J P P JOURNAL OF PROTEINS AND PROTEOMICS
7(1), 2016, pp. 47-60

Research Article

STATISTICAL INSIGHT INTO THE BINDING REGIONS IN DISORDERED HUMAN PROTEOME

Uttam Pal[#], **Mritunjoy Maity**[#], **Nitin Khot, Swagata Das, Supriya Das, Sandip Dolui and Nakul C Maiti***

Structural Biology and Bioinformatics Division, Indian Institute of Chemical Biology, 4, Raja S.C. Mullick Road
Kolkata 700032, India
[#]*These authors contributed equally*

Abstract: The human proteome contains a significant number of intrinsically disordered proteins (IDPs). They show unusual structural features that enable them to participate in diverse cellular functions and play significant roles in cell signaling and reorganization processes. In addition, the actions of IDPs, their functional cooperativity, conformational alterations and folding often accompany binding to a target macromolecule. Applying bioinformatics approaches and with the aid of statistical methodologies, we investigated the statistical parameters of binding regions (BRs) found in disordered human proteome. In this report, we detailed the bioinformatics analysis of binding regions found in the IDPs. Statistical models for the occurrence of BRs, their length distribution and percent occupancy in the parent proteins are shown. The frequency of BRs followed a Poisson distribution pattern with increasing expectancy with the degree of disorderedness. The length of the individual BRs also followed Poisson distribution with a mean of 6 residues, whereas, percentage of residues in BR showed a normal distribution pattern. We also explored the physicochemical properties such as the grand average of hydropathy (GRAVY) and the theoretical isoelectric points (pIs). The theoretical pIs of the BRs followed a bimodal distribution as in the parent proteins. However, the mean acidic/basic pIs were significantly lower/higher than that of the proteins, respectively. We further showed that the amino acid composition of BRs was enriched in hydrophobic residues such as Ala, Val, Ile, Leu and Phe compared to the average sequence content of the proteins. Sequences in a BR showed conformational adaptability mostly towards flexible coil structure and followed by helix, however, the ordered secondary structural conformation was significantly lower in BRs than the proteins. Combining and comparing these statistical information of BRs with other methods may be useful for high-throughput functional annotation of proteins, drug target identification and drug discovery linking protein disorder.

Keywords: IDPs, phylogenetic tree, ANCHOR, sequence analysis, GRAVY, pI, amino acid composition, secondary structure

Coloured figures and supplementary available on journal website

Introduction

Recent investigations and genome analysis revealed the unique presence of intrinsically disordered proteins (IDPs) in eukaryotes (Dunker *et al.*, 2000; Xie *et al.*, 2007a; Monsellier *et al.*, 2008) and more than 30% of amino acid residues in human proteome are believed to be in the disordered regions of proteins (Dosztányi *et al.*, 2010; Habchi *et al.*, 2014; van der Lee *et al.*, 2014; Peng *et al.*, 2014). High content of disorderedness in proteome suggests a functional role of such regions (Haynes *et al.*, 2006; Cumberworth *et al.*, 2013; Fuxreiter *et al.*, 2014). The presence of disorder regions in a protein is thought to confer large plasticity to interact efficiently with several targets, as compared with a globular protein with limited conformational flexibility (Wright and Dyson, 1999; Romero *et al.*, 2001; Dunker *et al.*, 2005). Thus, the disorderedness of proteins are believed to play significant roles in several

Corresponding Author: **Nakul C. Maiti**
E-mail: ncmaiti@iicb.res.in

Received: October 28, 2015
Accepted: December 30, 2016
Published: January 4, 2016

biochemical processes, and have been linked to various molecular recognition processes (Wright and Dyson, 1999; Uversky *et al.*, 2000; Dunker and Obradovic, 2001; Dunker *et al.*, 2002; Xie *et al.*, 2007b; Fong *et al.*, 2009) such as DNA binding, cell cycle regulation, membrane transport and other important cellular functions (Dunker *et al.*, 2002; Dyson and Wright, 2005; Tompa and Csermely, 2004; Xie *et al.*, 2007a). The disorderedness, thus, became an intense topic to know its composition, genomic distribution, cellular localization and energetic aspects linked to function and binding to a targeted partner-molecule.

IDPs lack a compact well defined three dimensional structures in their native state and may instead have a number of thermodynamically stable inter-converting states (Babu *et al.*, 2011; Edwards *et al.*, 2009; Orosz and Ovádi, 2011; Uversky *et al.*, 2000; Vucetic *et al.*, 2003). Some of these proteins are completely unfolded and some contain both the disordered and folded domains with the degree of disorderedness varying from protein to protein (Chen *et al.*, 2006; Dunker *et al.*, 2000). These proteins also have no consistency in their sizes and structurally they resemble the denatured states of ordered proteins (Ahmad *et al.*, 2005; Huang *et al.*, 2006; Uversky, 2002; Weinreb *et al.*, 1996). In a solution, even under physiological conditions, these proteins exist as flexible ensembles of rapidly inter-convertible native conformations (Ahmad *et al.*, 2005; Cohlberg *et al.*, 2002; Huang *et al.*, 2006; Uversky, 2002; Uversky *et al.*, 2000; Weinreb *et al.*, 1996). The binding of a disordered protein to a target molecule or its interaction-partner often causes folding and structural transformation, particularly when it binds to a structured partner/ protein. Figure 1 shows an example of the structural adaptability of a disordered protein α-synuclein under certain conditions. α-Synuclein, which remains in completely coil conformation in aqueous buffer attains predominantly alpha helical structure upon binding the membrane (Figure 1A). Figure 1A also highlights the flexible ensembles of rapidly inter-convertible native conformations of α-synuclein solved by NMR spectroscopy. Figure 1B shows how the binding to a target protein induces beta sheet structure in α-synuclein. Under pathological conditions, intermolecular interactions may even induce beta sheet structure leading to the aggregation of α-synuclein (Figure 1C). Structural alterations of the

