Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database

Alejandro N Estaña, Nathalie Sibille, Elise Delaforge, Marc Vaisset, Juan Cortés, Pau Bernadó

Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database

Alejandro Estaña[a,b], Nathalie Sibille[b], Elise Delaforge[b], Marc Vaisset[a], Juan Cortés[a,*], Pau Bernadó[b,**]

[a] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France*
[b] *Centre de Biochimie Structurale. INSERM, CNRS, Université de Montpellier, France*

Abstract

Intrinsically Disordered Proteins (IDPs) play fundamental roles in signaling, regulation and cell homeostasis by specifically interacting with their partners. The structural characterization of these interacting regions remains challenging and requires the integration of extensive experimental information. Here we present an approach that exploits the structural information encoded in tripeptide fragments from coil regions of high-resolution structures. Our results indicate that a simple building approach that disregards the sequence context provides a good structural representation of fully disordered regions. Conversely, the description of partially structured motifs calls for the consideration of sequence-dependent structural preferences. By using NMR Residual Dipolar Couplings and SAXS data for multiple IDPs we demonstrate that the appropriate combination of these two building strategies produces ensemble models that correctly describe the secondary structural classes and the population of partially structured regions. This study paves the way for the extension of structure prediction and protein design to disordered proteins.

Keywords: Intrinsically disordered proteins, Conformational sampling, Residual dipolar couplings, Protein fragment database.

*Corresponding authors
**Lead Contact
Email address: juan.cortes@laas.fr,pau.bernado@cbs.cnrs.fr

1. Introduction

Intrinsically Disordered Proteins or Regions (IDPs/IDRs) play crucial roles in multiple biological processes and are directly involved in several pathologies, including cancer and neurodegeneration (Uversky et al., 2008; Csizmok et al., 2016; Babu et al., 2011). The inherent plasticity of this family of proteins facilitates a range of functions that are complementary to those of their folded counterparts (Xie et al., 2007). In most cases, the activity of IDPs is manifested when interacting with globular partners to trigger signaling or metabolic cascades (Tompa et al., 2015). These interactions are mediated by Short Linear Motifs (SLiMs) that recognize regions of the partner surface in a highly specific manner (Van Roey et al., 2014). The presence of transiently formed structural motifs in SLiMs facilitates partner recognition and tunes the thermodynamics and kinetics of interactions (Mohan et al., 2006; Pancsa and Fuxreiter, 2012; Schneider et al., 2015). To understand these functional mechanisms, it is pivotal to identify and characterize these partially structured elements inserted into IDPs.

The relatively flat conformational energy landscape of IDPs has notably hampered their structural characterization. Experimental data obtained by Nuclear Magnetic Resonance (NMR) and Small-Angle X-ray Scattering (SAXS) provide information on conformational trends at the residue level, the presence of transient long-range contacts, and the overall size of the ensemble of conformations (Eliezer, 2009). However, the quantitative interpretation of these data requires the use of computational approaches that account for their ensemble averaging properties. These computational approaches are based on the construction of large conformational ensembles, which are subsequently refined by integrating the experimental data using restrained Molecular Dynamics (MD) simulations (Dedmon et al., 2005; Silvestre-Ryan et al., 2013), sub-ensemble selection (Ozenne et al., 2012b; Krzeminski et al., 2013; Bernadó et al., 2007), or Bayesian statistics (Fisher et al., 2010). Chemical Shifts (CSs) and Residual Dipolar Couplings (RDCs) measured in partially aligned media are the most

sensitive probes to quantify conformational restrictions at the residue level and to define secondary structural elements (Dyson and Wright, 2004; Jensen et al., 2009). Conversely, ensembles refined with SAXS data describe the overall properties of the protein in solution (Bernadó and Svergun, 2012; Receveur-Brechot and Durand, 2012). Consequently, conformational ensembles that simultaneously describe both sources of complementary information are excellent structural models of proteins in solution (Sibille and Bernadó, 2012; Cordeiro et al., 2017).

Multiple computational tools using different levels of description have been developed to characterize IDPs when no or limited experimental information is available. Current disorder prediction tools, which are based on the statistical analysis of protein sequences, provide rough estimations of partly structured regions in IDPs (Deng et al., 2015), although the exact secondary structure classes are poorly defined.

In principle, a more accurate characterization can be provided by MD-based methods. However, despite significant advances in the extension of MD methods to IDPs (Piana et al., 2015; Henriques et al., 2015), their applicability to exhaustively explore the conformational space of these proteins is still limited. Knowledge-based approaches have emerged as an alternative to overcome some of these limitations. These approaches usually describe the conformational properties of individual residues using the so-called coil libraries, which contain residue-specific $\{\phi, \psi\}$ angles from fragments of experimentally determined protein structures that do not form secondary structural elements (Smith et al., 1996; Feldman and Hogue, 2000; Jha et al., 2005; Bernadó et al., 2005; Fitzkee et al., 2005; Ting et al., 2010; Esteban-Martin et al., 2010; Shen et al., 2018). Despite their simplicity, coil models provide an accurate description of NMR parameters such as J-couplings (Smith et al., 1996; Shen et al., 2018) and RDCs (Bernadó et al., 2005; Jensen et al., 2009), and SAXS curves (Bernadó and Svergun, 2012) for flexible peptides and disordered proteins. To ensure a large conformational exploration, the most common methods sequentially append individual residues using peptide planes (Bernadó et al., 2005) or Cα atoms

(Feldman and Hogue, 2000) as building units. These coil models are normally used as background ensembles for the subsequent refinement using experimental data (Dedmon et al., 2005; Silvestre-Ryan et al., 2013; Ozenne et al., 2012b; Krzeminski et al., 2013; Bernadó et al., 2007; Fisher et al., 2010). Therefore, they are not supposed to identify secondary structural elements in IDPs, although conformations can be biased by including information from secondary structure predictors (Feldman and Hogue). This limitation is caused by the amino acid type specific conformational database that overlooks the sequence and structural context (Feldman and Hogue, 2000; Jha et al., 2005; Bernadó et al., 2005). Consequently, approaches such as TraDES and Flexible-Meccano provide realistic models for purely random coil regions, but do not capture structural features involving multiple consecutive residues. The omission of coordinated effects precludes the capacity of current approaches to predict structural classes and their populations, and hamper their application for advanced purposes.

Here we present a new approach to build atomistic models of IDPs that uses an extensive coil library of three-residue fragments (called tripeptides herein), which are the minimal fragments containing structural information (Huang et al., 2013). The exploitation of the structural information encoded in the library provides accurate descriptions of RDCs and SAXS datasets for multiple disordered proteins presenting distinct secondary structural motifs. This observation suggests that, by capturing conformational restrictions in turns, α-helices, and β-strands inserted in IDPs, our structural ensembles are realistic models of these proteins. The relative population, the internal coordination that transiently stabilizes these secondary structural elements, and the fluctuating behavior of these elements naturally emerge from our strategy. Our study seeks to extend structure prediction approaches to disordered chains, thereby enabling the identification of the structural perturbations that deleterious point mutations or alternative splicing exert on IDPs and IDRs.

