FEBS Letters 589 (2015) 2561–2569

ELSEVIER

FEBS Letters

journal homepage: www.FEBSLetters.org

In various protein complexes, disordered protomers have large per-residue surface areas and area of protein-, DNA- and RNA-binding interfaces

Zhonghua Wu[a], Gang Hu[a], Jianyi Yang[a], Zhenling Peng[b], Vladimir N. Uversky[c,d,e,*], Lukasz Kurgan[f,*]

[a] School of Mathematical Sciences and LPMC, Nankai University, Tianjin, PR China
[b] Center for Applied Mathematics, Tianjin University, Tianjin, PR China
[c] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA
[d] Department of Biology, Faculty of Science, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[e] Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation
[f] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

ARTICLE INFO

Article history:
Received 28 June 2015
Revised 31 July 2015
Accepted 3 August 2015
Available online 20 August 2015

Edited by Wilhelm Just

Keywords:
Intrinsic disorder
Protein–RNA interaction
Protein–DNA interaction
Protein–protein interaction
Binding surface
Binding interface

ABSTRACT

We provide first large scale analysis of the peculiarities of surface areas of 5658 dissimilar (below 50% sequence similarity) proteins with known 3D-structures that bind to proteins, DNA or RNAs. We show here that area of the protein surface is highly correlated with the protein length. The size of the interface surface is only modestly correlated with the protein size, except for RNA-binding proteins where larger proteins are characterized by larger interfaces. Disordered proteins with disordered interfaces are characterized by significantly larger per-residue areas of their surfaces and interfaces when compared to the structured proteins. These result are applicable for proteins involved in interaction with DNA, RNA, and proteins and suggest that disordered proteins and binding regions are less compact and more likely to assume extended shape. We demonstrate that disordered protein binding residues in the interfaces of disordered proteins drive the increase in the per residue area of these interfaces. Our results can be used to predict in silico whether a given protomer from the DNA, RNA or protein complex is likely to be disordered in its unbound form.

1. Introduction

Protein activity is inevitably related to the formation of complexes with other proteins, nucleic acids, membranes and lipids, polysaccharides, and various small molecules. The resulting protein–centric complexes can be stable or transient, can involve one or many binding partners, and can be formed by ordered, partially disordered or completely disordered proteins. It is recognized now that the intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDPRs) play numerous important roles in the assembly and maintenance of protein complexes [1–3]. The broad involvement of IDPs/IDPRs in functionality of protein complexes is related to the overall abundance of IDPs and hybrid proteins (i.e., proteins containing ordered domains and IDPRs [4]) in various proteomes [5–10], their functional complementarity to functions of ordered proteins and domains [11–23], their exceptional binding promiscuity and ability to be involved in various interactions with diverse partners [2,24], their functionality originating from a specific disordered form, from inter-conversion between disordered forms, and from transitions between disordered to ordered or ordered to disordered states [20,21,25–27], their known involvement in regulation and control of protein–protein interaction networks [17,28–32], their exceptional structural heterogeneity [11,33,34], their predisposition for template-dependent folding at binding [2,12,16,18,21], the capability to bind to multiple partners, gaining very different structures in the bound state [16,35], and the ability to possess unrelated, even opposite, functions [36]. As a result, intrinsic disorder plays a number of important roles in organization, maintenance, and control of various protein-based complexes [1–3].

Author contributions: V.N.U. and L.K. conceived the study. J.Y., Z.P., Z.W. and G.H. prepared the data. Z.W. and G.H. carried out the experiments. Z.W., G.H., J.Y., Z.P., V.N.U. and L.K. analyzed the results. Z.W., G.H., V.N.U. and L.K. drafted the paper. All authors read and approved the final manuscript.

* Corresponding authors at: Department of Molecular Medicine, University of South Florida, 12901 Bruce B. Downs Blvd. MDC07, Tampa, FL 33612, USA (V.N. Uversky). Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta T6G 2V4, Canada (L. Kurgan). Fax: +1 (780) 492 1811.

E-mail addresses: vuversky@health.usf.edu (V.N. Uversky), lkurgan@ece.ualberta.ca (L. Kurgan).

http://dx.doi.org/10.1016/j.febslet.2015.08.014
0014-5793/© 2015 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

The overall biological importance of protein-based complexes is reflected in their high prevalence among the PDB entries. Based on the analysis of the mechanisms of dimer formation it has been concluded that, as a first approximation, these simplest protein–protein complexes can be formed via two general mechanisms, depending on the coupling/decoupling of the processes of protein folding and assembly [37–39]. In the three-state scenario, the process of monomer folding precedes the assembly and as a result, the dimer is formed from the independently folded monomers. In the two-state scenario, assembly occurs concomitantly with folding; i.e., the monomers are not folded in the unbound state and the dimer is formed via the binding-induced folding of these intrinsically disordered protomers [37–39]. Using a set of 90 proteins for which experimental data were available in literature, Gunasekaran et al. revealed that the protomers forming two-state complexes can be differentiated from the protomers forming three-state dimers [39]. In that study, comparison of complexes formed by 60 IDPs (44 ribosomal proteins, eleven two-state dimers, and five IDPs for which the crystal structure is known in the complexed state) and 30 ordered proteins (14 three-state dimers and 16 monomeric proteins that are dimers in crystals) suggested that ordered monomers can be distinguished from disordered monomers on the basis of the per-residue surface and interface areas, which are significantly smaller for ordered proteins [39]. We extended that original study by analyzing substantially more complex-forming proteins with known 3D structures (5306 vs. 90). In addition to dimers, which were the major target of the original paper, we included higher order oligomers and complexes of proteins with DNA and RNA. To assure that our protomers were diverse, we reduced their pairwise sequence similarity to below 50%. We analyzed disordered proteins based on structured complexes with proteins, DNA and RNA where the disordered regions underwent disorder-to-order transition. Another new development of our study was the use of the computational annotation of disorder and order, where intrinsically disordered residues in the UniProt chains were identified using a majority vote consensus of five computational tools (three versions of Espritz method [40] that were optimized to predict disorder annotated from X-ray crystal structures, NMR structures, and from the Disprot database [41]; and two versions of IUPred tool [42] that predicts long and short disordered regions). Our analysis provided a strong support to the previously reported observation and clearly showed that in analyzed protein–protein, protein–DNA and protein–RNA complexes, disordered protomers have large per-residue surface areas and per-residue interface areas, suggesting that the known structure of a complex can be used to predict whether a given protein involved in the formation of a complex with DNA, RNA or protein is likely to be disordered in its non-bound monomeric form.

2. Materials and methods

We collected high-quality structures of dissimilar proteins in complex with proteins, DNAs, or RNAs and mapped their sequences into the corresponding complete proteins chains. We annotated intrinsic disorder in the complete chains and selected a subset of structured and disordered monomers. Using structures of complexes we computed surface areas of protein–protein, -DNA, and -RNA interfaces and the overall surface areas of the protein monomers.