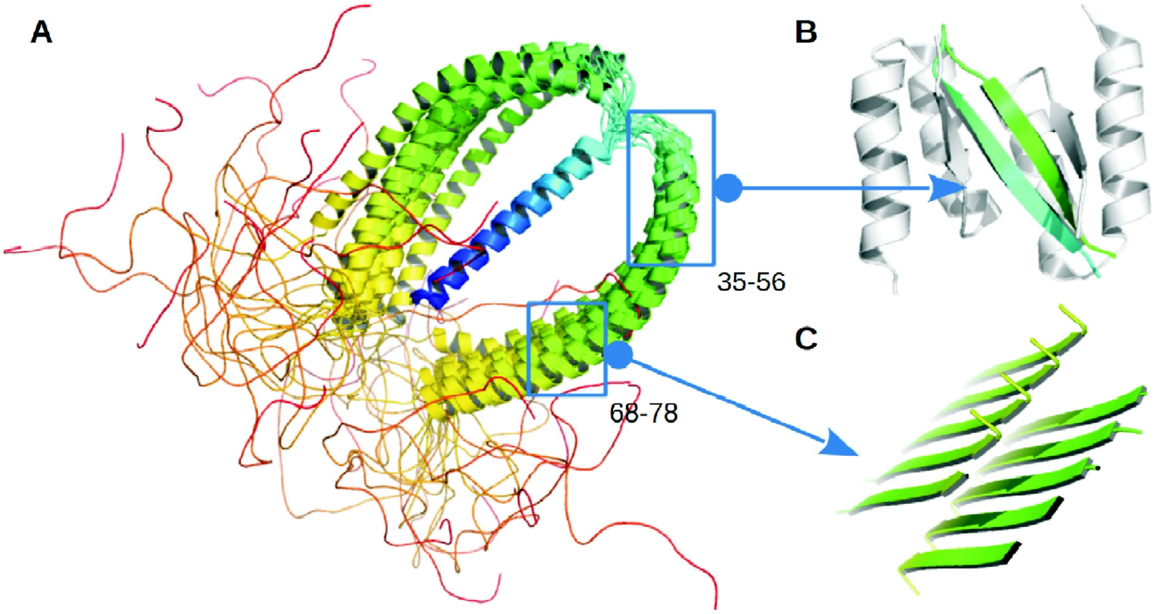

Figure 1: **Structure of a disordered protein, α-synuclein (UniProtKB: P37840). (A) Micelle bound α-synuclein solved by NMR spectroscopy shows predominantly alpha-helical structure (PDB: 2KKW). An ensemble of thermodynamically stable native conformations are shown. N-terminal to C-terminal of the protein is colored in rainbow (violet to red). (B) Binding with a protein induced beta sheet structure in α-synuclein (PDB: 4BXL). (C) α-synuclein takes beta sheet structure when self aggregates to form insoluble fibrils (PDB: 4ZNN, 4RIL).**

binding region may, therefore, render a protein function more effectively and with certain specificity (Dyson and Wright, 2002). Recent studies indicated that binding of the disordered proteins precedes global folding and the interactions follow a complex energy landscape. Conformational transformations in a disordered region are often much larger than the changes in globular proteins (Bordelon *et al.*, 2004; Dunker *et al.*, 2002; Gunasekaran *et al.*, 2003; Sugase *et al.*, 2007). As such, the IDPs can bind to several different partners or interfaces and perform diverse functions (Sharma *et al.*, 2014).

In human proteome and other species, a significant number of proteins are found, which are involved in cellular activity but lack any globular fold in their native state. It represents at least 30% of human proteome and they play a seminal role in cell signaling, memory storage and other cellular function (Nguyen Ba *et al.*, 2012). Therefore, it is important to understand the differences in the structure-function paradigm as it applies to globular proteins and to IDPs. Apart from functional role, numerous IDPs are associated with several human diseases, including cancer, cardiovascular disease, amyloidosis, neurodegenerative disorder and diabetes (Babu *et al.*, 2011; Xie *et al.*, 2007b). Some of these diseases or disorders are inextricably attached to IDPs. α-Synuclein, tau protein, amyloid beta (Aβ) are among many IDPs involved in diseases like Parkinson's disease (PD) and Alzheimer's disease (AD). Also, it was observed that α-synuclein and tau binding often lead and accelerate the aggregation of the proteins and formation of amyloid. Understanding intermolecular interaction or binding among IDPs and other candidates like small organic molecules and small peptides, therefore, is an interesting area to explore. The IDPs, as such could be attractive targets for designing drug molecules that may modulate the protein-protein interactions (Apetri *et al.*, 2006).

Some regions of the IDPs are prone to interact with target molecules and act as binding regions (BRs) or functional part of the proteins. Short functional regions in the disordered region, based on computational analysis and other prediction methods, were detected and termed as molecular recognition features (MoRFs) (Cheng *et al.*, 2007; Disfani *et al.*, 2012; Mohan *et al.*, 2006; Vacic *et al.*, 2007). Computational studies and experimental investigations further verified that BRs in IDPs are exposed and often considered as a primary contact site for the interaction and binding (Csizmók *et al.*, 2005). These regions frequently showed structural propensities similar to the structure they attained upon complex formation with the partner molecule (Dancheck *et al.*, 2008; Fuxreiter *et al.*, 2004). In the present investigation, we aimed to derive the statistical parameters of the physicochemical properties linked to BRs in the intrinsically disordered human proteome. Elucidating binding regions (BRs) and associated statistical knowledge is crucial to address functional and binding roles of the proteins. Our result provided the models of statistical distributions on different aspects of the BRs in IDPs such as the occurrence of BRs, their length, percent occupancy in the parent proteins and the correlations with the degree of disorderedness of the proteins.

We selected all the experimentally validated and annotated proteins with different degrees of disorderedness from IDEAL (Intrinsically Disordered proteins with Extensive Annotations and Literature; release 21 March 2014) and DisProt databases (release 6.02) (Fukuchi *et al.*, 2012; Sickmeier *et al.*, 2007). Several computational methods are available to predict the binding regions in a disordered protein region (Sharma *et al.*, 2014) such as MoRFpred, DISOPRED3 and ANCHOR. Among them, MoRFpred and DISOPRED3 are developed to predict short protein-binding regions in disorder region, which are implicated in molecular recognition processes (Jones and Cozzetto, 2015). We used ANCHOR method to detect the BRs in disordered protein dataset. Short segments of unfolded proteins that showed propensity to interact with some target molecules (mainly proteins) with a possibility of structural recognition are key to the detection of binding region by ANCHOR method (Dosztányi *et al.*, 2009; Mészáros *et al.*, 2009). The method utilizes a statistical potential matrix based on pairwise interaction energy from known coordinates using a dataset of globular proteins (Dosztányi *et al.*, 2005, 2009; Mészáros *et al.*, 2009). ANCHOR is independent of amino acid composition, although it was reported that the construction of the algorithm for the prediction of interaction energy implies its sensitivity to amino acid composition (Mészáros *et al.*, 2009). We found more than 3000 binding regions in the human disordered proteome which were partly or fully disordered in their native state.

Different statistical models were invoked to describe the distribution pattern of frequency and length of the BRs, hydrophobicity and many other properties which are very crucial for the binding regions for their functional activity. The statistical patterns and associated parameters so derived could be used to predict the behavior and property of new proteins that contain certain degree of disorderedness. These information could be useful in medicine and of interest in protein disorder as a possible target for designing drug molecules. Partial or fully disorder states play critical role in cell signaling and also linked to several disease formation and therefore the binding regions could be the targets for designing drug molecules to arrest/stop progression of disease linked to protein disorder.

Materials and Methods

Compilation of Dataset

Information of the intrinsically disordered human proteins was obtained from IDEAL (Intrinsically Disordered proteins with Extensive Annotations and Literature; release 21 March 2014) database and DisProt database, release 6.02 (Fukuchi *et al.*, 2012; Sickmeier *et al.*, 2007). We retrieved 471 unique protein sequences from UniProt (release 2014_07) after ID mapping. IDEAL database entries have extensive annotations of disorderedness. DisProt also lists the IDPs detected by experimental methods such as fluorescence, circular dichroism, FTIR, sensitivity to proteolysis etc. Therefore, our dataset comprised experimentally determined and extensively annotated IDPs and represented human disordered proteome. Sequences were obtained from UniProt in FASTA format and then converted to strings of one letter amino acid codes for further analysis.

Sequence Comparisons and Phylogenetic Analysis

Global alignment of the sequences was performed by Clustal Omega, which is a multiple sequence alignment program available at EMBL-EBI web server (*http://www.ebi.ac.uk/Tools/msa/clustalo/*). It uses seeded guide trees and hidden Markov model profile-profile techniques to generate alignments between sequences. Phylogenetic tree was generated from the sequence alignment using the neighbor-joining algorithm to construct trees from the distance matrix by neighbor joining method. The tree was rendered at the Interactive Tree Of Life (iTOL) server, which is an online tool for the display and manipulation of phylogenetic trees (http://itol.embl.de/).

