Cell Biochem Biophys (2013) 67:1193–1205
DOI 10.1007/s12013-013-9638-0

ORIGINAL PAPER

Intrinsically Semi-disordered State and Its Role in Induced Folding and Protein Aggregation

Tuo Zhang · Eshel Faraggi · Zhixiu Li · Yaoqi Zhou

Published online: 31 May 2013

Abstract Intrinsically disordered proteins (IDPs) refer to those proteins without fixed three-dimensional structures under physiological conditions. Although experiments suggest that the conformations of IDPs can vary from random coils, semi-compact globules, to compact globules with different contents of secondary structures, computational efforts to separate IDPs into different states are not yet successful. Recently, we developed a neural-network-based disorder prediction technique SPINE-D that was ranked as one of the top performing techniques for disorder prediction in the biannual meeting of critical assessment of structure prediction techniques (CASP 9, 2010). Here, we further analyze the results from SPINE-D prediction by defining a semi-disordered state that has about 50 % predicted probability to be disordered or ordered. This semi-disordered state is partially collapsed with intermediate levels of predicted solvent accessibility and secondary structure content. The relative difference in compositions between semi-disordered and fully disordered regions is highly correlated with amyloid aggregation propensity (a correlation coefficient of 0.86 if excluding four charged residues and proline, 0.73 if not). In addition, we observed that some semi-disordered regions participate in induced folding, and others play key roles in protein aggregation. More specifically, a semi-disordered region is amyloidogenic in fully unstructured proteins (such as alpha-synuclein and Sup35) but prone to local unfolding that exposes the hydrophobic core to aggregation in structured globular proteins (such as SOD1 and lysozyme). A transition from full disorder to semi-disorder at about 30–40 Qs is observed in the poly-Q (poly-glutamine) tract of huntingtin. The accuracy of using semi-disorder to predict binding-induced folding and aggregation is compared with several methods trained for the purpose. These results indicate the usefulness of three-state classification (order, semi-disorder, and full-disorder) in distinguishing non-folding from induced-folding and aggregation-resistant from aggregation-prone IDPs and in locating weakly stable, locally unfolding, and potentially aggregation regions in structured proteins. A comparison with five representative disorder-prediction methods showed that SPINE-D is the only method with a clear separation of semi-disorder from ordered and fully disordered states.

Keywords Intrinsically disordered proteins · Induced folding · Amyloid formation · Poly-Q · SOD1

T. Zhang · E. Faraggi · Z. Li · Y. Zhou (✉)
School of Informatics, Indiana University Purdue University, Indianapolis, IN 46202, USA
e-mail: yqzhou@iupui.edu

T. Zhang · E. Faraggi · Z. Li · Y. Zhou
Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA

Introduction

The origin of protein aggregation and amyloid formation is poorly understood for intrinsically disordered proteins (IDPs) that do not have a fixed three-dimensional structure in physiological conditions. Some IDPs are resistant to protein aggregation while others are directly involved in amyloid formation [1]. Similarly, some IDPs can have a fixed structure under some physiological conditions, for example, when interacting with other molecules (folders) while others are so-called nonfolders that do not fold into a unique structure under any known conditions [2]. What makes some IDPs foldable or aggregation prone is an open

question, although such divergent behaviors of IDPs are likely related to their inherently diverse types of conformations ranging from random coils, semi-compact globules, to compact globules with varying content of secondary structures [2–4]. Differences in structural shapes of IDPs led to proposed multi-state concepts such as "protein trinity" (order, collapsed, and extended disorder) [5] and "protein quartet" (folded structure, molten globule, pre-molten globule, and coil) [6]. The latter states have a one-to-one correspondence to surface-molten solid, ordered globule, disordered globule, and coil discovered for a model three-helix bundle protein [7]. However, it is not clear whether these states are discrete (i.e., separable) or continuous (inseparable) based on sequence information alone [8]. Clustering disordered sequences into groups was not successful [9]. A neural network method [10] was developed by iteratively partitioning disordered sequences into separate "flavors" for different predictors. The resulting three flavors of disorder, however, do not naturally separate extended from collapsed disordered proteins. All other methods developed so far (>50) are dedicated to a two-state prediction of order and disorder [11].

Recently, we developed a sequence-based prediction method with integrated neural networks for disorder (SPINE-D) [12] that was ranked as one of the top five performing methods according to area under the curve in the critical assessment of structure prediction techniques (CASP 9) [12, 13]. For a given protein sequence, SPINE-D predicts the probability of each amino-acid residue in the sequence to be disordered. Here, we found that defining a semi-disordered state about the 50 % disorder probability predicted by SPINE-D is useful for identifying semi-collapsed and semi-structured regions compared with ordered and fully disordered regions. The semi-disordered state is associated with folders and aggregation-prone regions in disordered proteins and weakly stable or locally unfolded regions in structured proteins.

Results

Defining Semi-disorder

SPINE-D was trained by a large database of 4,229 (4,157 + 72) non-redundant proteins with 90 % ordered residues and 10 % disordered residues [12]. This unbalanced dataset led to a threshold of 0.06 for predicted probability when optimized for the highest accuracy. That is, residues are assigned as ordered if the predicted probability is less than 0.06 and residues are assigned as disordered if the predicted disorder probability is greater than 0.06. However, for a two-state classification, a perfect threshold should be at a probability of 0.5 when there is an equal probability of being ordered and disordered. To change the low threshold of 0.06 to a more natural threshold of 0.5 as required by CASP, we linearly scaled from 0–0.06 to 0–0.5 and 0.06–1 to 0.5–1. The simple linear scaling was employed because it is parameter free. Such rescaled probability of SPINE-D was employed with success for disorder prediction in CASP 9 [12, 13].

Separate scaling for ordered and disordered regions led to an unintended discontinuity for the distribution of predicted disorder probabilities at the probability $P = 0.5$ as shown in Fig. 1b for all three datasets SL477, DX4080, and Control703 that respectively represent re-annotated disordered proteins from Disprot [12, 14, 15], high-resolution X-ray structures (ordered proteins) with residues without coordinates as disordered residues [12], and a negative control set of stably folded monomeric proteins without cofactors and without missing coordinates (see materials and methods). The distributions are based on predicted disorder probabilities for all sequence regions of proteins regardless if they were annotated or not annotated with disorder or order. This discontinuity, not observed before

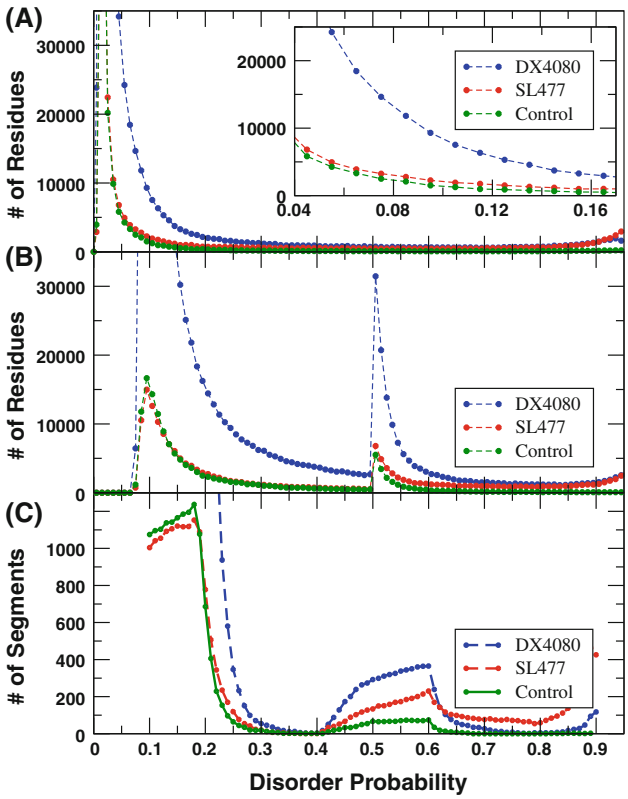

Fig. 1 The distribution of disorder probability predicted by SPINE-D at residue level before (**a**) and after scaling (**b**) and at long segment level (>30 amino acid residues) (**c**) for three datasets (DX4080, SL477, and Control703). The insert in (**a**) shows the fine detail around the disorder probability of 0.06. The negative control set (stable monomeric proteins) does not have a peak for fully disordered residues or regions, indicating the usefulness of separating semi-disorder from full disorder

scaling (Fig. 1a), occurs because the population of amino acid residues in an ordered region (0–0.06) is diluted into a wider range between 0 and 0.5, while the population of amino acid residues in the disordered region (0.06–1) is concentrated to a narrower range between 0.5 and 1.

