Intrinsic Disorder in Human RNA-binding Proteins

Bi Zhao[1], Akila Katuwawala[1], Christopher J. Oldfield[2], Gang Hu[3], Zhonghua Wu[4], Vladimir N. Uversky[5], Lukasz Kurgan[1*]

[1]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA
[2]Department of Microbiology and Immunology, Virginia Commonwealth University, Richmond, VA 23284, USA
[3]School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin 300071, China
[4]School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China
[5]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

[*]corresponding author: lkurgan@vcu.edu

Abstract

Although RNA-binding proteins (RBPs) are known to be enriched in intrinsic disorder, no previous analysis focused on RBPs interacting with specific RNA types. We fill this gap with a comprehensive analysis of the putative disorder in RBPs binding to six common RNA types: messenger RNA (mRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), non-coding RNA (ncRNA), ribosomal RNA (rRNA), and internal ribosome RNA (irRNA). We also analyze the amount of putative intrinsic disorder in the RNA-binding domains (RBDs) and non-RNA-binding-domain regions (non-RBD regions). Consistent with previous studies, we show that in comparison with human proteome, RBPs are significantly enriched in disorder. However, closer examination finds significant enrichment in predicted disorder for the mRNA-, rRNA- and snRNA-binding proteins, while the proteins that interact with ncRNA and irRNA are not enriched in disorder and the tRNA-binding proteins are significantly depleted in disorder. We show a consistent pattern of significant disorder enrichment in the non-RBD regions coupled with low levels of disorder in RBDs, which suggests that disorder is relatively rarely utilized in the RNA-binding regions. Our analysis of the non-RBD regions suggests that disorder harbors posttranslational modification sites and is involved in the putative interactions with DNA. Importantly, we utilize experimental data from DisProt and independent data from Pfam to validate the above observations that rely on the disorder predictions. This study provides new insights into the distribution of disorder across proteins that bind different RNA types and the functional role of disorder in the regions where it is enriched.

1 Introduction

The RNA-binding proteins (RBPs) are involved in a wide spectrum of cellular functions including regulation of gene expression, post-transcriptional regulations, and protein synthesis [1, 2]. RBPs interact with several different types of RNAs including messenger RNA (mRNA), transfer RNA (tRNA), small nuclear RNA (snRNA), non-coding RNA (ncRNA), ribosomal RNA (rRNA), and internal ribosome RNA (irRNA) [3-5]. Many techniques have been developed and utilized to identify and characterize interactions between RBPs and RNAs. The X-ray crystallography and nuclear magnetic resonance produce structures of protein-RNA complexes that are used to study these interactions and the interacting interfaces at the atomic level [6-8]. However, these structures are available for a relatively few RBPs. Other methods, such as next-generation sequencing and protein mass spectrometry, are applied to study

protein-RNA interactions at the coarser sequence level. They are used on a large scale to identify binding regions in the sequences of RBPs and their corresponding transcripts [3, 4, 9-11]. However, they do not provide information about structural states of these sequences. Interestingly, sequences of RBPs include ordered regions that fold into well-defined tertiary structures and intrinsically disordered regions (IDRs) that lack well-defined 3-D structure, are highly flexible and form heterogeneous ensembles of interconverting conformations [12-15]. Recent studies suggest that proteins with IDRs are common in nature [16-18] and that RBPs are significantly enriched in IDRs [19-22]. The IDRs in RBPs were shown to carry out many functional roles [20, 23-26]. For instance, IDRs are prone to include posttranslational modification (PTM) sites that facilitate regulation of protein-RNA interactions [20, 23, 26]. They also contribute to the formation of ribonucleoprotein granules and ribosomal assembly through their involvement in the underlying protein-protein interactions [20, 23, 25-27]. Furthermore, IDRs serve as flexible linkers that enable cooperative interactions between multiple RNA-binding regions [25, 28]. Finally, there are also examples of IDRs that directly interface with the interacting RNA [25].

While the broad family of RBPs was already shown to be enriched in IDRs [19-22], studies that investigate RBPs interacting with specific RNA types are lacking. We only found works focusing on protein-mRNA interactions which demonstrate that these RBPs are also highly disordered [24, 29]. However, the amount and potential enrichment of disorder in RBPs that interact with the other RNA types is still an open question. We also note the lack of a systematic genome-wide analysis of functions of IDRs in RBPs, particularly examining presence of disorder in the RNA-binding domains (RBDs), which are the sequence regions that directly interface with RNA. To this end, we quantify the amount and enrichment of predicted disorder in RBPs that interact with the six main types of RNAs: mRNA, tRNA, snRNA, ncRNA, rRNA, and irRNA, in the human proteome. We also investigate functions of disorder in RBPs by analyzing presence and amount of putative disorder in RBDs and in the other sequence regions of human RBPs. This way, we aim to answer an intriguing question whether the disorder is directly implicated in the protein-RNA interaction or whether it facilitates other functions of RBPs. Moreover, we use experimental disorder annotations to validate and confirm these results. Our approach follows past studies that similarly rely on disorder predictions that are often supported by a smaller scale analysis of the experimental data [20-22, 24].

2 Materials and Methods

2.1 Datasets

We study disorder in RBPs in the human proteome. We select this proteome because of the high coverage by the experimental annotations of disorder (i.e., the most populated organism in the DisProt database [30]) and RBPs [3-5], and very high completeness level; i.e., its BUSCO (Benchmarking Universal Single-Copy Orthologs) score is 99.5% [31]. We collect human proteome from version 2021_01 of UniProt database (Proteome ID: UP000005640). We exclude protein fragments and peptides by removing sequences annotated in UniProt with term “Fragment” or having < 30 amino acids. Next, we annotate RBPs in the remaining set of human protein by collecting the RBP encoding genes released across three recent large-scale studies of human RBPs [3-5]. After mapping these three gene sets to the UniProt accession numbers we manage to annotate 1,544 RBPs. This number is in line with a recent study that estimated the number of human RBPs to be 1,511 via a comprehensive computational prediction [32]. Next, we utilize the annotations of RBDs, which are the sequence elements that directly bind RNAs [33, 34], to further screen the list of human RBPs. Importantly, the knowledge of the location of RBDs is necessary for us to study the disorder functions based of the potential co-location of IDRs and RBDs. We identify RBDs among the 1,544 proteins using version 33.1 of the Pfam database [35], resulting in the

final set of 1,112 RBPs that were identified as human RBPs in recent studies [3-5] and have at least one RBD. Furthermore, we analyze potential impact of the removal of 1,544 – 1,112 = 423 RBPs that lack annotations of RBDs. Using the data introduced in Section 2.2, we find no statistically significant difference in the amount of disorder between these 423 proteins and the 1,112 RBPs, for which we identified RBDs. This suggests that the use of the 1,112 RBPs, which facilitates examination of the co-location of IDRs and RBDs, should not affect our analysis of the disorder enrichment and functions. We divide the 1,112 RBPs into subgroups based on the type of RNAs that they interact with, which we collected from the source studies [3-5]. We remove the subgroups with low counts of RBPs (≤30 proteins) since their numbers would be insufficient to perform reliable statistical analysis. Consequently, we annotate RBPs that interact with six RNA types: mRNA (480 RBPs), irRNA (148), tRNA (121), rRNA (82), ncRNA (80), and snRNA (54).

