Tuning intrinsic disorder predictors for virus proteins

Gal Almog[1,*], Abayomi S Olabode[1,*,†], Art FY Poon[1,2,3]

∗ denotes equal contribution
† corresponding author

[1]Department of Pathology & Laboratory Medicine, Western University, London, Canada;
[2]Department of Applied Mathematics, Western University, London, Canada;
[3]Department of Microbiology & Immunology, Western University, London, Canada

Abstract

Many virus-encoded proteins have intrinsically disordered regions that lack a stable folded three-dimensional structure. These disordered proteins often play important functional roles in virus replication, such as down-regulating host defense mechanisms. With the widespread availability of next-generation sequencing, the number of new virus genomes with predicted open reading frames is rapidly outpacing our capacity for directly characterizing protein structures through crystallography. Hence, computational methods for structural prediction play an important role. A large number of predictors focus on the problem of classifying residues into ordered and disordered regions, and these methods tend to be validated on a diverse training set of proteins from eukaryotes, prokaryotes and viruses. In this study, we investigate whether some predictors outperform others in the context of virus proteins. We evaluate the prediction accuracy of 21 methods, many of which are only available as web applications, on a curated set of 126 proteins encoded by viruses. Furthermore, we apply a random forest classifier to these predictor outputs. Based on cross-validation experiments, this ensemble approach confers a substantial improvement in accuracy, *e.g.*, a mean 36% gain in Matthews correlation coefficient. Lastly, we apply the random forest predictor to SARS-CoV-2 ORF6, an accessory gene that encodes a short (61 AA) and moderately disordered protein that inhibits the host innate immune response.

Introduction

For almost a century, it was assumed that proteins required a properly folded and stable three-dimensional or tertiary structure in order to function [1–3]. More recently, it has become evident that many proteins and protein regions are disordered, which are referred to as intrinsically disordered proteins (IDPs) and intrinsically disordered protein regions (IDPRs), respectively. Both IDPs and IDPRs can perform important biological functions despite lacking a properly folded and stable tertiary structure [2, 4].

These kinds of proteins are an important area of research because they play major roles in cell regulation, signalling, differentiation, survival, apoptosis and proliferation [5, 6]. Some are also postulated to be involved in disease etiology and could represent potential targets for new drugs [7, 8]. Virus-encoded IDPs facilitate multiple functions such as adaptation to new or dynamic host environments, modulating host gene expression to promote virus replication, or counteracting host-defense mechanisms [9–11]. IDPRs may be more tolerant of non-synonymous mutations than ordered protein regions [12], which may partly explain why virus genomes can tolerate high mutation rates [13, 14]. Viruses also have very compact genomes with overlapping reading frames [15, 16], in which mutations may potentially modify multiple proteins. This may confer viruses a greater capacity to acquire novel functions and interactions [17]. Overlapping regions tend to be more structurally disordered when compared to non-overlapping regions [18].

Several experimental techniques are available to detect IDPs and IDPRs. The most common methods identify either protein regions in crystal structures that have unresolvable coordinates (X-ray crystallography) or regions in nuclear magnetic resonance (NMR) structures that have divergent structural conformations [6, 19, 20]. Other experimental techniques include circular dichroism (CD) spectroscopy and limited proteolysis (LiP) [1]. The challenge, however, is that these methods are very labour-intensive and difficult to scale up to track the rapidly accumulating number of

unique protein sequences in public databases [6, 19]. At the time of this writing, over 60 million protein sequences have been deposited in the Uniprot database, yet only 0.02% of these sequences have been annotated for disorder [6]. As a result, numerous computational techniques that could potentially predict intrinsic disorder in protein sequences have been developed. These techniques work based on the assumptions that compared to IDPs and IDPRs, ordered proteins have a different amino acid composition as well as levels of sequence conservation [21, 22]. To date about 60 predictors for intrinsic disorder in proteins have been developed [1, 23, 24], which can be broadly classified into three major categories. The first category, the scoring function-based methods, predict protein disorder solely based on basic statistics of amino acid propensities, physio-chemical properties of amino acids and residue contacts in folded proteins to detect regions of high energy. A second category is characterized by the use of machine learning classifiers (*e.g.*, regularized regression models or neural networks) to predict protein disorder based on amino acid sequence properties. The third category are meta-predictors that predict disorder from an ensemble of predictive methods from the other two categories [1, 6, 25].

Different predictors of intrinsic disorder are developed on a variety of methodologies and will inevitably vary with respect to their sensitivities and biases in application to different protein sequences. As a result, it has been relatively difficult to benchmark these methods to identify a single disorder prediction method that can be classified as the most accurate relative to the others [24]. The DisProt database is a good resource for obtaining experimental data that has been manually curated for disorder in proteins, and can be used for benchmarking the performance of disorder predictors. As of April 27th, 2020, the Disprot protein database contained $n = 3500$ proteins of which 126 were virus-encoded proteins that have been annotated for intrinsic disorder as a presence-absence characteristic at the amino acid level [26, 27]. Previously, Tokuriki *et al.* [14] reported preliminary evidence that when compared to non-viruses, viral proteins possess many distinct biophysical properties including having shorter disordered regions. We are not aware

of a published study that has previously benchmarked predictors of intrinsic disorder specifically for viral proteins. Here, we report results from a comparison of 21 disorder predictors on viral proteins from the DisProt database to firstly determine which methods work best for viruses, and secondly to generate inputs for an ensemble predictor that we evaluate alongside the predictors used individually.

Methods

Data collection

The Database of Protein Disorder (DisProt) [28] was used to collect virus protein sequences annotated with intrinsically disordered regions, based on experimental data derived from various detection methods; *e.g.*, X-ray crystallography, NMR spectroscopy, CD spectroscopy (both far and near UV) and protease sensitivity. DisProt records include the amino acid sequence and all disordered regions annotated with the respective detection methods as well as specific experimental conditions. At the time of our study, DisProt contained 3,500 author-verified proteins, of which all viral proteins were collected for the present study. A total of 126 virus proteins were obtained, derived from different detection methods. Similarly, a set of 126 non-viral proteins was sampled at random without replacement from the protein database for comparison.

We evaluated a number of disorder prediction programs and web applications. From the methods tested, we selected a subset of predictors favouring those that were developed more recently, are actively maintained, and performed well in previous method comparison studies [1, 29]. Where alternate settings or different versions based on training data were available for a given predictor, we tested all combinations. Our final set of 21 prediction methods tested were: SPOT-Disorder2 [30], PONDR-FIT [31], IUPred2 (short and long) [32], PONDR (VLXT, XL1-XT, CAN-XT, VL3-BA, and VSL2 variants) [33], Disprot (VL2 and variants VL2-V, -C and -S; VL3, VL3H, and VSLB) [34], CSpritz (short and long) [35], and ESpritz (variants trained on X-ray, NMR, and Disprot

data) [36]. Although several other predictor models have been released online, the respective web services were unavailable or broken over the course of our data collection.