Calculation of Disorderedness and Binding Regions

Disorderedness of the proteins was computed using the IUPred program (Dosztányi *et al.*, 2005). The ANCHOR method was engaged to detect the binding regions in the IDPs. ANCHOR analyzed the input sequences of unfolded protein and predicted the binding regions based on certain scoring values (Dosztányi *et al.*, 2009; Mészáros *et al.*, 2012). BR sequences were obtained from the protein sequences using the position and length parameters.

Sequence Analysis

Length, amino acid composition, charged residues, total charge, and molecular weight were calculated from the sequence data. The GRAVY value for a BR or protein was calculated as the average of hydropathy values (Kyte and Doolittle, 1982) of all the amino acids. Isoelectric points were calculated using the Compute pI/MW tool (Bjellqvist *et al.*, 1994; Gasteiger *et al.*, 2005) at ExPASy Bioinformatics Resource Portal. Computational algorithm PSIPRED was used to predict the conformation propensity for each protein from their amino acid sequence (Jones, 1999). Larger proteins were segmented into domains using DomPred prior to secondary structure prediction. Percentage of residues in a protein with preference for a particular conformation was measured by taking a ratio of the total number of residues preferring a particular conformation to the protein sequence length. Secondary structure compositions of BRs were obtained from the parent protein analysis using the position and length parameters.

Statistical Analysis

All the statistical analysis was performed in Wolfram Mathematica 10. Cramér-von Mises test (Anderson and Darling, 1952) was used to test the normality of the data. For the normally distributed data, mean, standard deviation (SD) and standard error of mean (SEM) were calculated. Significance of the mean differences was established with Student's t-test and the null hypotheses were rejected at the 5 percent level of significance. Probability values of less than 0.0005 were considered as highly significant and denoted by *** in the graphs. Likewise, the probability values in between 0.0005 and 0.005 were considered as very significant and denoted by ** in the graphs and the rest were denoted by a single star. Poisson distribution was fitted to the BR frequency and length data (discrete random variables). The probability mass function (PMF) is given by:

$$f(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!}, \tag{1}$$

where e is Euler's number and μ is the expected value of the random variable x. The cumulative distribution function (CDF) is given by:

$$g(x;\mu) = e^{-\mu}\sum_{i=0}^{\lfloor x \rfloor}\frac{\mu^{i}}{i!}, \tag{2}$$

where $|x|$ is the floor function.

Generalized Poisson distribution function or the Poisson-Consul distribution is given by the equation 3.

$$f(x;\mu) = \frac{e^{-x\lambda-\mu}(x\lambda+\mu)^{-1+x}}{x!}, \tag{3}$$

where λ is any real number between 0 and 1.

Normal distribution with mean (μ) and standard deviation (σ) were fitted to the normally distributed data. The probability distribution function is given by:

$$P(x) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(x-\mu)^{2}}{2\sigma^{2}}}, \tag{4}$$

The cumulative distribution function is given by:

$$D(x) = \frac{1}{2}\left[1+erf\left[\frac{x-\mu}{\sqrt{2\sigma}}\right]\right], \tag{5}$$

where erf is the error function.

PDF and CDF of the skewed normal distribution were described by the following two equations, respectively.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^{2}}{2\sigma^{2}}}, \tag{6}$$

$$D(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^{2}}{2\sigma^{2}}}\,Erfc\left[\frac{-\alpha(x-\mu)}{\sigma\sqrt{2\pi}}\right], \tag{7}$$

For the bimodal distributions, a mixture of distributions was fitted to the data. In regression analysis straight lines passing through the origin were fitted to the data.

Results and Discussion

The protein dataset (Table S1) comprised experimentally determined disordered proteins obtained from IDEAL and DisProt databases. These proteins were extensively annotated IDPs and represented human disordered proteome. Based on the content of structural disorderedness, the proteins were grouped into three categories as suggested in previous reports (Das *et al.*, 2014; Schad *et al.*, 2013). In the dataset, the total number of largely disordered proteins (LDP, structural disorderedness >70%) was 71. 163 proteins with disorderedness ranging from 30 to 70% were grouped as moderately disordered proteins (MDP). Rest of the proteins, having less than 30% disorderedness, was grouped as partially disordered protein (PDP).

We compared all the sequences of human disordered proteome in the dataset using the multiple sequence alignment tool, Clustal Omega. A phylogenetic tree was derived from this alignment to study the similarity and evolutionary distances between the sequences.

Closely similar proteins were grouped together and the dissimilar proteins got separated in different groups. The relationships among the proteins are shown in cladograms (Figure 2). The detailed tree with branch length information and leaf labels is given in the supporting information (Scheme S1). Figure 2 shows the cladograms of the tree with 471 leafs in the rooted and unrooted modes. Some important disease related disordered proteins such as α-synuclein, BRCA1, p53 and amyloid-β are marked. The tree suggests that the α-synuclein is closer to BRCA1 in sequence similarity than p53 or amyloid-α. Proteins with different degrees of disorderedness were also color coded to show their distribution among the different clades of disordered proteins. It was found that all the clades have proteins with varying degree of disorderedness.

We used ANCHOR algorithm to detect the BRs in disordered protein dataset. ANCHOR largely depends on the pair-wise energy estimation method that is used in IUpred algorithm. IUpred was used to quantitate the disorderedness of the proteins. Binding regions were selected by ANCHOR algorithm by identifying region in the polypeptide chain that are in disordered regions and not supported by favorable intra chain interactions to attain folded structure. ANCHOR detected 3494 binding regions (BRs) in 471 unique human proteins with different degree of disorderedness (Table S1 and S2). Table S1 lists the number of BRs in each protein, their total lengths and percent occupancy in the parent protein, whereas, Table S2 lists the details of the individual binding regions. Most of the proteins contained multiple binding regions.

Figure 3A and 3B shows the probability distributions of BR frequency in all the three groups of protein, whereas, Figure 3C and 3D shows the probability distribution of finding a BR per 100 residues of a protein. Interestingly, the number of BRs did not always follow a normal distribution. Instead, it shows a Poisson distribution pattern suggesting that the occurrence of BRs in a protein is a stochastic process, which satisfies the Markov property (Durbin, 1998; Nguyen Ba *et al.*, 2012). Figure 3 shows the fitted poison distributions for BRs in whole protein and also with respect to 100 residues in each group of proteins. The Poisson distribution analysis provided the expectation values (μ), which represent the occurrence rate of the event (here number of BRs). The expected value of BRs was 3 in MDP and LDP. In PDP the expected BR frequency was 1 (Table 1). However, the expected values of BRs per 100 residues of a protein were found to be 0, 1 and 2 for PDP, MDP and LDP, respectively. Interestingly we observed that the percentage of residues in BR followed a normal distribution pattern. In PDP the normal distribution was positively skewed. We observed a shift in the modal class with the increasing degree of disorderedness in proteins. In the LDPs the percentage of residues in BRs were very high compared to the MDPs or PDPs. The increase in the number of expectation values with protein disorderedness was within the scope of ANCHOR algorithm, which was used to detect the BRs in the disordered protein dataset, however, we described here the detailed statistics of BR frequencies, provided the quantitative parameters and showed how the BR frequencies correlates with the disorderedness of the protein, which would be useful for understanding the origin of binding regions in disordered as well as ordered proteins.

