Insights into the evolutionary forces that shape the codon usage in the viral genome segments encoding intrinsically disordered protein regions

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

Insights into the evolutionary forces that shape the codon usage in the viral genome segments encoding intrinsically disordered protein regions

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

[#]Correspondence:

E-mail address: naveen.kumar4@icar.gov.in (Naveen Kumar); Telephone: +917552759204; Fax: +917552758842

Abstract

Intrinsically disordered regions/proteins (IDRs/IDPs) are abundant across all the domains of life, where they perform important regulatory roles and supplement the biological functions of structured proteins/regions (SRs). Despite the multi-functionality features of IDRs, several interrogations on the evolution of viral genomic regions encoding IDRs in diverse viral proteins remain unreciprocated. To fill this gap, we benchmarked the findings of two most widely used and reliable intrinsic disorder prediction algorithms (IUPred2A, and ESpritz) to a dataset of 6,108 reference viral proteomes to unravel the multi-faceted evolutionary forces that shape the codon usage in the viral genomic regions encoding for IDRs and SRs. We found persuasive evidence that the natural selection predominantly governs the evolution of codon usage in regions encoding IDRs by most of the viruses. In addition, we confirm not only that codon usage in regions encoding IDRs is less optimized for the protein synthesis machinery (tRNAs pool) of their host than for those encoding SRs, but also that the selective constraints imposed by codon bias sustain this reduced optimization in IDRs. Our analysis also establishes that IDRs in viruses are likely to tolerate more translational errors than SRs. All these findings hold true irrespective of the disorder prediction algorithms used to classify IDRs. In conclusion, our study offers a novel perspective on the evolution of viral IDRs and the evolutionary adaptability to multiple taxonomically divergent hosts.

Keywords: Viral proteome, intrinsically disordered regions; disorder prediction algorithms; evolutionary forces, CpG contents, translation adaptation.

Introduction

By virtue of genetic code redundancy, more than one codon (synonymous codons) codes for the same amino acid. Increasing pieces of evidence suggest that in any given organism, synonymous codons are not used arbitrarily, a phenomenon called codon usage bias, and results in species-specific codon usage bias [1-7]. The origin of the codon usage bias is mostly explained with the help of widely accepted 'selection-mutation-drift theory', which posits that the mutational bias and the natural selection are the two leading evolutionary forces that shape the codon usage bias in a species [8].

Mutational biases are likely to accrue certain types of mutations unevenly, resulting in interspecies differences in the complete genome. Such mutation biases may arise from errors during DNA replication [9, 10], transcription-mediated mutational biases [9, 11, 12], methylation of CpG dinucleotide to form 5-methylcytosine followed by deamination resulting in C-T substitution [13], and uneven DNA repair [14]. On the contrary, natural selection can influence the synonymous codon usage patterns by selecting specific codon subsets to match the most abundant host tRNAs (or translational selection) [15-17]. This phenomenon is predominantly observed in highly expressed genes. Besides, other factors that may influence the codon usage bias include regulatory structural RNA elements, secondary RNA structure, and viral RNA packaging [18-20].

Defying the classical structure-function paradigm, intrinsically disordered regions (IDRs, i.e. regions that fail to acquire a defined secondary or tertiary structure under physiological conditions) perform important biological functions such as signaling, recognition and regulation [21-25]. IDRs are abundant in nature, and their prevalence among the three kingdoms of life (i.e., bacteria, archaea, and eukaryotes) differs significantly, with IDRs being enriched in eukaryotes

and in complex life forms [21, 26-30]. Viral proteins have been reported to possess distinct structural features such as a high occurrence of IDRs and lower van der Waals contact densities [31, 32]. Furthermore, increasing evidences suggest that IDRs in viruses also play important roles in both virus replication and adaptation to the host [22, 25, 30, 33-39]. These unique features of viral proteins might provide increased structural malleability needed for interaction with various components of the host immune system and quickly adapt to the host environment [33, 40-42]. Since a defined structure is not a prerequisite for IDRs function, IDRs are more tolerant of mutations. As such, IDRs might provide a unique strategy for tolerating the typically high mutation rates observed in viruses, and especially in RNA viruses [25, 42].

While the functional importance of IDRs in viruses has already been established, to date, there is a limited understanding of the evolutionary forces shaping viral IDRs. Herein, we benchmarked the findings of two most widely used and accurate intrinsic disorder predictors (IUPred2A and ESpritz, both of which utilizes distinct algorithms for the prediction of IDRs, viz., IUPred2A estimates total pairwise interaction energy from the amino acid compositions, while the latter employs bi-directional recursive neural networks for the predictions for IDRs) to a dataset comprising 6,108 reference viral proteomes encompassing 283,000 viral proteins. In addition, a systematic analysis of the evolutionary forces (natural selection and mutational bias) that shape the codon usage bias in virus genomic regions encoding IDRs and structural regions (SRs) was performed using selective bioinformatics tools to measure the effective number of codons (ENc) plots [43], neutrality plots [44], and tRNA adaptation index (tAI) [15]. Our results suggest that the codon usage in regions encoding IDRs are strongly influenced by the natural selection in most of the viruses. Moreover, IDRs contribute significantly to viral protein functionality and evolutionary adaptability to multiple taxonomically divergent hosts.

Materials and Methods

Sequence Dataset

The reference viral proteomes (n = 6,108) used in this study were retrieved from the UniProt database Release 2018_1 [45]. The viral proteomes considered in the present study account for more than 283,000 viral proteins. The categorization of the viral genomes corresponding to the viruses in reference proteomes was performed on the basis of the Baltimore Classification that relies on the nature of the viral nucleic acid to group them. The categorization resulted in splitting the reference viruses into 10 groups, viz. ssDNA (n = 741), dsDNA (n = 2,596), dsDNA-RT (n = 76), ssRNA(-) (n = 413), ssRNA(+) (n = 597), dsRNA (n = 189), ssRNA-RT (n = 55), virophage (n = 7), satellite (n = 59), and unclassified (n = 1,375) viruses. The taxonomy of all the viruses and their associated hosts were retrieved from the UniProt database and further cross-checked from virus-host DB [46]. The hosts were classified based on the RH Whittaker five kingdom classification system [47] as shown in Table S1.

Identification of Intrinsically Disordered Regions (IDRs)

Two different disorder predictors (IUPred2A, and ESpritz) were employed for the prediction of IDRs across the diverse virus proteomes. Since these two prediction algorithms are using principally different attributes and approaches for disorder prediction, this allowed us to capture flavors of disorder in our dataset. Not only do these two disorder predictors depend on very distinct ID prediction algorithms, they have also been shown to provide robust predictions with a favorable trade-off between speed and accuracy [48-51], being also able to outperform several other disorder predictors [52-54]. Therefore, we utilized these two prediction algorithms to provide information about disorder in our dataset.

IUPred2A is one of the commonly used methods for predicting protein disorder and it is based on capturing the basic biophysical properties of IDRs. This predictor is based on the assumption that IDRs have a specific amino acid composition that does not allow the formation of enough favorable inter-residue interactions to stabilize a well-defined structural state, with said interaction capacity of each residue being captured by an energy estimation scheme [55, 56]. In addition, IUPred2A offers two prediction types, long and short regions of disorder, with the former being an acclaimed option for predicting biologically relevant disordered regions, and the latter being recommended for short proteins, such as those of viruses. The long and short disorder prediction types allow predicting IDRs of at least 30 and 10 consecutive residues, respectively. These two prediction types additionally contribute to the flavors of disorders in our dataset. IUPred2A provides a *per* residue ID probability score for the protein sequence that ranges from 0 to 1. Residues having an ID probability score ≥ 0.5 are defined as disordered, and the content in IDRs for each protein is calculated as the ratio between the number of predicted disordered residues and the total number of residues in the protein. A similar approach is used for the calculation of the content of IDRs in a viral proteome, calculated as the ratio between the total number of residues predicted to be disordered in a given proteome and the total number of residues in that proteome.

