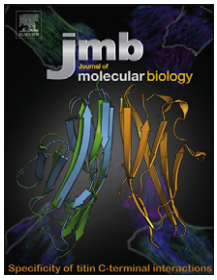

The s2D Method: Simultaneous Sequence-Based Prediction of the Statistical Populations of Ordered and Disordered Regions in Proteins

Pietro Sormanni [1], Carlo Camilloni [1], Piero Fariselli [2] and Michele Vendruscolo [1]

*1 - **Department of Chemistry,** University of Cambridge, Cambridge CB2 1EW, UK*
*2 - **Department of Computer Science,** University of Bologna, 40127 Bologna, Italy*

Correspondence to Michele Vendruscolo: *mv245@cam.ac.uk*
http://dx.doi.org/10.1016/j.jmb.2014.12.007
Edited by A. Keating

Abstract

Extensive amounts of information about protein sequences are becoming available, as demonstrated by the over 79 million entries in the UniProt database. Yet, it is still challenging to obtain proteome-wide experimental information on the structural properties associated with these sequences. Fast computational predictors of secondary structure and of intrinsic disorder of proteins have been developed in order to bridge this gap. These two types of predictions, however, have remained largely separated, often preventing a clear characterization of the structure and dynamics of proteins. Here, we introduce a computational method to predict secondary-structure populations from amino acid sequences, which simultaneously characterizes structure and disorder in a unified statistical mechanics framework. To develop this method, called *s2D*, we exploited recent advances made in the analysis of NMR chemical shifts that provide quantitative information about the probability distributions of secondary-structure elements in disordered states. The results that we discuss show that the *s2D* method predicts secondary-structure populations with an average error of about 14%. A validation on three datasets of mostly disordered, mostly structured and partly structured proteins, respectively, shows that its performance is comparable to or better than that of existing predictors of intrinsic disorder and of secondary structure. These results indicate that it is possible to perform rapid and quantitative sequence-based characterizations of the structure and dynamics of proteins through the predictions of the statistical distributions of their ordered and disordered regions.

Introduction

Proteins folded in their native states experience conformational fluctuations, which in many cases facilitate their functions [1–6]. It has also been recently recognized that many proteins or protein regions, respectively known as intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDRs), are highly dynamical and do not populate stable three-dimensional structures, even under physiological conditions, despite being still perfectly functional [7–12].

Because of their involvement in key biological processes, such as regulation and signaling, and their connections with conformational diseases and some cancer types [9,10], IDPs have increasingly become the center of focussed attention [7–12]. As IDPs do not conform to the traditional sequence–structure–function paradigm, a distinction is commonly made between two classes of proteins—"ordered" and "intrinsically disordered". Yet, the boundary between these two classes is not well defined, as many structured proteins contain at least some disordered regions, while disordered proteins tend to contain some structured parts. Furthermore, the local stability of different structural regions varies extensively. In the course of their lifetimes, most proteins populate partially or fully disordered states, which are often crucial in determining their functional or dysfunctional behaviors [7–13].

Detailed experimental characterization of structure and disorder in proteins remains a difficult and costly task, which is associated with the risk of obtaining inconclusive results especially when not employing a broad range of different techniques [7–12,14]. As a consequence, to offer guidance for experiments,

to allow for large-scale studies and to facilitate the functional annotation of proteins, a variety of sequence-based predictors of protein disorder has been developed [14–25]. Predictions of intrinsic disorder can be presented together with predictions of secondary or tertiary structure, as performed by some Web servers [22,26,27], to provide qualitative insights about the type and the stability of the structures formed by a given protein.

In this context, since structure and disorder represent closely intertwined properties of protein molecules, it would be desirable to have a predictor that provides them at the same time. The development of such type of predictors has been hindered by the great challenges in obtaining experimentally structural information about disordered states. Here, we exploit recent advances in the use of nuclear magnetic resonance (NMR) chemical shifts, which are making it possible to determine quantitatively the populations of secondary-structure elements directly from the measured chemical shifts [28]. We thus introduce a prediction method, named *s2D*, which offers a sequence-based estimate of the secondary-structure populations at the individual residue. As a consequence, the *s2D* method is simultaneously a predictor of intrinsic disorder and of secondary structure. More specifically, the *s2D* method is trained on the secondary-structure populations calculated with the *δ2D* method [28] from measured backbone chemical shifts, and thus, it gives information about conformational properties of disordered states.

Results

In this work we present the *s2D* method of predicting secondary-structure populations, which consists in a combination of artificial neural network trained with extreme learning machines (ELMs) (see Materials and Methods). The predictor is trained using a dataset of protein sequences with the corresponding populations of α-helix, β-strand and random coil calculated primarily from measured NMR chemical shifts at each residue position (see "The *δ2D* and *s2D* datasets"). In this context, the three secondary-structure types predicted by the *s2D* method can be regarded as orthogonal since they produce well-distinct patterns of NMR backbone chemical shifts [28,29]. As a result, disordered states that have significant populations of either α-helix or β-strand can be distinguished from fully random coil states without transient secondary-structure elements, thus allowing the *s2D* method to provide a detailed characterization of the conformational properties of disordered states.

10-Fold cross-validation of the *s2D* method

In order to assess the performance of the *s2D* method, we applied a 10-fold cross-validation procedure to the framework described in "The architecture of the *s2D* method" (Materials and Methods). Given the 10 non-homologous subsets generated as described in "The *δ2D* and *s2D* datasets", we trained the *s2D* predictor 10 times. Each time, a different subset was selected for testing, while all sequences belonging to the other nine subsets were used for training. This procedure ensured that the local sequence identity between sequences in the training and in the testing set was always below 25%. The method performance was assessed using the Pearson's coefficients of correlation between the predicted and the observed populations of the three secondary-structure types considered (R_H, R_E, R_C) and the corresponding mean square errors (MSE_H, MSE_E, MSE_C) and mean absolute errors (MAE_H, MAE_E, MAE_C). Differently from the model selection (see "Model selection"), here we calculated these performance indicators (*R*, *MSE* and *MAE*) only on secondary-structure populations obtained from NMR data, while the few entries extrapolated from X-ray structures were ignored in the testing, as their secondary-structure populations are inherently less accurate (see "The *δ2D* and *s2D* datasets").

The results are reported in Table 1, and the similarity of the performance indicators on the training and on the testing sets indicates that there is no significant over-fitting. More importantly, the small values of the standard deviations, calculated on the results of the 10 different training rounds, suggest that the results in the first column of Table 1 are representative of the performance of the *s2D* predictor on most protein sequences. Consequently, on average, the *s2D* method can predict the populations of secondary-structure elements of a protein molecule—as a monomer in solution—with

Table 1. Performance indicators from the 10-fold cross-validation of the *s2D* method

	μ_{test}	σ_{test}	μ_{train}	σ_{train}
R_H	0.817	0.007	0.836	0.001
R_E	0.77	0.02	0.789	0.002
R_C	0.71	0.02	0.737	0.002
MSE_H	0.038	0.002	0.0340	0.0004
MSE_E	0.024	0.001	0.0217	0.0003
MSE_C	0.041	0.001	0.0371	0.0004
MAE_H	0.140	0.004	0.1321	0.0008
MAE_E	0.113	0.003	0.1063	0.0008
MAE_C	0.158	0.003	0.1491	0.0009

We report the Pearson's correlation coefficients (*R*), the mean square errors (*MSE*) and the mean absolute error (*MAE*) calculated for the populations of α-helix (H), β-strand (E) and disordered random coil (C). μ is the mean calculated on the results of the 10 different rounds of training and testing and σ is the corresponding standard deviation. Training and testing sets were extracted from the *s2D* dataset, but all reported parameters are calculated only on entries whose secondary-structure populations were obtained from NMR chemical shifts (i.e., entries in the *δ2D* dataset).

