Structural propensity database of proteins

Kamil Tamiola[1*], **Matthew M Heberling**[1], **Jan Domanski**[1]

***For correspondence:** kamil@peptone.io (KT)

[1]Peptone - The Protein Intelligence Company, Amsterdam, Hullenbergweg 280, 1101BV, The Netherlands

Abstract An overwhelming amount of experimental evidence suggests that elucidations of protein function, interactions, and pathology are incomplete without inclusion of intrinsic protein disorder and structural dynamics. Thus, to expand our understanding of intrinsic protein disorder, we have created a database of secondary structure (SS) propensities for proteins (dSPP) as a reference resource for experimental research and computational biophysics. The dSPP comprises SS propensities of 7,094 unrelated proteins, as gauged from NMR chemical shift measurements in solution and solid state. Here, we explain the concept of SS propensity and analyze dSPP entries of therapeutic relevance, α-synuclein, MOAG-4, and the ZIKA NS2B-NS3 complex to show: (1) how propensity mapping generates novel structural insights into intrinsically disordered regions of pathologically relevant proteins, (2) how computational biophysics tools can benefit from propensity mapping, and (3) how the residual disorder estimation based on NMR chemical shifts compares with sequence-based disorder predictors. This work demonstrates the benefit of propensity estimation as a method that reports both on protein structure, lability, and disorder.

Introduction

Protein sequence is believed to hold the key to a perplexing mystery in the life sciences–the protein folding problem (***Dobson, 2003***). Therefore, immense efforts have been devoted to unraveling the sequence-structure relationship in polypeptides (***Baker, 2000***; ***Bowie, 2005***; ***Huang et al., 2016***). Although the fundamental forces of protein folding are known (***Dobson and Karplus, 1999***; ***Karplus and Weaver, 1976***; ***Karplus and Kuriyan, 2005***), complexity has hampered development of accurate folding prediction methods (***Moult et al., 2014***). Computational analysis of public protein databases, especially the Protein Data Bank (PDB) (***Varadi et al., 2015***), has played an integral role in shaping our fundamental understanding of protein structure, and for the advancement of protein design and structure prediction methodology (***Mackenzie and Grigoryan, 2017***). With accumulating structural data, it has become possible to mine for more complete and complex observations, which capture recurring structural features of proteins along with their sequence preferences. However, as explained in the seminal works of Dyson, Wright (***Dyson and Wright, 2004***) and Dobson (***Dobson, 2003***), naturally occurring protein disorder severely limits three-dimensional structure determination using X-ray crystallography, which renders only a rudimentary knowledge of the conformational state of disordered protein regions. Consequently, databases of protein structures are devoid of representative experimental data for intrinsically disordered proteins (IDPs) (***He et al., 2009***).

IDPs are abundant and control a vast array of biologically important processes, effectively complementing the functional spectrum of ordered proteins (***Dobson, 2003***; ***Xie et al., 2007***; ***Vucetic et al., 2003***). The prevalence of functional protein disorder (***Wright and Dyson, 2014***) demands reevaluation of the classical paradigm that a given protein sequence corresponds to a defined

structure and function. Importantly, literature suggests that the biophysical features of IDPs and their protein interactions vary tremendously, and that there may be no common mechanism that can explain the different binding modes observed experimentally. Disordered proteins appear to make combined use of features such as pre-formed structure and flexibility, depending on the individual system and the functional context ***Mollica et al.*** (***2016***).

With no simple physical model that relates residual disorder to protein sequence and function, machine learning (ML) offers hope for unraveling the biophysical features of disordered protein regions (***Varadi et al., 2015***; ***Dosztányi and Tompa, 2017***; ***Hanson et al., 2016***). However, the quality and annotation level of input data will dictate the broad applicability of ML-based prediction tools, which are currently hindered through the incomplete implementation of two fundamentals factors: (1) sensitivity to intrinsic protein disorder at the residue level and (2) experimental conditions. For the latter, fundamental laws of equilibrium thermodynamics prove that experimental conditions influence protein structure and dynamics (***Finkelstein and Badretdinov AYa, 1997***). Numerous protein models in PDB database contain ambiguous disordered regions, where more than one structure of the same protein sequence "disagrees" in terms of the presence or absence of missing residues. A thorough survey of intrinsically disordered protein regions (IDPRs) suggests that ambiguity is a natural result for many proteins crystallized under different conditions. It is likely that structural ambiguity arises because many of intrinsically disordered protein regions were conditionally or partially disordered (***DeForte and Uversky, 2016***). Although specialized databases of intrinsically disordered proteins (IDPs) exist (***Varadi and Tompa, 2015***; ***Piovesan et al., 2017***; ***Yu et al., 2017***), they contain approximately only 900 fully annotated proteins with binary assignment of structural disorder, as gauged from coarse experimental techniques (***Dosztányi and Tompa, 2017***). However, there is a constant need for comprehensive and residue-specific datasets, which would enhance our understanding of intrinsic protein disorder and propel the development of better predictive methods.

Among the available experimental techniques, NMR spectroscopy has proven to be unique in its capacity to study disordered and folded proteins with atomic detail, both in solid and solution states***Felli and Pierattelli*** (***2015***). NMR chemical shifts are perfectly suited to help answer such questions, since they reflect the conformational preferences of polypeptide chains with atomic resolution***Wishart and Case*** (***2001***) and display exquisite sensitivity to local dynamics***Berjanskii and Wishart*** (***2007***); ***Wishart et al.*** (***1992***); ***Marsh et al.*** (***2006***). Furthermore, chemical shifts are easy to measure experimentally and can be efficiently assigned to individual atoms in the protein molecule.

To advance computational methodologies that are sensitive to experimental conditions and intrinsic disorder at the residue level, we have constructed a database of structural propensities of proteins (dSPP). Our repository is derived from a subset of 7094 NMR resonance assignments of unrelated proteins in solution and solid state near physiological conditions. The transpiring chemical shifts are perfectly suited to address the above issues in computational predictions, since they reflect conformational preferences of polypeptide chains (***Wishart and Case, 2001***) with high sensitivity to local dynamics (***Berjanskii and Wishart, 2007***; ***Wishart et al., 1992***; ***Marsh et al., 2006***). The dSPP database makes use of an enhanced version of the structural propensity method, which has been specifically developed for IDPs (***Tamiola and Mulder, 2012***), thus providing optimal sensitivity to residual disorder for folded and unstructured polypeptides.

To demonstrate the value in dSPP, we first compare the structural propensity mappings from dSPP with corresponding 3D structures of therapeutically-relevant database entries, α-synuclein (αS), its aggregation modifying protein (MOAG-4), and NS2B-N3S protease complex from the Zika virus (ZIKV). Subsequently, we explain the relative benefit of structural propensity mapping in machine learning methodology. Lastly, empirically derived disorder scores from ZIKV are compared with theoretical disorder predictions from six state-of-the-art tools (***Hanson et al., 2016***; ***Ishida and Kinoshita, 2007***; ***Linding et al., 2003***; ***Ward et al., 2004***). Our work concludes with discussing how structural propensity data can propel development of structure and disorder prediction tools with higher accuracy and computational efficiency.

Results

NMR Assignment Data

A subset of 11286 protein resonance assignment entries was retrieved from the Biological Magnetic Resonance Data Bank (***Ulrich et al., 2008***). Within the downloaded records, 3286 contained more than one resonance assignment. Upon stringent filtering, 5860 assignment records were rejected because of a suboptimal length, 2426 were omitted due to the lack of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$, and 335 entries were removed that contain abnormal backbone resonance assignments due to the use of paramagnetic agents, non-protein compounds, and assignment errors. The final dataset consisted of 7094 protein resonance assignments. The average level of resonance assignment completeness (Supplementary Table 2) for $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C^{O}$, and ^{15}N was found to be 79%.

Experimental Conditions

Experimental parameters having profound biophysical effects on protein structure were assessed for the dSPP entries: temperature, pH, and ionic strength. Figure 1 plots the distributions of experimental conditions in dSPP. Panel 1a shows that temperature centers around 296 K $\pm$ 10 K(min., 283 K; max., 323 K). Panel 1b shows that a majority of database entries were studied in a physiological pH range with a mean pH of 6.9 $\pm$ 0.7. Furthermore, panel 1c shows a dominant contribution of low-salt resonance assignments that center around an ionic strength of approx. 82 mM $\pm$ 87 mM. Despite efforts towards standardization of data referencing (***Wishart et al., 1995***), NMR spectra are known to contain systematic referencing errors (***Zhang et al., 2003***). Panel d reflects this notion by reporting a mean offset correction of 0.36 ppm and standard deviation of 0.02 ppm, which follows a unimodal distribution.

Protein Sequence Statistics

As demonstrated in Figure 2a, dSPP contains 7094 protein sequences with the mean length of 119 $\pm$ 61 residues distributed in a unimodal fashion. Upon sequence homology and residue conservation analyses (Methods), the mean sequence conservation among aligned dSPP members is 0.11. Based on Figure 2b, it is apparent that amino acids are not distributed uniformly. Other protein sequence databases generate the same trend, which has been extensively studied from the perspective of sequence conservation analysis (***Valdar, 2002***). Dominant residues in dSPP are leucine, alanine, glutamate, and glycine; whereas cysteine and tryptophan are the least abundant.

Structural Propensity Statistics

Structural propensity is derived from the differences between experimental backbone chemical shifts and empirically predicted shielding constants observed in IDPs of similar composition (***Tamiola et al., 2010***). Therefore, structural propensity (Ψ) is a measure of the departure of an individual polypeptide residue from canonical SSs towards a 'random-coil' state. Calculation of the residue-specific structural propensity score is described in the Methods section. We assume the 'random-coil' state can be modeled after the ensemble behavior and characteristics of IDPs in solution (***Tamiola and Mulder, 2012***). The structural propensity adopts real number values. A residual score of -1.0 indicates a fully formed β-sheet, whereas a propensity of 1.0 suggests that 100% of ensemble members at a given polypeptide position adopt an α-helical conformation. Importantly, a near-zero score (0.0) indicates residual behavior and conformational sampling observed in IDPs. Fractional propensity scores are quantitative indicators of structural lability. Therefore, a score of -0.5 or 0.5 signifies that 50% of ensemble members sample conformations that are neither β-sheet nor α-helix, respectively.