2. Results

2.1. Computational models

A tripeptide coil database was built from high-resolution, experimentally determined protein structures (see Method Details). Tripeptides capture the conformational variability of the 20 proteinogenic amino acids while accounting for the effects of the closest neighboring residues. Using this tripeptide database and a simple steric term to avoid atom overlap, we generated ensembles of 100,000 conformations for several IDPs using the different building strategies explained below. N-HN RDCs and SAXS curves were computed from the resulting ensembles using standard methods (see Method Details) and were compared with the experimental datasets. RDCs for MAPK Kinase 7 (MKK7) (Kragelj et al., 2015), the fragment 955-1097 of the Erythrocyte binding antigen 181 (eba181) (Blanc et al., 2014), p15 (De Biasio et al., 2014), sic1 (Mittag et al., 2010), Measles virus ntail (ntailMV) (Jensen et al., 2011), Sendai virus ntail (ntailSV) (Jensen et al., 2008), the unique domain of the src kinase (src) (Pérez et al., 2009), K18 fragment of Tau protein (K18) (Mukrasch et al., 2007), and full-length Tau protein (Schwalbe et al., 2014) were used to probe the residue-specific sampling of the models, including the presence of partially-formed secondary structural elements. The agreement of the different building strategies with the experimental data was quantified using Q-factors (Cornilescu et al., 1998) (Table S1). Moreover, SAXS curves for p15 (De Biasio et al., 2014), src (Arbesú et al., 2017), and Tau (Mylonas et al., 2008) were used to probe the overall size and shape of the ensembles constructed.

2.2. The coil model describes disordered regions in IDPs

As a first approach, we built the conformations by randomly selecting $\{\phi, \psi\}$ values from the database in a residue-specific manner without taking into account the neighboring residues. Only residues preceding prolines were specifically selected from the database, since the Ramachandran distributions of these residues differ considerably (MacArthur and Thornton, 1991; Ting et al., 2010).

This building mode, which we call single-residue-based sampling (SRS), can be considered a Flory model since the sequence context of the building units is not used. The RDC profiles computed using the SRS strategy nicely reproduced the experimental ones for large sections of all the proteins (Fig. 7, blue lines). Conversely, other regions displaying large (positive or negative) RDCs were not properly reproduced by SRS ensembles. Not surprisingly, this lack of agreement was observed in known α-helical regions with positive RDCs (ntailMV, ntailSV and MKK7), extended regions with strongly negative N-HN RDCs (p15), and turns displaying sharp positive peaks (eba181, K18 and Tau). Note that inaccuracies in the representation of partially structured regions have also been observed when using similar building strategies, such as Flexible-Meccano (Bernadó et al., 2005; Ozenne et al., 2012a). The proteins with highly populated secondary structural elements, such as ntailMV, ntailSV and MKK7 present large Q-factors (around 100).

2.3. Structural information encoded in the tripeptide database identifies partially formed secondary structural elements

We generated large conformational ensembles using a three-residue-based sampling strategy (TRS) that selects $\{\phi, \psi\}$ values for each residue i, taking into account the amino acid type and the conformation of the neighboring residues $i-1$ and $i+1$ (see Method Details). In general, RDCs derived from the TRS strategy adopted less negative or even positive values compared to those obtained from the SRS strategy (Fig. 7, green lines). In some cases, such as for eba181 and ntailMV, almost the entire RDC profile remained positive. We attribute this systematic deviation towards positive values to an overpopulation of α-helical conformations in the tripeptide database, as previously observed when using coil libraries derived from globular proteins (Jha et al., 2005; Schweitzer-Stenner and Toal, 2016). Interestingly, some local features observed in the experimental profiles, which were not reproduced by the SRS strategy, were captured by the TRS strategy. Theoretical RDCs for α-helical regions in ntailMV, ntailSV and MKK7 were systematically more positive than those

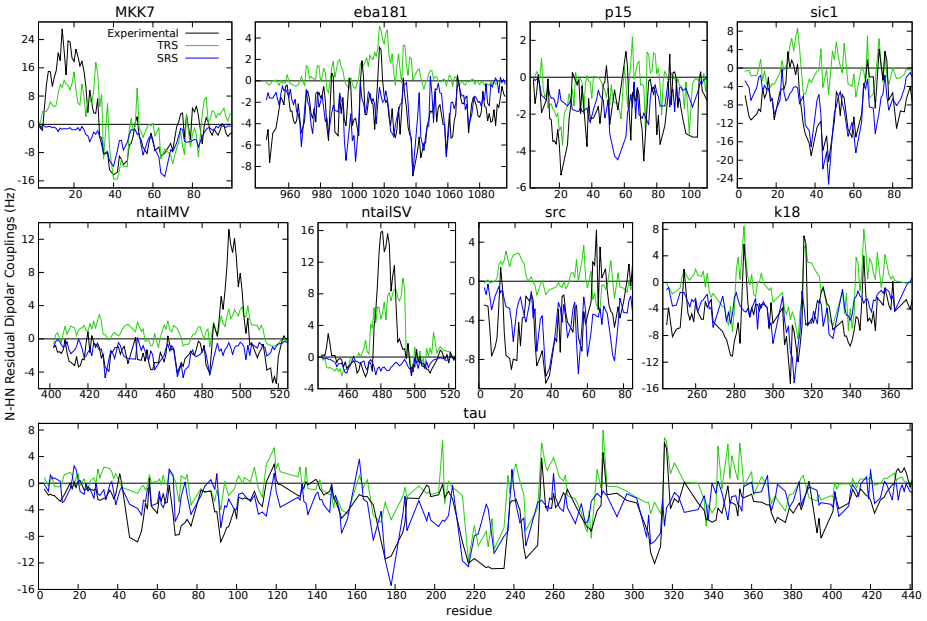

Figure 1: Experimental N-HN RDCs (black solid lines) for the nine proteins analyzed compared with the theoretical RDCs computed using the SRS (blue solid line) and TRS (green solid line) sampling strategies. To facilitate visual analysis, RDCs from the SRS method were scaled considering only the regions defined as random coil in the hybrid approach

.

corresponding to their flanking regions. In fact, these were the only three cases for which the Q-factor for the TRS was better than that of the SRS. Moreover, turns in K18 and Tau were naturally pinpointed by the TRS strategy, producing sharp peaks in the RDC profile. Note that more negative RDC values were also observed in some cases, such as the N-terminus of p15. These observations indicate that some tripeptide sequences in the database are enriched in particular conformational classes that are present in solution.

2.4. A hybrid sampling strategy simultaneously describes structural properties of disordered and partially ordered regions

The satisfactory description of disordered and partially structured regions achieved with the SRS and TRS strategies, respectively, prompted us to apply

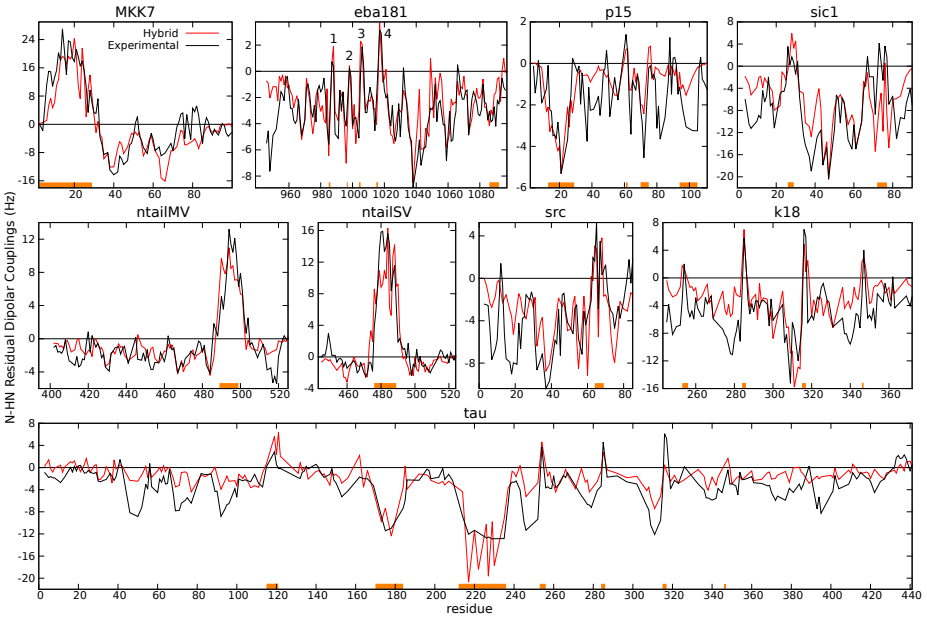

Figure 2: Experimental N-HN RDCs (black solid lines) for the nine IDPs studied compared with those computed using the hybrid SRS-TRS sampling strategy (red solid lines). Fragments highlighted in orange correspond to regions considered partially structured, for which the TRS was applied (see Table S2 for details).

