Taxonomic landscape of the dark proteomes: whole-proteome scale interplay between structural darkness, intrinsic disorder, and crystallization propensity

Gang Hu[1#], Kui Wang[1#], Jiangning Song[2,3], Vladimir N. Uversky[4,5] and Lukasz Kurgan[6*]

[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin, PR China
[2]Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, Australia
[3]Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, Australia
[4]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, USA
[5]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Russia
[6]Department of Computer Science, Virginia Commonwealth University, Richmond, USA

Emails GH: huggs@nankai.edu.cn
KW: wangkui@nankai.edu.cn
JS: jiangning.song@monash.edu
VNU: vuversky@health.usf.edu
LK: lkurgan@vcu.edu

* Corresponding Author
Phone: +1-804-827-3986; Fax: +1-804-828-2771; email: lkurgan@vcu.edu
401 West Main Street, Room E4225, Richmond, Virginia 23284-3019

[#] These authors contributed equally

Short Title: Taxonomic landscape of structural darkness and intrinsic disorder

Abstract

Growth rate of the protein sequence universe dramatically exceeds the speed of expansion for the protein structure universe, generating an immense dark proteome that includes proteins with unknown structure. We perform a whole-proteome scale analysis of 5.4 million proteins from 987 proteomes in the three domains of life and viruses to systematically dissect an interplay between structural coverage, degree of putative intrinsic disorder, and predicted propensity for structure determination. We found that Archaean and Bacterial proteomes have relatively high structural coverage and low amounts of disorder, whereas Eukaryotic and Viral proteomes are characterized by a broad spread of structural coverage and higher disorder levels. Our analysis reveals that dark proteomes (i.e., proteomes containing high fractions of proteins with unknown structure) have significantly elevated amounts of intrinsic disorder and are predicted to be difficult to solve structurally. Although the majority of dark proteomes are of viral origin, many dark viral proteomes have at least modest crystallization propensity and only a handful of them are enriched in the intrinsic disorder. We mapped the disorder, structural coverage and propensity for structural determination onto a novel proteome-level sequence similarity network to analyze the interplay of these characteristics in the taxonomic landscape.

Statement of Significance

An interplay between structural darkness, intrinsic disorder, and crystallization propensity was investigated for 5,381,183 proteins from 987 complete proteomes covering all domains of life (Bacteria, Archaea, and Eukaryota) and viruses. We show that the bacterial proteomes are characterized by high structural coverage (~60%), whereas the majority of the dark proteomes (i.e., proteomes containing large fraction of proteins with unknown structure) are of viral and eukaryotic origin. Although eukaryotic dark proteomes possess low crystallization propensity combined with high disorder content, the majority of dark viral proteomes have moderate crystallization propensities and only some of them possess high intrinsic disorder predisposition. This indicates that structural information can be retrieved for many viral proteomes in the future. Therefore, our data provide strong support to the important concept that although intrinsic disorder represents an essential contributing factor to structural darkness, not all structurally dark proteins and proteomes are necessarily enriched in intrinsic disorder. Finally, mapping intrinsic disorder predisposition, currently known structural coverage, and propensity for structural determination onto a proteome-level sequence similarity network provided taxonomic landscape of the dark proteomes and showed the existence of an intricate relationship between the structural darkness and intrinsic disorder across the three domains of life and viruses.

Keywords:
Dark proteomes; Intrinsic disorder; X-ray crystallography; Eukaryota; Bacteria; Archaea; Viruses

Introduction

Advances and investments in the sequencing technologies [1] and structural genomics efforts [2, 3] in the last several decades have resulted in large and ever growing protein universe [4, 5, 6]. The structure determination efforts do not scale up to the high-throughput of the sequencing, but it is believed that this gap can be partially bridged with the help of computational tools that provide accurate predictions of protein structures and functions [7, 8]. Knowledge of protein structure is fundamental to the discovery and characterization of protein functions (at least for proteins that conform classical structure-function paradigm) [9], rational drug design [10], elucidation of the side effects of drugs [11], drug repurposing and repositioning [12], and to the understanding of protein evolution [13], to name a few application areas. Given the importance of protein structure, numerous studies have focused on estimating current and future structural coverage of protein sequences [6, 14]. Our aim is to shed light on the other, dark side of the protein sequence space.

Two definitions of the "dark proteome" were coined in recent years. One defines dark proteome as a collection of intrinsically disordered proteins and intrinsically disordered protein regions [15]. These proteins and regions are usually not amenable to structure determination by conventional methods, such as X-ray crystallography and electron microscopy, because of their highly flexible nature and they are also typically inaccessible to homology modelling. The other defines the dark proteome as the structurally unknown part of the protein sequence space; i.e., collection of sequences and sequence fragments for which structure is currently unknown [16, 17]. Both definitions share the focus on the structurally undetermined protein sequences. The latter is more comprehensive and oblivious to the root causes while the former narrows the dark proteome down to the intrinsically unstructured regions. Furthermore, the aforementioned studies of the dark proteome emphasize its importance and the fact that it is largely understudied.

Recently, the generic structural darkness has been investigated using a set of about 0.5 million protein sequences [17]. This study has analyzed protein sequences that lack structural information and concluded that about half of amino acids in Eukaryota are dark compared to a much smaller fraction of about 14% in Archaea and Bacteria [17]. A particularly relevant aspect of this work is the observation that while intrinsically disordered residues are mostly structurally dark, most of the structurally dark residues are not disordered. Our aim is to analyze the interplay of structural darkness, intrinsic disorder and propensity for structural characterization on a much larger scale and at a lower level of granularity. We analyze these relations at the whole-organism level by considering 5.4 million proteins from close to 1000 complete proteomes that cover the three domains of life and viruses.

There are several related large proteome-scale analyses in the literature. They include studies of the abundance and functional roles of intrinsic disorder [18-23], structural coverage of protein sequences [24, 25], and propensity for structural determination via the most frequently used experimental tool of structural biology, X-ray crystallography [4]. However, none of these studies provides a holistic view that brings these multiple interconnected aspects together. The defining aspect of our analysis is not the scale, but rather the comprehensive scope that, for the first time, analyzes the three interlinked structural characteristics together. Furthermore, we analyze these three key characteristics (structural darkness, intrinsic disorder predisposition, and propensity for structural characterization) in the context of the taxonomic relationships among the diverse set of the considered organisms. This is accomplished by

developing and using a novel proteome similarity network. The network associates the proteomes based on the proteome-level sequence similarity used as a proxy for the taxonomic relations. Our main focus is to annotate and systematically analyze and characterize dark proteomes, defined as the proteomes with the highest levels of dark proteins; i.e., proteins lacking structure.

Materials and methods

Proteome dataset

Proteomes, defined as collections of proteins encoded by the genome of a specific organism, were collected from the UniProt resource [26]. We collected complete proteomes which are based on the fully sequenced genomes. We used one proteome per species, by selecting the proteome that had the largest number of sequences among all proteomes with the same taxonomic identifier. We removed the proteomes with less than 10 proteins to ensure that statistical analysis is robust. In total we considered 987 complete proteomes that included 5,381,183 protein chains. They cover 64 Archaean species (152,622 proteins), 553 Bacteria (1,971,973 proteins), 201 Eukaryotic organisms (3,245,432 proteins), and 169 viruses (12,439 sequences). A complete list of the considered species is provided in the Supplementary Materials.

Portrayal of structural characteristics

We used computational tools to estimate the current structural coverage (the fraction of proteins for which structure is available), the content of intrinsic disorder (the fraction of residues that are predicted to be intrinsically disordered), and the projected propensity for structural coverage (the likelihood that a given proteome can be solved structurally) at the whole-proteome scale. The latter estimates the likelihood to solve protein structures using the presently dominant approach that relies on the X-ray crystallography. Over 90% of proteins structures in the Protein Data Bank (PDB) [27] were solved with this technology; i.e., 118,425 out of the 131,108 protein structures as of June 2018.