The second disorder predictor, ESpritz is an ensemble of protein disorder predictors based on bidirectional recursive neural networks and trained on three different flavors of disorder (DisProt disorder, X-ray disorder, and NMR mobility) [57]. Similar to IUPred2A, ESpritz can produce fast and accurate sequence-only disorder predictions and therefore is suitable for disorder annotation of large datasets. A short-disorder prediction (trained on the missing atoms from the Protein Data Bank X-ray crystallography structures) with a 5% false positive rate implemented

in ESpritz, was run on the dataset of the reference viral proteomes. Like IUPred2A, ESpritz provides a *per* residue ID probability score and an approach similar to that described above was used for the calculation of the content of IDRs in a viral proteome.

In addition, we assessed and validated the performance of two distinct disorder predictors (IUPred2A, and ESpritz) used in our study against the experimental disorder content information of viral proteins. To achieve this, we retrieved viral proteins (n = 79) whose disordered region annotations have been achieved experimentally from the Disprot database Version: 8.1 (https://www.disprot.org/) and estimated the sensitivity and specificity of disorder content prediction. The summary of viral proteins dataset retrieved from the Disprot database is provided in Table S2A.

Assessing the factors driving the evolution of codon usage in IDRs and SRs of viral proteins

The nucleotide sequences encoding viral IDRs and SRs, as predicted by IUPred2A, were extracted separately from the reference virus genomes derived from the NCBI GenBank Release 230.0 (n = 6,108). The virus nucleotide sequences (n = 646) containing either internal stop codons or non-translatable codons or both, were discarded and the remaining sequences from 5,462 reference viruses were used for further analysis.

ENc-GC3s plots

The effective number of codons (ENc) represents the magnitude of codon usage bias within a gene. The ENc values range from 20 to 61, where the smaller the ENc value the greater the extent of codon usage bias and *vice-versa*. The plotting of ENc values against the GC3s (frequency of either a guanine or cytosine at the third codon position of the synonymous codons,

excluding Met, Trp, and stop codons) provides a qualitative estimation of the driving factors (mutation bias and natural selection) that shape the codon usage patterns [43]. In the codon usage table, there are two amino acids (Met, and Trp) with only one codon choice (AUG, and UGG, respectively), nine amino acids with two codon choices, one with three, five with four, and three with six codon choices that make up five distinct synonymous families (1, 2, 3, 4, and 6). The overall contributions made by each synonymous family to codon usage bias thus make up the ENc. The ENc values were calculated for IDRs and SRs coding sequences using the formula described in equation (1).

$$ENc = 2 + \frac{9}{\overline{F_2}} + \frac{1}{\overline{F_3}} + \frac{5}{\overline{F_4}} + \frac{3}{\overline{F_6}} \quad (1)$$

where $F_{(i = 2,3,4,6)}$ is the mean of F_i (homozygosity frequency) values for the i-fold (synonymous family type) degenerate amino acid. The F_i values were calculated using equation (2).

$$\overline{F_i} = \frac{n \sum_{j\text{-}1}^{i} (\frac{n_j}{n})^2 \text{ - } 1}{n \text{ - } 1} \quad (2)$$

where n represents the total number of occurrences of the codons for that amino acid and n_j is the total number of occurrences of the j^{th} codon for that amino acid.

In the ENc-GC3s plot, the ENc values occupy the ordinate, while the GC3s values (frequency of either a guanine or cytosine at the third codon position of the synonymous codons, excluding Met, Trp, and stop codons) occupy the abscissa [43]. In those cases where the calculated ENc values cluster on or just below the standard/expected curve (functional relation between expected ENc and GC3s), the codon usage is constrained only by G+C mutational bias. By contrast, clustering of the calculated ENc values far below the standard curve indicates a predominant role

of natural selection in shaping the codon usage bias. Expected ENc values were calculated by using equation (3).

$$ENc_{expected} = 2 + s + \frac{29}{s^2 + (1 - s^2)} \quad (3)$$

where, 's' is the frequency of G + C at the third codon position of synonymous codons (i.e. GC3s).

Neutrality plots

The role of the mutational bias in shaping the evolution of synonymous codon usage has been shown to be related to a higher or lower GC content of the genomes. GC content changes have been observed more frequently in GC3 (nucleotides G + C at the third codon position), one of the most neutral nucleotides of the genome [44]. Therefore, the quantitative contributions of the mutational bias and the natural selection that influenced the codon usage patterns of the IDR/SR coding sequences of a virus was assessed by using neutrality plots.

The neutrality plot was constructed with GC3 as abscissa and GC12 (a sum of nucleotides G + C at the first, GC1 and second, GC2 codon positions) as ordinate, where each dot represents an independent IDR/SR of a virus. The regression line slopes of this plot give an estimation of the evolutionary rates of the mutational bias - natural selection equilibrium. For example, a regression line with a slope of zero indicates an insignificant influence of the mutational bias in shaping the codon usage patterns, while a slope of one is indicative of complete neutrality [44].

tRNA adaptation index

The tRNA adaptation index (tAI) is a widely used tool to measure the translation efficiency which takes into account the adaptation of codons to the intracellular tRNA pool of the host and the efficiency of each codon–anticodon pairing [15, 58]. The tAI for the IDRs and SRs coding sequences from 1637 viruses (the host tRNA genes information for the rest of the viruses are not available) with respect to their hosts was estimated (Table S3). The absolute adaptiveness value of the i^{th} codon was calculated using equation (4)

$$W_i = \sum_{j=1}^{n_i} (1 - S_{ij}).\, tGCN_{ij} \qquad (4)$$

where, n_i is the number of tRNA isoacceptors that recognize the i^{th} codon, $tGCN_{ij}$ is the gene copy number of the j^{th} tRNA that identifies the i^{th} codon, and S_{ij} is a selective constraint on the efficiency of the interaction between the i^{th} codon and the j^{th} tRNA [15].

The codon relative adaptiveness value (wi) was calculated by dividing each Wi by the maximum Wi value over all codons [15]. The tAI of a IDR/SR is the geometric mean of the wi values of its codons. The frequencies of host tRNA genes specific for each codon were retrieved from the GtRNAdb database [59]. In the case of multiple hosts for a particular virus, a reservoir or clinical host was considered for the analysis.

Investigation of the factors driving the evolution of codon usage in the core and the non-core regions of the viral proteins

In order to gain insight into the viral proteins segments encoding for structured regions (SRs), we divided SRs into core regions (CRs) consisting of the α-helix and the β-sheet, and non-core regions (NCRs) forming random coils or loops. All the accessible experimentally solved protein structures (n = 1077) were downloaded from the Protein Databank (https://www.rcsb.org/) and

filtered for redundant protein sequences (Table S4). The protein sequences and corresponding secondary structures (n = 737) were extracted using STRIDE secondary structure assignment tool [60]. The nucleotide sequences corresponding to CRs and NCRs were extracted by mapping UniProt identifiers to respective gene identifiers. The same set of protein sequences (n = 737) was considered for IDRs prediction using IUPred2A and ESpritz. The factors driving the evolution of codon usage in the CRs, NCRs and IDRs of these viral protein sequences were investigated by using the selected genetic tools as described in the previous sections of the manuscript.

Statistical analyses

Statistical analyses were carried out using GraphPad Prism 7.01 (GraphPad Software, San Diego, CA, USA). One-Way Analysis of Variance (ANOVA) with Bonferroni correction was used to compare the differences between tAI and CpG dinucleotide contents of IDRs and SRs. While in the case of ENc-GC3s plots, which shows the functional relationship between ENc and GC3s, we first estimated the signed distances of each IDRs and SRs of every virus genome type from the standard curve and thereafter, employed the Wilcoxon signed rank test, a non-parametric test, to compute the differences between the mean distances of IDRs and SRs from the standard curve. In all the statistical analyses, a p-value less than 0.01 was considered statistically significant. Additionally, we calculated Cohen's term d — commonly used to measure the effect size that is independent of group size — for each of the tAI values, CpG contents, and IDRs/SRs distances from the standard curve in the ENc-GC3s plots, and thereafter, classified the effect size as small (d = 0.2), medium (d = 0.5) and large (d ≥ 0.8) [61]. In the case of linear regression analyses, the effect size was estimated by Cohen's f^2 where f^2 = 0.02, 0.15, and ≥ 0.35 denotes small, medium and large size effects, respectively [62]. All the graphs were generated by using GraphPad Prism 7.01.

