Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments

Alejandro N Estaña, Amélie Barozet, Assia Mouhand, Marc Vaisset, Christophe Zanon, Pierre Fauret, Nathalie Sibille, Pau N Bernadó, Juan Cortés

▶ **To cite this version:**

Alejandro N Estaña, Amélie Barozet, Assia Mouhand, Marc Vaisset, Christophe Zanon, et al.. Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments. Journal of Molecular Biology, 2020, 342 (19), pp.5447-5459. 10.1016/j.jmb.2020.07.026 . hal-02920302

HAL Id: hal-02920302
https://laas.hal.science/hal-02920302

Submitted on 24 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments

Alejandro Estaña[a,b], Amélie Barozet[a], Assia Mouhand[b], Marc Vaisset[a], Christophe Zanon[a], Pierre Fauret[a], Nathalie Sibille[b], Pau Bernadó[b,*], Juan Cortés[a,1,*]

[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 key functional roles facilitated by their inherent plasticity. In most of the cases, IDPs recognize their partners through partially-structured elements inserted in fully-disordered chains. The identification and characterization of these elements is fundamental to understand the functional mechanisms of IDPs. Although several computational methods have been developed to identify order within disordered chains, most of the current secondary structure predictors are focused on globular proteins and are not necessarily appropriate for IDPs. Here, we present a comprehensible method, called Local Structural Propensity Predictor (LS2P), to predict secondary structure elements from IDP sequences. LS2P performs statistical analyses from a database of three-residue fragments extracted from coil regions of high-resolution protein structures. In addition to identifying scarcely populated helical and extended regions, the method pinpoints short stretches triggering β-turn formation or promoting α-helices. The simplicity of the method enables a direct connection between experimental observations and structural features encoded in IDP sequences.

Keywords: Intrinsically disordered proteins, short linear motifs, molecular recognition elements, secondary structure prediction, structural database.

*Corresponding authors
Email addresses: pau.bernado@cbs.cnrs.fr, juan.cortes@laas.fr
[1] Lead contact

1. Introduction

Intrinsically Disordered Proteins (IDPs) have emerged as key actors in multitude of relevant biological processes such as signalling, regulation and homeostasis [1, 2, 3]. Moreover, malfunction of IDPs has been linked to a large proportion of cancers and neurodegenerative and cardiovascular diseases [4]. IDPs perform highly specialized functions despite they are devoid of permanent secondary or tertiary structure. Indeed, their malleability enables biological tasks that are out of reach for their globular counterparts [5]. In most cases, function is manifested when these flexible proteins interact with globular partners to trigger signaling or metabolic cascades [6]. These interactions are normally of low or moderate affinity, giving rise to fuzzy complexes where the IDP remains flexible upon binding [7, 8]. These interactions are often mediated by Short Linear Motifs (SLiMs) or Molecular Recognition Elements (MoREs) that specifically recognize the surface of the partner [9, 10, 11, 12]. The presence of partially-structured elements in SLiMs tunes the thermodynamics and kinetics of the interaction, often assisted by their flanking regions [13]. Structural and electrostatic changes induced by post-translational modifications can also modulate the affinity of the interaction and represent efficient mechanisms of regulation [14, 15].

The identification and characterization of partially-structured elements in IDPs is complex and requires extensive experimental work, mainly using Nuclear Magnetic Resonance (NMR). In particular, NMR Chemical Shifts (CSs) and Residual Dipolar Couplings (RDCs) are sensitive to small populations of secondary structural elements [16, 17, 18]. Computational tools represent a good complement or an alternative to experimental studies to localize such structurally biased elements. For over 40 years, numerous methods have been developed to predict secondary structure in proteins from their amino acid sequence (see for instance [19]). However, current secondary structure predictors are in general trained and evaluated on folded/globular proteins, and thus are

not necessarily appropriate to identify partially structured regions in IDPs. Numerous methods have also been proposed to predict structural disorder from protein sequence (see [20, 21] and references therein). Most of the available disorder predictors focus on the identification of disordered regions in predominantly folded proteins. In general, they only provide a binary output (i.e. ordered/disordered) or a residue-specific disorder probability, but do not identify structural classes. Since they aim at providing different information, traditionally, secondary structure and disorder predictors have been developed independently from each other. One exception is the s2D method [22], which predicts secondary structure populations and disorder in a unified framework. s2D, as the work presented here, relies on a more holistic view of IDPs by exploring structural descriptors that span the continuum between ordered and disordered proteins [23, 24, 12].

In contrast to the most recent approaches, which are based on intricate machine-learning techniques, here we present an extremely simple strategy to identify secondary structural propensities from protein sequences. As machine-learning-based approaches, our method exploits structural information contained in databases. However, instead of training a machine-learning model or architecture, our approach performs simple statistical operations. These operations are based on a classification of the conformational preferences of three-residue fragments extracted from coil regions of experimentally determined high-resolution protein structures. Although small, tripeptides have been shown to encode relevant sequence-dependent structural information [25], and are valuable building-blocks to model unfolded states and disordered proteins or regions [26, 27, 28]. Furthermore, statistical analyses of three-residue fragments have also been used as key components of knowledge-based potentials and protein fold recognition methods [29, 30].

We have evaluated the performance of our method, called Local Structural Propensity Predictor (LS2P), using a benchmark of nine well-characterized IDPs. LS2P accurately predicts previously identified helical and extended regions in the benchmark. Moreover, small stretches forming β-turns or pro-

moting α-helices emerge from the analysis of the preferred structural classes of the tripeptides within the local sequence context. The main advantage of our strategy with respect to most machine-learning-based methods for secondary structure prediction, especially those using neural networks, is that it enables a comprehensible connection between amino acid sequence and structural preferences. LS2P is publicly available through a web server at: `https://moma.laas.fr/applications/LS2P`.

2. Theory

The prediction method proposed in this work, LS2P, exploits statistical information about the structural preferences of three-residue fragments, called *tripeptides* from now on. This information was extracted from a structural database constructed from coil regions in high-resolution protein structures. Details about the tripeptide database construction can be found in the Materials and Methods section.

To simplify the structural classification, the conformational space of each residue r_i was subdivided according to the values of the Ramachandran angles, ϕ and ψ, into three regions $S = \{\alpha, \beta, \gamma\}$. These regions, represented in Figure 1, are defined as follows [31]:

$$\alpha : -180^\circ < \phi \leq 0^\circ \ , \ -120^\circ < \psi \leq 50^\circ$$
$$\beta : -180^\circ < \phi \leq 0^\circ \ , \ 50^\circ < \psi \leq 240^\circ$$
$$\gamma : 0^\circ < \phi \leq 180^\circ \ .$$

For a given tripeptide, combining these three structural classes at the single residue level leads to 27 structural classes $\mathscr{S}$: $\alpha\alpha\alpha, \alpha\alpha\beta, \alpha\alpha\gamma, \alpha\beta\alpha, \ldots, \gamma\gamma\gamma$. The number of conformations per class was retrieved from the database and stored for each of the 8,000 tripeptide types. These numbers are used by the LS2P predictor as explained below.

