J Biomol NMR (2013) 57:65–72
DOI 10.1007/s10858-013-9768-0

ARTICLE

Resonance assignment for a particularly challenging protein based on systematic unlabeling of amino acids to complement incomplete NMR data sets

Peter Bellstedt · Thomas Seiboth · Sabine Häfner · Henriette Kutscha · Ramadurai Ramachandran · Matthias Görlach

Received: 8 May 2013 / Accepted: 3 August 2013 / Published online: 14 August 2013
© Springer Science+Business Media Dordrecht 2013

Abstract NMR-based structure determination of a protein requires the assignment of resonances as indispensable first step. Even though heteronuclear through-bond correlation methods are available for that purpose, challenging situations arise in cases where the protein in question only yields samples of limited concentration and/or stability. Here we present a strategy based upon specific individual unlabeling of all 20 standard amino acids to complement standard NMR experiments and to achieve unambiguous backbone assignments for the fast precipitating 23 kDa catalytic domain of human aprataxin of which only incomplete standard NMR data sets could be obtained. Together with the validation of this approach utilizing the protein GB1 as a model, a comprehensive insight into metabolic interconversion ("scrambling") of NH and CO groups in a standard *Escherichia coli* expression host is provided.

Keywords Resonance assignment · Unlabeling · Selective isotope labeling · Reverse labeling · Aprataxin

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9768-0) contains supplementary material, which is available to authorized users.

P. Bellstedt (✉) · T. Seiboth · S. Häfner · H. Kutscha · R. Ramachandran · M. Görlach
Biomolecular NMR Spectroscopy, Leibniz Institute for Age Research, Fritz Lipmann Institute, Beutenbergstr. 11, 07745 Jena, Germany
e-mail: pbell@fli-leibniz.de

Introduction

Assignment of NMR resonances and here assignment of backbone nuclei constitutes the essential first step in the process of biomolecular structure determination. Frequently, heteronuclear triple resonance experiments like the HNCACB (Wittekind and Mueller 1993) in combination with the CBCA(CO)NH (Grzesiek and Bax 1992) are performed to identify and assign intra- and interresidue resonances. In addition, approaches for linking adjacent amide residues based on NH(COCA)NH type experiments (Bracken et al. 1997; Kumar et al. 2010) are available. To ameliorate the inherent sensitivity problem of NMR spectroscopy, improved hardware (higher field strength, cryogenic probes) and new pulse sequences yielding higher s/n ratios (Mori et al. 1995; Lescop et al. 2007; Felli and Brutscher 2009; Marion 2010) and/or requiring less acquisition time through non-uniform sampling and/or projection reconstruction techniques (Kupče and Freeman 2004; Coggins et al. 2010; Qiang 2011) are continuously developed. All these approaches require isotope labeling and, as second prerequisite, a protein sample of relatively high concentration and considerable stability. In cases where stability or concentration of the sample is limited, experiments with a reduced number of coherence transfer steps (e.g., HSCQ, HNCO, HNCA), which are inherently more sensitive and, therefore, need shorter recording times, may still be feasible. Such experiments, however, may pose a problem with respect to resonance assignment due to the limited dispersion of CA and CO chemical shifts. This in turn results in potentially ambiguous assignments already for medium sized proteins. Such ambiguity may be resolvable in part by using experiments selective for amino acid type (Schubert et al. 1999; Schubert 2001) or the limited information linked to characteristic resonance

"regions" (e.g., glycine in [^{1}H,^{15}N]-based correlation experiments) together with a given protein sequence as the latter restricts the potential amino acid *i, i+1* pairs. Furthermore, specific isotope labeling in an unlabeled background allows to identify the residue type. Specific isotope labeling of many amino acids, however, is a costly proposition if performed in vivo and may preclude the NMR analysis of a challenging protein altogether. On the other hand, specific incorporation of [^{14}N, ^{12}C]-amino acids in an uniformly labeled background is much less costly and can easily be achieved using standard bacterial systems for heterologous protein expression. Such "unlabeling" approaches have so far been employed sporadically and were restricted to a limited subset of amino acids (Shortle 1994; Kelly et al. 1999; Krishnarjuna et al. 2011; Banigan et al. 2013). The complicating issue in specific (un)labeling, however, constitutes the scrambling (interconversion) of the label into the intentionally unlabeled amino acids and *vice versa*. Even though genetically modified strains can be used to avoid scrambling (Waugh 1996; O'Grady et al. 2012), such bacterial strains are not considered standard recombinant systems and would e.g., require re-engineering to allow for T7 polymerase based (Studier and Moffatt 1986) overexpression. Furthermore, genetic engineering of metabolic pathways to minimize scrambling is restricted and affects a limited number of amino acids only (Rasia et al. 2012). Virtually scrambling-free synthesis of a target protein may be achievable in cell-free protein synthesis systems (Morita et al. 2004; Su et al. 2011). Eukaryotic systems have also been used in combination with amino acid selective (un)labeling (Strauss et al. 2003; Tanio et al. 2009), but frequently suffer from lower yield and higher overall costs. Here we systematically evaluated the unlabeling approach for all 20 natural amino acids and for two recombinant proteins expressed in the standard T7 polymerase based *Escherichia coli* BL21(DE3) system. The first immunoglobulin binding domain of protein G (GB1) served as a model system to assess the feasibility of this strategy. In a second step we successfully applied this approach to the 23 kDa catalytic domain of human aprataxin. Mutations in aprataxin are linked to the neurodegenerative disease ataxia with oculomotor apraxia 1 (AOA1; Moreira et al. (2001); Date et al. (2001)). Aprataxin is involved in DNA repair (Gueven et al. 2004; Ahel et al. 2006; Rass et al. 2007; Rass et al. 2008) and it might even constitute a novel drug target in colorectal cancer therapy (Dopeso et al. 2010). Despite its biomedical significance, the NMR spectroscopic analysis of human aprataxin was severely hampered by its propensity to denature in short time and to aggregate already at low protein concentrations. Based on our systematic unlabeling approach, and by including the information of the amino acid type(s) during the assignment process, here we present an assignment strategy based on linking CO and CA resonances and a partial backbone resonance assignment for aprataxin's catalytically active domain in order to provide vital chemical shift information for e.g., NMR-based compound screening.