The profile of structural propensities in Figure 3a shows a dominance by disordered and near-helical segments in dSPP entries. The low abundance of near-β residual propensities is directly related to a sparse representation of all-β proteins in the BMRB database. Figure 3b provides fundamental evidence for conformational preferences of individual polypeptide residues in solution

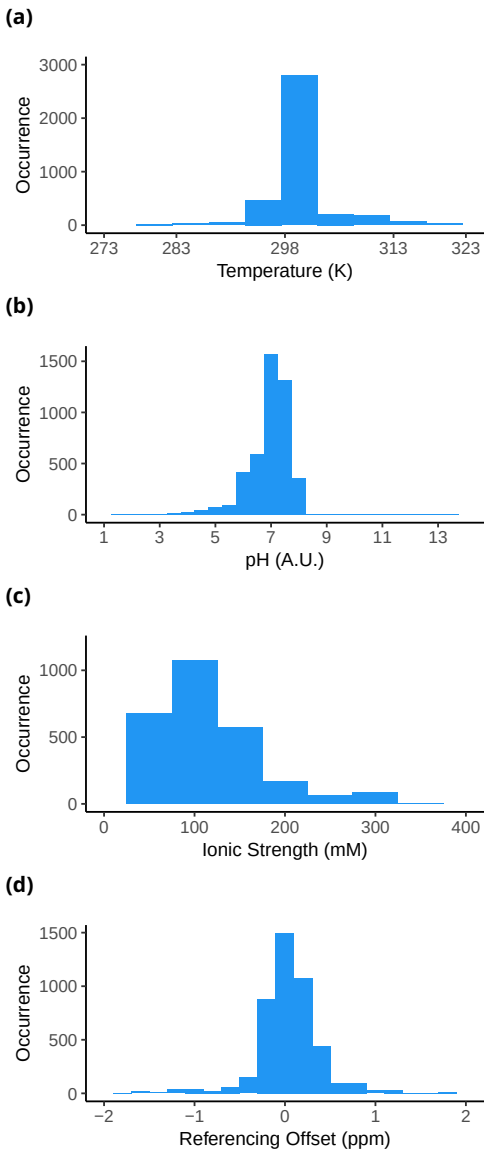

Figure 1. Statistical summary of experimental conditions in dSPP. Frequency distribution plots for (a) temperature, (b) pH, (c) ionic strength, and (d) resonance assignment referencing offset.

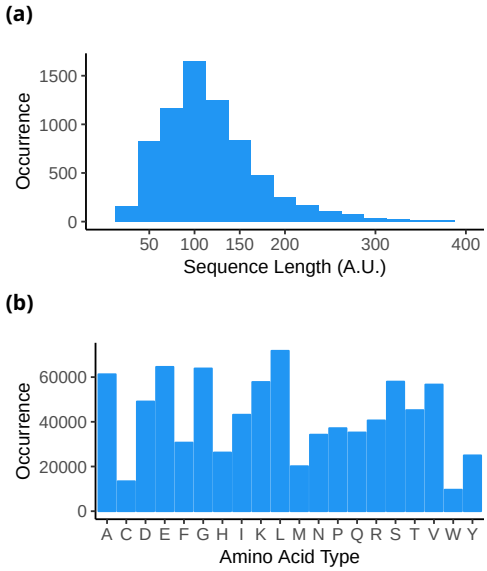

Figure 2. Statistical summary of protein sequences in dSPP. Frequency distribution plots for (a) protein sequence length and (b) amino acid composition.

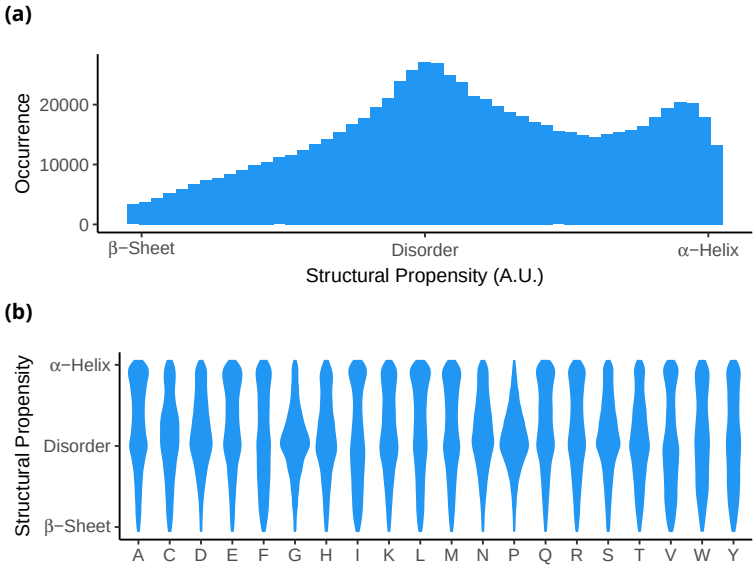

Figure 3. Statistical summary of structural propensity distribution in dSPP. Frequency distribution plots for (a) collective structural propensity and (b) normalized, residue-specific distribution of structural propensities. The widths of individual plots are directly proportional to the distribution density at the given structural propensity value. .

near physiological conditions. The skewed propensities towards disordered segments by glycine, proline, serine, and threonine is known (***Dwyer, 2006***; ***Kumari et al., 2015***). Interestingly, a difference in structural preference is observed for aspartate and glutamate. Although both residues were reportedly abundant in disordered proteins (***Dyson and Wright, 2005***; ***Uversky, 2016***, ***2014***), our analysis reveals a clear preference of aspartate (smaller side-chain) to populate disordered states, whereas glutamate displays a preference to populate more compact, α-helical conformations.

Examples of Structural Propensity Mapping

Structural propensities are directly mapped to atomistic structures derived from either NMR restraints or X-ray crystallography, where both have supplementary dynamics studies and extensive biophysical characterization. The models αS, MOAG-4, and ZIKV NS2B-NS3 complex are prime targets for drug discovery. Subsequently, we demonstrate the practical advantage of structural propensity over canonical structure-based classification methods and demonstrate how six seminal disorder predictors score against experimentally derived structural propensities for ZIKV NS2B-NS3 complex.

Intrinsically disordered α-synuclein

α-synuclein (αS) is a 140-residue IDP with high net charge and low hydrophobic content that has been implicated in a vast array of highly debilitating neurodegenerative conditions; most notably the Parkinson's disease (PD) (***McCormack and Di Monte, 2009***; ***Luk et al., 2012***). αS is believed to be involved in the regulation of the homeostasis of synaptic vesicles during neurotransmitter release (***Cooper et al., 2006***) and it has been suggested to play a crucial role in the interactions with vesicular membranes in both physiological and pathological contexts (***Cooper et al., 2006***). Since a hallmark of PD is the formation of abnormal intracellular protein aggregates of αS, referred to as Lewy bodies (***Cooper et al., 2006***), αS and its biophysical characterization have become the focal point of research on IDPs and neuropathology. Figure 4a shows the structural ensemble model of αS derived from experimental restraints; NMR chemical shifts and paramagnetic relaxation enhancement of NMR measurements (***Fusco et al., 2016***). As evidenced by supporting experimental data, under near-physiological conditions αS exists as a highly labile entity that inter-converts between a multitude of conformations. This notion is reflected in ensemble averaged secondary structure (SS) fraction depicted in Figure 4b, which demonstrates a lack of persistent canonical SS preference in αS ensemble. The structural propensity for αS, adopted from dSPP entry dSPP19351_0, is given in Figure 4c. Our structural propensity analysis displays good qualitative agreement with the ensemble model of αS. However, upon closer inspection, propensity scores for αS reveal the existence of three structural domains (***Luk et al., 2012***): 1) an N-terminal domain (residues 1-60) that supports regions of transient α-helical propensity; 2) a central hydrophobic region, known for historical reasons as the "non-Amyloid β component" (NAC) (residues 61-95) that is itself highly amyloidogenic and forms the core of the amyloid fibrils (***Vilar et al., 2008***); and 3) a C-terminal, acidic and proline-rich segment with residual tendencies to adopt extended structures (residues 96–140).

Partially disordered Modifier of Aggregation 4 (MOAG-4)

Although biophysical features and ensemble properties of αS have been elucidated, the regulatory forces behind αS aggregation in the cellular environment remain elusive. A study of αS aggregation behavior in *C. elegans* models led to the discovery of a regulator of protein aggregation; 'modifier of aggregation 4' (MOAG-4) (***van Ham et al., 2010***). It has been shown that inactivation of MOAG-4 leads to suppression of protein aggregation and associated toxicity in *C. elegans* models for αS (***van Ham et al., 2010***). Importantly, the human ortholog of MOAG-4, EDRK-rich factor (SERF) 1A, accelerates the aggregation of a broad range of amyloidogenic proteins *in vitro* in the initial stage of the process (***Falsone et al., 2012***). In a recent and extensive study, ***Yoshimura et al.*** (***2017***) investigated the kinetic and structural effects of MOAG-4 on the aggregation of αS. Figure 5a shows the structural

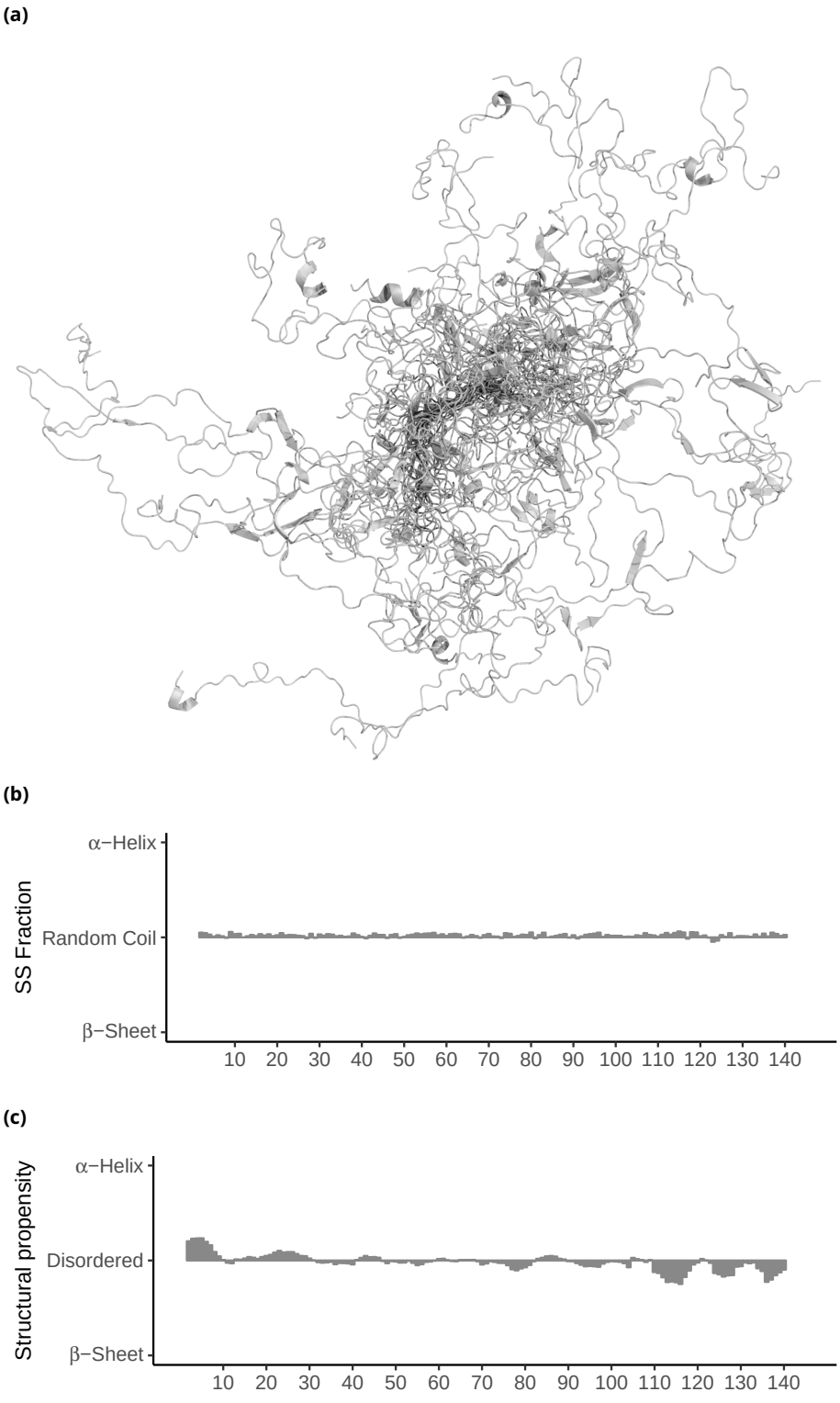

Figure 4. Structural ensemble of human α-synuclein. (a) Superposition (using 10-30) of fifty low-energy conformers from ensemble of αS (PED9AAC). (b) Residue-specific secondary structure (SS) fraction computed from models present in (PED9AAC) ensemble. (c) Structural propensity mapping for αS from dSPP database (dSPP19351_0).