For each residue r_i in the sequence, the secondary structure propensity is calculated using statistical information for the tripeptide t_i centered at this

residue and for its neighbors: t_{i-2}, t_{i-1}, t_{i+1} and t_{i+2}. Let n^i denote the total number of structures present in the tripeptide database for t_i. The number of structures for each one of the 27 structural classes is indicated using the corresponding Greek letters in subscript. For instance, $n^i_{\beta\gamma\alpha}$ is the number of structures of t_i with the first residue of the tripeptide in the β region, the second in γ and the third in α. We use lower-case Latin letters, for instance x or y, as variables when the three structural classes have to be considered for one or several residues. This notation is used below within summation equations.

For a tripeptide t_i, independently of the rest of the sequence, the number of structures present in each of the 27 structural classes with respect to the total number of structures already gives an idea of its conformational preferences. For example, for the particular case $\mathscr{S} = \beta\gamma\alpha$, and considering t_i independently of the rest of the sequence:

$$p(\beta\gamma\alpha)_i = \frac{n^i_{\beta\gamma\alpha}}{\sum\limits_{w,x,y\in S} n^i_{wxy}} \tag{1}$$

However, in order to better take into account the sequence context, the compatibility of the structural preferences of t_i with those of the neighboring tripeptides has to be considered. This is illustrated in Figure 1. In this particular case, the probability of t_i to adopt a $\beta\gamma\alpha$ conformation depends on the probability of the last two residues of t_{i-1} to adopt a $\beta\gamma$ conformation, of the first two residues of t_{i+1} to adopt a $\gamma\alpha$ conformation, of the last residue of t_{i-2} to adopt a β conformation, and of the first residue of t_{i+2} to adopt a α conformation. The structural preferences conditioned by the neighbors can be easily computed operating with the numbers of structures in the tripeptide database. For the example of $\mathscr{S} = \beta\gamma\alpha$, the equation can be written as:

$$p(\beta\gamma\alpha)_i = \frac{\sum\limits_{t,u,y,z\in S} n^{i-2}_{tu\beta}\; n^{i-1}_{u\beta\gamma}\; n^{i}_{\beta\gamma\alpha}\; n^{i+1}_{\gamma\alpha y}\; n^{i+2}_{\alpha yz}}{\sum\limits_{t,u,v,w,x,y,z\in S} \left(n^{i-2}_{tuv}\; n^{i-1}_{uvw}\; n^{i}_{vwx}\; n^{i+1}_{wxy}\; n^{i+2}_{xyz}\right)} \tag{2}$$

To compute the propensity of tripeptide t_i to adopt a particular structural class e.g. $\mathscr{S} = \beta\gamma\alpha$ with respect to the observations in our database, $p(\beta\gamma\alpha)_i$ is

divided by the overall probability to observe this structural class in all tripeptides:

$$p(\beta\gamma\alpha)_{\text{all}} = \frac{n^{\text{all}}_{\beta\gamma\alpha}}{N} \tag{3}$$

where "all" implies the sum for the 8,000 tripeptide sequences, and N is the total number of tripeptide structures in the database. Thus, the structural propensity can be written as:

$$P(\beta\gamma\alpha)_i = \frac{p(\beta\gamma\alpha)_i}{p(\beta\gamma\alpha)_{\text{all}}} \tag{4}$$

Note that $P(\mathscr{S})_i$ values do not correspond to the estimations of the population for the structural classes of the tripeptides found in the protein. They are an indicator of the structural propensities along the IDP sequence. Values larger than 1.0 for a given structural class indicate that this class is favored for a given tripeptide in the local sequence context, while values below 1.0 indicate the unlikelihood of this class.

3. Results

3.1. Identification of secondary structure propensities in IDPs: An overall picture

A benchmark set of nine structurally well-characterized IDPs were used to evaluate the performance of our approach. Concretely, MAPK Kinase 7 (MKK7) [32], the fragment 945-1097 of the Erythrocyte binding antigen 181 (EBA-181) [33], p15 [34], Sic1 [14], Measles virus ntail (ntailMV) [35], Sendai virus ntail (ntailSV) [36], the unique domain of the src kinase (USrc) [37], K18 construct of Tau protein (K18) [38], and full-length Tau protein [39] were used in our study. Predictions of secondary structure propensities by LS2P were compared to the NMR RDCs, which are extremely sensitive to conformational preferences at the residue level [17]. Previous structural analyses of these nine proteins were also considered for this evaluation. In addition, we applied five commonly used disorder predictors: DisEMBL-coils [40], DISOPRED-3 [41], IUPred2A [42], PONDR-VLTX [43], and SPOT-Disorder2 [44]. The obtained

disorder probability profiles are presented in SI (panel (a) in Figures S3 to S11). Overall, they agree with experimental studies, showing a high level of disorder for the proteins in our benchmark set. Discrepancies between the different predictors highlight the difficulty to structurally characterize these nine proteins using computational methods.

First, we analyzed the number of structures in our database for all the tripeptides of the benchmark (see Figure S1 for details). The average number of structures per tripeptide ranges from 637 to 792 for Sic1 and Tau, respectively. The minimum number of structures found for a tripeptide type is 22, which corresponds to the tripeptide ^{942}Met-His-Met^{944} in EBA-181, while the tripeptide Gly-Gly-Gly, which appears in several proteins of the benchmark, is the most represented tripeptide with 2,560 structures. These observations indicate that we have sufficient samples for the vast majority of the tripeptides, so that reliable statistics can be retrieved from the analysis. We also performed a comparative analysis of the number of structures for each of the 27 structural classes of tripeptides found in the nine IDPs (see Figure S2 for details). This analysis shows that, although the tripeptides are extracted from protein fragments excluding α-helices and β-strands according to the DSSP classification [45], a large proportion of the tripeptides in the database adopt fully helical or extended conformations (according to our classification based on ϕ and ψ dihedral angles), with a similar percentage around 20%. Nevertheless, when we compare the proportion of the tripeptide sequences in our benchmark with the overall proportion (i.e. computed from the 8,000 tripeptide sequences), one can observe that, in general, $\alpha\alpha\alpha$ and $\beta\beta\beta$ structures are relatively less frequent for the tripeptides found in our IDPs, while "rare" structural classes, such as $\gamma\gamma\gamma$, are statistically more frequent. This highlights the specific amino acid sequences found in IDP that, in turn, define their structural preferences.

In order to illustrate the application of LS2P, results for two representative cases, MKK7 and EBA-181, are presented in more detail here (Figure 2), while results for the other proteins are shown in SI (Figures S5 to S11). From a structural point of view, MKK7 and EBA-181 present very different features. While

MKK7 involves relatively long regions with helical or extended propensities, EBA-181 is almost fully disordered, only presenting short partially-structured fragments. Note that disorder predictors also provide significantly different results for both proteins (see Figures S3.a and S4.a). Whereas most of the predictors agree on a relatively low disorder probability at the N-terminus and C-terminus of MKK7, the consensus is less clear for EBA-181, and three over five predictors return a very high degree of disorder for this protein.

The N-terminus of MKK7 (residues 5-30) presents an α-helical structure that is characterized by the positive values of the RDC profile. The α-helical propensity is well predicted by the LS2P method. Then, LS2P predicts the region starting at residue 26 to be highly extended, in line with their negative RDC values. The rest of the sequence appears, according to LS2P, as preferentially extended, although some α-helical propensity is observed at the C-terminus. Moreover, some MKK7 stretches around residues 54, 65 and 80 are dominated by less abundant structures involving γ-type conformations.