To obtain disorder predictions from the methods that were only accessible as web applications, *i.e.*, with no source code or compiled binary standalone distribution, we wrote Python scripts to automate the process of submitting protein sequence inputs and parsing HTML outputs. We used Selenium in conjunction with ChromeDriver (v81.0.4044.69) [37] to automate the web browsing and form submission processes. For each predictor, we implemented a delay of 90 seconds between consecutive protein sequence queries to avoid overloading the webservers hosting the respective predictor algorithms with repeated requests. Due to issues with the Disprot webserver, we were only able to obtain predictions for the non-viral protein data set for 13 of the predictors.

We converted each DisProt record to a binary vector corresponding to ordered/disordered state of residues in the amino acid sequence. To compare results between disorder prediction algorithms, we dichotomized continuous-valued residue predictions, *i.e.*, intrinsic disorder probability, by locating the threshold that maximized the Matthews correlation coefficient (MCC) for each predictor applied to the DisProt training data. This optimal threshold was estimated using Brent's root-finding algorithm as implemented by the *optim* function in the R statistical computing environment (version 3.4.4). In addition, we calculated the accuracy, specificity and sensitivity for each predictor from the contingency table of DisProt residue labels and dichotomized predictions.

Ensemble classifier training and validation

To assess whether the accuracy of existing predictors could be further improved on the virus-specific data set, we trained an ensemble classifier on the outputs of all predictors as features. Specifically, we used the random forest method implemented in the *scikit-learn* (version 0.23.1) Python module [38], which employs a set of de-correlated decision trees and averages their respective outputs to obtain an ensemble prediction [39]. To reduce bias, random forests fit the same

decision trees to many bootstrap samples of the training data, and each committee of trees 'votes' for a particular classification [40]. By splitting the trees based on different samples of features, random forests reduce the correlation between trees and the overall variance.

We split the viral protein data into random testing and training subsets, with 30% of protein sequences reserved for testing. Due to class imbalance in the data (*i.e.*, only a minority of residues are labeled as disordered), we used stratified random sampling using the 'StratifiedShuffleSplit' function in the *scikit-learn* module. This function stratifies the data by label so that a constant proportion of labels is maintained in the training subset. Continuous-valued outputs from each predictor were normalized to a zero mean and unit variance. Thus, we did not apply the dichotomizing thresholds to these features (predictor outputs) when training the random forest classifier.

We used 5-fold cross validation to tune the four hyper-parameters of the random forest classifier; namely: (1) the number of decision trees; (2) the maximum depth of any given decision tree; and the minimum number of samples required to split (3) an internal node or (4) a leaf node. To further minimize the effect of class imbalance in our data, we used over-sampling to balance the data with synthetic cases [41]. As suggested in [42], we applied an over-sampling procedure at every iteration of the cross-validation analysis to avoid over-optimistic results. We used the Python package *imbalanced-learn* [43] to over-sample the minority class (residues in intrinsically disordered regions) using the synthetic minority oversampling technique (SMOTE) [44]. SMOTE generates new cases by sampling the original data at random with replacement, evaluates each sample's k nearest neighbours in the feature space, and then generates new synthetic samples along the vectors joining the sample to one of the neighbouring points. Over-sampling enables decision trees to be more generalizable by amplifying the decision region of the minority class.

Using the optimized tuning parameters, we fit the final model on all of the training data. We applied this final model to generate predictions on the reserved testing data and calculated the MCC,

sensitivity, specificity and accuracy. We repeated this process 10 times with randomly generated seeds to split the data into training and testing subsets, and averaged these performance metrics across replicates.

Comparison to non-viral data

To characterize how the performance of individual disorder predictors might vary among proteins from viruses and non-viruses, we computed the root mean square error (RMSE) for all continuous-valued predictions relative to the Disprot label (0, 1). We visualized this error distribution using principal component analysis (PCA). As well, we trained a support vector machine (SVM) on the RMSE values to determine whether the virus/non-virus labels were separable in this space. We used the default radial basis kernel with the C-classification SVM method implemented in the R package *e1071* [45], with 100 training subsets sampled at random without replacement for half of the data, and the remaining half for validation.

Data availability

We have released the Python scripts for automating queries to the disorder prediction web servers under a permissive free software license at https://github.com/PoonLab/Floppy/.

Results and Discussion

Viral and non-viral proteins have similar levels of disorder

We obtained 126 viral and 126 randomly selected non-viral protein sequences from the DisProt database. The sequences were already annotated manually by a panel of experts for the presence or absence of disorder at each amino acid position, based on experimental data [27]. Supplementary Tables S1 and S2 summarize the composition of the viral and non-viral protein datasets, respectively. The viral protein data set represents 22 virus families and 48 species. Not surpris-

ingly, human immunodeficiency virus type 1 was disproportionately represented in these data with 16 entries corresponding to seven different gene products. Similarly, the non-viral protein data set was predominated by 75 human proteins, followed by 23 proteins from the yeast *Saccharomyces cerevisiae*. We found no significant difference in amino acid sequence lengths between viruses and all other organisms (Wilcoxon rank sum test, $P = 0.60$), with median lengths of 355 [interquartile range, IQR: 145–846] and 395 [203–729] amino acids, respectively. Furthermore, the dispersion in sequence lengths was significantly greater among viral proteins relative to the nonviral proteins (Ansari-Bradley test, $P = 0.0028$). There was no significant difference in the proportion of residues in disordered regions between the viral and non-viral data (Wilcoxon $P = 0.97$). The mean proportions were 0.30 (interquartile range, IQR [0.07-0.42]) for viral and 0.30 [0.07-0.47] for non-viral proteins, and similar numbers of proteins exhibited complete disorder (13 and 9, respectively).

Divergent predictions of disorder in viral proteins

Our first objective was to benchmark the performance of different predictors of intrinsic protein disorder to determine which predictor conferred the highest accuracy for viral proteins. These predictors generate continuous-valued outputs that generally correspond to the estimated probability that the residue is in an intrinsically disordered region. To create a uniform standard for comparison to the binary presence-absence labels, we optimized the disorder prediction thresholds as a tuning parameter for each predictor for the viral and non-viral datasets, respectively (Supplementary Tables S3 and S4). Put simply, residues with values above the threshold were classified as disordered. We used both the Matthews correlation coefficient (MCC, ranging from -1 to $+1$ [46]) and area under the receiver-operator characteristic curve (AUC, ranging from 0 to 1) to quantify the performance of each predictor.