We estimated the current structural coverage of proteomes based on an approach proposed in [24]. For each of the 5.38 million protein sequences we ran three rounds of PSI-BLAST [28] searches against the sequences of protein structures collected from the PDB. A given sequence is considered to be solved structurally if it registers a hit in PDB with the E-value <0.001 that has >50 amino acids in length. In other words, this method categorizes a given protein as solved structurally if it contains at least one long segment of residues (representing at least one domain) that is sufficiently similar to a sequence of an already solved protein structure. Consequently, this approach is particularly effective for annotating dark proteins that lack structure along their entire sequence. We defined the structural coverage of a proteome as the fraction of the structurally solved sequences among all sequences in this proteome. The remaining part of the proteome is considered as dark. This approach is sufficiently computationally efficient to process our large dataset with over 5 million proteins. We acknowledge that more precise alternatives to accurately estimate the number of proteins for which structure is available can be found. They include tools such as I-TASSER [29], HHpred [30], and MODELLER [31] that can better detect remote homology. However, it would be virtually impossible to scale these tools up to the size of our dataset. To the best of our knowledge, the largest such attempt is the MODBASE, which includes results for 76 species [8] and relies on a combination of PSI-BLAST and MODELLER. Furthermore, we did not require high levels of accuracy given that we aggregated and analyzed (in relative terms) the abundance of dark

proteins at the whole proteome scale. Furthermore, methods that rely on PSI-BLAST were shown to provide accurate results. For instance, a similar PSI-BLAST-based method failed to find templates (similar sequences that are structured) for only 3 out of 120 target proteins in CASP9 [32]. Nevertheless, we should keep in mind that the structural coverage computed by our approach is slightly underestimated, as it fails to annotate some of the remotely homologous structures.

Multiple studies suggest that intrinsic disorder can be accurately predicted from protein sequences [33, 34, 35]. Consensus approaches that combine results of multiple disorder predictors were shown to provide more accurate predictions when compared to single predictors [36-38]. Analysis in [37] reveals that consensus-based predictors more accurately predict the residue-level annotation of disorder, with 2% improvement in the AUC (area under ROC curve) and 4% increase in MCC (Matthews correlation coefficient). They also provide a more accurate estimate the overall amount of disorder in a given protein, reducing the error by about 4%. Recent research also shows that consensus-based predictions provide better detection of long disordered regions (20 or more consecutive residues) [38]. These advantages come at the cost of a substantial increase in the computational time, given that multiple vs. single prediction have to be computed. We utilized a consensus of five predictions produced by two popular tools, IUPred [39] and ESpritz [40]. These tools were selected based on their competitive predictive quality [33, 35], short runtime (for each individual predictor) and complementary designs. To the latter point, the five predictions were generated based on two versions of IUPred tool that were designed to predict long (30 or more consecutive residues) and short disordered regions, and three versions of ESpritz that focus on three main types of annotations of disordered residues: using DisProt database [41], crystal structures, and structures defined by nuclear magnetic resonance. The consensus requires that at least 3 out of 5 predictions must indicate intrinsic disorder to predict a given residue as disordered. The same consensus was applied in several related studies [19, 42]. This is arguably a more accurate methodology that the one used in the related study of the dark proteome where a single tool for disorder predictions, IUPred, was utilized [17]. Our approach is also similar to the consensus-derived annotations of putative disorder in both MobiDB [43] and D^2P^2 [44] databases. The putative disorder is annotated at the amino acid level allowing us to quantify the amount of disorder per protein and per proteome. The disorder content of a proteome is defined as a fraction of residues predicted as disordered among all the residues in that proteome.

We emphasize again that protein structures determined by X-ray crystallography account for the significant majority (over 90%) of the currently solved protein structures [2, 45]. Therefore, we focus here on prediction of the predisposition of protein sequence for crystallization as a means for the evaluation of its propensity for structural determination. Recent studies reveal that these predictors produce quite accurate propensities for crystallization [46, 47, 48]. Although a modest correlation between these predicted propensities for crystallization and the resolution of the crystallizable proteins was noted for several tools [46], prediction of the resolution is relatively inaccurate and outside the scope of this study. The expected propensity for structural coverage was estimated using a recently published fDETECT method [49]. fDETECT was selected for our analysis since this is currently the fastest tool capable of making an accurate prediction of the crystallization propensity [48, 49], having sufficiently short runtime to process the 987 proteomes. Moreover, fDETECT does not use information about disorder to make predictions of the propensity for structural determination, which allows us to estimate these two factors independently. This tool was previously used to analyze the prospective structural coverage at the multi proteome-scale [4]. The crystallization propensity generated by fDETECT is a real number that quantifies the likelihood that a diffraction-quality crystal can be produced for a given protein sequence. The

projected proteome-level propensity for structural coverage is quantified with the median predicted crystallization propensity over all proteins in a given proteome. It is noteworthy that this number does not quantify the actual projected structural coverage, but rather a relative propensity for structural determination. In other words, we can use the proteome-level crystallization propensity to assess the relative difficulty of gaining structural information for proteomes.

Construction of the proteome similarity network

We studied the relations between the three structural characteristics (i.e., structural darkness, intrinsic disorder predisposition, and propensity for structural characterization) in a large set of 987 proteomes. To facilitate this analysis, we created a proteome similarity network that links the proteomes based on their mutual sequence similarity. This proteome-level sequence similarity for a given pair of proteomes is measured as an average similarity between the most similar pairs of sequences from the two proteomes. The corresponding proteome similarity network was subsequently constructed to cluster the proteomes based on their taxonomic classification, allowing us to visualize a taxonomic landscape of the structural characteristics. Details concerning computation of this network are provided in the Supplement.

Table 1. The overall values of the structural coverage, disorder content and crystallization propensity for X-ray crystallization across the domains and kingdoms/phyla of life. The domains of life are arranged according to their overall structural coverage.

Domain of life	Kingdom/phylum of life	Number of Species	Structural Coverage (%)	Disorder Content (%)	Crystallization Propensity
Bacteria	**All**	**553**	**59.7**	**5.8**	**0.317**
	Proteobacteria	255	61.2	5.8	0.306
	Firmicutes	76	59.8	4.6	0.360
	Actinobacteria	70	59.1	10.0	0.325
	Other	152	57.4	4.4	0.312
Archaea	**All**	**64**	**53.8**	**5.8**	**0.372**
	Euryarchaeata	43	54.6	6.9	0.366
	Crenarchaeota	17	51.7	3.0	0.392
	Other	4	53.8	5.0	0.356
Eukaryotes	**All**	**201**	**47.7**	**19.6**	**0.136**
	Fungi	84	45.6	21.5	0.140
	Metazoa	69	55.0	18.3	0.133
	Other	48	41.0	18.2	0.134
Viruses	**All**	**169**	**31.5**	**9.4**	**0.235**
	Archaeal viruses	10	10.1	6.8	0.425
	Bacterial viruses	28	18.2	9.4	0.388
	Eukaryotic viruses	131	35.9	9.7	0.188

Results and discussion

Structural characteristics at the top taxonomic classification levels

Table 1 provides a statistical summary of the results at the top taxonomic levels. Remarkably, bacterial

proteins are characterized by the highest value of the current structural coverage at about 60%. Based on our definition of the structural coverage, this means that at least a partial structure is available for 60% of the bacterial proteins, while the remaining 40% are dark. The highest structural coverage of Bacterial proteomes stems from the availability of the largest number of structures for these proteins, relative to their proteome sizes. More specifically, among about 219 thousand structures in the PDB that are assigned to the three domains of life and viruses, 120.5 thousand are in Eukaryota, 79.0 thousand in Bacteria, 11.8 thousand in viruses and 7.5 thousand in Archaea, respectively (source: PDB as of June 2018). Furthermore, Bacteria have considerably smaller proteomes than the proteomes of the most structurally populated Eukaryota. In our dataset, the average proteome size is 16,146 proteins in Eukaryota compared to 3,566 for Bacteria and 2,385 for Archaea. The current structural coverage for Archaea is 54%, for Eukaryota is 48% and for viruses is only 32%. The smaller amount of structural coverage in Eukaryota when compared to Archaea and Bacteria is in agreement with the prior, smaller-scale study that analyzed structural darkness at the protein level [17]. Our analysis reveals that viral proteomes are predominantly structurally dark. Interestingly, while structural coverage of the kingdoms/phyla generally follows the coverage at the domain level for Bacteria, Achaea, and Eukaryota, one exception is Metazoa that has a much higher coverage than an average eukaryotic proteome (55% vs. 48%). This observation has roots in the skewed distribution of protein structures among Eukaryota, where Metazoa is covered by 87 thousand structures (with 54 thousand structures in *Homo sapiens* alone) compared to 23 thousand in Fungi and just 10.5 thousand in all other Eukaryota combined. The current structural coverage in viruses spans a wide range between 10% for archaeal viruses, 18% for bacteriophages, and 36% for eukaryotic viruses. These differences stem from substantial differences in sequences, structures, and mechanisms underlying their interactions with the host organisms [50]. However, even the most structurally covered eukaryotic viruses substantially lag behind the structural coverage of proteomes in the three domains of life. This reveals that the viruses are universally biased to be structurally dark, irrespective of the taxonomic classification of their host organisms.