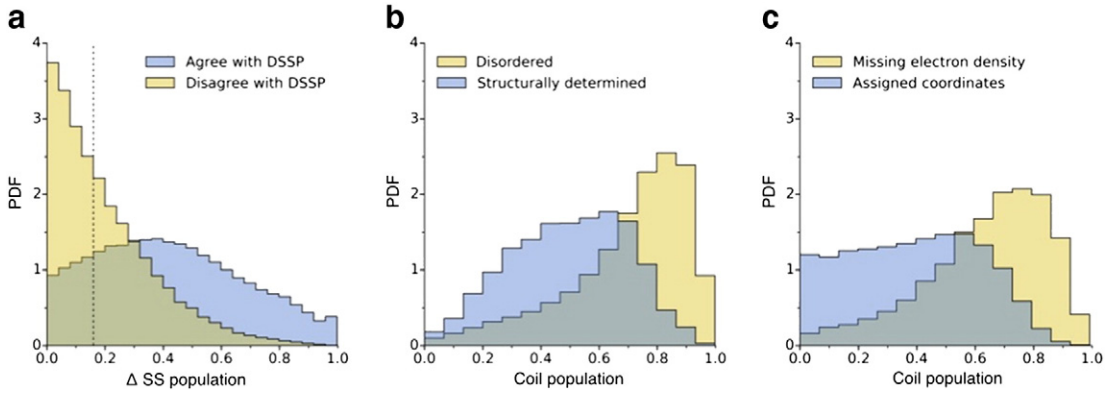

Fig. 1. Validation of the *s2D* method on structured and disordered proteins. (a) Normalized distributions of the difference between the highest and second-highest secondary-structure populations of the residues for which the *s2D* prediction agrees with the DSSP assignment (blue) and for which it does not agree (yellow). The broken line represents the median of the second distribution. (b) Distribution of the predicted coil population of the residues belonging to regions annotated as disordered in the DisProt database [35] (yellow) and to regions annotated as structurally determined (blue). (c) Distribution of the predicted coil population of the residues with assigned coordinates (blue) and with missing electron density (yellow) in a dataset of 1304 protein sequences from Ref. [37].

an average error of about 14% and a correlation coefficient of about 0.77.

Validation of the *s2D* method on structured proteins

In order to obtain an independent validation of the performance of the *s2D* method on structured proteins, we ran the predictor on a dataset containing 1833 gapless protein sequences with less than 25% sequence identity between themselves. This dataset was extracted from X-ray structures and was used in Ref. [30] to test the accuracy of the three-state predictor of secondary-structure SPINE X. Here, we assigned the secondary structure to each Protein Data Bank (PDB) chain with the DSSP program [31].

The secondary-structure type predicted by the *s2D* method to be the most populated at each residue position was used for comparison with the DSSP assignment of the X-ray structures. The analysis reveals that the predicted most populated secondary structure corresponds to the one assigned by DSSP in 78.6 ± 0.2% of the cases, where the error is the standard error calculated from the *Q3* score on the different sequences. This *Q3* score (accuracy of the three-state prediction) increases slightly to 79.0 ± 0.2% when 355 membrane proteins are removed from the dataset of 1833 protein chains. Similarly, the median segment overlap scores [32] is 73 (or 73.5 without membrane proteins), which reflect the great sensitivity of these scores to individual residues (e.g., the ones at the end of observed secondary-structure segments) whose predicted most populated state is not the one of the crystal structure, even when this is only marginally more populated than the one in the crystal. Overall, the performance of the *s2D* method on this dataset is comparable with that of existing three-state secondary-structure predictors trained on larger and less noisy databases [30].

To investigate the origins of the specific disagreements between the *s2D* and DSSP results, we calculated the differences between highest and second-highest secondary-structure populations predicted by the *s2D* method. We found that these differences are below 0.16 for 50% and below 0.05 for 20% of the residues for which the *s2D* predictions and the DSSP assignment differ (Fig. 1a). By contrast, when we consider the residues for which the *s2D* prediction agrees with the DSSP assignment, this is the case for only 17% and 5%, respectively. These results are likely to reflect the known fact that some structural motifs, which are favored energetically but not entropically in solution, are stabilized under crystallization conditions [33,34]. Differently from existing three-state secondary-structure predictors, the *s2D* method is not biased toward predicting such motifs, as it is trained mostly on solution-based NMR measurements.

Finally, we treated the *s2D* predictions of the populations of α-helix, β-strand and disordered random coil as three independent binary classifiers and performed a receiver operating characteristic (ROC) curve analysis using the DSSP assignments as true outcomes (Fig. 2a). The area under the curve is 0.96 for α-helix, 0.93 for β-strand and 0.89 for random coil.

Validation of the *s2D* method on IDPs

We then applied the *s2D* method to the protein sequences in the DisProt database [35], which contains IDPs or proteins with intrinsically

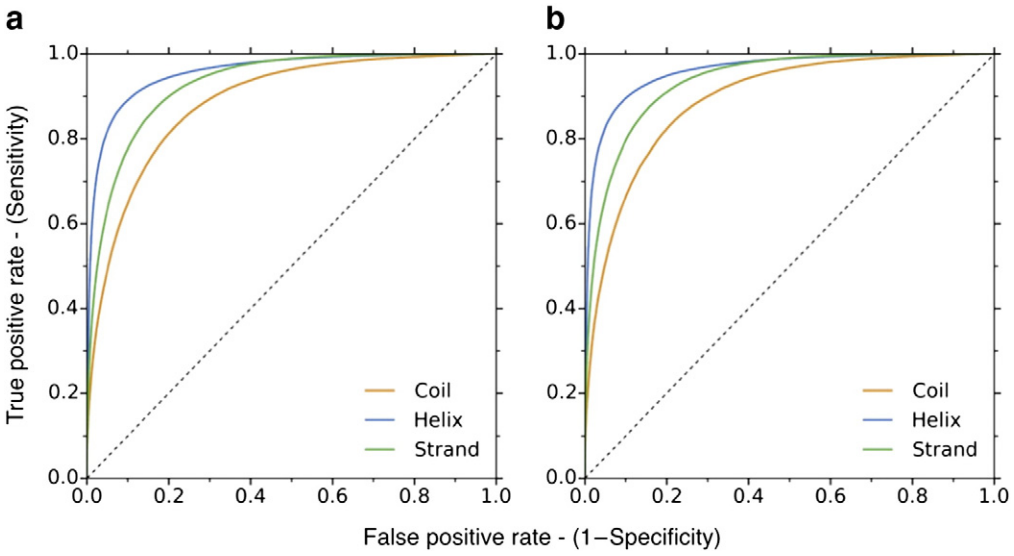

Fig. 2. ROC curves calculated on two validation datasets. (a and b) ROC curves calculated for the predicted populations of α-helix (blue), β-strand (green) and random coil (orange) considered as three independent binary classifiers. The calculations were performed using as true outcomes the DSSP [31] assignments of the residues in the dataset of 1883 protein sequences described in the "Validation of the *s2D* method on structured proteins" (a) and in the dataset of 1034 described in the "Validation of the *s2D* method on structured proteins with disordered regions" (b). Residues with missing electron density in the second dataset were considered random coil. The broken line is the line of no discrimination and represents the ROC curve of a random guess.

disordered regions. The predictions of the *s2D* method were compared to the regions annotated as disordered in DisProt. *s2D* predicted 86.5 ± 0.8% of the residues annotated as disordered to preferentially populate random coils and 88.8 ± 0.7% to have both α-helix and β-strand populations smaller than 0.5.

Since the DisProt database also contains information about structurally determined regions in its sequences, we compared the random-coil population of the residues in regions annotated as disordered with the one of residues in regions annotated as structurally determined (Fig. 1b). This comparison is not ideal as it is hindered by two facts. First, some residues within structurally determined regions may be found in loops of the structures and would thus be classified as coil by the *s2D* method. Second, the annotations in the DisProt database (version 6.02) are not free from noise, probably as a result of the different experimental methods used for disorder and structure determination. In some entries, structurally determined regions overlap with regions annotated as disordered (e.g., DP00064, DP00702 and so on; such entries were neglected in the analysis), and in some others, the regions annotated as disordered are in sharp disagreement with protein structures deposited in the PDB (e.g., DP00355 and 2L7B, DP00701 and chain B of 1SC5 and so on; such entries are retained in the analysis). Nevertheless, the two distributions in Fig. 1b, corresponding to the *s2D* predictions of the coil population of residues in the two classes, are well resolved with a median difference of 0.25.