2.2 Annotation of Intrinsic Disorder, Posttranslational Modifications Sites, DNA-binding Residues, and Pfam Domains

We collect experimentally annotated disorder from the DisProt database [30]. We secure disorder annotations for 57 RBPs based on mapping human proteins included in DisProt. This dataset is available in the **Supplementary Dataset S1**. Given the sparsity of the experimental data, we also collect putative disorder annotations for the entire human proteome. The experimental annotations are used to validate the proteome-scale result that rely on the disorder predictions. The use of the disorder predictions is motivated by fact that they were shown to be very accurate [36-40], particularly when applied to the nucleic acid binding proteins [41]. Numerous recent studies of disorder function and abundance similarly rely on the disorder predictions [20-22, 24, 27, 42-47]. Arguably the most accurate option is to utilize consensus-based predictors, defined as the methods that combine results produced by multiple "base" disorder predictors. The consensus prediction is designed to provide higher predictive performance compared to the base predictors used individually [48-50]. Two databases that provide convenient access to several disorder predictions are available: D^2P^2 [51] and MobiDB [52, 53]. We use the pre-computed consensus disorder predictions from the larger and more up-to-date MobiDB database [54]. This consensus utilizes majority vote approach to combine results generated by a comprehensive collection of ten disorder predictors: PONDR-VSL2B [55], GlobPlot [56], JRONN [57], two versions of DisEMBL that predict hot loops and disordered regions identified using X-ray structures [58], two versions of IUPred for predicting short and long IDRs [59, 60], and three versions of ESpritz that were developed using annotations of disorder generated form the X-ray structures, NMR structures, and using data from the DisProt database [61]. Our approach improves over the related studies that analyze enrichment of disorder in the RNA-binding proteins using either a single disorder predictor [20, 21, 24] or a consensus of five methods [22].

We also collect functional features that are potentially associated with the presence of disorder including the PTM sites as well as DNA-binding and protein-binding regions. These annotations can be collected at the residue level (i.e., for each amino acid in the protein sequence) and as the domain level (i.e., for protein domains). The residue and domain level annotations are derived using fundamentally different approaches. This allows us to cross check whether the amount and enrichment of disorder in RBDs, DNA-binding domains and PTM sites generated based on the residue-level data agree with the results derived from the domain-level data.

We use the D^2P^2 database to collect experimental residue-level annotations of the PTM sites that were originally sourced from the PhosphoSitePlus resource [62]. Besides utilizing PTM sites collectively, we analyze disorder for the four most numerous PTM types: phosphorylation, ubiquitination, acetylation and methylation, which are in sufficient numbers to perform reliable statistical analysis. We also collect

putative annotations of the disordered residues that bind DNA and that bind proteins that we generate with the DisoRDPbind predictor [63, 64]. This is the only tool capable of producing these residue-level annotations. DisoRDPbind was shown to produce accurate predictions [40], is sufficiently fast to process the human proteome, and its pre-computed results can be conveniently collected from the DescribePROT database [65].

We rely on Pfam to collect the domain-level data. We combine the Pfam clan data (Pfam version 33.1) [35] and Gene Ontology (GO) data from Pfam2GO (version 2020_06) [66] to identify and categorize domains in RBPs. We categorize these domains into four types: RBDs, DNA-binding domains, domains with PTM sites (PTM domains), and other domains. We derive annotations of the PTM domains using the function description in the Pfam clan data, which lists specific PTM types. We do not annotate the protein-binding domains since the residue-level results do not suggest enrichment in the disorder. The residue and domain-level annotations of the PTMs, putative disorder and DNA binding for the RBPs and the other human proteins are available in the **Supplementary Datasets S2** for the human RBPs and **Supplementary Datasets S3** for the other human proteins.

2.3 Computational and Statistical Analysis

We quantify the amount of disorder in a given protein or protein region/domain with the disorder content, which is defined as the number of disordered residues in that protein/region divided by the length of the sequence of that protein/region. We also compute the content of DNA-binding and protein-binding residues and domains, and PTM sites and domains in disordered regions (i.e., the number of PTM sites in disordered regions divided by the number of disordered residues in the disordered regions) to investigate potential cellular functions of disordered regions.

We investigate enrichment of disorder by assessing significance of differences in the disorder content computed for RBPs and subgroups of RBPs that interact with specific RNA types against corresponding reference disorder content values computed for the other human proteins. We similarly study enrichment of DNA-binding domains/residues, protein-binding residues, PTM domains/residues in disordered regions located in non-RBD regions, which are defined as sequence regions in RBPs that exclude RBDs. We assess this enrichment for RBPs and the subgroups of RBPs by evaluating significance of differences when comparing the corresponding binding/PTM content values against reference content computed for the other human proteins. The reference human proteins and regions are selected at random and length-matched to the length of the RBPs, RBDs, and disordered regions in non-RBD regions. The length-matching is motivated by the fact that sequences of RBPs are on average substantially longer than the sequences of other human protein (**Supplementary Table S1**). This accommodates for the known bias in the disorder content across proteins that have different sequence length [67]. Similar procedure was used in several recent studies [42, 45, 68, 69]. We perform these tests based on bootstrapping 50% of the proteins/domains 100 times. We compare the corresponding content measurements using the student *t*-test if the data are normal; otherwise, we use the Wilcoxon signed-rank test. We test normality with the Anderson-Darling test at the 0.05 significance.

3 Results

3.1 Putative Intrinsic Disorder is Enriched for Some of the RNA-binding Proteins

We quantify the amount and enrichment of the putative disorder in all human RBPs and the RBPs that interact with several specific RNA types. Figure 1 shows distributions of the disorder content for proteins, domains, and regions outside the domains in the human proteome, the complete set of 1,112 RBPs, the

RBPs that interact with mRNA, irRNA, tRNA, rRNA, ncRNA, and snRNA and for their corresponding human reference sets. We summarize the corresponding numeric results in the **Supplementary Table S1**. The median putative disorder content in the human proteome is 0.13 (**Figure 1(a)**), which agrees with recent studies that reported the disorder content of 0.15 [70] and 0.13 [71]. The median disorder content for RBPs is over 30% higher and equals 0.17 (**Figure 1(b)**). We find that this enrichment is statistically significant (p-value < 0.05) and consistent with studies, which similarly showed that RBPs are significantly enriched in the intrinsic disorder [19, 20].