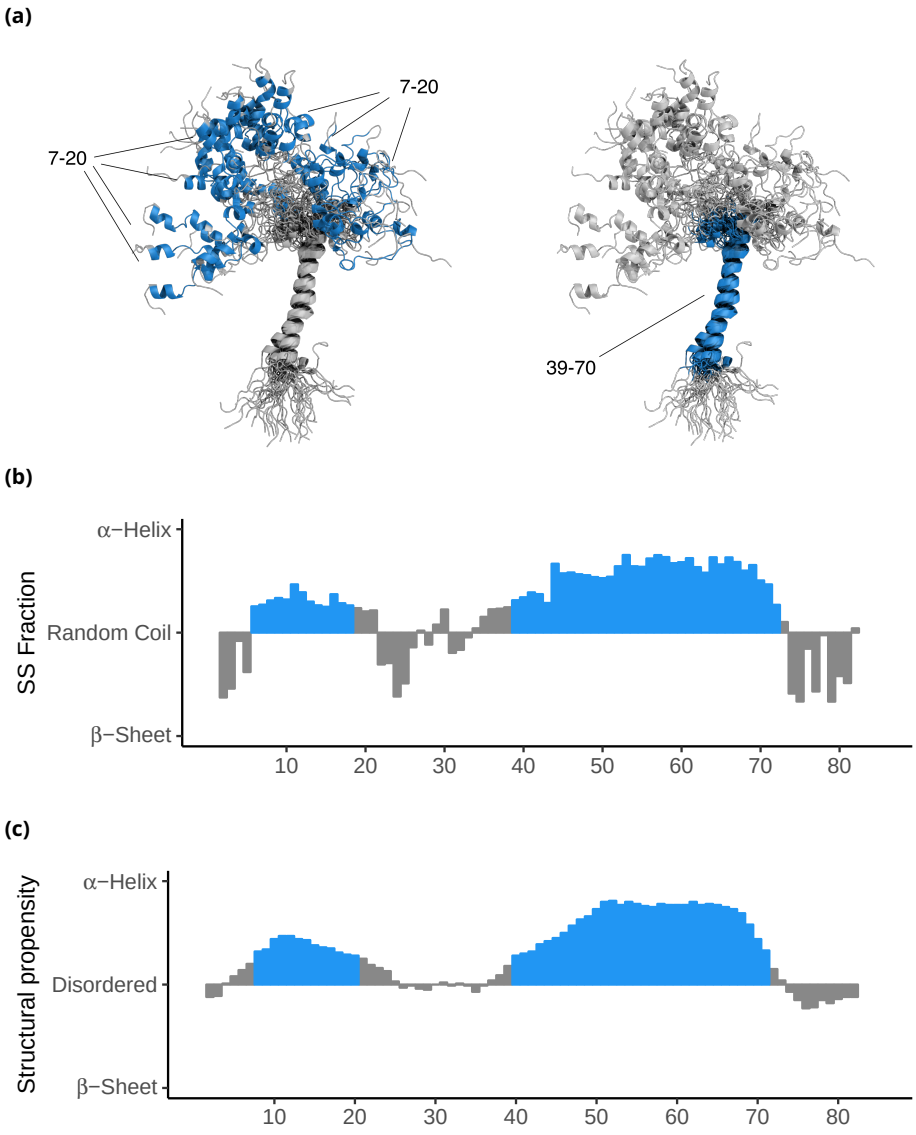

Figure 5. Structural ensemble of MOAG-4. (a) Fifty low energy conformers from ensemble of MOAG-4. (b) Ensemble-averaged secondary structure (SS) fraction computed from models present in PED5AAA ensemble. (c) Structural propensity from dSPP database (dSPP18841_0). The segments of MOAG-4 with SS fraction and propensity higher than 0.25 are marked in blue.

ensemble of MOAG-4 ***Yoshimura et al.*** (***2017***), calculated from NMR backbone chemical shifts (***Camilloni et al., 2013***) and cross-validated by prediction of experimental NOEs (***Yoshimura et al., 2017***). As evidenced by ensemble-average fractional SS analysis in Figure 5b, MOAG-4 is a partially disordered polypeptide with two distinct regions of α-helical propensity: lowly-populated, transient α-helix comprised of residues 7-20, and well-defined helical segment made of residues 39-70. The structural propensity mapping from NMR resonance assignments for MOAG-4 is available in dSPP as dSPP18841_0 and shown on Figure 5c. Computed structural propensity displays an excellent quantitative agreement with fractional SS, clearly demonstrating transient character of the 7-20 helical segment relative to the defined α-helix of 39-70 fragment.

NS2B-NS3 protease from Zika virus

The Zika virus (ZIKV) is a highly contagious representative of pathogenic flaviviruses and is linked to fetal microcephaly and neurological complications in adults, such as Guillain-Barré syndrome, acute myelitis, and meningoencephalitis (***Petersen et al., 2016***). The flavivirus NS2B-NS3 protease is a main target for antiviral therapeutics due to its role in ZIKV replication (***Luo et al., 2015***). Recently, NMR resonance assignments and crystal structures of the NS2B peptide cofactor complexed with NS3 protease from ZIKV (PDB: 5GJ4) were reported (***Zhang et al., 2016***). Structural studies of the NS2B-NS3 complex have been complemented with an analysis of ^{15}N T_1, T_2 NMR relaxation times and hetNOE experiments in solution (***Zhang et al., 2016***). The structural propensity and experimental conditions for the NS2B peptide and NS3 protease are available in dSPP under accession numbers dSPP26928_0 and dSPP26928_1, respectively. The structural propensity analyses of the NS2B-NS3 complex in Figures 6a and 6b show that both NS2B and NS3 display a heterogeneous distribution of structural disorder. Both N- and C-termini of NS3 protease resemble structural disorder found in IDPs, which are in an excellent agreement with the reported ^{15}N T_1, T_2, and hetNOE NMR experiments. It has been shown that residues 1-17 and 170-177 of the respective N- and C-termini in the NS3 protease are highly dynamic in solution, as evidenced by low T_1 and hetNOE values (< 0.6) (***Petersen et al., 2016***). As a result of extensive dynamics, both termini are missing in the X-ray structure. Conversely, although NS2B induces a closed complex with NS3 protease in solution, our structural propensity analysis indicates that the C-terminal of NS2B and residues around the P2 catalytic pocket in NS3 exhibit a large degree of disorder, which is contrary to the X-ray structure. Additionally, residues 80-95 of NS2B, which form a β-hairpin in the crystal structure, were found to be highly dynamic in solution, as evidenced by severe NMR spectral broadening and negative hetNOEs (***Zhang et al., 2016***). Thus, the NS2B-NS3 complex displays completely different dynamics near physiological conditions (ionic strength: 170 mM; pH: 7.3; pressure: 1 atm; temperature: 310 K) in solution-state NMR experiments compared to the conditions in X-ray crystallography (0.2M Sodium Malonate pH 4.0; 20% PEG 3350; flash-frozen in liquid N_2).

Structural propensity as an input for machine learning

The relative benefit of machine learning as a computational mean for protein structure and disorder prediction is its ability to accept arbitrary input and output data types (***Wang et al., 2016***). ML methods aimed specifically at structural disorder prediction are predominantly trained on datasets of binary-encoded, multi-class SS types. Protein disorder is inferred from missing structural coordinates, sequence conservation, similarity to known disordered proteins, and 3D contact maps (***Wang et al., 2017***; ***Jones, 1999***; ***Ishida and Kinoshita, 2007***). The assignment of c canonical classes of SSs for a peptide of N amino acids translates into a 'one-hot' encoded $c \times N$ tensor X_c which is then