The MKK7 fragment analyzed involves three MAPK binding domains that have been structurally characterized by NMR: D1 (residues 25-34), D2 (residues 38-47) and D3 (residues 70-79) [32]. Our secondary structure prediction is in very good agreement with the structural conformation found for these motifs. D1 lies in the transition between helical and extended conformations at the N-terminus of MKK7, and this dual behavior was captured by the ensemble refinement done in the original study. D2, which is inserted in the long extended region of MKK7 according to LS2P, was experimentally shown to sample β-strand and polyproline-II (PPII) conformations. Conversely, predictions of the D3 indicate that this region has no special enrichment neither in helical nor in extended conformations, with the exception of residues 73 and 74, in line with the original experimentally-derived ensemble model.

Structural investigations of EBA-181 have shown that the fragment involving residues 945-1097, which is part of the RIII–V region, behaves essentially as a random coil with the presence of several turn motifs or short single-turn α-helices [33]. These short helical elements, corresponding to positive RDCs

around residues 987–988, 998, 1006–1007 and 1016–1019, are correctly identified by LS2P (Figure 2). Note that the identification of turns will be described in more detail below. Other short regions present some propensity to adopt extended conformations, in particular regions around prolines P945, P949, P1003, P1039, P1040 and P1044. These short extended regions are also well predicted by LS2P. When analyzing the enrichment of the 27 structural groups, we observe that most of them are present along the sequence and only regions around 958, 1050 and the C-terminus seem to be highly enriched in less common structural classes (Figure S4). As mentioned above and illustrated in Figure S2, sequences allowing more heterogeneous conformations seem to be an indicator of disorder and absence of secondary structural elements.

These results for MKK7 and EBA-181, which showcase two structurally diverse types of IDPs, demonstrate the performance of LS2P. The following sections will describe more specifically the ability of LS2P to identify different types of secondary structural elements within IDPs.

3.2. Identification of α-helical elements in IDPs

In addition to the previously described examples, our benchmark contains other examples of IDPs involving relatively long fragments with helical propensity. The two most prominent ones are ntailMV and ntailSV. These two proteins have similar sequences and perform the same function by interacting with the phosphoprotein in two related viruses through a highly stable α-helix [46].

LS2P identifies several regions displaying an enrichment in helical conformations in both N-tail proteins, including the experimentally characterized functional α-helix. Figure 3 shows the experimental RDC profiles and the results of the LS2P predictor of this region for both proteins. When comparing with s2D predictions (note that the comparison with s2D is discussed in detail in a different section below), we observe different levels of agreement (see Figures S7 and S8). In ntailSV, both algorithms identify four α-helices and, interestingly, two of these regions (around residues 450 and 515) display positive RDCs, suggesting the presence of helical populations in solution. Conversely, only the

functional helix is identified by s2D for ntailMV. The most surprising result is that our approach predicts that the two functional helices, especially ntailSV, contain a non-negligible proportion of extended conformations in the middle of the functional α-helix. This observation is in contrast with the experimental data [36, 35] and the predictions done with s2D. This contradictory observation underlines a fundamental difference between both methods. While s2D was trained using data from NMR experiments and captures propensities in longer protein stretches, LS2P is only based on local conformational bias. Despite this potential limitation, LS2P was able to identify the long α-helix at the N-terminus of MKK7 (Figure 2), suggesting that N-tail helices present some specific features. It has been shown that the functional helices of both N-tail proteins are highly stabilized by N-capping serine and aspartic acid residues placed upstream of the helix [36, 35]. The inspection of the conformational propensities in these regions identifies several residues with a strong propensity for $\beta\alpha\alpha$ and $\beta\beta\alpha$ structural classes. Concretely, tripeptides centered at residues 485, 488 and 491 in ntailMV, and 473, 474 and 479 in ntailSV display a strong enrichment in these conformational classes. We speculate that this structural feature, which is identified by LS2P, promotes and stabilizes helical conformations in both N-tail proteins.

3.3. Identification of extended regions in IDPs

Several regions are identified as extended ($\beta\beta\beta$) in the analysis of the benchmark set. Note that the current implementation of LS2P does not discriminate between β-strand-type and PPII-type conformations, both of them being classified as "extended". Note also that the possible presence of hydrogen bonds to stabilize parallel or anti-parallel β-strands is not considered as this constitutes an uncommon situation in IDPs. Instead, extended regions are identified only on the basis of local structural preferences along the amino acid sequence.

Protein Tau is a good example to illustrate the ability of LS2P to predict the propensity of some regions within IDPs to adopt extended conformations. The method identifies extended regions described in previous studies [38, 39]: at the

N-terminal region of Tau (around residue 50, in particular), within the proline-rich region (residues 212-232), and inside the pseudorepeat domains contained in the K18 fragment (residues 275-282, 307-313 and 338-346, approximately). All the regions correspond to negative RDCs (see Figures 4, S10 and S11). LS2P also identifies extended regions in other proteins, such as p15 [34] (Figure S5) and Sic1 [14, 47] (Figure S6), in good agreement with the original studies and the RDC profiles.

Relying only on the local sequence, there are two main factors that induce extended conformations. One of them is the presence of prolines, which enriches neighboring residues in extended conformations, as it is the case for the proline-rich region in Tau or the short extended regions in EBA-181 (see above). Amino acid bulkiness is another property that has a strong effect on the conformational preferences of neighboring residues [48]. Amino acids with large side chains enrich extended conformations in neighboring residues as a conformational mechanism to avoid steric clashes. To illustrate the importance of amino acid bulkiness in the identification of extended conformations, we computed the bulkiness profile for the proteins in the benchmark set, using averaged values over a five-residue window as proposed in [49]. Note that in our case we did not increase the theoretical volume of prolines. Figure 4 shows experimental RDCs, the predicted extended propensity and the bulkiness profile for the K18 construct of Tau. We observe a correlation between the regions having highly negative RDCs, displaying an enhanced population of $\beta\beta\beta$ propensity, and large bulkiness. This correlation suggests that LS2P properly identifies regions with extended conformations and that our statistical approach, despite its local nature, captures the steric influence exerted by flanking bulky residues.

3.4. Identification of β-turns in IDPs

β-turns represent the third most abundant secondary structure in proteins [50, 51]. Based on a standard definition, β-turns are constituted of four consecutive residues, with a distance between the Cα atoms of the first and the fourth residues smaller than 7 Å. Different β-turn types can be defined based on the ϕ

and ψ values adopted by the two central residues, which are normally derived from high-resolution crystallographic structures [51]. However, the coarse subdivision of the conformational space used by LS2P hampers the possibility to precisely discriminate among all these β-turn types. Moreover, in IDPs, turns are only partially formed, which complicates their identification and classification. Experimentally, turns can be identified based on RDCs, which display anomalous values with respect to the neighboring residues [38, 33, 39]. Based on this fuzzy description, β-turns identified in IDPs were characterized by two consecutive residues with nearly helical conformations, preceded and succeeded by residues with more extended conformations. Such a description broadly fits the definition of type I and some type IV β-turn sub-types [51].

