

Chapter 4

Ensemble Calculation for Intrinsically Disordered Proteins Using NMR Parameters

Jaka Kragelj, Martin Blackledge and Malene Ringkjøbing Jensen

Abstract Intrinsically disordered proteins (IDPs) perform their function despite their lack of well-defined tertiary structure. Residual structure has been observed in IDPs, commonly described as transient/dynamic or expressed in terms of fractional populations. In order to understand how the protein primary sequence dictates the dynamic and structural properties of IDPs and in general to understand how IDPs function, atomic-level descriptions are needed. Nuclear magnetic resonance spectroscopy provides information about local and long-range structure in IDPs at amino acid specific resolution and can be used in combination with ensemble descriptions to represent the dynamic nature of IDPs. In this chapter we describe sample-and-select approaches for ensemble modelling of local structural propensities in IDPs with specific emphasis on validation of these ensembles.

Keywords Structure · Dynamics · Conformational ensembles · Experimental validation

1 Introduction

Structural biology is an important branch of the life sciences. The number of protein structures deposited in the Protein Data Bank (PDB)¹ is already exceeding 100000 and underlines the enormous effort that has been invested in solving ever-newer protein structures. The description of protein motion can be seen as the next logical step stemming from this wealth of structural data, strongly supported by the fact that proteins display functional dynamics occurring on a broad range of timescales

¹ www.pdb.org.

M. R. Jensen (✉) · J. Kragelj · M. Blackledge
IBS, University Grenoble Alpes, 38044 Grenoble, France
e-mail: malene.ringkjobering-jensen@ibs.fr

IBS, CNRS, 38044 Grenoble, France

IBS, CEA, 38044 Grenoble, France

(Karplus and Kuriyan 2005; Mittermaier and Kay 2006; Henzler-Wildman and Kern 2007; Bernadó and Blackledge 2010). Nuclear magnetic resonance (NMR) spectroscopy is uniquely suited to probing protein dynamics at atomic resolution as a number of experimental parameters report on motions occurring on different time scales ranging from pico- to millisecond (Mittermaier and Kay 2009; Salmon et al. 2011; Göbl et al. 2014).

Protein motion comes in many flavours and can span from local backbone and side chain dynamics in globular, folded proteins (Lindorff-Larsen et al. 2005; Bouvignies et al. 2005; Lange et al. 2008; Salmon et al. 2009; Salmon et al. 2012; Guerry et al. 2013) through the concerted motion of entire domains in multi-domain proteins (Bertini et al. 2007; Yang et al. 2010; Rózycki et al. 2011; Francis et al. 2011; Deshmukh et al. 2013; Huang et al. 2014) to intrinsically disordered proteins (IDPs), which represent the most extreme case of protein flexibility (Dyson and Wright 2002; Dunker et al. 2008; Tompa 2012). One way of representing the dynamics of a protein is to capture its characteristics—or more accurately, to explain the experimental NMR data, which depend on the underlying dynamics—with an ensemble of protein structures (Fig. 4.1).

In this chapter we will focus on atomic resolution ensemble descriptions of IDPs on the basis of experimental NMR data with a special emphasis on mapping local conformational propensities. The determination of a single set of three-dimensional atomic coordinates would have little meaning for these conformationally heterogeneous molecules, and ensemble descriptions are therefore necessary in order to build molecular models of IDPs that accurately capture the dynamic behaviour of the polypeptide chains. Special care has to be taken at each step of the ensemble generation protocol to ensure the validity of the obtained ensembles. The way in which ensemble generation protocols are tested and the factors that influence the modelling of the ensembles are therefore important questions that need to be addressed. In this chapter we will discuss these issues with a focus on the application of sample-and-select approaches to mapping local conformational propensities in IDPs.

2 Local Structure in IDPs can be Described by the Dihedral Angle Distributions of Amino Acids

It is expected that a single residue in an IDP will adopt many conformations over the time and ensemble average, and therefore undergoes exchange among many different dihedral angles. The distribution of dihedral angles sampled by a residue may at first seem like a simplistic representation of residual structure but is in reality very practical. The well-known secondary structures, the α -helix and β -sheet, are defined by hydrogen-bonding criteria and also have their own characteristic dihedral angles that are commonly used for annotating secondary structure elements in proteins (Fig. 4.2a, 4.2c) (Kabsch and Sander 1983; Frishman and Argos 1995).

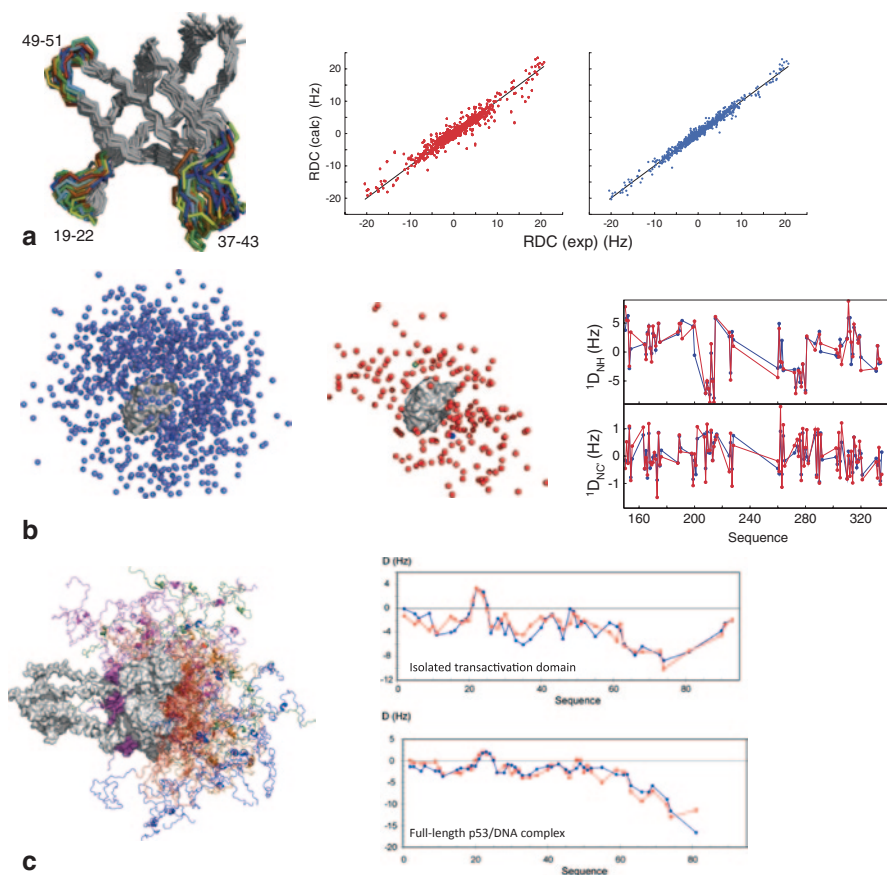


Fig. 4.1 Interpreting NMR data with molecular ensembles to map conformational dynamics in proteins. **a** Dynamics of the SH3 domain from CD2AP derived from NMR residual dipolar couplings (RDCs) measured in multiple, complementary alignment media. An ensemble is shown of the SH3 domain derived from selection of conformational ensembles on the basis of experimental RDCs. The agreement between experimental and back-calculated RDCs is shown for the derived final ensemble (*blue*) and the starting pool of structures from which the ensemble was selected (*red*). Reprinted in part with permission from (Guerry et al. 2013). Copyright 2013 Wiley-VCH. **b** Dynamics of the two-domain splicing factor U2AF65 derived from RDCs and paramagnetic relaxation enhancements induced by *S*-(1-oxyl-2,2,5,5-tetramethyl-2,5-dihydro-1 H-pyrrol-3-yl) methyl methanesulfonylthioate (MTSL) spin labels attached at different positions in the two-domain protein. Ensembles of the two-domain protein are shown, where the grey surface represents the location of the domain RRM1, while the location of the second domain RRM2 is shown as spheres positioned at the centre of mass of RRM2. Ensembles are shown representing the initial pool of structures sampling all conformational space (*blue*) and the space occupied by RRM2 after refinement against experimental data (*red*). The agreement between experimental RDCs (*red*) and those back-calculated from the derived ensemble (*blue*) is also shown. Reprinted in part with permission from (Huang et al. 2014). Copyright 2014 American Chemical Society. **c** Conformational ensemble of the intrinsically disordered N-terminal transactivation domain of p53 in the context of the full-length p53-DNA complex (Wells et al. 2008). The ensemble was obtained on

