

Mapping the Potential Energy Landscape of Intrinsically Disordered Proteins at Amino Acid Resolution

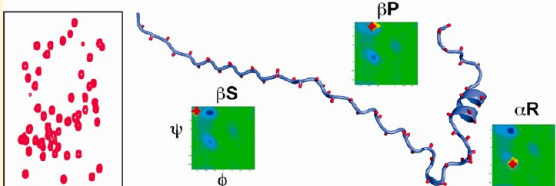
Valéry Ozenne,[†] Robert Schneider,[†] Mingxi Yao,[†] Jie-rong Huang,[†] Loïc Salmon,[†] Markus Zweckstetter,[‡] Malene Ringkjøbing Jensen,[†] and Martin Blackledge^{*,†}

[†]CEA, CNRS, and UJF-Grenoble 1, Protein Dynamics and Flexibility, Institut de Biologie Structurale Jean-Pierre Ebel, 41 Rue Jules Horowitz, Grenoble 38027, France

[‡]Department of NMR-Based Structural Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, and German Center for Neurodegenerative Diseases (DZNE), 37077 Göttingen, Germany

S Supporting Information

ABSTRACT: Intrinsically disordered regions are predicted to exist in a significant fraction of proteins encoded in eukaryotic genomes. The high levels of conformational plasticity of this class of proteins endows them with unique capacities to act in functional modes not achievable by folded proteins, but also places their molecular characterization beyond the reach of classical structural biology. New techniques are therefore required to understand the relationship between primary sequence and biological function in this class of proteins. Although dependences of some NMR parameters such as chemical shifts (CSs) or residual dipolar couplings (RDCs) on structural propensity are known, so that sampling regimes are often inferred from experimental observation, there is currently no framework that allows for a statistical mapping of the available Ramachandran space of each amino acid in terms of conformational propensity. In this study we develop such an approach, combining highly efficient conformational sampling with ensemble selection to map the backbone conformational sampling of IDPs on a residue specific level. By systematically analyzing the ability of NMR data to map the conformational landscape of disordered proteins, we identify combinations of RDCs and CSs that can be used to raise conformational degeneracies inherent to different data types, and apply these approaches to characterize the conformational behavior of two intrinsically disordered proteins, the K18 domain from Tau protein and N_{TAIL} from measles virus nucleoprotein. In both cases, we identify the enhanced populations of turn and helical regions in key regions of the proteins, as well as contiguous strands that show clear and enhanced polyproline II sampling.



INTRODUCTION

The realization that a large fraction of proteins encoded in eukaryotic genomes contain a significant level of functional disorder^{1–4} has engendered considerable interest in the development of experimental and analytical techniques to describe this disorder.^{5–8} The conformational plasticity of intrinsically disordered proteins (IDPs) endows them with unique capabilities to act in functional modes not achievable by folded, globular proteins. A number of different scenarios have been identified for the binding of IDPs to their partner proteins, including folding-upon-binding⁹ or the formation of dynamic, so-called fuzzy complexes¹⁰ where the IDP samples various states on the surface of the partner. However, a number of open questions remain, for example, it is unclear how the intrinsic structural propensity is defined by the primary sequence of an IDP, and how this propensity is related to the thermodynamics and kinetics of the interaction and the conformation adopted in the complex. A full understanding of how IDPs carry out their function in the absence of a stable tertiary fold requires a description of the potential energy landscape sampled by each amino acid in the protein. In order to achieve this end, ensemble representations of a continuum of rapidly interconverting structures have emerged as a convenient

tool for representing the structural and dynamic properties of IDPs and their complexes.^{11–19} In this context, the determination of representative descriptions of the behavior of IDPs remains one of the major challenges for the study of the molecular basis of biological function in these highly disordered systems.

Nuclear magnetic resonance (NMR) spectroscopy represents a tool of choice to address this challenge, providing experimental measurement of site-specific ensemble averages over all conformers sampled up to the millisecond time scale. Of these, the chemical shift (CS) is the most accessible, reporting on the local chemical and electronic environment, as well as medium and long-range interactions.^{20–23} Unfortunately, this conformational dependence is poorly defined at a theoretical level. A popular empirical alternative is to compile experimental CSs measured in folded proteins for which three-dimensional coordinates are available and to establish conformational dependences on this basis.^{24,25} This approach has led to the observation that secondary structural elements such as α -helices and β -sheets can be readily identified on the

Received: July 15, 2012

Published: August 20, 2012

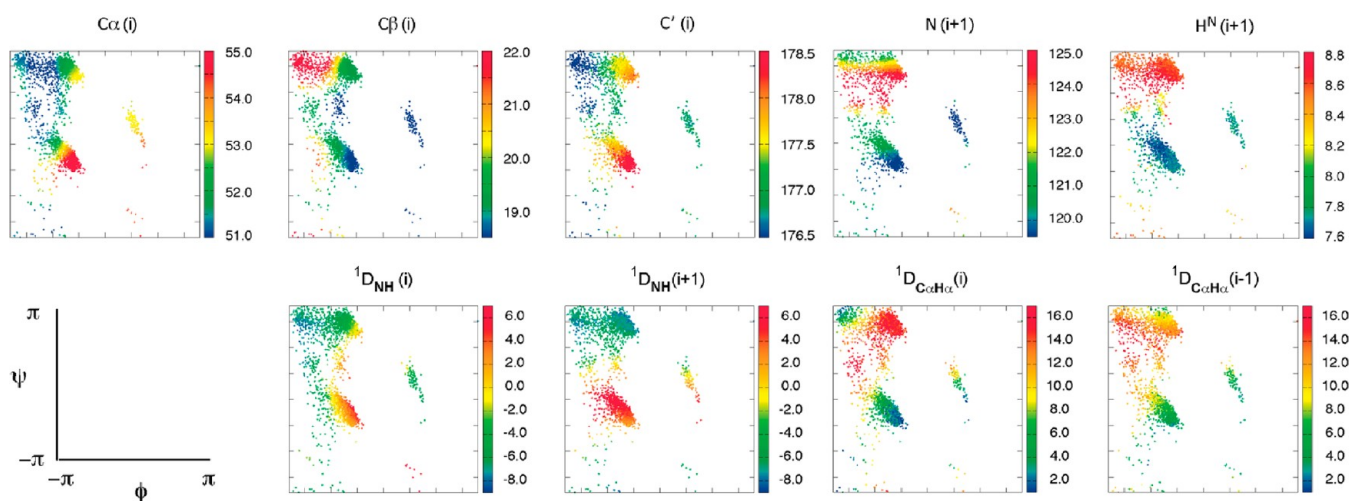


Figure 1. Dependence of primary experimental data on backbone dihedral angle sampling. (A) Distribution of predicted chemical shifts (in ppm) for the central residue $i = 8$ and its neighbor $i = 9$ of a poly-alanine 15-mer chain as function of the conformational sampling $\{\phi, \psi\}$ of residue i . (B) Ensemble averaged backbone RDCs for the poly-alanine 15-mer chain plotted against average $\{\phi, \psi\}$ values of residue i . Values are shown in hertz (Hz) in all cases, assuming an arbitrary level of overall alignment.

basis of the ^{13}C backbone CS.^{20,26–28} Structural restraints based on CS have also been introduced into structure determination algorithms, and the power of CS prediction using database-dependent approaches was further exemplified via their combination with molecular modeling to achieve full structure determination.^{29–31}

The application of CS to the study of disordered systems, where deviation of the shift from its coil value—the secondary shift—is expected to be smaller than in a folded protein, requires a more subtle approach.^{21,27,32,33} Nevertheless, the strong and complementary dependence of $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ shifts on the presence of α -helix and β -sheet conformations has led to the development of simple and accurate algorithms for the determination of the propensity of regions of the protein to form secondary structure in solution.³⁴ Recently CSs have been combined with ensemble selection algorithms^{14,15,35,36} or expressed as the population weighted average of generic CSs from three regions of Ramachandran space (α -helix, β -sheet and polyproline II) and a random coil shift,³⁷ to solve for the populations of these regions. Residual dipolar couplings (RDCs), measured under conditions of weak molecular alignment, are sensitive to the reorientational sampling properties of internuclear bond-vectors, and are therefore also sensitive reporters of the local conformational behavior of IDPs.^{16,38–41} Most applications of RDCs to the studies of disordered systems have exploited the particular ability of RDCs to identify the presence of α -helical and turn elements in otherwise disordered systems,^{42–45} while the combination of different RDCs measured throughout the peptide plane can also detect enhanced sampling of more extended backbone conformations (either β -sheet or polyproline II).^{15,39,46}

Despite intense contemporary interest in this question, it remains unclear how accurately NMR CSs and RDCs can be used to uniquely define backbone conformational sampling in intrinsically disordered proteins, principally because no analytical or numerical framework for the determination of the potential energy landscape of unfolded proteins at amino acid specific resolution is yet available. This question is of additional importance because of the proposed relevance, derived from vibrational spectroscopy and circular dichroism as well as homonuclear NMR, of the polyproline II (PPII) region

of Ramachandran space for the behavior of disordered proteins.^{47–49} The development of a method that unequivocally maps the population of the entire backbone conformational space sampled by each amino acid is therefore of considerable importance.