Next, we investigate whether the enrichment in the putative disorder content is consistent across RBPs that interact with specific types of RNAs. **Figure 1(b)** provides distributions of the protein-level putative disorder content in the human RBPs, the six common types of RBPs, and their corresponding human reference sets. The median putative disorder content in the most disorder-enriched and the largest subgroup of RBPs that interact with mRNA is 0.25 (**Supplementary Table S1**), which nearly doubles the median disorder content in the human proteome. This is in line with the published studies that similarly find that mRNA-binding proteins are enriched in disorder [24, 29]. However, we are the first to study the enrichment for the other five types of RBPs that bind ncRNA, irRNA, rRNA, snRNA, and tRNA. We find that RBPs that interact with rRNA and snRNA are also significantly enriched in disorder (p-value<0.05). However, the putative disorder content is on par with the expected/proteome-wise content values for the ncRNA-binding (content = 0.11) and irRNA-binding (content = 0.13) proteins and is actually significantly depleted in RBPs that bind tRNA (content = 0.08; p-value < 0.05). Our analysis reveals that putative intrinsic disorder is unevenly distributed across different subgroups of RBPs. To summarize, we show that the overall statistically significant enrichment of RBPs in putative disorder is driven by the significant enrichment for the mRNA-, rRNA- and snRNA-binding proteins, while the protein-level disorder content for the tRNA-binding proteins is significantly depleted.

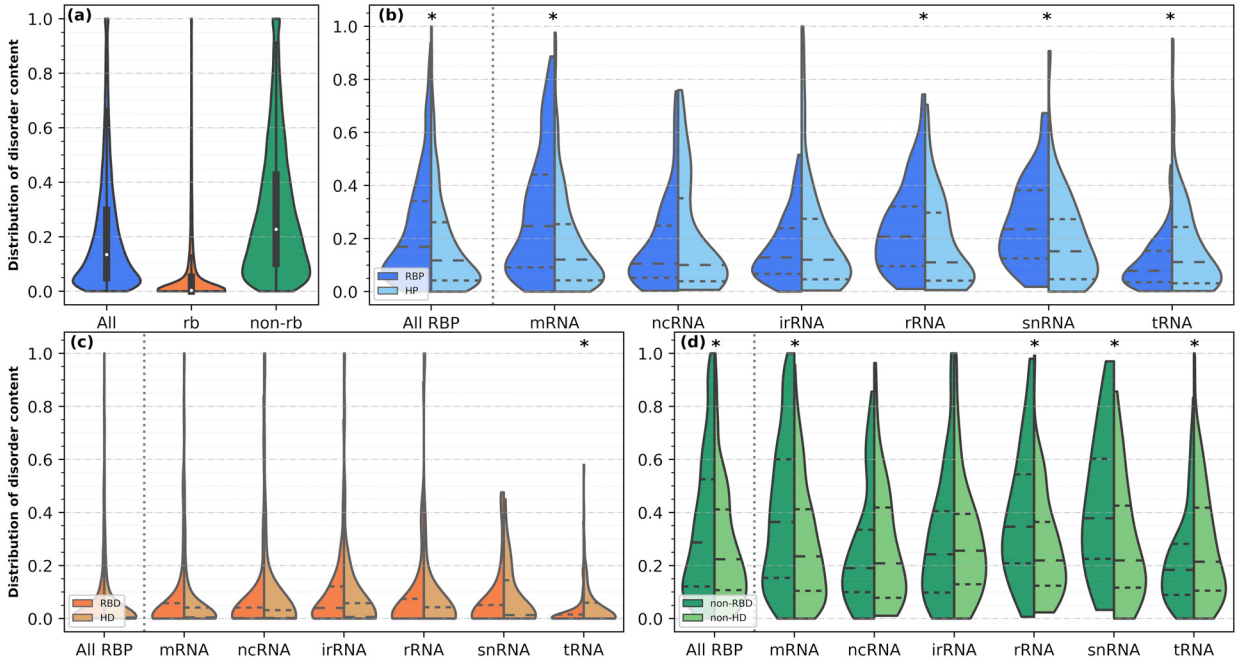

Figure 1. Distributions of the putative disorder content in the human proteome, human RBPs, and across subgroups of RBPs that bind specific RNA types. The blue/orange/green violin plots in panel (a) show the distribution of the protein-level disorder content calculated across human proteome (HP)/Pfam domains in human proteome (HD)/non-domain regions in human proteome (non-HD). The box-plots inside the violin plots give the 95 percentile, 75 percentile, median, 25 percentile, and 5 percentile values for a given distribution. The dark-shaded parts of the violin plots show the distribution of the putative disorder

content for RBPs and the six RBP types calculated at the protein level in panel (b), for RBDs in panel (c), and for non-RBD regions (i.e., the sequence regions in RBPs that exclude RBDs) in panel (d). The light-shaded parts of these violin plots give the putative disorder content for the corresponding reference human data, which we sample to equalize the sequence length. The dashed lines inside the violin plots denote the 75 percentile, median, and 25 percentile values for a given distribution. "*" at the top of a given plot denotes that a given distribution is significant different compared to the corresponding sampled reference human proteome distribution (p-value<0.05). We detail the statistical tests in Section 2.3.

3.2 Putative Intrinsic Disorder is Not Enriched in the RNA-binding Domains

While the putative intrinsic disorder is significantly enriched in the mRNA-, rRNA- and snRNA-binding proteins, it is unclear whether it is present in RBDs. We analyze the putative disorder content in RBDs in the complete set of RBPs and in each of the six subgroups of RBPs (**Figure 1(c)** and **Supplementary Table S1**). The median of the putative disorder content values in RBD across all RBPs equals 0.0. The median disorder content in RBD across the RBPs that bind different types of RNA is also 0.0 for all RNA types, except for irRNA where the content is also low and equals 0.04. We quantify significance of the differences between the disorder content of RBDs and the domains in a corresponding sampled (length-matched) reference set of human proteins. We find that the putative intrinsic disorder of RBDs in the complete set of RBPs and for each of the six subgroups of RBPs is not significantly enriched when compared to the reference disorder content computed for domains in human proteins that exclude RBPs and RBD regions (p-value > 0.05). This result combined with the overall enrichment of putative disorder in RBPs suggests that disorder is localized outside of RBDs, which we dub non-RBD regions. Indeed, our empirical analysis of the putative disorder content in the non-RBD regions of RBPs shows that their median disorder content is high and equals 0.29 and is statistically significantly higher than the putative disorder content in RBDs (p-value < 0.05).

