Molecular Recognition Features (MoRFs) in three domains of life

Jing Yan,[1] A. Keith Dunker,[2,3*] Vladimir N. Uversky,[4,5,6*] and Lukasz Kurgan[7,1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada, T6G 2V4
[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, USA, 46202
[3]Indiana University School of Informatics, Indianapolis, USA, 46202
[4]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA, 33612
[5]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation, 142292
[6]Biology Department, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah, Kingdom of Saudi Arabia, 21589
[7]Department of Computer Science, Virginia Commonwealth University, Richmond, U.S.A., 23284

[*]Corresponding authors

LK: Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA, USA 23284
E-mail: lkurgan@vcu.edu; phone: 804-827-3986; fax: 804-828-2771

VNU: University of South Florida, 12901 Bruce B. Downs Blvd. MDC07, Tampa, FL, USA 33647
E-mail: vuversky@health.usf.edu; Tel.: 813-974-5816; Fax: 813-974-3757

AKD: Indiana University, 410 W. 10th Street, Suite 5000, Indianapolis, IN, USA 46202
E-mail: kedunker@iu.edu; Tel.: 317-278-9220; Fax: 317-278-9217

This is the author's manuscript of the article published in final edited form as:

Yan, J., Dunker, A. K., Uversky, V. N., & Kurgan, L. (2016). Molecular recognition features (MoRFs) in three domains of life. Molecular BioSystems, 12(3), 697–710. http://doi.org/10.1039/C5MB00640F

Abstract

Intrinsically disordered proteins and protein regions offer numerous advantages in the context of protein-protein interactions when compared to the structured proteins and domains. These advantages include ability to interact with multiple partners, to fold into different conformations when bound to different partners, and to undergo disorder-to-order transitions concomitant with their functional activity. Molecular Recognition Features (MoRFs) are widespread elements located in disordered regions that undergo disorder-to-order transition upon binding to their protein partners. We characterize abundance, composition, and functions of MoRFs and their association with the disordered regions across 868 species spread across Eukaryota, Bacteria and Archaea. We found that although disorder is substantially elevated in Eukaryota, MoRFs have similar abundance and amino acid composition across the three domains of life. The abundance of MoRFs is highly correlated with the amount of intrinsic disorder in Bacteria and Archaea but only modestly correlated in Eukaryota. Proteins with MoRFs have significantly more disorder and MoRFs are present in many disordered regions, with Eukaryota having more MoRF-free disordered regions. MoRF-containing proteins are enriched in the ribosome, nucleus, nucleolus and microtubule and are involved in translation, protein transport, protein folding, and interactions with DNAs. Our insights into the nature and function of MoRFs enhance our understanding of the mechanisms underlying the disorder-to-order transition and protein-protein recognition and interactions. The fMoRFpred method that we used to annotate MoRFs is available at http://biomine.ece.ualberta.ca/fMoRFpred/

Keywords: intrinsically disordered protein; intrinsically disordered region; MoRF; Molecular Recognition Feature; protein-protein interactions; protein-protein recognition.

Introduction

The protein structure-function paradigm, where a specific sequence folds into a specific structure that is responsible for a unique function, served as cornerstone of protein science for more than a century [1, 2]. Research of the past decade and a half has broadened this view of a protein functionality by adding a new player, the class of intrinsically disordered proteins (IDPs), members of which fail to form rigid 3D structures under physiological conditions, either along their entire lengths or in localized regions, but still possess numerous important biological functions [3-9]. Sequences of these IDPs and disordered parts of hybrid proteins possessing ordered domains and intrinsically disordered regions (IDRs) are characterized by a number of specific features that distinguish them from those of ordered proteins and domains and make these IDPs and IDRs predictable [3, 5, 6, 10-12].

Application of computational tools developed for sequence-based intrinsic disorder prediction revealed the wide spread occurrence of IDPs and hybrid proteins containing both structured regions and IDRs within all three domains of life [13-23]. The lack of unique structure under physiological conditions provides IDPs and IDRs with a remarkable set of advantages for certain functions compared to the structured proteins, since the resulting plasticity allows them to efficiently interact with a variety of different targets [4-6, 9]. IDPs are typically involved in pathways that carry out signaling, regulation, and/or control [8, 24, 25], which nicely complements the functional repertoire of ordered proteins that have primarily evolved to carry out small molecule binding, transport and catalytic functions [11]. Several illustrative biological activities of IDPs and IDRs include various roles in transcription and translation, regulation of cell division, signal transduction, storage of small molecules, sites for protein phosphorylation and other posttranslational modifications, chaperone action, and regulation of the self-assembly of large multi-protein complexes such as the ribosome [4-7, 9-11, 24-40].

IDPs and IDRs offer advantages in the context of protein-protein interactions when compared to the structured proteins and domains. One of these functional advantages is the ability of many IDPs/IDRs to undergo disorder-to-order transitions concomitant with their functional activity [4, 6, 8, 24-28, 31, 38, 41-45]. Furthermore, the structural flexibility of a disordered protein or region enables it to interact with numerous partners and to fold into different conformations when bound to different partners [3, 43, 46, 47]. Also, partner selection by IDRs can be modulated by post-translational modifications (PTMs) [47, 48], and such partner binding sites (with or without PTMs) can be added, deleted, or modulated by alternative splicing (AS) of an IDR's pre-mRNA [49]. Thus, tissue-specific PTMs [50, 51] and AS [52, 53] can lead to the "rewiring" of protein-protein interaction networks in different cell types [54, 55]. IDPs and their modulation by PTMs and AS provide a robust mechanism that enables context-dependent signaling [32] that is likely of fundamental importance for cellular differentiation [34-36, 56]. The plasticity of these interactions provides additional functional advantages particularly for signaling and regulation.

A further point is that protein-protein interaction (PPI) networks include proteins, called "hubs," that bind to large numbers of partners while most proteins in the networks bind to only a few [57, 58]. Other networks that have a similar architecture arise because the hubs have special features that facilitate their association with multiple partners and to new partners that come along over time; that is, "the rich get richer" [57]. This raised the question for PPI networks, what are these

special features [59]? As pointed out above, IDPs and IDRs can readily bind to multiple partners, so based on these observations, we proposed that the special features that enabled the evolution of complex networks containing hubs was IDPs and IDRs [25]. This proposal has been supported by a number of subsequent studies [47, 60-65].

Given the importance of IDPs and IDRs for signaling, regulation, and control via PPI pathways and networks, as well as via IDP and IDR involvement in gene pathways and networks, computational methods have been developed to identify the partner binding sites. One approach depends on the identification of short sequence motifs [66], linear motifs [67], eukaryotic linear motifs (ELMs) [68] or short linear motifs (SLiMs) [69]. This approach depends on identifying over-represented sequence patterns found among a collection of different sequences that bind to a common protein partner [47, 66, 68, 69].

An alternative approach was provided by the discovery that some disorder predictors identified localized regions having increased structural propensity. These regions were initially thought to be prediction errors, but instead many of these regions were found to be binding sites for protein partners [70]. Interestingly, this approach actually pre-dated the motif-based methods for finding binding partners. The Protein Data Bank (PDB) contains more than 10,000 complexes containing short peptides bound to globular protein partners. Studies of hundreds of these showed that a large number of these peptides are located in IDPs or IDRs that are predicted to be considerably longer than the segments found in the PDB. Curiously, those that form α-helix or β-sheet upon complex formation were often found to be associated with a local region of predicted structure due to a localized increase in hydrophobicity, whereas those that formed irregular or random structure upon binding, rarely gave strong predictions of localized structure. To indicate their specialized functions within the longer IDRs, these binding segments were called molecular recognition features (MoRFs) [71].

These partner-binding regions contain higher local concentrations of large hydrophobic side chains, especially aromatics, as compared to the flanking IDRs. Furthermore, the PDB structures showed these hydrophobic groups to be mostly buried in the interfaces between the IDPs or IDRs and their partners. Even though the random-structured MoRFs (forming coils and/or turns upon binding), or γ-MoRFs, show weaker predictions of structure and reduced hydrophobicity compared to the helix-forming, or α-MoRFs, and the sheet-forming, or β-MoRFs, the γ-MoRF hydrophobic side chains are, if anything, more selectively buried in their respective interfaces [47, 72].

Using collections of MoRFs from the PDB ranging from 5 to 25 residues in length, predictors of α-MoRFs were developed [47, 73, 74]. This work focused initially on α-MoRFs because they typically give strong predictions of local structure flanked by predictions of IDRs. Binding regions longer than ~ 30 residues arising from IDRs have been called disordered binding domains [75]; these often contain subregions that are identified by MoRF predictors (unpublished observations). Eventually, predictors that could recognize all types of binding regions, including those that form irregular structure upon binding, were developed. They include ANCHOR [76, 77], DISOPRED3 [78], MoRFCHiBi [79], MoRFpred [80], and fMoRFpred, which was developed as part of this study with the aim to offer accurate predictions in high throughput. These methods use the PDB protein complexes for training.

While their details differ, these predictors identify binding sites by their increased hydrophobicity and reduced propensity for disorder compared to the flanking regions of disorder. Although these feature-based algorithms and the motif-based approach for finding partner binding sites are completely different, both approaches typically identify binding sites that are located in IDRs [67, 81]. In recognition of this, the MoRFpred algorithm includes sequence similarity to any of the MoRFs in its training set as one of the inputs. This input serves as a surrogate for the use of motifs. Others have studied such regions from a different perspective. Many proteins appear in the PDB more than once with the structure determinations carried out under (slightly) different conditions. Such proteins often have regions that are structured under one set of conditions but are disordered under the other. These ambiguous [82] or dual-personality [83] or semi-disordered [84] regions exhibit disorder predictions (and hydrophobicities) that are intermediate between the extremes observed for structured and disordered proteins [83, 84].

Experimentalists can use binding-site predictors to speed-up the process of PPI discovery [85-87], and, indeed, MoRF predictions have been used for this purpose [88-90]. In [90], the yeast-two hybrid method [91] was followed by mutational analysis to identify both the partners of the MoRFs and the MoRF residues essential for partner binding, suggesting we are now in a position to study MoRF-partner interactions by high throughput methods. The first step for such high throughput studies of disorder-based PPIs is MoRF prediction (and/or binding motif identification) on a large scale.

However, to date only a few studies have investigated properties and abundance of a larger set of MoRFs. In 2006, Mohan and colleagues performed analysis of secondary structure, amino acid composition, aromaticity and charge, and a limited functional analysis of a relatively small set of 372 MoRFs derived from PDB [71]. In 2007, geometric and physiochemical properties of the surface of the corresponding binding regions for 258 MoRFs collected from the PDB were examined [72]. A recent study investigated MoRFs in a small set of 289 membrane proteins from PDB [92]. There were only two studies that analyzed MoRFs on genomic scale. The prevalence of α-MoRFs generated by the α-MoRF predictor was estimated in 82 genomes from Eukaryota, Bacteria and Archaea and the authors observed that a median eukaryotic genome has greater fraction of proteins with α-MoRF propensities than median archaeal and bacterial genomes [74]. More recently, analysis of 736 complete proteomes that took advantage of the ANCHOR method was performed; however, it was limited to the characterization of abundance and length of these binding regions [76]. To this end, herein we present our analysis of the 868 complete proteomes from the three domains of life. We consider multiple perspectives including 1) abundance of MoRFs and their types; 2) relation between abundance of MoRFs and IDRs; 3) enrichment of disorder in MoRF-containing proteins; 4) compositions of MoRF, intrinsically disordered and structured regions; and 5) functions of MoRF-containing proteins. Our analysis across different species and domains of life points to interesting and distinct differences between MoRF and generic IDRs. These data provide experimentalists with the starting points for the high throughput analysis of MoRF-based PPIs for any of these organisms.

Materials and Methods

We analyze putative MoRFs and IDRs in the complete proteome set from UniProt release of April 2013 [93]. This dataset includes 174,381 protein sequences from 72 species in Archaea, 2,025,100 sequences from 567 species in Bacteria and 3,645,837 sequences from 229 species in Eukaryota (Table 1).

