Proc Indian Natn Sci Acad **80** No. 5 December 2014 pp. 1085-1097

DOI: 10.16943/ptinsa/2014/v80i5/47975

Research Paper

Identification of C-Terminal Neighbours of Residues that have Only One $^{1}H^{\beta}$ Attached to $^{13}C^{\beta}$: (Ile, Thr and Val)-Specific (3,2)D-CB(CACO)NNH Experiment

K CHANDRA[1,2], S PATEL[1,3], H S ATREYA[2] and KANDALA V R CHARY[1,3,*]
[1]*Tata Institute of Fundamental Research, Mumbai, India*
[2]*NMR Research Centre, Indian Institute of Science, Bangalore, India*
[3]*Tata Institute of Fundamental Research, Center for Interdisciplinary Sciences, Hyderabad, India*

(Received on 25 March 2014; Revised on 9 July 2014; Accepted on 1 September 2014)

Sequence specific resonance assignment is the first step in determining the three-dimensional (3D) structure of proteins. Several double and triple-resonance NMR experiments were proposed in this direction for achieving unambiguous assignments. However, these methodologies are non-trivial for proteins with molecular weight beyond 15 kDa. Common problems in this process are spectral overlaps and rapid relaxation rates of the nuclei that result in broad cross peaks. Further, Pro residues, which lack $^{1}H^{N}$, complicate the scenario. This led to identify maximum possible starting points along the polypeptide chain of a given protein to aid in the sequence-specific resonance assignment procedure. In this paper, we propose a fast NMR methodology namely Ile, Thr and Val (ITV) specific (3, 2) D-CB(CACO)NNH, which shows the spectral signatures of the C-terminal sequential neighbors of Ile, Thr and Val residues in highly resolved manner. This aids in the generation of additional starting points along a given polypeptide chain, apart from the known canonical Ala, Gly, Ser and Thr residues, and thus accelerate resonance assignment process significantly.

Key Words: Isotope Labeling; NMR; Automated Assignments; Sequence Specific Resonance Assignments

Introduction

Advent of labeling schemes to produce ^{13}C- or/and ^{15}N-enriched proteins paved the way to determine their 3D structures of medium size proteins (upto a M^{r} = 50 kDa) by 3D triple resonance NMR. In this endeavor, ^{1}H, ^{13}C and ^{15}N sequence-specific resonance assignments remain the first step. Till date, several double- and triple-resonance experiments were proposed to carry out these assignments in isotope-labeled proteins. However, these methodologies are non-trivial for proteins with molecular weights even beyond 15 kDa. Common problems in this process are spectral overlaps and rapid relaxation rates of the nuclei that result in broad spectral signatures. Further, Pro residues, which lack $^{1}H^{N}$, complicate the situation. Hence maximum possible number of starting points is required along the polypeptide chain of a given protein to facilitate the sequence-specific resonance assignment procedure.

It is a well-known fact that, spectral signatures of Ala, Gly, Ser and Thr residues are easily identifiable residues because of their characteristic $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts. The $^{13}C^{\alpha}$ chemical shifts of Gly resonate around 45.4±1.8 ppm, which is well separated from the $^{13}C^{\alpha}$ chemical shifts of the rest of the residues. On the other hand, the $^{13}C^{\beta}$ spins of Ala, Ser and Thr resonate with their characteristic

**Author for Correspondence: E-mail: chary@tifr.res.in; Tel: +91-22-2278 2489*

chemical shifts around 19.1±2.7, 63.7±2.4 and 69.6±4.7 ppm [see BioMagResBank (*www.bmrb.wisc.edu*)], respectively, thus leading to their unambiguous identification (Atreya *et al*., 2000; Atreya *et al*., 2002). Thus, in principle, four triple-resonance experiments HNCACB, CBCA(CO)NH, HNCO and HN(CA)CO, should suffice for the resonance assignment of $^{1}H^{N}$, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and $^{13}C'$ spins (Atreya *et al*., 2000; Atreya *et al*., 2002). However, in practice, the success rate in any automated resonance assignment procedure turns out to be low, primarily because of incomplete spectral signatures or/and spectral overlap. Hence, the quest has been to identify spectral signatures arising from residues other than Ala, Ser, Thr and Gly. Based on the analysis of BMRB data of 9000 odd proteins, the amino acid residues (Ala, Gly, Ser and Thr), which can be easily identified from their respective $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts constitute around 30% percent of the total residues and thus can act as various starting points in any manual/automated resonance assignment procedure. With further identification of Ile and Val residues, the percentage of starting points can be pushed upto 37%. Additionaly, new di- and tri-peptide segments are generated, that can be uniquely mapped onto the protein primary sequence. The uniqueness of such mapping increases from 68% to 98% in opting for the identification of tripeptide segments (Supplementary Table S1) (Supplementary Table S2). With this in the backdrop, residue specific labeling (Shi *et al*., 2004; Volk *et al*., 2008; Ye *et al*., 2005) or unlabeling (Atreya & Chary, 2001; Krishnarjuna *et al*., 2011) procedures were used for selective identification of various residue types. The disadvantage of these methodologies has been the requirement of more than one sample and the associated problem of amino-acid cross-metabolism. To overcome this, residue type selective triple resonance experiments (Dotsch *et al*., 1996a; Schubert *et al*., 2001a, 2001b, 2001c; Schubert *et al*., 1999) were proposed, requiring only one $^{13}C/^{15}N$ doubly labeled sample. For example, selective CBCA(CO)NH pulse sequences were designed to select the C-terminal sequential neighbours of residues without $^{13}C^{\gamma}$ (Ala, Asp, Asn, Cys, Gly and Ser) (Dotsch & Wagner, 1996) and residues which have only one $^{1}H^{\beta}$ attached to $^{13}C^{\beta}$ (Ile, Thr and Val) (Dotsch *et al*., 1996b). However, such simplified spectra could not be used for the identification of peaks as belonging to a specific residue type within the selected set of residues. As an extension of this methodology, we proposed an Ala, Asp, Asn, Cys, Gly, and Ser specific (3, 2)D-CB(CACO)NNH pulse sequence (Barnwal *et al*., 2008), where in the chemical shifts of the underlined nuclei were jointly sampled, to identify the peaks as arising from six amino-acid residues that lack $^{13}C^{\gamma}$ (Ala, Asp, Asn, Cys, Gly and Ser).

In furthering the quest to identify spectral signatures arising from residues other than those mentioned above, we propose here Ile, Thr and Val (ITV) specific (3, 2)-D CB(CACO)NNH pulse sequence, a rapid NMR methodology, that shows the spectral signatures of the C-terminal sequential neighbors of amino acid residues of Ile, Thr and Val, which have single $^{1}H^{\beta}$ attached to $^{13}C^{\beta}$ in a given protein. This experiment is a variation of the TVI-CBCACONH pulse sequence proposed earlier by Wagner and co-workers to edit the Ile, Thr and Val spin systems and it is based on the methodology of G-matrix Fourier transform (GFT) NMR spectroscopy as described below.