This population around $P = 0.5$ in Fig. 1b is not created by isolated residues but mostly by segments in which all residues have P around 0.5. In Fig. 1c, we counted the number of long segments (>30 residues) within a given disorder probability plus/minus 0.1. There is a significant population with long sequence regions with semi-disordered probability, separated from ordered ($P \sim 0$) and fully disordered ($P \sim 1$) states. Based on Fig. 1c, we define three states for residues: $0 \leq P < 0.4$ as the ordered state, $0.4 \leq P \leq 0.7$ as the semi-disordered state, and $0.7 < P \leq 1$ as the fully disordered state. The negative control set (stable monomeric proteins) does not have a peak for fully disordered residues or regions. This indicates the usefulness of separating semi-disorder from full-disorder. This definition of three states is somewhat arbitrary. We did not make any attempts to optimize the definition for these states. A slightly different definition will not significantly change the results presented here.

Although this population of the semi-disordered state arose from separate linear scaling, rescaling the threshold for order/disorder transition to 50 % probability itself is physically meaningful. Thus, it is of interest to investigate whether this semi-disordered state is a purely mathematical artifact or a physically meaningful state for proteins.

Characterization of the Semi-disordered State

To characterize a semi-disordered state, we compare fractions of secondary structures (helical and strand residues) predicted by SPINE-X [16] and fractions of exposed residues predicted by Real-SPINE 3 [17] for long ordered, semi-disordered, and fully disordered regions (>30 residues) of the proteins in the DX4080 set. Here, we employed predicted secondary structures and solvent accessibility for all proteins because not all proteins or regions have structures to calculate secondary structure or solvent accessibility. Figure 2 shows that ordered regions occupy the upper left corner with low fraction of exposed residues and high content of secondary structures while fully disordered regions mostly locate at the bottom right corner (highly exposed with little secondary structures). The semi-disordered regions are located somewhat in between. That is, it is semi-collapsed with some secondary structures. Thus, a semi-disordered state correctly captures protein regions that are semi-collapsed or semi-structured, based on current state-of-the-art techniques for predicting secondary structure and solvent accessibility.

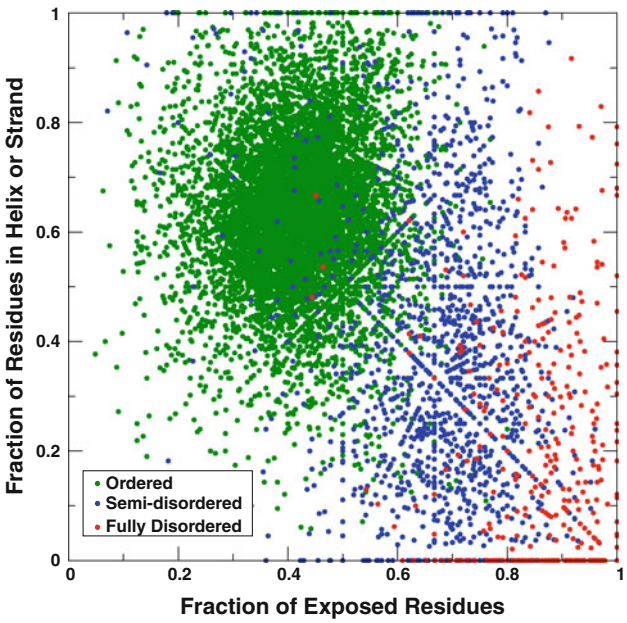

Fig. 2 Ordered (*green*), semi-disordered (*blue*), and fully disordered (*red*) regions in term of fraction of exposed residues (*x-axis*) and fraction of residues with secondary structures (*y-axis*) based on SPINE-D results of the DX4080 dataset. A residue is defined as exposed if its predicted solvent accessibility is greater than 25 %. Secondary structures and solvent accessibility are predicted by SPINE-X and Real-SPINE 3, respectively (Color figure online)

The Semi-disordered State in Disordered Proteins

In order to have a better understanding of the above-defined semi-disorder, it is necessary to investigate the occurrence of the semi-disordered state in disordered proteins at the individual protein level. Here, we defined a *wholly disordered protein* as a protein without any predicted ordered residues (i.e., only semi-disordered and fully disordered residues). For convenience, we denote f_o, f_{sd}, and f_{fd} as the fraction of ordered residues, the fraction of semi-disordered residues, and the fraction of fully disordered residues for a given protein, respectively. $f_o + f_{sd} + f_{fd} = 1$. For a predicted disordered protein, $f_o = 0$ and $f_{sd} + f_{fd} = 1$. Here, we will analyze wholly disordered proteins in all three datasets mentioned above. Fig. 3a shows a Gibbs-triangle diagram where each protein is a point and the position of the protein is determined by f_o, f_{sd}, and f_{fd}. All predicted disordered proteins ($f_o = 0$ and $f_{sd} + f_{fd} = 1$) are located on the right edge of the triangle that mixes semi-disordered and fully disordered residues in Fig. 3a.

Wholly Disordered Proteins in the Monomer Control Set

Most proteins in the monomer control set (in green) locate near the line that mixes ordered and semi-disordered residues and the majority of proteins in the control set

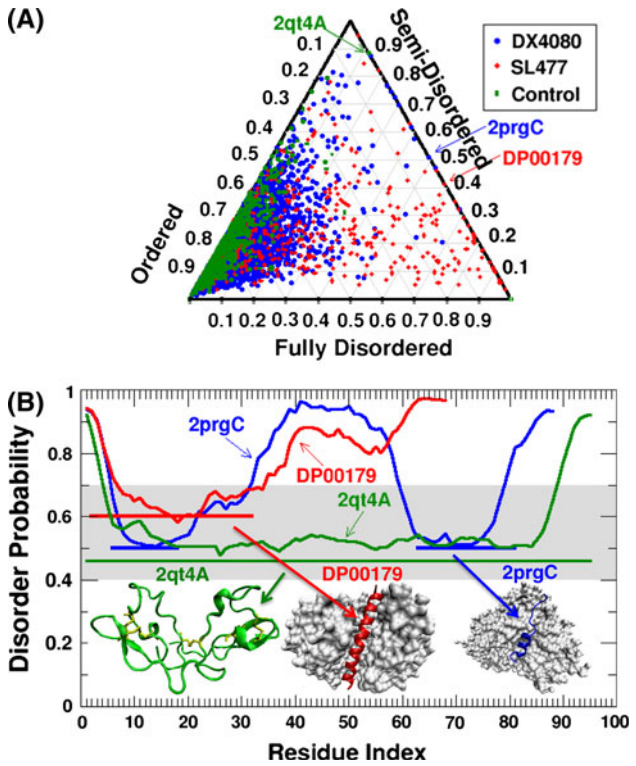

Fig. 3 (**a**) The Gibbs triangle diagram of the fractions of residues in three states (ordered, semi-disordered, fully disordered residues) for all proteins in the three datasets as labeled. Each protein is a point and its position is determined by three fractions of residues. (**b**) Disorder probability profiles with zero ordered residues ($f_o = 0$) for the chain A of the PDB ID 2qt4 (2qt4A) in the control set, for DP00179 (chain B in PDB ID 1DPJ) in SL477, and for chain C of PDB ID 2prg (2prgC) in DX4080. The semi-disordered regions correspond to structured regions (*horizontal lines*) stabilized by disulfide bonds (2qt4A), by binding-induced folding (DP00179 and 2prgC). Only one structured region of 2prgC bound with its target is visible in this figure. The *gray area* indicates the region defined as semi-disordered

(674/703 = 96 %) are predicted to have more than 50 % ordered residues. Such dominance of ordered residues over semi- or fully disordered residues further validates the two-state accuracy of SPINE-D in distinguishing ordered from disordered residues. There are only two proteins with $f_o = 0$. One (PDB ID 2pne) is a snow flea antifreeze protein (sfAFP) predicted with $f_{fd} = 1$ and $f_o = 0$. The protein unfolds at room temperature [18]. Its X-ray determined structure is stabilized by two disulfide bonds and solved only in the presence of the mirror image form of sfAFP [19]. The second one is the antiviral lectin scytovirin (PDB ID: 2qt4, $f_o = 0$, $f_{sd} = 0.88$, $f_{fd} = 0.12$). As shown in Fig. 3b, this protein is made of a long semi-disordered region (except near the terminals) and is stabilized by five disulfide bonds with little secondary structures (12 % in short helices and 12 % short beta sheets) [20]. Thus, the instability or marginal stability of these two proteins is correctly predicted by SPINE-D: a fully disordered state for sfAFP that has no stable structure at room temperature (2pne) and a semi-disordered state for antiviral lectin scytovirin that is stabilized by five disulfide bonds (2qt4A, Fig. 3b). The existence of semi-disordered regions (also fully disordered regions, to a much lesser extent) in some stably folded monomeric proteins suggests that they can participate in folding into unique structure in the presence of sequence regions encoded for structures.