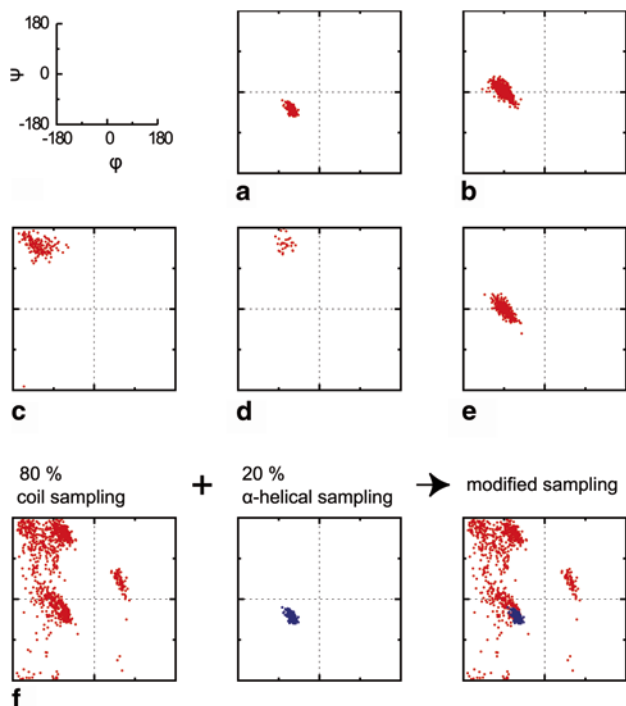


Fig. 4.2 Dihedral angle distributions characteristic of different secondary structure types. **a** Central residues in α -helices. **b** Last residues in α -helices that are C-capped with Schellman loops. **c** Residues in β -strands. **d** Residues in PPII conformations. **e** Residues in type I β -turns. **f** Modification of the dihedral angle sampling of a given residue can be achieved by combining the random coil distribution with an over-sampling of other regions of Ramachandran space (in this case the α -helical region). Dihedral angles for **(a)**, **(b)** and **(e)** were extracted from the database embedded in the structure motivator application (Leader and Milner-White 2012). Dihedral angles in **(c)** were extracted from parallel and anti-parallel β -strands from structures with the following PDB codes: 1MLD, 1QCZ, 2CMD, 1XH3, 1OGT and 3GP6. Dihedral angles in **(d)** were extracted from non-proline residues of peptide ligands bound to SH3 domains in a PPII conformation (1BBZ, 1CKA, 1CKB, 1SSH, 1W70, 2DRK, 2DRM, 2O88, 2O9V, 2W0Z, 2W10, 3EG1 and 315R)

Other structural motifs can also be identified by their specific ϕ/ψ angles. Apart from the α -helix and β -sheet, poly-L-proline II (PPII) is the only secondary structure that forms linear groups of residues that all adopt the same conformation (Fig. 4.2d) (Hollingsworth et al. 2009). This conformation is particularly interesting as it has been proposed to be significantly populated in IDPs and unfolded states of proteins (Shi et al. 2006; Schweitzer-Stenner 2012). Residues within β -turns also adopt spe-

the basis of experimental RDCs (local conformational sampling) and small angle X-ray scattering data (long-range behaviour). The agreement is shown between experimental RDCs (*red*) and those back-calculated from the model ensemble (*blue*) for both the isolated transactivation domain and in the context of the full-length p53/DNA complex. Reprinted in part with permission from (Wells et al. 2008). Copyright 2008 National Academy of Sciences, USA

cific dihedral angles and differ for each β -turn type (I, II, I', II') (Fig. 4.2e). Residues of both N-terminal and C-terminal helix capping motifs have unique dihedral angle distributions (Shen and Bax 2012), and it has been shown that within α -helices of structured proteins, the central residues display different distributions than the C-terminal residues (Fig. 4.2a, 4.2b) (Leader and Milner-White 2011).

Since each structural motif has distinct dihedral angle distributions, we can use them to describe the conformational energy surface of each residue within the disordered protein chain and more importantly also as a metric for the presence of residual secondary structure in IDPs. An increase in sampling of dihedral angles corresponding to the α -helical region will, if sampled at a high enough propensity, give rise to transiently populated α -helices, even in the absence of cooperative effects. IDPs can therefore in general be described as random coils (*i.e.* a peptide chain without specific secondary or tertiary structure), with deviations from this model corresponding to the presence of residual secondary structure (Fig. 4.2f). In order to map the dihedral angle distributions in an IDP we can exploit a number of different NMR parameters as described below.

3 NMR Parameters for Characterizing Local Conformational Propensities in IDPs

NMR is a powerful technique for studying IDPs at atomic resolution and provides many experimental parameters that inform us about local conformational propensities (Jensen et al. 2014). Chemical shifts are the most readily accessible parameters and as a single NMR resonance is usually observed for each nucleus in the spectra of IDPs, the chemical shifts report on the population-weighted average over all conformations sampled in solution up to the millisecond time scale. Chemical shifts are sensitive to the backbone dihedral angle distributions and can, therefore, be interpreted in terms of local conformational propensities. A simple analysis of chemical shifts in IDPs involves the calculation of secondary structure propensities (Marsh et al. 2006; Camilloni et al. 2012; Tamiola and Mulder 2012). This usually relies on characteristic shifts for α -helix, β -sheet and random coil derived from experimental chemical shifts of folded proteins with known three-dimensional structure or from a collection of assigned IDPs (Zhang et al. 2003; De Simone et al. 2009; Tamiola et al. 2010). When deriving conformational propensities it is important to correctly reference the experimental chemical shifts as systematic offsets may lead to erroneous estimates of the amount of secondary structure. It is possible to verify whether the chemical shift is correctly referenced using the secondary structure propensity (SSP) algorithm, which reports the potential reference offset based on the observation that C^α and C^β secondary chemical shifts are inversely correlated (Marsh et al. 2006).

Scalar couplings measured between nuclei of the protein backbone are also important structural probes in proteins and can be used to map dihedral angle distributions in IDPs. In the same way as chemical shifts, as long as the exchange rate is fast,

the scalar couplings represent a population-weighted average over all conformations sampled in solution. The dependence of scalar couplings on the main chain torsion angles can be described using a so-called Karplus relationship (Karplus 1959) that is generally parameterized against experimental scalar couplings measured in proteins of known structure (Smith et al. 1996). One of the commonly measured scalar couplings, the three-bond coupling constant ${}^3J_{\text{HNH}\alpha}$, depends on the backbone dihedral angle ϕ , allowing one to distinguish between α -helical (${}^3J_{\text{HNH}\alpha} < 5$ Hz) and β -sheet conformations (${}^3J_{\text{HNH}\alpha} > 8$ Hz) (Vuister and Bax 1993). Other scalar couplings such as ${}^3J_{\text{CaCa}}$, ${}^3J_{\text{NH}\alpha}$, ${}^3J_{\text{NC}\beta}$ and ${}^3J_{\text{NN}}$ report on the ψ angle and in principle provide a more accurate measure of PPII conformations (Graf et al. 2007; Hagarman et al. 2010).

Residual dipolar couplings (RDCs) are obtained by partially aligning the protein molecules in the magnetic field using, for example, a liquid crystal (Rückert and Otting 2000), filamentous phages (Hansen et al. 1998), polyacrylamide gels (Sass et al. 2000), or bicelles (Tjandra and Bax 1997). The inter-nuclear dipolar coupling, which is efficiently averaged to zero by the isotropic rotational tumbling of the molecules in solution, will no longer average to zero and a small part of the dipolar coupling will be measurable (Tolman et al. 1995; Tjandra and Bax 1997). RDCs report on bond vector orientations with respect to a common reference frame and have been used extensively for structure determination of folded proteins as reporters on the relative orientations of secondary structure elements (Prestegard et al. 2004; Blackledge 2005). Since the first measurement of RDCs in an unfolded protein (Shortle and Ackerman 2001) we have significantly advanced in our understanding and interpretation of RDCs in IDPs (Jensen et al. 2009). It is now clear that the RDCs carry contributions from the dihedral angle distribution of the amino acid of interest as well as its nearest neighbours, and the measurement of a single RDC value does therefore not provide a direct “read-out” of residue specific sampling in the same way as chemical shifts and scalar couplings (Huang et al. 2013). In addition, a contribution from the local flexibility of the chain (bulkiness) to the RDCs should be taken into account together with a length-dependent baseline that reflects the polymeric nature of the unfolded chain (Salmon et al. 2010; Huang et al. 2013). In the case of IDPs there is a preference for alignment media that rely on steric interactions between the protein and the medium such that the alignment tensor, and thereby the RDCs, can be predicted directly from the shape of each protein conformation (Zweckstetter and Bax 2000) and averaged over the ensemble.

4 Sample-and-Select Approaches

One way of obtaining representative ensemble descriptions of IDPs on the basis of experimental NMR data is to apply a two-step procedure involving the initial generation of a large pool of structures representing all of the conformational space available to the polypeptide chain (Fig. 4.3). Experimental data are then included in the second step where a set of structures (an ensemble) that agrees with the data is selected, for example using a genetic algorithm.