Results and Discussion

Fig. 1 shows the proposed radio-frequency (RF) pulse scheme, which we have called it as, ITV specific (3, 2)D-CB(CACO)NNH experiment. As described by Wagner and coworkers, the experiment starts with an INEPT transfer of magnetization from ^{1}H to their directly bonded ^{13}C spins, resulting in ^{13}C coherence, which is anti-phase with respect to its directly attached ^{1}H. During the following delay δ, this coherence is refocused with respect to proton coupling(s) ($2\tau_2$ in the pulse scheme presented here). After this delay δ, ^{13}C magnetization evolves into in-phase magnetization with respect to the directly attached protons. Ignoring all other couplings, five terms are generated with the following modulation coefficients:

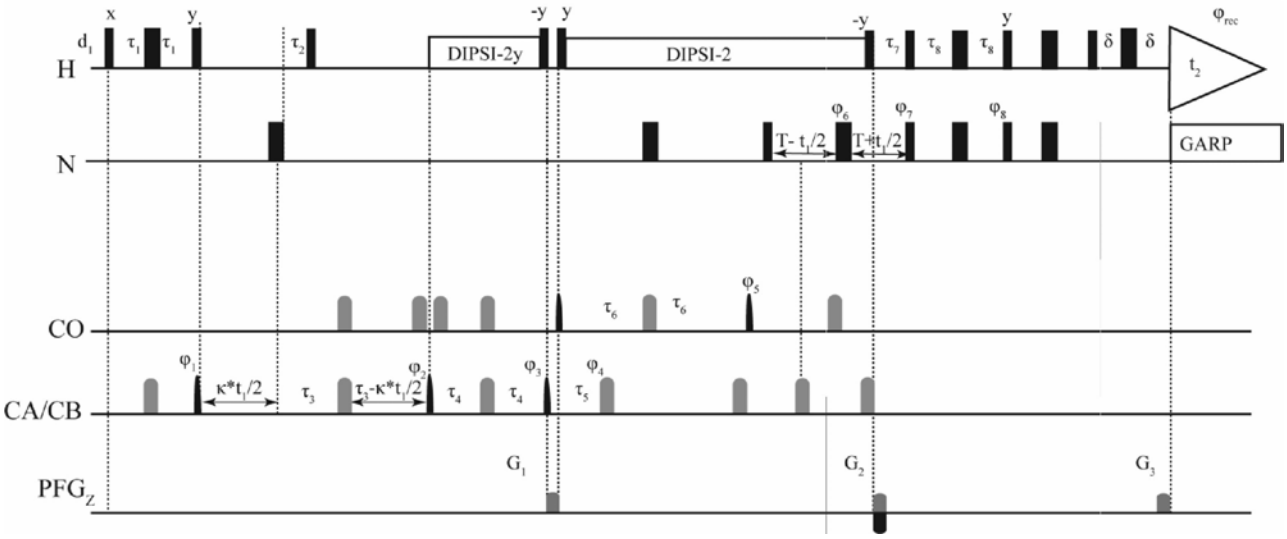


Fig. 1: Pulse sequence for ITV-specific (3, 2)D CB(CACO)NNH. Thin and thick vertical bars indicate rectangular 90° and 180° pulses, respectively, and phases are indicated above the pulses. Where no r.f. phase is marked, the pulse is applied along x axis. High-power 90° pulse lengths for ubiquitin are: 19.01 μs for ^{1}H, 45 μs for ^{15}N and 15 μs for ^{13}C. The scaling factor κ was either 0.5 or 1.0. DIPSI was employed to decouple ^{1}H during the hetero-nuclear magnetization transfers. The ^{1}H r.f. carrier is placed at the position of the solvent line (4.69 ppm). The ^{15}N carrier position is set at 118 ppm. The ^{13}C carrier is kept at 43 ppm throughout the experiment. For Hahellin, high-power 90° pulse length was 8.84 μs for ^{1}H and the ^{1}H r.f. carrier was placed at the position of the solvent line at 4.77 ppm. GARP was employed to decouple ^{15}N during acquisition. All pulsed z-field gradients (PFGs) are sine.1000 shaped with gradient recovery delay of 100 μs. The duration and strengths of the PFGs are: G1 (1 ms, 30 G/cm); G2 (1 ms, 80G/cm); G3 (1 ms, 8.1G/cm). The delays are: τ_1 = 2.75 ms, τ_2 = 1.9 ms, τ_3 = 4.5 ms, τ_4 = 7.2 ms, τ_5 = 4.4 ms, τ_6 = 12.4 ms, τ_7 = 5.5 ms, τ_8 = 2.3 ms and d_1 = 1 s. Phase cycling: φ_1 = x,-x; φ_2 = x; φ_3 = x, -x; φ_4 = (8x, -8x); φ_5 = x; φ_6 = (4x, -4x); φ_7=(2x, -2x); φ_8=(-2y, 2y) and φ_{rec} = 8(x, -x). Quadrature detection along ^{15}N is accomplished using the sensitivity-enhanced scheme. GFT NMR phase-cycle: φ_1 = x, y yields, in conjunction with quadrature detection in ^{15}N, two data sets, which are linearly combined employing a G-matrix transformation with the G-matrix. The number of complex points, t_{1max} and the digital resolution along the indirect dimension were 240, 7.4 ms, 67.5 Hz/pt for the experiment with the scaling factor (κ) = 0.5; and 360, 6.2 ms, 81.1 Hz/pt for the experiment with κ = 1. The same parameters were 2048; 127 ms; 3.91 Hz/pt for the ω_2 dimension in both the experiments with the total measurement time of 1 hr 34 mins

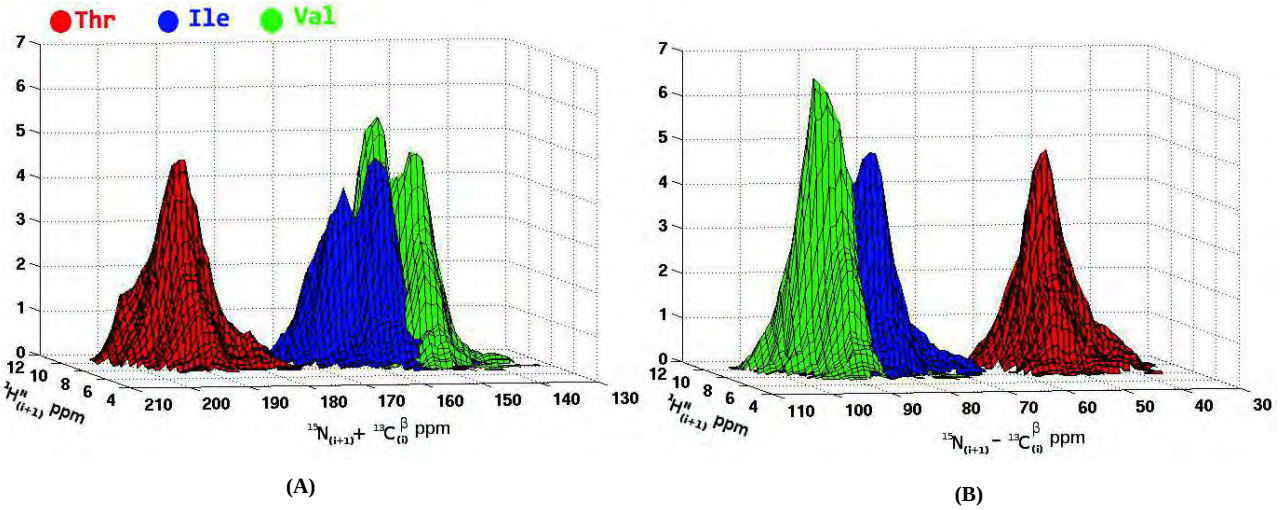
$$^{13}C^{\beta}[\sin(\pi J_{CH}\delta) + \sin(\pi J_{CH}\delta)\cos(\pi J_{CH}\delta) + \sin(\pi J_{CH}\delta)\cos^2(\pi J_{CH}\delta)] + {}^{13}C^{\alpha}[\sin(\pi J_{CH}\delta) + \sin(\pi J_{CH}\delta)\cos(\pi J_{CH}\delta)]$$

