potential conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Bertelsen EB, Chang L, Gestwicki JE, Zuiderweg ER (2009) Solution conformation of wild-type *E. coli* Hsp70 (DnaK) chaperone complexed with ADP and substrate. Proc Natl Acad Sci USA 106:8471–8476. doi:10.1073/pnas.0903503106

Cervantes CF, Handley LD, Sue SC, Dyson HJ, Komives EA (2013) Long-range effects and functional consequences of stabilizing mutations in the ankyrin repeat domain of IκBα. J Mol Biol 425:902–913. doi:10.1016/j.jmb.2012.12.012

Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. J Biomol NMR 6:277–293. doi:10.1007/BF00197809

Grzesiek S, 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. doi:10.1007/BF00178261

Hefke F, Bagaria A, Reckel S, Ullrich SJ, Dötsch V, Glaubitz C, 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. doi:10.1007/s10858-010-9462-4

Hembram DS, Haremaki T, Hamatsu J, Inoue J, Kamoshida H, Ikeya T, Mishima M, Mikawa T, Hayashi N, Shirakawa M, 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 438:653–659. doi:10.1016/j.bbrc.2013.07.127

Ikura M, Kay LE, Bax A (1990) A novel approach for sequential assignment of ^{1}H, ^{13}C, and ^{15}N spectra of proteins: heteronuclear triple-resonance three-dimensional NMR spectroscopy. Application to calmodulin. Biochemistry 29:4659–4667. doi:10.1021/bi00471a022

Jaipuria G, Krishnarjuna B, Mondal S, Dubey A, Atreya HS (2012) Amino acid selective labeling and unlabeling for protein resonance assignments. Adv Exp Med Biol 992:95–118. doi:10.1007/978-94-007-4954-2_6

Johnson BA, Blevins RA (1994) NMR view: A computer program for the visualization and analysis of NMR data. J Biomol NMR 4:603–614. doi:10.1007/bf00404272

Kainosho M, 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 21:6273–6279. doi:10.1021/bi00267a036

Kigawa T, Yabuki T, Yoshida Y, Tsutsui M, Ito Y, Shibata T, Yokoyama S (1999) Cell-free production and stable-isotope labeling of milligram quantities of proteins. FEBS Lett 442:15–19. doi:10.1016/S0014-5793(98)01620-2

Kigawa T, Yabuki T, Matsuda N, Matsuda T, Nakajima R, Tanaka A, Yokoyama S (2004) Preparation of *Escherichia coli* cell extract for highly productive cell-free protein expression. J Struct Funct Genomics 5:63–68. doi:10.1023/B:JSFG.0000029204.57846.7d

Krishnarjuna B, Jaipuria G, Thakur A, D'Silva P, Atreya HS (2011) Amino acid selective unlabeling for sequence specific resonance

assignments in proteins. J Biomol NMR 49:39–51. doi:10.1007/s10858-010-9459-z

Löhr F, Reckel S, Karbyshev M, Connolly PJ, Abdul-Manan N, Bernhard F, Moore JM, Dötsch V (2012) Combinatorial triple-selective labeling as a tool to assist membrane protein backbone resonance assignment. J Biomol NMR 52:197–210. doi:10.1007/s10858-012-9601-1

Maslennikov I, Choe S (2013) Advances in NMR structures of integral membrane proteins. Curr Opin Struct Biol 23:555–562. doi:10.1016/j.sbi.2013.05.002

Maslennikov I, Klammt C, Hwang E, Kefala G, Okamura M, Esquivies L, Mörs K, Glaubitz C, Kwiatkowski W, Jeon YH, 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 USA 107:10902–10907. doi:10.1073/pnas.1001656107

Matsuda T, Koshiba S, Tochio N, Seki E, Iwasaki N, Yabuki T, Inoue M, Yokoyama S, Kigawa T (2007) Improving cell-free protein synthesis for stable-isotope labeling. J Biomol NMR 37:225–229. doi:10.1007/s10858-006-9127-5

Parker MJ, Aulton-Jones M, Hounslow AM, Craven CJ (2004) A combinatorial selective labeling method for the assignment of backbone amide NMR resonances. J Am Chem Soc 126:5020–5021. doi:10.1021/ja039601r

Schneider WM, Tang Y, Vaiphei ST, Mao L, Maglaqui M, Inouye M, Roth MJ, Montelione GT (2010) Efficient condensed-phase production of perdeuterated soluble and membrane proteins. J Struct Funct Genomics 11:143–154. doi:10.1007/s10969-010-9083-x

Seki E, Matsuda N, Yokoyama S, 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 377:156–161. doi:10.1016/j.ab.2008.03.001

Shi J, Pelton JG, Cho HS, Wemmer DE (2004) Protein signal assignments using specific labeling and cell-free synthesis. J Biomol NMR 28:235–247. doi:10.1023/b:jnmr.0000013697.10256.74

Sobhanifar S, Reckel S, Junge F, Schwarz D, Kai L, Karbyshev M, Löhr F, Bernhard F, Dötsch V (2010) Cell-free expression and stable isotope labelling strategies for membrane proteins. J Biomol NMR 46:33–43. doi:10.1007/s10858-009-9364-5

Staunton D, Schlinkert R, Zanetti G, Colebrook SA, Campbell ID (2006) Cell-free expression and selective isotope labelling in protein NMR. Magn Reson Chem 44:S2–S9. doi:10.1002/mrc.1835

Suzuki M, Mao L, Inouye M (2007) Single protein production (SPP) system in *Escherichia coli*. Nat Protoc 2:1802–1810. doi:10.1038/nprot.2007.252

Trbovic N, Klammt C, Koglin A, Löhr F, Bernhard F, Dötsch V (2005) Efficient strategy for the rapid backbone assignment of membrane proteins. J Am Chem Soc 127:13504–13505. doi:10.1021/ja0540270

Wu PS, Ozawa K, Jergic S, Su XC, Dixon NE, Otting G (2006) Amino-acid type identification in ^{15}N-HSQC spectra by combinatorial selective ^{15}N-labelling. J Biomol NMR 34:13–21. doi:10.1007/s10858-005-5021-9

Yokoyama J, Matsuda T, Koshiba S, Tochio N, Kigawa T (2011) A practical method for cell-free protein synthesis to avoid stable isotope scrambling and dilution. Anal Biochem 411:223–229. doi:10.1016/j.ab.2011.01.017