Results

The magnitude of mutational bias in shaping the codon usage in genomic regions encoding IDRs in double-stranded viral genomes is significantly higher than in regions encoding SRs

Previous studies emphasized that the mutational bias and the natural selection are the key factors driving the evolution of codon usage patterns in viral genomes [4]. Therefore, ENc plots were generated to investigate the influence of these factors on the viral IDRs and SRs. In the ENc–GC3s plot, the clustering of points over the standard curve suggests the absolute role of the mutational bias in shaping the codon usage patterns, whereas the below-curve clustering is indicative of the foremost influence of natural selection. So, the signed distances of each IDRs and SRs of every virus genome type from the standard curves are considered to examine the influence of natural selection and mutational bias. The variable size of datasets in different virus groups are further taken care by performing statistical analysis for effect-size. The mean signed distances of IDRs in ssDNA ($d = 1.012$, $p < 0.0001$), ssRNA(+) ($d = 0.529$, $p = 0.0014$), and ssRNA(-) ($d = 0.460$, $p < 0.0001$) viral genomes from the standard curve are significantly greater than that of SRs, whereas these differences are insignificant in other viral genomes (dsDNA, dsDNA-RT, dsRNA, ssRNA-RT, satellite, virophage, and unclassified) (Figure 1). These findings suggest that although the evolution of codon usage bias in IDRs of ssDNA, ssRNA(-), and ssRNA(+) viral genomes is primarily governed by the natural selection, nevertheless, the influence of mutational bias is not completely negligible. Furthermore, since the points falling on, or just below, the standard curve also indicate an optimal codon usage, the codon usage, especially in the IDRs of ssDNA, ssRNA(-), and ssRNA(+) viral genomes was found to be sub-optimal as compared to that of SRs.

We next employed the neutrality plots to explain the magnitude of mutational bias and natural selection in driving the codon usage bias. In these plots, a significant correlation between GC12

and GC3 coupled with a regression slope close to 1 indicates the prominent role of mutational bias while a non-significant or negative correlation with a regression slope close to zero indicates the predominant influence of natural selection in governing the codon usage patterns. It is clear from Figure 1 and Table S5 that the contribution of the mutational bias in influencing the codon usage in IDRs of dsDNA (r = 0.843, p < 0.0001, f^2 = 0.168, 28.3%), dsDNA-RT (r = 0.653, p < 0.0001, f^2 = 0.234, 47.1%), and dsRNA (r = 0.669, p < 0.0001, f^2 = 0.713, 45.4%) is remarkably high. By contrast, in the rest of the viral genome types, a non-significant contribution of mutational bias in influencing the codon usage in IDRs was observed. These results are in concordance with ENc–GC3s plots. IDRs, as predicted by ESpritz, also experienced a prominent influence of mutational bias in governing the codon usage of double-stranded viral genomes (Table S6). We also investigated the magnitude of mutational bias and natural selection in driving the codon usage bias in the viral genomic segments encoding for core regions (CRs), and non-core regions (NCRs). We noted that the contribution of natural selection in dictating the evolution of codon usage in the NCRs (80.2%) is higher than that of CRs (75.7 %) (p < 0.01) (Figures S1A and S1B).

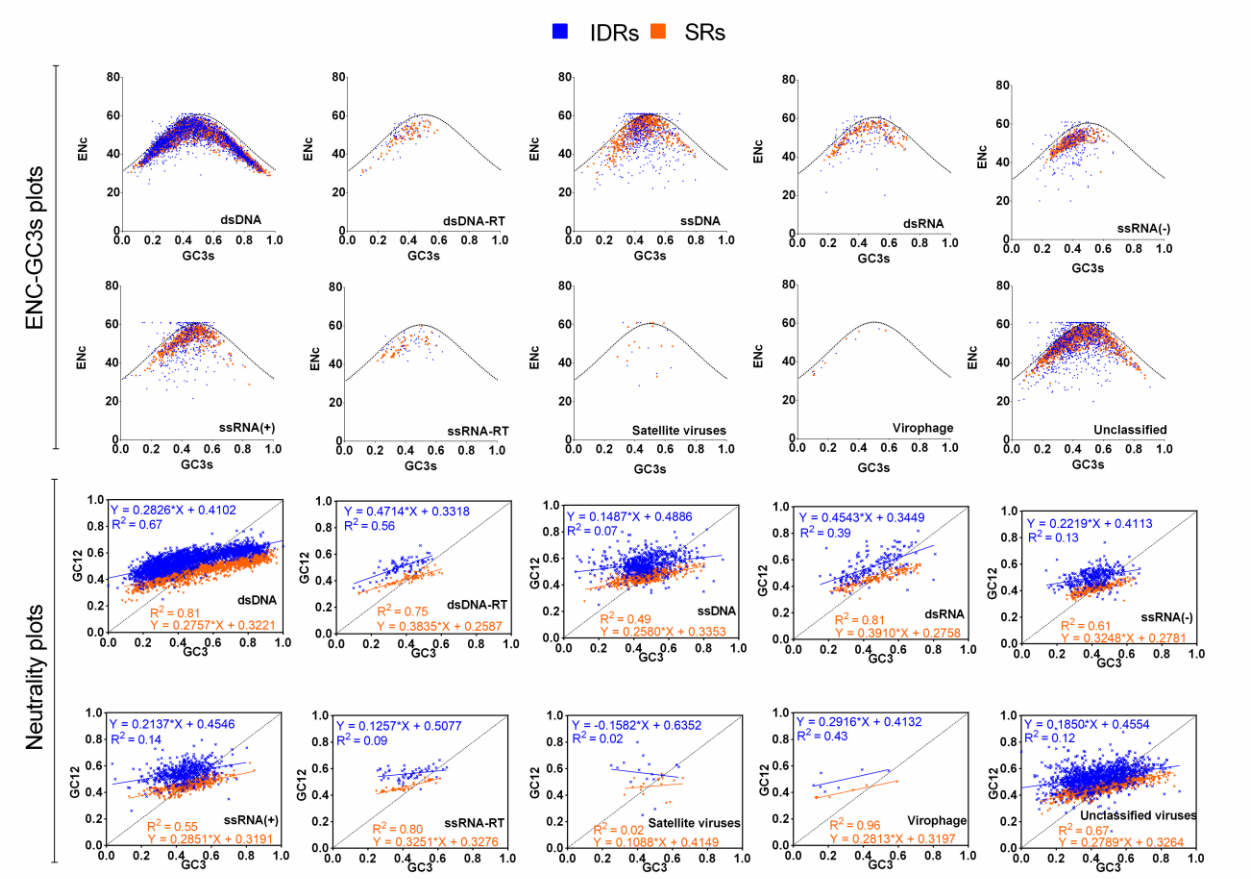

Figure 1. ENc–GC3s and neutrality plots for viral genomic segments encoding for intrinsically disordered regions (IDRs) and structured regions (SRs). In the ENc–GC3s plots, the black dotted line represents the standard curve, where the codon usage bias was determined by the GC3s compositions only. In the case of neutrality plots, the linear regressions of GC12 against GC3 for both the IDRs and SRs are shown. The IDRs and SRs encoded from each virus genome type are color coded, i.e., blue and orange, respectively.