3.3 Enrichment of the Putative Intrinsic Disorder in the non-RBD Regions is Consistent with the Enrichment for the RNA-binding Proteins

Figure 1(d) shows the distributions of the putative disorder content for the non-RBD regions in RBPs and in subgroups of RBPs that interact with specific RNA types. The putative median disorder content for the non-RBD segments in RBPs equals 0.29, compared to the median content of 0.22 for the reference non-domain regions in the human proteome (**Supplementary Table S1**). The difference between these content measurements is statistically significantly (p-value < 0.05). The putative disorder content values for the non-RBD regions across the six subgroups of RBPs vary between 0.18 (tRNA-binding proteins) to 0.38 (snRNA-binding proteins). Our analysis suggests that the enrichment of putative disorder in the non-RBD regions for the mRNA-, rRNA- and snRNA-binding proteins is significant when compared to the corresponding human reference set (p-value < 0.05). In contrast, the putative disorder is significantly depleted for the tRNA-binding proteins (p-value < 0.05), and the difference in the disorder content is not significant for the ncRNA- and irRNA-binding proteins. These results are consistent with the findings for the full sequences of RBPs, suggesting that putative disorder is primarily located in the non-RBD sequence segments.

3.4 Patterns in the Enrichment of the Putative Intrinsic Disorder are Consistent with the Enrichment of the Experimentally Annotated Disorder

The above analysis relies on the putative disorder produced by an accurate consensus prediction that we collect from the popular MobiDB resource [52, 54]. We validate these results based on the analysis of the experimental disorder that we collect from DisProt for a subset of 57 RBPs [30]. We compare the results for the generic set of RBPs. The relatively small size of the experimental dataset prevents us from performing this analysis across the six subgroups of RBPs. We observe that the native median disorder content in the RBDs and in the non-RBD regions across the 57 RBPs is 0.03 and 0.35, respectively. The

difference in the disorder content in RBDs against in non-RBD regions is statistically significantly (p-value<0.05), which demonstrates that the native disorder in enriched in the sequences of RBPs outside of the RNA-binding regions. These values are very similar to the median content for the complete set of 1,112 RBPs computed using the putative disorder, which are 0.00 and 0.29, respectively, and which are also significantly different (p-value < 0.05). Further analysis using the putative disorder for the same set of 57 RBPs provides consistent results, with the median disorder content in RBDs of 0.00 and in non-RBD regions of 0.41, where the corresponding difference is statistically significant (p-value < 0.05). The increase in the putative disorder content for the 57 RBPs from DisProt compared to the content for the complete set of 1,112 RBPs can be explained by the fact that DisProt specifically focuses on the proteins that have IDRs. To sum up, our analysis of both putative and experimental disorder annotations reveals a consistent pattern of the statistically significant disorder enrichment in the non-RBD regions and low amounts of disorder in RBDs. This motivates us to investigate potential functions of disorder in the non-RBD regions of the RBP sequences.

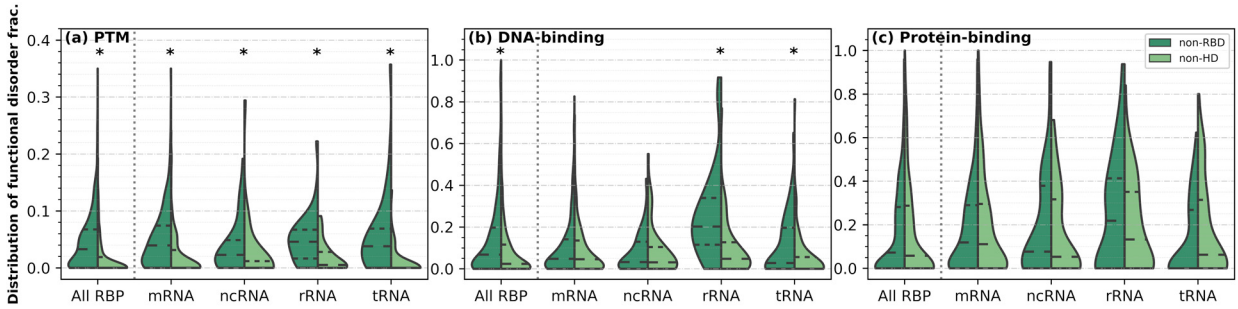

Figure 2. Distributions of the content of experimental PTM sites (panel (a)), putative DNA-binding residues (panel (b), and putative protein-binding residues (panel (c)) in the putative disordered regions located in non-RBD regions in the human RBPs and across several subgroups of RBPs that bind specific RNA types. We exclude the disordered non-RBD regions < 10 residues long and cover the four subgroups of RBPs that have at least 30 proteins with these disordered regions to ensure that statistical analysis is robust. The dark-shaded parts of the violin plots show the distribution of the content values in non-RBD regions for RBPs and the four RBP types that satisfy the above criteria. The light-shaded parts of the violin plots show the content values for the corresponding non-domain regions in the reference human set (non-HD), which we sample to equalize the sequence length. The dashed lines inside the violin plots give the 75 percentile, median, and 25 percentile values for a given distribution. "*" at the top of a given plot denotes that a given distribution is significant different compared to the corresponding sampled reference human proteome distribution (p-value<0.05). We detail the statistical tests in Section 2.3.

3.5 Putative Intrinsic Disorder in non-RBD Regions Facilitates DNA-binding and Hosts PTM sites

We investigate functions of disorder in the non-RBD regions in RBPs by evaluating potential enrichment of the content of the putative DNA-binding and protein-binding residues and experimental PTM sites among intrinsically disordered regions located in the non-RBD regions in RBPs. **Figure 2** provides distributions of the protein-level content of these three types of functional residues localized in disordered regions in RBPs, in the mRNA-, ncRNA-, rRNA- and tRNA-binding proteins, and in the human proteome that we use as a reference set. We limit our analysis to the four types of RBPs that include sufficient amount of data to warrant robust statistical analysis (i.e., at least 30 proteins that have sufficient amount of disorder, as we detail in the caption for Figure 2). **Supplementary Table S2** provides the corresponding median values, including the medians for each length-matched reference content based on the human proteome.

The putative DNA-binding residues constitute about 7% of disordered regions located in non-RBD regions and they are significantly enriched compared to the reference proteome-level data (p-value<0.05). The analysis over the four RNA types shows significant enrichment for the RBPs that interact with rRNA

and tRNA where putative DNA-binding residues (**Figure 2(b)**) account for 3.2% (tRNA-binding proteins) and 20.2% (rRNA-binding proteins) of their disordered regions (p-value<0.05). In contrast, **Figure 2(c)** reveals that putative protein-binding residues are not enriched among the disordered residues, when compared to the reference data (p-value > 0.05). They account for 7.2% of the putative disordered residues in the non-RBD regions in RBPs, when compared to their content of 5.8% among the putative disordered regions in the corresponding human reference set. However, the PTM sites (**Figure 2(a)**), which comprise 3.3% of the putative disordered regions in the non-RBD segments, are significantly enriched and this holds true for the mRNA-, ncRNA-, rRNA- and tRNA-binding proteins (p-value < 0.05).