These quantities were significantly correlated (Spearman's $\rho = 0.95$, $P = 5.2 \times 10^{-6}$) and identified ESpritz.Disprot, CSpritz.Long and SPOT.Disorder2 as the most effective predictors for the

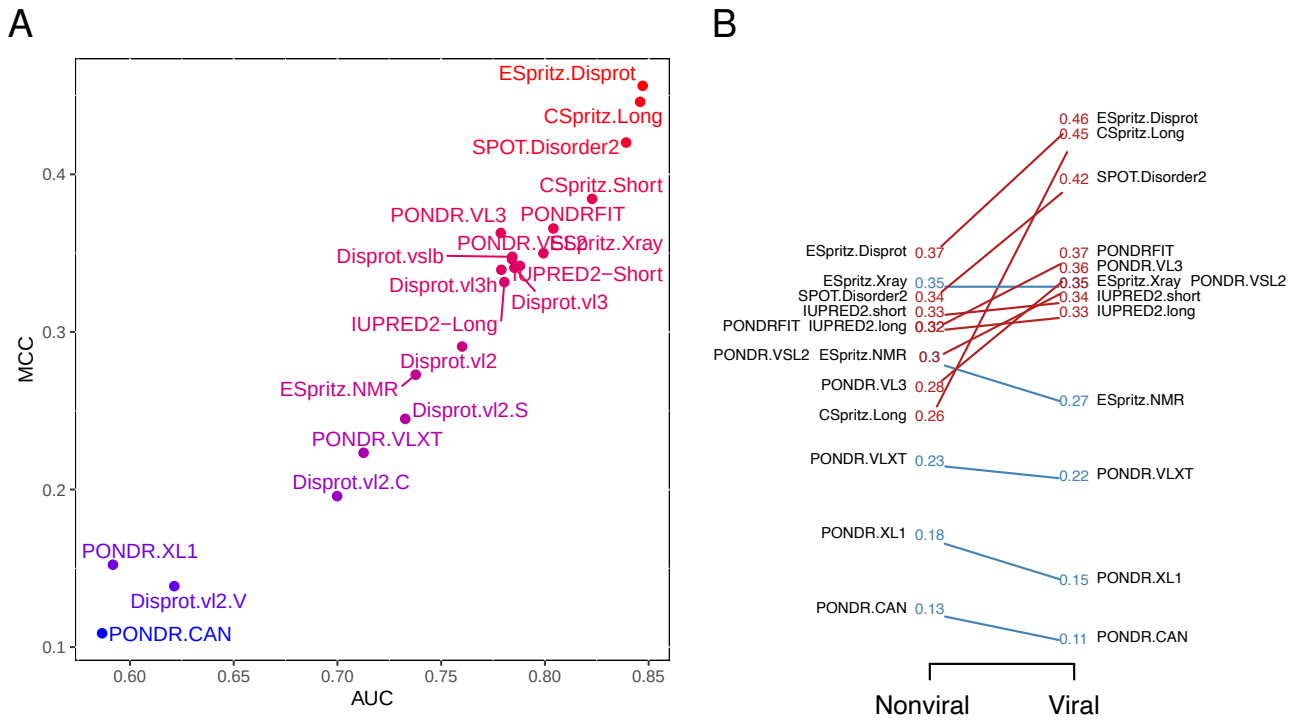

Figure 1: Performance of predictors on viral data set. (A) Scatterplot of MCC and AUC values for 21 predictors applied to the viral protein data set. (B) Slopegraph comparing the MCC values for 13 predictors applied to both non-viral and viral data sets. Because the three variants of the ESpritz model obtained identical MCC values, the corresponding labels were merged. Two labels (PONDRFIT, PONDR.VSL2) were displaced to prevent overlaps on the left and right sides, respectively.

viral proteins (Figure 1A). ESpritz.Disprot obtained the highest overall values for both MCC and AUC (0.46 and 0.85, respectively). We note that SPOT.Disorder2 has recently been reported to exhibit a high degree of prediction accuracy for proteins of varying length [47]. In contrast, the predictors Disprot-VL2-V, PONDR-XL1 and PONDR-CAN performed very poorly on the viral dataset with MCC < 0.2 and AUC < 0.65. VL2-V is a 'flavour' of the VL2 predictors which were allowed to specialize on different subsets of a partitioned training set; for example, V tended to call higher levels of disorder in proteins of Archaebacteria [34]. Similarly, PONDR-XL1 was optimized to predict longer disordered regions and PONDR-CAN was trained specifically on calcineurins (a protein phosphatase) that is known to perform poorly on other proteins [48].

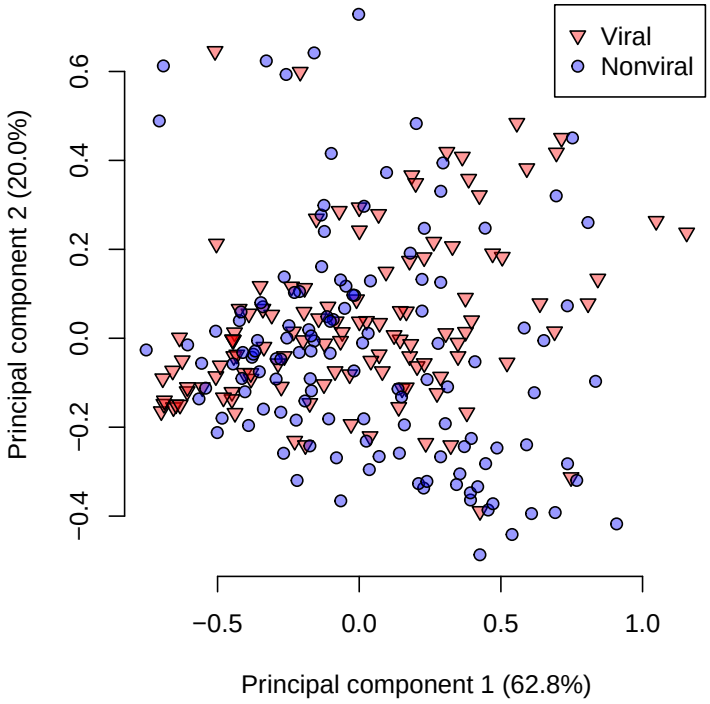

Figure 2: Principal components analysis plot of the root mean squared errors (RMSEs) for 13 disorder predictors on viral (red, triangles) and non-viral (blue, circles) protein sequences. The percentages of total variance explained by the first two principal components are indicated in parentheses in the respective axis labels.

Figure 1B compares the MCC values for non-viral and viral protein data sets. Predictors exhibited substantially less variation in MCC for the non-viral data — put another way, the majority of predictors were more accurate at predicting disorder in viral proteins. The entire set of MCC, AUC, sensitivity and specificity values for both data sets are summarized in Supplementary Tables S3 and S4. To examine potential differences among predictors in greater detail, we calculated the RMSE for each protein and predictor and used a principal components analysis to visualize the resulting matrix (Figure 2). The PCA indicated that the different predictors did not exhibit markedly divergent error profiles at the level of entire proteins. However, a support vector machine classifier trained on a random half of these data obtained, on average, an AUC of 0.75 ($n = 100$, range $= 0.65 - 0.83$), indicating that the viral and non-viral protein labels were appreciably separable

with respect to these RMSE values.

Ensemble prediction

Ensemble classifiers are expected to perform better than their constituent models because they can reduce overfitting of the data by the latter [49]. Although multiple predictive models of protein disorder employ an ensemble approach, none of them has been trained specifically on viral protein data. We trained a random forest classifier on the outputs of the predictors used in our study using 10 random training subsets of the viral protein data. Next, we validated the performance of this ensemble model in comparison to these individual predictors to determine if training on viral data conferred a significant advantage. We found that the ensemble classifier performed substantially better, with a mean MCC of 0.72 (range $0.62 - 0.86$). This corresponded to a roughly 27% improvement relative to ESpritz.Disorder, the best performing disorder predictor on these data (Figure 1).