2.1. Dataset

We used the BioLiP resource [43] to collect 182 114 proteins that interact with proteins, 6292 that interact with DNAs, and 21 608 with RNAs and to annotate the corresponding binding residues. We combined together interactions with the same ligand type for each protein. Next, we mapped protein sequences from BioLiP into Protein Data Bank (PDB) [44] using PDB identifiers in order to collect the structures and compute surface. We also mapped these proteins into UniProt [45] with SIFTS [46] to obtain complete protein chains that we used to annotate intrinsic disorder. After mapping, the number of protein-binding, DNA-binding, and RNA-binding proteins for which we obtained structures and complete sequences was 163 589, 5913, and 20 731, respectively. Next, we filtered out peptides (chains with less than 30 residues), proteins with lower quality structures (resolution worse than 3 Å), and proteins for which mapping between the PDB sequence and the UniProt sequence was potentially questionable. This mapping is necessary to annotate disorder on the protein surface and for the binding interface. The quality of mapping from PDB to UniProt that was performed with SIFTS was quantified with coverage and identity. The coverage is defined as a ratio of the number of residues from the PDB chain that were successfully mapped into the UniProt sequence and the total number of residues in the PDB chain. The identity is defined as a ratio of identical residues among the residues that we mapped. Only protein chains with coverage and identity both >95% are used. As a result, we retained 114 024, 4177, and 4992 protein-, DNA-, and RNA-binding chains. We further exclude proteins for which binding interfaces are relatively too small to be able to have sufficient amount of data to characterize abundance of disorder in the interface for each protein. We calculated the number of binding residues and content of binding residues (i.e., number of binding residues divided by the total number of residues) for each protein chain. We removed proteins for which the number of binding residues < median that equals 20, 29, and 33 for DNA-, RNA- and protein-binding, and for which content of binding residues <0.05. Consequently, the number of remaining proteins was 55 949 for protein binding, 1744 for DNA binding, and 2611 for RNA binding. Lastly, we extracted a subset of these proteins that share low sequence similarity for each type of ligand. This removes bias toward certain types of folds that could be over-represented in PDB. Protein chains were clustered with BLASTClust at the 50% identity; choice of this cut-off is motivated by an observation that function of proteins is conserved at higher that this sequence identity [47,48]. We selected a representative chain from each cluster that has the highest content of binding residues. The final dataset includes 5306, 213, and 139 protein-, DNA-, and RNA-binding proteins.

2.2. Computational analysis

We annotated intrinsically disordered residues in the UniProt chains using a majority vote consensus of three versions of Espritz method [40] that were optimized to predict disorder annotated from X-ray crystal structures, NMR structures, and from the Disprot database [41]; and two versions of IUPred [42] that predicts long and short disordered regions. A recent assessment demonstrates that these predictors are characterized by good predictive performance [49]. Use of the consensus was motivated by empirical observation that this leads to improved predictive performance when compared to the use of a single method [49–51]. Moreover, this consensus was recently used in several related studies [10,52,53].

For each protein chain we calculated content of binding residues to quantify the size of the DNA-, RNA-, and protein-binding interfaces. We also computed content of disordered residues (i.e., number of putative disorder residues divided by total number of considered residues) for the complete UniProt sequences and the corresponding PDB chains and binding residues. Using these values we extracted a subset of structured and disordered proteins for which the disorder content is very small and relatively large, respectively. In other words, we consider proteins on both extremes of disorder content and separate them by excluding proteins with medium amount of disorder. More specifically, the structured proteins are defined as proteins for which disorder content in the UniProt chain, PDB chain and in binding residues is below 0.05. The disordered proteins have the three disorder content values >0.3 for the protein- and RNA-binding proteins and >0.15 for the DNA-binding proteins; the lower value for the DNA-binding proteins is motivated by low number of proteins that have content >0.3. In total, we include 35 (2649), 28 (30), and 16 (76) disordered (structured) protein-, RNA-, and DNA-binding proteins, respectively. Among the protein-binding proteins, 2448 and 12 are structured protomers interacting with structured and disordered partners, respectively, while 17 and 12 are disordered protomers interacting with structured and disordered partners, respectively.

We calculated solvent accessible surface area for structures of all proteins, structures of the considered three types of ligands, and structures of complexes with RNA, DNA and protein molecules using PyMOL (The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC). Given that surface area of a protein monomer is denoted by $S_{monomer}$, surface area of ligands of the same type that bind to this monomer by $S_{ligands}$, and surface area of the complex with these ligands as $S_{complex}$, the surface area of the interface between this monomer and ligand was calculated as

$$S_{interface} = (S_{monomer} + S_{ligands} - S_{complex})/2$$

We also computed normalized per-residue area of the surface of the protein monomers which is defined as the surface area of a given monomer $S_{monomer}$ divided by the number of residues included in the structure of this monomer. Similarly, we computed the normalized per-residue area of the interface that equals to the surface area of the interface $S_{interface}$ divided by the number of binding residues. This way we quantify packing of the residues on the surface and in the interface. Compared to Ref. [39] where the interface area is normalized by dividing by the total number of residues in the protein, our normalization decouples the relationship of the area with the packing of the binding residues from the influence of the content of the binding residues. In other words, higher (lower) value of our normalized surface area indicates lower (higher) degree of packing of the binding residues, while in case of Ref. [39] it could be due to lower degree of packing and/or larger content of binding residues (higher degree of packing and/or smaller content).

We note that the proteins included in our datasets were crystallized which was necessary to quantify properties of their surface. This means that the disordered proteins were stabilized, likely primarily by their interaction with the ligands, to obtain structures. Although they are all crystallizable, our dataset includes a diverse sets of disordered proteins that share low similarity in their sequences, and thus our conclusions should generalize into all disordered proteins.