The first three terms arise from $^{13}C^{\beta}$ spins attached to a lone $^1H^{\beta}$ (as in Ile, Thr and Val), two $^1H^{\beta}$ protons and three $^1H^{\beta}$ protons (as in Ala), respectively. The last two terms arise from $^{13}C^{\alpha}$ spins attached to a lone $^1H^{\alpha}$ and two $^1H^{\alpha}$ protons (as in Gly), respectively. Thus, setting δ to $1/2^1J(^{13}C^{\beta}\text{-}^1H^{\beta})$ and manipulating specific delays (τ_2 and τ_3) in the pulse sequence, selection of the magnetization arising from the first term alone is possible as the corresponding magnetization is sine modulated. The signals arising from ^{13}C-^{1}H pairs of methylene and methyl groups are suppressed by selecting δ as $1/2^1J(^{13}C^{\beta}\text{-}^1H^{\beta})$ and also by employing proton DIPSI decoupling. The variation in the proposed pulse sequence, ITV specific (3, 2)D-CB(CACO)NNH from the earlier one TVI-CBCACONH, arises from the joint sampling of chemical shifts of the underlined nuclei ($^{13}C^{\beta}$ and ^{15}N) and hence the 3D spectral information is encoded in the form of a 2D spectrum rapidly. The phase-sensitive joint-sampling of $^{13}C^{\beta}$ and ^{15}N chemical shifts is achieved by co-incrementing their respective chemical-shift evolution periods, with the $^{13}C^{\beta}$ chemical shifts scaled by a tunable factor κ (chosen as 0.5 in the present experiment) relative to ^{15}N. Two 2D data sets are acquired by altering the phase of the first $^{13}C^{\alpha/\beta}$ pulse (φ_1) by 90°. Thus, among the two acquired datasets,

one corresponds to the cosine modulation, represented by $\cos(\Omega_{C\beta}t_1)$ [which is equal to $\kappa^*\{\exp(i\Omega_{C\beta}t_1) + \exp(-i\Omega_{C\beta}t_1)\}$, where $\Omega_{C\beta}$ is the chemical shift of $^{13}C^{\beta}$ of i^{th} residue], while the other dataset corresponds to the sine modulation, represented by $i(\sin(\Omega_{C\beta}t_1)$ [which is equal to $\kappa^*\{\exp(i\Omega_{C\beta}t_1)-\exp(-i\Omega_{C\beta}t_1)\}$], as the phase of the second data set is shifted by 90°. Thus, the G-matrix Fourier transformation of such data set results in two sub-spectra each comprising of peaks along the shared dimension (t_1) at a given linear combination of chemical shifts namely $\omega_1 = \{\Omega(^{15}N_{i+1}) \pm \kappa^*\Omega(^{13}C^{\beta}_i)\}$. An additional 2D [^{15}N-^{1}H] HSQC recorded with the same sample provides central peak information $\omega_1 = \Omega(^{15}N_{i+1})$, which is needed to analyze the data. Each cross-peak seen in the spectrum thus essentially arises from the $^{13}C^{\beta}$ spins of Ile, Thr and Val (i^{th} residue) and correlates to the ^{15}N/^{1}H pair of their C-terminal neighbour ($i+1^{th}$ residue). Thus, each correlation peak seen in the spectrum provides the chemical shift information of $^{13}C^{\beta}_i$ (of Ile, Thr and Val) and $(^{15}N/^{1}H)_{i+1}$. Further, the fact that the $^{13}C^{\beta}$ of Ile, Thr and Val have distinct chemical shifts (38.4±2.0, 69.5±2.4 and 32.4±1.8 ppm, respectively) as shown in the supplementary Fig. S1, it is straight forward to distinguish and assign each peak as arising from Ile, Thr and Val and hence in assigning the ^{15}N/^{1}H chemical shifts of their C-terminal neighbors.

We demonstrate the utility of the proposed ITV-specific (3, 2)D-CB(CACO)NH, by recording the spectra on uniformly ^{13}C/^{15}N doubly labeled (u-^{13}C/^{15}N-labeled) Ubiquitin sample (1 mM in concentration) and Ca^{2+}-bound form of Hahellin (0.8 mM), which is a putative member of βγ crystallin family, from a marine bacterium *Hahella chejuensis* (S. Patel, 2014; Srivastava & Chary, 2011; Srivastava *et al.*, 2008, 2010), an intrinsically unstructured protein in the absence of Ca^{2+} and undergoes a drastic conformational change upon Ca^{2+}-binding and acquires a typical βγ-crystallin fold (Srivastava & Chary, 2011; Srivastava *et al.*, 2008, 2010). NMR experiments with both the proteins were performed at 25°C on a Bruker 800 MHz NMR spectrometer equipped with a cryogenically cooled triple-resonance probe-head and pulse-field gradients. The total measurement times were 1 hr 34 min for Ubiquitin and 2 hr 05 min for Hahellin that included the recording of both the cosine and sine modulated data (Table S3). The data were processed using TOPSPIN (Version 2.1) and addition and subtraction were done in the frequency domain.

Peak patterns in ITV specific (3, 2)D-CB(CACO)NNH: We have simulated the two sub-



(A) **(B)**

Fig. 2: Simulation of the additive (A) and subtractive (B) sub-spectra of ITV-specific (3, 2)D CBCA(CO)NH spectrum, comprising of peaks at $(\omega_1, \omega_2) = [\Omega(^{13}C^{\beta}_i \pm {}^{15}N_{i+1}), \Omega(^{1}H^{N}_{i+1})]$

spectra of (3, 2)D CB(CACO)NNH (Figs. 2A and 2B), based on the statistical analysis of known chemical shift data of 365 proteins available in the BioMagResBank (*www.bmrb.wisc.edu*). As evident from the simulations presented here (Figs. 2A and 2B), phase-sensitive modulation (joint-sampling) of $^{13}C^{\beta}$ and ^{15}N chemical shifts achieved by co-incrementing their respective chemical-shift evolution periods, with the $^{13}C^{\beta}$ shifts scaled by a tunable factor κ (chosen as 0.5 in the present experiment) relative to ^{15}N, results in a significant resolution in the spectral signatures arising from Ile ($^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}$, $^{1}H^{N}{}_{i+1}$), Thr ($^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}$, $^{1}H^{N}{}_{i+1}$) and Val ($^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}$, $^{1}H^{N}{}_{i+1}$) correlations (Figs. 2A and 2B). As evident from Fig. 2A and 2B, both the additive and subtractive sub-spectra comprising of peaks at $(\omega_1, \omega_2) = [\Omega(^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}), \Omega(^{1}H^{N}{}_{i+1})]$, completely resolved the Thr $[\Omega(^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}), \Omega(^{1}H^{N}{}_{i+1})]$ correlations, though the spectral signatures arising from Ile $[\Omega(^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}), \Omega(^{1}H^{N}{}_{i+1})]$ and Val $[\Omega(^{13}C^{\beta}{}_{i}\pm^{15}N_{i+1}), \Omega(^{1}H^{N}{}_{i+1})]$ marginally overlap. However, as mentioned above once the $^{13}C^{\beta}$ chemical shift values are determined, it is easy to distinguish and assign these resonances unambiguously as arising from Ile, Thr or Val residues as their respective $^{13}C^{\beta}$ chemical shifts are unique (Fig. S1).

