Journal of Biomolecular NMR (2021) 75:273–291
https://doi.org/10.1007/s10858-021-00374-w

ARTICLE

CheSPI: chemical shift secondary structure population inference

Jakob Toudahl Nielsen[1] · **Frans A. A. Mulder**[1]

Received: 18 February 2021 / Accepted: 11 June 2021 / Published online: 19 June 2021

Abstract
NMR chemical shifts (CSs) are delicate reporters of local protein structure, and recent advances in random coil CS (RCCS) prediction and interpretation now offer the compelling prospect of inferring small populations of structure from small deviations from RCCSs. Here, we present CheSPI, a simple and efficient method that provides unbiased and sensitive aggregate measures of local structure and disorder. It is demonstrated that CheSPI can predict even very small amounts of residual structure and robustly delineate subtle differences into four structural classes for intrinsically disordered proteins. For structured regions and proteins, CheSPI provides predictions for up to eight structural classes, which coincide with the well-known DSSP classification. The program is freely available, and can either be invoked from URL www.protein-nmr.org as a web implementation, or run locally from command line as a python program. CheSPI generates comprehensive numeric and graphical output for intuitive annotation and visualization of protein structures. A number of examples are provided.

Keywords NMR · Protein · Order · Disorder · Chemical shifts

Introduction

NMR chemical shifts are very sensitive to the local structure of proteins, and can be measured and assigned routinely with great precision for both structured and unstructured proteins (Felli and Pierattelli 2012, Brutscher et al. 2015). The relationship between chemical shifts and local protein structure is well-established for folded proteins e.g. through simple index methods (Wishart et al. 1992; Wishart and Sykes 1994a, b) statistic and probabilistic methods (Eghbalnia et al. 2005: Wang et al. 2007) or by methods relying more on sequence homology through neural networks or database searches (Jones 1999; Labudde et al. 2003; Shen and Bax 2013, 2015), and super-secondary structure predictions (Hafsa et al. 2015). Furthermore, CSs have also been used to aid the structural and dynamical characterization of proteins (Wishart and Sykes 1994a, b; Wishart and Case 2001; Berjanskii and Wishart 2007; Cavalli et al. 2007; Mielke and Krishnan 2009; Kjaergaard and Poulsen 2012; Robustelli et al. 2012).

In stark contrast, intrinsically disordered proteins and regions (IDPs and IDRs) display no or very little regular secondary structure, are not folded into a globular structure, but rather constitute a dynamical equilibrium between several conformations with less regularity. NMR spectroscopy is an ideal technique to study these dynamic IDPs (Tompa 2009; Uversky and Longhi 2010). Intrinsically disordered polypeptides display population-averaged chemical shifts that provide an operational definition of random coil chemical shifts. Deviations from RCCSs contain information about structural composition, but also pose a challenge to deconvolute induced from intrinsic structure. Fortunately, however, RCCSs have been predicted with increasing accuracy over time (Braun et al. 1994, Wishart et al. 1995, Schwarzinger et al. 2001, Simone et al. 2009, Tamiola et al. 2010, Kjaergaard et al. 2011), culminating in the most accurate predictor to date, known as POTENCI (Nielsen and Mulder 2018). As a result, deviations from RCCSs (the secondary chemical shifts, SCSs) are now common parameters used to identify and quantitate order/disorder in IDPs (Berjanskii and Wishart 2007; Kjaergaard and Poulsen 2012; Nielsen and Mulder 2016, Sormanni et al. 2017). The improved accuracy has also permitted a benchmark of the performance of disorder prediction methods (Nielsen and Mulder 2019) and CSs

✉ Jakob Toudahl Nielsen
jtn@inano.au.dk

✉ Frans A. A. Mulder
fmulder@chem.au.dk

[1] Interdisciplinary Nanoscience Center (iNANO) and Department of Chemistry, Aarhus University, Gustav Wieds Vej 14, 8000 Aarhus C, Denmark

were recently used to train the disorder predictor ODINPred (Dass et al. 2020).

A yet more challenging task is to quantify the statistical composition of structural states for IDPs, since accurate reference experimental data with *unique* structural interpretation do not exist; As this problem is *ill-posed*, astronomical numbers of ensembles could be constructed that all give rise to the experimentally observed averages. One possible avenue to plausible solutions is to use physics-based models of protein conformational sampling (by e.g. molecular mechanics force fields) coupled to parametrization of chemical shifts, and possibly other NMR observables, to conformation(Ozenne et al. 2012, Varadi et al. 2015). Alternative empirical approaches adopt a heuristic treatment of CSs and secondary structure relationships in folded proteins to IDPs for the inference of secondary structure populations of α-helix, β-strand, random coil, and polyproline II (Camilloni et al. 2012). A more robust reductionist approach was taken with SSP (secondary structure propensity) (Marsh et al. 2006; Tamiola and Mulder 2012) where linear combinations of SCSs were aggregated into a scale between -1 (sheet) and 1 (helix) and interpreted as a structural propensity. Although such an approach implicitly remedies correlated CSs, information potentially contained in the individual CSs risks being lost by the reduction to a single value. For example, a rigid loop between a sheet and a helix will display near-zero propensity, and risks being falsely interpreted as disordered. Furthermore, neither of the above methods can so-far discriminate between disordered and ordered turns, whereas such information is potentially contained in the particular combination of SCSs.

Here, we introduce the linear analysis of signed SCSs, introducing **Che**mical shift **S**econdary structure **P**opulation **I**nference (CheSPI). This approach extends the previously-introduced CheZOD Z-score for quantifying local order and disorder in proteins (Nielsen and Mulder 2016, 2020) derived from the statistical analysis of sums of squared SCSs. Technically, CheSPI applies multivariate analysis and dimension reduction techniques to generate linear combinations (CheSPI components) of SCSs that optimally describing the variance: The first CheSPI component ensures the optimal distinction between secondary structure classes, whereas the second component accounts for the variance within the classes (which was found to be closely related to the local structure such as backbone conformation, flexibility and hydrogen bonding). CheSPI components offer an accurate and comprehensive quantification of local dynamic and structural composition in structured as well as disordered proteins, being sensitive to local protein structure as well as dynamics. CheSPI components are presented to the user as a color scale, which conveys the information in a simple, intuitive, and visually appealing manner. As shown herein, CheSPI colors can be used to annotate 3D structures, and thereby highlight important and detailed structural changes in proteins.

The power of CheSPI to discriminate between secondary structure classes was exploited to derive estimates for the populations of helix, extended structure, turn, and "non-folded" structures (CheSPI populations) through statistical inference, and these populations were validated through comparison to simulated ensembles for four distinct IDPs (vide infra). Contemporary methods for inferring secondary structure from NMR data typically provide only three-class predictions. A notable exception is CSI3.0, which offers a four-state prediction including turns, as well as the distinction between internal and external strands. CheSPI takes a stride further, and extends secondary-structure inference to encompass the prediction of the eight structural classes (SS8) defined by the popular DSSP classification algorithm (Kabsch and Sander 1983); α-helix (H), 3_{10}-helix (G), π-helix (I), extended β-strand (E), bridge (B; isolated single residue β-strand), turn (T), and—if no well-defined pattern can be found—"bend" (S) and "coil" (C).

Availability

CheSPI is available at www.protein-nmr.org and source code can be obtained from https://github.com/protein-nmr. The CheSPI analysis output is summarized in text files as well as in a combined plot containing three panels visualizing for each residue along the sequence: (i) CheSPI colors and CheZOD Z-scores, (ii) CheSPI populations, and (iii) CheSPI 8-state secondary structure predictions.

Results

Derivation of CheSPI components for secondary structure

NMR chemical shifts (CSs) are very sensitive to the local structure and dynamics in peptides. To analyze this correspondence, we used two previously derived sequence databases with CSs for proteins deposited in the BMRB database. The first database contains primarily disordered residues used to parametrize POTENCI (Nielsen and Mulder 2018), whereas the other, derived from the RefDB database (Neal et al. 2003), contains primarily structured residues. Secondary chemical shift (SCSs) were derived by subtracting random coil shifts derived by POTENCI corresponding to the backbone atoms and Cβ, and Hβ. For the disordered database, residue data were labeled with "D" for disorder if their CheZOD Z-score (derived by CheZOD (Nielsen and Mulder 2016)) was less than 5.0, or else "O" for order. Analysis with DSSP (Kabsch and Sander 1983) was performed

for all proteins in the structured database, and here residue data were labeled using the 8-class DSSP secondary structure designations. Subsequently, a supervised modelling approach was applied for dimensionality reduction, which optimizes the discrimination between the different classes (or equivalently, maximizes their separation). This procedure, called Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA), resembles PCA analysis (Worley and Powers 2016), but quantifies the variance *between* the classes with the first principal component whereas the variation *within* the classes is captured in the other dimensions (Trygg and Wold 2002, Bylesjö et al. 2006, Bradley and Robert 2013). The optimal weights for the principal components were derived using Simca-P (Wu et al. 2010) as defined in Eqs. 1–3, Methods.

Figure 1 shows the loading plot that visualizes the optimized weights, scaled according to the gyromagnetic ratio for each nucleus. The relative magnitude of the optimized weights reflects the well-documented sensitivity of SCSs towards secondary structure (Hα/C'/Cα /Cβ » N/Hβ/H_N), with Hα the hydrogen most sensitive to structure, and the three carbon-13 nuclei displaying a similar importance. In addition, N/H_N/Hα/Cβ display opposite SCSs compared to C'/Cα/Hβ. HSCSs correlate with secondary structure as commented in Supplementary Results 1 and Supplementary Fig. S1. For the second component, all weights are positive and display the following sensitivity: (H_N > C'/Cα/Cβ > N > Hα > Hβ). It is noteworthy, that the two chemical shifts with the largest magnitudes, H_N and C', were previously found to undergo the largest chemical shift changes upon backbone hydrogen bonding (Nielsen et al. 2012). The third CheSPI component did not add any significant value to classification.

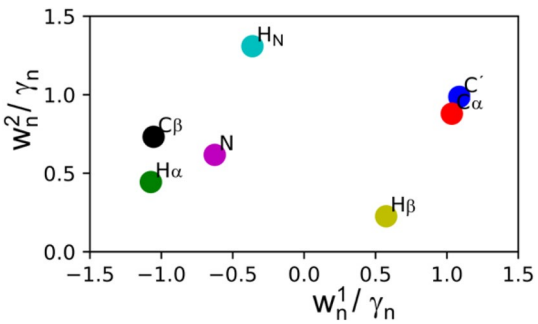

Fig. 1 CheSPI component loading plot showing the weights (Eq. 1, Methods) scaled by gyromagnetic ratio for each nucleus

CheSPI components discriminate between local folded structures

CheSPI components were calculated for the 809 proteins in the structured database using the weights optimized by the OPLS-DA procedure (Eqs. 1–3, Methods). Figure 2 shows two-dimensional histograms of the combined observation of the first two CheSPI components for the three canonical secondary structure types helix, sheet, and coil as determined by DSSP. It is clear how the first component (P^1) offers a near-perfect discrimination between helix and sheet whereas coil, placed around the middle of the score plot, overlaps with the helix and sheet classes, but has a lower average value for the second CheSPI component (P^2) (see also Supplementary Results 1 and Figure S1). For reference, it is noted that disordered residues have near-zero values for both of the first CheSPI components, as expected.

