Comprehensive Intrinsic Disorder Analysis of 6108 Viral Proteomes: From the Extent of Intrinsic Disorder Penetrance to Functional Annotation of Disordered Viral Proteins

Naveen Kumar, Rahul Kaushik, Chandana Tennakoon, Vladimir N Uversky, Sonia Longhi, Kam y J Zhang, Sandeep Bhatia

Comprehensive intrinsic disorder analysis of 6,108 viral proteomes: From the extent of intrinsic disorder penetrance to functional annotation of disordered viral proteins

Naveen Kumar[1¶], Rahul Kaushik[2¶], Chandana Tennakoon[3¶], Vladimir N. Uversky[4,5], Sonia Longhi[6], Kam Y. J. Zhang[2], Sandeep Bhatia[1]*

[1]Diagnostic & Vaccine Group, ICAR- National Institute of High Security Animal Diseases, Bhopal 462022, India

[2]Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research, RIKEN, 1-7-22 Suehiro, Yokohama, Kanagawa 230-0045, Japan

[3]The Pirbright Institute, Woking GU24 0NF, UK

[4]Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

[5]Institute for Biological Instrumentation of the Russian Academy of Sciences, Federal Research Center 'Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences', Moscow region, 142290 Pushchino, Russia

[6]Aix Marseille Univ and CNRS, Laboratoire Architecture et Fonction des Macromolecules Biologiques (AFMB), UMR 7257, Marseille, France

¶These authors contributed equally to this work.

*To whom correspondence should be addressed: Naveen Kumar, Diagnostics & Vaccines Group, ICAR- National Institute of High Security Animal Diseases, Bhopal 462022, India; Email: naveen.kumar4@icar.gov.in; navyog.yadav84@gmail.com; Phone: +91-7552759204; Fax: +91-755-2758842.

ABSTRACT

Much of our understanding of proteins and proteomes comes from the traditional protein structure-function paradigm. However, in the last two decades, both computational and experimental studies have provided evidence that a large fraction of functional proteomes across different domains of life consists of intrinsically disordered proteins thus triggering a quest to unravel and decipher protein intrinsic disorder. Unlike structured/ordered proteins, intrinsically disordered proteins/regions (IDPs/IDRs) do not possess a well-defined structure under physiological conditions, and exist as highly dynamic conformational ensembles. In spite of this peculiarity, these proteins have crucial roles in cell signaling and regulation. To date, studies on the abundance and function of IDPs/IDRs in viruses are rather limited. In order to fill this gap, we carried out an extensive and thorough bioinformatics analysis of 283,000 proteins from 6,108 reference viral proteomes. We analyzed protein intrinsic disorder from multiple perspectives, such as abundance of IDPs/IDRs across diverse virus types, their functional annotations, and subcellular localization in taxonomically divergent hosts. We show that the content of IDPs/IDRs in viral proteomes varies broadly as a function of virus genome types and

taxonomically divergent hosts. We have combined the two most commonly used and accurate IDPs predictor results with Charge-Hydropathy (CH) versus Cumulative Distribution Function (CDF) plots to categorize the viral proteins according to their IDRs content and physico-chemical properties. Mapping of gene ontology on disorder content of viral proteins reveals that IDPs are primarily involved in key virus-host interactions, and host antiviral immune response down-regulation, which is reinforced by the post-translational modifications tied to disorder-enriched viral proteins. The present study offers detailed insights into the prevalence of the intrinsic disorder in viral proteomes and provides appealing targets for the design of novel therapeutics.

Keywords: Intrinsically disordered proteins, Viruses, Gene ontology, Post-translational modifications, Subcellular localization

INTRODUCTION

Proteins are among the most important biological macromolecules that govern the majority of cellular processes required for the sustenance of life. To carry out those diverse and essential functions, many proteins acquire unique well-defined structures. This phenomenon is known as the classic protein structure-function paradigm, which states that a unique protein function is tied to its unique three-dimensional (3D) structure, which is encoded in the unique amino acid sequence[1, 2]. Nevertheless, many biologically active proteins or protein regions fail to acquire distinctive 3D structures under physiological conditions either completely or partially. Such proteins and regions are known as intrinsically disordered proteins/regions (IDPs/IDRs)[3]. Importantly, such proteins/regions are not “broken”, since the lack of a specific 3D structure

does not prevent them from performing a broad spectrum of crucial and diverse biological functions. Instead, intrinsic disorder defines the multi-functionality of these proteins and makes them essential for the control of cellular signaling networks, regulation of various biological processes, and disease-related pathways[4-17]. Indeed, IDRs are abundant in different domains of life and proteomes have been extensively analyzed in terms of the abundance of IDPs/IDRs. Such studies include the bioinformatics characterization of six archaeans, 13 bacterial, and five eukaryotic proteomes[18], 332 prokaryotic proteomes[19], 53 archaean proteomes[20], as well as disorder-focused analysis of 965 proteomes including 59 archaean, 110 eukaryotic, 471 bacterial, and 325 viral proteomes[21]. Besides, in 2012, Xue et al. had carried out the most comprehensive study on disorder abundance in 3,484 proteomes including 73 archaeans, 951 bacterial, 67 eukaryotic, and 2,393 viral proteomes[22]. The trend of disorder abundance is quite similar in all these studies, viz., eukaryotes are steadily predicted to have much higher disorder content than prokaryotes.

Conversely, viruses are ubiquitous parasites capable of infecting representatives of all domains of life. Despite having a much smaller proteome size, they easily hijack the complex cellular processes of their hosts across all the domains of life. Due to the very limited size of the viral proteome, and to achieve a steady and successful perpetuation, proteins encoded by viruses are destined to be multi-functional. Therefore, studying IDPs/IDRs in viruses is of great interest. Previous studies have established that IDPs/IDRs in viruses play diverse roles[23-29], such as adaptation to their host[30], oncogenicity[26, 31-33], binding to host cells[34], and virus replication and pathogenesis[35-41]. Although previous studies systematically analyzed specific structural and functional features of intrinsic disorder in viral proteins[42-44], and in spite of the obvious

importance of intrinsic disorder for the function of many viral proteins, comprehensive analyses of IDPs/IDRs encoded by viruses and their functional annotations are relatively limited.

To fill this gap, we performed a comprehensive and detailed analysis of 283,000 viral proteins from 6,108 reference viral proteomes, utilizing the two most commonly used and accurate intrinsic disorder predictors (IUPred2A[45] and Espritz[46]). We gathered the proteome-level analyses to describe intrinsic disorder abundances and differences across the different types of viral genomes, diverse proteome sizes, and distinct cellular replication sites in the taxonomically divergent hosts. Furthermore, we classified viral proteins by Charge-Hydropathy (CH) *versus* Cumulative Distribution Function (CDF) plot, and subsequently, we evaluated the correlation between the disorder content of viral proteins and their functional annotations (*i.e.* gene ontologies, and post-translational modifications (PTMs)), and subcellular localization.

MATERIALS AND METHODS

Reference virus proteome dataset

In this study, we retrieved proteomes of a total of 6,108 reference viruses from the UniProt database[47] that include more than 283,000 viral proteins. On the basis of the Baltimore Classification, the genomes of these reference viruses were classified into ten groups based on the nature of their nucleic acids viz., dsDNA, (n = 2,596), ssDNA (n = 741), dsRNA (n = 189), ssRNA(-) (n = 413), ssRNA(+) (n = 597), dsDNA-RT (n = 76), ssRNA-RT (n = 55), satellite (n = 59), virophage (n = 7), and unclassified (n = 1,375) viruses. The host information of these reference viruses were retrieved from virus-host DB[48] and categorized into eight taxonomically distinct groups, Animalia (n = 1,882), Archaea (n = 150), Eubacteria (n = 2,227), Fungi (n = 79), Plantae (n = 757), Protista (n = 41), Environment (n = 59), and Unclassified (n = 916) (Table S1).

