J Biomol NMR (2009) 44:185–194
DOI 10.1007/s10858-009-9327-x

ARTICLE

CSSI-PRO: a method for secondary structure type editing, assignment and estimation in proteins using linear combination of backbone chemical shifts

Monalisa Swain · Hanudatta S. Atreya

Received: 13 March 2009 / Accepted: 7 May 2009 / Published online: 16 June 2009

Abstract Estimation of secondary structure in polypeptides is important for studying their structure, folding and dynamics. In NMR spectroscopy, such information is generally obtained after sequence specific resonance assignments are completed. We present here a new methodology for assignment of secondary structure type to spin systems in proteins directly from NMR spectra, without prior knowledge of resonance assignments. The methodology, named Combination of Shifts for Secondary Structure Identification in Proteins (CSSI-PRO), involves detection of specific linear combination of backbone $^{1}H^{\alpha}$ and $^{13}C'$ chemical shifts in a two-dimensional (2D) NMR experiment based on G-matrix Fourier transform (GFT) NMR spectroscopy. Such linear combinations of shifts facilitate editing of residues belonging to α-helical/β-strand regions into distinct spectral regions nearly independent of the amino acid type, thereby allowing the estimation of overall secondary structure content of the protein. Comparison of the predicted secondary structure content with those estimated based on their respective 3D structures and/or the method of Chemical Shift Index for 237 proteins gives a correlation of more than 90% and an overall rmsd of 7.0%, which is comparable to other biophysical techniques used for structural characterization of proteins. Taken together, this methodology has a wide range of applications in NMR spectroscopy such as rapid protein structure determination, monitoring conformational changes in protein-folding/ligand-binding studies and automated resonance assignment.

Keywords Protein secondary structure · CSI · GFT NMR · Protein folding · 3D structure

Introduction

The process of determining high-resolution three-dimensional (3D) structure of proteins by NMR spectroscopy is often preceded by the elucidation of its various secondary structural elements (Wüthrich 1986). Estimation of secondary structure is also carried out in studies investigating protein folding pathways (Dyson and Wright 2004), proteomics projects where the feasibility of carrying out structural studies of proteins is assessed based on its 'foldability' (Montelione et al. 2000; Page et al. 2005) and automated assignment approaches where the information on secondary structure is used for spin-system linking and mapping (Choy et al. 1997; Baran et al. 2004). There are currently three methods used widely to obtain this information: (1) the method of chemical shift index (CSI), wherein, the deviation of the observed chemical shift of a given residue from its random-coil value provides information on its secondary structure (Wishart and Sykes 1994), (2) the method based on the observation of specific pattern/intensity of cross peaks in the NMR spectrum obtained using nuclear Overhauser effect spectroscopy (NOESY) (Wüthrich 1986) and (3) the value of three bond scalar coupling constant ($^{3}J_{HNH\alpha}$) involving backbone dihedral angle, φ, which correlates with the secondary structure

Electronic supplementary material The online version of this article (doi:10.1007/s10858-009-9327-x) contains supplementary material, which is available to authorized users.

M. Swain · H. S. Atreya (✉)
NMR Research Centre, Indian Institute of Science, Bangalore 560012, India
e-mail: hsatreya@sif.iisc.ernet.in

M. Swain
Solid State and Structural Chemistry Unit, Indian Institute of Science, Bangalore 560012, India

(Pardi et al. 1984; Wüthrich 1986). Using these methods, the various secondary structural elements in a given protein is determined *after* the process of sequence specific resonance assignments is completed. That is, the amino acid type corresponding to a given spin-system is known before the knowledge of its secondary structure is obtained.

There is currently no method known to assign directly, based on chemical shifts, the secondary structure type to which a particular spin-system in the protein belongs without the knowledge of the amino acid type and independent of the primary sequence. Such a method would then be independent of the choice of the random-coil shifts and can be used for estimation of the secondary structure content prior to sequence specific resonance assignments. Towards this end, a method was proposed recently wherein the chemical shift of a given type of nucleus ($^1H^N/^1H^\alpha$) averaged over all the residues in the protein was found to correlate with its secondary structure content (Mielke and Krishnan 2005). However, this method does not help in predicting the secondary structure type for any given spin-system in the protein.

We have recently proposed a method utilizing the measurement of ($^3J_{HNH\alpha}$) values for each spin-system for reliably estimating the secondary structure content in proteins (Barnwal et al. 2007). Taking clue from this method, we explored the different possibilities of identifying secondary structure based on chemical shifts by direct inspection of the NMR spectra. A further motivation was to devise NMR experiments wherein residues belonging to different secondary structure types (i.e., α-helix, β-strand) are edited into different regions of the spectrum, thereby simplifying the process of resonance assignments. To this end, it is found that linear combinations of certain backbone chemical shifts facilitate the distinction between the different secondary structure types. The most important nuclei for such combination are found to be $^1H^\alpha$ and $^{13}C'$, the chemical shifts of which are well known to correlate with secondary structure (Wishart and Sykes 1994). As proposed here, a linear combination of these chemical shifts gives an (nearly) amino-acid independent way of estimating secondary structures with more accuracy than those obtained using their individual chemical shift values. This methodology is substantiated with extensive statistical analysis using a database of 237 proteins containing sequence specific ^1H and ^{13}C chemical shift assignments for a total of about 25,000 residues obtained from the BioMagResBank (BMRB) (http://www.bmrb.wisc.edu) and the TALOS database (Cornilescu et al. 1999). The TALOS database contains both chemical shifts and high-resolution crystal structures of non-homologues proteins. This database therefore facilitates the comparison of secondary structures in proteins estimated using linear combination of chemical shifts with those observed in the 3D structures.

In order to detect such a linear combination directly in a NMR spectrum and thereby facilitate secondary structure based editing and estimation, we have devised a new 2D NMR experiment based on the principle of G-matrix Fourier transform (GFT) projection NMR spectroscopy (Kim and Szyperski 2003; Atreya and Szyperski 2005). GFT NMR is based on phase sensitive joint sampling of two or more chemical shifts in a single dimension, thereby providing high dimensional spectral information rapidly with high precision. GFT NMR has found several applications in protein resonance assignment and structure determination (Atreya and Szyperski 2004, 2005; Atreya et al. 2005, 2007; Kim and Szyperski 2003; Eletski et al. 2005; Shen et al. 2005; Szyperski and Atreya 2006). Joint sampling in GFT NMR provides signals with linear combination of chemical shifts, which can be scaled relative to each other (Szyperski and Atreya 2006). This helps in the detection of the desired linear combination of chemical shifts. The experiment specifically devised for the current study is (3,2)D HA(CA)CO(N)H, wherein for the nuclei shown underlined, the chemical shift evolution periods (of $^1H^\alpha$ and $^{13}C'$) are jointly co-incremented. This results, after G-matrix transformation, in two sub-spectra each comprising of peaks at a given linear combination of chemical shifts along the indirect dimension: ω_1:$\Omega(^{13}C') \pm \kappa * \Omega(^1H^\alpha)$ where 'κ' is scaling factor which is decided based on the desired choice of linear combination needed. Further, out of the two linear combinations, the one which facilitates editing of residues into different secondary structural type can be selectively detected using the method of combination shift selective GFT NMR (Swain and Atreya 2008). We have named this approach: CSSI-PRO (Combination of Shifts for Secondary Structure Identification in Proteins). This methodology is demonstrated using experimental data acquired for Ubiquitin and using simulated data for three different proteins. The potential different applications of this methodology are discussed.