To examine the relative contribution of the different predictors used as inputs for the ensemble method, we evaluated the feature importance of each input (Figure 3) — roughly the prevalence of that feature among the decision trees comprising the random forest. We observed that the individual accuracy of a predictor did not necessarily correspond to its feature importance. Specifically, the best predictors (ESpritz.Disprot, CSpritz.Long and SPOT-Disorder.2) tended to be assigned higher importance values. On the other hand, both Disprot-VL2.C and Disprot-VL2.V also displayed high importance despite having some of the worst accuracy measures when evaluated individually (Figure 1).

Example: SARS-CoV-2 accessory protein 6

To illustrate the use of our ensemble model on a novel protein, we applied this model and the 21 individual predictors to the accessory protein encoded by ORF6 in the novel 2019 coronavirus that was first isolated in Wuhan, China (designated SARS-CoV-2). ORF6 is one of the eight accessory

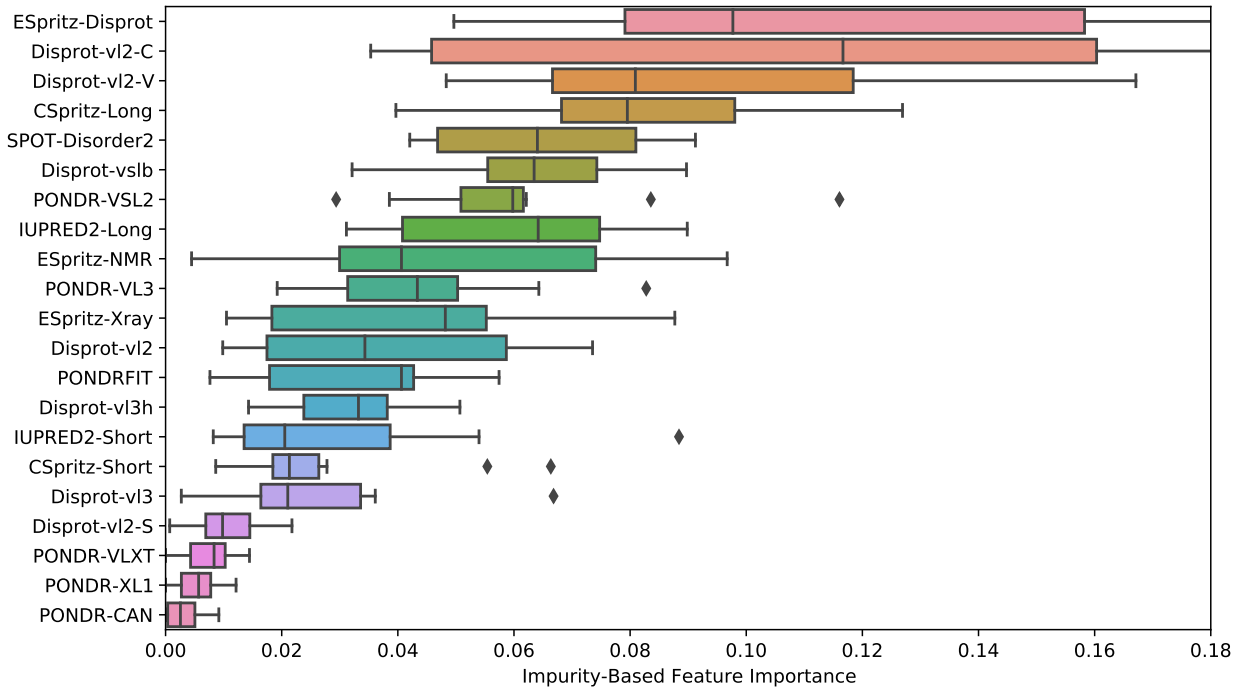

Figure 3: Box plot of the average decrease in Gini impurity by each feature in the random forest, for 10 random runs of the random forest model. Vertical line indicates the median, the box is the interquartile range (IQR; range from first to third quartiles). The left whisker extends to the first datum greater that Q1 $-$ 1.5$\times$IQR and the right whisker extends to the last datum smaller than Q3 + 1.5$\times$IQR. Individual points are outliers that lie outside this range.

genes of this virus. Its protein product is involved in antagonizing interferon activity thereby suppressing host immune response [50]. The protein is predicted to be highly disordered, particularly in its C-terminal region that contains short linear motifs involved in numerous biological activities [51]. We used a heatmap (Figure 4) to visually summarize results from the ensemble method and individual predictors, mapped to the ORF6 amino acid sequence. Overall, most predictors assigned a higher probability of disorder in the C-terminal region of the protein, with the conspicuous exception of PONDR-XL1 and PONDR-CAN, which did not predict any disordered residues in this region. We also observed considerable variation among predictors around this overall trend. Although the PONDR-XL1 predictor is documented to omit the first and last 15 residues from disorder predictions, we observed that only 14 residues were reported this way — this treatment was

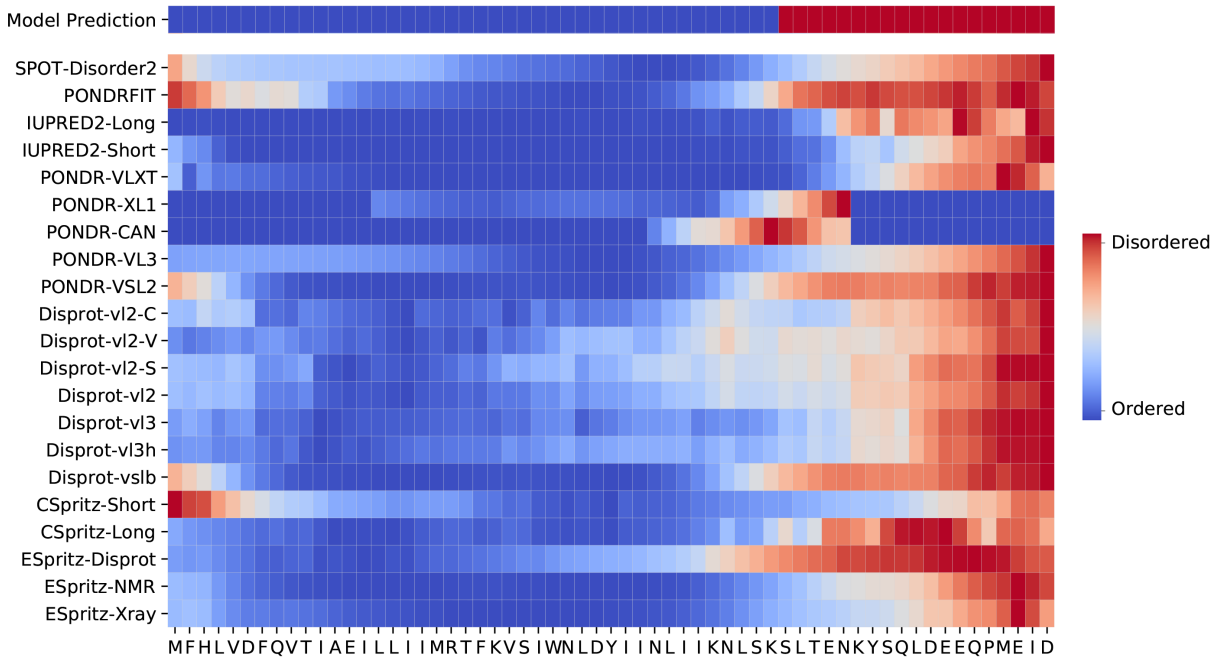

Figure 4: Disorder Predictions for novel ORF6 in SARS-CoV-2. The first row represents the random forest model predictions, with subsequent rows corresponding to individual predictors. The entire protein length is represented on the x-axis, each grid is an amino acid. Red squares indicate disordered predictions and blue squares indicate ordered predictions.

also obtained for PONDR-CAN, although it was not a documented behaviour of that predictor.