The panels A and B in Figs. 3 and 4 show the additive and subtractive (3, 2)D-CB(CACO)NNH spectra for Ubiquitin and Hahellin (Srivastava & Chary, 2011; Srivastava *et al.*, 2008, 2010), respectively, of cosine- and sine-modulated sub-spectral data. Each peak in the spectrum is characterized by the chemical shift information of $^{13}C^{\beta}{}_{i}$, $^{1}H^{N}{}_{i+1}$ and $^{15}N_{i+1}$. From the information about the $^{13}C^{\beta}{}_{i}$ thus derived, it is easy to mark each peak as arising specifically from Val, Ile or Thr as mentioned earlier. It is worth mentioning here that the additional peaks that are observed in the spectrum are those arising from Ala residues, though the delay δ is tuned to $(1/2)^{1}J(^{13}C^{\beta}\text{-}^{1}H^{\beta})$, for the selection of the magnetization arising from Ile, Thr and Val residues. The break-through peaks arising from Ala are in phase with those arising from Ile, Thr and Val, because of the cosine-square dependence of their β-carbon (CH_3 group) magnetization. These spectral signatures do not cause any problem in spectral analysis, as they can be readily distinguished from their unique $^{13}C^{\beta}$ chemical shifts and their appearance in a non-overlapping region of the (3, 2)D-CB(CACO)NNH spectrum as shown in Fig. 3. During the pulse sequence, the magnetization arising from $^{13}C^{\alpha}$ spins are also selected as they are sine modulated with respect to $^{1}J(^{13}C^{\alpha}\text{-}^{1}H^{\alpha})$ and later using a delay, which is almost equal to $1/2^{1}J(^{13}C^{\alpha}\text{-}^{13}C')$ $(2\tau_3)$. However, the optimized value for this delay was found to be 8.8 ms, which is much less than $1/2^{1}J(^{13}C^{\alpha}\text{-}^{13}C')$ and this explains as the reason as to why the breakthrough peaks arise from $^{13}C^{\alpha}$ nuclei. However, these spectral signatures can be easily distinguished from those arising from Ala, Ile, Thr and Val residues based on their chemical shift values.

It is worth mentioning here that, in principle, one can record a cosine modulated ITV-specific (3, 2)D-CB(CACO)NNH spectrum alone and a 2D [^{15}N-^{1}H]-HSQC, to derive the information about the chemical shifts of $^{13}C^{\beta}$ and $^{1}H^{N}/^{15}N$ of all the Ile, Thr and Val residues and their C-terminal neighbors, respectively, present in a given protein. Fig. 3C shows the cosine modulated ITV-specific (3, 2)D-CB(CACO)NNH sub-spectrum of Ubiquitin. For each individual amide-pair $\{\Omega(^{1}H_{i+1})/\Omega(^{15}N_{i+1})\}$ observed along the ω_1 dimension, two peaks are seen at $\{\Omega(^{15}N_{i+1}) \pm \kappa^{*}(\gamma_C/\gamma_N)^{*}\Omega(^{13}C^{\beta}{}_{i}\text{-}^{13}C_{offset})\}$. Thus, with the implicit knowledge of the $^{13}C_{offset}$, one can derive the information of the $\Omega(^{15}N_{i+1})/\Omega(^{1}H_{i+1})$for each observed amide-pair and that of $\Omega(^{13}C^{\beta}{}_{i})$ belonging to Ala, Ile, Thr and Val residues. The information of $^{1}H^{N}{}_{i+1}$ and $^{15}N_{i+1}$ chemical shifts thus derived help in an unambiguous identification the C-terminal neighbours of these four residues and hence can be mapped on to the 2D [^{15}N-^{1}H]-HSQC as shown in the supplementary Figs. S2A (for Ubiquitin) and S2B (for Hahellin), respectively.

Additionally, the ITV-specific (3, 2)D-CB(CACO)NNH experiment described here has added advantage of substantially reduced acquisition

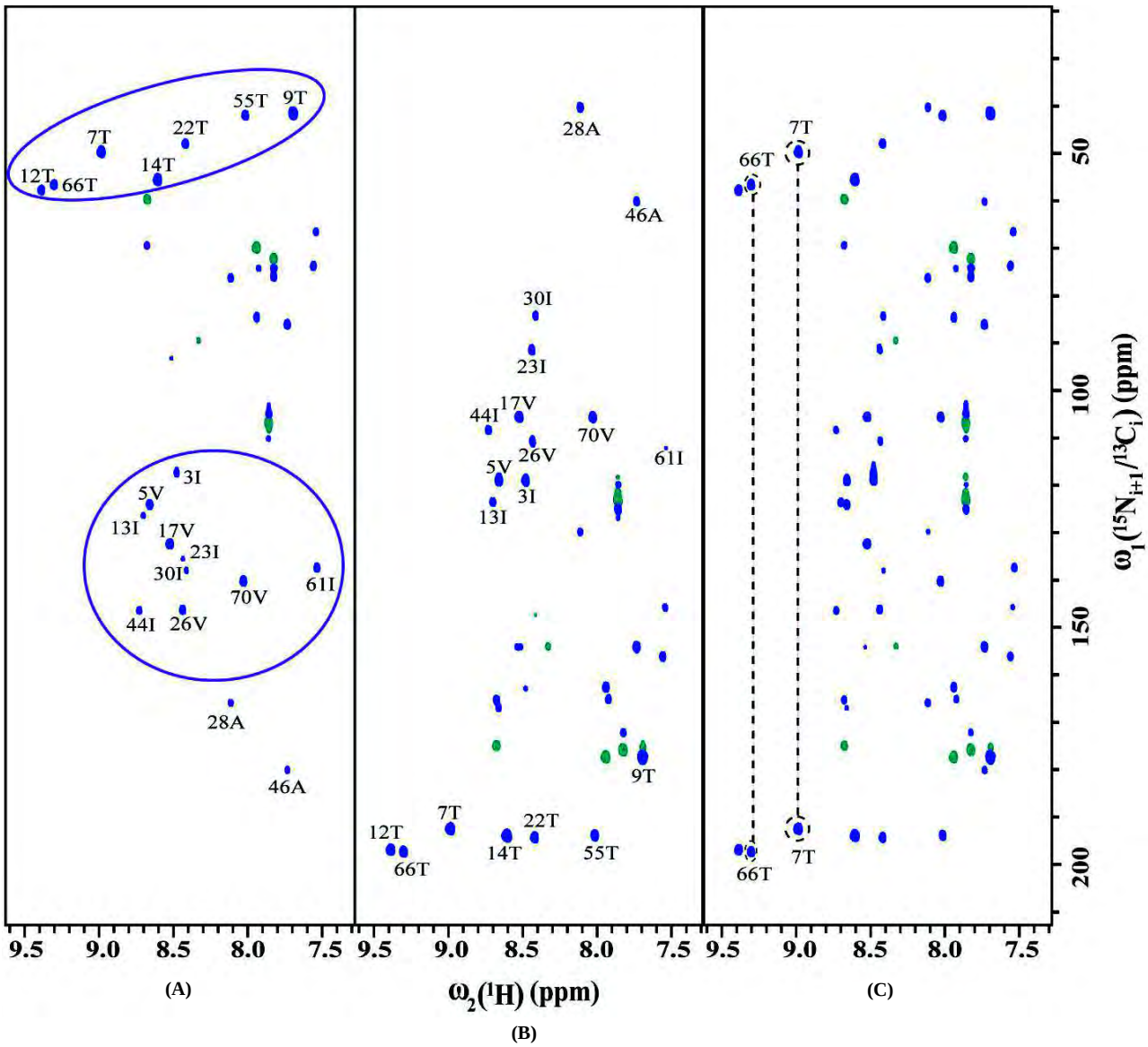


Fig. 3: (A) Subtracted ITV-specific (3, 2)D-CBCA(CO)NNH spectrum of ubiquitin. The assignments are labeled on the spectra. (B) Added ITV-specific (3, 2)D-CBCA(CO)NNH spectrum of Ubiquitin with the assignments labeled. (C) Cosine modulated ITV-specific (3, 2)D-CBCA(CO)NNH spectrum (φ_1 = x, -x) of Ubiquitin. Along the ^{15}N-dimension for each amide-pair ($\Omega\ ^1H_{i+1}$, $\Omega\ ^{15}N_{i+1}$), two peaks show-up at ($\Omega\ ^{15}N_{i+1}$) ± 0.5*γ_C/γ_N*($\Omega\ ^{13}C_i$ - $\Omega\ ^{13}C_{offset}$). The peak pairs corresponding to two Thr are highlighted. The centre of the each peak pair derives the chemical shift of $^{15}N_{i+1}$, which in turn help in determining the corresponding $^1H_{i+1}$ chemical shift. The knowledge of $^{15}N_{i+1}$ and $^1H_{i+1}$ chemical shifts can then be easily mapped on to the corresponding 2D [^{15}N, ^{1}H] HSQC of the protein