Prediction of Intrinsically Disordered Regions (IDRs)

It is largely recommended to employ more than one disorder predictor for the identification of IDRs to improve the accuracy and reliability of predictions. Hence, we used two different disorder predictors namely, IUPred2A[45] and Espritz[46], where the former estimates total pairwise interaction energy from the amino acid compositions, and the latter relies on bi-directional recursive neural networks, a sequence-based machine learning algorithm, for the prediction of IDRs. The additional advantage of using IUPred2A is that it can predict two different types of IDRs, namely short (at least 10 consecutive residues) and long disorder (at least 30 consecutive residues). Using Espritz on a dataset of reference viral proteome, we predicted IDRs by a short-disorder prediction method with a 5% false-positive rate[46]. Both disorder predictors offer a *per* residue intrinsic disorder (ID) probability score —that ranges from 0 to 1— for each viral protein sequence. Residues with an ID probability score ≥ 0.5 are classified as disordered. The total content of ID was estimated for each viral protein sequence as inputs as the fraction of the total number of residues with an ID probability score ≥ 0.5 out of the total residues in that protein sequence. A similar methodology was then used for the estimation of the total content of ID in a virus proteome. In addition, the viral proteins were classified depending on their ID content (as the fraction of disordered residues over the total residues) as ordered (ID content < 0.1), moderately disordered (ID content, 0.1 to 0.3), and highly disordered (ID content > 0.3) proteins. After doing so, the fraction of disordered proteins (ID content > 0.3) over total proteins for each of the viral proteome was also calculated as a metric to determine the ID content.

Sub-classification of structure-disorder tendencies of viral proteins by the CH-CDF plot

All the proteins (n = 283,160) of 6,108 reference viral proteomes were subjected to binary classification into ordered (structured) proteins and IDPs using Charge-Hydropathy (CH)[14, 49],

and Cumulative Distribution Function (CDF) classifiers[49, 50]. Both the classifiers provide a binary classification of a protein sequence as ordered or disordered, which leads to four possibilities when implemented in combination[50-53]. These four possibilities are usually indicated as four quadrants of the CH-CDF plot. Quadrant-I (Q-I) corresponds to proteins that are classified as disordered by the CH classifier (CH > 0) and ordered by the CDF classifier (CDF > 0). Quadrant-II (Q-II) represents the proteins that are classified as ordered by both the CH and CDF classifiers (CH < 0 and CDF > 0). Quadrant-III (Q-III) represents proteins that are ordered according to the CH classifier (CH < 0) and disordered as per the CDF classifier (CDF < 0). Quadrant-IV (Q-IV) accounts for proteins classified as disordered by both the classifiers (CH > 0 and CDF < 0).

Correlation of disorder contents with functional annotations and subcellular locations of viral proteins

We first extracted the gene ontology terms (GO Terms), post-translational modifications (PTMs), and subcellular locations of individual viral proteins from the UniProt database with the help of bash shell scripts in an automated mode. A further data mining was performed on the raw information to derive the metadata and thereafter, the viral proteins were segregated in accordance with the CH-CDF plots (unusual/rare, structured, mixed, and disordered viral proteins). The average/baseline disorder content for each category of CH-CDF classified viral proteins were calculated and significant differences in the ID content of a given GO term/PTMs/subcellular location from corresponding baselines were designated as ID enriched or depleted.

Statistical analyses

Statistical analyses were performed using GraphPad Prism 7.01 (GraphPad Software, San Diego, CA, USA). We used Dunn's multiple comparisons test, a non-parametric test, to compare the difference in the ID content among the viruses. In addition, one-way analysis of variance (ANOVA) with the Bonferroni method for multiple comparisons was used to assess the difference between the mean IDRs content of viruses that replicate in the nucleus and that of viruses replicating in the cytoplasm. In all the statistical analyses, a p-value less than 0.01 was considered statistically significant. All the graphs were generated by using GraphPad Prism 7.01.

RESULTS

ID content in the viral proteomes is governed by both virus genome types and host species

On comparing the two different types of ID predictions (namely, short and long disorder) implemented in IUPred2A on the virus proteome dataset, an extremely high significant positive correlation was observed ($r = 0.980$, $p < 0.0001$) (Figure 1A). Likewise, we also detected a highly significant positive correlation between Espritz and IUPred2A short disorder ($r = 0.758$, $p < 0.0001$) (Figure 1B), and Espritz and IUPred2A long disorder ($r = 0.743$, $p < 0.0001$) (Figure 1C) indicating that the observations and interpretations of results based on these two different ID predictors/types are comparable.

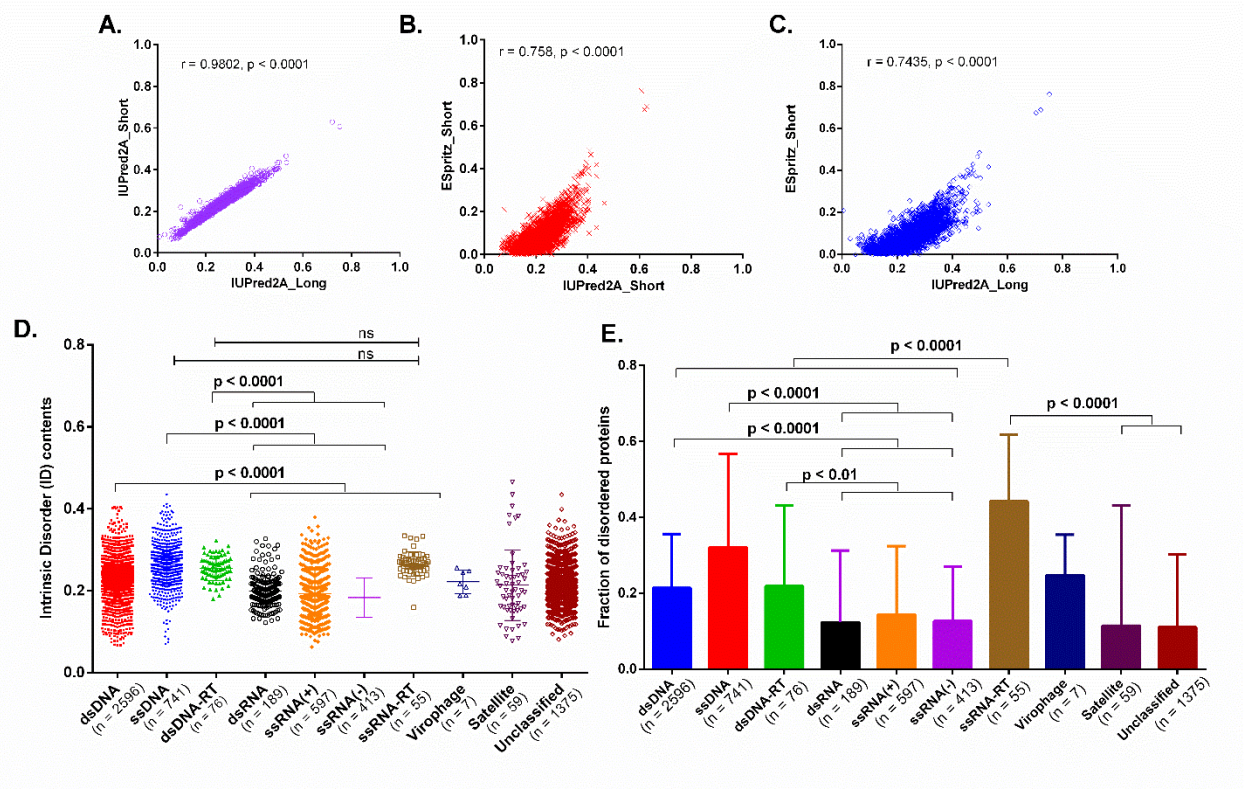

Figure 1. The diversity of intrinsically disordered (ID) contents across the diverse reference viral proteomes. The correlation analyses between (A) the long and short IDRs types predictions implemented in IUPred2A algorithm, (B) the short IDRs predictions type implemented in IUPred2A and the short-disorder prediction method of Espritz, and (C) the long IDRs predictions type implemented in IUPred2A with the short-disorder prediction method of Espritz are presented. (D) shows the extent of variability in the ID contents (IUPred2A_short method) across the various virus genome types. (E) represents the fraction of disordered proteins (ID content > 0.3) over total proteins for each of the viral proteomes across diverse virus genome types. One-way analysis of variance (ANOVA) with the Bonferroni method for multiple comparisons was used to estimate the differences in the ID contents of reference virus genome types. The error bars show the standard deviations and a p-value less than 0.01 was considered statistically significant.