Materials and methods

Chemical shift statistics for α-helix, β-strand and random coil

Using method of chemical shift index

The methodology described above for estimating secondary structure in proteins was evaluated using a collection of chemical shifts from 237 proteins ranging in 50–370 amino acids ($\sim$25,000 amino acids) downloaded from the BMRB (130 proteins) and TALOS (107 proteins) databases. The PDB codes and the BMRB deposition numbers of all the proteins used in the analysis are given in Table S1 of Supporting Information. To obtain the secondary structure

information for each amino acid residue in a given protein, the method of CSI (Wishart and Sykes 1994) was used. An in-house script was written to calculate the 'secondary shifts' (deviation of observed chemical shifts from the random-coil value) for each residue using the known assignments. The random-coil value tabulated by Wishart and Sykes (1994) was used for calculating these secondary shifts. A consensus approach was taken in which a given residue was assigned one of the three secondary structure types (i.e., α-helix, β-strands or random coil) only if indicated by two or more of $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C'$ secondary shifts. Next, for each residue in a given protein, a linear combination (LC) of the backbone chemical shifts of the following form was calculated:

$$\Omega_{LC} = x * \Omega(^{13}C^{\alpha}) + y * \Omega(^{13}C') - z * \Omega(^{1}H^{\alpha}) * 4.0 \quad (1)$$

where 'Ω' refers to the deviation of the shift from an arbitrarily chosen chemical shift reference for the given nucleus (i.e., $\Omega = \delta_{observed} - \delta_{reference}$). This reference shift is equivalent to the carrier frequency that is generally used for detecting these nuclei in NMR experiments. Thus, 56, 176 and 4.77 ppm were chosen as the reference shift for $^{13}C^{\alpha}$, $^{13}C'$ and $^{1}H^{\alpha}$, respectively. Note that the changing these chemical shift references only shifts Ω_{LC} and does not change the nature of chemical shift distribution, which is discussed below. The multiplication factors x, y and z are adjustable weighting factors ($x, y, z \geq 0$). A weighting factor of zero implies no contribution to Ω_{LC}. The contribution of $^{1}H^{\alpha}$ to the linear combination is opposite in sign as compared to ^{13}C due to the generally observed fact that the secondary shift for all the 20 amino acids in both α-helix and β-strands are opposite in sign (Wishart and Sykes 1994). Further, the contribution from $^{1}H^{\alpha}$ is scaled by 4.0 to take into account the fourfold higher gyromagnetic ratio of $^{1}H^{\alpha}$ relative to ^{13}C. The chemical shifts of $^{13}C^{\beta}$ were not used in the analysis due to their strong dependence on the amino acid type (Atreya et al. 2000) (see Figure S1 of Supporting Information).

For each residue belonging to a given secondary structure type identified based on CSI, its respective Ω_{LC}^{X} value (X = α-helix, β-strand, random coil) was collected. Following this, for each type of secondary structure, a distribution of Ω_{LC}^{X} (i.e., percentage of residues having a given Ω_{LC}) was generated.

Estimation of overall secondary structure content in a given protein

The overall secondary structure content of a given protein was estimated as follows. The Ω_{LC} distribution (Eq. 1) calculated over all the proteins for each of the three secondary structure types was fit to a Gaussian to estimate the mean (μ) and standard deviation (σ):

$$\Omega_{LC}^{X} = A_X * \exp\left(-(\Omega - \mu_X)^2/2\sigma_X^2\right) \quad (X = \alpha\text{-helix, } \beta\text{-strand, random coil}) \quad (2)$$

Next, a Ω_{LC} distribution was then calculated for each protein in the database separately, smoothened using a 3-point or a 5-point moving average method and fit to the sum of three Gaussians given in Eq. 2 as:

$$\Omega_{LC}^{protein} = F_{\alpha-helix} * \exp\left(-(\Omega-\mu_{\alpha-helix})^2/2\sigma_{\alpha-helix}^2\right) + F_{\beta-strand} * \exp\left(-(\Omega-\mu_{\beta-strand})^2/2\sigma_{\beta-strand}^2\right) + F_{randomcoil} * \exp\left(-(\Omega-\mu_{randomcoil})^2/2\sigma_{randomcoil}^2\right) \quad (3)$$

where μ_X and σ_X (X = *α-helix*, *β-strand*, *random coil*) obtained from Eq. 2 were held constant and $F_{\alpha\text{-helix}}$, $F_{\beta\text{-strand}}$ and $F_{\text{random coil}}$ were the adjustable parameters in the fit. The area under the curve for each type of secondary structure calculated using F_X, μ_X and σ_X (X = α-helix, β-strand, random coil) was then used as a measure of the fraction of α-helix, β-strand and random coil content, respectively, in the protein. The fitting was carried out using the commercial software program: SIGMA-PLOT.

Estimation of accuracy

In order to rigorously assess the methodology for estimating secondary structure content in proteins, three sets of proteins were generated (Fig. 1):

I. A set consisting of 107 non-homologous proteins from the TALOS database with high-resolution 3D structures. In this database, secondary structure type for each residue is assigned using the DSSP method (Kabsch and Sander 1983). This set served as a check of how the secondary structure identified in proteins using CSSI-PRO compared

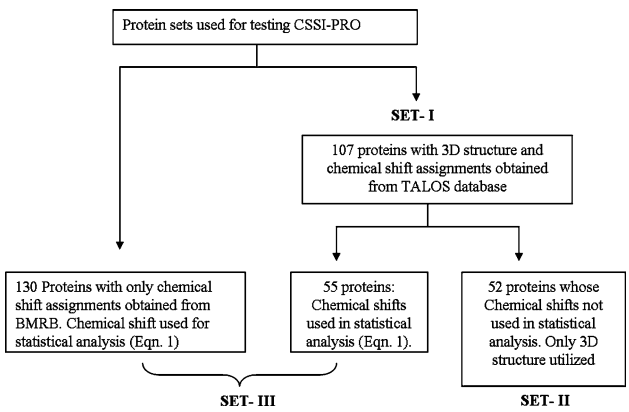

Fig. 1 A schematic diagram depicting the different sets of proteins obtained from the BMRB and TALOS databases that were used to test the CSSI-PRO methodology

with those estimated based on their respective 3D structures.