Materials and methods

Expression and purification of GB1

Uniformly [^{15}N, ^{13}C]-labeled GB1 (T2Q mutant) was expressed in *E. coli* BL21(DE3) and M9 media containing 1 g/l $^{15}NH_4Cl$ and 2 g/l [^{1}H, ^{13}C]-glucose. For selectively unlabeled samples of GB1, 1 g/l (except for Cys: 0.1 g/l) of the respective [^{14}N, ^{12}C]-amino acid (Sigma Aldrich) was added to the medium 15 min prior to induction with 0.3 mM IPTG for 3 h at 30 °C. Based on published protocols (Franks et al. 2005) the harvested cells were disrupted by heating to 80 °C for 15 min in phosphate-buffered saline (200 mM NaCl, 50 mM KH_2PO_4/K_2HPO_4, pH 7), debris was removed by centrifugation (16,000×*g*, 4 °C, 30 min) and the supernatant was subjected to size exclusion chromatography (Sephadex 75, GE Heathcare; 100 mM NaCl, 50 mM KH_2PO_4/K_2HPO_4, pH 7). GB1 containing fractions were combined and supplemented with 10 % D_2O concentrated to 500 µM using a Vivaspin 20 concentrator (3.5 kDa cut-off, GE Healthcare). Approx. 30 mg of pure protein was obtained from 250 ml M9 media inoculated with a cell mass corresponding to an OD_{600nm} of 0.7 in 1 l LB media.

Expression and purification of aprataxin

The pET-15b expression plasmid coding for residues 161–356 of human aprataxin (UniProt id: Q7Z2E3-1), comprising its enzymatically active histidine triad (HIT) domain and zinc finger motif, was transformed into *E. coli* BL21(DE3). The His_6-tagged protein was expressed in M9 media supplemented with 1 g/l $^{15}NH_4Cl$ and 1.5 g/l [^{1}H, ^{13}C]-glucose and a final concentration of 10 µM $ZnSO_4$. Unlabeling and protein expression were induced as for GB1 with 0.3 mM IPTG for 3 h at 30 °C. Harvested cells were disrupted by French Press and ultrasonification, debris was removed by centrifugation (15,400×*g*, 4 °C, 30 min) and the supernatant was subjected to a affinity chromatography on Ni-NTA agarose (Qiagen) followed by thrombin cleavage and dialysis (50 mM NaCl, 10 mM Tris/HCl pH 7.5, 4 °C over night). Pure protein was obtained as flow through of the final anion exchange chromatography (DEAE fast-flow sepharose; GE Heathcare). The final NMR samples were concentrated and exchanged into NMR buffer (150 mM NaCl, 10 mM dTris/

HCl pH 7.5, 10 % D_2O) using a Vivaspin 20 concentrator (10 kDa cut-off, GE Healthcare). The protein yield was approx. 5 mg of pure protein obtained from 200 ml M9 media inoculated with a cell mass corresponding to an OD_{600nm} of 0.7 in 1 l LB media. A scheme of the purification process is given in Figure S4.

NMR experiments

Chemical shift correlation experiments were performed on a Bruker 750 MHz Avance III NMR system equipped with a triple resonance probe. Sample temperature was set to 25 °C (GB1) or 29 °C (aprataxin), respectively. 3D HNCO and 3D HN(CA)CO data for GB1 were obtained from a 2.3 mM uniformly [^{15}N,^{13}C]-labeled sample. For backbone resonance assignment of aprataxin 3D HNCO (experimental time: 16 h), 3D HN(CA)CO (2 d 18 h), 3D HNCA (1 d 20 h), 3D HN(CO)CA (1 d 20 h) and 3D HNCACB (3 d 17 h) were collected with freshly prepared 0.5 mM uniformly [^{15}N,^{13}C]-labeled samples for each of the mentioned experiments. The initial protein concentration dropped within several hours from 0.5 to approx. 0.1 mM. After one day the protein concentration was around 50 μM. The set of the aforementioned standard NMR experiments yielded NMR data (example shown in Figure S6) inadequate to allow for complete and unambiguous assignments. Hence, these experiments were complemented with the information obtained *via* the unlabeling approach presented here to achieve unambiguous resonance assignment for the fast precipitating 23 kD catalytic domain of aprataxin. Unlabeling: [^{1}H,^{15}N]-HSQC spectra were recorded using a standard sequence (Bruker: "hsqcfpf3gpphwg"). [^{13}CO]-edited [^{1}H,^{15}N]-correlation spectra were collected as 2D versions of a 3D HNCO experiment (Bruker: "hncogpwg3d") similar to the approach presented by (Parker et al. 2004). To achieve higher resolution in t_2, normally restricted by the constant time character, this pulse sequence was modified to allow for free evolution in t_2. Water suppression was achieved via presaturation (Jesson et al. 1973). The modified HNCO sequence is referred to as HN(CO) here. The experimental time for the collection of each of the [^{1}H,^{15}N]-HSQC spectra was approx. 1 h utilizing a maximum acquisition time of 85 ms (GB1) / 100 ms (aprataxin) in the direct, and 91 ms (GB1) / 50 ms (aprataxin) in the indirect dimension, respectively. The HN(CO) spectra of each of the specifically unlabeled samples of GB1 and aprataxin were recorded within 2 h and 6 h, respectively.