Furthermore, our analysis demonstrates that Eukaryota have substantially higher levels of intrinsic disorder, particularly when compared to Bacteria and Archaea (Table 1). This is in agreement with several recent studies that have concluded that intrinsic disorder is significantly enriched in eukaryotes when compared to prokaryotes [18-22]. A side-by-side comparison with a prior article that has quantified the abundance of disorder using a different collection of organisms reveals that these estimates are comparable: 19.6% that we estimated in the current analysis vs. 20.5% [19] and 18.9% [18] that were previously estimated in Eukaryota, 5.8% vs. 8.5% and 5.7% in Bacteria, 5.8 vs. 7.4% and 3.8% in Archaea, and 9.4% vs. 13.2% for viruses (the other study did not consider viral proteomes). The enrichment in Eukaryota is partially due to the documented involvement of the intrinsically disordered regions in cell division, gene regulation, and cellular differentiation [18, 51]. Furthermore, combination of alternative splicing and disorder was suggested to enable tissue-specific modulation of protein functions that is necessary for cell differentiation and evolution of multicellular organisms [52].

Similar to the two structural properties, the crystallization propensities are also distinct across taxonomic domains (Table 1). The highest propensity is found in Archaea, followed by Bacteria, viruses, and Eukaryota. This is in close agreement with a recent study that estimated the crystallization propensity in a similarly large and diverse set of proteomes [4]. Our estimate for the overall crystallization propensity in Archaea is 0.37 vs. 0.39 in the earlier study, 0.32 vs. 0.33 in Bacteria, and 0.14 vs. 0.14 in Eukaryota. As a point of reference, crystallization propensities for the structurally solved proteins in PDB has median value of 0.50 with the first and third quartiles equal to 0.27 and 0.81, respectively [4]. One

possible explanation for the high crystallization propensity in Archaea is that many of these organisms are thermophiles. Proteins in thermophilic organisms are particularly stable and consequently are more easily purified and processed for crystallization [53]. Bacteria are the most commonly used host organisms for expression of recombinant proteins for structure determination efforts; i.e., *E.coli* is used as the protein expression organism for 85.4% of the structures in PDB (120,977 out of 141,702 structures as of June 2018). This use of bacteria as factories for production of large number of copies of the recombinant DNA needed for the robust production of recombinant proteins is determined by several factors. These include rapid multiplication of bacteria [54], easiness to grow allowing obtaining high cell density cultures [55], easiness to maintain and manipulate in laboratory (culture media can be made from readily available and inexpensive components [56]), and presence of the plasmids, which are extrachromosomal elements of bacteria that can be used as carriers of recombinant DNA into cells that can be readily isolated from bacteria [57]. The fact that bacteria are able to produce massive amounts of recombinant proteins, still maintaining production of their own proteins at levels suitable for normal survival suggests that bacterial proteins are easier to be expressed compared to proteins from other domains of life, providing a plausible explanation for the relatively high crystallization propensities for these organisms. In contrast, Eukaryotes have higher levels of intrinsic disorder, which is one of the key characteristics that makes structural determination of these proteins via crystallization more challenging [48, 58, 59]. Interestingly, the crystallization propensities for viral proteins are in step with the propensities of the host proteomes in their respective domains of life. The crystallization propensities for archaeal, bacterial, and eukaryotic viruses are 0.42, 0.39 and 0.19, respectively (Table 1). This parallels the crystallization propensities of archaeal, bacterial, and eukaryotic proteins (0.37, 0.32, and 0.14, respectively). Similar results were also observed in [4].

Supplementary Figure S2 visualizes the interplay of the three structural characteristics that are aggregated at the level of domains of life (see results in bold font in Table 1). Eukaryota has the richest content of disorder coupled with the moderate levels of current structural coverage and lowest crystallization propensity. Bacteria and Archaea share similarly low disorder content. However, Bacteria have somewhat higher current structural coverage, which is compensated with lower propensity for crystallization when compared to Archaea. The viral proteomes are characterized by moderate levels of disorder and a wide range of current structural coverage and crystallization propensity, which is dependent on the taxonomic classification of their hosts. Eukaryotic viruses have the highest structural coverage coupled with the lowest propensity for crystallization. On the other side of the spectrum, archaeal viruses are featured by the lowest structural coverage and the highest propensity for crystallization.

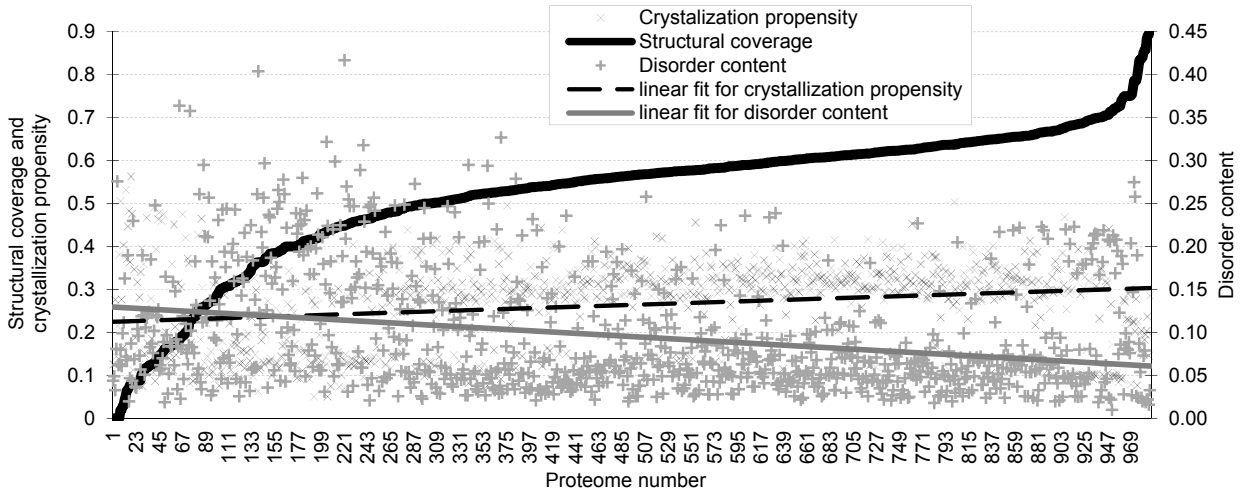

Figure 1. Relations between the current structural coverage, disorder content, and crystallization propensity for the 987 considered proteomes. The proteomes are arranged in the ascending order based on their corresponding structural coverage. The solid gray and dashed black lines show a linear fit into the disorder content and crystallization propensity values, respectively.

Proteome-level interplay between structural coverage, disorder content and crystallization propensity

Figure 1 compares the values of structural coverage, disorder content, and propensity for crystallization over the set of 987 considered proteomes. Interestingly, the values of disorder content and crystallization propensity show relatively weak relations with the structural coverage. The corresponding Pearson correlation coefficients (PCCs) of these two features are -0.23 and 0.17, respectively. This is represented by the relatively low value of slope for the linear fit into these data shown as solid gray and dashed black lines in Figure 1. The negative sign of the correlation between structural coverage and disorder content is in agreement with the fact that the presence of the disordered regions negatively impacts many of the steps involved in the protein crystallization process [58, 60]. In fact, the PCC between disorder content and crystallization propensity is strong and equals -0.65. We test veracity of these correlations by comparing results based on the applied here consensus disorder predictions with the use of another popular disorder predictor, VSL2 [61]. We collected VSL2 predictions from the D2P2 database [44] for a subset of 631 proteomes that are in common between this resource and our list of 987 complete proteomes. PCC over the 631 proteomes between the consensus and VSL2 predictions is 0.89, which suggests that these disorder estimates are in close agreement. This test confirms the weak relationships between disorder content and structural coverage (PCC = -0.35 for the consensus prediction vs. -0.36 for VSL2), and between crystallization propensity and structural coverage (PCC = 0.09 vs. 0.27). They also confirm strong negative correlation between disorder content and crystallization propensity (PCC = -0.69 for the consensus prediction vs. -0.76 for VSL2). The lower value of the correlation between disorder content and structural coverage can be explained by the underrepresentation of disorder in PDB, the public resource that defines current structural coverage. In other words, the high levels of current structural coverage are fueled by the focus on structured proteins (i.e. proteins lacking disorder) and structured fragments of hybrid proteins that have ordered domains and disordered regions. The latter is realized by

solving structures of structured domains or sequence fragments in the disorder-rich proteins, rather than complete protein structures.