3. Results

3.1. Per-residue surface and interface areas

The values of the normalized per-residue surface area and per-residue interface area for the 5658 protein-, DNA- and RNA-binding proteins are shown in Fig. 1A. The per-residue surface area values are in the same range as in [39] that analyzed a smaller set of 90 disordered and structured protein-binding monomers. Our per-residue area of interface is larger since we divided by the number of binding residues while the interface area was normalized by the number of all residues in the corresponding protein in the article by Nussinov and colleagues [39]. However, the values of both types of areas are spread in a similarly wide range when comparing these two articles. The per-residue surface and interface area values are correlated with Pearson Correlation Coefficient (PCC) equal 0.43, 0.56, and 0.77 for protein–protein, protein–DNA and protein–RNA interactions. The high correlation for the RNA-binding proteins can be explained by the high fraction of the binding residues in these proteins that equals 27%, compared to 22% and 14% for the protein-binding and DNA-binding proteins, respectively. We selected proteins with larger than median number of binding residues for each of the three ligands which explains why all three fractions are relatively high. Fig. 1B shows that the per-residue areas of RNA- and protein-binding interfaces are larger than those

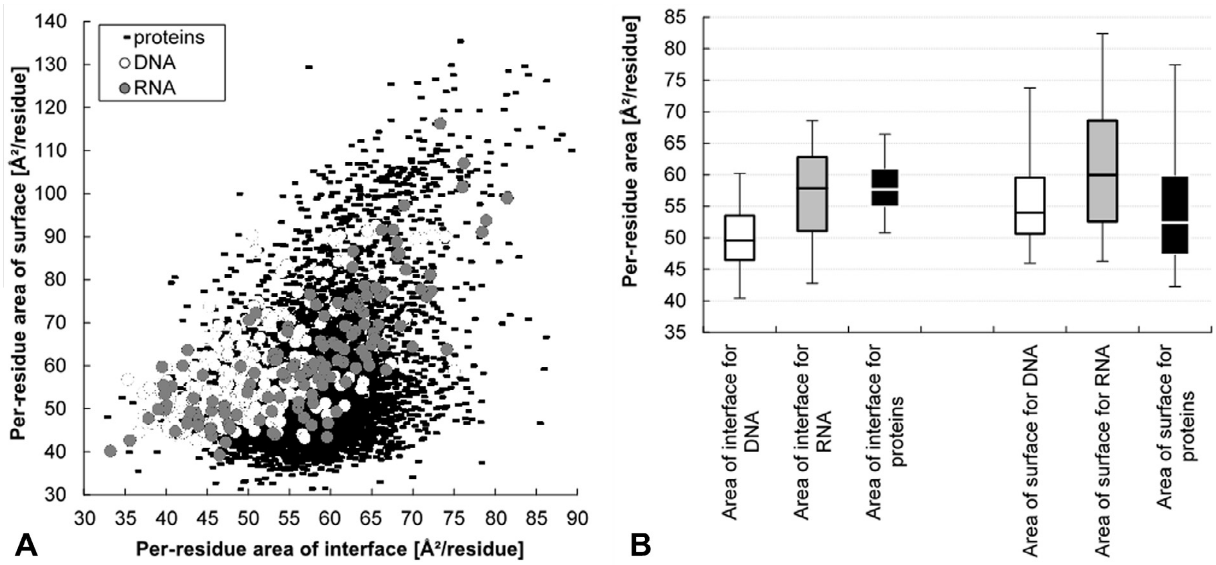

Fig. 1. Normalized per-residue surface area and normalized per-residue interface area for protomers interacting with RNAs, DNAs, and proteins. Panel A shows scatterplot of all values of areas and Panel B summarizes distribution of these values, where white, gray, and black symbols and box denote interactions with DNA, RNA, and proteins, respectively. The box plots show 30th centile, 50th centile (median) and 70th centile, and the two error bars show 10th and 90th centiles of the corresponding distributions of values of the two types of areas.

for the DNA-binding interfaces, while the per-residue surface area of the RNA-binding proteins is larger than that for the other two types of binding. However, in spite of these differences, the values of areas of surface and interface substantially overlap between the three sets of proteins.

We observe that the interfaces of protein complexes with RNA and DNA have lower per-residue area compared to the corresponding surfaces (Fig. 1B), which suggests that these binding regions are more compact, particularly compared to the protein-binding interfaces. Although this is not true across the entire set of the protein-binding proteins, Fig. 2 which analyzes per-residue surface and interface area separately for disordered and structured monomers, reveals that per-residue surface area of the disordered protein-binding proteins is larger than the per-residue interface area. Similarly, the differences between the per-residue surface and interface areas for the DNA- and RNA-binding proteins are driven by the disordered proteins; i.e., the ordered nucleic acid binding proteins have similar per-residues surface and interface area. More discussion of data shown in Fig. 2 are provided below (see Section 3.3.).

3.2. Relation between protein size and surface or interface area

The average disorder content (fraction of disordered residues) of the RNA-, DNA-, and protein-binding proteins (and corresponding interfaces) equals 0.17 (0.21), 0.06 (0.07), and 0.04 (0.04), respectively. This shows that intrinsic disorder is present in the interfaces and that it is more abundant in proteins interacting with nucleic acids. This is in agreement with several studies which observed enrichment of disorder in the DNA- and RNA-binding