Data analysis

Amide resonance assignments of GB1 are based on literature data (Franks et al. 2005) and were reproduced with the CCPN software package (Vranken et al. 2005) using 3D HNCO and 3D HN(CA)CO data. Backbone resonances of aprataxin were analyzed using the same software package and the spectra recorded as stated in section "NMR experiments" Unlabeling: Absolute peak intensities were extracted from [^{1}H,^{15}N]-HSQC and HN(CO) data using TOPSPIN V2.1. All spectra were recorded and processed identically as described in the respective figure legends (Figure S1, S2, S8).

Results and discussion

In an ideal case, the incorporation of a specific [^{14}N,^{12}C]-amino acid in an otherwise uniformly [^{15}N, ^{13}C]-labeled protein renders the respective nuclei unobservable in heteronuclear chemical shift correlation experiments. By comparing [^{1}H,^{15}N]-HSQC data of a uniformly labeled sample with data of samples where only one individual amino acid was present in its unlabeled form during expression, one can link each of the amide signals to the respective amino acid type (example shown in Fig. 1a). A ^{13}CO editing step was used to additionally eliminate amide signals arising from $i + 1$ ^{15}N amide directly following a [^{14}N, ^{12}C]-amino acid (Fig. 1 b). The main prerequisite for a straightforward analysis is that the respective [^{14}N,^{12}C]-amino acid used for unlabeling is not interconverted to other amino acids (also referred to as "isotope scrambling"), which obviously influences the (un)labeling pattern. However, in biological systems used for heterologous protein expression (e.g., *E. coli*), this precondition is met only for some amino acids (Muchmore et al. 1989). The amount of isotope scrambling strongly depends on the type of amino acid and on the chemical group to be looked at (e.g., NH, CO, CA, side chain).

Here, we first systematically evaluate the amount of ^{14}NH and ^{12}CO scrambling for a number of amino acids for the *E. coli* BL21(DE3) standard expression system and GB1 as a model protein. Secondly we present an approach for sequential resonance assignment applied to the 23 kDa catalytic domain of human aprataxin, for which essentially only HN, CO and CA chemical shift information could be obtained. In contrast to a recent study (Krishnarjuna et al. 2011) utilizing a 2D $\{^{12}CO_i - {}^{15}N_{i+1}\}$-filtered HSQC experiment, the approach presented here is based on standard experiments, as we found this to be more straightforward for normalizing peak intensities for the assessment of amino acid scrambling (see below).

Classification of amino acids based on amount of ^{14}N scrambling

In order to validate our approach, we assessed the level of ^{14}N scrambling for each of the 20 standard amino acids.