We further studied the distribution pattern of the length of BRs with respect to degree of protein disorderedness. We observed that the content of BRs did not follow a normal distribution. Generalized Poison distribution formula fitted the data much better than the normal distribution. Figure 3E and 3F displays the length distribution

Table 1
Poisson/Poisson-Consul distribution parameters (μ and λ) for BR frequency, BR frequency per 100 residues and the length of the BRs

Variable	*Group*	*μ*	*λ*
BR frequency	PDP	1.41	0.56
	MDP	3.82	0.67
	LDP	3.83	0.70
BR frequency per 100 residues of protein	PDP	0.49	–
	MDP	1.65	–
	LDP	2.42	–
BR length	PDP	6.40	0.49
	MDP	6.38	0.66
	LDP	6.44	0.75

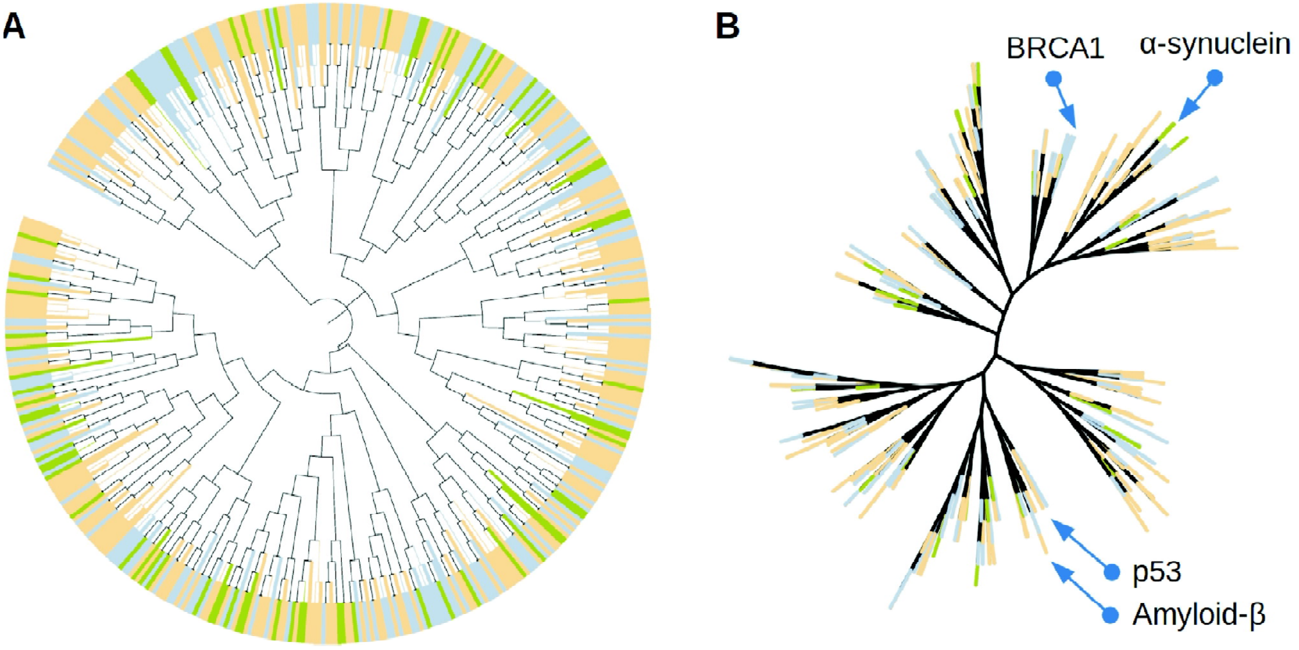

Figure 2: Human disordered proteome tree. (A) A 471 leaf tree with colored clades rendered in circular mode. The tree is shown without branch length information. (B) The same tree is shown in unrooted mode. Some important disease related disordered proteins are marked. Color strip dataset was used to define branch colors for different group of disordered proteins: partially disordered proteins (PDP), gold; moderately disordered proteins (MDP), blue; largely disordered proteins (LDP), green.

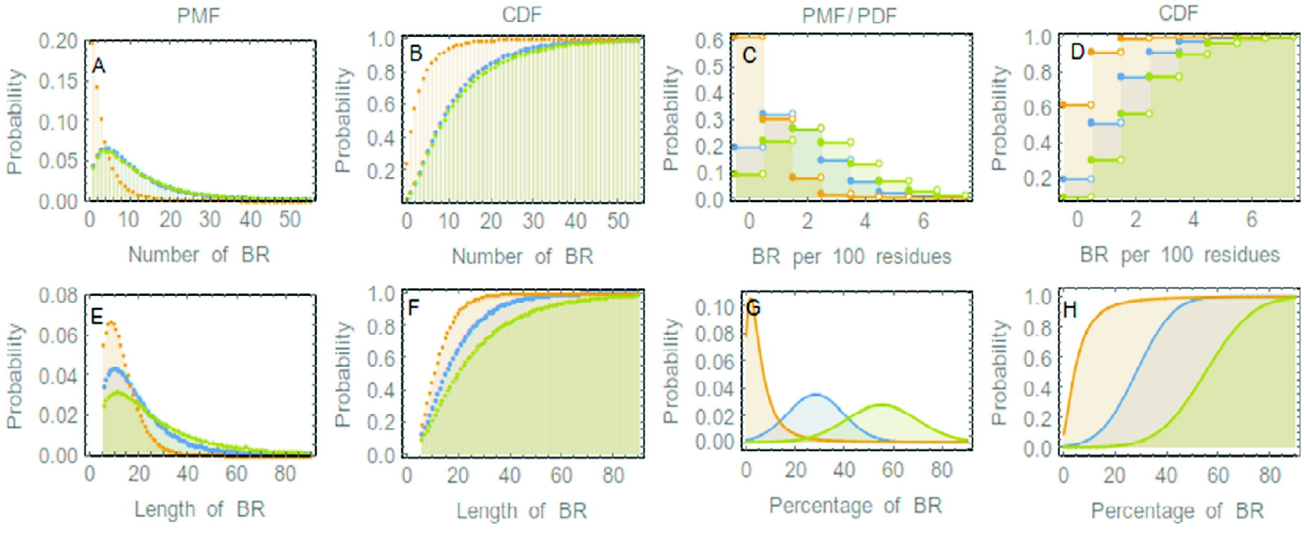

Figure 3: Frequency and length distribution of binding regions (BRs). (A) Probability of occurrence of a BR (BR frequency) in different group of disordered proteins. Probability mass function (PMF) of the fitted Generalized Poisson Distribution is shown. (B) Cumulative distribution function (CDF) of the BR frequency. (C) Probability of occurrence of a BR per 100 residues of a protein. (D) CDF of the BR frequency per 100 residues of a protein. (E) Probability distribution (PMF) of individual BR lengths in different group of disordered proteins. (F) CDF of the BR length distribution. (G) The distribution of BR content (percent occupancy) in a protein. PDF of the fitted skewed/normal distribution is shown. (H) CDF of the BR content distribution. Color key: gold, partially disordered proteins (PDP); blue, moderately disordered proteins (MDP); green, largely disordered proteins (LDP).

of the individual BRs and Figure 3G and 3H shows the percent occupancy in different group of proteins. We observed an expectation value of the BR length of 6 residues in all the three groups of protein (Table 1). However, the spread of the distributions increased with the increase in disorderedness.