a hybrid building approach. In this approach, residues belonging to a partially structured region defined *a priori* were incorporated into the model using the TRS strategy, while the rest of the chain was built with the SRS strategy. For the nine proteins tested, we defined the partially structured regions on the basis of the experimental N-HN RDCs and previously reported structural analyses (see Table S2). In this regard, SRS-derived RDCs were compared with the experimental ones, and those regions presenting a systematic deviation were initially assigned as partially structured. The exact borders of these regions were subsequently refined by testing multiple alternatives. The Q-factors, revealed excellent agreement between the simulated and the experimental RDC profiles for all the proteins tested (Fig. 8 and Table S1). This metric thereby indicates that the hybrid strategy, which simultaneously describes disordered and partially structured regions, notably improved the SRS and TRS chain

building approaches. However, the level of Q-factor improvement depended on the percentage of the sequence involved in secondary structural elements (Table S1). In highly disordered proteins such as eba181, the improvement of the hybrid method with respect to the SRS approach was modest, with Q-factors of 56.01 and 46.90 for the SRS and hybrid strategies respectively. Conversely, a considerable improvement in the Q-factor was observed in proteins with long and highly populated α-helices, such as MKK7, ntailMV and ntailSV, whose Q-factors decreased from 100.36, 98.89 and 110.62 for SRS to 45.20, 47.23 and 43.97 with the hybrid strategy, respectively.

Computed RDCs for the α-helical regions of MKK7, ntailMV and ntailSV nicely reproduced the experimentally observed bell-shape and the saw-teeth. Importantly, the description of the positive RDCs did not compromise that of the disordered regions as the model captured their relative intensity. Other characteristic features observed in the experimental RDC profiles, such as turns in eba181, K18 and Tau (see below), the broken helix in the 60-75 fragment of src caused by two consecutive glycine residues (Pérez et al., 2009), and the sharp inverse γ-turn of W61 of p15 (De Biasio et al., 2014), naturally emerged when using the hybrid approach. Remarkably, this building method did not require the specification of either the type or the population of secondary structures. Protein Tau is a particularly challenging example due to its size and the presence of multiple structural features, which have been extensively studied by NMR (Mukrasch et al., 2007; Ozenne et al., 2012b; Schwalbe et al., 2014). Seven regions of Tau were defined as structured using the hybrid approach, four of them being the well described turns found in the repeat region corresponding to the K18 construct (Mukrasch et al., 2007; Ozenne et al., 2012b). The presence of highly positive RDC values found in these four turns were captured by the hybrid approach in both proteins (Fig. 8), thereby indicating the realistic conformational representation of their sub-sequences in the database.

CSs were used to further validate the conformational ensembles built with the hybrid SRS-TRS strategy. In this regard, averaged Cα, Cβ, CO and NH CSs for ntailMV were computed from the ensembles using the program SPARTA+

(Shen and Bax, 2010) and then compared with the experimental ones (Fig. S4). The simulated CSs were in good agreement with the experimental ones, and they clearly captured deviations from the purely random coil behavior represented by the SRS ensemble. These observations substantiate the results obtained when using RDCs.

2.5. Comparison to SAXS data

SAXS accurately probes the overall properties of conformational ensembles in solution, thus complementing the residue-specific information provided by RDCs and CSs (Cordeiro et al., 2017; Sibille and Bernadó, 2012). Simulated SAXS profiles were computed from the ensembles using standard procedures (see Method Details). Overall, excellent agreement between experimental and simulated profiles was observed for the three proteins, with χ^2 of 1.93, 1.04, and 1.52 for src, p15 and Tau, respectively (Fig. S1). For src and Tau, these values were notably better than those obtained with the SRS (χ^2 of 2.70 and 2.02) and the TRS (χ^2 of 2.58 and 2.15) sampling approaches. For p15, the profiles achieved the three sampling strategies showed an excellent correlation with the experimental profile, with χ^2 near 1.0. These results strongly suggest that the ensembles built with the hybrid approach properly describe the overall properties of IDPs.

2.6. Prediction of local conformations and secondary structural elements

The previous sections demonstrate that the ensembles built with the hybrid approach are realistic models of IDPs in solution. Next, we explored the structural features of the resulting models using the helical region in ntailMV, the extended region at the N-terminus of p15, and the turns in eba181 and K18 as examples.

For ntailMV, the hybrid strategy notably enriched the structured region in α-helical conformations while it was depleted in extended (β-S) and polyproline-II (β-P) (Fig. 9a). This structural enrichment in helical conformations induced

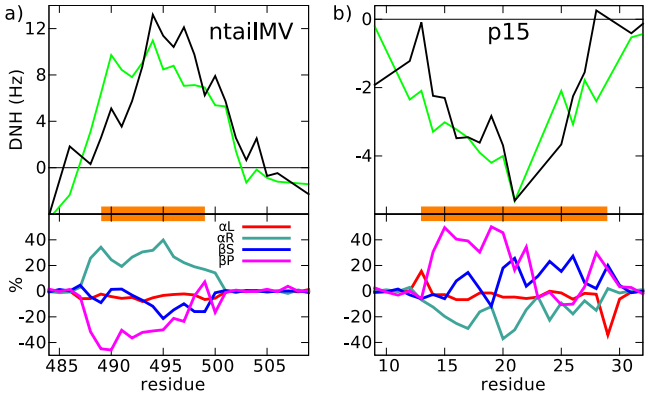

Figure 3: Experimental (black) and hybrid building model (green) N-HN RDCs for two fragments of (a top) ntailMV and (b top) p15. Fragments highlighted in orange were considered partially structured and built using the TRS strategy. In bottom panels, the percentage of enrichment of secondary structure classes present in the ensemble built with the hybrid strategy compared with that built with the SRS strategy. Secondary structure classes were identified using definitions in related work (Ozenne et al., 2012b). Concretely, $[\beta S : -100 > \phi; -120 > \psi > 50]$, $[\beta P : 0 > \phi > -100; -120 > \psi > 50]$, $[\alpha R : 0 > \phi; 50 > \psi > -120]$, $[\alpha L : \phi > 0]$.

positive RDC values in this region. The conformational analysis of the ensemble built for the N-terminus of p15 indicated a strong enrichment in extended conformations, β-S and β-P, whereas α-helical ones were depleted (Fig. 9b). Interestingly, neither β-S nor β-P were homogeneously populated along the segment, and either one or the other became dominant depending on the specific sequence.