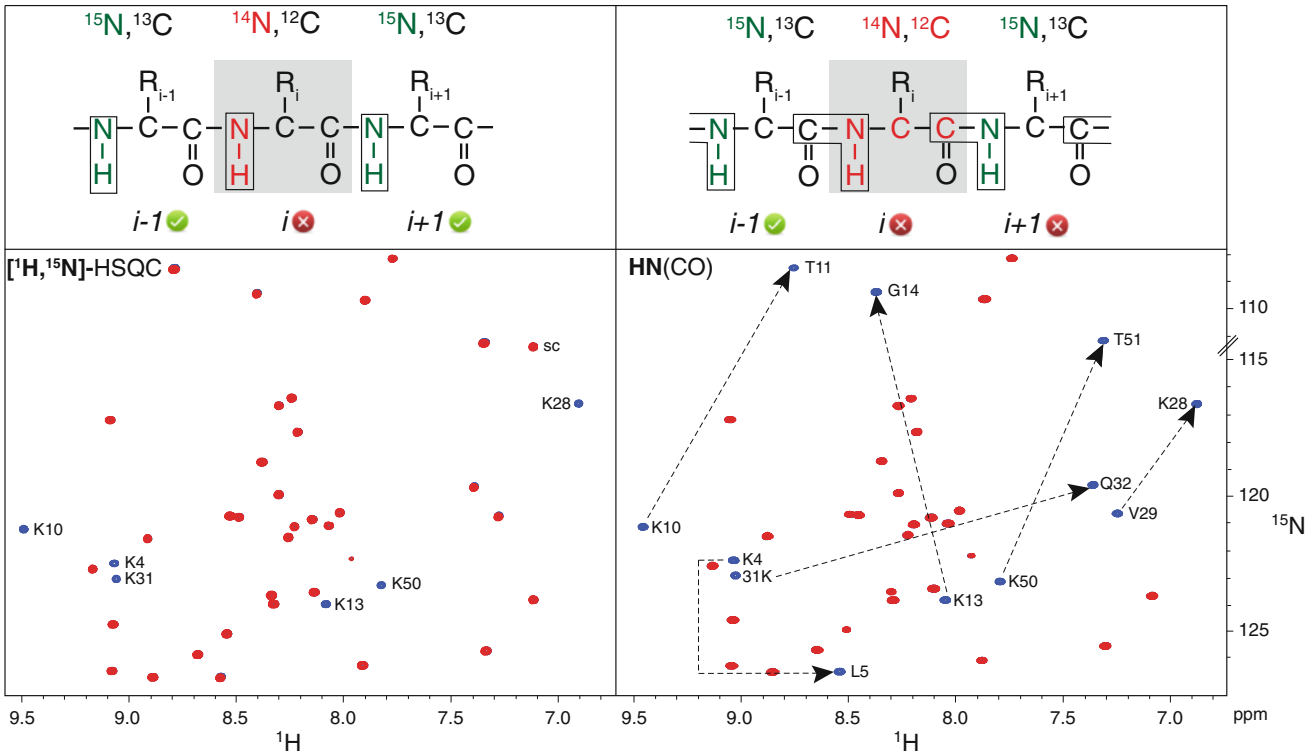

Fig. 1 Identification of unlabeled amino acid *i* via [^{1}H,^{15}N]-HSQC and of *i* and *i+1* via HN(CO) using [^{14}N,^{12}C]-lysine in otherwise uniformly [^{15}N,^{13}C]-labeled GB1 protein. Resonances of the reference spectrum are colored in *blue*, resonances from the specific [^{14}N,^{12}C]-lysine sample in *red*. Missing amide signals in the [^{1}H,^{15}N]-HSQC result from specific unlabeling of lysine amide groups. In the HN(CO) spectra also amide groups are not observable, which directly follow a lysine. The respective *i* and *i+1* pairs are indicated with *dotted arrows*. Assignment is based on standard 3D NMR data (as stated in materials and method). sc: side chain, not observable in HN(CO) spectra

We compared the signal intensity of each of the amide signals of uniformly labeled GB1 with the signal intensity of the respective peak in each of the 20 samples with a single [^{14}N,^{12}C]-amino acid added prior to protein expression (referred to as unlabeled samples). Peak intensities were extracted from [^{1}H,^{15}N]-HSQC data obtained with equal protein concentration, equal experimental and processing parameters. Peak intensities for the same type of amino acid were grouped and averaged to analyze if and how unlabeling of a respective amino acid also effects other ("undesired") amino acids (Figure S1). Based on the resulting normalized, grouped and averaged ^{15}N peak intensities we classified each amino acid with respect to the amount of amide group scrambling (Table 1). Arg, Cys, Ser and His are not present in GB1, but were used for unlabeling to analyze the effect towards other amino amino acids. For 7 out of the 20 amino acids tested (Table 1, 1st column), no scrambling for amide groups was found, allowing a straightforward and unambiguous identification with the outlined unlabel approach. Specific unlabeling of Ala and Gly, respectively, additionally results in a approx. 50 % signal reduction for Val and Trp, respectively (Figure S1). Since the reduction of signal intensity is significantly different (complete loss of signal for Ala and Gly vs. approx. 40–50 % reduction in signal intensity for Val and Trp), this finding can safely be exploited to identify both of the respective amino acids using one singly unlabeled protein sample only. In fact, in the case of the [^{14}N]-Gly sample where also Trp can be determined this is quite important since specific unlabeling of Trp itself leads to uniform scrambling which renders identification of the targeted amide impossible. It is worth emphasizing, that although unlabeling of Gln leads to a general scrambling (Figure S1), the intensity drop is significantly greater for Gln amide signals as compared to all other amino acids. Therefore although Gln is considered to be metabolized by the bacteria, in our hands a quantitative analysis allowed a clear identification of Gln amide signals in GB1. The amide groups of 6 out of 20 amino acids (Table 1, 3rd column) are scrambled among a smaller group of other amino acids and, therefore, only permit the identification of a distinct group of amino acids instead of a single one. Amide signal intensities of Ile, Leu and Val are reduced to the same extent irrespective of which of the three amino acids was present in unlabeled form during protein expression (Figure S1). Likewise, the amide groups of Phe

Table 1 Classification of amino acids with respect to amount of ^{14}N scrambling based on the analysis of [^{1}H,^{15}N]-HSQC peak intensities for GB1 (Figure S1) and the catalytic domain of aprataxin (APTX) (Figure S8)

Unlabeled target only	Target + one more	Distinct group	Uniform scrambling
Arg[a]	Ala	Ile, Leu, Val	Asp
Asn		Phe, Tyr	Glu
Cys[a]		Thr	Ser[a]
Gln		Gly[b]	Trp
His[a]			
Lys			
Met			

[a] Respective amino acid is not present in GB1, but unlabeling data was assessed for scrambling with respect to other amino acid types