Concluding remarks

Intrinsically disordered protein regions play an essential role in many viral functions [11]. It is therefore important to predict these regions accurately in order to make biological inferences from sequence variation. In this study, we found that predictive models of intrinsic disorder were more divergent in performance when evaluated on viral proteins than non-viral proteins. We note that many of these predictors could only be accessed through web applications, and some services become unavailable at different points of our study. Although we obtained more accurate predictions — or at least, predictions that were more concordant with an expert-curated database of intrinsic protein disorder [27] — using an ensemble 'machine learning' method, the erratic availability of

the constituent predictors presents a significant obstacle to the practical utility of such approaches. Hence, we encourage researchers in the field of disorder prediction to support open science by releasing their source code or compiled binaries for local execution.

Acknowledgments

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC, RGPIN-2018-05516) and from the Canadian Institutes of Health Research (CIHR, PJT-155990).

Supplementary Tables

Table S1: Summary of viral protein data

Family	Organism	Protein	Disprot ID	Length
Alphafusellovirus	Sulfolobus spindle-shape virus 1	Protein F-112	DP00847	112
Alphatectivirus	Enterobacteria phage PRD1	Protein P16	DP01012	117
Betapolyomavirus	JC polyomavirus	Agnoprotein	DP01186	71
	Simian virus 40	Large T antigen	DP01618	708
		Major capsid protein VP1	DP00182	362
Chordopoxvirinae	Molluscum contagiosum virus subtype 1	Viral CASP8 and FADD-like apoptosis regulator	DP02042	241
	Myxoma virus	M156R	DP00849	102
		Probable host range protein 2-3	DP01983	203
	Vaccinia virus	Protein F1	DP01539	222
		Protein K7	DP02194	149
Deltavirus	Hepatitis delta virus genotype I	Small delta antigen	DP00965	195
Firstpapillomavirinae	Human papillomavirus type 16	Protein E6	DP01615	158
		Protein E7	DP00024	98
		Regulatory protein E2	DP01428	365
	Human papillomavirus type 45	Protein E7	DP01780	106
		Protein E7	DP00947	106
	Human papillomavirus type 51	Protein E6	DP02256	151
Flaviviridae	Bovine viral diarrhea virus	Genome polyprotein	DP00675	3988
	Dengue virus type 1	Genome polyprotein	DP01929	3392
	Dengue virus type 2	Genome polyprotein	DP01930	3391
			DP01245	3388
			DP00876	3391
	Dengue virus type 3	Genome polyprotein	DP02204	3390
	Dengue virus type 4	Genome polyprotein	DP01931	3387
	Hepatitis C virus genotype 1a	Genome polyprotein	DP00588	3011
	Hepatitis C virus genotype 1b	Genome polyprotein	DP01142	3010

Table S1: Summary of viral protein data

Family	Organism	Protein	Disprot ID	Length
Flaviviridae	Hepatitis C virus genotype 1b	Genome polyprotein	DP00615	3010
	Hepatitis C virus genotype 2a	Genome polyprotein	DP01031	3033
	Hepatitis GB virus B	Genome polyprotein	DP00674	2864
	Kunjin virus	Genome polyprotein	DP02051	3433
	Murray valley encephalitis virus	Genome polyprotein	DP02212	3434
	West Nile virus	Genome polyprotein	DP02203	3433
			DP00673	3430
	Zika virus	Genome polyprotein	DP01256	3419
Herpesviridae	Epstein-Barr virus	Latent membrane protein 2A	DP01060	118
	Human herpesvirus 1	Envelope glycoprotein B	DP02128	904
		Major viral transcription factor ICP4	DP01305	1298
		TAP transporter inhibitor ICP47	DP02208	88
		Tegument protein VP16	DP02291	490
			DP01642	490
		Thymidine kinase	DP00419	376
	Human herpesvirus 2	Tegument protein VP16	DP00087	490
	Human herpesvirus 8	Kaposi's sarcoma-associated herpes-like virus ORF73 homolog	DP02334	1162
	Human herpesvirus 8	LANA	DP01621	1117
	Human herpesvirus 8 type P	Viral macrophage inflammatory protein 2	DP00685	94
Inovirus	Enterobacteria phage fd	Attachment protein G3P	DP00034	424
Mastadenovirus	Human adenovirus A serotype 12	Early E1A protein	DP01151	266
	Human adenovirus C serotype 2	Early E1A protein	DP01928	289
	Human adenovirus C serotype 5	DNA-binding protein	DP00003	529
		Early E1A protein	DP01150	289
		Pre-protein VI	DP00808	250
Mimivirus	Acanthamoeba polyphaga mimivirus	Probable uracil-DNA glycosylase	DP01481	370

Table S1: Summary of viral protein data

Family	Organism	Protein	Disprot ID	Length
Mimivirus	Acanthamoeba polyphaga mimivirus	Tyrosine–tRNA ligase	DP00726	346
Myoviridae	Enterobacteria phage T4	Baseplate central spike complex protein gp5	DP00284	575
		Deoxycytidylate deaminase	DP00583	193
		Fibritin	DP01616	487
		RNA polymerase-associated protein Gp33	DP00898	112
	Escherichia phage P1	Antitoxin phd	DP00288	73
		Recombination enhancement function protein	DP00932	186
Myoviridae	Escherichia phage P2	Integrase	DP00850	337
Negarnaviricota	Hendra virus	Nucleoprotein	DP00698	532
		Phosphoprotein	DP00700	707
	Human respiratory syncytial virus A	Phosphoprotein	DP00447	241
			DP00895	241
	Influenza A virus	Hemagglutinin	DP00566	566
		Matrix protein 2	DP01016	96
		Nuclear export protein	DP00871	121
	Influenza B virus	Nucleoprotein	DP01405	560
	Lassa virus	RING finger protein Z	DP00820	99
	Measles virus	Nucleoprotein	DP00160	523
			DP00640	525
		Phosphoprotein	DP00133	507
	Nipah virus	Glycoprotein G	DP00686	602
		Nucleoprotein	DP00697	532
		Phosphoprotein	DP00699	709
	Rabies virus	Phosphoprotein	DP01759	297
	Sendai virus	Nucleoprotein	DP00629	524
		Phosphoprotein	DP00939	568
	Vesicular stomatitis Indiana virus	Phosphoprotein	DP01395	265
			DP01394	265
			DP01393	265
			DP01391	265
	Zaire ebolavirus	Hexameric zinc-finger protein VP30	DP00627	288
		Polymerase cofactor VP35	DP00998	340