Only high-throughput methods that find putative MoRF and IDRs could be used given the size of the UniProt dataset. We apply fMoRFpred predictor to find putative MoRFs; the design and predictive performance of this method are discussed in the Supplement. fMoRFpred uses a similar design to the popular MoRFpred method (see Supplement) [80]. In short, each residue in the input protein sequence is represented by 20 features that are derived from structural, physicochemical and biochemical properties of this and its neighboring residues; these features were empirically selected as the most predictive from a comprehensive group of over 7000 features. The predictive properties include putative annotation of intrinsic disorder and secondary structure, estimated B-factor, structural stability, and unfolding energy. They allow identifying MoRF regions since these regions are enclosed inside of longer disordered regions, may fold into secondary structures upon binding, and are characterized by a relatively high flexibility (B-factor) and lower structural stability as compared to the structured regions. The prediction is performed with Support Vector Machine model that uses the features as inputs and which was trained using a large dataset with annotated MoRFs to optimize separation of its output values between MoRF and non-MoRF residues. fMoRFpred is shown to provide accurate estimates of abundance of MoRFs via comprehensive tests that utilize several benchmarking datasets, which include chains with low similarity to the training proteins. This means that it can be used to accurately predict MoRFs on the whole proteome scale. It is also characterized by a relatively low runtime, which allows for the genome-scale predictions on a single desktop computer. The predictive quality of MoRF predictors was also validated against experimental results in a few applications [90, 94], supporting the claim that they provide accurate results. A webserver-based implementation of fMoRFpred is available at http://biomine.ece.ualberta.ca/fMoRFpred/.

IDRs were predicted using a consensus of five high-throughput predictors: two versions of IUPred [95] designed to find long and short IDRs and three version of Espritz [96] that predict intrinsic disorder annotated based on X-ray crystal structures, the NRM-derived structures, and the Disprot database. Thus, the consensus considers two types of IDRs (short regions and longer domains) and three dominant types of annotations. These methods were shown to provide good predictive performance in a recent large-scale assessment [97]. A given residue is predicted as disordered if at least three of the five methods predict so; otherwise it is predicted as structured (ordered). The same consensus was recently used in related works [23, 98]. Use of the consensus is an improvement over some prior studies where only one or at most two methods were used [14, 16, 99]. MoRF and disorder predictions were filtered by removing segments with less than 4 consecutive residues, which is in agreement with other studies [23, 80, 100]. The secondary structure, which is used to define different types of MoRFs, was predicted using the fast version of the PSIPRED method [101].

We characterize abundance of MoRF and disordered residues and regions and aggregate this information by species and by domains of life. We compute content of MoRF and disordered

residues which is defined as the number of MoRF or disordered residues in a given sequence, species or domain of life divided by the total number of residues. We analyze normalized number and size of MoRFs and IDRs. The number of regions was normalized per sequence (the count was divided by the total number of proteins in a given dataset) while the size of the regions was normalized by dividing their length by the length of the corresponding proteins. We also investigate MoRFs at the whole protein level. We compute the disorder content and fraction of fully disordered protein (proteins composed of only disordered residues) in proteins that contain MoRFs, that contain IDRs, and in all proteins in a given species or domain of life. Finally, we calculate content of amino acid defined as the count of residues for a given amino acid type divided by the total number of residues. These content values were compared between MoRFs, IDRs, structured (ordered) regions, and a generic (drawn at random from protein sequences) set of regions.

We assess statistical significance of differences between content values or normalized counts between two protein sets. We select at random 1000 samples (proteins or residues) from a given set of proteins (e.g., Eukaryotic proteins with MoRFs), calculate a given characteristic (content or count) and repeat that 10 times. The resulting vector of 10 values is compared with the corresponding vector of 10 values computed from the second proteins set (e.g., Eukaryotic proteins with IDRs). We determine normality of these values with the Anderson-Darling test at the 0.05 significance. We use the *t*-test for normal distributions; otherwise we use the Wilcoxon rank-sum test. We assume that the difference is significant if p-value < 0.01. We also report average and standard deviation over the 10 repetitions if data are normal, and median with 25^{th} and 75^{th} centiles otherwise.

We carried analysis of functional annotations of proteins that have MoRFs based on the Gene Ontology (GO) terms collected from the UniProt resource. We utilize statistical test to find annotations that are significantly enriched in these proteins when compared with a generic set of proteins from the same domain of life. We consider annotations of biological processes and cellular components that indicate cellular localization of the MoRF-including proteins. We compute significance of enrichment for each annotation that occurs at least 20 times in the proteins with MoRFs (to assure statistically sound estimates) and which has the rate of occurrence (defined as number of occurrences divided by the number of proteins) that is higher than the rate in the whole domain of life. We select 50% of the MoRF-including proteins at random ten times and compute the rates of occurrence for these 10 sets of proteins. Next, we select 10 times the same number of proteins with matching chain sizes (with tolerance of 10%) at random from the entire domain of life and calculate the corresponding rates. The matching is motivated by a bias in disorder content related to chain sizes [98], which in turn influences abundance of MoRFs. We compare the two sets of 10 rates of occurrence using either the *t*-test or the Wilcoxon rank-sum test, depending on the normality of these samples. A given GO term is assumed to be enriched in proteins that have MoRFs if the rate of occurrence in these proteins is higher by at least 20% compared with the proteins drawn at random and the p-value < 0.01.

Results and discussion

Overall disorder status of proteins in three domains of life

In order to provide background needed for the subsequent analysis, we analyzed the peculiarities of the distributions of protein length and correlation between the disorder content and protein length for the considered close to 6 million proteins of the 868 species from Archaea, Bacteria and Eukaryota. Results of these analyses are shown in Figures 1A and 1B, respectively, and they demonstrate that eukaryotic proteins are different from bacterial and archaean proteins, being typically longer and noticeably more disordered. These observations are in agreement with the results of previous studies [13-23]. We emphasize the peculiar shape of the disorder content *versus* protein length plot for the eukaryotic proteins. While short proteins in three domains of life are consistently predicted to have significant amount of disorder, longer eukaryotic proteins have a much different profile of disorder content when compared to the almost coinciding plots for the bacterial and archaean proteins (see Figure 1B). The amount of predicted disorder in the bacterial and archaean proteins decreases as protein length increases and reaches a plateau for proteins longer than about 300 residues. However, in eukaryotes the disorder content reaches a minimum for proteins with the length range between 250 and 500 residues and then it substantially increases and reaches a plateau for proteins with length at about 1000 or more residues. This peculiar shape of the disorder content *versus* length of eukaryotic proteins has been described earlier in a study that used smaller dataset (110 complete eukaryotic proteomes) [98]. In other words, this analysis revealed that medium sized eukaryotic proteins (length between 250 and 500 residues) possess smaller amount of predicted disorder than shorter and longer proteins.

Abundance of MoRF and intrinsically disordered regions in the three domains of life

We estimate the abundance of putative MoRFs and IDRs across the 868 species from Archaea, Bacteria and Eukaryota. The abundance is based on content of residues located in the IDRs and in the MoRFs; i.e., fraction of disordered and MoRF residues among all residues in a given species or domain of life. The results shown in Table 1 suggest that MoRF residues have similar content of about 1% across all three domains of life. This is in contrast to the content of the disordered residues that vary widely with lowest values in Bacteria and substantially higher values in Eukaryota (Figure 1B), which was also shown in other studies [14, 16, 23].

Statistical tests reveal that the differences in the per species disorder content between different domains of life are significant (*t*-test; degrees of freedom = 9; *p*-value<0.01); i.e., eukaryotic organisms have significantly larger disorder content than species in Archaea, which in turn have significantly larger content values than species in Bacteria. The disorder content in Archaea is bimodal (see Figure 2A, blue triangles), with some species at the low end and others above the high end of the bacterial range. The Archaea with high predicted disorder are mostly halophiles [21] that live in saturated salt and have high internal salt concentrations. To accommodate these high salt concentrations, the non-membrane proteins develop an excess of negative charges on their surfaces, leading to a stabilizing shell of cations, and a reduced amount of hydrophobic residues [102]. This leads to high prediction of disorder [40] even though they are structured in their

high salt environment. Indeed, many enzymes and other proteins from halophiles become disordered if they are transferred from high salt to the typical "physiological" range [102].

The content of MoRF residues is not significantly different between the species from three domains of life. Interestingly, we found that the content of disordered residues and content of MoRF residues are strongly correlated in Bacteria and Archaea. The corresponding Pearson Correlation Coefficients (PCCs) equal 0.89 and 0.98, respectively. However, the abundance of disorder and MoRFs in Eukaryotes is characterized by only modest correlation of 0.43.

A plot of content of intrinsically disordered residues *vs*. MoRF residues for the considered species is shown in Figure 2A. The disorder content in Bacteria and Archaea is constrained to a relatively narrow range of up to about 15% with the MoRF content ranging between 0.5 and 2%. A linear trend where more disorder implies proportionally more MoRFs is evident for species from these two domains. In Eukaryotes, the disorder content is on average higher and varies more widely between about 5 and 30%, which is agreement with prior results [14, 23, 103]. Surprisingly, the content of MoRF residues is constrained to the range that is similar to the range in Bacteria and Archaea and its linear relation with the disorder content is weaker. Overall, Figure 2A reveals that content of MoRF residues is similar across the species from each of the three domains of life. On the other hand, Figure 2B represents a histogram of the fraction of proteins containing different number of MoRFs per protein and shows that, on average, eukaryotic proteomes have substantially more multi-MoRF proteins.

Figure 3 shows differences in the relation between the abundance of MoRFs and protein length in the three domains of life. Figure 3A illustrates that the relations for the bacterial and archaean proteins have very similar shapes, where short proteins have more MoRFs per protein than long proteins. In these two domains of life, the smallest per-protein counts of MoRFs (~0.4 MoRF per protein) are found for the medium-length proteins (200-400 residues), whereas these numbers slightly increase to ~0.5 for the longer proteins. The corresponding relation for the eukaryotic proteins is very different. Although short eukaryotic proteins have more MoRFs (0.7 MoRFs per protein) and although this number decreases to ~0.55 for the medium-length proteins (~200 residues), the number of MoRFs per protein increases steadily for proteins longer than 200 residues. Moreover, the long eukaryotic proteins which are abundant in this domain of life (Figure 1A) clearly contain more MoRFs than short ones (Figure 3A). One average, proteins that are 1000 or more residues long have one MoRF region, which means that one of their protein domains is involved in protein-protein interactions via intrinsic disorder. Although the number of MoRFs per proteins is higher for the large proteins, the content of MoRF residues decreases as the protein size grows. Figure 3B visualizes the corresponding relation between the number of MoRFs versus protein length plots calculated on the per-residue basis. This trend is very different from the trend of disorder content for eukaryotes (Figure 1B). Although the disorder content in eukaryotic proteins is higher for long proteins, the content of MoRF residues is lower. This analysis supports the idea that proteins in all domains of life are characterized by similar abundance of MoRFs.

Figure 4 shows distribution of the per-species MoRF and disorder content values grouped by the second level in taxonomy that corresponds to kingdoms or phyla. Boxplots, which show spread of content values in species from a given kingdom/phylum, are grouped and colored by the

corresponding domain of life and sorted in the descending order by the median content of MoRFs. We observe a clear trend in Archaea and Bacteria where the median disorder content in a given kingdom/phylum follows the median MoRF content. This is not the case in Eukaryotes, where additionally the content of MoRFs is substantially lower than the content of disordered residues.

Abundance of different types of MoRFs in the three domains of life

The abundance of different types of MoRFs including α-MoRFs, β-MoRFs, γ-MoRFs, and complex-MoRFs (which fold into a mixture of helices and strands upon binding) in the three domains of life is summarized in Figure 5. We show that MoRFs fold into secondary structures with similar proportions irrespective of the taxonomic classification. The largest fraction of MoRFs become structured as coils (71% to 79% of MoRFs depending on the domain of life). Between 16% and 21% of MoRFs establish α-helix conformation upon binding. The higher proportion of α-MoRFs in Eukaryota compared to Archaea and Bacteria is consistent with prior observations [73, 74]. We note that the overall content of α-MoRF residues in Eukaryota which we estimate to be 0.22% is similar the estimate of 0.28% from the contribution that analyzed these types of MoRFs [73, 74]. Figure 5 also reveals that β- and complex-MoRFs account for a relatively low fraction of 6 to 8% of MoRFs.