[b] In GB1, [^{14}N]-Gly labeling additionally results in the partial unlabeling of Trp amide groups (Figure S1). However, based on the analysis of aprataxin data (Figure S8), [^{14}N]-Gly labeling also leads to unlabeling of Cys and Ser amide groups, not present in GB1

Table 2 Classification of amino acids with respect to amount of ^{12}CO scrambling based on the analysis of HN(CO) peak intensities for GB1 (Figure S2)

No scrambling	Some scrambling	Not analysed[a]
Arg[b], Cys[b]	Ala, Gly	Asp, Asn
His[b], Ile	Thr, Pro[b]	Gln, Glu
Leu, Lys	Ser[b]	Trp
Met, Phe		
Tyr, Val		

[a] Excluded from analysis due to the labeling status of the respective *i+1* amide group, see text for further explanation

[b] Respective amino acid is not present in GB1, but unlabeling data was assessed for scrambling with respect to other amino acid types

and Tyr are scrambled among each other. Although a clear identification of the desired and originally unlabeled amino acid within the latter two groups is not possible, information from that spectra is quite useful to narrow down the possible amino acid types. This in turn can be essential if used in conjunction with sequence information to resolve ambiguities during the assignment process. Adding unlabeled Asp, Glu, Ser and Trp (Table 1, 4th column) results in equal reduction in signal intensity indicative of uniform scrambling and precluding spectral identification (Figure S1). The classification of amino acids with respect to the amount of ^{14}N presented here (Table 1), confirms previously published results (Shortle 1994; Hiroaki et al. 2011; Krishnarjuna et al. 2011) that were limited to a subset or a combination of amino acids only. Here we provide quantitative insight into the ^{14}N isotopic scrambling pattern (Figure S1) individually for each of the 20 amino acids in *E. coli* BL21(DE3) as a standard host for heterologous protein expression.

Classification of amino acids based on amount of ^{12}CO scrambling

A labeling scheme for specific ^{13}CO *labeling* of amino acids has been presented (Takeuchi et al. 2007). However, from the experimental setup and the biochemical point of view specific *labeling* is different from specific *unlabeling*. Unlabeling is achieved by adding the desired [^{14}N,^{12}C]-amino acid only to the expression media. In contrast, for specific *labeling* mostly not only the targeted amino acid (or a precursor) is added to the expression media in its labeled form, but also the other amino acids in their unlabeled form to minimize scrambling through influencing regulation of the amino acid metabolism. To analyze the effect of specific *unlabeling* on the labeling state of carbonyls among the different amino acids, we compared the peak intensities in HN(CO) data recorded with a fully labeled reference sample with data recorded with 20 differently unlabeled samples in the same way as described for the analysis of amide scrambling. Since the resonance intensity of the HN(CO) is not only affected by ^{12}C incorporation into the *i* carbonyl position, but also by the amide of the *i+1* amino acid, a perfect analysis would require an unlabeling approach with [^{15}N,^{12}C]-amino acids, which was far beyond the intent of this study to obtain vital information for the assignment process of the 23 kDa domain of aprataxin. For this reason we only analyzed carbonyls of amino acids *i*, which are followed by an *i+1* amide group, of which signal intensity reduction was <20 % as determined from [^{1}H,^{15}N]-HSQC data. With that restriction we were able to evaluate 15 out of 20 amino acids (all except Asp, Asn, Gln, Glu, Trp) with respect to CO scrambling (Figure S2, summary listed in Table 2). By comparing the extent of ^{14}N incorporation (Figure S1) with the extent of ^{12}C incorporation (Figure S2) following specific unlabeling with an amino acid one can directly extract information about the $^{15}N/^{12}C$ (or $^{14}N/^{13}C$) labeling status. Most of the amino acids analyzed are not prone to ^{12}CO scrambling confirming the observation (Takeuchi et al. 2007), that in general carbonyls are less affected by scrambling than the amide groups. Whereas e.g., the amide group of Ile, Leu and Val scrambles among these 3 amino acids (Tables 1, 3rd column, Figure S1), the carbonyls are *not* affected (Figure S2), permitting unambiguous identification of the *i+1* amide in the HN(CO) spectrum.

Backbone resonance assignment strategy

We have shown that by quantitative analysis of peak intensities for [^{1}H,^{15}N]-HSQC and HN(CO) spectra in