A highly relevant feature of the hybrid strategy is its ability to identify turns from sequences. Four turns have been localized in eba181 based on their positive RDCs (Blanc et al., 2014), however the sizes of these RDCs differed (Fig. 8). While turns 3 (DASL) and 4 (DDAK) presented highly positive values, turns 1 (DPEK) and 2 (DPNT) were only slightly positive thereby suggesting distinct structural features. Fig. 10 shows the conformations adopted by the residues involved in the four turns. In all turns, residue $i + 1$ adopted an α-helical conformation. However, while residue i in turns 1 and 2 was mainly extended

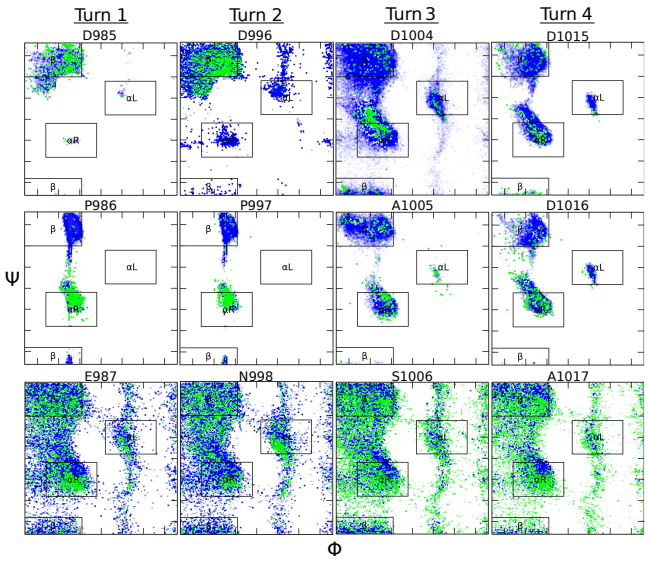

Figure 4: Conformational sampling for the four turns identified in eba181. Each column displays the Ramachandran plots for the three first residues in turns 1 to 4 when using the SRS (blue) or the hybrid (green) sampling approaches.

due to the following proline, it was α-helical in turns 3 and 4. This structural difference most probably explains the different RDC values of the four turns. According to current definitions (de Brevern, 2016), the four turns can be considered β-turns, types *I* and *VIII* being compatible with the conformation of the residue $i+1$. Nevertheless, the sequence composition clearly suggests that turns 1 and 2 with D and P in positions i and $i+1$, respectively, are type *I* β-turns (de Brevern, 2016). In another example, the four turns identified in K18 were enriched in α-helical conformations in their two central residues (Fig. S2), an observation that is in line with the original study (Mukrasch et al., 2007). However, residues in position $i+1$ (L253, L248, L315 and F346) sampled the region $\{\phi = -90, \psi = 0\}$ whereas residues $i+2$ (K254, N285, S316, and K347) adopted mainly an α-helical conformation with $\{\phi = -60, \psi = -30\}$. Although resembling type *I* β-turns, they did not adopt the canonical conformation (de Brevern, 2016).

2.7. Coordinated formation of structural elements

We further studied how secondary structural elements are formed within the conformational ensembles using the helical region in MKK7 as an example (Fig. 11a). The Secondary Structure-map (SS-map) (Iglesias et al., 2013), which allows the quantification of multiple structured elements within conformational ensembles, was used for this analysis. According to the SS-map, the ensemble of the N-terminal region of MKK7 presented scarcely populated helical regions of virtually all sizes from 4 up to 28 residues. Although the helix encompassing the whole 28-residue-long region was found in the ensemble, its population was extremely low, and shorter α-helices were preferred. In this regard the most populated helices (around 5%) involved eight and nine residues in non-overlapping segments of the protein. Interestingly, the N-terminal region of this fragment seemed more prone to form long α-helices expanding up to 15 residues. The continuum of multiple overlapping helical sections observed in the ensemble of MKK7, which induces the bell-shape of the resulting RDC profile, highlights the conformational complexity of helical regions in IDPs.

We tested two alternative procedures to introduce helicity into ensembles generated using a Flory model (i.e. the SRS strategy in our implementation) that are frequently used to describe NMR data (De Biasio et al., 2014; Ozenne et al., 2012b; Pérez et al., 2009; Wells et al., 2008; Bernadó and Blackledge, 2009). Firstly, a 25% increase in α-helical conformations was imposed for each of the residues within the region, but no structural coordination between residues was forced (Fig. 11b). Secondly, a canonical α-helix spanning the 28-residue-long region was introduced in 25% of the conformations (Fig. 11c). When the helical tendency was increased at the residue level, the resulting ensemble displayed multiple short helices spanning the whole region. However, the population of longer helices decreased dramatically. Consequently, resulting RDCs were positive but with values close to zero and they did not display residue-specific features. When a canonical α-helix was forced within the complete region no shorter helices spontaneously formed in the remaining 75% of the ensemble. As a result of this conformational homogeneity, the RDC profile adopted large

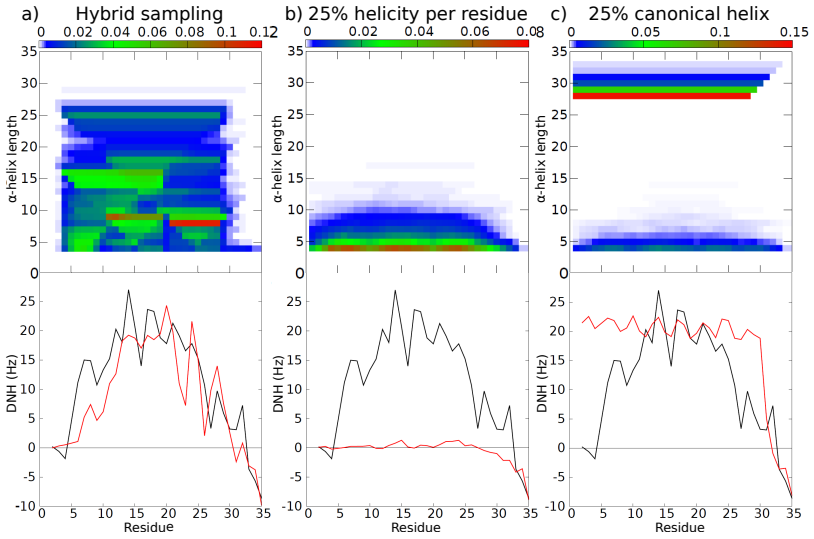

Figure 5: Structural analysis of the helical region in MKK7. (Top panels) Length and encompassing residues of the α-helices found in ensembles computed using (a) the hybrid sampling and two theoretical models imposing (b) 25% of enhanced helicity per residue, and (c) 25% population of a canonical α-helix in the 28-residue long segment. Colors from white to red indicate the population of helical segments found in the ensembles. (Bottom panels) Theoretical RDCs calculated from the above described ensembles (red lines) compared with the experimental ones (black lines).

positive values with the saw-teeth shape induced by the continuous α-helix. However, RDCs did not present the overall bell-shape observed experimentally.

To further evaluate the ensembles generated with the aforementioned procedures, we also used two-dimensional plots that display the deviation with respect to a canonical α-helix (see Fig. S3 and the associated explanations in SI). In these plots, each conformation is represented by a point with the x coordinate corresponding to the distance between the first N and last C backbone atoms of the 28-residue fragment, and the y coordinate corresponding to the average distance between H-bond donor and acceptor atoms within this fragment. The wide structural heterogeneity found in ensembles built with the hybrid approach was clearly highlighted using this representation. In contrast, distributions produced by the two other approaches were less likely from a physical point of view.

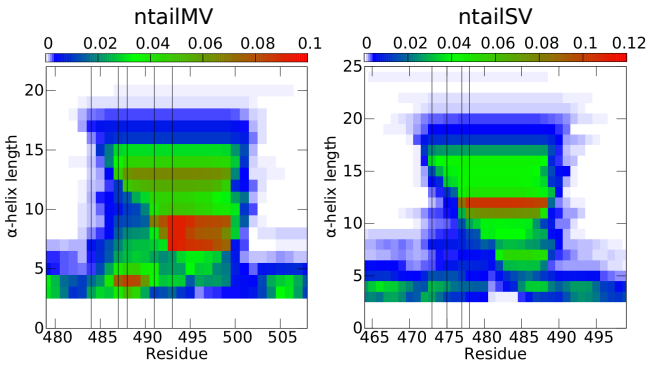

Figure 6: SS-map analysis for the helical regions in ntailMV and ntailSV displaying the length and the composing residues of the α-helices found in ensembles generated with the hybrid sampling strategy for both proteins. Color from white to red indicates the population of these helices. Vertical lines indicate aspartic acids and serines in the sequence that act as helix N-capping residues. Concretely, D484, D487, S488, S491, and D493 are highlighted for ntailMV, and D473, D475, S477, and D478 and highlighted for ntailSV.