II. The above set of 107 proteins was divided into two sets: (A) and (B) consisting of 55 and 52 proteins each. One set (A) of 55 proteins was included in calculating the distribution of Ω_{LC} (Eq. 1) described above for residues belonging to a given secondary structure type based on the CSI. The second set (B) of 52 proteins was treated as test cases. That is, the secondary structures content of these proteins were predicted using CSSI-PRO but their chemical shifts were not used in calculating the distribution, Ω_{LC} (Eq. 1). Thus, this set of proteins served as an unbiased independent check of the methodology, since their chemical shift data was not utilized in the statistical analysis.

III. A set consisting of 185 proteins, which included 55 of the TALOS proteins set (A) above and 130 of proteins from the BMRB database containing sequence specific ^{1}H and ^{13}C resonance assignments. The chemical shifts of all proteins in this set were used in calculating the distribution, Ω_{LC} (Eq. 1). This set served as a check of how the secondary structure identified using CSSI-PRO compared with those estimated using CSI. The different sets of proteins selected are schematically depicted in Fig. 1. Using these sets of proteins, two figure of merits were used:

(1) A root mean square deviation (rmsd) for individual secondary structural elements was calculated as follows:

$$\mathrm{RMSD}_X(\zeta) = \mathrm{sqrt}\left[\sum_{i=1,\ldots,N\zeta} |F(\text{CSSI-PRO})^{\zeta}_{X,i} - F^{\zeta}_{X,i}|^2 / N_{\zeta}\right] \quad (4)$$

N_{ζ} is the number of proteins considered for the analysis in the set ζ: Set I, II or III, $X = \alpha$-helix, β-strand or random coil, $F(\text{CSSI - PRO})^{\zeta}_{X,i}$ refers to the percentage (F) of secondary structural element X in the 'ith' protein of the database ζ calculated using CSSI-PRO and $F^{\zeta}_{X,i}$ refers to the percentage (F) of secondary structural element X in the 'ith' protein in the database ζ estimated based on the respective 3D structure (for set I and II) and based on the method of CSI (for set III).

(2) A linear Pearson's correlation coefficient was calculated for the different set of proteins to estimate the extent of correlation between the secondary structure content predicted using CSSI-PRO and those based on the 3D structure and CSI.

NMR spectroscopy

In order to observe the desired linear combination of shifts directly in a NMR spectrum for secondary structure based editing and estimation, we have devised a GFT (3,2)D HA(CA)CO(N)H experiment, where the chemical shifts of $^{1}H^{\alpha}$ and $^{13}C'$ are jointly sampled. This experiment has been derived from (5,2)D HACACONH (Kim and Szyperski 2003). The specific pair of chemical shifts ($^{1}H^{\alpha}$ and $^{13}C'$) was chosen due to the fact that their combination was found to be the most appropriate for editing of residues in α-helix/β-strands independent of the amino acid type (discussed below). Figure 2 shows radio-frequency (RF) pulse scheme for the proposed (3,2)D HA(CA)CO(N)H experiment. For the nuclei shown underlined ($^{1}H^{\alpha}$ and $^{13}C'$), chemical shifts are jointly sampled (Kim and Szyperski 2003; Atreya and Szyperski 2005; Szyperski and Atreya 2006). Joint-sampling of $^{1}H^{\alpha}$ and $^{13}C'$ chemical shifts is achieved by co-incrementing their respective chemical-shift evolution periods with the $^{1}H^{\alpha}$ shifts scaled by a factor 'κ' relative to $^{13}C'$ (Szyperski and Atreya 2006). Further, each of these shifts is phase sensitively sampled. This results, after G-matrix transformation, in two subspectra each comprising of peaks at a given linear combination of chemical shifts along the indirect dimension (t_1) : $\omega_1 : \Omega(^{13}C'_{i-1}) \pm \kappa\Omega(^{1}H^{\alpha}_{i-1})$ with the direct dimension (t_2) encoding the chemical shift, $\omega_2 : \Omega(^{1}H^{N}_{i})$. Here, '$i$' refers to the residue number along the polypeptide chain.

In the present study, since only one of the two linear combinations $\omega_1 : \Omega(^{13}C'_{i-1}) \pm \kappa\Omega(^{1}H^{\alpha}_{i-1})$ helps in secondary structure type editing and estimation (see Results and Discussion below), the method of combination shift selective GFT NMR (Swain and Atreya 2008) was used to selectively detect the desired linear combination: $(^{13}C'_{i-1}) - \kappa * \Omega(^{1}H^{\alpha}_{i-1})$ along the GFT dimension. The method involves selecting the appropriate cosine/sine modulations of chemical shifts and forming the desired linear combination by phase cycling of the radiofrequency pulses and receiver (Swain and Atreya 2008). Thus, the G-matrix transformation is implemented within the pulse sequence. This helps to avoid acquiring the second undesired linear combination (i.e., $(^{13}C'_{i-1}) + \kappa^{*}\Omega(^{1}H^{\alpha}_{i-1})$). However, for secondary structure assignment of spin systems discussed below, this combination can be retained to obtain the central peak information (i.e., $^{13}C'_{i-1}$ chemical shift). The appropriate phase cycling required to achieve this selection is described in the legend of Fig. 2.

NMR data collection

To demonstrate the utility of the proposed method, we recorded the spectra for u-^{13}C/^{15}N-doubly labeled Ubiquitin (1.0 mM). NMR experiments were performed at 25°C on a Bruker 700 MHz spectrometer equipped with a cryogenic probe. The scaling factor, κ, in (3,2)D *HA*(CA)*CO*(N)H was set to 1.0. The total measurement time for the spectra was 20 min. The data was processed with NMRPipe (Delaglio et al. 1995) and analyzed using XEASY (Bartels et al. 1995). Since the G-matrix transformation is applied within the pulse sequence, no