In this study, we develop an approach to address the ability of primary experimental NMR data, specifically CSs and RDCs, to map the conformational behavior of IDPs on an amino acid specific basis. To achieve this aim, we combine the ensemble selection algorithm ASTEROIDS,¹⁵ with *flexible-meccano*^{50,51} and SPARTA²⁵ to systematically map the sensitivity of different CSs and RDCs to determine the population distribution of each backbone dihedral angle in the protein. This approach provides clear insight into conformational propensities that can be distinguished on the basis of experimental data, and simultaneously identifies regions of Ramachandran space whose populations cannot be resolved. Finally, we propose combinations of RDCs and CSs that can be used to raise these degeneracies and determine populations of all regions of Ramachandran space. The approach is applied to the two experimental cases, N_{TAIL} , the intrinsically disordered C-terminal domain of the nucleoprotein from measles virus, and the K18 domain of the protein Tau, an IDP that is implicated in the development of Alzheimer's disease. In both systems, we identify turn and helical regions as well as the presence of contiguous regions exhibiting enhanced PPII sampling.

RESULTS AND DISCUSSION

Variation of Backbone Chemical Shifts over $\{\phi, \psi\}$ Space. One of the advantages of using CSs as structural probes is that resonances from different nuclei exhibit complementary dependences on backbone dihedral angles $\{\phi, \psi\}$. In principle, this complementarity may allow for a site-specific mapping of the conformational sampling in disordered proteins. The predicted dihedral angle dependence of five experimentally measurable CSs is shown in Figure 1 for an alanine sequence. The conformers were generated using *flexible-meccano* on the basis of the statistical coil model, and the chemical shifts were predicted for each conformer using the program SPARTA.²⁵ To simplify the subsequent discussion, we divide the

Ramachandran plot into four regions: β -sheet (β S), PPII (β P), α -helical (α R) and left handed helix (α L) (Figure 2). We note

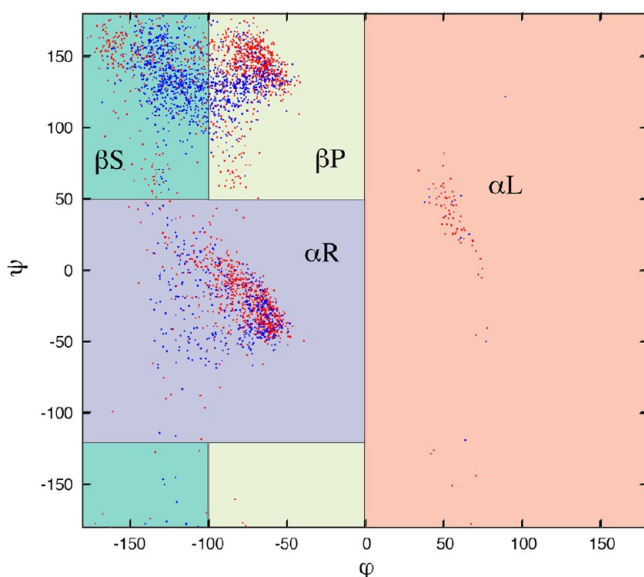


Figure 2. Definition of the regions of Ramachandran space used throughout the study. Points shown are from valine (blue) and alanine (red) residues in statistical coil conformations.

that this definition of conformational space avoids the appearance of bias when mapping specific conformations due to the arbitrary definition of an additional sampling regime termed ‘random coil’ that represents the remaining sampling. In this study, the entire Ramachandran space is mapped in terms of population distributions, or described in terms of these four regions, obviating the need to define an additional ‘random coil’ region.

Well-known dependences are immediately identifiable from Figure 1, with higher values of $^{13}\text{C}^\alpha$ and $^{13}\text{C}^\beta$ shifts uniquely populating α R and β S conformations, respectively. The determination of the populations in other regions of Ramachandran space appears less straightforward. Thus, similar shifts are predicted in the β P and the upper left α R region for $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$ and $^{13}\text{C}'$, making it difficult, on the basis of the ^{13}C CSs alone, to map the populations in these regions. This degeneracy is partially raised by considering the influence of the $\{\phi, \psi\}$ sampling on the CSs of the neighboring amino acids. In particular, ^{15}N and $^1\text{H}^\text{N}$ shifts of the following residue provide additional differentiation of the β P and upper α R regions.

The prediction for the alanine peptide shown in Figure 1 is relevant for this specific sequence. While overall features will be retained for different sequences, considerable variation is observed as a function of the identity of the three amino acids. To develop a better understanding of the ability of ensemble descriptions to define conformational propensities on the basis of CSs, we have therefore performed explicit simulations using synthetic data derived from specific conformational sampling regimes.

Ensemble Mapping of Conformational Propensities from Chemical Shifts. Conformationally biased ensembles obeying specific sampling properties were generated using the *flexible-meccano* algorithm, and averaged CSs were predicted from these ensembles using the program SPARTA. These synthetic data were then used as the target for the ASTEROIDS approach to select subensembles in agreement with these values

(see Methods). Subensembles are selected from a pool of 20000 structures calculated using the amino acid specific potential energy surfaces derived from the statistical coil model. An iterative procedure is then used to modify the potentials to enhance the sampling as a function of each selection until convergence is achieved. It is important to note here that the *flexible-meccano*/ASTEROIDS approach is used as a means to describe the potential energy landscape sampled by the protein backbone. Repetition of the selection procedure (Supporting Information [SI], Figure S2) determines ensembles containing different structures, which are therefore not unique in this sense; however, the backbone sampling characteristics do not vary from one ensemble to another, which are therefore converged and unique in terms of conformational substates and their populations. This also demonstrates that pool sampling is sufficiently complete.

The modulation of the predicted CSs when sampling a specific conformational propensity is compared to statistical coil values in Figure 3a. Three regimes that are significantly different from the statistical coil model were tested, comprising a higher tendency to sample the β S, β P or α R regions (see Methods). Simple inspection reveals that while well-known deviations are seen for ^{13}C shifts in the presence of β S and α R propensity, these CSs are hardly modified by the presence of raised β P population. This is evidently because the mean values of the statistical coil shifts are essentially indistinguishable from β P values (Figure 1). The uncertainties for each CS as determined from predictions for folded proteins are also shown on this Figure 3a.²⁵ It is notable that the expected changes for ^{15}N and $^1\text{H}^\text{N}$ shifts in the presence of enhanced β P sampling are relatively small compared to this uncertainty.