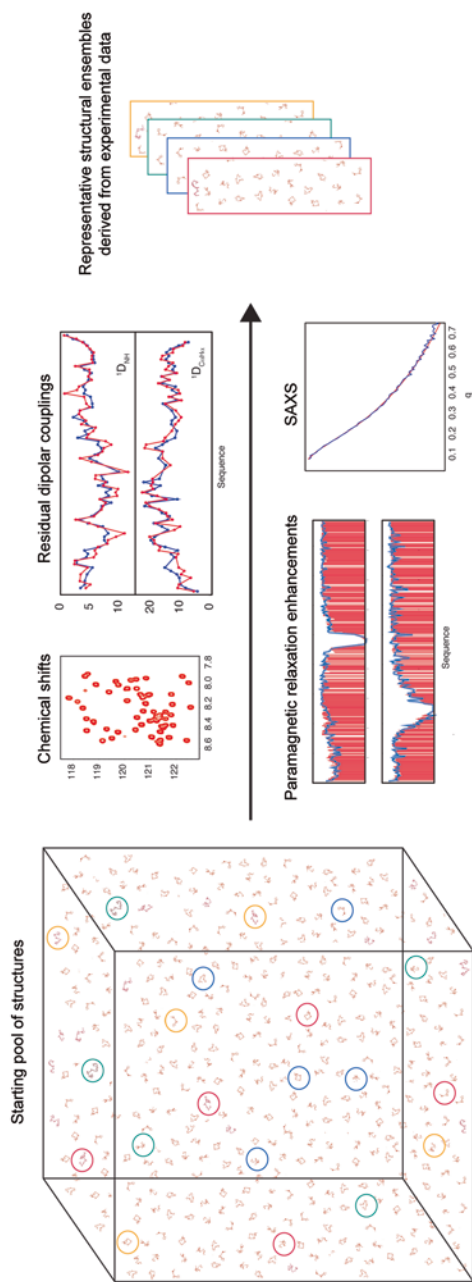


Fig. 4.3 Overview of sample-and-select approaches. Initially a large pool of structures is generated that represents the entire conformational space available to the protein under investigation. Experimental data such as chemical shifts, residual dipolar couplings, paramagnetic relaxation enhancements and small angle X-ray scattering (SAXS) are exploited in a second step to refine this conformational space by selecting sub-ensembles that agree with the experimental data. Reprinted in part with permission from (Jensen et al. 2014). Copyright 2014 American Chemical Society

Different approaches can be used to generate the initial pool of structures, but for a number of reasons it is important that the generated pool covers the entire conformational space of the molecule. Generally, a starting pool can be generated using molecular dynamics approaches or statistical coil generators.

5 Sampling Space Using Molecular Dynamics Simulations

For classical molecular dynamics (MD) simulations, sufficient sampling remains a problem when studying IDPs, even when the simulations are run over long time-scales of several hundreds of microseconds (Lindorff-Larsen et al. 2012). Other types of MD simulations address this problem and provide a better sampling; one example is replica exchange molecular dynamics (REMD), which artificially enhances the sampling by exchanging copies of the simulated protein that evolve under different conditions (Hansmann 1997; Sugita and Okamoto 1999). In its simplest form the protein is exchanged between two different temperature reservoirs where at higher temperatures the sampling rate is faster but not physical. When the protein evolves at lower temperatures it can get trapped in a local or global minimum with the end result being an inefficient sampling of the conformational space. Exchanging the protein copy to a reservoir with higher temperature facilitates the sampling of other minima because the energy barriers between them are easier to overcome.

Other approaches include enhanced sampling techniques such as metadynamics, in which a term is added to the force field that penalizes the conformations that have already been explored by the molecule (Leone et al. 2010). The energy penalties accumulate as the protein explores an energy minimum and after some time the protein is forced to explore other minima. Metadynamics can be combined with REMD to enhance the sampling rate even further (Piana and Laio 2007). Accelerated molecular dynamics (AMD) is another approach for enhancing sampling that can be used for IDPs. In this method the free energy surface is modulated by a scaling factor that affects the energy barriers between minima and therefore increases the chance of barrier crossing (Voter 1997; Hamelberg et al. 2004; Pierce et al. 2012).

6 Sampling Space Using Statistical Coil Generators

A starting pool of conformers that is subsequently used in selections can also be produced with a statistical coil generator (Feldman and Hogue 2000; Jha et al. 2005a; Bernadó et al. 2005b; Ozenne et al. 2012a). In this approach a protein molecule is built starting from either end of the chain by adding amino acid after amino acid with a ϕ/ψ angle that is randomly chosen from a database of dihedral

angles (the statistical coil library). Each newly added amino acid is checked for steric clashes between the backbone atoms and between simplified representations of side chains. In case of steric clash the newly placed residue is rejected and rebuilt until a suitable conformation is found. Force field bond and angle potentials are not included during the generation of conformers, and the steric clash model is very simple and only defines a certain radius of exclusion for each atom. This approach allows the conformational space to be sampled roughly but efficiently and generates a pool of many different combinations of ϕ/ψ angles for consecutive amino acids.

Statistical coil libraries, which are used for generating the structures, are assembled with the help of databases of high-resolution crystal structures (Serrano 1995; Jha et al. 2005b). The conformational preferences of amino acids in folded proteins differ from those of disordered proteins as most of the residues in folded proteins reside within secondary structure elements, while IDPs are expected to more closely resemble the loop regions. If the α -helices, β -sheets and β -turns are removed from the initial data set of high-resolution crystal structures, only motifs from non-regular loops remain in the database. These loop residues are not restrained by secondary structure hydrogen bonding criteria, unlike for example α -helices, and when a large number of the loop residues are taken into consideration, the potential contributions from the long-range tertiary contacts mostly average out. If we extract the ϕ/ψ angle distributions from the database of loop regions, we obtain a library of amino acid specific distributions of ϕ/ψ angles. These ϕ/ψ angle distributions represent a valid starting point for describing the conformational free energy surface of amino acids within IDPs and can be used in conjunction with statistical coil generators for building ensembles of IDPs. One of these statistical coil generators, Flexible-Meccano, is freely available² and is provided with a graphical interface that allows the testing of different sampling regimes by manually modifying the ϕ/ψ sampling of selected amino acids (Ozenne et al. 2012a). Flexible-Meccano calculates NMR observables such as chemical shifts, RDCs, scalar couplings and paramagnetic relaxation enhancements (PREs) from the generated ensembles that allow direct comparison with experimental data (Fig. 4.4).

The statistical coil libraries still have room for improvement in terms of the inclusion of neighbour residue effects, which would be analogous to what has been carried out for random coil chemical shift tabulations (Wishart et al. 1995; Wang and Jardetzky 2002b; Wang and Jardetzky 2002a; De Simone et al. 2009; Tamiola et al. 2010). In fact the neighbour residue correction is often used in the statistical coil libraries for pre-proline residues, because the neighbour effect of prolines on the preceding residue is particularly pronounced. This is due to steric hindrance between the $^{\delta}\text{CH}_2$ side chain group of the proline and the NH and $\text{C}^{\beta}\text{H}_2$ atoms of the preceding residues (MacArthur and Thornton 1991). Again, similarly to random coil chemical shifts, the statistical coil databases could be improved by refining them with the help of experimental data from IDPs themselves (Tamiola et al. 2010).

² www.ibs.fr/science-213/scientific-output/software.

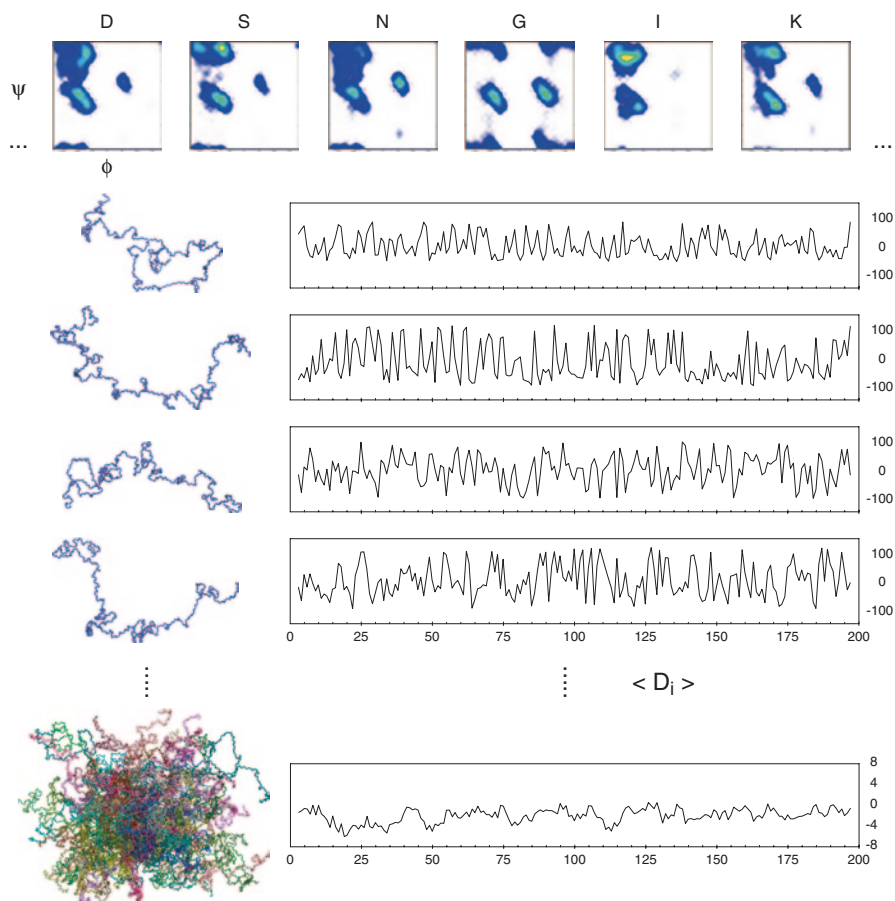


Fig. 4.4 Generation of conformational ensembles of IDPs using a statistical coil generator. Ramachandran plots (ϕ/ψ distributions) are shown for the amino acids D, N, S, G, I and K as derived from loop regions of high-resolution crystal structures. These distributions are used to construct conformations of the protein for a given primary sequence by starting from either the C- or N-terminal end of the protein and building amino acid after amino acid according to randomly chosen ϕ/ψ pairs of the statistical coil library. For each copy of the molecule, experimental NMR data can be calculated and the ensemble-average over multiple copies of the protein can be compared to experimental data. Reprinted in part with permission from (Jensen et al. 2014). Copyright 2014 American Chemical Society