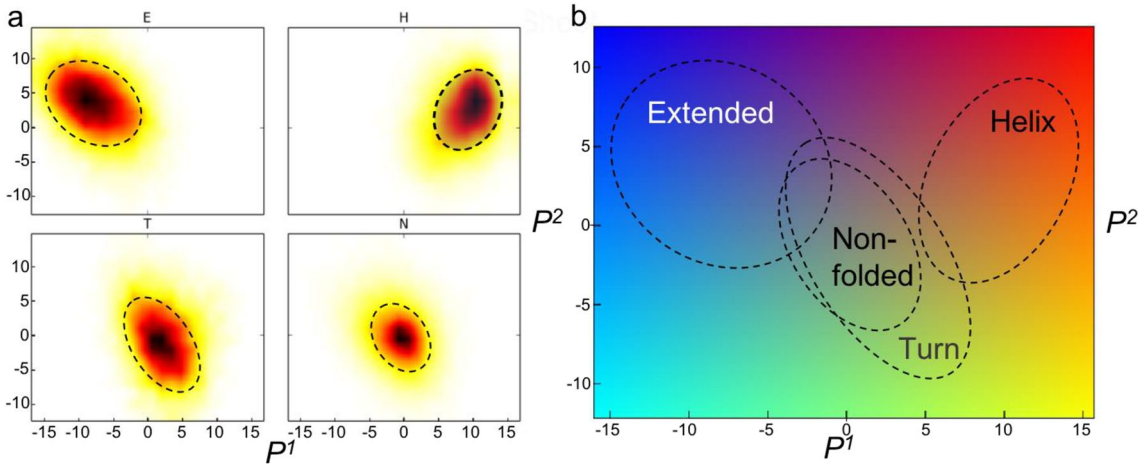

Fig. 2 **a** Experimental distributions of the first two CheSPI components (Eq. 1–3, Methods) for the secondary structure classes; helix (H), extended (E), and turn (T) and non-folded backbone conformations (N, everything not helix, sheet, or turn) with ellipsoids marked by dashed lines. **b** overlay of ellipsoids from (a) merged with CheSPI color scheme

Encouraged by the strong relationship to local structure, CheSPI components were converted to a color scale, using linear combinations of the first two components (see Eqs. 4 and 5, Methods). This CheSPI color scale provides a visual interpretation of the CheSPI components and an intuitive overview of the local structure and dynamics of proteins. The CheSPI application produces both bar plots and a script for 3D structure visualization based on CheSPI colors and several examples are discussed below (see Discussion). On this scale, well-formed strands and helices are defined by blue and red colors, respectively, coil can display multiple colors depending on context, turns are in green, and disordered residues are grey. The variation in CheSPI components along the secondary elements, as described above, is reflected in hues changing from red through orange to yellow at the C-terminal ends of helices, and green at the ends of β-strands, which sometimes have lighter blue or purple CheSPI colors (see Fig. 4a for an example of CheSPI colors).

To further demonstrate the power of the CheSPI components to discriminate between different local structures and describe variation within structures, we analyzed the eight DSSP classes. These were further separated into subclasses based on hydrogen bonding or local backbone structure as visualized in Fig. 3 (see also Fig. S2 for individual histograms showing the variability within each class). Average values of the first two CheSPI components for all subclasses are visualized in Fig. 3. First, it is seen how the first CheSPI component (P^1) roughly segregates the average values for the 8 DSSP classes with E < B < C < S < T < G < I < H, with the more extended conformations having the most negative values, in general. Secondly, the second CheSPI component (P^2) shows negative values for turns and positive ones for helices and strands. Notably, P^2 accounts for the variation within the same DSSP class. For the helix and turn classes, it appears that in case of H_N, hydrogen bonding reduces the value of P^2, whereas for C', H-bonding increases it. This tendency reflects the larger contribution from H_N and C'

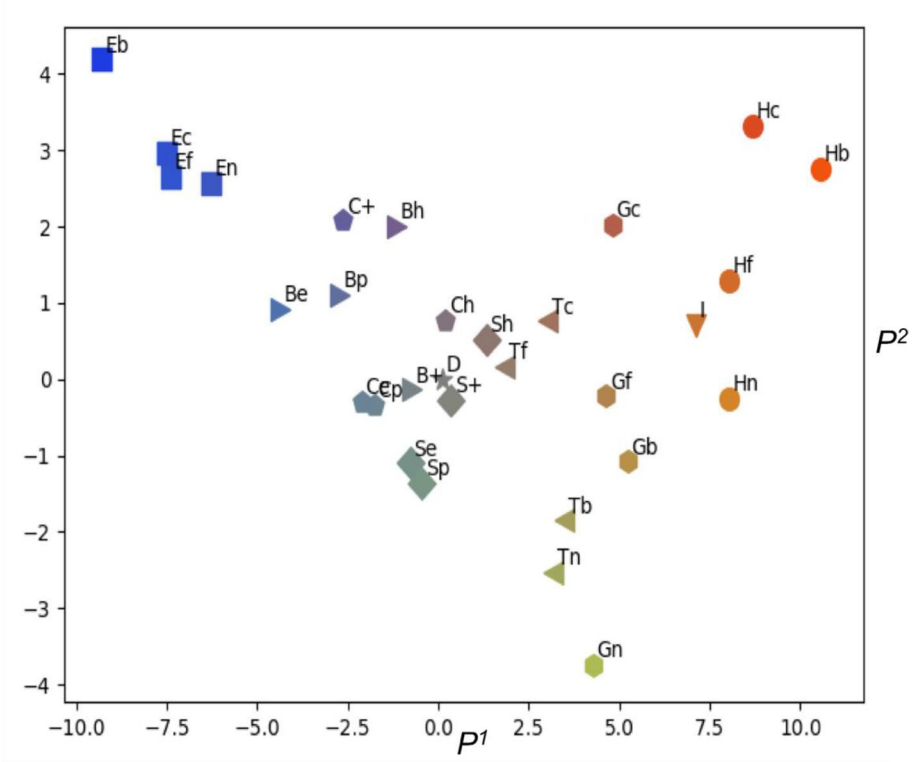

Fig. 3 Average values of the first two CheSPI components for all 8 DSSP secondary structure classes with subdivisions. Average values were derived from the set of structured proteins. The average points are labelled with two letters (except I-helix and disorder), where the first letter indicates the DSSP classification ("C" indicates coil). The major classes E(β-sheet)/H(α-helix)/G(3_{10}-helix)/T(turn) are labeled to indicate presence of hydrogen bonds (HB) using: n/c/b/f corresponding to NH hydrogen bonding only, C' hydrogen bonding only, both NH and C' hydrogen bonding, and none (free), respectively. For the remaining classes, B(bridge), S(bend) and C(coil), points are labeled according to the local backbone structure with: +/h/p/e corresponding to "positive"/"helical"/"poly-proline II"/"extended", respectively, using the definitions in Table 1. The eight classes are shown with different marker symbols: circle, square, hexagon, triangle pointing left (left triangle), right triangle, down triangle, diamond, and pentagon, respectively for classes H/E/G/T/B/I/S/C and visualized using CheSPI color fills corresponding to CheSPI components (see Eq. 4 and 5, Methods). The corresponding point for disorder is indicated at the origin with a star, for reference

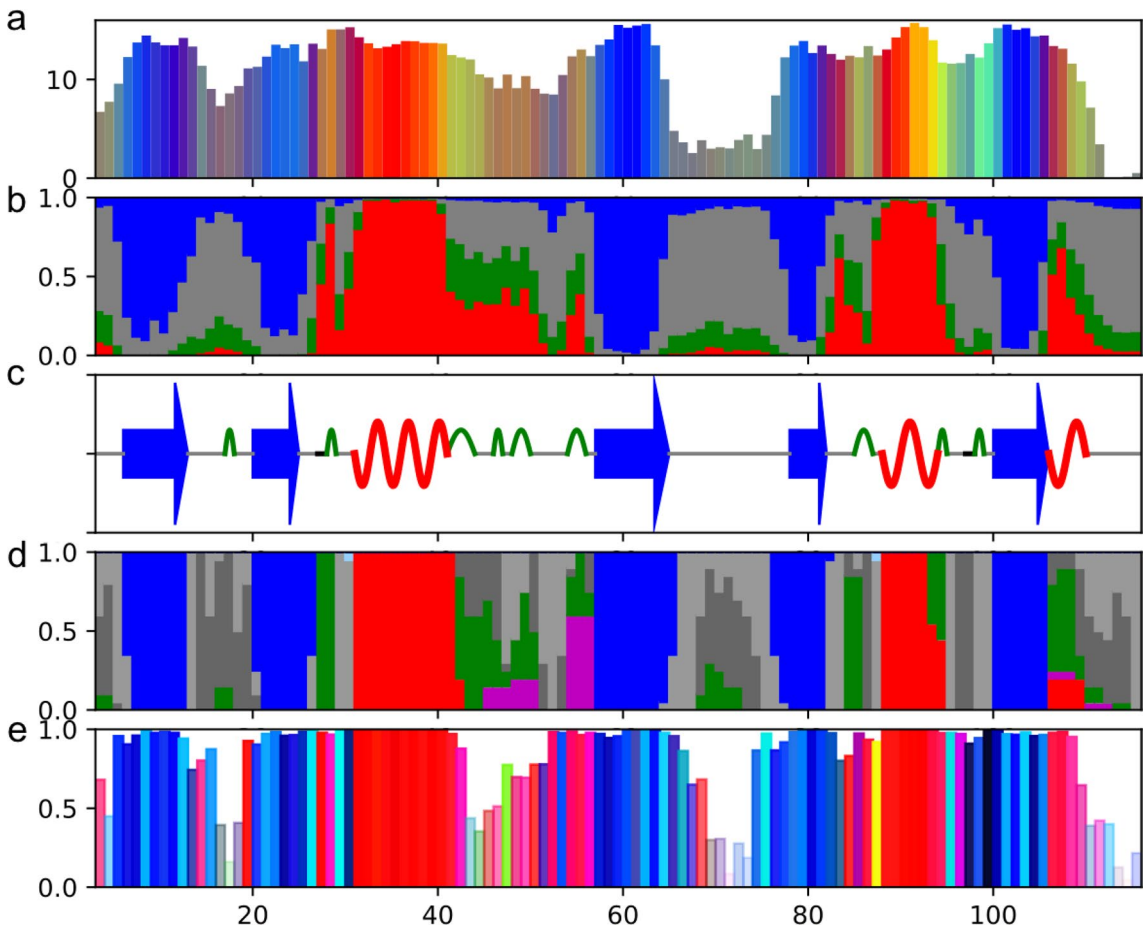

Fig. 4 Output multi-panel plot produced by CheSPI using chemical shifts from BMRB id 6635 and validation from structural ensemble for PLCε-RA2 (Bunney et al. 2006) (PDB id 2BYF). **a** Bar plot with CheSPI colors (Eq. 5, Methods) with bar height equal to the CheZOD Z-scores. **b** Stacked bar plot of CheSPI populations of "extended" (blue), "helical" (red), "turn" (green), and "non-folded" (grey), local structures, corresponding to DSSP classes of sheet and bridge (E/B), all helix-types (G/H/I), turn (T) and, finally, the remaining bend and coil (S/-), respectively (see text). **c** Cartoon of the most confident CheSPI prediction of eight class DSSP secondary structure using red curved lines for α-helix (H), magenta for 3_{10}-helix (G), blue arrows for sheets (E) and bridges (B), green arcs for turns (T), and grey and black lines for coil (C) and bend (S), respectively. **d** Stacked bar plot visualizing the observed conformations in the structural ensemble 2BYF for the 8 DSSP classes using same colors for helices, turns, and lighter and darker grey colors for coil and bend, respectively. **e** Bar plot for backbone angle conformation and variation. The heights of the bars were set to the geometrical average of the squared angular order parameters (Hyberts et al. 1992) for ϕ and Ψ backbone torsion angles (Eqs. 14–16, Methods) where values close to unity indicate local order and values close to zero indicate structural disorder (large angular variation). The colors were taken from a 2D-color-scale (see Fig. S3) based on the position in Ramachandran plot of pairs of ϕ and Ψ backbone torsion angles using trigonometric averages of the ensemble values (see Eq. 16, Methods). With this scale, backbone angles in the helical domain of the Ramachandran map appear in red as before, and extended β-sheet-like conformations have blue colors. Furthermore, left-twisted β-strands as well as fragments with PPII structure appear with cyan colors, whereas conformations with positive ϕ have yellow and green colors, and finally, other conformation referred to elsewhere as "forbidden" in Ramachandran space are shown in black. Transparency is added to the bars using the above local angular order parameters as the "alpha value". See also Figure S1

chemical shifts in the definition of P^2 as described above. Conversely, the effect of hydrogen bonding is less pronounced for P^2 in the case of strands. Here, strands with both H_N and C' hydrogen bonding ("Eb") have larger magnitudes for both CheSPI components. Such "Eb" strands correspond to the inner strands of β-sheets and are identified by deeper blue CheSPI color (see Discussion below). The variation in P^2 within strands likely relates to their variability in local structure, including twists, bends, and bulges, as witnessed by examples given below. Finally, the classes with lower tendencies to form hydrogen bonds, B (bridge), S(bend) and C(coil), have CheSPI components closer to zero. The more extended conformations "extended" and "poly-proline II" (PPII) (see legend to Fig. 3) show lower values for P^1, while the more compact conformations, "helical" and "positive Ψ", display larger values for P^2. Markedly, "extended" and PPII conformations have almost indistinguishable CheSPI components for bend and coil, whereas for bridges, "extended"

displays slightly lower values for P^2 in agreement with the closer resemblance to canonical β-sheet.