Notably, DNA viruses were found to have a significantly higher mean ID content (0.233-0.264) compared to RNA viruses (0.183-0.199) ($p < 0.0001$), except for ssRNA-RT viruses (Figure 1D). In addition, on classifying the viral proteins depending on their ID content (as the fraction of disordered residues over the total residues) as ordered (ID content < 0.1), moderately disordered (ID content, 0.1 to 0.3), and highly disordered (ID content > 0.3), we found that most of the viral proteomes are enriched in moderately disordered proteins (Table S2). On investigation of the fraction of disordered proteins over total proteins as a metric to the ID content in virus proteomes, the finding that DNA viruses carry high mean ID content than that of RNA viruses remains consistent ($p < 0.01$ - 0.0001) (Figure 1E). Besides, of all virus genome types, ssRNA-RT viruses have an exceptionally high fraction of disordered proteins ($p < 0.0001$). However, viruses sharing the same genome type show a marked variation in their mean ID content as a function of their host (Table S3). While investigating the influence of cellular sites of replication that different viruses prefer, we observed that the members of dsDNA viral families that use the nucleus as a site of viral genome replication show a higher mean ID content (0.179 - 0.318), as compared to those that use cytoplasm (such as *Poxviridae* = 0.138±0.044) ($p < 0.01$ to < 0.0001) (Figure 2A). Of note, ssRNA(-) family members that replicate in the nucleus (especially *Orthomyxoviridae*) show a significantly higher mean ID content compared to those that replicate in the cytoplasm ($p < 0.0001$) (Figure 2B). This finding is consistent even after comparing pooled ssRNA(-) family members that replicate in the nucleus against those of the cytoplasm ($p < 0.0001$) (Figure 2C). In general, viruses of either DNA or RNA genome type that prefer the nucleus as a site of replication show a higher mean content in IDRs. These results

suggest that virus genome type and host species together govern the IDRs contents in viral proteomes.

Relationships between ID contents and viral proteomes size

We next investigated the relationships between ID contents and virus proteomes size. A significant negative correlation between ID content and virus proteome size was observed for dsDNA ($r = -0.229$, $p < 0.0001$), dsRNA ($r = -0.268$, $p = 0.0002$), ssRNA(+) ($r = -0.518$, $p < 0.0001$) and unclassified viruses ($r = -0.229$, $p < 0.001$), while a significant positive correlation was found for ssRNA-RT ($r = 0.271$, $p = 0.045$) and satellite viruses ($r = 0.465$, $p = 0.0002$) (Table S4). In order to ascertain whether the observed varying relationships between ID content and virus proteome size could be tied to differences in hosts, we performed a host-wise correlation analysis. Interestingly, a significant negative correlation between ID content and virus proteome size was observed for dsDNA viruses infecting Animalia ($r = -0.521$, $p < 0.0001$) and Plantae ($r = -0.452$, $p = 0.023$), but a significant positive correlation was observed for Archaea ($r = 0.370$, $p < 0.0014$) and Eubacteria ($r = 0.070$, $p < 0.0022$) (Figure 3A).

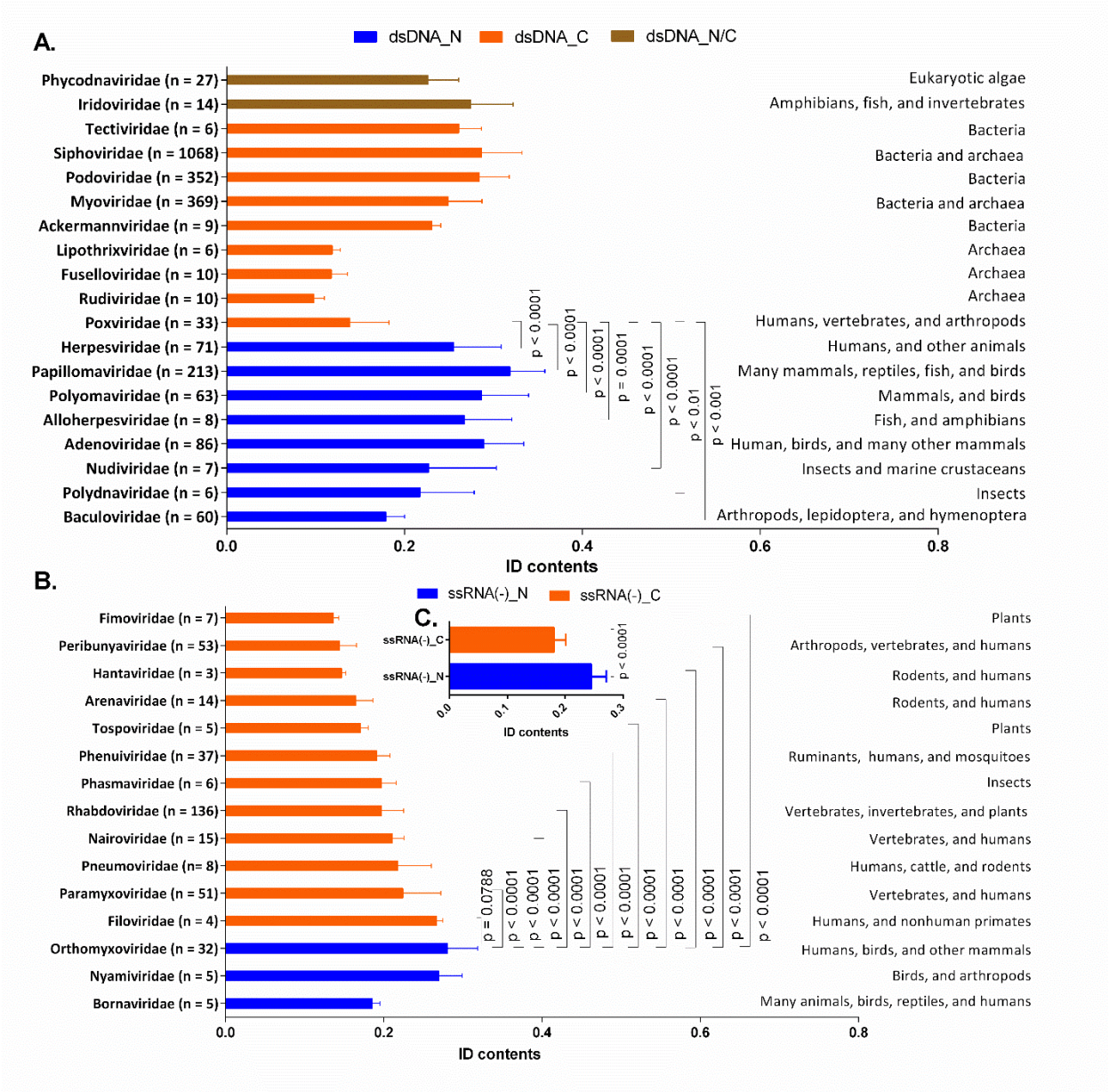

Figure 2. Mean IDRs contents within the members of (A) dsDNA and (B) ssRNA(-) virus families, whose replication primarily occurs in the nucleus and cytoplasm, respectively. (C) shows the differences between the overall mean ID content of ssRNA(-) viruses that replicate in the nucleus with that of the cytoplasm by using an unpaired t-test with Welch's correction. The error bars show the standard deviations and a p-value less than 0.01 was considered statistically significant.