Validation of the *s2D* method on partially structured proteins

As a further validation of the *s2D* method, we applied it to a dataset of 1304 protein sequences extracted from high-resolution (≤2.5 Å and $R \leq 0.25$) X-ray structures containing missing residues (remark 465 in the PDB annotations). These regions of missing electron density have been associated with intrinsic disorder because they possess a similar amino acid composition to IDPs [36] and because disordered regions are not expected to be detectable in the crystals as they fail to form stable structures. The dataset that we used was assembled in Ref. [37] (additional file 1 therein) and contains sequences with less than 25% sequence identity with each other and without His tags or leading/trailing segments. The 1304 chains comprise 318,431 residues with assigned coordinates and 14,737 disordered ones. The latter are separated in 1954 short regions (≤30 residues) and 54 long regions (>30 residues), which contain 19% of the disordered residues.

Secondary structures were determined for the residues with assigned coordinates using the DSSP program [31] and the performance of the *s2D* method on these residues was assessed as in "Validation of the *s2D* method on structured proteins". The *Q3* score

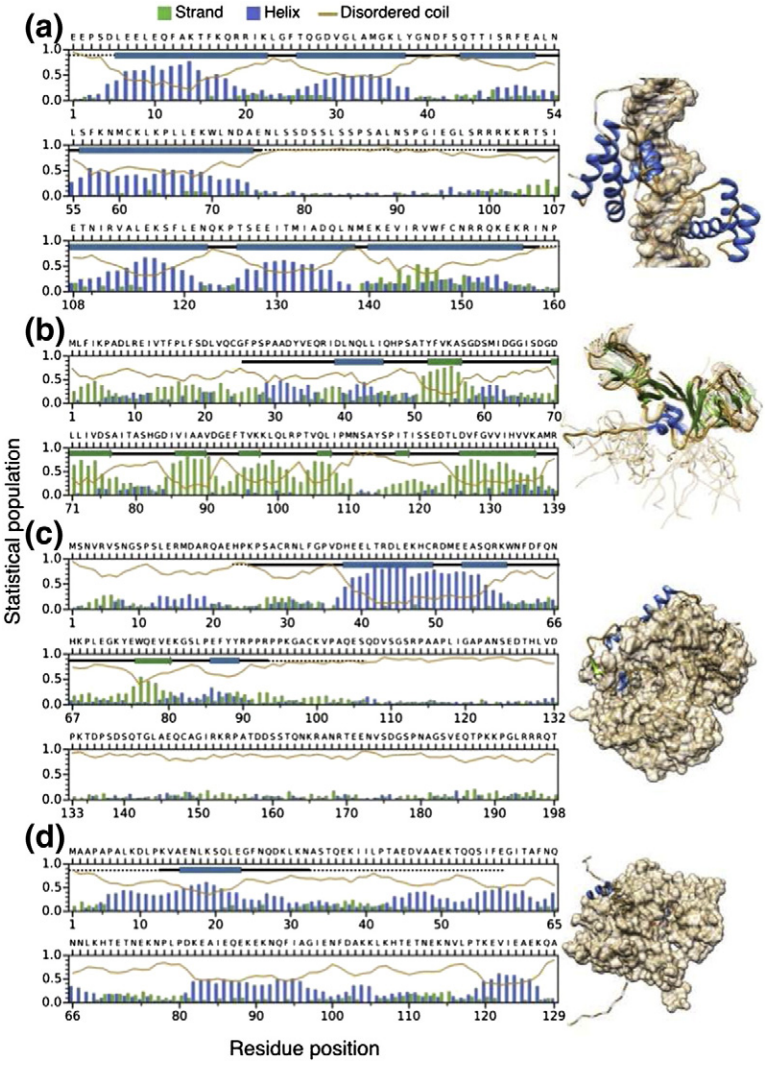

Fig. 3. Application of the *s2D* method to four proteins that form fuzzy complexes. In all panels, the predictions of the *s2D* method (left) are superimposed with a representation of the DSSP assignment of the protein structures shown to the right. In the DSSP assignment, a continuous black line denotes coil, a broken line denotes residues of missing electron density, a blue rectangle denotes an α-helix and a green arrow denotes a β-strand. Absence of the DSSP assignments denotes sequence regions present in UniProt but not in the seqres field of the PDB files. (a) The Oct-1 transcription factor (P14859) binds to DNA acting as a clamp. The two binding domains (PDB file 1hf0) are connected by a linker that is disordered also in the bound state, which contributes to binding as its shortening has been shown to reduce the affinity [75]. (b) The UmuD protein (P0AG11) forms a dimer (PDB file 1i4v) after RecA-facilitated self-cleavage. UmuD yields a random-coil signal in CD experiments at physiologically relevant concentrations [76]. (c) The p27Kip1 cell cycle kinase inhibitor (P46527) binds to the cyclin–Cdk2 complex (PDB file 1jsu). (d) The *s2D* prediction for the WH2 domain of ciboulot (O97428). The structure shown is the PDB file 3u9z, but this domain has been reported to interact with actin in a polymorphic way, binding in different locations [77].

on this dataset was 79.3 ± 0.2% (median segment overlap score, 74.5), and the behavior presented in Fig. 1a was observed again (data not shown). Of the residues with missing electron density, 83.2 ± 0.5% were predicted to mostly populate random-coil populations, and 87.3 ± 0.4% had both predicted α-helix and β-strand populations smaller than 0.5. Such percentages are changed to 84.6% and 89.1%, respectively, for residues found in short disordered regions, as well as to 76.4% and 79.5%, respectively, for residues found in long ones. The smaller amount of predicted random coil in long disordered regions is compatible with the idea that such regions might populate some structured motifs, but their overall degree of disorder would be too high for successful crystallization.

The distribution of the random-coil population of residues with assigned coordinates, which include crystallized loops, is distinct from the one of residues with missing electron density and the median difference is 0.27 (Fig. 1c). In addition, we performed a ROC curve analysis as in "Validation of the *s2D* method on structured proteins", including the residues with missing electron density as disordered random coil (Fig. 2b). The areas under the curve on this dataset are 0.96, 0.94 and 0.89 for α-helix, β-strand and random coil, respectively.

s2D analysis of IDPs that fold upon binding

The *s2D* method enables the prediction of the populations of secondary-structure elements of protein molecules as free monomers in solution (see "The *δ2D* and *s2D* datasets"). Many IDPs, however, perform their function by transiently populating relatively structured states [7,12,38]. Notable examples of IDPs in this group are the ones that fold upon binding, where a disordered protein or protein region can show well-defined secondary structures in the bound state [39,40].

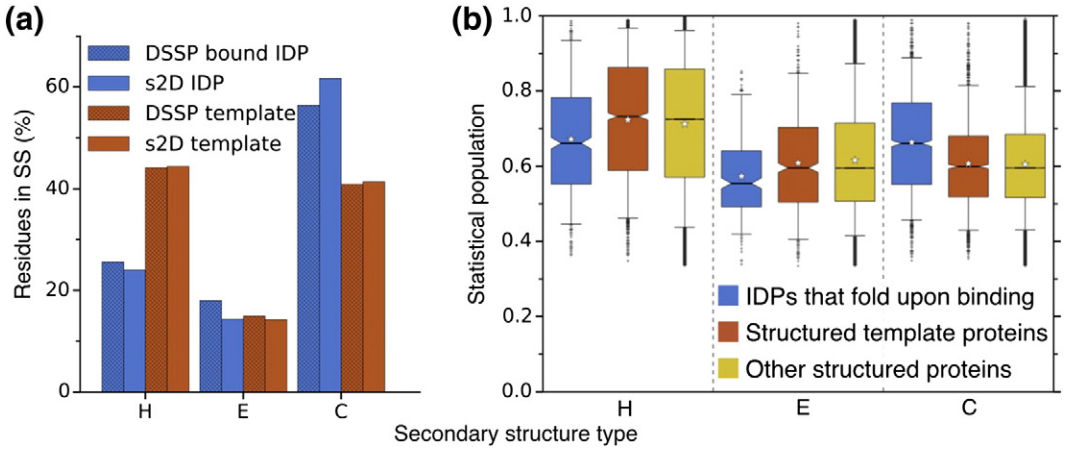

Fig. 4. Application of the *s2D* method to IDPs that fold upon binding. (a) Bar plot of the percentage of residues from the dataset of IDPs containing preformed structural elements from Ref. [44] in the different secondary-structure types. IDPs that fold upon binding are shown in blue, while the corresponding template proteins are shown in red. Hatched columns correspond to the DSSP assignments of the structures of the bound complexes and plain columns to the *s2D* predictions. While the *s2D* predictions were carried out using the full protein sequence employed in the structure determination experiments, the analysis reported here considered only those residues whose atom coordinates were actually solved in the experiment and for which a DSSP assignment was available. (b) Residues are divided in three classes (*x*-axis) according to the secondary-structure type predicted to be the most populated by the *s2D* method; these are α-helix (H), β-strand (E) or random coil (C). The box plots are the distributions of the predicted populations in each class, further divided into residues belonging to the IDPs that folds upon binding (blue), to their template proteins (red) and to the dataset of structured proteins employed in "Validation of the *s2D* method on structured proteins" (yellow). In each box the horizontal black line denotes the median of the distribution; the white star, the mean. Notches represent the standard error about the median calculated with 10^5 bootstrap cycles, whiskers extend from the 5th to the 95th percentile of the distribution and boxes from the lower to the upper quartile.