Prediction of secondary structure populations

Intrinsically disordered proteins are in a dynamic equilibrium between different local conformations. Encouraged by the discrimination power of the CheSPI components (Figs. 2 and 3), we derived an estimate for the population of different local structural states based on the relative likelihood for observing a given combination of the first two CheSPI components as defined in Eq. 6, in Methods. **Che**mical shift **S**econdary structure **P**opulations **I**nference (CheSPI) are provided for classes referred to here as "extended", "helical", "turn", and "non-folded". This classification is based on inference from statistics of CheSPI components measured in the structured proteins set for DSSP classes for strand and bridge (E/B), all helix-types (G/H/I), turn (T) and, finally, the remaining non-folded bend and coil (S/-) classes, respectively.

We now turn our attention to a small number of examples for demonstrating the utility of CheSPI. First, we analyze in detail the NMR solution structural ensemble of the phospholipase c epsilon RA 2 domain (Bunney et al. 2006) (PLCε-RA2, henceforth) as summarized in Fig. 4 (see also Figure S1). PLCε-RA2 contains 5 β-strands, two α-helices and two shorter 3_{10}-helices which are modelled in some of the members of the deposited NMR ensemble (PDB id 2byf). The α-helices are predicted with close to 100% population throughout. Although for the β-strands, populations of about 90% "extended" were predicted for the central residues, relatively lower estimates are obtained at both ends of the strands, echoing the fall-off in CheSPI component amplitudes at β-strand ends described above. At the same time, significant populations for "extended" were predicted in loop segments for residues flanking the β-strands, mirroring the gradual change from strand character to flexible coil, bracketing β-sheets. More interestingly, disordered regions are characterized by a composition of different local structures, and the termini and the long loop region (residues 66–76) and classified as disordered, when judged by CheZOD Z-scores < 8.0. For these regions, larger "non-folded" populations (> 70%) were indeed predicted by CheSPI. Of note, residues 68–76 have missing density in the corresponding X-ray structure (Bunney et al. 2006) (pdb id: 2C5L) suggesting that they may be dynamically or statically disordered. In the corresponding NMR structure ensemble (Fig. 5) (Bunney et al. 2006), a mixture of bend and coil conformations as well as a few turns are observed for this long loop. Averaging of the local backbone conformations in this loop leads to small angular order parameters (Hyberts et al. 1992; Nielsen and Mulder 2019) (Fig. 4e) indicative of increased local disorder. Average chemical shifts for the individual conformations lead to observations very close to random-coil chemical shifts for these residues, and thereby low CheZOD Z-scores. Interestingly, for the other long loop (residues 45–56), CheSPI estimated intermediate populations of α-helix (ca. 30–40% for residues 45–50 and 54–55), and comparison with the NMR ensemble reveals fractional populations of two 3_{10}-helices for residues 45–50 (15–20% of the members in the ensemble) and residues 54–56 (60%).

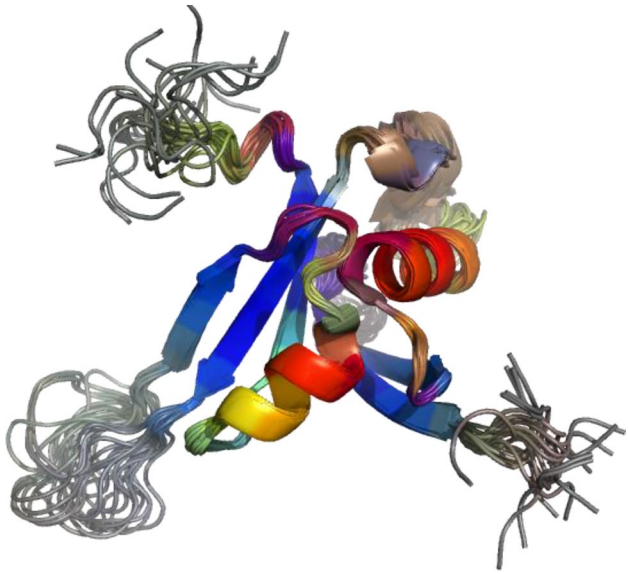

Fig. 5 Structure visualization using CheSPI colors for the NMR solution structural ensemble of PLCε-RA) (Bunney et al. 2006) using chemical shifts from BMRB id 6624 and PDB id, 2BYF

Validation of canonical secondary structure populations in disordered protein ensembles

Although the range of conformations in an ensemble of structures determined by NMR spectroscopy in solution is evidence of the dynamics of the system, the generated models also reflect the precision of the structure, and depend on the type, quality and number of geometrical constraints as well as the structure determination protocol. More advanced computational protocols (Bernadó et al. 2005, Jensen et al. 2010, Marsh and Forman-Kay 2012, Ozenne et al. 2012, Camilloni et al. 2013, Varadi et al. 2015) use extensive conformational sampling, culled by data from NMR spectroscopy and small-angle X-ray scattering (SAXS). To gauge how well CheSPI-derived populations compare with the composition of local structure, we made comparisons for four protein systems: (i) The K18 domain from Tau, a human intrinsically unstructured protein implicated in Alzheimer's disease pathology (Cleveland et al. 1977). K18 was previously investigated by NMR chemical shifts, residual dipolar couplings (RDCs), and paramagnetic relaxation

enhancements (PREs) (Mukrasch et al. 2007), and an ensemble of structures was computed (Ozenne et al. 2012) using a combination of the ASTEROIDS (Jensen et al. 2010) and Flexible Meccano (Bernadó et al. 2005) protocols (see also Discussion); (ii) The unfolded state of drkN SH3. A structural ensemble of this small domain has been generated with the ENSEMBLE software (Marsh and Forman-Kay 2012) based on an extensive amount of experimental data (Marsh and Forman-Kay 2009) including CSs, RDCs, SAXS, PREs, and ^{15}N R_2 relaxation data. Here we compare CheSPI populations based on the assigned chemical shifts (Lee et al. 2015) (iii) The PaaA2 antitoxin. This protein contains two partially formed helices, and was modelled based on a combination of NMR data, filtering with SAXS, and cross-validation with RDCs (Sterckx et al. 2014). These three ensemble structures were taken from the pE-DB protein ensemble database for IDPs (Varadi et al. 2014). (iv) The oncogene protein E7 of human papillomavirus type 16 (Kukic et al. 2019), which contains both ordered and disordered domains with an ensemble of conformations

Table 1 Definitions of local backbone conformations

	ϕ	Ψ
positive ϕ	$0° < \phi < 150°$	all
helical	$\phi < 0°$ or $\phi > 150°$	$-120° < \Psi < 50°$
poly-proline II	$-105° < \phi < -45°$	$115° < \Psi < 175°$
extended[a]	n.a.[a]	n.a.[a]

[a]All which is neither positive ϕ, helical, or poly-proline II

similulated by replica-averaged metadynamics (RAM) simulations (Camilloni et al. 2013) based on assigned CSs and RDCs. In all cases, the local secondary structure was calculated by DSSP for each member of the ensemble. The latter was stratified into populations of helix (H/G/I DSSP classes), β-strand (E/B), and a coil class. "Coil" was further divided according to the backbone conformations as described above (see legend to Table 1).

A good correlation between "observed" fractions of formed helix conformations in the ensembles is seen in Fig. 6, with Peason correlation coefficient 0.915, and a similar degree of correlation for strand structures (R = 0.911). For comparison, the δ2D algorithm predicts populations of helix and strand with correlations to the observed fraction of populations in the ensembles of 0.85 and 0.78, respectively (see also Fig. S4 for the corresponding correlation scatter plot). Furthermore, ncSPC yields comparable correlations of 0.87 and 0.79 for helix and strand (Tamiola and Mulder 2012), respectively, when interpreting positive secondary structure propensities (SSPs) as helix fraction and the absolute of the negative SSPs as strand fraction (Fig. S4). The first CheSPI component is closely related to the ncSPC scale and gives correlations of 0.91 and 0.89 to helix and strand populations, respectively, which is close to the aggrement between CheSPI populations and observed fractions. When considering residues with local "helical" backbone structure as helix, and similarily, residues with local "extended" backbone structure as strand, the correlations between predicted and observed population decrease (between R = 0.63

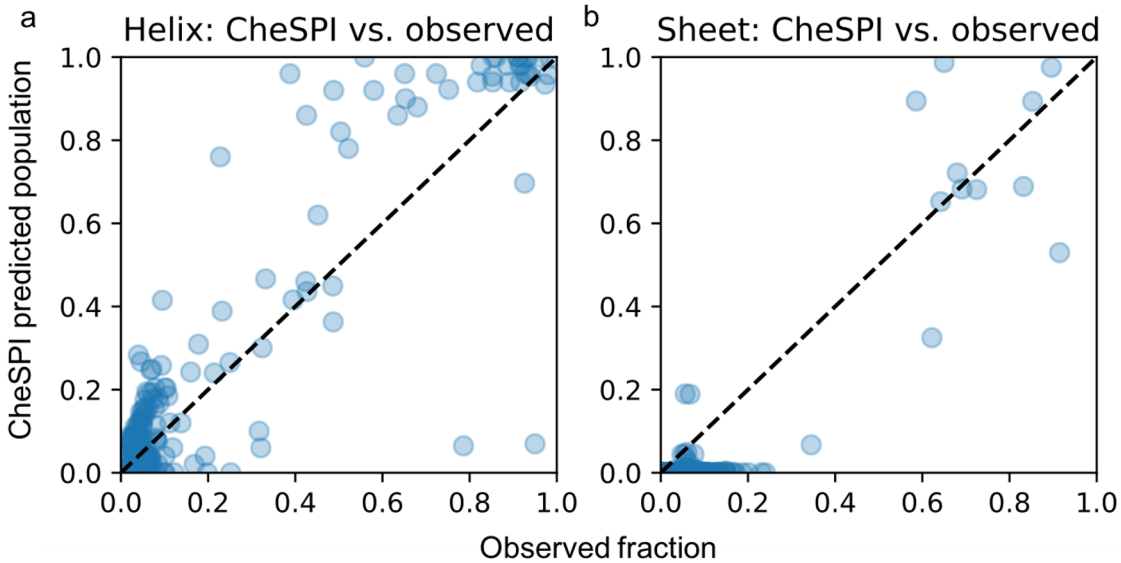

Fig. 6 Correlation between **a** helical and **b** sheet populations predicted by CheSPI and observed in experimental ensembles. Each residue data is indicated with a blue disk. Populations were using data from (BMRB id, pE-DB id) = K18 Tau: (19253, 6AAC), PaaA2: (18841 5AAA), DrkN SH3 unfolded; (25501, 8AAC), E7: (BMRB id´s 19442 and 26069 for residues 3–45 and 46–97, respectively, using coordinate data provided by the authors)

and 0.77) but the ranking of the performance of the methods is preserved.