The codon usage in genomic regions encoding IDRs in viruses is less optimized to the protein synthesis machinery of their corresponding hosts than that of genomic regions encoding SRs

We first classified the reference virus proteomes, for which host tRNA gene copy number information was available in the GtRNAdb database (n = 1637), into IDRs and SRs based on the two types of IDRs predictions (viz. short and long disorder) implemented in the IUPred2A program, and computed the tAIs of codons encoding IDRs and SRs with respect to their corresponding hosts. The computed tAIs showed a highly significant correlation between the two types of disorder (short and long) used for predicting IDRs ($r = 0.978$, $p < 0.0001$) and SRs ($r =$

0.998, p < 0.0001) indicating a consistency in results between the two disorder types (Figures 2A and 2B). This consistency is also maintained when a different IDRs predictor, e.g. ESpritz, was used, which showed a highly significant correlation between the computed tAIs of IDRs (r = 0.921, p < 0.0001) and SRs (r = 0.998, p < 0.0001), with the IUPred2A short disorder type (Figures S2A and S2B).

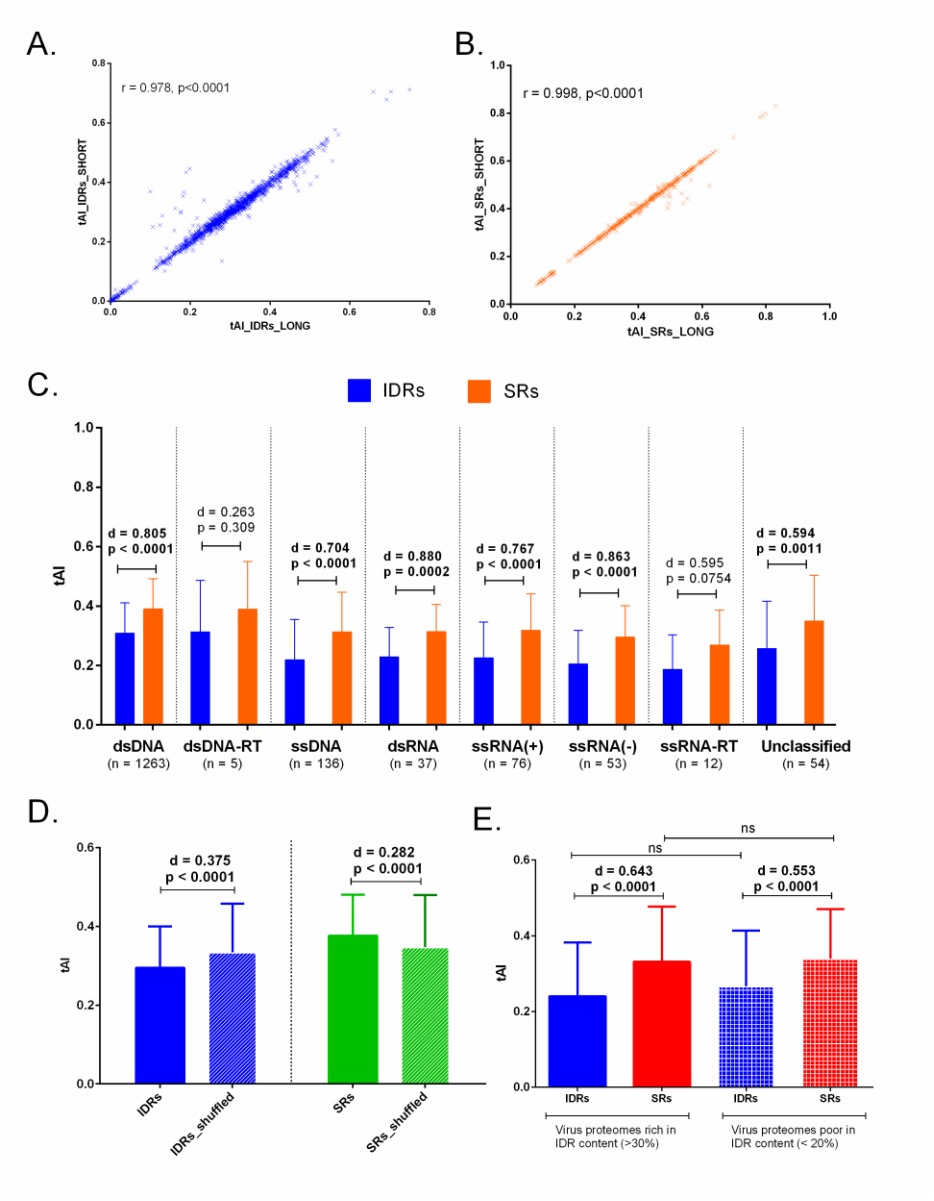

Figure 2. tRNA adaptation index (tAI) analyses of viral genomic segments encoding for intrinsically disordered regions (IDRs) and structured regions (SRs). (A), and (B) represent the correlation analyses between the long and

short disorder methods of IDRs predictions implemented in the IUPred2A for IDRs and SRs, respectively. (C) shows the comparison of the tAI of the viral genomic segments encoding IDRs and SRs of different virus genome types. (D) shows the difference between the tAI of real coding sequences of IDRs and SRs in comparison to the tAI of shuffled sequences of IDRs and SRs. (E) shows the comparison of the tAI of IDRs with respect to SRs in both IDRs-rich and IDRs-poor viral proteomes. One-Way Analysis of Variance (ANOVA) with Bonferroni correction was used to compare the differences between tAIs of IDRs and SRs. 'd' denotes the Cohen's term d that measures the effect size, independent of group size. The error bars correspond to the standard deviation.

The calculations revealed that the mean tAIs of viral genome regions encoding IDRs is lower than that of the regions encoding SRs (0.289±0.115, and 0.373±0.112, respectively, $p < 0.0001$). These results indicate that the codon usage is less optimized in the regions encoding IDRs than in those encoding SRs. In order to check whether this phenomenon depends on the virus genome type, we further investigated the tAIs in different viral genomes. Importantly, the mean tAIs of the regions encoding IDRs was found to be significantly lower than those of the regions encoding SRs in dsDNA, ssDNA, ssRNA(+), ssRNA(-) and unclassified viruses ($d = 0.594$ to 0.880, $p < 0.001$ to < 0.0001), while no significant difference was observed in dsRNA, dsDNA-RT and ssRNA-RT viruses (Figure 2C). These results were consistently obtained irrespective of whether IDRs were predicted by the IUPred2A short disorder type or by ESpritz (Figures S2C and S2D, respectively). In addition, we noted that reduced codon optimization in the viral genomic regions encoding IDRs was maintained even when compared to the viral genomic regions encoding CRs ($p < 0.01$) and NCRs ($p < 0.001$) (Figure S1C).

To investigate whether the reduced optimization in IDRs comes actually from codon bias or an intrinsic bias due to the amino acid bias associated with IDRs, we generate a synthetic dataset (by shuffling or randomizing codons) encoding IDRs and SRs where the amino acid composition is held fixed [63]. By doing this, it is possible to study the constraints associated with real and randomized coding sequences. We found that the randomized coding sequence in IDRs is significantly optimized (increased tAI) as compared to the real coding sequence ($d = 0.375$, $p < 0.0001$) (Figure 2D). Conversely, in the case of SRs, the randomized coding sequence is sub-

optimized (reduced tAI) in comparison to the real coding sequence ($d = 0.282$, $p < 0.0001$). These results imply that the selective constraints imposed by codon bias maintain the reduced optimization in IDRs.

It is known that the expression level of proteins rich in IDRs is tightly controlled and codons of poorly expressed proteins tend to be less optimized [64, 65]. If this is the case, then IDRs in IDR rich-proteins are expected to be less codon-optimized. To test this hypothesis, we divided the viral proteomes into two large categories, i.e. those that are enriched in IDRs (> 30%), and those that are poor in IDRs (< 20%). We showed that the reduced optimization in IDRs compared to that of SRs is a common trend in both IDRs-rich and IDRs-poor viral proteomes ($d = 0.553$-0.643, $p < 0.0001$) (Figure 2E). This finding is further supported by the non-significant difference in codon optimization (tAI) between the IDRs-rich and IDRs-poor viral proteomes. Furthermore, we tested the aforementioned hypothesis on the viral proteins level, where we categorized them into two groups, i.e. those that are enriched in IDRs (> 50%), and those that are poor in IDRs (< 20%). We showed that the results — reduced optimization of IDRs in comparison to that of SRs in both the IDRs-rich and IDRs-poor viral proteins — remained same ($d = 0.572$-0.857, $p < 0.0001$) (Figure S3). These results imply that, in comparison to SRs, the reduced optimization in IDRs is maintained irrespective of whether the viral proteins/proteomes are rich or poor in IDRs contents.