To illustrate how the *s2D* method performs in such cases, we ran it on four IDPs that have been reported to form fuzzy complexes (Fig. 3), that is, complexes that show some residual (e.g., after binding) degree of disorder but are generally amenable for structural studies [41,42]. The predictions of the *s2D* method qualitatively resemble the secondary-structure elements observed in the bound states (Fig. 3). However, these are predicted to be populated at a relatively low extent and, in some cases, the secondary-structure type observed in the bound structure is less populated than random coil in the *s2D* predictions (e.g., third, fourth and seventh α-helix in Fig. 3a; first α-helix in Fig. 3b; the β-strand and the last α-helix in Fig. 3c). These results are compatible with the view that the free-energy landscapes of monomeric IDPs already contain conformations very similar to the bound structures, which are subsequently stabilized upon binding by the presence of the binding partner [43–45].

To further investigate this matter, we used a dataset of 22 IDPs for which a bound structure was available [44]. These proteins contain disordered segments that undergo disorder-to-order transitions upon binding, which have been referred to as molecular recognition features/molecular recognition elements [46–48] or preformed structural elements and prestructured motifs [44,49]. We ran the *s2D* method on all IDPs and on all protein template structures on which IDPs are bound. Figure 4a compares the number of residues in each secondary-structure type according to the *s2D* method, with that obtained from the DSSP assignments of the structures of the complexes. While the agreement is remarkable for the template globular proteins (in red), the *s2D* method slightly underestimates the number of residues in α-helix or β-strand for the bound IDPs (in blue), as expected from the fact that these proteins have been reported as disordered in their monomeric state. However since many IDP residues were predicted to mostly populate structured states (either α-helix or β-strand), we asked whether the predicted populations of such structural motifs were, on average, smaller than those of corresponding motifs in structured proteins. To address this problem, we grouped residues according to the secondary-structure type predicted to be the most populated by the *s2D* method, and we looked at the population of such types in IDPs that fold upon binding, in the template proteins and in the structured proteins of the dataset of "Validation of the *s2D* method on structured proteins", which we employ here as an additional control. The results reveal that structured motifs in IDPs that fold upon binding are predicted to be significantly less populated than the corresponding

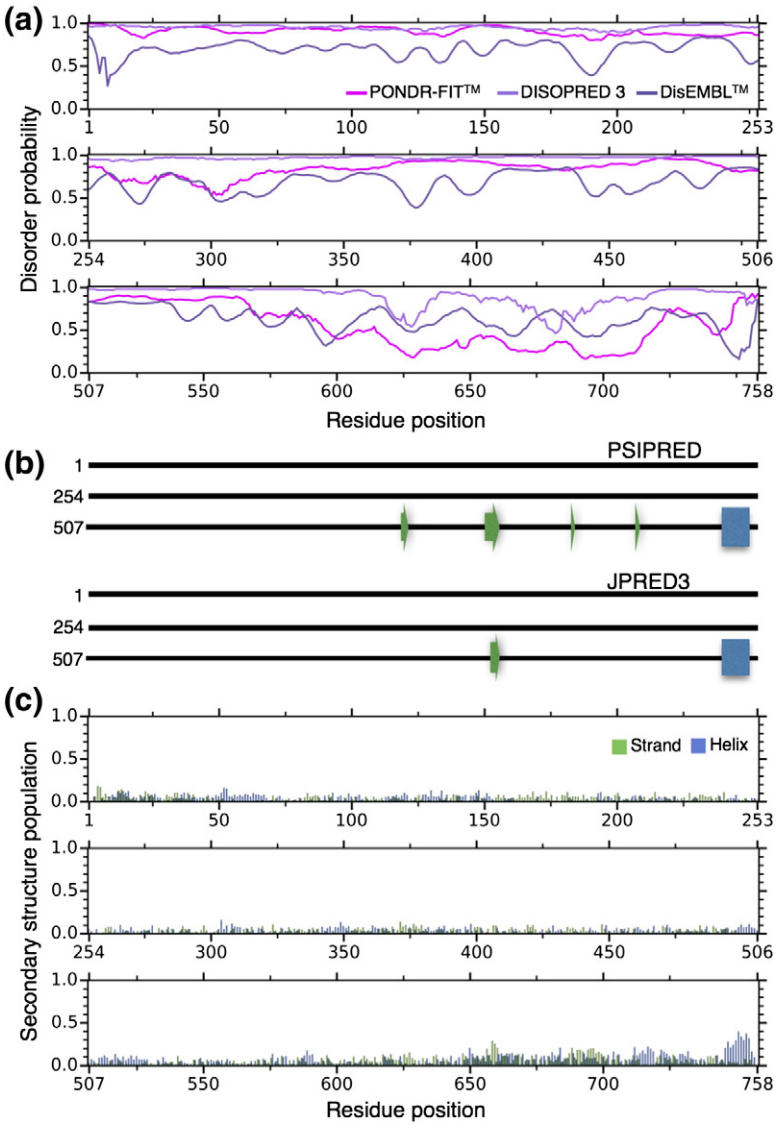

Fig. 5. Analysis of the IDP tau. (a) Results obtained using three different disorder predictors: PONDR-FIT™ [25] (magenta), DISOPRED 3 [18,73] (lilac) and DisEMBL™ [19] (purple). (b) Results obtained with two different secondary-structure predictors: PSIPRED version 3.5 [71,72] (top) and JPRED3 [78] (bottom). Boxes represent α-helices (blue) and arrows represent β-strands (green). (c) Secondary-structure population predicted with the *s2D* method, colored blue for α-helices and green for β-strands.

motifs in structured proteins, while their disordered regions tend to have a larger coil population than the loops in structured proteins (Fig. 4b).

Taken together, these results are compatible with the existence of preformed structural motifs that facilitate the disorder-to-order transition upon binding, and they suggest that the *s2D* method tends to predict the conformational properties of a "mixed" state, where the structural motifs of the bound state are present in the prediction, albeit at a relatively low population.

Applications of the *s2D* method

In order to present the opportunities offered by the *s2D* method of predicting secondary-structure populations, we applied it to well-characterized IDPs and we compared its results with those of existing predictors of secondary structure and of intrinsic disorder. Confronting the prediction of intrinsic disorder with the one of three-state secondary structures (Figs. 5a and b and 6a and b; Fig. S1a and b) reveals the challenge of combining these two types of predictions in order to estimate the extent to which structural elements are populated, also because these two types of predictions are sometimes contradictory.

This problem is illustrated, for instance, by the case of the C-terminal region of tau, which is predicted to contain one α-helix and some short β-strands by PSIPRED and by JPRED3 (Fig. 5b), two of the most commonly used three-state secondary-structure predictors, while it is characterized as highly disordered according to the predictors of

intrinsic disorder that we tested (Fig. 5a). Similarly, the N-terminal region of α-synuclein is predicted to be in an α-helix conformation by PSIPRED and by JPRED3 (Fig. S1b), while it is disordered according to two out of the three predictors of intrinsic disorder that we tested (Fig. S1a). PONDR-FIT™ and DISOPRED3 predict a disorder probability of over 70%, and DisEMBL™ [19] sets it to about 25%. A similar situation is observed for the central and C-terminal regions of Aβ42 (Fig. 6a and b). In particular, residues 5–10 and 37–41 are predicted to have a high propensity for being unstructured by the disorder predictors (Fig. 6a) but to form β-strands by the secondary-structure predictors (Fig. 6b).

A problem in combining the use of disorder and secondary-structure prediction methods is that one does not obtain direct information about the stability of the predicted secondary-structure elements or about the conformational properties of the disordered states, even in those regions where the two types of predictions are not in contrast. The *s2D* method is designed instead to provide such information.