In order to better understand how the BR frequencies, BR frequency per 100 residues of a protein, BR length and percent occupancy correlates with the protein disorderedness we performed regression analysis as shown in the Figure 4. Frequency versus disorderedness plots suggested linear regressions. Therefore, straight lines passing through origins were fitted to the data. BR expectancy per 100 residues produced discrete densities over the continuous axis of protein disorderedness, which gave a much clearer understanding about how the expected number of BRs per 100 residues changes depending on the protein disorderedness. Analysis showed that the length of the individual BRs did not correlate with the disorderedness of the protein, which was evident from the very low R^2 value of the linear model fit (Figure 4F). However, the percent occupancy of the BRs in a protein increased linearly along with the protein disorderedness with a coefficient of 0.62.

Using the hydropathy indexes (Wu *et al.*, 2006) of individual amino acids, grand average hydropathy (GRAVY) values of the BRs and the parent proteins (Tables S1 and S2) were derived. The calculated GRAVY indexes of all the proteins were predominantly negative and varied between 0 to -2, approximately (Figure 5 and Table 2 and S3). It was expected as the proteins were rich in polar and charged residues. Mean GRAVY decreased with the increase in the degree of disorderedness (PDP, MDP and LDP). However, the spread of GRAVY values was very high (ranging 2 to -2, approximately) for BRs, mean nearing neutrality (Figure 5).

Distribution of the isoelectric points (pI) of BRs and the parent proteins are shown in the Figure 6. Statistical analysis showed that theoretical pI values followed a bimodal distribution for both the proteins and BRs (Figure 6 and Table 3). pIs were mostly distributed either in acidic or in basic regions, but rarely at the neutral pH. On both sides, they followed a normal distribution. Such multimodal distributions of pIs for the whole proteome are known in the literature (Kiraga *et al.*, 2007; Nandi *et al.*, 2005; Taylor *et al.*, 2002). pI distribution of the BRs closely followed that of the parent proteins in terms of the density. However, the mean pI of BRs in the acid ranges was found to be significantly less than their parent proteins and in the basic ranges significantly higher.

The amino acid composition of the binding regions is shown in Figure 7 and compared to the protein sequence composition. ANCHOR is independent of amino acid composition, although it was reported that the construction of the algorithm for the prediction of interaction energy implies its sensitivity to amino acid composition (Mészáros *et al.*, 2009). In most disordered regions the functional amino acid residues remain unknown (Nguyen Ba *et al.*, 2012). We have found that the BRs mostly differ from their parent proteins in the content of charged or polar amino acids. Charged amino acids such as Glu, Lys, Arg, and Asp are present in significantly lower amounts in the BRs so are the uncharged polar residues: Thr and Asn. Hydrophobic amino acids such as Leu, Ala, Val, Ile and Phe are more abundant in the BRs (Maity and Maiti, 2012). Ser alone in the uncharged-polar group is present in significantly higher number in the BRs. It is the smallest amino acid in this group having least bulky side-chain. Hydrophobic and hydrogen bonding interactions are the major players in the

Table 2
Fitted Normal Distribution parameters for GRAVY

Groups	*PDP*		*MDP*		*LDP*	
	μ	*σ*	*μ*	*σ*	*μ*	*σ*
Protein	-0.41	0.25	-0.66	0.25	-0.98	0.42
BR	-0.35	0.74	-0.05	0.81	0.30	0.91

Table 3
Mean acidic and basic pIs of proteins and BRs. Values are given as μ ± σ

Groups	*pI acidic*	*pI basic*
Protein	5.68±0.71	9.03±1.05
BR	4.91±1.05	9.63±1.24

t test (Protein vs BR): pI acidic $3.53295*10^{-76}$; pI basic $1.47157*10^{-29}$

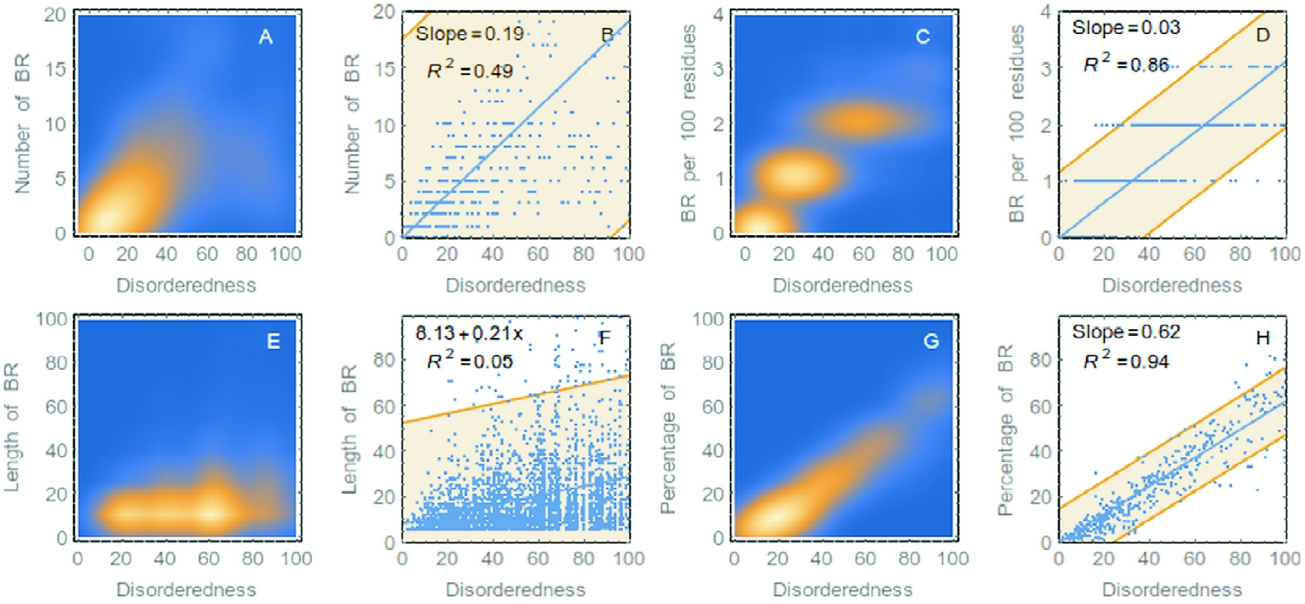

Figure 4: **Correlation of BR frequency/occupancy with protein disorderedness. Distribution of BR frequency with protein disorderedness (A) and the fitted linear model (B). Distribution of BR frequency per 100 residues of protein with the protein disorderedness (C) and the fitted linear model (D). Distribution of BR lengths with protein disorderedness (E) and the fitted linear model (F). Distribution of BR occupancy in a protein with the protein disorderedness (G) and the fitted linear model (H). Confidence level bands at 95% are shown.**

