J. Chem. Sci., Vol. 117, No. 1, January 2005, pp. 3–21.

NMR of unfolded proteins

AMARNATH CHATTERJEE, ASHUTOSH KUMAR, JEETENDER CHUGH, SUDHA SRIVASTAVA, NEEL S BHAVESH and RAMAKRISHNA V HOSUR*
Department of Chemical Sciences, Tata Institute of Fundamental Research Homi Bhabha Road, Mumbai 400 005, India
e-mail: hosur@tifr.res.in

MS received 29 November 2004; revised 3 January 2005

Abstract. In the post-genomic era, as more and more genome sequences are becoming known and hectic efforts are underway to decode the information content in them, it is becoming increasingly evident that flexibility in proteins plays a crucial role in many of the biological functions. Many proteins have intrinsic disorder either wholly or in specific regions. It appears that this disorder may be important for regulatory functions of the proteins, on the one hand, and may help in directing the folding process to reach the compact native state, on the other. Nuclear magnetic resonance (NMR) has over the last two decades emerged as the sole, most powerful technique to help characterize these disordered protein systems. In this review, we first discuss the significance of disorder in proteins and then describe the recent developments in NMR methods for their characterization. A brief description of the results obtained on several disordered proteins is presented at the end.

Keywords. Unfolded proteins; NMR; high throughput procedures.

1. Introduction

It has been the central dogma for long that the function of a protein is related to its significant and unique state, which is a well-defined three-dimensional structure called native structure. Pursuing this goal, three-dimensional structures of several hundreds of proteins have been solved using X-ray crystallography, NMR spectroscopy and more recently electron microscopy. The Protein Data Bank (PDB) is the single depositary for all these data. In February 2003 the number of entries for proteins in PDB was more than 18000.[1]

Though the above paradigm about folded structures and functions remains valid, recent structural and genomic data have clearly shown that not all proteins have unique folded structures under normal physiological conditions. There are about 15,000 proteins in the Swiss Protein Database that have been predicted to contain disordered regions of at least 40 consecutive amino acids[2,3]. PONDR VL-XT, an algorithm in a series of *p*redictors *o*f *n*atural *d*isordered *r*egions (PONDRs) has been applied to about 30 genomes to assess for ≥ 40 consecutive residue disorders in proteins. The results were: bacteria = 6–33%, archea = 9–37%, eukaryotes = 35–51%. With the growth in the information on unstructured proteins, their role in biological specificity, transport, regulation and disease is being realized.[4–10] A more complex signaling and regulation network is perceived to be a possible reason for the high occurrence of disorder in eukaryotes.[11,12] A special term, 'natively unfolded', was coined to describe the properties of *tau* proteins[13] and since then a large number of proteins belonging to this special class have been reported.[9] A detailed study on 91 known 'natively unfolded' proteins has revealed that the majority of natively unstructured proteins have 50 to 100 residues. A typical natively unfolded protein is characterized by: (a) a specific amino acid sequence with low overall hydrophobicity and high net charge; (b) hydrodynamic properties typical of a random coil in poor solvent or pre molten globule (PMG) conformation; (c) low level of ordered structure; (d) the absence of a tightly packed core; (e) high conformational flexibility; (f) ability to adopt relatively rigid conformation in the presence of natural ligand; and (g) a 'turn out' response to environmental changes with structural complexity increase at high temperature or at extreme pH.[10]

*For correspondence

In view of these, Dunker *et al* proposed a protein trinity paradigm,[11] according to which biological function was thought to be an interplay between three thermodynamic states, namely ordered, molten globule and random coil. In this view, not just the ordered state, but any of the three states can be the native state of a protein.[11] Recently a Protein Quartet Model (figure 1) has been proposed for generalization of the structure–function paradigm.[9] According to this, a biological function arises as a result of interplay between four specific conformational forms, namely, ordered forms, molten globules, premolten globules, and random coils. It would not be an exaggeration to assume an ensemble existence of all the four states at any particular time, their relative abundance being governed by basic thermodynamics. Upon ligand binding or some signaling modification, concentration of one state may increase at the expense of the others. This can explain the fast regulatory steps involved in various biological functions.

From the functional point of view, intrinsically disordered proteins have advantages like increased binding specificity at the expense of thermodynamic stability, regulation by proteolysis and the ability to recognize a range of proteins. The large occurrence of unstructured proteins in multicellular metazoans[14] has made researchers envisage that unstructured proteins may be less sensitive to environmental perturbations and, therefore, may impart stability to complex regulatory networks that might otherwise be sensitive to changes in cellular conditions.[7] Two kinds of natively unfolded proteins can be distinguished: first, those in which the unfolded protein performs biological functions in its natively unfolded state[4,6,15,16] and, second, those in which a natively unfolded protein acquires a compact structure upon binding to its ligand or any co-factors (synergistic folding).[17] Excellent reviews have been written on this subject recently, covering all the systems, so far discovered.[5,7,8,12] With the increasing number of intrinsically unfolded proteins and detailed information on their biological function, need for an appropriate database was felt. Thus, a web site http://DisProt.wsu.edu has been created by a group in Washington state university for deposition of data and advanced search on the pattern of PDB.[12]

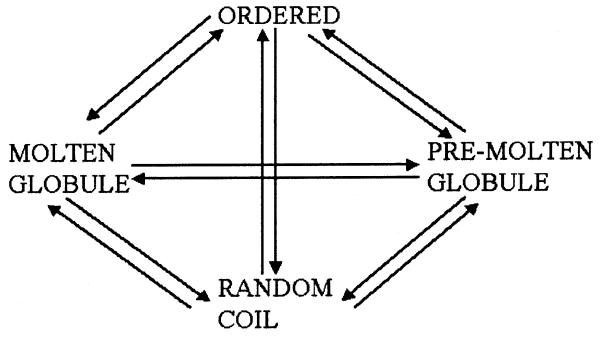

Figure 1. Protein quartet model proposed for generalization of the structure–function paradigm.[9]

Detailed characterization of the unfolded state and consequent identification of the folding initiation sites in a given protein provide valuable insight into its folding mechanism.[18] Well-formed or transient residual structures in the unfolded state are possible candidates for folding initiation sites.[19] Unfolded or partly unfolded states of globular proteins can be created by use of denaturants, which disrupt the noncovalent interactions and propel them to lose their biological activity. This state is referred to as the 'denatured state'. However, often, denaturation is not accompanied by complete unfolding of the protein, and the denatured state is an ensemble of conformations between native and completely unfolded states.[9] In the denatured state floppy molecular chains explore a far smaller number of different shapes than is theoretically available to them, between 100 and 1,000 whereas many millions are possible. They are like dancers executing a series of practiced moves rather than thrashing about at random.

2. NMR methods for investigating unfolded or partly folded proteins

In addition to its normal application for determination of 3D structure at atomic resolution of folded proteins in solution, nuclear magnetic resonance (NMR) spectroscopy is unparalleled in its ability to provide detailed structural and dynamical information on unfolded and flexible proteins.[20] The proton chemical shift dispersion and line width give first-hand information on the state of proteins; poor proton chemical shift dispersion and broad lines are indicative of disorder. Figure 2 shows the amide proton and ^{15}N dispersions in the HSQC spectra for a properly folded Sumo-1 (a) and for the same protein unfolded by 8 M urea (b). Clearly, the amide proton dispersion in the unfolded protein is very narrow compared to that in the folded protein. The ^{15}N chemical shift dispersion is, however, fairly good in both the states.

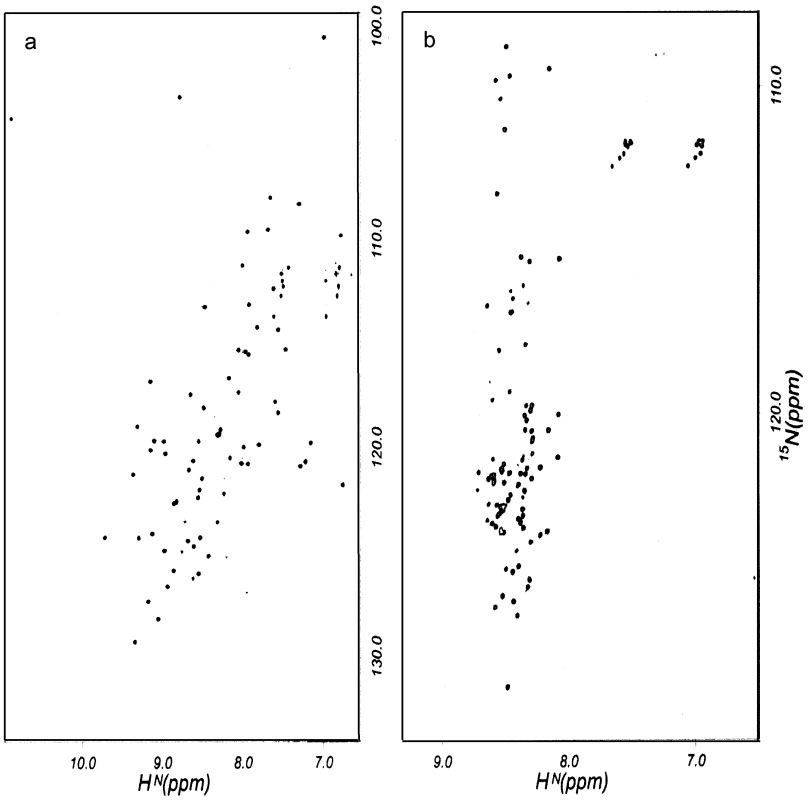

Figure 2. Amide proton and ^{15}N dispersions in the HSQC spectra for a properly folded SUMO (**a**) and for the same protein unfolded by 8 M urea (**b**).

Structural investigations on proteins by NMR are, currently, almost invariably carried out using multidimensional multinuclear experiments on doubly (^{15}N, ^{13}C), and, often, triply (^{15}N, ^{13}C, ^{2}H) labeled protein samples, produced by recombinant means (reviewed in refs [21–23]). Commendable successes have been achieved in the last decade and structures of several hundreds of proteins have been determined to-date to very high resolution. While the detailed description of the procedures is beyond the scope of this article, briefly, the structure determination protocol involves the following general steps: (i) Resonance assignment of the backbone and side chain atoms by separate sets of triple resonance experiments which employ sequential walk through the backbone and along the side chains on the basis of one bond correlations. The three-dimensional spectra are scanned plane by plane and the peaks are identified in a self-consistent manner and this leads to the assignment of peaks to specific residues along the sequence of the polypeptide chain. The major advantage of this procedure is that these correlations are structure-independent, are very sensitive to the large magnitudes of the one-bond couplings, rely on dispersions in ^{13}C and ^{15}N chemical shifts which have generally very wide range, and thus lead to unambiguous assignments in large proteins as well. (ii) Extraction of structural parameters such as coupling constants, inter-proton distances, and orientational parameters of individual N–H or C–H bond vectors in the molecule. (iii) Calculation of the structural ensembles consistent with the NMR structural parameters, by making use of the distance geometry and simulated annealing protocols.