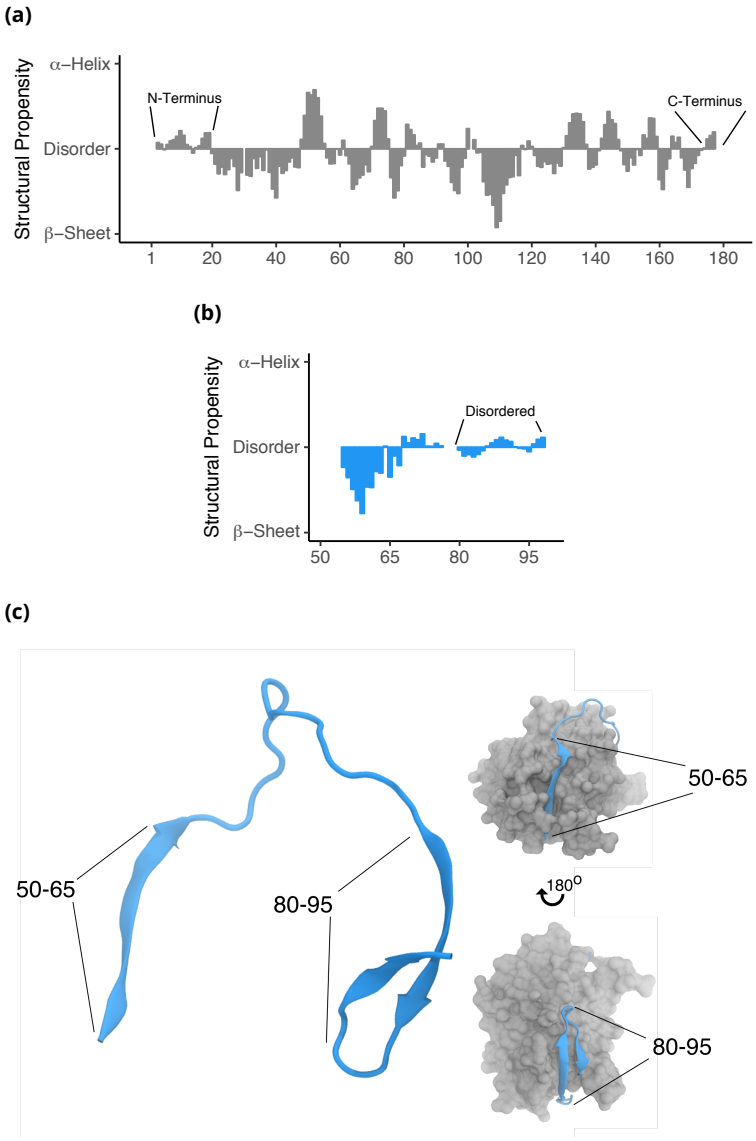

Figure 6. Structural propensity analysis for ZIKV NS2B-NS3 protease complex. Structural propensity plots for (a) NS3 protease and (b) NS2B peptide. The disordered segments of NS2B and NS3 are marked accordingly. (c) Comparative assessment of structural propensity and X-ray crystallography model (PDB: 5GJ4) for ZIKV NS2B-NS3 complex. The structure of NS2B polypeptide is given in blue, whereas NS3 is depicted in gray. Left-hand panel depicts cartoon representation of NS2B with 50-65 and 80-95 *beta*-hairpin segment, as gauged from X-ray model. The right-hand side panel demonstrates close-complex of NS2B-NS3.

Table 1. Non-parameteric single-way ANOVA (Kruskal-Wallis rank sum test) analyses of disorder prediction distributions for ZIKV (NS2B and NS3) proteins using SPOT, PrDOS, DisEMBL Loops, DisEMBL Hot Loops, DisEMBL Remark465 and Disopred 3 against absolute experimental propensity Ψ_{abs}. χ^2 denotes power divergence of predicted data with respect to experimental propensities, and p is the probability of distribution similarity. The ANOVA test was performed assuming one degree of freedom.

Protein	Method	χ^2	p
NS2B	SPOT	60.92	6.0×10^{-15}
	PrDOS	50.12	1.5×10^{-12}
	DisEMBL Loops	44.51	2.6×10^{-11}
	DisEMBL Hot Loops	82.43	2.2×10^{-16}
	DisEMBL Remark 465	44.23	2.9×10^{-11}
	Disopred 3	26.40	1.3×10^{-6}
NS3	SPOT	238.91	2.2×10^{-16}
	PrDOS	196.81	2.2×10^{-16}
	DisEMBL Loops	118.0	2.2×10^{-16}
	DisEMBL Hot Loops	254.8	2.2×10^{-16}
	DisEMBL Remark 465	205.0	2.2×10^{-16}
	Disopred 3	45.55	1.5×10^{-11}

used as input for ML, as given by Equation 1.

$$X_c(N,c) = \begin{bmatrix} \{0 & 0 & 0 & 1 & 1 & 1 & \cdots & 0\}_{\text{helix}} \\ \{0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0\}_{\text{beta}} \\ \{1 & 1 & 1 & 0 & 0 & 0 & \cdots & 1\}_{\text{coil}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \\ \{1 & 1 & 1 & 0 & 0 & 0 & \cdots & 1\}_{\text{c}} \end{bmatrix} \tag{1}$$

The structural propensity approach offers a computationally effective alternative to binary-type SS class assignments. The structural features of a polypeptide chain from Equation 1, can be encoded using a $1 \times N$ tensor X_Ψ shown in Equation 2.

$$X_\Psi(N) = \left[\{0.00 \quad 0.22 \quad 0.37 \quad 0.55 \quad 0.75 \quad 1.00 \quad \cdots \quad 0.00\}_\Psi\right] \tag{2}$$

The computational gain due to replacement of binary-class assignment by structural propensity is directly proportional to the number of structural classes c. Thus, structural propensity representation of a triple-class tensor reduces parameter search space by three-fold, which translates into shorter training time and better convergence (***Qian and Sejnowski, 1988***).

Comparison of disorder predictions and absolute experimental propensity scores

We have predicted structural disorder probabilities for the ZIKV peptides using SPOT (***Hanson et al., 2016***), PrDOS (***Ishida and Kinoshita, 2007***), variants of DisEMBL; Loops, Hot Loops, Remark465 (***Linding et al., 2003***) and Disopred 3 (***Ward et al., 2004***). The disorder probability predictions were compared against the absolute structural propensity score Ψ_{abs}, which reports on a normalized probability that residues within the polypeptide chain sample 'random-coil' conformations. Ψ_{abs} is computed from $\Psi_{abs} = 1 - |\Psi|$, and shown in Figure 7 in blue. Major discrepancies in the magnitude of predicted disorder probabilities are apparent for all theoretical methods. Smoothly interpolated disorder predictions by SPOT match only the fragments of N- and C- terminally disordered segments of NS2B-NS3. Compared to the experimentally derived Ψ_{abs} scores, SPOT displays the smallest sensitivity to structural detail among all computational methods. SPOT cannot discern stable SSs from confirmed domains of structural disorder in all tested polypeptides. Relative to SPOT, PrDOS produces more dispersed probabilities, yet still at great variance with Ψ_{abs}. The systematic

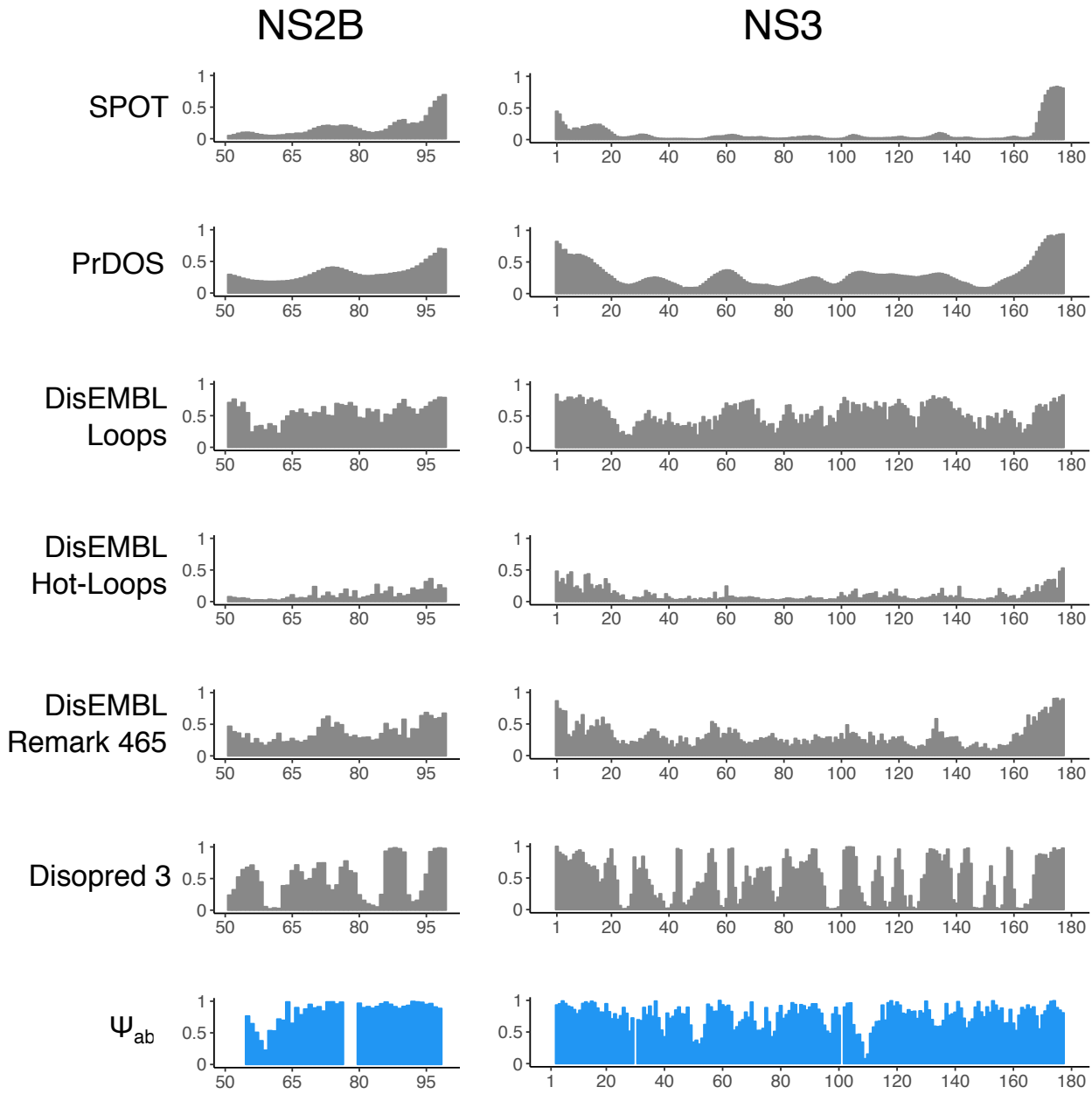

Figure 7. Prediction of probability of structural disorder using SPOT, PrDOS, DisEMBL Loops, Hot-Loops, Remark 465, Disopred 3 and absolute experimental structural propensity score Ψ_{abs}, shown in blue.

underestimation of structural disorder using DisEMBL Hot Loops method is also apparent, which is accentuated in Figure 7 where the method assigns near-zero disorder probability for the entire NS3 protein. This finding not only contrasts with the Ψ_{abs} score and NMR data, but disagrees profoundly with the trends reported by remaining predictors. Remark 465 and Loops display limited sensitivity to residual structure. As shown in Figure 7, DisEMBL Loops consistently produces probability distribution, which qualitatively resembles experimental Ψ_{abs}, yet differs significantly in the magnitudes of predicted disorder probabilities.

Table 1 contains the outcome of non-parametric single-way ANOVA analyses of computed disorder probability distributions with respect to experimental absolute structural propensity score Ψ_{abs}. The Disopred 3 method consistently produced the distribution of disorder scores with the smallest disparity between the experimental propensity. The method managed to properly assign structural disorder to N- and C- terminal domains of NS3, as well as properly identify β-extended 55-65 segment of NS2B. However, Disopred 3 systematically underestimated structural disorder in both NS2B and NS3 peptides. This notion was reflected by low value of χ^2 and p. A general inspection of power divergence χ^2 and distribution similarity probability p clearly suggest that all theoretical tools yield disorder probability distributions that are different from experimental absolute structural propensity scores for NS2B and NS3 proteins.