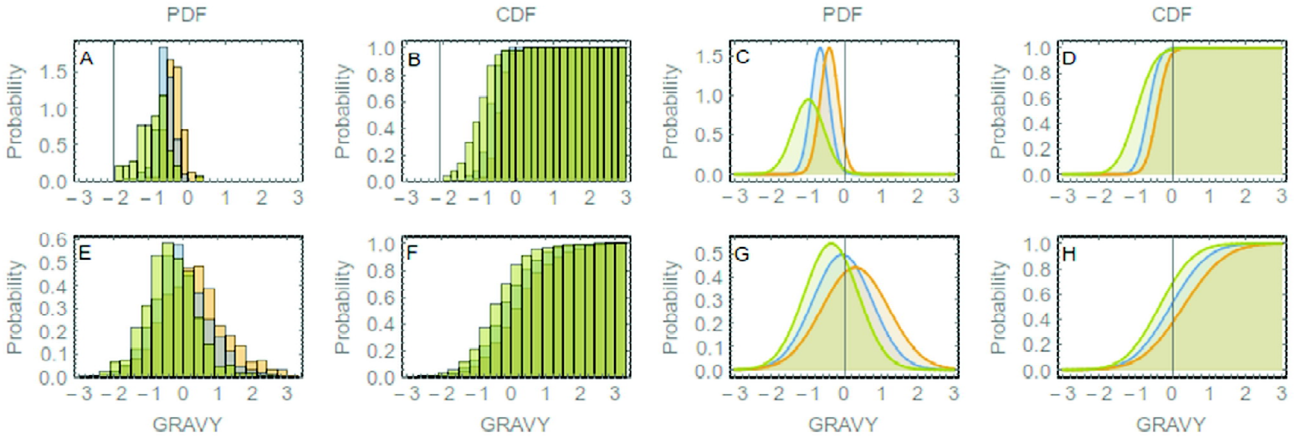

Figure 5: **GRAVY distribution of the whole proteins versus BRs. Row 1: GRAVY of the whole proteins; row 2: GRAVY of the BRs. Column 1: histogram of PDF; column 2: histogram of CDF; column 3: fitted normal distributions (PDF); column 4: fitted normal distributions (CDF). Color key: gold, PDP; blue, MDP; green, LDP.**

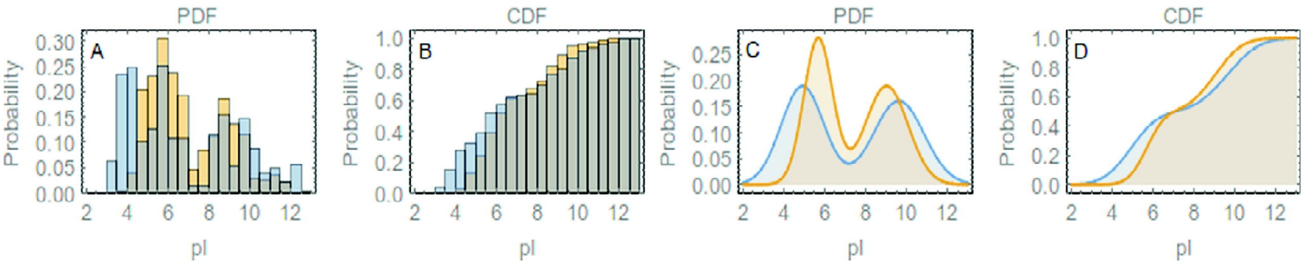

Figure 6: **Distribution of isoelectric points (pI) in proteins and BRs. (A) Histograms of pI densities. (B) Histograms of cumulative densities of pI. (C) Fitted bimodal distributions. (D) CDF of the fitted distributions to pI densities. Color key: gold, protein; blue, BR.**

ligand-protein or protein-protein binding (Jiang *et al.*, 2002). Contribution of electrostatic/ionic interactions is much less compared to them. Therefore, the less abundance of charged and uncharged-polar residues and the prominence of hydrophobic residues in the BRs are desirable and justified. Figure 7 also shows the comparison of mean residue molecular weights (MRW) of protein and BRs. Although the mean MRW is similar in both the BRs and Protein, the spread is very high for the BRs.

Elucidating binding regions and associated structure formation on these binding regions in IDPs is significant as this is the starting point for investigations into higher-order structure and, thus, functions of IDPs. The conformations (extended/β-strand, helix and coil), the residues in proteins and in BRs prefer to adopt are shown in Figure 8. It should be noted that ANCHOR analysis is independent of the adopted secondary structure (Mészáros *et al.*, 2009) and, therefore, the result was not biased by the algorithm. We found that the ordered secondary structure decreased with the increase in disorderedness of the protein. Coil conformation was found to be the most preferred conformation (Maity and Maiti, 2012) in all the three groups of disordered proteins, followed by helix and then the extended or β-strand/sheet. BR structural propensity followed a similar trend in all three groups as well. However, in the PDP and MDP groups the BR structure propensity toward coil conformation was significantly higher than that of the proteins in that group. In MDP, the propensity toward helix was significantly lower for BRs. Although the trend is visible, such statistical significance could not be established in other groups. However, when PDP, MDP and LDP data were combined, we observed significantly less propensity toward helix and higher propensity toward coil in the BR residues (Figure 8D). Propensity toward extended conformation was also significantly lower in BRs. The overall structural content of BR sequences was: extended ~6%, helix ~18% and coil conformation ~76% indicating that binding region was dominated with sequences that preferred to be flexible. In total protein, however, structural preference of the sequences was: extended ~9%, helix ~27% and coil ~64%. The analysis showed that very few residues preferred β-sheet/strand conformation and both the BRs and the parent protein molecules are rich in sequences, most of which preferred coil/random conformation and the propensity for coil conformation is more in BR sequences.

Experimental and computational studies highlighted widespread roles of protein disorder in biological processes (Dunker *et al.*, 2002; Dyson and Wright, 2005; Gsponer *et al.*, 2008; Krishnan

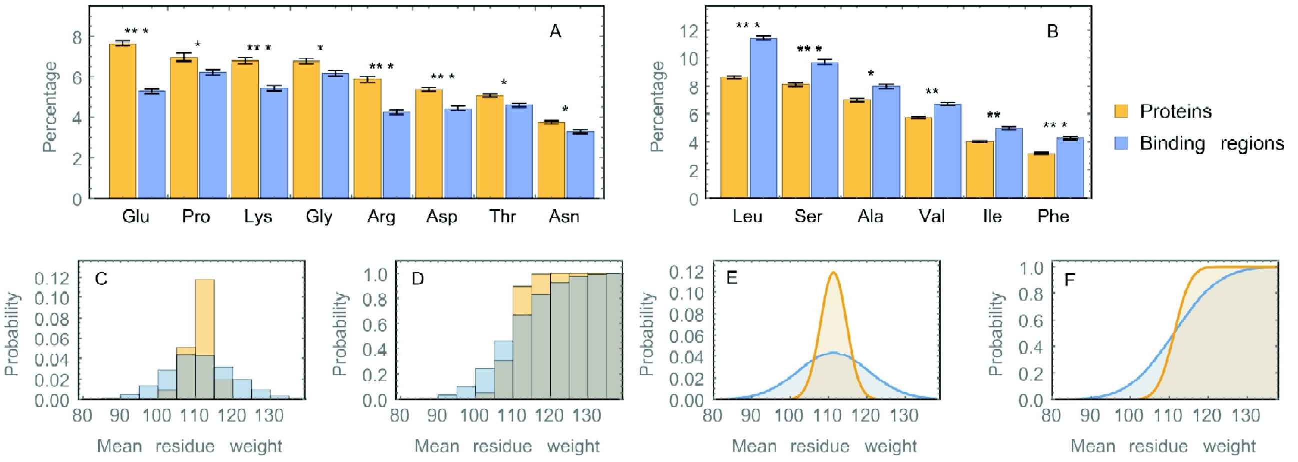

Figure 7: **Comparison of amino acid composition between whole protein and the BRs. (A) Amino acids that are more abundant in whole protein. (B) Amino acids that are more abundant in BRs. Significant variations were marked with asterisks. ***, p-value <0.0005; **, p-value <0.005 but not <0.0005; *, p-value <0.05 but not <0.005. (C) Comparison of mean residue molecular weight (MRW) distribution of BR and protein. (D) CDF of MRW histogram. (E) Fitted normal distributions of MRW (PDF). (F) Fitted normal distributions of MRW (CDF). Color key: gold, protein; blue, BR.**