Prediction of secondary structure according to eight-class DSSP

CheSPI provides estimates for the populations of local structure types, whereas folded proteins are more commonly described by segments of completely formed regular structure. The classical DSSP algorithm assigns each residue to one of eight classes (Kabsch and Sander 1983), which are more informative about the local structure than the traditional coarse-grained three-state canonical division of secondary structure. Unfortunately, to date no tool exists that can infer this 8-class DSSP structure from NMR chemical shifts and sequence alone. Therefore, we extended the CheSPI analysis to the prediction of secondary structure segments and 8-class DSSP secondary structure (SS8) using only protein sequence and assigned NMR chemical shifts as input. As presented above, CheSPI reveals clear trends of secondary structure in its principal components. To quantify this dependence, we thus defined a linear approximation back-calculation of the CheSPI components (PCs) based on sums of contributions from the (four) nearest neighboring SS8 and amino acid types (see Eqs. 7–9 Methods), e.g. the first component would decrease if the center residue was strand and to smaller degree if the neighboring residues were strand. All weights were parametrized by linear regression, similar to the POTENCI implementation (Nielsen and Mulder 2018), using observed PCs derived from the structured database as targets and observed SS8 and sequence as input variables (see Methods). By this procedure, the first PC was predicted with an average error of 1.97, 3.63, and 2.94 for helix, strand, and coil, respectively (the full span of PCs is almost one order of magnitude larger). This prediction of PCs was used together with the average error to estimate a likelihood of observing any PC given an SS8 assignment and the sequence (Eq. 10, Methods). Bayes' theorem was then used to "invert" the probabilities, i.e. to give the probability of an SS8 given the observed PCs. By this procedure, the prior probability for the secondary structure is "updated" by multiplying with the likelihood of observing the PCs given the secondary structure and sequence (Eq. 11, Methods). Finally, the predicted secondary structure is the combination of SS8 states for all residues that give the maximal posterior probability (Eq. 12, Methods). Unfortunately, SS8 states for a residue cannot be optimized univariately, since the value of predicted PCs depends on the nature of the neighboring SS8s. Therefore, the posterior probability maximization algorithm was implemented using a genetic algorithm (as in POTENCI) to produce populations of SS8 assignments for the full sequence. The SS8 predictions are rapidly calculated (2–3 s). Finally, the SS8 assignment from the population with the lowest energy is considered the best prediction, and the variation within the population at each site, along with the agreement between predicted and observed PCs, is used to estimate a confidence for each SS8 prediction.

For PLCε-RA2 (Fig. 4), CheSPI detects all secondary structure elements and identifies the borders of these elements (Fig. 4c) with high accuracy when compared to the observed secondary structures (Fig. 4d). The disordered stretches are predicted as coil (majority). The fractionally-occupied 3_{10}-helices are predicted as turns by CheSPI, which is not very surprising given the similarity between 3_{10}-helices and type I β-turns (Shapovalov et al. 2019).

To derive a systematic evaluation of CheSPI SS8 predictions, a validation set was generated by considering all entries in the BRMB database published after the newest version of the RefDB database, which was used to derive our first structured database for optimizing the CheSPI weights and perform inference. We aimed for a small validation set, keeping only entries with (i) less than 30% sequence identity to all sequences in the structured database, (ii) all backbone chemical shifts available including Hβ, (iii) a high quality X-ray structure for the corresponding sequence ($R < 2.0$ Å) with exact sequence identity to the NMR study, (iv) no biasing conditions in either the derivation of the NMR assignments or the X-ray structure [e.g. standard buffer conditions as before (Nielsen and Mulder 2016) and no large ligands present]. This procedure yielded 13 protein entries (see Table S1) with assigned chemical shifts, which were used to generate CheSPI SS8 predictions and comparison to the observed secondary structure classes, as calculated with DSSP from the high-quality X-ray structures. CheSPI achieves a good accuracy with 68.6% (between 53.2% and 80.6%) correct 8-class predictions (Q8), and 84.6% correct for the classical three-class predictions (Q3), which is an improvement by 2.7% relative to the 81.7% (Q3) for CSI 3.0 (see Table S1). Furthermore, considering only the CheSPI predictions with the highest confidence (28% of cases), CheSPI performs with 94% for SS8 accuracy. High confidence is typically found in the middle of the secondary structure elements and in long disordered loops, whereas lower confidence is more likely to be observed at the borders between regular secondary structure and loop elements.

Discussion

In this paper, we have introduced CheSPI components—derived from NMR secondary chemical shifts —that provide an aggregate descriptor of local structure and dynamics for both structured and disordered proteins. CheSPI components estimate the populations of secondary structure, and are visualized using color, rather than the previously-published SSP and ncSPC procedures (Marsh et al. 2006; Tamiola

et al. 2010; Tamiola and Mulder 2012), which present a scale bar to differentiate only the two most common secondary structure *propensities* (SSPs). The first CheSPI component is similar to the SSP scale in its power to discriminate between different secondary structures, and gives comparable values (see below). CheSPI, however, supersedes ncSPC by the introduction of a second component that affords to account for the variation *within* structural classes—and thereby offers a far more comprehensive discrimination of local structure and dynamics in proteins. Alternatively, secondary structure populations can also be predicted by δ2D in order to stratify residues as "helix", "beta", "poly-proline II" and "coil". CheSPI takes this differentiation further by its sensitivity to discriminate distinct non-folded and folded "turn" coil types from NMR chemical shifts. This advance is possible, as CheZOD Z-scores facilitate the appropriate classification of non-canonical local structure and dynamics.

To feature the potential of CheSPI for detailed structural analysis using NMR chemical shifts, we provide a few examples below. These examples demonstrate that small, but important changes in solution structures can be characterized from NMR chemical shifts, which may otherwise be difficult or impossible to capture.

Metal binding and aggregation of Cu/Zn superoxide dismutase 1

As a first example, we focus on CS data for the protein Cu/Zn superoxide dismutase 1 (SOD1) (Milani et al. 2011) in the *apo* and metal-bound *holo* states. The misfolding of SOD1 is linked to familial amyotrophic lateral sclerosis (ALS) (Rosen et al. 1993, Robberecht et al. 1994). SOD1 contains a double β-sandwich structure, where two long loops that are disordered in the *apo* form become structured upon metal binding (Rakhit and Chakrabartty 2006, Teilum et al. 2009, Sirangelo and Iannuzzi 2017). In Fig. 7, CheSPI analysis reveals a clear difference between the *apo* and *holo* forms. CheZOD Z-scores (Fig. 7a, b) confirm that the bound form is structured, whereas the two largest loops (IV and VIII) are unstructured in the *apo* form. The first CheSPI component (related to secondary structure propensity) remains close to zero and doesn't change much between the two states of the protein. On the other hand, the second CheSPI component changes to negative for the bound state, indicative of the formation of turn structure. This change from non-folded to folded coil is apparent in the changes from grey to green on the CheSPI color scheme for loops IV and VIII (Fig. 7c,d). Structure determination (Banci et al. 2002) clearly shows how metal-ion coordination induces folding of these two large loops (Figs. 7h and 8), which become enriched in turn structure. Using NMR relaxation dispersion measurements, Teilum and co-workers identified a weakly populated exited state of apo-SOD1, which is believed to trigger deviant oligomerization (Teilum et al. 2009). They showed that the largest structural changes between the *apo* ground and exited states involves the flexible loops IV, VI, and VII, as well as β-strands 4, 5, 7, and 8. While native dimers form through association of pairs of β1, it was discussed how excited-state exposure of edge strands 5 and 8, which are protected by turn structures in the metal bound form, could initiate the oligomerization process. Extensive aggregation is avoided by negative design (Richardson and Richardson 2002) of β5 and β8, which are more twisted and less hydrogen-bonded (Fig. 8b,c), contain less β-sheet in the structural ensemble, which is slightly higher for the *apo* form (Fig. 7h), and have fewer canonical β-sheet backbone angles (Fig. 7i). This is reflected in the lower populations of extended structure for these strands estimated by CheSPI (panels e–g). In contrast, β1 is fully formed in the ensemble with canonical backbone angles (Fig. 7h, i). Non-native inter-molecular contacts were identified for several residues in loop IV, in particular residues H63 and F64 (no data available for residues 65–66) through paramagnetic relaxation enhancement (Teilum et al. 2009). Intriguingly, CheSPI analysis pinpoints a small segment, residues 61–66, with noteworthy local order within this loop (Fig. 7a, c). Residues 61–63 show significant "extended" CheSPI populations, whereas residues 64–65 have elevated helical population (Fig. 7e). Whereas H63 forms a β-bridge in the metal-bound structure, flanked by residues with extended conformations, residues 65–66 mostly populate helical backbone conformations in the ensemble structure (Fig. 7h, i). Similarly, residues 132–137 form a helix within loop VII of the *holo* form, with a pronounced peak of helical populations that is mirrored in the *apo* structure (Fig. 7e). CheSPI analysis suggests that the locally ordered residues 65–66 might initiate non-native oligomerization through contacts within preformed extended structure.

Poly-proline II formation in an antifreeze protein

Poly-proline II helix (PPII) conformations (which are often, but not necessarily, rich in prolines) are relatively rare in folded proteins, although they appear to be important for certain molecular recognition events (Adzhubei et al. 2013). In contrast, PPII has been suggested to be more prevalent in IDPs (Shi et al. 2006) although it is not a generic conformation for IDPs, but part of the statistical composition of local structural states (Jha et al. 2005, Makowska et al. 2006). δ2D predicts populations of the classical folded α-helix and β-sheet states, and populations of either so-called "coil" or "PPII", which are both considered as disordered by δ2D. Similarly, CheSPI predicts "helical" and "extended" populations, but separates the remaining states into "turn" (folded DSSP turn class) or "non-folded" (combined bend and coil DSSP states). In order to scrutinize the relationship

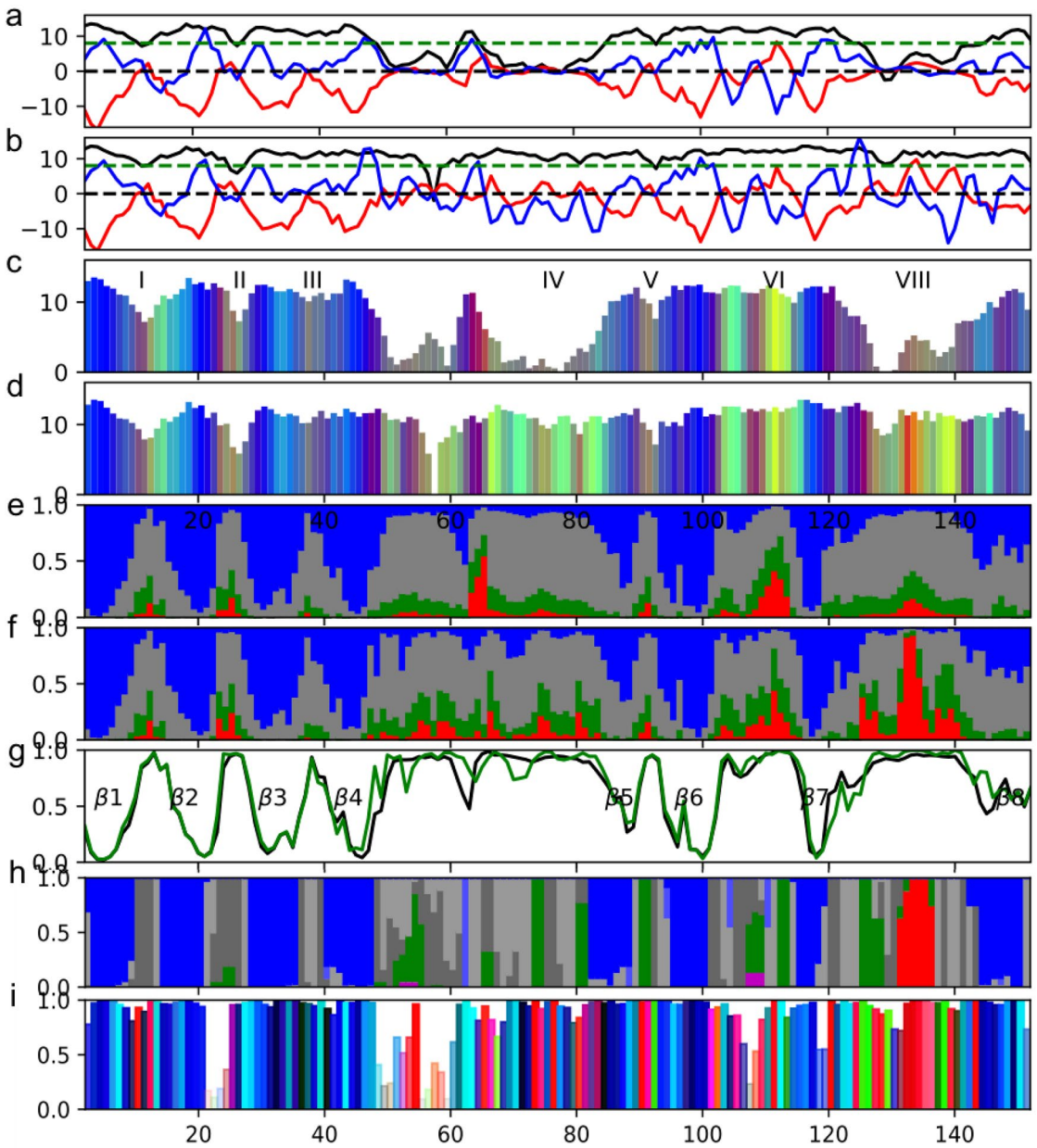

Fig. 7 CheSPI analysis of Cu/Zn superoxide dismutase 1 (SOD1). **a** and **b** CheSPI components and CheZOD Z-scores for apo-SOD1 (see legend to Figure S1b) and using CSs from entry with BMRB id 15711 and Cu/Zn-bound state of SOD1 (BMRB 4402), respectively. **c** and **d** CheSPI colored bar plot for apo-SOD1 and Cu/Zn-bound SOD1, respectively (see legend to Fig. 4a). Loops are labeled using Roman numbers. **e** and **f** CheSPI populations for apo-SOD1 and Cu/Zn-bound SOD1, respectively (see legend to Fig. 4b). **g** plot of CheSPI predictions for "extended" populations for apo-SOD1 (black curve) and Cu/Zn-bound SOD1 (green), respectively. β-strands are labeled consecutively with Greek letter and Arabic numbers. (h) Local structure observed population in structural ensemble for and Cu/Zn-bound SOD1 (PDB id 1ba9) (see legend to Fig. 4d). (i) Average backbone angles and angular order parameters for Cu/Zn-bound SOD1 (PDB id 1ba9) (see legend to Fig. 4e)