In the case of tau, the *s2D* method predicts the whole protein to be disordered, with some structural motifs moderately populated in the C-terminal region. Interestingly, the C-terminal α-helix predicted by both PSIPRED and JPRED3 (residues 746–755) is present in the *s2D* prediction as well, but it is populated at only about 40% in agreement with the observation that this region is disordered according to the predictors of intrinsic disorder that we tested.

Analogously, in the case of α-synuclein, the *s2D* method predicts the whole sequence to be disordered, with the highest amount of random-coil population in the residues at the C-terminus (residues 95–140; Fig. S1c). This prediction is consistent with experimental measurements on the monomeric state of α-synuclein in solution, by both circular dichroism (CD) [50] and NMR spectroscopy [51], performed also *in vivo* in *Escherichia coli* cells [52]. In addition, the *s2D* method predicts some amount (≈30%) of α-helix population for the first residues of the N-terminus (residues 2–22) and a similar amount of β-strand and α-helix population for residues in the central part (residues 37–95). These predictions are in agreement with recent results that characterized the distinct roles of the different regions of α-synuclein in the process of association with lipid membranes [53]. The N-terminal α-helix (residues 1–25) was shown to form on the membrane acting as a stable anchor that strongly binds the monomer to the lipid layer, while a central α-helix (residues 37–95) is more transiently populated on the membrane and modulates the affinity [53]. In the light of these results, one may expect to find a degree of α-helical population also in the membrane-free solution state of α-synuclein, as predicted by the *s2D* method (Fig. S1c). Likewise, the region of small β-strand population matches the residue range that

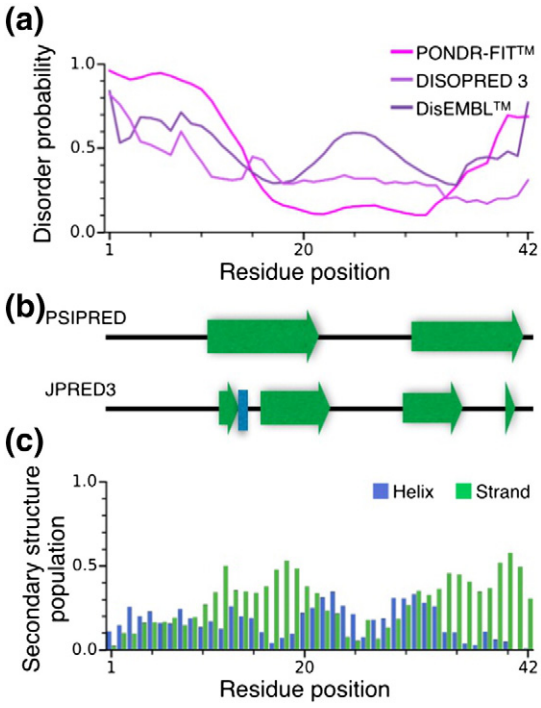

Fig. 6. Analysis of the intrinsically disordered Aβ42 peptide. (a) Results obtained using three different disorder predictors: PONDR-FIT™ [25] (magenta), DISOPRED 3 [18,73] (lilac) and DisEMBL™ [19] (purple). (b) Results obtained with two different secondary-structure predictors: PSIPRED version 3.5 [71,73] (top) and JPRED3 [78] (bottom). Boxes represent α-helices (blue) and arrows represent β-strands (green). (c) Secondary-structure population predicted with the *s2D* method, colored blue for α-helices and green for β-strands.

includes the proposed five strands of the β-sheet sandwich core of the fibrillar structure [54]. Thus, the predicted populations of β-strand and α-helix are in qualitative agreement with existing experimental data. While the populations of the structural motifs predicted by the *s2D* method slightly overestimate those calculated from the chemical shifts [28]—by an amount compatible with the results in Table 1—they are of similar magnitude to those reported in a recent study that combines NMR chemical sifts, residual dipolar couplings and small-angle scattering to obtain an ensemble description of α-synuclein [55]. However, while the *s2D* predictions suggest a higher degree of random coil in the C-terminal region (Fig. S1c, observed also in the chemical shift analysis [28] and in solid-state NMR [53]), the populations reported in Ref. [55] are more uniform along the sequence.

Similarly, in agreement with the *s2D* predictions, the Aβ42 peptide has been reported to be mostly disordered at physiological conditions, as assessed by different techniques including CD, NMR and molecular dynamics simulations [56]. However, most

kind of secondary structures, including α-helices and β-strands, can become significantly populated as a result of changes in conditions [57,58] (e.g., temperature and pH) or formulation [59,60] (e.g., presence of trifluoroethanol or micelles), suggesting that states other than random coil might be marginally populated even at physiological conditions.

To illustrate the applicability of the *s2D* method to large-scale analysis, we used it to investigate the human cytosolic proteome. The analysis reveals that a large number of proteins (or protein domains) for which a crystal structure is not yet available have a predicted degree of disorder similar to that found in proteins successfully crystallized. In particular, the distribution of mean coil population of protein domains (or whole protein sequences in some instances) that have been employed in crystallographic experiments has a large overlap with that of proteins that have not yet been crystallized (blue and yellow distributions in Fig. 7b). In addition, the distribution of the mean coil population across the human proteome is fully consistent with previously reported results stating that IDPs, or proteins highly enriched in disordered regions, constitute a large fraction of the proteome itself [18,61].

The analysis presented in Fig. 7 was carried out extracting from the human reference proteome of the UniProt Web site protein sequences with assigned cytosolic subcellular location. Non-X-ray structures were excluded from the mapping to the PDB provided by UniProt [62]. Such mapping relies the sequence that was used in the crystallography experiment and not on those residues for which coordinates have actually been assigned. This implies that some of the sequences considered as "crystallized" might comprise a large number of residues with missing electron density (remark 465). This likely explains the tail toward disordered states observed in the yellow distribution in Fig. 7b.

Discussion

In this work, we have described the *s2D* method of predicting secondary-structure populations of proteins from their amino acid sequences. A 10-fold cross-validation procedure was applied, showing that the *s2D* method can predict secondary-structure populations with a mean absolute error of about 0.14 and a mean coefficient of correlation of 0.77.

We validated the predictions provided by the *s2D* method on three independent datasets containing protein sequences not employed in training or testing: a first dataset of structured proteins, a second one of disordered proteins (the DisProt database [35]) and a third one of structured proteins enriched in disordered regions, as assessed by regions of missing electron density in the X-ray structures. The results show that the *s2D* method predicts the secondary-structure populations assigned to the X-ray structures with a *Q3* score above 79% and can identify disordered regions with an accuracy of about 85–88%, depending on the definition of disorder employed. We suggest that the scores from these validations may represent lower bounds of the actual performance because some secondary-structure elements that are favored

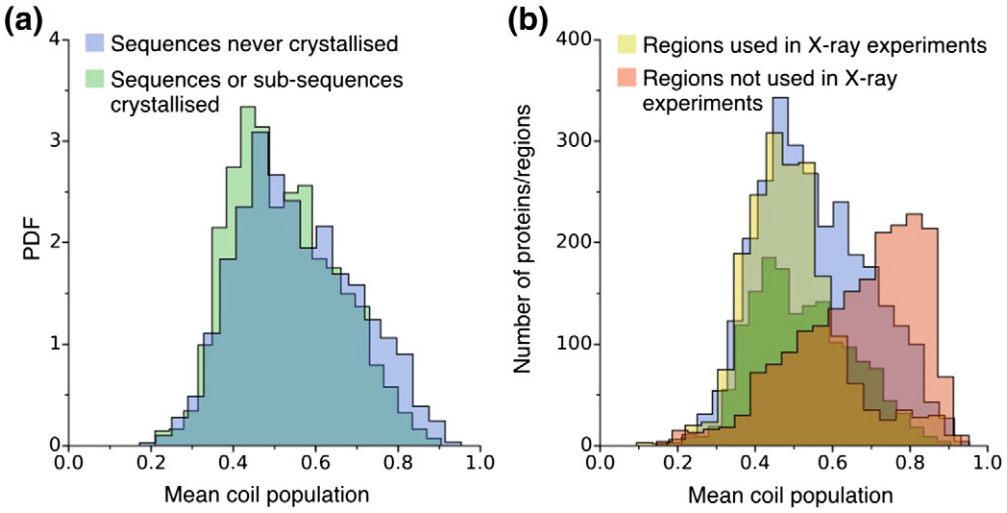

Fig. 7. *s2D* analysis of the human cytosolic proteome. (a) Probability distribution of the mean random-coil population of protein sequences that have been crystallized fully or in part (green) superimposed to the one of sequences that have never been crystallized (blue). (b) Histogram reporting the two distributions in (a) (here non-normalized) and the distributions of the mean coil population of subsequences of proteins from the green distribution that have been used in an X-ray experiments (yellow) and that have not (orange).