Wholly Disordered Proteins in the SL477 Set

In SL477, there are a total of 30 proteins predicted with $f_o = 0$. All but one are annotated as entirely disordered proteins (without any ordered residues) by experimental means [15]. Thus, there is excellent agreement between predicted and annotated disordered proteins with $f_o = 0$. The only protein (DP00179) annotated with an ordered region has about half of the residues annotated as ordered and about half annotated as disordered. As shown in Fig. 3b, the annotated ordered region of DP00179 (yeast protein IA3) is predicted as semi-disordered and has a single helical structure stabilized by its inhibiting target aspartic proteinase A [21]. That is, predicted semi-disordered region has an exact match to the induced folding region meaningfully separated from the region that is fully disordered.

Wholly Disordered Proteins in the DX4080 Set

In DX4080, there are nine proteins with $f_o = 0$. For eight proteins (pdb ID:1meyG, 1ohhH, 1qqp4, 1urqA, 2pxbA, 2prgC, 3f5hB, and 3k29A), their structured regions all contain long semi-disordered regions. One example (2prgC) is shown in Fig. 3b, and the rest are shown in Fig. 4. Only one protein, called vasopressin V1a receptor (PDB ID: 1ytvN), contains a short *fully* disordered region at the N-terminal that is folded into a turn after it binds to maltose-binding periplasmic protein. Because SPINE-D was trained to predict disorder at terminal regions, we removed such effect (dashed line) by employing the sequence that is made of three vasopressin V1a receptor sequences and taking the result from the center sequence. The terminal fully disordered region becomes semi-disorder. Thus, all structured regions are semi-disordered. These nine proteins result from induced folding due to the presence of co-factors such as proteins, DNA, or ligands. That is, induced folding occurs at predicted semi-disordered regions for these proteins. This result further confirms the accuracy of $f_o = 0$ from SPINE-D predictions because all these proteins should have been annotated as disordered proteins (semi-disordered + fully disordered) in an isolated monomeric form. More importantly, the connection between induced folding and semi-disordered regions is consistent with what was found for two proteins in SL477.

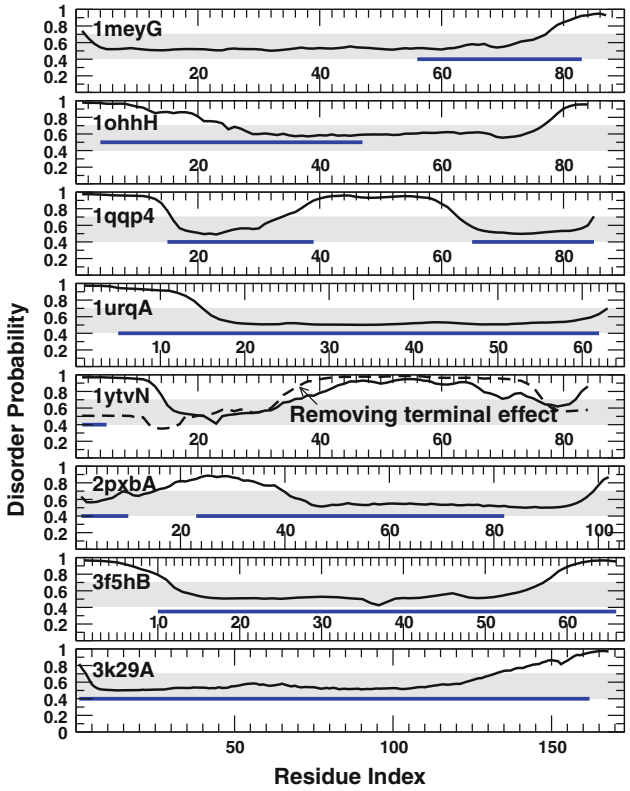

Fig. 4 Structured regions (*blue bar*) by induced folding of disordered proteins are compared with their semi-disordered regions (probability profile within the *gray region*) in eight additional proteins with predicted $f_o = 0$ in the DX4080 dataset (PDB IDs as labeled). Only one structured region (1ytvN) corresponds to a fully disordered region at the N-terminal end of chain N of 1ytv. But it is semi-disordered after removing the terminal effect (*dashed line*). The N-terminal region of chain G of 1mey (consensus zinc finger) does not have coordinates but the same region in identical chains C and F does. Thus, the whole chain G made of mostly the semi-disordered state can be labeled as structured from residue 1 to 85 after binding with DNA in a trimeric form (Color figure online)

Quantifying the Link Between Semi-disorder and Induced Folding

The above result was based on a limited number of examples. To quantify the relation between semi-disorder and induced folding, we employ the ANCHOR dataset [22] that is a collection of binding regions in disordered proteins that fold upon binding. The dataset was divided into long (28 complexes) and short (46 complexes) according to the size of disordered regions (30 residues). In this dataset, each residue is annotated as either in binding (positive) or non-binding (negative) regions. To examine if a residue in a semi-disordered state is a potential binding residue, we define true positive (TP) if an annotated binding residue is predicted as semi-disorder, true negative (TN) if a non-binding residues is predicted as non-semi-disorder, false positive (FP) if a non-binding residue is predicted as semi-disorder, and false negative (FN) if a binding residue is predicted as non-semi-disorder. This allows us to calculate sensitivity [TP/(TP + FN)], specificity [TN/(TN + FP)], and Matthews correlation coefficient $[\mathrm{MCC} = (\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}) / \sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}]$ without any training. Here, we assess the performance on the residue level, rather than on the region level to avoid the difficulty of defining true/false negatives/positives at the region level without introducing additional parameters.

Table 1 compares the results of SPINE-D with those from ANCHOR [22] and MoRFpred [23], two recently developed techniques that were trained to predict binding in disordered regions. The accuracy of all three methods is low with the average sensitivity and specificity (balanced accuracy) between 56 and 72 %. MoRFpred, trained for the short dataset, has the highest MCC value of 0.29 for the short dataset while SPINE-D has the highest MCC value of 0.15 for the long dataset. The result confirms a weak but positive association between a semi-disordered state and the binding-induced folding region, for binding residues in long disordered regions, in particular.

Semi-disorder and Protein Aggregation: Illustrative Examples

The connection between semi-disorder and binding-induced folding also suggests the potential role of semi-disorder in protein aggregation because protein aggregation can be viewed as "folding" coupled with self-association. Here, we started with several known aggregation-prone proteins to examine if there is a connection between semi-disorder and aggregation.