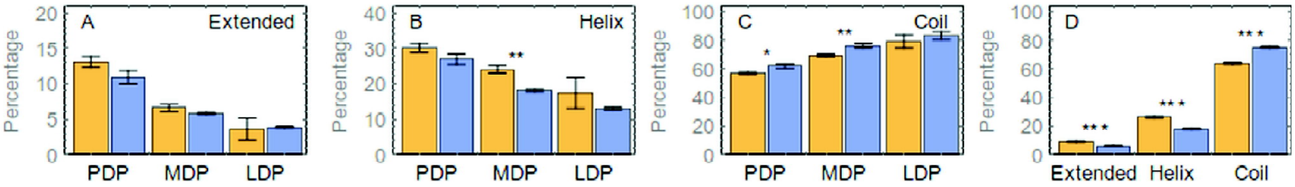

Figure 8: **Comparison of the secondary structure propensity of the BRs to that of whole proteins. (A) Propensity for extended conformation in three groups of disordered proteins. (B) Propensity for helix and (C) propensity for coil conformation. (D) PDP, MDP and LDP combined. Significant variations were marked with asterisks. ***, p-value <0.0005; **, p-value <0.005 but not <0.0005; *, p-value <0.05 but not <0.005. Color key: gold, protein; blue, BR.**

et al., 2014; Wright and Dyson, 1999). Recent discovery showed that some protein phase separation leads to formation of membrane less organelles/component which have important roles in cellular function; IDPs have significant role in the formation of such assembly and localization of many signaling proteins to act efficiently. Protein disorder is also linked with several diseases and, therefore, the disordered proteins are considered important drug targets in rational drug design (Cheng *et al.*, 2006; Lao *et al.*, 2014). However, the disordered protein regions do not act as isolated domains and the surrounding segments in addition to its length also govern its function and stability. Thus, it is very important to characterize the large number of disordered regions along with total protein to realize their greater role in cellular activity and to develop new strategies for drug design targeting specific regions in IDPs.

Conclusion

Our analysis provided the content, composition and statistical behavior of the binding regions in disordered proteins and some of its physicochemical aspects such as isoelectric point and hydrophobicity. We have shown the distributions of BR lengths and their percent occupancy in the parent proteins. We have described the correlations of BR occurrence frequencies, lengths and percentages with the degree of disorderedness of the parent protein. Statistical models for the occurrence of BRs in disordered proteins were derived. Some parameters followed poison distribution and some others showed normal distribution. Theoretical pI values followed a bimodal distribution. The statistical analysis further illustrated how the linked parameters differed along with the content of protein disorderedness. The report also shows that the BRs contained amino acids optimal for hydrophobic and the hydrogen bonding type of interactions with the target molecule/protein. Hydrophobicity of the BRs was widespread and the pIs were more acidic or more basic than that of the parent proteins. The structural disposition of BRs towards the more flexible coil conformation was also discussed. It would be interesting to test the binding and functional efficacy of the regions with some of the target molecules.

References

Ahmad, A., Uversky, V.N., Hong, D., and Fink, A.L. (2005). Early events in the fibrillation of monomeric insulin. J. Biol. Chem. *280*, 42669–42675.

Anderson, T.W., and Darling, D.A. (1952). Asymptotic Theory of Certain “Goodness of Fit” Criteria Based on Stochastic Processes. Ann. Math. Stat. *23*, 193–212.

Apetri, M.M., Maiti, N.C., Zagorski, M.G., Carey, P.R., and Anderson, V.E. (2006). Secondary Structure of á-Synuclein Oligomers: Characterization by Raman and Atomic Force Microscopy. J. Mol. Biol. *355*, 63–71.

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

Bjellqvist, B., Basse, B., Olsen, E., and Celis, J.E. (1994). Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. Electrophoresis *15*, 529–539.

Bordelon, T., Montegudo, S.K., Pakhomova, S., Oldham, M.L., and Newcomer, M.E. (2004). A disorder to order transition accompanies catalysis in retinaldehyde dehydrogenase type II. J. Biol. Chem. *279*, 43085–43091.

Chen, J.W., Romero, P., Uversky, V.N., and Dunker, A.K. (2006). Conservation of Intrinsic Disorder in Protein Domains and Families: II. Functions of Conserved Disorder. J. Proteome Res. *5*, 888–898.

Cheng, Y., LeGall, T., Oldfield, C.J., Mueller, J.P., Van, Y.-Y.J., Romero, P., Cortese, M.S., Uversky, V.N., and

Dunker, A.K. (2006). Rational drug design via intrinsically disordered protein. Trends Biotechnol. *24*, 435–442.

Cheng, Y., Oldfield, C.J., Meng, J., Romero, P., Uversky, V.N., and Dunker, A.K. (2007). Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry (Mosc.) *46*, 13468–13477.

Cohlberg, J.A., Li, J., Uversky, V.N., and Fink, A.L. (2002). Heparin and other glycosaminoglycans stimulate the formation of amyloid fibrils from alpha-synuclein in vitro. Biochemistry (Mosc.) *41*, 1502–1511.

Csizmók, V., Bokor, M., Bánki, P., Klement, E., Medzihradszky, K.F., Friedrich, P., Tompa, K., and Tompa, P. (2005). Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2. Biochemistry (Mosc.) *44*, 3955–3964.

Cumberworth, A., Lamour, G., Babu, M.M., and Gsponer, J. (2013). Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochem. J. *454*, 361–369.

Dancheck, B., Nairn, A.C., and Peti, W. (2008). Detailed structural characterization of unbound protein phosphatase 1 inhibitors. Biochemistry (Mosc.) *47*, 12346–12356.

Das, S., Pal, U., Das, S., Bagga, K., Roy, A., Mrigwani, A., and Maiti, N.C. (2014). Sequence Complexity of Amyloidogenic Regions in Intrinsically Disordered Human Proteins. PLoS ONE *9*, e89781.

Disfani, F.M., Hsu, W.-L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N., and Kurgan, L. (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.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics *25*, 2745–2746.

Dosztányi, Z., Mészáros, B., and Simon, I. (2010). Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief. Bioinform. *11*, 225–243.

Dunker, A.K., and Obradovic, Z. (2001). The protein trinity—linking function and disorder. Nat. Biotechnol. *19*, 805–806.

Dunker, A.K., Obradovic, Z., Romero, P., Garner, E.C., and Brown, C.J. (2000). Intrinsic protein disorder in complete genomes. Genome Inform. *11*, 161–171.

Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M., and Obradoviæ, Z. (2002). Intrinsic disorder and protein function. Biochemistry (Mosc.) *41*, 6573–6582.

Dunker, A.K., Cortese, M.S., Romero, P., Iakoucheva, L.M., and Uversky, V.N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. *272*, 5129–5148.

Durbin, R. (1998). Biological sequence analysis: probabilistic models of proteins and nucleic acids (Cambridge UK: Cambridge university press).

Dyson, H.J., and Wright, P.E. (2002). Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol. *12*, 54–60.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. *6*, 197–208.