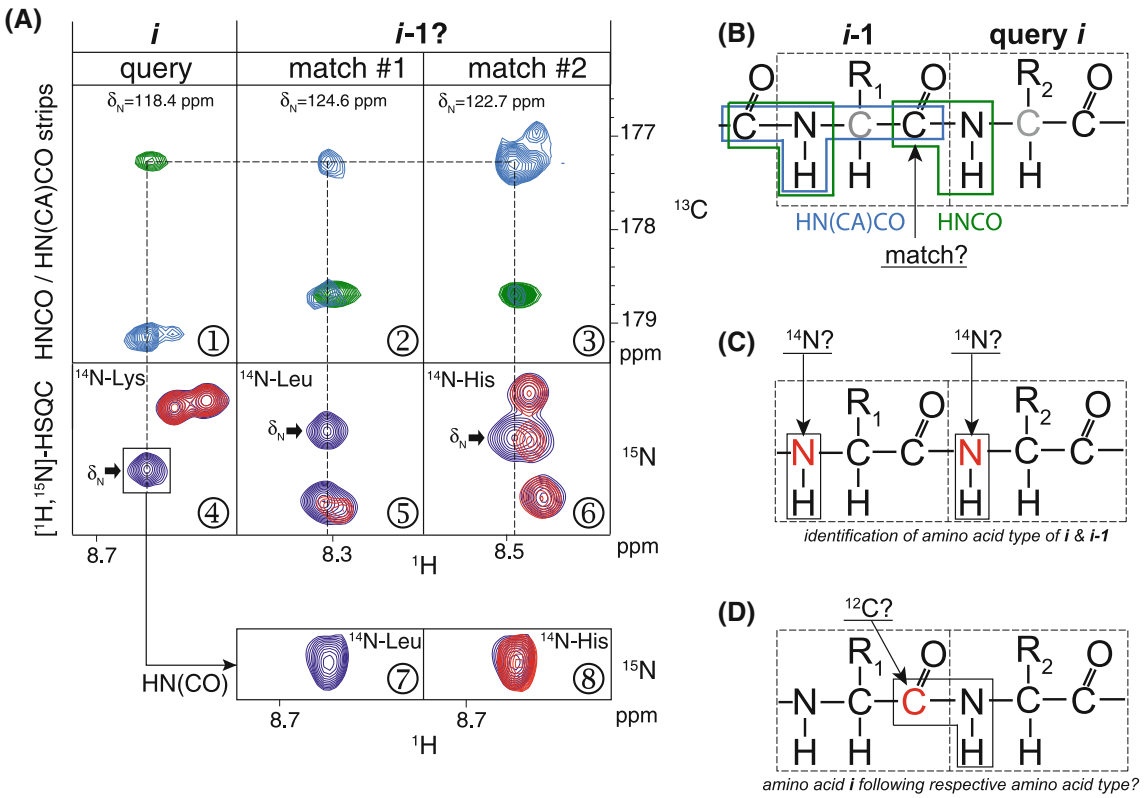

Fig. 2 Assignment strategy for aprataxin supported by amino acid type information obtained from specific unlabeling. This strategy is illustrated here for one assignment step: one "query" amino acid *i* (**a**, *box 1*) could be linked to different potential *i-1* "matches" with equally well corresponding carbon chemical shifts (**a**, *box 2, 3*). Resonances in *light blue* represent HN(CA)CO, resonances in *green* HNCO data. HN(CA)CO is *i, i-1* specific, HNCO *i-1* specific (see scheme in **b**). The 20 different specifically unlabeled samples of aprataxin have to be evaluated together with the sequence information to predict the amino acid type of the chosen query amino acid *i*. Here, the amide peak intensity is reduced only in the [^{14}N]-Lys sample (**a**, *box 4*) and, therefore, the chosen query amino acid *i* is most probably a lysine. *Blue contours* represent data from the fully labeled reference sample, contours in *red* data from the unlabeled sample which were recorded and processed identically. Next, the amino acid types of the potentially matching amino acids have to be identified in the same way (scheme in **c**). In **a**, *box 5 and 6* show the region of [^{1}H,^{15}N]-HSQC spectra where the reduction of signal intensity is the highest observed for the amide group to be evaluated in all 20 unlabeled samples. The particular amino acid used for the respective unlabeling is indicated inside the *panels*. Potentially matching amino acid #1 is most probably a leucine (**a**, *box 5*), matching amino acid #2 a histidine (**a**, *box 6*). The obtained information about amino acid type of the query (*i*) as well as potentially matching amino acids (*i-1*) are used to assess if combinations of *i* and *i-1* are present in the protein sequence. This is the first step, at which *i-1* candidates can be excluded. Here, based on the protein sequence (Figure S3 b), His can be excluded from further consideration as *i-1* candidate, since no His-Lys combination is present in the catalytic domain of aprataxin. Finally, the HN(CO) spectrum for the specific unlabeled amino acid is used to cross check if the initially chosen query *i* is following a Leu residue (scheme in **d**). Only in the [^{14}N]-Leu sample there is a complete reduction in signal intensity in the HN(CO) spectrum (**a**, *box 7* vs. *box 8*), indicating that the chosen *i* is most probably adjacent to a Leu and therefore the query should be linked to matching amino acid #1, delivering an unambiguously assigned Leu-Lys fragment

combination with amino acid-selective unlabeling, in many cases one can extract information about the amino acid type(s) of the respective resonance. Likewise, it is possible to identify peaks that originate from *i+1* amide groups directly following an unlabeled amino acid *i* by analysis of HN(CO) spectra. In cases where concentration and/or stability of a protein in question restricts the NMR data collection to experiments with a minimal number of coherence transfer steps, additional information has to be included to unambiguously assign these backbone resonances. Here we applied this strategy to the backbone resonance assignment of human aprataxin, the accessibility of which for NMR structure determination is severely limited by its aggregation propensity even after buffer optimization (Figure S5). Therefore, the residue specific assignment process of aprataxin was essentially based on NH, CO and CA chemical shifts only. However, owing to limited spectral dispersion and the resulting overlap, it was essential to employ the unlabeling approach to achieve unambiguous backbone resonance assignment for this 23 kDa protein domain.