While **Figure 2(a)** shows the per-protein distribution of content across all PTMs, we also breakdown this analysis by several major types of PTMs. Given the scarcity of the PTM sites, we reanalyze the data at the residue-level. We subsample 10% of the disordered residues in the non-RBD regions for RBPs and for each of the four subgroups of RBPs and calculate content of specific PTM types among these residues. We repeat that 100 times and compare the resulting distributions of the content values to the reference data computed in the same way in the human proteome. We compare the resulting residue-level content values for different types of PTMs including phosphorylation, ubiquitination, acetylation, and methylation among the disordered regions in the non-RBD segments in **Table 1.** We find that the residue-level analysis is consistent with the protein-level analysis and reveals that PTM sites in the disordered regions that exclude RNA-binding domains are significantly enriched in RBPs (p-value < 0.05). Moreover, we show that the four major types of PTM sites (phosphorylation, ubiquitination, acetylation and methylation) are consistently significantly enriched in the disordered regions localized in non-RBD segments across all groups of RBPs; the only exception are the methylation sites that lack enrichments for tRNA-binding proteins.

Table 1. Residue-level content of PTMs and specific types of PTMs in putative disordered regions in the non-RBD segments in human RBPs and across several subgroups of RBPs that bind specific RNA types. We subsample 10% of the disordered residues in the non-RBD regions for RBPs and for each of the four subgroups of RBPs and calculate content of specific PTM types among these residues. We repeat that 100 times and compare the resulting distributions of the content values to the reference data computed in the same way in the human proteome. We report median and the 75th percentile of the PTM content for each considered protein set. The bold and red font represents significant enrichment in PTMs when compared to the corresponding reference set (p-value<0.05). We detail the statistical tests in Section 2.3.

RNA type	Median *(75 percentile)* of the content of PTM sites in putative disordered non-RBD regions				
	All PTMs	Phosphorylation	Ubiquitination	Acetylation	Methylation
All RNAs	**0.054** ***(0.059)***	**0.045** ***(0.049)***	**0.003** ***(0.004)***	**0.003** ***(0.004)***	**0.003** ***(0.004)***
mRNA	**0.061** ***(0.068)***	**0.053** ***(0.058)***	**0.003** ***(0.004)***	**0.002** ***(0.004)***	**0.004** ***(0.005)***
ncRNA	**0.040** ***(0.045)***	**0.035** ***(0.040)***	**0.001** ***(0.002)***	**0.003** ***(0.005)***	**0.001** ***(0.001)***
rRNA	**0.046** ***(0.048)***	**0.037** ***(0.040)***	**0.004** ***(0.005)***	**0.004** ***(0.005)***	**0.001** ***(0.001)***
tRNA	**0.042** ***(0.046)***	**0.033** ***(0.037)***	**0.006** ***(0.007)***	**0.003** ***(0.004)***	0.000 ***(0.000)***

The above results imply that disordered regions outside of the RBDs in RBPs are involved in DNA-binding and host PTM sites. We validate these results using an independent source of data collected from Pfam. We annotate DNA-binding and PTM Pfam domains in RBPs and map them into the non-RBD segments that have disordered residues. We identified these domains in 29 RBPs and we list them in **Supplementary Table S3**. We use these data to investigate whether the prevalence of these domains in RBPs is significantly enriched when compared against the sampling of these domains in the human proteome. To do that, we randomly select half of all Pfam domains in RBPs and calculate the fraction of the annotated above PTM or DNA-binding domains among them. We perform the same calculation for the other human proteins (considering PTM and DNA-binding Pfam domains that are colocalized with disorder) and bootstrap this process 100 times. The medians of the DNA-binding and PTM Pfam domain

fractions are 0.62 and 0.10, which are significantly enriched in RBPs compared to the medians of 0.17 and 0.03, respectively, for the reference human proteome data (p-value < 0.05). This result is consistent with the analysis shown in **Figure 2**. Altogether, our analysis suggests that disorder is enriched in non-RBD regions of RBPs where it facilitates DNA-binding and harbors PTM sites.

3.6 Analysis of Highly-disordered RNA-binding Proteins

We conduct detailed analysis of several illustrative examples of highly disordered human RBPs interacting with different forms of RNAs, such as mRNA, rRNA, and snRNA (**Figure 3**).

mRNA-binding protein E1B-AP5. Heterogeneous nuclear ribonucleoprotein U-like protein 1 (HNRNPUL1; UniProt ID: Q9BUJ2) is an 856-residue-long RBP acting as a transcriptional regulator [72] and playing a role in the mRNA processing and transport, as well as in the nucleocytoplasmic transport of adenovirus [73]. The transcriptional activity of E1B-AP5 is regulated by the formation of complex with the bromodomain-containing protein BRD7 [72]. E1B-AP5 is known to have several functional regions, such as SAP domain (residues 3-37, named after SAF-A/B, Acinus and PIAS, three proteins known to contain it) that can serve as a DNA-binding motif [74], B30.2/SPRY domain (residues 191-388) that functions through protein-protein interaction, and the RGG box (residues 612-658) that contains five RGG repeats and is involved in the RNA binding and is needed for transcription repression [75]. The disorder content of this protein is 0.48, with the putative IDRs located at both termini of its sequence. In line with the conclusions of this study, the functional disorder profile generated for this protein using the MobiDB [52, 54] and D^{2}P^2 [51] resources (**Figure 3(a)**) illustrates that the DNA-binding SAP domain is disordered. Furthermore, functionality of this protein is regulated by a multitude of different PTMs that are preferentially located within the putative IDRs.