Edwards, Y.J., Lobley, A.E., Pentony, M.M., and Jones, D.T. (2009). Insights into the regulation of intrinsically disordered proteins in the human proteome by analyzing sequence and gene expression data. Genome Biol. *10*, R50.

Fong, J.H., Shoemaker, B.A., Garbuzynskiy, S.O., Lobanov, M.Y., Galzitskaya, O.V., and Panchenko, A.R. (2009). Intrinsic Disorder in Protein Interactions: Insights From a Comprehensive Structural Analysis. PLoS Comput Biol *5*, e1000316.

Fukuchi, S., Sakamoto, S., Nobe, Y., Murakami, S.D., Amemiya, T., Hosoda, K., Koike, R., Hiroaki, H., and Ota, M. (2012). IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. Nucleic Acids Res. *40*, D507–D511.

Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004). Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J. Mol. Biol. *338*, 1015–1026.

Fuxreiter, M., Tóth-Petróczy, Á., Kraut, D.A., Matouschek, A.T., Lim, R.Y.H., Xue, B., Kurgan, L., and Uversky, V.N. (2014). Disordered Proteinaceous Machines. Chem. Rev. *114*, 6806–6843.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M.R., Appel, R.D., and Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In The Proteomics Protocols Handbook, J.M. Walker, ed. (Humana Press), pp. 571–607.

Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. Science *322*, 1365–1368.

Gunasekaran, K., Tsai, C.-J., Kumar, S., Zanuy, D., and Nussinov, R. (2003). Extended disordered proteins: targeting function with less scaffold. Trends Biochem. Sci. *28*, 81–85.

Habchi, J., Tompa, P., Longhi, S., and Uversky, V.N. (2014). Introducing Protein Intrinsic Disorder. Chem. Rev. *114*, 6561–6588.

Haynes, C., Oldfield, C.J., Ji, F., Klitgord, N., Cusick, M.E., Radivojac, P., Uversky, V.N., Vidal, M., and Iakoucheva, L.M. (2006). Intrinsic Disorder Is a Common Feature of Hub Proteins from Four Eukaryotic Interactomes. PLoS Comput Biol *2*, e100.

Huang, K., Maiti, N.C., Phillips, N.B., Carey, P.R., and Weiss, M.A. (2006). Structure-specific effects of protein topology on cross-beta assembly: studies of insulin fibrillation. Biochemistry (Mosc.) *45*, 10278–10293.

Jiang, L., Gao, Y., Mao, F., Liu, Z., and Lai, L. (2002). Potential of mean force for protein-protein interaction studies. Proteins 46, 190–196.

Jones, D.T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. *292*, 195–202.

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

Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M.R., and Cebrat, S. (2007). The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics *8*, 163.

Krishnan, N., Koveal, D., Miller, D.H., Xue, B., Akshinthala, S.D., Kragelj, J., Jensen, M.R., Gauss, C.-M., Page, R., Blackledge, M., *et al.* (2014). Targeting the disordered C terminus of PTP1B with an allosteric inhibitor. Nat. Chem. Biol. *10*, 558–566.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lao, B.B., Drew, K., Guarracino, D.A., Brewer, T.F., Heindel, D.W., Bonneau, R., and Arora, P.S. (2014). Rational Design of Topographical Helix Mimics as Potent Inhibitors of Protein–Protein Interactions. J. Am. Chem. Soc. *136*, 7877–7888.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., *et al.* (2014). Classification of Intrinsically Disordered Regions and Proteins. Chem. Rev. *114*, 6589–6631.

Maity, M., and Maiti, N.C. (2012). Sequence Composition of Binding Sites in Natively Unfolded Human Proteins. J. Proteins Proteomics 3, 117–125.

Mészáros, B., Simon, I., and Dosztányi, Z. (2009). Prediction of Protein Binding Regions in Disordered Proteins. PLoS Comput Biol *5*, e1000376.

Mészáros, B., Dosztányi, Z., and Simon, I. (2012). Disordered Binding Regions and Linear Motifs—Bridging the Gap between Two Models of Molecular Recognition. PLoS ONE *7*, e46829.

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

Monsellier, E., Ramazzotti, M., Taddei, N., and Chiti, F. (2008). Aggregation Propensity of the Human Proteome. PLoS Comput Biol *4*, e1000199.

Nandi, S., Mehra, N., Lynn, A.M., and Bhattacharya, A. (2005). Comparison of theoretical proteomes: Identification of COGs with conserved and variable pI within the multimodal pI distribution. BMC Genomics *6*, 116.

Nguyen Ba, A.N., Yeh, B.J., van Dyk, D., Davidson, A.R., Andrews, B.J., Weiss, E.L., and Moses, A.M. (2012). Proteome-Wide Discovery of Evolutionary Conserved Sequences in Disordered Regions. Sci. Signal. *5*, rs1.

Orosz, F., and Ovádi, J. (2011). Proteins without 3D structure: definition, detection and beyond. Bioinforma. Oxf. Engl. *27*, 1449–1454.

Peng, Z., Yan, J., Fan, X., Mizianty, M.J., Xue, B., Wang, K., Hu, G., Uversky, V.N., and Kurgan, L. (2014). Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci. *72*, 137–151.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. Proteins *42*, 38–48.

Schad, E., Kalmar, L., and Tompa, P. (2013). Exon-phase symmetry and intrinsic structural disorder promote modular evolution in the human genome. Nucleic Acids Res. *41*, 4409–4422.

Sharma, A., Dehzangi, A., Lyons, J., Imoto, S., Miyano, S., Nakai, K., and Patil, A. (2014). Evaluation of Sequence Features from Intrinsically Disordered Regions for the Estimation of Protein Function. PLoS ONE *9*, e89890.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., *et al.* (2007). DisProt: the Database of Disordered Proteins. Nucleic Acids Res. *35*, D786–D793.

Sugase, K., Dyson, H.J., and Wright, P.E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature *447*, 1021–1025.

Taylor, J.P., Hardy, J., and Fischbeck, K.H. (2002). Toxic Proteins in Neurodegenerative Disease. Science *296*, 1991–1995.

Tompa, P., and Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. FASEB J. *18*, 1169–1175.

Uversky, V.N. (2002). What does it mean to be natively unfolded? Eur. J. Biochem. FEBS *269*, 2–12.

Uversky, V.N., Gillespie, J.R., and Fink, A.L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins *41*, 415–427.

Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. J. Proteome Res. *6*, 2351–2366.

Vucetic, S., Brown, C.J., Dunker, A.K., and Obradovic, Z. (2003). Flavors of protein disorder. Proteins *52*, 573–584.

Weinreb, P.H., Zhen, W., Poon, A.W., Conway, K.A., and Lansbury, P.T., Jr (1996). NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry (Mosc.) *35*, 13709–13715.

Wright, P.E., and Dyson, H.J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. *293*, 321–331.

Wu, S., Wan, P., Li, J., Li, D., Zhu, Y., and He, F. (2006). Multi-modality of pI distribution in whole proteome. PROTEOMICS *6*, 449–455.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Uversky, V.N., and Obradovic, Z. (2007a). Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions. J. Proteome Res. *6*, 1882–1898.

Xie, H., Vucetic, S., Iakoucheva, L.M., Oldfield, C.J., Dunker, A.K., Obradovic, Z., and Uversky, V.N. (2007b). Functional Anthology of Intrinsic Disorder. 3. Ligands, Post-Translational Modifications, and Diseases Associated with Intrinsically Disordered Proteins. J. Proteome Res. *6*, 1917–1932.