We initially consider two scenarios for selection on the basis of CSs, simulating data sets comprising either $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$ and $^{13}\text{C}'$ or ‘full’ CS sets including $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N and $^1\text{H}^\text{N}$. Figure 3b presents the ability of ASTEROIDS to reproduce conformational tendencies present throughout the protein when using these different combinations of CSs in the target function. In all cases, the simulated data are well reproduced by the selected ensemble (Supporting Information Figure S1). When using CSs from $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$ and $^{13}\text{C}'$ the ASTEROIDS algorithm accurately reproduces the propensity of enhanced conformational sampling in the β S and α R regions (see also Table 1). The population of the β P region is however poorly reproduced, with additional sampling of the upper α R region that appears to compensate for insufficient sampling of β P. Figure 3c shows the comparison of the average Ramachandran space of the five amino acids from each strand (β S, α R and β P) and from the coil regions in between these strands, for the target and selected ensembles. This further highlights the degeneracy of the upper α R and β P regions when only $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$ and $^{13}\text{C}'$ CSs are used in the selection. As expected from consideration of Figure 1, the addition of ^{15}N and $^1\text{H}^\text{N}$ improves this situation considerably; however, the dependence of these shifts on additional factors such as temperature, ionic strength and pH, renders them potentially volatile in terms of conformational mapping. To determine the levels of confidence that can be derived from different CSs, we have therefore applied the same approach to simulated data with Gaussian-based noise levels reflecting the relative accuracy of predictions for the different nuclei (see Methods). The results are summarized in Table 1, and demonstrate that the accuracy of the determination of the populations of β S and α R regions is

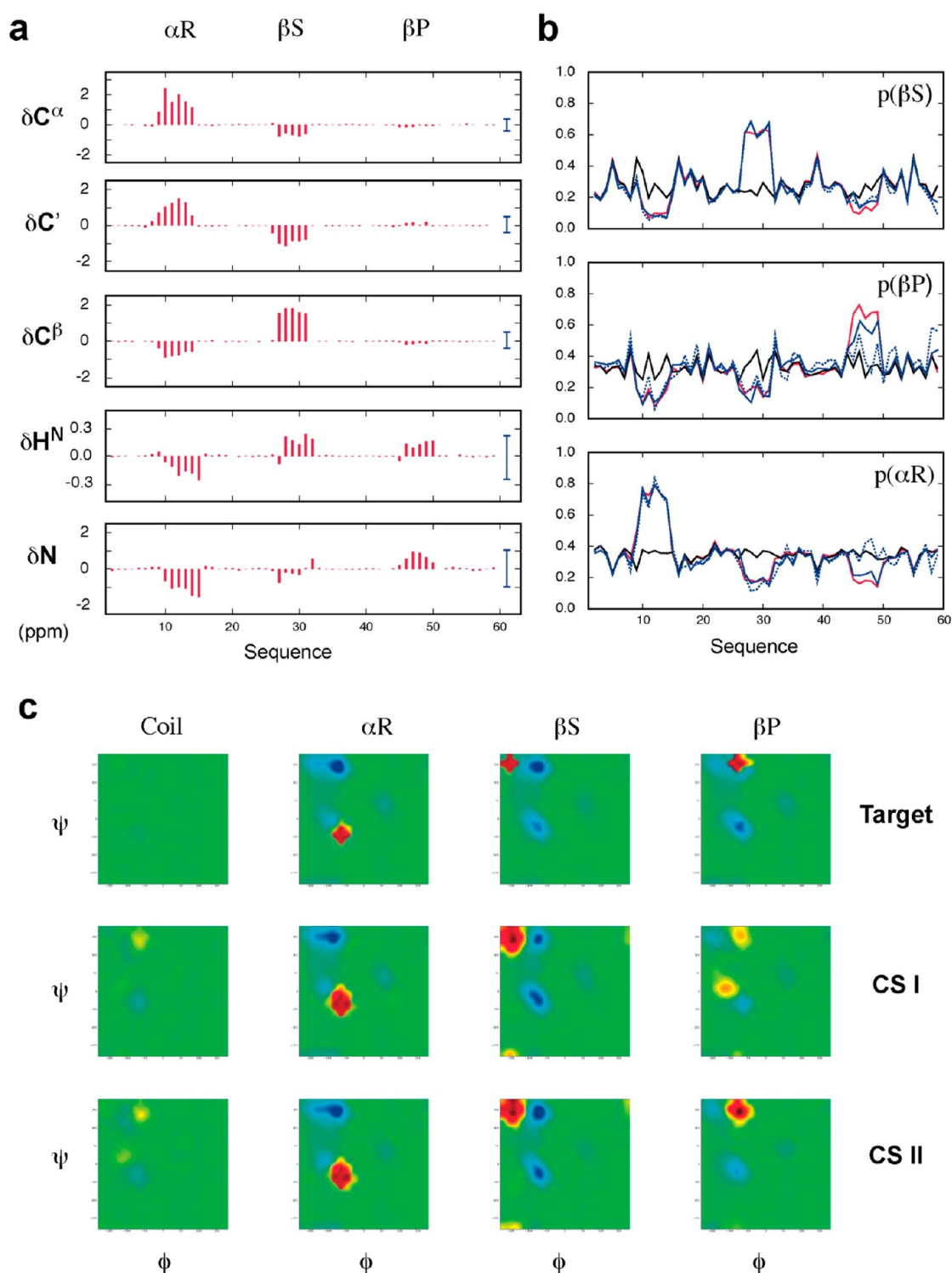


Figure 3. Mapping of conformational space in disordered systems using CSs. (a) Modification of predicted chemical shifts for enhanced conformational propensities in different regions of Ramachandran space compared to statistical coil values. Three regimes that are significantly different from the statistical coil model were tested, comprising a higher tendency to sample the βS , αR and βP regions. Blue error bars indicate the average accuracy to which each chemical shift is predicted for folded proteins. (b) Reproduction of conformational sampling by an ASTEROIDS-selected ensemble comprising 200 conformers obtained by targeting the synthetic chemical shift data set shown in panel a. The pool from which the structures were selected was created using the standard coil library of *flexible-meccano*. Selection carried out using $^{13}C^\alpha$, $^{13}C^\beta$ and $^{13}C'$ chemical shifts or $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$ and ^{15}N , $^1H^N$. Red: populations of conformational space in the target ensemble. Blue: populations in the selected ensemble (dashed line ^{13}C shifts only, solid line all shifts). Black: populations in the starting (statistical coil) ensemble. (c) Ramachandran plots showing the difference compared to statistical coil for the regions of the model peptide sampling coil, αR , βP , and βS regions. Top line, target ensemble; middle line, selection using only ^{13}C CS, bottom line, selection using all CSs. Red, increased sampling; blue, reduced sampling compared to statistical coil.

Table 1. Ability of CSs To Reproduce Conformational Sampling in the Presence and Absence of Noise

Δ^a	βS^b	αR^b	βP^b
Coil ^c	0.45	0.45	0.41
CS I ^d	0.065	0.07	0.35
CS II ^e	0.06	0.08	0.08
CS I σ^f	0.11	0.13	0.41
CS II σ^g	0.18	0.17	0.27
RDC ^h	0.12	0.11	0.27
RDC CS ⁱ	0.07	0.05	0.06
RDC CS σ^j	0.13	0.13	0.19
RDC CS σ^k	0.10	0.13	0.15

^aAll values in the table show average absolute differences between target and selection, averaged over the five amino acid regions experiencing selective enhanced sampling. ^bPopulations averaged over the five amino acids oversampling these regions. ^cDifference between target population and statistical coil average. ^dDifference between target population and selection using $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$ CSs. ^eDifference between target population and selection using $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N , $^1\text{H}^\text{N}$ CSs. ^{f,g}As in *d*, *e* in the presence of Gaussian weighted noise using errors estimated from 25% of the rmsd's of SPARTA predictions of CSs from folded proteins. ^hDifference between target population and selection using $^1\text{D}_{\text{N-H}}$, $^2\text{D}_{\text{C-HN}}$, $^1\text{D}_{\text{Ca-H}\alpha}$ and $^1\text{D}_{\text{Ca-C}}$ RDCs. ⁱDifference between target population and selection using RDCs listed in *h* and $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$ CSs. ^jDifference between target population and selection using $^1\text{D}_{\text{N-H}}$ and $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$ CSs in the presence of noise. ^kAs in *i* in the presence of noise.

significantly more robust to the presence of noise than βP , mainly due to the higher predictive imprecision of ^{15}N and $^1\text{H}^\text{N}$ shifts.

These calculations highlight two important points concerning the use of CSs to map local conformational sampling in disordered systems. The first concerns the inherent degeneracy of CSs for the upper αR and βP regions, which is partially raised by the ^{15}N and $^1\text{H}^\text{N}$ shifts. Second, and more importantly, the expected ^{13}C CSs in the presence of enhanced βP sampling are strongly degenerate with the statistical coil values that are expected from intrinsic sampling in the absence of specific conformational propensity.

Variation of Residual Dipolar Couplings over $\{\phi, \psi\}$ Space. RDCs measured in disordered systems have also been shown to depend strongly on the nature of the backbone conformational sampling. This is illustrated in Figure 1 where different ensemble averaged backbone RDCs are plotted against average $\{\phi, \psi\}$ values (see Methods). The sensitivity of RDCs both to the conformational sampling of the amino acid of interest and its immediate neighbors complicates interpretation of this representation, and underlines the importance of using the ASTEROIDS approach to select ensembles of entire structures. Nevertheless, the most commonly measured RDCs, $^1\text{D}_{\text{N-H}}$ and $^1\text{D}_{\text{Ca-H}\alpha}$ clearly exhibit the expected sensitivity to αR , but also show degeneracy between βS and βP , either for the amino acid of interest or an immediate neighbor. Expected values for RDCs simulated from the sequence containing additional populations of βS , βP and αR presented above are shown in Figure 4a. In this case, all three additional propensities modulate the expected values of RDCs, averaging to different values than the statistical coil, although this modulation is similar for βS and βP .