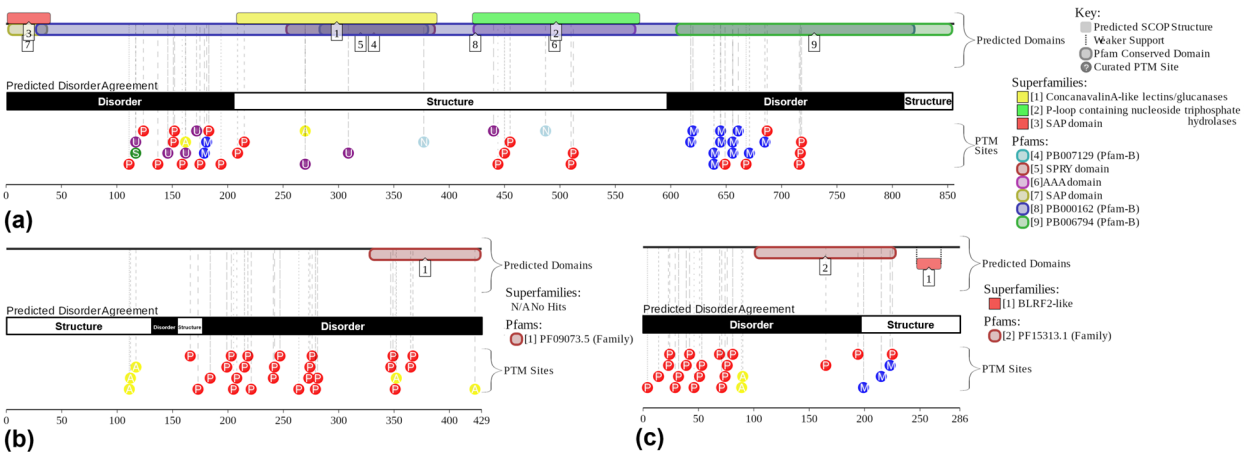

Figure 3. Functional profiles for illustrative representatives of human RBPs interacting with mRNA, rRNA, and ncRNA. Panel (a) shows the mRNA-binding protein E1B-AP5 (UniProt ID: Q9BUJ2). Panel (b) shows the rRNA-binding protein p49/STRAP (UniProt ID: Q8NEF9). Panel (c) shows the ncRNA-binding protein HEXIM2 (UniProt ID: Q96MH2). Horizontal bar in the middle shows disordered regions (in black) and structured regions (in white) predicted using the consensus of the ten disorder predictors from MobiDB. The colored and numbered bars above the disorder predictions show the positions of the (mostly structured) SCOP domains that are generated using the SUPERFAMILY predictor. The colored circles at the bottom show location of various PTMs assigned using PhosphoSitePlus, which is a comprehensive resource of the experimentally determined PTM sites. We collect the domain and PTM annotations from the D^{2}P^2 database.

rRNA-binding protein p49/STRAP. Serum response factor-binding protein 1 (SRFBP1) is also known as SRF-dependent transcription regulation-associated protein or p49/STRAP (UniProt ID: Q8NEF9). This 429-residue-long protein plays a role in the maturation of a precursor small subunit (SSU) rRNA

molecule into a mature SSU-rRNA molecule. SRFBP1 is a transcription cofactor of serum response factor (SRF) which regulates cytoskeletal and muscle-specific genes [76]. Levels of SRFBP1 increase with normal aging and are sensitive to glucose levels, being involved in redistribution of cytoskeletal F-actin during glucose deprivation, thereby playing a role in the glucose-deprivation associated cytoskeletal changes [77]. Recently, SRFBP1 was identified as a host factor for all seven genotypes of the hepatitis C virus (HCV) entry [78]. The currently available information about human SRFBP1 sequence is rather limited. This protein contains a conserved N-region that overlaps with the SRF-binding domain and C-terminally located highly conserved BUD22 domain, which in the yeast cellular morphogenesis acts as a regulator of the budding selection, polarity, and RNA biogenesis [79, 80]. The MobiDB and D^2P^2-generated functional disorder profile of the human SRFBP1 (**Figure 3(b)**) illustrates the high overall intrinsic disorder status of this protein, with the putative protein-level disorder content of 0.60, and presence numerous PTM sites. The fact that these sites are localized in IDRs suggest that disorder is involved in the PTM-driven regulation of its functionality.

snRNA-binding protein HEXIM2. Hexamethylene bis-acetamide-inducible protein 2 (HEXIM2; UniProt ID: Q96MH2) is a 286-residue-long protein that binds to 7SK snRNA [81] and the positive transcription elongation factor via a short sequence motif (residues 140-143) [82]. This nuclear protein contains a coiled-coil domain (residues 207-277) that overlaps with the region responsible for the interaction with cyclin-T1 (CCNT1) and formation of the homooligomers (likely homodimers) or heterooligomers with HEXIM1 [82, 83]. Based on the data reported in IntAct database [84], HEXIM2 has 69 protein binding partners and interacts with RNA and ssDNA. **Figure 3(c)** shows that the human HEXIM2 is a highly disordered protein (putative disorder content of 0.76), which is regulated by multiple PTMs located within its predicted disordered regions and is likely to utilize intrinsic disorder for the protein-protein and protein-nucleic acids interactions.

4. Summary and Conclusions

We analyze the peculiarities of intrinsic disorder distribution within the amino acid sequence of 1,112 human RNA-binding proteins. We also investigate at the differences in the intrinsic disorder predispositions for the RBPs that interact with mRNA, irRNA, tRNA, rRNA, ncRNA, and snRNA. In agreement with previous studies, we find that on average, RBPs are significantly more disordered than general human proteins. However, disorder predisposition is not equally distributed among the RBPs interacting with different RNA classes. Our analysis shows that although RBPs interacting with mRNAs, rRNAs, and snRNAs are significantly enriched in putative intrinsic disorder, the ncRNA- and irRNA-binding proteins are not enriched in disorder, whereas the tRNA-binding proteins are significantly depleted in putative disorder. Furthermore, we discover that the predicted disorder is heterogeneously distributed within the individual RBPs. The amount of putative intrinsic disorder in the non-RBD regions significantly exceed the disorder levels of RBDs, reflecting an interesting notion that high levels of disorder in RBPs are mostly used not for the interaction with RNA, but for the interactions with DNA and to host numerous and diverse types of the PTM sites. We also validate these observations that rely on the disorder predictions using a smaller collection of experimental disorder data from DisProt and based on an independent data from Pfam. This work provides an important snapshot of the peculiarities of functional disorder in RBPs and demonstrates that RBPs interacting with different functional RNAs are noticeably diversified in their functional utilization of intrinsic disorder. It also indicates that careful attention should be paid to small details even while looking at the big picture.