The conventional approaches mentioned above have the following limitations in the context of high speed demand from genomic data analysis and the difficulties encountered with natively unfolded proteins: (a) Repeated scanning through the ^{15}N planes of the 3D spectra to locate peaks at the desired chemical shifts is a very time-consuming process; (b) equivalence of carbon chemical shifts produces ambiguities; (c) residue-type identification along the chain required to remove ambiguities is a slow process and

requires many NMR experiments; (d) the sequential walk stops whenever a proline is encountered and then a new start has to be made from another point along the sequence. However, *a priori*, no known starting points are available in this exercise and they have to be chosen almost randomly; their sequence-specific identification comes at a much later stage. As a consequence, assignments are often made in short stretches and they have to be put together in a self-consistent manner to get complete assignments;[24] (e) finally, in such a procedure, the assignment of a residue, say X, in the stretch PXP, where X is flanked on both sides by prolines becomes impossible. NOE based procedures can be used,[25] but these require full-side chain proton assignments, which is a long drawn process. For this reason, today there are only a handful of detailed investigations on unfolded proteins as compared to several hundreds on folded proteins. Consequently, it is of immense current interest and importance to develop methods which circumvent the above problems and facilitate rapid analysis of unfolded proteins. Pulse sequences have been developed exploiting the better chemical shift dispersions of ^{15}N and ^{13}CO. In this context, the recent new developments made in our laboratory, which include development of pulse sequences and novel strategies based on ^{15}N dispersion for resonance assignments and extraction of structural parameters, have been the most successful. These have opened up new frontiers and possibilities for the investigation of unfolded as well as partially folded proteins in a high throughput fashion. These will be described in some detail for the sake of completeness and ready reference.

2.1 *HNN and HN(C)N pulse sequences*

These are three-dimensional triple resonance experiments, which establish correlations between amide protons and ^{15}N atoms of neighbouring residues along a polypeptide chain. Figure 3 traces the magnetization transfer pathways through the two experimental sequences.

The amide magnetization originating from the *i*th residue is partly transferred in the end to the amides of $i - 1$ and $i + 1$ residues. What remains on i itself results in the diagonal peak and what gets transferred to the two neighbouring residues results in cross peaks in the ^{15}N plane of the *i*th residue. Thus the peaks appear at the following coordinates in the two spectra.

HNN:

$F_1 = N_i$; $(F_3, F_2) = (H_i, N_i), (H_{i-1}, N_{i-1}), (H_{i+1}, N_{i+1})$

$F_2 = N_i$; $(F_3, F_1) = (H_i, N_i), (H_i, N_{i-1}), (H_i, N_{i+1})$

HN(C)N:

$F_1 = N_i$; $(F_3, F_2) = (H_i, N_i), (H_{i-1}, N_{i-1})$

$F_2 = N_i$; $(F_3, F_1) = (H_i, N_i), (H_i, N_{i+1})$.

It is clear that if any of the neighbouring residues is a proline, then that peak will not appear in the spectrum.

2.2 *Peak patterns for proline and glycine neighbours*

We consider here triplets of residues since the HNN and HN(C)N experiments display correlations among three consecutive residues at a time. In the HNN spectrum every F_1–F_3 and F_2–F_3 plane contains the diagonal peak ($F_1 = F_2 = i$) and two sequential peaks to $i - 1$ and $i + 1$ residues. On the other hand, in the HN(C)N spectrum, the F_2–F_3 plane contains the diagonal ($F_1 = F_2 = i$) peak and one sequential peak to $i - 1$ residue, whereas the $F_1 - F_3$ plane contains the diagonal peak and one sequential peak to $i + 1$ residue. Thus, although the HN(C)N sequence generates only i and $i - 1$ correlations, the F_1–F_3 and F_2–F_3 planes taken together help in identifying a triplet of consecutive residues.

Analytical expressions for the intensities and signs of the various diagonal and cross peaks have been

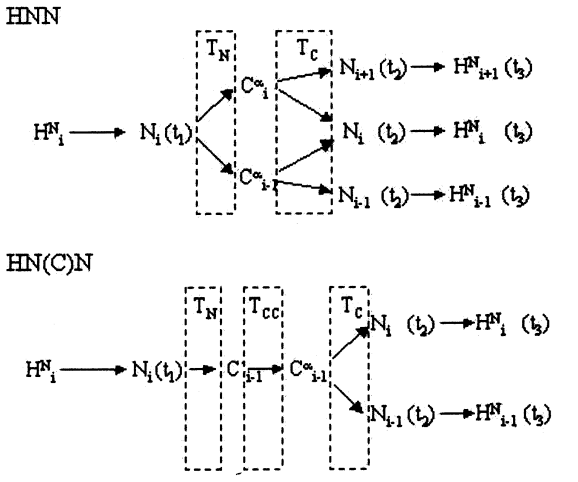

Figure 3. Schematic diagram showing the magnetization transfer pathway in the HNN and HN(C)N experiments. T_N, T_C, T_{CC} are the delays during which the transfers indicated by the arrows take place in the pulse sequences.[26]

derived earlier following product operator methods.[26] It turns out that the evolutions of the magnetization components are slightly different for glycine and non-glycine residues, because of the absence of the C^b carbon in the former. This results in different combinations of positive and negative signs for the various self- and cross-peaks in the different planes of the 3D spectra. Further, the absence of an amide proton for a proline results in the absence of the corresponding peak. In addition, ^{15}N chemical shifts also display certain residue-type dependence.[27] Glycines are distinctly upfield compared to others. Interestingly, the average values of the shifts for the different residue types are similar in both folded and unfolded proteins, though the spreads are more in the folded proteins. Thus there are different patterns of peaks for different triplets of residues containing glycines and prolines; we hasten to add that chemical shift-based distinction is only a guideline.

Four categories of triplets of residues may be distinguished: (I) PXZ, ZXP, PXP, (II) PXG, PGX, PGG, GXP, GGP, XGP, PGP, (III) XGZ, GXZ, ZXG, GGZ, XGG, GXG, GGG and (IV) ZXZ′ where X, Z and Z′ can be any residue other than proline and glycine. Category I has prolines but no glycines, category II has combinations of glycines and prolines, category III has glycines but no prolines, and category IV is a general one, not containing glycines and prolines, and has been included to be able to distinguish the special patterns from the general pattern. The expected peak patterns for each of the above cases in the F_1–F_3 planes of the HNN and HN(C)N spectra, at the F_2 chemical shift of the central residue, under the optimally chosen experimental conditions of magnetization transfer delays (T_N = 28–32 ms, T_C = 28–32 ms, see figure 3) are schematically shown in the figure 4. In each of the planes, the peaks occur aligned at the amide (F_3) chemical shift of the central residue. The choice of the relative ^{15}N chemical shifts of Z, Z′ and X residues is quite arbitrary. In reality, the positions of the positive and negative peaks can get altered as per the relative chemical shifts. The important things to consider are: (i) the sign of the self or the diagonal (F_1 = F_2) peak, and (ii) the signs of the sequential peaks relative to that of the diagonal peak. The patterns in figure 4 can be readily understood from the following salient features of the HNN and HN(C)N spectra.

In the HNN experiment which generates correlations from *i* to both *i* – 1 and *i* + 1 residues, the sign of the self peak for glycine is always opposite to that of any other residue; the actual signs depend upon how the spectra are phased. We have chosen the diagonal of glycine to have a negative sign. In order to emphasize this sign distinction, the diagonal peaks are shown with a different symbol (square) in the figure. This enables unambiguous discrimination between triplets having G and triplets having X as the central residues. For example, the patterns for PGX and PXG in HNN seem similar, but in the former the diagonal peak is negative and the sequential is positive, whereas the reverse is true in the latter case. The sign of the sequential peak at *i* – 1 position will be positive or negative depending upon whether that residue is glycine or otherwise. Similarly, the sign of the sequential peak to *i* + 1 residue will be positive or negative depending on whether the *i*th residue is glycine or otherwise.

In the HN(C)N experiment which generates *i* to *i* – 1 correlations the signs of the self and sequential

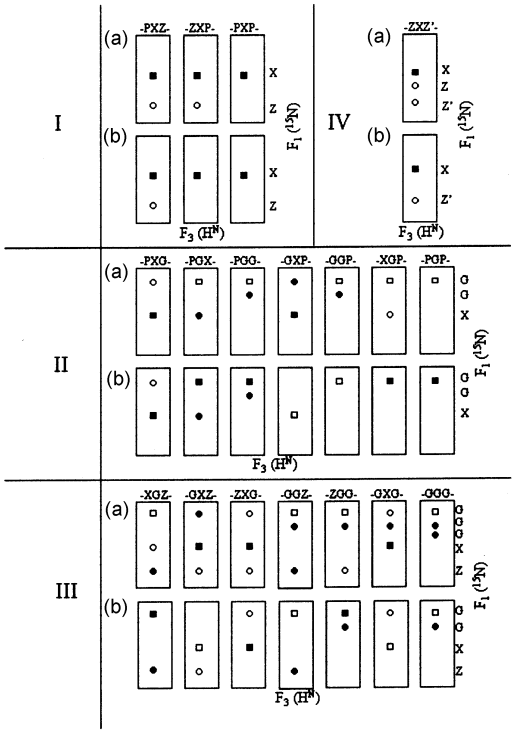

Figure 4. Schematic patterns in the F_1–F_3 planes at the F_2 chemical shift of the central residue in the triplets mentioned on the top of each panel, in the HNN (**a**) and HN(C)N (**b**) spectra for various special triplet sequences of categories I–IV (see text). X, Z, Z′ is any residue other than glycine and proline. Squares are the diagonal peaks and circles are the sequential peaks. Filled and open symbols represent positive and negative signals respectively. In all cases the peaks are aligned at the F_3 (H^N) chemical shift of the central residue.[28]

peaks are always opposite. The actual signs are dictated by whether the $i-1$ residue is a glycine or otherwise, and of course by the phasing of the spectra. Again we have chosen the diagonal to be negative for a glycine at $i-1$ position. Consequently, in the F_1–F_3 plane at F_2 chemical shift of residue i, the signs of the i and $i+1$ peaks are dictated by the nature of the residues at $i-1$ and i positions respectively.