An ASTEROIDS analysis was performed on the same system, using $^1\text{D}_{\text{N-H}}$, $^2\text{D}_{\text{C-HN}}$, $^1\text{D}_{\text{Ca-H}\alpha}$ and $^1\text{D}_{\text{Ca-C}}$ RDCs in the selection procedure. Figure 4b, 4c and table 1 present the

ability of a combination of these four RDC types to define the conformational potentials. The ASTEROIDS-selected ensemble accurately reproduces the propensity of enhanced conformational sampling in the αR region, and in the extended region (βS and βP together). However the data do not distinguish between these extended regions, in particular the enhanced βP population is not correctly determined. Similarly, upper and lower αR regions are found to be degenerate when using only RDCs.

From the above it is evident that combination of CSs and RDCs should raise the upper $\alpha R/\beta P/\text{coil}$ and $\beta S/\beta P$ degeneracies observed for ^{13}C CSs and RDCs respectively, and thereby allow for a more accurate mapping of Ramachandran space. In the following we test this hypothesis and identify generally accessible and conformationally informative combinations of CS and RDCs that can be usefully applied to the study of a large number of disordered proteins.

Ensemble Mapping of Conformational Propensities by Combining CSs and RDCs. An ASTEROIDS analysis of the same system as illustrated earlier was performed combining $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$ and $^{13}\text{C}'$ CSs with $^1\text{D}_{\text{N-H}}$, $^2\text{D}_{\text{C-HN}}$, $^1\text{D}_{\text{Ca-H}\alpha}$ and $^1\text{D}_{\text{Ca-C}}$ RDCs (SI Figure S3). In this case (Figure 5), a more precise mapping of Ramachandran space is achieved, raising all degeneracies identified for CSs and RDCs alone. Removal of some RDCs, so that only $^1\text{D}_{\text{N-H}}$ RDCs are included, still provides good reproduction of all regions of conformational space. As shown in Table 1, the populations are still correctly reproduced in the presence of significant levels of noise (equivalent to 0.5 Hz error for the $^1\text{D}_{\text{N-H}}$ RDCs).

The combination of $^1\text{D}_{\text{N-H}}$ RDCs and ^{13}C CSs, with or without ^{15}N and $^1\text{H}^\text{N}$ CSs, therefore represents a tractable solution for many experimental studies that is evidently information rich, while remaining robust with respect to uncertainty of experimental conditions, spectral calibration, noise and prediction error. We have therefore applied this approach to two experimental systems.

Application to the Disordered Domain of the Nucleoprotein from Measles Virus. $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N and $^1\text{H}^\text{N}$ CSs and $^1\text{D}_{\text{N-H}}$ RDCs were used to define the conformational sampling of the 125 amino acid intrinsically disordered C-terminal domain of the nucleoprotein of measles virus (Figure 6a). In addition to characterizing the molecular recognition element that comprises a high population of helix as described recently,^{52,53} the 105 unfolded amino acids appear to indicate the presence of a lower population of βS in localized regions of this domain, compared to the statistical coil description (Figure 6b). This reduction is mainly due to higher βP population, in particular for the three continuous regions (435–445), (448–453) and (518–524), where close to 50% of conformers populate this region of Ramachandran space. Figure 8 shows the reproduction of the $^1\text{D}_{\text{N-H}}$ RDCs when only $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N and $^1\text{H}^\text{N}$ CSs are used, testifying that the analysis is both predictive, and not noticeably prone to overfitting.

Application to the K18 Domain of Tau Protein. The same method was applied to the 130 amino acid K18 domain of Tau protein using $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N and $^1\text{H}^\text{N}$ CSs and $^1\text{D}_{\text{N-H}}$ RDCs (Figure 7a). This domain contains four highly homologous repeat sequences, so that the sampling profile necessarily exhibits a repetitive nature. In this case the βS population is again depleted compared to the statistical coil (Figure 7b). The four previously described type I β -turns and

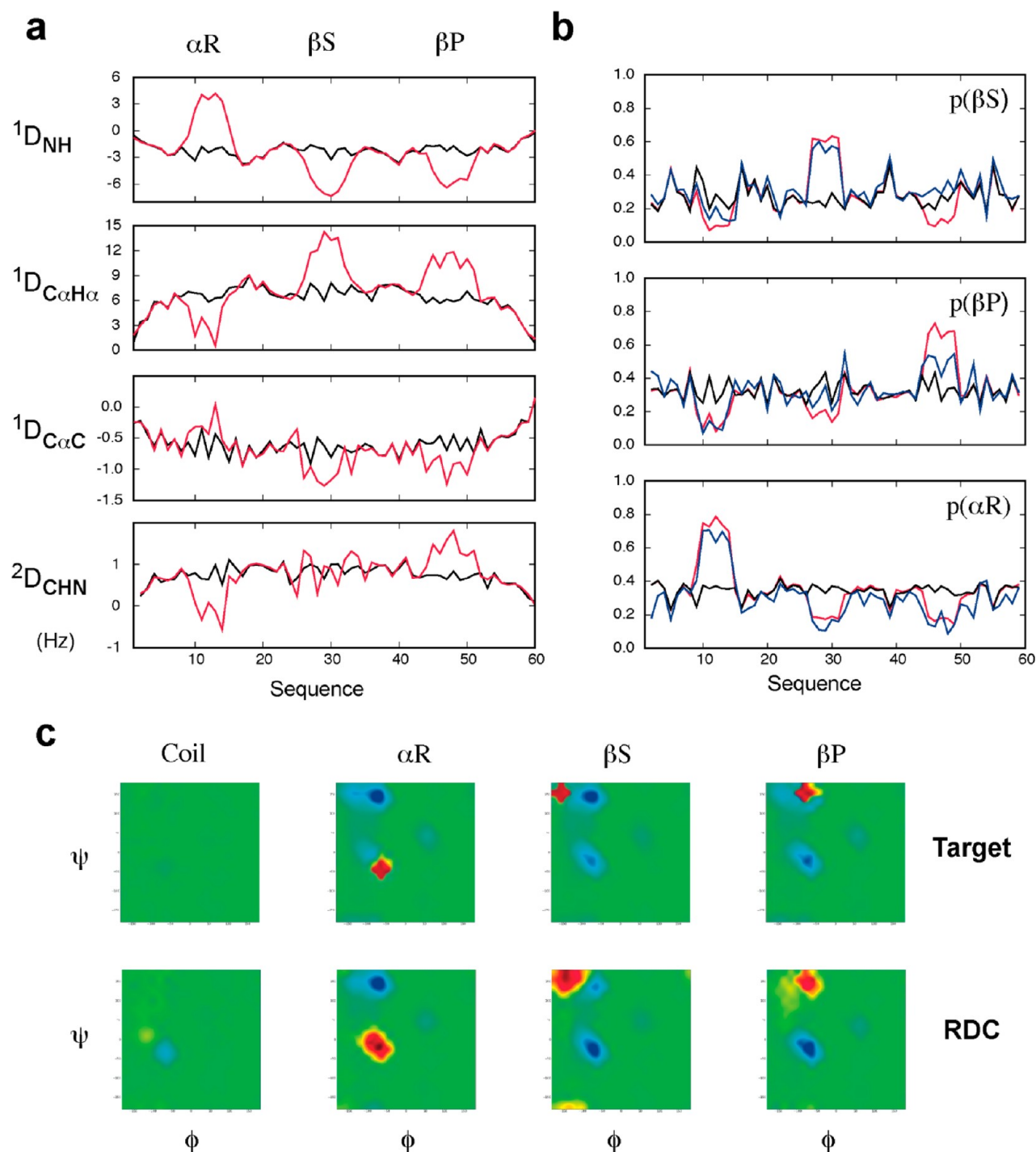


Figure 4. Mapping of conformational space in disordered systems using RDCs. (a) Modification of predicted RDCs for enhanced conformational propensities in different regions of Ramachandran space compared to expected values for statistical coil sampling (see Figure 3). An arbitrary level of alignment was assumed for the absolute scaling of the RDCs. (b) Amino acid specific difference in population between the ASTEROIDS selection and target using simulated RDC data shown in panel a. Red: populations in the target ensemble. Blue: populations in the selected ensemble. Black: populations in the starting (statistical coil) ensemble. (c) Ramachandran plots showing the difference compared to statistical coil for the regions of the model peptide sampling coil, αR , βP , and βS regions. Top line, target ensemble; bottom line, selection using simulated RDC data shown in panel a. Color coding as in Figure 3.