time and flexible scaling factor to resolve overlap problems (if any) by increasing the dispersion of peaks and thus facilitating a fairly high spectral resolution, which is depicted in the supplementary Fig. S3. Further, the proposed ITV-specific (3, 2)D-CB(CACO)NNH experiment has several other advantages, which is characteristic of GFT-NMR (Atreya *et al*., 2007; Atreya & Szyperski, 2004, 2005; Barnwal *et al*., 2008; Barnwal *et al*., 2008; Barnwal *et al*., 2007; Kim & Szyperski, 2003; Rout *et al*., 2010) or RD-NMR(Atreya & Szyperski, 2005; Chandra *et al*., 2012). For example, for Ubiquitin both cosine- and sine-modulated ITV-specific (3, 2)D-CB(CACO)NNH experiments were recorded with a resolution of 67.5 Hz/pt along the shared (^{15}N/^{13}C) dimension, which is spread over 200 ppm. With the

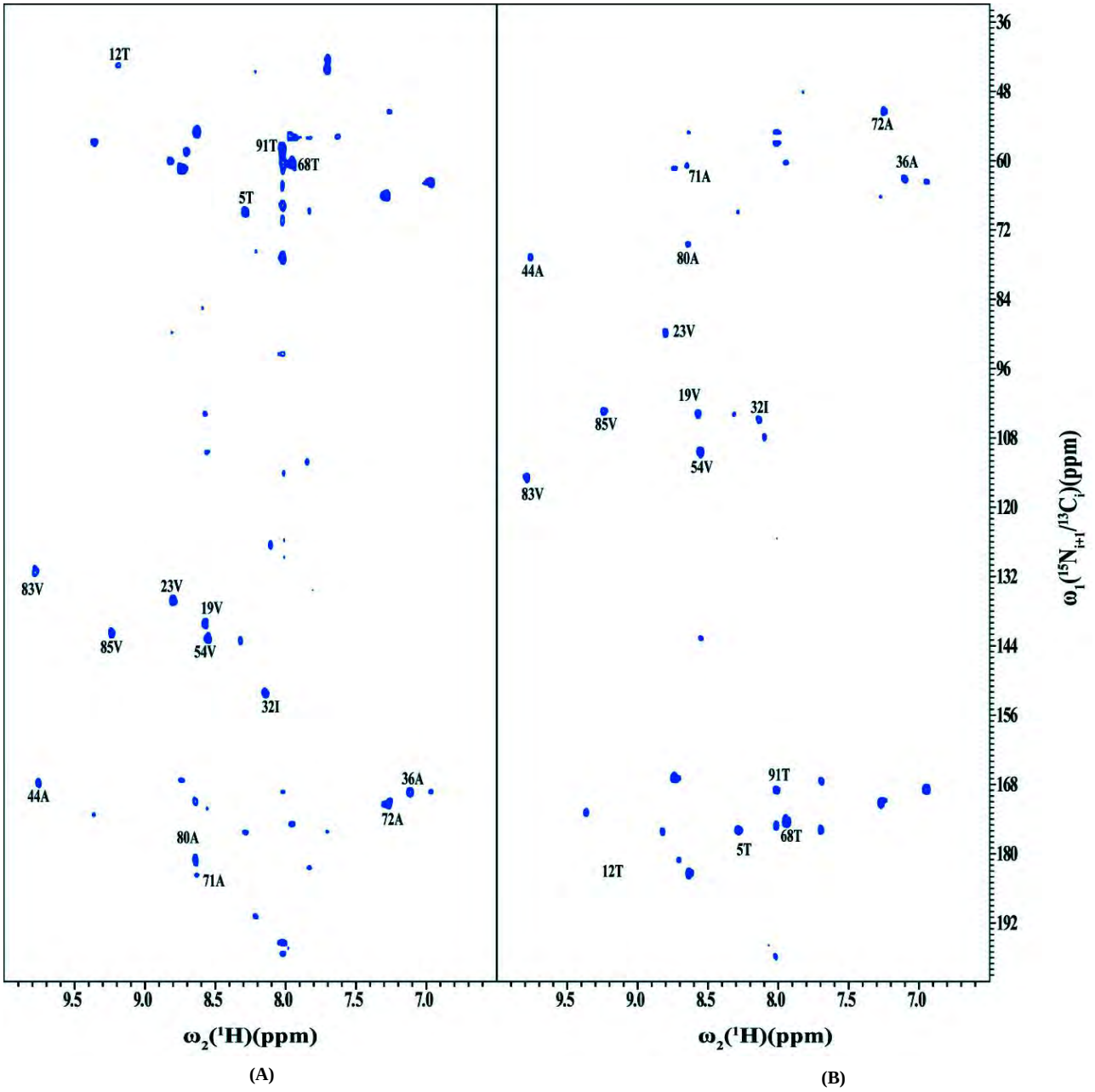


Fig. 4: (A) Subtracted ITV-specific (3, 2)D-CBCA(CO)NNH spectrum of Hahellin. The assignments are labeled on the spectrum. (B) Added ITV-specific (3, 2)D-CBCA(CO)NNH spectrum of Hahellin with the assignments labeled. Along the ^{15}N-dimension for each amide-pair ($\Omega\ ^{1}H_{i+1}$, $\Omega\ ^{15}N_{i+1}$), two peaks show-up at ($\Omega\ ^{15}N_{i+1}$) ± 0.5*γ_C/γ_N *($\Omega\ ^{13}C_i$ - $\Omega\ ^{13}C_{offset}$)

number of scans set to 12 for each experiment, the total experimental time was 1 hr 34 min. For Hahellin, both cosine- and sine-modulated ITV-specific (3, 2)D-CB(CACO)NNH experiments were recorded with a number of scans set to 16 and the total experimental time was 2 hr 5 min. In comparison, to record a normal 3D-CB(CACO)NNH with 40.5 Hz/pt resolution along ^{15}N-dimension (acquisition time 12.3 ms) and 67.3 Hz/pt resolution along ^{13}C-dimension (acquisition time 7.4 ms) one needs a minimum of 14 hr of experimental time even for a two scan experiment. The fact that the overall experimental time is substantially reduced in a GFT experiment, we chose the number of scans as 12/16 to compensate

the sensitivity issue, which resulted in an excellent average S/N ratio of 21 for Ubiquitin and 19 for Hahellin. Supplementary Figs. S4A and 4B show the histogram of the observed S/N ratio as a function of percentage of residues of Ubiquitin and Hahellin, respectively. Thus, the present dataset can be recorded with even less number of scans to further reduce the experimental time, and still get all the expected spectral information. This method can be combined with TROSY approach (Salzmann *et al*., 1998, 1999; Yang & Kay, 1999) for application to large systems (> 30 kDa) to overcome the line-broadening problems. Adoption of methods like longitudinal relaxation optimized technique and non-uniform sampling for fast data collection (Hyberts *et al*., 2012; Hyberts *et al*., 2012; Kazimierczuk *et al*., 2012) can easily be employed in the present case, which can further reduce the experimental time by an order of magnitude. Since the data is acquired in the form of a 2D spectrum, high spectral/digital resolution can still be achieved easily without spending much instrument time. Most importantly, there is no need to collect TVI-selective 3D-CBCA(CO)NH to retrieve the information about the corresponding C-terminal amide pairs.