In particular, the discontinuity in the conformational space produced by imposing a given percentage of a canonical helix is unrealistic. Indeed, although this last procedure can yield good agreement between computationally generated ensembles and experimental data in some cases, these ensembles are inaccurate representations of the conformational heterogeneity expected in partially structured regions in IDPs.

A SS-map analysis was also performed in the helical regions of ntailSV and ntailMV (Fig. 12). As in the case of MKK7, the co-existence of multiple overlapping short α-helices was observed. However, in contrast to MKK7, these two proteins displayed a triangular shape in the SS-map, in agreement with their similar amino acid sequence and function. This shape arises from the presence at the N-terminus of the motif of multiple residues with a strong tendency to trigger the formation of α-helical segments. The most prevalent initial residues of the detected helices in our ensembles were aspactic acid and serine. These two amino acids have been identified as helix N-capping amino acids, which stabilize α-helices with their side chain by forming a hydrogen bond at positions 2 or 3 in the helix (Jensen et al., 2008; Lovell et al., 2003). This observation

suggests that the N-capping properties of these amino acids are encoded in the tripeptide database and that their capacity to initiate helical motifs naturally emerges in ensembles built with the hybrid strategy.

3. Discussion

Partially structured motifs are key elements to trigger signaling events and to regulate transcription and metabolic pathways (Tompa et al., 2015). The localization and characterization of these motifs inserted within fully disordered fragments have been the focus of intense research (Tompa et al., 2015; Van Roey et al., 2014; Mohan et al., 2006). Here we present an approach that exploits the structural information encoded in tripeptide fragments extracted from coil regions of experimentally determined protein structures to build realistic structural ensembles of IDPs/IDRs, including scarcely populated structured motifs. Although Flory models, which do not consider the sequence context, generate conformational ensembles with the capacity to reproduce diverse experimental data for disordered chains, they fail to predict and model partially structured elements. Our results demonstrate that the tripeptide database, which accounts for this sequence context, contains structural features that are subsequently found experimentally in solution. Whereas libraries involving larger fragments have been shown to be powerful tools for the prediction of probable (stable) conformations of globular proteins and peptides (Han and Baker, 1996; Kolodny et al., 2002; Rohl et al., 2004; Baeten et al., 2008; Shen et al., 2014; Mackenzie et al., 2016), our results highlight that our extensive database of three-residue fragments is enough to represent the conformational variability and local structural propensities in IDPs. Moreover, representing the conformational variability of disordered chains requires a broad sampling of structures, which would not be guaranteed using databases of larger fragments. In this regard our tripeptide database emerges as optimal for this purpose.

The general agreement between experimental and simulated RDCs implies that the residue-specific structural information encoded in our tripeptide database

is coherent with the conformational behavior of IDPs in solution. This is a remarkable observation as the database has been derived from coil regions of crystallographic structures, which are susceptible to experience packing contacts and/or reduced mobility. Therefore, the sequence context is a major determinant of structural propensities, regardless of the state (globular/disordered) or the environment (crystal/solution). However, for some sequences, a less accurate agreement between the experimental and simulated RDC profiles has been observed. We attribute this local lack of agreement to the limited conformational coverage of these sequences in our database. With the increasing number of experimentally determined high-resolution protein structures, we expect that more extensive and higher quality tripeptide databases will be built in the future, which will further improve the quality of conformational ensembles generated with our method.

Our approach relies on the discrimination between disordered and partially structured regions to subsequently apply the SRS and TRS sampling strategies, respectively. Here we have used the experimental RDCs and previous studies of the considered proteins to define both regions. In the absence of RDCs, other experimental data and bioinformatics predictions can be used to identify partially structured motifs. CSs, which are the primary information derived from NMR, are also very sensitive to small conformational bias at the residue level (Tamiola et al., 2010; Schwarzinger et al., 2001). Partially structured motifs can also be discriminated from fully disordered regions by their faster NMR transverse relaxation rates (Jensen et al., 2011; De Biasio et al., 2014). Multiple bioinformatics tools based on different principles identify regions prone to forming structures (Deng et al., 2015). Another interesting source to distinguish structured elements is sequence conservation analysis. In IDPs, motifs involved in protein-protein interactions present slower mutational rates when compared to non-functional regions (Ota and Fukuchi, 2017). The tripeptide database can also be used to identify structured regions. In several examples, such as ntailSV, MKK7, and p15, the TRS sampling strategy pinpointed partially structured regions and turns when yielding larger RDC values (either positive or negative)

than the rest of the chain. This observation is caused by the conformational enrichment that these sequences present in the database, which biases the sampling and narrows the RDC averaging. In this context, and in the absence of experimental information, a simple TRS ensemble can provide insights into structurally relevant motifs within IDPs.

Partially structured motifs are not permanently folded in IDPs. They can be seen as an equilibrium between conformations hosting distinct smaller structured elements that are in continuous exchange driven by their extension or shortening. In other words, these sequences lack the internal coordination to form permanent secondary structural motifs and, as a consequence, are susceptible to partial unfolding events. Recognition processes exploit this structural heterogeneity to efficiently achieve the desired biological tasks. Binding affinities of the co-existing conformers are modulated by the entropic penalty caused by the folding of the recognition motif fragment that remains disordered in the unbound state (Pancsa and Fuxreiter, 2012). Moreover, recognition kinetics studies have demonstrated the existence of transiently populated encounter complexes, and different conformational states of the recognition element most probably present distinct energy barriers to achieve the final bound form (Schneider et al., 2015; Sugase et al., 2007; Delaforge et al., 2018). In the context of RDCs, the coexistence of multiple partially folded helical elements in the same region leads to the bell-shaped RDC profile and the saw-teeth, which report on the prevalence of the different helical fragments. Importantly, this structural heterogeneity is nicely captured by our hybrid sampling strategy, thereby highlighting the correspondence between the information encoded in the database and the conformational sampling of IDPs in solution. This feature is exemplified by the helix N-capping properties that we observed in the ensembles of ntailMV and ntailSV.

In summary, we have developed a method to build realistic conformational ensembles of IDPs and IDRs that describes scarcely populated secondary structural elements embedded in otherwise fully disordered regions. Our strategy is based on an extensive database of tripeptide structures and on the sepa-

ration between disordered and partially structured regions within the chain. Conformationally-biased ensembles generated with our approach will be better starting models for programs that integrate experimental data to derive structural models of IDPs. This will be specially relevant for strategies such as those based on the maximum entropy principle, aiming at minimizing the structural perturbation exerted to the initial ensemble to fit the experimental data (Esteban-Martin et al., 2010; Rozycki et al., 2011). Moreover our approach detects binding motifs involved in partner recognition that are, in most cases, linked to biological tasks. Our approach has the potential to anticipate structural effects caused by point mutations with an eventual role in disease, and the insertion or deletion of disordered fragments originating from alternative splicing processes. In this regard, we believe that our approach is the first step towards extending structural bioinformatics and protein design to disordered proteins.

Acknowledgments

The authors thank M. Zweckstetter (U. Göttingen), M. Blackledge and M.R. Jensen (IBS-Grenoble) for sharing experimental data. This work was supported by the European Research Council under the H2020 Programme (2014-2020) *chemREPEAT* [648030], and Labex EpiGenMed (ANR-10-LABX-12-01) awarded to P.B., and the ANR GPCteR (ANR-17-CE11-0022-01) to N.S. The CBS is a member of the French Infrastructure for Integrated Structural Biology-FRISBI (ANR-10-INSB-05). We used the HPC resources of the CALMIP supercomputing center under the allocation 2016-P16032.