between CSs and PPII, and in particular the CheSPI/CheZOD signatures possibly related to PPII, we analyzed spectroscopic and structural data for the left-handed helical bundle of Hypogastrura harveyi "snow flea" antifreeze protein (sfAFP), which is rich in Gly-Gly-X repeats (Graham and Davies 2005), and for which both an X-ray structure (Pentelute et al. 2008) as well as extensive NMR data including CSs are available (Pentelute et al. 2008). sfAFP has a compact structure of a bundle of six PPII helices connected by hydrogen bonds alternating between inter-strand neighbors with a three-residue periodicity (Fig. 9h). Relaxation measurement confirmed the rigid backbone—except for

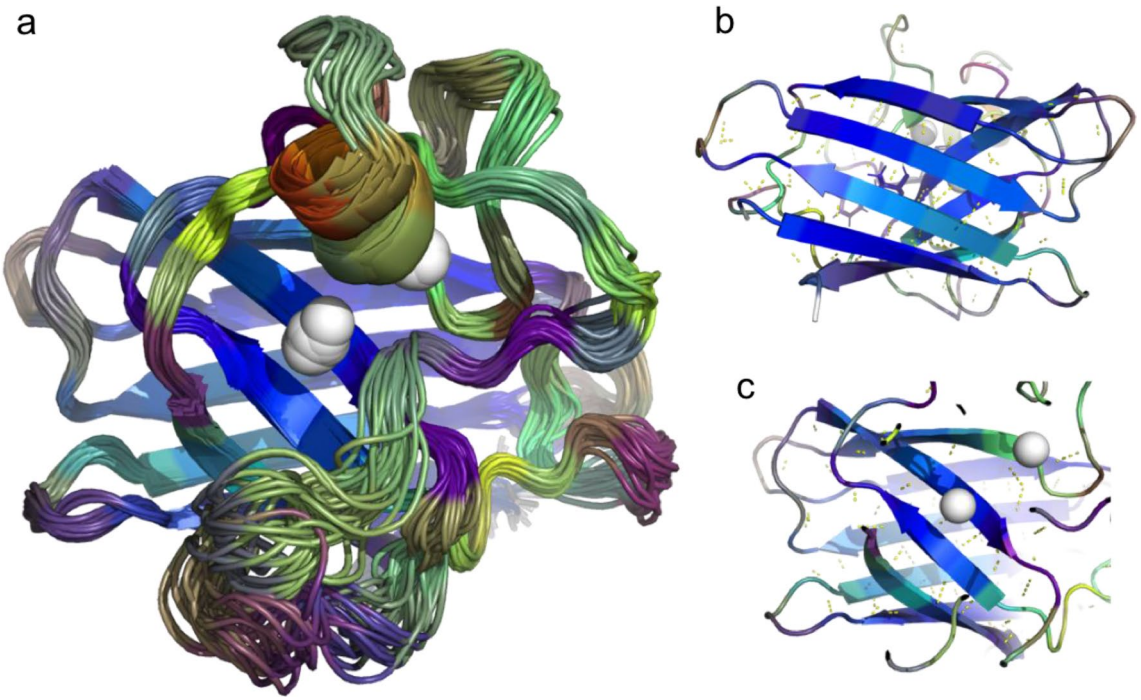

Fig. 8 NMR structure ensemble of the Cu/Zn bound SOD1 colored with CheSPI colors derived from BMRB id 4402. **a** from "top" looking down the metal binding pocket. **b** from bottom facing strands 1, 2, 3, and 6. **c** from top "clipping through" the long loops to observe the top sheet of strands 4, 5, 7, and 8

residues 25–31, which had elevated backbone dynamics. CheSPI analysis reveals some fluctuation in SCSs although with the absence of a common sign when scaled with the weight for the first component (Fig. 9a). This is reflected in Z-scores that typically lie between 10 and 12, which would indicate order, although the score is not as high as for fully formed standard helices and sheets (these lie around 13–15 typically). Residues 25–35 form an exception, and appear to be clearly more disordered, with Z-scores below 5. Furthermore, the PPII stretches feature CheSPI components much closer to zero (i.e. closer to random coil values with averages around ca.–2.0 and 0.0 for the first two components, and likewise near-zero secondary structure propensities by ncSPC) (Fig. 9b) than for ordered helices and sheets resulting in paler CheSPI colors closer to cyan-grey (Fig. 9c, i). The above ranges for CheZOD Z-scores and CheSPI components may be considered hallmarks of PPII helices, but with the current algorithm, CheSPI predicts primarily non-folded conformations for sfAFP and typically around 10–25% extended structure for the PPII stretches (Fig. 9d). In comparison, δ2D predicts around 25% PPII for these stretches (Fig. 9g), which is very similar to the PPII populations predicted for the IDPs tested here and for the case of human Tau protein discussed below (Fig. 10). The peptide segments forming PPII are apparent from the figure with exclusively coil DSSP classes (barring a few bends) and backbone angles in the PPII domain (Fig. 9e, f). It could be suspected that the helical bundle in sfAFP might have peculiar and specific interactions, such as variations in twist along the PPII helices, as reflected here in the varying CheSPI colors, and both standard backbone and unusual H-C' hydrogen bonds (Pentelute et al. 2008), that could potentially affect the CSs and thus the resulting CheSPI components. To address this systematically, our database of structured proteins was searched for consecutive stretches of three residues in PPII conformation (and other conformations, for reference) within DSSP "coil" states. Distributions of the CheSPI components for PPII in the database were found to be similar to the sfAFP case but were also found to be rather similar for stretches of three consecutive "extended" conformations (first CheSPI component was -0.97 and -1.64 in the former and latter case, respectively, i.e., with difference within one standard deviation, and the second component close to zero) (see Figure S5 and see also Fig. 2). This was also the case for the PPII helix, residues 95–99, in the *P. aeruginosa* protein PA1324, protein (BMRB id 6343, see Fig. S6). Hence, it remains very challenging to discriminate between "extended" (β-strand like) and PPII stretches using chemical shifts alone.

Aggregation nuclei of the protein Tau

IDPs are most fittingly interpreted as a statistical ensemble of local structural states. Here we demonstrate the ability of

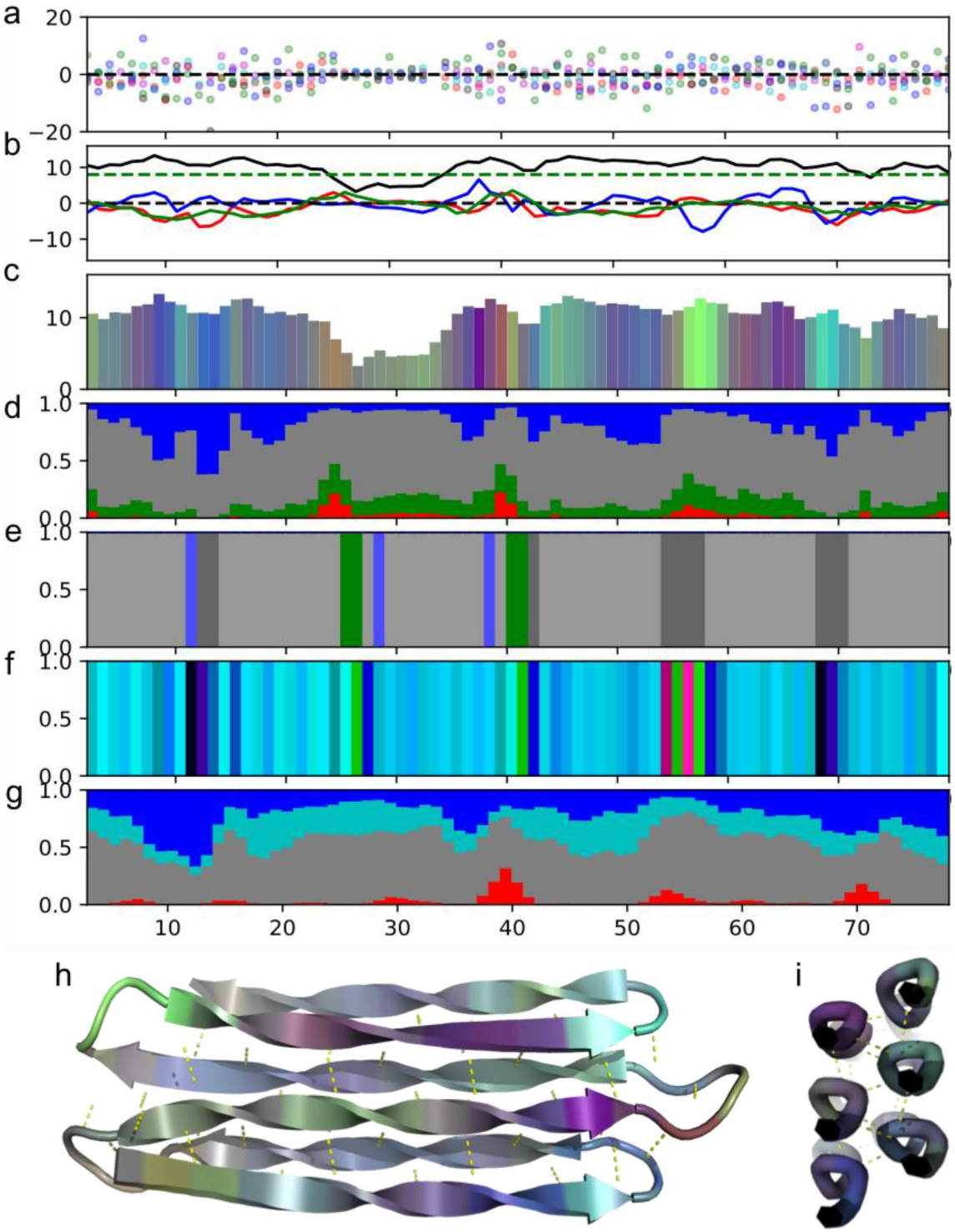

Fig. 9 CheSPI analysis of PPII helical bundle protein, the Hypogastrura harveyi "snow flea" antifreeze protein (sfAFP) **a**–**d** CheSPI output derived from assigned chemical shifts for BMRB id 27473 (see legend to Figure S1 and Fig. 4a,b) and the ncSPC secondary structure propensity multiplied with 8 is shown with a green broken line in panel b. **e** DSSP classification and **f** local backbone conformation in structure of sfAFP determined by X-ray crystallography with PDB id 2pne (see legend to Fig. 4d, e). **g** Predictions by δ2D visualized with stacked by plot showing helix, coil, PPII, and β-sheet using red, grey, cyan and red blue, respectively. **h**, **i** X-ray structure of sfAFP (PDB id 2pne) colored with CheSPI colors (as in panel c), hydrogen bonds are highlighted with yellow dashes. PPII and extended stretches are highlighted with standard β-strand cartoon rendering in **h** for visual purposes