A recent study illustrated this point by comparing the amount of disorder in the full protein chains collected from UniProt and protein sequences for which structures are included in PDB for the same set of 25 thousand proteins [33]. The disorder content in the PDB chains was estimated to be 5.5%, compared with 11.4% in the complete UniProt sequences [33]. This clearly shows a substantial bias toward solving structures of less disordered regions in the disordered proteins, which in turn lowers the value of the correlation. In an earlier related study, it was also pointed out that the vast majority of PDB proteins are shorter than their corresponding UniProt sequences and/or contain numerous residues, which are not observed in maps of electron density [62]. For example, only ~7% of structures in PDB correspond to proteins with complete amino acid sequences, and only ~25% of the PDB structures represent proteins with >95% of their sequences [62]. Furthermore, many PDB structures were shown to contain ambiguous regions; i.e., regions where more than one structure of the same protein sequence "disagree" in terms of their structural description. This ambiguity fits the nature of intrinsically disordered regions [62, 63].

However, the left side of Figure 1 also shows a large group of proteomes that have higher disorder content and lower structural coverage. The low positive correlation between structural coverage and propensity for crystallization can be explained by the disproportional focus on the eukaryotic proteins that are relatively difficult to be structurally solved (as indicated by low values of crystallization propensity). Eukaryotic proteins account for about 55% of the structures in PDB (120.5 out of 219 thousand structures). To compare, UniProt includes 29 million eukaryotic proteins compared to 79 million in Bacteria, 2.5 million in Achaea and 3.5 million viral proteins. Therefore, over half of structures target eukaryotic proteins that constitute only 25% of the currently known proteins.

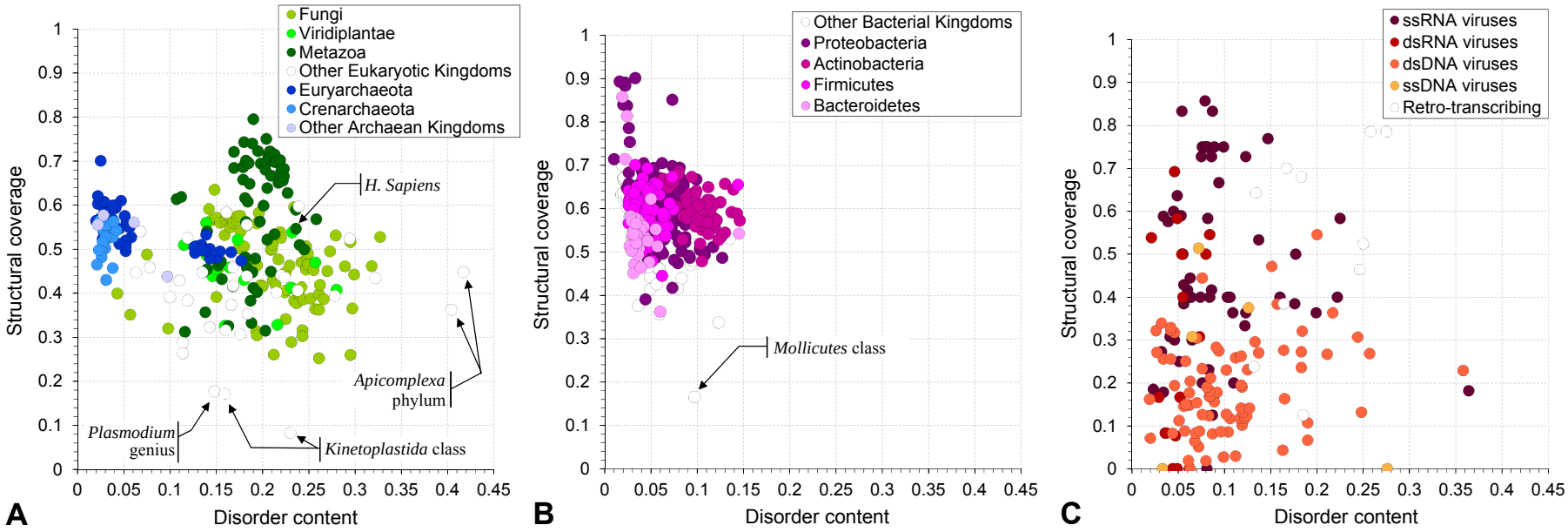

Figure 2. Relations between structural coverage, content of intrinsic disorder, and crystallization propensity (CP) across the proteomes within each domain of life. Panel A shows the Eukaryotic (colored in green) and Achaean (in blue) kingdoms/phyla. Panel B focuses on the Bacteria (in pink). Panel C shows the viruses (in red) categorized by the type of their genomes.

Figure 2 visualizes relationship between disorder content and structural coverage in the context of taxonomic classification. The disorder content and structural coverage values of proteomes in the same domain of life are clustered together (Figures 2A and 2B), except for viruses (Figure 2C). Nevertheless, we also observe distinct patterns across different domains, kingdoms, and phyla of life. The eukaryotic proteomes are characterized by a wide range of intrinsic disorder content (between a few and over 40%), and an even wider range of the current structural coverage that spans between about 10 and 80%. Figure 2A reveals that the animal proteomes have high current structural coverage, among which the puffer fish (*T. rubripes*) had the highest value of 80%, combined with the modest content of disorder, as opposed to the other Eukaryota. The proteome of *H. Sapiens* secures the structural coverage of 55% and disorder content of 24%; the latter value is close to a previous estimate of 25% [22]. Eukaryotes with the lowest structural coverage of below 20% include single-cell flagellates (from *Kinetoplastida* class) and a malaria parasite (*Plasmodium* genius). Eukaryotic species with the largest disorder content are in the *Apicomplexa* phylum. In contrast, the species in Archaea possess much narrower ranges of structural coverage (between 40 and 70%) and intrinsic disorder content (between a few and 20%), with the *Crenarchaeota* phylum having the disorder content consistently below 5%. The highest structural coverage is observed for thermophilic methanogen *M. fervidus*. Compared to Archaea, the proteomes in Bacteria are characterized by a similarly narrow range of the disorder content values combined with much broader range of structural coverage, between 17% for the parasitic bacteria in the *Mollicutes* class to about 90% in several *Proteobacteria* (Figure 2B). Interestingly, viruses have by far the largest spread in terms of the current structural coverage and disorder content (Figure 2C). The structural coverage values for the viral proteomes range between 0% (i.e. entirely dark proteomes) and over 80% for a few single-stranded RNA viruses, including Influenza A virus that has the structural coverage of 86%. Similarly, the disorder content values range between 2% and 36% for the viral proteomes, with the plant polerovirus having the largest amount of disorder content. This is in line with the conclusions of previous studies also reporting a wide spread of disorder predisposition across the viral proteomes [19, 20].

A comprehensive view of the relations between the three investigated structural characteristics and the taxonomic classification is summarized in Figure 3. Figure 3A shows the proteome similarity network (PSN), where nodes represent the considered 987 proteomes and edges correspond to the sequence similarities between the proteomes. The nodes in the network are colored according to their taxonomic classification where green, blue, pink, and red correspond to Eukaryota, Archaea, Bacteria, and viruses, respectively. The proteomes with >40% similarity are connected with black edges, those with between 20 and 40% similarity are colored with gray edges, while those with <20% similarity are rendered transparent. PSN is self-organized into multiple clusters according to the taxonomic classification, with various bacterial phyla localized in the lower left corner, eukaryotic kingdoms at the top, archaeal phyla in the middle, and viruses on the outside and in the vicinity of the domain of their host organisms. The constructed PSN facilitates convenient analysis of the taxonomic landscape of disorder content (Figure 5B), crystallization propensity (Figure 3C), and structural coverage (Figure 3D).