Relation between MoRF and intrinsically disordered regions in the three domains of life

We analyze abundance, size and localization in the sequence of the IDRs and divide them into those that include one or multiple MoRFs and those that are free of MoRFs. Figures 6A and 6B show the number of IDRs per protein and fraction of IDRs that have one, multiple, and no embedded MoRFs for each domain of life. These characteristics are similar in Archaea and Bacteria with on average approximately one IDR without MoRFs per protein, one MoRF region in every other protein, and 30% of the IDRs having MoRFs. However, Eukaryotic proteins are different and have substantially more intrinsically disordered regions without MoRFs (close to 2.5 per protein) and a similar number of regions with MoRFs. Moreover, nearly 80% of IDRs in Eukaryota have no MoRFs. This shows that Eukaryotic species have evolved to introduce additional, MoRF-free disordered regions when compared with the other two domains of life. Figure 6C compares sizes of IDRs that are normalized by the size of the corresponding proteins. We observe that these values are similar between the three domains of life, which suggests that the difference in the number of disordered regions without MoRFs is not driven by the difference in the size of the disordered regions. Also, IDRs that have MoRFs are longer (Figure 6C) and this is particularly true for the small fraction of regions that ranges between 0.3% (in Bacteria) and 1.2% (in Eukaryota) (Figure 6B) that include multiple MoRFs. Figure 6D summarizes localization of IDRs in the protein sequences. While most IDRs in Archaea and Bacteria are localized at the termini of the sequence, this bias is reversed in Eukaryota where 60% of the disordered regions are inside the chain. The IDRs that include MoRFs are located almost exclusively at the termini in Archaea and Bacteria and similarly in Eukaryota only about 10% of these regions are located inside the protein sequence.

To sum up, our analysis reveals that the intrinsically disordered regions that have MoRFs have similar characteristics across the three domains of life, while the enrichment in disorder in Eukaryotes is driven by inclusion of MoRF-free disordered regions which are biased to be localized inside the protein chains.

Proteins with MoRFs are significantly enriched in disorder

Figure 7 shows distribution of disorder content among proteins that include MoRFs (solid lines), intrinsically disordered region(s) (dashed lines) and all proteins (dotted lines) partitioned according to their taxonomic domain (denoted by colors). Comparison of distributions for the proteins with MoRFs and with IDRs (solid vs dashed lined of the same color in Figure 7) reveals that the former are depleted for chains with low disorder content (< 0.1) and enriched in chains with higher disorder content (>0.1). This enrichment is higher (relative to the value for the proteins with IDRs) as the amount of disorder increases.

Overall, the MoRF-including proteins are characterized by a substantially higher amount of disorder compared to proteins with IDRs and all proteins universally across the three domains of life. This result is consistent with Figure 6C that shows that IDRs that include MoRFs are longer compared with the regions that do not. We analyze statistical significance of the differences in the disorder content and in the fraction of fully disordered proteins (i.e., proteins composed entirely of disordered residues) between the three sets of proteins. For each domain, we compare the disorder content and the fraction of fully disordered proteins that is calculated 10 times, each time using 1000 randomly chosen proteins from the set of all proteins, proteins with MoRFs, and proteins with IDRs. Figure 8 shows the average and standard deviation of these 10 measurements for the disorder content (panel A) and for the fraction of the fully disordered proteins (panel B) and the *p*-values associated with the differences between proteins with MoRFs, with IDRs and all proteins. The enrichment of the disorder in the MoRF-containing proteins is significant when compared with all proteins and with proteins with the IDRs. Moreover, the fraction of fully disordered proteins among proteins that has MoRFs, which is between 1% in Archaea and Bacteria and close to 2% in Eukaryota, is also significantly higher. The differences in the disorder content and in the fraction of fully disordered proteins are significant in all three domains of life.

Amino acid composition of MoRF and disordered regions

Biases in the amino acid composition in MoRFs and two of their types: α-MoRFs and β-MoRFs, in IDRs and in structured regions are summarized in Figure 9. The amino acids are sorted by the average (over the three domains of life) differences in composition between the MoRF residues and a generic set of residues selected at random; note that similar prior plots for IDRs/IDPs were sorted by the flexibility index [104, 105]. Figure 9 shows the averages and standard deviations of the differences between composition of MoRF/disordered/structured residues and the composition of the generic residues over the 10 repetitions of measurements of content for each of the 20 amino acid types; details are provided in Materials and Methods section. Solid bars indicate amino acids that have significantly different (enriched or depleted) composition for a given set of residues. The patterns of the enrichment and depletion of amino acids in the MoRFs (Figure 9C)

and in the IDRs (Figure 9D) are consistent across different domains of life. The correlation coefficients between these difference values are relatively high and equal 0.82, 0.81, and 0.72 for Archaea, Bacteria and Eukaryota, respectively. Several amino acids are enriched (Proline and Glutamine) and depleted (Cysteine, Phenylalanine, Isoleucine, Leucine, and Valine) in both IDRs and MoRFs in each of the three domains. However, we also found that Methionine and Threonine that are enriched in IDRs are no longer enriched in the MoRFs.

The observed biases in the amino acid composition in the MoRFs in comparison with structured regions are in line with the notion that the majority of MoRFs are γ-MoRFs (Figure 5). This MoRF type should not be too different from "general" IDRs which also often lack propensity to form secondary structures. This is supported by the fact that according to our analysis both IDRs and MoRFs are enriched in a couple of major disorder-promoting residues (Proline and Glutamine), typically contain in abundance some other disorder-promoting residues (Glutamic acid and Serine), and are depleted in major order-promoting residues (Cysteine, Phenylalanine, Isoleucine, Leucine, and Valine). However, analysis of the α- and β-MoRFs reveals a more substantial difference from IDRs (Figure 9A). The α-MoRFs are significantly enriched in Arginine, Glutamate, Lysine, and Glutamine which are enriched to a lesser extend in IDRs and have a relatively high propensity to form helical conformations. The β-MoRFs are enriched in Valine, Isoleucine and Methionine. The former two are depleted in IDRs and all three amino acid types have a relatively high propensity for formation of strands. Both, the α- and β-MoRFs are depleted in Proline and Glycine, which are considered as disorder-promoting residues [10, 106], are known as major structure breaker residues and are commonly found at the ends of regular secondary structure elements [107]. For example, since Proline peptide bonds exhibit structural features that differ substantially from those of other residues, also because they do not contain backbone amide hydrogen atoms at physiological pH, they do not form stabilizing hydrogen bonds in α-helices or β-sheets [108, 109]. Curiously, eukaryotic β-MoRFs are moderately enriched in Tryptophan, Tyrosine, and Phenylalanine; i.e., residues known to be commonly involved in specific interactions [110]. Also, aromatic residues are crucial for folding and stability of proteins, and Tryptophan-Tryptophan pairs were, for example, shown to contribute more than any other hydrophobic interaction to the stability of β-hairpins [111]. These results suggests that β-MoRFs might use aromatic residues to be specifically zipped to their binding partners.

The residues in the structured regions (Figure 9E) are characterized by lack of significant differences, which could be explained by the fact that majority of residues (at least 80% as shown in Table 1) are structured. Moreover, the biases are consistent between different domains of life. The correlation coefficients between the differences for the same group of residues between Eukaryota, Bacteria, and Archaea are high are range between 0.76 and 0.96.

Functional analysis of proteins with MoRFs

Using Gene Ontology (GO) annotations associated with proteins that have MoRFs we extracted cellular component and biological process that are significantly enriched (*t*-test or Wilcoxon test (see Figure 10); degrees of freedom = 9; p-value < 0.01) in these proteins. This analysis was performed separately for each domain of life (Figure 10). We compared rate of occurrence of a given annotation (defined as number of occurrences divided by the number of proteins) between proteins with MoRFs and proteins selected at random from the same domain of life; details are

given in Materials and Methods section. We only consider annotations with a large rate of occurrence in the MoRF-containing proteins (> 0.25%) for which the relative increase when compared with the rate in the random protein is at least 20%.

The results reveal that the ribosomes in all three domains of life are enriched in proteins with MoRFs. Furthermore, these proteins are also very common in the nucleus, nucleolus and microtubule in the Eukaryota. This is consistent with the results that were obtained for α-MoRFs in the human genome, which were found to be enriched in the ribosome and cytoskeleton [74]. Analysis of biological processes shows that MoRF-containing proteins are involved in translation, protein transport, protein folding, and interactions with DNAs. They are also enriched among eukaryotic proteins that are associated with the regulation of transcription. This is in line with the results of a smaller study that considered about 200 proteins with MoRFs and pointed to their enrichment in DNA binding and regulation of transcription [71]. Importantly, we found that similar GO terms are enriched across the three domains of life, which is consistent with our other finding related to the similar levels of abundance of MoRFs in nature.

Conclusions

The protein-protein interactions are crucial for many biological processes which rely on protein-centric recognition, regulation and signaling interactions. Therefore, understanding molecular mechanisms underlying such interactions is directly linked to gaining critical insights into signaling and regulation within biological systems. Furthermore, on the practical side, better understanding of the molecular mechanisms defining these interactions might enable the development of small molecule therapies that could be used to modulate protein-protein interactions and thereby target various human diseases [112-116].

IDPs/IDRs are known to be promiscuous binders that play different roles in regulation of the function of their binding partners and in promotion of the assembly of supra-molecular complexes [45]. The conformational plasticity associated with intrinsic disorder provides IDPs/IDRs with a wide spectrum of exceptional functional advantages over the functional modes of ordered proteins and domains [4, 6, 9, 25, 27, 28, 31, 32, 37, 38, 43, 46]. Many IDPs/IDRs are known to contain specific identification regions via which they are involved in various regulation, recognition, signaling and control pathways [25, 31]. IDPs/IDRs can form highly stable complexes, and can be involved in signaling interactions where they constantly cycle between bound and unbound forms, thus acting as dynamic and sensitive “on-off” switches. The ability of these proteins to return to the highly flexible conformations after the completion of a particular function, and their predisposition to gain different conformations depending on the environmental peculiarities, are unique physiological properties of IDPs which allow them to exert different functions in different cellular contests according to a specific conformational state [9, 117]. The action of IDPs is further modulated by extensive posttranslational modifications [6, 48] and by alternative splicing [49]. IDPs/IDRs are commonly involved in various human diseases where they often play central roles [10]. As a result, IDPs and hybrid proteins possessing ordered domains and IDRs represent attractive but very difficult drug targets [118-122].

An important first step in developing new drugs targeting protein-protein interactions is the ability to predict such interactions from sequence and structure. For ordered proteins, combination of structural knowledge with evolutionary information provides means for the successful predictions of both binding regions and binding partners from known protein structure [123-126]. The situation is more complicated with IDPs/IDRs, since they do not have unique structures suitable for structure-based rational drug design. However, the known ability of many IDPs/IDRs to undergo a disorder-to-order transition upon binding to their partners [4, 6, 8, 24-28, 31, 38, 41-44] combined with the fact this disorder-to-order-transition-based recognition is commonly mediated by short specific MoRF elements [71-74], simplifies the task of finding such disorder-based binding regions from sequence alone.

In this work, we developed a novel accurate and fast computational tool, fMoRFpred, for finding all types of MoRFs. This tool and the consensus of five high-throughput predictors of intrinsic disorder were applied to analyze putative MoRFs and MoRF-free IDRs in over 800 species. Various sequence features of these regions in the three domains of life were compared. Functions of MoRF-containing proteins were also analyzed based on the Gene Ontology (GO) terms collected from the UniProt resource.

