DOI 10.1002/pmic.201500177 *Proteomics* 2016, *16*, 1486–1498

RESEARCH ARTICLE

Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea

Chen Wang[1,2], *Vladimir N. Uversky*[3,4,5*] *and Lukasz Kurgan*[1,2]

[1] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA
[2] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada
[3] Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA
[4] Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation
[5] Department of Biology, Faculty of Science, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia

Intrinsically disordered proteins (IDPs) are abundant in various proteomes, where they play numerous important roles and complement biological activities of ordered proteins. Among functions assigned to IDPs are interactions with nucleic acids. However, often, such assignments are made based on the guilty-by-association principle. The validity of the extension of these correlations to all nucleic acid binding proteins has never been analyzed on a large scale across all domains of life. To fill this gap, we perform a comprehensive computational analysis of the abundance of intrinsic disorder and intrinsically disordered domains in nucleiomes (~548 000 nucleic acid binding proteins) of 1121 species from Archaea, Bacteria and Eukaryota. Nucleiome is a whole complement of proteins involved in interactions with nucleic acids. We show that relative to other proteins in the corresponding proteomes, the DNA-binding proteins have significantly increased disorder content and are significantly enriched in disordered domains in Eukaryotes but not in Archaea and Bacteria. The RNA-binding proteins are significantly enriched in the disordered domains in Bacteria, Archaea and Eukaryota, while the overall abundance of disorder in these proteins is significantly increased in Bacteria, Archaea, animals and fungi. The high abundance of disorder in nucleiomes supports the notion that the nucleic acid binding proteins often require intrinsic disorder for their functions and regulation.

Received: May 11, 2015
Revised: February 26, 2016
Accepted: March 29, 2016

Keywords:
DNA-binding proteins / Intrinsically disorder / Intrinsically disordered proteins / Nucleiome / RNA-binding proteins / Systems biology

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

A significant portion of any given proteome is ascribed to biologically active proteins that do not have unique 3D structures as a whole or in part [1–6]. The dynamic conformational ensembles of such intrinsically disordered proteins

Correspondence: Dr. Lukasz Kurgan, 401 West Main Street, Room E4225, Richmond, VA 23284, USA
E-mail: lkurgan@vcu.edu
Fax: +1-804-828-2771

Abbreviations: DC, Disorder content; **IDP**, Intrinsically disordered protein; **IDR**, Intrinsically disordered regions; **LDR**, Long disordered region; **MNC**, mean net charge; **RDC**, relative disorder content

*Additional corresponding author: Vladimir N. Uversky
E-mail: vuversky@health.usf.edu
Fax: 813-974-3757
Colour Online: See the article online to view Fig. 1 in colour.

Significance of the study

Intrinsically disordered proteins and protein regions are abundant across all domains of life and their functions complement the functional repertoire of structured (ordered) proteins. In spite of the functional importance of protein-nucleic acids interactions and the fact that intrinsically disordered proteins were shown to be very common in some classes of RNA- and DNA-binding proteins, the question on how abundant the disorder is in the nucleiome (whole complement of proteins involved in interaction with nucleic acids) remains open. We report a first-of-its-kind comprehensive analysis of the abundance of intrinsic disorder and intrinsically disordered domains in nucleiomes of over 1100 species from Archaea, Bacteria and Eukaryota. We investigate intrinsic disorder from multiple perspectives that include enrichment in the overall amount of disorder and the amount of disordered domains, relation to the nucleic acid binding and enrichment in a comprehensive set of functional classes of RNA- and DNA-binding proteins. Our results provide a strong support to the notion that the nucleic acid binding proteins often require intrinsic disorder for their functions and regulation.

(IDPs) or hybrid proteins with ordered and intrinsically disordered regions (IDRs) [2, 7–11] are highly heterogeneous, ranging from collapsed-disordered (molten globule-like) to partially collapsed-disordered (pre-molten globule-like) and to extended-disordered (coil-like) forms [12, 13]. Functions of these proteins complement functional repertoire of ordered proteins [14–17], with IDPs/IDPRs being commonly involved in various cellular regulation, recognition, signaling and control pathways [18–20].

Furthermore, intrinsic disorder was shown to be very common in several classes of RNA- and DNA-binding proteins [4, 9, 10, 21]. For example, IDPs were shown to be overrepresented in the nuclei of the *Saccharomyces* cells, where they are potentially involved in the regulation and control of transcription [3]. IDPs are also very common in nuclei of human cells [22]. Some other illustrative examples of intrinsically disordered DNA- or RNA-binding proteins include transcription factors [23–25], histones [26], ribosomal proteins [27] and proteins involved in the formation and action of human [28] and yeast spliceosomes [29]. One of the most recent studies investigates conformational characteristics and evolutionary conservation for a generic set of RNA-binding proteins [30]. However, these studies focus on specific families/classes of DNA- or RNA-binding proteins or specific organisms.

Although the aforementioned studies draw an impressive picture of the prevailing abundance of disorder in several classes of nucleic acid binding proteins, the question still remained open on how abundant disorder is in the nucleiome, which is defined as whole complement of proteins involved in interaction with nucleic acids (DNA and RNA). In other words, despite the fact that the high levels of intrinsic disorder were found in transcription factors, histones, ribosomal and spliceosomal proteins, the validity of the extension of these correlations to the entire nucleiome in a comprehensive set of species from all domains of life has never been analyzed before. To fill this gap, we perform a comprehensive computational analysis of the abundance of intrinsic disorder in the nucleiome comprising ~548 000 nucleic acid binding proteins from 1121 complete proteomes from three domains of life, Archaea, Bacteria and Eukaryota. Our study is the first to systematically and on large scale analyze and compare side-by-side the abundance and functional roles of disorder in DNA- and RNA-binding proteins across a diverse set of species that comprehensively cover all domains of life. The key factors that define our analysis are: (i) inclusion of both DNA- and RNA-binding proteins over the same set of organisms; (ii) comprehensive coverage of all domains of life; and (iii) analysis at the species and domain levels. Overall, this analysis reveals that the entire nucleiome is enriched in intrinsic disorder, since relatively to other proteins in corresponding proteomes, DNA- and RNA-binding proteins contain significantly more intrinsic disorder, often in the form of disordered domains.

2 Materials and methods

2.1 Datasets and annotation of DNA- and RNA-binding proteins

We collect 1674 complete proteomes and the corresponding 20 619 215 proteins from the release 2015_08 of UniProt [31]. We remove viruses and categorized the remaining proteomes into the Archaea, Bacteria and Eukaryota domains of life. To reduce bias towards certain overpopulated in the UniProt species and to reduce redundancy, we aggregate all the taxonomic identifiers to the second last level of the taxonomy, which most commonly is the genus. We select one largest proteome from each group of these identifiers, consequently sampling species at the genus level. We annotate DNA- and RNA-binding proteins in these proteomes based on the annotations from the Gene Ontology (GO) database [32] that are linked in the UniProt. We include proteins that have molecular function specified as "DNA binding" or "RNA binding", and the direct child terms of the "DNA binding" and "RNA binding" terms in the gene ontology network. To assure that the number of DNA- and RNA-binding proteins is sufficiently large to perform statistical analysis, we remove the proteomes

where the total number of DNA-binding proteins and RNA-binding proteins is below 20. The resulting dataset includes 7 912 445 proteins from 1121 complete proteomes (species). The number of species and proteins in each domain of life and in the eukaryotic kingdoms that are included in our dataset, and the number and fraction of the DNA- and RNA-binding proteins aggregated into the domains and kingdoms of life are shown in Supporting Information Table S1. To summarize, the nucleiome analyzed in our study includes 548 091 DNA- and RNA-binding proteins. The complete list of considered species is given in Supporting Information Table S2.