Author Contributions

J.C. and P.B. designed the research; A.E., N.S., J.C. and P.B. conceived the methods; A.E., and M.V. and J.C. implemented the methods; N.S., E.D. and P.B. collected the experimental data; A.E. performed the computational

experiments; A.E., N.S., J.C. and P.B analyzed the data; A.E., J.C. and P.B. wrote the paper.

STAR Methods

Contact for Reagent and Resource Sharing

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Pau Bernadó (pau.bernado@cbs.cnrs.fr).

Method Details

Experimental NMR and SAXS data

Details on the RDCs and SAXS data analyzed in this study can be found in the original articles cited in Section 2.1.

Structural database

The tripeptide database was built from a curated database of high-resolution experimentally determined protein structures. We used the SCOPe (Fox et al., 2014) 2.06 release, with entries having less than 95% sequence identity to each other. A total of 8,907,065 of three-residue fragments were extracted from these protein structures and classified on the basis of their sequence (8,000 tripeptide classes).

Conformations sampled by residues were assigned using the program DSSP (Kabsch and Sander, 1983), which allowed us to filter out fragments corresponding to α-helices and β-strands.

More precisely, we removed all tripeptides containing at least one residue involved in these types of secondary structures (i.e. DSSP types H, G, I, E and B) from the database. This applied to approximately 60% of the total number of tripeptides extracted from the SCOPe database. The remaining 40% of the tripeptides (3,645,381), which contained residues in loop/coil regions (i.e. DSSP types L, T, S), were included in the coil database.

Sampling methods

Conformations were built incrementally from N- to C-termini in a residue-by-residue manner. When placing a new residue, its backbone angles $\{\phi, \psi, \omega\}$ were extracted from the coil database. An all-atom model was used for the backbone, whereas a simplified model was used for the side-chains, considering a pseudo-atom placed at the Cβ position for each residue, as previously proposed (Levitt, 1976; Bernadó et al., 2005; Ozenne et al., 2012a).

When placing a new residue, collisions with the previously built residues were tested.

In case of collision, a new configuration of the residue was sampled and tested. This was repeated until a valid configuration was found or a maximum number trials of 100 ($n_{fail}^{col} = 100$) was reached. In these cases a backtracking search process was applied, which consisted of removing the last three residues and restarting sampling from this point. When the backtracking process resulted unsuccessful, the chain construction was restarted from the beginning.

Single-residue-based sampling (SRS): This strategy is similar to the one used in Flexible-Meccano (Bernadó et al., 2005; Ozenne et al., 2012a). The backbone angles of each residue are sampled disregarding the neighboring residues. In this strategy, when the residue type is alanine, the angles are randomly selected among all tripeptide conformations of type X-Ala-Z, X and Z being any of the 20 amino acid types (i.e. 400 tripeptide sequence types). The process is slightly different when the Z residue is a proline. In this case, the conformation is selected from sequences X-Ala-Pro.

Three-residue-based sampling (TRS): This strategy takes into account the sequence of the neighboring residues $i-1$ and $i+1$ when sampling the conformation of residue i. In other words, when the amino acid types of residues $i-1$, i, $i+1$ are X, Y, Z, respectively, the conformation of residue i is sampled from the corresponding class X-Y-Z in the tripeptide database. In addition, the conformation of these two neighbors is considered in order to restrict sampling to the most structurally probable regions. For this purpose, sampling of residue i is constrained to a subset of conformations of the tripeptide class X-Y-Z, such

that the backbone angles of residue $i-1$ are within a given angular range ($\pm 20°$) around its current conformation, which was built in the previous step. Since the conformation of residue $i+1$ is not sampled in this building step, the structural restriction requires a back-step test. Once the conformation of residue i has been built, the conformation of the tripeptide formed by residues $i-2$, $i-1$, and i is checked to be present in the database of the corresponding sequence, considering the aforementioned angular tolerance. As for collision tests, this structural test can also fail. In this case, a backtracking process is also applied, with $n_{fail}^{str} = 250$.

Hybrid Sampling: The two sampling strategies SRS and TRS were combined in the hybrid strategy. Based on experimental RDCs and on additional information from previous studies, TRS is applied to sample partially structured regions while SRS is used for the disordered regions.

Computation of experimental properties from ensembles

Alignment properties and associated RDCs for each conformation were computed by exploiting the similarity between the radius of gyration and the alignment tensors as previously described (Almond and Axelsen, 2002; Bernadó et al., 2005). Reported RDCs correspond to averages over 100,000 conformations of each ensemble. Computational RDCs were homogeneously scaled to minimize discrepancy with the experimental ones. The agreement of the resulting RDCs with the experimental ones was evaluated using the Q-factor (Cornilescu et al., 1998): $Q = \texttt{rms}(D_{\text{meas}} - D_{\text{calc}})/\texttt{rms}(D_{\text{meas}})$, where D_{meas} and D_{calc} are the experimental and computed RDCs, respectively.

Ensemble-averaged SAXS data were computed from 2,000 randomly selected conformations from the ensembles generated with the three sampling strategies. Side-chains for each conformation were introduced with SCWRL4 (Krivov et al., 2009) before computation of its associated theoretical SAXS profile with CRYSOL (Svergun et al., 1995) using default parameters. The ensemble-averaged curve was compared with the experimental one by optimizing a scaling and a shift parameter, using χ^2 as a figure of merit. Averaged

Cα, Cβ, CO and NH chemical shifts were computed from ensembles of 5,000 conformations with SPARTA+ (Shen and Bax, 2010). Side-chains for each conformation were introduced with SCWRL4 (Krivov et al., 2009) before the calculation. Random coil chemical shifts were computed using POTENCI (Nielsen and Mulder, 2018) and subtracted from the computed ones to facilitate the interpretation.

Declaration of Interests

The authors declare no conflict of interest.

References

Almond, A., Axelsen, J.B. (2002). Physical interpretation of residual dipolar couplings in neutral aligned media. J. Am. Chem. Soc. 124, 9986–9987.

Arbesú, M., Maffei, M., Cordeiro, T.N., ao M.C. Teixeira, J., Pérez, Y., Bernadó, P., Roche, S., Pons, M. (2017). The unique domain forms a fuzzy intramolecular complex in Src family kinases. Structure 25, 630 – 640.e4.

Babu, M.M., van der Lee, R., de Groot, N.S., Gsponer, J. (2011). Intrinsically disordered proteins: Regulation and disease. Curr. Opin. Struct. Biol. 21, 432 – 440.

Baeten, L., Reumers, J., Tur, V., Stricher, F., Lenaerts, T., Serrano, L., Rousseau, F., Schymkowitz, J. (2008). Reconstruction of protein backbones from the brix collection of canonical protein fragments. PLOS Comput. Biol. 4, 1–11.

Bernadó, P., Blackledge, M. (2009). A self-consistent description of the conformational behavior of chemically denatured proteins from nmr and small angle scattering. Biophys. J. 97, 2839–2845.

Bernadó, P., Blanchard, L., Timmins, P., Marion, D., Ruigrok, R.W.H., Blackledge, M. (2005). A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. Proc. Natl. Acad. Sci. USA 102, 17002–17007.

Bernadó, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., Svergun, D.I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. J. Am. Chem. Soc. 129, 5656–5664.

Bernadó, P., Svergun, D.I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. Mol. BioSyst. 8, 151–167.