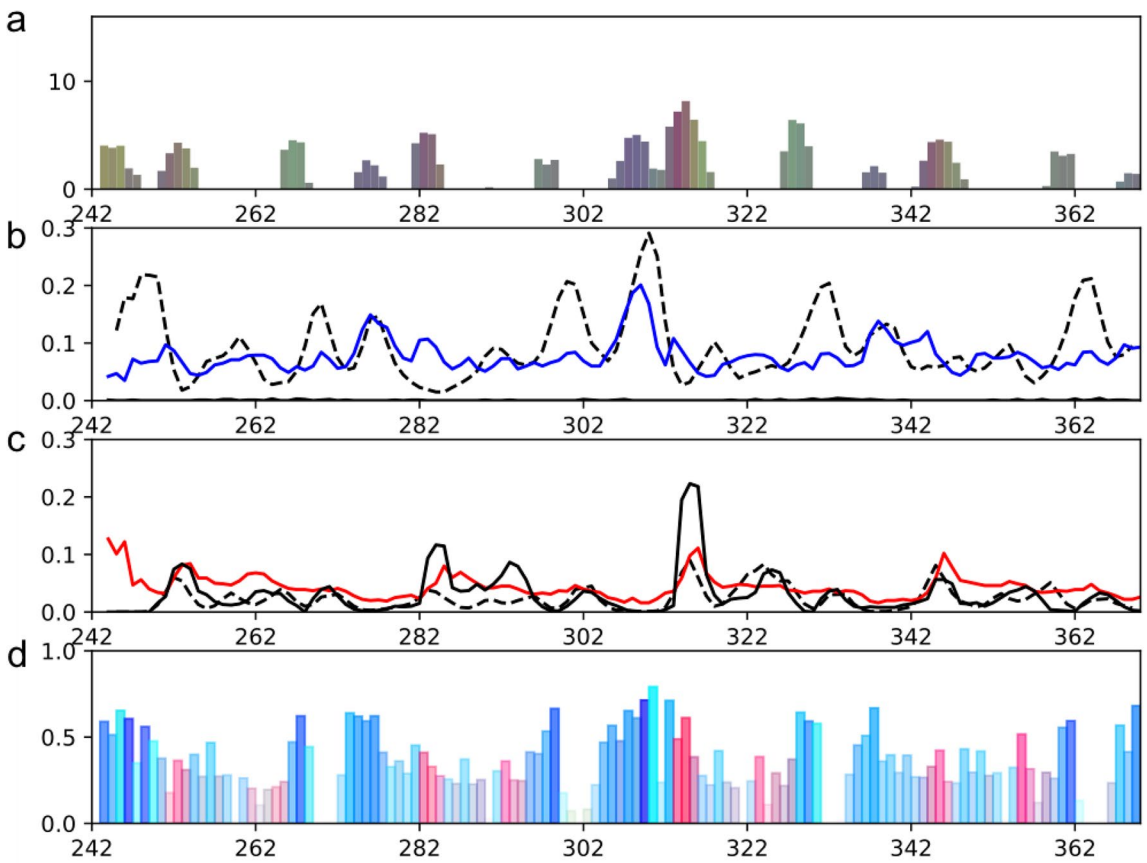

Fig. 10 CheSPI output panels for K18-Tau and comparison to structural ensemble and δ2D. **a** CheSPI colors based on assigned CSs from BMRB id 19253. **b** CheSPI extended populations (blue curve), δ2D predictions of beta-strand (black broken curve), fractions of beta-strand or bridge formed in the ensemble structure of K18 Tau (black curve very close to zero) (PED id 6AAC). **c** CheSPI helical populations (red curve), δ2D predictions of helix (black broken curve), fractions of helix formed in the ensemble structure of K18 Tau (black curve). **d** Average local backbone conformations in K18 Tau ensemble (see legend to Fig. 4e). See also Fig. S7

CheSPI to quantitatively infer the local structural composition from chemical shifts for a well-studied IDP, the K18 domain from human Tau (K18-Tau). Tau is intrinsically disordered, implicated in the regulation of microtubule organization, and prone to aggregation with a pathology related to Alheimer's disease. Tau aggregates as neurofibrillar tangles, containing paired helical filaments (PHFs) that adopt cross-β and β-helix conformation akin to other amyloidogenic proteins (Berriman et al. 2003, Barghorn et al. 2004). The K18 domain is directly involved in aggregation and consist of four imperfect 31–32 residue repeats, R1-R4. Hexapeptide segments within each repeat, residues, 275–280, 306–311, and 337–342, (HPF6, HPF6*, and HPF6**) are nucleation sites for aggregation, having a local β-sheet structure when forming multimers, and where HPF6 is at the core of the cross-β structure (Bergen et al. 2000, Eliezer et al. 2005, Daebel et al. 2012, Fitzpatrick et al. 2017). K18-Tau was studied by NMR spectroscopy (Mukrasch et al. 2007) and chemical shifts, RDCs, and PREs were used to calculate an ensemble of structures, accounting for the statistical composition of local structures, as outlined above (Ozenne et al. 2012). An algorithm for efficient sampling of conformational space was applied, while simultaneously satisfying the diverse experimental constraints. Indeed, the simulations confirmed the unstructured nature of K18-Tau with very little regular secondary structure and revealed a mixture of compositions of primarily PPII and "extended" conformations, with fewer turns and helical formations. Although K18-Tau is largely disordered, the ensemble shows subtle sequence-specific variations in the local conformational composition with some similarities between its four pseudo-repeats.

Analysis by CheSPI (summarized in Fig. 10 and Fig. S7) reveals findings that agree very well with the earlier observations outlined above. SCSs display limited scatter (although with subtle variations) leading to low CheZOD scores < 8.0 indicative of disorder (Fig. S7a). Concomitantly, the first two CheSPI components show values close to zero, again with some variation. For comparison, the ncSPC-derived secondary structure propensity shows a close correspondence

with the first CheSPI component (Fig. S7b). CheSPI predicts a preponderance of non-folded populations, albeit with important local biases (Fig. S7d). Firstly, elevated extended conformations are predicted by CheSPI for the hexapeptide HPF6(*/**) segments as shown in Fig. 10b. It is interesting, that segments that are responsible for aggregation and part of the core cross-β structure are already more extended in the unfolded conformation. Secondly, shorter 3–4 residue segments following the HPF6s had higher CheSPI populations for helical structure (Fig. 10c). Indeed, turn structures were assigned to these segments measured from RDCs (Mukrasch et al. 2007) DLKN (residues 253–256), DLSN, DLSK, and DKFD in repeats R1-R4). These turns are of type beta I, and two such consecutive turns correspond to a short 3_{10} helix (Pal et al. 2003). In the ASTEROIDS ensemble structure of K18-Tau (Fig. S7f), an excess of helical conformations and backbone angles were actually observed for these residues, in particular for the segment 313–315 in R3. Regular β-strand formation was not observed in the ensemble. However, a higher content of extended and PPII backbone conformations was encountered for residues in the HPF6 segments having also the highest CheSPI extended populations (Fig. 10d and Fig. S7e). For comparison, δ2D (Fig. 10, black broken curve and Fig. S7g) identifies the same maximum for the helix conformation – but only with confidence for R3. Furthermore, δ2D identifies higher β-like conformation for the HPF6 segments, but also for other positions in the sequence that were found to have mixtures of helical, extended, and positive ϕconformations (e.g. residues 298–303) in the ensemble derived from experimental data. The level of PPII conformation for K18-Tau predicted by δ2D (Fig. S7g) was also almost constant throughout the sequence and even slightly higher than the predictions for the PPII helical bundle discussed above.

Misfolding of alpha-synuclein

The protein alpha-synuclein (aS) is disordered under native conditions, but prone to misfolding forming cytotoxic aggregates implicated in the pathogenesis of Parkinson's disease (Singleton et al. 2003, Stefanis 2012). One of the physiological functions of aS is its binding to synaptic vesicles where it adopts a semi-folded α-helical conformation (Davidson et al. 1998, Jensen et al. 1998). This spurred a range of studies related to the binding of lipids and engineered membrane mimics (Jensen et al. 1998, Tofaris and Spillantini 2005). aS is comprised of seven 11-residue adjoined pseudo-repeats of amphiphilic character (I–VII) with a small flanking N-terminal four-residue insertion (between IV and V), and a longer acidic C-terminal region. A new variant with shuffled repeats, referred to as SaS, was designed previously to study the effect of sequence on vesicle binding and aggregation (Rao et al. 2008). aS and SaS was studied together with beta synuclein (bS) by NMR spectroscopy, analyzing the effect of interaction with sodium lauroyl sarcosinate (SLAS) micelles (Rao et al. 2009) and a structural model of aS-bound SLAS micelles was later derived based on NMR and EPR data (Rao et al. 2010). CheSPI analysis confirms aS to be disordered under native conditions (Fig. 11a, e). In contrast, when interacting with SLAS micelles, all studied synuclein variants form structures with high helical content in the amphiphilic repeat region (Fig. 11b-d, f–h). Comparison with the ensemble structure model for aS reveals helix formation for residues 1–91 with partial interruption of the helical structure around repeat III and a small helix kink around residues 60–65 (Fig. S8c). The helix interruption region corresponds to the region with lowest CheSPI helical populations (repeat III, Fig. 11f) and the helix kink is located at a position in the sequence with less "canonical" CheSPI colors (end of repeat V, Fig. 11b), i.e. the colors transition to greener, which is found at the C-terminal end of a helix (see Results and Fig. 4a) suggesting partial disruption of hydrogen bonding. Concomitantly, lower helical populations were found for the end of repeat V. A similar dip in CheSPI-derived helical populations and green colors were found at the end of repeat VII, although, in this case, no significant local irregularities were identified in the model structure (Fig. 11b, f). bS differs from aS by 14 mostly conservative mutations in the first 95 residues and the deletion of residues 73–83, disrupting repeats VI and VII. It is seen by the similarities of the CheSPI color profiles (Fig. 11b, d) that bS, despite the sequence modification, retains the local structural and dynamical properties with e.g. similar transition to greener colors and deletion of residues 73–83 to display similar signatures for the end of the helical region. In analogy to the bS case, SaS and aS share the same positions for the last repeat VII, and similar CheSPI color profiles are observed, indicating near-identical local structural and dynamical properties. Furthermore, repeats I, II, IV and the insertion form canonical helices in aS as indicated by red CheSPI colors and close to 100% CheSPI helical population. Concurrently, the same repeats also show strong signatures of helix structure when repositioned in the SaS sequence (Fig. 11c,g). On the other hand, the full repeat III and the end of repeat V, which feature lowered CheSPI helical populations (about 50%), and have partially disrupted helical structure or a kink in the SLAS-bound model, also indicate lowered helical strength for SaS – but in this case about 75% CheSPI helical populations for repeat III and ca. 30% for the end of repeat V when repositioned in SaS. aS and its variants binds micelles due to the amphiphilic nature of their sequences (see Figure S8a, b). Repeats III and V contain more charged and hydrophilic residues and fewer amino acids with hydrophobic side chains, when compared to their neighbor counterparts, suggesting a lower micelle binding affinity (Fig. S8a,b), hence explaining the lower CheSPI helical populations. This exemplifies how the local structure in proteins and their interactions with substrates is mostly driven by the local sequence.