Table S1: Summary of viral protein data

Family	Organism	Protein	Disprot ID	Length
Nidovirales	Human SARS coronavirus	Nucleoprotein	DP00948	422
Orthohepadnavirus	Hepatitis B virus	Large envelope protein	DP01806	445
Parvovirinae	Adeno-associated virus	Capsid protein VP1	DP01984	733
Picornavirales	Enterovirus D68	VP4	DP00986	69
	Foot-and-mouth disease virus	Genome polyprotein	DP00573	2332
	Mengo encephalomyocarditis virus	Genome polyprotein	DP01129	2293
Podoviridae	Bacillus phage phi29	Capsid assembly scaffolding protein	DP02261	98
	Salmonella phage P22	Transcriptional repressor arc	DP01512	53
Potyviridae	Potato virus Y	Polyprotein	DP01039	594
Reoviridae	Reptilian orthoreovirus	Membrane fusion protein p14	DP01043	125
Retroviridae	Equine infectious anemia virus	Protein Tat	DP00764	78
	HIV-1	Protein Nef	DP00919	208
	HIV-1	Protein Tat	DP01295	72
	HIV-1	Protein Tat	DP01087	101
	HIV-1 subtype B	Envelope glycoprotein gp160	DP00976	856
		Envelope glycoprotein gp160	DP00978	843
		Gag-Pol polyprotein	DP00410	1435
		Gag polyprotein	DP00101	500
		Gag polyprotein	DP00148	512
		Protein Nef	DP01843	206
		Protein Nef	DP00048	206
		Protein Nef	DP00189	206
		Protein Rev	DP00424	116
		Protein Tat	DP00929	86
		Protein Vif	DP00875	192
	HIV-1 subtype C	Protein Tat	DP01003	101
	HIV-1 subtype D	Protein Tat	DP00842	86
	Mason-Pfizer monkey virus	Gag polyprotein	DP01625	657
	Moloney murine leukemia virus	Gag-Pol polyprotein	DP00651	1738

Table S1: Summary of viral protein data

Family	Organism	Protein	Disprot ID	Length
Siphoviridae	Bacillus phage SPP1	39 protein	DP00750	126
	Escherichia phage HK022	Excisionase	DP01013	72
	Escherichia phage lambda	Antitermination protein N	DP00005	107
		DNA-packaging protein FI	DP01336	132
		Head-tail connector protein FII	DP01762	117
		Regulatory protein cro	DP00741	66
Solemoviridae	Southern cowpea mosaic virus	Capsid protein	DP00064	279
Togaviridae	Chikungunya virus	Nonstructural polyprotein	DP01469	2474
			DP01468	2474
			DP01466	2474
		Polyprotein P1234	DP01188	2474
	Semliki forest virus	Structural polyprotein	DP00999	1253
	Sindbis virus subtype Ockelbo	Structural polyprotein	DP00066	1245
Tombusviridae	Carnation mottle virus	Capsid protein	DP02071	348
Tymovirales	Pepino mosaic virus	Coat protein	DP01059	237

Table S2: Summary of non-viral protein data

Family	Organism	Protein	Disprot ID	Length
Chordata	Homo sapiens	60S acidic ribosomal protein P2	DP00793	115
		60S ribosomal protein L4	DP01654	427
		Amyloid-beta precursor protein	DP01280	770
		Anaphase-promoting complex subunit 15	DP01454	121
		Androgen receptor	DP00492	920
		Antigen peptide transporter 2	DP02210	686
		Apoptosis-stimulating of p53 protein 2	DP01164	1128
		ATM interactor	DP01288	823
		ATP-dependent RNA helicase DDX19B	DP01560	479
		Axin-1	DP00959	862
		Beta-adducin	DP00241	726
		Brain acid soluble protein 1	DP00930	227
		Breast cancer type 2 susceptibility protein	DP01869	3418
		Calmodulin regulator protein PCP4	DP00592	62
		cAMP-dependent protein kinase inhibitor alpha	DP00934	76
		C-C motif chemokine 26	DP00696	94
		Cellular tumor antigen p53	DP00086	393
		Cyclin-T1	DP01462	726
		Cysteine protease ATG4B	DP01326	393
		Cystic fibrosis transmembrane conductance regulator	DP00012	1480
		Cytoplasmic protein NCK1	DP01114	377
		DnaJ homolog subfamily C member 24	DP00865	149
		DNA repair protein XRCC4	DP00152	336
		E3 ubiquitin-protein ligase PPP1R11	DP00219	126

Table S2: Summary of non-viral protein data

Family	Organism	Protein	Disprot ID	Length
Chordata	Homo sapiens	E3 ubiquitin-protein ligase XIAP	DP01773	497
		Epidermal growth factor receptor	DP00309	1210
		ETS domain-containing protein Elk-4	DP01329	431
		Eukaryotic initiation factor 4A-III	DP02069	411
		Eukaryotic translation initiation factor 1A, X-chromosomal	DP00903	144
		F-box only protein 4	DP01884	387
		Filamin-binding LIM protein 1	DP01310	373
		Geminin	DP00901	209
		Glycosylphosphatidylinositol-anchored high density lipoprotein-binding protein 1	DP01327	184
		Heterogeneous nuclear ribonucleoprotein F	DP01736	415
		Heterogeneous nuclear ribonucleoproteins A2/B1	DP01109	353
		Homeobox protein Nkx-3.1	DP00683	234
		Hypoxia-inducible factor 1-alpha	DP00262	826
		Immunoglobulin alpha Fc receptor	DP00311	287
		Integrin beta-2	DP01848	769
		Isoform 11 of E3 ubiquitin-protein ligase Mdm2	DP01133	497
		Isoform 2 of Protein max	DP01097	151
		Kinetochore protein NDC80 homolog	DP01576	642
		Kinetochore scaffold 1	DP01269	2342
		Mast/stem cell growth factor receptor Kit	DP02247	976
		M-phase inducer phosphatase 3	DP02126	473

Table S2: Summary of non-viral protein data

Family	Organism	Protein	Disprot ID	Length
Chordata	Homo sapiens	Natriuretic peptides B	DP00551	134
		Neurogenic locus notch homolog protein 1	DP01104	2555
		Nuclear inhibitor of protein phosphatase 1	DP00937	351
		Nuclear pore complex protein Nup133	DP02164	1156
		Nuclear pore complex protein Nup153	DP01799	1475
		Nuclear receptor coactivator 2	DP01880	1464
		Nuclear receptor coactivator 3	DP00343	1424
		Nucleophosmin	DP01474	294
		P antigen family member 5	DP01473	130
		Peroxisome proliferator-activated receptor gamma	DP00718	505
		Polyglutamine-binding protein 1	DP01308	265
		Protein jagged-1	DP00418	1218
		Protein max	DP00084	160
		Protein regulator of cytokinesis 1	DP02316	620
		Protein SMG7	DP01844	1137
		Prothymosin alpha	DP01677	111
		Proto-oncogene c-Fos	DP00078	380
		Ras-related protein Rap-2a	DP00167	183
		Replication protein A 32 kDa subunit	DP01361	270
		Serine/threonine-protein kinase PAK 4	DP01184	591
		Signal recognition particle 19 kDa protein	DP00570	144
		SOSS complex subunit C	DP01943	104
		Stonin-2	DP01368	905
		T-cell surface glycoprotein CD3 gamma chain	DP00508	182
		Thymidylate synthase	DP00073	313
		TP53-regulated inhibitor of apoptosis 1	DP01835	76