Following this definition, turns could be predicted from the results of LS2P by identifying consecutive (overlapping) tripeptides with high propensities for $\beta\alpha\alpha$ and $\alpha\alpha\beta$ structural classes. As shown in Figure 5.a, this concatenation of classes is found by LS2P for the well-characterized turns in the K18 construct of Tau [38], involving residues 252-255, 283-286, 314-317, and 345-348. In addition to the aforementioned signature $\beta\alpha\alpha$-$\alpha\alpha\beta$ for the two middle residues, these β-turns can present a $\alpha\alpha\alpha$ peak for the second of these middle residues, which can be higher than the $\alpha\alpha\beta$ propensity (see turn IV of K18 in Figure 5.a). In such cases, the $\alpha\alpha\beta$ propensity increases for the next residues in the sequence. For the four turns in K18, the extended-helical-extended transition is reinforced by the high propensity for $\beta\beta\alpha$ and $\alpha\beta\beta$ structural classes for the N- and C-flanking tripeptides, respectively (Figure 5.a). Therefore, LS2P results suggest that the concatenation of specific structural classes is a good indication of the presence of stable turns in IDPs.

The concatenation of $\beta\beta\alpha$, $\beta\alpha\alpha$, $\alpha\alpha\beta$ and $\alpha\beta\beta$ propensities is also found in the four turns described for EBA-181 (see Figure 5.b). Interestingly, LS2P rationalizes the differences between these turns found in a previous study [27]. While the first two turns (around residues 987 and 998) follow the above-described concatenation of classes, the last two turns (around residues 1006 and 1017) present a short segment enriched in $\alpha\alpha\alpha$. As a consequence of this structural

difference, these two last turns present more positive RDCs. Note that the $\beta\alpha\alpha$-$\alpha\alpha\beta$ propensity for pairs of consecutive residues is also high in other regions of EBA-181, such as residues 971-972 and 1031-1032 (see Figure S8). Although less intense than the previously described turns, these two regions also display specific features in the RDC profile.

3.5. Comparison with state-of-the-art methods for structural propensity prediction

As mentioned in the introduction, the vast majority of secondary structure predictors aims at identifying structural elements within globular proteins. These methods usually fail to recognize partially-structured regions in IDPs, especially when the structural propensity is relatively low [22, 23]. On the other hand, disorder predictors, which aim at identifying regions lacking secondary structure, do not provide information about structural propensities at the frontier between order and disorder. A remarkable exception is the s2D method [22], which was especially conceived to simultaneously predict secondary structure and disorder propensities, and which is particularly well suited to the structural investigation of IDPs. Here, we compare the performance of s2D and LS2P to predict secondary structure propensities for the nine proteins considered in this work.

LS2P and s2D agree in many cases, particularly when the structural elements are known to have relatively high propensity to be formed in solution. This is the case for instance for the helical region at the N-terminus of MKK7 and for the helical regions in ntailSV and ntailMV (see Figures S3, S7 and S8). As mentioned before, s2D performs better than LS2P in some of these cases due to the underlying principles of the method, which operates in terms of longer sequence fragments. Both methods also agree on the prediction of the extended regions in K18 (see Figure S10). In all these cases, disorder predictors tend to provide a relatively low disorder probability.

However, s2D generally fails to identify transient secondary structure in several cases in which LS2P successfully provides this information. This is for

instance the case for p15, illustrated in Figure 6 (see also Figure S5). Based on the negative RDC values observed for the 15–24 and 94–104 segments in p15, [34] suggested a low population (about 8%) of β-strand conformations. While LS2D clearly identifies these two extended regions, s2D predictions are unclear. Sic1, which has been shown to concatenate regions with significant propensity to adopt extended or helical conformations [14, 47], is another example of better performance of LS2P (Figure S6). In this case, s2D does not identify any secondary structural preference in the protein. s2D also fails to identify small structural motifs such as turns or short helices, whereas LS2P is able to find them, as it has been illustrated for EBA-181 and K18 (Figures S4 and S10).

Another difference between LS2P and s2D concerns the effect of the overall sequence in the prediction. Whereas the simple principle implemented in LS2P only operates in terms of local structural preferences, s2D considers mean secondary structure propensities for the entire protein. Although this makes sense within the machine-learning approach implemented in s2D and can lead to improved predictions in some cases, it also can produce unreliable results in other cases. For instance, the results provided by s2D for the K18 construct of Tau are very different than these for the same region within the full-length protein (see Figures S10 and S11). Conversely, LS2P provides the same results in both cases due to its local-sequence focus.

4. Discussion

In this work, we have investigated the ability to predict secondary structure propensities within IDPs using local sequence-dependent information encoded in small protein fragments extracted from coil regions in high-resolution protein structures. We have developed an extremely simple statistical approach based on a coarse classification of tripeptide conformations. In contrast with nowadays popular neural-network-based secondary structure predictors, this approach enables a comprehensible connection between sequence and structural propensities. Moreover, thanks to this simplicity, the proposed predictor LS2P is

very computationally inexpensive, enabling the fast scanning of large databases or complete proteomes.

Results presented show that LS2P is able to predict in a robust manner the main secondary structural elements: α-helices and extended conformations. These are detected regardless of the length of the secondary structural element, even though the method operates from structural preferences of three-residue fragments. This highlights the importance of the local sequence context, which implicitly encodes the cooperative formation of structural elements along the polypeptide chain. This is a clear advantage with respect to state-of-the-art secondary structure predictors, including those suited to IDPs, such as s2D, which mainly identify relatively long and highly populated secondary structure elements. In addition to the detection of the canonical secondary structures, LS2P enables the identification of short structural elements. For instance, conformational classes inducing helical N-capping have been identified in the N-tail proteins. Concretely, classes $\beta\beta\alpha$ and $\beta\alpha\alpha$ are enriched in residues preceding α-helical regions that become more stable than predicted by LS2P.

Another unique feature of our approach is the detection of certain types of β-turns [51]. Provided with the correct definition of turn-types found in literature, we can connect amino acid sequence and the structural classes predicted by LS2P with the presence of β-turns. However, the coarse description of the Ramachandran space used by LS2P precludes the discrimination between certain β-turn types [51].

Despite the good overall performance of the method, it should be noted that LS2P may predict structural propensity in some regions for which there is no experimental evidence of secondary structure (i.e. false positives), and it may also fail to predict structural propensities in a few cases (i.e. false negatives). Nevertheless, it is hard to assess the degree of (in)accuracy of LS2P due to the difficulty to precisely characterize structural properties of IDPs from experiments. As an example of false positive prediction, LS2P predicts helical propensity at the C-terminal region (residues 95-99) of MMK7, whereas the RDC profile does not indicate such a structural propensity. Note that s2D also

predicts this helical region (as shown in Figure S3) and that disorder profiles obtained by most of the predictors are relatively low in this region. However, it is well known that RDCs are much less informative at the sequence termini [52], and thus it is not possible to discard that this helical propensity exists in reality. An example of false negative prediction is the low-populated helical structure involving residues 60-75 in USrc, which has been characterized by NMR experiments [37]. Interestingly, s2D also fails to predict this helical region and disorder predictors provide discrepant results, as shown in Figure S9. A possible explanation of this under-performance of both predictors is that transient structural elements in IDPs are not necessarily canonical secondary structure elements, and sometimes are a concatenation of small partially-stable elements. In these circumstances, the local focus of LS2P could be advantageous with respect to methods that operate in terms of longer sequence fragments, such as s2D. Actually, although LS2P does not identify helical propensity around residues 60-75 in USrc, it predicts a concatenation of $\beta\gamma\gamma$ and $\gamma\gamma\beta$ propensity at residues 65 and 66, respectively. This approximately fits the description of type I' β-turns [51], and could explain the positive RDC profile in this region. Note that these two consecutive residues are glycines, which are frequently found at the central positions of β-turns [50].