The annotations of the DNA- and RNA-binding proteins that we use include both manually annotated proteins and proteins that are annotated automatically in UniProt using sequence similarity to the manually annotated entries. The coverage drops dramatically if we would only use the manually annotated entries. Out of the considered 1121 proteomes only 1% of eukaryotic species has over 50% of its proteins manually annotated while such level of coverage is not even present in Archaea and Bacteria (Supporting Information Fig. S1A). Moreover, only 6, 3 and 2% of the species in Archaea, Eukaryota and Bacteria, respectively, have over 20% of their proteins reviewed. Similar lack of reviewed entries is true for the proteins annotated as DNA- and RNA-binding. About 30, 15 and 5% of species in Archaea, Bacteria and Eukaryota, respectively, have at least 20% of their all DNA- and RNA-binding proteins annotated manually and these numbers drop to 6, 1 and 1% when we require at least 50% coverage (Supporting Information Fig. S1B). More importantly, our analysis shown in the Results section reveals that results based on a limited set of the reviewed proteomes and proteins are very similar to the results on the complete set of all proteomes and proteins.

2.2 Annotation and computational characterization of intrinsic disorder

Given the large scale of our analysis, we perform the annotation of disordered regions with high-throughput predictors. We utilize a consensus of five such predictors including two versions of IUPred [33] designed to find long and short IDRs and three versions of Espritz [34] that predict intrinsic disorder annotated based on structures solved via X-ray crystallography, the NMR-derived structures and the Disprot database [35]. These methods were shown to provide good predictive performance in a recent large-scale assessment with AUC values around 0.77 [36]. We use majority vote consensus where a given residue is predicted as disordered if majority of the methods (three or more out of the five methods) predict it as disordered; otherwise the residue is predicted as structured. The motivation to use of the consensus comes from empirical observations which reveal that this leads to further increase in the predictive performance when compared to the use of individual predictors [36–38]. The same consensus was recently used in related works [6, 27, 39]. The predictions were filtered by removing disordered segments with less than four consecutive residues, which is in agreement with other studies [6, 27, 40]. We note that such annotations of disordered regions can be collected from two databases: MobiDB [41, 42] and D^2P^2 [43].

We use the putative disorder to compute disorder content (DC) per protein, which is defined as the fraction of disordered residues in the given protein chain. The DC for a proteome is the average of the DC values of its proteins. We calculate relative disorder content (RDC), which quantifies the DC in the DNA- and RNA-binding proteins relative to the overall DC in their corresponding proteome. RDC is calculated per protein as:

$$\begin{aligned}\mathrm{RDC}_{\mathrm{protein}} = {} & \mathrm{sign}\left\{\mathrm{DC}_{\mathrm{bind}} - \mathrm{DC}_{\mathrm{proteome}}\right\} \\ & \times \left(\max\left\{\mathrm{DC}_{\mathrm{bind}}, \mathrm{DC}_{\mathrm{proteome}}\right\} / \right. \\ & \left. \min\left\{\mathrm{DC}_{\mathrm{bind}}, \mathrm{DC}_{\mathrm{proteome}}\right\} - 1\right) * 100\%\end{aligned}$$

where $\mathrm{DC}_{\mathrm{bind}}$ is the DC of a given DNA- or RNA-binding protein; $\mathrm{DC}_{\mathrm{proteome}}$ is the DC of the corresponding proteome; and sign is the sign of the difference between $\mathrm{DC}_{\mathrm{proteome}}$ and $\mathrm{DC}_{\mathrm{bind}}$ where the value is + (–) when $\mathrm{DC}_{\mathrm{bind}} > \mathrm{DC}_{\mathrm{proteome}}$ ($\mathrm{DC}_{\mathrm{bind}} < \mathrm{DC}_{\mathrm{proteome}}$) and $\mathrm{RDC}_{\mathrm{protein}} = 0$ if $\mathrm{DC}_{\mathrm{bind}} = \mathrm{DC}_{\mathrm{proteome}}$.

We also compute RDC per proteome as:

$$\begin{aligned}\mathrm{RDC}_{\mathrm{proteome}} = {} & \mathrm{sign}\left\{\mathrm{aveDC}_{\mathrm{bind}} - \mathrm{DC}_{\mathrm{proteome}}\right\} \\ & \times \left(\max\left\{\mathrm{aveDC}_{\mathrm{bind}}, \mathrm{DC}_{\mathrm{proteome}}\right\} / \right. \\ & \left. \min\left\{\mathrm{aveDC}_{\mathrm{bind}}, \mathrm{DC}_{\mathrm{proteome}}\right\} - 1\right) * 100\%\end{aligned}$$

where $\mathrm{aveDC}_{\mathrm{bind}}$ is the average DC of all DNA- or all RNA-binding proteins in a given proteome.

The $\mathrm{RDC}_{\mathrm{protein}}$ ($\mathrm{RDC}_{\mathrm{proteome}}$) values are defined as a ratio between DC of the binding chain (or a set of binding chains) and its (their) proteome and it quantifies the amount of enrichment (if value is positive) or depletion (if negative) of the disorder in the binding protein(s) relative to the content of the proteome. The values of RDC vary between negative and positive infinity and have intuitive interpretation, e.g., given $\mathrm{DC}_{\mathrm{bind}} = 0.5$ and $\mathrm{DC}_{\mathrm{proteome}} = 0.25$, $\mathrm{RDC}_{\mathrm{protein}} = 100\%$ which shows that the DC of the protein is 100% higher than the DC of its proteome; given $\mathrm{DC}_{\mathrm{bind}} = 0.25$ and $\mathrm{DC}_{\mathrm{proteome}} = 0.5$, $\mathrm{RDC}_{\mathrm{protein}} = -100\%$ which means that the DC of the protein is 100% lower than that of its proteome.

We measure correlation between the average DC in proteome and the average DC in its DNA- or RNA-binding proteins over a given domain or kingdom of life to investigate whether enrichment of depletion of disorder in the binding proteins is consistent. The correlation is quantified with the Pearson correlation coefficient (PCC):

$$\begin{aligned}\mathrm{PCC} = {} & \mathrm{covariance}\left(\mathrm{aveDC}_{\mathrm{bind}}, \mathrm{DC}_{\mathrm{proteome}}\right) / \\ & \left(\sigma_{\mathrm{aveDCbind}} * \sigma_{\mathrm{DCproteome}}\right)\end{aligned}$$

where $\mathrm{aveDC}_{\mathrm{bind}}$, $\mathrm{DC}_{\mathrm{proteome}}$ are computed over proteomes in a given domain or kingdom and σ is the standard deviation.

We annotate long disordered regions (LDRs) which are defined as having 30 or more consecutive disordered residues [3, 44–46]. These regions are recognized as a distinct class of biologically functional IDP domains [3, 47, 48], which means that proteins with LDRs are likely to carry functions through disorder. We use these annotations to compute fraction of proteins with LDRs in DNA- or RNA-binding proteins in a given proteome, in entire proteomes and in domains and kingdoms of life. The latter results are summarized in Supporting Information Table S1.

We compute mean net charge (MNC) of proteins and LDRs, which can be approximated as [4, 49]:

$$\mathrm{MNC} = [(N_R + N_K) - (N_E + N_D)]\,/\mathrm{length}$$

where N_R, N_K, N_E and N_D are the counts of arginine, lysine, glutamic acid and aspartic acid residues and length is the length of a given protein sequence of a protein or LDR.

2.3 Evaluation of the annotations of disorder

We compare the consensus of the five methods that is utilized in this study with two other consensuses to demonstrate that our results are not affected by a potential noise coming from the use of different types of annotations of disorder by various prediction methods. The first alternative consensus is based on the ten methods that are included in MobiDB database [41, 42]: two versions of IUPred [33], three versions of Espritz [34], two flavors of DisEMBL [50], GlobPlot [51], VSL2b [52] and JRONN [53]). The second consensus includes these ten methods and another recently published method: DynaMine [54, 55], totalling in 11 predictors. We compare the predictions of the three consensuses on two popular model genera: *Escherichia* and *Drosophila* (Supporting Information Fig. S2). The results show that the three consensuses generate similar distributions of DC values for these genera, which suggests that our putative annotations of disorder are not affected by the selection of methods in the consensus.