Table 1 Predicting binding residues in short and long disordered regions (the short and long ANCHOR set) by ANCHOR, MoRFpred, and the semi-disorder from SPINE-D

Method	Short disordered region			Long disordered region		
	Sensitivity	Specificity	MCC	Sensitivity	Specificity	MCC
ANCHOR	0.64	0.71	0.14	0.45	0.66	0.06
MoRFpred[a]	0.50	0.94	0.29	0.17	0.94	0.10
SPINE-D	0.32	0.80	0.05	0.42	0.81	0.15

[a] MoRFpred failed to make predictions for 2 proteins in the short set

Huntingtin

One example of protein aggregation involves the protein huntingtin that contains a region with repeated glutamines (Qs). Individuals with 37 or more glutamines in their huntingtin protein are likely to develop Huntington's disease during their lifetime, and the severity of the disease is monotonically related to the number of glutamines [24]. Figure 5 shows that as the number of glutamines increases roughly beyond 20, there is a significant increase in fraction of glutamines in the semi-disordered state along with a large reduction in the average disorder probability for the glutamines. That is, the poly-Q region experiences a transition from a fully disordered state (0–24 glutamines) to >30 % semi-disordered (35–100 glutamines), with a monotonic increase in fraction of Qs in the semi-disordered state.

Alpha-Synuclein

Alpha-Synuclein, a classical example of IDPs, was recently found to have a tetrameric structure for the first 100 residues in physiological conditions [25, 26]. This induced folding and/or aggregation-prone region corresponds to a semi-disordered region as shown in Fig. 6a. The separation of two domains around residue 100 is consistent with compaction ratios obtained from combined NMR experiments and replica exchange molecular dynamics simulations [27] as well as the partial condensation in the central region (30–100) from molecular dynamics simulation with restraints from spin-label NMR experiments [28]. A compaction ratio was defined as the average end-to-end distance relative to the end-to-end distance calculated from random coil ensembles. The medium compaction ratio of about 0.5 for the N-terminal and NAC regions indicates that they are semi-collapsed and the high compaction ratio of about 0.8 for the C-terminal of alpha-synuclein suggests that it is random-coil-like and accessible. The accessible C-terminal is also consistent with the fact that the region is not directly involved in the mechanism of aggregation and accessible to single-domain camelid antibody [29], and its truncation promotes aggregation [30]. That is, the amyloidogenic and induced-folding region of alpha-synuclein is semi-disordered.

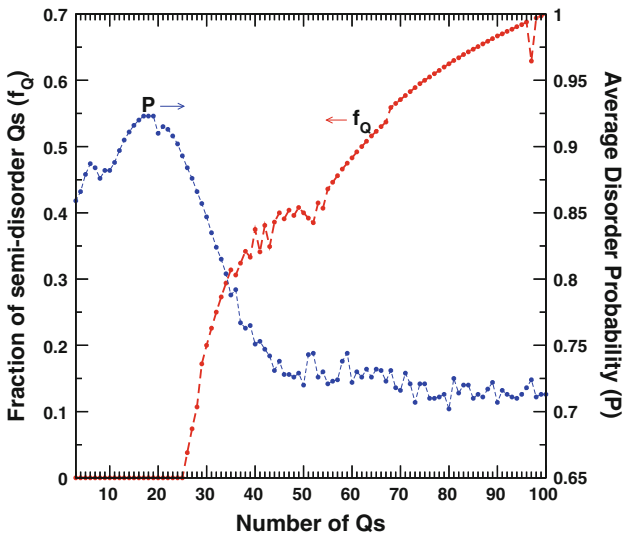

Fig. 5 Transition of the polyglutamine tract of huntingtin from a fully disordered to a partially semi-disordered state. Fraction of glutamines (Qs) in a semi-disordered state (f_Q in *red*) and the average disorder probability (*P*, in *blue*) in the poly-Q region as a function of the number of glutamines in the poly-Q tract of huntingtin (Color figure online)

Yeast Sup35

The overlap between amyloidogenic and semi-disordered regions in alpha-synuclein is further observed for amyloidogenic yeast Sup35. In Fig. 6b, the disorder probability profile for Sup35 predicted by SPINE-D is compared with

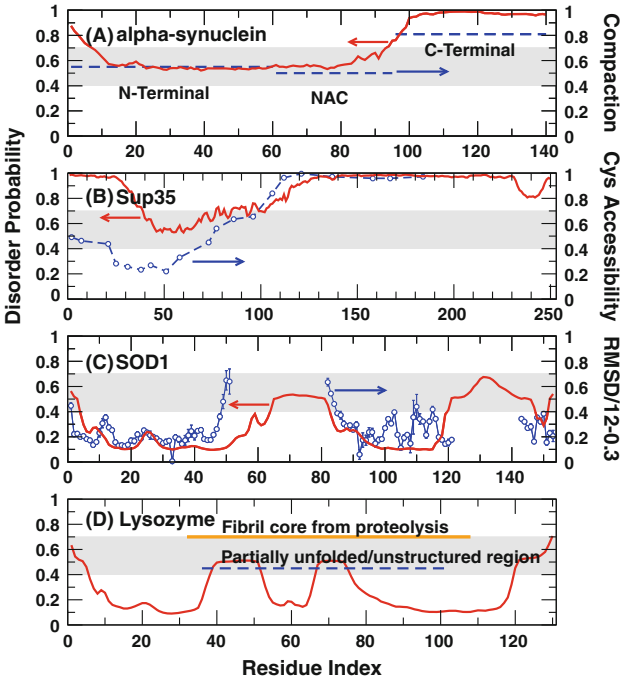

Fig. 6 Semi-disordered state in unstructured (alpha-synuclein and Sup35) and structured proteins (SOD1 and human lysozyme). Predicted disordered probability profiles (*P* in *red*) compared with compaction ratios for three different regions at normal pH from combined NMR experiments and replica exchange molecular dynamics simulations of alpha-synuclein (in *blue*) (**a**), the measured Cys accessibility profile (scaled by the largest accessibility of 82.2 %, in *blue*) of yeast Sup 35 (**b**), root mean squared distance (RMSD) from native by molecular dynamics simulations of SOD1 (**c**), and the unstructured regions in a partially unfolded state detected by H/D exchange (*blue*) and the fibril core region from proteolysis (*orange*) (**d**). In (**c**), open regions in *blue line* correspond locally unfolded regions of SOD1. RMSD values are rescaled and shifted to facilitate comparison. The *gray bar* indicates the region defined as semi-disordered in disorder probability ($0.4 \leq P \leq 0.7$) (Color figure online)

the measured Cys accessibility profile of amyloid fibrils at different substitution position in Sup35 [31]. The Cys accessibility profile indicates that amyloid fibrils are made of the N-terminal domain while the C-terminal domain remains fully accessible. The amyloidogenic N-terminal and accessible C-terminal domains of Sup35 match nicely to the semi- and fully disordered regions identified by SPINE-D.

Cu, Zn Superoxide Dismutase

The above results are for IDPs with predicted disorder probabilities >0.5 for all residues. Does a semi-disordered state play a role for aggregation of structured proteins?. In Fig. 6c, we applied SPINE-D to Cu, Zn superoxide dismutase (SOD1). ApoSOD1 has a well-defined crystal structure but has locally unfolded regions in solution based on experiments [32, 33] and simulations [34]. Such locally unfolded regions from molecular dynamics simulations [34] are in excellent agreement with semi-disordered regions predicted from SPINE-D as shown in Fig. 6c. Because stable, ordered regions are found in the fibrillar core of wild-type SOD1 [35], its semi-disordered regions play the key role for opening up the hydrophobic core for aggregation [32, 33, 35].

Human Lysozyme

In Fig. 6d, the disordered probability profile is shown for another structured protein: human lysozyme. Its semi-disordered regions (residues 39–52 and 67–75) are within the unstructured region of a partially unfolded state detected by H/D exchange (residues 36–102) [36] and the fibril core region according to proteolysis (residues 32–108) [37].

Acylphosphate

As a control, we also examined the disorder probability profile of acylphosphate from hyperthermophilic archaeon *Sulfolobus solfataricus* (Sso AcP). This stable protein does not have detectable aggregation except in the presence of a mild destabilizing co-solvent such as 20 % trifluoroethanol [38]. As shown in Fig. 7, this protein does not have any semi- or fully disordered residues except in the terminal regions. The unstructured region for the first 12 residues from the NMR experiment [39], in close agreement with the mix of semi- and fully disordered residues from 1 to 15 at the N-terminal from SPINE-D (or residues 1–12 after removing terminal effect), was shown to play the key role in promoting aggregation from protein engineering experiments [40]. Thus, the semi-disordered state promotes aggregation even for highly stable proteins that do not aggregate under normal physiological conditions.