Thus, we could demonstrate the utility of the proposed experiment on u-$^{13}C/^{15}N$-labeled Ubiquitin and Hahellin samples and identify the chemical shifts of all the $^{13}C^{\beta}$ spins belonging to Ala, Ile, Thr and Val and correlate them with their $^{15}N/^{1}H$ pair belonging to their C-terminal neighbors. This experiment can aid in and accelerate resonance assignment process and increase the precision of assignment in automated assignment strategies like TATAPRO (Atreya *et al*., 2002; Atreya *et al*., 2000), as it will help in the generation of additional reference points apart from the known conventional ones, namely Ala, Gly, Ser and Thr residues.

Acknowledgement

The facilities provided by the National Facility for High Field NMR, supported by the Department of Science and Technology (DST), New Delhi, Department of Biotechnology, New Delhi, Council of Scientific and Industrial Research, New Delhi, and Tata Institute of Fundamental Research, are gratefully acknowledged. The help provided by Abhinav Dubey in preparing the supplementary tables is gratefully acknowledged. KVRC thanks DST for JC Bose Fellowship. KC thanks DST for INSPIRE Faculty Schemes.

References

Atreya HS and Chary KVR (2001). Selective 'unlabeling' of amino acids in fractionally ^{13}C labeled proteins: An approach for stereospecific NMR assignments of CH_3 groups in Val and Leu residues. *J Biomol NMR* **19**(3) 267-272

Atreya HS, Chary KVR and Govil G (2002). Automated NMR assignments of proteins for high throughput structure determination: TATAPRO II. *Current Science* **83**(11) 1372-1376

Atreya HS, Garcia E, Shen Y and Szyperski T (2007). J-GFT NMR for precise measurement of mutually correlated nuclear spin-spin couplings. *J Am Chem Soc* **129**(3) 680-692

Atreya HS, Sahu SC, Chary KV and Govil G (2000). A tracked approach for automated NMR assignments in proteins (TATAPRO). *J Biomol NMR* **17**(2) 125-136

Atreya HS and Szyperski T (2004). G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignment. *PNAS* **101**(26) 9642-9647

Atreya HS and Szyperski T (2005). Rapid NMR data collection. *Nuclear Magnetic Resonance of Biological Macromolecules, Part C* **394** 78-108

Barnwal RP, Atreya HS and Chary KV (2008). Chemical shift based editing of CH_3 groups in fractionally ^{13}C-labelled proteins using GFT (3, 2)D CT-HCCH-COSY: stereospecific assignments of CH_3 groups of Val and Leu residues. *J Biomol NMR* **42**(2) 149-154

Barnwal RP, Rout AK, Atreya HS and Chary KV (2008). Identification of C-terminal neighbours of amino acid residues without an aliphatic ^{13}Cgamma as an aid to NMR assignments in proteins. *J Biomol NMR* **41**(4) 191-197

Barnwal RP, Rout AK, Chary KV and Atreya HS (2007). Rapid measurement of $^{3}J(^{1}H^{N}\text{-}^{1}H^{\alpha})$ and $^{3}J(^{15}N\text{-}H^{\beta})$ coupling constants in polypeptides. *J Biomol NMR* **39**(4) 259-263

Chandra K, Jaipuria G, Shet D and Atreya HS (2012). Efficient

sequential assignments in proteins with reduced dimensionality 3D HN(CA)NH. *J Biomol NMR* **52**(2) 115-126

Dotsch V, Oswald RE and Wagner G (1996a). Amino-acid-type-selective triple-resonance experiments. *J Magn Reson Series B* **110**(1) 107-111

Dotsch V, Oswald RE and Wagner G (1996b). Selective identification of threonine, valine and isoleucine sequential connectivities with a TVI-CBCACONH experiment. *J Magn Reson Series B* **110**(3) 304-308

Dotsch V and Wagner G (1996). Editing for amino-acid type in CBCACONH experiments based on the $^{13}C^{\beta}$-^{13}C gamma coupling. *J Magn Reson Series B* **111**(3) 310-313

Hyberts SG, Arthanari H and Wagner G (2012). Applications of non-uniform sampling and processing. *Top Curr Chem* **316** 125-148

Hyberts SG, Arthanari H and Wagner G (2012). Applications of Non-Uniform Sampling and Processing. *Novel Sampling Approaches in Higher Dimensional NMR* **316** 125-148

Kazimierczuk K, Misiak M, Stanek J, Zawadzka-Kazimierczuk A and Kozminski W (2012). Generalized Fourier Transform for Non-Uniform Sampled Data. *Novel Sampling Approaches in Higher Dimensional NMR* **316** 79-124

Kim S and Szyperski T (2003). GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. *J Am Chem Soc* **125**(5) 1385-1393

Krishnarjuna B, Jaipuria G, Thakur A, D'Silva P and Atreya HS (2011). Amino acid selective unlabeling for sequence specific resonance assignments in proteins. *J Biomol NMR* **49**(1) 39-51

Rout AK, Barnwal RP, Agarwal G and Chary KV (2010). Root-mean-square-deviation-based rapid backbone resonance assignments in proteins. *Magn Reson Chem* **48**(10) 793-797

S Patel AKS, V Ramanujan and Chary KVR (2014). Conformational Propensities and Dynamics of a βγ-crystallin, an Intrinsically Disordered Protein *Phys Chem Chem Phys*

Salzmann M, Pervushin K, Wider G, Senn H and Wuthrich K (1998). TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. *Proc Natl Acad Sci USA* **95**(23) 13585-13590

Salzmann M, Pervushin K, Wider G, Senn H and Wuthrich K (1999). [^{13}C]-constant-time [^{15}N,^{1}H]-TROSY-HNCA for sequential assignments of large proteins. *J Biomol NMR* **14**(1) 85-88

Schubert M, Oschkinat H and Schmieder P (2001a). Amino acid type-selective backbone ^{1}H-^{15}N correlations for Arg and Lys. *J Biomol NMR* **20**(4) 379-384

Schubert M, Oschkinat H and Schmieder P (2001b). MUSIC and aromatic residues: Amino acid type-selective ^{1}H-^{15}N correlations, III. *J Magn Reson* **153**(2) 186-192

Schubert M, Oschkinat H and Schmieder P (2001c). MUSIC, selective pulses, and tuned delays: Amino acid type-selective ^{1}H-^{15}N correlations, II. *J Magn Reson* **148**(1) 61-72

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

Shi J X, 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**(3) 235-247

Srivastava A K and Chary K V R (2011). Conformational heterogeneity and dynamics in a beta gamma-Crystallin from Hahella chejuensis. *Biophysical Chemistry* **157**(1-3) 7-15

Srivastava A K, Sharma Y and Chary K V R (2008). Sequence specific ^{1}H, ^{13}C, and ^{15}N resonance assignments of Hahellin from Hahella chejuensis, a putative member of the beta gamma-crystallin superfamily. *Biomol NMR Assign* **2**(2) 151-153

Srivastava A K, Sharma Y and Chary K V R (2010). A Natively Unfolded beta gamma-Crystallin Domain from Hahella chejuensis. *Biochemistry* **49**(45) 9746-9755

Volk J, Herrmann T and Wuthrich K (2008). Automated sequence-specific protein NMR assignment using the memetic algorithm MATCH. *J Biomol NMR* **41**(3) 127-138

Yang D and Kay L E (1999). Improved $^{1}H^{N}$-detected triple resonance TROSY-based experiments. *J Biomol NMR* **13**(1) 3-10

Ye Y M, Lee H W, Yang W, Shealy S and Yang J J (2005). Probing site-specific calmodulin calcium and lanthanide affinity by grafting. *J Am Chem Soc* **127**(11) 3743-3750.