7 Selection of Ensembles on the Basis of Experimental NMR Data

Once the initial pool of structures has been generated, NMR parameters such as scalar couplings, chemical shifts and RDCs can be calculated for each member of the pool. The selection of sub-ensembles proceeds by calculating the averages

of the NMR parameters over a given sub-ensemble and comparing them to experimental data. Different approaches have been proposed in the literature for deriving representative ensembles such as ENSEMBLE, which assigns weights to the different conformations of the pool (Marsh and Forman-Kay 2009; Krzeminski et al. 2013), and ASTEROIDS, which relies on a genetic algorithm to select sub-ensembles (Nodet et al. 2009; Jensen et al. 2010), as well as ensemble optimization on the basis of Bayesian weighting (Fisher et al. 2010; Fisher and Stultz 2011). It is important to note that in cases where the IDPs possess transiently populated secondary structures, it is not possible to select an ensemble that matches all the experimental data directly from a pool of statistical coil conformers. The reason for this is that the probability of finding continuous stretches of secondary structure is too low. Therefore, the sample-and-select protocol is often repeated multiple times in an iterative procedure, where the sampling pool is regenerated using the information (local conformational sampling) obtained from ensembles selected in the previous iteration. In this way, the sampling pool is enriched at each step with conformational preferences characteristic of the protein under investigation.

8 Ensemble Representations of the IDP Tau from Chemical Shifts and RDCs

The combination of the statistical coil generator Flexible-Meccano (Bernadó et al. 2005b; Ozenne et al. 2012a) and the ensemble selection algorithm ASTEROIDS has allowed quantitative insight into residue-specific conformational sampling in a number of IDPs involved in neurodegenerative diseases (Bernadó et al. 2005a; Mukrasch et al. 2007; Schwalbe et al. 2014). The protein Tau is a 441 amino acid protein that is intrinsically disordered and undergoes a conformational transition to a pathological form of the same protein. The NMR spectra of Tau have been fully assigned (Narayanan et al. 2010), allowing insight into the conformational preferences of this protein at atomic resolution. A complete set of chemical shifts and $^1D_{NH}$ RDCs were obtained for the protein Tau in order to accurately map α -helical, β -strand and PPII populations. Figure 4.5a shows the agreement between experimental data and those back-calculated from selected ASTEROIDS ensembles. The ensemble selections were repeated five times and the conformational sampling of each residue along the sequence of Tau is conveniently represented by their dihedral angle distributions (Fig. 4.5b).

In general, we can learn a lot from these ensembles as they provide quantitative insight into the sampling in different regions of Ramachandran space. Specifically, it is seen that the aggregation nucleation sites in Tau overpopulate the PPII region, suggesting that these conformations represent precursors of aggregation (Fig. 4.5b) (Schwalbe et al. 2014). In addition to these observations, the presence of turn-like motifs can be identified in each of the Tau repeat regions (R1-R4). These turn motifs were also studied in detail previously using AMD simulations of small peptides

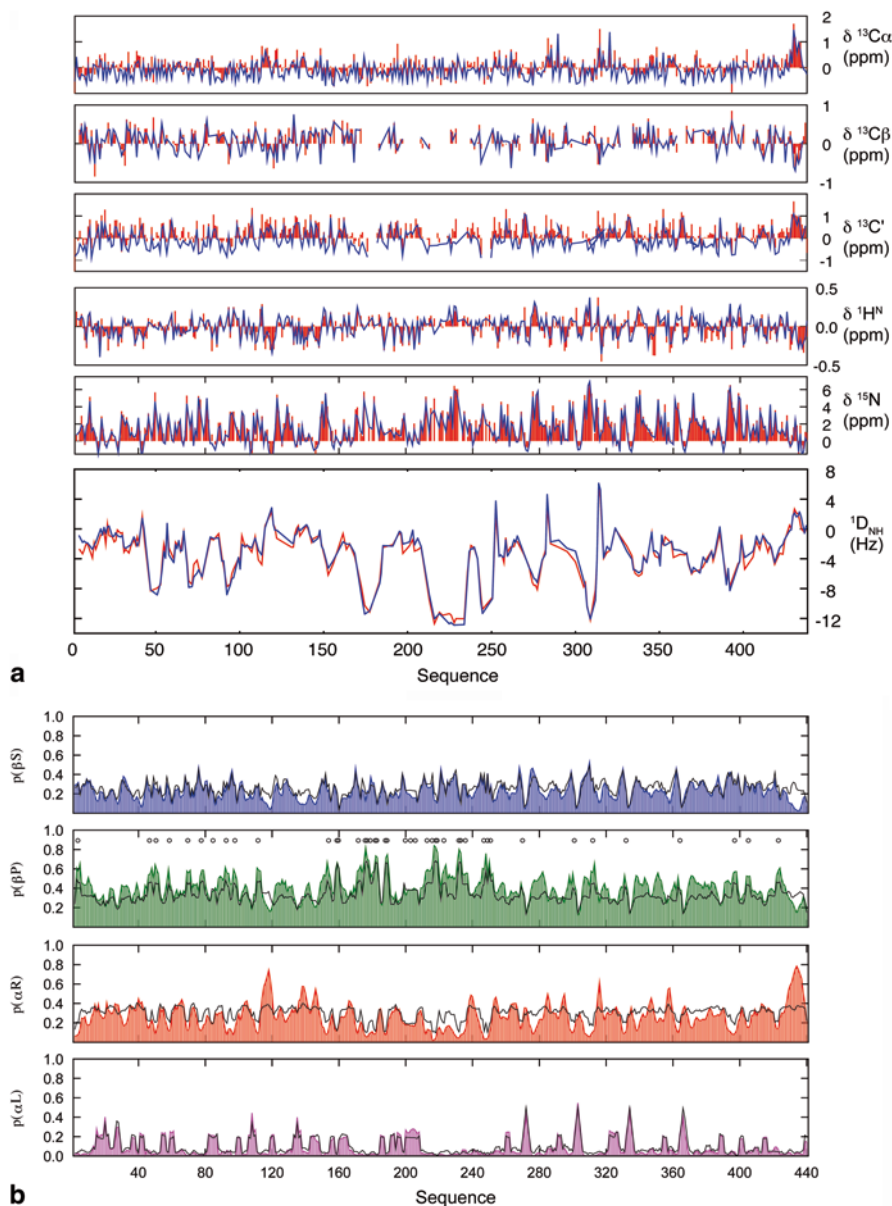


Fig. 4.5 Ensemble representations of the intrinsically disordered Tau protein on the basis of chemical shifts and RDCs. **a** Agreement between experimental (*red*) and back-calculated secondary chemical shifts and RDCs (*blue*) from selected ASTEROIDS ensembles of Tau. **b** Site specific conformational sampling in Tau derived from the selected ensembles (*blue*, *green*, *red*,

of Tau, where it was shown that the AMD derived ϕ/ψ sampling corresponded to type I β -turns (Mukrasch et al. 2007). When this AMD sampling was incorporated into a model ensemble of the smaller K18 construct of Tau, the agreement between experimental and back-calculated $^1D_{NH}$ RDCs improved significantly, proving that these regions indeed adopt type I β -turns as predicted by AMD.

9 The Reference Ensemble Method

The study described above combines different data types to map local conformational sampling. The accuracy with which conformational propensities can be determined depends on the amount of experimental data available for a given system. Assuming that we want to map the population of α -helix, β -strand and PPII conformations for each amino acid of the protein, it would be useful to determine a minimum dataset that would allow this. The α -helical and β -strand propensities can be well characterized with the help of carbon (C^α , C^β , C') chemical shifts, but a residue sampling a statistical coil distribution and a residue sampling exclusively PPII specific dihedral angles have approximately the same carbon chemical shift (Fig. 4.6a). We therefore cannot use carbon chemical shifts to distinguish between the two mentioned sampling regimes. Similarly, most of the RDC types display degeneracy between β -strand and PPII conformations. The $^1D_{NH}$ RDCs are negative for both increased β -strand propensities and increased PPII propensities (Fig. 4.6b).

Selection against synthetic data from a reference ensemble can help reveal such degeneracies and determine the minimum dataset necessary for accurate mapping of the conformational energy landscape. In the reference ensemble approach, an ensemble of structures is generated using either an MD simulation or a statistical coil generator. These structures constitute the target ensemble for which a synthetic dataset is calculated. If our ensemble selection protocol is working without bias and we have sufficient and complementary data types, we should be able to regenerate the local conformational sampling preferences by targeting the synthetic dataset using the sample-and-select approach.