the four triglycine sequences account for the eight regions of significantly increased αR population. The turns are found to be populated between 15 and 25%, spanning very similar ranges to those determined using a combination of accelerated molecular dynamics and RDCs.⁴² Outside these localized regions, a higher population of βP is observed, in particular in

the aggregation nucleation sites, between residues (256–261), (275–282), (307–313) and (338–346). These strands, the central two of which mediate binding to microtubules and have been identified as aggregation nucleation sites important for the formation of Tau oligomers, have previously been proposed to sample extended populations.⁴² The results shown here clearly

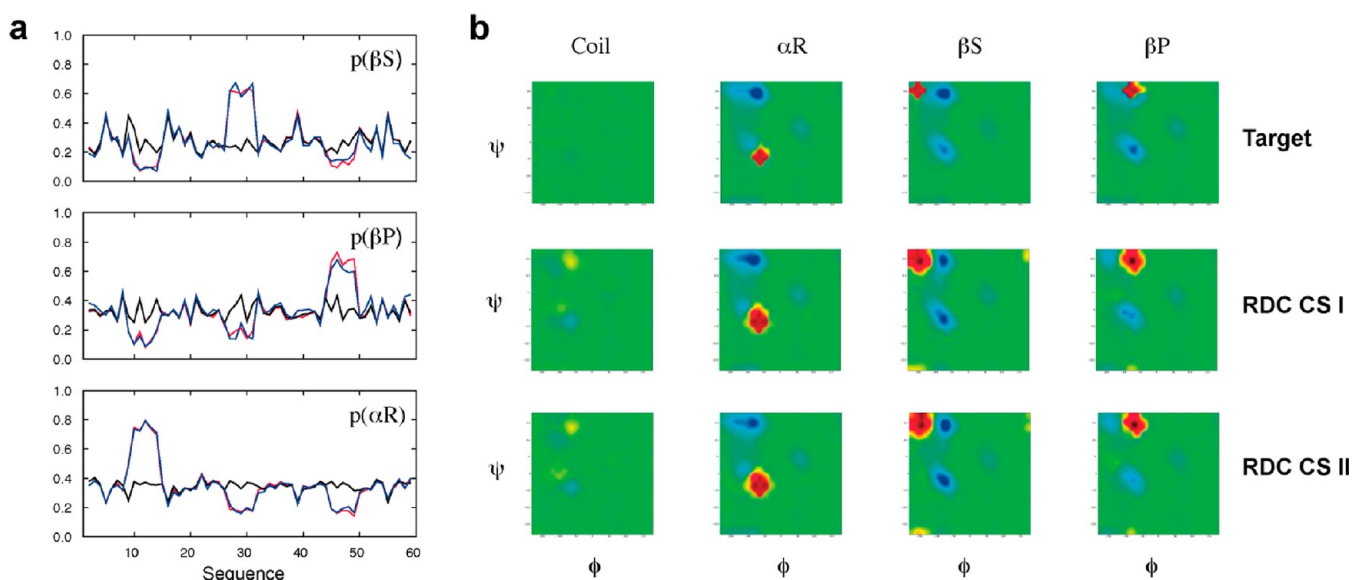


Figure 5. Mapping of conformational space in disordered systems using a combination of RDCs and CSs. (a) Amino acid specific difference in population between the target and the ASTEROIDS selection on the basis of simulated CS and RDC data shown in Figures 3a and 4a. Red: populations in the target ensemble. Blue: populations in the selected ensemble. Black: populations of different regions of conformational space in the starting (statistical coil) ensemble. (b) Ramachandran plots showing the average difference compared to statistical coil for the regions of the model peptide sampling coil, αR , βP , and βS regions. Top line, target ensemble; middle line, selection using ^{13}C CS and $^1\text{D}_{\text{NH}}$ RDCs; bottom, selection using ^{13}C CS and all RDCs shown in Figure 4. Color coding as in Figure 3.

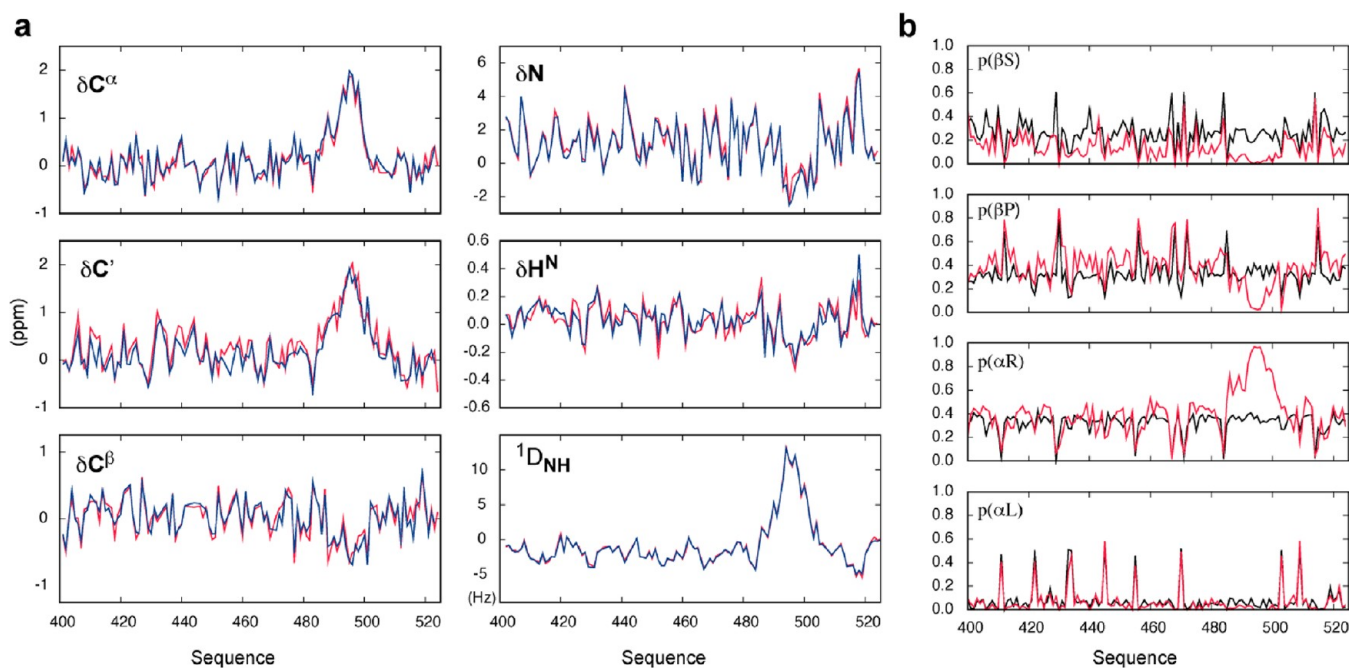


Figure 6. Characterization of intrinsically disordered proteins using RDCs and CSs. ASTEROIDS CS-RDC approach applied to experimental data from the disordered C-terminal domain, N_{TAIL} , of the nucleoprotein from measles virus. (a) Reproduction of experimental data (red experimental, blue ensemble average). (b) Population of different regions of conformational space for each amino acid in the N_{TAIL} sequence (red selected ensemble, black statistical coil).

indicate that this extended sampling is due to strongly enhanced sampling of the βP region of conformational space over a continuous range of 6–9 amino acids. Figure 8 shows the reproduction of the $^1\text{D}_{\text{N-H}}$ RDCs when only $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N and $^1\text{H}^{\text{N}}$ CSs are used; the ‘free’ data are again closely reproduced.

The amino acid conformational potentials for the region 273–287 of K18 are shown in Figure 9, in comparison to the

statistical coil sampling. The raised βP sampling in the region 275–282 is evident, as is the partially populated β -turn that immediately follows this. We note that this conformational sampling, determined in this case uniquely from the experimental data, is very similar to that predicted by accelerated molecular dynamics simulation in a previous study,⁴² populating enhanced αR in Leu284 and Ser285 to very similar levels.