CpG dinucleotide content in viral genomic regions encoding IDRs is higher than in regions encoding SRs

The frequency of the CpG dinucleotide in IDRs and SRs of different viral genome types was estimated by dividing the number of CpG dinucleotides by that of total dinucleotides from the genome base compositions. The reverse transcribing (dsDNA-RT, and ssRNA-RT) and ssRNA(-

) viruses showed the most severe CpG depletion among all the viruses (Figure 3A). The viruses infecting Animalia, Plantae, and Protista showed a comparatively high CpG depletion with respect to viruses infecting Archaea, Fungi, and Bacteria ($p < 0.01$ to < 0.0001).

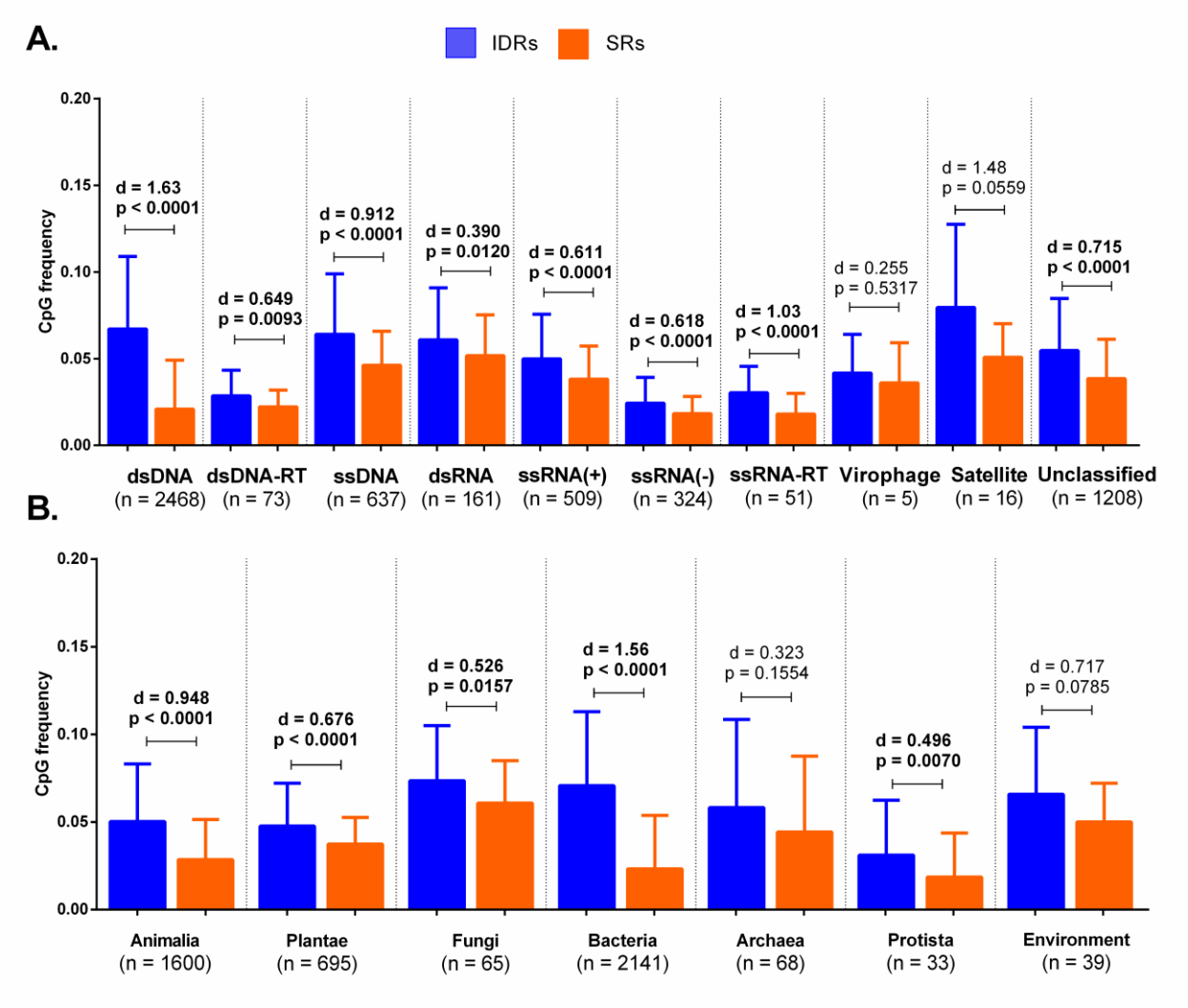

Figure 3. Comparison of CpG dinucleotide contents in the regions encoding IDRs and SRs of different virus genome types (A) and in viruses infecting taxonomically divergent hosts (B). One-Way Analysis of Variance (ANOVA) with Bonferroni correction was used to compare the differences between CpG content of IDRs and SRs. The error bars correspond to the standard deviation.

Interestingly, the abundance of the CpG dinucleotide in regions encoding IDRs is significantly higher compared to that of regions encoding SRs, a finding consistent for all virus genome types except dsDNA-RT, ssRNA-RT, virophage and satellite viruses ($d = 0.390$-1.63, $p = 0.0003$ to $<$

0.0001). Even, the CpG dinucleotide content in the IDRs is substantially higher compared to the viral genomic segments encoding for core (CRs, $p < 0.0001$) and non-core regions (NCRs, $p < 0.0001$) of SRs (Figure S1D). A higher content in the CpG dinucleotide in regions encoding IDRs compared to SRs is observed in viruses infecting Bacteria ($d = 1.56$, $p < 0.0001$), Animalia ($d = 0.948$, $p < 0.0001$), Plantae ($d = 0.676$, $p < 0.0001$), and Archaea ($d = 0.323$, $p < 0.001$), while no significant difference was observed for viruses infecting Fungi, Protista and isolated from the environment (Figure 3B). Although the CpG dinucleotide contents in the IDRs predicted by ESpritz show a minor discrepancy as compared to those predicted by IUPred2A (Figures S4A and S4B), the overall results do not affect the interpretations.

Discussion

Despite the functional importance of IDRs in viruses, a deep understanding of the evolutionary forces acting on them is lacking so far. To fill this gap in knowledge, we herein performed a comprehensive analysis of the abundance of IDRs in 6,108 proteomes from representative viruses belonging to 10 different genome types. We explored IDRs in these viral proteomes from multiple perspectives using selected genetic tools that enabled us to assess the evolutionary forces that shape the codon usage bias in IDRs/SRs.

Intrinsic disorder (ID) in proteins, in fact, is not a single state, but rather a set of biophysical features that lead to a variety of conformational states (known as flavors of disorder) [66]. As a result, IDPs/IDRs are characterized by high spatiotemporal heterogeneity and exist as dynamic structural ensembles. As a consequence, despite the fact that structure and disorder are often treated as binary states, they actually sit on a structural continuum [67]. Therefore, a correlation between protein structure and function is described by a "protein structure-function continuum" model, where a given protein exists as a dynamic conformational ensemble containing multiple

proteoforms characterized by a broad spectrum of structural features and possessing various functional potentials [68]. ID in proteins can be precisely defined in terms of conformational ensembles and can be captured by various experimental methods. Since only a small number of viral proteins have been characterized experimentally to capture the conformational ensembles, computational tools continue to be methods of choice that have allowed the large-scale disorder predictions.