A study by Ozenne et al. demonstrated how useful this approach can be when applied to IDPs (Ozenne et al. 2012b). Initially, an ensemble of a model protein of 60 amino acids of arbitrary sequence was obtained using the statistical coil generator Flexible-Meccano, where three distinct regions of the protein over-sampled the α -helical, β -strand and PPII region of Ramachandran space (50% additional sampling in each region compared to the statistical coil). Different types of ensemble-averaged chemical shifts and RDCs were calculated for this ensemble

magenta) compared to standard statistical coil distributions (*black*). Populations are reported for four different regions of Ramachandran space corresponding to right- (αR , *red*) and left-handed α -helix (αL , *magenta*), β -strand (βS , *blue*) and PPII conformations (βP , *green*). Circles indicate the presence of proline residues. Reprinted in part with permission from (Schwalbe et al. 2014). Copyright 2014 Elsevier

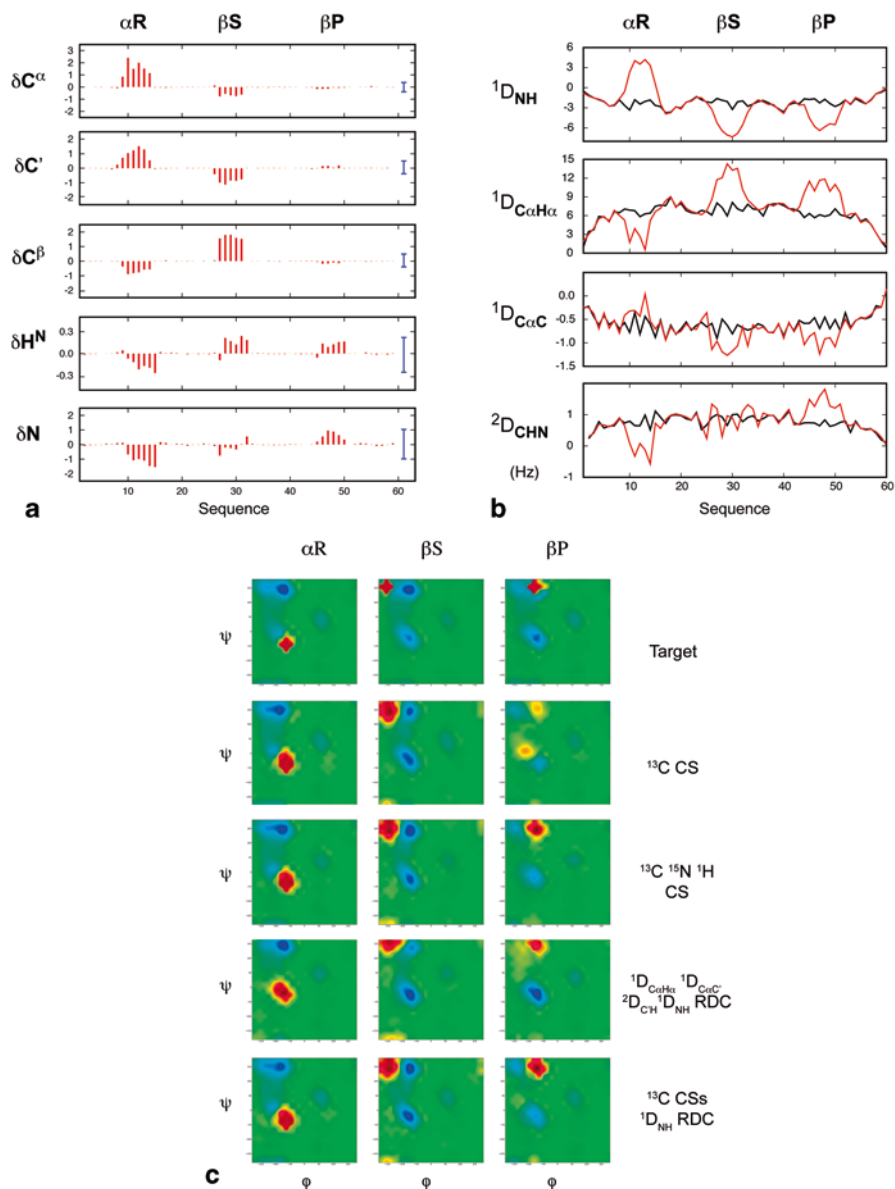


Fig. 4.6 Testing the accuracy with which experimental data can map local conformational propensities in IDPs using the reference ensemble method. **a** Synthetic chemical shift dataset calculated for an ensemble of a model protein of 60 amino acids of arbitrary sequence. Three different regions of the protein over-sample the α -helical, β -strand and PPII region, while the remaining regions sample statistical coil conformations. The difference is shown between the predicted chemical shifts for this ensemble and the ensemble-averaged chemical shifts for a statistical coil ensemble. **b** Synthetic RDC dataset calculated for the model protein over-sampling the three different regions of Ramachandran space (*red*) compared to RDCs predicted for a statistical coil ensemble of the same protein (*black*). **c** Selection of sub-ensembles using ASTEROIDS on the basis of different

and used as targets in a selection protocol using the genetic algorithm ASTEROIDS by starting from a statistical coil pool, i.e. a pool without any particular secondary structure preferences. After the ensemble selection using only data for carbon chemical shifts (C^α , C^β , C'), the ϕ/ψ sampling was reproduced in the regions with enhanced α -helical and β -strand sampling, but not in the region with enhanced PPII sampling (Fig. 4.6c). The study further showed that inclusion of either backbone ^{15}N and $^1\text{H}^{\text{N}}$ chemical shifts or $^1\text{D}_{\text{NH}}$ RDCs in the selection procedure allowed a reproduction of the PPII sampling in the third biased region of the target ensemble, while inclusion of both backbone chemical shifts and RDCs represents a robust and accurate way to map the local conformational sampling of IDPs (Fig. 4.6c). Calibration of ensemble generation protocols against a synthetic target can therefore tell us if we are able to reproduce the sampling of a synthetic ensemble, and consequently also a real ensemble with the same characteristics. The reference ensemble method can also be used in a quantitative way by adding Gaussian noise to the synthetic dataset to determine the accuracy with which the conformational space of IDPs can be mapped using different data types (Ozenne et al. 2012b).

10 Taking into Account Cooperatively Formed Secondary Structures in IDPs

When an IDP contains a longer stretch of a cooperatively formed structure, such as an α -helix, the sample-and-select approach does not work as efficiently. An α -helix can be stabilized by many cooperative interactions (Muñoz and Serrano 1995; Doig 2002). For example, the effect of helix capping can span several amino acids further down the protein sequence and can affect the stability of the helix as a whole. Apart from capping interactions and the regular backbone-backbone i to $i+3$ hydrogen bonding pattern, many other stabilizing interactions are present between i and $i+3$ residues and between residues even further away. As helices in IDPs can span more than ten residues, we expect that amino acids that are far apart in the primary sequence should contribute together to the formation of the helix.

Statistical coil generators take into account amino acid type conformational preferences that are mainly local. As a consequence the sampling in the statistical coil library can correctly sample α -helical conformations in selected regions of the protein; however, the chance of building a long helix without an interruption is relatively small. For example, with a statistical coil library with 80% helical sampling, the probability of forming a helical element consisting of six consecutive amino acids is 0.8^6 , which is around 25%. The probability of forming a longer α -helix with

combinations of the synthetic chemical shift and RDC datasets. Ramachandran plots of the target (*top line*) and the results of the selections employing different data types are shown. Reprinted in part with permission from (Ozenne et al. 2012b). Copyright 2012 American Chemical Society

a high enough population to fit the data is therefore very low. Approaches using MD simulations experience a similar problem when it comes to long cooperatively folded helices (or other secondary structures), and breaks in helices are often observed throughout the simulations.

A solution to this problem is to generate many different starting ensembles where each ensemble incorporates an α -helix with a different start and end point, calculate the ensemble-averaged NMR data for each of these ensembles, and subsequently find the best combination of ensembles with corresponding populations that agree with the experimental data. Essentially this corresponds to enriching the initial starting pool with cooperatively formed α -helices in specific regions of the protein that are known to over-sample the α -helical region of Ramachandran space.

This approach was developed and applied to the C-terminal intrinsically disordered domain, N_{TAIL} , of Sendai virus nucleoprotein, which undergoes induced α -helical folding of its molecular recognition element upon binding to its partner protein PX (Jensen et al. 2008). The $^1D_{\text{NH}}$ RDCs measured in N_{TAIL} were positive within the molecular recognition element and showed a characteristic dipolar wave pattern consistent with the formation of cooperatively formed α -helices (Fig. 4.7a) (Jensen and Blackledge 2008). The experimental RDCs were fitted with models of increasing complexity, i.e. starting from a statistical coil model and increasing the number of helical ensembles until a satisfactory fit was obtained (Fig. 4.7a). For each model the populations of the helical elements were optimized to best agree with the experimental data. Data reproduction evidently improves as the number of helical ensembles increases, and a standard F-test was therefore used to test for the statistical significance of this improvement. It was found that three helical ensembles with different populations in exchange with a disordered form of the protein are needed to describe the experimental RDCs (Fig. 4.7b). Interestingly, all the selected helical ensembles are preceded by aspartic acids or serines, which are the most common N-capping residues in helices of folded proteins (Fig. 4.7c, 4.7d). An N-capping residue stabilizes a helix by forming a hydrogen bond between its side chain and the backbone amides at position 2 or 3 in the helix (Fig. 4.7c). Importantly, this indicates that the helices preferentially being populated in solution in N_{TAIL} are stabilized by N-capping interactions, and that the helical formation is being promoted by strategically placed aspartic acids and serines in the primary sequence. The partial pre-structuration of N_{TAIL} in its free state suggests that the interaction with PX occurs through conformational selection, where one of the helices is selected by the partner protein in order to form the complex (Hammes et al. 2009).