Comparison of the patterns in HNN and HN(C)N spectra enables ready discrimination of the $i-1$ and $i+1$ neighbours and this provides directionality to the assignment. The patterns in HN(C)N resolve some of the possible ambiguities in the HNN patterns and vice-versa. For example, ZXP and PXZ can be readily distinguished from the HN(C)N spectrum, whereas they look similar in the HNN spectrum. Similarly, PGG and GGP patterns are similar in HNN, but are distinctly different in HN(C)N. PGP and XGP would look similar in HN(C)N, but they can be readily distinguished from the HNN spectrum, and so on.

2.3 *Assignment strategies based on HNN and HN(C)N*

The correlations and the peak patterns in HNN and HN(C)N spectra allow development of efficient assignment protocols in labeled proteins. Figure 5 shows schematically the protocols we have derived from these two experiments. While an F_1–F_3 plane at the F_2 chemical shift of residue i in the HNN spectrum displays self and sequential correlations to ^{15}N chemical shifts of $i-1$ and $i+1$ residues at the amide position of i, the F_2–F_3 plane at the F_1 chemical shift of i displays the three correlations at their respective amide positions; note that both F_1 and F_2 dimension have ^{15}N chemical shifts. In any given F_1–F_3 plane distinction between the peaks of $i-1$ and $i+1$ residues can be obtained by comparing with the identical plane in the HN(C)N spectrum. This leads to a new strategy for rapid assignment of the amide and ^{15}N chemical shifts of the individual residues, sequence specifically, as shown in figure 5a.[28] In figure 5b a simplified version is shown where H^N identification from the F_2–F_3 planes has been implicitly assumed. The sequential walk can start from any of the triplet fixed points, identifiable as per figure 4.

Figure 5c shows a similar protocol that is based exclusively on the HN(C)N experiment. The essential ingredients of the sequential walk protocol are indicated as, 'start', 'continue', 'check' and 'break' in the figure. A vertical line makes a connection from the diagonal ($F_1 = F_2$) of the ith residue to the sequential peak identifying the ^{15}N chemical shift of the $i+1$ residue and then the adjoining horizontal line connecting to the diagonal of the $i+1$ residue identifies the amide chemical shift of that residue. The sequential assignment proceeds from the N-terminal toward the C-terminal of the protein.[29]

The HNN and HN(C)N sequences described above have been extremely successful in enabling rapid resonance assignments in many different types of proteins. Table 1 lists the proteins which have been assigned either completely or partially under different experimental conditions in our laboratory. Detailed characterizations of these proteins will be the subject matter of future investigations. It is interesting to see that the methodology has been successful for some folded proteins as well, even though it was originally thought that due to faster transverse relaxation rates in folded proteins, the signal losses during the pulse sequence may be too large. All the above success has been achieved within a period of 2 years, which must be looked at in the background of previous efforts of 15 years which had resulted in the assignment of less than 15 proteins. As illustrations for the experimental HNN and HN(C)N spectra and the assignments, we show the sequential connectivity patterns in a few proteins in figure 6. The fingerprint HSQC spectra of these proteins along with the assignments are shown in figure 7.

2.4 *Extraction of y constraints from HN(C)N spectra*

One- and two-bond heteronuclear couplings, $^1J(N–C^a)$ and $^2J(N–C^a)$ have assumed importance in recent years due to the observation that these show correlations to the y torsion angles along the protein backbone.[30–34] Wirmer and schwalbe[31] described a Karplus-type relation: $^1J(N–C^a) = A\cos^2 y + B\cos y + C$, using data on two proteins, staphylococcal nuclease[30] and ubiquitin[31] and the empirical constants A, B, C have values 9·5098, –0·9799, 1·7040 respectively. The correlation coefficient was 0·85. Wienk *et al*[32] observed similar correlations in three proteins, flavodoxin, xylanase, and DFPase. These suggest that $^1J(N–C^a)$ can be very useful for restraining the y torsion angle in structural calculations. From the extensive data on the different proteins published, it appears that for $^1J(N–C^a)$ greater than 11·0 Hz, y can be restrained to the range (120°, 200°) and for couplings less than 11 Hz the correlation is rather

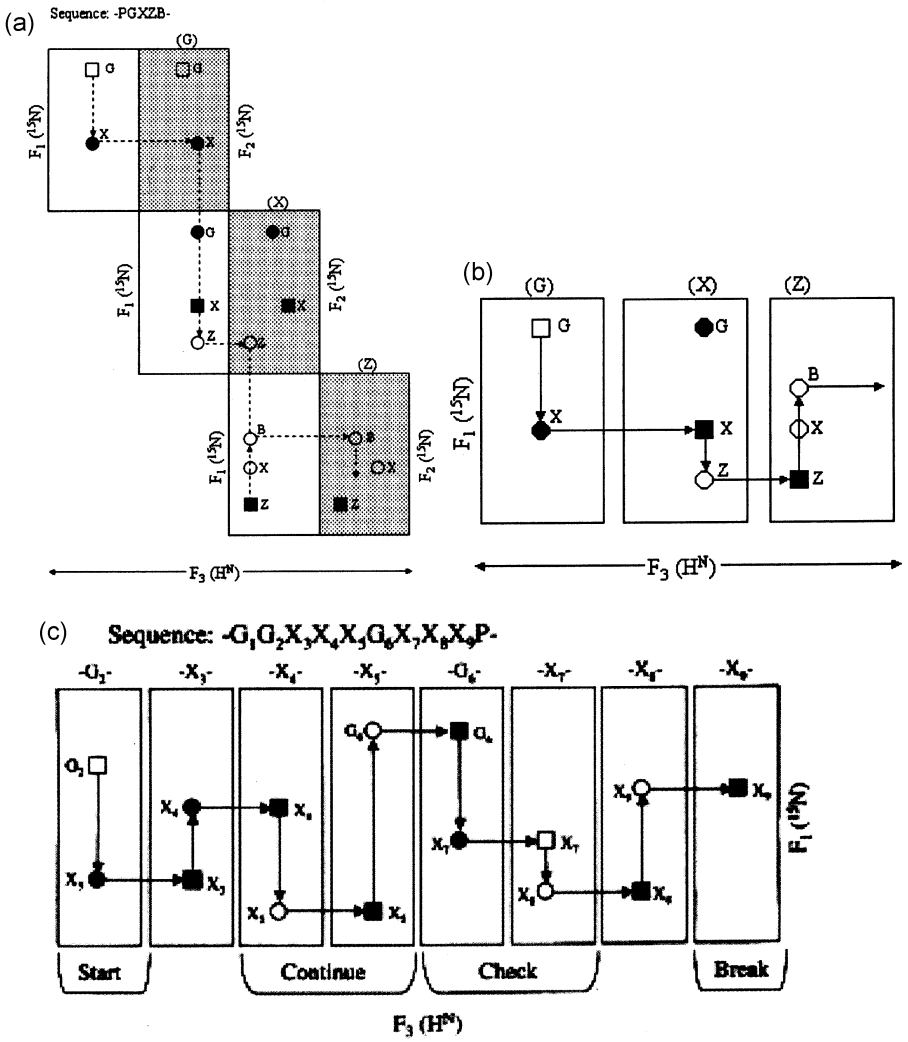

Figure 5. (**a**) The protocol for sequential walk through the HNN spectrum using an illustrative sequence PGXZB where X, Z, B can be any residue other than glycine and proline. Pairs of F_1–F_3 and F_2–F_3 (light shaded) planes belonging to the three residues, G, X and Z are stacked in an appropriate alignment so that the F_1–F_3 plane of one residue stacks over the F_2–F_3 plane of the neighbouring residue. Squares are diagonal peaks and circles are sequential peaks. Filled and open symbols are positive and negative peaks respectively. Note that the sequence chosen includes the PGX and GXZ special triplets and the signs of the peaks have been drawn accordingly. The dashed line indicates the sequential walk. The vertical line at the amide position of a particular residue goes from the diagonal peak to the sequential peak in the F_1–F_3 plane and identifies the ^{15}N chemical shift of the sequentially connected residue (G to X, X to Z and Z to B in the G, X and Z planes respectively), whereas, the horizontal line going from the F_1–F_3 plane to the F_2–F_3 plane of a given residue enables identification of the amide chemical shift of the sequentially connected residue. Note that the diagonal peaks in one plane become sequential peaks in the vertically neighbouring plane. (**b**) A simplified schematic diagram in which the H^N identification from the F_2–F_3 planes is implicitly assumed, and the sequential walk is shown in the F_1–F_3 planes.[28] (**c**) HN(C)N based protocol for sequential walk through the polypeptide chain. An arbitrary amino acid sequence is chosen to illustrate the start, continue, check and break points during the sequential walk. The peak patterns are drawn as per the schematic in figure 3 and the residue identified on the top of each strip identifies the central residue of the triplet.[29]