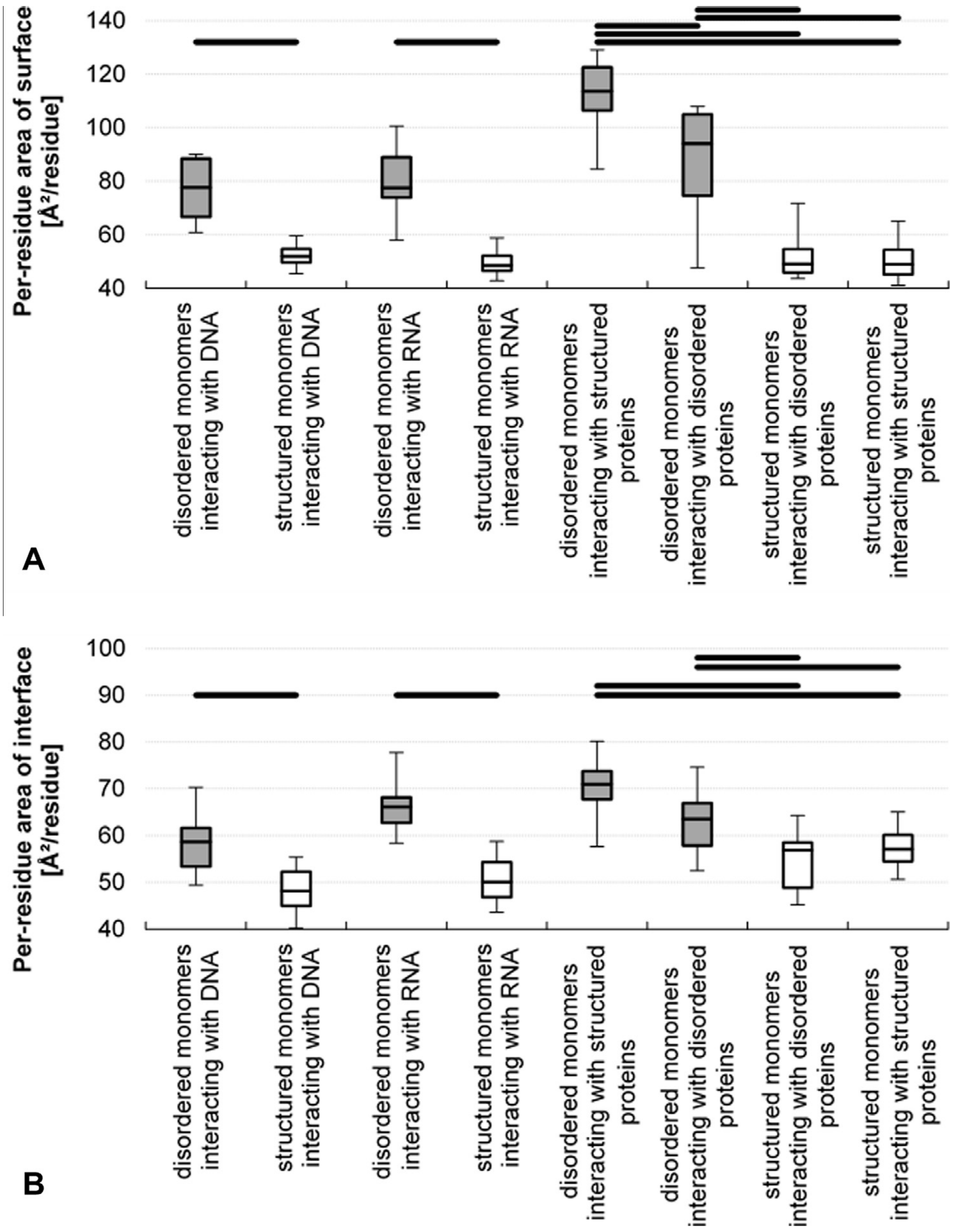

Fig. 2. Normalized per-residue surface area (panel A) and normalized per-residue interface area (panel B) for structured and disordered protein monomers interacting with RNAs, DNAs, and structured and disordered proteins. The box plots show 30th centile, 50th centile (median) and 70th centile, and the two error bars show 10th and 90th centiles of the corresponding distributions of values of areas. Gray (white) denotes disordered (structured) monomers. We compare values of area between structured and disordered proteins that bind to the same type of ligand (DNA, RNA, disordered proteins, and structured proteins). Horizontal bars at the top denote that the disordered and structures protein sets that they connect are characterized by statistically significant different values of surface or interface area (P-value < 0.05).

Table 1
Pearson Correlation Coefficient (PCC) between the area of surface or area of interface and the protein size quantified with the number of residues for the protein-, DNA-, and RNA-binding proteins. We analyze all proteins together and the subsets of disordered and structured monomers.

Scope	Ligand	Considered proteins		
		All	Disordered	Structured
Area of surface	Protein	0.95	0.96	0.95
	DNA	0.98	0.86	0.98
	RNA	0.99	0.94	0.98
Area of interface	Protein	0.42	0.38	0.40
	DNA	0.40	0.26	0.44
	RNA	0.00	0.89	0.58

proteins [53–57]. This also prompted our analysis that compares binding interfaces between structured and disordered proteins.

Table 1 represents values of the Pearson Correlation Coefficient (PCC) between the surface area or interface area and the protein size quantified as the number of residues in a corresponding polypeptide chain. As it could be expected, correlations between the surface area and the protein size are high, which means that larger proteins are characterized by the proportionally increased surface areas. This is consistent across the datasets of protein-, RNA-, and DNA-binding proteins as well as for the corresponding subsets of disordered and ordered monomers. The PCC between the chain length and the surface of the disordered and ordered protein-binding proteins equals at least 0.95, which is consistent with the earlier study reported these coefficients at 0.79 and higher [39].

Interestingly, the correlation coefficients are much lower for the interface areas (Table 1). PCC values for the protein- and DNA-binding proteins are at about 0.4, which indicates that a modest trend is present. Fig. 3A and B show the corresponding scatterplots. They also reveal that disordered proteins (cross-shaped markers) are smaller than structured proteins (circle-shaped markers) but they are all located on the same trend line. This can be seen by the close proximity of the linear fits into these sets (see lines in Fig. 3A and B). The relation between surface area and protein size for the protein–RNA binding proteins is different. Although there is no correlation when considering the complete set of RNA binders, we observe modest and high degree of correlation for the structured and disordered RNA binders, respectively (Table 1). Fig. 3C visualizes these relations. The disordered monomers are again smaller than structured proteins. Furthermore, Fig. 3C shows