11 Choosing an Appropriate Ensemble Size

A scoring function that measures the agreement between the experimental and simulated data for the model ensemble is applied during the ensemble selection procedure. A measure commonly used is chi square ($\chi^2 = \sum (s_i - m_i)^2 / \sigma_i$) where s_i represents the back-calculated data from the ensemble, m_i represents the

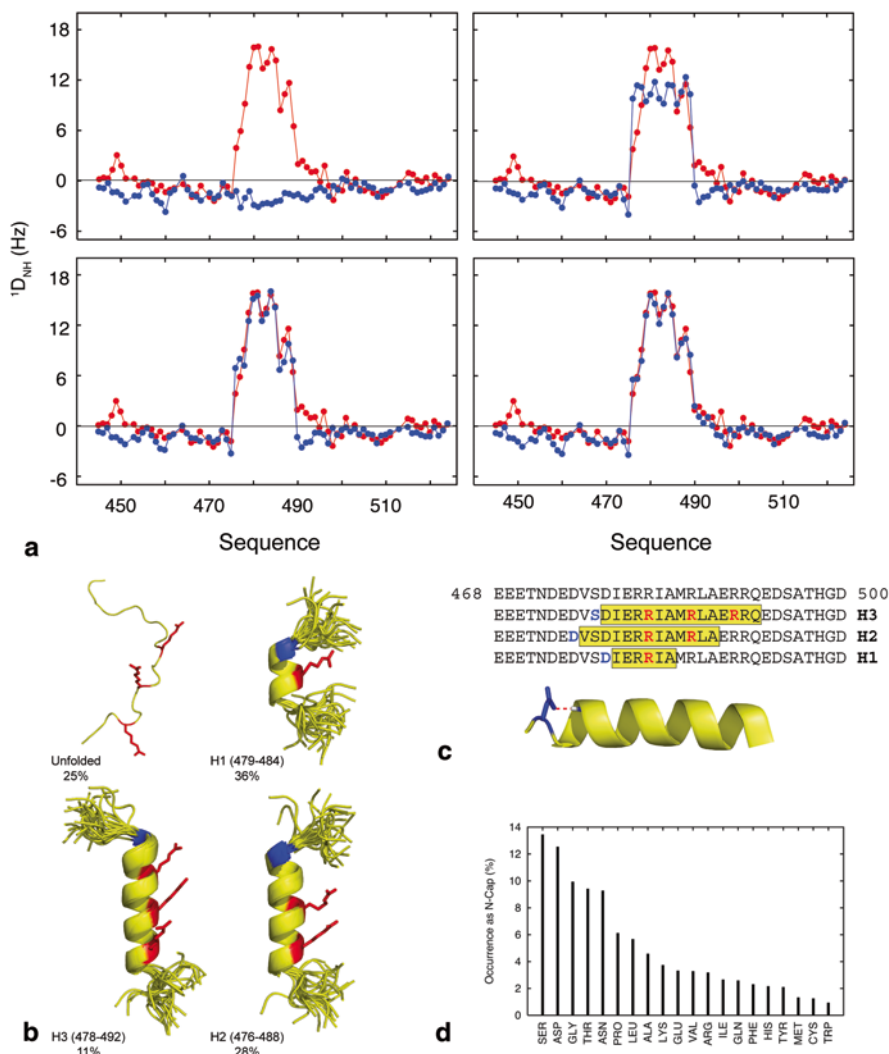


Fig. 4.7 Analysis of cooperatively formed α -helices within the molecular recognition element of the C-terminal domain, N_{TAIL} , of the Sendai virus nucleoprotein. **a** Reproduction of experimental $^1D_{NH}$ RDCs in N_{TAIL} for models with an increasing number, N , of helical ensembles: $N=0$ (top, left), $N=1$ (top, right), $N=2$ (bottom, left), $N=3$ (bottom, right). Experimental RDCs are shown in red, while back-calculated RDCs from the different models are shown in blue. **b** Molecular representation of the equilibrium of the molecular recognition element of N_{TAIL} in solution. The four different helical states are presented as a single structure for the completely disordered form and as twenty randomly selected conformers for the three helical states. The molecular recognition arginines are displayed in red, while N-capping residues are shown in blue. **c** The amino acid sequence of the molecular recognition element of N_{TAIL} showing that the selected helical elements are all preceded by aspartic acids or serine residues. The cartoon representation illustrates an N-capping aspartic acid side chain-backbone interaction. **d** The occurrence of different amino acid types as N-capping residues in helices of folded proteins. Reprinted in part with permission from (Jensen et al. 2008). Copyright 2008 American Chemical Society

measured data, and σ_i is the experimental error associated with the different NMR parameters. One has to take care in order not to over-fit the experimental data. Over-fitting happens when the difference between the simulated and experimental data is minimized during the fitting process not in order to improve the physical model describing the system, but because the model is modified to fit the random error and noise contributions. A good fit therefore always means a good fit within the defined experimental error.

Ensemble size also influences the goodness of the fit and the ensemble should not be too small or too large. The ensemble obtained in the selection procedure is not accurate if it is composed of too few structures and therefore does not represent the conformational heterogeneity present in solution. In this case, with too few conformers in the ensemble, we say that we are over-restraining or also under-fitting. On the other hand, as we increase the ensemble size, the number of parameters (e.g. dihedral angles) that can be independently adjusted increases and the total χ^2 value will therefore decrease. The fit may improve because of an improvement in our model, but also because inaccuracies in the model are compensated by newly added structures.

There are tests that can help us decide on the ensemble size that we should choose for ensemble selection. Most commonly a plot of final χ^2 against ensemble size is used to determine the appropriate size for a given set of experimental data. The fit does not improve significantly above a certain ensemble size, and the increase in the number of degrees of freedom introduced by selecting a larger ensemble is no longer justifiable.

An alternative method, and in principle a more correct one, is to use cross-validation procedures where a part of the experimental data is left out of the ensemble selection procedure. Ensembles of different sizes are selected and the “passive” data are back-calculated from the selected ensembles and compared to the experimental data. The optimal reproduction of the passive data will normally occur for the most appropriate ensemble size. This procedure has for example been used to obtain the most appropriate ensemble size (200 structures) for describing the local conformational sampling of urea-denatured ubiquitin on the basis of multiple types of RDCs (Nodet et al. 2009).

12 Ensemble Size in Relation to Convergence Properties of NMR Parameters

When optimizing the size of the selected ensembles, one also needs to consider the convergence characteristics of the different NMR parameters when averaged over the sub-ensembles. We say that convergence of a parameter has been reached when the addition of one more conformer to the ensemble does not perturb the calculated average parameter within a predefined limit. The convergence of parameters is particularly important as the use of too few structures in the selected ensembles will force the fitting procedure to accommodate fluctuations in the averaged NMR parameters that do not necessarily correspond to specific conformational propensities, thereby potentially leading to incorrect residue-specific conformational sampling.

The number of conformers needed for a certain simulated parameter to converge depends on its variance. This is the reason for the different convergence properties of RDCs and chemical shifts. Chemical shifts are sensitive to the local chemical environment and are affected by main and side chain dihedral angles, amino acid identity, ring current effects and hydrogen bonding. When chemical shifts are predicted in IDPs the most important factor is the dihedral angle distribution. Carbon (C^α , C^β , C') and proton H^α chemical shifts depend mostly on the ϕ/ψ angles of the residue of interest, while the chemical shifts of the nitrogen (N) atom and the amide proton (H^N) depend mostly on the ψ angle of the preceding residue. The fact that chemical shifts can be predicted from local structure only makes them a well-behaved parameter when it comes to convergence. Sufficient sampling of the ϕ/ψ space of a single amino acid can even be achieved with only a few hundred structures. As a consequence, when selecting ensembles against experimental chemical shifts, 100–200 structures are sufficient for achieving convergence of the predicted chemical shifts (Fig. 4.8a, 4.8b, 4.8c). Scalar couplings also report on local conformational features of the polypeptide chain, and similarly to chemical shifts, a hundred conformers in the model ensemble suffice for achieving convergence.