Discussion

X-ray crystallography and NMR spectroscopy have proven to be pivotal techniques in the determination of precise molecular models of proteins. However, polypeptides owe their complexity to an intricate combination of structure and dynamics (***Ishima and Torchia, 2000***). A dynamic view of protein structures can rationalize effects from experimental conditions or structural features and disorder that result in altered function and pathology. Understanding the link between protein structure and dynamics is particularly important in proteins with intrinsically disordered domains, which are prevalent in nature, have been implicated in numerous human pathologies, and are difficult to study by conventional structural biology methods (***Vucetic et al., 2003***; ***Dobson, 2003***; ***Uversky et al., 2008***; ***Wright and Dyson, 2014***; ***Uversky, 2014***). We have shown here how the inclusion of NMR-derived structural propensity can enhance biological understanding from static crystal structures, provide extensive structural characterization for IDPs involved in human pathologies, and assess the accuracy of computational methods that aim to discern stable SSs from disordered regions. This approach transpires from our dSPP public database, which should facilitate the inclusion of NMR-based SS descriptors in existing structure and disorder prediction methods and serve as a resource for structural interpretation of protein NMR chemical shifts.

The exquisite sensitivity of SS propensity to structural detail at the residue level is reflected in our assessment of intrinsically disordered α-synuclein (***Fusco et al., 2016***) and partially folded MOAG-4 (***Yoshimura et al., 2017***). We demonstrate that αS displays detectable propensity variation throughout its polypeptide chain, which corresponds to the presence of known micro-domains involved in αS function and aggregation behavior (***Luk et al., 2012***). Importantly, our propensity calculations agree well with the fractional SS estimation from structural ensemble models of αS computed using NMR chemical shifts and paramagnetic distance restraints (***Fusco et al., 2016***), reinforcing the notion that structural propensity reports on ensemble average properties of αS. Furthermore, the comparative assessment of structural features of MOAG-4 ensembles (***Yoshimura et al., 2017***) and SS propensity mapping clearly show that our method succeeds in structural characterization of polypeptides with heterogeneous distribution of structured and highly disordered domains. Additionally, ***Yoshimura et al.*** (***2017***) demonstrated that SS propensity could be used as a powerful agent to characterize structural effects of inter-molecular αS and MOAG-4 interactions, thus extending the application of SS propensity analysis to protein-protein complexes. Our SS propensity mapping of the ZIKV NS2B-NS3 protease complex shows that NS2B at near-physiological conditions exhibits a high level of structural disorder commonly found in intrinsically unstructured proteins (***Tamiola and Mulder, 2012***; ***Tamiola et al., 2010***). Very good agreement between our analysis and

hetNOEs NMR experiments (***Zhang et al., 2016***) suggests that the ordered β-hairpin motif in the NS2B X-ray model is a transient structure, which may obscure dynamics of the P2 interaction site at the NS2B-NS3 interface. As this interaction site has been designated as a potential drug design target (***Zhang et al., 2016***), NMR-derived structural propensities and relaxation techniques seem to be the most optimal tools to investigate changes in structure and dynamics of NS2B-NS3 upon drug binding under physiological conditions. The ultimate benefit of SS propensity as an experimental proxy for residual structure and disorder can be judged from our assessment of disorder predictors SPOT (***Hanson et al., 2016***), PrDOS (***Ishida and Kinoshita, 2007***), variants of DisEMBL; Loops, Hot Loops and Remark465 (***Linding et al., 2003***), and Disopred 3 (***Ward et al., 2004***). We show how absolute SS propensity can be used as a benchmark for residual disorder probability predictions, complementing existing approaches (***Moult et al., 2014***). Our analyses indicate that established disorder prediction methods suffer from insufficient sensitivity to disordered regions among folded domains. Our findings align very well with an extensive study by ***DeForte and Uversky*** (***2016***) who have shown that intrinsically disordered regions (IDRs) gauged from missing coordinates in PDB database may not be representative of intrinsic protein disorder. Thus, methods based of IDRs, including DisEMBL Loops, Hot Loops and Remark465 severely lack sensitivity to disorder behavior experimentally observed in IDPs. We postulate that our normalized SS propensity score could help to refine existing prediction methodologies and serve as an alternative to multi-dimensional, binary representation of protein structure classes for input in ML methods. This reduces numerical complexity and computational effort in development and training (***Dosztányi and Tompa, 2017***), and opens a possibility to include structural lability in predictive algorithms.

However, as a structure characterization technique based on chemical shift analyses, propensity mapping is subject to the theoretical and practical limitations of protein NMR spectroscopy (***Felli and Pierattelli, 2015***). The size of studied proteins and consequential overlap in spectral resonances, signal broadening, extensive chemical exchange, dynamics on different time scales (***Ishima and Torchia, 2000***), and the complex nature of electronic contributions to measured chemical shifts (***Wishart et al., 1992***; ***Berjanskii and Wishart, 2006***) all limit the submission of NMR resonance assignments to public repositories. Furthermore, the current implementation our structural propensity model is fine-tuned to detect deviations from canonical $\alpha-$ or $\beta-$ structures towards the disordered state only. It is expected that an expansion of the structural propensity concept to other SS classes (***Kabsch and Sander, 1983***) could further refine the computed SS propensity scores.

By transforming NMR resonance assignments of 7094 proteins at different experimental conditions into a database of structural propensities scores, we have created a resource that can enhance biological understanding of proteins with known NMR resonance assignments and propel the development of computational methods that aim to discern stable SSs from disordered regions. We hope that our structural propensity repository with a fully automated update cycle will benefit the machine learning community by providing a simple descriptor of SS class that is sensitive to structural disorder.

Methods

NMR Resonance Assignment Data

A subset of protein resonance assignment records with $^{1}H^{N}$, $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$, $^{13}C^{O}$, and ^{15}N nuclei was retrieved from Biological Magnetic Resonance Data Bank (***Ulrich et al., 2008***). Entries with sequence length of less than 30 amino acid residues and containing non-standard amino acids, nucleic acids, and paramagnetic agents were excluded from further analysis. The final screening for resonance assignment data was performed by selecting the resonance assignments which contained at least $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei, simultaneously with the absolute referencing offset smaller than $|\Delta_r| < 2.0$ ppm.

Secondary Chemical Shift Calculations

The sequence-dependent deviations of experimental resonance assignments from the 'random coil' chemical shifts, known as secondary chemical shifts, were calculated using the ncIDP chemical shift library (***Tamiola et al., 2010***). In our procedure, the 'random-coil' chemical shift for a nucleus $n \in \{\ ^1H^N\ ,\ ^1H^\alpha\ ,\ ^{13}C^\alpha\ ,\ ^{13}C^\beta\ ,\ ^{13}C^O\ ,\ ^{15}N\ \}$ of amino acid residue a, within a tripeptide $x-a-y$, is expressed as,

$$\delta^n_{calc}(x,a,y) = \Delta^n_p(x) + \delta^n(a) + \Delta^n_n(y) \tag{3}$$

where $\delta^n(a)$ is the 'random-coil' chemical shift in the $G{-}a{-}G$ reference sequence, and $\Delta^n_p(x)$ and $\Delta^n_n(y)$, are the neighbor corrections due to preceding ('p') and next ('n') residue, respectively. Consequently, the secondary chemical shift for a nucleus n, residue i is computed from,

$$\delta^n(i) = \delta^n_{exp}(i) - \delta^n_{calc}(i) \tag{4}$$

where $\delta^n_{exp}(i)$ is an experimental resonance assignment belonging to residue i.

Secondary Chemical Shift for Canonical Secondary Structures

The secondary chemical shifts, expected for fully formed α- or β-structures, were calculated from,

$$\delta^n(i,SS) = \delta^n_{SS}(i) - \delta^n_{calc}(i) \tag{5}$$

where δ_{SS} is the expected chemical shift for α- or β-structures taken from a chemical shift library by Wang and Jardetzky (***Wang and Jardetzky, 2002***); and $\delta^n_{calc}(i)$ denotes the sequence-specific 'random-coil' chemical shift computed from Equation 3.

NMR Resonance Referencing Offset Corrections

The relative difference between the experimental secondary chemical shifts computed for $^{13}C^\alpha$, $^{13}C^\beta$, and the expected secondary shifts of fully formed α- or β-structure (***Wang and Jardetzky, 2002***) was used as a measure of a mean referencing error of resonance assignments (***Marsh et al., 2006***). In the current implementation, the effects of fractional deuteration on $^{13}C^\alpha$, $^{13}C^\beta$ were not treated explicitly, but assumed to contribute to mean referencing offset. The mean chemical shift referencing offset Δ_ϵ was computed by minimizing,

$$\Delta_\epsilon = min \begin{cases} + \left| \sum_{i=1}^{N} (\delta^{^{13}C^\beta}(i) - \frac{\delta^{^{13}C^\alpha}(i)\ \delta^{^{13}C^\beta}(i,\alpha)}{\delta^{^{13}C^\alpha}(i,\alpha)}) \right| & \text{if } \delta^{^{13}C^\alpha}(i) - \delta^{^{13}C^\beta}(i) \geq 0 \\ + \left| \sum_{i=1}^{N} (\delta^{^{13}C^\alpha}(i) - \frac{\delta^{^{13}C^\beta}(i)\ \delta^{^{13}C^\alpha}(i,\beta)}{\delta^{^{13}C^\beta}(i,\beta)}) \right| & \text{if } \delta^{^{13}C^\alpha}(i) - \delta^{^{13}C^\beta}(i) < 0 \end{cases} \tag{6}$$

where $\delta^n(i,\alpha)$ and $\delta^n(i,\beta)$ denote the secondary chemical shift in a fully formed α- or β-structure, respectively.

Structural Propensity Calculations

The neighbor-corrected Structural Propensity Scores***Tamiola and Mulder*** (***2012***) Ψ were computed as,

$$\Psi(k,w) = \frac{\sum_n \sum_{j=k-w}^{k+w} C\ \theta^n(SS)\ \frac{\delta^n(j)}{\delta^n(j,SS)}}{\sum_n \sum_{j=k-w}^{k+w} \theta^n(SS)\ \frac{\delta^n(j,SS)}{\sigma^n(j,SS)}} \tag{7}$$

where $\delta^n(j)$ is the secondary chemical shift of type n for the j–th residue, $\delta^n(j,SS)$ represents the expected secondary chemical shift in canonical secondary structure of type SS, and $\sigma^n(j,SS)$ is the standard deviation of the expected secondary chemical shift taken from the database by Wang and Jardetzky (***Wang and Jardetzky, 2002***). The parameter $\theta^n(SS)$ reflects the relative sensitivity of the chemical shift n to secondary structure of type SS. Normalized values of $\theta^n(SS)$ are given in Table 2.

Table 2. Normalized weight parameters $\theta^n(SS)$, reflecting relative sensitivity of chemical shifts to the canonical secondary structures. $\theta^n(SS)$ are given in arbitrary units.