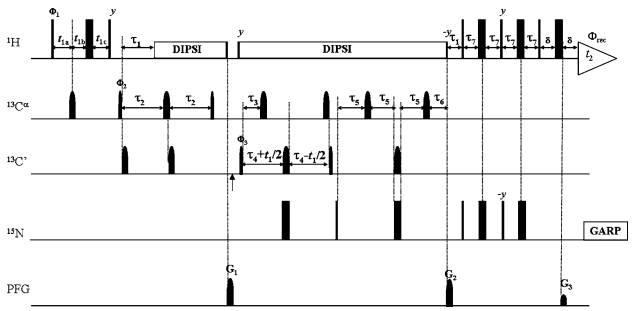

Fig. 2 R.f. pulse scheme of GFT (3,2)D HA(CA)CO(N)H employing the States method for quadrature detection. Rectangular 90° and 180° pulses are indicated by thin and thick vertical bars, respectively, and phases are indicated above the pulses. Where no r.f. phase is marked, the pulse is applied along x. High-power 90° pulse lengths are: 8.0 μs for ^{1}H, 14.5 μs for ^{13}C and 38 μs for ^{15}N. The scaling factor: $\kappa = 1.0$. The ^{1}H r.f. carrier is placed at the position of the solvent line at 4.77 ppm. The ^{15}N carrier position is set to 119.5 ppm. The ^{13}C carrier is initially placed at 54 ppm and switched to 176 ppm before the first 90° r.f. pulse on ^{13}CO (indicated by an arrow). All ^{13}C pulses are of Guassian cascade shape (Cavanagh et al. 2007) with pulsewidth 320 μs for 90° (on- or off-resonance) and 220 μs for 180° (on- or off-resonance). DIPSI (Cavanagh et al. 2007) (r.f. = 3.2 kHz) is used for ^{1}H decoupling during ^{13}C-^{13}C and ^{13}C-^{15}N polarization transfer. GARP (Cavanagh et al. 2007) is employed to decouple ^{15}N (r.f. = 1.50 kHz) during acquisition. The chemical shift evolution of ^{1}H is carried out in a semi-constant time manner (Cavanagh et al. 2007) with the initial delay periods: $t_{1a} = 1.8$ ms, $t_{1b} = 3$ μs, $t_{1c} = 1.8$ ms and incremental delay: $\Delta t_{1a} = 1/2\mathrm{SW_H}$, $\Delta t_{1c} = -(t_{1b}/$(no. of complex points)) and $\Delta t_{1b} = \Delta t_{1a} - \Delta t_{1c}$. All pulsed z-field gradients (PFGs) are sine-bell shaped shaped with gradient recovery delay of 200 μs. The duration and strengths of the PFGs are: G1 (1.0 ms, 15 G/cm); G2 (1.0 ms, 40 G/cm); G3 (1.0 ms, 4 G/cm). The delays are: $\tau_1 = 5.5$ ms, $\tau_2 = 3.6$ ms, $\tau_3 = 4.4$ ms, $\tau_4 = 12.4$ ms, $\tau_5 = 6.2$ ms, $\tau_6 = 0.7$ ms, $\tau_7 = 2.7$ ms and $\delta = 1.2$ ms. Phase cycling: (i) for acquiring selectively the linear combination: $\Omega(^{13}\mathrm{C}^{\alpha}) + \Omega(^{1}\mathrm{H}^{\alpha})$: $\phi_1 = x, y; \phi_2 = x, -x; \phi_3 = x, y;$ ϕ(receiver) $= x, -x$; (ii) for acquiring selectively the linear combination $\Omega(^{13}\mathrm{C}^{\alpha}) - \Omega(^{1}\mathrm{H}^{\alpha})$: $\phi_1 = x, y; \phi_2 = x, -x; \phi_3 = x, -y;$ ϕ(receiver) $= x, -x$

pre-processing of the data is required and the linear combination of chemical shifts can be directly observed in the spectrum.

In addition to acquiring experimental data, GFT spectra were simulated for three proteins of different molecular weights and comprising different secondary structure types: (1) Calmodulin (predominantly α-helical) (Torizawa et al. 2004), (2) Urea-unfolded Ubiquitin (predominantly random coil) (Wolfgang et al. 2001) and (3) M-crystallin (predominantly β-strand) (Barnwal et al. 2006).

Results

Figure 3 shows a plot of chemical shift distribution of $^{1}\mathrm{H}^{\alpha}$ and $^{13}\mathrm{C}'$ for the 20 amino acid residues in α-helix, β-strand and random coil (the chemical shift values have been obtained from Wang and Jardetzky 2002). While some separation of shifts within each type of nuclei in the different secondary structures is noticed, it is evident that the overlap of the distributions (within $^{1}\mathrm{H}^{\alpha}$ and $^{13}\mathrm{C}'$) is too high to determine the secondary structure type directly from the shifts without a-priori knowledge of the amino acid type. In the case of $^{13}\mathrm{C}^{\alpha}$, this is further aggravated by the fact that out of 20 amino acids, Val, Ile, Pro, Ser and Thr have substantially downfield $^{13}\mathrm{C}^{\alpha}$ shifts and Gly has a substantially upfield $^{13}\mathrm{C}^{\alpha}$ chemical shift (Cavanagh et al. 2007) and hence their distribution in the three secondary structure types stand out distinct from those of other residues (shown in Figure S1 of Supporting Information).

The chemical shift distribution of a linear combination of $^{1}\mathrm{H}^{\alpha}$ and $^{13}\mathrm{C}'$ shifts: $\Omega(^{13}\mathrm{C}') - \Omega(^{1}\mathrm{H}^{\alpha}) * 4.0$ for the 20 amino acid residues in α-helix, β-strand and random coil are shown in Fig. 3d. A significant separation in distribution for the three secondary structure types is obtained compared to their individual shift distributions (Fig. 3a, b). The separation results from the fact that the secondary shifts of $^{13}\mathrm{C}'$ and $^{1}\mathrm{H}^{\alpha}$ in α-helix and β-strand are generally observed to be opposite in sign for all the 20 amino acids (Fig. 3a, b). Thus, when the difference of these shifts is taken (after appropriate relative scaling), the correlation between the secondary shifts and the secondary structural elements is enhanced. That is, the linear combination results in distinct values for three different secondary structures nearly independent of the amino acid types. In addition, contribution to $^{13}\mathrm{C}'$ and $^{1}\mathrm{H}^{\alpha}$ shifts from the neighboring amino acid residue (Schwarzinger et al. 2001) are of the same sign and almost cancel out when the scaled difference is taken. On the other hand, the linear combination $\Omega(^{13}\mathrm{C}') + \Omega(^{1}\mathrm{H}^{\alpha}) * 4.0$ (Fig. 3c) does not contain information on the secondary structure because the secondary shifts of $^{13}\mathrm{C}'$ and $^{1}\mathrm{H}^{\alpha}$ cancel out resulting in a value which is close to sum of the individual $^{13}\mathrm{C}'$ and $^{1}\mathrm{H}^{\alpha}$ random coil value for all amino acids. Other similar combinations of $^{13}\mathrm{C}^{\alpha}$, $^{13}\mathrm{C}'$ and $^{1}\mathrm{H}^{\alpha}$ shifts were calculated and are shown in Figure S2 of Supporting Information. It is evident that in any shift combination where $^{13}\mathrm{C}^{\alpha}$ shifts are involved, secondary structure determination requires an a-priori knowledge of the amino acid type. Thus, out of all the combinations, $\Omega(^{13}\mathrm{C}') - \Omega(^{1}\mathrm{H}^{\alpha}) * 4.0$ was found to the most appropriate for editing of spin systems according to secondary structure type and hence chosen for further analysis.