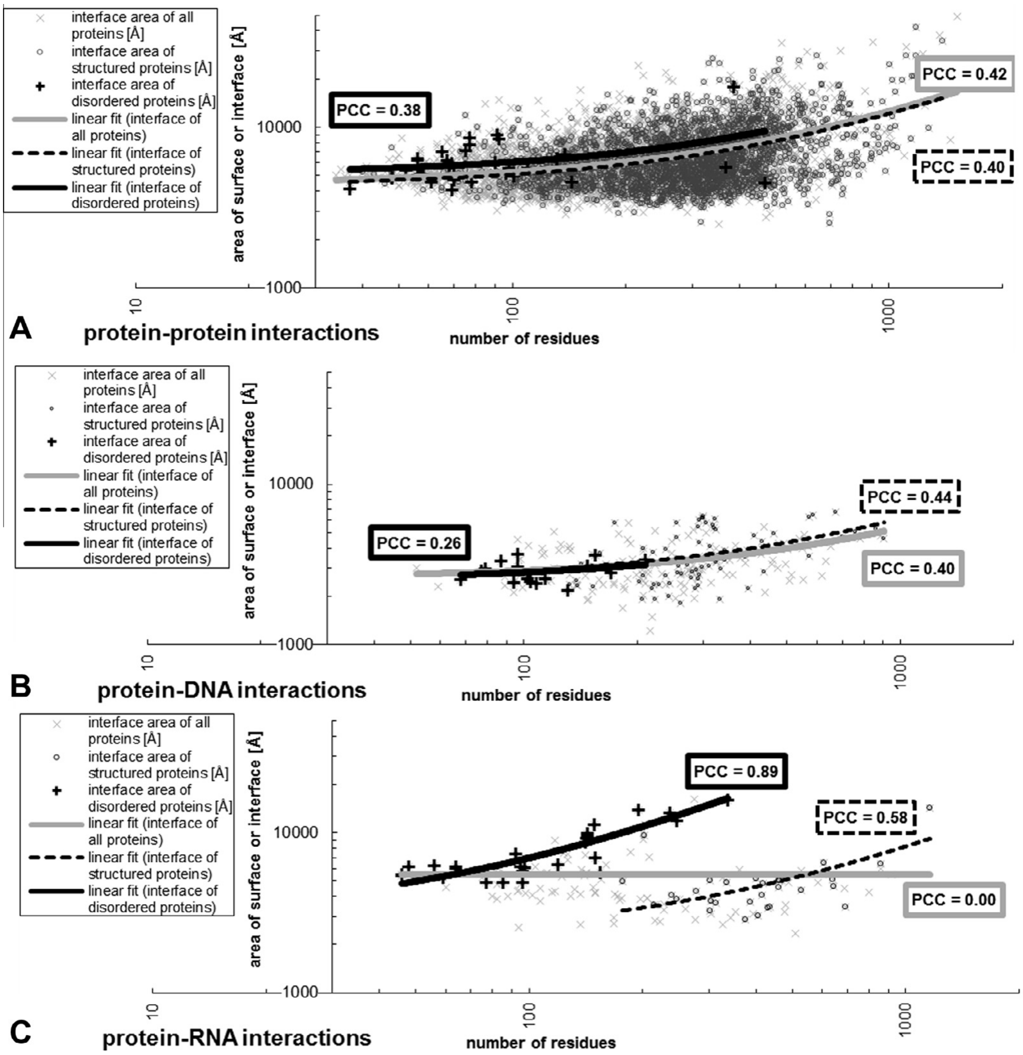

Fig. 3. Relation between the surface area of binding interfaces and number of residues for the protein-binding (panel A), DNA-binding (panel B), and RNA-binding proteins. Both axes are in logarithmic scale. Each panel shows the relations over all considered binders (x-shaped markers) and for the subsets of structured (circle-shaped markers) and disordered (cross-shaped markers) monomers; we note that the set of all binders includes the structured and disordered proteins. Lines correspond to linear fit into the corresponding scatters (sets of proteins) and are accompanied by the corresponding Person Correlation Coefficients (PCCs).

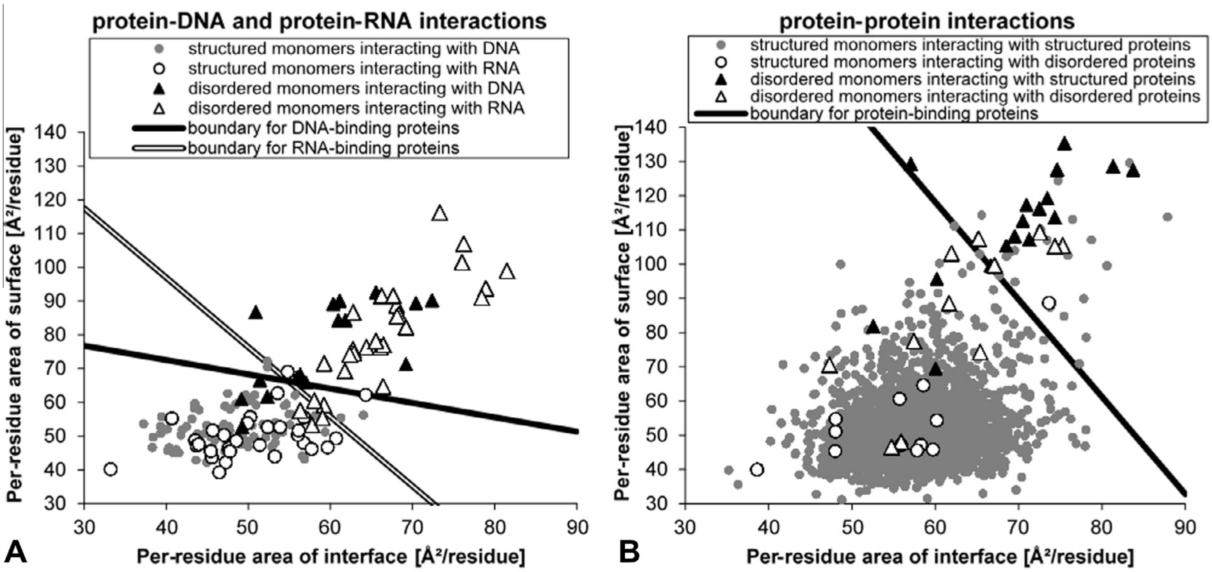

Fig. 4. Normalized per-residue surface area vs. normalized per-residue interface area for structured and disordered protein monomers interacting with RNAs and DNAs (panel A) and with structured and disordered proteins (panel B). Circles and triangles denote structured and disordered monomers, respectively. Hollow and solid markers in panel A denote interactions with RNAs and DNAs, respectively. Hollow and solid markers in panel B denote interactions with disordered and structured proteins, respectively. The lines divide disordered and structured monomers for the DNA-binding (black line in panel A), RNA-binding (hollow line in panel A) and protein-binding (panel B) proteins.

that the surface areas for each of these two protein sets (ordered and disordered) are linearly correlated with the protein sizes while the range of the surface areas for the complete set of RNA binders is similar, irrespective of the protein size.

3.3. Comparison of surface and interface areas between structured and disordered proteins

Since structured and disordered monomers are different in size but their interface areas are similar (Fig. 3), we analyzed their normalized per-residue binding surfaces. Fig. 4 shows relation between the per-residue areas of interfaces and the per-residue areas of surfaces for the protein-, DNA-, and RNA-binding proteins. We compared results between structured and disordered protein monomers and we divided protein–protein interactions into four groups where structured monomers interact with either structured or disordered proteins and disordered monomers interact with structured or disordered proteins (Fig. 4B).

Structured proteins have smaller per-residue areas of surface and interface (circle-shaped markers in Fig. 4) and they are clustered together when compared to the disordered proteins that are distributed over a wide range of values of both normalized areas. This observation is true across the three types of binding and it suggests that structured proteins and their binding interfaces are more compact. The relatively large per-residue interface areas that are linearly correlated with the per-residue surface areas of the disordered monomers (PCC values equal 0.68, 0.84, 0.68 and 0.76 for interactions with DNA, RNA, structured proteins and disordered proteins, respectively) also suggest that these proteins have extended rather than globular shapes and that packing of residues in their interfaces is less dense. We fit lines that divide structured and disordered monomers in the 2-D planes in Fig. 4 using logistic regression [58] based on the implementation in the WEKA platform [59]. These lines provide accurate separation between disordered and ordered monomers where the former (latter) monomers are located above (below) the line. In particular, accuracy of the classification for the DNA-binding proteins based on dividing line equals 93.5%; 12 out of 16 disordered monomers and 74 out of 76 structured monomers are predicted correctly. The accuracy for the RNA-binding monomers is 91.4% with 25 out of 28 disordered and 28 out of 30 structured monomers predicted correctly. Finally, 19 out of 29 disordered and 2438 out of 2460 structured protein-binding proteins are predicted correctly, which corresponds to the accuracy of 98.7%. This demonstrates that the per-residue surface and interface areas of the protein–protein and protein–nucleic acids complexes can be used to accurately determine whether the corresponding monomers are disordered in the unbound state.