The strategy to include information obtained via amino acid selective unlabeling during the assignment process is presented in Fig. 2 and consists of 4 principal steps. *Step 1:*

prediction of amino acid type of the query amino acid *i* via analysis of [^{1}H,^{15}N]-HSQC spectra of reference and specifically unlabeled samples, *step 2:* prediction of amino acid type of potentially matching *i-1* amino acids in the same way, *step 3:* check protein sequence if combination of *i* and *i-1* occurs, if not exclude respective *i-1*; *step 4:* cross check in HN(CO) of unlabeled *i-1* candidate as predicted in step 2, if the NH of *i* is vanished as consequence of ^{13}C editing. Where available additional HNCA and incomplete HNCACB data were included in the analysis as well. Following our approach, 78 % of the backbone nuclei of the the catalytic domain of human aprataxin could be assigned successfully (Table S2, Figure S7). Missing assignments result from non-observability of at least one of the resonances for a given amino acid required for unambiguous data analysis. Together with our observation of aprataxin's propensity to aggregate and to precipitate we attribute this to the highly unfavorable dynamics of the protein backbone. The assigned chemical shifts have been deposited in the BMRB (BioMagResBank (Ulrich et al. 2007)) under accession number 19182.

As, in contrast to GB1 (Table S1), each of the 20 natural amino acids is present in aprataxin, we used this work to provide for a more detailed analysis of ^{14}N scrambling. As expected, the labeling scheme of amide groups following amino acid-type specific unlabeling of aprataxin (Figure S8, Table 1) are in agreement with the results extracted from the GB1 data. Strikingly, specific [^{14}N]-Gly unlabeling leads also to complete unlabeling of Cys and Ser amide groups not present in GB1. This nicely illustrates the underlying bacterial amino acid metabolism (Figure S9).

In summary we provide a detailed analysis of ^{14}N and ^{12}CO scrambling using *E. coli* BL21(DE3) as a standard system for heterologous protein expression. In addition, we present a feasible strategy to include the information of amino acid type(s) during the assignment process to resolve potential ambiguities if analysis is hampered due to an incomplete standard NMR data set.

References

Ahel I, Rass U, El-Khamisy SF, Katyal S, Clements PM, McKinnon PJ, Caldecott KW, West SC (2006) The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. Nature 443(7112):713–716

Banigan JR, Gayen A, Traaseth NJ (2013) Combination of 15N reverse labeling and afterglow spectroscopy for assigning membrane protein spectra by magic-angle-spinning solid-state NMR: application to the multidrug resistance protein EmrE. J Biomol NMR 55(4):391–399

Bracken C, Palmer AG, Cavanagh J (1997) (H)N(COCA)NH and HN(COCA)NH experiments for 1H-15N backbone assignments in 13C/15N-labeled proteins. J Biomol NMR 9(1):94–100

Coggins BE, Venters RA, Zhou P (2010) Radial sampling for fast NMR: concepts and practices over three decades. Prog Nucl Magn Reson Spectrosc 57(4):381–419

Date H, Onodera O, Tanaka H, Iwabuchi K, Uekawa K, Igarashi S, Koike R, Hiroi T, Yuasa T, Awaya Y, Sakai T, Takahashi T, Nagatomo H, Sekijima Y, Kawachi I, Takiyama Y, Nishizawa M, Fukuhara N, Saito K, Sugano S, Tsuji S (2001) Early-onset ataxia with ocular motor apraxia and hypoalbuminemia is caused by mutations in a new HIT superfamily gene. Nat Genet 29(2): 184–188

Dopeso H, Mateo-Lozano S, Elez E, Landolfi S, Ramos Pascual FJ, Hernández-Losa J, Mazzolini R, Rodrigues P, Bazzocco S, Carreras MJ, Espín E, Armengol M, Wilson AJ, Mariadason JM, Ramon Y, Cajal S, Tabernero J, Schwartz S, Arango D (2010) Aprataxin tumor levels predict response of colorectal cancer patients to irinotecan-based treatment. Clin Cancer Res 16(8): 2375–2382

Felli IC, Brutscher B (2009) Recent advances in solution NMR: fast methods and heteronuclear direct detection. ChemPhysChem 10(9-10):1356–1368

Franks WT, Zhou DH, Wylie BJ, Money BG, Graesser DT, Frericks HL, Sahota G, Rienstra CM (2005) Magic-angle spinning solid-state NMR spectroscopy of the β1 immunoglobulin binding domain of protein G (GB1): 15N and 13C chemical shift assignments and conformational analysis. J Am Chem Soc 127(35):12,291–12,305

Grzesiek S, Bax A (1992) Correlating backbone amide and side chain resonances in larger proteins by multiple relayed triple resonance NMR. J Am Chem Soc 114(16):6291–6293

Gueven N, Becherel OJ, Kijas AW, Chen P, Howe O, Rudolph JH, Gatti R, Date H, Onodera O, Taucher-Scholz G, Lavin MF (2004) Aprataxin, a novel protein that protects against genotoxic stress. Hum Mol Gen 13(10):1081–1093

Hiroaki H, Umetsu Y, Nabeshima Yi, Hoshi M, Kohda D (2011) A simplified recipe for assigning amide NMR signals using combinatorial 14N amino acid inverse-labeling. J Struct Funct Genomics 12(3):167–174

Jesson JP, Meakin P, Kneissel G (1973) Homonuclear decoupling and peak elimination in Fourier transform nuclear magnetic resonance. J Am Chem Soc 95(2):618–620