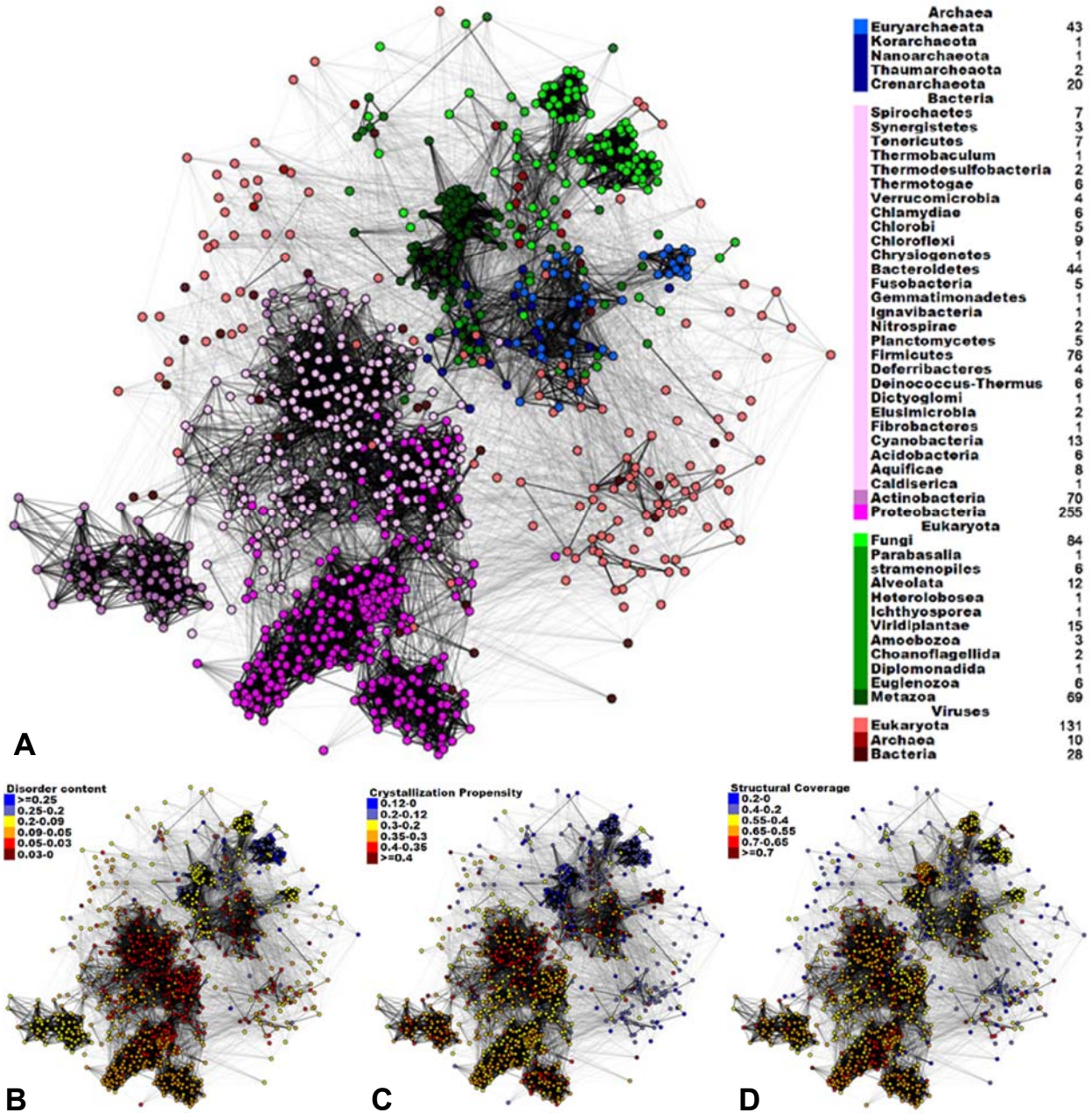

Figure 3. Proteome similarity network (PSN) for the considered 987 complete proteomes (species). Nodes denote complete proteomes. The length of an edge represents the average similarity between the corresponding proteomes where edges with >40% similarity are black, between 20 and 40% similarity are gray, and <20% similarity are transparent. Panel A gives PSN with color-coded taxonomic annotation. Panels B, C and D show PSN with color-coded annotations of the disorder content, propensity for structure determination, and current structural coverage, respectively. The networks were generated with the NAViGaTOR software [64].

Many proteomes with high structural coverage also tend to have high propensity for crystallization and low amounts of disorder (red and orange nodes in Figures 3B, 3C and 3D), which suggests that further increase in their structural coverage should be relatively easy to attain. These proteomes primarily include Bacteria with the exception of the *Actinobacteria* phylum that has the modest amounts of

disorder content. Eukaryotic proteomes have the moderate to high structural coverage coupled with low propensity for crystallization and the moderate to high levels of disorder content. Interestingly, Metazoan proteomes form two distinct clusters (dark green nodes in Figure 3A). While both clusters share similarly low propensity for crystallization, the upper cluster has comparatively higher structural coverage (Figure 3D) and lower disorder content (Figure 3B) when compared to the lower cluster. A similar observation can be made for the Fungi kingdom, which also forms two major clusters in Figure 3A. Both clusters share low crystallization propensity, while they trade lower disorder content and higher structural coverage (upper cluster) for higher disorder content and lower structural coverage (lower cluster). This suggests that sequence similarity at the level of proteomes is associated with the similarity in structural coverage and disorder content. The archaeal proteomes share higher sequence similarity with the eukaryotic proteomes while exhibiting less similarity to the bacterial proteomes (Figure 3A). However, they have higher propensity for crystallization, and many of them have lower disorder content when compared to the eukaryotic proteomes. Overall, this analysis reveals an interesting interplay between the three considered structural properties, sequence similarity, and taxonomic classification.

Analysis of dark proteomes

The current structural coverage measured in this study provides high-confidence annotations of dark proteins that lack structural coverage along the entire sequence. We aggregated this information at the proteome level to comprehensively annotate dark proteomes that are largely composed of dark proteins. Figure 4A shows distributions of the proteome-level structural coverage across all the proteomes (gray bars) and specific domains of life and viruses (colored bars). We defined the dark proteomes as the proteomes having the lowest quintile of the structural coverage. The left-most bar plot shows the range of structural coverage for the dark proteomes, which spans an interval between 0 and 43%. Sliding horizontally along this range of the structural coverage values reveals that the dark proteomes exclude archaeal species while including about 4% of bacterial organisms, 35% of eukaryotic organisms and 75% of viral proteomes.

Figure 4B compares the disorder content values of the dark proteomes and the proteomes located in the other four quintiles of the current structural coverage. The gray bars show statistics over all proteomes, while colored bars focus on specific domains of life and viruses. The bars at the top show *p*-values that quantify the statistical significance of the differences in the disorder content between the dark proteomes and each of the other four quintiles. Overall, disorder content is significantly lower for the proteomes in the top three quintiles when compared to the dark proteomes (gray bars in Figure 4B; p-values < 0.01). The dark archaeal proteomes are also significantly enriched in disorder compared to the remaining archaeal proteomes (p-values ≤ 0.05), while the differences in the disorder content for bacterial, eukaryotic, and viral proteomes are not significant. This means that the significant enrichment in disorder for the dark proteomes predominantly stems from the differences in disorder content across the domains of life. In other words, the dark proteomes are significantly enriched in the disorder, since they are primarily composed of viral and eukaryotic species. In contrast, only 10 bacterial and none of archaeal organisms are among the 198 dark proteomes.