Blanc, M., Coetzer, T.L., Blackledge, M., Haertlein, M., Mitchell, E.P., Forsyth, V.T., Jensen, M.R. (2014). Intrinsic disorder within the erythrocyte binding-

like proteins from plasmodium falciparum. Biochim. Biophys. Acta 1844, 2306 – 2314.

de Brevern, A.G. (2016). Extension of the classical classification of beta-turns. Sci. Rep. 6, 33191.

Cordeiro, T.N., Herranz-Trillo, F., Urbanek, A.N., Estaña, A.N., Cortés, J., Sibille, N., Bernadó, P. (2017). Small-angle scattering studies of intrinsically disordered proteins and their complexes. Curr. Opin. Struct. Biol. 42, pp.15 – 23.

Cornilescu, G., Marquardt, J.L., Ottiger, M., Bax, A. (1998). Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J. Am. Chem. Soc. 120, 6836–6837.

Csizmok, V., Follis, A.V., Kriwacki, R.W., Forman-Kay, J.D. (2016). Dynamic protein interaction networks and new structural paradigms in signaling. Chem. Rev. 116, 6424–6462.

De Biasio, A., Ibáñez de Opakua, A., Cordeiro, T.N., Villate, M., Merino, N., Sibille, N., Lelli, M., Diercks, T., Bernadó, P., Blanco, F.J. (2014). p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins. Biophys. J. 106, 865 – 874.

Dedmon, M.M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., Dobson, C.M. (2005). Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. J. Am. Chem. Soc. 127, 476–477.

Delaforge, E., Kragelj, J., Tengo, L., Palencia, A., Milles, S., Bouvignies, G., Salvi, N., Blackledge, M., Jensen, M.R. (2018). Deciphering the dynamic interaction profile of an intrinsically disordered protein by nmr exchange spectroscopy. J. Am. Chem. Soc. 140, 1148–1158.

Deng, X., Gumm, J., Karki, S., Eickholt, J., Cheng, J. (2015). An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions. Int. J. Mol. Sci. 16, 15384–15404.

Dyson, H.J., Wright, P.E. (2004). Unfolded proteins and protein folding studied by NMR. Chem. Rev. 104, 3607–3622.

Eliezer, D. (2009). Biophysical characterization of intrinsically disordered proteins. Curr. Opin. Struct. Biol. 19, 23 – 30.

Esteban-Martin, S., Fenwick, R.B., Salvatella, X. (2010). Refinement of ensembles describing unstructured proteins using nmr residual dipolar couplings. J. Am. Chem. Soc. 132, 4626–4632.

Feldman, H.J., Hogue, C.W.Probabilistic sampling of protein conformations: New hope for brute force? Proteins: Struct., Funct., and Bioinf. 46, 8–23.

Feldman, H.J., Hogue, C.W.V. (2000). A fast method to sample real protein conformational space. Proteins: Struct., Funct., and Bioinf. 39, 112–131.

Fisher, C.K., Huang, A., Stultz, C.M. (2010). Modeling intrinsically disordered proteins with bayesian statistics. J. Am. Chem. Soc. 132, 14919–14927.

Fitzkee, N.C., Fleming, P.J., Rose, G.D. (2005). The protein coil library: A structural database of nonhelix, nonstrand fragments derived from the pdb. Proteins 58, 852–854.

Fox, N.K., Brenner, S.E., Chandonia, J.M. (2014). SCOPe: Structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucl. Acids Res. 42, D304–D309.

Han, K.F., Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. Proc. Natl. Acad. Sci. USA 93, 5814–5818.

Henriques, J., Cragnell, C., Skepö, M. (2015). Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment. J. Chem. Theory Comput. 11, 3420–3431.

Huang, J.R., Ozenne, V., Jensen, M.R., Blackledge, M. (2013). Direct prediction of NMR residual dipolar couplings from the primary sequence of unfolded proteins. Angew. Chem. Int. Edit. 52, 687–690.

Iglesias, J., Sanchez-Martínez, M., Crehuet, R. (2013). SS-map: Visualizing cooperative secondary structure elements in protein ensembles. Intrinsically Disord. Proteins 1, e25323.

Jensen, M.R., Communie, G., Ribeiro, E.A., Martinez, N., Desfosses, A., Salmon, L., Mollica, L., Gabel, F., Jamin, M., Longhi, S., Ruigrok, R.W.H., Blackledge, M. (2011). Intrinsic disorder in measles virus nucleocapsids. Proc. Natl. Acad. Sci. USA 108, 9839–9844.

Jensen, M.R., Houben, K., Lescop, E., Blanchard, L., Ruigrok, R.W.H., Blackledge, M. (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, M.R., Markwick, P.R., Meier, S., Griesinger, C., Zweckstetter, M., Grzesiek, S., Bernadó, P., Blackledge, M. (2009). Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings. Structure 17, 1169 – 1185.

Jha, A.K., Colubri, A., Freed, K.F., Sosnick, T.R. (2005). Statistical coil model of the unfolded state: Resolving the reconciliation problem. Proc. Natl. Acad. Sci. USA 102, 13099–13104.

Kabsch, W., Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577–2637.

Kolodny, R., Koehl, P., Guibas, L., Levitt, M. (2002). Small libraries of protein fragments model native protein structures accurately. J. Mol. Biol. 323, 297 – 307.

Kragelj, J., Palencia, A., Nanao, M.H., Maurin, D., Bouvignies, G., Blackledge, M., Jensen, M.R. (2015). Structure and dynamics of the MKK7-JNK signaling complex. Proc. Natl. Acad. Sci. USA 112, 3409–3414.

Krivov, G.G., Shapovalov, M.V., Dunbrack Jr., R.L. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins 77, 778–795.

Krzeminski, M., Marsh, J.A., Neale, C., Choy, W.Y., Forman-Kay, J.D. (2013). Characterization of disordered proteins with ensemble. Bioinformatics 29, 398–399.

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. J. Mol. Biol. 104, 59 – 107.

Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., Richardson, J.S., Richardson, D.C. (2003). Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins 50, 437–450.

MacArthur, M.W., Thornton, J.M. (1991). Influence of proline residues on protein conformation. J. Mol. Biol. 218, 397 – 412.

Mackenzie, C.O., Zhou, J., Grigoryan, G. (2016). Tertiary alphabet for the observable protein structural universe. Proc. Natl. Acad. Sci. USA 113, E7438–E7447.

Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., Tyers, M., Forman-Kay, J.D. (2010). Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an {SCF} ubiquitin ligase. Structure 18, 494 – 506.

Mohan, A., Oldfield, C.J., Radivojac, P., Vacic, V., Cortese, M.S., Dunker, A.K., Uversky, V.N. (2006). Analysis of molecular recognition features (MoRFs). J. Mol. Biol. 362, 1043 – 1059.

Mukrasch, M.D., Markwick, P., Biernat, J., von Bergen, M., Bernadó, P., Griesinger, C., Mandelkow, E., Zweckstetter, M., Blackledge, M. (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.

Mylonas, E., Hascher, A., Bernadó, P., Blackledge, M., Mandelkow, E., Svergun, D.I. (2008). Domain conformation of tau protein studied by solution small-angle X-ray scattering. Biochemistry 47, 10345–10353.

Nielsen, J.T., Mulder, F.A.A. (2018). Potenci: prediction of temperature, neighbor and ph-corrected chemical shifts for intrinsically disordered proteins. J. Biomol. NMR 70, 141–165.

Ota, H., Fukuchi, S. (2017). Sequence conservation of protein binding segments in intrinsically disordered regions. Biochem. Biophys. Res. Comm. 494, 602 – 607.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.r., Jensen, M.R., Segard, S., Bernadó, P., Charavay, C., Blackledge, M. (2012)a. 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., Huang, J.r., Salmon, L., Zweckstetter, M., Jensen, M.R., Blackledge, M. (2012)b. Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. J. Am. Chem. Soc. 134, 15138–15148.