Inaccurate predictions of LS2P in some regions can be due to biases or lack of information in the tripeptide database. Indeed, in the same way that machine-learning-based methods strongly rely on the data-set used for training, results provided by LS2D depend on the quality of the tripeptide database. Our current database was constructed from coil regions in a large set of protein structures, mostly determined by X-ray crystallography. The available information in this database can be inaccurate (e.g. due to biases induced by experimental conditions) or limited for sequences that are seldom observed in globular proteins but that may appear in IDPs. With the growing number of high-resolution structures and NMR data-sets deposited in specific repositories, we expect to enrich our database and achieve more robust predictions in the future. A more extensive structural database would also enable to further refine the structural

classes with respect to the three classes per residue α, β, γ considered in this work. In particular, it would be interesting to discriminate between β-strand and PPII conformations, which have been reported to be very common in IDPs [53, 54].

In summary, we presented a novel method to identify partially structured regions in IDPs, which we have made available through a web server. Although the structural analysis of IDP sequences still represent a challenge for current algorithms, LS2P presents some advantageous features. The most important of them is the simplicity of the statistical approach used to compute local structural propensities, enabling an easy connection between sequence and conformational properties at the residue level. This new tool paves the way to systematic studies of large IDP data-sets in order to better understand the connection between structural changes and functional effects induced by point mutations.

5. Materials and Methods

5.1. Tripeptide database

The tripeptide database was built from a curated database of high-resolution experimentally determined protein structures. More precisely, we used protein domains from the SCOPe [55] 2.06 release. In order to remove highly-redundant sequences, we used the 95% sequence-identity-filtered subset of these domains. This subset consists of PDB-style files for 28,011 domains. DSSP [45] was employed to assign secondary structure labels to each residue in these files.

Each structure file was processed by passing a sliding window of size 3 along the amino acid sequence. Each resulting tripeptide was added to the database if none of its 3 residues had a DSSP code of H, I or E. In other words, none of the residues of the tripeptide participates in a α-helix, π-helix or β-strand. An additional treatment was applied when the provided domain structure file originated from NMR data. For each structural file that contained more than one model, a distance filter was applied to corresponding tripeptides in each model to avoid redundancy in the database. A tripeptide structure was consid-

ered sufficiently distant from another one already extracted from the same file, and was thus added to the database, if it met at least one of the two following criteria: the RMSD on ω, ϕ and ψ angles is above 0.2 radians, or one of the nine dihedral angles differs by more than 0.6 radians. In total, 2,972,319 tripeptides were extracted. The tripeptide backbone dihedral angles were collected in the database and indexed by its corresponding amino acid sequence (i.e. 8,000 tripeptide classes). These values were used for the classification into the 27 structural classes.

5.2. LS2P

The principle of the method is explained in the Theory section of the manuscript. The code implementing this method is freely available (see below).

5.3. s2D

The s2D method was used through the dedicated web server:
`http://www-mvsoftware.ch.cam.ac.uk`.
Only the protein sequence is required as input. The server provides a file with the populations of α-helix, β-strand and random coil for each residue. This information was used to generate the plots presented in this study. Similar plots can be directly obtained from the s2D web server.

Availability

LS2P is publicly available through a web server at:
`https://moma.laas.fr/applications/LS2P`.
The code of LS2P (in Python) and the data (number of structures for each tripeptide type and structural class extracted from high-resolution experimentally determined protein structures) are available upon request to the Lead Contact.

CRediT author statement

Alejandro Estaña: Methodology, Data curation, Software, Writing - Original Draft. **Amélie Barozet**: Methodology, Writing - Review & Editing. **Assia Mouhand**: Investigation, Data curation. **Marc Vaisset**: Data curation, Software. **Christophe Zanon**: Software. **Pierre Fauret**: Software. **Nathalie Sibille**: Investigation, Writing - Review & Editing. **Pau Bernadó**: Conceptualization, Investigation, Supervision, Writing - Original Draft, Review & Editing. **Juan Cortés**: Conceptualization, Methodology, Software, Supervision, Writing - Original Draft, Review & Editing.

Acknowledgments

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 France-BioImaging (FBI) and the French Infrastructure for Integrated Structural Biology (FRISBI), 2 national infrastructures supported by the French National Research Agency (ANR-10-INBS-04-01 and ANR-10-INBS-05, respectively).

Declaration of Interests

The authors declare no conflict of interest.

References

[1] V. Csizmok, A. V. Follis, R. W. Kriwacki, J. D. Forman-Kay, Dynamic protein interaction networks and new structural paradigms in signaling, Chem. Rev. 116 (11) (2016) 6424–6462.

[2] M. M. Babu, R. van der Lee, N. S. de Groot, J. Gsponer, Intrinsically disordered proteins: Regulation and disease, Curr. Opin. Struct. Biol. 21 (3) (2011) 432 – 440.

[3] P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation, Nat. Rev. Mol. Cell Biol. 16 (2015) 18–29.

[4] V. N. Uversky, C. J. Oldfield, A. K. Dunker, Intrinsically disordered proteins in human diseases: Introducing the D2 concept, Ann. Rev. Biophys. 37 (1) (2008) 215–246.

[5] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, Z. Obradovic, Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, J. Proteome Res. 6 (5) (2007) 1882–1898.

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

[7] M. Fuxreiter, Fuzziness: linking regulation to protein dynamics, Mol. BioSyst. 8 (2012) 168–177.

[8] T. N. Cordeiro, N. Sibille, P. Germain, P. Barthe, A. Boulahtouf, F. Allemand, R. Bailly, V. Vivat, C. Ebel, A. Barducci, W. Bourguet, A. le Maire, P. Bernadó, Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression, Structure 27 (8) (2019) 1270 – 1285.e6.

[9] A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, V. N. Uversky, Analysis of molecular recognition features (MoRFs), J. Mol. Biol. 362 (5) (2006) 1043 – 1059.

[10] K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, N. E. Davey, Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation, Chem. Rev. 114 (13) (2014) 6733–6778.

[11] R. Pancsa, M. Fuxreiter, Interactions via intrinsically disordered regions: What kind of motifs?, IUBMB Life 64 (6) (2012) 513–520.

[12] N. E. Davey, The functional importance of structure in unstructured protein regions, Curr. Opin. Struct. Biol. 56 (2019) 155 – 163.

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

[14] T. Mittag, S. Orlicky, W.-Y. Choy, X. Tang, H. Lin, F. Sicheri, L. E. Kay, M. Tyers, J. D. Forman-Kay, Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor, Proc. Natl. Acad. Sci. USA 105 (46) (2008) 17772–17777.

[15] A. Bah, J. D. Forman-Kay, Modulation of intrinsically disordered protein function by post-translational modifications, J. Biol. Chem. 291 (13) (2016) 6696–6705.