We assessed significance of differences of values of the per-residue surface areas and per-residue interface areas between structured and disordered proteins, separately for each type of binding. We used Student's *t*-test given that the corresponding distributions are normal and Wilcoxon test otherwise. A given difference was assumed significant if *P*-value < 0.05. The distribution type was verified using Anderson–Darling test at *P*-value of 0.05. The corresponding distributions of values of these areas and statistical significance are summarized in Fig. 2. The gray (white) box plots and error bars visualize the distributions by giving 10th, 30th, 50th (median), 70th, and 90th centiles for the disordered (structured) protein monomers. The horizontal bars at the top denote that the disordered and structures protein sets that they connect have significantly different values of surface or interface areas. Fig. 2A shows that disordered monomers have significantly higher per-residues area of surface for the DNA-, RNA-, and protein-binding proteins when compared to the structured monomers. The corresponding distributions have little overlap. Fig. 2B reveals that the same is true for the per-residue area of the binding regions. Interestingly, the disordered monomers interacting with structured proteins have larger per-residue area of surface and interface compared to the disordered monomers that bind disordered proteins. It is likely that these differences can be explained by the different morphologies of resulting complexes, where in the former case, the disordered binders wrap around the ordered partners thereby possessing large surface and interface, whereas in the latter case, two disordered partners are typically co-folded to form relatively compact structures. Fig. 5 visualizes structures of selected structured and disordered protein monomers interacting with DNAs, proteins and RNAs. The per residue surface and interface areas are shown in green and blue, respectively, where

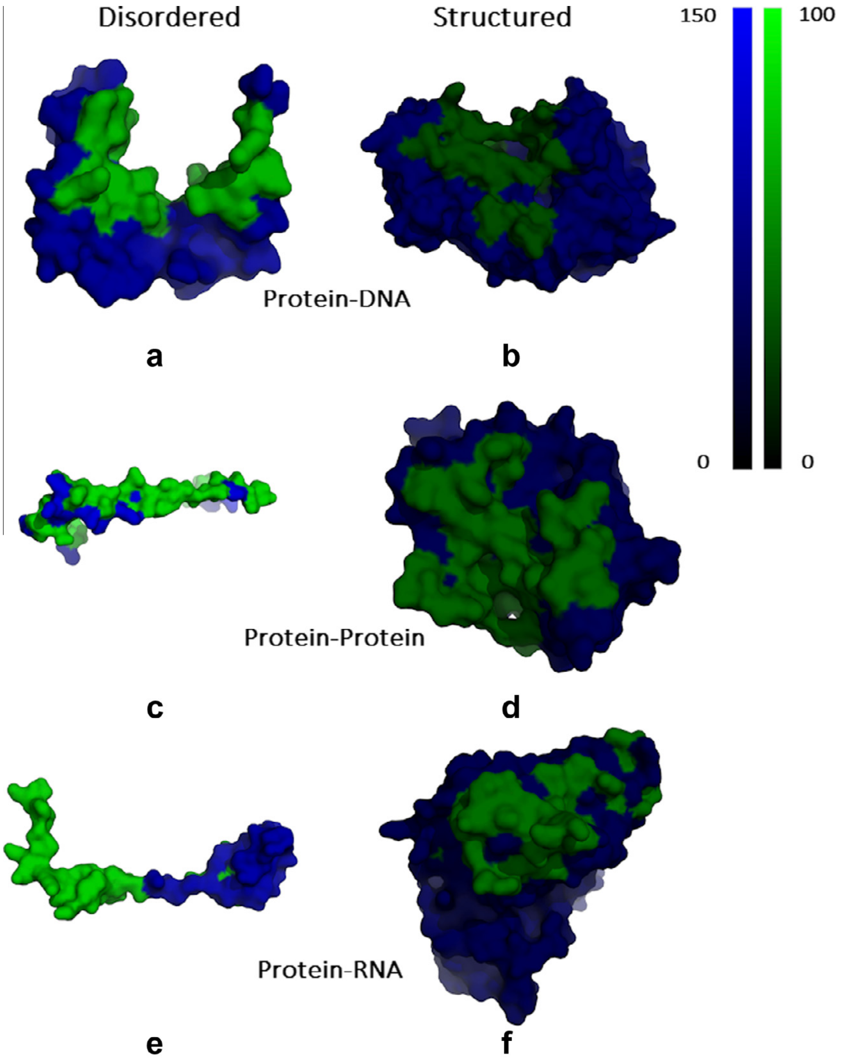

Fig. 5. Structures with color-coded normalized per-residue surface and interface areas for selected structured and disordered protein monomers interacting with DNAs, proteins and RNAs. The interfaces are shown in green and the remainder of the surface in blue. Lighter shade denotes larger per residue area, with the corresponding scale shown in top right corner. Panels a and b are disordered and structured proteins, transcription factor SOX-2 (PDB ID: 1gt0 chain D) and HHAI methylase (PDB ID: 2uz4 chain A), interacting with DNA, respectively. Panels c and d are disordered and structured proteins, calpastatin (PDB ID: 3df0 chain C) and sentrin-specific protease 8 (PDB ID: 2kbr chain A), interacting with proteins, respectively. Finally, panels e and f are disordered and structured proteins, 30S ribosomal protein S5 (PDB ID: 4rbk chain 5) and GTPase Era (PDB ID: 3r9x chain A), interacting with RNA, respectively. Structures were drawn with PyMOL.

lighter shade denotes larger area. In agreement with our conclusions, the normalized interface and surface areas of the disordered proteins are larger compared to the structured proteins (colors are darker for the structured proteins). We also note that structured proteins have globular structures while disordered proteins are elongated and their binding surface suggests that the ligands wrap about them. Although the values of areas for the disordered monomers are significantly larger, we observe continuity of the values between disordered and structured proteins (Fig. 2). This was predicted in [39] for the protein–protein interactions and we show that such continuity is also true for the DNA- and RNA-binding.

3.4. Disordered proteins binding regions

The disordered protein-binding proteins by definition have high disorder content of over 0.3 in the binding interface. We provide further support for our claim that these disordered residues in the interface drive the binding and the increase in the per residue area of the interface when compared with the interfaces of the structured protein monomers. We applied ANCHOR method [60,61] to predict disordered protein binding regions that undergo disorder-to-order transition upon binding to a protein partner. Next, we computed content of the corresponding putative disordered protein-binding residues among the interface residues and in the whole proteins and compared it against the per residue area of the interface for the disordered protein-binding monomers (Fig. 6). The per residue surface area is correlated with the content of the disordered protein-binding residues (PCC = 0.44), which reveals that larger fraction of binding residues results in larger area. This in turn suggests that increase in the normalized area of the surface of the disordered proteins results from the inclusion of a larger number the disordered protein binding residues.