Pancsa, R., Fuxreiter, M. (2012). Interactions via intrinsically disordered regions: What kind of motifs? IUBMB Life 64, 513–520.

Pérez, Y., Gairí, M., Pons, M., Bernadó, P. (2009). Structural characterization of the natively unfolded N-terminal domain of human c-Src kinase: Insights

into the role of phosphorylation of the unique domain. J. Mol. Biol. 391, 136 – 148.

Piana, S., Donchev, A.G., Robustelli, P., Shaw, D.E. (2015). Water dispersion interactions strongly influence simulated structural properties of disordered protein states. J. Phys. Chem. B 119, 5113–5123.

Receveur-Brechot, V., Durand, D. (2012). How random are intrinsically disordered proteins? a small angle scattering perspective. Curr. Protein. Pept. Sci. 13, 55–75.

Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D. (2004). Protein structure prediction using Rosetta, in: Numerical Computer Methods, Part D. Academic Press. volume 383 of *Method. Enzymol.*, pp. 66 – 93.

Rozycki, B., Kim, Y.C., Hummer, G. (2011). Saxs ensemble refinement of escrt-iii chmp3 conformational transitions. Structure 19, 109 – 116.

Schneider, R., Maurin, D., Communie, G., Kragelj, J., Hansen, D.F., Ruigrok, R.W.H., Jensen, M.R., Blackledge, M. (2015). Visualizing the molecular recognition trajectory of an intrinsically disordered protein using multinuclear relaxation dispersion NMR. J. Am. Chem. Soc. 137, 1220–1229.

Schwalbe, M., Ozenne, V., Bibow, S., Jaremko, M., Jaremko, L., Gajda, M., Jensen, M., Biernat, J., Becker, S., Mandelkow, E., Zweckstetter, M., Blackledge, M. (2014). Predictive atomic resolution descriptions of intrinsically disordered htau40 and alpha-synuclein in solution from NMR and small angle scattering. Structure 22, 238 – 249.

Schwarzinger, S., Kroon, G.J.A., Foss, T.R., Chung, J., Wright, P.E., Dyson, H.J. (2001). Sequence-dependent correction of random coil nmr chemical shifts. J. Am. Chem. Soc. 123, 2970–2978.

Schweitzer-Stenner, R., Toal, S.E. (2016). Construction and comparison of the statistical coil states of unfolded and intrinsically disordered proteins

from nearest-neighbor corrected conformational propensities of short peptides. Mol. BioSyst. 12, 3294–3306.

Shen, Y., Bax, A. (2010). Sparta+: a modest improvement in empirical nmr chemical shift prediction by means of an artificial neural network. J. Biomol. NMR 48, 13–22.

Shen, Y., Maupetit, J., Derreumaux, P., Tufféry, P. (2014). Improved PEP-FOLD approach for peptide and miniprotein structure prediction. J. Chem. Theory Comput. 10, 4745–4758.

Shen, Y., Roche, J., Grishaev, A., Bax, A. (2018). Prediction of nearest neighbor effects on backbone torsion angles and nmr scalar coupling constants in disordered proteins. Prot. Sci. 27, 146–158.

Sibille, N., Bernadó, P. (2012). Structural characterization of intrinsically disordered proteins by the combined use of NMR and SAXS. Biochem. Soc. Trans. 40, 955–962.

Silvestre-Ryan, J., Bertoncini, C., Fenwick, R., Esteban-Martin, S., Salvatella, X. (2013). Average conformations determined from pre data provide high-resolution maps of transient tertiary interactions in disordered proteins. Biophys. J. 104, 1740 – 1751.

Smith, L.J., Bolin, K.A., Schwalbe, H., MacArthur, M.W., Thornton, J.M., Dobson, C.M. (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.

Sugase, K., Dyson, H.J., Wright, P.E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature 447, 1021.

Svergun, D., Barberato, C., Koch, M.H.J. (1995). CRYSOL – a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. J. Appl. Crystallogr. 28, 768–773.

Tamiola, K., Acar, B., Mulder, F.A.A. (2010). Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J. Am. Chem. Soc. 132, 18000–18003.

Ting, D., Wang, G., Shapovalov, M., Mitra, R., Jordan, M.I., Dunbrack, Jr, R.L. (2010). Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model. PLOS Comput. Biol. 6, 1–21.

Tompa, P., Schad, E., Tantos, A., Kalmar, L. (2015). Intrinsically disordered proteins: Emerging interaction specialists. Curr. Opin. Struct. Biol. 35, 49 – 59.

Uversky, V.N., Oldfield, C.J., Dunker, A.K. (2008). Intrinsically disordered proteins in human diseases: Introducing the D2 concept. Ann. Rev. Biophys. 37, 215–246.

Van Roey, K., Uyar, B., Weatheritt, R.J., Dinkel, H., Seiler, M., Budd, A., Gibson, T.J., Davey, N.E. (2014). Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem. Rev. 114, 6733–6778.

Wells, M., Tidow, H., Rutherford, T.J., Markwick, P., Jensen, M.R., Mylonas, E., Svergun, D.I., Blackledge, M., Fersht, A.R. (2008). Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain. Proc. Natl. Acad. Sci. USA 105, 5762–5767.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., Obradovic, Z. (2007). Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J. Proteome Res. 6, 1882–1898.

Figure 7: Experimental N-HN RDCs (black solid lines) for the nine proteins analyzed compared with the theoretical RDCs computed using the SRS (blue solid line) and TRS (green solid line) sampling strategies. To facilitate visual analysis, RDCs from the SRS method were scaled considering only the regions defined as random coil in the hybrid approach.

Figure 8: Experimental N-HN RDCs (black solid lines) for the nine IDPs studied compared with those computed using the hybrid SRS-TRS sampling strategy (red solid lines). Fragments highlighted in orange correspond to regions considered partially structured, for which the TRS was applied (see Table S2 for details).

Figure 9: Experimental (black) and hybrid building model (green) N-HN RDCs for two fragments of (a top) ntailMV and (b top) p15. Fragments highlighted in orange were considered partially structured and built using the TRS strategy. In bottom panels, the percentage of enrichment of secondary structure classes present in the ensemble built with the hybrid strategy compared with that built with the SRS strategy. Secondary structure classes were identified using definitions in related work (Ozenne et al., 2012b). Concretely, $[\beta S : -100 > \phi; -120 > \psi > 50]$, $[\beta P : 0 > \phi > -100; -120 > \psi > 50]$, $[\alpha R : 0 > \phi; 50 > \psi > -120]$, $[\alpha L : \phi > 0]$.

Figure 10: Conformational sampling for the four turns identified in eba181. Each column displays the Ramachandran plots for the three first residues in turns 1 to 4 when using the SRS (blue) or the hybrid (green) sampling approaches.

Figure 11: Structural analysis of the helical region in MKK7. (Top panels) Length and encompassing residues of the α-helices found in ensembles computed using (a) the hybrid sampling and two theoretical models imposing (b) 25% of enhanced helicity per residue, and (c) 25% population of a canonical α-helix in the 28-residue long segment. Colors from white to red indicate the population of helical segments found in the ensembles. (Bottom panels) Theoretical RDCs calculated from the above described ensembles (red lines) compared with the experimental ones (black lines).

775

Figure 12: SS-map analysis for the helical regions in ntailMV and ntailSV displaying the length and the composing residues of the α-helices found in ensembles generated with the hybrid sampling strategy for both proteins. Color from white to red indicates the population of these helices. Vertical lines indicate aspartic acids and serines in the sequence that act as helix N-capping residues. Concretely, D484, D487, S488, S491, and D493 are highlighted for ntailMV, and D473, D475, S477, and D478 and highlighted for ntailSV.