Kelly MJ, Krieger C, Ball LJ, Yu Y, Richter G, Schmieder P, Bacher A, Oschkinat H (1999) Application of amino acid type-specific 1H- and 14N-labeling in a 2H-, 15N-labeled background to a 47 kDa homodimer: potential for NMR structure determination of large proteins. J Biomol NMR 14(1):79–83

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(1):39–51

Kumar D, Paul S, Hosur RV (2010) BEST-HNN and 2D-(HN)NH experiments for rapid backbone assignment in proteins. J Magn Reson 204(1):111–117

Kupče E, Freeman R (2004) Projection-reconstruction technique for speeding up multidimensional NMR spectroscopy. J Am Chem Soc 126(20):6429–6440

Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. J Magn Reson 187(1):163–169

Marion D (2010) Combining methods for speeding up multidimensional acquisition. Sparse sampling and fast pulsing methods for unfolded proteins. J Magn Reson 206(1):81–87

Moreira MC, Barbot C, Tachi N, Kozuka N, Uchida E, Gibson T, Mendonça P, Costa M, Barros J, Yanagisawa T, Watanabe M, Ikeda Y, Aoki M, Nagata T, Coutinho P, Sequeiros J, Koenig M (2001) The gene mutated in ataxia-ocular apraxia 1 encodes the new HIT/Zn-finger protein aprataxin. Nat Genet 29(2):189–193

Mori S, Abeygunawardana C, Johnson MO, Vanzijl PCM (1995) Improved sensitivity of HSQC spectra of exchanging protons at short interscan delays using a new fast HSQC (FHSQC) detection scheme that avoids water saturation. J Magn Reson 108(1):94–98

Morita EH, Shimizu M, Ogasawara T, Endo Y, Tanaka R, Kohno T (2004) A novel way of amino acid-specific assignment in 1H-15N HSQC spectra with a wheat germ cell-free protein synthesis system. J Biomol NMR 30(1):37–45

Muchmore DC, McIntosh LP, Russell CB, Anderson DE, Dahlquist FW (1989) Expression and nitrogen-15 labeling of proteins for proton and nitrogen-15 nuclear magnetic resonance. Methods Enzymol 177:44–73

O'Grady C, Rempel BL, Sokaribo A, Nokhrin S, Dmitriev OY (2012) One-step amino acid selective isotope labeling of proteins in prototrophic *Escherichia coli* strains. Anal Biochem 426(2): 126–128

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(16): 5020–5021

Qiang W (2011) Signal enhancement for the sensitivity-limited solid state NMR experiments using a continuous, non-uniform acquisition scheme. J Magn Reson 213(1):171–175

Rasia RM, Brutscher B, Plevin MJ (2012) Selective isotopic unlabeling of proteins using metabolic precursors: application to NMR assignment of intrinsically disordered proteins. ChemBioChem 13(5):732–739

Rass U, Ahel I, West SC (2007) Actions of aprataxin in multiple DNA repair pathways. J Biol Chem 282(13):9469–9474

Rass U, Ahel I, West SC (2008) Molecular mechanism of DNA deadenylation by the neurological disease protein aprataxin. J Biol Chem 283(49):33,994–34,001

Schubert M (2001) MUSIC, selective pulses, and tuned delays: amino acid type-selective 1H–15N correlations, II. J Magn Reson 148(1):61–72

Schubert M, Smalla M, Schmieder P, Oschkinat H (1999) MUSIC in triple-resonance experiments: amino acid type-selective (1)H-(15)N correlations. J Magn Reson 141(1):34–43

Shortle D (1994) Assignment of amino acid type in 1H-15N correlation spectra by labeling with 14N-amino acids. J Magn Reson 105(1):88–90

Strauss A, Bitsch F, Cutting B, Fendrich G, Graff P, Liebetanz J, Zurini M, Jahnke W (2003) Amino-acid-type selective isotope labeling of proteins expressed in Baculovirus-infected insect cells useful for NMR studies. J Biomol NMR 26(4):367–372

Studier F, Moffatt B (1986) Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. J Mol Biol 189(1):113–130

Su XC, Loh CT, Qi R, Otting G (2011) Suppression of isotope scrambling in cell-free protein synthesis by broadband inhibition of PLP enymes for selective 15N-labelling and production of perdeuterated proteins in H2O. J Biomol NMR 50(1):35–42

Takeuchi K, Ng E, Malia TJ, Wagner G (2007) 1-13C amino acid selective labeling in a 2H15N background for NMR studies of large proteins. J Biomol NMR 38(1):89–98

Tanio M, Tanaka R, Tanaka T, Kohno T (2009) Amino acid-selective isotope labeling of proteins for nuclear magnetic resonance study: proteins secreted by Brevibacillus choshinensis. Anal Biochem 386(2):156–160

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2007) BioMagResBank. Nucleic Acids Res 36(Database):D402–D408

Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, Laue ED (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59(4):687–696

Waugh D (1996) Genetic tools for selective labeling of proteins with alpha-15N-amino acids. J Biomol NMR 8(2):184–192

Wittekind M, Mueller L (1993) HNCACB, a high-sensitivity 3D NMR experiment to correlate amide-proton and nitrogen resonances with the alpha-and beta-carbon resonances in proteins. J Magn Reson 101(2):201–205