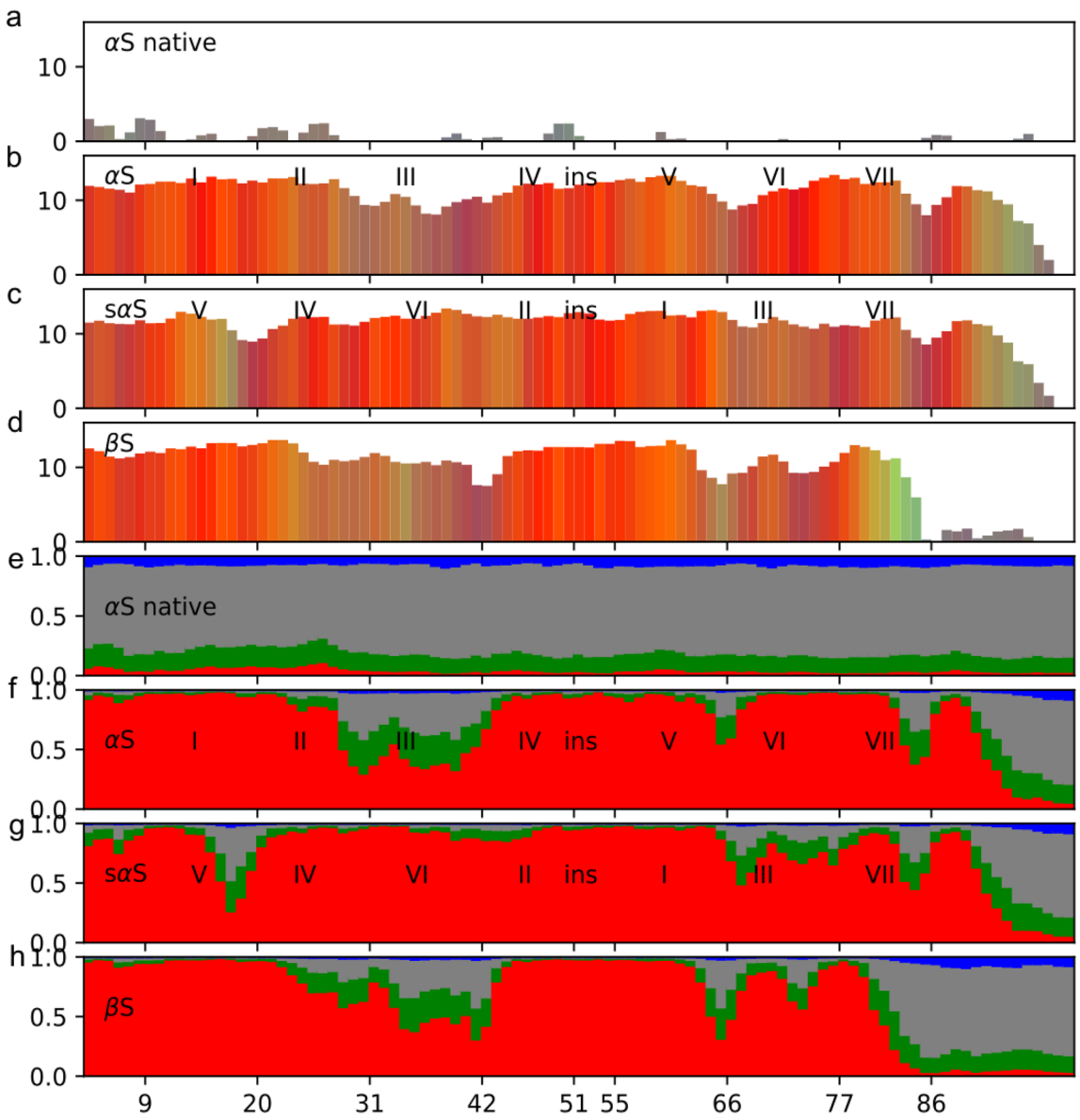

Fig. 11 Alpha Synuclein variants. Only residues up to 100 are shown **a–d** CheSPI color bar plot (see legend to Fig. 4c). **e–h** CheSPI populations stacked bar plot (see legend to Fig. 4d). The individual panels show results for the following alpha synuclein variants and conditions: αSyn native disordered (BMRB id 16300 panels a and e), αSyn SLAS-bound (BMRB id 16302 panels b and f), "shuffled" αSyn (sαS, see text) SLAS-bound (BMRB id 16303 panels c and g) βSyn SLAS-bound (BMRB id 16304 panels d and h)

The relocated repeats in SaS are modulated by the surrounding sequence/structure so that repeat III is presumably less helically interrupted whereas the end of repeat V is more kinked compared to aS.

Conclusions

We have introduced the software CheSPI for the comprehensive inference of structural and dynamical properties of proteins from assigned NMR chemical shifts and sequence. CheSPI can be applied to decompose chemical shifts to reduced dimensions and visualize protein secondary structure preferences using CheSPI colors. CheSPI provides predictions for the fine-grained conformations of local structure through estimated probabilities for the eight commonly recognized DSSP classes. A strong correlation was observed for Q8 (to recognize the eight classes solely from chemical shift data) and this was even stronger for Q3. It was demonstrated with a small number of examples how CheSPI can quantify and display the degree of protein disorder, and detect small populations of local structures in IDPs.

Methods

CheSPI components: the chemical shift principal components

The CheSPI component of order, k, for residue i, is computed as the weighted sum of truncated secondary chemical shifts (SCSs), Δ, for a 3-residue window as:

$$P_i^k = \frac{8.0}{\sqrt{N}} \sum_{j=-1,0,1} \sum_n w_n^k \Delta_n(i+j), \tag{1}$$

where

$$\Delta_n(i) = \Lambda\left(\delta_{obs}^n(i) - \delta_{ref}^n(i), \tau_n\right), \tag{2}$$

and

$$\Lambda(x,\tau) = \begin{cases} -\tau \; for \; x < -\tau \\ x \; for -\tau \le x \le \tau \\ \tau \; for \; x > \tau \end{cases} \tag{3}$$

and τ_n are nuclei-specific values used to truncate SCSs to avoid extreme values caused by assignment errors or typos. Here N is the total number of available experimental CSs for the residue triplet and $\delta_{obs}^n(i)$ and $\delta_{ref}^n(i)$ are the observed and reference CSs, respectively, for residue, i, and nuclei, n, where the reference CSs are the "random coil" CSs computed by POTENCI (Nielsen and Mulder 2018). Note that the universal weight of 8.0 was chosen arbitrarily to obtain component values comparable to the CheZOD Z-score ranges. The weights, w_n^k, were derived by an Orthogonal Partial Least Squares Discriminant Analysis (OPLSDA) using SIMCA Umetrics (Wu et al. 2010). This process identifies the linear combinations of the CSCs that best discriminates between the different secondary structure classes.

CheSPI colors

The first two CheSPI components are visualized as a unique color, first scaling a component, x, to be between 0 and 1 and truncated between threshold values $\pm\tau$ using.

$$f(x,\tau) = [\Lambda(x,\tau) + \tau]/2\tau \tag{4}$$

The color is then defined in terms of an RGB fraction vector, C, as:

$$C = \left[f\left(P_i^1, 12\right), 1 - f\left(P_i^2, 8\right), 1 - f\left(P_i^1, 12\right)\right]. \tag{5}$$

This definition leads to primarily blue colors for sheets, red for helices, and green colors for turns whereas disordered states with principal components close to zero correspond to grey colors.

CheSPI populations: secondary structure populations inferred from CheSPI components

Helix, sheet and coil all have rather distinct distributions of CheSPI components as seen for the correlated distribution for the 809 proteins with positive and negative values for the first component for helices and sheets, respectively, whereas there is more overlap for coil states with average values for both components near zero. The population of a 3-state secondary structure type, s, is calculated based on the density,ρ_s, from the experimental correlated distribution of the two first CheSPI components.

$$p_i(s) = \frac{\rho_s\left(P_i^1, P_i^2\right)}{\sum_{z=S,H,C} \rho_z\left(P_i^1, P_i^2\right)} \tag{6}$$

The densities were estimated from histogram distributions from the 809 proteins set.

Prediction of 8-state DSSP secondary structure classes from sequence and CheSPI components

A back-calculation prediction model was defined for the CheSPI component, P, from local primary sequence and 8-state DSSP secondary structure class. The model is linear with a constant term corresponding to the DSSP class and corrections for the sequence, C, and secondary structure, D, values in a sliding window of four residues in each direction as:

$$P_i^k(A,S) = \mathbf{b}_{s_i}^k + C_i^k\left[A, \tau\left(S_i\right)\right] + D_i^k\left[S, \tau\left(S_i\right)\right] \tag{7}$$

where

$$C_i^k(A,t) = \sum_{n=-4}^{n=4} {}_n^t\mathbf{c}_{A_{i+n}}^k \tag{8}$$

and

$$D_i^k(S,t) = \sum_{n=-4,\; n\neq 0}^{n=4} {}_n^t\mathbf{d}_{S_{i+n}}^k \tag{9}$$

where A_i denotes the amino acid sequence, and S_i and $t = \tau\left(S_i\right)$ the 8- and 3-state secondary structure types, respectively, all at residue position i, and τ maps the 8-state classes to 3 classes helix/sheet/coil. The correction term, C, constitutes 540 (=9*20*3*2) predetermined constants for each CheSPI component whereas the secondary term, D, constitutes 384 (=8*8*3*2) constants. These constants were derived from a multi-linear regression fit using the large set

of 809 protein sequences with known secondary structure. Some constants were set to zero in order to limit the number of free parameters (428 and 182 non-zero constants were used for the above two terms). The optimal balance between free parameters and goodness of fit was derived by minimizing Akaike's Information Criterion (AIC) and varying the number of adjustable parameters using a genetic algorithm as described previously when parametrizing POTENCI (Nielsen and Mulder 2018).

The DSSP 8-state secondary structure classes are predicted using comparison between observed and back-calculated CheSPI components and by applying Bayes Theorem. First, based on the observed CheSPI components, Q, the likelihood, L, of observing the principal components, given a certain secondary structure and the sequence, is calculated as:

$$L_i(Q|A,S) = \prod_{(k=1,2)} \phi\left[P_i^k(A,S) - Q_i^k, \sigma(S,k)\right], \tag{10}$$

where ϕ (x, s) is the normal distribution density function with mean 0 evaluated at x with variance s^2, and σ is the standard deviation of the prediction errors measured for the training set of the secondary structure S. The likelihood can be evaluated as a product of marginal probabilities for the individual component because they are orthogonal by definition of the OPLS-DA procedure. Secondly, the posterior probability for the secondary structure is calculated by multiplying the above likelihood with a prior probability for the secondary structure:

$$P_i^{post}(S|A,Q) = L_i(Q|A,S) * \mathrm{P_i^{pri}}(S|A) \tag{11}$$

in other words, we use the CheSPI component data to update the prior probabilities for the secondary structure. The CheSPI application offers two procedures for estimating the prior probabilities: (*i*) simple per residue type frequencies for secondary structure types or (*ii*) secondary structure prediction based on sequences alone using Xraptor (Wang et al. 2011, Källberg et al. 2012) Subsequently, the posterior probabilities for all 8 secondary structure classes are normalized to sum to unity.

Finally, the secondary structure predicted by CheSPI is identified as the configuration with maximum combined posterior probability for all residues:

$$S_{opt} = \max_S \left[\prod_{i=1}^{N} P_i^{post}(S|A,Q)\right] \tag{12}$$

This problem has a large dimensionally and cannot be optimized exhaustively, and therefore, the secondary structure optimization was implemented with a genetic algorithm solver as was the case for POTENCI (Nielsen and Mulder 2018). The algorithm is initiated with random secondary structure conformations sampled based on the secondary structure predictions from sequence.

Definition of backbone torsion angle averages order parameters

The dihedral angle order parameter, S, of Hyberts, Wagner and co-workers (Hyberts et al. 1992) is defined by averaged trigonometric values for backbone torsion angles:

$$s_\theta = \frac{1}{N}\left[\sum_{i=1}^{N} \sin(\theta_i)\right] and\, c_\theta = \frac{1}{N}\left[\sum_{i=1}^{N} \cos(\theta_i)\right] \tag{13}$$

$$S_\theta^2 = s_\theta^2 + c_\theta^2 \tag{14}$$

$$S_{bb}^2 = \sqrt{S_\phi^2 S_\psi^2} \tag{15}$$

for an ensemble of N structures, where θ_i is the value of a particular dihedral angle θ in the i^{th} member of the ensemble and S_{bb} is the combined order parameter for the backbone torsion angles ϕ and Ψ. The corresponding averaged angles are found by renormalizing with the order parameter and calculating inverse cosine:

$$\bar{\theta} = \begin{cases} \cos^{-1}\left(c_\theta/S_\theta\right) if\ s_\theta > 0 \\ -\cos^{-1}\left(c_\theta/S_\theta\right) if\ s_\theta < 0 \end{cases} \tag{16}$$

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10858-021-00374-w.

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request. Python code for CheSPI is available for download at GitHub: https://github.com/protein-nmr.