Nucleus	α-helix	β-sheet
$^1H^N$	0.15	0.30
$^1H^\alpha$	1.00	1.00
$^{13}C^O$	0.5	0.25
$^{13}C^\alpha$	1.00	1.00
$^{13}C^\beta$	1.00	1.00
^{15}N	0.125	0.250

Table 3. Parameters for the histogram analysis and plotting of NMR resonance assignment data.

Property	Samples	Bin width [Unit]
Temperature	4509	5 [K]
pH	4509	0.5 [A.U.]
Ionic Strength	4509	50 [mM]
Referencing Offset	4509	0.2 [ppm]
Sequence Length	7094	25 [A.U.]
Structural Propensity	770653	0.04 [A.U.]

Secondary structure type discrimination in Equation 7 is achieved by an inclusion of constant C, which is given by Equation 8.

$$\delta^n(j,SS) = \begin{cases} \delta^n(j,\alpha) \ \wedge \ C = 1 & \text{if } \delta^n(j)\ \delta^n(j,\alpha) > 0 \\ \delta^n(j,\beta) \ \wedge \ C = -1 & \text{if } \delta^n(j)\ \delta^n(j,\beta) > 0 \end{cases} \tag{8}$$

Sequence Conservation Analysis

Protein sequence data obtained from NMR resonance assignment records were fed to MUSCLE sequence alignment program (***Edgar, 2004***). A column-ordered mean sequence conservation score was used as a measure of sequence similarity across dSPP entries (***Valdar, 2002***). The sequence conservation score was computed assuming BLOSUM62 matrix (***Eddy, 2004***).

Statistical Analysis

The statistical analysis of protein resonance assignment data was performed in R (version 3.3.2) (***R Core Team, 2013***). In order to avoid data over-binning, the distribution analyses in Figures 1, 2 and 3 were done with variable, sample-dependent bin width. The exact parameters of the histogram analysis are given in Table 3. The residue-specific distributions of structural propensities, depicted on Figure 3b, were computed from the individual kernel density plots of structural propensities with the fixed kernel size of 0.04. The non-parametric single-way ANOVA analysis of theoretical disorder probability distributions was performed using Kruskal-Wallis Test available in R software.

Secondary structure fraction

The ensemble average secondary structure (SS) fraction was computed from 1000 and 946 PDB files available in αS (PED9AAC) (***Fusco et al., 2016***) and MOAG-4 ensembles (kindly provided by Dr. Predrag Kukic, University of Cambridge, UK), respectively. We have used dihedral angle analysis tools of GROMACS (***Berendsen et al., 1995***) to extract residue-specific Φ and Ψ angles. Subsequently, we have computed how many times residues in each ensemble member visited canonical SSs, α-helix and β-sheet, defined by Φ, Ψ angles of $< -48°, -34° >$ and $< -140°, 130° >$, respectively. The

final fractional SS was computed from arithmetic averages of collected dihedral angle statistics for both ensembles.

Software Implementation and Availability

The dSPP database was implemented using reactive MeteorJS web application framework with Numerical Python and TensorFlow wrappers. The database is available at https://peptone.io/dspp, both as an interactive application with contextual search and standalone download in JSON format.

Acknowledgments

Authors acknowledge Dr. Wenwei Zheng (NIDDK, US), Dr. Ruud Scheek (University of Groningen, NL) and Dr. Xavier Periole (Aarhus University, DK) for insightful comments and editorial suggestions. Authors thank Alison Lowndes, Carlo Ruiz and Dr. Adam Grzywaczewski, (NVIDIA Corporation) for facilitating collaboration and access to DGX-1 supercomputer. Jon Wedell (BMRB) is greatly acknowledged for technical support with NMR resonance assignment retrieval from BMRB. We thank Dr. Frans A.A. Mulder (Aarhus University, DK) and Dr. Predrag Kukic (University of Cambridge, UK) for providing structural ensemble models of MOAG-4. Lastly, we want to greatly acknowledge Mark Berger (NVIDIA Corporation) for overwhelming support throughout the execution of this project.

References

Baker D. A surprising simplicity to protein folding. Nature. 2000 5; 405(6782):39–42. http://www.nature.com/doifinder/10.1038/35011000, doi: 10.1038/35011000.

Beck DAC, Alonso DOV, Inoyama D, Daggett V. The intrinsic conformational propensities of the 20 naturally occurring amino acids and reflection of these propensities in proteins. Proceedings of the National Academy of Sciences of the United States of America. 2008 8; 105(34):12259–64. http://www.ncbi.nlm.nih.gov/pubmed/18713857http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2527899, doi: 10.1073/pnas.0706527105.

Berendsen HJC, van der Spoel D, van Drunen R. GROMACS: A message-passing parallel molecular dynamics implementation. Computer Physics Communications. 1995 9; 91(1-3):43–56. doi: 10.1016/0010-4655(95)00042-E.

Berjanskii M, Wishart DS. NMR: prediction of protein flexibility. Nat Protocols. 2006 2; 1(2):683–688.

Berjanskii MV, Wishart DS. The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. Nucleic acids research. 2007 7; 35(Web Server issue):531–7. http://www.ncbi.nlm.nih.gov/pubmed/17485469http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1933179, doi: 10.1093/nar/gkm328.

Berjanskii MV, Wishart DS. Application of the random coil index to studying protein flexibility. Journal of Biomolecular NMR. 2008 1; 40(1):31–48. http://link.springer.com/10.1007/s10858-007-9208-0, doi: 10.1007/s10858-007-9208-0.

Berman HM. The Protein Data Bank. Nucleic Acids Research. 2000 1; 28(1):235–242. doi: 10.1093/nar/28.1.235.

Bhattacharjee N, Biswas P. Position-specific propensities of amino acids in the β-strand. BMC structural biology. 2010 9; 10(1):29. doi: 10.1186/1472-6807-10-29.

Biljan I, Ilc G, Giachin G, Raspadori A, Zhukov I, Plavec J, Legname G. Toward the molecular basis of inherited prion diseases: NMR structure of the human prion protein with V210I mutation. J Mol Biol. 2011 9; 412(4):660–673. http://dx.doi.org/10.1016/j.jmb.2011.07.067, doi: 10.1016/j.jmb.2011.07.067.

Bowie JU. Solving the membrane protein folding problem. Nature. 2005 12; 438(7068):581–589. http://www.nature.com/doifinder/10.1038/nature04395, doi: 10.1038/nature04395.

Busia A, Collins J, Jaitly N. Protein Secondary Structure Prediction Using Deep Multi-scale Convolutional Neural Networks and Next-Step Conditioning. ArXiv e-prints. 2016 11; http://arxiv.org/abs/1611.01503.

Camilloni C, Cavalli A, Vendruscolo M. Replica-Averaged Metadynamics. Journal of Chemical Theory and Computation. 2013 12; 9(12):5610–5617. http://pubs.acs.org/doi/abs/10.1021/ct4006272, doi: 10.1021/ct4006272.

Chou CC, Wang AHJ. Structural D/E-rich repeats play multiple roles especially in gene regulation through DNA/RNA mimicry. Mol BioSyst. 2015; 11(8):2144–2151. http://xlink.rsc.org/?DOI=C5MB00206K, doi: 10.1039/C5MB00206K.

Cooper AA, Gitler AD, Cashikar A, Haynes CM, Hill KJ, Bhullar B, Liu K, Xu K, Strathearn KE, Liu F, Cao S, Caldwell KA, Caldwell GA, Marsischky G, Kolodner RD, Labaer J, Rochet JC, Bonini NM, Lindquist S. -Synuclein Blocks ER-Golgi Traffic and Rab1 Rescues Neuron Loss in Parkinson's Models. Science. 2006 7; 313(5785):324–328. http://www.ncbi.nlm.nih.gov/pubmed/16794039http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1983366http://www.sciencemag.org/cgi/doi/10.1126/science.1129462, doi: 10.1126/science.1129462.

DeForte S, Uversky VN. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. Protein Science. 2016 3; 25(3):676–688. http://www.ncbi.nlm.nih.gov/pubmed/26683124http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4815412http://doi.wiley.com/10.1002/pro.2864, doi: 10.1002/pro.2864.

Dobson CM. Protein folding and misfolding. Nature. 2003 12; 426(6968):884–90. http://www.ncbi.nlm.nih.gov/pubmed/14685248, doi: 10.1038/nature02261.

Dobson CM, Karplus M. The fundamentals of protein folding: bringing together theory and experiment. Current Opinion in Structural Biology. 1999 2; 9(1):92–101. http://www.ncbi.nlm.nih.gov/pubmed/10047588http://linkinghub.elsevier.com/retrieve/pii/S0959440X99800128, doi: 10.1016/S0959-440X(99)80012-8.

Dosztányi Z, Tompa P. Bioinformatics Approaches to the Structure and Function of Intrinsically Disordered Proteins. In: *From Protein Structure to Function with Bioinformatics* Dordrecht: Springer Netherlands; 2017.p. 167–203. http://link.springer.com/10.1007/978-1-4020-9058-5_5http://link.springer.com/10.1007/978-94-024-1069-3_6, doi: 10.1007/978-94-024-1069-3_6.

Dwyer DS. Nearest-neighbor effects and structural preferences in dipeptides are a function of the electronic properties of amino acid side-chains. Proteins: Structure, Function, and Bioinformatics. 2006 2; 63(4):939–948. http://doi.wiley.com/10.1002/prot.20906, doi: 10.1002/prot.20906.

Dyson HJ, Wright PE. Unfolded proteins and protein folding studied by NMR. Chemical reviews. 2004 8; 104(8):3607–22. http://www.ncbi.nlm.nih.gov/pubmed/15303830, doi: 10.1021/cr030403s.

Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nature reviews Molecular cell biology. 2005 3; 6(3):197–208. http://www.ncbi.nlm.nih.gov/pubmed/15738986, doi: 10.1038/nrm1589.

Ebert MC, Pelletier JN. Computational tools for enzyme improvement: why everyone can – and should – use them. Current Opinion in Chemical Biology. 2017 4; 37:89–96. http://linkinghub.elsevier.com/retrieve/pii/S1367593117300248, doi: 10.1016/j.cbpa.2017.01.021.

Eddy SR. Where did the BLOSUM62 alignment score matrix come from? Nature Biotechnology. 2004 8; 22(8):1035–1036. http://www.nature.com/doifinder/10.1038/nbt0804-1035, doi: 10.1038/nbt0804-1035.

Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004 3; 32(5):1792–1797. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh340, doi: 10.1093/nar/gkh340.

Englander SW, Mayne L. The nature of protein folding pathways. Proceedings of the National Academy of Sciences of the United States of America. 2014 11; 111(45):15873–80. http://www.ncbi.nlm.nih.gov/pubmed/25326421http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4234557, doi: 10.1073/pnas.1411798111.