Moreover, we also test predictive quality of these consensuses on the same genera by comparing their predictions with the results generated by the ANCHOR method [56, 57], which predicts protein binding residues located in the disordered regions. The predictions of the disordered residues should significantly overlap with the binding residues that are annotated with ANCHOR. To test that we compute the fraction of binding residues that are predicted by ANCHOR and that are also predicted as disordered by each of the three consensuses. We compute a ratio between this fraction and the fraction of randomly generated binding residues that are also predicted as disordered; we generate the same number of binding residues and regions as generated by ANCHOR and place them randomly in protein sequences excluding the locations of the ANCHOR's predictions. The ratio equal 1 would indicate that disorder is equally abundant in the randomly picked residues and in the disordered protein binding regions coming from ANCHOR, while ratios > 1 would indicate that disorder is enriched among the ANCHOR-generated binding residues. On the *Escherichia* we obtain ratios of 6.02, 4.13 and 3.20 for the consensuses with 5, 10 and 11 methods, respectively, while for *Drosophila* the ratios are 3.47, 2.64 and 2.39, respectively. These high values demonstrate the expected substantial enrichment of intrinsic disorder in the disordered protein binding regions. This suggests that the considered consensuses generate accurate predictions, which in in agreement with the prior empirical evaluations of these methods [36, 58].

2.4 Analysis of statistical significance

We assess statistical significance of the relationship between enrichment or depletion of intrinsic disorder in the DNA- or RNA-binding proteins relative to the disorder in the corresponding proteomes for each domain or kingdom of life. We quantify this using paired tests that evaluate significance of differences between $\mathrm{aveDC}_{\mathrm{bind}}$ and $\mathrm{DC}_{\mathrm{proteome}}$ (between the DC the DNA- or RNA-binding proteins and the whole proteome) paired for the same proteomes and over all proteomes in a given domain or eukaryotic kingdom. We also evaluate the significance of difference in the fraction of DNA- or RNA-binding proteins with LDRs between the DNA- or RNA-binding proteins and the whole proteome, again paired for the same proteomes and over all proteomes in a given domain or eukaryotic kingdom. First we test normality of the values of $\mathrm{aveDC}_{\mathrm{bind}}$, $\mathrm{DC}_{\mathrm{proteome}}$ and the fractions using Anderson–Darling test at p-value of 0.05. We use the Student's paired t-test for the normal data; otherwise we apply the Wilcoxon signed-rank test.

2.5 Functional annotations of the DNA- and RNA-binding proteins

We functionally annotate the considered DNA- and RNA-binding proteins using the GO annotations collected from the release 2015_08 of UniProt. We only consider reviewed UniProt entries and aggregate the annotations by the corresponding domains of life. We generate a list of all molecular functions from GO and Enzyme Commission (EC) numbers associated with the RNA- and DNA-binding proteins in the UniProt for each of the three domains of life. To allow for statistically sound estimates of the abundance of intrinsic disorder we include a given functional annotation in our analysis if the count of proteins that it covers > 20. In case of the enzyme types, we focus the analysis of eukaryotic species since the coverage of the EC numbers in the Archaea and Bacteria is relatively low. Moreover, we also collected annotations of four specific functional classes of RNA- and DNA-binding proteins for which the count of proteins > 20: transcription factors, histones, ribosomal proteins and splicing factors. These were identified by the corresponding keywords in the

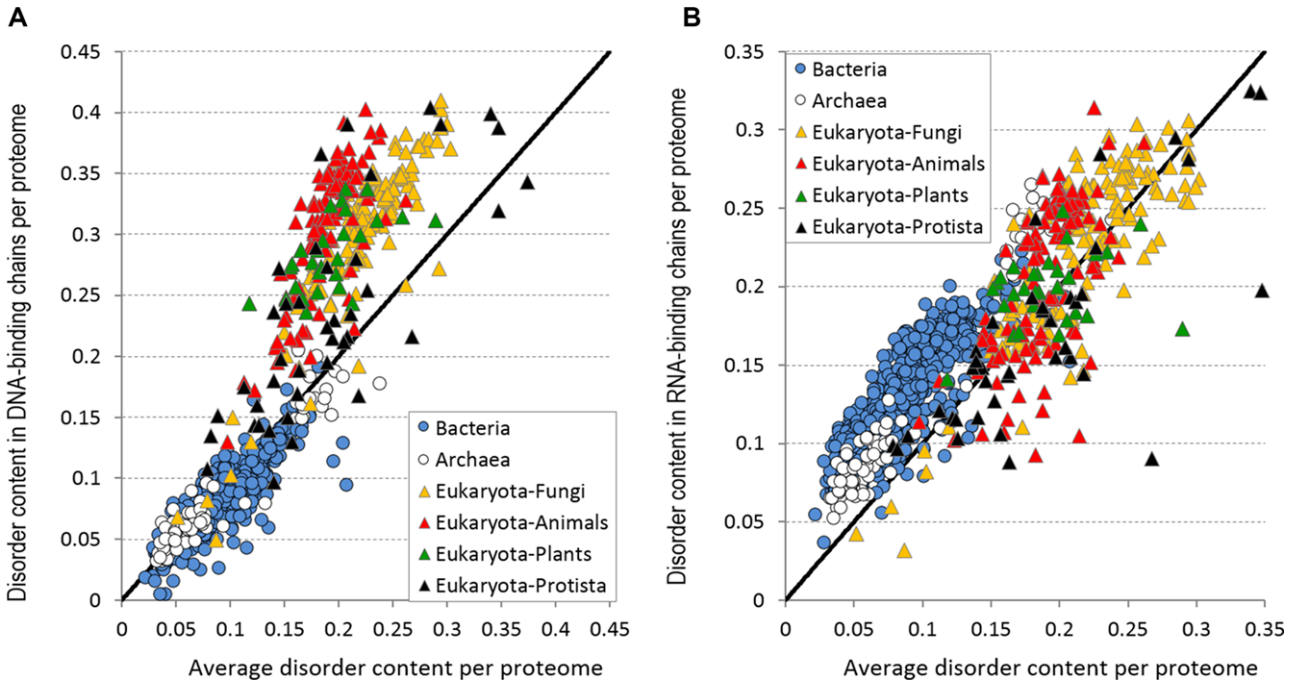

Figure 1. Relationship between average disorder content in an entire proteome (species) and average disorder content in the corresponding DNA-binding (panel A) and RNA-binding proteins (panel B) for the considered 1121 proteomes. Each marker represents a proteome (species). The black diagonal line represents the positions where the average disorder content in the DNA-binding or RNA-binding proteins equals to the average disorder content in the entire proteome. Markers above (below) the diagonal line correspond to species that have enriched (depleted) disorder content in the binding proteins when compared to all proteins.

protein name. In total, we consider 36 molecular functions in Archaea (17 for DNA-binding and 19 for RNA-binding proteins), 85 molecular functions in Bacteria (34 for DNA-binding and 51 for RNA-binding proteins), 194 molecular functions in Eukaryota (117 for DNA-binding and 77 for RNA-binding proteins), 23 enzyme types (10 for DNA-binding and 13 for RNA-binding proteins) and four functional classes in Eukaryota. Some these functional annotations are common for both DNA- and DNA-binding proteins and across the three domains of life.