Table S2: Summary of non-viral protein data

Family	Organism	Protein	Disprot ID	Length
Chordata	Homo sapiens	Transcription elongation regulator 1	DP01893	1098
		Transcription initiation factor TFIID subunit 6	DP01262	677
		Tyrosine-protein kinase Lck	DP01580	509
		Ubiquitin carboxyl-terminal hydrolase 7	DP00941	1102
	Mus musculus	Amelogenin, X isoform	DP01477	210
		BH3-interacting domain death agonist	DP01661	195
		Dehydrodolichyl diphosphate synthase complex subunit Nus1	DP01304	297
		Dystroglycan	DP00491	893
		Fermitin family homolog 1	DP00655	677
		Mediator of RNA polymerase II transcription subunit 1	DP02151	1575
		Phorbol-12-myristate-13-acetate-induced protein 1	DP01281	103
		Protein BEX1	DP01183	128
		Protein kinase C alpha type	DP01105	672
		Transcription regulator protein BACH2	DP01009	839
		Tumor suppressor ARF	DP00335	169
	Rattus norvegicus	Calcium/calmodulin-dependent protein kinase type 1	DP01958	374
		Calpain-2 catalytic subunit	DP01996	700
		Calpastatin	DP01994	713
		Cyclic AMP-responsive element-binding protein 1	DP00080	341
		Neuroendocrine protein 7B2	DP01557	210
		Olfactory marker protein	DP00279	163
		Rab proteins geranylgeranyltransferase component A 1	DP00458	650

Table S2: Summary of non-viral protein data

Family	Organism	Protein	Disprot ID	Length
Chordata	Rattus norvegicus	Seminal vesicle secretory protein 4	DP00527	112
		Synaptosomal-associated protein 25	DP00068	206
		Vesicle-associated membrane protein 2	DP00622	116
Dikarya	Millerozyma farinosa	Salt-mediated killer protoxin 1	DP00180	222
	Saccharomyces cerevisiae	Acetyl-CoA carboxylase	DP00557	2233
		Autophagy-related protein 13	DP01732	738
		Cold sensitive U2 snRNA suppressor 1	DP01978	436
		DNA-directed RNA polymerases I, II, and III subunit RPABC2	DP00771	155
		DNA topoisomerase 2	DP00076	1428
		Dolichyl-diphosphooligosaccharide–protein glycosyltransferase subunit STT3	DP01195	718
		Eukaryotic initiation factor 4F subunit p150	DP00082	952
		H/ACA ribonucleoprotein complex subunit CBF5	DP02055	483
		Histone H2A.Z-specific chaperone CHZ1	DP01135	153
		Mitochondrial distribution and morphology protein 35	DP02325	86
		Pre-mRNA-splicing factor 18	DP02073	251
		Protein SAN1	DP01136	610
		Protein STE50	DP01515	346
		Regulatory protein ADR1	DP00077	1323
		Ribosome biogenesis protein ERB1	DP00900	807
		Ribosome biogenesis protein NSA1	DP02195	463

Table S2: Summary of non-viral protein data

Family	Organism	Protein	Disprot ID	Length
Dikarya	Saccharomyces cerevisiae	Securin	DP00256	373
		Suppressor protein STM1	DP00994	273
		U6 snRNA-associated Sm-like protein LSm7	DP01261	115
		Ubiquitin-conjugating enzyme E2 1	DP02193	215
		Ubiquitin-like modifier-activating enzyme ATG7	DP02249	630
		UV excision repair protein RAD23	DP01629	398
		Vacuolar-sorting protein SNF8	DP01604	233
	Schizosaccharomyces pombe	YTH domain-containing protein mmi1	DP01975	488
Ecdysozoa	Caenorhabditis elegans	ATP-dependent RNA helicase laf-1	DP01113	708
	Drosophila melanogaster	Chromatin accessibility complex 16kD protein, isoform A	DP00811	140
		FACT complex subunit Ssrp1	DP00720	723
		Transcription initiation factor TFIID subunit 1	DP00081	2129
Streptophyta	Arabidopsis thaliana	Auxin-responsive protein IAA7	DP01121	243
		Calvin cycle protein CP12-2, chloroplastic	DP00534	131
		Dehydrin COR47	DP00657	265

Table S3: Optimized thresholds and accuracy of intrinsic disorder prediction in viral proteins for each predictor analysed. MCC = Matthews correlation coefficient; AUC = area under the receiver operator characteristic curve.

Predictor	MCC	AUC	Specificity	Sensitivity	Threshold
ESpritz.Disprot	0.46	0.85	0.64	0.90	0.29
CSpritz.Long	0.45	0.85	0.57	0.92	0.34
SPOT.Disorder2	0.42	0.84	0.65	0.88	0.13
CSpritz.Short	0.38	0.82	0.67	0.85	0.11
PONDRFIT	0.37	0.80	0.57	0.88	0.50
PONDR.VL3	0.36	0.78	0.56	0.88	0.54
ESpritz.Xray	0.35	0.80	0.58	0.87	0.06
Disprot.vslb	0.35	0.78	0.54	0.88	0.65
PONDR.VSL2	0.35	0.78	0.54	0.88	0.65
IUPRED2.short	0.34	0.79	0.57	0.87	0.47
Disprot.vl3	0.34	0.79	0.53	0.88	0.68
Disprot.vl3h	0.34	0.78	0.55	0.87	0.63
IUPRED2.long	0.33	0.78	0.54	0.87	0.49
Disprot.vl2	0.29	0.76	0.53	0.85	0.57
ESpritz.NMR	0.27	0.74	0.52	0.84	0.36
Disprot.vl2.S	0.24	0.73	0.53	0.81	0.54
PONDR.VLXT	0.22	0.71	0.50	0.80	0.62
Disprot.vl2.C	0.20	0.70	0.57	0.73	0.52
PONDR.XL1	0.15	0.59	0.38	0.82	0.69
Disprot.vl2.V	0.14	0.62	0.29	0.87	0.52
PONDR.CAN	0.11	0.59	0.25	0.88	0.62

Table S4: Optimized thresholds and accuracy of intrinsic disorder prediction in non-viral proteins for each predictor analysed. MCC = Matthews correlation coefficient; AUC = area under the receiver operator characteristic curve.