References

1. Re, A., et al., *RNA-protein interactions: an overview.* Methods Mol Biol, 2014. **1097**: p. 491-521.
2. Glisovic, T., et al., *RNA-binding proteins and post-transcriptional gene regulation.* FEBS Lett, 2008. **582**(14): p. 1977-86.
3. Van Nostrand, E.L., et al., *A large-scale binding and functional map of human RNA-binding proteins (vol 583, pg 711, 2020).* Nature, 2021. **589**(7842): p. E5-E5.
4. Gerstberger, S., M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins.* Nat Rev Genet, 2014. **15**(12): p. 829-45.
5. Sundararaman, B., et al., *Resources for the Comprehensive Discovery of Functional RNA Elements.* Molecular Cell, 2016. **61**(6): p. 903-913.
6. Han, K. and C. Nepal, *PRI-Modeler: extracting RNA structural elements from PDB files of protein-RNA complexes.* FEBS Lett, 2007. **581**(9): p. 1881-90.
7. Barik, A., A. Mishra, and R.P. Bahadur, *PRince: a web server for structural and physicochemical analysis of Protein-RNA interface.* Nucleic Acids Res, 2012. **40**(W1): p. W440-W444.
8. Lewis, B.A., et al., *PRIDB: a Protein-RNA interface database.* Nucleic Acids Res, 2011. **39**(Database issue): p. D277-82.
9. Konig, J., et al., *Protein-RNA interactions: new genomic technologies and perspectives (vol 13, pg 77, 2012).* Nature Reviews Genetics, 2012. **13**(3): p. 221-221.
10. Ascano, M., et al., *Identification of RNA-protein interaction networks using PAR-CLIP.* Wiley Interdisciplinary Reviews-Rna, 2012. **3**(2): p. 159-177.
11. Mann, M., *Functional and quantitative proteomics using SILAC.* Nature Reviews Molecular Cell Biology, 2006. **7**(12): p. 952-958.
12. Habchi, J., et al., *Introducing protein intrinsic disorder.* Chem Rev, 2014. **114**(13): p. 6561-88.
13. van der Lee, R., et al., *Classification of intrinsically disordered regions and proteins.* Chem Rev, 2014. **114**(13): p. 6589-631.
14. A. Keith Dunker, M.M.B., Elisar Barbar, Martin Blackledge, Sarah E. Bondos, Zsuzsanna Dosztányi, H. Jane Dyson, Julie Forman-Kay, Monika Fuxreiter, Jörg Gsponer, Kyou-Hoon Han, David T. Jones, Sonia Longhi, Steven J. Metallo, Ken Nishikawa, Ruth Nussinov, Zoran Obradovic, Rohit V. Pappu, Burkhard Rost, Philipp Selenko, Vinod Subramaniam, Joel L. Sussman, Peter Tompa & Vladimir N Uversky, *What's in a name? Why these proteins are intrinsically disordered.* Intrinsically Disordered Proteins, 2013. **1**(1): p. e24157
15. Lieutaud, P., et al., *How disordered is my protein and what is its disorder for? A guide through the "dark side" of the protein universe.* Intrinsically Disord Proteins, 2016. **4**(1): p. e1259708.
16. Xue, B., A.K. Dunker, and V.N. Uversky, *Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life.* J Biomol Struct Dyn, 2012. **30**(2): p. 137-49.
17. Schad, E., P. Tompa, and H. Hegyi, *The relationship between proteome size, structural disorder and organism complexity.* Genome Biology, 2011. **12**(12).
18. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cell Mol Life Sci, 2015. **72**(1): p. 137-51.
19. Corley, M., M.C. Burns, and G.W. Yeo, *How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms.* Molecular Cell, 2020. **78**(1): p. 9-29.
20. Jarvelin, A.I., et al., *The new (dis)order in RNA regulation.* Cell Commun Signal, 2016. **14**: p. 9.
21. Varadi, M., et al., *Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins.* PLoS One, 2015. **10**(10): p. e0139731.
22. Wang, C., V.N. Uversky, and L. Kurgan, *Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea.* Proteomics, 2016. **16**(10): p. 1486-98.

23. Hentze, M.W., et al., *A brave new world of RNA-binding proteins.* Nat Rev Mol Cell Biol, 2018. **19**(5): p. 327-341.
24. Castello, A., et al., *Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins.* Cell, 2012. **149**(6): p. 1393-1406.
25. Basu, S. and R.P. Bahadur, *A structural perspective of RNA recognition by intrinsically disordered proteins.* Cell Mol Life Sci, 2016. **73**(21): p. 4075-84.
26. Calabretta, S. and S. Richard, *Emerging Roles of Disordered Sequences in RNA-Binding Proteins.* Trends Biochem Sci, 2015. **40**(11): p. 662-672.
27. Peng, Z., et al., *A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome.* Cell Mol Life Sci, 2014. **71**(8): p. 1477-504.
28. Hudson, B.P., et al., *Recognition of the mRNA AU-rich element by the zinc finger domain of TIS11d.* Nat Struct Mol Biol, 2004. **11**(3): p. 257-64.
29. Castello, A., et al., *System-wide identification of RNA-binding proteins by interactome capture.* Nat Protoc, 2013. **8**(3): p. 491-500.
30. Hatos, A., et al., *DisProt: intrinsic protein disorder annotation in 2020.* Nucleic Acids Res, 2020. **48**(D1): p. D269-D276.
31. Simão, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-3212.
32. Chowdhury, S., J. Zhang, and L. Kurgan, *In Silico Prediction and Validation of Novel RNA Binding Proteins and Residues in the Human Proteome.* Proteomics, 2018: p. e1800064.
33. Nicastro, G., I.A. Taylor, and A. Ramos, *KH-RNA interactions: back in the groove.* Curr Opin Struct Biol, 2015. **30**: p. 63-70.
34. Lunde, B.M., C. Moore, and G. Varani, *RNA-binding proteins: modular design for efficient function.* Nat Rev Mol Cell Biol, 2007. **8**(6): p. 479-90.
35. Mistry, J., et al., *Pfam: The protein families database in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D412-D419.
36. Katuwawala, A., C.J. Oldfield, and L. Kurgan, *Accuracy of protein-level disorder predictions.* Brief Bioinform, 2020. **21**(5): p. 1509-1522.
37. Necci, M., et al., *A comprehensive assessment of long intrinsic protein disorder from the DisProt database.* Bioinformatics, 2018. **34**(3): p. 445-452.
38. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder.* Bioinformatics, 2015. **31**(2): p. 201-8.
39. Peng, Z.L. and L. Kurgan, *Comprehensive comparative assessment of in-silico predictors of disordered regions.* Curr Protein Pept Sci, 2012. **13**(1): p. 6-18.
40. Necci, M., et al., *Critical assessment of protein intrinsic disorder prediction.* Nat Methods, 2021. **18**(5): p. 472-481.
41. Katuwawala, A. and L. Kurgan, *Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins.* Biomolecules, 2020. **10**(12).
42. Zhao, B., et al., *IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell.* Cell Mol Life Sci, 2020.
43. Hu, G., et al., *Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity.* Proteomics, 2018: p. e1800243.
44. Kulkarni, P. and V.N. Uversky, *Intrinsically Disordered Proteins: The Dark Horse of the Dark Proteome.* Proteomics, 2018. **18**(21-22).
45. Yan, J., et al., *Structural and functional analysis of "non-smelly" proteins.* Cell Mol Life Sci, 2019.
46. Hu, G., et al., *Functional Analysis of Human Hub Proteins and Their Interactors Involved in the Intrinsic Disorder-Enriched Interactions.* Int J Mol Sci, 2017. **18**(12).
47. Giri, R., et al., *Understanding COVID-19 via comparative analysis of dark proteomes of SARS-CoV-2, human SARS and bat SARS-like coronaviruses.* Cell Mol Life Sci, 2020.