3 Results

3.1 Enrichment in intrinsic disorder in the DNA- and RNA-binding proteins

We compare the DC computed per proteome ($DC_{proteome}$) with the content computed for the DNA- or RNA-binding proteins ($aveDC_{bind}$) in the same proteome. The results for the considered 1121 proteomes are grouped by their taxonomic annotations into Archaea, Bacteria, Eukaryota and four eukaryotic kingdoms including animals, plants, fungi and Protista (Fig. 1). Points located above the black diagonal line indicate species for which abundance of intrinsic disorder in the nucleic acid binding proteins is enriched compared to their overall abundance of disorder. Our analysis suggests that DNA-binding proteins are enriched in disorder in proteomes from all eukaryotic kingdoms (Fig. 1A). The RNA-binding proteins are enriched in proteomes from Bacteria, Archaea and in most of the Eukaryota, in particular, in the majority of animal and fungi proteomes (Fig. 1B). The amount of enrichment is substantial and can be as high as 100%, which we observe for some animal and plant proteomes for the DNA-binding proteins and for some bacterial proteomes for the RNA-binding proteins.

We aggregate these results for each domain and eukaryotic kingdom of life, compute the change in the DC in the RNA- and DNA-binding proteins relative to the content in their proteomes, assess statistical significance of these changes and compute correlation between the DC in proteomes and in the corresponding nucleic acid binding proteins (Table 1). The results reveal that the median increase of the DC in the DNA-binding proteins equals 46% in Eukaryota, with animal and plant proteomes having the largest median increases at 63 and 49%, respectively. The median DC in these proteins in proteomes from Eukaryota is 0.3 with 0.32 in animals, 0.31 in fungi and 0.28 in plants. This means that about a third of residues in these DNA-binding proteins are disordered. The enrichment in the disorder in the DNA-binding proteins in Eukaryota and all considered here eukaryotic kingdoms is statistically significant (p-value < 0.01). Interestingly, the correlations between DC in the DNA-binding proteins and

Table 1. The content and enrichment of the disorder in the DNA- and RNA-binding proteins in the considered 1121 proteomes. All measurements are computed at the proteome (species) level and grouped into the corresponding domains of life and eukaryotic kingdoms. We report the 20th centile (c20), 50th centile (median) and 80th centile (c80) over all species in a given domain/kingdom. We include values of disorder content in the RNA- and DNA-binding proteins, relative disorder content (content in the binding proteins relative to the content in a given species), fraction of proteins with LDRs, relative fraction of proteins with LDRs and Pearson Correlation Coefficient (PCC) values and statistical significance of differences between the overall disorder content and disorder content in the binding proteins, and between the overall fraction of proteins with LDRs and fraction of the binding proteins with LDRs. The Eukaryotic kingdoms are sorted by the median value of the relative disorder content

Binding type	Domain or kingdom of life	Disorder content in binding proteins median[c20, c80]	Relative disorder content median[c20, c80]	PCC	Significance of differences in content (*p*-value)	Fraction of binding proteins with LDRs median[c20, c80]	Relative fraction of proteins with LDRs median[c20, c80]	PCC	Significance of differences in fraction (*p*-value)
DNA	Archaea	0.06 [0.05, 0.14]	–4% [–17%, 12%]	0.95	0.0720	0.05 [0.02, 0.19]	17% [–19%, 87%]	0.96	8.63E–06
	Bacteria	0.07 [0.05, 0.09]	–9% [–24%, 5%]	0.88	7.84E–42	0.06 [0.03, 0.11]	–16% [–60%, 15%]	0.89	6.77E–13
	Eukaryota	0.30 [0.24, 0.34]	46% [29%, 65%]	0.75	1.30E–107	0.70 [0.57, 0.80]	72% [52%, 88%]	0.84	4.05E–151
	Animals	0.32 [0.25, 0.35]	63% [46%, 76%]	0.89	1.15E–59	0.69 [0.56, 0.79]	79% [70%, 90%]	0.88	3.34E–72
	Plants	0.28 [0.26, 0.31]	49% [36%, 66%]	0.76	6.21E–52	0.62 [0.55, 0.70]	88% [58%, 102%]	0.91	3.19E–65
	Fungi	0.31 [0.25, 0.35]	37% [29%, 48%]	0.60	2.09E–16	0.76 [0.66, 0.84]	68% [53%, 83%]	0.70	4.75E–20
	Protista	0.22 [0.15, 0.32]	18% [4%, 55%]	0.77	1.39E–05	0.54 [0.39, 0.71]	44% [27%, 77%]	0.82	1.04E–12
RNA	Archaea	0.09 [0.07, 0.19]	46% [28%, 69%]	0.97	2.45E–28	0.06 [0.03, 0.25]	48% [8%, 154%]	0.97	5.04E–13
	Bacteria	0.12 [0.10, 0.15]	56% [32%, 88%]	0.83	2.25E–298	0.14 [0.09, 0.21]	102% [43%, 184%]	0.81	3.38E–225
	Eukaryota	0.21 [0.16, 0.26]	4% [–10%, 18%]	0.76	0.0015	0.54 [0.43, 0.64]	34% [15%, 50%]	0.81	3.66E–84
	Animals	0.21 [0.16, 0.25]	13% [–10%, 25%]	0.84	0.0052	0.52 [0.41, 0.61]	37% [25%, 49%]	0.85	7.17E–43
	Fungi	0.20 [0.18, 0.21]	6% [–7%, 19%]	0.63	0.0001	0.49 [0.42, 0.55]	54% [16%, 71%]	0.85	2.23E–41
	Plants	0.24 [0.18, 0.27]	4% [–8%, 14%]	0.32	0.2659	0.60 [0.49, 0.67]	31% [15%, 47%]	0.47	3.44E–10
	Protista	0.16 [0.12, 0.20]	–5% [–22%, 8%]	0.81	0.0201	0.44 [0.27, 0.59]	18% [0%, 47%]	0.79	0.0004