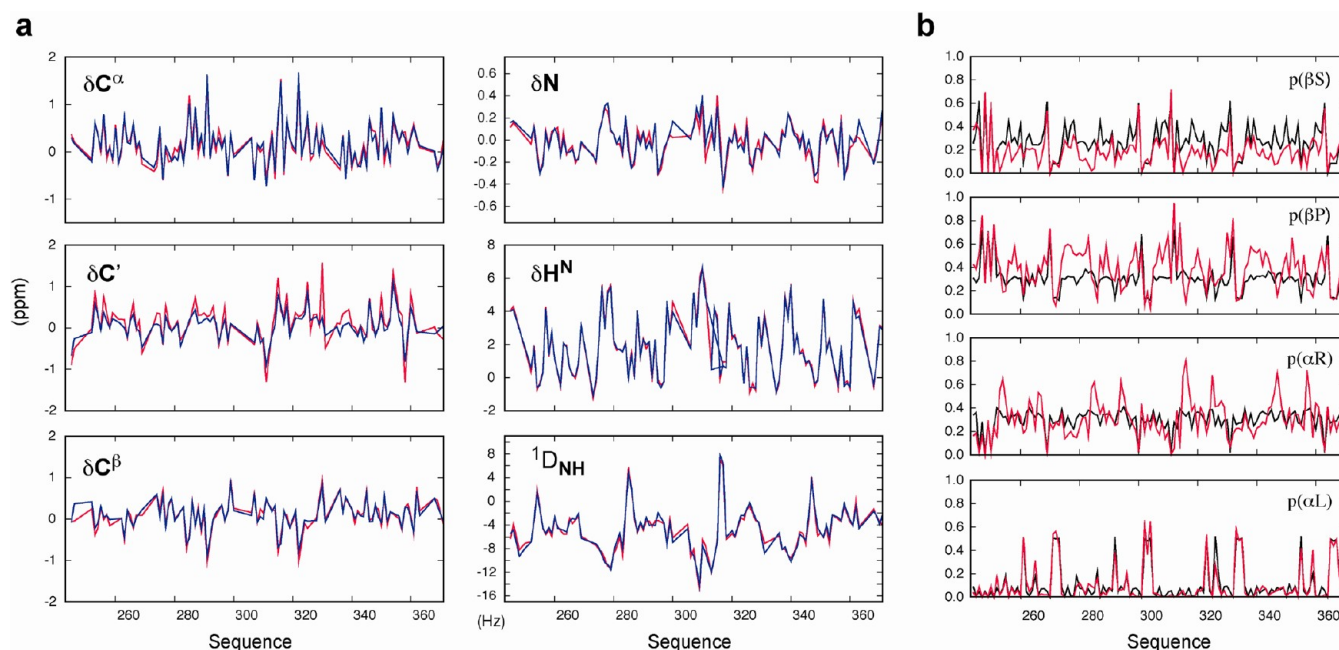


Figure 7. ASTERIODS CS-RDC approach applied to experimental data from the K18 fragment of Tau protein. (a) Reproduction of experimental data (red experimental, blue ensemble average). (b) Population of different regions of conformational space for each amino acid in the K18 sequence (red selected ensemble, black statistical coil).

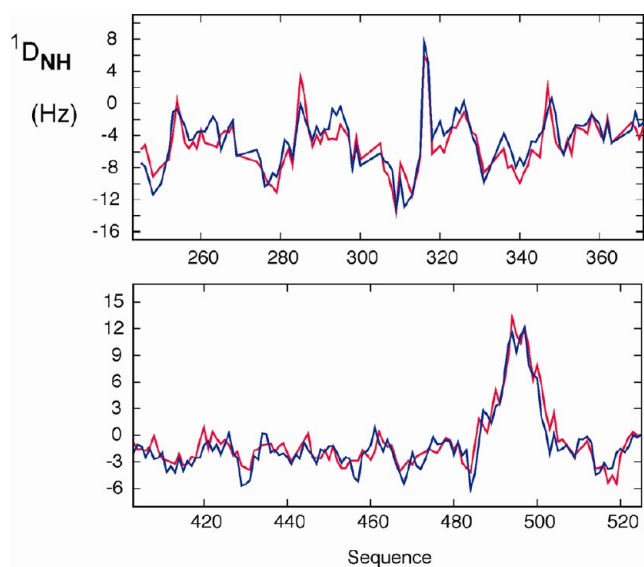


Figure 8. Cross validation of data not used in the ensemble selection procedure. Top: K18 fragment of Tau protein. Bottom: Disordered C-terminal domain, N_{TAIL} , of the nucleoprotein from measles virus. In both cases, back-calculated $^1D_{NH}$ values (blue) from the ensemble selected against $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$, ^{15}N and $^1H^N$ CSs are compared to the experimental data (red).

Finally, we note that this entire study was repeated using the program SPARTA+,⁵⁴ and the results concerning both experimental systems are essentially indistinguishable in terms of conformational sampling (data not shown), indicating that the analysis is robust at least with respect to the differences between these two prediction programs.

CONCLUSION

It is becoming increasingly clear that intrinsic disorder plays a central role in the function of a significant fraction of both

eukaryotic and prokaryotic proteins. The development of an atomic resolution description of the conformational behavior of disordered proteins is a fundamental requirement if we are to understand their biological activity on a molecular level, and NMR represents potentially the most powerful source of this information. However, the actual resolution to which the amino acid specific potential energy surface can be mapped from experimental data remains obscure. Although dependences of some NMR parameters on structural propensities in disordered systems are known, so that sampling regimes are often inferred from experimental observations, there is currently no framework that allows for a statistical mapping of the available Ramachandran space of each amino acid in terms of conformational propensity. In this study, we address this question by combining highly efficient conformational sampling with ensemble selection to systematically investigate the ability of different sources of NMR data to map the backbone conformational sampling of IDPs on a residue specific level.

The results provide clear insight into conformational propensities that can be distinguished on the basis of experimentally available data. While backbone ^{13}C chemical shifts can be used to accurately determine the populations of βS and αR regions of Ramachandran space, clear degeneracies exist, in particular concerning the βP region, which is degenerate with average values predicted for random statistical coil sampling. This degeneracy can be raised by ^{15}N and $^1H^N$ shifts, although the prediction accuracy of these shifts is lower. Extending our analysis to commonly measured RDCs confirms the ability of this kind of measurement to distinguish between extended and helical bias, but also identifies a distinct degeneracy, this time between the βS and βP regions.

We demonstrate that a simple combination of RDCs and CSs raises inherent degeneracies to accurately resolve backbone conformational propensities. On the basis of these results, we propose a robust and generally applicable approach for the mapping of conformational potentials uniquely from exper-