Semi-disorder and Protein Aggregation: Quantification

To quantify the relation between aggregation and semi-disorder beyond above examples, we employed the AmyPDB dataset [41]. The AmyPDB dataset contains 31 amyloid families, including 25 amyloid precursors and 6 prions [41]. Among them, 12 proteins have annotated amyloidogenic regions: yeast prion protein (URE2), podospora small s protein, human amyloid beta A4 protein (A4), atrial natriuretic factor (ANF), apolipoprotein A-1 (APOA1), beta 2 microglobulin (B2MG), islet amyloid polypeptide (IAPP), integral membrane protein 2B (ITM2B), lactadherin (MFGM), major prion protein precursor (PrP), serum amyloid A (SAA), and tau protein. This dataset of 12 proteins was enlarged with four additional proteins shown in Fig. 6. We define true positive if a residue annotated in aggregation regions is predicted as semi-disorder, true negative if a residue not in aggregation regions is predicted as non-semi-disorder, false positive if a residue not in aggregation regions is predicted as semi-disorder, and false negative if a residue in aggregation region is predicted as non-semi-disorder. This allows us to calculate sensitivity, specificity, and MCC values as in the case of binding prediction.

Table 2 compares the accuracy of SPINE-D with three methods dedicated to predict protein aggregation. The three methods are Fold-amyloid [42] based on expected

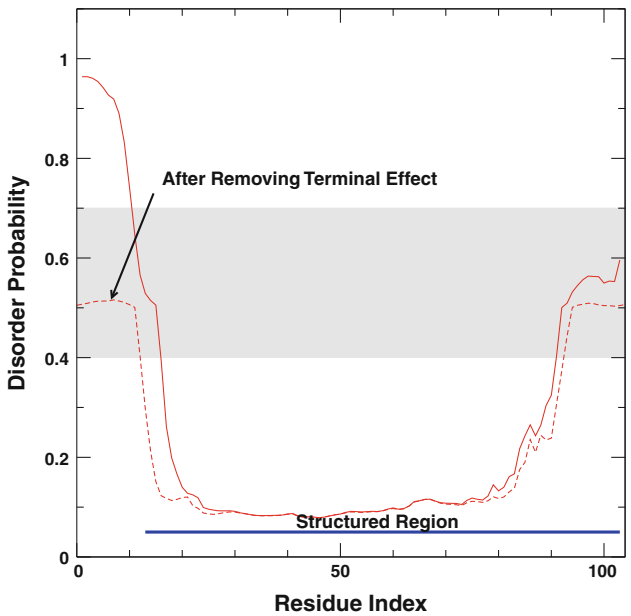

Fig. 7 The disorder probability profile of acylphosphate from hyperthermophilic archaeon *Sulfolobus solfataricus* (Sso AcP) predicted from SPINE-D (*red line*). The semi-disordered residues from 1 to 12 at the N-terminal after removing terminal effect agree with the unstructured region for the first 12 residues from the NMR experiment (PDB #1Y9O) (*blue*) (Color figure online)

probability of hydrogen bonds formation and expected packing density of residues, Waltz [43] based on the sequence diversity of amyloid hexa-peptides, and Aggrescan [44] based on aggregation-propensity scale. The accuracy of three methods (Fold-Amyloid, Aggrescan and Waltz-Best performance) is poor with the average sensitivity and specificity (balanced accuracy) around 50 % and the MCC value between −0.04 and 0.01 for this dataset. Only Waltz-high sensitivity and SPINE-D have some ability to predict aggregation regions (MCC = 0.18 and 0.11, respectively). This highlights the challenge of predicting aggregation. The MCC value given by SPINE-D can be improved from 0.11 to 0.15 if the definition of aggregation-prone residues covers both semi-disorder and full-order (0–0.7). This suggests the importance of both ordered and semi-disordered regions in protein aggregation. We note that many methods for predicting protein aggregations are built on the dataset of aggregation-prone and non-aggregation peptides (for example, [45]). Such a dataset is not useful for examining the relation between semi-disorder and aggregation because SPINE-D is only applicable to protein sequences.

Semi-disorder and Residue Aggregation Propensity

The role of semi-disorder in protein aggregation, however, seems to contradict observed anti-correlation between disorder propensity and amyloid aggregation propensity of 20 amino-acid residue types [46, 47]. To explain this observation, we applied SPINE-D to 4080 non-redundant high-resolution X-ray structures (DX4080) and obtained the compositions of the 20 amino acid residues that are ordered ($0 \le P < 0.4$), C_r^{o}, semi-disordered ($0.4 \le P \le 0.7$), C_r^{sd}, or fully disordered ($0.7 < P \le 1$), C_r^{fd} ($r = 1,\ldots 20$) to compare with residue amyloid aggregation propensity from empirical fit to experimental aggregation rates of unstructured polypeptide chains [46]. We confirmed the anti-correlation between the propensities for full disorder $(C_r^{\mathrm{fd}} - C_r^{\mathrm{o}})/C_r^{\mathrm{o}}$ and the propensity for amyloid aggregation with a correlation coefficient of −0.77. However, the amino acid residues gained in changing from the fully disordered to the semi-disordered state $(C_r^{\mathrm{sd}} - C_r^{\mathrm{fd}})/C_r^{\mathrm{fd}}$ is highly correlated with amyloid aggregation propensity. As shown in Fig. 8, the correlation coefficient is 0.86 without Pro and four charged residues (Arg, Asp, Glu, and Lys) and 0.74 for all residues. The highest enrichment of a residue in a semi-disordered region over the fully disordered region is 185 % for the strongest aggregation-prone residue Trp and more than 100 % for the second and third strongest aggregation-prone residues Phe and Cys. This strong positive correlation supports the capability of semi-disordered regions to promote aggregation. Changing from the semi-disordered state to the ordered one continues to enrich residues with high amyloid aggregation propensity but with a much smaller enrichment factor (36 % for Trp, 52 % for Phe, and 41 % for Cys). The correlation coefficient is 0.79 for all 20 residue types and 0.87 without Pro and charged residues. Thus, only the fully disordered state is aggregation-resistant. Both ordered and semi-disordered regions can participate in aggregation as demonstrated in Figs. 6 and 7.

Discussion

The disorder probability predicted by SPINE-D was rescaled for CASP 9 so that the threshold for disorder is at 50 % being disordered or ordered. Although the simple linear scaling was somewhat arbitrary, the resulting population of semi-disordered residues appears to be physically meaningful. This is reflected from the fact that these semi-disordered residues can be characterized as semi-collapsed (according to predicted solvent accessible surface area) and semi-structured (according to predicted secondary structure content). Furthermore, the semi-disordered regions made of semi-disordered residues are found capable of induced folding and protein aggregation.

This article established a quantitative connection between semi-disorder and induced folding. Previously, the observed connection between induced folding and a dip in disorder probability [48, 49] has motivated development of neural network-based alpha-MoRF predictors [50] and SVM-based MoRF-predictor [23] with predicted disorder

Table 2 Predicting residues in aggregation regions for 12 proteins in the AmyPDB dataset and 4 proteins from Fig. 6

Method	Sensitivity	Specificity	MCC
Fold-amyloid	0.16	0.81	−0.03
Aggrescan	0.23	0.76	−0.01
Waltz-best perform (high sensitivity)[a]	0.14 (0.54)	0.93 (0.66)	0.01 (0.18)
SPINE-D (order+semi-disorder)[b]	0.38 (0.84)	0.73 (0.32)	0.11 (0.15)

[a] Two options in Waltz server were used: best performance and high sensitivity (in parentheses)

[b] Results from SPINE-D are obtained by employing predicted semi-disordered residues in aggregation regions. The numbers in parentheses are resulted from assigning both ordered and semi-disordered residues (0–0.7) as aggregation prone