Our work demonstrates that MoRFs are similarly abundant across the three domains of life and are enriched in the same amino acid types. In fact ~21% of IDRs in Eukaryota and ~29% in Bacteria and Archaea have MoRFs and these MoRF-containing regions are substantially longer than the MoRF-free disordered regions. In Bacteria and Archaea, there is a strong correlation between the abundance of MoRFs and the amount of intrinsic disorder in corresponding proteomes. This correlation is much less pronounced in eukaryotic proteins that have twice as many MoRF-free IDRs compared to Archaea and Bacteria. This observation can be explained by the fact that eukaryotic proteins are noticeably more disordered than bacterial and archaean proteins suggesting that the enrichment in disorder in Eukaryotes is driven by inclusion of MoRF-free disordered regions which have a bias to be localized inside the protein chains. One possibility that might explain the higher incidence of MoRF-free IDRs in eukaryotes is that eukaryotic IDRs have many additional protein interaction sites that are different from MoRFs. They could be found via alternative approaches that rely on short conserved regions that were identified by their binding to the same partner. These have been called eukaryotic linear motifs (ELMs) [68] or short linear motifs (SLiMs) [127], both of which are typically found in IDRs [128].

Importantly, our analysis enriches current knowledge of the PPI networks which treat proteins as whole entities. We show that these interactions are relatively often driven by disordered regions that fold upon binding; that some proteins, particularly in Eukaryota (Figure 2B), have multiple such MoRF regions; and that one average every long eukaryotic protein has at least one MoRF (Figure 3A).

Moreover, our functional analysis reveals that, in all three domains of life, MoRF-containing proteins are commonly found in ribosomes and are involved in translation, protein transport, protein folding, and interactions with DNA. Eukaryotic MoRF-containing proteins can also be found in the nucleus, nucleolus, and microtubule and can be related to the regulation of transcription. Our large scale analysis of the abundance and peculiarities of MoRFs provides new insights into the nature and function of MoRFs and enhances our knowledge of the mechanisms

underlying the disorder-to-order transition related to the protein-protein recognition and interaction.

References

1. E. Fischer, *Ber. Dt. Chem. Ges.*, 1894, **27**, 2985-2993.
2. U. R. Lemieux and U. Spohr, *Adv. Carbohydrate Chem. Biochem.*, 1994, **50**, 1-20.
3. A. K. Dunker, E. Garner, S. Guilliot, P. Romero, K. Albrecht, J. Hart, Z. Obradovic, C. Kissinger and J. E. Villafranca, *Pac Symp Biocomput*, 1998, 473-484.
4. P. E. Wright and H. J. Dyson, *J Mol Biol*, 1999, **293**, 321-331.
5. V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins*, 2000, **41**, 415-427.
6. A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner and Z. Obradovic, *J Mol Graph Model*, 2001, **19**, 26-59.
7. P. Tompa, *Trends Biochem Sci*, 2002, **27**, 527-533.
8. G. W. Daughdrill, G. J. Pielak, V. N. Uversky, M. S. Cortese and A. K. Dunker, in *Handbook of Protein Folding*, eds. J. Buchner and T. Kiefhaber, Wiley-VCH, Verlag GmbH & Co. KGaA, Weinheim, Germany, 2005, pp. 271-353.
9. V. N. Uversky and A. K. Dunker, *Biochimica et biophysica acta*, 2010, **1804**, 1231-1264.
10. A. Campen, R. M. Williams, C. J. Brown, J. Meng, V. N. Uversky and A. K. Dunker, *Protein Pept Lett*, 2008, **15**, 956-963.
11. P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky and A. K. Dunker, *Biophys J*, 2007, **92**, 1439-1456.
12. P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins*, 2001, **42**, 38-48.
13. A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Inform Ser Workshop Genome Inform*, 2000, **11**, 161-171.
14. J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton and D. T. Jones, *J Mol Biol*, 2004, **337**, 635-645.
15. V. N. Uversky, *J Biomed Biotechnol*, 2010, **2010**, 568068.
16. B. Xue, A. K. Dunker and V. N. Uversky, *J Biomol Struct Dyn*, 2012, **30**, 137-149.
17. P. Romero, Z. Obradovic, C. R. Kissinger, J. E. Villafranca, E. Garner, S. Guilliot and A. K. Dunker, *Pac Symp Biocomput*, 1998, 437-448.
18. Z. P. Feng, X. Zhang, P. Han, N. Arora, R. F. Anders and R. S. Norton, *Mol Biochem Parasitol*, 2006, **150**, 256-267.
19. P. Tompa, Z. Dosztanyi and I. Simon, *J Proteome Res*, 2006, **5**, 1996-2000.
20. C. A. Galea, A. A. High, J. C. Obenauer, A. Mishra, C. G. Park, M. Punta, A. Schlessinger, J. Ma, B. Rost, C. A. Slaughter and R. W. Kriwacki, *Journal of proteome research*, 2009, **8**, 211-226.
21. B. Xue, R. W. Williams, C. J. Oldfield, A. K. Dunker and V. N. Uversky, *BMC Syst Biol*, 2010, **4 Suppl 1**, S1.
22. P. V. Burra, L. Kalmar and P. Tompa, *PLoS One*, 2010, **5**, e12069.
23. Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky and L. Kurgan, *Cellular and molecular life sciences : CMLS*, 2014, DOI: 10.1007/s00018-014-1661-9.

24. L. M. Iakoucheva, C. J. Brown, J. D. Lawson, Z. Obradovic and A. K. Dunker, *J Mol Biol*, 2002, **323**, 573-584.
25. A. K. Dunker, M. S. Cortese, P. Romero, L. M. Iakoucheva and V. N. Uversky, *FEBS Journal*, 2005, **272**, 5129-5148.
26. A. K. Dunker and Z. Obradovic, *Nat. Biotechnol.*, 2001, **19**, 805-806.
27. A. K. Dunker, C. J. Brown and Z. Obradovic, *Adv. Protein Chem.*, 2002, **62**, 25-49.
28. A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva and Z. Obradovic, *Biochemistry*, 2002, **41**, 6573-6582.
29. V. N. Uversky, *Protein Sci.*, 2002, **11**, 739-756.
30. V. N. Uversky, *Eur. J. Biochem.*, 2002, **269**, 2-12.
31. V. N. Uversky, C. J. Oldfield and A. K. Dunker, *J. Mol. Recognit.*, 2005, **18**, 343-384.
32. A. K. Dunker, I. Silman, V. N. Uversky and J. L. Sussman, *Curr. Opin. Struct. Biol.*, 2008, **18**, 756-764.
33. A. K. Dunker, C. J. Oldfield, J. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic and V. N. Uversky, *BMC Genomics*, 2008, **9 Suppl 2**, S1.
34. S. Vucetic, H. Xie, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1899-1916.
35. H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky and Z. Obradovic, *J. Proteome Res.*, 2007, **6**, 1882-1898.
36. H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, Z. Obradovic and V. N. Uversky, *J. Proteome Res.*, 2007, **6**, 1917-1932.
37. M. S. Cortese, V. N. Uversky and A. K. Dunker, *Prog Biophys Mol Biol*, 2008, **98**, 85-106.
38. H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell. Biol.*, 2005, **6**, 197-208.
39. P. Tompa, *FEBS Lett.*, 2005, **579**, 3346-3354.
40. Z. Peng, C. J. Oldfield, B. Xue, M. J. Mizianty, A. K. Dunker, L. Kurgan and V. N. Uversky, *Cell Mol Life Sci*, 2014, **71**, 1477-1504.
41. G. E. Schulz, in *Molecular Mechanism of Biological Recognition*, ed. M. Balaban, Elsevier/North-Holland Biomedical Press, New York, 1979, pp. 79-94.
42. B. W. Pontius, *Trends Biochem Sci*, 1993, **18**, 181-186.
43. H. J. Dyson and P. E. Wright, *Curr Opin Struct Biol*, 2002, **12**, 54-60.
44. K. W. Plaxco and M. Gross, *Nature*, 1997, **386**, 657, 659.
45. M. Fuxreiter, A. Toth-Petroczy, D. A. Kraut, A. T. Matouschek, R. Y. Lim, B. Xue, L. Kurgan and V. N. Uversky, *Chem Rev*, 2014, **114**, 6806-6843.
46. C. J. Oldfield, J. Meng, J. Y. Yang, M. Q. Yang, V. N. Uversky and A. K. Dunker, *BMC Genomics*, 2008, **9 Suppl 1**, S1.
47. W. L. Hsu, C. J. Oldfield, B. Xue, J. Meng, F. Huang, P. Romero, V. N. Uversky and A. K. Dunker, *Protein science : a publication of the Protein Society*, 2013, **22**, 258-273.
48. V. Pejaver, W. L. Hsu, F. Xin, A. K. Dunker, V. N. Uversky and P. Radivojac, *Protein Sci*, 2014, **23**, 1077-1093.
49. P. R. Romero, S. Zaidi, Y. Y. Fang, V. N. Uversky, P. Radivojac, C. J. Oldfield, M. S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic and A. K. Dunker, *Proc Natl Acad Sci U S A*, 2006, **103**, 8390-8395.
50. R. H. Oakley and J. A. Cidlowski, *J Biol Chem*, 2011, **286**, 3177-3184.
51. K. Vaidyanathan and L. Wells, *J Biol Chem*, 2014, **289**, 34466-34471.
52. A. Kalsotra and T. A. Cooper, *Nat Rev Genet*, 2011, **12**, 715-729.

53. M. Gabut, P. Samavarchi-Tehrani, X. Wang, V. Slobodeniuc, D. O'Hanlon, H. K. Sung, M. Alvarez, S. Talukder, Q. Pan, E. O. Mazzoni, S. Nedelec, H. Wichterle, K. Woltjen, T. R. Hughes, P. W. Zandstra, A. Nagy, J. L. Wrana and B. J. Blencowe, *Cell*, 2011, **147**, 132-146.
54. J. D. Ellis, M. Barrios-Rodiles, R. Colak, M. Irimia, T. Kim, J. A. Calarco, X. Wang, Q. Pan, D. O'Hanlon, P. M. Kim, J. L. Wrana and B. J. Blencowe, *Mol Cell*, 2012, **46**, 884-892.
55. M. Buljan, G. Chalancon, A. K. Dunker, A. Bateman, S. Balaji, M. Fuxreiter and M. M. Babu, *Curr Opin Struct Biol*, 2013, **23**, 443-450.
56. A. K. Dunker, S. E. Bondos, F. Huang and C. J. Oldfield, *Semin Cell Dev Biol*, 2015, **37**, 44-55.
57. H. Jeong, S. P. Mason, A. L. Barabasi and Z. N. Oltvai, *Nature*, 2001, **411**, 41-42.
58. A. L. Barabasi and Z. N. Oltvai, *Nat Rev Genet*, 2004, **5**, 101-113.
59. J. Hasty and J. J. Collins, *Nature*, 2001, **411**, 30-31.
60. D. Ekman, S. Light, A. K. Bjorklund and A. Elofsson, *Genome biology*, 2006, **7**, R45.
61. P. M. Kim, A. Sboner, Y. Xia and M. Gerstein, *Mol Syst Biol*, 2008, **4**, 179.
62. M. Higurashi, T. Ishida and K. Kinoshita, *Protein Sci*, 2008, **17**, 72-78.
63. A. Patil, K. Kinoshita and H. Nakamura, *Protein Sci*, 2010, **19**, 1461-1468.
64. E. Barbar and A. Nyarko, *Semin Cell Dev Biol*, 2015, **37**, 20-25.
65. E. A. Cino, R. C. Killoran, M. Karttunen and W. Y. Choy, *Sci Rep*, 2013, **3**, 2305.
66. J. C. Obenauer, L. C. Cantley and M. B. Yaffe, *Nucleic Acids Res*, 2003, **31**, 3635-3641.
67. M. Fuxreiter, P. Tompa and I. Simon, *Bioinformatics*, 2007, **23**, 950-956.
68. H. Dinkel, K. Van Roey, S. Michael, N. E. Davey, R. J. Weatheritt, D. Born, T. Speck, D. Kruger, G. Grebnev, M. Kuban, M. Strumillo, B. Uyar, A. Budd, B. Altenberg, M. Seiler, L. B. Chemes, J. Glavina, I. E. Sanchez, F. Diella and T. J. Gibson, *Nucleic acids research*, 2014, **42**, D259-266.
69. K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson and N. E. Davey, *Chem Rev*, 2014, **114**, 6733-6778.
70. E. Garner, P. Romero, A. K. Dunker, C. Brown and Z. Obradovic, *Genome Inform Ser Workshop Genome Inform*, 1999, **10**, 41-50.
71. A. Mohan, C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker and V. N. Uversky, *J Mol Biol*, 2006, **362**, 1043-1059.
72. V. Vacic, C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky and A. K. Dunker, *J Proteome Res*, 2007, **6**, 2351-2366.
73. C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 12454-12470.
74. Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2007, **46**, 13468-13477.
75. P. Tompa, M. Fuxreiter, C. J. Oldfield, I. Simon, A. K. Dunker and V. N. Uversky, *BioEssays : news and reviews in molecular, cellular and developmental biology*, 2009, **31**, 328-335.
76. B. Meszaros, I. Simon and Z. Dosztanyi, *PLoS computational biology*, 2009, **5**, e1000376.
77. Z. Dosztanyi, B. Meszaros and I. Simon, *Bioinformatics*, 2009, **25**, 2745-2746.
78. D. T. Jones and D. Cozzetto, *Bioinformatics*, 2014, DOI: 10.1093/bioinformatics/btu744.
79. N. Malhis and J. Gsponer, *Bioinformatics*, 2015, **31**, 1738-1744.