Our study has a few limitations, since it depends on the disorder predictions, accuracies of which are not perfect. However, scarcity of the experimentally proven disorder information in viral proteins precludes the development of dedicated and the most precise disorder predictors specific for viruses. Furthermore, although viral disordered proteins are expected to undergo function-related structural transitions in their host's diverse and complex microenvironments, currently available disorder predictors are not entirely capable of relating to such biological microenvironments. Therefore, in the absence of accurate tools for unambiguous evaluation of intrinsic disorder in viral proteins, we can only rely on the careful use of currently available disorder predictors.

We accept the fact that none of the disorder predictors is perfect, and the resulting mispredictions might affect the reliability of the subsequent analyses and mislead the results for the evolutionary forces acting on IDRs. Therefore, we designed and conducted our experiments based on the utilization of two different disorder predictors (IUPRed2A, and ESpritz) for the analysis of the diverse virus proteomes. The benchmarking of these two disorder predictors against the experimentally validated disorder content information of viral proteins has also shown that the performance of IUPred2A (Specificity = 86.72%, and Sensitivity = 62.71%) and ESpritz (Specificity = 89.90%, and Sensitivity = 69.10%) is considerably high (Table S2B). Therefore,

these tools are frequently used as stand-alone predictors or in combination with other tools to provide information about disorder. Nevertheless, by evaluating the three key components of the findings, the magnitude of mutational bias in shaping the codon usage, translation efficiency and CpG dinucleotide content, we showed that our results remain consistently the same on a set of IDRs predicted by two principally different disorder predictors. These results also support the robustness and reliability of our findings and interpretations. Therefore, we consider that this approach allows us to create a more reliable depiction of the evolutionary forces acting on IDRs.

In the first place, we investigated the contribution of mutational bias and natural selection in the evolution of codon usage in the virus genomic regions encoding IDRs and SRs, using selected genetic tools, such as the ENc-plot, and the neutrality plot. The results showed that the codon usage in regions encoding IDRs of viruses possessing a single-stranded genome (ssDNA, ssRNA(-), and ssRNA(+)) is sub-optimal and primarily governed by the natural selection. Of note, a significantly higher mutational bias was observed in regions encoding IDRs of viruses possessing a double-stranded genome (dsDNA, dsDNA-RT, and dsRNA). Overall, the natural selection dictates the evolution of codon usage in regions encoding IDRs in all viruses, with the notable exception of viruses with a double-stranded genome. In addition, the evolution of codon bias in the segments encoding the core (composed of α-helix and β-sheet) and non-core (random coils or loops) regions of viral proteins are primarily governed by the natural selection, however, degree of extent varies. Previous studies have shown that both purifying selection and mutational bias are primarily responsible for the rapid evolution of IDRs in comparison to globular proteins, which is in concordance with our findings [69-74].

Many viruses possess the ability to infect multiple taxonomically divergent hosts for their efficient transmission in nature [75]. However, maintaining and adopting a multiple-host cell

cycle strategy seems to be intrinsically challenging for a virus, because these taxonomically divergent hosts do possess species-specific codon usage reflecting differences in intracellular tRNA pools. Our study has shown that codon usage in the viral genomic regions encoding IDRs is less optimized to the tRNA pool of their corresponding hosts than that of regions encoding SRs including the CRs and NCRs. This peculiar feature, however, is not limited to viruses, but has also been detected in eukaryotic genomic regions coding for IDRs [7, 76]. In eukaryotes, the reduced optimization of IDRs or reduced IDR translation efficiency (primarily due to non-optimal codons usage) has been shown to be important for both protein structure and biological function(s), where delaying the translation of IDRs may allow sufficient time for the proper folding of SRs or structured domains (SDs) [7]. Furthermore, protein expansion in the hosts is largely due to indels in regions encoding IDRs rather than in regions encoding SRs [77]. Because IDRs tend to arise later than SRs in the evolution of modern proteins, the codons in the genetic regions encoding IDRs tend to be less optimized than those encoding SRs [76].

Additionally, to examine whether the low optimization of the codon usage in the regions encoding IDRs arises from codon bias or from an intrinsic bias due to the amino acid bias associated with IDRs, we generated a synthetic dataset by shuffling the codons encoding IDRs and SRs while keeping the amino acid compositions fixed. In line with previous studies [71, 76], we showed that the poor codon usage optimization of regions encoding IDRs is pronounced in viruses because the codon usage patterns in regions encoding IDRs are more selectively constrained than in those encoding SRs. Furthermore, the codons in regions encoding IDRs are less optimized in both IDRs-rich and IDRs-poor viral proteins, thus ruling out the hypothesis that the regions encoding IDRs in IDRs-rich proteins tend to be less codon-optimized. We next investigated whether the low codon usage optimization in the regions encoding IDRs depends on

the lengths of the latter. To this end, we have compared disorder prediction results provided by the long disorder method of IUPred2A, conceived to identify IDRs longer than 30 amino acids, to those provided by the short disorder method (conceived for the identification of IDRs of 10 amino acids). The strong correlation between the short and long disorder prediction types indicates that the length of IDRs does not affect the results. The consistency in the results is conserved even when using a different IDRs predictor. Taken together, based on these findings, we speculate that the sub-optimal codon usage likely provides an opportunity to use divergent host species-specific tRNA pools, with this being especially true for the regions encoding IDRs of RNA viruses that typically have a broad host-specificity (i.e. they can infect multiple hosts). Concomitantly, this sub-optimal codon usage, with ensuing delayed translation of IDRs, would also ensure proper folding and function of viral SRs/SDs. This sub-optimal adaptation to the codon usage of the host would help the virus to be maintained among the multiple taxonomically divergent hosts (Figure S5).

Certain dinucleotides, such as CpG, are known to be over- and under-represented in the genomes of living organisms, thus creating distinct nucleotide compositional patterns or codon usage patterns [78]. The genomes of organisms, especially of the Animalia and Plantae, where the DNA methylation is extensive, employ a unique enzymatic mechanism to suppress CpG [79, 80]. In such organisms, including the Human, the methylated cytosine in a CpG dinucleotide is more prone to mutate into thymines through spontaneous deamination, creating mutation hotspots, and thus contributes to shaping the codon usage bias [71]. Furthermore, the regulatory activity of histone methyltransferases (which catalyze the methylation of histones and thus contribute to the regulation of gene transcription) has been shown to be mediated by their IDRs [81]. In contrast, although Prokaryotes and Archaea genomes do undergo methylation, such methylation

frequently occurs at a different site, i.e. N6-methyladenine, thereby explaining why these genomes show little CpG dinucleotide depletion [82].

Similarly, a CpG dinucleotide depletion has also been observed in viruses and appears to have functional roles in improving virus replication [83], escaping the host antiviral immune response [17, 84], and mimicking the hosts' CpG usage [85, 86]. Consistent with these studies, the genomes of viruses infecting Prokaryotes and Archaea show little CpG dinucleotide depletion [87-89]. Of note, high CpG depletion, especially in reverse transcribing viruses (dsDNA-RT and ssRNA-RT), may be due to the host-driven methylation pressure as these viruses produce DNA intermediates during their genome replication [90, 91]. Our results provide a link between virus CpG dinucleotide content and the methylation capabilities of the corresponding hosts.

In particular, the CpG dinucleotide depletion in RNA viruses provides them with an alternative mechanism to escape the host antiviral innate immune system. The unmethylated CpG is in fact a PAMP (Pathogen Associated Molecular Pattern) being recognized by Toll-like receptor 9 (TLR9), a type of intracellular pattern recognition receptor [86, 92]. In addition, dsRNA viruses that involve DNA intermediates during their viral genome replication seem to be affected primarily by the host methylation. Therefore, the host-driven CpG selective pressure on RNA viruses shapes their codon usage patterns. Nevertheless, the genomic regions encoding IDRs show significantly less CpG dinucleotide depletion compared to genomic regions encoding SRs including the CRs and NCRs. The less CpG dinucleotide depletion in the genome regions encoding IDRs is not only restricted to viruses, but it has also been observed in the human proteome [71]. Overall, these findings suggest that the host-driven methylation (most likely in DNA viruses) or CpG selective pressure (most likely in RNA viruses) contribute more

significantly in shaping the codon usage patterns in the regions encoding SRs than in those encoding IDRs.