energetically but not entropically are stabilized under crystallization conditions [33,34]. Such motifs would thus be observed in the crystals but would not be highly populated in solution, where the NMR measurements used to train the *s2D* method were performed. The results in Fig. 2a are consistent with this possibility for a large fraction of the residues for which the *s2D* prediction is in disagreement with the crystal structure assignment. Similarly, possible structural preferences of disordered states are not embodied in these validation datasets. Therefore, there is currently no fully reliable way to tell whether the small populations of secondary structure, which *s2D* sometimes predicts in disordered states, are accurately or erroneously predicted. Taking these considerations into account, for structured proteins, the performance of the *s2D* method on these validation datasets is comparable to that of the state-of-the-art predictors of secondary structures [30]. At the same time, for disordered proteins, its performance is also comparable to that the state-of-the-art predictors of intrinsic disorder [22].

Our motivation for developing the *s2D* method is that, although disorder and structure are intimately related concepts, it is often not immediate to obtain an overall view of the conformational properties of a protein molecule by looking simultaneously at existing predictions of secondary structure and intrinsic disorder. In fact, most existing secondary-structure predictors have been built by exploiting static information extracted from X-ray structures of native states of proteins, and thus, they may not be ideally suited to account for protein dynamics. Conversely, the majority of the existing disorder predictors have been developed by defining disorder as the "absence of structure" (e.g., from the missing electron densities in X-ray structures) and may not contain specific characterizations of the conformational properties of disordered states. These differences result for many proteins in large gaps between the two types of prediction, which prevents a clear quantification of the structure and dynamics of the molecule under scrutiny. To address this problem, the *s2D* method bridges the gap by predicting directly the populations of secondary-structure elements, which allows for a unified prediction of structure and disorder.

We anticipate that, with the progressive increase in the size of the Biological Magnetic Resonance Data Bank (BMRB) database and the growing accuracy of chemical shifts analysis techniques, it will become possible to train the *s2D* method on a larger and less noisy database. This development would both improve the performance of the method and allow for a more detailed characterization of the populations of different secondary-structure elements, which may include other types of secondary structures, such as polyproline II. Similarly, improvements in the resolution of chemical shift analysis will allow for a more finely tuned characterization of the coil states, distinguishing between highly dynamic states typical of fully disordered regions and relatively static states found in some globular proteins [29,63] (Fig. S2).

Quite generally, the possibility of performing sequence-based predictions of statistical populations of ordered and disordered regions will enable a rapid and quantitative assessment of the structure and dynamics of proteins and proteomes.

Materials and Methods

Artificial neural networks with ELMs

The *s2D* predictor relies on the machine-learning algorithm of single-hidden layer feedforward neural networks (SLFNs) trained with ELMs [64]. An important aspect of ELMs is that, differently from many other learning methods, their universal approximation capability has been proven [65]. In addition, a major practical advantage of ELMs is that their speed during the learning phase allows a large number of different models to be tested.

Unlike other learning algorithms, ELMs randomly assign the weights of the connections between the input and the hidden layer (the hidden weights, W) and keep them fixed during training. Only the weights of the connections between the hidden layer and the output layer (the output weights, β) are actually trained. Since the hidden weights are held fixed, the output weights can be analytically determined, yielding a learning speed that can be thousands of times faster than the one of conventional approaches for training feedforward neural networks [65].

The output vector $\mathbf{o} = \{o_1, \cdots, o_m\}$ of a SLFN can be mathematically expressed as

$$o_j = \sum_{i=1}^{N_h} \beta_{ji} h_i = \sum_{i=1}^{N_h} \beta_{ji} g(\boldsymbol{w}_i \cdot \boldsymbol{x}) \quad (1)$$

where the dot denotes the inner product, $\mathbf{h}$ is the hidden-neuron vector containing N_h hidden neurons that includes a bias neuron with $h_b = 1$, g is the activation function, $\mathbf{x} = \{x_1, \cdots, x_n\}^T$ is the input vector formed of n input neurons and $\mathbf{w}_i$ is the weight vector connecting the ith hidden neuron with the input neurons. The idea behind training is that, given N arbitrary distinct samples (x_k, t_k), where $\boldsymbol{t}_k = \{t_{1k}, \cdots, t_{mk}\}^T$ represents the observed output (target of the training) corresponding to the input $\boldsymbol{x}_k = \{x_{1k}, \cdots, x_{nk}\}^T$, there exist two matrices β and W such that $\sum_{k=1}^{N} \|\boldsymbol{o}_k - \boldsymbol{t}_k\| = 0$.

In the context of ELMs, the weights in the matrix W are held fixed and, in practice, the output weights β, represented by a matrix of size $m \times N_h$, can be determined by fitting the target output $T = [\boldsymbol{t}_1, \cdots, \boldsymbol{t}_N]_{m \times N}$ using a least-square approach [66] so that

$$\beta = TH^{-1} \quad (2)$$

where H^{-1} is the Moore–Penrose generalized inverse of the matrix $H = [\mathbf{h}_1, \cdots, \mathbf{h}_N]_{N_h \times N}$ [67]. It can be shown that this choice for H^{-1} minimizes both $\|T - \beta H\|$ and $\|\beta\|$ [65].

The *δ2D* and *s2D* datasets

To train and test the *s2D* predictor, we assembled a dataset of protein sequences with assigned NMR chemical shifts, which were then used to calculate the corresponding secondary-structure populations using the *δ2D* method [28]. We downloaded the full release of the BMRB [68] and filtered it in order to obtain a database representative of the conformational properties of protein molecules when present as free monomers in solution. Molecules other than proteins, proteins in complex and chemical shifts measured at extreme experimental conditions were left out. Included entries corresponded to experiments carried out in solution, at temperatures between 10 and 42 °C and at a pH between 5.5 and 8. These threshold values were chosen as a trade-off between measurements at physiological conditions and number of entries in the resulting dataset.

Moreover, since we wanted the *s2D* method to predict the secondary-structure populations of monomeric states in solution, we excluded entries with samples containing micelles, denaturants or other compounds that might affect the behavior of the protein. The remaining entries were used for the *δ2D* calculation.

As some of the sequences had some residues with an insufficient number of chemical sifts assigned, the resulting *δ2D* secondary-structure populations had some gaps. We analytically continued the secondary-structure population profiles for sequences where at least 70% of the residues had enough assigned chemical shifts and no individual gap was longer than four consecutive amino acids. Sequences not matching any of these two criteria were discarded.

This procedure resulted in the *δ2D* dataset, which contained 2223 protein sequences with their corresponding secondary-structure population profiles. Since in this dataset there is a degree of noise due to the variety of experimental conditions included, and of sequence similarity, it would be desirable to increase its size to be able to extract sufficiently large training and test sets for the *s2D* predictor. In addition, entries in the *δ2D* dataset have secondary-structure populations biased toward disordered random coil (Fig. S3a), with only 18% of the residues preferentially in β-strand conformations. While this probably reflects the relevance of NMR in studying disordered states of proteins, it may bias the training of *s2D* toward disordered states. Consequently, we sought to increase the size of the dataset, with particular attention to incorporating more residues in a β-strand conformation.

Having exhausted the available protein sequences with assigned chemical shifts, we decided to exploit the large number of X-ray structures present in the PDB. In order to match the secondary structures in the crystal state to the secondary-structure populations characteristic of solution states, we considered that the structures formed by proteins belonging to thermophilic organisms would be highly stable in the temperature range of the *δ2D* dataset. Thus, we started with 10,649 PDB chains of proteins from thermophilic organisms and filtered them for gapless chains, belonging to non-membrane and non-nucleotide-bound proteins, with less than 25% sequence identity with each other and at least 18% of the residues in β-strand conformations. This yielded 448 sequences (Fig. S4 and Supplementary File 1), which were added to the *δ2D* dataset assigning a population of 1 to their secondary-structure elements (determined with the DSSP program [31]), thus resulting in the *s2D* dataset of 2671 sequences.