Predictor	MCC	AUC	Specificity	Sensitivity	Threshold
ESpritz.Disprot	0.36	0.79	0.65	0.78	0.38
ESpritz.Xray	0.35	0.77	0.73	0.70	0.06
SPOT.Disorder2	0.34	0.77	0.70	0.73	0.24
IUPRED2.short	0.33	0.76	0.72	0.69	0.42
IUPRED2.long	0.32	0.76	0.72	0.69	0.47
PONDRFIT	0.32	0.75	0.74	0.67	0.46
PONDR.VSL2	0.30	0.74	0.76	0.63	0.62
PONDR.VL3	0.28	0.74	0.78	0.59	0.51
ESpritz.NMR	0.28	0.73	0.65	0.70	0.39
CSpritz.Long	0.26	0.71	0.83	0.51	0.21
PONDR.VLXT	0.23	0.69	0.70	0.60	0.41
PONDR.XL1	0.18	0.64	0.56	0.67	0.62
PONDR.CAN	0.13	0.60	0.51	0.65	0.34

References

[1] Necci M, Piovesan D, Dosztányi Z, Tompa P, Tosatto SC. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics. 2018;34(3):445–452.

[2] Uversky VN. Intrinsically disordered proteins and their 'mysterious' (meta) physics. Frontiers in Physics. 2019;7:10.

[3] Lichtenthaler FW. 100 Years "Schlüssel-Schloss-Prinzip": What made Emil Fischer use this analogy? Angewandte Chemie International Edition in English. 1995;33(23-24):2364–2374.

[4] Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. Journal of molecular biology. 1999;293(2):321–331.

[5] Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC bioinformatics. 2012;13(1):111.

[6] Katuwawala A, Oldfield C, Kurgan L. Accuracy of protein-level disorder predictions. Brief Bioinform. 2019;46:48.

[7] Uversky VN, Oldfield CJ, Dunker AK. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu Rev Biophys. 2008;37:215–246.

[8] Hu G, Wu Z, Wang K, N Uversky V, Kurgan L. Untapped potential of disordered proteins in current druggable human proteome. Current drug targets. 2016;17(10):1198–1205.

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

[10] Xue B, Blocquel D, Habchi J, Uversky AV, Kurgan L, Uversky VN, et al. Structural disorder in viral proteins. Chemical reviews. 2014;114(13):6880–6911.

[11] Mishra PM, Verma NC, Rao C, Uversky VN, Nandi CK. Intrinsically disordered proteins of viruses: Involvement in the mechanism of cell regulation and pathogenesis. Progress in Molecular Biology and Translational Science. 2020;.

[12] Walter J, Charon J, Hu Y, Lachat J, Leger T, Lafforgue G, 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.

[13] Sanjuán R, Nebot MR, Chirico N, Mansky LM, Belshaw R. Viral mutation rates. Journal of virology. 2010;84(19):9733–9748.

[14] Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfik DS. Do viral proteins possess unique biophysical features? Trends in biochemical sciences. 2009;34(2):53–59.

[15] Cotmore SF, Tattersall P, Kerr J, Bloom M, Parrish R, Linden C. Structure and organization of the viral genome. Parvoviruses Hodder Arnold, London, United Kingdom. 2005;p. 73–94.

[16] Holmes EC. The evolutionary genetics of emerging viruses. Annu Rev Ecol Evol Syst. 2009;40:353–372.

[17] Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. Genome research. 2007;17(10):1496–1504.

[18] Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. Journal of virology. 2009;83(20):10719–10736.

[19] Ferreon ACM, Moran CR, Gambin Y, Deniz AA. Single-molecule fluorescence studies of intrinsically disordered proteins. In: Methods in enzymology. vol. 472. Elsevier; 2010. p. 179–204.

[20] DeForte S, Uversky VN. Resolving the ambiguity: Making sense of intrinsic disorder when PDB structures disagree. Protein Science. 2016;25(3):676–688.

[21] Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. Journal of molecular graphics and modelling. 2001;19(1):26–59.

[22] Uversky VN. What does it mean to be natively unfolded? European Journal of Biochemistry. 2002;269(1):2–12.

[23] Liu Y, Wang X, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Briefings in bioinformatics. 2019;20(1):330–346.

[24] Atkins JD, Boateng SY, Sorensen T, McGuffin LJ. Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. International journal of molecular sciences. 2015;16(8):19040–19054.

[25] Li J, Feng Y, Wang X, Li J, Liu W, Rong L, et al. An overview of predictors for intrinsically disordered proteins over 2010–2014. International journal of molecular sciences. 2015;16(10):23446–23462.

[26] Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, et al. DisProt 7.0: a major update of the database of disordered proteins. Nucleic acids research. 2017;45(D1):D219–D227.

[27] Hatos A, Hajdu-Soltész B, Monzon AM, Palopoli N, Álvarez L, Aykac-Fas B, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic acids research. 2020;48(D1):D269–D276.

[28] DisProt;. Available from: https://www.disprot.org/.

[29] Nielsen JT, Mulder FA. Quality and bias of protein disorder predictors. Scientific reports. 2019;9(1):5137.

[30] Hanson J, Paliwal KK, Litfin T, Zhou Y. SPOT-Disorder2: Improved Protein Intrinsic Disorder Prediction by Ensembled Deep Learning. Genomics, proteomics & bioinformatics. 2019;17(6):645–656.

[31] Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics. 2010;1804(4):996–1010.

[32] Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic acids research. 2018;46(W1):W329–W337.

[33] Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC bioinformatics. 2006;7(1):208.

[34] Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. Proteins: Structure, Function, and Bioinformatics. 2003;52(4):573–584.

[35] Walsh I, Martin AJ, Di Domenico T, Vullo A, Pollastri G, Tosatto SC. CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs. Nucleic acids research. 2011;39(suppl_2):W190–W196.

[36] Walsh I, Martin AJ, Di Domenico T, Tosatto SC. ESpritz: accurate and fast prediction of protein disorder. Bioinformatics. 2012;28(4):503–509.

[37] ChromeDriver: WebDriver for Chrome;. Available from: https://chromedriver.chromium.org/.

[38] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825–2830.

[39] Breiman L. Random forests. Machine learning. 2001;45(1):5–32.

[40] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.; 2001.

[41] Japkowicz N. The class imbalance problem: Significance and strategies. In: Proc. of the IntÕl Conf. on Artificial Intelligence. vol. 56. Citeseer; 2000. .

[42] Hemmerich J, Asilar E, Ecker GF. COVER: conformational oversampling as data augmentation for molecules. Journal of Cheminformatics. 2020;12(1):1–12.

[43] Lemaître G, Nogueira F, Aridas CK. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. The Journal of Machine Learning Research. 2017;18(1):559–563.

[44] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority oversampling technique. Journal of artificial intelligence research. 2002;16:321–357.

[45] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011;2(3):1–27.

[46] Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PloS one. 2017;12(6):e0177678.

[47] Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics. 2017;33(5):685–692.

[48] Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins: Structure, Function, and Bioinformatics. 2001;42(1):38–48.

[49] Attia A. Ensemble Prediction of Intrinsically Disordered Regions in Proteins. BMC Bioinformatics. 2012;13(1):111.

[50] Yuen CK, Lam JY, Wong WM, Mak LF, Wang X, Chu H, et al. SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists. Emerging Microbes & Infections. 2020;p. 1–29.

[51] Giri R, Bhardwaj T, Shegane M, Gehi BR, Kumar P, Gadhave K. Dark Proteome of Newly Emerged SARS-CoV-2 in Comparison with Human and Bat Coronaviruses. bioRxiv. 2020;.