Conclusion

Out study showed that the evolution of codon usage in viral IDRs is primarily dictated by the natural selection. The non-optimal codon usage (leading to poor optimization to host protein biosynthesis machinery) in viral IDRs seems to reflect the need to adapt to divergent host species-specific tRNA pools, while concomitantly allowing proper folding and function of viral SRs. Furthermore, the genomic regions encoding IDRs are comparatively more enriched in CpG than those encoding SRs, and therefore, experience comparatively less pressure imposed by the host-driven methylation or CpG selective pressure, making them hot-spots for mutations. Therefore, IDRs in viruses likely accept more translational errors than SRs.

Key points

- The study offers benchmarking of two distinct disorder prediction algorithms on a dataset comprising 6,108 reference viral proteomes to unravel the evolutionary forces acting on intrinsically disordered regions (IDRs).
- The natural selection predominantly governs the evolution of codon usage in regions encoding IDRs.
- The codon usage in regions encoding IDRs is less optimized to the protein synthesis machinery of their host than in those encoding structured regions (SRs).
- The selective constraints imposed by codon bias maintain reduced codon optimization in IDRs.

- The viral genomic regions encoding IDRs are comparatively more enriched in CpG than those encoding SRs, and therefore, IDRs in viruses likely accept more translational errors than SRs.

Acknowledgments

All the authors acknowledge and thank their respective Institutes.

Data availability Statement

All the datasets and codes generated in this study have been deposited at GitHub and can be accessed at https://github.com/naveenkumar1984/Viral_Proteins_IDP_Analysis.

Conflicts of interest

The authors declare no conflict of interest.

Funding

This study received no financial support for this study.

Naveen Kumar is a research scientist at National Institute of High Security Animal Disease, Bhopal, India. He is an evolutionary biologist specialized in virus evolution and with a background in synthetic biology, molecular, and clinical virology. His current research interests focus on protein structure modelling, viral phylodynamics, and development of nanoparticles-based diagnostics and vaccines.

Rahul Kaushik is a research scientist at Laboratory for Structural Bioinformatics, RIKEN Center for Biosystems Dynamics Research, Japan. He works in the area of complex biological functions of proteins, development of methods for protein structure prediction and applying

design principles to create proteins with novel architectures, new biological functions or effective therapeutics.

Chandana Tennakoon is a bioinformatician at the Pirbright institute, UK. He works on developing efficient methods to analyze sequencing data while having a special interest in analyzing viral data.

Vladimir N. Uversky is a professor at the Department of Molecular Medicine, Morsani College of Medicine, University of South Florida, USA. He works in the field of protein physics, with the focus on protein structure, folding, misfolding, and non-folding, which is related to the discovery of intrinsically disordered proteins (IDPs), analysis of their abundance in nature, characterization of their exceptional structural and functional plasticity, understanding of their vital roles in various biological processes, and establishing their involvement in the pathogenesis of multiple human diseases.

Sonia Longhi is leader of the team "Structural Disorder & Molecular Recognition" at the Architecture and Function of Biological Macromolecules (AFMB) laboratory, UMR 7257 CNRS & Aix-Marseille University, France. She works on the identification, characterization and elucidation of the functional role of intrinsically disordered regions within proteins relevant in terms of human health, with a focus on proteins of the replicative complex of human pathogenic viruses, such as measles virus and the recently emerged Nipah and Hendra viruses.

Kam Y. J. Zhang is a team leader at Laboratory for Structural Bioinformatics, Center for Biosystems Dynamics Research, RIKEN, Japan. He works in the area of complex biological functions of proteins, development of methods for protein structure prediction and applying

design principles to create proteins with novel architectures, new biological functions or effective therapeutics.

Sandeep Bhatia is a principal scientist at National Institute of High Security Animal Disease, Bhopal, India. He is a recipient of ‘National Fellow award’, a prestigious award for carrying out cutting-edge research offered by Indian Council of Agricultural Research. He works on discovery and characterization of viruses, relevant to poultry health, with a focus on influenza viruses. He has vast experiences in the development of next generation diagnostics, and vaccines.

References

1. Clarke, B., Darwinian evolution of proteins. Science 1970; 168(3934):1009-11.

2. Hanson, G. and J. Coller, Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol 2018; 19(1): 20-30.

3. Ikemura, T., Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985; 2(1): 13-34.

4. Kumar, N., et al., Evolution of Codon Usage Bias in Henipaviruses Is Governed by Natural Selection and Is Host-Specific. Viruses 2018; 10(11).

5. Plotkin, J.B. and G. Kudla, Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 2011; 12(1): 32-42.

6. Shabalina, S.A., N.A. Spiridonov, and A. Kashina, Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res 2013; 41(4): 2073-94.

7. Zhou, M., et al., Nonoptimal codon usage influences protein structure in intrinsically disordered regions. Mol Microbiol 2015; 97(5): 974-87.

8. Bulmer, M., The selection-mutation-drift theory of synonymous codon usage. Genetics 1991; 129(3): 897-907.

9. Cui, P., et al., Distinct contributions of replication and transcription to mutation rate variation of human genomes. Genomics Proteomics Bioinformatics 2012; 10(1): 4-10.

10. Lobry, J.R., Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 1996; 13(5): 660-5.

11. Comeron, J.M., Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics 2004; 167(3): 1293-304.

12. Green, P., et al., Transcription-associated mutational asymmetry in mammalian evolution. Nat Genet 2003; 33(4): 514-7.

13. Kaufmann, W.K. and R.S. Paules, DNA damage and cell cycle checkpoints. Faseb j 1996; 10(2): 238-47.

14. Roth, A., M. Anisimova, and G.M. Cannarozzi, Measuring codon usage bias, in Codon Evolution: Mechanisms and Models, G.M. Cannarozzi and A. Schneider, Editors. 2012, Oxford University Press.

15. dos Reis, M., R. Savva, and L. Wernisch, Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 2004; 32(17): 5036-44.

16. Hershberg, R. and D.A. Petrov, Selection on codon bias. Annu Rev Genet 2008. 42: 287-99.

17. Kumar, N., et al., Revelation of Influencing Factors in Overall Codon Usage Bias of Equine Influenza Viruses. PLoS One 2016; 11(4): e0154376.

18. Marsh, G.A., et al., Highly conserved regions of influenza a virus polymerase gene segments are critical for efficient viral RNA packaging. J Virol 2008; 82(5): 2295-304.

19. Simmonds, P., A. Tuplin, and D.J. Evans, Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. RNA 2004; 10(9): 1337-51.

20. Weill, L., et al., A new type of IRES within gag coding region recruits three initiation complexes on HIV-2 genomic RNA. Nucleic Acids Res 2010; 38(4):1367-81.

21. Dunker, A.K., et al., The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics 2008; 9(S2): S1.

22. Dyson, H.J. and P.E. Wright, Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 2005; 6(3): 197-208.

23. Habchi, J., et al., Introducing protein intrinsic disorder. Chem Rev 2014;114(13): 6561-88.

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

25. Xue, B., et al., Structural disorder in viral proteins. Chem Rev 2014; 114(13): 6880-911.

26. Dunker, A.K., et al., Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform 2000; 11: 161-71.

27. Peng, Z., et al., Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. Cell Mol Life Sci 2015; 72(1): 137-51.

28. Uversky, V.N., The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. J Biomed Biotechnol 2010; 2010: 568068.

29. Ward, J.J., et al., Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 2004; 337(3): 635-45.

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

31. Tokuriki, N., et al., Do viral proteins possess unique biophysical features? Trends Biochem Sci 2009; 34(2): 53-9.

32. Xue, B., et al., Viral disorder or disordered viruses: do viral proteins possess unique features? Protein Pept Lett 2010; 17(8): 932-51.

33. Charon, J., et al., First Experimental Assessment of Protein Intrinsic Disorder Involvement in an RNA Virus Natural Adaptive Process. Mol Biol Evol 2018; 35(1): 38-49.