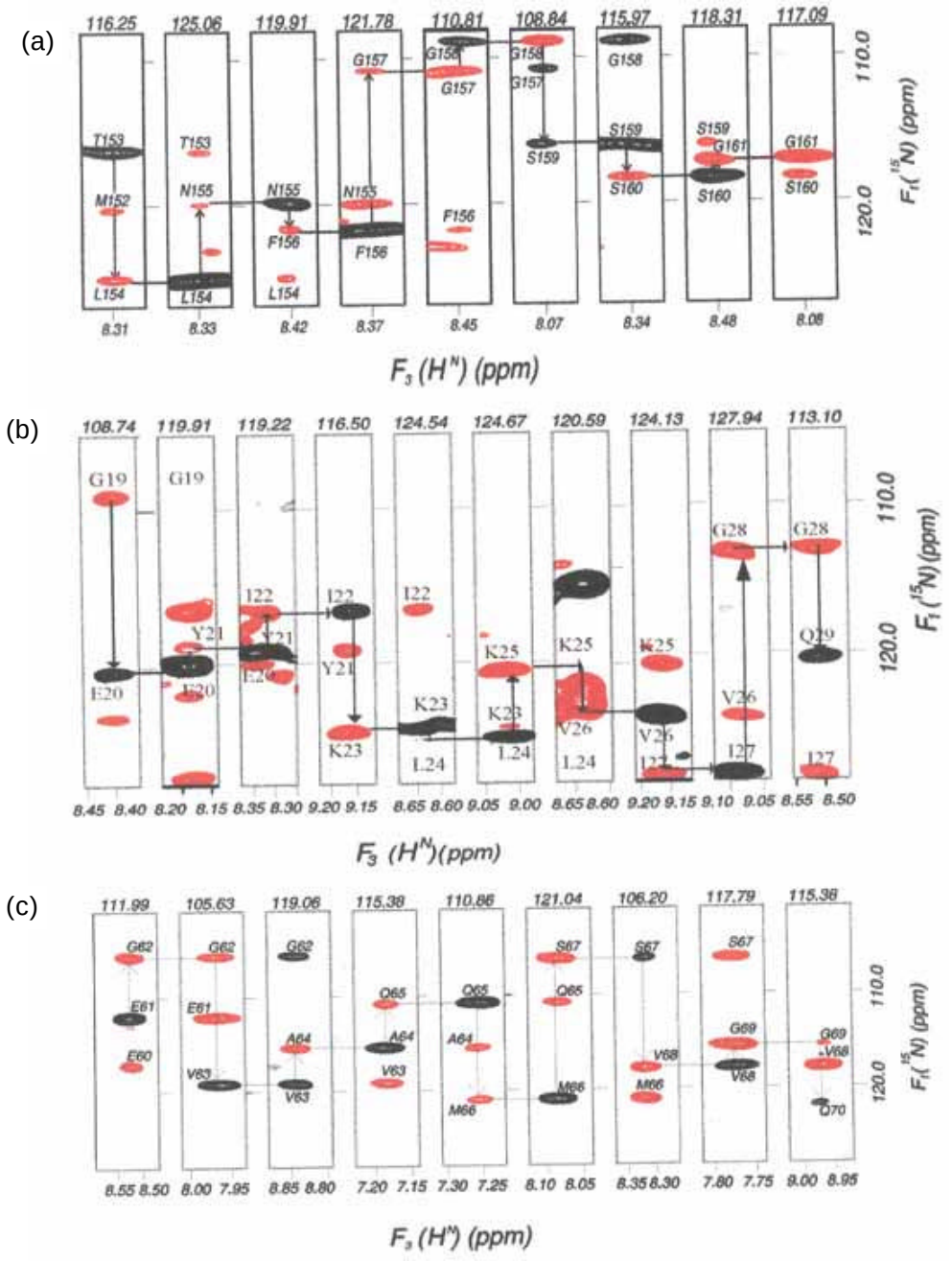

Figure 6. Sequential connectivity patterns for a few proteins: Illustrative sequential walks through the F_1–F_3 planes of HNN spectra are shown. (**a**) Unfolded TFR-PR-Linker precursor protein of HIV-1 Protease,[105] (**b**) folded SUMO[106] and (**c**) folded FKBP. Black and red contours are positive and negative peaks respectively. The numbers at the top of the F_1–F_3 strips indicate the F_2 chemical shifts.

loose in the range (–60°, 120°) (see figure 5 in ref. [32]). This information, in conjunction with the *f* dihedral constraints one would derive from the vicinal H^N–H^a coupling constants and the NOE distance information, augments a great deal the inputs for structural calculations.

We have developed an approach based on the HN(C)N experiment for estimation of one bond N-

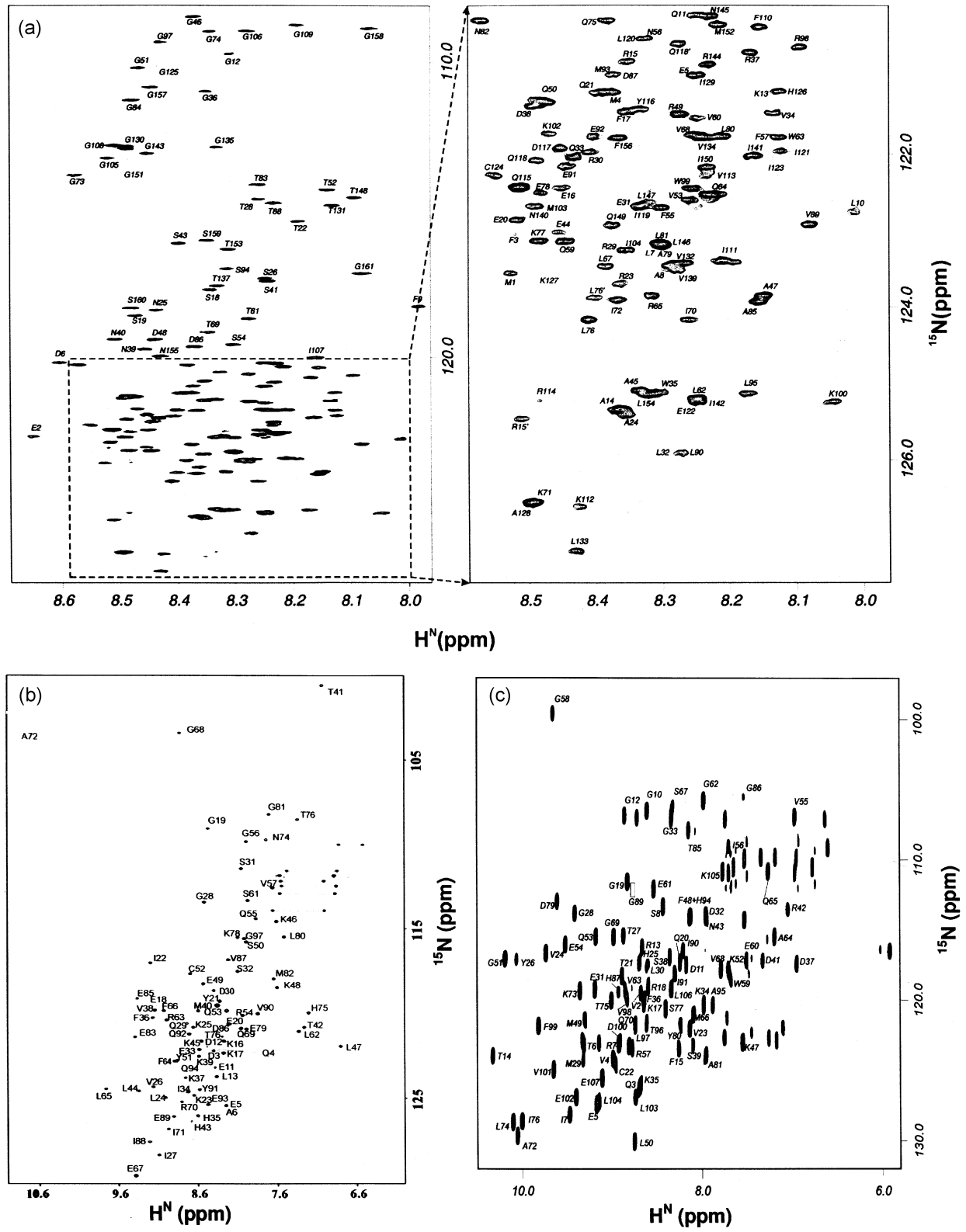

Figure 7. HSQC spectra for the unfolded protein, TFR-PR-Linker precursor protein of HIV-1 protease[105] (**a**), folded proteins, SUMO[106] (**b**) and FKBP[29](**c**).

Table 1. Proteins investigated by HNN and HN(C)N pulse sequences.

No.	Protein
1	HIV-1 protease tethered dimer: 22 kD
2	HIV-1 protease tethered dimer, unfolded: 22 kD
3	HIV-1 PR precursor: 18 kD (intrinsically unfolded)
4	FK 506 binding protein (FKBP): 12 kD
5	Small ubiquitin like modifier (SUMO): 10 kD
6	Small ubiquitin like modifier (SUMO), unfolded: 10 kD
7	Barstar, unfolded: 10 kD
8	Barstar, aggregated (molten globule at low pH): 160 kD
11	GTPase Effector Domain (GED): 15 kD, oligomerises
12	LC8: dyenin light chain, 8 kD folded
13	LC8 unfolded

C^a *J*-couplings in medium size labeled proteins, which seem to show good correlations with y torsion angles along the protein backbone. The approach uses the ratio of the intensities of the sequential and diagonal peaks in the F_2–F_3 planes of the HN(C)N spectrum. While the theoretical details of this analysis have been published elsewhere,[29] suffice it to give the salient features here. Under the generally applicable experimental conditions, the ratio (R) of the intensities of the cross peak to the diagonal peak in the F_2–F_3 plane of the spectrum is given by:

$$R = \tan^2 2p\, p_{i-1} T,$$

where, $p_i = {}^1J(C_i^a\text{–}N_i)$ and the experiment is performed with the settings of delays as $T_N = t_{CN} = T$.

3. Structural propensities and residual structures in denatured proteins

3.1 *Secondary chemical shifts and CSI*

$^1H^a$ and ^{13}C chemical shifts of C′, C^a and C^b are useful indicators of secondary structure in folded proteins[35–38] as they are primarily determined by the backbone f, y dihedral angles.[39] In order to detect the presence of residual structures in unfolded proteins, chemical shift deviations from random coil values for $^{13}C^a$, $^{13}C^b$ and ^{13}C resonances are calculated. A positive deviation of $^{13}C^a$ and $^{13}C'$ chemical shifts from their random coil values is indicative of presence of helical segment while converse is true for $^{13}C^b$ chemical shifts. However, there are more than one set of random coil values published in the literature and these differ because of the experimental conditions used in arriving at those values. Wishart *et al*[40] used 1 M urea, pH 5 and 25°C for their experiments on the peptides chosen, whereas Schwarzinger *et al*[41,42] derived another set of values from peptide spectra recorded in 8M urea, pH 2·3 and 20°C so as to obtain a better data set for their apomyoglobin protein. Schwarzinger *et al*[42] also observed that corrections have to be applied for sequence effects and provided a set of rules for these corrections. $^{13}C^a$ and $^{13}C'$ secondary chemical shifts are better indicators of backbone conformational propensities than $^{13}C^b$ secondary shifts.