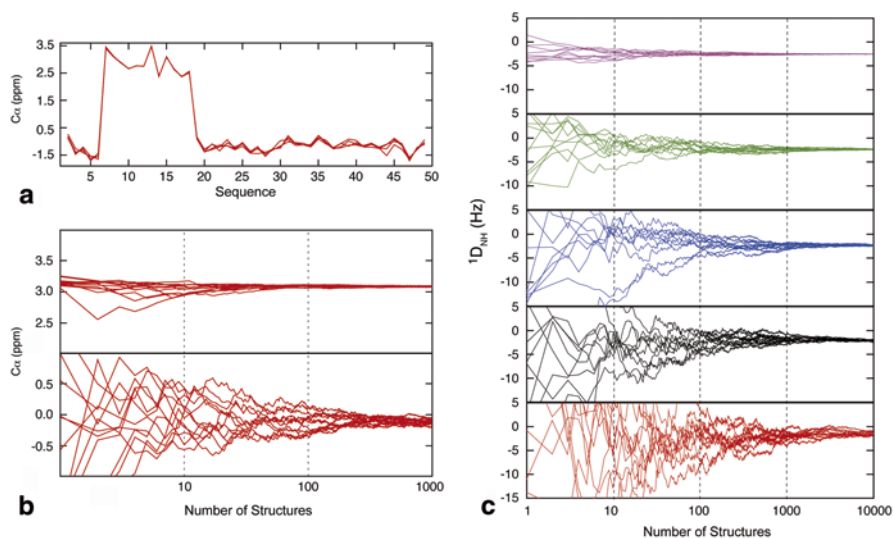


Fig. 4.8 Convergence of experimental NMR parameters over structural ensembles. **a** Secondary C^α chemical shifts averaged over 250 conformers generated using Flexible-Meccano for a model protein of 50 amino acids of arbitrary sequence. The results for five different ensemble averages are shown. Residues 7–18 populate the α -helical region of Ramachandran space, while the remaining residues adopt random coil conformations. **b** Ensemble-averaged secondary C^α chemical shifts for increasing ensemble size for residue 15 of the model protein. **c** Convergence of $^1D_{NH}$ RDCs over a structural ensemble with an increasing number of conformers for a model protein of 76 amino acids. Results are shown for the calculation using a global alignment tensor (*red*) and employing different sizes of short segments for calculating the alignment tensor: 25 (*black*), 15 (*blue*), 9 (*green*) and 3 (*pink*) amino acids. Reprinted in part with permission from (Nodet et al. 2009). Copyright 2009 American Chemical Society

As mentioned above, RDCs depend on both local and long-range structure, i.e. on the conformational sampling of the residue itself and immediate neighbours as well as intra-peptide long-range contacts. The large number of combinations of dihedral angle pairs that potentially all give different RDC values combined with the large range of RDCs calculated from a single structure make the convergence of the RDC average much slower. In addition, RDCs converge more slowly for longer polypeptide chains and for an IDP of 100 amino acids, more than 10,000 structures are needed in order to achieve convergence of the RDC average (Fig. 4.8d). In order to overcome this problem, we can divide the protein chain into shorter, uncoupled segments and predict the RDCs for the central amino acid of each segment (Marsh et al. 2008), thereby achieving sufficient convergence of the RDC average with only a few hundred structures (Fig. 4.8d). The disadvantage of this approach is that we remove any information about long-range structure from the predicted RDCs; however, this information can be reintroduced by multiplying the predicted RDCs by a baseline that takes into account the chain-like nature of the IDP (Nodet et al. 2009; Salmon et al. 2010). Our ability to separate the contribution to the RDCs from local conformational sampling and long-range interactions allows convergence of the RDC average with an ensemble of only a few hundred structures. The use of short segments for the calculation of RDCs therefore appears essential when using RDCs in ensemble selection procedures.

13 Validation of Ensemble Descriptions

Due to the under-determined nature of ensemble selections in general, it is useful to think about how we can potentially validate the structural ensembles that we derive from experimental NMR data. One way of doing this is to exploit the complementary nature of different data types and use cross-validation procedures where a part of the experimental data is left out of the ensemble selection and subsequently back-calculated from the selected ensembles. If the selected ensemble correctly reproduces the local conformational sampling, the agreement between the “passive” data and that back-calculated from the selected ensemble should be good and no systematic deviations should be observed. An example of this procedure is shown in Fig. 4.9 where experimental $^1D_{\text{NH}}$ RDCs measured in Tau protein are compared to the RDCs extracted from ASTEROIDS ensembles of Tau selected on the basis of chemical shift data alone. The agreement between the two sets of data is excellent and even the turn motifs in the repeat regions of Tau—where positive $^1D_{\text{NH}}$ RDCs are observed experimentally—are reproduced by the chemical shift ensemble. This type of procedure therefore validates the local conformational sampling of Tau derived from chemical shifts only.

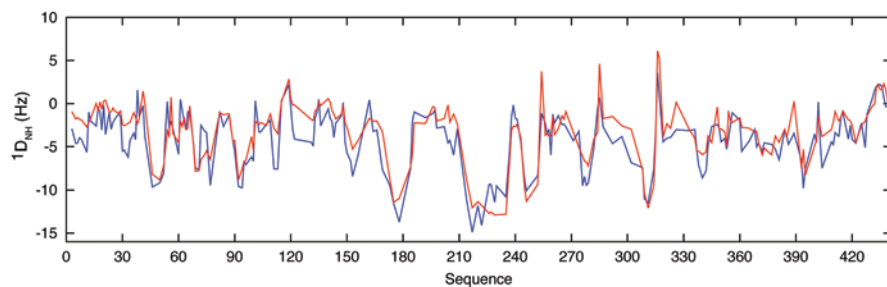


Fig. 4.9 Validation of structural ensembles of IDPs derived from experimental NMR data. An example is shown of cross-validation of experimental $^1D_{\text{NH}}$ RDCs of the IDP Tau. Experimental data are shown in red, while back-calculated RDCs from an ensemble of Tau selected by the genetic algorithm ASTEROIDS on the basis of experimental chemical shifts only are shown in blue

14 Conclusions and Outlook

Ensemble descriptions have in recent years emerged as the preferred tool for representing the structural and dynamic properties of IDPs and their functional complexes. Within such descriptions it is assumed that the protein adopts a continuum of rapidly interconverting structures, and the determination of these representative ensembles is one of the major challenges in the studies of IDPs. In this chapter we have described how different NMR data types can be combined with sample-and-select approaches to map local conformational propensities in IDPs. In particular, we have emphasized some of the pitfalls associated with these approaches such as under- and over-restraining, and we have discussed ways to validate the derived structural ensembles. Validating structural ensembles is particularly important if we are to use these ensembles in the future for the prediction of other, independent experimental observables or for the development of small molecules that can interfere with the biological function of IDPs.

References

- Bernadó P, Blackledge M (2010) Structural biology: proteins in dynamic equilibrium. *Nature* 468:1046–1048
- Bernadó P, Bertocini CW, Griesinger C et al (2005a) Defining long-range order and local disorder in native α -synuclein using residual dipolar couplings. *J Am Chem Soc* 127:17968–17969
- Bernadó P, Blanchard L, Timmins P et al (2005b) A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proc Natl Acad Sci U S A* 102:17002–17007
- Bertini I, Gupta YK, Luchinat C et al (2007) Paramagnetism-based NMR restraints provide maximum allowed probabilities for the different conformations of partially independent protein domains. *J Am Chem Soc* 129:12786–12794

- Blackledge M (2005) Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog Nucl Magn Reson Spectrosc* 46:23–61
- Bouvignies G, Bernadó P, Meier S et al (2005) Identification of slow correlated motions in proteins using residual dipolar and hydrogen-bond scalar couplings. *Proc Natl Acad Sci U S A* 102:13885–13890
- Camilloni C, De Simone A, Vranken WF et al (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. *Biochemistry* 51:2224–2231
- De Simone A, Cavalli A, Hsu S-TD et al (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. *J Am Chem Soc* 131:16332–16333
- Deshmukh L, Schwieters CD, Grishaev A et al (2013) Structure and dynamics of full-length HIV-1 capsid protein in solution. *J Am Chem Soc* 135:16133–16147
- Doig AJ (2002) Recent advances in helix-coil theory. *Biophys Chem* 101–102:281–293
- Dunker AK, Silman I, Uversky VN et al (2008) Function and structure of inherently disordered proteins. *Curr Opin Struct Biol* 18:756–764
- Dyson HJ, Wright PE (2002) Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 12:54–60
- Feldman HJ, Hogue CW (2000) A fast method to sample real protein conformational space. *Proteins* 39:112–131
- Fisher CK, Stultz CM (2011) Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol* 21:426–431
- Fisher CK, Huang A, Stultz CM (2010) Modeling intrinsically disordered proteins with bayesian statistics. *J Am Chem Soc* 132:14919–14927
- Francis DM, Rózycki B, Koveal D et al (2011) Structural basis of p38a regulation by hematopoietic tyrosine phosphatase. *Nat Chem Biol* 7:916–924
- Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. *Proteins* 23:566–579
- Göbl C, Madl T, Simon B et al (2014) NMR approaches for structural analysis of multidomain proteins and complexes in solution. *Prog Nucl Magn Reson Spectrosc* 80:26–63
- Graf J, Nguyen PH, Stock G, Schwalbe H (2007) Structure and dynamics of the homologous series of alanine peptides: a joint molecular dynamics/NMR study. *J Am Chem Soc* 129:1179–1189
- Guerry P, Salmon L, Mollica L et al (2013) Mapping the population of protein conformational energy sub-states from NMR dipolar couplings. *Angew Chem Int Ed Engl* 52:3181–3185
- Hagarman A, Measey TJ, Mathieu D et al (2010) Intrinsic propensities of amino acid residues in GxG peptides inferred from amide I' band profiles and NMR scalar coupling constants. *J Am Chem Soc* 132:540–551
- Hamelberg D, Mongan J, McCammon JA (2004) Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys* 120:11919–11929
- Hammes GG, Chang Y-C, Oas TG (2009) Conformational selection or induced fit: a flux description of reaction mechanism. *Proc Natl Acad Sci U S A* 106:13737–13741
- Hansen MR, Mueller L, Pardi A (1998) Tunable alignment of macromolecules by filamentous phage yields dipolar coupling interactions. *Nat Struct Biol* 5:1065–1074
- Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* 281:140–150
- Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. *Nature* 450:964–972
- Hollingsworth SA, Berkholz DS, Karplus PA (2009) On the occurrence of linear groups in proteins. *Protein Sci* 18:1321–1325
- Huang J, Ozenne V, Jensen MR et al (2013) Direct prediction of NMR residual dipolar couplings from the primary sequence of unfolded proteins. *Angew Chem Int Ed Engl* 52:687–690
- Huang J, Warner LR, Sanchez C et al (2014) Transient electrostatic interactions dominate the conformational equilibrium sampled by multidomain splicing factor U2AF65: a combined NMR and SAXS study. *J Am Chem Soc* 136:7068–7076