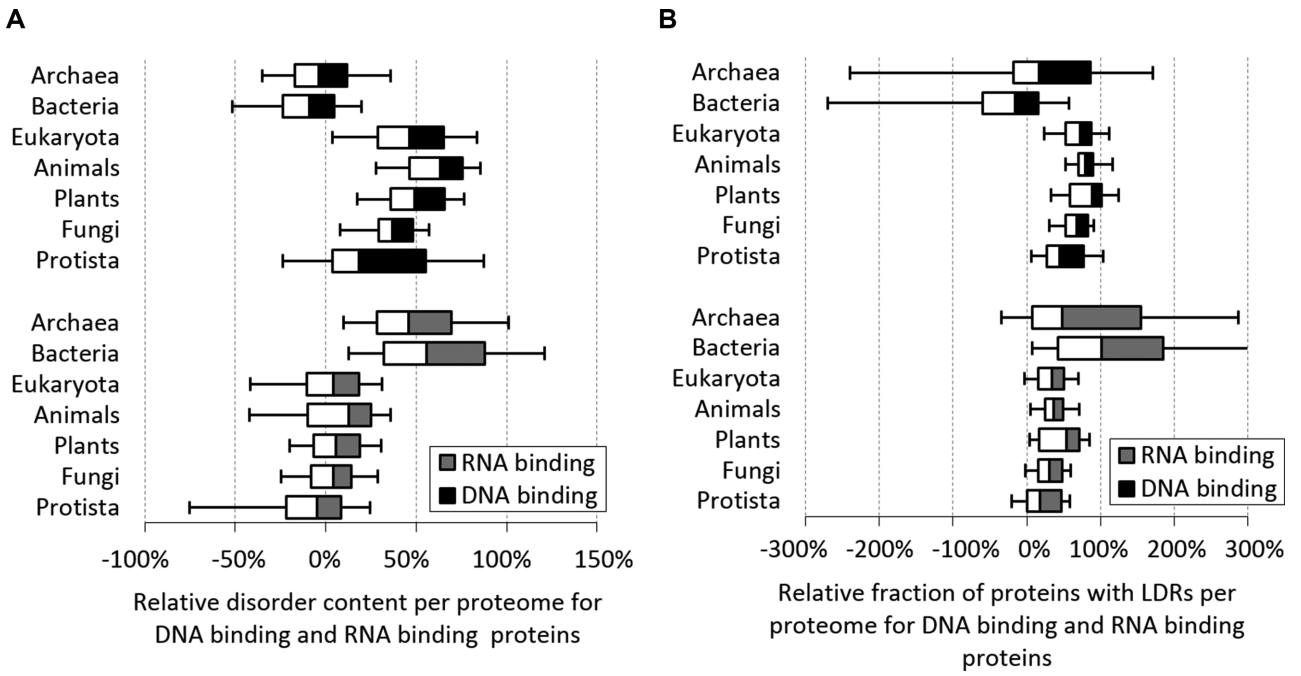

Figure 2. Relative disorder content (panel A) and relative fraction of proteins with LDRs (panel B) for the DNA- and RNA-binding proteins in Archaea, Bacteria, Eukaryota, Animals, Plants, Fungi and Protista. The box plots include 5th centile, 20th centile, 50th centile (median), 80th centile and 95th centile.

in the whole proteomes are very high, 0.6 or higher, in all domains of life; this can be observed in Fig. 1A. This means that proteomes that have lower or higher overall amount of disorder will also have similarly decreased or increased DC in the DNA-binding proteins. Results for the RNA-binding proteins show that organisms from Archaea and Bacteria are characterized by substantially larger amounts of disorder in these proteins compared to their overall amount of disorder. The median enrichment values of disorder in the RNA-binding proteins among these species are 46 and 56%, respectively, and these increases are statistically significant (p-value < 0.01). The enrichment in Eukaryota is relatively small at 4% (p-value $= 0.002$), with the exception of animal species that have the median enrichment at 13% (p-value $=$ 0.005). Similarly to the results for the DNA-binding proteins, the DC in the proteomes is highly correlated with the disorder in the RNA-binding proteins; see also Fig. 1B. The overall distribution of the RDC values for the RNA- and DNA-binding proteins in the considered domains and kingdoms of life is shown in Fig. 2A. It confirms that DNA-binding proteins are enriched in disorder in virtually all Eukaryotes while having no bias in the Archaea and Bacteria (black bars). Also, the RNA-binding proteins are universally and strongly enriched in disorder in Bacteria and Archaea and slightly enriched in the animals (gray bars).

We analyze further details of the enrichment in the intrinsic disorder in DNA- and RNA-binding proteins using histograms of the RDC for the nucleic acid binding proteins (Supporting Information Fig. S3). The x-axis shows RDC where positive values denote enrichment and negative depletion in DC in the nucleic acid binding proteins relative to the content of their proteomes. The first and last intervals on the x-axis denote DNA- and RNA-binding proteins that are structured and highly disordered, respectively. The abundance values for each interval, which are in the form of error bars, give the distribution of the fraction of proteins from specific proteomes in a given domain of life including the 5th, 20th, 50th (median), 80th and 95th centiles. Our analysis demonstrates that the differences in the DC range from a strong depletion to a strong enrichment. We note a bias towards enrichment in the DC for the RNA-binding proteins and for the DNA-binding proteins in Eukaryota; this is visible as a shift towards larger fractions of binding proteins (higher values on the y-axis) for the RDC values above 0 in Supporting Information Fig. S3. Majority of these DNA- and RNA-binding proteins have their relative enrichment in the DC below 250% (the median fraction of the binding proteins drops to low values for the x-axis values over 250%; see Supporting Information Fig. S3), although some binding proteins are enriched in disorder by as much as over 500%. This is true for the RNA-binding proteins in Archaea and Bacteria and stems from the overall lower amounts of disorder in these two domains of life compared to Eukaryota. However, a relatively substantial fraction of RNA-binding proteins (between a few and about 15%) and DNA-binding proteins (several to about 15% in Archaea and Bacteria and up to about 10% in Eukaryota) are structured. Thus, although overall we observe enrichment in the disorder in the nucleic acid binding

proteins, some of them are fully structured. Interestingly, both RNA- and DNA-binding proteins from animals (gray lines in Supporting Information Fig. S3E and F) are characterized by histograms that indicate increase in the enrichment (depletion) that is higher (lower) than the overall values for Eukaryota over the entire range of the RDC values.

3.2 Enrichment in intrinsically disordered domains in the DNA- and RNA-binding proteins

Overall, about 39% of proteins in Eukaryota, 10% in Archaea and 9% in Bacteria have LDRs; i.e., IDRs having 30 or more consecutive disordered residues, which constitute disordered domains (Supporting Information Table S1). We analyze differences in the abundance of proteins with LDRs between the DNA- and RNA-binding binding proteins and the overall proteomes. Table 1 shows that similarly to the DC, the DNA-binding proteins are significantly enriched in LDRs in Eukaryota (p-value < 0.01). The overall median enrichment equals 72%, and the median enrichment for animals and plants is 79 and 88%, respectively. About 68% of DNA-binding proteins in the eukaryotic species have LDRs compared to only about 12% in Archaea and 8% in Bacteria (Supporting Information Table S1). Interestingly, the RNA-binding proteins across all domains of life are significantly enriched in LDRs, see Table 1 (p-value < 0.01). The median enrichment values range between 18% in Protista and 102% in Bacteria, and they are at 37% and 54% for animals and fungi, respectively. Our analysis reveals that about 50% of RNA-binding proteins in Eukaryota have disordered domains, 15% in Bacteria and 11% in Archaea (Supporting Information Table S1). In spite of having relatively small magnitude, the latter two numbers are significantly higher than those for the corresponding complete proteomes, see Table 1 (p-value < 0.01). Figure 2B shows the distribution of the fraction of proteins that have LDRs and further confirms the notion that nearly all species in the three domains of life are enriched in the disordered domains in the RNA-binding proteins (gray bars), while this enrichment for the DNA-binding proteins holds only for the Eukaryotes (black bars). Taken together, we show that the amount of disordered domains is significantly enriched in the RNA-binding proteins across all three domains of life while the DC in these proteins is also significantly enriched in Bacteria, Archaea and animals. Furthermore, DNA-binding proteins in Eukaryota are significantly enriched in the disorder and in the disordered domains.

3.3 Comparison of results based on reviewed and a combined set of reviewed and automatically annotated DNA- and RNA-binding proteins

Since our analysis is based on both manually and automatically (based on sequence alignment) annotated nucleic acid binding proteins, we compare these results with a small subsets of manually reviewed annotations. The coverage rates of reviewed proteins are very low in most proteomes (Supporting Information Fig. S1) and thus to secure sufficient amount of data we select the proteomes with the coverage rates of at least 20%. Consequently, our analysis includes 31 such proteomes with five in Archaea, 18 in Bacteria and eight in Eukaryota. We compare the distribution of the per proteome values of DC and fraction of proteins with LDRs in the DNA- and RNA-binding proteins in each domain of life when considering the reviewed and all (reviewed and automatically annotated) proteins (Supporting Information Fig. S4); these data can be compared with Table 1 that is based on the complete set of 1121 proteomes. The values are very similar between the analysis using reviewed and all entries. For instance, Supporting Information Fig. S4A shows that the median DC in the DNA-binding proteins in species in Archaea for both reviewed and all proteins is 5.6% (6% in Table 1), in Bacteria for both is also equal 5.6% (7% in Table 1) and in Eukaryota it equals 30.1 and 33.8% for all and reviewed entries, respectively (30% in Table 1). The content in the RNA-binding proteins in species in Archaea for the reviewed and all proteins is 9.0 and 10.5%, respectively (9% in Table 1), in Bacteria is 11.0 and 9.7%, respectively (12% in Table 1), and in Eukaryota is 25.0 and 25.1%, respectively (21% in Table 1). We observe similar results for the fractions of proteins with LDRs. For instance, in Eukaryota 69.4 and 71.9% of the reviewed and all DNA-binding proteins, respectively, have LDRs (70% in Table 1) and 58.0 and 58.6% of the reviewed and all DNA-binding proteins, respectively, have LDRs (54% in Table 1). Overall, these numbers for the reviewed and all binding proteins from the selected 31 proteomes and from the all binding proteins in the full set of 1121 proteomes are similar and preserve the relative (between the domains of life) similarities and differences. This suggests that the inclusion of the automatically assigned annotations does not affect the overall conclusions.