80. F. M. Disfani, W. L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker, V. N. Uversky and L. Kurgan, *Bioinformatics*, 2012, **28**, i75-83.
81. B. Meszaros, Z. Dosztanyi and I. Simon, *PLoS One*, 2012, **7**, e46829.
82. T. Le Gall, P. R. Romero, M. S. Cortese, V. N. Uversky and A. K. Dunker, *J Biomol Struct Dyn*, 2007, **24**, 325-342.
83. Y. Zhang, B. Stec and A. Godzik, *Structure*, 2007, **15**, 1141-1147.
84. T. Zhang, E. Faraggi, Z. Li and Y. Zhou, *Cell Biochem Biophys*, 2013, **67**, 1193-1205.
85. J. C. Obenauer and M. B. Yaffe, *Methods Mol Biol*, 2004, **261**, 445-468.
86. T. Ehrenberger, L. C. Cantley and M. B. Yaffe, *Methods Mol Biol*, 2015, **1278**, 57-75.
87. A. Valencia and F. Pazos, in *Protein-protein interactions and networks*, eds. A. Panchenko and T. M. Przytycka, Springer-Verlag, London, 2008, pp. 67-81.
88. A. J. Callaghan, J. P. Aurikko, L. L. Ilag, J. Gunter Grossmann, V. Chandran, K. Kuhnel, L. Poljak, A. J. Carpousis, C. V. Robinson, M. F. Symmons and B. F. Luisi, *J Mol Biol*, 2004, **340**, 965-979.
89. J. M. Bourhis, K. Johansson, V. Receveur-Brechot, C. J. Oldfield, K. A. Dunker, B. Canard and S. Longhi, *Virus Res*, 2004, **99**, 157-167.
90. P. T. Dolan, A. P. Roth, B. Xue, R. Sun, A. K. Dunker, V. N. Uversky and D. J. LaCount, *Protein Sci*, 2015, **24**, 221-235.
91. P. T. Dolan, C. Zhang, S. Khadka, V. Arumugaswami, A. D. Vangeloff, N. S. Heaton, S. Sahasrabudhe, G. Randall, R. Sun and D. J. LaCount, *Mol Biosyst*, 2013, **9**, 3199-3209.
92. I. Kotta-Loizou, G. N. Tsaousis and S. J. Hamodrakas, *Biochimica et biophysica acta*, 2013, **1834**, 798-807.
93. C. UniProt, *Nucleic acids research*, 2014, **42**, D191-198.
94. X. Fan, B. Xue, P. T. Dolan, D. J. LaCount, L. Kurgan and V. N. Uversky, *Mol Biosyst*, 2014, **10**, 1345-1363.
95. Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon, *J Mol Biol*, 2005, **347**, 827-839.
96. I. Walsh, A. J. Martin, T. Di Domenico and S. C. Tosatto, *Bioinformatics*, 2012, **28**, 503-509.
97. I. Walsh, M. Giollo, T. Di Domenico, C. Ferrari, O. Zimmermann and S. C. Tosatto, *Bioinformatics*, 2014, DOI: 10.1093/bioinformatics/btu625.
98. M. Howell, R. Green, A. Killeen, L. Wedderburn, V. Picascio, A. Rabionet, Z. L. Peng, M. Larina, B. Xue, L. Kurgan and V. N. Uversky, *J Biol Syst*, 2012, **20**, 471-511.
99. A. Lobley, M. B. Swindells, C. A. Orengo and D. T. Jones, *PLoS computational biology*, 2007, **3**, e162.
100. B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano and K. Fidelis, *Proteins*, 2014, **82 Suppl 2**, 127-137.
101. D. T. Jones, *J Mol Biol*, 1999, **292**, 195-202.
102. T. Allers, *Bioengineered bugs*, 2010, **1**, 288-290.
103. J. Yan, M. J. Mizianty, P. L. Filipow, V. N. Uversky and L. Kurgan, *Biochim Biophys Acta*, 2013, **1834**, 1671-1680.
104. M. Vihinen, E. Torkkila and P. Riikonen, *Proteins*, 1994, **19**, 141-149.
105. A. K. Dunker, C. J. Brown and Z. Obradovic, *Advances in protein chemistry*, 2002, **62**, 25-49.
106. R. M. Williams, Z. Obradovi, V. Mathura, W. Braun, E. C. Garner, J. Young, S. Takayama, C. J. Brown and A. K. Dunker, *Pac Symp Biocomput*, 2001, 89-100.

107. F.-X. Theillet, L. Kalmar, P. Tompa, K.-H. Han, P. Selenko, A. K. Dunker, G. W. Daughdrill and U. V.N., *Intrinsically Disordered Proteins*, 2013, **1**, e24360.
108. S. C. Li, N. K. Goto, K. A. Williams and C. M. Deber, *Proc Natl Acad Sci U S A*, 1996, **93**, 6676-6681.
109. P. Y. Chou and G. D. Fasman, *Annual review of biochemistry*, 1978, **47**, 251-276.
110. P. Chakrabarti and U. Samanta, *J Mol Biol*, 1995, **251**, 9-14.
111. C. M. Santiveri and M. A. Jimenez, *Biopolymers*, 2010, **94**, 779-790.
112. A. G. Cochran, *Chemistry & biology*, 2000, **7**, R85-94.
113. S. K. Sharma, T. M. Ramsey and K. W. Bair, *Current medicinal chemistry. Anti-cancer agents*, 2002, **2**, 311-330.
114. D. C. Fry, *Current pharmaceutical design*, 2012, **18**, 4679-4684.
115. F. Falchi, F. Caporuscio and M. Recanatini, *Future medicinal chemistry*, 2014, **6**, 343-357.
116. L. Jin, W. Wang and G. Fang, *Annual review of pharmacology and toxicology*, 2014, **54**, 435-456.
117. V. N. Uversky, *Chemical Society reviews*, 2011, **40**, 1623-1634.
118. Y. Cheng, T. LeGall, C. J. Oldfield, J. P. Mueller, Y. Y. Van, P. Romero, M. S. Cortese, V. N. Uversky and A. K. Dunker, *Trends Biotechnol*, 2006, **24**, 435-442.
119. S. J. Metallo, *Curr Opin Chem Biol*, 2010, **14**, 481-488.
120. A. K. Dunker and V. N. Uversky, *Curr Opin Pharmacol*, 2010, **10**, 782-788.
121. V. N. Uversky, *Expert Opin Drug Discov*, 2012, **7**, 475-488.
122. G. Hu, Z. Wu, K. Wang, V. N. Uversky and L. Kurgan, *Current drug targets*, 2015.
123. O. Lichtarge, H. R. Bourne and F. E. Cohen, *J Mol Biol*, 1996, **257**, 342-358.
124. P. Fariselli, F. Pazos, A. Valencia and R. Casadio, *European journal of biochemistry / FEBS*, 2002, **269**, 1356-1361.
125. J. F. Xia, S. L. Wang and Y. K. Lei, *Protein Pept Lett*, 2010, **17**, 1069-1078.
126. A. Valencia and F. Pazos, *Curr Opin Struct Biol*, 2002, **12**, 368-373.
127. R. J. Edwards, N. E. Davey and D. C. Shields, *PloS one*, 2007, **2**, e967.
128. B. Uyar, R. J. Weatheritt, H. Dinkel, N. E. Davey and T. J. Gibson, *Molecular bioSystems*, 2014, **10**, 2626-2642.

Tables

Table 1. Abundance of MoRF and intrinsically disordered regions in Archaea, Bacteria and Eukaryota. + (=) indicates that content of MoRF or disordered residues for a domain of life in a given row is (is not) significantly higher compared to the domain in a given column (*t*-test (all data were normal); degrees of freedom = 9; *p*-value <0.01). The last column lists Pearson Correlation Coefficient (PCC) between the per species content of MoRF and intrinsically disordered residues in a given domain of life.

Domain of life	# species	# proteins	MoRFs residues			Disordered residues			PCC
			content [%]	significance		content [%]	significance		
				Archaea	Bacteria		Archaea	Bacteria	
Archaea	72	174,381	1.0			6.8			0.98
Bacteria	567	2,025,100	0.9	=		5.8	+		0.89
Eukaryota	229	3,645,837	1.0	=	=	19.1	+	+	0.43