- Jensen MR, Blackledge M (2008) On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. *J Am Chem Soc* 130:11266–11267
- Jensen MR, Houben K, Lescop E et al (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. *J Am Chem Soc* 130:8055–8061
- Jensen MR, Markwick PRL, Meier S et al (2009) Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. *Structure* 17:1169–1185
- Jensen MR, Salmon L, Nodet G et al (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc* 132:1270–1272
- Jensen MR, Zweckstetter M, Huang J-R et al (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. *Chem Rev* 114:6632–6660
- Jha AK, Colubri A, Freed KF et al (2005a) Statistical coil model of the unfolded state: resolving the reconciliation problem. *Proc Natl Acad Sci U S A* 102:13099–13104
- Jha AK, Colubri A, Zaman MH et al (2005b) Helix, sheet, and polyproline II frequencies and strong nearest neighbor effects in a restricted coil library. *Biochemistry* 44:9691–9702
- Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
- Karplus M (1959) Contact electron-Spin coupling of nuclear magnetic moments. *J Chem Phys* 30:11–15. doi:10.1063/1.1729860
- Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. *Proc Natl Acad Sci U S A* 102:6679–6685
- Krzeminski M, Marsh JA, Neale C et al (2013) Characterization of disordered proteins with ENSEMBLE. *Bioinformatics* 29:398–399
- Lange OF, Lakomek N-A, Farès C et al (2008) Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* 320:1471–1475
- Leader DP, Milner-White EJ (2011) The structure of the ends of α -helices in globular proteins: effect of additional hydrogen bonds and implications for helix formation. *Proteins* 79:1010–1019
- Leader DP, Milner-White EJ (2012) Structure Motivator: a tool for exploring small three-dimensional elements in proteins. *BMC Struct Biol* 12:26. doi:10.1186/1472-6807-12-26
- Leone V, Marinelli F, Carloni P et al (2010) Targeting biomolecular flexibility with metadynamics. *Curr Opin Struct Biol* 20:148–154
- Lindorff-Larsen K, Best RB, Depristo MA et al (2005) Simultaneous determination of protein structure and dynamics. *Nature* 433:128–132
- Lindorff-Larsen K, Trbovic N, Maragakis P et al (2012) Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J Am Chem Soc* 134:3787–3791
- MacArthur MW, Thornton JM (1991) Influence of proline residues on protein conformation. *J Mol Biol* 218:397–412
- Marsh JA, Forman-Kay JD (2009) Structure and disorder in an unfolded state under non-denaturing conditions from ensemble models consistent with a large number of experimental restraints. *J Mol Biol* 391:359–374
- Marsh JA, Singh VK, Jia Z et al (2006) Sensitivity of secondary structure propensities to sequence differences between α - and γ -synuclein: implications for fibrillation. *Protein Sci* 15:2795–2804
- Marsh JA, Baker JMR, Tollinger M et al (2008) Calculation of residual dipolar couplings from disordered state ensembles using local alignment. *J Am Chem Soc* 130:7804–7805
- Mittermaier A, Kay LE (2006) New tools provide new insights in NMR studies of protein dynamics. *Science* 312:224–228
- Mittermaier AK, Kay LE (2009) Observing biological dynamics at atomic resolution using NMR. *Trends Biochem Sci* 34:601–611
- Mukrasch MD, Markwick PRL, Biernat J et al (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. *J Am Chem Soc* 129:5235–5243

- Muñoz V, Serrano L (1995) Helix design, prediction and stability. *Curr Opin Biotechnol* 6:382–386
- Narayanan RL, Dürr UHN, Bibow S et al (2010) Automatic assignment of the intrinsically disordered protein Tau with 441-residues. *J Am Chem Soc* 132:11906–11907
- Nodet G, Salmon L, Ozenne V et al (2009) Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *J Am Chem Soc* 131:17908–17918
- Ozenne V, Bauer F, Salmon L et al (2012a) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28:1463–1470
- Ozenne V, Schneider R, Yao M et al (2012b) Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J Am Chem Soc* 134:15138–15148
- Piana S, Laio A (2007) A bias-exchange approach to protein folding. *J Phys Chem B* 111:4553–4559
- Pierce LCT, Salomon-Ferrer R, Augusto F et al (2012) Routine access to millisecond time scale events with accelerated molecular dynamics. *J Chem Theory Comput* 8:2997–3002
- Prestegard JH, Bougault CM, Kishore AI (2004) Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* 104:3519–3540
- Różycki B, Kim YC, Hummer G (2011) SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure* 19:109–116
- Rückert M, Otting G (2000) Alignment of biological macromolecules in novel nonionic liquid crystalline media for NMR Experiments. *J Am Chem Soc* 122:7793–7797
- Salmon L, Bouvignies G, Markwick PRL et al (2009) Protein conformational flexibility from structure-free analysis of NMR dipolar couplings: quantitative and absolute determination of backbone motion in ubiquitin. *Angew Chem Int Ed Engl* 48:4154–4157
- Salmon L, Nodet G, Ozenne V et al (2010) NMR characterization of long-range order in intrinsically disordered proteins. *J Am Chem Soc* 132:8407–8418
- Salmon L, Bouvignies G, Markwick P et al (2011) Nuclear magnetic resonance provides a quantitative description of protein conformational flexibility on physiologically important time scales. *Biochemistry* 50:2735–2747
- Salmon L, Pierce L, Grimm A et al (2012) Multi-timescale conformational dynamics of the SH3 domain of CD2-associated protein using NMR spectroscopy and accelerated molecular dynamics. *Angew Chem Int Ed Engl* 51:6103–6106
- Sass HJ, Musco G, Stahl SJ et al (2000) Solution NMR of proteins within polyacrylamide gels: diffusional properties and residual alignment by mechanical stress or embedding of oriented purple membranes. *J Biomol NMR* 18:303–309
- Schalwe M, Ozenne V, Bibow S et al (2014) Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α -synuclein in solution from NMR and small angle scattering. *Structure* 22:238–249
- Schweitzer-Stenner R (2012) Conformational propensities and residual structures in unfolded peptides and proteins. *Mol Biosyst* 8:122–133
- Serrano L (1995) Comparison between the phi distribution of the amino acids in the protein database and NMR data indicates that amino acids have various phi propensities in the random coil conformation. *J Mol Biol* 254:322–333
- Shen Y, Bax A (2012) Identification of helix capping and β -turn motifs from NMR chemical shifts. *J Biomol NMR* 52:211–232
- Shi Z, Chen K, Liu Z et al (2006) Conformation of the backbone in unfolded proteins. *Chem Rev* 106:1877–1897
- Shortle D, Ackerman MS (2001) Persistence of native-like topology in a denatured protein in 8 M urea. *Science* 293:487–489
- Smith LJ, Bolin KA, Schalwe H et al (1996) Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. *J Mol Biol* 255:494–506
- Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett* 314:141–151

- Tamiola K, Mulder FAA (2012) Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. *Biochem Soc Trans* 40:1014–1020
- Tamiola K, Acar B, Mulder FAA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J Am Chem Soc* 132:18000–18003
- Tjandra N, Bax A (1997) Direct measurement of distances and angles in biomolecules by NMR in a dilute liquid crystalline medium. *Science* 278:1111–1114
- Tolman JR, Flanagan JM, Kennedy MA et al (1995) Nuclear magnetic dipole interactions in field-oriented proteins: information for structure determination in solution. *Proc Natl Acad Sci U S A* 92:9279–9283
- Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* 37:509–516
- Voter AF (1997) Hyperdynamics: accelerated molecular dynamics of infrequent events. *Phys Rev Lett* 78:3908–3911
- Vuister GW, Bax A (1993) Quantitative J correlation: a new approach for measuring homonuclear three-bond J(HNH. α) coupling constants in ^{15}N -enriched proteins. *J Am Chem Soc* 115:7772–7777
- Wang Y, Jardetzky O (2002a) Investigation of the neighboring residue effects on protein chemical shifts. *J Am Chem Soc* 124:14075–14084
- Wang Y, Jardetzky O (2002b) Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci* 11:852–861
- Wells M, Tidow H, Rutherford TJ et al (2008) Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A* 105:5762–5767
- Wishart DS, Bigam CG, Holm A et al (1995) ^1H , ^{13}C and ^{15}N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *J Biomol NMR* 5:67–81
- Yang S, Blachowicz L, Makowski L et al (2010) Multidomain assembled states of Hck tyrosine kinase in solution. *Proc Natl Acad Sci U S A* 107:15757–15762
- Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. *J Biomol NMR* 25:173–195
- Zweckstetter M, Bax A (2000) Prediction of sterically induced alignment in a dilute liquid crystalline phase: aid to protein structure determination by NMR. *J Am Chem Soc* 122:3791–3792