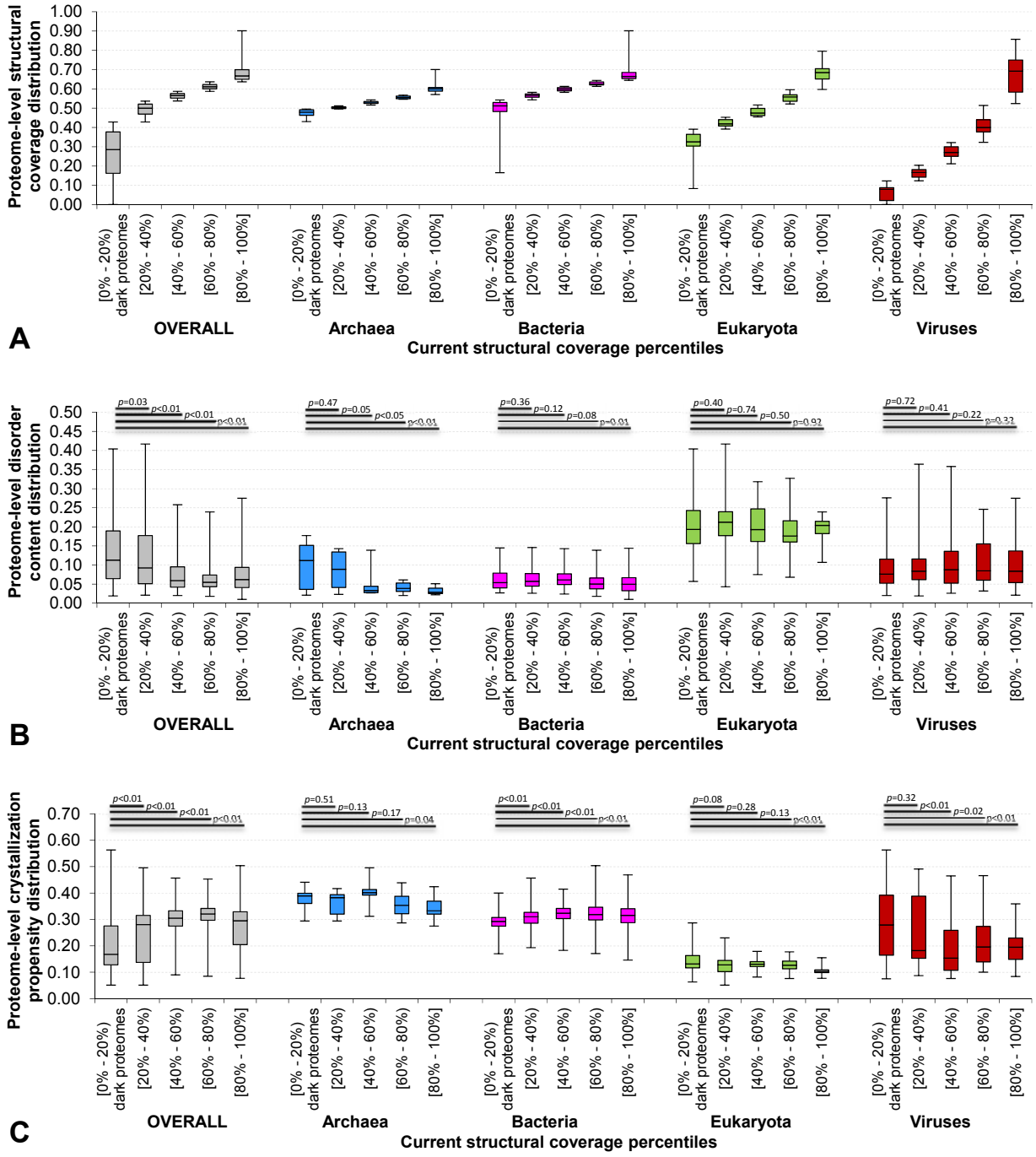

Figure 4. Distributions of the values of the current structural coverage (panel A), content of intrinsic disorder (panel B), and crystallization propensity (panel C) across all considered species (gray bars) and species in Archaea (blue), Bacteria (pink), Eukaryote (blue) and viruses (red). The x-axes group the species into the five consecutive quintiles of the structural coverage distribution. Correspondingly, the ranges in the panel A are disjoint. Each distribution is depicted by a box plot that shows the minimum, first quartile, median, third quartile and the maximum. *P*-values are shown at the top of panels B and C. They were generated with the Wilcoxon rank sum test that is used to quantify statistical significance of the differences in the intrinsic disorder content (panel B) and in the crystallization propensity (panel C) between the dark proteomes (the bottom quintile for structural coverage defined by the gray bars in panel A) and the proteomes in the other four quintiles.

Figure 4C contrasts the propensity for crystallization between the dark proteomes and the remaining proteomes. We observe that the dark proteomes have significantly lower propensity for crystallization when compared to the proteomes in each of the other four quintiles of the current structural coverage (gray bars in Figure 4C; p-values < 0.01). The same observation also holds true for the bacterial proteomes (pink bars in Figure 4C; p-values < 0.01). However, the crystallization propensity is similar between the dark archaeal/eukaryotic proteomes and the archaeal/eukaryotic proteomes that had higher levels of current structural coverage (green and blue bars in Figure 4C). This trend is actually reversed for the viral proteomes, where the dark viruses have significantly higher propensity for crystallization than the remaining viruses (red bars in Figure 4C; p-values ≤ 0.02). This suggests that although the majority of the dark proteomes are viruses, many of viral dark proteomes should be relatively easy to characterize structurally using X-ray crystallography, which currently is the method of choice for structural determination. Furthermore, the significant drop in the crystallization propensity for the dark proteomes (gray bars in Figure 4C) is mostly a result of the inclusion of a large number of eukaryotic species (62 out of 198 dark proteomes), which generally have low propensity for crystallization.

Supplementary Figure S3 provides a graphical overview of the three structural characteristics and taxonomic classification for the dark proteomes. A detailed discussion of this figure is included in the Supplement while here we summarize these observations. The figure reveals that the majority of the dark proteomes with the lowest current structural coverage are viruses. Virtually all dark proteomes that are depleted in disorder are viruses and most of these have at least moderate values of crystallization propensity. On the other hand, majority of dark eukaryotic proteomes are enriched in intrinsic disorder and have low propensity for crystallization. There are only a few dark bacterial proteomes and they have modest levels of both crystallization propensity and intrinsic disorder.

Overall, we have shown that the dark proteomes are primarily viruses and Eukaryota. The dark eukaryotic proteomes have significantly elevated disorder content and low propensity for crystallization, while the majority of the dark viral proteomes have at least moderate crystallization propensity and only a handful of them are enriched in the intrinsic disorder. This large-scale analysis sheds light on the interplay between key structural characteristics measured at the whole-proteome scale on the taxonomic landscape. It improves our understanding of the relationships between structural darkness, intrinsic disorder predisposition, and propensity for structural characterization across the three domains of life and in viruses.

Summary

It is recognized now that a significant fraction of the protein sequence space corresponds to the dark proteome; i.e., a set of sequences and sequence fragments for which structure is currently unknown and which are inaccessible to homology modelling. In order to systematically examine an interplay between structural darkness, intrinsic disorder, and crystallization propensity in the light of the taxonomic landscape, we computationally investigated 5,381,183 proteins from 987 complete proteomes that cover all three domains of life (Bacteria, Archaea, and Eukaryota) and viruses. Our analysis revealed that with their current structural coverage of almost 60%, bacterial proteomes are rather well structurally characterized in comparison with the proteomes of other domains of life and viruses, whereas viral proteomes are the most structurally dark. In fact, none of the 198 dark proteomes (i.e., proteomes having the lowest quintile of the structural coverage) are of archaeal origin, whereas 4% of bacterial proteomes,

35% of eukaryotic proteomes, and 75% of viral proteomes are structurally dark at the moment. Our analysis also showed that Eukaryota have substantially higher levels of intrinsic disorder, particularly when compared to Bacteria and Archaea, and viruses are characterized by the broadest distribution of the intrinsic disorder content values. We also show that the dark proteomes are significantly enriched in the disorder, being primarily composed of viral and eukaryotic species. As far as crystallizability is concerned, the highest predisposition for structural characterization is found in Archaea, followed by Bacteria, viruses, and Eukaryota. Furthermore, the crystallization propensities for viral proteins are correlated with the crystallizabilities of the host proteomes in their respective domains of life. In fact, the crystallization propensities for archaeal, bacterial, and eukaryotic viruses are 0.42, 0.39 and 0.19, respectively, which aligns with the crystallization propensities of archaeal, bacterial, and eukaryotic proteins (0.37, 0.32, and 0.14, respectively). Importantly, although the majority of the dark proteomes belong to viruses, many of the dark viral proteomes should be relatively easy to characterize structurally by X-ray crystallography. We also built proteome similarity network (PSN) that illustrates a comprehensive view of the relations between the three investigated structural characteristics and the taxonomic classification. Utilization of this network for visualization of the taxonomic landscape of disorder content, crystallization propensity, and structural coverage revealed the existence of an interesting interplay between the three considered structural properties, sequence similarity, and taxonomic classification.