3.4 Peculiarities of charge distribution

Figure 3 summarizes the MNC differences of proteins from the whole genomes and RNA- and DNA-binding proteins in Archaea, Bacteria and Eukaryota (black bars). As expected, the nucleic acid binders are characterized by higher net charge. This is because the protein-nucleic acid binding events involve interactions between the positively charged amino acids of the nucleiome members with the negatively charged phosphate backbone of DNA or RNA. This also agrees with the fact that net charge was found predictive for the DNA- and RNA-binding proteins and sites [59–61]. Furthermore, we investigate the values of the MNC in the LDRs localized in the DNA- and RNA-binding proteins compared to the LDRs in the whole genomes (gray bars in Fig. 3). Similar to the analysis for the whole proteins, there is a visible increase in the net charge for the disordered regions in the RNA-binding

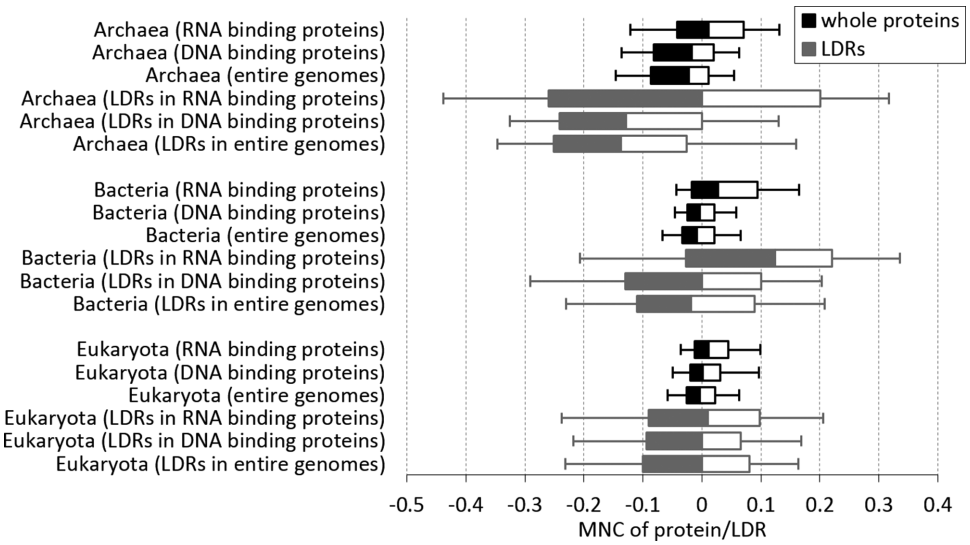

Figure 3. Mean net charge (MNC) of proteins (black bars) and LDRs (gray bars) in the complete proteomes and DNA- and RNA-binding proteins in Archaea, Bacteria and Eukaryota. The box plots include 5th centile, 20th centile, 50th centile (median), 80th centile and 95th centile.

proteins in Archaea and Bacteria and to a lesser extend in Eukaryota. This means that the disordered domains are not only enriched in the RNA-binding proteins in the three domains of life (Fig. 2B) but they are possibly directly involved in the binding. However, in spite of the substantial enrichment of the LDRs in the DNA-binding proteins in Eukaryota (Fig. 2B), these regions have similar net charge when compared with the overall population of LDRs (Fig. 3).

3.5 Analysis of functional annotations in the nucleiome

We analyze the enrichment in the disordered content for a set of nucleic acid binding proteins with well-annotated (manually reviewed) in the UniProt functional classes. We compute the median RDC per protein in each functional annotation (molecular functions from GO, EC numbers and the four specific functional classes) and each type of binding and select the annotations that are substantially depleted or enriched in disorder, i.e., the RDC $>$ 50% or $<$ –50%, respectively, when compared to the content in the same domain of life. In Eukaryota, we raise the threshold to 200% given the large number of enriched and depleted in disorder annotations in this domain of life. Consequently, out of 36 considered molecular functions in Archaea, 19 are either enriched or depleted in disorder with eight for the DNA-binding proteins (Fig. 4A) and 11 for the RNA-binding proteins (Fig. 4B). Similarly, in Bacteria we selected 30 out of 85 considered molecular functions, 11 for the DNA-binding proteins (Fig. 4C) and 19 for of the RNA-binding proteins (Fig. 4D). Finally, using the higher threshold, in Eukaryota we found 46 enriched or depleted molecular functions, 34 in the DNA-binding proteins (Fig. 4E) and 12 in the RNA-binding proteins (Fig. 4F), compared to the 194 considered annotations. We also found nine out of 23 considered types of enzymes in Eukaryota that are depleted or enriched in disorder, four in the DNA-binding proteins and five in RNA-binding proteins (Fig. 4G).

In Archaea (Fig. 4A and B) we found 3′-5′ exonucleases, polymerases, ribosomal proteins and ribonucleases to be enriched in disorder, while the 5′-3′ exonucleases, endonucleases, proteins associated with magnesium and ATP binding, and some of the ligases are depleted in disorder. In Bacteria (Fig. 4C and D), single stranded DNA-binding and double stranded RNA-binding proteins, ribosomal proteins, phosphatases and some mRNA-binding proteins are enriched in disorder. The list of molecular functions in Bacteria that are depleted in disorder includes magnesium binding proteins, 5′-3′ exonucleases, some of the hydratases, polymerases, ligases, nucleotidyltransferases, sulfotransferases, adenylyltransferases and methyltransferases. In Eukaryota (Fig. 4E and F) majority of the molecular functions are enriched in disorder and they include transcription factor and chromatic binding proteins, beta-catenin, ubiquitin, p53 and heparin binders, some of the kinases and methyltransferases and snRNA binders. The annotations that are depleted in disorder in Eukaryota include some of the ligases and ribonucleoside binding proteins. Analysis of the various types of enzymes (Fig. 4G) is consistent with the analysis based on the GO's molecular functions. We found that some of the methyltransferases and lyases are enriched in disorder, while a class of enzymes that act on ester bonds and some of the polymerases and ligases are depleted in disorder. Our analysis of the four specific functional classes that include histones, transcriptions factors, splicing factors and ribosomal proteins reveals that they are all substantially enriched in disorder in Eukaryota, with the median RDC of 317, 195, 355 and 47%, respectively.