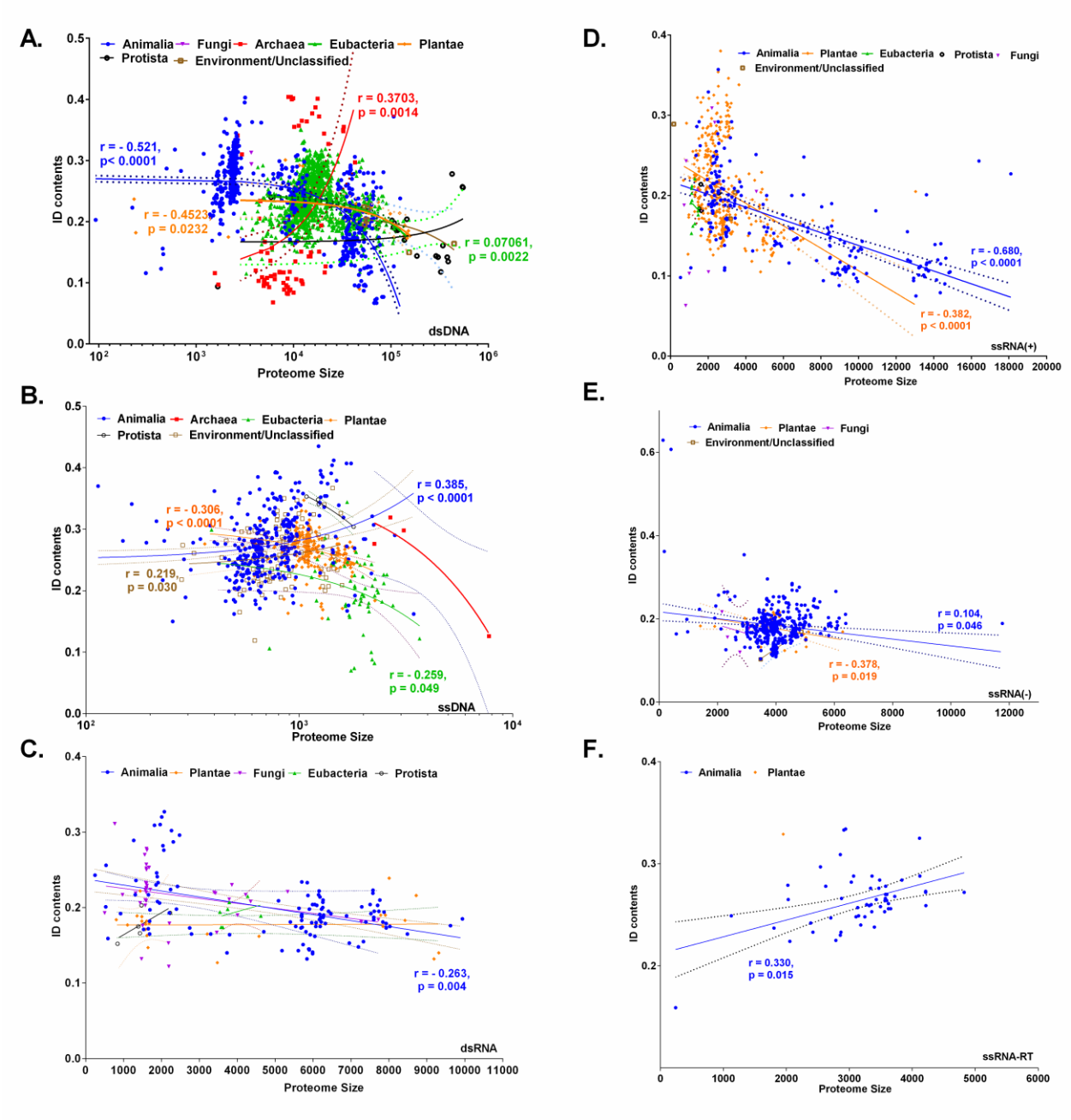

Figure 3. ID contents diversity in diverse virus genome types as a function of their proteome size. (A) to (F) show a detailed correlation analysis between ID contents and proteome size of viruses infecting taxonomically divergent hosts.

It may be noted that the different viral genome types infecting Animalia exhibit both a positive (ssDNA, ssRNA(-), and ssRNA-RT) and a negative (dsDNA, dsRNA, and ssRNA(+)) correlation between the ID content and virus proteome size (Figure 3A-F and Table S4). Importantly, double-stranded viral genomes infecting Animalia have a tendency to show a negative correlation while single-stranded viral genome infecting Animalia show a positive correlation tendency between ID content and proteome sizes, except for ssRNA(+). Conversely, in the case of viruses infecting Plantae, a significant negative correlation between ID content and virus proteome size was maintained by most of the viruses irrespective of their genome types.

Concordance between ID content and position proteins within the CH-CDF plot

CH-CDF plot[51-53] has been recently employed for the sub-classification of vertebrate host proteins into structured, disordered, mixed, and rare or unusual proteins. To this end, viral proteins (n = 283,160) encoded by the diverse viral types were sub-classified into structured, disordered, mixed, and rare proteins by using CH-CDF plots (Figures 4A-J). In the CH-CDF plots, the viral proteins that fit in quadrants, I, II, III, and IV are classified as unusual/rare proteins, structured, mixed (having both ordered and disorder properties or native molten globules), and disordered, respectively. Furthermore, quantification of ID within the viral proteins (in terms of ID %) was correlated (based on a set of ID thresholds) to the distribution of proteins across the four quadrants of the CH-CDF plot (Figure 4K and Table S5). The CH-CDF plot classifies 8.2% (23,439 out of 283,160) of viral proteins as disordered proteins, where all of them showed more than 10% ID content. While considering viral proteins with an ID content of more than 50%, 81.8% of the resulting viral proteins (3,804 out of 4,648) occupy quadrant of highly disordered proteins indicating a sharp rise in concordance between these two methods (Table S5). The concordance between these two methods has further improved to more than

91.0% by increasing the ID content threshold to more than 60% (Figure 4K). Therefore, an ID threshold of >50% can be used as a supplement with the CH-CDF plot to improve the reliability of prediction results.