This procedure increased the number of sequences by about 20% and helped remove biases toward disordered states by increasing the number of residues in β-strand conformations by 83% and in α-helix conformations by 47% (Fig. S3). In the *s2D* dataset, we introduced some additional noise since assigning a population of 1 to the secondary-structure elements is likely to overestimate their stability to some extent despite having used only highly stable proteins from thermophilic organisms. However, we anticipate that the growing size of the BMRB database will soon allow building a new, larger *δ2D* dataset derived solely from NMR chemical shifts and containing minimal amount of noise, which will further improve the performance of the *s2D* method.

In order to extract a subset of sequences from the *s2D* dataset to be used for testing, as performed in Ref. [66], we clustered the sequences by local similarity (computed with BLAST [69]) with a 25% sequence identity cutoff. From these clusters, we created 10 different groups of sequences that internally confined local sequence homology (Supplementary File 1), which were then used to perform the 10-fold cross-validation procedure.

The architecture of the *s2D* method

The *s2D* predictor is built by using a combination of artificial neural networks trained with ELMs. This procedure consists in a three-step iterative prediction of secondary-structure populations, which allows to include in the input for the final prediction both global and local information about the protein sequence.

For each protein target, a sequence profile is generated using the position-specific scoring matrix (PSSM) obtained with the PSI-BLAST program [70], and one additional parameter is assigned to each position along the sequence to represent the identity of the amino acid actually found at that position (Table S1). Consequently, each residue in a protein sequence is represented by 22 input values (21 come from the PSSM and 1 from Table S1).

The PSI-BLAST search is carried out with a procedure similar to the one described in Ref. [71]. It is performed against the UniRef90 dataset [72], which was filtered to remove trans-membrane segments, coiled-coil and low-complexity regions using the *pfilt* program distributed with PSIPRED [73]. The PSSM information is extracted from the ASN.1 checkpoint file rather than from the actual output of PSI-BLAST, as BLAST+ [69] by default saves PSSM data after rounding them down, hence loosing precision.

(i) The first step of the *s2D* prediction is performed using two SLFNs. The two networks are identical in architecture, but the first employs a sliding window of 11 amino acids and the second a sliding window of 15 amino acids, for a total of 242 and 330 input neurons, respectively. Each network has 3 output neurons, corresponding to the secondary-structure populations of α-helix, β-strand and random coil of the central residue in the sliding window. Vacant locations in the windows around residues near the termini of a protein are assigned all zeros as input values. Both networks have 4000 hidden nodes. For comparison, since the first layer of weights is randomly

assigned in the context of ELMs, an SLFN with 4000 hidden nodes trained with ELMs is equivalent, in terms of number of free parameters, to an SLFN trained with standard learning algorithms (e.g., back--propagation) with 242 input and 3 output nodes (like ours) and 49 hidden nodes.

(ii) The second step consists in predicting the mean secondary-structure populations of the entire protein (mean populations of α-helix, β-strand and random coil, calculated from the assigned chemical shifts), using an N-to-1 network. N-to-1 networks are aimed at encoding a whole sequence into a single object, overcoming the machine-learning problem of the variable length of biological sequences [74]; their formulation with ELMs is described in Ref. [66]. In the framework of the *s2D* method, the N-to-1 network uses as input, for each amino acid in the sequence, the 22 values previously described and the 6 additional values corresponding to the outputs of the two networks employed in the step (i). This network has 150 hidden nodes and predicts three output values: the mean populations of α-helix, β-strand and random coil of the protein sequence, as calculated from the chemical shifts using the *δ2D* method. This approach is superior to simply calculating by averaging over the sequence the mean populations from the predictions of the two networks in step (i). The N-to-1 network, exploiting the information contained in the sequence profile, is able to overcome possible inaccuracies of the first prediction, yielding more precise mean populations.

(iii) The third and final step of the *s2D* prediction employs again an SLFN. However, differently from the first step, this network has a sliding window of 5 amino acids and 3600 hidden nodes and uses the same input values per residue as the N-to-1 network of step (ii), plus the mean populations predicted by the N-to-1 network itself. Consequently, this network has access to the predicted secondary-structure populations of all five residues covered by its window and thus can correct the predictions for the central one using the predicted populations of the residues at its sides. Moreover, the N-to-1 provides it with valuable information about the global properties of the protein sequence under scrutiny, which can be exploited to partially account for the stabilizing effect of the tertiary structure. For instance, chances are that the structural motifs of a protein that is mostly disordered are not stabilized by tertiary contacts, while the opposite can be conjectured for highly structured proteins.

Model selection

The ELM models described above depend on specific parameters (amino acid representation, scaling parameter of the random weights, number of hidden neurons, the size of the input window) that influence the overall predictive capacity of the *s2D* method and need to be optimized. Moreover, different combinations of networks in the three-step iterative prediction can affect the performance of the *s2D* method, and the best combination needs to be selected.

Although we did not systematically search through the vast number of possible combinations of networks and network topologies, many combinations were tried (data not shown). Model selection was carried out by training the networks on 9 of the 10 subsets of sequences described at the end of "The *δ2D* and *s2D* datasets" and by using the remaining one as a benchmark (testing set). This approach ensures that the local sequence identity between training and testing proteins is always below 25%.

The comparison between the different network combinations and topologies that were tested was performed using as performance indicators the Pearson's correlation coefficients between the predicted and the observed populations of the three secondary-structure types considered (R_H, R_E, R_C) and the corresponding mean square errors (MSE_H, MSE_E, MSE_C) and mean absolute errors (MAE_H, MAE_E, MAE_C), all evaluated on the sequences in the testing set.

While for the two networks of step (i) and for the final network of step (iii) we found that different sets of random weights W did not have any significant influence on the performance indicators and we observed a weak dependence for the N-to-1 network of step (ii), consistently with previous reports for this type of networks [66]. These results are likely to reflect the fact that the stochastic nature of ELMs becomes more apparent with a smaller number of free parameters. Thus, in order to take account of the dependence on the initial weights, 5 sets of random weights were generated for each tried N-to-1 architecture. The N-to-1 network parameters were then tuned using the mean values of the performance indicators, and the network that best performed on the testing set was used to train the final network of step (iii).

The same training and testing sequences were used in model selection for all tried combinations of networks and network topologies, and only the selected model (described in "The architecture of the *s2D* method") was subjected to a thorough 10-fold cross-validation procedure (see "10-Fold cross-validation of the *s2D* method").

Availability

The *s2D* method is available as a Web server[†] and executable and source code can be downloaded from the same website under the GNU General Public License.

Funding

This work was supported by the Biotechnology and Biological Sciences Research Council (UK).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2014.12.007.

Received 23 September 2014;
Received in revised form 10 December 2014;

Accepted 12 December 2014
Available online 20 December 2014

Keywords:
intrinsically disordered proteins;
alpha-helix;
beta-sheet;
random coil

†at http://www-mvsoftware.ch.cam.ac.uk

Abbreviations used:
IDP, intrinsically disordered protein; ROC, receiver operating characteristic; BMRB, Biological Magnetic Resonance Data Bank; SLFN, single-hidden layer feedforward neural network; ELM, extreme learning machine; PDB, Protein Data Bank; PSSM, position-specific scoring matrix.