[16] H. J. Dyson, P. E. Wright, Unfolded proteins and protein folding studied by NMR, Chem. Rev. 104 (8) (2004) 3607–3622.

[17] M. R. Jensen, P. R. Markwick, S. Meier, C. Griesinger, M. Zweckstetter, S. Grzesiek, P. Bernadó, M. Blackledge, Quantitative determination of the conformational properties of partially folded and intrinsically disordered proteins using NMR dipolar couplings, Structure 17 (9) (2009) 1169 – 1185.

[18] S. Milles, N. Salvi, M. Blackledge, M. R. Jensen, Characterization of intrinsically disordered proteins and their dynamic complexes: From in vitro to cell-like environments, Prog. Nucl. Magn. Reson. Spectrosc. 109 (2018) 79 – 100.

[19] Q. Jiang, X. Jin, S.-J. Lee, S. Yao, Protein secondary structure prediction: A survey of the state of the art, J. Mol. Graph. Model. 76 (2017) 379–402.

[20] Y. Liu, X. Wang, B. Liu, A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction, Brief. Bioinform. 20 (1) (2017) 330–346.

[21] J. T. Nielsen, F. A. A. Mulder, Quality and bias of protein disorder predictors, Sci. Rep. 9 (2019) 5137.

[22] P. Sormanni, C. Camilloni, P. Fariselli, M. Vendruscolo, The s2D method: Simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins, J. Mol. Biol. 427 (4) (2015) 982–996.

[23] P. Sormanni, D. Piovesan, G. T. Heller, M. Bonomi, P. Kukic, C. Camilloni, M. Fuxreiter, Z. Dosztanyi, R. V. Pappu, M. M. Babu, S. Longhi, P. Tompa, A. Dunker, V. N. Uversky, S. C. E. Tosatto, M. Vendruscolo, Simultaneous quantification of protein order and disorder, Nat. Chem. Biol. 13 (4) (2017) 339–342.

[24] S. DeForte, V. N. Uversky, Order, disorder, and everything in between, Molecules 21 (8) (2016) 1090.

[25] J.-R. Huang, V. Ozenne, M. R. Jensen, M. Blackledge, Direct prediction of NMR residual dipolar couplings from the primary sequence of unfolded proteins, Angew. Chem. Int. Edit. 52 (2) (2013) 687–690.

[26] A. K. Jha, A. Colubri, K. F. Freed, T. R. Sosnick, Statistical coil model of the unfolded state: Resolving the reconciliation problem, Proc. Natl. Acad. Sci. USA 102 (37) (2005) 13099–13104.

[27] A. Estaña, N. Sibille, E. Delaforge, M. Vaisset, J. Cortés, P. Bernadó, Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database, Structure 27 (2) (2019) 381–391.e2.

[28] A. Barozet, K. Molloy, M. Vaisset, T. Siméon, J. Cortés, A reinforcement-learning-based approach to enhance exhaustive protein loop sampling, Bioinformatics 36 (4) (2020) 1099–1106.

[29] M. R. Betancourt, Knowledge-based potential for the polypeptide backbone, J. Phys. Chem. B 112 (16) (2008) 5058–5069.

[30] A. D. Solis, Deriving high-resolution protein backbone structure propensities from all crystal data using the information maximization device, PloS one 9 (6) (2014) e94334.

[31] V. Ozenne, R. Schneider, M. Yao, J.-R. Huang, L. Salmon, M. Zweckstetter, M. R. Jensen, M. Blackledge, Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution, J. Am. Chem. Soc. 134 (36) (2012) 15138–15148.

[32] J. Kragelj, A. Palencia, M. H. Nanao, D. Maurin, G. Bouvignies, M. Blackledge, M. R. Jensen, Structure and dynamics of the MKK7-JNK signaling complex, Proc. Natl. Acad. Sci. USA 112 (11) (2015) 3409–3414.

[33] M. Blanc, T. L. Coetzer, M. Blackledge, M. Haertlein, E. P. Mitchell, V. T. Forsyth, M. R. Jensen, Intrinsic disorder within the erythrocyte binding-like proteins from plasmodium falciparum, Biochim. Biophys. Acta 1844 (12) (2014) 2306 – 2314.

[34] A. De Biasio, A. Ibáñez de Opakua, T. N. Cordeiro, M. Villate, N. Merino, N. Sibille, M. Lelli, T. Diercks, P. Bernadó, F. J. Blanco, p15PAF is an intrinsically disordered protein with nonrandom structural preferences at sites of interaction with other proteins, Biophys. J. 106 (4) (2014) 865 – 874.

[35] M. R. Jensen, G. Communie, E. A. Ribeiro, N. Martinez, A. Desfosses, L. Salmon, L. Mollica, F. Gabel, M. Jamin, S. Longhi, R. W. H. Ruigrok, M. Blackledge, Intrinsic disorder in measles virus nucleocapsids, Proc. Natl. Acad. Sci. USA 108 (24) (2011) 9839–9844.

[36] M. R. Jensen, K. Houben, E. Lescop, L. Blanchard, R. W. H. Ruigrok, M. Blackledge, 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 (25) (2008) 8055–8061.

[37] Y. Pérez, M. Gairí, M. Pons, P. Bernadó, 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 (1) (2009) 136 – 148.

[38] M. D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernadó, C. Griesinger, E. Mandelkow, M. Zweckstetter, M. Blackledge, Highly populated turn conformations in natively unfolded Tau protein identified from residual dipolar couplings and molecular simulation, J. Am. Chem. Soc. 129 (16) (2007) 5235–5243.

[39] M. Schwalbe, V. Ozenne, S. Bibow, M. Jaremko, L. Jaremko, M. Gajda, M. Jensen, J. Biernat, S. Becker, E. Mandelkow, M. Zweckstetter, M. Blackledge, Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering, Structure 22 (2) (2014) 238 – 249.

[40] R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson, R. B. Russell, Protein disorder prediction: implications for structural proteomics, Structure 11 (11) (2003) 1453–1459.

[41] D. T. Jones, D. Cozzetto, DISOPRED3: precise disordered region predictions with annotated protein-binding activity, Bioinformatics 31 (6) (2015) 857–863.

[42] B. Mészáros, G. Erdős, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucl. Acids Res. 46 (W1) (2018) W329–W337.

[43] Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker, Predicting intrinsic disorder from amino acid sequence, Proteins 53 (Suppl 6) (2003) 566–572.

[44] J. Hanson, K. K. Paliwal, T. Litfin, Y. Zhou, SPOT-Disorder2: Improved

protein intrinsic disorder prediction by ensembled deep learning, Genom. Proteom. Bioinf. 17 (6) (2019) 645–656.

[45] W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (12) (1983) 2577–2637.

[46] M. R. Jensen, P. Bernado, K. Houben, L. Blanchard, D. Marion, R. W. H. Ruigrok, M. Blackledge, Structural disorder within sendai virus nucleoprotein and phosphoprotein: Insight into the structural basis of molecular recognition, Protein & Peptide Letters 17 (8) (2010) 952–960.

[47] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, M. Tyers, J. D. Forman-Kay, Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an {SCF} ubiquitin ligase, Structure 18 (4) (2010) 494 – 506.

[48] J. Zimmerman, N. Eliezer, S. R., Characterization of amino acid sequences in proteins by statistical methods, J. Theor. Biol. 21 (X) (1968) 170 – 201.