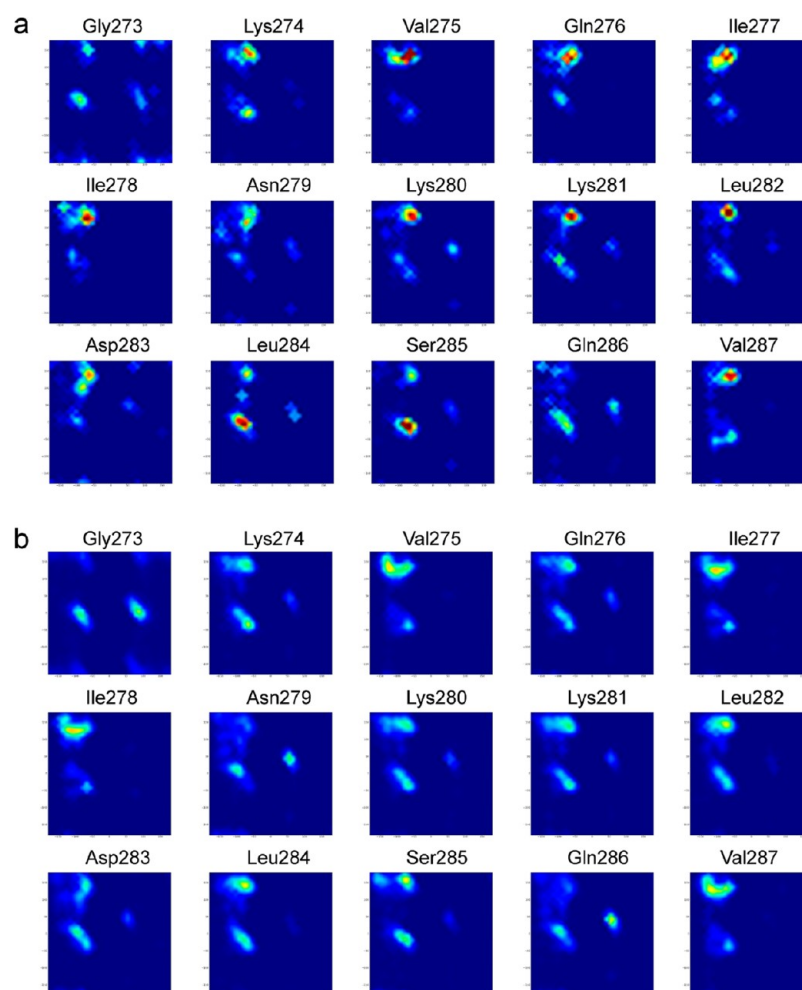


Figure 9. Ramachandran plots showing the amino acid specific conformational potentials in the 273–287 section of K18. (a) Selection from $^1\text{D}_{\text{N-H}}$, $^{13}\text{C}^\alpha$, $^{13}\text{C}^\beta$, $^{13}\text{C}'$, ^{15}N and $^1\text{H}^{\text{N}}$ CSs using the ASTERIODS approach for which the results are shown in Figure 7. (b) Conformational sampling from the statistical coil model. Dark blue represents lowest population, and red represents maximal sampling.

imental data, that is applied to two different biological systems. In both cases, we detect an increase of conformational sampling in the βP region compared to the standard statistical coil description, supporting previous experimental indications from vibrational spectroscopy and circular dichroism for the importance of this region in IDPs. Although the approach is amino acid specific, in many cases these regions are continuous, strongly suggesting that the observation is physically meaningful, but also suggesting that this is not simply a general feature, rather dependent on an underlying dependence on primary sequence. Using these approaches, a more extensive study of a broad range of experimentally available IDPs is currently underway in our laboratory, to determine whether general trends can be identified relating primary sequence composition to backbone conformational behavior.

More generally we are confident that the results from this study will pave the way to a more accurate understanding of the conformational propensities of disordered proteins in solution, and thereby provide hitherto inaccessible insight into the relationship between primary sequence and protein function in this fascinating family of proteins.

METHODS

Calculation of Average Chemical Shifts and RDCs in Ramachandran Space. The information content of the different

chemical shifts was investigated by generating a 50 000-strong ensemble of poly-alanine pentadecapeptide chains using the ensemble generation algorithm *flexible-meccano*.^{50,51} For each conformer, the CSs were calculated using the prediction algorithm SPARTA,²⁵ and conformers were clustered into bins with a radius of 1° according to the $\{\phi, \psi\}$ values of the central amino acid (residue 8). The CSs within each cluster were then averaged and plotted against the $\{\phi, \psi\}$ value of the central amino acid.

Similarly, the information content of different types of RDCs was investigated. An ensemble consisting of 1 000 000 conformers of the poly-alanine pentadecapeptide was created using *flexible-meccano*. RDCs were predicted using PALES⁵⁵ for each conformer and averaged in a similar way as described above for the CSs. The averaged RDCs of the central or neighboring amino acids were plotted against the $\{\phi, \psi\}$ sampling of the central amino acid.

Generation of Synthetic CS and RDC Data Sets in the Presence of Specific Conformational Sampling Regimes. To test the ability of different experimental CSs and RDCs to map conformational space, ensemble selections were carried out using ASTERIODS targeting synthetic data sets. A model protein of 60 amino acids of arbitrary sequence was chosen sampling the statistical coil model except for three regions of five amino acids, where enhanced propensity was introduced in the αR (aa 10–14), βS (aa 27–31) or βP (aa 45–49) regions. Each propensity was introduced such that 50% of the conformers in each strand populate the Ramachandran region of interest, and the remaining 50% populate the statistical coil. An ensemble comprising 10 000 conformers of this model protein was generated using *flexible-meccano*, and CSs were

predicted for each conformer using SPARTA. The CSs were subsequently averaged over the ensemble and used as the target for the ASTEROIDS protocol.

To generate the synthetic RDC data set, an ensemble comprising 100 000 conformers of the same sequence was generated. A global alignment tensor was calculated for each conformer using an in-house written routine based on steric exclusion volume and the RDCs were calculated using this tensor. The RDCs were subsequently averaged over the ensemble and used as the target for the ASTEROIDS protocol.

To test the robustness of the ASTEROIDS protocol for mapping conformational space using CSs and RDCs, Gaussian-based noise was added to the synthetic CS and RDC data sets. The noise levels were based on the relative accuracy of SPARTA predictions for the different nuclei²⁵ and the predicted range of each dipolar coupling type. The following noise levels were applied: C^α (0.22 ppm), C^β (0.24 ppm), C' (0.25 ppm), N (0.6 ppm), H^N (0.12 ppm), $^1D_{N-H}$ (0.5 Hz), $^2D_{C'-HN}$ (0.25 Hz), $^1D_{Ca-Ha}$ (1 Hz) and $^1D_{Ca-C'}$ (0.25 Hz).

Ensemble Selections Using ASTEROIDS. Initially, a large pool of statistical coil conformers (20 000) was generated using *flexible-meccano*^{50,51} and the genetic algorithm ASTEROIDS was used to select a subset of conformers in agreement with the experimental (or synthetic) data as described previously.¹⁵ This procedure was repeated in an iterative manner in order to enhance the presence of conformational propensities of interest within the pool. Thus, in each step, a new pool was generated using the residue-specific $\{\phi, \psi\}$ angles derived from the selected ASTEROIDS ensembles in the previous iteration. Five independent ensemble selections comprising 200 conformers were carried out at each iteration step and iterations were continued until convergence. RDCs were calculated from a given member of an ensemble using the local alignment window (LAW) of 15 amino acids in length combined with a generic baseline as described previously.^{15,36} The alignment tensor was calculated for each LAW using an in-house written routine based on steric alignment. A uniform scaling was applied to the entire predicted set to best reproduce the experimental data. CSs were calculated for each structure using the program SPARTA, and random coil values for calculation of secondary shifts were taken from RefDB.²⁷

Experimental Data: C-Terminal Domain of Measles Virus Nucleoprotein. Experimental CSs of the intrinsically disordered C-terminal domain of Measles virus nucleoprotein were obtained previously at 25 °C in a buffer consisting of 50 mM sodium phosphate at pH 6.5, 50 mM NaCl, 1 mM EDTA and 0.02% NaN₃.⁵³ $^1D_{N-H}$ RDCs were measured previously under the same conditions in a liquid crystal composed of poly-ethylene glycol and 1-hexanol.⁵²

Experimental Data: K18 Construct of Tau Protein. Experimental CSs of the K18 construct of Tau were obtained as described previously.⁵⁶ CS prediction using SPARTA relies on a database of 200 high-resolution structures for which nearly complete sets of chemical shift assignments are available. These CS assignments were obtained at temperatures above 20 °C with the vast majority lying between 20 and 30 °C. To avoid any bias, we calculated the CSs of K18 corresponding to 25 °C by comparing the 5 °C assignment of K18 to the 25 °C assignment of full-length Tau⁵⁷ and subsequently applying a uniform shift to each nucleus type independently. These new experimental data were used as the target for the ASTEROIDS protocol. $^1D_{N-H}$ RDCs of the K18 construct were measured previously in stretched polyacrylamide gels.⁴²

■ ASSOCIATED CONTENT

📄 Supporting Information

Figures showing the reproduction of synthetic data from the fits shown in Figures 3–5. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

martin.blackledge@ibs.fr

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge the Commissariat à l'énergie atomique, the CNRS and the Université Joseph Fourier (Grenoble). This work was supported financially by the ANR under the following projects: ProteinDisorder (JCJC 2010), TAUSTRUCT (MALZ 2010) and by FINOVI.