Supplement

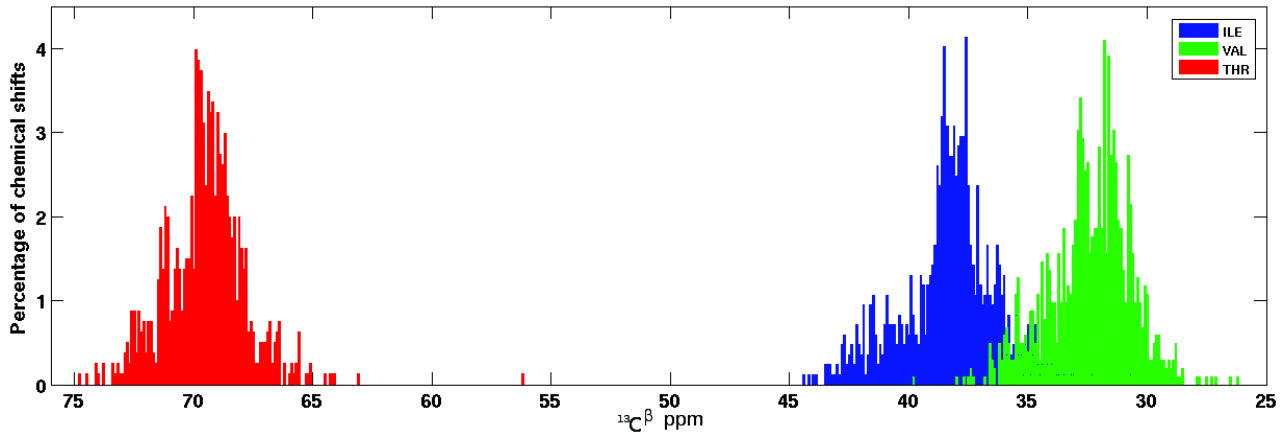
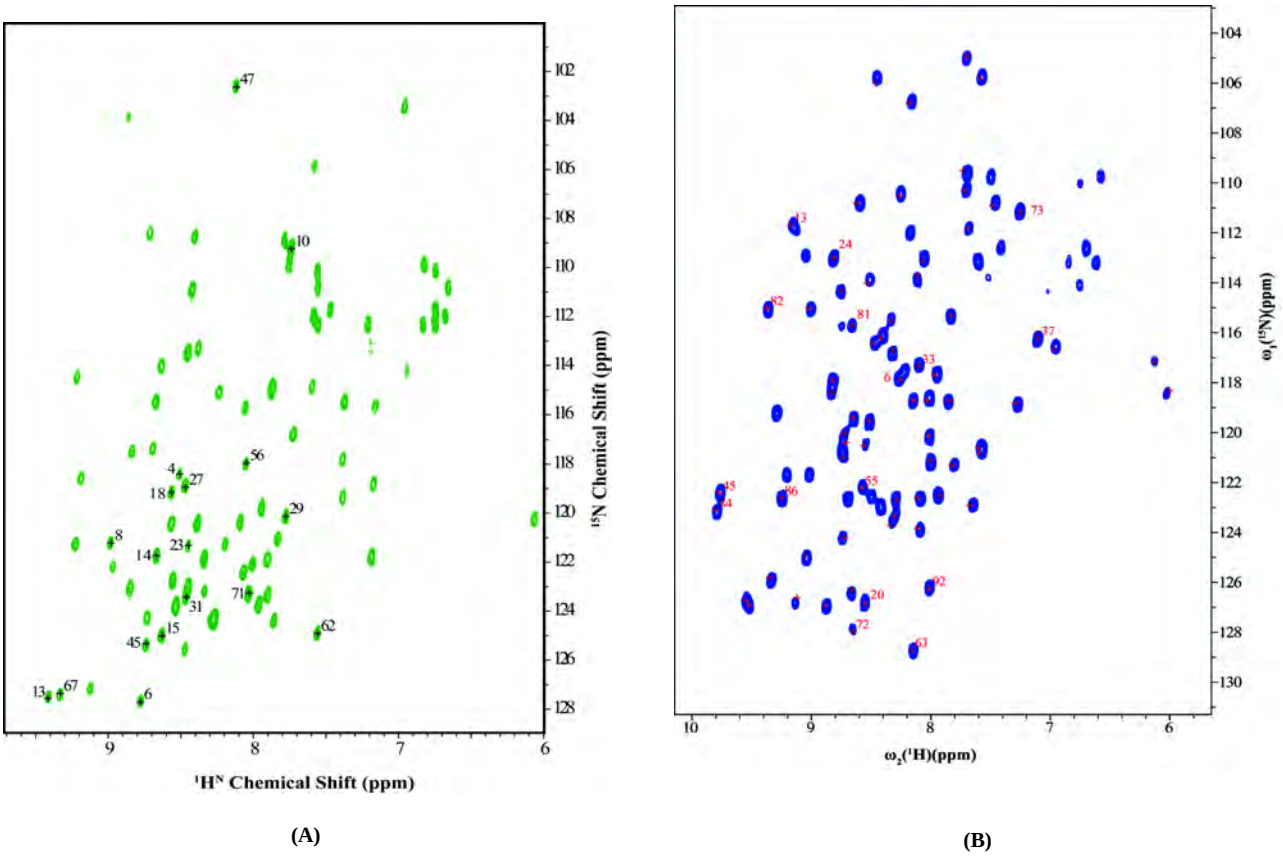


Fig. S1: Histogram of $^{13}C^{\beta}$ chemical shifts of Ile, Thr and Val residues of 365 proteins available in the BioMagResBank (BMRB; http://www.bmrb.wisc.edu)



(A) **(B)**

Fig. S2: 2D [^{15}N-^{1}H]-HSQC of Ubiquitin (A). The C-terminal neighbors of Val, Ile and Thr residues for which $^{1}H^{N}$ and ^{15}N chemical shifts could be obtained from the cosine modulated ITV-specific (3, 2)D CBCA(CO)NNH sub-spectrum are shown with their sequence number in black. 2D [^{15}N-^{1}H]-HSQC of Hahellin (B). The C-terminal neighbours of Val, Ile and Thr residues for which $^{1}H^{N}$ and ^{15}N chemical shift could be obtained from the cosine modulated ITV-specific (3, 2)D CBCA(CO)NNH sub-spectrum are shown with their sequence number in red

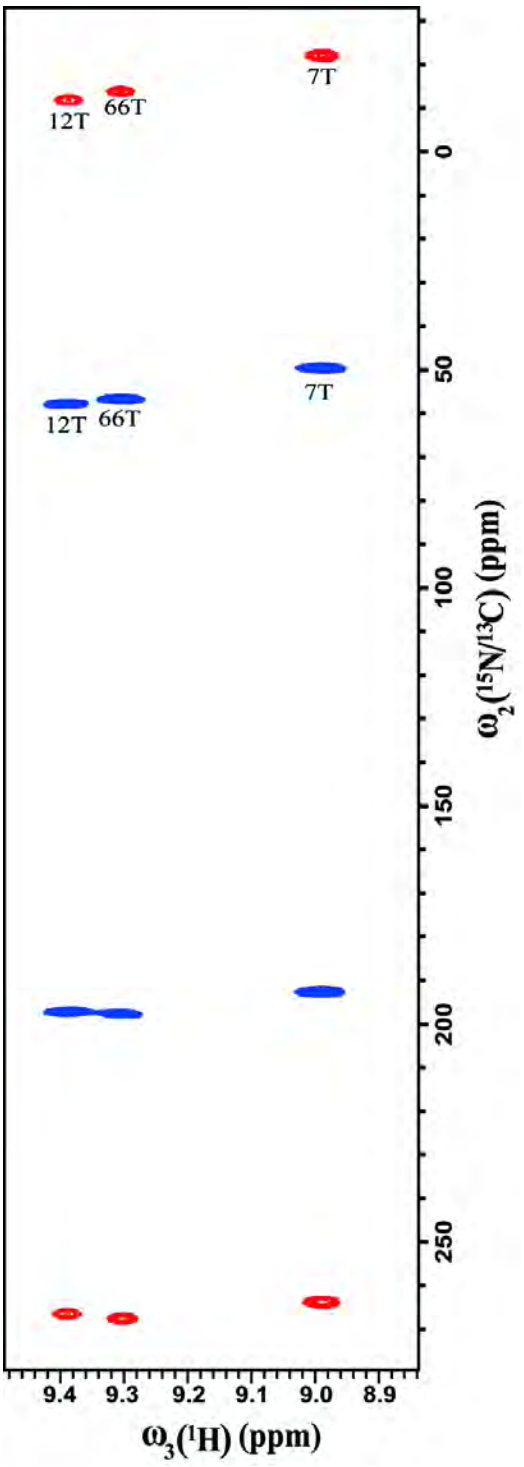


Fig. S3: The cosine modulated ITV-specific (3, 2)D CBCA(CO)NNH sub-spectrum of Ubiquitin recorded with two scaling factors, 0.5 (shown in blue) and 1 (red). Increase in the chemical shift difference along the shared dimension between the two peaks corresponding to T7, T12 and T66, from 1.13 to 2.09 ppm, with the increased scaling factor (0.5 to 1) is highlighted. Such scaling resolves partially overlapping peaks