[49] M.-K. Cho, H.-Y. Kim, P. Bernadó, C. O. Fernandez, M. Blackledge, M. Zweckstetter, Amino acid bulkiness defines the local conformations and dynamics of natively unfolded α-synuclein and Tau, J. Am. Chem. Soc. 129 (11) (2007) 3032–3033.

[50] J. S. Richardson, The anatomy and taxonomy of protein structure, Vol. 34 of Advances in Protein Chemistry, Academic Press, 1981, pp. 167 – 339.

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

[52] M. Louhivuori, K. Pääkkönen, K. Fredriksson, P. Permi, J. Lounila, A. Annila, On the origin of residual dipolar couplings from denatured proteins, J. Am. Chem. Soc. 125 (50) (2003) 15647–15650.

[53] Z. Shi, K. Chen, Z. Liu, N. R. Kallenbach, Conformation of the backbone in unfolded proteins, Chemical Reviews 106 (5) (2006) 1877–1897.

[54] M. Wells, H. Tidow, T. J. Rutherford, P. Markwick, M. R. Jensen, E. Mylonas, D. I. Svergun, M. Blackledge, A. R. Fersht, Structure of tumor suppressor p53 and its intrinsically disordered n-terminal transactivation domain, Proc. Natl. Acad. Sci. USA 105 (15) (2008) 5762–5767.

[55] N. K. Fox, S. E. Brenner, J.-M. Chandonia, SCOPe: Structural classification of proteins-extended, integrating SCOP and ASTRAL data and classification of new structures, Nucl. Acids Res. 42 (D1) (2014) D304–D309.

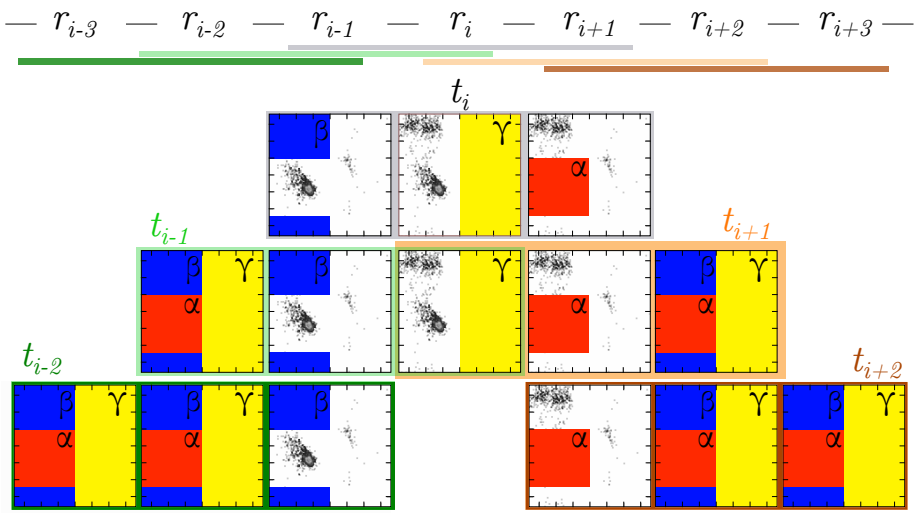

Figure 1: Ramachandran plots of residues in t_i, and in the neighboring tripeptides t_{i-2}, t_{i-1}, t_{i+1} and t_{i+2}. Colored regions correspond to the case where t_i is in the structural class $\mathscr{S} = \beta\gamma\alpha$. Notice that overlapping residues in consecutive tripeptides must be in the same structural class.

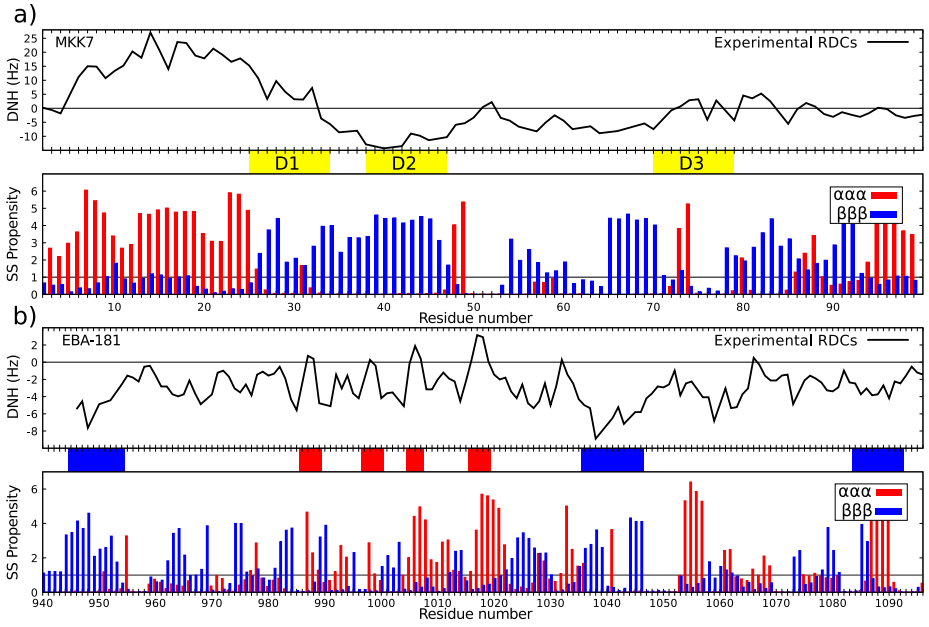

Figure 2: Experimental RDC profiles and secondary structure propensities predicted by LS2P for MKK7 (top) and EBA-181 (bottom). The plots only show helical ($\alpha\alpha\alpha$) and extended ($\beta\beta\beta$) propensities. The other structural classes are not displayed here, but are shown in Figures S3 and S4. The three binding domains in MKK7 are highlighted in yellow. For EBA-181, short helical and extended regions described in the literature are colored in red and blue, respectively.