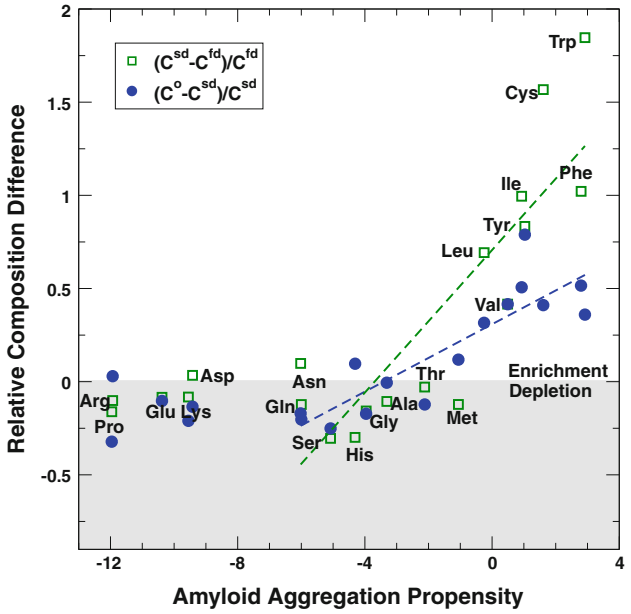

Fig. 8 Strong positive correlation between amyloid aggregation propensity at pH 7 and relative difference in compositions of amino acid residue types between semi-disordered and fully disordered regions $[(C_r^{sd} - C_r^{fd})/C_r^{fd}$, *green squares*] or between ordered and semi-disordered regions $[(C_r^{o} - C_r^{sd})/C_r^{sd}$, *blue circles*] generated from the DX4080 dataset. C_r^{o},C_r^{sd}, and $C_r^{fd}$are compositions of amino acid residues for ordered, semi-disordered, and fully disordered states, respectively. *Above* and *below zero* of $[(C_r^{sd} - C_r^{fd})/C_r^{fd}]$ or $[(C_r^{o} - C_r^{sd})/C_r^{sd}]$ indicates enrichment or depletion relative to fully disordered regions or semi-disordered regions, respectively

as input (trained on short disorder-to-ordered transitions). ANCHOR, on the other hand, predicts binding residues in disordered regions by predicting the inter-protein interaction strength based on the average composition of amino acid residues in globular proteins [22]. This study provides an alternative approach to characterize induced folding in the absence of specific training (Table 1).

The connection between semi-disorder and induced folding, however, is more complicated than simply assigning semi-disordered regions as induced folding, the assumption made in Table 1. It is complicated because a semi-disordered region may be folded by interacting with itself or other molecules (induced folding), but induced folding regions do not have to be semi-disordered. They can be made of ordered residues that are too few to stabilize a solid-like structure by themselves [51] or consist of fully disordered residues that fold in the presence of a perfectly matching partner. This explains the low accuracy in direct assignment of semi-disorder as induced folding shown in Table 1. Such low accuracy is also observed in other techniques, indicating room for further improvement by more specific training with SPINE-D output as input.

The ability of semi-disordered regions to aggregate is confirmed by enrichment of aggregation-prone residues in semi-disordered regions, relative to that in fully disordered regions. It is also evidenced by the overlap between known amyloidogenic and semi-disordered regions for 18 proteins studied here. Recently, Sikirzhytski et al. [52] showed that a de novo designed fibrillogenic polypeptide YE8 is made of a largely semi-disordered region from the SPINE-D prediction. Thus, for IDPs, it is the absence or existence of semi-disorder that leads to some IDPs being resistant to protein aggregation while others being aggregation-prone [1]. For structured proteins, aggregation can occur at either semi-disordered or ordered regions, or both. This is because both regions are enriched with amino-acid residues with high propensity for aggregation as shown in Fig. 8. Semi-disordered regions in structured proteins, however, are induced to fold by other structure-encoded regions. Thus, they are likely the weakly stable part of protein structures. Such instability is confirmed by the overlap between the semi-disordered regions and locally unfolded regions in SOD1 (Fig. 6c), human lysozyme (Fig. 6d), and Sso AcP (Fig. 7). This instability of semi-disorder can initiate aggregation in structured proteins by local unfolding [53] (or as meta-stable states/regions [54–56]) and exposes self-complementary amyloidogenic segments protected by evolution [57].

The ability of using semi-disorder *alone* to predict aggregation, however, is weak as shown in Table 2. This reflects the complex interplay between inter and intraprotein interactions. Not all predicted semi-disordered regions are amyloidogenic. For example, the APOA1 protein has two long semi-disordered regions (Residues 25–107 and 153–226). This protein is a six-helix bundle in which helices 1 and 2 (25–107, the amyloidogenic region) are slightly more accessible than helix 4 (153–226). The former has a residue solvent accessibility (RSA) of 0.57 for 57 exposed residues (RSA $>$ 0.25) compared with 0.52 in the second region with the same number of exposed residues. Non-amyloidogenic semi-disordered regions may also exist simply because the method was not trained to predict amyloid formation. Our sequence-based prediction relies mostly on local sequence interactions. Nonlocal interactions (interactions between residues that are not sequence neighbors) determine the winner of the competition between intramolecular (folding or misfolding) and inter-molecular interactions (aggregation). Incorporation of both inter and intra molecular interactions and combining the detection of the semi-disordered state with the models based on physicochemical properties, neural networks, and structural profiles [57–61] will likely lead to further improvement in accuracy of predicting amyloidogenic regions.

One interesting question is the relationship between predicted semi-disorder/disorder with energetically frustrated regions in proteins. Ferreiro et al. [62] found that some proteins contain highly frustrated interactions near binding sites that are less frustrated upon complex

formation. Although this local frustration index [63] is limited to proteins with known structures and yet to be applied to induced-folding proteins, it is likely that induced folding corresponds to the transition from frustrated (unable to fold) to minimally frustrated (foldable) interactions. Interestingly, local frustrated regions correspond to flexible regions that are described by temperature B-factor and simulated results of root-mean-squared fluctuation [64]. Similar result is obtained in Fig. 6 except that semi-disorder corresponds to locally unfolded regions where root-mean-squared fluctuation is significantly larger than what typically observed in structured proteins. That is, semi-disorder and full-disorder likely have strongly frustrated interactions. The quantitative relation between predicted disordered probability and flexibility can be examined by correlating disorder probabilities with temperature B-factors from X-ray structure determination. For a dataset of high-resolution and non-redundant 766 protein structures collected by Yuan et al. [65], we found that the average correlation coefficient for these 766 proteins given by SPINE-D is 0.39 ± 0.19. Thus, there is a positive relationship between protein disorder and structural flexibility, despite that SPINE-D was not trained for temperature B-factor prediction.

This study highlights the ability of SPINE-D in separating semi-disorder from ordered and fully disordered states. It would be of interest to know if other methods have similar capability. We selected six representative methods that cover three categories of disorder prediction methods, including: methods that only use amino acid propensity/energy associated with disorder, e.g., IUPred short/long disorder predictor [66]; method based on machine learning approaches, e.g., Dispro [67], Disopred2 [68] and meta servers that combine multiple disorder predictors, e.g., MD [69] and MFDp [70]. The distributions of predicted disorder probabilities for the SL477 dataset are shown in Fig. 9a. All have two state distributions. It is clear that SPINE-D is unique because its training on an unbalanced dataset requires rescaling the disorder probability. As an illustrative example, we apply these five techniques (IU-short and IU-long have similar results, only IU-long is shown) to Sup 35. As Fig. 9b shows, all these methods do not have a clear separation into two domains at residue 100, unlike SPINE-D predictions and experimental Cys accessibility.