Figure Legends

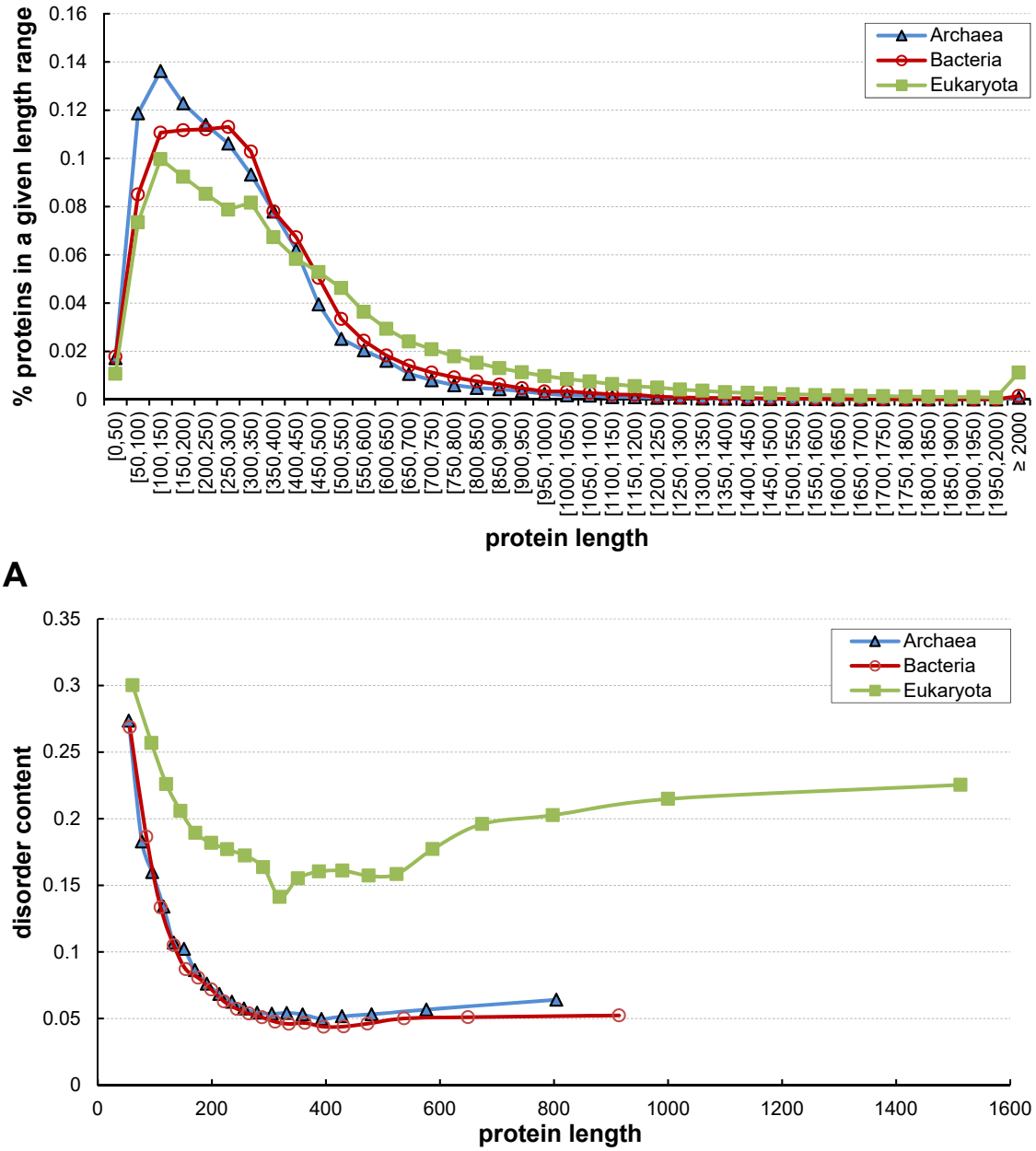
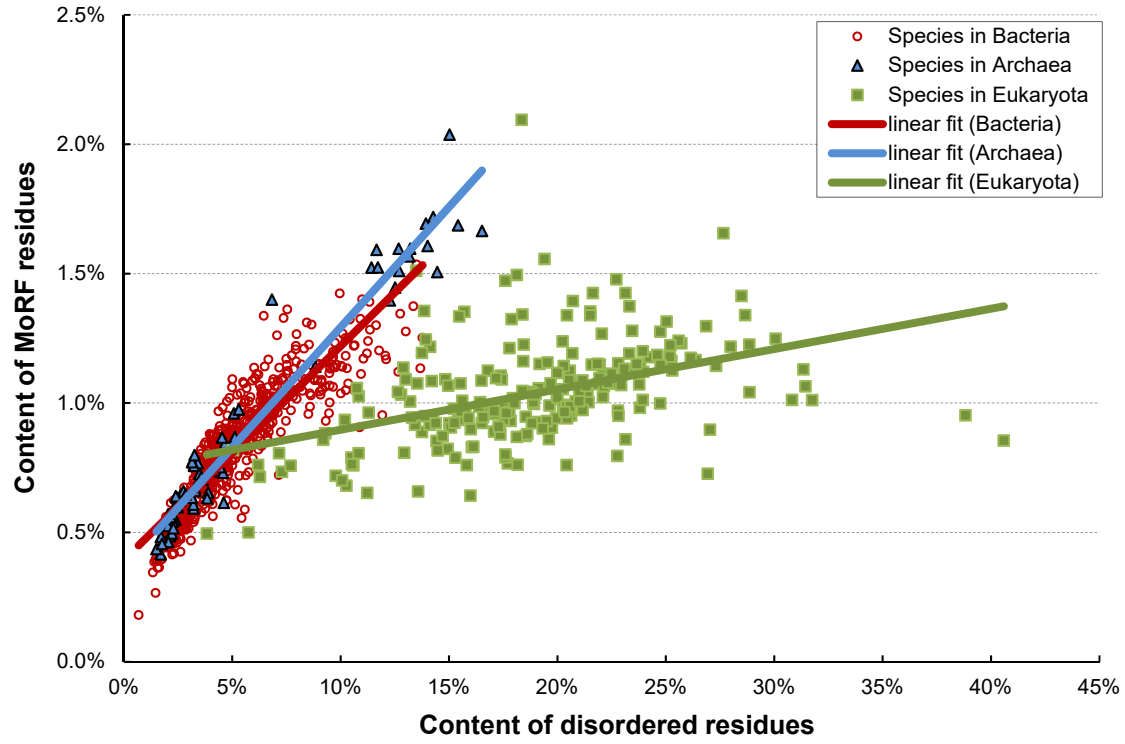
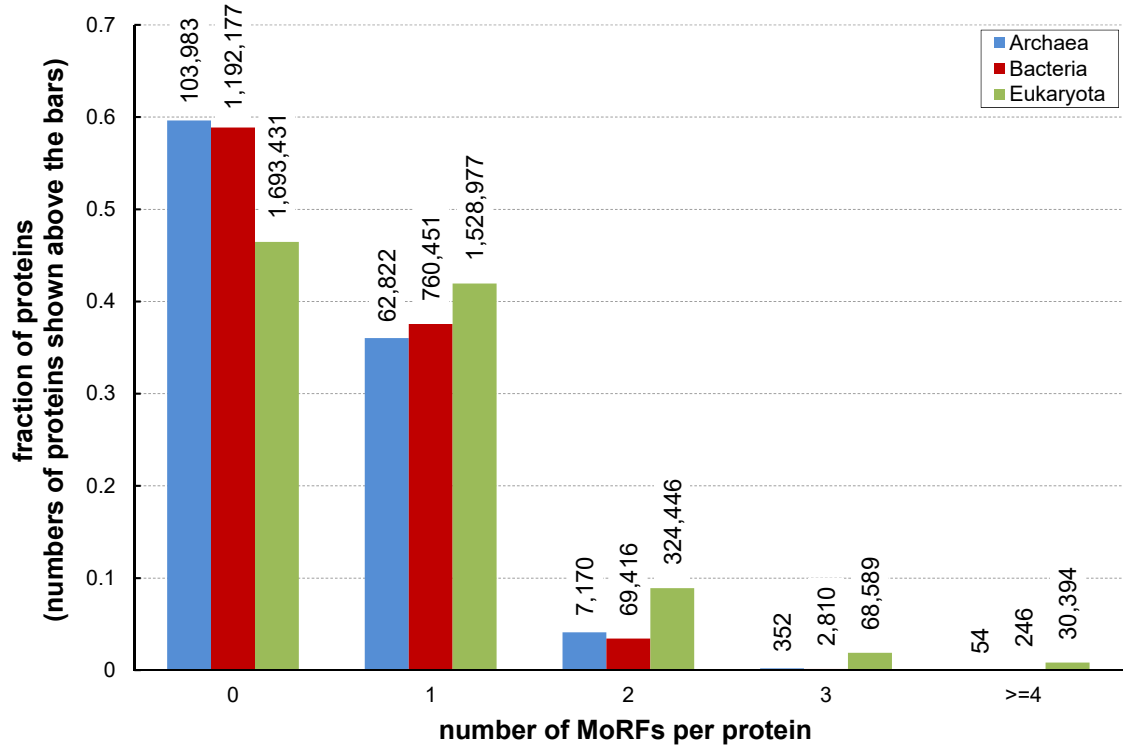


Figure 1. **A**. Distribution of protein length for 5,845,314 proteins of the 868 species from Archaea, Bacteria and Eukaryota. **B**. Disorder content *versus* protein length plots for proteins in the three domains of life. The proteins were sorted by length of their polypeptide chain and binned by every 5% of the sorted proteins. Each plot represents disorder content over all proteins in a given bin and given domain of life *versus* median length in the bin.



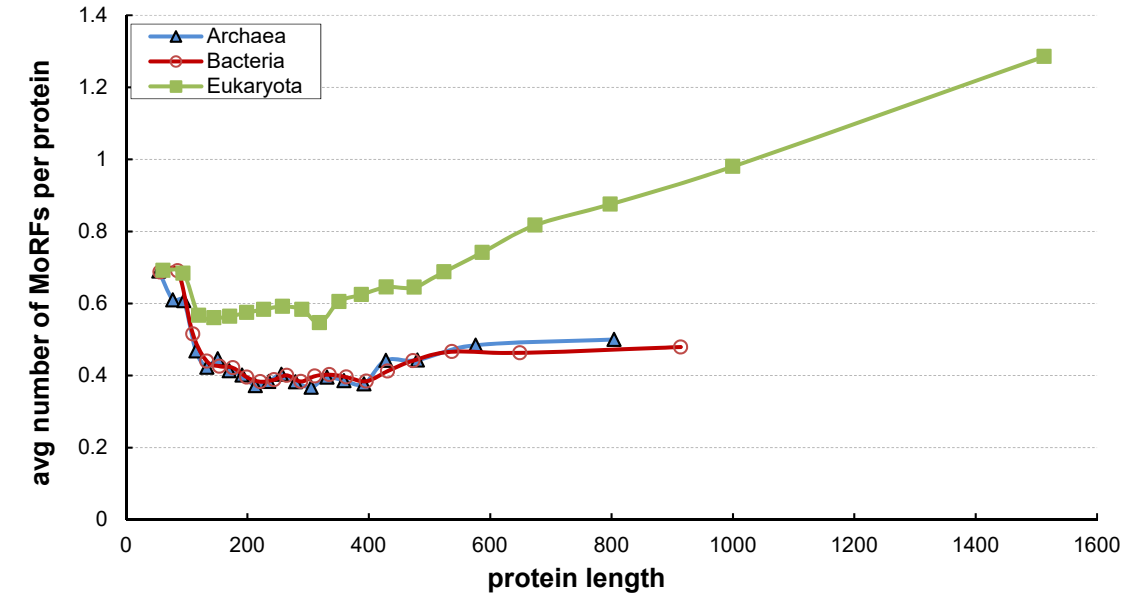
A



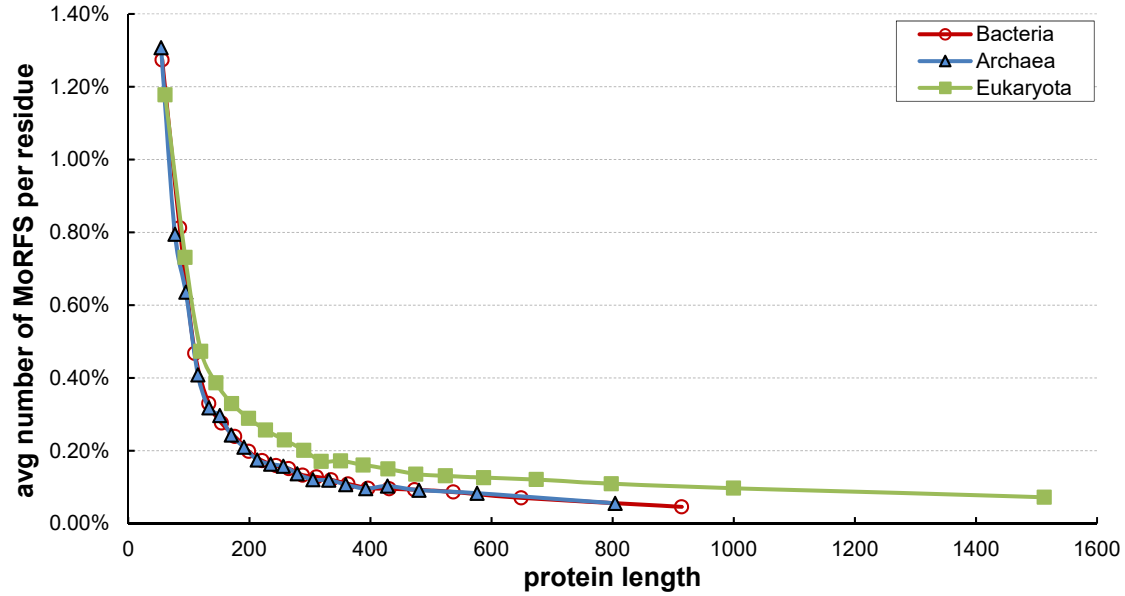
B

Figure 2. **A**. Relation between content of MoRFs and intrinsically disordered residues for species in the three domains of life. Markers represent individual species and lines show linear fit

between content of disordered residues and content of MoRF residues in a given domain of life. **B**. Histogram of the fraction of proteins (corresponding numbers of proteins are shown above the bars) *versus* number of MoRFs per protein in the three domains of life.