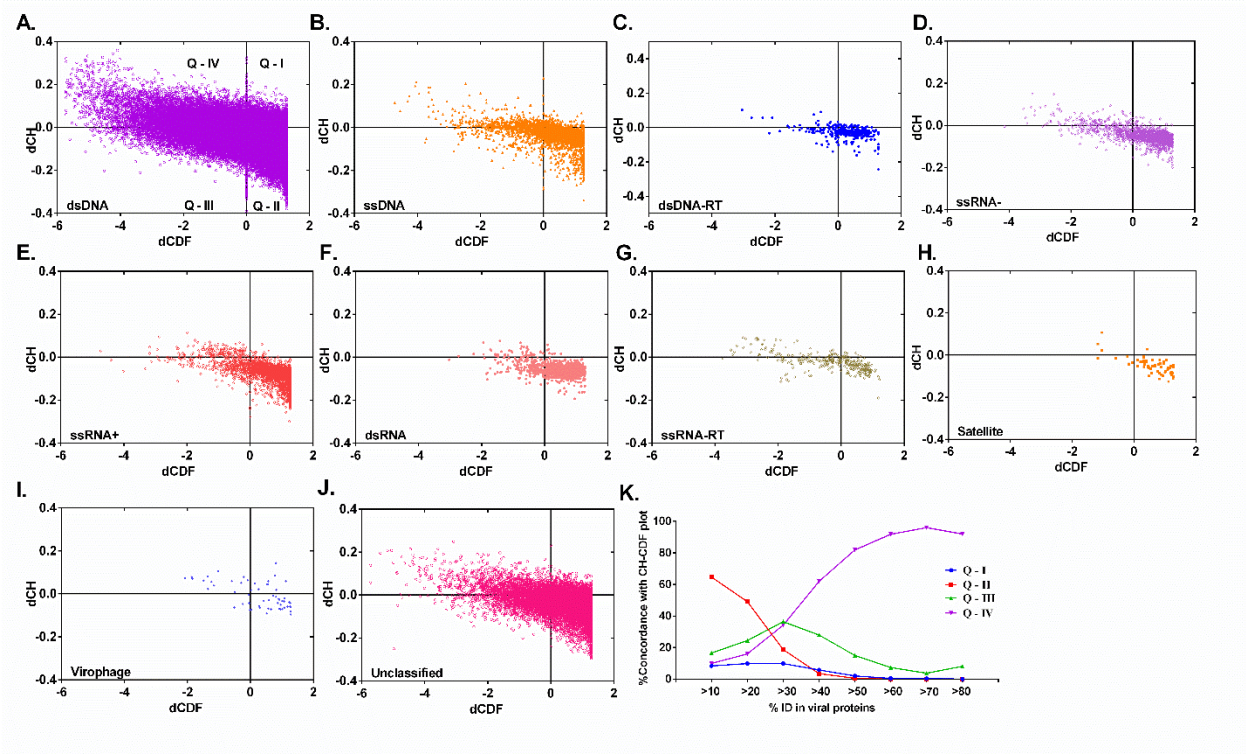

Figure 4. Charge-Hydropathy-Cumulative Distribution Function (CH-CDF) plots for the viral proteins encoded by (A) dsDNA, (B) ssDNA, (C) dsDNA-RT, (D) ssRNA(-), (E) ssRNA(+), (F) dsRNA, (G) ssRNA-RT, (H) satellite, (I) virophage, and (J) unclassified viruses. The quadrants shown in panel A are the same for the rest of the panels. (K) shows the concordance between the percentage of ID content (predicted by the IUPred2A_short method) in the viral proteins with the distribution of proteins in the four quadrants of the CH-CDF plot.

Functional annotations of viral proteins and association with their disorder contents

Given that the viral proteins sub-classified into the four quadrants of the CH-CDF plot have different characteristics, we assessed whether these viral proteins show distinctive functional

properties or specific subcellular locations. The specific GO terms (n = 29,813) of viral proteins were retrieved and sub-classified into the four quadrants of the CH-CDF plot. Since several GO terms were overlapping the four quadrants of the CH-CDF plots, we selected only those GO terms explicitly located in a particular quadrant of the CH-CDF plot (Figure 5). Of note, disordered proteins are primarily involved in key virus-host interactions, such as regulation of viral and host transcription, key steps in viral morphogenesis, and down-regulation of the host antiviral immune response. Proteins having both structured and unstructured regions (or behaving as native molten globules) affect virus replication, and modulate signaling pathways to down-regulate the host adaptive immune response, while ordered proteins are characterized by the GO terms featuring primarily metabolic and biosynthetic processes.

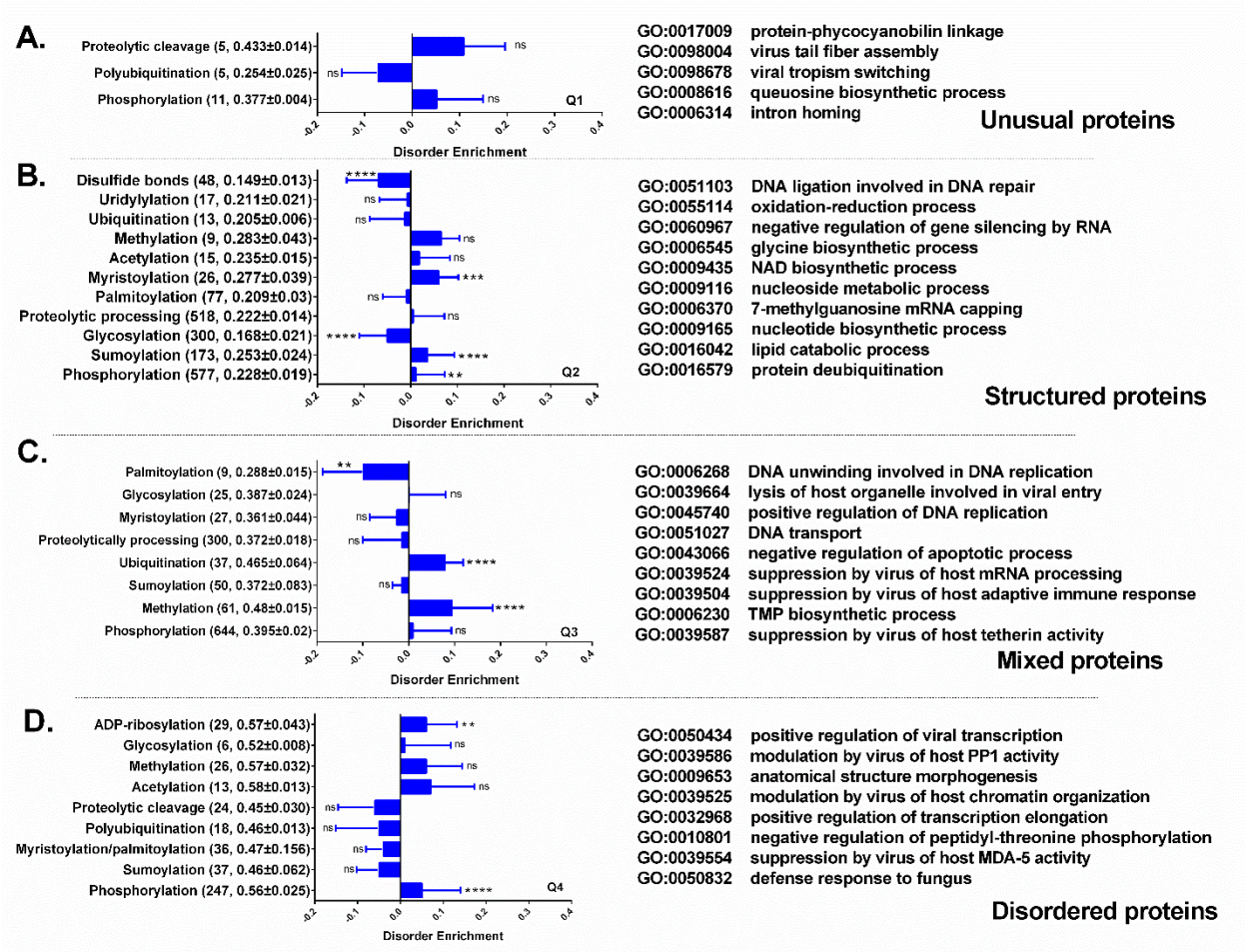

Figure 5. Functional annotations of the viral proteins sub-classified into the four quadrants of the CH-CDF plot. The gene ontology (GO) terms and post-translations modifications of unusual (A), structured (B), mixed (C), and disordered (D) viral proteins are shown. For a given PTM, the significant differences in the ID contents in the viral proteins classified by CH-CDF plot from their quadrant-wise mean ID content (baselines) are presented. The multiple t-tests with Holm-Sidak correction were used to estimate the significant differences. The error bars show the standard deviations and a p-value less than 0.01 was considered statistically significant.