48. Peng, Z. and L. Kurgan, *On the complementarity of the consensus-based disorder prediction.* Pac Symp Biocomput, 2012: p. 176-87.
49. Necci, M., et al., *MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins.* Bioinformatics, 2017. **33**(9): p. 1402-1404.
50. Yan, J., M. Marcus, and L. Kurgan, *Comprehensively designed consensus of standalone secondary structure predictors improves Q3 by over 3%.* J Biomol Struct Dyn, 2014. **32**(1): p. 36-51.
51. Oates, M.E., et al., *D(2)P(2): database of disordered protein predictions.* Nucleic Acids Res, 2013. **41**(Database issue): p. D508-16.
52. Piovesan, D., et al., *MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins.* Nucleic Acids Res, 2018. **46**(D1): p. D471-D476.
53. Piovesan, D., et al., *MobiDB: intrinsically disordered proteins in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D361-D367.
54. Necci, M., et al., *MobiDB-lite 3.0: fast consensus annotation of intrinsic disorder flavours in proteins.* Bioinformatics, 2020.
55. Peng, K., et al., *Length-dependent prediction of protein intrinsic disorder.* BMC Bioinformatics, 2006. **7**: p. 208.
56. Linding, R., et al., *GlobPlot: Exploring protein sequences for globularity and disorder.* Nucleic Acids Res, 2003. **31**(13): p. 3701-8.
57. Yang, Z.R., et al., *RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins.* Bioinformatics, 2005. **21**(16): p. 3369-76.
58. Linding, R., et al., *Protein disorder prediction: implications for structural proteomics.* Structure, 2003. **11**(11): p. 1453-9.
59. Meszaros, B., G. Erdos, and Z. Dosztanyi, *IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding.* Nucleic Acids Res, 2018. **46**(W1): p. W329-W337.
60. Dosztanyi, Z., et al., *IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content.* Bioinformatics, 2005. **21**(16): p. 3433-4.
61. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder.* Bioinformatics, 2012. **28**(4): p. 503-9.
62. Hornbeck, P.V., et al., *PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse.* Nucleic Acids Res, 2012. **40**(Database issue): p. D261-70.
63. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic Acids Res, 2015. **43**(18): p. e121.
64. Peng, Z., et al., *Prediction of Disordered RNA, DNA, and Protein Binding Regions Using DisoRDPbind.* Methods Mol Biol, 2017. **1484**: p. 187-203.
65. Zhao, B., et al., *DescribePROT: database of amino acid-level protein structure and function predictions.* Nucleic Acids Res, 2021. **49**(D1): p. D298-D308.
66. Mitchell, A., et al., *The InterPro protein families database: the classification resource after 15 years.* Nucleic Acids Res, 2015. **43**(Database issue): p. D213-21.
67. Howell, M., et al., *Not That Rigid Midgets and Not So Flexible Giants: On the Abundance and Roles of Intrinsic Disorder in Short and Long Proteins.* Journal of Biological Systems, 2012. **20**(4): p. 471-511.
68. Meng, F., et al., *Functional and structural characterization of osteocytic MLO-Y4 cell proteins encoded by genes differentially expressed in response to mechanical signals in vitro.* Sci Rep, 2018. **8**(1): p. 6716.
69. Ghadermarzi, S., et al., *Sequence-Derived Markers of Drug Targets and Potentially Druggable Human Proteins.* Front Genet, 2019. **10**: p. 1075.
70. Afanasyeva, A., et al., *Human long intrinsically disordered protein regions are frequent targets of positive selection.* Genome Res, 2018. **28**(7): p. 975-982.

71. Zhao, B., et al., *IDPology of the living cell: intrinsic disorder in the subcellular compartments of the human cell.* Cell Mol Life Sci, 2021. **78**(5): p. 2371-2385.
72. Kzhyshkowska, J., et al., *Regulation of transcription by the heterogeneous nuclear ribonucleoprotein E1B-AP5 is mediated by complex formation with the novel bromodomain-containing protein BRD7.* Biochem J, 2003. **371**(Pt 2): p. 385-93.
73. Gabler, S., et al., *E1B 55-kilodalton-associated protein: a cellular protein with RNA-binding activity implicated in nucleocytoplasmic transport of adenovirus and cellular mRNAs.* J Virol, 1998. **72**(10): p. 7960-71.
74. Aravind, L. and E.V. Koonin, *SAP - a putative DNA-binding motif involved in chromosomal organization.* Trends Biochem Sci, 2000. **25**(3): p. 112-4.
75. Kiledjian, M. and G. Dreyfuss, *Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box.* EMBO J, 1992. **11**(7): p. 2655-64.
76. Zhang, X., et al., *Overexpression of p49/STRAP alters cellular cytoskeletal structure and gross anatomy in mice.* BMC Cell Biol, 2014. **15**: p. 32.
77. Williams, E.D., et al., *p49/STRAP, a Serum Response Factor Binding Protein (SRFBP1), Is Involved in the Redistribution of Cytoskeletal F-Actin Proteins during Glucose Deprivation.* J Nutr Health Aging, 2017. **21**(10): p. 1142-1150.
78. Gerold, G., et al., *Quantitative Proteomics Identifies Serum Response Factor Binding Protein 1 as a Host Factor for Hepatitis C Virus Entry.* Cell Rep, 2015. **12**(5): p. 864-78.
79. Ni, L. and M. Snyder, *A genomic study of the bipolar bud site selection pattern in Saccharomyces cerevisiae.* Mol Biol Cell, 2001. **12**(7): p. 2147-70.
80. Huh, W.K., et al., *Global analysis of protein localization in budding yeast.* Nature, 2003. **425**(6959): p. 686-91.
81. Peterlin, B.M., J.E. Brogie, and D.H. Price, *7SK snRNA: a noncoding RNA that plays a major role in regulating eukaryotic transcription.* Wiley Interdiscip Rev RNA, 2012. **3**(1): p. 92-103.
82. Dulac, C., et al., *Transcription-dependent association of multiple positive transcription elongation factor units to a HEXIM multimer.* J Biol Chem, 2005. **280**(34): p. 30619-29.
83. Yik, J.H., et al., *Compensatory contributions of HEXIM1 and HEXIM2 in maintaining the balance of active and inactive positive transcription elongation factor b complexes for control of transcription.* J Biol Chem, 2005. **280**(16): p. 16368-76.
84. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D358-63.