A



B

Figure 3. Frequencies of MoRFs in proteins from the three domains of life. Plot **A**. Number of MoRFs per protein *versus* protein length. The proteins were sorted by length of their polypeptide chain and binned by every 5% of the sorted proteins. We plot the average number of MoRFs per protein over all proteins in a given bin (total number of MoRFs divided by the number of proteins in a given bin) *versus* median length in the bin. Plot **B**. Number of MoRFs per residue *versus* protein length. The proteins were sorted by length of their polypeptide chain and binned by every 5% of the sorted proteins. We plot the number of MoRFs per residue over all proteins in a given bin (total number of MoRFs divided by the number of residues in a given bin) and median value of length in the bin.

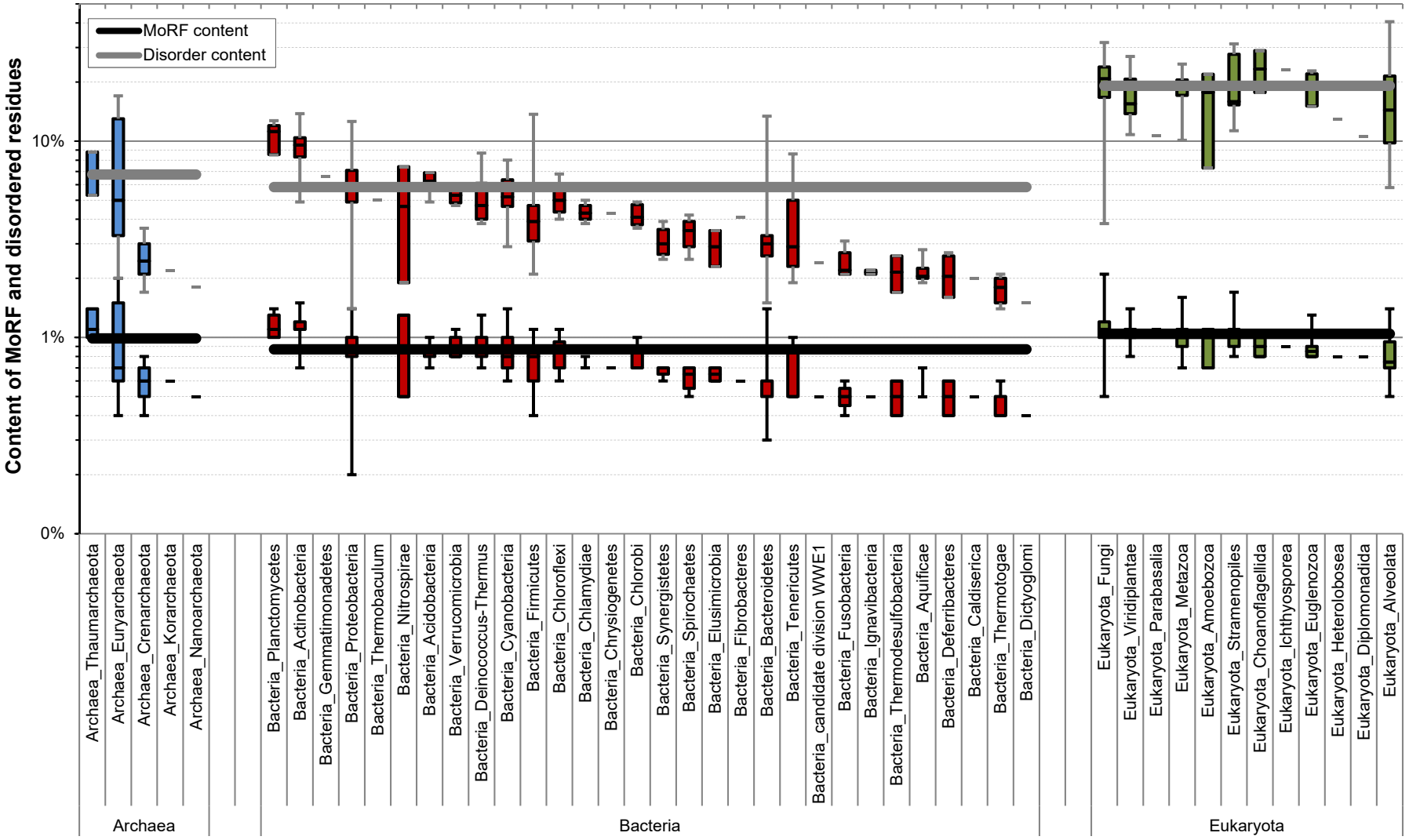


Figure 4. Content of MoRF and intrinsically disordered residues (y-axis in logarithmic scale) across species grouped into kingdoms/phyla (*x*-axis) in the three domains of life. The box plots for each kingdom show the minimal, 25th centile, median, 75th centile and maximal MoRF (lower box-plots shown using black lines) and disorder (upper boxplots shown using gray lines) content

over species in a given kingdom. Solid horizontal lines represent the content of MoRF (in black) and disordered (in gray) residues in the entire domain of life. The kingdoms are sorted in the descending order by their median content of MoRFs.

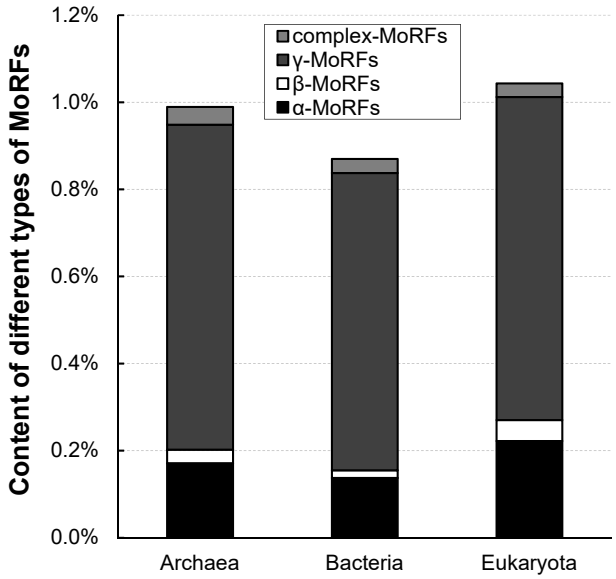


Figure 5. Content of the four types of MoRFs in the three domains of life.

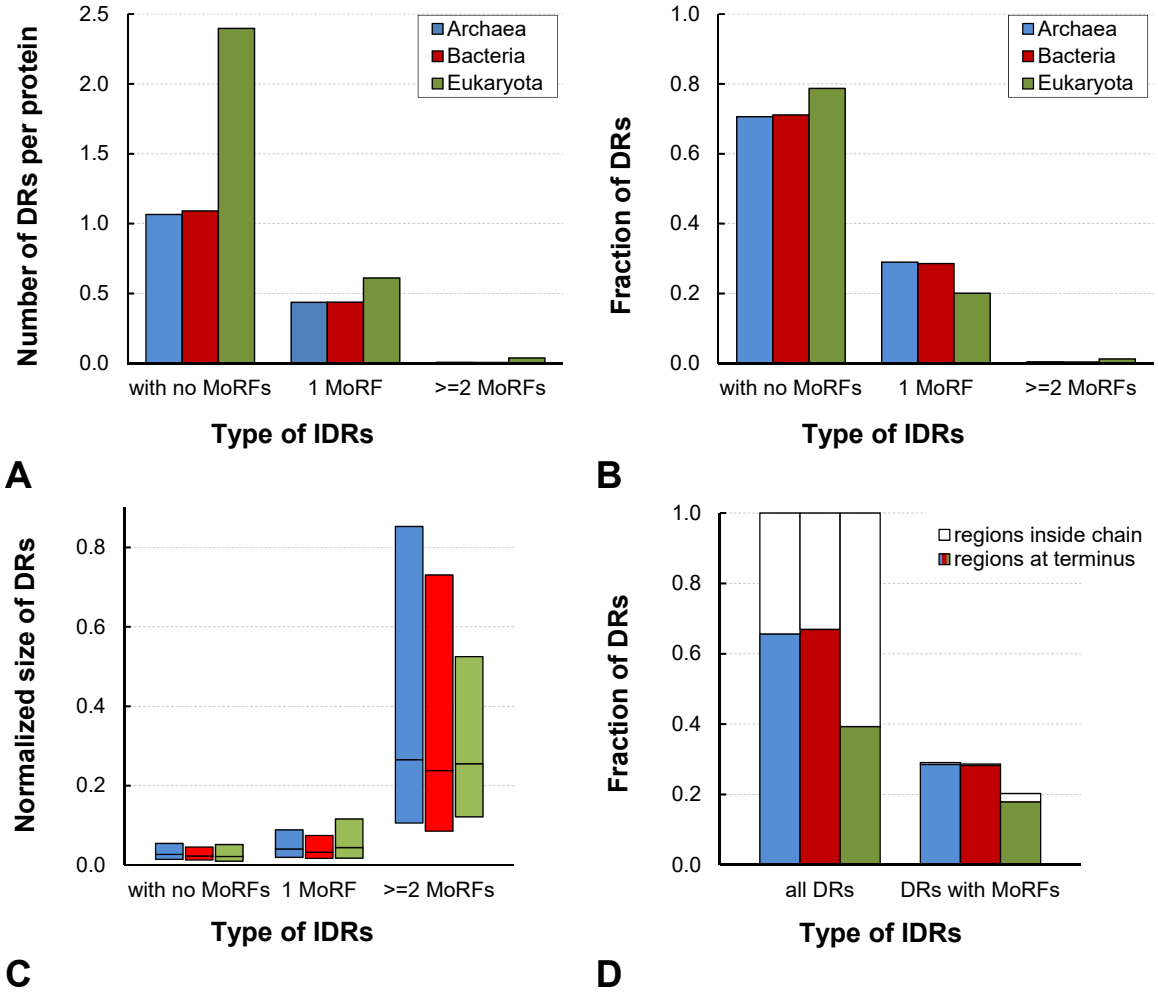


Figure 6. Analysis of intrinsically disordered regions (IDRs) in the three domains of life. We categorize IDRs into those that include no MoRFs, one and multiple MoRFs. Panels **A** and **B** shows the number of IDRs per proteins and fraction of the IDRs, respectively. Panel **C** gives boxplots (25th centile, median, and 75 centile) of the normalized size of the IDRs. Panel **D** summarizes fraction of IDRs that are localized at the sequence terminus vs. inside the sequence for all IDRs and IDRs that include at least one MoRF region.

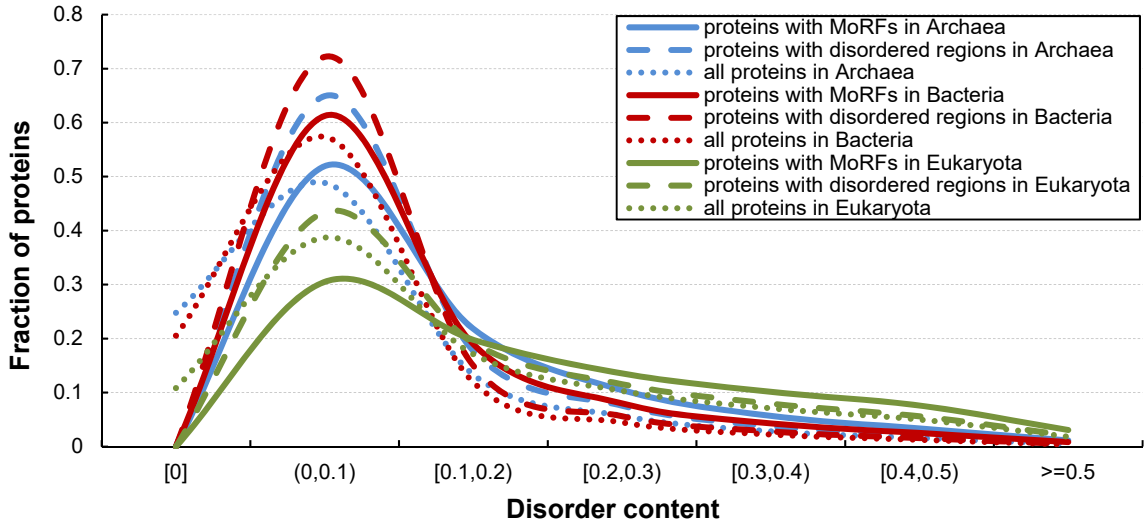


Figure 7. Distribution of disorder content values for proteins that include MoRFs, proteins that include intrinsically disordered regions, and all proteins in the three domains of life. The overall range of disorder content was divided into 10 intervals shown on the *x*-axis. The left-most point where the disorder content is 0 corresponds to fully structured proteins.