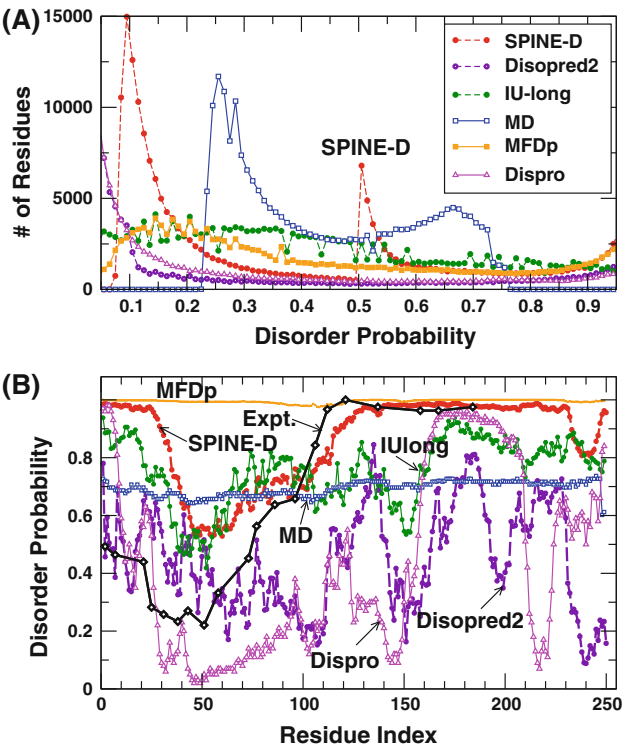

Fig. 9 (**a**) Distributions of predicted disorder probabilities for five online servers (Disopred2, IU-long, MD, MFDp, and Dispro) in addition to SPINE-D as labeled. (**b**) Disorder probability of yeast Sup 35 predicted by the above five methods in addition to SPINE-D as labeled. No methods except SPINE-D (in *red*) separated a collapsed N-terminal region and an extended C-terminal region in agreement with the experimental Cys accessibility data (*in black*) (Color figure online)

Materials and Methods

Datasets

In addition to DX4080 [non-redundant, high-resolution (<2 Å) X-ray structures, 25 % sequence identity or less between each other], we employed the SL dataset of 477 non-redundant proteins (25 % sequence identity cutoff) that was built by re-annotating manually annotated disordered proteins in the Disprot database so that it includes reliable disorder and order contents [15]. This dataset contains fully disordered proteins based on various experimental methods. The sequences in SL477 are 25 % sequence identity or less from the sequences in DX4080. As a control, we built a set of stably folded monomeric proteins by searching the PDB based on the following criteria: (a) X-ray determined structures without DNA, RNA, hybrid or other ligands; (b) having only one chain (both biological assembly and asymmetric unit); (c) high resolution (≤3.0 Å) with size ≥50 residues; and d) no missing residues (except terminal regions) or abnormal amino acid types. A total of 703 proteins are obtained after removing redundant chains at 30 % sequence identity.

SPINE-D Server

SPINE-D is a neural-network-based predictor trained on a non-redundant set of 4157 X-ray structures and 72 fully disordered proteins from the Disprot database v5.0 [14]. It

only requires an input of protein sequence is available at http://sparks-lab.org. For huntingtin, the calculation was started with three Qs and the sequence profile of the middle Q is employed to expand the poly-Q tract. More methodology details can be found in Ref. [12].

Amino-Acid Composition Calculations

Application of SPINE-D to DX4080 leads to residues in ordered ($P < 0.4$), semi-disordered ($0.4 \leq P \leq 0.7$), and fully disordered ($P > 0.7$) sets. The fractions of each residue type in these three states (amino-acid compositions) are obtained as $C_r^{o}, C_r^{sd}, C_r^{fd}$ ($r = 1,\ldots,20$), respectively. Relative composition differences between semi-disorder and full-disorder $[(C_r^{sd} - C_r^{fd})/C_r^{fd}]$ and between order and semi-disorder $[(C_r^{o} - C_r^{sd})/C_r^{sd}]$ are compared with experimentally measured aggregation propensity. We would like to emphasize that all analyses are not from annotated disorder/ordered regions, secondary structure, or ASA, but are based on predicted disorder probabilities, predicted secondary structure, and predicted solvent-accessible surface area because secondary structure, solvent accessibility, and semi-disorder annotation are unknown for unstructured regions.

Other Methods

We have used five representative on-line servers for generating disorder predictions: Dispro from http://www.ics.uci.edu/~baldig/dispro.html; DISOPRED2 from http://bioinf.cs.ucl.ac.uk/disopred/; MD from http://www.predictprotein.org/; IUpred Long/short from http://iupred.enzim.hu/; and MFDp from http://biomine-ws.ece.ualberta.ca/MFDp.html.

Acknowledgments Helpful discussions with Keith Dunker, Lukasz Kurgan, Quyen Hoang and Yuedong Yang are gratefully acknowledged. We also would like to thank Feng Ding for many helpful comments and providing us his simulation data for comparison. Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R01GM085003. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends in Biochemical Sciences, 27*, 527–533.
2. Rauscher, S., & Pomes, R. (2010). Molecular simulations of protein disorder. *Biochemistry and Cell Biology, 88*, 269–290.
3. Dunker, A. K., et al. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling, 19*, 26–59.
4. Uversky, V. N. (2002). Natively unfolded proteins: A point where biology waits for physics. *Protein Science, 11*, 739–756.
5. Dunker, A. K., & Obradovic, Z. (2001). The protein trinity - linking function and disorder. *Nature Biotechnology, 19*, 805–806.
6. Uversky, V. N. (2002). Natively unfolded proteins: A point where biology waits for physics. *Protein Science, 11*, 739–756.
7. Zhou, Y., & Karplus, M. (1997). Folding thermodynamics of a model three-helix-bundle protein. *Proceedings of National Academy of Science United States of America, 94*, 14429–14432.
8. Tompa, P., & Fuxreiter, M. (2008). Fuzzy complexes: Polymorphism and structural disorder in protein–protein interactions. *Trends in Biochemical Sciences, 33*, 2–8.
9. Dunker, A. K., Silman, I., Uversky, V. N., & Sussman, J. L. (2008). Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology, 18*, 756–764.
10. Vucetic, S., Brown, C. J., Dunker, A. K., & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins, 52*, 573–584.
11. He, B., et al. (2009). Predicting intrinsic disorder in proteins: An overview. *Cell Research, 19*, 929–949.
12. Zhang, T., et al. (2012). SPINE-D: Accurate prediction of short and long disordered regions by a single neural-network based method. *Journal of Biomolecular Structure and Dynamics, 28*, 799–813.
13. Monastyrskyy, B., Fidelis, K., Moult, J., Tramontano, A., & Kryshtafovych, A. (2011). Evaluation of disorder predictions in CASP9. *Proteins, 79*(S10), 107–118.
14. Sickmeier, M., et al. (2007). DisProt: The database of disordered proteins. *Nucleic Acids Research, 35*, D786–793.
15. Sirota, F. L., et al. (2010). Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics, 11*(Suppl 1), S15.
16. Faraggi, E., Zhang, T., Yang, Y., Kurgan, L., & Zhou, Y. (2011). SPINE X: Improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *Journal of Computational Chemistry, 33*, 259–263.
17. Faraggi, E., Xue, B., & Zhou, Y. (2009). Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins, 74*, 847–856.
18. Lin, F. H., Graham, L. A., Campbell, R. L., & Davies, P. L. (2007). Structural modeling of snow flea antifreeze protein. *Biophysical Journal, 92*, 1717–1723.
19. Pentelute, B. L., et al. (2008). X-ray structure of snow flea antifreeze protein determined by racemic crystallization of synthetic protein enantiomers. *Journal of the American Chemical Society, 130*, 9695–9701.
20. Moulaei, T., et al. (2007). Atomic-resolution crystal structure of the antiviral lectin scytovirin. *Protein Science, 16*, 2756–2760.
21. Li, M., et al. (2000). The aspartic proteinase from *S. cerevisiae* folds its own inhibitor into a helix. *Natural Structural Biology, 7*, 113–117.
22. Meszaros, B., Simon, I., & Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. *PLOS Computational Biology, 5*, e1000376.
23. Disfani, F. M., et al. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics, 28*, i75–i83.
24. Walker, F. O. (2007). Huntington's disease. *Lancet, 369*, 218–228.
25. Bartels, T., Choi, J. G., & Selkoe, D. J. (2011). Alpha-Synuclein occurs physiologically as a helically folded tetramer that resists aggregation. *Nature, 477*, 107–110.