Falsone SF, Meyer NH, Schrank E, Leitinger G, Pham CLL, Fodero-Tavoletti MT, Holmberg M, Dulle M, Scicluna B, Gesslbauer B, Rückert HM, Wagner GE, Merle DA, Nollen EA, Kungl AJ, Hill AF, Cappai R, Zangger K. SERF Protein Is a Direct Modifier of Amyloid Fiber Assembly. Cell Reports. 2012 8; 2(2):358–371. doi: 10.1016/j.celrep.2012.06.012.

Felli IC, Pierattelli R, editors. Intrinsically Disordered Proteins Studied by NMR Spectroscopy, vol. 870 of Advances in Experimental Medicine and Biology. Switzerland: Springer International Publishing; 2015.

Finkelstein AV, Badretdinov AYa. Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. Folding & design. 1997; 2(2):115–21. http://www.ncbi.nlm.nih.gov/pubmed/9135984.

500 **Fuentes G**, Nederveen AJ, Kaptein R, Boelens R, Bonvin AM. Describing partially unfolded states of proteins from
501 sparse NMR data. J Biomol NMR. 2005 1; 33(3):175–186. http://www.ncbi.nlm.nih.gov/pubmed/16331422, doi:
502 10.1007/s10858-005-3207-9.

503 **Fusco G**, De Simone A, Arosio P, Vendruscolo M, Veglia G, Dobson CM. Structural Ensembles of Membrane-
504 bound α-Synuclein Reveal the Molecular Determinants of Synaptic Vesicle Affinity. Scientific Reports. 2016 7;
505 6(1):27125. http://www.nature.com/articles/srep27125, doi: 10.1038/srep27125.

506 **Gal Y**. Uncertainty in Deep Learning. PhD thesis, University of Cambridge; 2016.

507 **Goedert M**, Spillantini MG, Del Tredici K, Braak H. 100 years of Lewy pathology. Nature Reviews Neurology.
508 2012 11; 9(1):13–24. doi: 10.1038/nrneurol.2012.242.

509 **van Ham TJ**, Holmberg MA, van der Goot AT, Teuling E, Garcia-Arencibia M, Kim He, Du D, Thijssen KL, Wiersma
510 M, Burggraaff R, van Bergeijk P, van Rheenen J, Jerre van Veluw G, Hofstra RMW, Rubinsztein DC, Nollen EAA.
511 Identification of MOAG-4/SERF as a Regulator of Age-Related Proteotoxicity. Cell. 2010 8; 142(4):601–612. doi:
512 10.1016/j.cell.2010.07.020.

513 **Hanson J**, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term
514 memory recurrent neural networks. Bioinformatics. 2016 12; 25(5):3389–3402. doi: 10.1093/bioinformat-
515 ics/btw678.

516 **He B**, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell
517 Research. 2009 8; 19(8):929–949. http://www.ncbi.nlm.nih.gov/pubmed/19597536http://www.nature.com/
518 doifinder/10.1038/cr.2009.87, doi: 10.1038/cr.2009.87.

519 **Henzler-Wildman K**, Kern D. Dynamic personalities of proteins. Nature. 2007 12; 450(7172):964–972. http:
520 //www.nature.com/doifinder/10.1038/nature06522, doi: 10.1038/nature06522.

521 **Huang PS**, Boyken SE, Baker D. The coming of age of de novo protein design. Nature. 2016 9; 537(7620):320–327.
522 http://www.nature.com/doifinder/10.1038/nature19946, doi: 10.1038/nature19946.

523 **Ilc G**, Giachin G, Jaremko M, Jaremko L, Benetti F, Plavec J, Zhukov I, Legname G. NMR Structure of the Human
524 Prion Protein with the Pathological Q212P Mutation Reveals Unique Structural Features. PLoS ONE. 2010 7;
525 5(7):e11715. http://dx.plos.org/10.1371/journal.pone.0011715, doi: 10.1371/journal.pone.0011715.

526 **Ishida T**, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. Nucleic Acids
527 Research. 2007 5; 35(Web Server):W460–W464. doi: 10.1093/nar/gkm363.

528 **Ishima R**, Torchia DA. No Title. Nature Structural Biology. 2000 9; 7(9):740–743. http://www.nature.com/
529 doifinder/10.1038/78963, doi: 10.1038/78963.

530 **Jansen C**, Parchi P, Capellari S, Vermeij AJ, Corrado P, Baas F, Strammiello R, van Gool WA, van Swieten JC,
531 Rozemuller AJM. Prion protein amyloidosis with divergent phenotype associated with two novel nonsense
532 mutations in PRNP. Acta neuropathologica. 2010 2; 119(2):189–97. http://www.ncbi.nlm.nih.gov/pubmed/
533 19911184http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2808512, doi: 10.1007/s00401-009-
534 0609-x.

535 **Jones DT**. Protein secondary structure prediction based on position-specific scoring matrices. Jour-
536 nal of molecular biology. 1999 9; 292(2):195–202. http://www.ncbi.nlm.nih.gov/pubmed/10493868, doi:
537 10.1006/jmbi.1999.3091.

538 **Kabsch W**, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and
539 geometrical features. Biopolymers. 1983 12; 22(12):2577–637. http://www.ncbi.nlm.nih.gov/pubmed/6667333,
540 doi: 10.1002/bip.360221211.

541 **Karplus M**, Kuriyan J. Molecular dynamics and protein function. Proceedings of the National
542 Academy of Sciences of the United States of America. 2005 5; 102(19):6679–85. http://www.ncbi.nlm.
543 nih.gov/pubmed/15870208http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1100762, doi:
544 10.1073/pnas.0408930102.

545 **Karplus M**, Weaver DL. Protein folding dynamics: the diffusion-collision model and experimental data. Protein
546 science. 1994 4; 3(4):650–68. http://www.ncbi.nlm.nih.gov/pubmed/8003983http://www.pubmedcentral.nih.
547 gov/articlerender.fcgi?artid=PMC2142854, doi: 10.1002/pro.5560030413.

548 **Karplus M**, Weaver DL. Protein-folding dynamics. Nature. 1976 4; 260(5550):404–406.

549 **Koday MT**, Nelson J, Chevalier A, Koday M, Kalinoski H, Stewart L, Carter L, Nieusma T, Lee PS, Ward AB, Wilson
550 IA, Dagley A, Smee DF, Baker D, Fuller DH. A Computationally Designed Hemagglutinin Stem-Binding Protein
551 Provides In Vivo Protection from Influenza Independent of a Host Immune Response. PLoS pathogens. 2016
552 2; 12(2):e1005409. doi: 10.1371/journal.ppat.1005409.

553 **Koehl P**, Levitt M. Structure-based conformational preferences of amino acids. Proceedings of the National
554 Academy of Sciences of the United States of America. 1999 10; 96(22):12524–9. http://www.ncbi.nlm.nih.gov/
555 pubmed/10535955http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC22969.

556 **Kovač V**, Zupančič B, Ilc G, Plavec J, Čurin Šerbec V. Truncated prion protein PrP226* - A structural view on
557 its role in amyloid disease. Biochemical and biophysical research communications. 2017 2; 484(1):45–50.
558 http://www.ncbi.nlm.nih.gov/pubmed/28109886, doi: 10.1016/j.bbrc.2017.01.078.

559 **Krois AS**, Ferreon JC, Martinez-Yamout MA, Dyson HJ, Wright PE. Recognition of the disordered p53 trans-
560 activation domain by the transcriptional adapter zinc finger domains of CREB-binding protein. Proceed-
561 ings of the National Academy of Sciences of the United States of America. 2016 3; 113(13):1853–62. doi:
562 10.1073/pnas.1602487113.

563 **Kumari B**, Kumar R, Kumar M. Low complexity and disordered regions of proteins have different structural and
564 amino acid preferences. Molecular bioSystems. 2015 2; 11(2):585–94. doi: 10.1039/c4mb00425f.

565 **Lee H**, Ren J, Nocadello S, Rice AJ, Ojeda I, Light S, Minasov G, Vargas J, Nagarathnam D, Anderson WF, Johnson
566 ME. Identification of novel small molecule inhibitors against NS2B/NS3 serine protease from Zika virus.
567 Antiviral Research. 2017; 139:49–58. doi: 10.1016/j.antiviral.2016.12.016.

568 **Linding R**, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for
569 structural proteomics. Structure (London, England : 1993). 2003 11; 11(11):1453–9. http://www.ncbi.nlm.nih.
570 gov/pubmed/14604535.

571 **Luk KC**, Kehm V, Carroll J, Zhang B, O'Brien P, Trojanowski JQ, Lee VMY. Pathological α-synuclein transmis-
572 sion initiates Parkinson-like neurodegeneration in nontransgenic mice. Science (New York, NY). 2012 11;
573 338(6109):949–53. doi: 10.1126/science.1227157.

574 **Luo D**, Vasudevan SG, Lescar J, The flavivirus NS2B-NS3 protease-helicase as a target for antiviral drug develop-
575 ment; 2015. doi: 10.1016/j.antiviral.2015.03.014.

576 **Mackenzie CO**, Grigoryan G. Protein structural motifs in prediction and design. Current Opinion in Structural
577 Biology. 2017 6; 44:161–167. doi: 10.1016/j.sbi.2017.03.012.

578 **Maltsev AS**, Ying J, Bax A. Impact of N-Terminal Acetylation of α-Synuclein on Its Random Coil and Lipid Binding
579 Properties. Biochemistry. 2012 6; 51(25):5004–5013. doi: 10.1021/bi300642h.

580 **Marsh JA**, Singh VK, Jia Z, Forman-Kay JD. Sensitivity of secondary structure propensities to sequence differences
581 between α- and γ-synuclein: Implications for fibrillation. Protein Science. 2006 12; 15(12):2795–2804. http:
582 //doi.wiley.com/10.1110/ps.062465306, doi: 10.1110/ps.062465306.

583 **McCormack AL**, Di Monte D. Enhanced α-Synuclein Expression in Human Neurodegenerative Diseases: Patho-
584 genetic and Therapeutic Implications. Current Protein and Peptide Science. 2009; 10(5):476–482.

585 **Mollica L**, Bessa LM, Hanoulle X, Jensen MR, Blackledge M, Schneider R. Binding Mechanisms of Intrinsically
586 Disordered Proteins: Theory, Simulation, and Experiment. Frontiers in molecular biosciences. 2016; 3:52. doi:
587 10.3389/fmolb.2016.00052.

588 **Moult J**, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein
589 structure prediction (CASP)–round x. Proteins. 2014 2; 82 Suppl 2:1–6. doi: 10.1002/prot.24452.

590 **Mukrasch MD**, Bibow S, Korukottu J, Jeganathan S, Biernat J, Griesinger C, Mandelkow E, Zweckstetter M.
591 Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. PLoS Biology. 2009 2; 7(2):e1000034.
592 http://dx.plos.org/10.1371/journal.pbio.1000034, doi: 10.1371/journal.pbio.1000034.