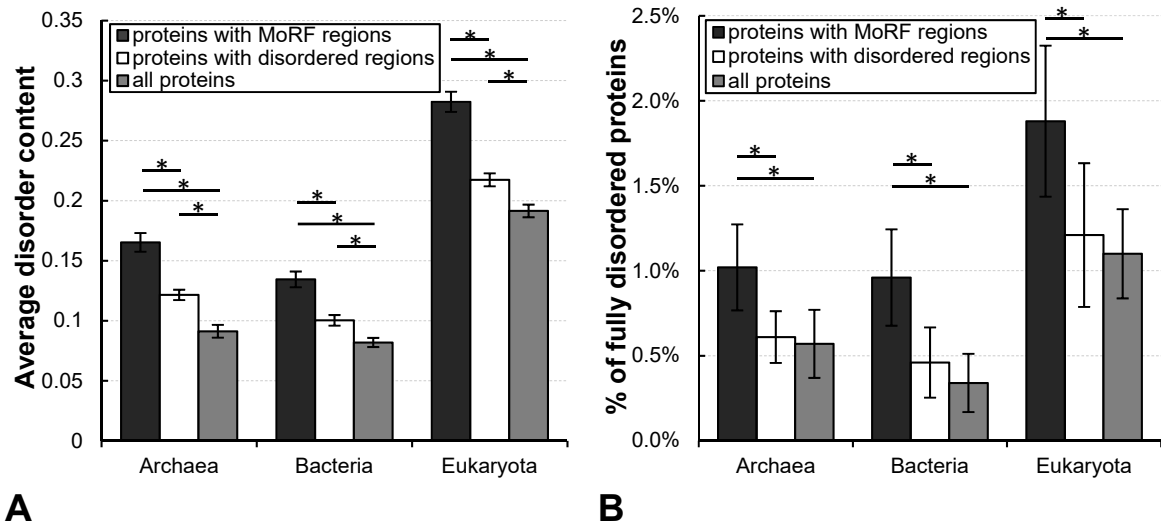


Figure 8. Average disorder content (panel **A**) and fraction of fully disordered proteins (panel **B**) computed for the proteins with the MoRFs (dark gray bars), with the intrinsically disordered regions (white bars) and for all proteins (light gray bars) in the three domains of life. Since all data were normal, the bars and the error bars show the average and the corresponding standard deviations based on 10 measurements that utilize 1000 randomly chosen proteins. * indicates that the difference is significant (*t*-test; degrees of freedom = 9; *p*-value < 0.01; see Materials and methods for details).

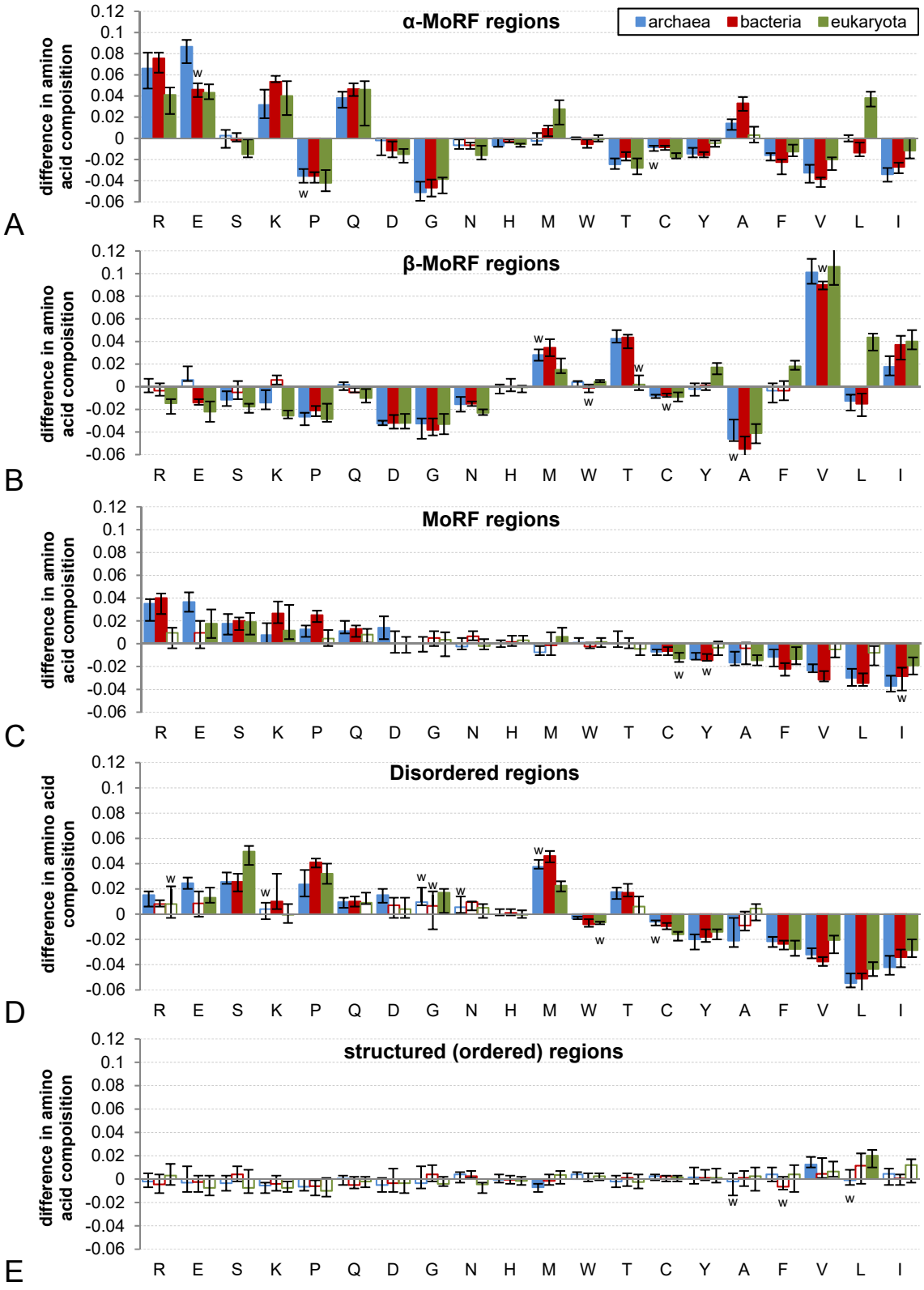


Figure 9. Differences in the amino acid composition between residues in the α-MoRF (panel **A**), β-MoRF (panel **B**), all MoRFs (panel **C**), intrinsically disordered regions (panel **D**), structured

regions (panel **E**) and generic (randomly selected) residues in the three domains of life. The bars and the error bars show the median and the corresponding 25^{th} and 75^{th} centiles based on 10 measurements with 1000 randomly chosen the α-MoRF/β-MoRF/MoRF/disordered/structured residues. Solid (hollow) bars indicate that the differences in the composition is (is not) statistically significant (*t*-test or Wilcoxon test; degrees of freedom = 9; *p*-value < 0.01; see Materials and methods for details). Data for which Wilcoxon test was used are annotated with w next to the error bar; lack of annotation indicates that data were normal.

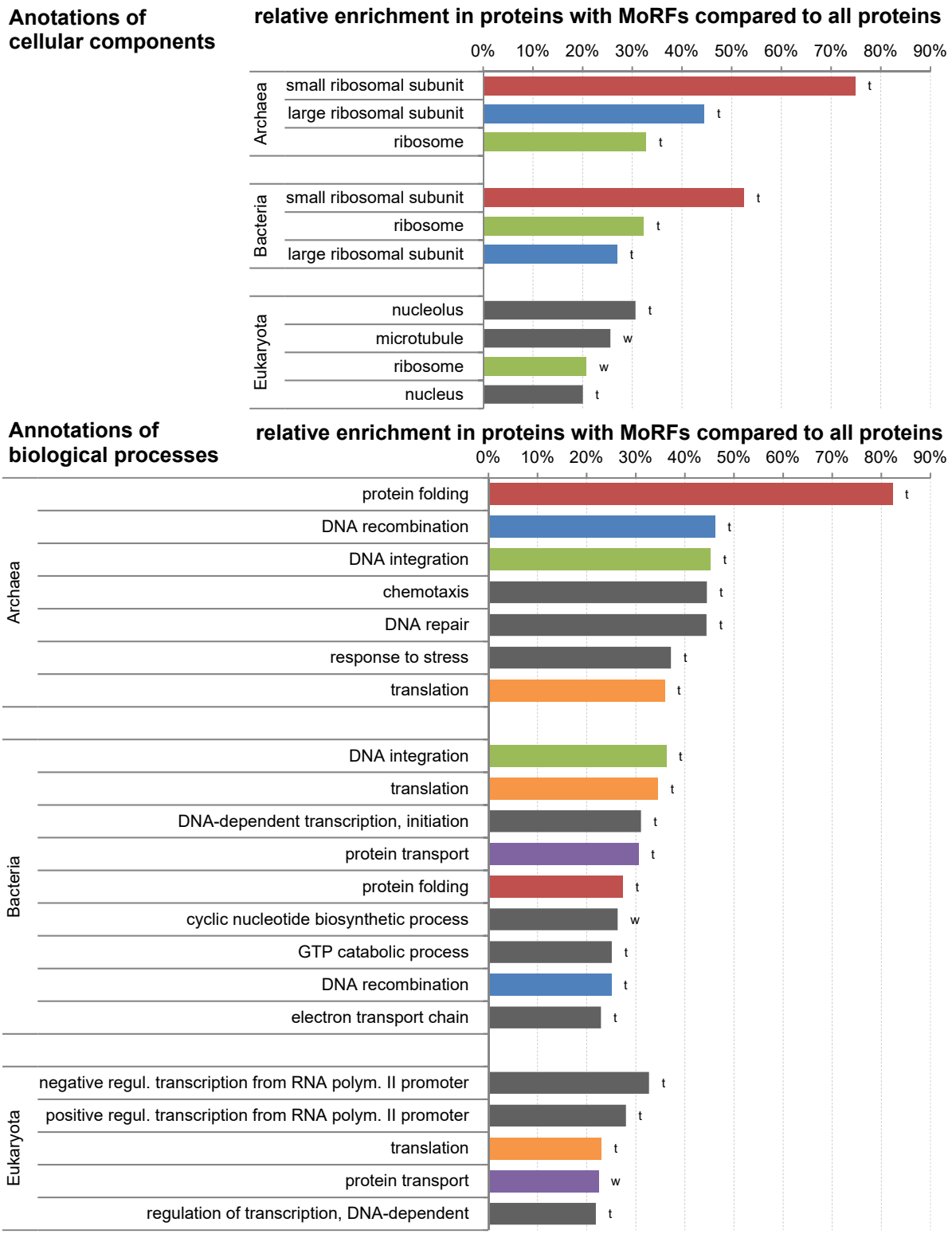


Figure 10. Cellular component and biological process GO terms that are significantly enriched in MoRF proteins (*t*-test or Wilcoxon test; degrees of freedom = 9; *p*-value <0.01). The *y*-axis gives the enriched GO terms. The *x*-axis shows the relative difference between rates of

occurrence of a given term in MoRF-containing proteins and a random set of proteins from the same domain of life. Colored bars are used to denote the same GO terms that appear in different domains of life. Data for which *t*-test and Wilcoxon test was used are annotated with t and w next to the bars, respectively.