In summary, we show that the significant majority of the dark proteomes are of viral and eukaryotic origin. In Eukaryota, the dark proteomes are characterized by high disorder content and low crystallization propensity. On the other hand, the majority of the dark viral proteomes have moderate crystallization propensities and only some of them possess high intrinsic disorder predispositions. Furthermore, our comprehensive, whole-proteome-scale analysis illuminates the existence of an intricate relationship between the structural darkness and intrinsic disorder, thereby providing a taxonomic landscape view of an interplay between structural darkness, intrinsic disorder predisposition, and propensity for structural characterization across the three domains of life.

Acknowledgments

This research was supported in part by the Qimonda Endowment and the National Science Foundation grant 1617369 to Lukasz Kurgan.

The authors have declared no conflict of interest.

References

[1] E. R. Mardis, Annu Rev Anal Chem (Palo Alto Calif) 2013, 6, 287; Y. Kodama, M. Shumway, R. Leinonen, C. International Nucleotide Sequence Database, Nucleic Acids Res 2012, 40, D54.

[2] M. Grabowski, E. Niedzialkowska, M. D. Zimmerman, W. Minor, J Struct Funct Genomics 2016, 17, 1.

[3] B. H. Dessailly, R. Nair, L. Jaroszewski, J. E. Fajardo, A. Kouranov, D. Lee, A. Fiser, A. Godzik, B. Rost, C. Orengo, Structure 2009, 17, 869.

[4] M. J. Mizianty, X. Fan, J. Yan, E. Chalmers, C. Woloschuk, A. Joachimiak, L. Kurgan, Acta Crystallogr D Biol Crystallogr 2014, 70, 2781.

[5] A. Godzik, Curr Opin Struct Biol 2011, 21, 398; I. S. Povolotskaya, F. A. Kondrashov, Nature 2010, 465,

922.
[6] K. Khafizov, C. Madrid-Aliste, S. C. Almo, A. Fiser, Proc Natl Acad Sci U S A 2014, 111, 3733; M. Levitt, Proc Natl Acad Sci U S A 2009, 106, 11079.
[7] J. Haas, S. Roth, K. Arnold, F. Kiefer, T. Schmidt, L. Bordoli, T. Schwede, Database (Oxford) 2013, 2013, bat031; D. Petrey, T. S. Chen, L. Deng, J. I. Garzon, H. Hwang, G. Lasso, H. Lee, A. Silkov, B. Honig, Curr Opin Struct Biol 2015, 32, 33.
[8] U. Pieper, B. M. Webb, G. Q. Dong, D. Schneidman-Duhovny, H. Fan, S. J. Kim, N. Khuri, Y. G. Spill, P. Weinkam, M. Hammel, J. A. Tainer, M. Nilges, A. Sali, Nucleic Acids Res 2014, 42, D336.
[9] I. Friedberg, Brief Bioinform 2006, 7, 225; F. Pazos, M. J. Sternberg, Proc Natl Acad Sci U S A 2004, 101, 14754; O. C. Redfern, B. Dessailly, C. A. Orengo, Curr Opin Struct Biol 2008, 18, 394.
[10] V. Lounnas, T. Ritschel, J. Kelder, R. McGuire, R. P. Bywater, N. Foloppe, Comput Struct Biotechnol J 2013, 5, e201302011; P. J. Gane, P. M. Dean, Curr Opin Struct Biol 2000, 10, 401; A. Jazayeri, J. M. Dias, F. H. Marshall, J Biol Chem 2015, 290, 19489; K. Lundstrom, Mol Biotechnol 2006, 34, 205.
[11] L. Xie, L. Xie, P. E. Bourne, Curr Opin Struct Biol 2011, 21, 189; G. Hu, K. Wang, J. Groenendyk, K. Barakat, M. J. Mizianty, J. Ruan, M. Michalak, L. Kurgan, Bioinformatics 2014, 30, 3561.
[12] F. Moriaud, S. B. Richard, S. A. Adcock, L. Chanas-Martin, J. S. Surgand, M. Ben Jelloul, F. Delfaud, Brief Bioinform 2011, 12, 336; B. Karaman, W. Sippl, Curr Med Chem 2018; H. K. Ho, L. Zhang, K. Ramamohanarao, S. Martin, Methods Mol Biol 2013, 932, 87; R. G. Govindaraj, M. Naderi, M. Singha, J. Lemoine, M. Brylinski, NPJ Syst Biol Appl 2018, 4, 13.
[13] E. L. Wise, I. Rayment, Acc Chem Res 2004, 37, 149; I. G. Choi, S. H. Kim, Proc Natl Acad Sci U S A 2006, 103, 14056.
[14] A. Grant, D. Lee, C. Orengo, Genome Biol 2004, 5, 107; L. Holm, C. Sander, Science 1996, 273, 595; S. Nepomnyachiy, N. Ben-Tal, R. Kolodny, Proc Natl Acad Sci U S A 2014, 111, 11691; R. Kolodny, L. Pereyaslavets, A. O. Samson, M. Levitt, Annu Rev Biophys 2013, 42, 559.
[15] P. Lieutaud, F. Ferron, A. V. Uversky, L. Kurgan, V. N. Uversky, S. Longhi, Intrinsically Disord Proteins 2016, 4, e1259708; A. Bhowmick, D. H. Brookes, S. R. Yost, H. J. Dyson, J. D. Forman-Kay, D. Gunter, M. Head-Gordon, G. L. Hura, V. S. Pande, D. E. Wemmer, P. E. Wright, T. Head-Gordon, J Am Chem Soc 2016, 138, 9730.
[16] N. Perdigao, A. C. Rosa, S. I. O'Donoghue, BioData Min 2017, 10, 24.
[17] N. Perdigao, J. Heinrich, C. Stolte, K. S. Sabir, M. J. Buckley, B. Tabor, B. Signal, B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, S. I. O'Donoghue, Proc Natl Acad Sci U S A 2015, 112, 15898.
[18] J. J. Ward, J. S. Sodhi, L. J. McGuffin, B. F. Buxton, D. T. Jones, J Mol Biol 2004, 337, 635.
[19] Z. Peng, J. Yan, X. Fan, M. J. Mizianty, B. Xue, K. Wang, G. Hu, V. N. Uversky, L. Kurgan, Cell Mol Life Sci 2015, 72, 137.
[20] B. Xue, A. K. Dunker, V. N. Uversky, J Biomol Struct Dyn 2012, 30, 137.
[21] Z. Peng, M. J. Mizianty, L. Kurgan, Proteins 2014, 82, 145.
[22] J. Yan, M. J. Mizianty, P. L. Filipow, V. N. Uversky, L. Kurgan, Biochim Biophys Acta 2013, 1834, 1671.
[23] Z. Peng, B. Xue, L. Kurgan, V. N. Uversky, Cell Death Differ 2013, 20, 1257.
[24] D. Vitkup, E. Melamud, J. Moult, C. Sander, Nat Struct Biol 2001, 8, 559.
[25] J. Hou, S. R. Jun, C. Zhang, S. H. Kim, Proc Natl Acad Sci U S A 2005, 102, 3651; M. Osadchy, R. Kolodny, Proc Natl Acad Sci U S A 2011, 108, 12301; N. O'Toole, S. Raymond, M. Cygler, J Struct Funct Genomics 2003, 4, 47; K. Yura, A. Yamaguchi, M. Go, J Struct Funct Genomics 2006, 7, 65.
[26] C. The UniProt, Nucleic Acids Res 2017, 45, D158; C. UniProt, Nucleic Acids Res 2015, 43, D204.
[27] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, Nucleic Acids Res 2000, 28, 235; S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, S. Velankar, Methods Mol Biol 2017, 1607, 627.