593 **Oates ME**, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kur-
594 gan L, Dunker AK, Gough J. D2P2: database of disordered protein predictions. Nucleic Acids Research.
595 2013 1; 41(D1):D508–D516. https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1226, doi:
596 10.1093/nar/gks1226.

Oldfield CJ, Dunker AK. Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. Annual Review of Biochemistry. 2014 6; 83(1):553–584. http://www.annualreviews.org/doi/10.1146/annurev-biochem-072711-164947, doi: 10.1146/annurev-biochem-072711-164947.

Petersen LR, Jamieson DJ, Powers AM, Honein MA. Zika Virus. New England Journal of Medicine. 2016 4; 374(16):1552–1563. http://www.nejm.org/doi/10.1056/NEJMra1602113, doi: 10.1056/NEJMra1602113.

Phoo WW, Li Y, Zhang Z, Lee MY, Loh YR, Tan YB, Ng EY, Lescar J, Kang C, Luo D. Structure of the NS2B-NS3 protease from Zika virus after self-cleavage. Nature Communications. 2016 11; 7:13410. http://www.nature.com/doifinder/10.1038/ncomms13410, doi: 10.1038/ncomms13410.

Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z, Elofsson A, Gasparini A, Hatos A, Kajava AV, Kalmar L, Leonardi E, Lazar T, Macedo-Ribeiro S, Macossay-Castillo M, Meszaros A, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic Acids Research. 2017 1; 45(D1):D219–D227. doi: 10.1093/nar/gkw1056.

POPE 5 1996 Montpellier, POPE 6 1997 Norwich, Perspectives on Protein Engineering Conference (POPE) (5 1996 03 02-06 Montpellier), Perspectives on Protein Engineering Conference (POPE) (6 1997 06 28-07 02 Norwich). Perspectives on protein engineering a comprehensive on-line access tool for Web-based resources, papers from Perspectives on Protein Engineering '96 [i.e. POPE 5 & POPE 6], databank guides from Brookhaven & EBI ; works with any WWW browser. BIODIGM Ltd; 1997.

Prusiner SB. Novel proteinaceous infectious particles cause scrapie. Science. 1982 4; 216(4542):136–44. http://www.ncbi.nlm.nih.gov/pubmed/6801762.

Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. Journal of Molecular Biology. 1988 8; 202(4):865–884. http://www.ncbi.nlm.nih.gov/pubmed/3172241, doi: 10.1016/0022-2836(88)90564-5.

R Core Team, R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013. http://www.r-project.org/.

Rose PW, Prlić A, Altunkaya A, Bi C, Bradley AR, Christie CH, Costanzo LD, Duarte JM, Dutta S, Feng Z, Green RK, Goodsell DS, Hudson B, Kalro T, Lowe R, Peisach E, Randle C, Rose AS, Shao C, Tao YP, et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. Nucleic Acids Res. 2016; 13(D1):e1002140.

Sharaf NG, Brereton AE, Byeon IJL, Andrew Karplus P, Gronenborn AM. NMR structure of the HIV-1 reverse transcriptase thumb subdomain. Journal of Biomolecular NMR. 2016 12; 66(4):273–280. doi: 10.1007/s10858-016-0077-2.

Shirota M, Kinoshita K. Analyses of the general rule on residue pair frequencies in local amino acid sequences of soluble, ordered proteins. Protein science. 2013 6; 22(6):725–33. doi: 10.1002/pro.2255.

Spratt DE, Julio Martinez-Torres R, Noh YJ, Mercier P, Manczyk N, Barber KR, Aguirre JD, Burchell L, Purkiss A, Walden H, Shaw GS. A molecular explanation for the recessive nature of parkin-linked Parkinson's disease. Nature Communications. 2013 6; 4:1983. doi: 10.1038/ncomms2983.

Sterckx YGJ, Volkov AN, Vranken WF, Kragelj J, Jensen MR, Buts L, Garcia-Pino A, Jové T, Van Melderen L, Blackledge M, van Nuland NAJ, Loris R. Small-Angle X-Ray Scattering- and Nuclear Magnetic Resonance-Derived Conformational Ensemble of the Highly Flexible Antitoxin PaaA2. Structure. 2014 6; 22(6):854–865. doi: 10.1016/j.str.2014.03.012.

Tamiola K, Acar B, Mulder FAA. Sequence-Specific Random Coil Chemical Shifts of Intrinsically Disordered Proteins. Journal of the American Chemical Society. 2010 12; 132(51):18000–18003. http://pubs.acs.org/doi/abs/10.1021/ja105656t, doi: 10.1021/ja105656t.

Tamiola K, Mulder FAA. Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. Biochemical Society Transactions. 2012 10; 40(5):1014–1020. http://biochemsoctrans.org/lookup/doi/10.1042/BST20120171, doi: 10.1042/BST20120171.

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL. BioMagResBank. Nucleic acids research. 2008 1; 36(Database issue):402–8. doi: 10.1093/nar/gkm957.

Uversky V, Longhi S. Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure and Conformation. Wiley; 2010.

Uversky VN. Introduction to intrinsically disordered proteins (IDPs). Chemical reviews. 2014 7; 114(13):6557–60. doi: 10.1021/cr500288y.

Uversky VN. p53 Proteoforms and Intrinsic Disorder: An Illustration of the Protein Structure-Function Continuum Concept. International journal of molecular sciences. 2016 11; 17(11):1874. doi: 10.3390/ijms17111874.

Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys. 2008 2; 37:215–246. doi: 10.1146/annurev.biophys.37.032807.125924.

Valdar WSJ. Scoring residue conservation. Proteins. 2002 8; 48(2):227–41. doi: 10.1002/prot.10146.

Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki RW, Pierattelli R, Sussman J, Svergun DI, Uversky VN, Vendruscolo M, Wishart D, Wright PE, Tompa P. pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Research. 2014 1; 42(D1):D326–D335. doi: 10.1093/nar/gkt960.

Varadi M, Tompa P. The Protein Ensemble Database. In: *Advances in experimental medicine and biology*, vol. 870; 2015.p. 335–349. doi: 10.1007/978-3-319-20164-1_11.

Varadi M, Vranken W, Guharoy M, Tompa P. Computational approaches for inferring the functions of intrinsically disordered proteins. Frontiers in Molecular Biosciences. 2015 8; 2:45. doi: 10.3389/fmolb.2015.00045.

Vilar M, Chou HT, Lührs T, Maji SK, Riek-Loher D, Verel R, Manning G, Stahlberg H, Riek R. The fold of alpha-synuclein fibrils. Proceedings of the National Academy of Sciences of the United States of America. 2008 6; 105(25):8637–42. doi: 10.1073/pnas.0712179105.

Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins. 2003 9; 52(4):573–84. http://www.ncbi.nlm.nih.gov/pubmed/12910457, doi: 10.1002/prot.10437.

Wang S, Peng J, Ma J, Xu J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports. 2016 1; 6(1):18962. doi: 10.1038/srep18962.

Wang S, Sun S, Li Z, Zhang R, Xu J. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology. 2017 1; 13(1):e1005324. doi: 10.1371/journal.pcbi.1005324.

Wang Y, Jardetzky O. Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc. 2002 12; 124(47):14075–14084.

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. Journal of Molecular Biology. 2004 3; 337(3):635–645. doi: 10.1016/j.jmb.2004.02.002.

Williamson MP. Using chemical shift perturbation to characterise ligand binding. Progress in nuclear magnetic resonance spectroscopy. 2013 8; 73(Pt 12 Pt 1):1–16. doi: 10.1016/j.pnmrs.2013.02.001.

Wishart DS, Case DA. Use of chemical shifts in macromolecular structure determination. Methods in enzymology. 2001; 338:3–34. http://www.ncbi.nlm.nih.gov/pubmed/11460554.

Wishart DS, Sykes BD, Richards FM. The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry. 1992 2; 31(6):1647–51. http://www.ncbi.nlm.nih.gov/pubmed/1737021.

Wishart DS, Bigam CG, Yao J, Abildgaard F, Dyson HJ, Oldfield E, Markley JL, Sykes BD. 1H, 13C and 15N chemical shift referencing in biomolecular NMR. Journal of biomolecular NMR. 1995 9; 6(2):135–40. http://www.ncbi.nlm.nih.gov/pubmed/8589602.

Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews Molecular Cell Biology. 2014 12; 16(1):18–29. http://www.nature.com/doifinder/10.1038/nrm3920, doi: 10.1038/nrm3920.

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. Journal of Proteome Research. 2007 5; 6(5):1882–1898. http://pubs.acs.org/doi/abs/10.1021/pr060392u, doi: 10.1021/pr060392u.

Xue B, Oldfield CJ, Hsu W, Uversky V, Dunker AK. Development of New Predictors of Intrinsically Disordered Proteins and Residues. Biophysical Journal. 2010 1; 98(3):256a. doi: 10.1016/j.bpj.2009.12.1391.

695 **Yoshimura Y**, Holmberg MA, Kukic P, Andersen CB, Mata-Cabana A, Falsone SF, Vendruscolo M, Nollen EAA,
696 Mulder FAA. MOAG-4 Promotes the Aggregation of α-Synuclein by Competing with Self-Protective Electrostatic
697 Interactions. Journal of Biological Chemistry. 2017 5; 292(20):jbc.M116.764886. doi: 10.1074/jbc.M116.764886.

698 **Ytreberg FM**, Borcherds W, Wu H, Daughdrill GW. Using chemical shifts to generate structural ensembles for
699 intrinsically disordered proteins with converged distributions of secondary structure. Intrinsically disordered
700 proteins. 2015 1; 3(1):e984565. doi: 10.4161/21690707.2014.984565.

701 **Yu JF**, Dou XH, Sha YJ, Wang CL, Wang HB, Chen YT, Zhang F, Zhou Y, Wang JH. DisBind: A database of classified
702 functional binding sites in disordered and structured regions of intrinsically disordered proteins. BMC
703 Bioinformatics. 2017 12; 18(1):206. doi: 10.1186/s12859-017-1620-1.

704 **Zahn R**, Liu A, Lührs T, Riek R, von Schroetter C, López García F, Billeter M, Calzolai L, Wider G, Wüthrich K. NMR
705 solution structure of the human prion protein. Proceedings of the National Academy of Sciences of the
706 United States of America. 2000 1; 97(1):145–50. http://www.ncbi.nlm.nih.gov/pubmed/10618385.

707 **Zhang H**, Neal S, Wishart DS. RefDB: a database of uniformly referenced protein chemical shifts. Journal of
708 biomolecular NMR. 2003 3; 25(3):173–95. http://www.ncbi.nlm.nih.gov/pubmed/12652131.

709 **Zhang Z**, Li Y, Loh YR, Phoo WW, Hung AW, Kang C, Luo D. Crystal structure of unlinked NS2B-NS3 protease from
710 Zika virus. Science. 2016 12; 354(6319):1597–1600. doi: 10.1126/science.aai9309.