3.2 *H^N–H^a coupling constants*

Coupling constants provide valuable secondary structural information in proteins. The $^3J_{HN,Ha}$ coupling constant is sensitive to the dihedral angle f, and thus provides a probe for backbone conformational preferences.[43] b-structures are characterized by large H^N–H^a coupling constant values in the range 8–10 Hz, while a-helical structures are characterized by values in the range 3–5 Hz. In unfolded proteins, however, the heterogeneity and conformational averaging leads to average values of 6–7·5 Hz[44]. Nonetheless, values significantly different from these average random coil values would indicate definite propensities for the structures.

The most common method for measuring the H^N–H^a coupling constants relies on the HNHA experiment[45] or the HNCA-*J* experiment.[46] In the former the coupling constants are derived from the ratios of the diagonal to cross-peak intensities in the different ^{15}N planes of the 3D spectrum, and in the latter they are measured from peak multiplet structures. While these work well for folded proteins with well-resolved peaks, they have serious problems for unfolded proteins where chemical shift dispersion is poor and reliable estimation of the peak intensities or separations is difficult. Relaxation losses also contribute to the uncertainties in intensity measurements in HNHA. Recently these coupling constants have been measured from a high resolution ^1H–^{15}N HSQC spectrum, where this information is contained in the fine structure of the correlation peaks.[47–49] This requires less time for data acquisition and analyses and is well-suited for unfolded proteins where lines are very sharp.

3.3 *Amide proton temperature coefficient*

The temperature dependence of the H^N chemical shift, that is temperature coefficients, provides an es-

timate of the involvement of the amide proton in hydrogen bonding.[50–51] Random coil temperature coefficients determined for residue X in a series of GGXGG unstructured peptide models at pH 5 over the temperature range 278 to 318 K are around –8 ppb/K.[52] Lowered temperature coefficients of the residues showing a deviation of ≥ 1 ppb/K, indicates the involvement of these residues, at least transiently, in hydrogen bonding. The H^N temperature coefficients cannot distinguish between intra- and inter-molecular hydrogen bonds, which can complicate the analyses in aggregated form of unfolded proteins.

3.4 *Hydrogen exchange*

An H → D HX experiment checks if any of the assigned amide protons shows protection from hydrogen exchange. This is an extremely useful parameter to understand the compactness and hydrogen bonding in disordered or partially folded proteins. The exchange reaction is generally described by the following model:

$$(\text{N–H})_{\text{cl}} \underset{k_{\text{cl}}}{\overset{k_{\text{op}}}{\rightleftharpoons}} (\text{N–H})_{\text{op}} \xrightarrow{k_{\text{rc}}} (\text{N–D}).$$

The protein is believed to be in two sets of conformations, closed and open, in equilibrium with each other. However, solvent exchange occurs only from the open state of the protein. The exchange characteristics are a reflection on the solvent accessibilities of the individual amide protons in the folded protein, on the one hand, and on the folding/unfolding equilibria on the other.[53,54] The hydrogen-bonded amide protons exchange much slower compared to the non-hydrogen bonded ones. Thus hydrogen-deuterium (H/D) exchange studies give considerable insight into structure, stability, folding, dynamics and intermolecular interactions in protein systems in solution. Under conditions where $k_{\text{cl}} \gg k_{\text{op}}$, the measured exchange rate k_{ex} is $K_{\text{op}}k_{\text{rc}}$ where K_{op} is the equilibrium constant for structural opening and k_{rc} is the intrinsic exchange rate. Then the free energy change for structural opening is given by

$$\Delta G_{\text{HX}} = -RT \ln (k_{\text{ex}}/k_{\text{rc}}).$$

Several studies suggest that the most stable residues exchange following a global unfolding, while the residues with intermediate stability exchange due to local fluctuations in the native state or following a partial unfolding, and the least stable residues exchange directly with the solvent in the native state. Accordingly, the unfolding free-energy values measured at the various amide sites would actually reflect a series of different unfolding events through the structure of the protein.

3.5 *Nuclear Overhauser effect*

Nuclear Overhauser effect (NOE) provides interatomic distance information in unfolded proteins as it does in folded proteins. It provides information on the secondary structure and other long-range interaction present in proteins. The $d_{\alpha\text{N}}(i, i + 1)$, $d_{\text{NN}}(i, i + 1)$ and $d\alpha_{\text{N}}(i, i + 1)$ NOEs provide the information on f and y dihedral angles preferences while medium range $d_{\alpha\text{N}}(i, i + 2)$, $d_{\alpha\text{N}}(i, i + 3)$ and $d_{b\text{N}}(i, i + 3)$ NOEs indicate the presence of secondary structures. Though an uphill task to assign due to poor resonance dispersion, distributions of structures, anisotropic tumbling and spin diffusion, the long-range NOEs do point to definite presence of structural elements in an ensemble of disordered protein conformations. 3D NOESY-HSQC and 3D HSQC-NOESY-HSQC are routinely used for NOE connectivities, but overlap of proton resonances limits their use for unfolded proteins.[44] For overcoming these problems, NOESY-based triple resonance experiments have been developed, which exploit the dispersion of ^{15}N and ^{13}C resonances to help resolve ambiguities in aliphatic ^{1}H and ^{13}C chemical shifts.[55] However, quantitative interpretation of NOE in unfolded proteins is difficult due to conformational averaging.

3.6 *Residual dipolar couplings*

Residual dipolar couplings (RDC) between nuclei, observable in partially aligned molecules in dilute orienting media, cause significant improvement in the precision of structure determination of biological macromolecules in aqueous solutions. In addition to providing new constraints, which enhance the experimental structural input, the RDCs are unique in defining the orientations of specific dipole–dipole vectors in the molecular frame and thus provide long-range constraints. This becomes especially useful in defining domain orientations in multi-domain proteins. Historical developments, theoretical basis and the applications to different protein and nucleic acid segments have been recently reviewed.[56] Although there are complexities involved in convert-

ing RDCs measured in unfolded proteins into sets of orientational constraints,[57] the structural information given by RDCs is independent of distances. The approach has been successfully used for characterization of topology of urea denatured staphylococcal nuclease. The measurement of RDCs has indicated the presence of long range native-like structures unaffected by local dynamism. Many of the residues were found to maintain a relatively fixed orientation with reference to a single reference axis.

3.7 *Ensembles of structures in the unfolded state*

The various NMR parameters listed above, which provide useful residue-level insight into the secondary structures and structural propensities in unfolded states of proteins, are actually ensemble averages of many 3D structures. Choy and Forman-Kay[58] have developed an algorithm called ENSEMBLE for calculation of ensembles of structures. It optimises the population weights assigned to each structure on the basis of experimental properties derived from NOEs, *J*-couplings, ^{13}C chemical shifts, translational diffusion coefficients and tryptophan environment obtained from NMR and fluorescence spectroscopy. The approach was applied to drkN SH3 domain and five sets of state ensembles were calculated with all having similar average hydrodynamic properties. The results indicated presence of some residual native and non-native structures.

4. Dynamics in unfolded proteins

NMR is unequaled in its ability to provide information on residue specific dynamics in proteins. In unfolded proteins, a restricted motion indicates higher propensity to form a structure or the presence of residual structure. Longitudinal and transverse relaxation rates (R_1 and R_2) of backbone ^{15}N nuclei as well as the ^{15}N–^{1}H steady-state heteronuclear NOE are useful probes of protein backbone dynamics and overall molecular tumbling motions.[59–60] While all the three relaxation parameters are sensitive to motions on a picosecond to nanosecond time scale, the ^{15}N–^{1}H NOE is the most sensitive, to these high frequency motions of the protein backbone. On the other hand, transverse relaxation is very sensitive to slow time scale (micro- to millisecond) motions and conformational exchange, in addition. Negative values of heteronuclear NOEs indicate the occurrence of large-amplitude motions on a sub-nanosecond time scale, which are very frequent in unfolded proteins. Similarly, high R_2 values suggest the presence of significant conformational exchange contributions in proteins. A detailed analysis of relaxation data provides important insight into the possible nucleation sites of protein folding. It can also throw light on residues interacting with their ligands.

Progressive folding of a protein is associated with significant changes in its internal dynamics at all time scales (pico to milliseconds). The fully unfolded state is highly dynamic with motions occurring mostly on pico second time scales. Any restriction in the motions implies transient ordering of the polypeptide chain. As the protein starts to fold, more and more structure forming-breaking events (milli to microsecond time scale) occur, and these lead to increase in the slow motions. Thus a systematic monitoring of these graded changes in the motional characteristics, as also in the residual structures along the polypeptide chain, under different conditions of denaturation, provides very valuable information on the hierarchy of folding propensities in a protein.

Model-free approach is the commonly used approach for analysis of relaxation data of folded proteins.[61,62] This is based on the assumption of separability of internal and globular motions. Thus, dynamics is described in terms of an overall rotational correlation time t_m, an internal correlation time t_c, and an order parameter S^2 describing the amplitude of the internal motions. It provides correlation between dynamics and a set of intuitive physical parameters. However, the model-free approach has limitations with regard to unfolded proteins because of anisotropic tumbling and multitude of uncorrelated motions in these systems. On the other hand, another approach, namely, reduced spectral density analysis, which assumes a distribution of correlation times, is more appropriate for analysis of the ^{15}N relaxation data for unfolded proteins.[44,63] Three spectral density functions $J(0)$, $J(w_N)$ and $J(w_H)$ are calculated as described by Lefèvre *et al*.[64] Of these, $J(w_H)$ is largely determined by heteronuclear NOEs and is most sensitive to higher frequency motions of the protein backbone. $J(w_N)$ is dominated by R_1 and $J(0)$ is dominated by both R_1 and R_2. Thus, $J(0)$ is sensitive to both nanosecond time-scale motions and contributions from slower micro- to millisecond exchange processes. Assuming a linear correlation between $J(0)$ and $J(w_N)$ two empirical parameters a and b are calculated from the slope and y-intercepts. These are then used to calculate the time constants characterizing vari-

ous motions of the protein by solving the following cubic equation in t:[64]

$$2aw_N^2t^3 + 5bw_N^2t^2 + 2(a - 1)t + 5b = 0,$$

where w_N is the larmor frequency of ^{15}N nuclei. A similar analysis can be performed for $J(0)$ and $J(w_H)$ relation.

A general framework for studying dynamics of folded as well as unfolded states of protein has been presented by Prompers *et al*.[65] This does not require the separability assumption. The approach, named isotropic reorientational eigenmode dynamics (*i*RED), depicts correlated dynamics of different polypeptide parts together with mode-specific correlation times.