A distribution of $\Omega(^{13}\mathrm{C}') - \Omega(^{1}\mathrm{H}^{\alpha}) * 4.0$ values for residues in α-helix, β-strand and random coil obtained from the database of 185 proteins (set III) is shown in Fig. 4. For this analysis the secondary structure for each residue in a given protein was determined using the CSI approach (Wishart and Sykes 1994). The distribution thus obtained for each of the secondary structure type was fit to a Gaussian to determine the mean and the standard deviation

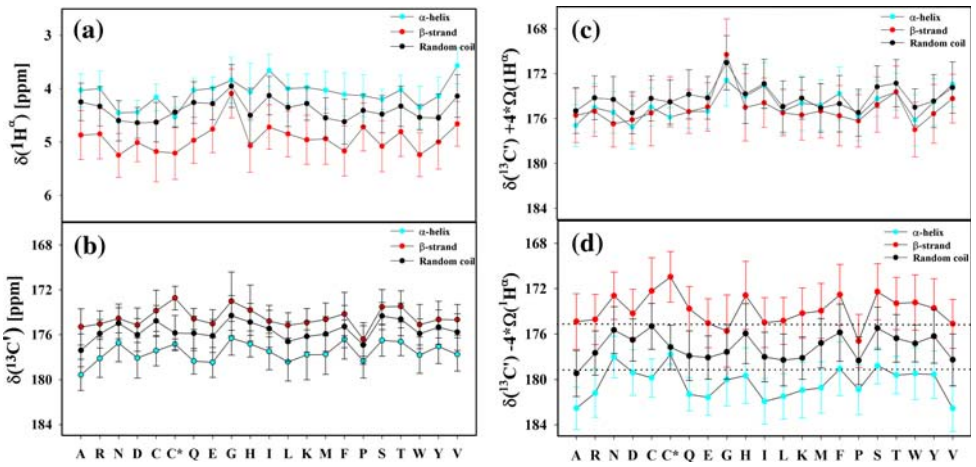

Fig. 3 The chemical shift distribution of $^{1}H^{\alpha}$ and $^{13}C'$chemical shifts and their linear combinations for the 20 amino acids in α-helix (*blue*), *β*-strand (*red*) and random coil (*black*). The average chemical shift value and the standard deviation (indicated as an *vertical error bar*) for each of the amino acids in the three secondary structural elements are shown. The amino acids are indicated by the single letter code (*C** indicates the oxidized state). **a** Shift distribution of $^{1}H^{\alpha}$, **b** shift distribution of $^{13}C'$, **c** shift distribution of linear combination $\delta(^{13}C') + \Omega(^{1}H^{\alpha}) * 4.0$, where '**Ω**' refers to the deviation of the shift from an arbitrarily chosen chemical shift reference for the given nucleus (i.e., $\Omega = \delta_{observed} - \delta_{reference}$). This reference shift is equivalent to the carrier frequency that is generally used for detecting these nuclei in NMR experiments. Thus, 4.77 ppm was chosen as the reference shift for $^{1}H^{\alpha}$, respectively and **d** shift distribution of linear combination $\Omega(^{13}C') - \Omega(^{1}H^{\alpha}) * 4.0$. The *dotted line* along 175 and 179 ppm in (**d**) indicates the chemical shift value, which demarcates the three different structural regions. Chemical shift values used for making the plots have been obtained from Wang and Jardetzky (2002)

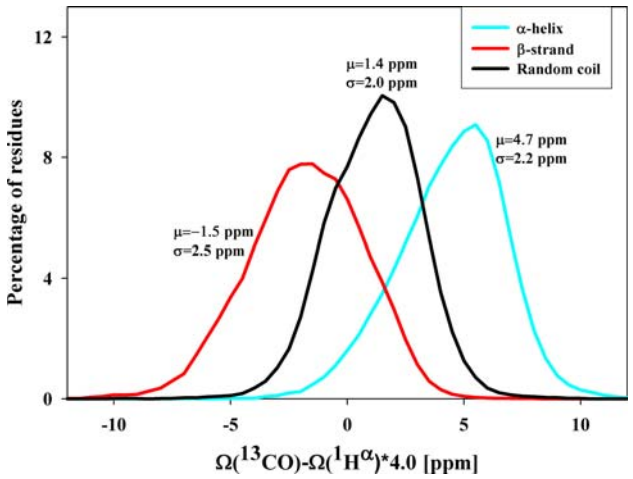

Fig. 4 A distribution of $\Omega(^{13}C')$-$\Omega(^{1}H^{\alpha})$4.0 values for residues in α-helix, *β*-strand and random coil obtained from the database of 190 proteins (set III; Fig. 1). The secondary structure for each residue in a given protein was determined using CSI. The distribution thus obtained for each of the secondary structure type was fit to a Gaussian (Eq. 2) to determine the mean and the standard deviation, which are indicated adjacent to the curves for each type of secondary structure

(Indicated in Fig. 4). Using these values, for a given protein in each of the three sets of proteins (generated as described in Material and Methods), the overall secondary content was estimated using a Gaussian fitting procedure described in Eq. 3. An example is shown in Figure S4 of Supporting Information for the protein Calmodulin. The fraction of α-helix, *β*-strand, random coil thus predicted in each protein was compared with secondary structure estimated for that protein based on its 3D structure (Fig. 5a–f) and the CSI approach (Fig. 5g–i). A linear correlation coefficient of >0.9 is obtained for α-helix and *β*-strands in all cases indicating a very strong correlation between the predicted and observed secondary structure. A lower correlation is seen for residues in the random coil regions due to the fact that linear combination of $^{1}H^{\alpha}$ and $^{13}C'$ for these residues lie in between those of α-helix and *β*-strands (Fig. 4). Hence, some of the residues in the random coil region can get assigned to either α-helix or *β*-strands. Notably, a high correlation obtained for the set of 52 proteins (Set II; Fig. 1) whose chemical shifts were not used in the statistical analysis gives an independent and unbiased test of the methodology. The RMSD between the predicted and observed secondary structure was calculated using Eq. 4 for α-helix, *β*-strands and random coil in each the three sets of proteins and are indicated in their respective plots in Fig. 5. On an average, a RMSD of 6–7% is seen for the different sets of proteins, indicating that the secondary structure content can be estimated with high accuracy using CSSI-PRO. Furthermore, the accuracy of estimation is comparable to those obtained using well-known biophysical techniques such as Circular Dichroism (Carolina and Miguel 2008) or bioinformatics based prediction methods for folded proteins (Mount 2004). A further evaluation of the Gaussian fitting approach was carried out by varying the carrier/offset frequency of $^{1}H^{\alpha}$ and $^{13}C'$. This shifts the distribution of $\Omega(^{13}C') - \Omega(^{1}H^{\alpha}) * 4.0$ in a protein (see Eq. 1) to left or right of the center depending on whether the offset frequency of $^{13}C'(^{1}H^{\alpha})$ is moved