References

[1] Frauenfelder H, Sligar SG, Wolynes PG. The energy landscapes and motions of proteins. Science 1991;254:1598–603.
[2] Fersht A. Structure and Mechanism in Protein Science. W. H. Freeman–Macmillan; 1999.
[3] Karplus M, Kuriyan J. Molecular dynamics and protein function. Proc Natl Acad Sci U S A 2005;102:6679–85.
[4] Mittermaier A, Kay LE. New tools provide new insights in NMR studies of protein dynamics. Science 2006;312:224–8.
[5] Vendruscolo M, Dobson CM. Structural biology. Dynamic visions of enzymatic reactions. Science 2006;313:1586–7.
[6] Boehr DD, Nussinov R, Wright PE. The role of dynamic conformational ensembles in biomolecular recognition. Nat Chem Biol 2009;5:789–96.
[7] Habchi J, Tompa P, Longhi S, Uversky VN. Introducing protein intrinsic disorder. Chem Rev 2014;114:6561–88.
[8] Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol 2008;18:756–64.
[9] Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: regulation and disease. Curr Opin Struct Biol 2011;21:1–9.
[10] Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys 2008;37:215–46.
[11] Tompa P. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 2005;579:3346–54.
[12] Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005;6:197–208.
[13] Dobson CM. Protein folding and misfolding. Nature 2003;426:884–90.
[14] He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Res 2009;19:929–49.
[15] Deng X, Eickholt J, Cheng J. A comprehensive overview of computational protein disorder prediction methods. Mol BioSyst 2012;8:114–21.
[16] Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins 2003;53:566–72.
[17] Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005;21:3433–4.
[18] 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. J Mol Biol 2004;337:635–45.
[19] Linding R, Jensen LJ, Diella F, Bork P, Gibson TJ, Russell RB. Protein disorder prediction: implications for structural proteomics. Structure 2003;11:1453–9.
[20] Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 2005;21:3435–8.
[21] Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins 2006;65:1–14.
[22] Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. Proteins 2013;82:127–37.
[23] Yan J, Mizianty MJ, Filipow PL, Uversky VN, Kurgan L. RAPID: fast and accurate sequence-based prediction of intrinsic disorder content on proteomic scale. Biochim Biophys Acta 1834;2013:1671–80.
[24] Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF. From protein sequence to dynamics and disorder with DynaMine. Nat Commun 2013;4:2741.
[25] Xue Bin, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 2010;1804:996–1010.
[26] Lieutaud P, Canard B, Longhi S. MeDor: a metaserver for predicting protein disorder. BMC Genomics 2008;9:S25.
[27] Yachdav G, Kloppmann E, Kajan L, Hecht M, Goldberg T, Hamp T, et al. PredictProtein—an open resource for online prediction of protein structural and functional features. Nucleic Acids Res 2014;42:W337–43.
[28] Camilloni C, De Simone A, Vranken WF, Vendruscolo M. Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. Biochemistry 2012;51:2224–31.
[29] De Simone A, Cavalli A, Hsu S-TD, Vranken W, Vendruscolo M. Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. J Am Chem Soc 2009;131:16332–3.
[30] Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y. SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. J Comput Chem 2012;33:259–67.
[31] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–637.
[32] Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–3.
[33] Acharya KR, Lloyd MD. The advantages and limitations of protein crystal structures. Trends Pharmacol Sci 2005;26:10–4.
[34] Swaminathan GJ, Holloway DE, Colvin RA, Campanella GK, Papageorgiou AC, Luster AD, et al. Crystal structures of oligomeric forms of the IP-10/CXCL10 chemokine. Structure 2003;11:521–32.
[35] Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, et al. DisProt: the database of disordered proteins. Nucleic Acids Res 2007;35:D786–93.

[36] Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK. Intrinsic disorder in the Protein Data Bank. J Biomol Struct Dyn 2007;24:325–42.
[37] Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinf 2006;7:208.
[38] van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. Chem Rev 2014;114: 6589–631.
[39] Pancsa R, Fuxreiter M. Interactions via intrinsically disordered regions: what kind of motifs? IUBMB Life 2012;64:513–20.
[40] Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol 2002;12:54–60.
[41] Tompa P, Fuxreiter M. Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem Sci 2008;33:2–8.
[42] Hazy E, Tompa P. Limitations of induced folding in molecular recognition by intrinsically disordered proteins. ChemPhysChem 2009;10:1415–9.
[43] Tsai C-J, Ma B, Sham YY, Kumar S, Nussinov R. Structured disorder and conformational selection. Proteins 2001;44: 418–27.
[44] Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J Mol Biol 2004;338:1015–26.
[45] Espinoza-Fonseca LM. Reconciling binding mechanisms of intrinsically disordered proteins. Biochem Biophys Res Commun 2009;382:479–82.
[46] Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, et al. Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 2007;6:2351–66.
[47] Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry 2005; 44:12454–70.
[48] Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, et al. Analysis of molecular recognition features (MoRFs). J Mol Biol 2006;362:1043–59.
[49] Lee S-H, Kim D-H, Han JJ, Cha E-J, Lim J-E, Cho Y-J, et al. Understanding pre-structured motifs (PreSMos) in intrinsically unfolded proteins. Curr Protein Pept Sci 2012;13:34–54.
[50] Uversky VN, Li J, Fink AL. Evidence for a partially folded intermediate in alpha-synuclein fibril formation. J Biol Chem 2001;276:10737–44.
[51] Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. J Am Chem Soc 2005;127:476–7.
[52] Waudby CA, Camilloni C, Fitzpatrick AWP, Cabrita LD, Dobson CM, Vendruscolo M, et al. In-cell NMR characterization of the secondary structure populations of a disordered conformation of α-synuclein within *E. coli* cells. PLoS One 2013;8:e72286.
[53] Fusco G, Gopinath T, Vostrikov V, Vendruscolo M, De Simone A, Dobson CM, et al. Direct observation of the three regions in α-synuclein that determine its membrane-bound behaviour. Nat Commun 2014;5:1–8.
[54] Vilar M, Chou H-T, Lührs T, Maji SK, Riek-Loher D, Verel R, et al. The fold of α-synuclein fibrils. Proc Natl Acad Sci U S A 2008;105:8637–42.
[55] Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, et al. Predictive atomic resolution descriptions of intrinsically disordered hTau40 and α-synuclein in solution from NMR and small angle scattering. Structure 2014;22: 238–49.
[56] Hamley IW. The amyloid beta peptide: a chemist's perspective. Role in Alzheimer's and fibrillization. Chem Rev 2012; 112:5147–92.
[57] Danielsson J, Jarvet J, Damberg P, Gräslund A. The Alzheimer beta-peptide shows temperature-dependent transitions between left-handed 3-helix, beta-strand and random coil secondary structures. FEBS J 2005;272: 3938–49.
[58] Barrow CJ, Zagorski MG. Solution structures of beta peptide and its constituent fragments: relation to amyloid deposition. Science 1991;253:179–82.
[59] Shao H, Jao S, Ma K, Zagorski MG. Solution structures of micelle-bound amyloid beta-(1-40) and beta-(1-42) peptides of Alzheimer's disease. J Mol Biol 1999;285:755–73.
[60] Barrow CJ, Yasuda A, Kenny PT, Zagorski MG. Solution conformations and aggregational properties of synthetic amyloid beta-peptides of Alzheimer's disease. Analysis of circular dichroism spectra. J Mol Biol 1992;225:1075–93.
[61] Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. J Mol Graph Model 2001;19:26–59.
[62] Gasteiger E, Jung E, Bairoch A. SWISS-PROT: connecting biomolecular knowledge via a protein database. Curr Issues Mol Biol 2001;3:47–55.
[63] Liu J, Tan H, Rost B. Loopy proteins appear conserved in evolution. J Mol Biol 2002;322:53–64.
[64] Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. Neurocomputing 2006;70: 489–501.
[65] Huang G-B, Chen L, Siew C-K. Universal approximation using incremental constructive feedforward networks with random hidden nodes. IEEE Trans Neural Netw 2006;17: 879–92.
[66] Savojardo C, Fariselli P, Casadio R. Improving the detection of transmembrane-barrel chains with N-to-1 extreme learning machines. Bioinformatics 2011;27:3123–8.
[67] Huang G-B. Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Trans Neural Netw 2003;14:274–81.
[68] Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Res 2007;36: D402–8.
[69] Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf 2009;10:421.
[70] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[71] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292: 195–202.
[72] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics 2007;23:1282–8.
[73] Buchan DWA, Minneci F, Nugent TCO, Bryson K, Jones DT. Scalable Web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Res 2013;41:W349–57.
[74] Mooney C, Wang YH, Pollastri G. SCLpred: protein subcellular localization prediction by N-to-1 neural networks. Bioinformatics 2011;27:2812–9.

[75] van Leeuwen HC, Strating MJ, Rensen M, de Laat W, van der Vliet PC. Linker length and composition influence the flexibility of Oct-1 DNA binding. EMBO J 1997;16:2043–53.
[76] Simon SM, Sousa FJR, Mohana-Borges R, Walker GC. Regulation of *Escherichia coli* SOS mutagenesis by dimeric intrinsically disordered *umuD* gene products. Proc Natl Acad Sci U S A 2008;105:1152–7.
[77] Renault L, Bugyi B, Carlier M-F. Spire and Cordon-bleu: multifunctional regulators of actin dynamics. Trends Cell Biol 2008;18:494–504.
[78] Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. Nucleic Acids Res 2008;36:W197–201.