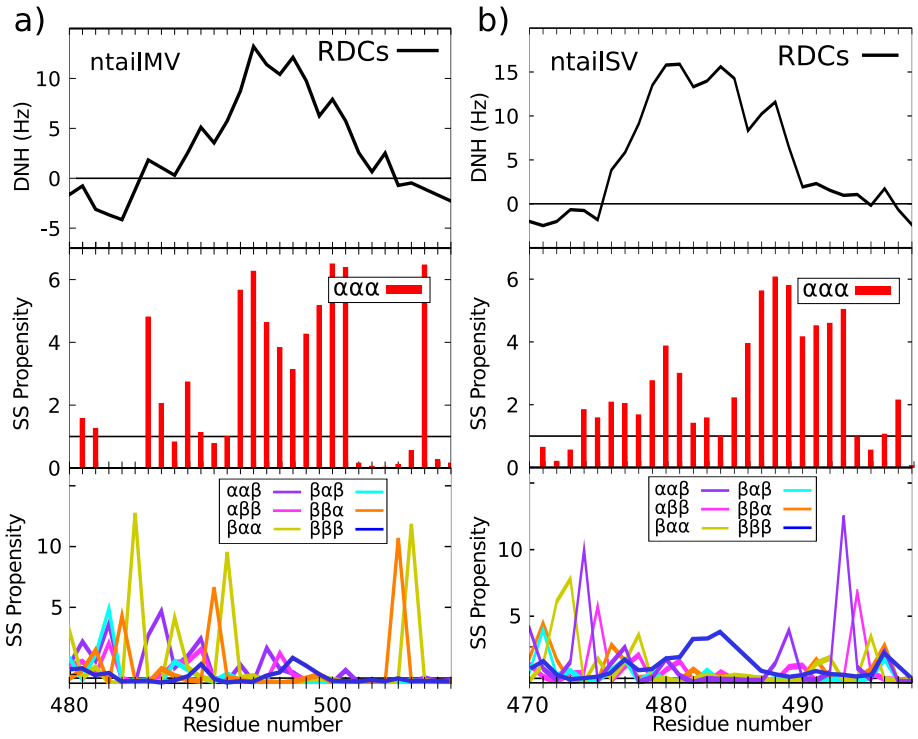

Figure 3: L2SP analysis of the helical region for ntailMV (a) and ntailSV (b). Top panels display the experimental RDC profiles. Middle panels show the predicted helical ($\alpha\alpha\alpha$) propensities. Bottom panels show other structural classes with significant propensities involving α and β conformations. In particular, the concatenation of high propensities for $\beta\beta\alpha$ and $\beta\alpha\alpha$ classes indicate the presence of N-capping residues.

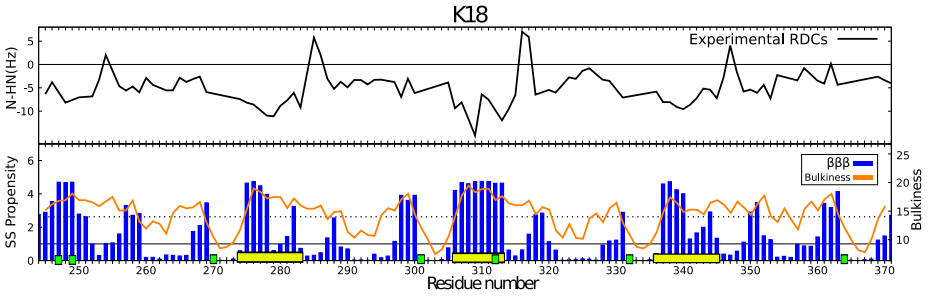

Figure 4: RDC profile (top, black line), $\beta\beta\beta$ propensity (bottom, blue bars) and bulkiness profiles (bottom, orange line) for the K18 construct of Tau protein. Prolines in the sequence are indicated with green squares. The dashed horizontal line at bulkiness = 14 indicates the threshold above which the sequence is considered bulky [49]. The three experimentally-characterized extended regions involving residues 275-282, 307-313 and 338-346 are highlighted in yellow.

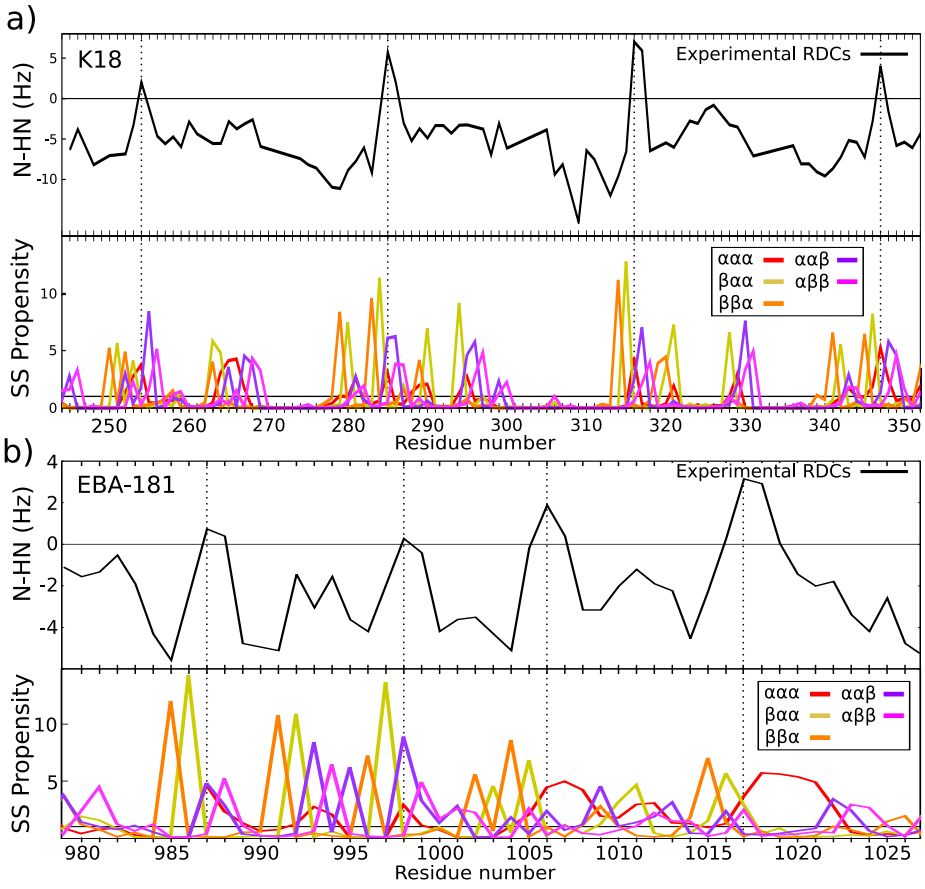

Figure 5: Experimental RDC profiles and secondary structure propensities predicted by LS2P for K18 (top) and central region in EBA-181 (bottom). The plots show concatenated $\beta\beta\alpha$, $\beta\alpha\alpha$, $\alpha\alpha\alpha$, $\alpha\alpha\beta$ and $\alpha\beta\beta$ propensities, which enable the identification of β-turns.

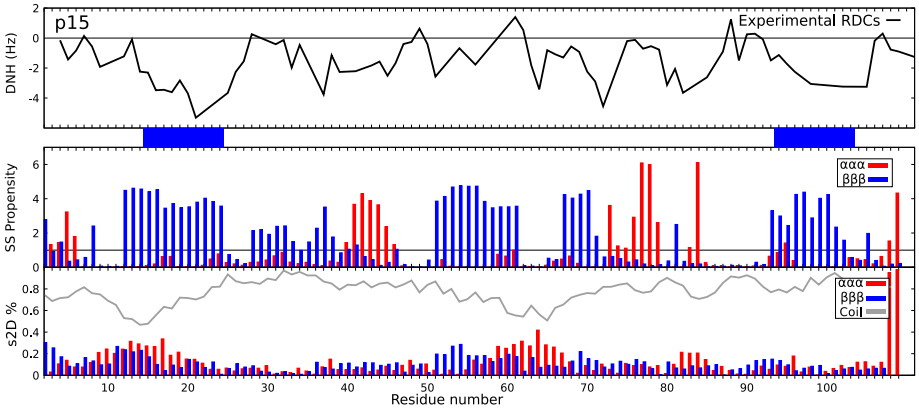

Figure 6: Experimental RDC profiles and secondary structure propensities predicted by LS2P (middle plot) and s2D (bottom plot) for p15. Extended regions described in the literature are represented as horizontal blue bars.