Cross-correlated dipole-dipole spin relaxation is used to explore molecular dynamics and structural properties.[66–69] It provides information directly on the dynamics of the interacting spins, if their relative orientations are known. More recently $^1H^b$–$^{13}C^b$ dipole–dipole cross-correlated spin relaxation has been used to probe dynamics of unfolded proteins.[70] The dipole-dipole cross-correlated spin relaxation rate for two ^1H–^{13}C dipoles, $\Gamma_{HC1,HC2}$, in the CH_2 group is obtained from the CBCA(CO)NNH spectrum. $\Gamma_{HC1,HC2}$ can be written in terms of spectral densities, which are evaluated in the context of a particular motional model.[71] $\Gamma_{HC1,HC2}$ is given by the relation

$$\Gamma_{HC1,HC2} = -(0{\cdot}25/T)\ln\{4 \times I_U(T) \times I_D(T)/I_C(T)^2\};$$

where I_U, I_D and I_C are the intensities of the upfield, downfield and central components of the $^{13}C^b$ triplet respectively, and T is the duration of the ^{13}C constant time evolution during which cross-correlated spin relaxation occurs.[70] A large $|\Gamma_{HC1,HC2}|$ is an indication of slow motion. Buried residues also exhibit large values.

5. Survey of the unfolded protein systems studied

The number of proteins which have been investigated in the unfolded state till date is rather small compared to that for folded proteins, because of the technical difficulties in the analysis of their spectra and complications in the interpretation of the NMR data. In the early days the same techniques as were used for the folded proteins were used for unfolded proteins as well and this had limited success for reasons already elaborated in earlier sections. An excellent review has appeared recently covering the state of affairs with regard to the NMR applications to study unfolded proteins.[72] It is envisaged that the newer methods, coupled with technological advances would facilitate larger number of investigations in the years to come. In the following, we briefly discuss the results obtained in the few proteins studied so far.

5.1 *The phage 434 repressor protein*

This was the first protein for which sequence specific backbone assignment was obtained in the unfolded state.[73,74] Comparing the assignment at 0 M urea and 7 M urea, coexistence of both the native and unfolded forms was observed at 4·2 M urea having exchange life times of ~ 1 s. Deviation from random roil chemical shifts indicated the presence of some residual structure. Using the high density of NOE constraints for segment 53 to 60 in the protein, structure calculation was done for the segment. The structure showed similar features as in folded protein and the hydrophobic cluster was preserved. The authors speculated that the segment could be a possible nucleation site for folding of this protein.[75]

5.2 *Fibronectin binding protein*

A 130-residue fragment (D1–D4) taken from a fibronectin-binding protein of *Staphylococcus aureus*, which contains four fibronectin-binding repeats and is unfolded but biologically active at neutral pH, has been studied extensively by NMR spectroscopy. The secondary chemical shifts, $^3J_{HN-Ha}$ couplings and absence of long range NOEs were indicative of lack of persistent secondary and tertiary structures. ^{15}N relaxation analysis was done to understand the differences in dynamical properties of residues in different domains. It was finally concluded that the sequence involved in binding has a high propensity for populating the extended conformation. This is likely to allow a number of both charged and hydrophobic groups to be presented to fibronectin for highly specific binding.[76]

5.3 *Lysozyme*

Structural and dynamical properties of oxidized and reduced forms of hen lysozyme in 8 M urea were

studied.[77] Resonances for the residues near the disulfide bridges were absent due to line broadening in oxidized form indicating slow motions. Deviation from random coil behaviour was observed for residues W62, W63, W108, W11 and W123 on the basis of amide and H^a chemical shifts and high R_{1r} relaxation rate, indicating the presence of hydrophobic clusters. A number of medium range NOEs were observed in the region corresponding to A, B, D, and C-terminal 3_{10} helices in the native protein. This suggested that the oxidized form could resemble a molten globule.[77] Hennig *et al* measured $^3J(C', C^g)$, $^3J(N, C^g)$, $^3J(C', C')$ and $^3J(C', C^b)$ coupling constants in denatured hen lysozyme in 8 M urea pH 2·0.[78] It provided insight into the side chain c_1 torsion angle population in the denatured state. The fractional populations of the –60°, 60° and 180° c_1 rotamers were derived from the measured coupling constants. This provides information on the side-chain conformational preferences, which depends on the variations in the electrostatic and steric properties of the side chains. H^N and H^a chemical shifts and R_2 rates indicated the presence of extensive hydrophobic clusters in the denatured state of lysozyme.[79]

5.4 *FK 506 binding protein*

Detailed structural characterization has been done on the FK 506 binding protein (FKBP) unfolded in urea and guanidine hydrochloride.[80] Measured C^a and ^1H chemical shifts, H^N–H^a coupling constants, chemical exchange and ^{15}N relaxation rates indicated extensive conformational averaging for FKBP in 6·3 M urea. The presence of medium range NOEs indicated presence of helical conformation for some residues. This disagreement between NOE and other parameters was explained considering the r^6 dependence of NOE; short ^1H–^1H distances corresponding to a particular secondary structure give strong NOEs even when present in a relatively small population of molecules while other parameters are averaged values. Medium-range NOEs patterns were different in FKBP in 2 M guanidine hydrochloride, possibly due to dependence of helix formation on the salt concentration.

5.5 *Apomyoglobin*

Structural and dynamic properties of sperm whale apomyoglobin have been extensively studied in different denaturing conditions.[81–83] Sequence specific resonance assignments for apomyoglobin denatured under low salt condition at pH 2·3 was obtained using (HCA)–CO(CA)NH spectrum which uses the superior $^{13}C'$ chemical shift dispersion. Presence of small populations of native-like helix in two out of three helical regions stabilized in the molten globule state was concluded on the basis of downfield-shifted sequence-corrected $^{13}C^a$ and $^{13}C'$ secondary shifts, motional restriction from ^{15}N relaxation data and slow amide proton exchange rate.[83]

Detailed studies have also been done on the pH 4 molten globule state of apomyoglobin[81] and also on its helix-destabilizing mutants.[83] Secondary chemical shifts, $H_N^i - H_{i+1}^N$ NOEs, temperature coefficients revealed the presence of large native like helical regions and this was supported by the dynamics data, especially R_2 and corresponding $J(0)$ values which are sensitive to slow motions.[83] Study of molten globule state at pH 4·1 of helix destabilizing mutant of apomyoglobin revealed significant differences in the distribution of helical segments in the protein. Though the overall helical content in other regions was the same, loss of helical structure was observed in regions where destabilizing mutation was made. The acid denatured state of apomyoglobin has also been studied by paramagnetic spin labeling. Nitroxides coupled to mutant cysteine residues were used as probes of chain compaction and long-range tertiary contacts. Even in the highly denatured form, the protein had transient compact states in which there were native-like contacts between the N- and C-terminal regions, though the central region had random coil conformation.[84]

In the urea (8 M) and low pH denatured state of apomyoglobin, clusters of small amino acids such as glycine and alanine led to increased backbone mobility, suggesting their roles as molecular hinges. Also local hydrophobic interactions persisted which caused some restriction of backbone motion on picosecond to nanosecond time scale.[85]

5.6 *Barnase*

Barnase, a small extracellular ribonuclease from *Bacilllus amyloliquefaciens* has been characterized in its pH, urea and temperature-denatured forms.[47,48] Transient structure formation in a small region was concluded in the pH denatured state by analysis of chemical shifts, NOEs and exchange rate.[47] Comparison of the different denatured states indicated

that pH/temperature and urea denatured barnase are more "unfolded" than the pH denatured protein. The residual native and non-native structures in helical and *b*-sheet regions were speculated as initiation points in the folding of barnase.

5.7 *Barstar*

Barstar, an intracellular inhibitor of barnase, gets reversibly denatured in presence of 3 M urea at 278 K.[86] The presence of some residual structure was inferred on the basis of chemical shifts, NOEs and coupling constants. The first and second helices were thought to be the potential initiation sites for the folding of barstar.

Structural characterization by NMR of barstar in 8 M urea has been reported at pH 6·5 and 25°C.[87] Complete backbone resonance assignments of the urea-unfolded protein were obtained using the recently developed three-dimensional NMR techniques of HNN and HN(C)N. The conformational propensities of the polypeptide backbone in the presence of 8 M urea, have been estimated by examining deviations of secondary chemical shifts from random coil values. Segments that are helical in native barstar, were found not to populate the helical conformation in the unfolded state. Similarly residues belonging to *b*-strands 1 and 2 of native barstar, do not appear to show any conformational preferences in the unfolded state. On the other hand, residues belonging to the *b*-strand 3 segment, show weak non-native helical conformational preferences in the unfolded state, indicating that this segment may possess a weak preference for populating a helical conformation in the unfolded state.

Barstar is also known to form a molten globule-like A form below pH 4. This form exists as a soluble 160 kDa aggregate of sixteen monomeric subunits, and appears to remain homogenous in solution. Its flexible region has been recently characterized.[88] New assignment methodology based on HNN and HN(C)N was used to obtain sequence specific assignment for 20 residues in its N terminal flexible region. Chemical shifts, temperature coefficients, exchange rate and dynamics data suggest that the A form of barstar is an aggregate with a rigid core, and the N-terminal 20 residues of each of the monomeric subunits in a highly dynamic random coil conformation which shows transient local ordering of structure. The N-terminal segment anchored to the aggregated core exhibits a free-flight motion.

5.8 *Annexin*

Detailed characterization of partially folded D2 domain of annexin I was done using ^{15}N relaxation data obtained at three magnetic fields, 500, 600 and 800 MHz.[89] Dynamical behaviour of different types of residual structures was explained on the basis of Lorentzian distribution of correlation times, which was used for representation of relaxation data. This approach yields a clearer picture of unfolded state dynamics than the model-free approach, which uses discrete correlation times. High values of the width of the distribution highlight the heterogeneous dynamical behaviour of the interconverting structures.

5.9 *Chymotrypsin inhibitor 2*

^{15}N relaxation and molecular dynamics (MD) have been used to get an insight into the folding pathway of chymotrypsin inhibitor 2 (CI2) at atomic resolution.[90] The unfolded state of CI2 at 6·4 M guanidium hydrochloride contains some residual native helical structure along with hydrophobic clustering in the centre of the chain. The lack of persistent non-native structure in the denatured state reduces barriers that must be overcome, leading to fast folding through a nucleation-condensation mechanism.