Similar to the GO terms, the viral proteins with known PTMs (n = 3,397) and subcellular localization (n = 29,001) were sub-classified into the four quadrants of the CH-CDF plot followed by a measurement of correlation with the ID contents. Figure 5 shows that the PTMs found in structured viral proteins (phosphorylation, sumoylation, and myristoylation), mixed/molten globular proteins (methylation, and ubiquitination), and disordered proteins (phosphorylation, and ADP-ribosylation) were associated with disorder enriched proteins. Of note, the ordered viral proteins having glycosylation, and disulfide bonds as PTMs were depleted in disorder. In addition, the majority of subcellular locations in each of the viral protein categories were not found to be tied to disorder enrichment/depletion except for disorder-depleted structured viral proteins that were found to be located in the virion membrane and apical cell membrane.

DISCUSSION

A substantial variation in ID distribution has been observed among the main kingdoms of life. Previous studies have shown that with the increase in the complexity of organisms, there is a notable increase in the length and frequency of IDRs[7, 18, 21, 22, 49, 54]. In fact, long IDRs (> 30

consecutive residues) are more common in eukaryotic proteins (45−50%) than in prokaryotic (7−30%) proteins[7, 18, 21, 22, 49, 54]. The broad variability of ID levels in different viruses found in our study is consistent with the results of previous studies[21, 22]. We analyzed whether disorder predictions are affected by the length of IDRs. To this end, we have compared the results of disorder prediction provided by the IUPred2A long method designed to identify IDRs longer than 30 amino acids, to those provided by the IUPred2A short method (developed for the identification of IDRs shorter than 10 amino acids). The strong correlation between the short and long disorder prediction types supports the conclusion that the length of IDRs does not affect the findings. Consistency in results is maintained even when using a different IDRs predictor, namely Espritz, which uses bi-directional recursive neural networks.

The propensity of viruses to co-evolve with their hosts or on the other hand, their propensity to undergo regular host-switching (cross-species transmission) can be measured by co-phylogenetic analysis, which compares the topology of virus and hosts phylogenetic trees[55]. Such studies have shown that dsDNA viruses exhibit typical virus-host co-evolution, contrary to RNA viruses that have a significant tendency for frequent host-switching[55-57]. This result is consistent with the fact that RNA viruses frequently cause acute (but transient) infections, while DNA viruses are mostly associated with chronic infections. These studies, therefore, support the notion that the higher levels of IDRs in proteins of DNA viruses relative to the RNA viruses could be the result of long-term co-evolution that would have endowed viral proteins with regulatory functions. Another plausible explanation comes from the observation that disorder-enriched regions are commonly found in those proteins involved in multiple protein-protein interactions and such manifold protein interactions were shown to be negatively correlated with their rate of evolution[58, 59]. Since the rate of evolution of DNA viruses is comparatively lower than that of

RNA viruses, DNA virus-encoded proteins appear to be more enriched in IDRs. Also, our study has shown that the proteomes of the viruses (either DNA or RNA) that prefer the nucleus as a site of replication are enriched in the IDRs. This may be consistent with previous research, which showed that in eukaryotes, DNA-binding proteins are significantly enriched in ID[60]. The high disorder levels of proteins in these viruses can lead to more efficient hijacking of numerous and complex cell processes/pathways at the advantage of the virus, which is confirmed by the GO terms analysis of highly disordered proteins.

Analyses of the correlation of the length of the viral proteomes concerning their ID contents showed that small viral proteomes appear to have a high content of disorder, and vice-versa, with a few exceptions. Since the viruses of a particular genome type infect taxonomically divergent hosts, these findings may have confounding effects. Therefore, we have expanded our analyses to the host-level associations. In particular, different viruses infecting Animalia showed both negative and positive correlations between ID content and proteome size; these differential findings may be attributed to the distinct virus-type specific proteome sizes; i.e., viruses having smaller proteome sizes (ssDNA, ssRNA(-), and ssRNA-RT) show a positive correlation, while viruses with larger proteome sizes (dsDNA, and dsRNA) exhibit negative correlation. Our findings are consistent with the previous research, which also showed a negative correlation between large proteome sizes of eukaryotes and bacteria and their disorder contents[21]. Furthermore, intrinsically disordered protein regions evolve more rapidly than that of ordered regions in both the eukaryotes[61, 62], and viruses[63]. This high rate of evolution of sequences in disorder regions not only modifies eukaryotic protein functions[62, 64], but also generates diversity in some viruses that enable them to interact with multiple proteins of different species[65].

Disorder regions, therefore, impart functional versatility as well as sequence malleability within the same protein.

Functional annotation shows that disordered proteins are primarily involved in key virus-host interactions, such as regulation of viral and host transcription, key steps in viral morphogenesis, and down-regulation of the host antiviral immune response. Besides, disorder-enriched viral proteins are preferentially associated with certain PTMs (phosphorylation, sumoylation, myristoylation, methylation, ubiquitination, and ADP-ribosylation), and these PTMs allow them to perform several regulatory functions and virus-host interactions. For instance, phosphorylation of viral proteins is known to affect various cellular signaling pathways, and viral proteins can be phosphorylated by the multiple cellular kinases, thereby providing an opportunity for a virus to expand its host and cellular tropism with varying kinases profiles[66]. On the other hand, sumoylation and ubiquitination are needed for the modulation of anti-viral defenses and viral replication[67, 68]; myristoylation is related to virus entry, assembly, structure, and budding[69]; methylation for modulation of protein-nucleic acid interaction[70]; and ADP-ribosylation for modulation of viral infection through unknown mechanisms[71].

Taking together the fact that a high ID content in the proteome of tardigrades (water bears) plays a critical role in survival against extreme desiccation[72], and that the ID content was found to be related to adaptability to complex and diverse environments[53, 73], we explored the relationship between the viral proteomes with a high ID content (>0.4), namely ssDNA (Torque teno virus), dsDNA (Papillomavirus and the *Haloarcula hispanica* SH1 virus), ssRNA(-) (Deltavirus) and satellite viruses (plant), with the environment of their respective hosts. Interestingly, the archaeal haloviruses that infect the halophilic archaeon *Haloarcula hispanica,* do not lose their infectivity even in the presence of high salt concentrations[74], making them

capable of surviving under the harsh environmental conditions of their hosts. Likewise, the Torque teno virus, a non-enveloped human DNA virus, is ubiquitously present in the environment due to its high stability[75, 76]. We noted that a high ID content in viruses tends to (i) support long-term survival of the virus under extreme environmental conditions, (ii) provide to their proteins multi-functionality that compensates their smaller proteome size (satellite viruses), and (iii) support their long-term co-evolution, especially in the case of DNA viruses.

AUTHOR INFORMATION

Corresponding Author

Naveen Kumar − *Diagnostics & Vaccines Group, ICAR- National Institute of High Security Animal Diseases, Bhopal 462022, India*; orcid.org/0000-0002-3326-5465; Email: naveen.kumar4@icar.gov.in; Phone: +91-7552759204; Fax: +91-755-2758842

CONFLICT OF INTEREST

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

All the authors acknowledge and thank their respective Institutes. The authors thank Dr. Christopher J. Oldfield, Virginia Commonwealth University, Richmond, USA for providing technical support in the generation of the CH-CDF plots.

ASSOCIATED CONTENT

Supporting Information

Table S1. Categorization of reference viral proteomes corresponding to their viral nucleic acid types and hosts; Table S2: Fractions of disordered, moderately disordered, and ordered viral proteins across the diverse reference virus proteomes; Table S3. Mean ID contents across the virus genome types infecting taxonomically divergent hosts; Table S4. Correlation analysis between ID contents and viral proteomes size; Table S5. Quantification of ID contents in the viral proteins (as the fraction of disordered residues over the total residues) and distribution of the viral proteins (n = 283,160) in the four quadrants of the CH-CDF Plot.

REFERENCES

1. Anfinsen, C. B.; Scheraga, H. A., Experimental and theoretical aspects of protein folding. *Adv. Protein Chem.* 1975, *29*, 205-300.

2. Redfern, O. C.; Dessailly, B.; Orengo, C. A., Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* 2008, *18* (3), 394-402.