Fig. 6 also shows that content values of disordered protein binding residues in the protein monomers are lower than in their interfaces (average of 0.33 vs. 0.44) and lack correlation with the values of the normalized area of surface (PCC = 0.18). Our calculations show that the content values in the structured protein-binding proteins and their interfaces both are very low and equal 0.003. These results confirm validity of the ANCHOR predictions since the disordered protein binding residues should be enriched in

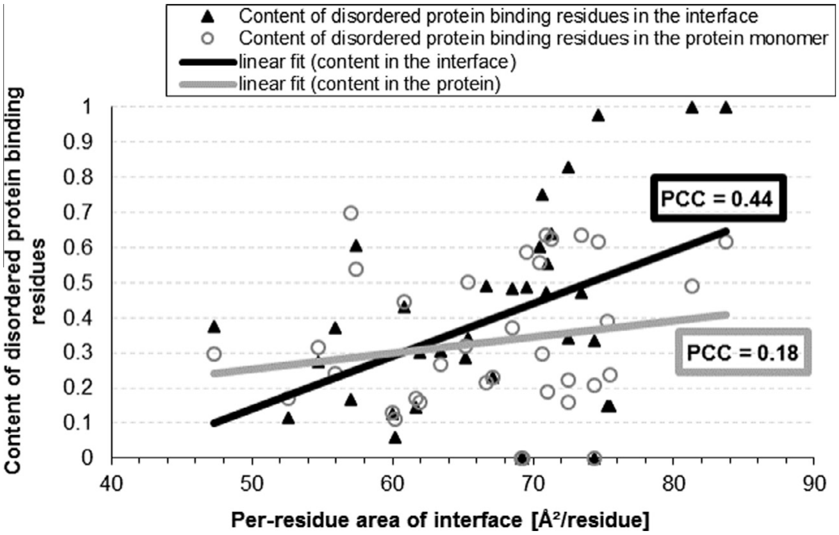

Fig. 6. Relation between the per residue surface area of binding interfaces and content of the disordered protein binding residues in the interface (triangle-shaped markers) and in the entire protein monomer (circle-shaped markers) for the protein-binding proteins. Lines correspond to linear fit into the corresponding scatters (sets of proteins) and are accompanied by the corresponding Person Correlation Coefficients (PCCs).

the interface of the disordered proteins compared to the whole proteins, and should not be present in the interfaces of structured proteins. Moreover, we also computed correlations between the per residues area of the interface and content of the disordered protein-binding residues in the protein monomers that bind RNA and DNA and their interfaces. As expected, these values are low (PCC values range between 0.18 and 0.27) because these interfaces are not for the protein binding.

4. Discussion

This study represents results of the first large scale analysis of the peculiarities of surface and interface areas for over 5600 dissimilar (below 50% sequence similarity) proteins with known 3D structures and involved in interaction with proteins, DNA, and RNAs, with the major focus on delineating differences between binding regions (interfaces) of ordered proteins and IDPs. The analysis of 5306, 213, and 139 protein-, DNA-, and RNA-binding proteins revealed that the size of the protein surface area is highly correlated with protein length when considering disordered, structured, and all binding proteins across protein, DNA and RNA interactions. On the other hand, the size of the interface area is only modestly correlated with the protein size, except for RNA-binding proteins, where larger disordered or structured proteins are characterized by larger binding interfaces. We also show that the IDPs with disordered interfaces are characterized by the significantly larger per-residue surface areas and per-residue interface areas when compared to those of structured proteins. Since these observations are applicable for all the binding proteins; i.e., for all the proteins involved in interaction with other proteins, DNA, and RNA, we can conclude that irrespectively of their binding partners, disordered proteins and binding regions are less compact and more likely to assume extended shape.

Our results can be used to predict, based on the known structure of a complex, whether a given protein involved in the formation of a complex with DNA, RNA or protein is likely to be disordered in its non-bound monomeric form. Our empirical analysis shows that this can be accomplished with high accuracy at well over 90%. Cases for which per-residue surface and interface areas are large and farther from the dividing line (Fig. 4) correspond to proteins for which unbound monomers are more likely to be disordered. On the other hand, proteins with relatively low values of these per-residue areas (about 60 or below), which are closer to the point of origin, are likely to be structured (Fig. 4).

Our results also point to the importance of the intrinsic disorder when analyzing and predicting protein–DNA, protein–RNA, and protein–protein interactions in silico. We show that binding regions are vastly different between structured and disordered monomers, which may require recalibration of the docking protocols and scoring functions (knowledge-based potentials) that are currently used to quantify binding affinity [62]. Some of the new methods that predict sites of protein–protein interactions already utilize information about putative intrinsic disorder [63].

Acknowledgments

This work was funded in part by the Natural Sciences and Engineering Research Council (NSERC) Canada [Discovery Grant 298328] to L.K., by the National Natural Science Foundation of China (NSFC) Grant 11501407 to Z.P., the NSFC Grant 11101226 to Z.W. and G.H., and NSFC Grant No. 11501306 and Nankai University start-up Grant ZB15006103 to J.Y.

References

[1] Fuxreiter, M., Toth-Petroczy, A., Kraut, D.A., Matouschek, A.T., Lim, R.Y., Xue, B., Kurgan, L. and Uversky, V.N. (2014) Disordered proteinaceous machines. Chem. Rev. 114, 6806–6843.

[2] Uversky, V.N. (2013) Intrinsic disorder-based protein interactions and their modulators. Curr. Pharm. Des. 19, 4191–4213.

[3] Uversky, V.N. (2015) The multifaceted roles of intrinsic disorder in protein complexes. FEBS Lett.

[4] Dunker, A.K. et al. (2013) What's in a name? Why these proteins are intrinsically disordered. Intrinsically Disordered Proteins 1, e24157.

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

[6] Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F. and Jones, D.T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337, 635–645.

[7] Uversky, V.N. (2010) The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. J. Biomed. Biotechnol. 2010, 568068.

[8] Schad, E., Tompa, P. and Hegyi, H. (2011) The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 12, R120.

[9] Xue, B., Dunker, A.K. and Uversky, V.N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. J. Biomol. Struct. Dyn. 30, 137–149.

[10] Peng, Z. et al. (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci. 72, 137–151.

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

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

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

[14] Campen, A., Williams, R.M., Brown, C.J., Meng, J., Uversky, V.N. and Dunker, A. K. (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept. Lett. 15, 956–963.

[15] Cortese, M.S., Uversky, V.N. and Dunker, A.K. (2008) Intrinsic disorder in scaffold proteins: getting more from less. Prog. Biophys. Mol. Biol. 98, 85–106.

[16] Oldfield, C.J., Meng, J., Yang, J.Y., Yang, M.Q., Uversky, V.N. and Dunker, A.K. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics 9 (Suppl. 1), S1.