5.10 *Fibronectin domain*

Meekhof *et al* have attempted to map the energy landscape for the third fibronectin type III domain from human tenascin (TNfn3) through measurement of ^{15}N backbone dynamics and other structural parameters in 5 M urea.[91] Secondary chemical shifts indicate local preferences for the *a*-regions. Notable clusters of protected amides were observed for acidic residues indicating intramolecular hydrogen bonding. Few nascent turn like structures were also observed. On the basis of the dynamics data it was concluded that deviation from random coil behaviour does not imply the formation of stable structural elements, but indicates conformational propensities only.

5.11 *Staphylococcal nuclease*

Extensive work has been done on the unfolded state of staphylococcal nuclease.[57,92–94] Measurement of residual dipolar couplings (D_{HN}) in urea denatured uniformly ^{2}H and ^{15}N labeled protein oriented in

strained polyacrylamide gels presents a high correlation among the dipolar couplings for individual residues indicating a native like spatial positioning and orientation of chain segments. A systematic decrease in correlation coefficient was observed with increase in urea concentration in the plots of D_{HN} in urea against D_{HN} in water. Some long range interaction similar to the native, were also found to be present in the 8 M urea denatured form. Steric repulsion between the residues or weakened long-range interactions are possible reasons for this partially fixed orientation. Earlier a *de novo* structure determination of the same protein using paramagnetic relaxation from 14 extrinsic spin labels revealed that many features of the folded arrangement of segment positions and orientations persist in this denatured state.[93]

5.12 *Plastocyanin*

Potential folding initiation sites were speculated in apo-plastocyanin on the basis of detailed structural and dynamic characterization of its unfolded state under non-denaturing conditions.[95] Though there was comparative uniformity in backbone motions, the nanosecond timescale motional restriction for certain segments in the protein was associated with transient formation of non-native helical structures. Certain bias towards β- and α-regions of conformational space was also observed which is sufficient to reduce the search time for folding to a level reasonably consistent with those observed in nature.

5.13 *N-terminal SH3 domain of drk*

The isolated N-terminal Src homology 3 (SH3) domain of Drosophila drk (drkN SH3 domain) exists in equilibrium between folded (F_{exch}) and unfolded states (U_{exch}) in aqueous buffer near neutral pH. This remarkable feature has been used extensively to study this protein by NMR.[96–102] The NMR spectra recorded on both states simultaneously, exhibited an approximate 1 : 1 ratio of protein conformations.[96] Interestingly, a stable turn at Leu-28 was observed in the unfolded state but not in the folded state. Comparison of this unfolded form with a denatured state in 2 M guanidine hydrochloride showed that, while both are highly disordered, these states are not identical and more residual structure is present under non-denaturing conditions.[98] The spectral density function evaluated at zero-frequency for the unfolded state of the domain indicated that residues in the middle of the protein sequence are considerably less mobile than those at the termini.[99] This suggested that the molecule does not behave as an extended polymer and that concerted motions of the central portions of the molecule occur, consistent with a reasonably compact conformation in this region. NMR studies for non-native structure in this region of the unfolded state of the drkN SH3 domain suggested that the free energy of the unfolded state plays a crucial role in stability, highlighting the importance of the unfolded states for understanding protein stability.[100] Denaturation as well as hydrogen-exchange experiments demonstrated a non-native burial of the Trp ring within the core of the unfolded state.[101] These findings supported the presence of non-native hydrophobic clusters, organised by Trp rings, within disordered states. Recently it has been confirmed, by NOESY experiments, that hydrophobic clustering does occur in the unfolded state (U_{exch}) of drkN SH3 domain.[102]

5.14 *HIV-1 protease*

NMR identification of local structural preferences in HIV-1 protease in the 'unfolded state' at 6 M guanidine hydrochloride has been reported.[49] Analyses of the chemical shifts revealed presence of local structural preferences many of which were native like, and there were also some non-native structural elements. Three bond H^N–H^α coupling constants that could be measured for some of the N-terminal and C-terminal residues were consistent with the native-like β structural propensities. Unusually shifted ^{15}N and amide proton chemical shifts of residues adjacent to some prolines and tryptophans also indicated presence of some structural elements. These conclusions were supported by amide proton temperature coefficients and NOE data. The preferred structural propensity was in residues at the dimer interface and this includes the cavity at the active site. A large number of these residues are hydrophobic in nature suggesting that hydrophobic clustering may be a strong driving force for the initial folding events of the protein.

It is also observed that glycines and alanines cause enhanced conformational flexibilities and thus may act as molecular hinges in the folding as suggested earlier.[85] As the denaturant concentration is progressively reduced, partially folded species start appearing and the protein begins to acquire autolytic activity. The investigations at 6 M and 5 M guanidine denaturing conditions, show that the dimerization domain, the flaps, and the hinges or the elbow

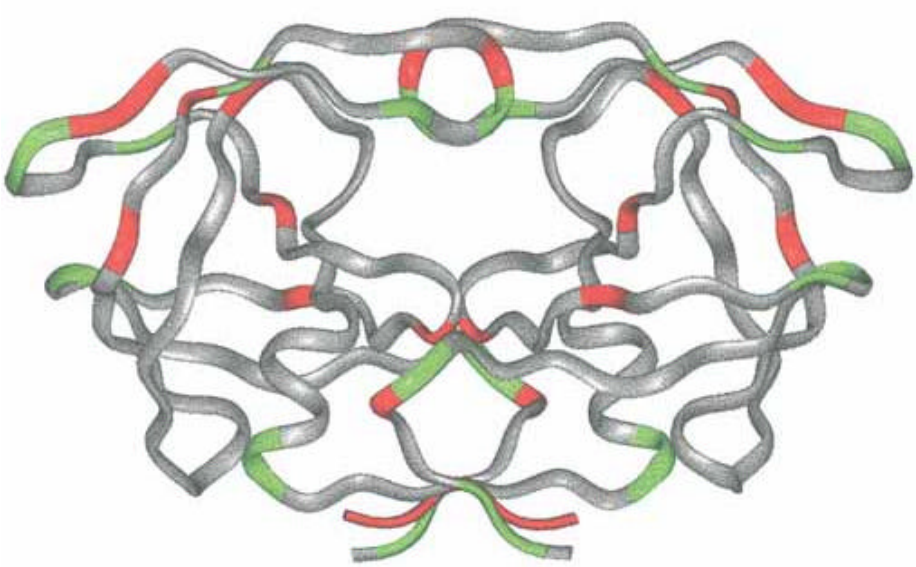

Figure 8. Propensities of conformational transitions or folding events identified by different colours on the basis of the magnitudes of the R_2 changes, displayed on the structure of the native protein. Green residues have higher folding propensities than the red ones.

of the protein have the highest propensity to conformational transitions.[103] The residue-wise graded propensities shown in colour-coded manner on the native structure of the protein in figure 8 also suggest a certain degree of local co-operativity in these conformational transitions.[103]

6. Future directions

In this review we have tried to describe the literature on structural studies on unfolded proteins using NMR as a tool. In the background of evolving concepts regarding the structure–function paradigm, which envisage involvement of various disordered states of proteins, elucidation of the folding pathways of proteins has become essential. In this context, total structural description of the species present at the two ends of the folding funnel, so also of the intermediates needs to be provided.

The NMR methodologies, currently available, have exploited mostly the nitrogen chemical shift dispersion for the study of unfolded proteins. However, it is known that the carbonyl chemical shift dispersion is also very good in both folded and unfolded proteins. Considering this, and also that the carbonyls have an advantage of longer transverse relaxation times, and hence reduced loss of magnetization, pulse sequences that use these nuclei need to be developed.

Recent years have witnessed a tremendous growth in the knowledge base for the role of misfolded proteins in many diseases like Alzheimer's, Parkinson's etc. Thus the study of unfolded proteins, misfolded proteins, and partially folded proteins is bound to provide useful insights into the origin of these diseases and hence help in design of drugs against them.

It is known that many unfolded proteins including intrinsically unfolded proteins (IUP's) have a tendency to aggregate and precipitate. Under such circumstances, solution state NMR studies are impractical and structural characterization has to be done in the solid state. Here, solid state NMR becomes indispensable. On the other hand, the methodologies in solid state NMR of proteins are not as well developed as those for solution state studies, and there are many challenges. Nevertheless, good progress has been made in recent years, and structural reports have appeared in few cases of folded proteins.[104] A significant effort is likely to be focused in this direction in the coming years.

Acknowledgements

We thank the Government of India for financial support to the National Facility for High Field NMR at TIFR.