3. Habchi, J.; Tompa, P.; Longhi, S.; Uversky, V. N., Introducing protein intrinsic disorder. *Chem. Rev.* 2014, *114* (13), 6561-88.

4. Dunker, A. K.; Brown, C. J.; Obradovic, Z., Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* 2002, *62*, 25-49.

5. Dunker, A. K.; Cortese, M. S.; Romero, P.; Iakoucheva, L. M.; Uversky, V. N., Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs j* 2005, *272* (20), 5129-48.

6. Dunker, A. K.; Lawson, J. D.; Brown, C. J.; Williams, R. M.; Romero, P.; Oh, J. S.; Oldfield, C. J.; Campen, A. M.; Ratliff, C. M.; Hipps, K. W.; Ausio, J.; Nissen, M. S.;

Reeves, R.; Kang, C.; Kissinger, C. R.; Bailey, R. W.; Griswold, M. D.; Chiu, W.; Garner, E. C.; Obradovic, Z., Intrinsically disordered protein. *J. Mol. Graph. Model.* 2001, *19* (1), 26-59.

7. Dunker, A. K.; Obradovic, Z., The protein trinity--linking function and disorder. *Nat. Biotechnol.* 2001, *19* (9), 805-6.

8. Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L., Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* 2008, *18* (6), 756-64.

9. Dyson, H. J., Making Sense of Intrinsically Disordered Proteins. *Biophys. J.* 2016, *110* (5), 1013-6.

10. Iakoucheva, L. M.; Brown, C. J.; Lawson, J. D.; Obradović, Z.; Dunker, A. K., Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 2002, *323* (3), 573-84.

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

12. Uversky, V. N., Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* 2002, *11* (4), 739-56.

13. Uversky, V. N.; Dunker, A. K., Understanding protein non-folding. *Biochim. Biophys. Acta* 2010, *1804* (6), 1231-64.

14. Uversky, V. N.; Gillespie, J. R.; Fink, A. L., Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins* 2000, *41* (3), 415-27.

15. Uversky, V. N.; Oldfield, C. J.; Dunker, A. K., Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* 2005, *18* (5), 343-84.

16. Uversky, V. N.; Oldfield, C. J.; Dunker, A. K., Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* 2008, *37*, 215-46.

17. Wright, P. E.; Dyson, H. J., Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* 2015, *16* (1), 18-29.

18. Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T., Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 2004, *337* (3), 635-45.

19. Burra, P. V.; Kalmar, L.; Tompa, P., Reduction in structural disorder and functional complexity in the thermal adaptation of prokaryotes. *PLoS One* 2010, *5* (8), e12069.

20. Xue, B.; Williams, R. W.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N., Archaic chaos: intrinsically disordered proteins in Archaea. *BMC Syst. Biol.* 2010, *4 Suppl 1* (Suppl 1), S1.

21. Peng, Z.; Yan, J.; Fan, X.; Mizianty, M. J.; Xue, B.; Wang, K.; Hu, G.; Uversky, V. N.; Kurgan, L., Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell. Mol. Life Sci.* 2015, *72* (1), 137-51.

22. Xue, B.; Dunker, A. K.; Uversky, V. N., Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* 2012, *30* (2), 137-49.

23. Uversky, V. N.; Longhi, S., *Flexible Viruses: Structural Disorder in Viral Proteins*. 1st edition ed.; John Wiley and Sons, Inc., United States: 2011.

24. Xue, B.; Mizianty, M. J.; Kurgan, L.; Uversky, V. N., Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. Cell Mol. Life Sci. 2012, 69 (8), 1211-59.

25. Fan, X.; Xue, B.; Dolan, P. T.; LaCount, D. J.; Kurgan, L.; Uversky, V. N., The intrinsic disorder status of the human hepatitis C virus proteome. Mol. Biosyst. 2014, 10 (6), 1345-63.

26. Xue, B.; Ganti, K.; Rabionet, A.; Banks, L.; Uversky, V. N., Disordered interactome of human papillomavirus. Curr. Pharm. Des. 2014, 20 (8), 1274-92.

27. Dolan, P. T.; Roth, A. P.; Xue, B.; Sun, R.; Dunker, A. K.; Uversky, V. N.; LaCount, D. J., Intrinsic disorder mediates hepatitis C virus core-host cell protein interactions. Protein Sci. 2015, 24 (2), 221-35.

28. Meng, F.; Badierah, R. A.; Almehdar, H. A.; Redwan, E. M.; Kurgan, L.; Uversky, V. N., Unstructural biology of the Dengue virus proteins. FEBS J. 2015, 282 (17), 3368-94.

29. Giri, R.; Kumar, D.; Sharma, N.; Uversky, V. N., Intrinsically Disordered Side of the Zika Virus Proteome. Front. Cell Infect. Microbiol. 2016, 6, 144.

30. Charon, J.; Barra, A.; Walter, J.; Millot, P.; Hébrard, E.; Moury, B.; Michon, T., First Experimental Assessment of Protein Intrinsic Disorder Involvement in an RNA Virus Natural Adaptive Process. *Mol. Biol. Evol.* 2018, *35* (1), 38-49.

31. Dyson, H. J.; Wright, P. E., How Do Intrinsically Disordered Viral Proteins Hijack the Cell? *Biochemistry* 2018, *57* (28), 4045-4046.

32. Tamarozzi, E. R.; Giuliatti, S., Understanding the Role of Intrinsic Disorder of Viral Proteins in the Oncogenicity of Different Types of HPV. *Int. J. Mol. Sci.* 2018, *19* (1), 198.

33. Uversky, V. N.; Roman, A.; Oldfield, C. J.; Dunker, A. K., Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs. *J. Proteome Res.* 2006, *5* (8), 1829-42.

34. Rieder, C. A.; Rieder, J.; Sannajust, S.; Goode, D.; Geguchadze, R.; Relich, R. F.; Molliver, D. C.; King, T. E.; Vaughn, J.; May, M., A Novel Mechanism for Zika Virus Host-Cell Binding. *Viruses* 2019, *11* (12), 1101.

35. Bourhis, J. M.; Canard, B.; Longhi, S., Structural disorder within the replicative complex of measles virus: functional implications. *Virology* 2006, *344* (1), 94-110.

36. Habchi, J.; Longhi, S., Structural disorder within paramyxovirus nucleoproteins and phosphoproteins. *Mol. Biosyst.* 2012, *8* (1), 69-81.

37. Nagano, Y.; Sugiyama, A.; Kimoto, M.; Wakahara, T.; Noguchi, Y.; Jiang, X.; Saijo, S.; Shimizu, N.; Yabuno, N.; Yao, M.; Gooley, P. R.; Moseley, G. W.; Tadokoro, T.; Maenaka, K.; Ose, T., The Measles Virus V Protein Binding Site to STAT2 Overlaps That of IRF9. *J. Virol.* 2020, *94* (17), e01169-20.

38. Prates, E. T.; Garvin, M. R.; Pavicic, M.; Jones, P.; Shah, M.; Demerdash, O.; Amos, B. K.; Geiger, A.; Jacobson, D., Potential pathogenicity determinants identified from structural proteomics of SARS-CoV and SARS-CoV-2. *Mol. Biol. Evol.* 2021, *38* (2), 702-15.

39. Song, Y.; Pei, Y.; Yang, Y. L.; Xue, J.; Zhang, G. Z., The Ntail region of nucleocapsid protein is associated with the pathogenicity of pigeon paramyxovirus type 1 in chickens. *J. Gen. Virol.* 2019, *100* (6), 950-957.

40. Sungsuwan, S.; Jongkaewwattana, A.; Jaru-Ampornpan, P., Nucleocapsid proteins from other swine enteric coronaviruses differentially modulate PEDV replication. *Virology* 2020, *540*, 45-56.