[17] Uversky, V.N., Oldfield, C.J. and Dunker, A.K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J. Mol. Recognit. 18, 343–384.

[18] Dunker, A.K., Brown, C.J., Lawson, J.D., Iakoucheva, L.M. and Obradovic, Z. (2002) Intrinsic disorder and protein function. Biochemistry 41, 6573–6582.

[19] Dunker, A.K., Brown, C.J. and Obradovic, Z. (2002) Identification and functions of usefully disordered proteins. Adv. Protein Chem. 62, 25–49.

[20] Dunker, A.K. et al. (2001) Intrinsically disordered protein. J. Mol. Graph. Model. 19, 26–59.

[21] Uversky, V.N. and Dunker, A.K. (2010) Understanding protein non-folding. Biochim. Biophys. Acta 1804, 1231–1264.

[22] Dunker, A.K. et al. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 9 (Suppl. 2), S1.

[23] Dunker, A.K. and Uversky, V.N. (2008) Signal transduction via unstructured protein conduits. Nat. Chem. Biol. 4, 229–230.

[24] Uversky, V.N. (2011) Multitude of binding modes attainable by intrinsically disordered proteins: a portrait gallery of disorder-based complexes. Chem. Soc. Rev. 40, 1623–1634.

[25] Dunker, A.K. and Obradovic, Z. (2001) The protein trinity-linking function and disorder. Nat. Biotechnol. 19, 805–806.

[26] Uversky, V.N. (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci. 11, 739–756.

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

[28] Patil, A. and Nakamura, H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett. 580, 2041–2045.

[29] Haynes, C. et al. (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput. Biol. 2, e100.

[30] Ekman, D., Light, S., Bjorklund, A.K. and Elofsson, A. (2006) What properties characterize the hub proteins of the protein–protein interaction network of *Saccharomyces cerevisiae*? Genome Biol. 7, R45.

[31] Dosztanyi, Z., Chen, J., Dunker, A.K., Simon, I. and Tompa, P. (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. J. Proteome Res. 5, 2985–2995.

[32] Singh, G.P., Ganapathi, M., Sandhu, K.S. and Dash, D. (2006) Intrinsic unstructuredness and abundance of PEST motifs in eukaryotic proteomes. Proteins 62, 309–315.

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

[34] Uversky, V.N. (2013) Unusual biophysics of intrinsically disordered proteins. Biochim. Biophys. Acta 1834, 932–951.

[35] Hsu, W.L., Oldfield, C., Meng, J., Huang, F., Xue, B., Uversky, V.N., Romero, P. and Dunker, A.K. (2012) Intrinsic protein disorder and protein–protein interactions. Pac. Symp. Biocomput. 2012, 116–127.

[36] Tompa, P. (2005) The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 579, 3346–3354.

[37] Teschke, C.M. and King, J. (1992) Folding and assembly of oligomeric proteins in *Escherichia coli*. Curr. Opin. Biotechnol. 3, 468–473.

[38] Xu, D., Tsai, C.J. and Nussinov, R. (1998) Mechanism and evolution of protein dimerization. Protein Sci. 7, 533–544.

[39] Gunasekaran, K., Tsai, C.J. and Nussinov, R. (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J. Mol. Biol. 341, 1327–1341.

[40] Walsh, I., Martin, A.J.M., Di Domenico, T. and Tosatto, S.C.E. (2012) ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 28, 503–509.

[41] Sickmeier, M. et al. (2007) DisProt: the database of disordered proteins. Nucl. Acids Res. 35, D786–D793.

[42] 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.

[43] Yang, J., Roy, A. and Zhang, Y. (2013) BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucl. Acids Res. 41, D1096–D1103.

[44] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) The Protein Data Bank. Nucl. Acids Res. 28, 235–242.

[45] Apweiler, R. et al. (2014) Activities at the universal protein resource (UniProt). Nucl. Acids Res. 42, D191–D198.

[46] Velankar, S. et al. (2013) SIFTS: structure integration with function, taxonomy and sequences resource. Nucl. Acids Res. 41, D483–D489.

[47] Addou, S., Rentzsch, R., Lee, D. and Orengo, C.A. (2009) Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J. Mol. Biol. 387, 416–430.

[48] Rentzsch, R. and Orengo, C.A. (2013) Protein function prediction using domain families. Bmc Bioinformatics 14.

[49] Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C., Zimmermann, O. and Tosatto, S.C. (2015) Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 31, 201–208.

[50] Fan, X. and Kurgan, L. (2014) Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J. Biomol. Struct. Dyn. 32, 448–464.

[51] Peng, Z. and Kurgan, L. (2012) On the complementarity of the consensus-based disorder prediction. Pac. Symp. Biocomput., 176–187.

[52] Howell, M. et al. (2012) Not that rigid midgets and not so flexible giants: on the abundance and roles of intrinsic disorder in short and long proteins. J. Biol. Syst. 20, 471–511.

[53] Peng, Z., Oldfield, C.J., Xue, B., Mizianty, M.J., Dunker, A.K., Kurgan, L. and Uversky, V.N. (2014) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. Cell. Mol. Life Sci. 71, 1477–1504.

[54] Peng, Z., Mizianty, M.J., Xue, B., Kurgan, L. and Uversky, V.N. (2012) More than just tails: intrinsic disorder in histone proteins. Mol. Biosyst. 8, 1886–1901.

[55] Peng, Z. et al. (2014) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell. Mol. Life Sci.

[56] 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.

[57] Liu, J., Perumal, N.B., Oldfield, C.J., Su, E.W., Uversky, V.N. and Dunker, A.K. (2006) Intrinsic disorder in transcription factors. Biochemistry 45, 6873–6888.

[58] Lecessie, S. and Vanhouwelingen, J.C. (1992) Ridge estimators in logistic-regression. Appl. Stat. J. R. Stat. Soc. Ser. C 41, 191–201.

[59] Frank, E., Hall, M., Trigg, L., Holmes, G. and Witten, I.H. (2004) Data mining in bioinformatics using Weka. Bioinformatics 20, 2479–2481.

[60] Dosztanyi, Z., Meszaros, B. and Simon, I. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25, 2745–2746.

[61] Meszaros, B., Simon, I. and Dosztanyi, Z. (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. 5, e1000376.

[62] Fornes, O., Garcia-Garcia, J., Bonet, J. and Oliva, B. (2014) On the use of knowledge-based potentials for the evaluation of models of protein-protein, protein-DNA, and protein-RNA interactions. Adv. Protein Chem. Struct. Biol. 94, 77–120.

[63] Aumentado-Armstrong, T.T., Istrate, B. and Murgita, R.A. (2015) Algorithmic approaches to protein–protein interaction site prediction. Algorithms Mol. Biol. 10, 7.