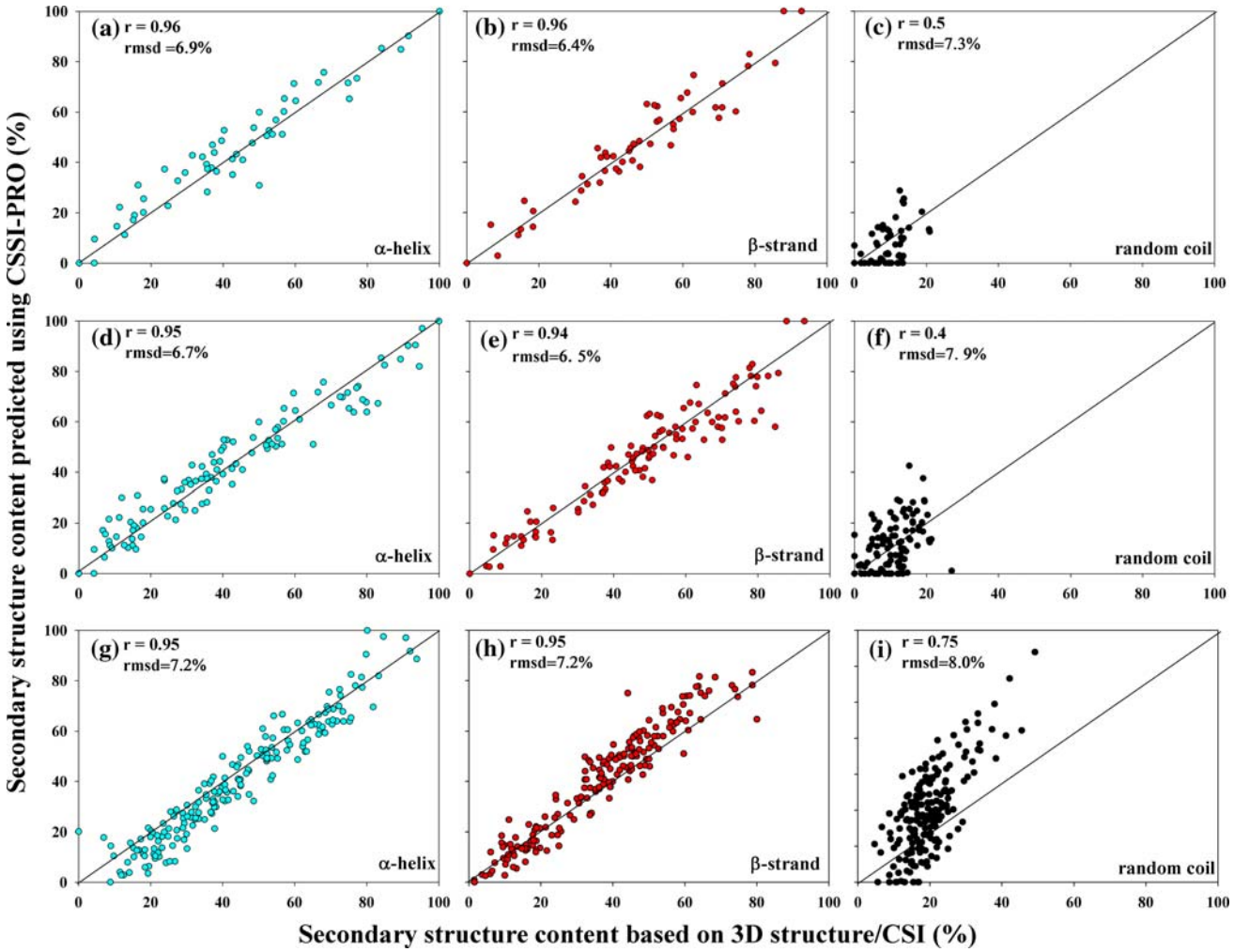

Fig. 5 Comparison of the secondary structure content in proteins estimated using CSSI-PRO (*y*-axis) (as illustrated in Figure S3 of Supporting Information) with those obtained using their respective 3D structure/CSI (*x*-axis). *Comparison plots* are shown for each of the three sets of proteins (Fig. 1) for each of the three secondary structure types (i.e., α-helix, β-strands or random coil): **a–c** SET I, **d–f** SET II and **g–i** SET III. For proteins in SETs I and II [(**a**)–(**f**)], the secondary structure content plotted on the *x*-axis was obtained from their 3D structure using the method of DSSP (Kabsch and Sander 1983). For proteins in SET III [(**g**)–(**i**)], the secondary structure content plotted on the *x*-axis was obtained using the CSI method (Wishart and Sykes 1994). The Pearson's linear correlation coefficient and the rmsd (calculated using Eq. 4) for each of the sets are indicated in the respective plots

downfiled (upfield) or vice versa. Up to a variation in the offset frequency of 0.5 ppm for $^{13}C'$ and 0.1 ppm for $^{1}H^{\alpha}$ no significant variation in the estimation of secondary structure content was noticed, indicating that the Gaussian fitting approach is robust to variations/errors in calibrating the offset frequency correctly.

In order to observe the desired linear combination of shifts discussed above directly in a NMR spectrum, we have devised a GFT (3,2)D HA(CA)CO(N)H experiment (Fig. 2), wherein the chemical shifts of $^{1}H^{\alpha}$ and $^{13}C'$ are jointly sampled with appropriate relative scaling factors. The spectra acquired for Ubiquitin is shown in Fig. 6. A clear separation of peaks belonging to the different secondary structure types is evident. Using the approach described in Eq. 3 and the experimental data (Fig. 6), the secondary structure content in Ubiquitin was estimated to be ~25% α-helix, ~54% β-strand and ~21% random coil which is comparable with ~29% α-helix, ~59% β-strand and ~12% random coil estimated based on its 3D structure.

In order to validate the CSSI-PRO approach, GFT spectra were simulated for proteins of different molecular weights and containing different secondary structure types. Figure 7 shows simulated spectra for three proteins: (1) Calmodulin (predominantly α-helical) (Torizawa et al. 2004), (2) Urea-unfolded Ubiquitin (predominantly random coil) (Wolfgang et al. 2001) and (3) M-crystallin (predominantly β-strand) (Barnwal et al. 2006). A separation of

residues in the different secondary structure similar to one observed for Ubiquitin is obtained for the three proteins, indicating the CSSI-PRO is a robust approach applicable to all types of proteins. Infact it is consistently observed, with very few exceptions, that all peaks downfield of 179 ppm and upfield of 175 ppm belong to residues in α-helical and β-strand structural elements, respectively (shown by dotted line in Fig. 3). This facilitates secondary structure type assignment of spin systems based on direct inspection of NMR spectra.