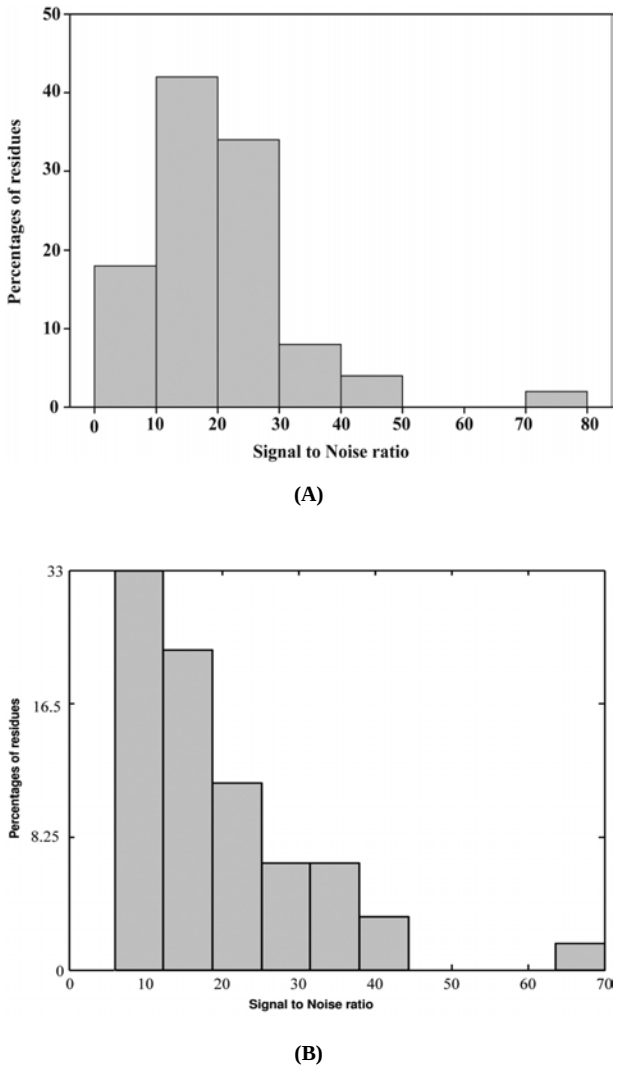


Fig. S4: Histogram displaying the signal to noise ratio versus the percentages of residues for Ubiquitin (A). Histogram displaying the signal to noise ratio versus the percentages of residues for Hahellin (B)

Table S1: This table depicts 12 possible tri-peptide segments among which one residue is either Val/Ile/Thr. For each case, individual dipeptide stretch was searched for each of the 9128 protein sequences and a total number of occurrences and percent of the occurrences averaged over all the proteins are reported. The uniqueness is calculated from the histogram and represents the percentage of single occurrence over all the sequences

VIT	VTI	IVT	ITV	TIV	TVI	Tri-peptide
0.016	0.017	0.016	0.022	0.011	0.023	Average % of occurrence
98.36	98.15	98.53	97.52	98.66	98.17	% of Uniqueness
153	177	135	229	122	170	Total number of occurrence
IAG	IGA	TAS	TSA	TAG	TGA	Tri-peptide
0.035	0.035	0.022	0.027	0.018	0.019	Average % of occurrence
97.50	97.84	97.43	97.49	98.08	97.89	% of Uniqueness
241	200	239	232	182	200	Total number of occurrence

Table S2A: Table S2A and S2B lists out 30 different dipeptides. Table S2A depicts the scenario when i^{th} residue is either Val/Ile/Thr and $i+1^{th}$ residue is any of the six amino acid residues (ASGVIT). Table S2B depicts the scenario when $i+1^{th}$ residue is either Val/Ile/Thr and i^{th} residue is any of the six amino acid residues (ASGVIT). The segments, which are common in Table S2A are omitted

	A	S	G	V	I	T	Amino acid
V	VA	VS	VG	VV	VI	VT	Dipeptide
	0.423	0.349	0.414	0.404	0.297	0.345	Average % of occurrence
	68.73	72.46	71.39	89.76	7.85	71.14	% of Uniqueness
	4189	3515	3575	4100	2845	3614	Total number of occurrence
I	IA	IS	IG	IV	II	IT	Dipeptide
	0.337	0.292	0.322	0.283	0.289	0.275	Average % of occurrence
	74.22	75.56	95.05	77.13	75.25	89.52	% of Uniqueness
	3108	2819	2632	2684	2441	2805	Total number of occurrence
T	TA	TS	TG	TV	TI	TT	Dipeptide
	0.339	0.293	0.417	0.386	0.248	0.302	Average % of occurrence
	73.41	76.82	71.28	69.62	78.42	78.62	% of Uniqueness
	3186	2656	3555	3814	2452	2643	Total number of occurrence

Table S2B:

	V	V	V	V	V	Amino acid
A	AV	SV	GV	IV	TV	Dipeptide
S	0.435	0.311	0.446	0.283	0.386	Average % of occurrence
G	69.15	73.86	68.87	77.14	69.62	% of Uniqueness
I	4069	3091	4005	2684	3814	Total number of occurrence
T						
	I	I	I	I		Amino acid
A	AI	SI	GI	TI		Dipeptide
S	0.383	0.269	0.388	0.248		Average % of occurrence
G	74.38	78.86	72.06	78.42		% of Uniqueness
T	3109	2429	3299	2452		Total number of occurrence
	T	T	T			Amino acid
A	AT	ST	GT			Dipeptide
S	0.352	0.314	0.396			Average % of occurrence
G	72.32	75.16	70.59			% of Uniqueness
	3388	2858	3593			Total number of occurrence

Table S3: Acquisition parameters of triple-resonance NMR experiments

Protein	**Experiment** (Scaling factor 0.5)	**Indirect dimension: Complex points; t_{max} (ms); Digital Resolution (Hz/Pt)**	**Measurement Time**
Ubiquitin	(3, 2)D-ITV-specific CB(CACO)NNH	$\omega_1(^{15}N/^{13}C)$: 240; 7.4; 67.5 $\omega_2(^{1}H)$: 2048; 127; 3.91 (800 MHz)	1 hr 34 min
	(3, 2)D-ITV-specific CB(CACO)NNH	$\omega_1(^{15}N/^{13}C)$: 360; 6.2; 81.1 $\omega_2(^{1}H)$: 2048; 127; 3.91 (800 MHz)	1 hr 34 min
Hahellin	(3, 2)D-ITV-specific CB(CACO)NNH	$\omega_1(^{15}N/^{13}C)$: 240; 7.4; 67.5 $\omega_2(^{1}H)$: 2048; 127; 3.91 (800 MHz)	2 hr 5 min