■ REFERENCES

- (1) Uversky, V. N. *Protein Sci.* **2002**, *11*, 739–756.
- (2) Dunker, A. K.; Brown, C. J.; Lawson, J. D.; Iakoucheva, L. M.; Obradović, Z. *Biochemistry* **2002**, *41*, 6573–6582.
- (3) Tompa, P. *Curr. Opin. Struct. Biol.* **2011**, *21*, 419–425.
- (4) Dyson, H. J.; Wright, P. E. *Chem. Rev.* **2004**, *104*, 3607–3622.
- (5) Dyson, H. J.; Wright, P. E. *Nat. Rev. Mol. Cell Biol.* **2005**, *6*, 197–208.
- (6) Meier, S.; Blackledge, M.; Grzesiek, S. *J. Chem. Phys.* **2008**, *128*, 052204.
- (7) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3–14.
- (8) Schneider, R.; Huang, J.; Yao, M.; Communie, G.; Ozenne, V.; Mollica, L.; Salmon, L.; Jensen, M. R.; Blackledge, M. *Mol. BioSyst.* **2012**, *8*, 58–68.
- (9) Wright, P. E.; Dyson, H. J. *Curr. Opin. Struct. Biol.* **2009**, *19*, 31–38.
- (10) Tompa, P.; Fuxreiter, M. *Trends Biochem. Sci.* **2008**, *33*, 2–8.
- (11) Smith, L. J.; Bolin, K. A.; Schwalbe, H.; MacArthur, M. W.; Thornton, J. M.; Dobson, C. M. *J. Mol. Biol.* **1996**, *255*, 494–506.
- (12) Lindorff-Larsen, K.; Kristjansdottir, S.; Teilmann, K.; Fieber, W.; Dobson, C.; Poulsen, F.; Vendruscolo, M. *J. Am. Chem. Soc.* **2004**, *126*, 3291–3299.
- (13) Kristjansdottir, S.; Lindorff-Larsen, K.; Fieber, W.; Dobson, C. M.; Vendruscolo, M.; Poulsen, F. M. *J. Mol. Biol.* **2005**, *347*, 1053–1062.
- (14) Marsh, J. A.; Forman-Kay, J. D. *J. Mol. Biol.* **2009**, *391*, 359–374.
- (15) Nodet, G.; Salmon, L.; Ozenne, V.; Meier, S.; Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2009**, *131*, 17908–17918.
- (16) Jensen, M. R.; Markwick, P. R. L.; Meier, S.; Griesinger, C.; Zweckstetter, M.; Grzesiek, S.; Bernadó, P.; Blackledge, M. *Structure* **2009**, *17*, 1169–1185.
- (17) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. *J. Am. Chem. Soc.* **2007**, *129*, 5656–5664.
- (18) Esteban-Martín, S.; Fenwick, R. B.; Salvatella, X. *J. Am. Chem. Soc.* **2010**, *132*, 4626–4632.
- (19) Huang, J.; Grzesiek, S. *J. Am. Chem. Soc.* **2010**, *132*, 694–705.
- (20) Wishart, D. S.; Sykes, B. D. *J. Biomol. NMR* **1994**, *4*, 171–180.
- (21) Schwarzing, S.; Kroon, G. J.; Foss, T. R.; Chung, J.; Wright, P. E.; Dyson, H. J. *J. Am. Chem. Soc.* **2001**, *123*, 2970–2978.
- (22) Wang, Y.; Jardetzky, O. *J. Am. Chem. Soc.* **2002**, *124*, 14075–14084.
- (23) Osapay, K.; Case, D. A. *J. Biomol. NMR* **1994**, *4*, 215–230.
- (24) Neal, S.; Nip, A. M.; Zhang, H.; Wishart, D. S. *J. Biomol. NMR* **2003**, *26*, 215–240.
- (25) Shen, Y.; Bax, A. *J. Biomol. NMR* **2007**, *38*, 289–302.
- (26) Yao, J.; Chung, J.; Eliezer, D.; Wright, P. E.; Dyson, H. J. *Biochemistry* **2001**, *40*, 3561–3571.
- (27) Zhang, H.; Neal, S.; Wishart, D. S. *J. Biomol. NMR* **2003**, *25*, 173–195.
- (28) Cornilescu, G.; Delaglio, F.; Bax, A. *J. Biomol. NMR* **1999**, *13*, 289–302.
- (29) Cavalli, A.; Salvatella, X.; Dobson, C.; Vendruscolo, M. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 9615–9620.
- (30) Shen, Y.; Lange, O.; Delaglio, F.; Rossi, P.; Aramini, J.; Liu, G.; Eletsky, A.; Wu, Y.; Singarapu, K.; Lemak, A.; Ignatchenko, A.; Arrowsmith, C.; Szyperski, T.; Montelione, G.; Baker, D.; Bax, A. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 4685–4690.

- (31) Berjanskii, M.; Tang, P.; Liang, J.; Cruz, J. A.; Zhou, J.; Zhou, Y.; Bassett, E.; MacDonell, C.; Lu, P.; Lin, G.; Wishart, D. S. *Nucleic Acids Res.* **2009**, *37*, W670–677.
- (32) De Simone, A.; Cavalli, A.; Hsu, S.-T. D.; Vranken, W.; Vendruscolo, M. *J. Am. Chem. Soc.* **2009**, *131*, 16332–16333.
- (33) Tamiola, K.; Acar, B.; Mulder, F. A. A. *J. Am. Chem. Soc.* **2010**, *132*, 18000–18003.
- (34) Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. *Protein Sci.* **2006**, *15*, 2795–2804.
- (35) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 1270–1272.
- (36) Salmon, L.; Nodet, G.; Ozenne, V.; Yin, G.; Jensen, M.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2010**, *132*, 8407–8418.
- (37) Camilloni, C.; De Simone, A.; Vranken, W. F.; Vendruscolo, M. *Biochemistry* **2012**, *51*, 2224–2231.
- (38) Mohana-Borges, R.; Goto, N. K.; Kroon, G. J. A.; Dyson, H. J.; Wright, P. E. *J. Mol. Biol.* **2004**, *340*, 1131–1142.
- (39) Meier, S.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 9799–9807.
- (40) Obolensky, O. I.; Schlepckow, K.; Schwalbe, H.; Solov'yov, A. V. *J. Biomol. NMR* **2007**, *39*, 1–16.
- (41) Louhivuori, M.; Pääkkönen, K.; Fredriksson, K.; Permi, P.; Lounila, J.; Annala, A. *J. Am. Chem. Soc.* **2003**, *125*, 15647–15650.
- (42) Mukrasch, M. D.; Markwick, P.; Biernat, J.; Bergen, M.; von; Bernadó, P.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M.; Blackledge, M. *J. Am. Chem. Soc.* **2007**, *129*, 5235–5243.
- (43) Jensen, M. R.; Houben, K.; Lescop, E.; Blanchard, L.; Ruigrok, R. W. H.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 8055–8061.
- (44) Jensen, M. R.; Blackledge, M. *J. Am. Chem. Soc.* **2008**, *130*, 11266–11267.
- (45) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 5762–5767.
- (46) Huang, J.; Gabel, F.; Jensen, M. R.; Grzesiek, S.; Blackledge, M. *J. Am. Chem. Soc.* **2012**, *134*, 4429–4436.
- (47) Shi, Z.; Chen, K.; Liu, Z.; Kallenbach, N. R. *Chem. Rev.* **2006**, *106*, 1877–1897.
- (48) Maiti, N. C.; Apetri, M. M.; Zagorski, M. G.; Carey, P. R.; Anderson, V. E. *J. Am. Chem. Soc.* **2004**, *126*, 2399–2408.
- (49) Woody, R. W. *J. Am. Chem. Soc.* **2009**, *131*, 8234–8245.
- (50) Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 17002–17007.
- (51) Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J.-R.; Jensen, M. R.; Segard, S.; Bernadó, P.; Charavay, C.; Blackledge, M. *Bioinformatics* **2012**, *28*, 1463–1470.
- (52) Jensen, M. R.; Communie, G.; Ribeiro, E. A., Jr; Martinez, N.; Desfosses, A.; Salmon, L.; Mollica, L.; Gabel, F.; Jamin, M.; Longhi, S.; Ruigrok, R. W. H.; Blackledge, M. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 9839–9844.
- (53) Gely, S.; Lowry, D. F.; Bernard, C.; Jensen, M. R.; Blackledge, M.; Costanzo, S.; Bourhis, J.-M.; Darbon, H.; Daughdrill, G.; Longhi, S. *J. Mol. Recognit.* **2010**, *23*, 435–447.
- (54) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13–22.
- (55) Zweckstetter, M. *Nat. Protoc.* **2008**, *3*, 679–690.
- (56) Mukrasch, M. D.; Biernat, J.; von Bergen, M.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *J. Biol. Chem.* **2005**, *280*, 24978–24986.
- (57) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. *PLoS Biol.* **2009**, *7*, e34.