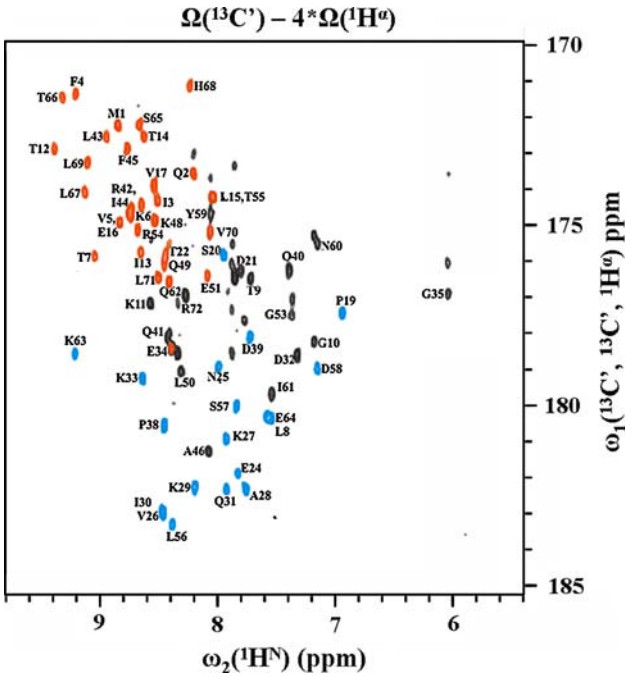

Fig. 6 The GFT (3,2)D *HA*(CA)*CO*(N)HN spectra acquired for Ubiquitin. The residues belonging to the different structure types (as obtained using the CSI method) are indicated: α-helix (*blue*), β-strands (*red*) or random coil (*black*). The assignments are indicated by the single letter code of the amino acid followed by the residue number

Discussion

In recent years, several different approaches have been proposed to delineate secondary structural elements in the protein. Among all methods, CSI is a widely used and well-established method (Wishart and Sykes 1994), although it has certain pre-requisites that need to be taken into account. First, amino acid residue should be sequence specifically assigned in order to use the method. Second, the prediction of secondary structures is strongly influenced by the choice of the reference 'random-coil' values for a given amino acid type. Third, in the case of unfolded proteins, CSI based methods require for all nuclei a correction in the chemical shift based on the neighboring amino acid in the primary sequence before interpreting the secondary structure (Schwarzinger et al. 2001). These constraints have led to several proposals for improving the reliability of the CSI based approach (Wang and Jardetzky 2002). Linear combinations of different chemical shifts have also been proposed to enhance the correlation of secondary shifts with secondary structure (Metzler et al. 1993; Barnwal and Chary 2008). The method of CSSI-PRO described here takes a different approach wherein the secondary structure type and content can be estimated prior to sequence specific resonance assignment without the knowledge of the amino acid type. Such rapid estimation of secondary structure can be combined with other known techniques to estimate secondary structures and thus will benefit NMR protein-folding studies, wherein structural changes in the protein is monitored at each stage of protein denaturation or re-naturation (Dyson and Wright 2004). In addition to protein-folding studies, the methodology has the following potential applications:

Secondary structure assignment of spin-system

Combination of $^{1}H^{\alpha}$ and $^{13}C'$ chemical shifts helps in resolving the spin-systems into different secondary

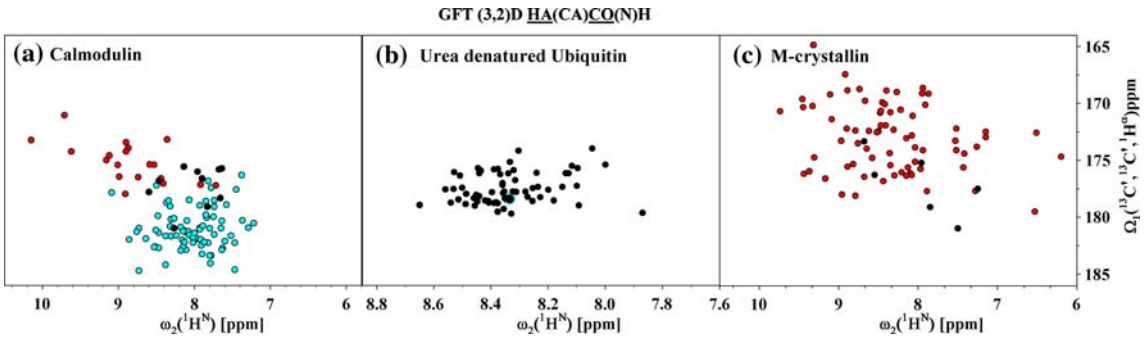

Fig. 7 Simulated GFT (3,2)D HA(CA)CO(N)HN spectra for three proteins **a** Calmodulin (predominantly α-helical), **b** Urea-unfolded Ubiquitin (predominantly random coil) and **c** M-crystallin (predominantly β-strand). The residues belonging to the different structure types (as obtained using the CSI method or/and the information available from the references) are indicated: α-helix (*blue*), β-strands (*red*) or random coil (*black*)

structural types (Figs. 3, 6 and 7). Thus, by direct inspection of the GFT spectra, it is possible to assign a given secondary structure type to a spin system. This can be done by considering spin systems with peaks in the GFT (3,2)D HA(CA)CO(N)H spectra (1) upfield of 175 ppm as belonging to β-strand, (2) downfield of 179 ppm as belong to α-helix and (3) between 175 and 179 ppm as either belonging to random coil or keeping it structurally unassigned. This gives more than 90% accuracy in secondary structure assignment. Such information is useful during automated resonance assignments (Choy et al. 1997; Baran et al. 2004) wherein ambiguity in spin-system matching can be resolved if the secondary structure type is known. In case of proteins with no $^{1}H^{N}$, ^{15}N chemical shift degeneracy such secondary structure type assignment can also be done using $^{1}H^{\alpha}$ and $^{13}C'$ shifts obtained from a conventionally acquired 3D HA(CACO)NH and 3D HNCO spectra, respectively. However, long minimal measurement times of these experiments precludes their use and a rapidly acquired (3,2)D HA(CA)CO(N)H (Fig. 2) becomes preferable. Note that only $^{1}H^{\alpha}$ and $^{13}C'$ are jointly detected in the GFT experiment and hence an additional experiment which has ^{15}N chemical shift information linked to $^{1}H^{\alpha}$ or/and $^{13}C'$ shifts is additionally required. This can be done in one of the two ways: (1) by recording a GFT (5,2)D HACACONH (Kim and Szyperski 2003) or (2) by recording a 3D HNCO, an experiment routinely acquired for resonance assignments of proteins. In both these experiments, the ^{15}N chemical shift information is linked to $^{13}C'_{i-1}$ shifts and the latter is detected in the GFT (3,2)D HA(CA)CO(N)H experiment described above as center peak (i.e., center of $\Omega\left(^{13}C'_{i-1}\right) - \Omega\left(^{1}H^{\alpha}_{i-1}\right)4.0$ and $\Omega\left(^{13}C'_{i-1}\right) + \Omega\left(^{1}H^{\alpha}_{i-1}\right) * 4.0$).

Screening of proteins in structural genomic projects

In structural genomic projects, a large number of proteins are screened at an early stage to estimate the feasibility of carrying out their structural studies by NMR spectroscopy (Montelione et al. 2000; Page et al. 2005; Yee et al. 2002). One of the criteria used for their selection is that they should have a well-defined fold, which in turn implies that they should be comprised of sufficient secondary structural elements such as α-helices and β-strands. CSSI-PRO can be used as one of the tools to determine the overall secondary structure content of the protein.