34. Goh, G.K., et al., Shell disorder analysis predicts greater resilience of the SARS-CoV-2 (COVID-19) outside the body and in body fluids. Microb Pathog 2020; 144: 104177.

35. Goh, G.K., A.K. Dunker, and V.N. Uversky, Protein intrinsic disorder and influenza virulence: the 1918 H1N1 and H5N1 viruses. Virol J 2009; 6: 69.

36. Kakisaka, M., et al., Intrinsically disordered region of influenza A NP regulates viral genome packaging via interactions with viral RNA and host PI(4,5)P2. Virology 2016; 496: 116-126.

37. Mishra, P.M., V.N. Uversky, and R. Giri, Molecular Recognition Features in Zika Virus Proteome. J Mol Biol 2018; 430(16): 2372-2388.

38. Redwan, E.M., A.A. AlJaddawi, and V.N. Uversky, Structural disorder in the proteome and interactome of Alkhurma virus (ALKV). Cell Mol Life Sci 2019; 76(3): 577-608.

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

40. Davey, N.E., G. Travé, and T.J. Gibson, How viruses hijack cell regulation. Trends Biochem Sci 2011; 36(3): 159-69.

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

42. Walter, J., et al., Comparative analysis of mutational robustness of the intrinsically disordered viral protein VPg and of its interactor eIF4E. PLoS One 2019; 14(2): e0211725.

43. Wright, F., The 'effective number of codons' used in a gene. Gene 1990; 87(1): 23-9.

44. Sueoka, N., Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci U S A 1988; 85(8): 2653-7.

45. Consortium, T.U., UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017; 45(D1): D158-d169.

46. Mihara, T., et al., Linking Virus Genomes with Host Taxonomy. Viruses 2016; 8(3): 66.

47. Whittaker, R.H. and L. Margulis, Protist classification and the kingdoms of organisms. Biosystems 1978; 10(1-2): 3-18.

48. Dosztányi, Z., et al., IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 2005; 21(16): 3433-4.

49. Peng, Z.L. and L. Kurgan, Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci 2012; 13(1): 6-18.

50. Walsh I., et al., Comprehensive large-scale assessment of intrinsic protein disorder. Bioinformatics 2015; 31(2): 201-8.

51. Necci, M., et al., A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics 2018; 34(3): 445-452.

52. Necci, M., et al., MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins. Bioinformatics 2017; 33(9): 1402-1404.

53. Fan, X. and L. Kurgan, Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus. J Biomol Struct Dyn 2014; 32(3): 448-64.

54. Almog, G., et al., Tuning Intrinsic Disorder Predictors For Virus Proteins. Virus Evol 2021; veaa106, https://doi.org/10.1093/ve/veaa106.

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

56. Dosztányi, Z., et al., The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005; 347(4): 827-39.

57. Walsh, I., et al., ESpritz: accurate and fast prediction of protein disorder. Bioinformatics 2012; 28(4): 503-9.

58. Ma, Y.P., et al., Multiple Evolutionary Selections Involved in Synonymous Codon Usages in the Streptococcus agalactiae Genome. Int J Mol Sci 2016; 17(3): 277.

59. Chan, P.P. and T.M. Lowe, GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res 2009; 37(Database issue): D93-7.

60. Heinig, M. and D. Frishman, STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. Nucleic Acids Res 2004; 32: W500-W502.

61. Sullivan, G.M. and R. Feinn, Using Effect Size-or Why the P Value Is Not Enough. Journal of graduate medical education 2012; 4(3): 279-282.

62. Selya, A.S., et al., A Practical Guide to Calculating Cohen's f(2), a Measure of Local Effect Size, from PROC MIXED. Frontiers in psychology 2012; 3: 111-111.

63. Rice, P., I. Longden, and A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000; 16(6): 276-7.

64. Ikemura, T., Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 1981; 151(3): 389-409.

65. Vavouri, T., et al., Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 2009; 138(1): 198-208.

66. Vucetic, S., et al., Flavors of protein disorder. Proteins 2003; 52(4): 573-84.

67. DeForte, S. and V.N. Uversky, Order, Disorder, and Everything in Between. Molecules 2016; 21(8): 1090.

68. Uversky, V.N., Protein intrinsic disorder and structure-function continuum. Prog Mol Biol Transl Sci. 2019; 166: 1-17.

69. Afanasyeva, A., et al., Human long intrinsically disordered protein regions are frequent targets of positive selection. Genome Res 2018; 28(7): 975-982.

70. Brown, C.J., et al., Evolution and disorder. Curr Opin Struct Biol 2011; 21(3): 441-6.

71. Forcelloni, S. and A. Giansanti, Evolutionary Forces and Codon Bias in Different Flavors of Intrinsic Disorder in the Human Proteome. J Mol Evol 2020; 88(2): 164-178.

72. Nilsson, J., M. Grahn, and A.P. Wright, Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. Genome Biol 2011; 12(7): R65.

73. Schlessinger, A., et al., Protein disorder--a breakthrough invention of evolution? Curr Opin Struct Biol 2011; 21(3): 412-8.

74. Xue, B., et al., Intrinsically disordered regions of p53 family are highly diversified in evolution. Biochim Biophys Acta 2013; 1834(4): 725-38.

75. Ebel, G.D., Promiscuous viruses-how do viruses survive multiple unrelated hosts? Curr Opin Virol 2017; 23: 125-129.

76. Homma, K., T. Noguchi, and S. Fukuchi, Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. Nucleic Acids Res 2016; 44(21): 10051-10061.

77. Light, S., et al., Protein expansion is primarily due to indels in intrinsically disordered regions. Mol Biol Evol 2013; 30(12): 2645-53.

78. Zu, W., et al., Genome-wide evolution analysis reveals low CpG contents of fast-evolving genes and identifies antiviral microRNAs. J Genet Genomics 2020; 47(1): 49-60.

79. Bird, A.P., DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 1980; 8(7): 1499-504.

80. Fros, J.J., et al., CpG and UpA dinucleotides in both coding and non-coding regions of echovirus 7 inhibit replication initiation post-entry. Elife 2017; 6.

81. Yang, J., M. Gao, and Y. Huang, Regulating the Activation of Ash1/Ash1L Histone Methyltransferase by Intrinsically Disordered Regions. Biophysical Journal 2020; 118: 62a.

82. Mohapatra, S.S. and E.G. Biondi, DNA Methylation in Prokaryotes: Regulation and Function, in Cellular Ecophysiology of Microbe. Handbook of Hydrocarbon and Lipid Microbiology., K. T, Editor. 2017, Springer.

83. Atkinson, N.J., et al., The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. Nucleic Acids Res 2014; 42(7): 4527-45.

84. Lytras, S. and J. Hughes, Synonymous Dinucleotide Usage: A Codon-Aware Metric for Quantifying Dinucleotide Representation in Viruses. Viruses 2020; 12(4).

85. Cheng, X., et al., CpG usage in RNA viruses: data and hypotheses. PLoS One 2013; 8(9): e74109.

86. Greenbaum, B.D., et al., Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog 2008; 4(6): e1000079.

87. Lobo, F.P., et al., Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One 2009; 4(7): e6282.

88. Simmonds, P., et al., Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla--selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. BMC Genomics 2013; 14: 610.

89. Upadhyay, M., et al., CpG dinucleotide frequencies reveal the role of host methylation capabilities in parvovirus evolution. J Virol 2013; 87(24): 13816-24.

90. Ellis, J., A. Hotta, and M. Rastegar, Retrovirus silencing by an epigenetic TRIM. Cell 2007; 131(1): 13-4.

91. Leung, D.C. and M.C. Lorincz, Silencing of endogenous retroviruses: when and why do histone marks predominate? Trends Biochem Sci 2012; 37(4): 127-33.

92. Vetsigian, K. and N. Goldenfeld, Genome rhetoric and the emergence of compositional bias. Proc Natl Acad Sci U S A 2009; 106(1): 215-20.