26. Wang, W., et al. (2011). A soluble alpha-synuclein construct forms a dynamic tetramer. *Proceedings of National Academy of Science United States of America, 108*, 17797–17802.
27. WU, K. P., Weinstock, D. S., Narayanan, C., Levy, R. M., & Baum, J. (2009). Structural reorganization of alpha-synuclein at low pH observed by NMR and REMD simulations. *Journal of Molecular Biology, 391*, 784–796.
28. Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M., & Dobson, C. M. (2005). Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society, 127*, 476–477.
29. De Genst, E. J., et al. (2010). Structure and properties of a complex of alpha-synuclein and a single-domain camelid antibody. *Journal of Molecular Biology, 402*, 326–343.
30. Li, W. X., et al. (2005). Aggregation promoting C-terminal truncation of alpha-synuclein is a normal cellular process and is enhanced by the familial Parkinson's disease-linked mutations. *Proceedings of National Academy of Science United States of America, 102*, 2162–2167.
31. Krishnan, R., & Lindquist, S. L. (2005). Structural insights into a yeast prion illuminate nucleation and strain diversity. *Nature, 435*, 765–772.
32. Shaw, B. F., et al. (2006). Local unfolding in a destabilized, pathogenic variant of superoxide dismutase 1 observed with H/D exchange and mass spectrometry. *Journal of Biological Chemistry, 281*, 18167–18176.
33. Nordlund, A., & Oliveberg, M. (2006). Folding of Cu/Zn superoxide dismutase suggests structural hotspots for gain of neurotoxic function in ALS: Parallels to precursors in amyloid disease. *Proceedings of National Academy of Science United States of America, 103*, 10218–10223.
34. Ding, F., Furukawa, Y., Nukina, N., & Dokholyan, N. V. (2012). Local unfolding of Cu, Zn superoxide dismutase monomer determines the morphology of fibrillar aggregates. *Journal of Molecular Biology*, *421*, 548–560.
35. Furukawa, Y., Kaneko, K., Yamanaka, K., & Nukina, N. (2010). Mutation-dependent polymorphism of Cu, Zn-superoxide dismutase aggregates in the familial form of amyotrophic lateral sclerosis. *The Journal of Biological Chemistry, 285*, 22221–22231.
36. Dumoulin, M., et al. (2005). Reduced global cooperativity is a common feature underlying the amyloidogenicity of lysozyme mutations. *Journal of Molecular Biology, 346*, 773–788.
37. Frare, E., et al. (2006). Identification of the core structure of lysozyme amyloid fibrils by proteolysis. *Journal of Molecular Biology, 361*, 551–561.
38. Plakoutsi, G., Taddei, N., Stefani, M., & Chiti, F. (2004). Aggregation of the acylphosphatase from *Sulfolobus solfataricus*: The folded and partially unfolded states can both be precursors for amyloid formation. *The Journal of Biological Chemistry, 279*, 14111–14119.
39. Corazza, A., et al. (2006). Structure, conformational stability, and enzymatic properties of acylphosphatase from the hyperthermophile *Sulfolobus solfataricus*. *Proteins, 62*, 64–79.
40. Soldi, G., Bemporad, F., & Chiti, F. (2008). The degree of structural protection at the edge beta-strands determines the pathway of amyloid formation in globular proteins. *Journal of the American Chemical Society, 130*, 4295–4302.
41. Pawlicki, S., Le Bechec, A., & Delamarche, C. (2008). AMYPdb: A database dedicated to amyloid precursor proteins. *BMC Bioinformatics, 9*, 273.
42. Garbuzynskiy, S. O., Lobanov, M. Y., & Galzitskaya, O. V. (2010). FoldAmyloid: A method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics, 26*, 326–332.
43. Maurer-Stroh, S., et al. (2010). Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nature Methods, 7*, 855.
44. Conchillo-Sole, O., et al. (2007). AGGRESCAN: A server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics, 8*, 65.
45. Thompson, M. J., et al. (2006). The 3D profile method for identifying fibril-forming segments of proteins. *Proceedings of the National Academy of Sciences United States of America, 103*, 4074–4078.
46. Pawar, A. P., et al. (2005). Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. *Journal of Molecular Biology, 350*, 379–392.
47. Linding, R., Schymkowitz, J., Rousseau, F., Diella, F., & Serrano, L. (2004). A comparative study of the relationship between protein structure and beta-aggregation in globular and intrinsically disordered proteins. *Journal of Molecular Biology, 342*, 345–353.
48. Mohan, A., et al. (2006). Analysis of molecular recognition features (MoRFS). *Journal of Molecular Biology, 362*, 1043–1059.
49. Oldfield, C. J., et al. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry United States of America, 44*, 12454–12470.
50. Garner, E., Romero, P., Dunker, A. K., Brown, C., & Obradovic, Z. (1999). Predicting binding regions within disordered proteins. *Genome Informatics Workshop on Genome Informatics, 10*, 41–50.
51. Zhou, Y. Q., Karplus, M., Wichert, J. M., & Hall, C. K. (1997). Equilibrium thermodynamics of homopolymers and clusters: Molecular dynamics and Monte Carlo simulations of systems with square-well interactions. *Journal of Chemical Physics, 107*, 10691–10708.
52. Sikirzhytski, V., et al. (2012). Fibrillation mechanism of a model intrinsically disordered protein revealed by 2D correlation deep UV resonance Raman spectroscopy. *Biomacromolecules, 13*, 1503–1509.
53. Chiti, F., & Dobson, C. M. (2009). Amyloid formation by globular proteins under native conditions. *Nature Chemical Biology, 5*, 15–22.
54. Kuwata, K., Kamatari, Y. O., Akasaka, K., & James, T. L. (2004). Slow conformational dynamics in the hamster prion protein. *Biochemistry United States of America, 43*, 4439–4446.
55. Saiki, M., Hidaka, Y., Nara, M., & Morii, H. (2012). Stem-forming regions that are essential for the amyloidogenesis of prion proteins. *Biochemistry United States of America, 51*, 1566–1576.
56. Tycko, R., Savtchenko, R., Ostapchenko, V. G., Makarava, N., & Baskakov, I. V. (2010). The alpha-helical C-terminal domain of full-length recombinant PrP converts to an in-register parallel beta-sheet structure in PrP fibrils: Evidence from solid state nuclear magnetic resonance. *Biochemistry United States of America, 49*, 9488–9497.
57. Goldschmidt, L., Teng, P. K., Riek, R., & Eisenberg, D. (2010). Identifying the amylome, proteins capable of forming amyloid-like fibrils. *Proceedings of the National Academy of Sciences United States of America, 107*, 3487–3492.
58. Chiti, F., Stefani, M., Taddei, N., Ramponi, G., & Dobson, C. M. (2003). Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature, 424*, 805–808.
59. Tartaglia, G. G., Cavalli, A., Pellarin, R., & Caflisch, A. (2004). The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein Science, 13*, 1939–1941.
60. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J., & Serrano, L. (2004). Prediction of sequence-dependent and

mutational effects on the aggregation of peptides and proteins. *Nature Biotechnology, 22*, 1302–1306.
61. Yoon, S., & Welsh, W. J. (2004). Detecting hidden sequence propensity for amyloid fibril formation. *Protein Science, 13*, 2149–2160.
62. Ferreiro, D. U., Hegler, J. A., Komives, E. A., & Wolynes, P. G. (2007). Localizing frustration in native proteins and protein assemblies. *Proceedings of the National Academy of Sciences United States of America, 104*, 19819–19824.
63. Jenik, M., et al. (2012). Protein frustratometer: A tool to localize energetic frustration in protein molecules. *Nucleic acids research*, *40*(W1), W348–W351.
64. Dixit, A., & Verkhivker, G. M. (2011). The energy landscape analysis of cancer mutations in protein kinases. *PLoS ONE, 6*, e26071.
65. Yuan, Z., Bailey, T. L., & Teasdale, R. D. (2005). Prediction of protein B-factor profiles. *Proteins, 58*, 905–912.
66. Dosztanyi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics, 21*, 3433–3434.
67. Hecker, J., Yang, J. Y., & Cheng, J. (2008). Protein disorder prediction at multiple levels of sensitivity and specificity. *BMC Genomics, 9*(Suppl 1), S9.
68. Ward, J. J., McGuffin, L. J., Bryson, K., Buxton, B. F., & Jones, D. T. (2004). The DISOPRED server for the prediction of protein disorder. *Bioinformatics, 20*, 2138–2139.
69. Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., & Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE, 4*, e4433.
70. Mizianty, M. J., et al. (2010). Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics, 26*, i489–i496.