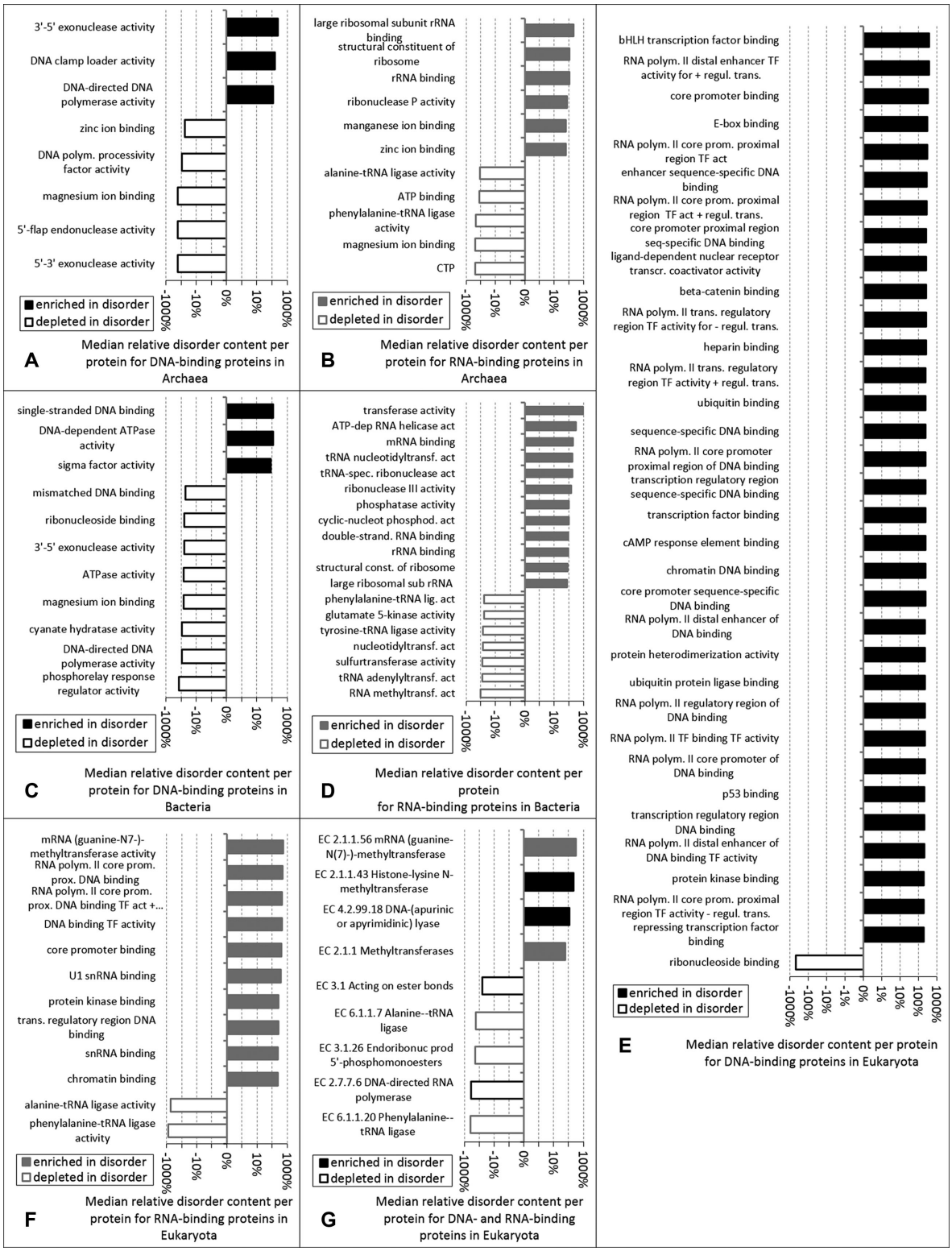

Figure 4. Median relative disorder content per protein for the enriched or depleted in disorder functional annotations of the DNA-binding and RNA-binding proteins. Panels A and B, C and D, E and F show results for the GO's molecular functions for the DNA- and RNA-binding proteins in Archaea, Bacteria and Eukaryota, respectively. Panel G shows results for the enzyme types for the DNA-binding and RNA-binding proteins in Eukaryota. Black and gray bars corresponding to the DNA-binding and RNA-binding proteins, respectively. Solid and hollow bars corresponding to functions where the disorder is enriched and depleted, respectively. The *x*-axis is in the logarithmic scale.

4 Discussion

Summarizing, for the first time, we performed a comprehensive computational analysis of the abundance of intrinsic disorder and intrinsically disordered domains in ~548 000 nucleic acid binding proteins found in 1121 complete proteomes of species from the three domains of life. These whole complements of proteins involved in interactions with nucleic acids (DNA and RNA) comprise nucleiomes of different species. On average, the nucleiome accounts for ~8.7% of the corresponding proteome. Although the 1121 analyzed proteomes contain, on average, ~16% of proteins with lLDRs, the corresponding fractions are significantly larger among the DNA- and RNA-binding proteins: 23.6 and 24.7% of them, respectively, contain LDRs. Overall, our study reveals that relative to other proteins in the corresponding proteomes, the DNA-binding proteins are characterized by the significantly increased DC and are significantly enriched in disordered domains in Eukaryotes (p-value < 0.01) but not in Archaea and Bacteria. The RNA-binding proteins are significantly enriched in disordered domains in Bacteria, Archaea and Eukaryota (p-value < 0.01), while the overall abundance of disorder in these proteins is significantly increased in Bacteria, Archaea and animals. We also show that proteins in nucleiomes are noticeably more charged than other proteins in corresponding proteomes and that the nucleiomal LDRs typically retain this bias toward higher content of charged residues. A very noticeable exception is the LDRs of the DNA-binding proteins in Eukaryota. In fact, although these proteins are characterized by the substantial enrichment in LDRs (Fig. 2B), their putative LDRs are characterized by the net charge similar that of other eukaryotic LDRs (Fig. 3). This finding points out to the intriguing possibility that the disordered domains of the eukaryotic DNA-binding proteins can have moonlighting functions other than DNA-binding and that these LDRs can also be used for regulation. This may also be true in reverse, meaning that the disordered domains that are not annotated as DNA-binding could have this as a "secondary" function.

Importantly, our analyses on disorder-function association show that among the highly disorder-enriched nucleic acid binding proteins are histones (over 300% enrichment), transcription factors (~200% enrichment), ribosomal proteins (~50% enrichment) and splicing factors (over 300% enrichment). These findings are in excellent agreement with the results of previous studies. In fact, earlier bioinformatics analyses revealed that many transcription factors are highly enriched in intrinsic disorder [23–25], with amino acid sequences of 401 human transcription factors being ~50% disordered [25]. Histone tails (N-terminal domains of core histones and C-terminal domains of linker histones) were long known to be important members of the IDP realm [62]. They contain multiple sites of various posttranslational modifications that modulate the structure of chromatin [63, 64] and constitute the basis of the histone code [64–68]. Interestingly, a recent large-scale analysis of the members of histone family, where 2007 non-redundant chains from 746 species were studied, revealed that majority of the histone family members were predicted to be mostly disordered, with intrinsic disorder extending far beyond the limits of the aforementioned tails [26]. Another recent bioinformatics analysis of the 3411 ribosomal proteins from 32 species revealed that >35% of these proteins are completely disordered and almost all remaining ribosomal proteins contain disordered domains [27]. Finally, computational analyses showed that ~50% of the 109 yeast spliceosomal proteins were predicted to be mostly disordered, 44 and 48% were expected to be moderately ($10\% \leq DC < 30\%$) and highly disordered (disorder $\geq 30\%$), respectively, and that only ~8% of the yeast spliceosomal proteins were expected to be highly ordered proteins containing less than 10% of disordered residues [29].

Concluding, our study provides compelling evidence for the high abundance of disorder in nucleiomes of species from three domains of life. Therefore, this work provides a strong support to the notion that the nucleic acid binding proteins often require intrinsic disorder for their functions and regulation.

This work was supported in part by the Discovery grant (298328) from the Natural Sciences and Engineering Research Council (NSERC) of Canada and Qimonda Endowed Chair position at the Virginia Commonwealth University to L.K.

The authors have declared no conflict of interest.

5 References

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

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

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

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

[5] Xue, B., Dunker, A. K., Uversky, V. N., 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*, 137–149.

[6] Peng, Z., Yan, J., Fan, X., Mizianty, M. J. et al., Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol. Life Sci.* 2015, *72*, 137–151.