New NMR experiments based on secondary structure type editing of spin-system

The editing of spin-systems into different secondary structural types (Figs. 6, 7) using linear combination of $^{1}H^{\alpha}$ and $^{13}C'$ shifts will be useful for designing experiments where such a linear combination can incorporated in one of the indirect dimensions that is used for resolving chemical shifts (similar to ^{15}N in triple resonance experiments). An example of such an experiment includes secondary structure edited-[^{1}H, ^{1}H] NOESY. In this experiment, the NOE cross peaks between ^{1}H spins of residues within a given secondary structural type all be clustered in same spectral region. This can be useful in automated NOESY assignment approaches (Güntert 2003) wherein search space for finding cross peaks between two ^{1}H spins is reduced resulting in more accurate assignments. Design of such experiments based on separation of secondary structure types is currently in progress in our laboratory.

Conclusions

We present here a new method, CSSI-PRO, for rapid estimation of secondary structure content in proteins and for assignment of secondary structure type to spin systems by direct inspection of NMR spectrum. The methodology involves detection of specific linear combination of backbone 1Hα and 13C$'$ chemical shifts based on GFT NMR spectroscopy. Such linear combinations of shifts facilitate editing of residues in α-helical/β-strands of the protein into distinct spectral regions independent of the amino acid type. These linear combinations can also be incorporated to design new secondary structure edited NMR experiments. Taken together, this methodology, if used as a stand-alone technique or in combination with other known methods to estimate secondary structure, will have a range of applications in protein structure determination, monitoring conformational changes in protein-folding studies and automated resonance assignment.

Acknowledgments The facilities provided by NMR Research Centre at IISc supported by Department of Science and Technology (DST), India is gratefully acknowledged. HSA acknowledges support from Department of Atomic Energy (DAE) BRNS, DST-SERC and DST-FAST TRACK research awards. MS acknowledges fellowship from Council of Scientific and Industrial Research (CSIR), India. We thank Dr. John Cort, Pacific Northwest National Laboratory, for providing the Ubiquitin plasmid and B. Krishnarjuna, IISc, for preparing the Ubiquitin sample.

References

Atreya HS, Szyperski T (2004) G-matrix Fourier transform NMR spectroscopy for complete protein resonance assignments. Proc Natl Acad Sci USA 101:9642–9647

Atreya HS, Szyperski T (2005) Rapid NMR data collection. Methods Enzymol 394:78–108

Atreya HS, Sahu SC, Chary KVR, Govil G (2000) A tracked approach for automated NMR assignments in proteins (TATAPRO). J Biomol NMR 17:125–136

Atreya HS, Eletsky A, Szyperski T (2005) Resonance assignment of proteins with high shift degeneracy based on 5D spectral information encoded in highly resolved G^2FT NMR experiments. J Am Chem Soc 127:4554–4555

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

Baran MC, Huang YJ, Moseley HNB, Montelione G (2004) Automated analysis of protein NMR assignments and structures. Chem Rev 104:3541–3555

Barnwal RP, Chary KVR (2008) An efficient method for secondary structure determination in polypeptides by NMR. Curr Sci 94:1302–1306

Barnwal RP, Jobby MK, Sharma Y, Chary KVR (2006) NMR assignment of M-crystallin: a novel Ca(+2) binding protein of the $\beta\gamma$-crystallin superfamily from methanosarcina acetivorans. J Biomol NMR 36(Suppl. 5):32–32

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

Bartels C, Xia TH, Billeter M, Guntert P, Wuthrich K (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 6:1–10

Carolina PI, Miguel AAN (2008) K2D2: estimation of protein secondary structure from circular dichroism spectra. BMC Struct Biol 8(25):1–5

Cavanagh C, Fairbrother WJ, Palmer AG, Rance M, Skelton NJ (2007) Protein NMR spectroscopy. Elsevier Academic Press, San Diego

Choy WY, Sanctuary BC, Zhu G (1997) Using neural network predicted secondary structure information in automatic protein NMR assignment. J Chem Inf Comput Sci 37:1086–1094

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13:289–302

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

Dyson HJ, Wright PE (2004) Unfolded proteins and protein folding studied by NMR. Chem Rev 104:3607–3622

Eletski A, Atreya HS, Liu G, Szyperski T (2005) Probing structure and functional dynamics of (large) proteins with aromatic rings: L-GFT-TROSY (4, 3)D *HC*CH NMR spectroscopy. J Am Chem Soc 127:14578–14579

Güntert P (2003) Automated NMR protein structure calculation. Prog NMR Spectrosc 43:105–125

Kabsch W, Sander C (1983) A dictionary of protein secondary structure. Biopolymers 22:2577–2637

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

Metzler WJ et al (1993) Characterization of the three-dimensional solution structure of human profillin: ^1H, ^{13}C, and ^{15}N NMR assignments and global folding pattern. Biochemistry 32:13818–13829

Mielke SP, Krishnan VV (2005) Estimation of protein secondary structure content directly from NMR spectra using an improved empirical correlation with average chemical shift. J Struct Func Genom 6:281–285

Montelione GT, Zheng D, Huang YJ, Gunsalus KC, Szyperski TS (2000) Protein NMR spectroscopy in structural genomics. Nat Struct Mol Biol 7:982–985

Mount DW (2004) Bioinformatics sequence and genome analysis. CSHL Press, USA

Page R, Peti W, Wilson IA, Stevens RC, Wüthrich K (2005) NMR screening and crystal quality of bacterially expressed prokaryotic and eukaryotic proteins in a structural genomics pipeline. Proc Natl Acad Sci USA 102:1901–1905

Pardi A, Billeter M, Wüthrich K (1984) Calibration of the angular dependence of the amide proton-C alpha proton coupling constants, $^3J_{HNH\alpha}$, in a globular protein. Use of $^3J_{HNH\alpha}$ for identification of helical secondary structure. J Mol Biol 180: 741–751

Schwarzinger S, Kroon GJA, Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123:2970–2978

Shen Y, Atreya HS, Liu G, Szyperski T (2005) G-matrix Fourier transform NOESY based protocol for high-quality protein structure determination. J Am Chem Soc 127:9085–9099

Swain M, Atreya HS (2008) A method to selectively observe a desired linear combination of chemical shifts in GFT projection NMR spectroscopy. Open Magn Reson J 1:96–104

Szyperski T, Atreya HS (2006) Principles and applications of GFT projection NMR spectroscopy. Magn Reson Chem 44:S51–S60

Torizawa T, Shimizu M, Taoka M, Miyano H, Kainosho M (2004) Efficient production of isotopically labeled proteins by cell-free synthesis: a practical protocol. J Biomol NMR 30:311–325

Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11:852–861

Wishart DS, Sykes BD (1994) Chemical shifts as a tool for structure determination. Methods Enzymol 239:363–392

Wolfgang P, Loma JS, Christina R, Harald S (2001) Chemical shifts in denatured proteins: resonance assignments for denatured ubiquitin and comparisons with other denatured proteins. J Biomol NMR 19:153–165

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley, New York

Yee A et al (2002) An NMR approach to structural proteomics. Proc Natl Acad Sci USA 99:1825–1830