References

Adzhubei AA, Sternberg MJ, Makarov AA (2013) Polyproline-II helix in proteins: structure and function. J Mol Biol 425:2100–2132

Banci L et al (2002) The solution structure of reduced dimeric copper zinc superoxide dismutase. Eur J Biochem 269:1905–1915

Barghorn S, Davies P, Mandelkow E (2004) Tau paired helical filaments from Alzheimer's disease brain and assembled in vitro are based on beta-structure in the core domain. Biochemistry 43:1694–1703

Berjanskii MV, Wishart DS (2007) The RCI server: rapid and accurate calculation of protein flexibility using chemical shifts. Nucleic Acids Res 35:W531–W537

Bernadó P et al (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc Natl Acad Sci USA 102:17002–17007

Berriman J et al (2003) Tau filaments from human brain and from in vitro assembly of recombinant protein show cross-beta structure. Proc Natl Acad Sci USA 100:9034–9038

Bradley W, Robert P (2013) Multivariate analysis in metabolomics. Curr Metabol 1:92–107
Braun D, Wider G, Wuethrich K (1994) Sequence-corrected 15N "random coil" chemical shifts. J Am Chem Soc 116:8466–8469
Brutscher B et al (2015) NMR methods for the study of instrinsically disordered proteins structure, dynamics, and interactions: general overview and practical guidelines. Adv Exp Med Biol 870:49–122
Bunney TD et al (2006) Structural and mechanistic insights into ras association domains of phospholipase C epsilon. Mol Cell 21:495–507
Bylesjö M et al (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. J Chemom 20:341–351
Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 51:2224–2231
Camilloni C, Cavalli A, Vendruscolo M (2013) Replica-averaged metadynamics. J Chem Theory Comput 9:5610–5617
Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci 104:9615
Cleveland DW, Hwo SY, Kirschner MW (1977) Physical and chemical properties of purified tau factor and the role of tau in microtubule assembly. J Mol Biol 116:227–247
Daebel V et al (2012) β-Sheet core of Tau paired helical filaments revealed by solid-state NMR. J Am Chem Soc 134:13982–13989
Dass R, Mulder FAA, Nielsen JT (2020) ODiNPred: comprehensive prediction of protein order and disorder. Sci Rep 10:14780
Davidson WS, Jonas A, Clayton DF, George JM (1998) Stabilization of alpha-synuclein secondary structure upon binding to synthetic membranes. J Biol Chem 273:9443–9449
De Simone A et al (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. J Am Chem Soc 131:16332–16333
Eghbalnia HR et al (2005) Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements. J Biomol NMR 32:71–81
Eliezer D et al (2005) Residual structure in the repeat domain of tau: echoes of microtubule binding and paired helical filament formation. Biochemistry 44:1026–1036
Felli IC, Pierattelli R (2012) Recent progress in NMR spectroscopy: toward the study of intrinsically disordered proteins of increasing size and complexity. IUBMB Life 64:473–481
Fitzpatrick AWP et al (2017) Cryo-EM structures of tau filaments from Alzheimer's disease. Nature 547:185–190
Graham LA, Davies PL (2005) Glycine-rich antifreeze proteins from snow fleas. Science 310:461
Hafsa NE, Arndt D, Wishart DS (2015) CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucleic Acids Res 43:W370–W377
Hyberts SG, Goldberg MS, Havel TF, Wagner G (1992) The solution structure of eglin c based on measurements of many NOEs and coupling constants and its comparison with X-ray structures. Protein Sci 1:736–751
Jensen PH et al (1998) Binding of alpha-synuclein to brain vesicles is abolished by familial Parkinson's disease mutation. J Biol Chem 273:26292–26294
Jensen MR, Salmon L, Nodet G, Blackledge M (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. J Am Chem Soc 132:1270–1272
Jha AK, Colubri A, Freed KF, Sosnick TR (2005) Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc Natl Acad Sci USA 102:13099
Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637
Källberg M et al (2012) Template-based protein structure modeling using the RaptorX web server. Nat Protoc 7:1511–1522
Kjaergaard M, Poulsen FM (2012) Disordered proteins studied by chemical shifts. Prog Nucl Magn Reson Spectrosc 60:42–51
Kjaergaard M, Brander S, Poulsen FM (2011) Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. J Biomol NMR 49:139–149
Kukic P et al (2019) The free energy landscape of the oncogene protein E7 of human papillomavirus type 16 reveals a complex interplay between ordered and disordered regions. Sci Rep 9:5822. https://doi.org/10.1038/s41598-019-41925-4
Labudde D, Leitner D, Krüger M, Oschkinat H (2003) Prediction algorithm for amino acid types with their secondary structure in proteins (PLATON) using chemical shifts. J Biomol NMR 25:41–53
Lee JH et al (2015) Heterogeneous binding of the SH3 client protein to the DnaK molecular chaperone. Proc Natl Acad Sci USA 112:E4206-4215
Makowska J et al (2006) Polyproline II conformation is one of many local conformational states and is not an overall conformation of unfolded peptides and proteins. Proc Natl Acad Sci USA 103:1744
Marsh JA, Forman-Kay JD (2009) Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. J Mol Biol 391:359–374
Marsh JA, Forman-Kay JD (2012) Ensemble modeling of protein disordered states: experimental restraint contributions and validation. Proteins 80:556–572
Marsh JA, Singh VK, Jia Z, Forman-Kay JD (2006) Sensitivity of secondary structure propensities to sequence differences between alpha- and gamma-synuclein: implications for fibrillation. Protein Sci 15:2795–2804
Mielke SP, Krishnan VV (2009) Characterization of protein secondary structure from NMR chemical shifts. Prog Nucl Magn Reson Spectrosc 54:141–165
Milani P, Gagliardi S, Cova E, Cereda C (2011) SOD1 Transcriptional and posttranscriptional regulation and its potential implications in ALS. Neurol Res Int 2011:458427
Mukrasch MD et al (2007) Highly populated turn conformations in natively unfolded tau protein identified from residual dipolar couplings and molecular simulation. J Am Chem Soc 129:5235–5243
Neal S, Nip AM, Zhang HY, Wishart DS (2003) Rapid and accurate calculation of protein H-1, C-13 and N-15 chemical shifts. J Biomol NMR 26:215–240
Nielsen JT, Mulder FAA (2016) There is diversity in disorder—"In all chaos there is a cosmos, in all disorder a secret order." Front Mol Biosci. https://doi.org/10.3389/fmolb.2016.00004
Nielsen JT, Mulder FAA (2018) POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. J Biomol NMR 70:141–165
Nielsen JT, Mulder FAA (2019) Quality and bias of protein disorder predictors. Sci Rep 9:5137
Nielsen JT, Mulder FAA (2020) Quantitative protein disorder assessment Using NMR chemical shifts. In: Kragelund BB, Skriver K (eds) Intrinsically Disordered proteins: methods and protocols. Springer, New York, pp 303–317
Nielsen JT, Eghbalnia HR, Nielsen NC (2012) Chemical shift prediction for protein structure calculation and quality assessment using an optimally parameterized force field. Prog Nucl Magn Reson Spectrosc 60:1–28

Ozenne V et al (2012) Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. J Am Chem Soc 134:15138–15148
Pal L, Chakrabarti P, Basu G (2003) Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. J Mol Biol 326:273–291
Pentelute BL et al (2008) X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. J Am Chem Soc 130:9695–9701
Rakhit R, Chakrabartty A (2006) Structure, folding, and misfolding of Cu, Zn superoxide dismutase in amyotrophic lateral sclerosis. Biochimica et Biophysica Acta Mol Basis Dis 1762:1025–1037
Rao JN, Dua V, Ulmer TS (2008) Characterization of alpha-synuclein interactions with selected aggregation-inhibiting small molecules. Biochemistry 47:4651–4656
Rao JN, Kim YE, Park LS, Ulmer TS (2009) Effect of pseudorepeat rearrangement on alpha-synuclein misfolding, vesicle binding, and micelle binding. J Mol Biol 390:516–529
Rao JN et al (2010) A Combinatorial NMR and EPR approach for evaluating the structural ensemble of partially folded proteins. J Am Chem Soc 132:8657–8668
Richardson JS, Richardson DC (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 99:2754–2759
Robberecht W et al (1994) Cu/Zn superoxide dismutase activity in familial and sporadic amyotrophic lateral sclerosis. J Neurochem 62:384–387
Robustelli P, Stafford KA, Palmer AG (2012) Interpreting protein structural Dynamics from NMR chemical shifts. J Am Chem Soc 134:6365–6374
Rosen DR et al (1993) Mutations in Cu/Zn superoxide dismutase gene are associated with familial amyotrophic lateral sclerosis. Nature 362:59–62
Schwarzinger S et al (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123:2970–2978
Shapovalov M, Vucetic S, Dunbrack RL Jr (2019) A new clustering and nomenclature for beta turns derived from high-resolution protein structures. PLoS Comput Biol 15:e1006844–e1006844
Shen Y, Bax A (2013) Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J Biomol NMR 56:227–241
Shen Y, Bax A (2015) Protein structural information derived from nmr chemical shift with the neural network program TALOS-N. Methods Mol Biol 1260:17–32
Shi Z, Chen K, Liu Z, Kallenbach NR (2006) Conformation of the backbone in unfolded proteins. Chem Rev 106:1877–1897
Singleton AB et al (2003) [alpha]-synuclein locus triplication causes Parkinson's disease. Science 302:841
Sirangelo I, Iannuzzi C (2017) The role of metal binding in the amyotrophic lateral sclerosis-related aggregation of copper-zinc superoxide dismutase. Molecules 22:1429
Sormanni P et al (2017) Simultaneous quantification of protein order and disorder. Nat Chem Biol 13:339–342
Stefanis L (2012) α-Synuclein in Parkinson's disease. Cold Spring Harb Perspect Med 2:a009399–a009399
Sterckx YGJ et al (2014) Small-angle X-Ray scattering- and nuclear magnetic resonance-derived conformational ensemble of the highly flexible antitoxin PaaA2. Structure 22:854–865
Tamiola K, Mulder FAA (2012) Using NMR chemical shifts to calculate the propensity for structural order and disorder in proteins. Biochem Soc Trans 40:1014–1020
Tamiola K, Acar B, Mulder FAA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J Am Chem Soc 132:18000–18003
Teilum K et al (2009) Transient structural distortion of metal-free Cu/Zn superoxide dismutase triggers aberrant oligomerization. Proc Natl Acad Sci USA 106:18273–18278
Tofaris GK, Spillantini MG (2005) Alpha-synuclein dysfunction in Lewy body diseases. Mov Disord 20(Suppl 12):S37-44
Tompa P (2009) Structure and function of intrinsically disordered proteins. Chapman and Hall/CRC, London
Trygg J, Wold S (2002) Orthogonal projections to latent structures (O-PLS). J Chemom 16:119–128
Uversky VN, Longhi S (2010) Instrumental analysis of intrinsically disordered proteins: assessing structure and conformation. Wiley, Hoboken
Varadi M et al (2014) pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic Acids Res 42:D326–D335
Varadi M, Vranken W, Guharoy M, Tompa P (2015) Computational approaches for inferring the functions of intrinsically disordered proteins. Front Mol Biosci. https://doi.org/10.3389/fmolb.2015.00045
von Bergen M et al (2000) Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306) VQIVYK(311)) forming beta structure. Proc Natl Acad Sci USA 97:5129–5134
Wang C-C, Chen J-H, Lai W-C, Chuang W-J (2007) 2DCSi: identification of protein secondary structure and redox state using 2D cluster analysis of NMR chemical shifts. J Biomol NMR 38:57–63
Wang Z, Zhao F, Peng J, Xu J (2011) Protein 8-class secondary structure prediction using conditional neural fields. Proteomics 11:3786–3792
Wishart DS, Case DA (2001) Use of chemical shifts in macromolecular structure determination. Resonance Biologica Macromol C 338:3–34
Wishart DS, Sykes BD (1994a) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J Biomol NMR 4:171–180
Wishart DS, Sykes BD (1994b) Chemical-shifts as a tool for structure determination. Methods Enzymol 239:363–392
Wishart DS, Sykes BD, Richards FM (1992) The chemical shift index: a fast and simple method for the assignment of protein secondary structure through NMR spectroscopy. Biochemistry 31:1647–1651
Wishart DS et al (1995) 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5:67–81
Worley B, Powers R (2016) PCA as a practical indicator of OPLS-DA model reliability. Curr Metabol 4:97–103
Wu Z, Li D, Meng J, Wang H (2010) Introduction to SIMCAP and its application. In: Esposito Vinzi V, Chin WW, Henseler J, Wang H (eds) Handbook of partial least squares concepts methods and applications. Springer, Berlin, pp 757–774