[7] Dunker, A. K., Garner, E., Guilliot, S., Romero, P. et al., Protein disorder and the evolution of molecular recognition: theory,

predictions and observations. *Pac. Symp. Biocomput*. 1998, 473–484.

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

[9] Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M. et al., Intrinsically disordered protein. *J. Mol. Graph Model* 2001, *19*, 26–59.

[10] Tompa, P., Intrinsically unstructured proteins. *Trends Biochem. Sci*. 2002, *27*, 527–533.

[11] Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S., Dunker, A. K., in: Buchner, J., Kiefhaber, T. (Eds.), *Handbook of Protein Folding*, Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany 2005, pp. 271–353.

[12] Dunker, A. K., Obradovic, Z., The protein trinity–linking function and disorder. *Nat. Biotechnol*. 2001, *19*, 805–806.

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

[14] Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z. et al., Intrinsic disorder and functional proteomics. *Biophys. J*. 2007, *92*, 1439–1456.

[15] Vucetic, S., Xie, H., Iakoucheva, L. M., Oldfield, C. J. et al., Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J. Proteome Res*. 2007, *6*, 1899–1916.

[16] Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J. et al., Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J. Proteome Res*. 2007, *6*, 1882–1898.

[17] Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J. et al., Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J. Proteome. Res*. 2007, *6*, 1917–1932.

[18] Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., Dunker, A. K., Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol*. 2002, *323*, 573–584.

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

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

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

[22] Frege, T., Uversky, V. N., Intrinsically disordered proteins in the nucleus of human cells. *Biochemistry Biophysics Reports* 2015, *1*, 33–51.

[23] Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W. et al., Intrinsic disorder in transcription factors. *Biochemistry* 2006, *45*, 6873–6888.

[24] Bhalla, J., Storchan, G. B., MacCarthy, C. M., Uversky, V. N., Tcherkasskaya, O., Local flexibility in molecular function paradigm. *Mol. Cell Proteomics* 2006, *5*, 1212–1223.

[25] Minezaki, Y., Homma, K., Kinjo, A. R., Nishikawa, K., Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. *J. Mol. Biol*. 2006, *359*, 1137–1149.

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

[27] Peng, Z., Oldfield, C. J., Xue, B., Mizianty, M. J. et al., A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol. Life Sci*. 2014, *71*, 1477–1504.

[28] Korneta, I., Bujnicki, J. M., Intrinsic disorder in the human spliceosomal proteome. *PLoS Comput. Biol*. 2012, *8*, e1002641.

[29] Coelho Ribeiro Mde, L., Espinosa, J., Islam, S., Martinez, O. et al., Malleable ribonucleoprotein machine: protein intrinsic disorder in the Saccharomyces cerevisiae spliceosome. *PeerJ*. 2013, *1*, e2.

[30] Varadi, M., Zsolyomi, F., Guharoy, M., Tompa, P., Functional Advantages of Conserved Intrinsic Disorder in RNA-Binding Proteins. *PLoS One* 2015, *10*, e0139731.

[31] Apweiler, R., Bateman, A., Martin, M. J., O'Donovan, C. et al., Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*. 2014, *42*, D191–D198.

[32] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. 2000, *25*, 25–29.

[33] Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I., The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol*. 2005, *347*, 827–839.

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

[35] Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V. et al., DisProt: the database of disordered proteins. *Nucleic Acids Res*. 2007, *35*, D786–D793.

[36] Walsh, I., Giollo, M., Di Domenico, T., Ferrari, C. et al., Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics* 2015, *31*, 201–208.

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

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

[39] Howell, M., Green, R., Killeen, A., Wedderburn, L. et al., 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*. 2012, *20*, 471–511.

[40] Monastyrskyy, B., Kryshtafovych, A., Moult, J., Tramontano, A., Fidelis, K., Assessment of protein disorder region predictions in CASP10. *Proteins* 2014, *82* Suppl 2, 127–137.

[41] Potenza, E., Di Domenico, T., Walsh, I., Tosatto, S. C. E., MobiDB 2.0: an improved database of intrinsically disordered

and mobile proteins. *Nucleic Acids Res*. 2015, *43*, D315–D320.

[42] Di Domenico, T., Walsh, I., Martin, A. J. M., Tosatto, S. C. E., MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics* 2012, *28*, 2080–2081.

[43] Oates, M. E., Romero, P., Ishida, T., Ghalwash, M. et al., D(2)P(2): database of disordered protein predictions. *Nucleic Acids Res*. 2013, *41*, D508–516.

[44] Lobley, A., Swindells, M. B., Orengo, C. A., Jones, D. T., Inferring function using patterns of native disorder in proteins. *PLoS Comput. Biol*. 2007, *3*, 1567–1579.

[45] Pancsa, R., Tompa, P., Structural disorder in eukaryotes. *PloS one* 2012, *7*, e34687.

[46] Peng, Z., Mizianty, M. J., Kurgan, L., Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins* 2014, *82*, 145–158.

[47] Pentony, M. M., Jones, D. T., Modularity of intrinsic disorder in the human proteome. *Proteins-Struct. Funct. Bioinf*. 2010, *78*, 212–221.

[48] Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I. et al., Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays* 2009, *31*, 328–335.

[49] Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J. et al., Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 2005, *44*, 1989–2000.

[50] Linding, R., Jensen, L. J., Diella, F., Bork, P. et al., Protein disorder prediction: implications for structural proteomics. *Structure* 2003, *11*, 1453–1459.

[51] Linding, R., Russell, R. B., Neduva, V., Gibson, T. J., GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. 2003, *31*, 3701–3708.

[52] Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., Obradovic, Z., Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006, *7*, 208.

[53] Yang, Z. R., Thomson, R., McNeil, P., Esnouf, R. M., RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 2005, *21*, 3369–3376.

[54] Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., Vranken, W. F., The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res*. 2014, *42*, W264–270.

[55] Cilia, E., Pancsa, R., Tompa, P., Lenaerts, T., Vranken, W. F., From protein sequence to dynamics and disorder with DynaMine. *Nat. Commun*. 2013, *4*, 2741.

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

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

[58] Peng, Z. L., Kurgan, L., Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci*. 2012, *13*, 6–18.

[59] Kuznetsov, I. B., Gou, Z. K., Li, R., Hwang, S. W., Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins-Struct. Funct. Bioinf*. 2006, *64*, 19–27.

[60] Terribilini, M., Lee, J. H., Yan, C. H., Jernigan, R. L. et al., Prediction of RNA binding sites in proteins from amino acid sequence. *Rna* 2006, *12*, 1450–1462.

[61] Ahmad, S., Sarai, A., Moment-based prediction of DNA-binding proteins. *J. Mol. Biol*. 2004, *341*, 65–71.

[62] Potoyan, D. A., Papoian, G. A., Energy landscape analyses of disordered histone tails reveal special organization of their conformational dynamics. *J. Am Chem. Soc*. 2011, *133*, 7405–7415.

[63] Cheung, P., Allis, C. D., Sassone-Corsi, P., Signaling to chromatin through histone modifications. *Cell* 2000, *103*, 263–271.

[64] Strahl, B. D., Allis, C. D., The language of covalent histone modifications. *Nature* 2000, *403*, 41–45.

[65] Rice, J. C., Allis, C. D., Code of silence. *Nature* 2001, *414*, 258–261.

[66] Dutnall, R. N., Cracking the histone code: one, two, three methyls, you're out! *Mol. Cell* 2003, *12*, 3–4.

[67] Margueron, R., Trojer, P., Reinberg, D., The key to development: interpreting the histone code? *Curr. Opin. Genet. Dev.* 2005, *15*, 163–176.

[68] Nightingale, K. P., O'Neill, L. P., Turner, B. M., Histone modifications: signalling receptors and potential elements of a heritable epigenetic code. *Curr. Opin. Genet. Dev.* 2006, *16*, 125–136.