41. Tsimbalyuk, S.; Cross, E. M.; Hoad, M.; Donnelly, C. M.; Roby, J. A.; Forwood, J. K., The Intrinsically Disordered W Protein Is Multifunctional during Henipavirus Infection, Disrupting Host Signalling Pathways and Nuclear Import. *Cells* 2020, *9* (8), 1913.

42. Brocca, S.; Grandori, R.; Longhi, S.; Uversky, V., Liquid-Liquid Phase Separation by Intrinsically Disordered Protein Regions of Viruses: Roles in Viral Life Cycle and Control of Virus-Host Interactions. *Int. J. Mol. Sci.* 2020, *21* (23), 9045.

43. Xue, B.; Blocquel, D.; Habchi, J.; Uversky, A. V.; Kurgan, L.; Uversky, V. N.; Longhi, S., Structural disorder in viral proteins. *Chem. Rev.* 2014, *114* (13), 6880-911.

44. Xue, B.; Williams, R. W.; Oldfield, C. J.; Goh, G. K.; Dunker, A. K.; Uversky, V. N., Viral disorder or disordered viruses: do viral proteins possess unique features? *Protein Pept Lett* 2010, *17* (8), 932-51.

45. Mészáros, B.; Erdos, G.; Dosztányi, Z., IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 2018, *46* (W1), W329-w337.

46. Walsh, I.; Martin, A. J.; Di Domenico, T.; Tosatto, S. C., ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics* 2012, *28* (4), 503-9.

47. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 2019, *47* (D1), D506-d515.

48. Mihara, T.; Nishimura, Y.; Shimizu, Y.; Nishiyama, H.; Yoshikawa, G.; Uehara, H.; Hingamp, P.; Goto, S.; Ogata, H., Linking Virus Genomes with Host Taxonomy. *Viruses* 2016, *8* (3), 66.

49. Oldfield, C. J.; Cheng, Y.; Cortese, M. S.; Brown, C. J.; Uversky, V. N.; Dunker, A. K., Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 2005, *44* (6), 1989-2000.

50. Xue, B.; Oldfield, C. J.; Dunker, A. K.; Uversky, V. N., CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions. *FEBS Lett.* 2009, *583* (9), 1469-74.

51. Huang, F.; Oldfield, C.; Meng, J.; Hsu, W. L.; Xue, B.; Uversky, V. N.; Romero, P.; Dunker, A. K., Subclassifying disordered proteins by the CH-CDF plot method. *Pac. Symp. Biocomput.* 2012, 128-39.

52. Huang, F.; Oldfield, C. J.; Xue, B.; Hsu, W. L.; Meng, J.; Liu, X.; Shen, L.; Romero, P.; Uversky, V. N.; Dunker, A., Improving protein order-disorder classification using charge-hydropathy plots. *BMC Bioinformatics* 2014, *15 Suppl 17* (Suppl 17), S4.

53. Mohan, A.; Sullivan, W. J., Jr.; Radivojac, P.; Dunker, A. K.; Uversky, V. N., Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol. Biosyst.* 2008, *4* (4), 328-40.

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

55. Geoghegan, J. L.; Duchêne, S.; Holmes, E. C., Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathog.* 2017, *13* (2), e1006215.

56. Kumar, N.; Bhatia, S.; Pateriya, A. K.; Sood, R.; Nagarajan, S.; Murugkar, H. V.; Kumar, S.; Singh, P.; Singh, V. P., Label-free peptide nucleic acid biosensor for visual detection of multiple strains of influenza A virus suitable for field applications. *Anal. Chim. Acta* 2020, *1093*, 123-130.

57. Kumar, N.; Malik, Y. S.; Sharma, K.; Dhama, K.; Ghosh, S.; Bányai, K.; Kobayashi, N.; Singh, R. K., Molecular characterization of unusual bovine rotavirus A strains having high genetic relatedness with human rotavirus: evidence for zooanthroponotic transmission. *Zoonoses Public Health* 2018, *65* (4), 431-442.

58. Bertolazzi, P.; Bock, M. E.; Guerra, C., On the functional and structural characterization of hubs in protein–protein interaction networks. *Biotechnol. Adv.* 2013, *31* (2), 274-286.

59. Fraser, H. B.; Hirsh, A. E.; Steinmetz, L. M.; Scharfe, C.; Feldman, M. W., Evolutionary rate in the protein interaction network. *Science* 2002, *296* (5568), 750-2.

60. Wang, C.; Uversky, V. N.; Kurgan, L., Disordered nucleiome: Abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics* 2016, *16* (10), 1486-98.

61. Brown, C. J.; Johnson, A. K.; Dunker, A. K.; Daughdrill, G. W., Evolution and disorder. Curr. Opin. Struct. Biol. 2011, 21 (3), 441-6.

62. Kastano, K.; Erdős, G.; Mier, P.; Alanis-Lobato, G.; Promponas, V. J.; Dosztányi, Z.; Andrade-Navarro, M. A., Evolutionary Study of Disorder in Protein Sequences. Biomolecules. 2020, 10 (10), 1413.

63. Gitlin, L.; Hagai, T.; LaBarbera, A.; Solovey, M.; Andino, R., Rapid evolution of virus sequences in intrinsically disordered protein regions. PLoS Pathog. 2014, 10 (12), e1004529.

64. Ahrens, J. B.; Nunez-Castilla, J.; Siltberg-Liberles, J., Evolution of intrinsic disorder in eukaryotic proteins. Cell. Mol. Life Sci. 2017, 74 (17), 3163-74.

65. Ortiz, J. F.; MacDonald, M. L.; Masterson, P.; Uversky, V. N.; Siltberg-Liberles, J., Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome Biol. Evol. 2013, 5 (3), 504-13.

66. Keating, J. A.; Striker, R., Phosphorylation events during viral infections provide potential therapeutic targets. *Rev. Med. Virol.* 2012, *22* (3), 166-81.

67. Lowrey, A. J.; Cramblet, W.; Bentz, G. L., Viral manipulation of the cellular sumoylation machinery. *Cell Commun Signal* 2017, *15* (1), 27.

68. Taylor, R. T.; Best, S. M., Assessing ubiquitination of viral proteins: Lessons from flavivirus NS5. *Methods* 2011, *55* (2), 166-71.

69. Maurer-Stroh, S.; Eisenhaber, F., Myristoylation of viral and bacterial proteins. *Trends Microbiol.* 2004, *12* (4), 178-85.

70. Hundt, J.; Li, Z.; Liu, Q., Post-translational modifications of hepatitis C viral proteins and their biological significance. *World J. Gastroenterol.* 2013, *19* (47), 8929-39.

71. Grunewald, M. E.; Fehr, A. R.; Athmer, J.; Perlman, S., The coronavirus nucleocapsid protein is ADP-ribosylated. *Virology* 2018, *517*, 62-68.

72. Boothby, T. C.; Tapia, H.; Brozena, A. H.; Piszkiewicz, S.; Smith, A. E.; Giovannini, I.; Rebecchi, L.; Pielak, G. J.; Koshland, D.; Goldstein, B., Tardigrades Use Intrinsically Disordered Proteins to Survive Desiccation. *Mol. Cell* 2017, *65* (6), 975-984.e5.

73. Pancsa, R.; Tompa, P., Structural disorder in eukaryotes. *PLoS One* 2012, *7* (4), e34687.

74. Demina, T. A.; Atanasova, N. S.; Pietilä, M. K.; Oksanen, H. M.; Bamford, D. H., Vesicle-like virion of Haloarcula hispanica pleomorphic virus 3 preserves high infectivity in saturated salt. *Virology* 2016, *499*, 40-51.

75. Brajão de Oliveira, K., Torque teno virus: a ubiquitous virus. *Rev Bras Hematol Hemoter* 2015, *37* (6), 357-8.

76. Griffin, J. S.; Plummer, J. D.; Long, S. C., Torque teno virus: an improved indicator for viral pathogens in drinking waters. *Virol J.* 2008, *5*, 112.