[28] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Nucleic Acids Res 1997, 25, 3389.
[29] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, Nat Meth 2015, 12, 7; J. Yang, Y. Zhang, Nucleic Acids Res 2015, 43, W174.
[30] J. Soding, A. Biegert, A. N. Lupas, Nucleic Acids Res 2005, 33, W244; A. Hildebrand, M. Remmert, A. Biegert, J. Soding, Proteins 2009, 77 Suppl 9, 128.
[31] B. Webb, A. Sali, Methods Mol Biol 2017, 1654, 39; A. Sali, L. Potterton, F. Yuan, H. van Vlijmen, M. Karplus, Proteins 1995, 23, 318.
[32] V. Mariani, F. Kiefer, T. Schmidt, J. Haas, T. Schwede, Proteins 2011, 79 Suppl 10, 37.
[33] I. Walsh, M. Giollo, T. Di Domenico, C. Ferrari, O. Zimmermann, S. C. Tosatto, Bioinformatics 2015, 31, 201.
[34] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, K. Fidelis, Proteins 2014, 82 Suppl 2, 127; F. Meng, V. N. Uversky, L. Kurgan, Cell Mol Life Sci 2017, 74, 3069; F. Meng, V. Uversky, L. Kurgan, Curr Protoc Protein Sci 2017, 88, 2 16 1; B. Monastyrskyy, K. Fidelis, J. Moult, A. Tramontano, A. Kryshtafovych, Proteins 2011, 79 Suppl 10, 107.
[35] Z. L. Peng, L. Kurgan, Curr Protein Pept Sci 2012, 13, 6.
[36] Z. Peng, L. Kurgan, Pac Symp Biocomput 2012, 176.
[37] X. Fan, L. Kurgan, J Biomol Struct Dyn 2014, 32, 448.
[38] M. Necci, D. Piovesan, Z. Dosztanyi, S. C. E. Tosatto, Bioinformatics 2017, 33, 1402.
[39] Z. Dosztanyi, V. Csizmok, P. Tompa, I. Simon, J Mol Biol 2005, 347, 827.
[40] I. Walsh, A. J. Martin, T. Di Domenico, S. C. Tosatto, Bioinformatics 2012, 28, 503.
[41] D. Piovesan, F. Tabaro, I. Micetic, M. Necci, F. Quaglia, C. J. Oldfield, M. C. Aspromonte, N. E. Davey, R. Davidovic, Z. Dosztanyi, A. Elofsson, A. Gasparini, A. Hatos, A. V. Kajava, L. Kalmar, E. Leonardi, T. Lazar, S. Macedo-Ribeiro, M. Macossay-Castillo, A. Meszaros, G. Minervini, N. Murvai, J. Pujols, D. B. Roche, E. Salladini, E. Schad, A. Schramm, B. Szabo, A. Tantos, F. Tonello, K. D. Tsirigos, N. Veljkovic, S. Ventura, W. Vranken, P. Warholm, V. N. Uversky, A. K. Dunker, S. Longhi, P. Tompa, S. C. Tosatto, Nucleic Acids Res 2016, D1, D219.
[42] I. Na, F. Meng, L. Kurgan, V. N. Uversky, Mol Biosyst 2016, 12, 2798; F. Meng, I. Na, L. Kurgan, V. N. Uversky, Int J Mol Sci 2016, 17; Z. Peng, C. J. Oldfield, B. Xue, M. J. Mizianty, A. K. Dunker, L. Kurgan, V. N. Uversky, Cell Mol Life Sci 2014, 71, 1477; G. Hu, Z. Wu, K. Wang, V. N. Uversky, L. Kurgan, Current drug targets 2016, 17, 1198; C. Wang, V. N. Uversky, L. Kurgan, Proteomics 2016, 16, 1486; Z. Peng, V. N. Uversky, L. Kurgan, Intrinsically Disordered Proteins 2016, 4, e1262225.
[43] T. Di Domenico, I. Walsh, A. J. M. Martin, S. C. E. Tosatto, Bioinformatics 2012, 28, 2080; E. Potenza, T. Di Domenico, I. Walsh, S. C. Tosatto, Nucleic Acids Res 2015, 43, D315.
[44] M. E. Oates, P. Romero, T. Ishida, M. Ghalwash, M. J. Mizianty, B. Xue, Z. Dosztanyi, V. N. Uversky, Z. Obradovic, L. Kurgan, A. K. Dunker, J. Gough, Nucleic Acids Res 2013, 41, D508.
[45] H. M. Berman, B. C. Narayanan, L. Di Costanzo, S. Dutta, S. Ghosh, B. P. Hudson, C. L. Lawson, E. Peisach, A. Prlic, P. W. Rose, C. H. Shao, H. W. Yang, J. Young, C. Zardecki, Febs Letters 2013, 587, 1036.
[46] J. Gao, Z. Wu, G. Hu, K. Wang, J. Song, A. Joachimiak, L. Kurgan, Curr Protein Pept Sci 2018, 19, 200.
[47] L. Kurgan, M. J. Mizianty, Nat. Science 2009, 1, 93; M. J. Mizianty, L. Kurgan, Biochem Biophys Res Commun 2009, 390, 10; G. Babnigg, A. Joachimiak, J Struct Funct Genomics 2010, 11, 71.
[48] H. Wang, L. Feng, G. I. Webb, L. Kurgan, J. Song, D. Lin, Brief Bioinform 2017, https://doi.org/10.1093/bib/bbx018.
[49] F. Meng, C. Wang, L. Kurgan, BMC Bioinformatics 2018, 18, 580.
[50] D. Prangishvili, Annu Rev Microbiol 2013, 67, 565; M. K. Pietila, T. A. Demina, N. S. Atanasova, H. M. Oksanen, D. H. Bamford, Trends Microbiol 2014, 22, 334; E. V. Koonin, V. V. Dolja, M. Krupovic,

Virology 2015, 479-480, 2.
[51] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, Z. Obradovic, J. Proteome Res. 2007, 6, 1882.
[52] 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, A. K. Dunker, Proc. Natl. Acad. Sci. U. S. A. 2006, 103, 8390; M. Buljan, G. Chalancon, A. K. Dunker, A. Bateman, S. Balaji, M. Fuxreiter, M. M. Babu, Curr Opin Struct Biol 2013, 23, 443; M. Buljan, G. Chalancon, S. Eustermann, G. P. Wagner, M. Fuxreiter, A. Bateman, M. M. Babu, Mol Cell 2012, 46, 871.
[53] M. Sadeghi, H. Naderi-Manesh, M. Zarrabi, B. Ranjbar, Biophys Chem 2006, 119, 256; F. E. Jenney, Jr., M. W. Adams, Extremophiles 2008, 12, 39.
[54] G. Sezonov, D. Joseleau-Petit, R. D'Ari, J Bacteriol 2007, 189, 8746.
[55] S. Y. Lee, Trends Biotechnol 1996, 14, 98; J. Shiloach, R. Fass, Biotechnol Adv 2005, 23, 345.
[56] G. L. Rosano, E. A. Ceccarelli, Front Microbiol 2014, 5, 172.
[57] B. Pope, H. M. Kent, Nucleic Acids Res 1996, 24, 536.
[58] C. J. Oldfield, B. Xue, Y. Y. Van, E. L. Ulrich, J. L. Markley, A. K. Dunker, V. N. Uversky, Biochim Biophys Acta 2013, 1834, 487.
[59] M. J. Mizianty, L. Kurgan, Bioinformatics 2011, 27, i24.
[60] D. E. Johnson, B. Xue, M. D. Sickmeier, J. Meng, M. S. Cortese, C. J. Oldfield, T. Le Gall, A. K. Dunker, V. N. Uversky, J Struct Biol 2012, 180, 201.
[61] K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker, Z. Obradovic, BMC Bioinformatics 2006, 7, 208.
[62] T. Le Gall, P. R. Romero, M. S. Cortese, V. N. Uversky, A. K. Dunker, J Biomol Struct Dyn 2007, 24, 325.
[63] S. DeForte, V. N. Uversky, Protein Sci 2016, 25, 676.
[64] K. R. Brown, D. Otasek, M. Ali, M. J. McGuffin, W. Xie, B. Devani, I. L. Toch, I. Jurisica, Bioinformatics 2009, 25, 3327.

Supplementary materials

The supplementary materials include three figures (S1, S2 and S3), additional details for the “Construction of the proteome similarity network” and “Analysis of dark proteomes” sections, and the complete list of the 987 considered species. The list includes information about the corresponding domain of life, taxonomic identifier, and taxonomic classification.