References

1. Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 *Nucleic Acids Res.* **28** 235
2. Romero P, Obradovic Z, Kissinger C, Villafranca J E, Garner E, Guilliot S and Dunker A K 1998 *Pac. Symp. Biocomput.* **3** 437
3. Dunker A K, Garner E, Guilliot S, Romero P, Albercht K, Hart J, Obradovic Z, Kissinger C and Villafranca J E 1998 *Pac. Symp. Biocomput.* **3** 473
4. Plaxco K W and Groβ M 1997 *Nature* (*London*) **386** 657
5. Wright P E and Dyson H J 1999 *J. Mol. Biol.* **293** 321
6. Adkins J N and Lumb K J 2002 *Protein: Struct. Funct. Genet.* **46** 1
7. Dyson H J and Wright P E 2002 *Curr. Opin. Struct. Biol.* **12** 54
8. Tompa P 2002 *Trends Biochem. Sci.* **27** 527
9. Uversky V N 2002 *Protein Sci.* **11** 739
10. Uversky V N 2002 *Eur. J. Biochem.* **269** 2
11. Dunker A K *et al* 2001 *J. Mol. Graph. Model.* **19** 26
12. Dunker A K, Brown C J, Lawson J D, Iakoucheva L M and Obradovic Z 2002 *Biochemistry* **41** 6573
13. Schweers O, Schönbrunn Hanebeck E, Marx A and Mandelkow E 1994 *J. Biol. Chem.* **269** 24290
14. Lee L, Stollar E, Chang J, Grossman J G, O'Brien R, Ladbury J, Carpenter B, Roberts S and Luisi B 2001 *Biochemistry* **40** 6580
15. Daughdrill G W, Chadsey M S, Karlinsey J E, Haughes K T and Dahlquist E W 1997 *Nature Struct. Biol.* **4** 285
16. Huang S, Ratliff K S and Matouschek A 2002 *Nature Struct. Biol.* **9** 301
17. Demarest S J, Martinez-Yamout M, Chung J, Chen H, Xu W, Dyson, H J, Evans R M and Wright P E 2002 *Nature* (*London*) **415** 549
18. Daggett V and Fersht A R 2003 *Trends Biochem. Sci.* **28** 18
19. Hammarström P and Carlsson U 2000 *Biochem. Biophys. Res. Commun.* **276** 393
20. Dyson H J and Wright P E 1998 *Nature Struct. Biol.* **5** (Suppl.) 499
21. Guntert P 1998 *Q. Rev. Biophys.* **31** 145
22. Ferentz A E and Wagner G 2000 *Q. Rev. Biophy.* **33** 29
23. Kennedy M A, Montelione G T, Arrowsmith C H and Markley J L 2002 *J. Struct. Funct. Genom.* **2** 155
24. Ikura M, Kay L E and Bax A A 1990 *Biochemistry* **29** 4659
25. Ikura M, Marion D, Kay L E, Shih H, Krinks M, Klee C B and Bax A 1990 *Biochem. Pharmacol.* **40** 153
26. Panchal S C, Bhavesh N S and Hosur R V 2001 *J. Biomol. NMR* **20** 135
27. Peti W, Smith L J, Redfield R and Schwalbe H 2001 *J. Biomol. NMR* **19** 153
28. Bhavesh N S, Panchal S C and Hosur R V 2001 *Biochemistry* **40** 14727
29. Chatterjee A, Bhavesh N S, Panchal S C and Hosur R V 2002 *Biochem. Biophys. Res. Commun.* **293** 427
30. Delaglio F, Torchia D A and Bax A 1991 *J. Biomol. NMR* **1** 439
31. Wirmer J and Schwalbe H 2002 *J. Biomol. NMR* **23** 47
32. Wienk H L, Martinez M M, Yalloway G N, Schmidt J M, Perez C, Ruterjans H and Lohr F 2003 *J. Biomol. NMR* **25**, 133
33. Edison A S, Markley J L and Weinhold F 1994 *J. Biomol. NMR* **4** 519
34. Edison A S, Weinhold F, Westler W M and Markley J L 1994 *J. Biomol. NMR* **4** 543
35. Spera S and Bax A 1991 *J. Am. Chem. Soc.* **113** 5490
36. Wishart D S, Sykes B D and Richards F M 1991 *J. Mol. Biol.* **222** 311
37. Wishart D S and Sykes B D 1994 *Methods Enzymol.* **239** 363
38. Wishart D S and Sykes B D 1994 *J. Biomol. NMR* **4** 171
39. de Dios A C, Pearson J G and Oldfield E 1993 *Science* **260** 1491
40. Wishart D S, Bigam C G, Holm A, Hodges R S and Sykes B D 1995 *J. Biomol. NMR* **5** 67
41. Schwarzinger S, Kroon G J A, Foss T R, Wright P E and Dyson H J 2000 *J. Biomol. NMR* **18** 43
42. Schwarzinger S, Kroon G J A, Foss T R, Chung J, Wright P E and Dyson H J 2001 *J. Am. Chem. Soc.* **123** 2970
43. Wuthrich K 1986 *NMR of proteins and nucleic acids* (New York: John Wiley)
44. Dyson H J and Wright P E 2001 *Methods Enzymol.* **339** 258
45. Vuister G W and Bax A 1993 *J. Am. Chem. Soc.* **115** 7772
46. Montelione G, Winkler M E, Rauenbuehler P and Wagner G 1989 *J. Magn. Reson.* **82** 198
47. Arcus V L, Vuilleumier S, Freund S M V, Bycroft M and Fersht A R 1994 *Proc. Natl. Acad. Sci. USA* **91** 9412
48. Arcus V L, Vuilleumier S, Freund S M V, Bycroft M and Fersht A R 1995 *J. Mol. Biol.* **254** 305
49. Bhavesh N S, Panchal S C and Hosur R V 2001 *FEBS Lett.* **509** 218
50. Rose G D, Gierasch L M and Smith J A 1985 *Adv. Protein. Chem.* **37** 1
51. Dyson H J and Wright P E 1991 *Annu. Rev. Biophys. Biophys. Chem.* **20** 519
52. Merutka G, Dyson H J and Wright P E 1995 *J. Biomol. NMR* **5** 14
53. Huyghues-Despointes B M P, Scholtz J M and Pace C N 1999 *Nature. Struct. Biol.* **6** 910
54. Englander S W and Krishna M M 2002 *Nature Struct. Biol.* **8** 741
55. Zhang O, Forman-Kay J D, Shortle D and Kay L E 1997 *J. Biomol. NMR* **9** 181
56. Alba E de and Tjandra N 2002 *Progr. NMR Spectrosc.* **40** 175
57. Shortle D and Ackerman M S 2001 *Science* **293** 487
58. Choy W Y and Forman-Kay J D 2001 *J. Mol. Biol.* **308** 1011

59. Kay L E, Torchia D A and Bax A 1989 *Biochemistry* **28** 8972
60. Kay L E 1998 *Nature Struct. Biol.* **5** (Suppl.) 513
61. Lipari G and Szabo A 1982 *J. Am. Chem. Soc.* **104** 4546
62. Lipari G and Szabo A 1982 *J. Am. Chem. Soc.* **104** 4559
63. Buevich A V and Baun J 1999 *J. Am. Chem. Soc.* **121** 8671
64. Lefèvre J F, Dayie K T, Peng J W and Wagner G 1996 *Biochemistry* **35** 2674
65. Prompers J J and Brüschweiler R 2002 *J. Am. Chem. Soc.* **124** 4522
66. Werbelow L G and Grant D M 1977 *Adv. Magn. Reson.* **9** 189
67. Vold R L and Vold R R 1978 *Prog. NMR Spectrosc.* **12** 79
68. Reif B, Hennig M and Griesinger C 1997 *Science* **276** 1230
69. Yang D, Konrat R and Kay L E 1997 *J. Am. Chem. Soc.* **119** 11938
70. Daiwen Y, Mok Y K, Muhandiram D R, Forman-Kay J D and Kay L E 1999 *J. Am. Chem. Soc.* **121** 3555
71. Daragan V A and Mayo K H 1995 *J. Magn. Reson. Ser.* **B107** 274
72. Dyson H J and Wright P E 2004 *Chem. Rev* **104** 3607
73. Neri D, Wider G and Wuthrich K 1992 *Proc. Natl. Acad. Sci. USA* **89** 4397
74. Neri D, Wider G and Wüthrich K 1992 *FEBS Lett.* **303** 129
75. Neri D, Billeter M, Wider G and Wüthrich K 1992 *Science* **257** 1559
76. Penkett C J, Redfield C, Jones J A, Dodd I, Hubbard J, Smith R A G, Smith L J and Dobson C M 1998 *Biochemistry* **37** 17054
77. Schwalbe H, Fiebig K M, Buck M, Jones J A, Grimshaw S B, Spencer A, Glaser S J, Smith L J and Dobson C M 1997 *Biochemistry* **36** 8977
78. Hennig M, Bermel W, Spencer A, Dobson C M, Smith L J and Schwalbe H 1999 *J. Mol. Biol.* **288** 705
79. Klein-Seetharaman J, Oikawa M, Grimshaw S B, Wirmer J, Duchardt E, Ueda T, Imoto T, Smith L J, Dobson C M and Schwalbe H 2002 *Science* **295** 1719
80. Logan T M, Thériault Y and Fesik S J 1994 *J. Mol. Biol.* **236** 637
81. Eliezer D, Chung J, Dyson H J and Wright P E 2000 *Biochemistry* **39** 2894
82. Cavagnero S, Nishimura C, Schwarzinger S, Dyson H J and Wright P E 2001 *Biochemistry* **40** 14459
83. Yao J, Chung J, Eliezer D, Wright P E and Dyson H J 2001 *Biochemistry* **40** 3561
84. Lietzow M, Jamin M, Dyson H J and Wright P E 2002 *J. Mol. Biol.* **322** 655
85. Schwarzinger S, Wright P E and Dyson H J 2002 *Biochemistry* **41** 12681
86. Wong K B, Freund S M V and Fersht A R 1996 *J. Mol. Biol.* **259** 805
87. Bhavesh N S, Juneja J, Udgaonkar J B and Hosur R V 2004 *Protein Sci.* **13** 3085
88. Juneja J, Bhavesh N S, Udgaonkar J B and Hosur R V 2002 *Biochemistry* **41** 9885
89. Ochsenbein F, Neumann J M, Guittet E and Heijenoort C V 2002 *Protein Sci.* **11** 957
90. Kazmirski S, Wong K B, Freund S M V, Tan Y J, Fersht A R and Daggett V 2001 *Proc. Natl. Acad. Sci. USA* **98** 4349
91. Meekhof A E and Freund S M V 1999 *J. Mol. Biol.* **286** 679
92. Gillespie J R and Shortle D 1997 *J. Mol. Biol.* **268** 158
93. Gillespie J R and Shortle D 1997 *J. Mol. Biol.* **268** 170
94. Ye K and Jinfeng W 2001 *J. Mol. Biol.* **307** 309.
95. Bai Y, Chung J, Dyson H J and Wright P E 2001 *Protein Sci.* **10** 1056
96. Zhang O, Kay L E, Olivier J P and Forman-Kay J D 1994 *J. Biomol. NMR* **4** 727
97. Farrow N A, Zhang O, Forman-Kay J D and Kay L E 1995 *Biochemistry* **34** 868
98. Zhang O, Forman-Kay J D 1995 *Biochemistry* **34** 6784
99. Farrow N A, Zhang O, Forman-Kay J D and Kay L E 1997 *Biochemistry* **36** 2390
100. Mok Y K, Elisseeva E L, Davidson A R and Forman-Kay J D 2001 *J. Mol. Biol.* **307** 913
101. Crowhurst K A, Tollinger M and Forman-Kay J D 2002 *J. Mol. Biol.* **322** 163
102. Crowhurst K A and Forman-Kay J D 2003 *Biochemistry* **42** 8687
103. Bhavesh N S, Sinha S, Mohan P M K and Hosur R V 2003 *J. Biol. Chem.* **278** 19980
104. McDermott A E 2004 *Curr. Opin. Struct. Biol.* **14** 554
105. Chatterjee A, Mridula P, Mishra R K, Mittal R and Hosur R V 2005 *J. Biol. Chem.* (in press)
106. Mishra R K, Jatiani S S, Kumar A, Simhadri V R, Hosur R V and Mittal R 2004 *J. Biol. Chem.* **279** 31445